Title
Analysis of Domain Knowledge for Machine Learning Prediction of Frequently Occurring Drug Side-Effects

Author
Liu, Han Jie

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analysis of Domain Knowledge for Machine Learning Prediction of Frequently Occurring Drug Side-Effects

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Bioengineering

by

Han Jie Liu

2023

ABSTRACT OF THE THESIS

Analysis of Domain Knowledge for Machine Learning Prediction of

Frequently Occurring Drug Side-Effects

by

Han Jie Liu

Master of Science in Bioengineering

University of California, Los Angeles, 2023

Professor Jennifer L. Wilson, Chair

Development of drugs often fails due to toxicity and intolerable side effects. Recent advancements in the scientific community have rendered it possible to leverage machine learning techniques to predict individual side effects with domain knowledge features, such as drug classification. While several factors can be used to anticipate drug effects including their targets, pathways, and drug classes, it is unclear which domain knowledge is most predictive and whether certain domain knowledge is more important than others for different side effects. The goal of this project is to understand the predictive values of drug targets, drug classification (level 2 ATC codes), and protein-protein interaction networks (PathFX targets and network proteins) for the prediction of 30 frequently occurring side effects. We compared the prediction accuracy for individual side effects of trained models across five domain knowledge combinations and discovered that level 2 ATC codes have the highest predictive value across the domain knowledge features. Logistic

regression coefficient analyses further suggest that side effects are significantly influenced by drug targets and drug classes, and not PathFX targets and network proteins. Our quantitative assessments may inform the development of safe and effective drugs by understanding the domain knowledge features underlying frequently occurring drug-induced side effects.

Keywords: Machine Learning, Drug Development, Domain Knowledge Features, Drug Targets, Level 2 ATC Codes, Protein-Protein Interaction Networks, Side Effects.

The thesis of Han Jie Liu is approved.

Aaron S. Meyer

William Hsu

Jennifer L. Wilson, Committee Chair

University of California, Los Angeles

2023

Table of Contents

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Jennifer Wilson for her guidance, mentorship, flexibility, patience, enthusiasm, and support throughout this project. I am grateful for all the positive energy and encouragement you have brought to me, which has served as a powerful source of motivation for me and has greatly enriched my academic experience as a Bioengineering graduate student at UCLA.

I would like to express my deep appreciation to Mayumi Prins, Janel Le Belle, Rana Khankan, and Katie Dixie for their generosity in offering me teaching assistant positions, which played a crucial role in funding my education. These experiences not only honed my scientific knowledge but strengthened my passion for mentorship. I am forever grateful for the opportunity to support, motivate, and inspire my students to pursue their goals.

I am grateful for the unconditional support of my mom, dad, and sister throughout this journey. Thank you for encouraging me to pursue my passion and supporting my decision to pursue an advanced degree in the biomedical research field.

Thank you to my girlfriend, Amy, for helping me navigate through my graduate school journey through your love and support. It made the journey less lonely and more meaningful with you by my side.

Thank you everyone for believing in me throughout my journey here at UCLA.

1. Introduction

The development of drugs often fails during clinical trials due to toxicity and intolerable side effects. Sun et al. (2022) analyzed clinical trial data from 2010 to 2017 and found that over 30% of drugs failed due to unmanageable toxicity. Furthermore, off-target toxicity from drugs can trigger dangerous side effects and cause clinical trial failure (Lin et al., 2019). For example, the kinase inhibitor Sunitinib is known to trigger cardiotoxicity through its interaction with proteins outside of what the drug was intended to bind (Force & Kolaja, 2011). Currently, there are strict guidelines and protocols set in place by the United States Food and Drug Administration (FDA) to ensure drug safety and efficacy. Despite this, many drugs that are approved on the market have intolerable adverse side effects documented. Notably, propranolol hydrochloride, despite receiving approval from the FDA in 2014 for the treatment of infantile hemangiomas (Kurta et al., 2018), has been associated with sleep disturbance, agitation, and bronchial hyperreactivity (Ji et al., 2018). These findings suggest that innovation in drug development related to improved safety and efficacy could advance therapeutic development.

Multiple data-driven resources have made it possible for the scientific community to better explore the relationship between drug target associations to various side effects. Kuhn et al. (2016) generated the Side Effect Resource (SIDER) database, which documents results from human clinical trials with ADRs on FDA-approved drugs. Separately, protein-protein interaction (PPI) networks, such as PathFX, seek to understand drug-induced effects by constructing drug pathways and integrating gene-disease phenotype associations from multiple databases. These drug pathways provide druggable targets and proteins downstream of targets associated with drug phenotypes. We also previously discovered that proteins downstream of druggable targets were more predictive of drug side effects as compared to drug targets, for severe ADRs listed on drugs'

labels (Wilson et al., 2022). DrugBank also contains domain knowledge about each drug, such as its Anatomical Therapeutic Chemical (ATC) classification and drug development group status (i.e., approved, experimental), which many have used for anticipating side effects from within-class drugs (Wishart et al., 2006).

Drug targets are often the starting place for predicting drug side effects. Campillos et al. (2008) demonstrated that shared drug side effect profiles were predictive of drug targets. Moreover, Xie et al. (2009) explored protein-drug interaction networks of Cholesteryl Ester Transfer Protein inhibitors and identified a panel of off-target interactions that influenced side effects. LaBute et al. (2014) trained an L1-regularized LR model based on UniProt ID numbers of drug targets to predict 85 side effects from SIDER grouped into 10 adverse drug reaction (ADR) phenotype groups and achieved a model area under the curve (AUC) of 0.61 – 0.74 during 10-fold cross-validation. However, drugs may have undocumented off-targets responsible for their effects, making drug targets alone insufficient for side effect prediction.

Additional domain knowledge could improve anticipation of side effects without knowing all off-targets. Huang et al. (2011) developed a logistic regression model by integrating ADR information, drug-target data, PPI networks, and gene ontology term annotations to predict cardiotoxicity and achieved a performance of 0.675 in performance accuracy, the median AUC of 0.771, and sensitivity of 0.632. However, this analysis was limited to predicting cardiotoxicity. They discovered that off-target proteins had more predictive power than documented on-target drug-protein interactions related to cardiotoxicity. Recently, Liang et al. (2020) trained a random forest (RF) model by sampling negative cases using the random walk with restart algorithm. Furthermore, they incorporated various domain knowledge, including drug fingerprint, ATC codes, literature association of drug-protein interactions, drug structure, and drug targets for the

prediction of drug side effects with RF model yielding nearly perfect performance (accuracy = 0.975).

Given recent successes with the integration of multiple drug data types and our previous discovery of the predictive utility of network proteins, we sought to measure the relative predictive value of drug targets, drug class, and drug network proteins for the prediction of frequently occurring individual side effects in SIDER. Since ATC codes have been leveraged in building models to predict drug side effects (Liang et al., 2020), incorporating such domain knowledge in machine learning (ML) may provide us further insights into specific drug classes that can influence frequently occurring individual side effects. Furthermore, by leveraging PathFX network proteins in our model, we sought to uncover certain proteins downstream of druggable targets that may influence certain side effects. The exploration of these three domain knowledge features has the potential to provide valuable insights for personalized medicine by identifying certain features that can influence drug side effects, which can assist clinicians in making informed decisions for prescribing medicine to patients. By understanding the predictive value of drug targets, drug class, and drug network proteins, we can inform the therapeutic development of safer and more effective drugs to enhance patient outcomes and minimize ADRs.

2. Methodology

2.1. Data Processing

2.1.1 SIDER 4.1

First, we downloaded SIDER 4.1 datasets (http://sideeffects.embl.de/download/) and prioritized two of them: 1) Medical Dictionary for Regulatory Activities (MedDRA) all side effects (meddra_all_se.tsv.gz) and 2) drug names (drug_names.tsv). The MedDRA all side effects dataset contains all side effects of FDA-approved drugs documented in MedDRA. The first and last columns of the MedDRA all side effects dataset were extracted, which represent the drug ID and its associated drug side effect, respectively. Then, we mapped each drug ID to the drug name using the drug names dataset. Last, we counted the occurrence of all side effects and extracted the drug names associated with the 30 most common side effects individually for further analysis.

2.1.2 Standardizing drug names to DrugBank IDs

DrugBank is a database that classifies specific drugs and their common synonyms under a DrugBank identifier (DBID), which consists of a DB prefix and suffix of 5 numbers. Standardizing drug names to their respective DBID can increase the accuracy of mapping drugs across datasets by mitigating data loss due to differences in naming and spelling. We downloaded a dataset that contains the common names and synonyms of a drug to its DBID (drugbank_vocabulary.csv). A default dictionary was generated by extracting the drug names as the key and its associated DBID as the value. The drug names from this dictionary were mapped to the drug names in SIDER 4.1 to standardize them to DBIDs.

2.1.3 Running PathFX on all drugs in DrugBank version 5.1.6

We analyzed all available drugs in DrugBank version 5.1.6 using PathFX on the Hoffman cluster to extract PathFX targets and network proteins. Briefly, PathFX generates a PPI network around drug targets based on the amount and quality of evidence supporting the PPIs. Next, PathFX uses a modified Fisher's exact test to discover biological phenotypes associated with the drug's network (full description in Wilson et al 2018). Importantly, PathFX can only generate a network when a drug has documented drug-binding proteins and those proteins are connected to the PathFX interactome. Of the 13474 drugs listed in DrugBank, PathFX generated a network file and phenotype association table for 7012 drugs - 2,232 of which are approved on the market and 4,780 which were experimental and not FDA-approved.

2.1.4 Extracting domain knowledge features to dictionaries

We sought to extract domain knowledge features associated with each DBID by storing them in dictionaries (key = DBID, value = domain knowledge feature) and subsequently appending the DBID (row) and domain knowledge features (columns) to generate the ML matrix. To assess the utility of domain knowledge for side-effect prediction, we considered 5 comparisons: 1) ATC level 2 codes only (ATC model), 2) DrugBank targets only (DT model), 3) DrugBank + PathFX targets and network proteins (DT/PathFX model), 4) DrugBank targets + ATC (DT/ATC model), and 5) DrugBank + PathFX targets and network proteins + ATC (DT/ATC/PathFX model). The level 2 ATC code consists of the first three characters of the ATC code. There are currently 94 distinct level 2 ATC codes, each one of them indicating the system of action of the drug and its associated pharmacological and therapeutic properties. For example, C08 are calcium channel blockers that influence the cardiovascular system. We extracted both the level 2 ATC codes and

drug targets and generated a set dictionary with its associated DBID from DrugBank version 5.1.6. All PathFX targets and network proteins were extracted from the "merge_neighborhood_.txt" files for all 7012 drugs using the os.walk function. Ultimately, these sets were merged using the union operator to generate the dictionaries for the five experimental conditions.

2.1.5 Matrix generation and filtering

For each of our five combinations of domain knowledge, we generated a ML input matrix where each row indicated a drug and the columns included a label of 1 (presence) or 0 (absence) of a domain-knowledge data type: 1) DrugBank target, 2) a PathFX target or network protein, or 3) a level 2 ATC code. We repeated this process for each of the 30 side effects and created 150 data matrices in total (30 side effects x 5 combinations of predictor variables). Since SIDER 4.1 only documents side effects observed in FDA-approved drugs, we generated a subset of the matrix by excluding drugs that were not FDA-approved (i.e., experimental drugs).

2.2 Machine learning model implementation

2.2.1 Exploring the confounding effects of unapproved drugs on logistic regression model performance

We ran the Logistic Regression (LR) model on an 80/20 train-test split with a random undersample of negative cases bootstrapped 100 times to evaluate its performance for predicting the 30 most common side effects in SIDER 4.1 using drug targets only. Since the matrix contains more negative cases than positive ones, bootstrapping the negative cases can expose the model to a broader range of negative instances to improve its generalization. This procedure was executed

for the matrix containing 1) all drugs and 2) FDA-approved drugs only. We then extracted the 10 most common distinct drug targets only in unapproved drugs and identified the logistic coefficients for those targets to evaluate their confounding effects on LR model performance.

2.2.2 Initial evaluation of classification models for selection

We selected six ML models from scikit-learn capable of performing binary classification for initial evaluation and selection. Specifically, we selected the LR model, Random Forest Classifier (RFC), Support Vector Machine, Decision Tree Classifier, Naive Bayesian Classifier, and K-Nearest Neighbor model on an 80/20 train-test split with random undersample of negative cases to evaluate its accuracy in predicting dizziness, the side-effect associated with the most drugs, on drug targets of FDA-approved drugs. Lastly, we selected the top two models based on their performance accuracy for subsequent analyses.

2.2.3 Running LR model and RFC

We used LR and RFC to evaluate the predictive value of experimental variables on 30 individual side effects for model selection of subsequent analyses. We bootstrapped negative samples 100 times using a random undersample with an 80/20 train-test split for every experimental group across all side effects. We compared the predictive value of both models on these side effects and selected the model that had the higher average accuracy across the 30 individual side effects.

2.2.4 Extracting LR coefficients

We extracted the LR coefficients from the trained model to understand which domain knowledge variables the model prioritized. In this project, the p variable of the LR model Equation (1) represents the probability that the side effect of interest will occur. The p threshold of our LR model is set to 0.5, in which any value greater than 0.5 will be classified with an output label of 1 (presence of side effect). The LR model assigns a coefficient to each variable based on the outcome variable as shown in Equation (1), where the β terms represent the coefficients and X represents the value of the predictor variable. Positive β terms suggest that an increase in the corresponding predictor variable leads to an increase in the outcome variable. Conversely, negative β terms imply that an increase in the corresponding predictor variable leads to a decrease or may not affect the outcome variable. The magnitude of the coefficient reflects the strength of the relationship between the predictor and outcome variable. These coefficients are then extracted to evaluate 1) the confounding effects of drug targets unapproved in the market and 2) the validity of the suggested drug-to-variable relationship for individual side effects.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n)}} \quad (1)$$

2.3. Statistical analysis

The initial unfiltered matrix contains both drugs approved on the market and experimental drugs, which may confound the model's performance. To evaluate the confounding effects of experimental drugs, we compared the performance of the two models using a dependent samples t-test with drug targets as the predictor variable.

We excluded all the experimental drugs in our matrix for subsequent analyses to ensure an accurate representation of the data. This is because SIDER does not document the side effects of experimental drugs, and therefore, the relationship between side effects and targets of unapproved drugs is not well established. Thus, including experimental drugs in our analyses could generate misleading results.

We performed an Analysis of Variance with Repeated Measures (ANOVA-RM) to test our hypothesis that there are between-group differences across the five combinations of domain knowledge for the prediction of individual side effects before proceeding further with subsequent analyses. Then, we performed a dependent samples t-test to assess specific between-group differences across the five combinations of domain knowledge. We benchmarked drug targets and evaluated the change in model performance for predicting individual side effects with the addition of domain knowledge independently for the following groups: 1) Drug Targets and PathFX proteins (DT/PathFX) model, 2) Drug Targets and ATC codes (DT/ATC) model, and 3) Drug Targets, ATC Codes, a d PathFX proteins (DT/PathFX/ATC) model. We chose a significance level of 0.05 for all our tests.

2.4 Software and code

The data collection, processing, and model training were conducted in Python version 3.7 using Jupyter Notebook version 6.3.0. The packages deployed for this project included: 1) Pandas, Numpy, and Pickle for data processing, 2) Matplotlib and Seaborn for data visualization, 3) Imbalanced-learn to balance the binary cases, 4) Scikit- learn for modeling processed data and evaluating results, and 5) Scipy for statistical analyses.

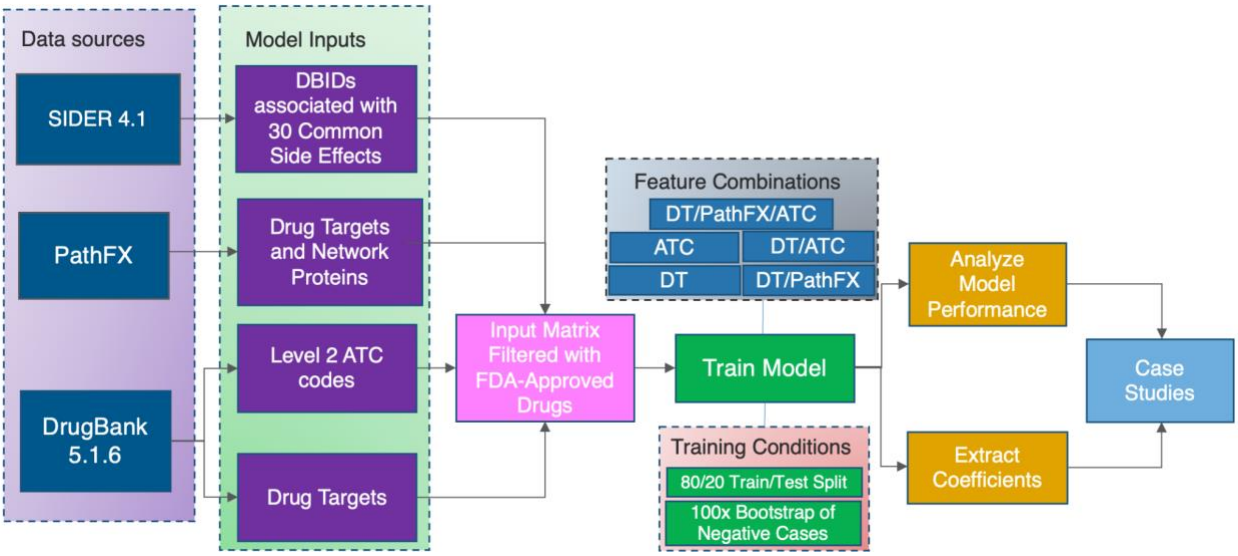

Figure 1. A high-level overview of our model construction and evaluation process to identify the predictive value of domain knowledge features on frequently occurring drug side effects.

3. Results

3.1 Data characteristics

SIDER 4.1 documented 309,848 side effects across 1,430 drugs. SIDER splits side effects based on their classification in the MedDRA as either 1) preferred terms (PTs), which is a distinct medical concept for the associated side effect (i.e., nausea), or 2) lowest level terms (LLTs), which parallels how information is communicated to patients (i.e., feeling queasy). Each LLT is linked to only one PT, whereas each PT has at least one LLT. Because of this, nearly all drugs to side effect combinations may be documented multiple times. We extracted the top 30 most common side effects based on both PTs and LLTs from SIDER, with dizziness having the highest count (n = 2826) and musculoskeletal discomfort having the 30th-most count (n = 1255) in our analysis. The side effect associated with the highest number and lowest number of unique drugs in our analysis is nausea (n = 1207) and arthralgia (n = 588), respectively. We next matched SIDER drugs to DBIDs for integration with other data sources. Of the 1,430 drugs listed in SIDER, 1,079 of them mapped to a DBID. The percentage of drugs matched to a DBID ranged from 79.7 to 86.7% per side effect. Our original ML input matrix consisted of 7,012 drugs with documented targets or PathFX network proteins – 2,232 approved drugs and 4,780 experimental, unapproved drugs. The percentage of DBIDs from SIDER that matched our ML input matrix, which consisted of DBIDs associated with a documented target or PathFX network proteins, ranged from 90.4 to 95.4% depending on the domain knowledge (some drugs did not have documented targets or PathFX networks). We curated a total of 88 level 2 ATC codes, 3,819 drug targets, and 6,467 drug targets with PathFX network genes which were included in our input matrix for further analyses.

3.2 Logistic regression and random forest outperform other ML models for initial side effect prediction

We first benchmarked six ML models on the most common side effect documented in SIDER: dizziness. We specifically modeled dizziness using 1) Logistic Regression (LR) model, 2) Random Forest Classifier (RFC), 3) Support Vector Machine, 4) Decision Tree Classifier, 5) Naive Bayesian Classifier, and 6) K-Nearest Neighbor model and discovered that RFC had the highest, and LR had the second highest performance as shown in Table 1. Thus, we considered these two models in subsequent additional analyses.

Binary Classification Model	**Precision**		**Recall**		**F-1 Score**		**Accuracy**
	Negatives (n = 156)	Positives (n = 162)	Negatives (n = 156)	Positives (n = 162)	Negatives (n = 156)	Positives (n = 162)	
Logistic Regression	0.64	0.67	0.69	0.62	0.66	0.65	0.65
Random Forest	0.64	0.67	0.68	0.64	0.66	0.65	0.66
Support Vector	0.63	0.65	0.63	0.65	0.63	0.65	0.64
Decision Tree	0.61	0.67	0.72	0.55	0.66	0.60	0.63
Naive Beysian	0.73	0.57	0.29	0.90	0.42	0.70	0.60
K-Nearest Neighbors	0.56	0.69	0.82	0.39	0.67	0.50	0.60

Table 1. Prediction performance of dizziness using FDA-approved drug targets with multiple ML models.

3.3 Logistic regression has the highest average prediction accuracy across side effects

We analyzed the top 30 most frequent side effects in SIDER, using targets alone to predict the occurrence of the side-effect compared to non-side-effect associated drugs using both approved and experimental drugs. We further completed these prediction tasks using RFC and LC models and measured their accuracy across side effects. The LR model had a higher average accuracy (0.67) compared to the RFC (0.66) for prediction across all 30 side effects. The side-effect with

the highest LR prediction accuracy was thrombocytopenia with a prediction accuracy of 0.71. The side-effect with the lowest LR prediction accuracy was infection with a prediction accuracy of 0.6.

3.4 Drug targets of unapproved drugs confounded LR performance

We analyzed the confounding effects of unapproved drug targets by predicting the 30 most common SIDER side effects on drug targets with LR on 100-repeat bootstrap for all drugs and approved drugs only. The model accuracy was higher when all drugs were included across all side effects compared to approved drugs only. Specifically, the mean model accuracy ranged from 0.768 to 0.833 in all drugs, and 0.612 to 0.702 in approved drugs. We hypothesized that drug targets for unapproved drugs were distinct from approved drugs and influenced model performance. Of the 3,819 drug targets curated, 1,276 of them are associated with unapproved drugs only. We extracted the regression coefficients of the 10 most common unapproved drug targets, and at least 5 of them were assigned a relatively negative coefficient number, suggesting that the model prioritized these targets for predicting frequently occurring side effects. Certain side effects, such as nausea, headache, and diarrhea, assigned negative coefficients for all 10 most common unapproved drug targets as shown in Table 2.

Target	Count	Nausea Coef.	Headache Coef.	Diarrhea Coef.
CCNA2	66	-0.29	-0.14	-0.35
PKIA	60	-0.43	-0.42	-0.28
BACE1	56	-0.29	-0.18	-0.42
map	46	-0.40	-0.44	-0.51
MMP3	44	-0.29	-0.03	-0.31
thyA	44	-0.47	-0.32	-0.06
CTSK	44	-0.49	-0.32	-0.31
NCOA1	42	-0.26	-0.14	-0.14
CELA1	34	-0.29	-0.33	-0.55
MMP8	30	-0.49	-0.35	-0.26

Table 2. Most frequent targets for experimental drugs and their regression coefficients in three example side effects: Nausea, Headache, and Diarrhea

3.5 ANOVA-RM suggests between-group differences across side effects

After comparing ML models and dropping unapproved drugs, we repeated LR analysis for all 30 side effects using 5 different combinations of domain knowledge (see methods 2.1.4): 1) Drug Targets and PathFX proteins (DT/PathFX) model, 2) Drug Targets and ATC codes (DT/ATC) model, 3) Drug Targets (DT) model, 4) ATC Codes, and 5) PathFX proteins (DT/PathFX/ATC) model. We then performed an ANOVA-RM across all experiment groups to assess between-group differences across side effects. The results show significant between-group differences across all groups for the prediction of 30 individual side effects, with F-values ranging from 20.5 to 140.8, and P-values from 9.87E-75 to 2.42E-15 as shown in Tables 3 and 4.

3.6 Incorporation of level 2 ATC codes improved prediction accuracy for both DT and DT/PathFX models

The results of the LR analysis showed that the average prediction accuracy for the DT model was 0.67, while the average prediction accuracy for the DT/ATC model was 0.70. Consequently, the average prediction for the DT/PathFX model was 0.66, while the average prediction accuracy for the DT/PathFX/ATC model was 0.68. We then performed a paired t-test to assess the effect of incorporating ATC codes with drug targets benchmarked with drug targets on predicting the 30 most common SIDER side effects using LR at the significance level of 0.05. Incorporation of level 2 ATC codes in the DT model significantly improved model performance across all side effects, with t-test statistics ranging from -15.26 to -3.52, and p-values from 9.58E-28 to 6.57E-04. Incorporation of level 2 ATC codes in the DT/PathFX model significantly improved performance across all side effects, with t-test statistics ranging from -11.60 to -2.36, and p-values from 3.78E-20 to 2.02E-02 (Table 3).

3.7 PathFX targets and network proteins improve LR model performance for prediction of seven side effects

We performed a paired t-test to evaluate the predictive power of PathFX targets and network proteins on the 30 most common SIDER side effects when benchmarked with drug targets alone. The addition of PathFX targets and network proteins improved LR model performance for seven side effects, which include: pruritus, vomiting, gastrointestinal disorder, dermatitis, insomnia, infection, and hypotension (Table 4).

3.8 DT/PathFX/ATC model improve LR model for prediction of six side effects compared to DT/ATC model

After comparing the performance of the DT/PathFX and DT model for predicting 30 side effects, we compared the performance of the DT/ATC model to the DT/PathFX/ATC model to assess the impact of ATC codes and determine if the same side effects would be affected. Interestingly, the DT/PathFX/ATC model only improved prediction for six out of the seven side effects listed in Table 4. One unique observation is in the case of dermatitis, where the DT/PathFX model exhibited higher prediction accuracy compared to the DT model. However, the DT/ATC model surpassed the DT/PathFX/ATC model for LR prediction of dermatitis, suggesting a stronger ATC class-driven effect.

3.9 Level 2 ATC codes are more predictive than drug targets for 17 side effects

We performed a paired t-test to compare the predictive power of level 2 ATC codes and DT for predicting the 30 most common SIDER side effects. ATC codes were shown to be more predictive than DT for 17 drug side effects, with the largest difference in prediction accuracy between ATC codes and drug targets occurring for the side effect, infection. DT was more predictive than ATC codes for only 8 side effects, with the largest difference in prediction accuracy between drug targets and ATC codes being for the side effect, arthralgia. There were no significant differences in predictive power between ATC codes and DT for 5 side effects, which include constipation, abdominal pain, diarrhea, musculoskeletal discomfort, and vomiting. Overall, the predictive power was shown to be similar between level 2 ATC codes and drug targets with t-test statistics ranging from -10.61 to 13.46, P-values from 4.19E-24 to 7.91E-01, and the average difference in accuracy being 0.01.

3.10 Three trends were identified across the four model conditions DT, DT/PathFX, DT/ATC, and DT/PathFX/ATC

Three distinct trends for LR prediction across the 30 side effects were identified by analyzing the model accuracy of the DT, DT/PathFX, DT/ATC, and DT/PathFX/ATC models as listed below.

- Trend 1 (6 side effects): DT/PathFX model accuracy is greater than the DT model. Both these model performances improve with the addition of ATC codes, with the DT/PathFX/ATC model demonstrating the highest performance.
- Trend 2 (23 side effects): DT model accuracy is greater than the DT/PathFX model. Both these model performances improve with the addition of ATC codes, with the DT/ATC model demonstrating the highest performance.
- Trend 3 (1 side effect): DT/PathFX model accuracy is greater than the DT model. Both these models improve with the addition of ATC codes, with the DT/ATC model demonstrating the highest performance.

Side Effect	DT model	ATC model	DT/PathFX model	DT/ATC model	DT/PathFX/ ATC model	F-Value	P-value
thrombocytopenia	0.71	0.70	0.67	0.73	0.69	85.36	3.15E-52
constipation	0.70	0.70	0.69	0.73	0.70	33.62	3.71E-24
somnolence	0.70	0.72	0.68	0.73	0.70	61.63	1.85E-40
tachycardia	0.70	0.69	0.67	0.72	0.68	48.88	2.08E-33
asthenia	0.69	0.72	0.69	0.73	0.71	90.37	1.61E-54
diarrhea	0.69	0.69	0.68	0.72	0.70	54.47	1.43E-36
dyspepsia	0.69	0.66	0.67	0.71	0.67	67.63	1.38E-43
arthralgia	0.69	0.65	0.67	0.74	0.69	132.52	1.01E-71
dizziness	0.68	0.68	0.65	0.69	0.68	38.68	2.46E-27
nausea	0.68	0.70	0.66	0.70	0.69	69.35	1.84E-44
rash	0.68	0.69	0.66	0.71	0.68	96.80	2.24E-57
abdominal pain	0.68	0.68	0.66	0.70	0.68	32.21	2.98E-23
headache	0.67	0.66	0.62	0.69	0.65	140.80	9.87E-75
dyspnoea	0.67	0.69	0.65	0.70	0.68	59.05	4.47E-39
anaphylactic shock	0.67	0.66	0.63	0.69	0.66	51.44	7.12E-35
paraesthesia	0.67	0.66	0.66	0.70	0.67	31.03	1.74E-22
urticaria	0.66	0.67	0.66	0.69	0.68	20.47	2.42E-15
body temperature increased	0.66	0.67	0.66	0.69	0.68	22.36	1.16E-16
dermatitis	0.66	0.68	0.68	0.70	0.68	71.06	2.51E-45
fatigue	0.66	0.67	0.64	0.69	0.67	60.49	7.54E-40
musculoskeletal discomfort	0.66	0.67	0.65	0.71	0.66	59.86	1.64E-39
hypersensitivity	0.64	0.66	0.61	0.66	0.62	110.46	3.82E-63
pain	0.64	0.66	0.63	0.66	0.65	33.01	9.05E-24
nervous system disorder	0.64	0.62	0.64	0.67	0.67	57.35	3.74E-38

Table 3. ANOVA-RM LR prediction of 23 side effects with higher performance in DT model than DT/PathFX model. Each cell represents the prediction accuracy from 100 bootstrapped samples. F-values indicate the ratio of variability between conditions to within conditions. P-values reflect the probability of obtaining the observed differences in means given the null hypothesis is true.

Side Effect	DT model	ATC model	DT/PathFX model	DT/ATC model	DT/PathFX/ ATC model	F-Value	P-value
vomiting	0.67	0.67	0.69	0.69	0.70	47.31	1.69E-32
dermatitis	0.66	0.68	0.68	0.70	0.68	71.06	2.51E-45
hypotension	0.66	0.70	0.68	0.69	0.71	45.80	1.28E-31
pruritus	0.64	0.66	0.66	0.67	0.68	37.13	2.24E-26
gastrointestinal disorder	0.64	0.66	0.67	0.66	0.69	38.19	4.96E-27
insomnia	0.64	0.69	0.67	0.68	0.68	47.39	1.51E-32
infection	0.61	0.69	0.63	0.65	0.66	125.66	3.81E-69

Table 4. ANOVA-RM LR prediction of 7 side effects with higher performance in DT/PathFX model than DT model. Each cell represents the prediction accuracy from 100 bootstrapped samples. F-values indicate the ratio of variability between conditions to within conditions. P-values reflect the probability of obtaining the observed differences in means given the null hypothesis is true.

3.11 Trend 1 case study: LR prediction of gastrointestinal disorder is enhanced when level 2 ATC codes and PathFX targets and network proteins is incorporated with drug targets

Gastrointestinal disorder LR prediction accuracy increased when using all domain area knowledge (accuracy = 0.69) compared with drug targets only (accuracy = 0.64) and drug targets with ATC codes (accuracy = 0.66) as shown in Figure 2. To better understand how LR prioritized domain knowledge, we extracted the top and bottom 30 LR coefficients for this side effect (Table 5) and discovered that 20/30 of the largest positive coefficients were level 2 ATC codes with the largest being A10. The 30 most negative LR targets had 7/30 coefficients that were level 2 ATC codes, and 1/30 that were proteins adjacent to drug targets (network downstream proteins).

Overall, ATC class association and certain drug targets were shown to be strong predictors of gastrointestinal disorder.

We sought literature support for the importance of features prioritized by the LR model with all domain knowledge included. We specifically emphasized the drug target, ATP binding cassette subfamily B member 11 (ABCB11), and the level 2 ATC code A10 because they had the highest coefficient values assigned in the DT/ATC/PathFX model, which had the highest performance. The evidence from the literature supports the relationship between the LR model coefficients of these variables. Chen et al. (2016) studied the effects of ingesting anti-tuberculosis drugs on Chinese individuals with the ABCB11 SNP rs2287616 and observed some adverse effects including gastrointestinal disorders, arthralgia, and pruritus. The Level 2 ATC code A10 is associated with drugs used in diabetes. An example of a drug associated with the A10 ATC code is Metformin, which is prescribed for individuals with diabetes to help control their blood sugar levels. This drug has commonly been associated with side effects of gastrointestinal disorder along with nausea, vomiting, and diarrhea, with a prevalence of 2-63% (Siavash et al., 2017).

Further, we evaluated the relationship of negative coefficients of the drug target folP and level 2 ATC code D01 on the side effect of gastrointestinal disorder. We selected folP because its coefficient values were consistently in the bottom 3 most negative values across the DT and DT/ATC/PathFX models. D01 was selected for further evaluation since it was the most negative level 2 ATC code listed in the DT/ATC/PathFX model. The folP gene encodes for Dihydropteroate synthase (DHPS), an enzyme involved in the synthesis of folate in bacteria. According to Yoshida et al. (2022), inhibition of DHPS activity by Dapsone improves gastrointestinal symptoms in children with immunoglobulin A vasculitis (Yoshida et al., 2022), which contradicts the relationship which the LR model identified. The ATC code D01 is associated with Antifungals for

dermatological use. Currently, the association between D01 drugs and gastrointestinal disorders is not well understood. However, the ATC code A03 is associated with drugs for functional gastrointestinal disorders, which is the 2nd most negative ATC code classified by the LR model.

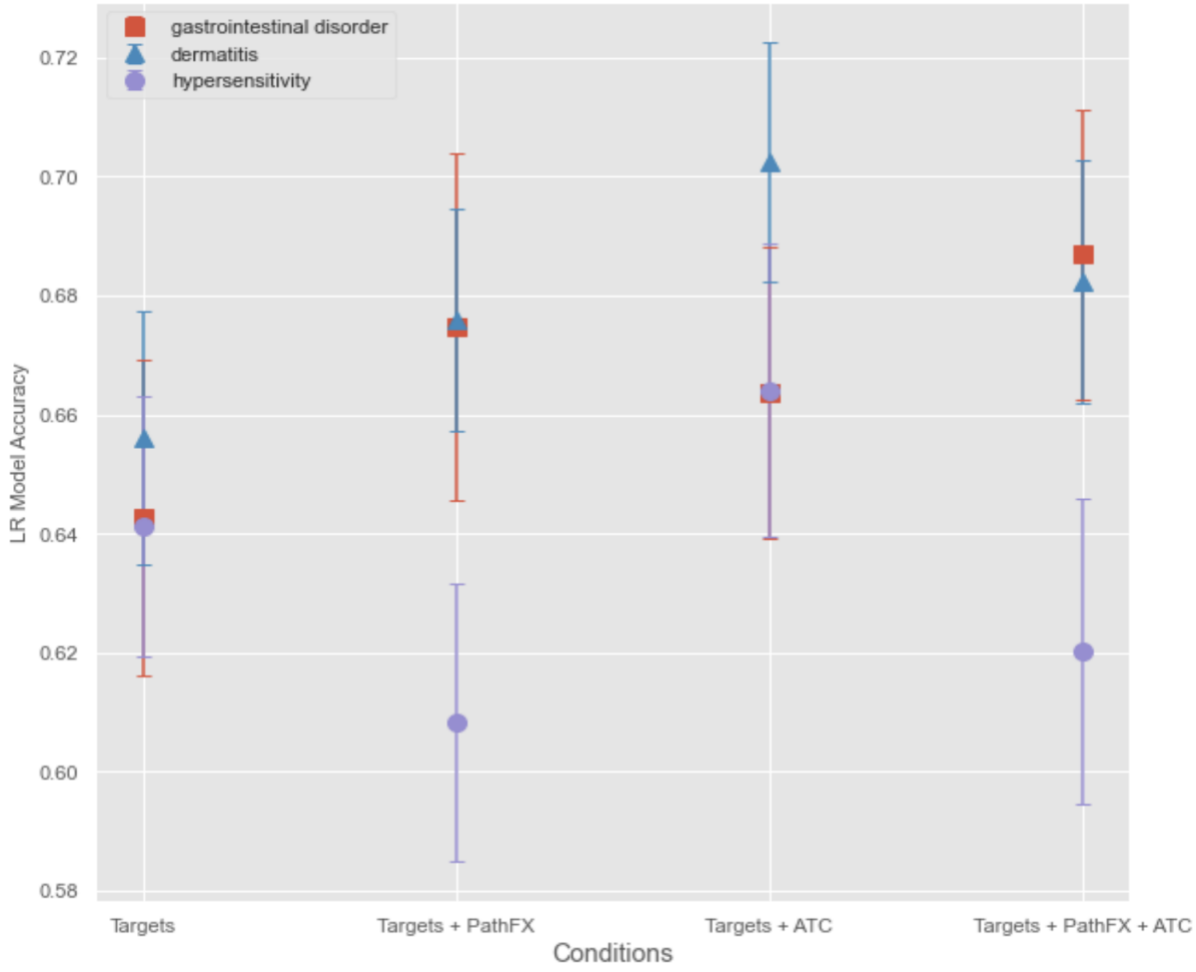

Figure 2. LR model accuracy for side effects of gastrointestinal disorder, dermatitis, and hypersensitivity across DT, DT/PathFx, DT/ATC, and DT/ATC/PathFX models. The square, triangle, and circle represent the mean prediction accuracy for side effects of gastrointestinal disorder, dermatitis, and hypersensitivity, respectively. Error bars represent one standard deviation of uncertainty.

Gastrointestinal Disorder							
Drug Targets Only (DT model)				Drug and PathFX Targets and ATC Codes (DT/ATC/PathFX model)			
Positive Features	**Coefficients**	**Negative Features**	**Coefficients**	**Positive Features**	**Coefficients**	**Negative Features**	**Coefficients**
XDH	1.56	NR3C1	-1.30	A10	1.97	SLC22A11	-1.17
DCK	1.28	CYP1B1	-1.03	ABCB11	1.59	folP	-1.05
HTR1B	1.20	folP	-0.97	A06	1.58	SLC10A1	-1.05
OPRK1	1.18	AR	-0.96	R03	1.38	GABRA1	-1.04
DPP4	1.14	ATP1A1	-0.95	A07	1.22	D01	-0.89
CYP3A4	1.14	CYP4A11	-0.93	J01	1.21	NR3C1	-0.80
SLC6A3	1.13	HTR2C	-0.90	L01	1.20	SLC12A3	-0.78
BCL2	1.11	ESRRG	-0.90	J05	1.13	A03	-0.76
PTGS2	1.10	CNR1	-0.90	J02	1.08	HTR6	-0.73
GNRHR	1.09	CYP2B6	-0.89	N03	1.08	pbpC	-0.68

Table 5. Top and bottom 10 LR coefficients from the DT and DT/ATC/PathFX model for the side effect of Gastrointestinal Disorder. Positive coefficients suggest that the feature is positively associated with the side effect. Conversely, negative coefficients imply the feature is preventative against or may not affect the side effect.

3.12 Trend 2 case study: Drug targets with ATC codes had the highest prediction accuracy for hypersensitivity

For hypersensitivity, the DT/ATC model had the highest prediction accuracy (accuracy = 0.66, Table 3). The DT/ATC/PathFX model had lower prediction accuracy than DT alone (accuracy = 0.62 compared to accuracy = 0.64, respectively) as shown in Figure 2. To better understand this trend, we extracted the top and bottom 30 LR coefficients for this side effect (Table 6) and counted the number of features that weren't drug targets in the 1) DT/ATC/PathFX and 2)

DT/ATC models. We discovered that 15/30 of the largest positive variables and 3/30 of the negative variables were level 2 ATC codes in the DT/ATC/PathFX model. However, in the DT/ATC model, only 14/30 of the largest positive coefficients and 2/30 of the most negative coefficients were level 2 ATC codes. Consistent with gastrointestinal disorder, these findings suggest that ATC class association and certain drug targets are strong predictors of hypersensitivity.

We again sought literature evidence to support features that were prioritized by the LR model. Specifically, we selected the drug target prostaglandin D2 (PGD), which was the 10th highest feature in the DT model, and ATC code J01, which was the 10th highest feature in the DT/ATC model, for further investigation. While limited studies have documented the direct effect of drug-induced hypersensitivity from PGD interactions, studies have shown that the PGD metabolite levels in urine are associated with the severity of hypersensitive reactions to ingested foods (Maeda et al., 2017). The ATC code J01 is associated with antibacterials for systemic use. Currently, there are drugs within the J01 ATC category that have been associated with hypersensitivity reactions, including Penicillins (Weiss & Adkinson, 1988), Cephalosporins (Moreno et al., 2008), and Sulfonamides (Slatore & Tilles, 2004).

We subsequently evaluated the influence of drug target ATP1A1 (ATPase Na+/K+ transporting subunit alpha-1) which was the 10th lowest feature in the DT/ATC model, and ATC code H02, which was the 9th lowest feature in the DT/ATC/PathFX model, on the side effect of hypersensitivity. Since ATP1A1 is involved in ion transport, it is not specifically associated with modulating hypersensitivity reactions to drugs. As such, there is currently no known association between ATP1A1 and drug-induced hypersensitivity reactions. The ATC code H02 is associated with corticosteroids for systemic use. There are several drugs within the H02 ATC category that

have been used to treat hypersensitivity, including Methylprednisolone (Ocejo & Correa, 2019) and Dexamethasone (Johnson et al., 2018).

Hypersensitivity											
Drug Targets Only (DT model)				Drug Targets and ATC Codes (DT/ATC model)				Drug and PathFX Targets and ATC Codes (DT/ATC/PathFX model)			
Positive Targets	**Coef.**	**Negative Targets**	**Coef.**	**Positive Targets**	**Coef.**	**Negative Targets**	**Coef.**	**Positive Targets**	**Coef.**	**Negative Targets**	**Coef.**
DNMT1	1.16	CYP3A43	-1.19	V08	1.53	CYP3A43	-1.09	DPP4	1.30	folP	-1.32
DCK	1.15	SLC16A10	-1.17	HTR1D	1.37	ADRA2C	-1.09	J05	1.22	CYP3A43	-0.93
rpsI	1.01	CYP2B6	-1.15	G01	1.26	TEK	-0.96	A02	1.22	AR	-0.90
CHRNA3	1.01	IFNAR2	-0.96	ADORA2A	1.26	M09	-0.89	MPO	1.20	HTR1E	-0.87
MTOR	1.00	ABCC10	-0.92	A04	1.12	TNF	-0.87	M03	1.18	ampC	-0.83
UGT1A9	0.98	IDH1	-0.83	DNMT1	1.11	ABCC10	-0.82	V03	1.12	A12	-0.82
GNRHR	0.97	ADRA2C	-0.82	XDH	1.09	ALK	-0.81	L01	1.11	GNRHR2	-0.80
ADRB2	0.97	NTRK1	-0.81	TSPO	1.07	CACNA1G	-0.77	J02	1.08	KCND3	-0.79
HTR3A	0.92	ADRA1D	-0.80	ABCB11	1.07	HSD3B1	-0.75	J01	1.06	H02	-0.78
PGD	0.91	SULT2A1	-0.80	J01	1.06	ATP1A1	-0.73	J04	1.01	HTR6	-0.77

Table 6. Top and bottom 10 LR coefficients from the DT, DT/ATC, and DT/ATC/PathFX model for the side effect of hypersensitivity. Positive coefficients suggest that the feature is positively associated with the side effect. Conversely, negative coefficients imply the feature is a preventative against or may not affect the side effect.

3.13 Trend 3 case study: LR prediction of dermatitis is increased when level 2 ATC codes are incorporated with drug targets

Dermatitis LR prediction accuracy is increased when level 2 ATC codes are incorporated with drug targets. Interestingly, both PathFX proteins and ATC codes improve LR performance (accuracy = 0.68) compared to drug targets alone (accuracy = 0.66), but do not improve accuracy as much as level 2 ATC codes and drug targets (accuracy = 0.70) as shown in Figure 2. To better understand this trend, we extracted the top and bottom 30 LR coefficients from the DT, DT/ATC, and DT/ATC/PathFX models for this side effect (Table 7) and counted the number of non-drug target features. We discovered that 21/30 of the largest positive variables and 6/30 of the negative variables were level 2 ATC codes when both PathFX proteins and ATC level 2 codes were included. However, when PathFX network proteins were eliminated from the LR model, only 17/30 of the largest positive coefficients and 6/30 of the most negative coefficients were level 2 ATC codes. Given its high absolute coefficient values, our findings suggest that ATC class association is more associated with dermatitis than individual drug targets and network proteins.

We sought literature support for the drug target, Gonadotropin-releasing hormone receptor (GNRHR), and the level 2 ATC code D07, both of which had positive coefficients for predicting the side effect of dermatitis. We selected GNRHR as the drug target of interest because it had the highest positive coefficient value amongst all targets in the DT/ATC model. Further, we selected D07 because it was assigned the highest coefficient across all level ATC codes for predicting dermatitis. While there are currently limited studies that demonstrate the relationship between GNRHR and drugs on the side effect of dermatitis, Han et al. (2023) recently administered the GnRH antagonist Relugolix which revealed lichenoid dermatitis with eosinophils 9 weeks post-treatment. Relugolix has been demonstrated to lower testosterone levels fast (Shore et al., 2020).

This effect may increase the risk of developing dermatitis, as previous studies show that male atopic dermatitis patients have lower testosterone levels when compared to controls (Gratton et al. 2022). The ATC code D07 is associated with Corticosteroids for dermatological preparations. This class of drugs, including Hydrocortisone (Sears et al., 1997), Betamethasone (Jensen et al., 2009), and Clobetasol (Alam et al., 2013), have been used for dermatitis treatment. However, contact sensitivity to such drugs could lead to adverse effects, such as stasis dermatitis, perineal dermatitis, and chronic actinic dermatitis (Coondoo et al., 2014).

We further investigated the drug target, folP, and level 2 ATC code, R06, both of which had negative coefficients for the prediction of dermatitis. We selected folP because its coefficient values were consistently in the bottom 3 most negative values across the DT/ATC and DT/ATC/PathFX models. R06 had the largest negative coefficient in the DT/ATC model. Dapsone, an FDA-approved for dermatitis, has shown to reduce inflammation associated with dermatological conditions by competitively inhibiting the action of DHPS (Kurien et al., 2022). Consistent with the gastrointestinal disorder case study, this result contradicts the relationship which the LR model identified. The ATC code R06 is associated with Antihistamines. Currently, there are several antihistamines that have been found to be effective in improving dermatitis symptoms, including Cetirizine (Hannuksela et al., 1993), Loratadine (Herman & Vender, 2003), and Fexofenadine (Kawashima et al., 2003).

Dermatitis											
Drug Targets Only (DT model)				Drug Targets and ATC Codes (DT/ATC model)				Drug and PathFX Targets and ATC Codes (DT/ATC/PathFX model)			
Positive Targets	**Coef.**	**Negative Targets**	**Coef.**	**Positive Targets**	**Coef.**	**Negative Targets**	**Coef.**	**Positive Targets**	**Coef.**	**Negative Targets**	**Coef.**
AGTR1	1.38	folP	-1.07	D07	2.24	R06	-1.12	D07	2.38	folP	-1.30
DCK	1.35	SLC47A2	-1.01	GNRHR	1.47	SLC16A10	-1.11	N03	1.81	V04	-1.09
ORM1	1.32	CFTR	-0.95	N04	1.44	SLC10A1	-1.07	N02	1.56	SLC22A11	-0.98
ABCC4	1.14	ABCG2	-0.95	C09	1.43	folP	-1.04	M03	1.52	SLC18A2	-0.95
TSPO	1.13	PPARA	-0.92	A04	1.40	ABCC10	-1.00	B01	1.38	CYP3A43	-0.94
HTR3A	1.12	SLC16A10	-0.91	J02	1.38	TNF	-0.94	C09	1.32	S02	-0.92
UL30	1.09	SLC18A2	-0.90	C03	1.38	V04	-0.92	G02	1.23	R06	-0.91
MPO	1.09	PGR	-0.89	N03	1.35	CYP2B6	-0.90	L02	1.21	JAK2	-0.87
FDPS	1.07	SLC10A1	-0.89	G04	1.26	SLC18A2	-0.88	HTR2B	1.16	FXYD2	-0.86
PDE3A	1.03	IFNAR2	-0.89	A08	1.25	PPARA	-0.88	J05	1.15	L03	-0.86

Table 7. Top and bottom 10 LR coefficients from the DT, DT/ATC, DT/ATC/PathFX model for the side effect of dermatitis. Positive coefficients suggest that the feature is positively associated with the side effect. Conversely, negative coefficients imply the feature is a preventative against or may not affect the side effect.

4. Discussion

Side effects in FDA-approved drugs continue to be a major concern despite the strict guidelines and protocols in place during the drug development and approval process. These side effects can significantly impact the quality of life for its users. Recent advancements in the scientific community have sought to address these issues through the development of various tools and resources such as the PathFX algorithm, SIDER, and DrugBank databases. Specifically, the PathFX algorithm identifies potential connections between drugs, targets, and downstream proteins associated with a phenotype. SIDER documents known side effects of FDA-approved drugs based on their classification in the MedDRA. The DrugBank database assigns a standardized ID to all drugs and provides extensive information about each one of them, such as its associated ATC code, drug target, and description. These resources provide crucial information in enhancing our understanding of the relationships between drugs and side effects, thereby facilitating future developments of safe and effective drugs.

This project analyzed the predictive value of certain domain knowledge features for the prediction of the 30 most common side effects from SIDER. We used the DT model as a benchmark to evaluate the predictive value of three domain knowledge combinations 1) DT/PathFX 2) DT/ATC, and 3) DT/PathFX/ATC. Our results showed the following key observations based on the three trends identified: 1) incorporation of PathFX targets and network proteins resulted in improved prediction for side effects for 7 out of 30 side effects, 2) level 2 ATC codes enhanced LR model performance for prediction of all 30 side effects, and 3) despite the DT model performing worse than the DT/PathFX model, the DT/PathFX/ATC model did not substantially improve model performance compared to the DT/ATC model for LR prediction of dermatitis. Overall, these observations suggest the following: 1) while PPI networks can be useful

for the prediction of certain side effects, they can be more effectively engineered to improve the prediction of frequently occurring side effects, 2) drug classification information positively impacted the accuracy of side effect predictions, and 3) incorporation of both PathFX targets and level 2 ATC codes may not significantly influence the prediction accuracy compared to level 2 ATC codes alone for prediction of certain side effects. We further extracted the top and bottom 30 LR coefficients of three individual side effects from each identified trend to gain a better understanding of the features that our models prioritized. The LR model prioritized both drug targets and level 2 ATC codes, further implying that drug-target interactions and drug classification may play a significant role in influencing the occurrence of side effects. Since drugs within the same class share common characteristics in terms of their mechanism of action, chemical structure, or intended therapeutic use, they could also share the same targets as well. These shared properties between level 2 ATC codes and drug targets could potentially explain the similarities in their predictive power. However, the LR model did not prioritize PathFX targets and network proteins, which suggests that pathway information and protein interactions may be relatively less influential in predicting individual side effects.

Previous studies have explored the use of ATC codes and drug target information to predict side effects. Kim et al. (2016) analyzed the utility of drug off-targets in predicting side effects by identifying relationships in the tissue protein-symptom matrix. While this study leveraged drug target information to uncover off-target tissue effects, it does not directly address the predictive power of drug target information for the prediction of individual side effects. Further, Zhao et al. (2018) evaluated the predictive power of five domain knowledge features, namely drug targets, ATC code, structure similarity, literature association of drug-protein interactions, and drug fingerprint similarity for the prediction of drug side effects with four ML models. The RFC model

achieved the highest performance when all five domain knowledge features were integrated, yielding an accuracy of 0.775. Despite achieving a higher prediction accuracy through the integration of multiple domain knowledge features, Zhao et al. (2018) did not specifically aim to assess its utility in predicting individual side effects. Lastly, Huang et al. (2011) trained an LR model that combined drug target data, PPI networks, and gene ontology annotations for the prediction of side effects of experimental drugs and achieved an accuracy of 0.675 for the prediction of cardiotoxicity. However, the study's claim of predicting cardiotoxicity with experimental drugs may be limited. First, Huang et al. (2011) used drugs from SIDER, which primarily documents the side effects of FDA-approved drugs. Second, their study depended on molecular docking information, and they did not incorporate protein structural information in their model. Last, they only trained their model to predict one type of side effect: cardiotoxicity. Our trained LR models for prediction across 30 side effects achieved similar performance as reported by Huang et al. (2011). Interestingly, our trained DT/ATC model surpassed Huang et al. (2011) with an average performance of 0.70 across 30 side effects. Overall, these findings suggest that incorporating drug targets, PPI networks, and ATC codes for predicting drug side effects may be useful for the prediction of side effects, and leveraging more domain knowledge features may help further strengthen model performance.

Although previous studies have leveraged drug targets, ATC codes, and PPI networks for the prediction of side effects, limited studies have assessed the predictive value of ATC codes, drug target information, and PathFX targets and network proteins for predicting individual side effects. Consistent with our hypothesis, this study showed that LR model performance changes with the inclusion of domain knowledge for prediction across 30 side individual side effects. LR coefficient analyses further suggest that side effects may be more heavily influenced by drug target

and classification information. Understanding the large predictive value of drug targets and level 2 ATC codes for the prediction of drug side effects can help researchers modify or select drug candidates to minimize the risk of adverse reactions by considering the potential side effects associated with them, thereby enabling the development of safer and more effective therapeutic interventions. Furthermore, discerning the relationship between drug side effects and domain knowledge features can inform the development of personalized precision medicine, which can enable healthcare providers to make informed decisions about drug selection to minimize the risk of side effects. Ultimately, an enhanced understanding of the specific features that influence individual drug side effects can guide future research to elucidate the specific molecular mechanisms underlying these effects.

There are some limitations to our method. First, the hyperparameters of our model were not fine-tuned to enable optimal model performance. While this study aimed to understand the predictive value of domain knowledge features for the prediction of side effects, refining the hyperparameters to improve the performance of our model may lead to a more accurate representation of coefficient assignments in the LR model. The suboptimal accuracy of our model may have led to our model inaccurately assigning negative coefficients. For example, the inhibition of folP expression was shown to improve side effects of gastrointestinal disorder and dermatitis case studies despite the assignment of a negative coefficient by the LR model. Second, our study results cannot be generalized across the human population as it does not consider the genetic variation of individuals which may further influence the expression of side effects. Rather, this study identifies specific domain knowledge features that have an influence over individual side effects. Third, we only sought literature evidence for the LR coefficient association of two positive and negative features for three side effects across the identified trends. Therefore, the evidence

curated may have been influenced by chance. Additional investigations are needed to further validate the coefficient associations identified by the LR model to enhance the reliability and applicability of our results. Fourth, our trained LR model only considers three areas of domain knowledge (drug targets, PathFX targets, network proteins, and level 2 ATC codes), which may limit its performance potential. Of the five domain knowledge features in Zhao et al. (2018), the exclusion of drug targets and ATC codes had the least impact on the overall model. This suggests that the inclusion of additional domain knowledge, such as drug similarity, literature association of drug-protein interactions, and protein structural information, can potentially improve the performance of our model. Lastly, while PathFX targets and network proteins hold promise in predicting a subset of side effects (Wilson et al., 2022), this study shows it may not be most suitable to predict frequently occurring side effects, suggesting a potential area for future development. Since different PPI networks may be harnessed to predict certain side effects, other PPIs may be more effectively engineered to predict frequently occurring side effects.

Appendices

Appendix 1: Discussion of PathFX network protein GNRHR2

While PathFX network proteins were not assigned high coefficient values by the LR model, one notable exception was GNRHR2 (Gonadotropin-Releasing Hormone Receptor 2). This was the only PathFX network identified among the 30 most negative coefficients across the three case studies. There is currently limited evidence that suggests drug-GNRHR2 interactions directly protects individuals against such side effects. However, this finding may be worth further investigation given the complexity of GNRHR2 activity. GNRHR2 helps mediate the effects of LH release, which in turn affects testosterone production. As mentioned above, previous studies show that low testosterone levels are associated with male atopic dermatitis (Gratton et al. 2022). Furthermore, GNRHR2 may have additional roles in regulating the immune system and digestive system cells as well. According to Desaulniers et al. (2017), both GNRH2 and GNRHR2 were found to be produced in both organs associated with the digestive system (i.e., stomach, small intestine, and large intestine) and immune system (i.e., spleen and bone arrow) in humans. Given that gastrointestinal disorders are predominantly linked to the digestive system and hypersensitivity reactions are primarily associated with the immune system, future research holds promise in elucidating the relationship between drug-GNRHR2 interactions and such side effects. It is worth noting that our study only evaluated the LR coefficients of three side effects. Therefore, assessing the strength of such PathFX network protein across the 30 side effects may provide a more comprehensive understanding of its overall impact, which may be a future direction for our research.

Appendix 2: Supplementary Files

Thesis files can be accessed through the SL Thesis Files Google Drive Folder Link.

Folder 1: datasets

Folder description: this folder contains all the datasets used to address the thesis question. Files are bolded.

1. **Pfx050120_dint.pkl**: consists of a pickled dictionary of all drugbank drugs with its associated targets
2. **meddra_all_se.tsv**: contains all drug to side effect combinations curated in SIDER 4.1 based on MedDRA classification
3. **drug_names.tsv**: dataset containing all drug names and its unique drug ID in SIDER 4.1
4. **Drugbank050120.xlsx**: includes drug names and its associated DBID documented in DrugBank 5.1.6 along with its drug type (i.e., biotech, small molecule), group (i.e., approved, investigational), ATCcodes, categories, and description
5. **drugbank_vocabulary.csv**: a curated list of all common names and synonyms associated with drugs in DrugBank

Folder 2: results

Folder description: this folder contains all the results generated. Files are bolded. Excel tab name and description listed under each file name.

1. **Characteristics**:
    a. data characteristics: contains 1) total count of side effects, 2) number of drugs associated with side effect, 3) number of drugs matched to DBID, 4) DBID match %, 5) matrix match count and 6) matrix match %

b. LR vs RFC: comparison of LR and RFC model prediction across 30 side effects with 100x bootstrap with random undersample of negative cases
c. average accuracy across 5 conditions: average accuracy of model performance for prediction of 30 side effects in 1) ATC model, 2) DT model, 3) DT/PathFX model, 4) DT/ATC model, and 5) DT/PathFX/ATC model
d. ANOVA-RM: repeated measures ANOVA across the 5 conditions sorted based on the highest to lowest prediction accuracy from the DT model

2. **LR_Coefficients**:
    a. experimental drug target coefficients: count of the 10 most common unapproved drug target with its assigned LR coefficients across 30 side effects trained on the unfiltered dataset
    b. gastrointestinal disorder: top and bottom 30 LR coefficients for this side effect
    c. hypersensitivity: top and bottom 30 LR coefficients for this side effect
    d. dermatitis: top and bottom 30 LR coefficients for this side effect
3. **all_ttest_results**:
    a. DT vs DT/PathFX: paired t-test table comparing the performance of DT vs DT/PathFX model
    b. DT/ATC vs DT/ATC/PathFX: paired t-test table comparing the performance of DT/ATC vs DT/ATC/PathFX model
    c. ATC effects: 3 paired t-test tables comparing the performance of: 1) ATC vs DT model, 2) DT vs DT/ATC model, and 3) DT/PathFX vs DT/ATC/PathFX model

File 1: MS_Thesis_Code.ipynb

File description: the Python script that I wrote to execute my thesis project.

References

Alam, M. S., Ali, M. S., Alam, N., Siddiqui, M. R., Shamim, M., & Safhi, M. M. (2013). In vivo study of clobetasol propionate loaded nanoemulsion for topical application in psoriasis and atopic dermatitis. *Drug invention today*, 5(1), 8-12.

Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, *321*(5886), 263-266.

Chen, R., Wang, J., Tang, S., Zhang, Y., Lv, X., Wu, S., ... & Zhan, S. (2016). Role of polymorphic bile salt export pump (BSEP, ABCB11) transporters in anti-tuberculosis drug-induced liver injury in a Chinese cohort. *Scientific reports*, 6(1), 1-7.

Coondoo, A., Phiske, M., Verma, S., & Lahiri, K. (2014). Side-effects of topical steroids: A long overdue revisit. *Indian dermatology online journal*, 5(4), 416.

Desaulniers, A. T., Cederberg, R. A., Lents, C. A., & White, B. R. (2017). Expression and role of gonadotropin-releasing hormone 2 and its receptor in mammals. *Frontiers in endocrinology*, 269.

Force, T., & Kolaja, K. L. (2011). Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes. *Nature reviews Drug discovery*, *10*(2), 111-126.

Gratton, R., Del Vecchio, C., Zupin, L., & Crovella, S. (2022). Unraveling the role of sex hormones on keratinocyte functions in human inflammatory skin diseases. *International journal of molecular sciences*, 23(6), 3132.

Han, J., Baek, P., Poplausky, D., Agarwal, A., Young, J. N., Mubasher, A., ... & Gulati, N. (2023). A Case of Lichenoid Drug Eruption Associated with Relugolix. *JAAD Case Reports*.

Hannuksela, M., Kalimo, K., Lammintausta, K., Mattila, T., Turjanmaa, K., Varjonen, E., & Coulie, P. J. (1993). Dose ranging study: cetirizine in the treatment of atopic dermatitis in adults. *Annals of allergy*, 70(2), 127-133.

Herman, S. M., & Vender, R. B. (2003). Antihistamines in the treatment of atopic dermatitis. *Journal of cutaneous medicine and surgery*, 7(6), 467-473.

Huang, L. C., Wu, X., & Chen, J. Y. (2011). Predicting adverse side effects of drugs. *BMC genomics*, *12*(5), 1-10.

Jensen, J. M., Pfeiffer, S., Witt, M., Bräutigam, M., Neumann, C., Weichenthal, M., ... & Proksch, E. (2009). Different effects of pimecrolimus and betamethasone on the skin barrier in patients with atopic dermatitis. *Journal of allergy and clinical immunology*, 124(3), R19-R28.

Ji, Y., Chen, S., Wang, Q., Xiang, B., Xu, Z., Zhong, L., ... & Qiu, L. (2018). Intolerable side effects during propranolol therapy for infantile hemangioma: frequency, risk factors and management. *Scientific reports*, *8*(1), 4264.

Johnson, D. B., Lopez, M. J., & Kelley, B. (2018). Dexamethasone.

Kawashima, M., Tango, T., Noguchi, T., Inagi, M., Nakagawa, H., & Harada, S. (2003). Addition of fexofenadine to a topical corticosteroid reduces the pruritus associated with atopic dermatitis in a 1-week randomized, multicentre, double-blind, placebo-controlled, parallel-group study. *British Journal of Dermatology,* 148(6), 1212-1221.

Kim, D., Lee, J., Lee, S., Park, J., & Lee, D. (2016). Predicting unintended effects of drugs based on off-target tissue effects. *Biochemical and biophysical research communications,* 469(3), 399-404.

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic acids research*, *44*(D1), D1075-D1079.

Kurien, G., Jamil, R. T., & Preuss, C. V. (2022). Dapsone. In StatPearls [Internet]. *StatPearls Publishing.*

Kurta, A., Cazeau, C., Dai, D., & Siegfried, E. (2018). Prescribing propranolol for hemangioma of infancy: assessment of dosing errors.

LaBute, M. X., Zhang, X., Lenderman, J., Bennion, B. J., Wong, S. E., & Lightstone, F. C. (2014). Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PloS one*, *9*(9), e106298.

Liang, H., Chen, L., Zhao, X., & Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Computational and Mathematical Methods in Medicine*, 2020, 1-16.

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. L., ... & Sheltzer, J. M. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science translational medicine*, *11*(509), eaaw8412.

Maeda, S., Nakamura, T., Harada, H., Tachibana, Y., Aritake, K., Shimosawa, T., ... & Murata, T. (2017). Prostaglandin D2 metabolite in urine is an index of food allergy. *Scientific Reports*, 7(1), 1-8.

Moreno, E., Macías, E., Dávila, I., Laffond, E., Ruiz, A., & Lorente, F. (2008). Hypersensitivity reactions to cephalosporins. *Expert opinion on drug safety*, 7(3), 295-304.

Ocejo, A., & Correa, R. (2019). Methylprednisolone.

Sears, H. W., Bailer, J. W., & Yeadon, A. (1997). Efficacy and safety of hydrocortisone buteprate 0.1% cream in patients with atopic dermatitis. *Clinical therapeutics*, 19(4), 710-719.

Shore, N. D., Saad, F., Cookson, M. S., George, D. J., Saltzstein, D. R., Tutrone, R., ... & Tombal, B. (2020). Oral relugolix for androgen-deprivation therapy in advanced prostate cancer. *New England Journal of Medicine*, 382(23), 2187-2196.

Siavash, M., Tabbakhian, M., Sabzghabaee, A. M., & Razavi, N. (2017). Severity of gastrointestinal side effects of metformin tablet compared to metformin capsule in type 2 diabetes mellitus patients. *Journal of research in pharmacy practice*, 6(2), 73.

Slatore, C. G., & Tilles, S. A. (2004). Sulfonamide hypersensitivity. *Immunology and Allergy Clinics*, 24(3), 477-490.

Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it?. *Acta Pharmaceutica Sinica B*.

Xie, L., Li, J., Xie, L., & Bourne, P. E. (2009). Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors. *PLoS computational biology*, 5(5), e1000387.

Weiss, M. E., & Adkinson, N. F. (1988). Immediate hypersensitivity reactions to penicillin and related antibiotics. *Clinical & Experimental Allergy*, 18(6), 515-540.

Wilson, J. L., Gravina, A., & Grimes, K. (2022). From random to predictive: a context-specific interaction framework improves selection of drug protein–protein interactions for unknown drug pathways. *Integrative Biology*, *14*(1), 13-24.

Wilson, J. L., Racz, R., Liu, T., Adeniyi, O., Sun, J., Ramamoorthy, A., ... & Altman, R. (2018). PathFX provides mechanistic insights into drug efficacy and safety for regulatory review and therapeutic development. *PLoS computational biology*, *14*(12), e1006614.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, *34*(suppl_1), D668-D672.

Yoshida, M., Nambu, R., Yasuda, R., Sakaguchi, H., Hara, T., Iwama, I., & Mizuochi, T. (2022). Dapsone for Refractory Gastrointestinal Symptoms in Children With Immunoglobulin A Vasculitis. *Pediatrics*, 150(3).

Zhao, X., Chen, L., & Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Mathematical biosciences*, 306, 136-144.