Proceedings of Machine Learning Research 209:100–116, 2023 Conference on Health, Inference, and Learning (CHIL) 2023

# Revisiting Machine-Learning based Drug Repurposing: Drug Indications Are Not a Right Prediction Target

**Siun Kim** shiuhn95@snu.ac.kr
*Seoul National University, South Korea*

**Jung-Hyun Won** jhwon@snu.ac.kr
*Seoul National University, South Korea*

**David Seung U Lee** dlee0880@snu.ac.kr
*Seoul National University, South Korea*

**Renqian Luo** renqianluo@microsoft.com
*Microsoft Research*

**Lijun Wu** lijun.wu@microsoft.com
*Microsoft Research*

**Yingce Xia** yingce.xia@microsoft.com
*Microsoft Research*

**Tao Qin** taoqin@microsoft.com
*Microsoft Research*

**Howard Lee** howardlee@snu.ac.kr
*Seoul National University, South Korea*

## Abstract

In this paper, we challenge the utility of approved drug indications as a prediction target for machine learning in drug repurposing (DR) studies. Our research highlights two major limitations of this approach: 1) the presence of strong confounding between drug indications and drug characteristics data, which results in shortcut learning, and 2) inappropriate normalization of indications in existing drug-disease association (DDA) datasets, which leads to an overestimation of model performance. We show that the collection patterns of drug characteristics data were similar within drugs of the same category and the Anatomical Therapeutic Chemical (ATC) classification of drugs could be predicted by using the data collection patterns. Furthermore, we confirm that the performance of existing DR models is significantly degraded in the realistic evaluation setting we proposed in this study. We provide realistic data split information for two benchmark datasets, Fdataset and deepDR dataset.

**Data and Code Availability** In this study, we used three publicly available drug-target affinity databases BindingDB, ChEMBL, and PDSP, and two benchmark DDA datasets, Fdataset, and the deepDR dataset. In addition, a new data partitioning proposed in this study and codes for this study are available at github.com/revisit-ML-based-DR

**Institutional Review Board (IRB)** This study does not require IRB approval.

## 1. Introduction

Drug repurposing (DR), or drug repositioning, refers to finding new indications (i.e., the use of a drug for treating a particular disease) and targets (e.g., protein receptors related to drugs' therapeutic effects) for approved drugs or drugs under development. DR has strategic advantages over de novo drug development (Ashburn and Thor, 2004; Nosengo et al., 2016). Due to evidence of safety profiles and pharmacokinetics (PK), DR offers lower development risk and bypasses time-consuming processes such as chemical optimization or formulation development (Pushpakom et al., 2019; Breckenridge and Jacob, 2019). A recent example of successful DR includes the repurposing of remdesivir, an antiviral originally developed against the Ebola virus, to treat COVID-19 (Beigel et al., 2020).

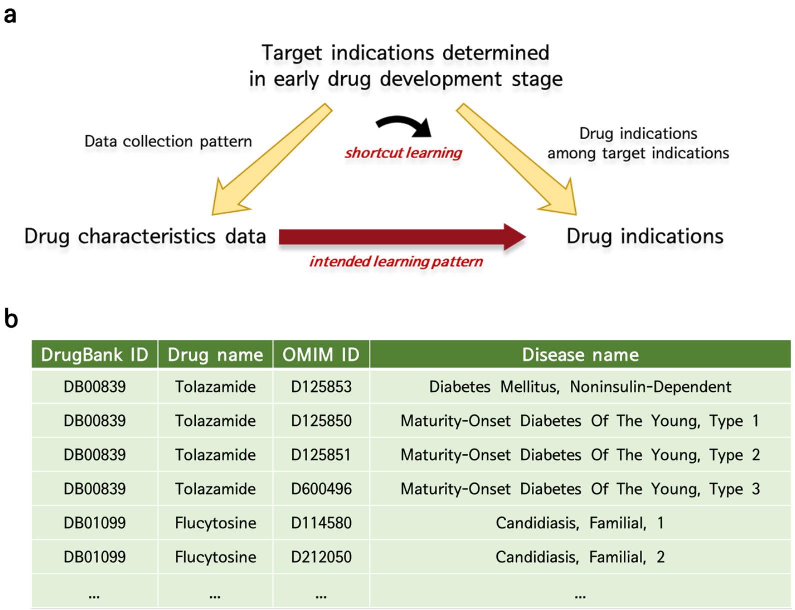
Although DR traditionally relied on serendipitous discovery of off-target effect, systematic machine learning (ML)-based approaches have been widely adopted by DR studies (Jarada et al., 2020). ML-based DR studies take one of the two task formulations: 1) predicting drug-target interaction (DTI) or 2) predicting drug-disease association (DDA). DR studies based on DTI assumes that drugs interact with the same protein targets may exhibit the same therapeutic effect. On the other hand, DR studies based on DDA assumes that drugs of similar characteristics (e.g., physicochemical properties) may be repurposed to the indications of their counterparts. Discovering potential DR signals by predicting unknown DTI has benefited from the advance of protein structure prediction models (Jumper et al., 2021). However, it was often pointed out that drug-target binding affinity may not always translate into therapeutic efficacy (Yu et al., 2021).

Thus, a growing number of DR studies formulate DR task as predicting DDA, in which data about drug characteristics (e.g., molecular structure, physicochemical properties, or drug-target affinity) are used to predict “approved drug indication" (*drug indications*, hereafter). Drug indications, which can be found in its product documents like drug labels and approval documents, specify the medical condition or disease for which the drug is intended or authorized to be used in treatment. Recent DR studies focusing on predicting DDA have demonstrated excellent performance with AUROCs frequently surpassing 0.9 in their validation set (Luo et al., 2016; Liu et al., 2016; Zeng et al., 2019; Cai et al., 2021).

However, we suspect that the performance of existing DDA prediction models for DR is overestimated. In this study, we challenge the conventional approach of utilizing drug indication as a prediction target for DR.

First, we argue that there is a confounder between drug indications and drug characteristics data (Fig. 1a). To support this claim, we show that the collection patterns of drug characteristics data are confounded with the Anatomical Therapeutic Chemical (ATC) classification of drugs and that predicting ATC class could be done by using only the data collection patterns (i.e., the specific types of data collected by drug developers to support the approval of the drug under development).

Second, we point out inappropriate normalization in DDA datasets (Fig. 1b). We find that the normalization of drug indications is inappropriate and lacks predetermined criteria for controlling the granularity of disease concepts in DDA datasets. When using a random split for model evaluation, this issue can lead to the overestimated performance of DR models. To address this issue, we propose a novel data partitioning


a
Target indications determined
in early drug development stage
Data collection pattern
shortcut learning
Drug indications
among target indications
Drug characteristics data
Drug indications
intended learning pattern
b

| DrugBank ID | Drug name | OMIM ID | Disease name |
|---|---|---|---|
| DB00839 | Tolazamide | D125853 | Diabetes Mellitus, Noninsulin-Dependent |
| DB00839 | Tolazamide | D125850 | Maturity-Onset Diabetes Of The Young, Type 1 |
| DB00839 | Tolazamide | D125851 | Maturity-Onset Diabetes Of The Young, Type 2 |
| DB00839 | Tolazamide | D600496 | Maturity-Onset Diabetes Of The Young, Type 3 |
| DB01099 | Flucytosine | D114580 | Candidiasis, Familial, 1 |
| DB01099 | Flucytosine | D212050 | Candidiasis, Familial, 2 |
| ... | ... | ... | ... |

Figure 1: Two major limitations of conventional DR task formulation as DDA prediction: a) target indication determined in the early stage of drug development as a confounder between drug indications and drug characteristics data, b) inappropriate normalization of indications in existing DDA datasets

method that enables a more realistic evaluation of DR model.

## 2. Target Indication as a Confounder

DDA prediction models assume that drug indications are predictable from drug characteristics data. However, we thought that this assumption ignores the complexity in the drug development process (Kaitin, 2010; Kinch et al., 2014). In reality, various industrial contexts such as market demands, existing treatments, and regulatory environment, in addition to the biological properties of the drug, contribute to determining drug indication.

Moreover, because it is unrealistic and infeasible from the business standpoint to investigate potential indications in all therapeutic areas, drug developers decide on which indications to pursue (i.e., target indications) from the early stage of drug development (Fig. 1a). Thus, drug developers tend to selectively collect drug characteristics data needed to support the druggability within the pre-meditated target indications (Hughes et al., 2011; Strovel et al., 2016). While almost all drug indications are among target indications, target indications may become a strong confounder between drug indications and drug characteristics data (and their collection pattern).

## 3. Disease Normalization for Drug Indication

DDA datasets contain drug indication information extracted from drug labels. The drug indications are normalized to disease concepts in medical ontologies such as Online Mendelian Inheritance in Men (OMIM) and Unified Medical Language System (UMLS) to use standard disease concepts from diverse disease characteristics databases. Thus, normalization of drug indication (*indication normalization*, hereafter) involves deciding how different disease concepts will be linked or combined to express corresponding drug indications (Luo et al., 2019).

Because standard medical ontologies used in DDA datasets are hierarchical, the options for disease normalization are either 1) opting for higher (broad) concepts or 2) using lower (specific) concepts. For example, noninsulin-dependent diabetes mellitus can be normalized into a higher concept “type 2 diabetes mellitus" or by using lower concepts “maturity-onset diabetes of the young" and its subtypes (Fig. 1b). Each option has a downside. The former can lead to information loss, while the latter can induce an overrepresentation issue, where a single drug indication is normalized by several, closely related disease concepts. We suspect that when developing DDA prediction models for DR, using lower concepts for disease normalization should be refrained. It is especially so if the model is developed by using random split, resulting in overestimated model performance.

## 4. Correlation Investigation and ATC Prediction

In this section, we first present t-SNE visualization and odds ratio measurements that we utilized to investigate the correlation between the drug indications and the collection patterns of drug characteristics data. Next, we present ATC prediction results using the data collection patterns as input to demonstrate that there is a shortcut in predicting drug indications.

### 4.1. Data

We utilized drug-target affinity data from BindingDB(Liu et al., 2007), PDSP Ki (PDSP), ChEMBL v31 (Mendez et al., 2019) databases to perform t-SNE visualization in order to explore a correlation between drug indications and the collection patterns of drug characteristics data. Additionally, we developed an ATC prediction model based on the drug-target affinity data to show that the data collection patterns could work as a shortcut in predicting drug indications.

To standardize the data, we normalized all drugs using the DrugBank identifier, and the ATC classification of each drug was extracted from the corresponding ATC codes in DrugBank[1]. Furthermore, we excluded the binding affinity values for proteins related to drug PK (i.e., absorption, distribution, metabolism and elimination of drugs in the body) because they would be present for all drugs regardless of indication.

### 4.2. Correlation Investigation

We performed a 2D t-SNE visualization to investigate the correlation between drug indications and the collection patterns of drug characteristic data. A binary vector was created for each drug to indicate the presence or the absence of affinity data for each target protein (see Appendix A.1). Thus, the dimension of the binary vector is the number of target proteins in each database. We adjusted the perplexity as the sole hyperparameter we adjusted in t-SNE, with values of 10, 30 and 50. The results were compared to t-SNE plots of randomly generated data which is a set of binary drug vectors randomly constructed to have the same overall collection frequency (i.e., number of '1' elements) in the binding affinity databases. The drugs were colored according to their ATC class.

[1] https://go.drugbank.com/

In addition, to assess the correlation between drug ATC classes and data collection patterns, we computed odds ratios for each ATC class-target protein combination. We used same binary vectors used for the t-SNE visualization. We computed the odds ratios by considering whether a drug belong to a particular ATC class or not, and whether affinity data for a specific target protein were available for that drug. Thus, a higher odds ratio greater indicates a stronger association between a particular ATC class and the availability of data for a specific target protein.

The t-SNE plots showed that drugs within the same ATC class are more likely to cluster (Fig. 2). This finding suggests that data collection patterns are more likely to be similar for drugs in the same category.

Moreover, the odds ratios were significantly higher in the actual binding affinity data compared to randomly generated data (Fig. 3). Notably, the majority of the ATC classes and target proteins with high odds ratios are clinically associated (Table. B.1). These results provide evidence in support of the proposed link between drug indications and data collection patterns through target indication, as presented in section 2.

### 4.3. ATC Prediction

We developed an ATC prediction model to assess whether there is a shortcut in predicting drug indications using drug characteristics data. The input vector of each drug was binary, indicating the presence or absence of affinity data for target proteins like in t-SNE visualization.

The model was trained on 869 drugs with known ATC class from the affinity databases. XGBoost, with the learning rate, max depth, and number of estimators as hyperparameters (see Appendix A.2), was used (Chen and Guestrin, 2016). Moreover, since the total number of target proteins with at least one binding affinity data was about 2500, which is much larger than the total number of drugs, it was essential to reduce the dimensionality of the input data. Therefore, we considered the number of target proteins used in ATC prediction as a hyperparameter, and obtained the highest model performance when using the top 10% of most frequently collected target proteins. Model validation was done with the leave-one-out (i.e., jackknife) method as widely used in previous studies (Chen et al., 2012; Cheng et al., 2017; Wang et al., 2019).

The model achieved an accuracy of 47.3%, lower than the existing benchmark models (Wang et al., 2019; Cheng et al., 2017; Chen et al., 2012) but substantially higher than a random guess of 7% for 14 ATC classes (Table 1). It is noteworthy that our model only used the binary vector as input features expressing the presence or absence of affinity data for target proteins. In contrast, NLSP-XGB-LPA (Lumini and Nanni, 2018) and iATC-mISF (Cheng et al., 2017) used similarity scores based on drug interaction, molecule structure, and molecular fingerprint as drug characteristics data, while Chen et al., 2012 chemical-chemical interaction and structural similarity.

Table 1: Leave-one-out accuracy for ATC class prediction

| **Models** | **Accuracy** |
|---|---|
| NLSP-XGB-LPA | 0.7808 |
| iATC-mISF | 0.6641 |
| Chen et al., 2012 | 0.4938 |
| Our model | 0.4730 |
| Random guess | 0.0714 |

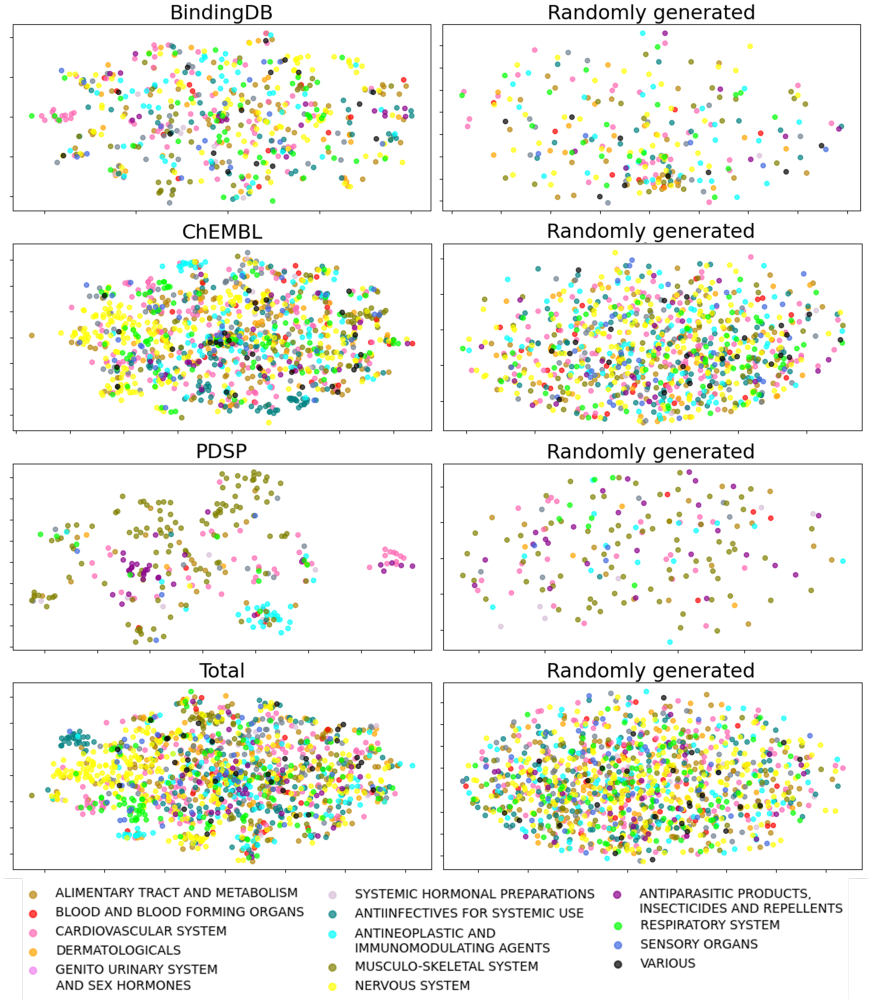


Figure 2: t-SNE visualization of drug characteristics databases based on the data collection pattern

## 5. Re-visiting Existing DR Datasets and Models

In this section, we showed the performance degradation of existing DDA prediction models in a realistic evaluation setting.

### 5.1. Data

We used two benchmark datasets for DDA prediction: Fdataset (Gottlieb et al., 2011) and DDA dataset collected from deepDR study (Zeng et al., 2019). The Fdataset (i.e., PREDICT), which extracts drug indication information from DailyMed and DrugBank, contains a total of 1933 DDAs between 593 drugs and 313 diseases. The deepDR dataset contains 6677 clinically reported DDAs between 1519 drugs and 1229 diseases. The indications in Fdataset and the deepDR datasets are normalized to disease concepts in OMIM and UMLS, respectively. Additionally, all drugs in both datasets are normalized using drug identifiers of DrugBank.

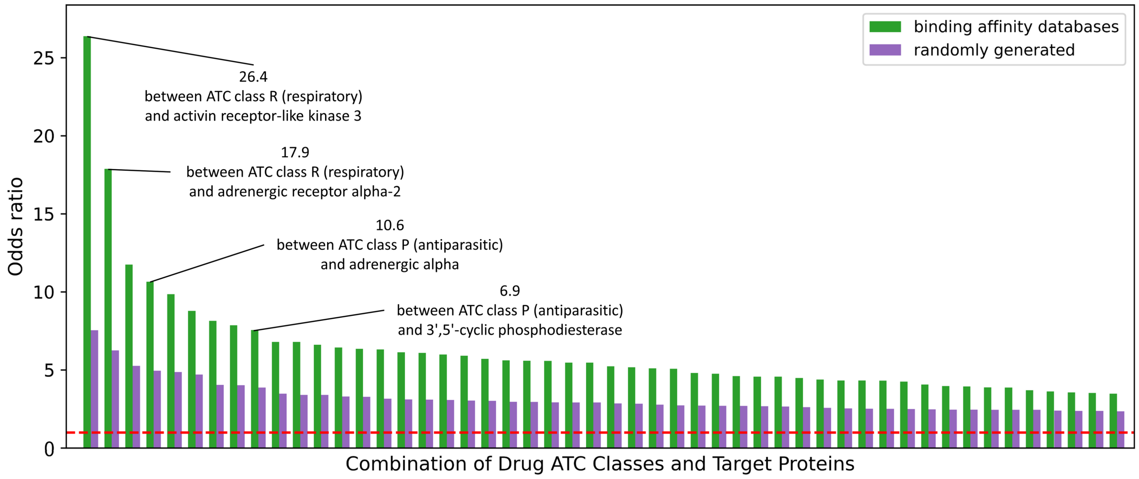


Figure 3: Distribution of odds ratios for ATC class-target protein combinations in drug characteristics databases (Top 50)

## 5.2. Evaluating Normalization Appropriateness

In this section, we evaluated whether drug indications in DDA datasets were appropriately normalized. To this end, we calculated similarity scores between drug indications for each drug in the datasets. We hypothesized that the presence of drug indication pairs with high similarity scores (i.e., $\geq$ 0.6 for ICD10 and $\geq$ 0.8 for MeSH) indicates the datasets are not appropriately normalized. The threshold was heuristically determined (see Appendix B.2). To calculate similarity scores, we mapped drug indications to disease concepts in the International Classification of Diseases 10th revision (ICD10) and Medical Subject Headings (MeSH) using automated mapping provided by UMLS metathesaurus and OMIM. Then, the remaining indications were manually mapped.

Similarity scores were calculated using the hierarchical structure of ICD10 and MeSH and the set-level similarity score based on the concept of information content (Sánchez et al., 2011; Jia et al., 2019, see Appendix A.3).

Results showed that 17.50% and 9.95% of drug indication pairs within a single drug had a closer match above the threshold in the Fdataset and deepDR dataset, respectively when measuring similarity scores based on ICD10 (Fig. 4). In other words, the DDA benchmark dataset currently includes several highly similar diseases as separate drug indications for a single drug, which can lead to overestimation when evaluating performance. Likewise, based on MeSH, similarity scores of 22.50% and 15.90% of drug indication pairs were above the threshold in Fdataset and the deepDR dataset, respectively. These results imply that close disease concepts were frequently used to express drug indications for a single drug in Fdataset and the deepDR datasets.

## 5.3. Re-evaluating Existing DR Models

In this section, we revisited three representative DDA prediction models in a more realistic evaluation setting. We used deepDR (Zeng et al., 2019), HNET-DNN (Liu et al.,

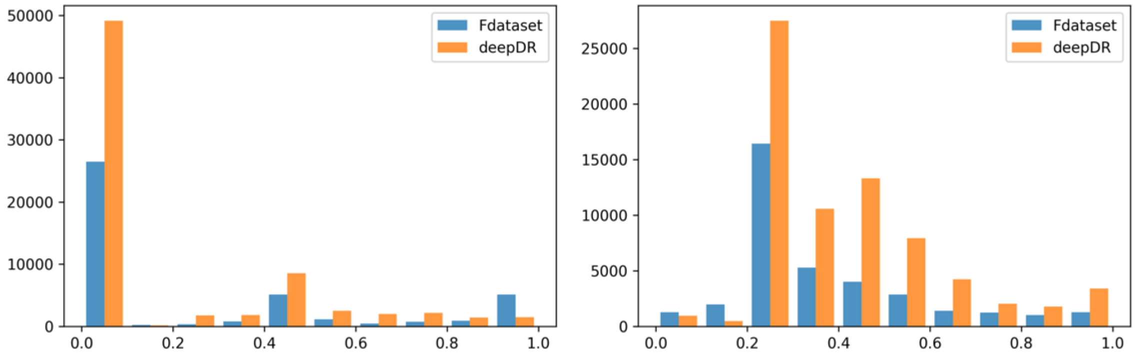


Figure 4: Distribution of similarity scores between drug indication pairs for each drug (Left: ICD10, right: MeSH)

2020), and TP-NRWRH (Liu et al., 2016)) models.

We categorized the drug indications into “original" and “expanded." We separated the original and expanded indications by performing agglomerative clustering based on the similarity scores (see Appendix A.3). Drug indications in the larger cluster were categorized as original and the other as expanded. However, when a similarity score between larger and smaller clusters was above a pre-determined threshold (0.1 for ICD10, 0.3 for MeSH), drug indications in the smaller cluster were classified as original. We conducted an iterative search for thresholds that divide original and expanded indications across the distinct therapeutic area.

We set the number of clusters, disease ontology used to measure similarity, and the linkage criterion as clustering hyperparameters (see Appendix A.3). Then, we evaluated the performance of existing DR models for every combination of hyperparameters in three different scenarios: scenario 1) train and test through random split; scenario 2) train and test on original indications; and scenario 3) train on original and test on expanded indications. The number of samples in the train and test set was kept constant.

The results revealed consistent performance degradation of the DR model in scenario 3 (Tables 2 and B.6) when comparing the performance in scenarios 1 and 2. Conversely, scenario 2, showed a slight improvement compared to scenario 1. We believe this is because more similar drug indications were contained in train and test datasets.

We propose scenario 3 as the most appropriate and realistic data partitioning method for evaluating DR models because finding drug indications outside of an initial therapeutic area is a genuine goal of generating DR signals through ML models. In addition, we expect that the problem of containing highly similar drug indications in both train and test datasets would be mitigated in scenario 3.

## 6. Related Work

### 6.1. DDA Prediction Models

Recent studies have focused to deal with two major huddles to improve the performance DDA prediction models: 1) reducing the dimensionality of drug features through unsupervised learning methods and 2) incorporating drug and disease characteristics data of different modalities from diverse databases (Luo et al., 2021; Liang et al., 2017). The di-

Table 2: AUC performances of existing DR models in different evaluation scenarios

| | **Scenario 1** | **Scenario 2** | **Scenario 3** |
|---|---|---|---|
| **deepDR** | | | |
| **ICD10** | 0.907 ± 0.018 | 0.915 ± 0.016 (▲0.007) | **0.840 ± 0.012 (▼0.068)** |
| **MeSH** | 0.926 ± 0.016 | 0.929 ± 0.014 (▲0.003) | **0.756 ± 0.035 (▼0.170)** |
| **HNET-DNN** | | | |
| **ICD10** | 0.926 ± 0.025 | 0.927 ± 0.025 (▲0.002) | **0.838 ± 0.024 (▼0.088)** |
| **MeSH** | 0.927 ± 0.017 | 0.934 ± 0.027 (▲0.007) | **0.866 ± 0.024 (▼0.060)** |
| **TP-NRWRH** | | | |
| **ICD10** | 0.942 ± 0.022 | 0.942 ± 0.012 (▲0.000) | **0.837 ± 0.008 (▼0.105)** |
| **MeSH** | 0.927 ± 0.019 | 0.945 ± 0.026 (▲0.018) | **0.750 ± 0.010 (▼0.177)** |

mension reduction is important in DDA prediction because sample size of drug indication data is limited compared to the high-dimensional drug characteristics data.

Indeed, autoencoder (Zeng et al., 2019; Jiang et al., 2020) and similarity networks (Liu et al., 2020; Jarada et al., 2021) have been widely used to obtain the dense features for drug and disease while utilizing characteristics data of multiple modalities. In addition, matrix factorization (Luo et al., 2018; Jarada et al., 2021) and matrix completion methods (Zhang et al., 2020) have been used to combine drug-drug and disease-disease networks or integrate other types of interaction data, such as drug-target and disease-gene interactions.

### 6.2. Shortcut Learning

Shortcut learning refers to a phenomenon in which a ML model learns a spurious correlation that does not generalize to real-world scenarios, resulting in poor performance outside the benchmark dataset (Geirhos et al., 2020). This has been observed in various tasks across vision and natural language processing, including image classification (Beery et al., 2018; Rosenfeld et al., 2018), medical imaging (Zech et al., 2018b; DeGrave et al., 2021), question answering (Jia and Liang, 2017) and argument reasoning (Niven and Kao, 2019). Unlike overfitting, shortcut learning exhibits relatively high performance on test data drawn from the same distribution as the train data, but demonstrates poor performance on out-of-distribution data (Geirhos et al., 2020).

Our study was motivated by a previous study that found the use of confounding variables as shortcut features in medical settings reduced the generalizability of model performance (Zech et al., 2018a). In this study, we tested the hypothesis that there is a high risk of problems caused by shortcut learning in DR settings, where the approved indication of a drug is determined not only by its biological characteristics but also by social factors, which act as confounders.

## 7. Discussion

In this study, we found that the performance of current DDA prediction models was significantly reduced when evaluated in a more realistic evaluation setting following the data partitioning method we proposed. We offer two possible explanations for our finding.

First, there may be a spurious correlation between drug indications and the drug characteristics data, possibly through the target indications of a drug and data collection pattern. Of note, we found that the collection patterns of drug characteristics data were correlated with the drug's ATC class through t-SNE visualization and that the ATC class could be predicted based on the data collection patterns. This correlation could result in the DR models relying on shortcuts rather than drug characteristics when predicting drug indications. When DR models rely on shortcuts, the DR models' performance will decrease in an evaluation setting that is more relevant to DR, where the models often need to predict new indications in a different therapeutic area.

Second, inappropriate normalization of drug indications to disease concepts in existing DDA datasets may have led to overestimated prediction performance. We found that in the benchmark DDA datasets, similar disease concepts were used to express drug indications of a drug (section 5.2). In this situation, the commonly used random split method for model evaluation is not appropriate because the presence of highly similar disease concepts within both the train and test sets increases the risk of overfitting and reduces the practical value of DR models.

Thus, we suggest that future DR models be trained and tested in a more realistic evaluation setting, as proposed in this study. In addition, we recommend establishing guidelines for creating DDA datasets, with attention to the granularity of disease concepts used to normalize drug indications.

In addition, to circumvent the spurious correlation due to data collection pattern, we propose that future DR studies prioritize using the drug characteristics data in which missingness is low. In the case of high-dimensional drug characteristics data with substantial missingness, it is advisable to ensure that the missing pattern is not associated with the drug indications to prevent shortcut learning.

We originally aimed to define 'original indications' as 'the group of indications for which marketing approval was first granted', and 'expanded indications' as 'subsequently granted approved indications that are clinically distinct from the original indications'. This was intended to replicate the real-world DR process, where drug developers should use information on already approved drug indications to predict new indications for drug approval. However, due to challenges in collecting approval dates for drug indications and disease normalization issues, we instead classified approved indications into major and minor groups based on disease similarity scores and clustering techniques. Despite the deviation from our initial plan, we still referred to these groups as original and expanded indications. We would like to note that the most realistic approach would involve dividing the indication data based on their approval dates, which we hope to achieve in future DDA datasets.

In fact, the difficulty in normalizing drug indications to disease concepts is due to the fact that a drug indication may not correspond to a single disease. Drug indications may contain information about the patient's condition, such as non-responsiveness to other medications, contraindications, or severity of the disease (FDA, 2018). In other words, drug indications provide information about the relevant clinical context in which the drug is being used. This fact makes it complex to map drug indications to disease concepts provided by medical ontologies, leading to irregularities in indication normalization. To avoid this issue, a novel approach to formulating DR problems, such as predicting investigational conditions in clinical trials (Brown and Patel, 2017) or recommending eligibility criteria for these clinical studies, could be considered.

## Acknowledgments

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program) (RS-2022-00155958) supervised by the IITP(Institute for Information, Communications Technology Planning and Evaluation.

## References

Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 3(8):673–683, 2004.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

John H Beigel, Kay M Tomashek, Lori E Dodd, Aneesh K Mehta, Barry S Zingman, Andre C Kalil, Elizabeth Hohmann, Helen Y Chu, Annie Luetkemeyer, Susan Kline, et al. Remdesivir for the treatment of covid-19. *New England Journal of Medicine*, 383(19):1813–1826, 2020.

Alasdair Breckenridge and Robin Jacob. Overcoming the legal and regulatory barriers to drug repurposing. *Nature reviews Drug discovery*, 18(1):1–2, 2019.

Adam S Brown and Chirag J Patel. A standard database for drug repositioning. *Scientific data*, 4(1):1–7, 2017.

Lijun Cai, Changcheng Lu, Junlin Xu, Yajie Meng, Peng Wang, Xiangzheng Fu, Xiangxiang Zeng, and Yansen Su. Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Briefings in Bioinformatics*, 22(6): bbab319, 2021.

Lei Chen, Wei-Ming Zeng, Yu-Dong Cai, Kai-Yan Feng, and Kuo-Chen Chou. Predicting anatomical therapeutic chemical (atc) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS one*, 7(4):e35254, 2012.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Xiang Cheng, Shu-Guang Zhao, Xuan Xiao, and Kuo-Chen Chou. iatc-misf: a multilabel classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics*, 33(3):341–346, 2017.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

FDA. Herceptin (trastuzumab) for injection, for intravenous use, 2018. URL https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/103792s5345lbl.pdf.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular systems biology*, 7(1):496, 2011.

Britta Haenisch, Jutta Walstab, Stephan Herberhold, Friedrich Bootz, Marion Tschaikin, René Ramseger, and Heinz Bönisch. Alpha-adrenoceptor agonistic activity of oxymetazoline and xylometazoline. *Fundamental & clinical pharmacology*, 24(6):729–739, 2010.

Marius M Hoeper, David B Badesch, H Ardeschir Ghofrani, J Simon R Gibbs, Mardi Gomberg-Maitland, Vallerie V McLaughlin, Ioana R Preston, Rogerio Souza, Aaron B Waxman, Ekkehard Grünig, et al. Phase 3 trial of sotatercept for treatment of pulmonary arterial hypertension. *New England Journal of Medicine*, 2023.

Friedrich Horak, Petra Zieglmayer, René Zieglmayer, Patrick Lemell, Ruji Yao, Heribert Staudinger, and Melvyn Danzig. A placebo-controlled study of the nasal decongestant effect of phenylephrine and pseudoephedrine in the vienna challenge chamber. *Annals of Allergy, Asthma & Immunology*, 102(2):116–120, 2009.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Tamer N Jarada, Jon G Rokne, and Reda Alhajj. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of cheminformatics*, 12(1):1–23, 2020.

Tamer N Jarada, Jon G Rokne, and Reda Alhajj. Snf-nn: computational method to predict drug-disease interactions using similarity network fusion and neural networks. *BMC bioinformatics*, 22(1):1–20, 2021.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.

Zheng Jia, Xudong Lu, Huilong Duan, and Haomin Li. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. *BMC medical informatics and decision making*, 19(1):1–11, 2019.

Han-Jing Jiang, Yu-An Huang, and Zhu-Hong You. Saerof: an ensemble approach for large-scale drug-disease association prediction by incorporating rotation forest and sparse autoencoder deep neural network. *Scientific reports*, 10(1):1–11, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kenneth I Kaitin. Deconstructing the drug development process: the new face of innovation. *Clinical Pharmacology & Therapeutics*, 87(3):356–361, 2010.

Michael S Kinch, Janie Merkel, and Sheila Umlauf. Trends in pharmaceutical targeting of clinical indications: 1930–2013. *Drug discovery today*, 19(11):1682–1685, 2014.

Christian Konrad, Sherry F Queener, Ronald C Wek, and William J Sullivan Jr. Inhibitors of eif2$\alpha$ dephosphorylation slow replication and stabilize latency in toxoplasma gondii. *Antimicrobial agents and chemotherapy*, 57(4):1815–1822, 2013.

Elaine M Leslie, Roger G Deeley, and Susan PC Cole. Multidrug resistance proteins: role of p-glycoprotein, mrp1, mrp2, and bcrp (abcg2) in tissue defense. *Toxicology and applied pharmacology*, 204(3): 216–237, 2005.

Xujun Liang, Pengfei Zhang, Lu Yan, Ying Fu, Fang Peng, Lingzhi Qu, Meiying Shao, Yongheng Chen, and Zhuchu Chen. Lrssl:

predict and interpret drug–disease associations based on data integration using sparse subspace learning. *Bioinformatics*, 33(8):1187–1196, 2017.

Hui Liu, Yinglong Song, Jihong Guan, Libo Luo, and Ziheng Zhuang. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC bioinformatics*, 17(17):269–277, 2016.

Hui Liu, Wenhao Zhang, Yinglong Song, Lei Deng, and Shuigeng Zhou. Hnet-dnn: inferring new drug–disease associations with deep neural network based on heterogeneous network features. *Journal of Chemical Information and Modeling*, 60(4):2367–2376, 2020.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35 (suppl_1):D198–D201, 2007.

Alessandra Lumini and Loris Nanni. Convolutional neural networks for atc classification. *Current pharmaceutical design*, 24 (34):4007–4012, 2018.

Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Xiaoqing Peng, Fang-Xiang Wu, and Yi Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.

Huimin Luo, Min Li, Shaokai Wang, Quan Liu, Yaohang Li, and Jianxin Wang. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, 34(11): 1904–1912, 2018.

Huimin Luo, Min Li, Mengyun Yang, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang. Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in bioinformatics*, 22(2): 1604–1619, 2021.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. Mcn: a comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92: 103132, 2019.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

Nicola Nosengo et al. Can you teach old drugs new tricks? *Nature*, 534(7607):314–316, 2016.

PDSP. Pdsp ki database - psychoactive drug screening program. https://pdsp.unc.edu/databases/kidb.php. Accessed: 2023-01-27.

Raquel Porto, Ana C Mengarda, Rayssa A Cajas, Maria C Salvadori, Fernanda S Teixeira, Daniel DR Arcanjo, Abolghasem Siyadatpanah, Maria de Lourdes Pereira, Polrat Wilairatana, and Josué de Moraes. Antiparasitic properties of cardiovascular agents against human intravascular parasite schistosoma mansoni. *Pharmaceuticals*, 14(7):686, 2021.

Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

David Sánchez, Montserrat Batet, and David Isern. Ontology-based information content computation. *Knowledge-based systems*, 24 (2):297–303, 2011.

Jeffrey Strovel, Sitta Sittampalam, Nathan P Coussens, Michael Hughes, James Inglese, Andrew Kurtz, Ali Andalibi, Lavonne Patton, Chris Austin, Michael Baltezor, et al. Early drug discovery and development guidelines: for academic researchers, collaborators, and start-up companies. *Assay Guidance Manual [Internet]*, 2016.

Madeleine Monique Uys, Mohammed Shahid, and Brian Herbert Harvey. Therapeutic potential of selectively targeting the $\alpha$2c-adrenoceptor in cognition, depression, and schizophrenia—new developments and future perspective. *Frontiers in Psychiatry*, 8:144, 2017.

Fien M Verhamme, Ken R Bracke, Guy F Joos, and Guy G Brusselle. Transforming growth factor-$\beta$ superfamily in obstructive lung diseases. more suspects than tgf-$\beta$ alone. *American journal of respiratory cell and molecular biology*, 52(6):653–662, 2015.

Xiangeng Wang, Yanjing Wang, Zhenyu Xu, Yi Xiong, and Dong-Qing Wei. Atc-nlsp: prediction of the classes of anatomical therapeutic chemicals using a network-based label space partition method. *Frontiers in pharmacology*, 10:971, 2019.

Jun-Lin Yu, Qing-Qing Dai, and Guo-Bo Li. Deep learning in target prediction and drug repositioning: Recent advances and challenges. *Drug Discovery Today*, 2021.

Keizo Yuasa, Fumika Mi-Ichi, Tamaki Kobayashi, Masaya Yamanouchi, Jun Kotera, Kiyoshi Kita, and Kenji Omori. Pfpde1, a novel cgmp-specific phosphodiesterase from the human malaria parasite plasmodium falciparum. *Biochemical Journal*, 392(1):221–229, 2005.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric K Oermann. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018a.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018b.

Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.

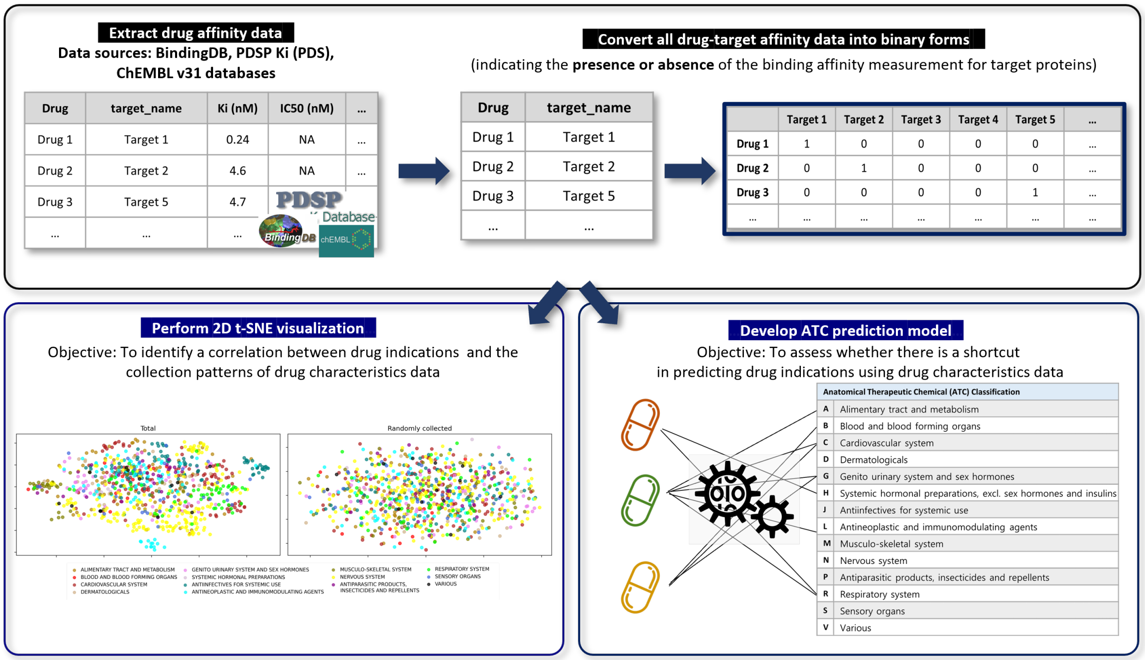
Wenjuan Zhang, Hunan Xu, Xiaozhong Li, Qiang Gao, and Lin Wang. Drimc: an improved drug repositioning approach using bayesian inductive matrix completion. *Bioinformatics*, 36(9):2839–2847, 2020.

# Appendix A. Supplementary Methods

We elaborate methodologies for 1) obtaining a binary vector expressing the presence or the absence of drug characteristics data for t-SNE visualization and ATC prediction and 2) splitting drug indications in DDA datasets based on similarity scores in this appendix section.

## A.1. Obtaining binary vectors for t-SNE visualization and ATC prediction

In this section, we present a detailed methodology for obtaining binary vectors that indicate the presence or the absence of drug-target affinity data in three databases used for t-SNE visualization and ATC prediction (Fig. A1).



**Extract drug affinity data**

Data sources: BindingDB, PDSP Ki (PDS), ChEMBL v31 databases

| Drug | target_name | Ki (nM) | IC50 (nM) | … |
|---|---|---|---|---|
| Drug 1 | Target 1 | 0.24 | NA | … |
| Drug 2 | Target 2 | 4.6 | NA | … |
| Drug 3 | Target 5 | 4.7 | | |
| … | … | … | | |

**Convert all drug-target affinity data into binary forms**

(indicating the **presence or absence** of the binding affinity measurement for target proteins)

| Drug | target_name |
|---|---|
| Drug 1 | Target 1 |
| Drug 2 | Target 2 |
| Drug 3 | Target 5 |
| … | … |

| | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 | … |
|---|---|---|---|---|---|---|
| Drug 1 | 1 | 0 | 0 | 0 | 0 | … |
| Drug 2 | 0 | 1 | 0 | 0 | 0 | … |
| Drug 3 | 0 | 0 | 0 | 0 | 1 | … |
| … | … | … | … | … | … | … |

**Perform 2D t-SNE visualization**

Objective: To identify a correlation between drug indications and the collection patterns of drug characteristics data

**Develop ATC prediction model**

Objective: To assess whether there is a shortcut in predicting drug indications using drug characteristics data

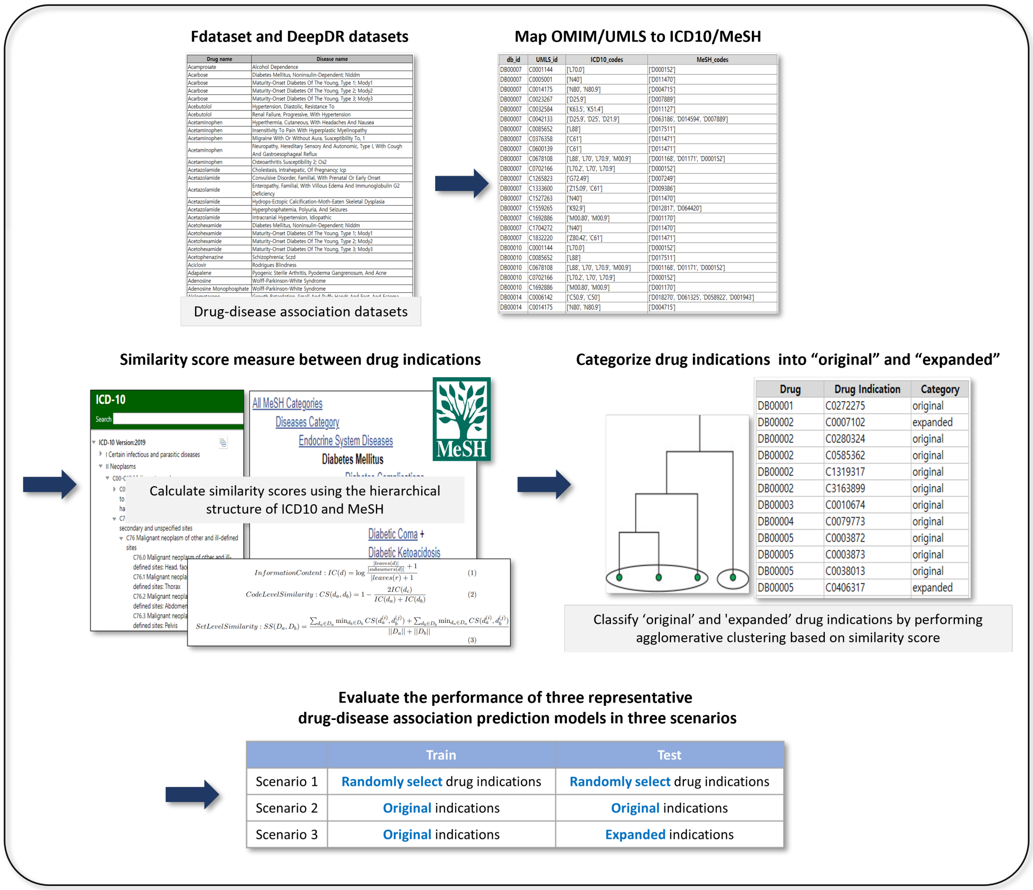
| Anatomical Therapeutic Chemical (ATC) Classification | |
|---|---|
| A | Alimentary tract and metabolism |
| B | Blood and blood forming organs |
| C | Cardiovascular system |
| D | Dermatologicals |
| G | Genito urinary system and sex hormones |
| H | Systemic hormonal preparations, excl. sex hormones and insulins |
| J | Antiinfectives for systemic use |
| L | Antineoplastic and immunomodulating agents |
| M | Musculo-skeletal system |
| N | Nervous system |
| P | Antiparasitic products, insecticides and repellents |
| R | Respiratory system |
| S | Sensory organs |
| V | Various |

PDSP
Database
BindingDB
chEMBL
Total
Randomly collected
ALIMENTARY TRACT AND METABOLISM
BLOOD AND BLOOD FORMING ORGANS
CARDIOVASCULAR SYSTEM
DERMATOLOGICALS
GENITO URINARY SYSTEM AND SEX HORMONES
SYSTEMIC HORMONAL PREPARATIONS
ANTIINFECTIVES FOR SYSTEMIC USE
ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS
MUSCULO-SKELETAL SYSTEM
NERVOUS SYSTEM
ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS
RESPIRATORY SYSTEM
SENSORY ORGANS
VARIOUS

Figure A1: Visualization of obtaining binary vectors representing the collection patterns of drug-target affinity data for t-SNE visualization and ATC prediction

## A.2. Hyperparamter settings for XGBoost model predicting drug's ATC class

We performed a grid search to find the best combination of hyperparameter settings for the XGBoost model that predict drugs' ATC class. The tuned hyperparameters included learning rate (range of [0.1, 0.05, 0.01]), max depth (range of [2, 6, 8]), and the number of estimators (range of [10, 100, 1000]).



Fdataset and DeepDR datasets

| Drug name | Disease name |
|---|---|
| Acamprosate | Alcohol Dependence |
| Acarbose | Diabetes Mellitus, Noninsulin-Dependent; Niddm |
| Acarbose | Maturity-Onset Diabetes Of The Young, Type 1; Mody1 |
| Acarbose | Maturity-Onset Diabetes Of The Young, Type 2; Mody2 |
| Acarbose | Maturity-Onset Diabetes Of The Young, Type 3; Mody3 |
| Acebutolol | Hypertension, Diastolic, Resistance To |
| Acebutolol | Renal Failure, Progressive, With Hypertension |
| Acetaminophen | Hyperthermia, Cutaneous, With Headaches And Nausea |
| Acetaminophen | Insensitivity To Pain With Hyperplastic Myelinopathy |
| Acetaminophen | Migraine With Or Without Aura, Susceptibility To, 1 |
| Acetaminophen | Neuropathy, Hereditary Sensory And Autonomic, Type I, With Cough And Gastroesophageal Reflux |
| Acetaminophen | Osteoarthritis Susceptibility 2; Os2 |
| Acetazolamide | Cholestasis, Intrahepatic, Of Pregnancy; Icp |
| Acetazolamide | Convulsive Disorder, Familial, With Prenatal Or Early Onset |
| Acetazolamide | Enteropathy, Familial, With Villous Edema And Immunoglobulin G2 Deficiency |
| Acetazolamide | Hydrops-Ectopic Calcification-Moth-Eaten Skeletal Dysplasia |
| Acetazolamide | Hyperphosphatemia, Polyuria, And Seizures |
| Acetazolamide | Intracranial Hypertension, Idiopathic |
| Acetohexamide | Diabetes Mellitus, Noninsulin-Dependent; Niddm |
| Acetohexamide | Maturity-Onset Diabetes Of The Young, Type 1; Mody1 |
| Acetohexamide | Maturity-Onset Diabetes Of The Young, Type 2; Mody2 |
| Acetohexamide | Maturity-Onset Diabetes Of The Young, Type 3; Mody3 |
| Acetophenazine | Schizophrenia; Sczd |
| Aciclovir | Rodrigues Blindness |
| Adapalene | Pyogenic Sterile Arthritis, Pyoderma Gangrenosum, And Acne |
| Adenosine | Wolff-Parkinson-White Syndrome |
| Adenosine Monophosphate | Wolff-Parkinson-White Syndrome |
| Alclometasone | Growth Retardation, Small And Puffy Hands And Feet, And Eczema |

Drug-disease association datasets

Map OMIM/UMLS to ICD10/MeSH

| db_id | UMLS_id | ICD10_codes | MeSH_codes |
|---|---|---|---|
| DB00007 | C0001144 | ['L70.0'] | ['D000152'] |
| DB00007 | C0005001 | ['N40'] | ['D011470'] |
| DB00007 | C0014175 | ['N80', 'N80.9'] | ['D004715'] |
| DB00007 | C0023267 | ['D25.9'] | ['D007889'] |
| DB00007 | C0032584 | ['K63.5', 'K51.4'] | ['D011127'] |
| DB00007 | C0042133 | ['D25.9', 'D25', 'D21.9'] | ['D063186', 'D014594', 'D007889'] |
| DB00007 | C0085652 | ['L88'] | ['D017511'] |
| DB00007 | C0376358 | ['C61'] | ['D011471'] |
| DB00007 | C0600139 | ['C61'] | ['D011471'] |
| DB00007 | C0678108 | ['L88', 'L70', 'L70.9', 'M00.9'] | ['D001168', 'D01171', 'D000152'] |
| DB00007 | C0702166 | ['L70.2', 'L70', 'L70.9'] | ['D000152'] |
| DB00007 | C1265823 | ['G72.49'] | ['D007249'] |
| DB00007 | C1333600 | ['Z15.09', 'C61'] | ['D009386'] |
| DB00007 | C1527263 | ['N40'] | ['D011470'] |
| DB00007 | C1559265 | ['K92.9'] | ['D012817', 'D064420'] |
| DB00007 | C1692886 | ['M00.80', 'M00.9'] | ['D001170'] |
| DB00007 | C1704272 | ['N40'] | ['D011470'] |
| DB00007 | C1832220 | ['Z80.42', 'C61'] | ['D011471'] |
| DB00010 | C0001144 | ['L70.0'] | ['D000152'] |
| DB00010 | C0085652 | ['L88'] | ['D017511'] |
| DB00010 | C0678108 | ['L88', 'L70', 'L70.9', 'M00.9'] | ['D001168', 'D01171', 'D000152'] |
| DB00010 | C0702166 | ['L70.2', 'L70', 'L70.9'] | ['D000152'] |
| DB00010 | C1692886 | ['M00.80', 'M00.9'] | ['D001170'] |
| DB00014 | C0006142 | ['C50.9', 'C50'] | ['D018270', 'D061325', 'D058922', 'D001943'] |
| DB00014 | C0014175 | ['N80', 'N80.9'] | ['D004715'] |

Similarity score measure between drug indications

ICD-10
Search
ICD-10 Version:2019
I Certain infectious and parasitic diseases
II Neoplasms
C00-C
C0
to
ha
C7
secondary and unspecified sites
C76 Malignant neoplasm of other and ill-defined sites
C76.0 Malignant neoplasm of other and ill-defined sites: Head, fac
C76.1 Malignant neopla defined sites: Thorax
C76.2 Malignant neopla defined sites: Abdomen
C76.3 Malignant neopla defined sites: Pelvis
All MeSH Categories
Diseases Category
Endocrine System Diseases
Diabetes Mellitus
Diabetic Coma +
Diabetic Ketoacidosis
MeSH

Calculate similarity scores using the hierarchical structure of ICD10 and MeSH

$$InformationContent: IC(d) = \log \frac{\frac{|leaves(d)|}{|subsumers(d)|} + 1}{|leaves(r) + 1} \quad (1)$$

$$CodeLevelSimilarity: CS(d_a, d_b) = 1 - \frac{2IC(d_c)}{IC(d_a) + IC(d_b)} \quad (2)$$

$$SetLevelSimilarity: SS(D_a, D_b) = \frac{\sum_{d_a \in D_a} \min_{d_b \in D_b} CS(d_a^{(i)}, d_b^{(j)}) + \sum_{d_b \in D_b} \min_{d_a \in D_a} CS(d_a^{(i)}, d_b^{(j)})}{||D_a|| + ||D_b||} \quad (3)$$

Categorize drug indications into “original” and “expanded”

| Drug | Drug Indication | Category |
|---|---|---|
| DB00001 | C0272275 | original |
| DB00002 | C0007102 | expanded |
| DB00002 | C0280324 | original |
| DB00002 | C0585362 | original |
| DB00002 | C1319317 | original |
| DB00002 | C3163899 | original |
| DB00003 | C0010674 | original |
| DB00004 | C0079773 | original |
| DB00005 | C0003872 | original |
| DB00005 | C0003873 | original |
| DB00005 | C0038013 | original |
| DB00005 | C0406317 | expanded |

Classify ‘original’ and 'expanded' drug indications by performing agglomerative clustering based on similarity score

Evaluate the performance of three representative drug-disease association prediction models in three scenarios

| | Train | Test |
|---|---|---|
| Scenario 1 | **Randomly select** drug indications | **Randomly select** drug indications |
| Scenario 2 | **Original** indications | **Original** indications |
| Scenario 3 | **Original** indications | **Expanded** indications |

Figure A2: Visualization of splitting drug indications in DDA datasets based on similarity scores

## A.3. Splitting drug indications in DDA datasets based on similarity scores

We used the set-level similarity proposed by Jia et al., 2019 as the similarity score for splitting drug indications into original and expanded as follows:

$$\text{InformationContent: IC(d)} = \log \frac{\frac{|leaves(d)|}{|subsumers(d)|}+1}{|leaves(r)+1} \quad (1)$$

$$\text{CodeLevelSimilarity: CS}(\mathrm{d}_a, \mathrm{d}_b) = 1 - \frac{2IC(d_c)}{IC(d_a)+IC(d_b)} \quad (2)$$

$$\text{SetLevelSimilarity: SS}(\mathrm{D}_a, \mathrm{D}_b) = \frac{\sum_{d_a \in D_a} \min_{d_b \in D_b} CS(d_a^{(i)}, d_b^{(j)}) + \sum_{d_b \in D_b} \min_{d_a \in D_a} CS(d_a^{(i)}, d_b^{(j)})}{||D_a||+||D_b||} \quad (3)$$

where $d$ denotes a single disease concept in ID10 or MeSH ontologies, $r$ is the root concept in each medical ontology, $d_c$ is the least common ancestor of disease concept $d_a$, and $d_b$. $D_a$ and $D_b$ are the sets of disease concepts for expressing a single drug indication. In addition, $leaves(d)$ denotes the leaf nodes that are a descendant of disease concept $d$, and

$subsumers(d)$ is the ancestors of disease concept $d$. On the other hand, we used MeSH tree numbers to calculate the similarity score in MeSH, because the unique identifiers in MeSH do not have a tree structure.

We performed splitting drug indications into original and expanded in all the combinations of clustering hyperparameters (Fig. A2), such as the number of clusters (2 or 3), disease ontology used to measure similarity (ICD10 or MeSH), and linkage criterion in the agglomerative clustering analysis (complete or average).

## Appendix B. Supplementary results

### B.1. Lists of ATC classes and target proteins with the highest odds ratios

In this section, we present lists of ATC classes and target proteins with the highest odds ratio measured in section 4.2 (Table B.1). Table B.1 highlights that most of the therapeutic area associated with the ATC classes and target proteins with high odds ratios are clinically associated. These results provide further evidence for the proposed connection between drug indications and data collection patterns discussed in section 2.

### B.2. Lists of drug indication pairs of similarity scores above the pre-determined thresholds

In this section, we present lists of drug indication pairs with similarity scores above the pre-determined threshold (Tables B.2, B.3, B.4, and B.5). We caution against dividing these similar disease concepts into train and test datasets, which are used to normalize drug indications, as this can lead to overestimated DR performance.

### B.3. Re-evaluation of existing DR models in three evaluation scenarios

In this section, we present the performances of three representative DR models in different evaluation settings under all the combinations of clustering hyperparameters (Table B.6). The degradation of performances in setting 3 was consistent in all the data split settings. The performances of DR models were lower when using the number of clusters as 2 rather than 3. This is because the size of the train dataset (i.e., the number of original indications - the number of expanded indications) is smaller when setting the number of clusters as 2.

Table B.1: Top 10 ATC classes and target proteins with the highest odds ratio and their clinical association

| **Odds ratio** | **ATC class** | **Target protein** | **Clinical association** |
|---|---|---|---|
| 26.4 | respiratory system | activin receptor-like kinase 3 (ALK3 or BMPR1a) | ALK3 is associated with development of respiratory disorder, including pulmonary arterial hypertension (PAH) and chronic obstructive disease (COPD) (Verhamme et al., 2015; Hoeper et al., 2023). |
| 17.9 | respiratory system | adrenergic receptor alpha-2 (ADRA2) | ADRA2 is a target protein of nasal preparation drugs such as oxymetazoline and xylometazoline (Haenisch et al., 2010). |
| 11.7 | respiratory system | alpha-glucosidase | not clear |
| 10.6 | antiparasitic products, insecticides and repellents | adrenergic alpha | Studies have suggested that adrenergic alpha agonists such as clonidine may inhibit parasitic growth and replication (Konrad et al., 2013; Porto et al., 2021). |
| 9.8 | nervous system | alpha-2c adrenergic receptor (ADRA2C) | ADRA2C is a target protein of drugs for treating schizophrenia, bipolar disorder, and major depressive disorder such as quetiapine and risperidone (Uys et al., 2017). |
| 8.8 | respiratory system | atp-binding cassette sub-family g member 2 (ABCG2 or BRCP) | ABCG2 is an efflux transporter protein important to transport toxic compounds across cell membranes in the respiratory system. (Leslie et al., 2005). |
| 8.1 | respiratory system | aurora kinase a (AURKA) | not clear |
| 7.9 | respiratory system | adrenergic alpha | Adrenergic alpha is a target protein of nasal decogestants such as phenylephrine and pseudoephedrine (Horak et al., 2009). |
| 7.5 | systemic hormonal preparations, excl. sex hormones and insulins | alpha-1d adrenergic receptor (ADRA1D) | not clear |
| 6.9 | antiparasitic products, insecticides and repellents | 3',5'-cyclic-AMP phosphodiesterase (cAMP-specific PDE) | The potential of PfPDE1, a cAMP-specific PDE from the human malaria parasite *Plasmodium falciparum*, as a target protein for malaria treatment has been studied. (Yuasa et al., 2005). |

Table B.2: Drug indication pairs in Fdataset of which similarity based on ICD10 was above a pre-determined threshold (0.6)

| UMLS Code 1 | Disease Name 1 | UMLS Code 2 | Disease Name 2 | Similarities |
|---|---|---|---|---|
| C0014859 | esophageal cancer (Esophageal Neoplasms) | C0027651 | Cancer (Neoplasm) | 0.600938012 |
| C0023418 | LEUKAEMIA (leukemia) | C0023448 | Lymphoblastic leukaemia NOS (Lymphoid leukemia) | 0.604412021 |
| C0036202 | Sarcoidosis | C1840264 | IMMUNE SUPPRESSION | 0.611315305 |
| C0011560 | Amyloid (Amyloid deposition) | C0085681 | Hyperphosphatemia | 0.615445301 |
| C0576224 | Small feet (Small foot) | C1300226 | MOTH-EATEN SKELETAL DYSPLASIA (Greenberg dysplasia) | 0.619091281 |
| C0003811 | cardiac arrhythmia | C1560249 | CARDIAC ARRHYTHMIA | 0.633847941 |
| C0011860 | Diabetes Mellitus, Non-Insulin-Dependent | C0011860 | NIDDM (Diabetes Mellitus, Non-Insulin-Dependent) | 0.633847941 |
| C0005744 | Blepharophimosis | C0005745 | Ptosis (Blepharoptosis) | 0.638848603 |
| C0270327 | Nocturnal Enuresis (Bedwetting) | C1833268 | ENUR2 (ENURESIS, NOCTURNAL, 2) | 0.666666667 |
| C0042875 | Vitamin E Deficiency | C1848533 | VED (AVED) | 0.666666667 |

Table B.3: Drug indication pairs in Fdataset of which similarity based on MeSH was above a pre-determined threshold (0.8)

| UMLS Code 1 | Disease Name 1 | UMLS Code 2 | Disease Name 2 | Similarities |
|---|---|---|---|---|
| C0014550 | Epilepsy, Myoclonic (Epilepsies, Myoclonic) | C0751785 | Unverricht (Unverricht-Lundborg Syndrome) | 0.801462118 |
| C0022350 | DJS (Jaundice, Chronic Idiopathic) | C1855980 | HYPERBILIRUBINEMIA, ROTOR TYPE | 0.802169339 |
| C0023448 | Lymphoblastic leukaemia NOS (Lymphoid leukemia) | C2063390 | acute lymphoma | 0.803133644 |
| C0027497 | Nausea | C1704628 | Hyperthermia | 0.803273592 |
| C0032460 | PCOS1 (Polycystic Ovary Syndrome) | C0085215 | PREMATURE OVARIAN FAILURE 1 (Ovarian Failure, Premature) | 0.804119097 |
| C0020437 | HYPERCALCAEMIA (Hypercalcemia) | C0270685 | Cerebral calcification | 0.808016006 |
| C0038454 | Stroke (Cerebrovascular accident) | C0852949 | Arteriopathy (Arteriopathic disease) | 0.810390489 |
| C0007758 | Cerebellar Ataxia | C0027066 | Myoclonus | 0.812515766 |
| C0398701 | Immunoglobulin G2 deficiency | C1840264 | IMMUNE SUPPRESSION | 0.816091392 |
| C0009917 | Contractures (Contracture) | C0026848 | Myopathy | 0.819916169 |

Table B.4: Drug indication pairs in the deepDR dataset of which similarity based on ICD10 was above a pre-determined threshold (0.6)

| **UMLS Code 1** | **Disease Name 1** | **UMLS Code 2** | **Disease Name 2** | **Similarities** |
|---|---|---|---|---|
| C0004771 | Bartonella Infections | C0014836 | Escherichia coli Infections | 0.600269 |
| C0010043 | Corneal Ulcer | C0022073 | Iridocyclitis | 0.601288 |
| C0009088 | Cluster Headache | C0393735 | Headache Disorders | 0.601349 |
| C0085437 | Meningitis, Bacterial | C0456107 | Neonatal meningitis | 0.604572 |
| C0346976 | Secondary malignant neoplasm of pancreas | C1282500 | Metastasis from malignant tumor of colon | 0.606009 |
| C0039614 | Tetanus | C1318973 | Staphylococcus aureus infection | 0.606237 |
| C1261256 | Pelvic inflammatory disease due to Mycoplasma hominis | C2733595 | Pulmonary Mycobacterium avium complex infection | 0.606237 |
| C0153246 | Tinea manus | C2349994 | Tinea barbae | 0.608611 |
| C0020598 | Hypocalcemia | C1704431 | Disorder of electrolytes | 0.615445 |
| C0018099 | Gout | C0268108 | Chronic gouty arthritis | 0.618304 |

Table B.5: Drug indication pairs in the deepDR dataset of which similarity based on MeSH was above a pre-determined threshold (0.8)

| **UMLS Code 1** | **Disease Name 1** | **UMLS Code 2** | **Disease Name 2** | **Similarities** |
|---|---|---|---|---|
| C0011616 | Contact Dermatitis | C0036508 | Seborrheic dermatitis | 0.800242518 |
| C0520777 | Chlamydial pelvic inflammatory disease | C1261256 | Pelvic inflammatory disease due to Mycoplasma hominis | 0.800252406 |
| C0085438 | Meningitis, Fungal | C0153256 | Candidal meningitis | 0.800669954 |
| C0004238 | Atrial Fibrillation | C0030590 | Paroxysmal supraventricular tachycardia | 0.800983448 |
| C0022602 | Actinic keratosis | C0043037 | Common wart | 0.802293872 |
| C0020461 | Hyperkalemia | C0020598 | Hypocalcemia | 0.802801151 |
| C0030920 | Peptic Ulcer | C0043515 | Zollinger-Ellison syndrome | 0.80563138 |
| C0027424 | Nasal congestion (finding) | C0035460 | Rhinitis, Vasomotor | 0.806782936 |
| C0006060 | Boutonneuse Fever | C0035021 | Relapsing Fever | 0.807744887 |
| C0014544 | Epilepsy | C0238111 | Lennox-Gastaut syndrome | 0.815127811 |

Table B.6: AUC performances of existing DR models in different evaluation settings under diverse clustering settings.

| | **Cluster 2** | | | **Cluster 3** | | |
|---|---|---|---|---|---|---|
| **Complete** | **Scenario 1** | **Scenario 2** | **Scenario 3** | **Scenario 1** | **Scenario 2** | **Scenario 3** |
| **deepDR** | | | | | | |
| **ICD10** | 0.922 | 0.916 | 0.840 | 0.920 | 0.914 | 0.900 |
| **MeSH** | 0.925 | 0.929 | 0.790 | 0.934 | 0.948 | 0.812 |
| **HNET-DNN** | | | | | | |
| **ICD10** | 0.931 | 0.937 | 0.798 | 0.927 | 0.947 | 0.873 |
| **MeSH** | 0.924 | 0.907 | 0.884 | 0.888 | 0.932 | 0.836 |
| **TP-NRWRH** | | | | | | |
| **ICD10** | 0.939 | 0.930 | 0.841 | 0.905 | 0.950 | 0.825 |
| **MeSH** | 0.911 | 0.945 | 0.750 | 0.926 | 0.940 | 0.833 |

| | **Cluster 2** | | | **Cluster 3** | | |
|---|---|---|---|---|---|---|
| **Average** | **Scenario 1** | **Scenario 2** | **Scenario 3** | **Scenario 1** | **Scenario 2** | **Scenario 3** |
| **deepDR** | | | | | | |
| **ICD10** | 0.884 | 0.905 | 0.849 | 0.944 | 0.904 | 0.872 |
| **MeSH** | 0.941 | 0.935 | 0.739 | 0.912 | 0.928 | 0.789 |
| **HNET-DNN** | | | | | | |
| **ICD10** | 0.917 | 0.937 | 0.837 | 0.931 | 0.934 | 0.849 |
| **MeSH** | 0.924 | 0.918 | 0.808 | 0.949 | 0.968 | 0.903 |
| **TP-NRWRH** | | | | | | |
| **ICD10** | 0.942 | 0.931 | 0.660 | 0.924 | 0.938 | 0.843 |
| **MeSH** | 0.921 | 0.942 | 0.749 | 0.931 | 0.912 | 0.815 |