

UCLA

Library Prize for Undergraduate Research

Title

Evaluating and Optimizing Machine Learning Models to Predict Drug Side Effects

Permalink

<https://escholarship.org/uc/item/7343k6z4>

Authors

Koduvayur, Vyas

Jabourian, Jenna

Zhang, Flicka

et al.

Publication Date

2025-11-05

Undergraduate

EVALUATING AND OPTIMIZING MACHINE LEARNING MODELS TO PREDICT DRUG SIDE EFFECTS

Vyas Koduvayur¹, Jenna Jabourian¹, Melody Mao¹, Flicka Zhang¹, Abeni Liu¹

¹*Department of Computational and Systems Biology, University of California, Los Angeles*

Keywords: Machine Learning, Pharmacovigilance, Adverse Drug Reactions, Predictive Modeling, Model Optimization, Feature Selection, Regularization

ABSTRACT

Predicting drug side effects is crucial for drug safety and development. In our study, we extend the work of previous literature that utilized machine learning (ML) approaches that integrate domain knowledge features—drug targets (DT), Anatomical Therapeutic Chemical (ATC) classification codes, and PathFX protein-protein interaction networks—to enhance side effect prediction. Building on a prior study that identified logistic regression (LR) as the best-performing model, we investigate two key extensions: (1) evaluating the impact of higher-resolution ATC levels (levels 3-5) on predictive accuracy and (2) exploring how to optimize the LR model for improved performance.

Our approach involves replicating the original study’s workflow by conducting quality control. We then test higher levels of ATC codes and modify model parameters. Results show that ATC levels 2, 3, and 4 have comparable performance, while level 5 underperforms substantially, likely due to overfitting. Additionally, fine-tuning the logistic regression model, such as making decision threshold adjustments and optimizing regularization types, showed promise regarding accuracy improvement.

Our findings ultimately highlight a tradeoff between feature granularity and model generalizability. By demonstrating how careful feature selection and parameter tuning can enhance drug side effect prediction, our study emphasizes the importance of balancing specificity and model robustness in ML-based approaches to advance drug side effect prediction. Our findings can ultimately inform future efforts that integrate computational approaches into pharmacological research.

INTRODUCTION

Drug side effect prediction is a critical component of pharmacology and drug development, as unforeseen toxicity and intolerable side effects continue to be some of the primary causes of drug failure and drug withdrawal once reaching the market [1]. For instance, retinoids, a class of medication that is used to treat severe recalcitrant nodular acne and also induce remission of leukemia [2], are associated with causing meibomian gland dysfunction, ultimately leading to eye discomfort and acute transient refractive changes [3].

Traditional methods for identifying side effects rely on preclinical studies, clinical trials, and post-market surveillance [4]. However, these approaches have limitations,

such as small sample sizes, lack of accrual, and time-consuming regulatory processes [5]. As a result, machine learning (ML) has emerged as a powerful tool to address these challenges and predict individual drug side effects by leveraging large-scale datasets to identify patterns that may not be apparent through conventional methods [6,7]. ML-based approaches hold great potential in advancing drug safety assessment, but previous studies have faced significant hurdles, especially due to a lack of standardized datasets that classify side effects based on the clinical trial stage at which they were identified [7]. This results in limitations in interpreting results due to data sparsity, and ultimately, makes it difficult to fully analyze adverse effects.

To address this issue, researchers in the scientific community created the Side Effect Resource (SIDER) database, which uses Natural Language Processing to document results from free public sources [8]. SIDER integrates data on drugs, targets, and side effects into a powerful dataset that can be utilized to better predict adverse effects. Additionally, another group of researchers compiled DrugBank, a data resource that contains comprehensive drug target information with over 4000 drug entries and over 14,000 target sequences associated with those drugs [9]. The database also includes domain knowledge for each drug, including its Anatomical Therapeutic Chemical (ATC) classification as well as its drug development group status, offering an alternative feature representation by categorizing drugs based on both therapeutic use and chemical properties [10]. This structured classification could strengthen the predictive power of ML models by integrating pharmacological context into side effect prediction.

Such models can be especially powerful when enriched with a combination of domain knowledge features such as drug targets, drug classifications, and protein-protein interaction (PPI) networks. For instance, a previous study that incorporated several domain knowledge features—including drug targets, PPI networks, and ATC codes—into an ML model achieved a high performance with 0.975 accuracy [11]. These successes suggest that it may be useful to evaluate the relative predictive power of the various domain knowledge features.

In this project, we build upon a study that utilized the SIDER database, drug targets (DT), ATC codes, and PathFX (a pathway modeling tool developed by the researchers that can identify drug-induced phenotypic effects) to predict the top 30 most frequently appearing drug-induced side effects [12]. The study evaluated five different domain knowledge feature sets utilizing several ML models—including the Random Forest Classifier and Logistic Regression (LR) models—and sought to identify the most optimal combinations for accurate side effect prediction [13]. Their principal conclusion was that LR models using level 2 ATC codes combined with DT performed most optimally for a majority of the side effects, although there were some exceptions. Their conclusion serves to reinforce the importance of using domain expertise in prediction models.

However, several key areas remain unexplored. For instance, although the study predominantly focused on analyzing the predictive value of level 2 ATC codes, the use of higher-resolution ATC levels (levels 3-5) could improve the model's ability to differentiate between drugs with similar broad classifications but distinct pharmacological effects [14]. Additionally,

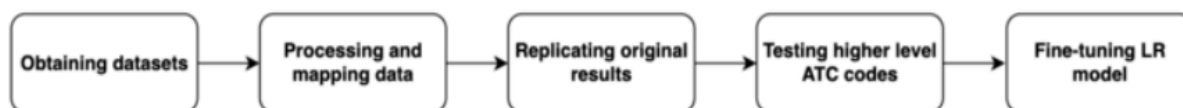


Figure 1: Summary of Workflow. After SIDER database, DrugBank dataset, and PathFx Pkl files are obtained, drug names are mapped to DrugBank ID and the top 30 most common side effects are extracted. Logistic regression (LR) model is tested across all ATC levels and results are analyzed. Afterwards, LR model's parameters are fine tuned and tested with best performing ATC and DT feature sets together.

the default parameters for their LR model were not optimized for this specific task, suggesting an avenue to adjust certain parameters and utilize additional data to enhance prediction performance by balancing sensitivity and specificity.

This study expands on the original research by addressing two key questions: (1) How does increasing ATC code granularity (from level 2 to 5) affect predictive power? (2) Can the parameters of the Logistic Regression model (LR) be fine-tuned for better prediction accuracy? We hypothesize that adjusting model parameters (*e.g.*, decision thresholds and regularization strength) can calibrate a sensitivity-specificity trade-off for better accuracy [15,16]. Additionally, we expect that higher-level ATC codes will provide more informative drug information, potentially improving predictive power but also increasing the risk of overfitting [17].

To address these questions, we replicate the original study, explore higher levels of ATC codes, and refine the Logistic Regression model to identify the optimal feature selection and model parameters for drug side effect prediction, with the goal of increasing the model's predictive power and ultimately working to improve drug safety and development.

METHODS

At a high level, our project involved obtaining the necessary datasets, mapping drug names to their drug bank IDs, replicating the original results, testing higher level ATC codes, and finally, fine-tuning the model. The workflow process is summarized above (**Figure 1**).

DrugBank ID	Name	Type	Group(s)	ATC Codes
DB00001	Lepirudin	biotech	approved	B01AE02
DB00002	Cetuximab	biotech	approved	L01XC06
DB00003	Dornase alfa	biotech	approved	R05CB13
DB00004	Denileukin diftitox	biotech	approved investigational	L01XX29
DB00005	Etanercept	biotech	approved investigational	L04AB01
DB00006	Bivalirudin	small molecule	approved investigational	B01AE06
DB00007	Leuprolide	small molecule	approved investigational	L02AE51 L02AE02
DB00008	Peginterferon alfa-2a	biotech	approved investigational	L03AB11 L03AB61
DB00009	Alteplase	biotech	approved	B01AD02 S01XA13
DB00010	Sermorelin	biotech	approved withdrawn	V04CD03 H01AC04

Table 1: Overview of drugbank.tsv (Mapped and Processed data). Each column represents a different drug, including information about its technical name, DrugBank ID, type, group, and ATC Code. Level 2 represents the first three characters (i.e., “B01” for Lepirudin). Level 3 represents the first four characters, Level 4 represents the first five characters, and Level 5 includes the entire ATC Code. The first 100 entries can be found in the supplementary data.

Obtaining Datasets and Replicating Initial Study

The first step of our project was to obtain the main code (written in Python) and datasets from the lab’s GitHub repository. Datasets from SIDER included the Medical Dictionary for Regulatory Activities (MedDRA) file containing side effects of Food and Drug Administration approved drugs. A file called “drug_names.tsv” was also included and contained all drug names and their unique SIDER drug IDs. Several DrugBank (DB) datasets were included in the repository as well. The file “drugbank_vocabulary.csv” contained common names and synonyms of drugs along with their associated DrugBank identifier (DBID). To access the ATC codes associated with DBIDs, access to DrugBank release version 5.1.6 was requested and approved for this project [10]. Included in the repository was a PathFX file containing a pickled dictionary of all DB drugs and associated targets. Lastly, several Excel files were included with a summary of the original study’s results.

Once all datasets were obtained, the next step was to process and map the data. First,

SIDER drug names were mapped to DBIDs to match drug names. Drug names were standardized to a single ID to ensure accurate mapping. A subset of the file with mapped drug names is displayed in tabular form (**Table 1**). From SIDER, the 30 most common side effects and their associated drug names were extracted.

After standardization and mapping were completed, the third step was to replicate the original results using the best-performing model (LR). To fully highlight the effects of using level 2 ATC codes (first three characters in ATC code) and DT, PathFX feature sets were omitted. Replication involved finding the LR mean accuracy for three feature combinations: (1) DT only, (2) ATC codes only, and (3) ATC codes with DT. To compare the performance of the LR model for these three feature sets, an analysis of variance (ANOVA) was conducted along with paired T-Tests to compare ATC codes only with ATC codes with DT.

Increasing Level of ATC Codes

The first main extension for this project was to perform a similar analysis using higher

level ATC codes, particularly levels 3, 4, and 5, by increasing the number of characters included in the whole ATC code. For all additional ATC code levels, accuracy comparisons were made using the ATC-only model. ANOVA was used to compare the performance of each extended ATC code level (levels 2-5). For deeper analysis, a pairwise comparison was conducted on each ATC level to determine the effects of individual ATC levels. Based on this pairwise comparison, the two ATC code levels with the most notable results, levels 2 and 4, were compared using a T-test for (1) level 2 ATC-only versus level 4 ATC-only and (2) DT with level 2 ATC versus DT with level 4 ATC. The use of a T-test was justified after performing a Shapiro-Wilk Test to prove normality and a Levene's Test to prove equal variance. Additionally, the mean difference in prediction accuracy for ATC only and ATC with DT for the top 30 side effects was compared for level 2 and level 4 ATC codes. This extension provided a deeper exploration into differences associated with ATC level choice and specific side effects.

Fine-tuning LR Model Parameters

Since the original study used default parameters for the LR model, the final phase of this project focused on fine-tuning these parameters to improve model accuracy. The first adjustment made was to the decision threshold of the model due to its ability to improve sensitivity and specificity. By default, a threshold of 0.5 classifies probabilities below 0.5 as "no side effect" and those above 0.5 as "side effect occurrence". To determine the most optimal decision threshold, we maximized Youden's J Statistic, calculated as the difference between True Positive Rate (TPR) and False Positive Rate (FPR). This approach ultimately finds the best balance between false positives and false negatives,

contributing to improving the overall accuracy. TPR and FPR values were obtained using the Receiver Operating Characteristic (ROC) curve.

Additionally, we adjusted the regularization strength (C) parameter of the LR model, which has a default value of 1. This parameter controls how the model penalizes large coefficients. A higher C value corresponds to less regularization and a more complex model, with the risk of overfitting, while a lower C value corresponds to more regularization and a simpler model, with the risk of underfitting [18]. To automate the optimization of this parameter, Grid Search, an exhaustive hyperparameter tuning method that evaluates multiple values to identify the best-performing one, was used. As higher C values can make it harder for the model to converge, the maximum iterations parameter was also increased from 100 to 300 to accommodate longer training times.

Improvement in model performance was assessed using the DT-only feature set. Once the parameter modifications were determined to correspond with an improvement in the accuracy of predictions, the ultimate prediction model for drug side effects was formulated. This involved the use of the fine-tuned LR model applied to the DT/ATC feature combination, which was determined in the original paper to demonstrate the highest performance. Level 4 ATC codes were used as they were determined previously in our first extension to perform generally on par with, if not better than, level 2 (used in the original study) for all top 30 side effects.

Code and Figures

Data mapping and model evaluation code were run in Python, and figures were generated with R.

Side Effect	LR: ATC Level 2	LR: ATC Level 3	LR: ATC Level 4	LR: ATC Level 5	F-Value	P-value
dizziness	0.675898305	0.66779661	0.679627119	0.539118644	1168.720528	4.73E-164
nausea	0.683143713	0.668892216	0.683263473	0.556317365	1232.911303	3.09E-167
headache	0.665696203	0.680221519	0.684778481	0.570506329	728.4721486	1.50E-136
rash	0.664735099	0.670496689	0.674503311	0.565496689	798.0560486	9.38E-142
vomiting	0.696447368	0.673815789	0.663223684	0.526282895	1532.945956	2.46E-180
...						
Average	0.6771841376	0.6722445646	0.6770792136	0.5515443844		

Table 2: Logistic Regression (LR) Model Performance Across Anatomical Therapeutic Chemical (ATC) Code Levels 2-5 (only) for Top 30 Side Effects. Model performance varies depending on the side effect, with level 2 and 3 having nearly identical results and level 4 performing marginally better for some side effects and marginally worse for others. Model using level 5 codes underperforms drastically. ANOVA test shows significant differences across all side effects. Full results can be found in the supplementary data.

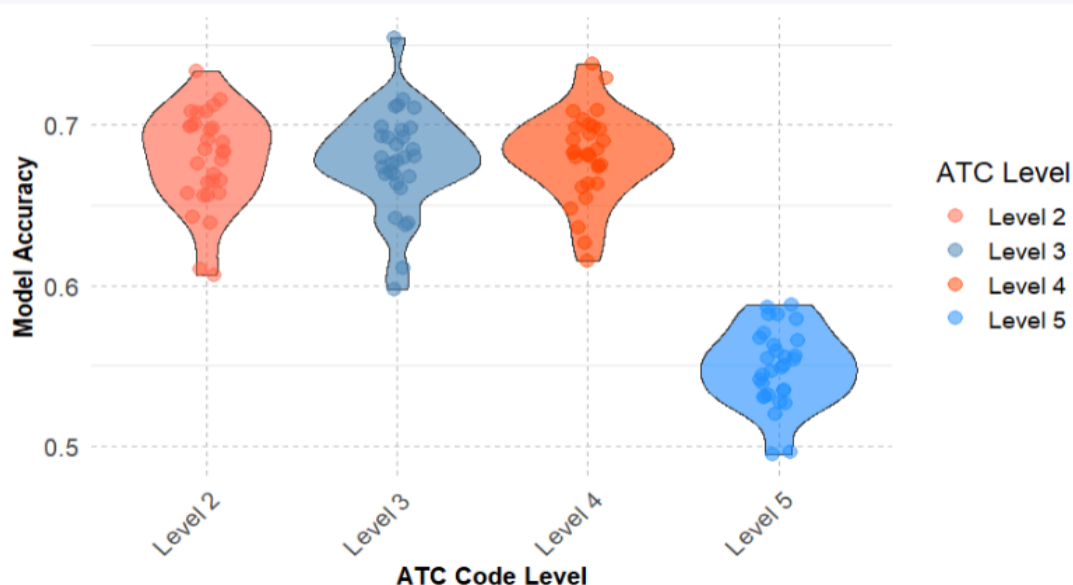


Figure 2: Logistic Regression (LR) Model Performance Across Anatomical Therapeutic Chemical (ATC) Code Levels 2-5 (only). Violin and jitter plot of model performance distribution across the top 30 side effects shows level 2, 3, and 4 perform with comparable accuracy, while level 5 substantially underperforms.

RESULTS

LR Models with Level 2, 3, 4, and 5 ATC Codes

The model performance by ATC levels was analyzed using ANOVA, with all thirty side

effects indicating significant differences across models. The impact of higher ATC code levels (levels 3-5) on predictive accuracy was recorded in tabular form (**Table 2**). Additionally, models were

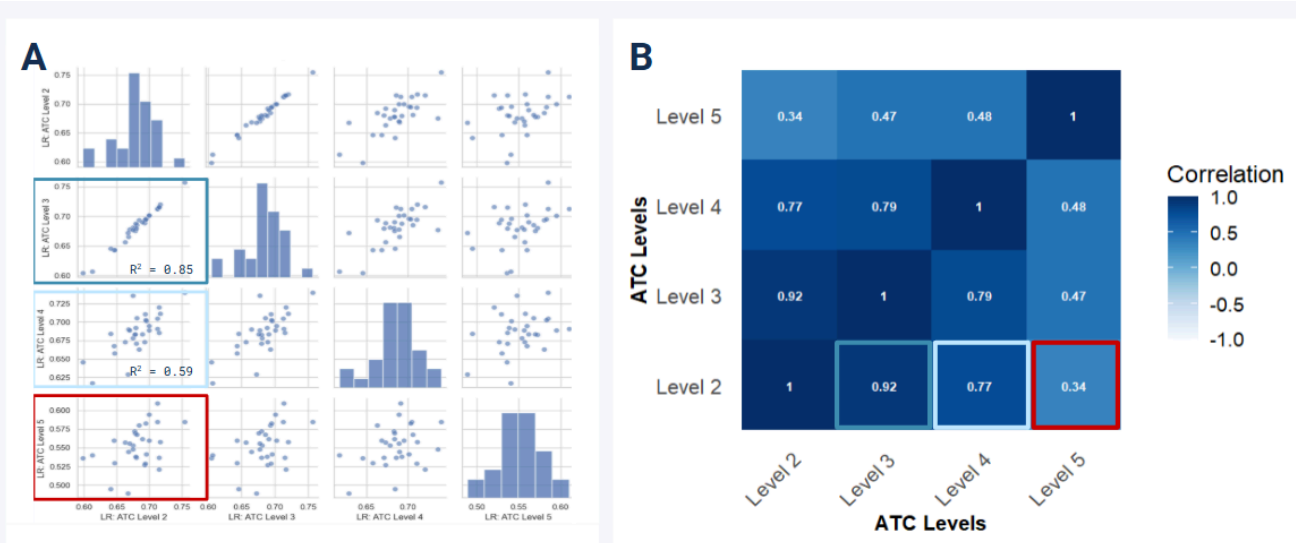


Figure 3: Pairwise Comparison of Logistic Regression (LR) Model Performance Across Anatomical Therapeutic Chemical (ATC) Code Levels 2-5 (only). (A) Pairwise scatterplots and histograms showing model performance distribution across ATC levels. (B) Heatmap of correlation matrix displaying the correlation coefficient (r). ATC levels 2 and 3 have a strong linear relationship and are highly correlated. ATC level 5 has a nonlinear relationship with other levels, indicating significant differences. Level 4 is correlated with level 2 and 3, albeit with a weaker r value.

analyzed using multiple visualizations. Violin plots showed that ATC levels 2, 3, and 4 were found to have comparable performance, whereas level 5 underperforms greatly (**Figure 2**).

Next, pairwise comparisons between each ATC model were visualized in scatter plot form and a correlation heatmap. Comparisons showed a strong linear correlation between levels 2 and 3. Level 5 exhibited a weak or non-linear relationship with all other levels, suggesting a significant difference in performance. Level 4 was found to have a linear, albeit weaker, relationship with level 2 and level 3 (**Figure 3A**). A correlation heat map further confirmed that levels 2 and 3 were very similar (correlation coefficient of 0.92), whereas level 5 had weak correlations with all other levels (**Figure 3B**). Level 4 and level 2 had a correlation of 0.77, indicating a moderately strong linear relationship but

also suggesting some differences between the models, prompting a need for further analysis comparing level 4 to level 2.

When investigating this relationship further, a mean difference bar chart showed that level 4 performed significantly better than level 2 for several side effects, particularly musculoskeletal discomfort, anaphylactic shock, and infection, which had a $>2\%$ difference (**Figure 4A**). Likewise, level 2 performed better for a little less than half of the top 30 side effects, particularly tachycardia, vomiting, and hypertension, with a mean difference $>2\%$.

Next, side-by-side box plots comparing the distributions of model accuracy for level 2 and level 4 were also generated, but the mean values were very similar. A T-test found no significant difference between the two (**Figure 4B**). We repeated the analysis by reintroducing DT along with ATC codes,

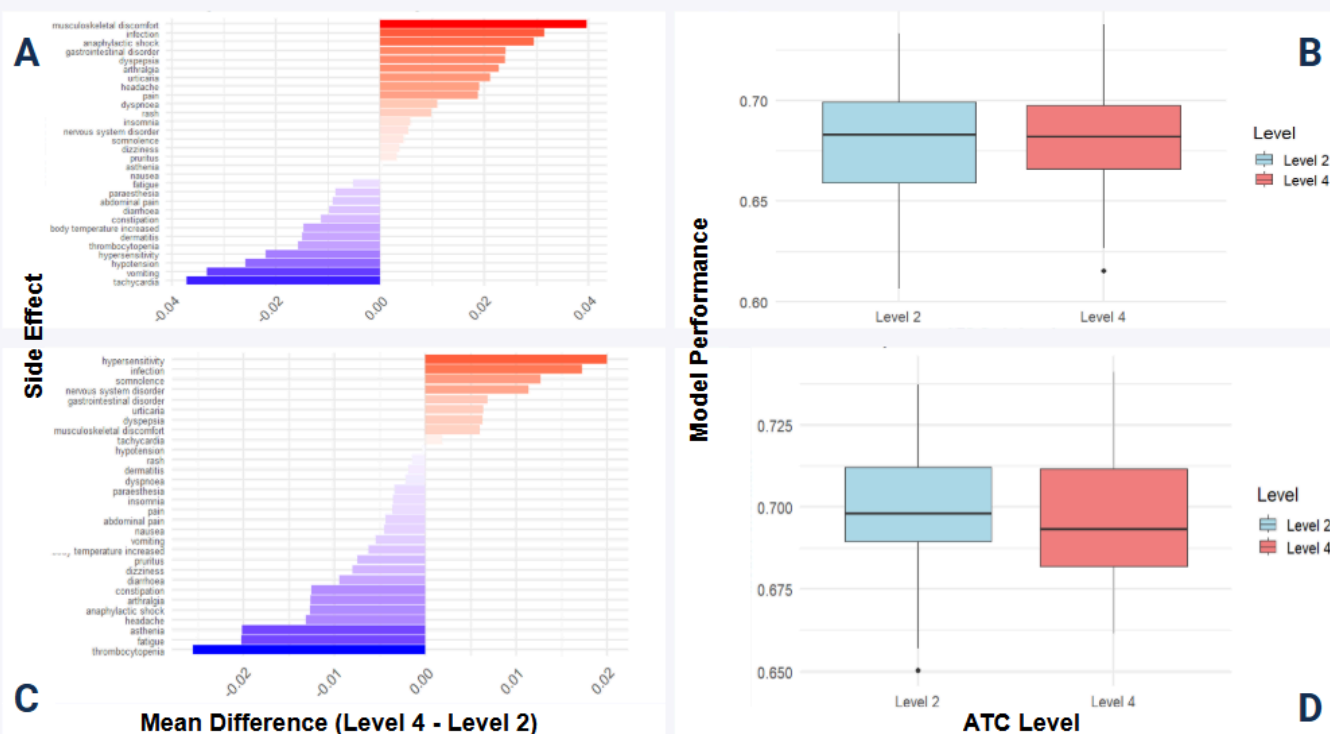


Figure 4: Comparing Logistic Regression (LR) Model Performance Across Anatomical Therapeutic Chemical (ATC) Code Levels 2 and 4 only and for ATC + Drug Targets (DT). (A) Mean difference (level 4 - level 2) plot, with red bars indicating side effects where level 4 performs better and blue bars indicating side effects where level 2 performs better. (B) Side by side boxplot comparing distributions of ATC level 2 v. 4; T-test p-value = 0.7618, indicating no significant difference. (C) Mean difference plot of DT + level 4 and DT + level 2. (D) Side by side boxplot of DT + level 4 and DT + level 2; p = 0.5798, also indicating no significant difference.

as this feature combination was shown to perform best in the study done by Liu *et al.* The mean difference plot showed that when coupled with DT, LR models using level 2 ATC codes seem to perform better for a greater number of side effects (**Figure 4C**), but a side-by-side boxplot and another T-test reaffirm that this difference isn't significant (**Figure 4D**). Overall, the results indicate that models using level 2, 3, and 4 ATC codes have comparable performance patterns when explaining the top 30 side effects, while the model with level 5 codes vastly underperform.

Optimizing LR Model Shows Improved Performance

The performance of parameter adjustments in the LR model was initially benchmarked

with the DT-only model. Observing the difference in mean accuracy between the original and tuned LR model (**Figure 5**) shows that the mean accuracy of the tuned LR model increased across all 30 side effects, with the highest improvements observed for infection (0.0352), nervous system disorder (0.0338), and insomnia (0.0291). On average, the accuracy of the fine-tuned LR model improved by 0.0186. A T-test comparing the tuned and untuned DT-only model confirmed that the difference was statistically significant ($p = 0.000535$).

Next, the fine-tuned LR model was applied to the best-performing domain knowledge feature combination (DT/ATC), incorporating level 4 ATC codes to create the highest-performing model. This model

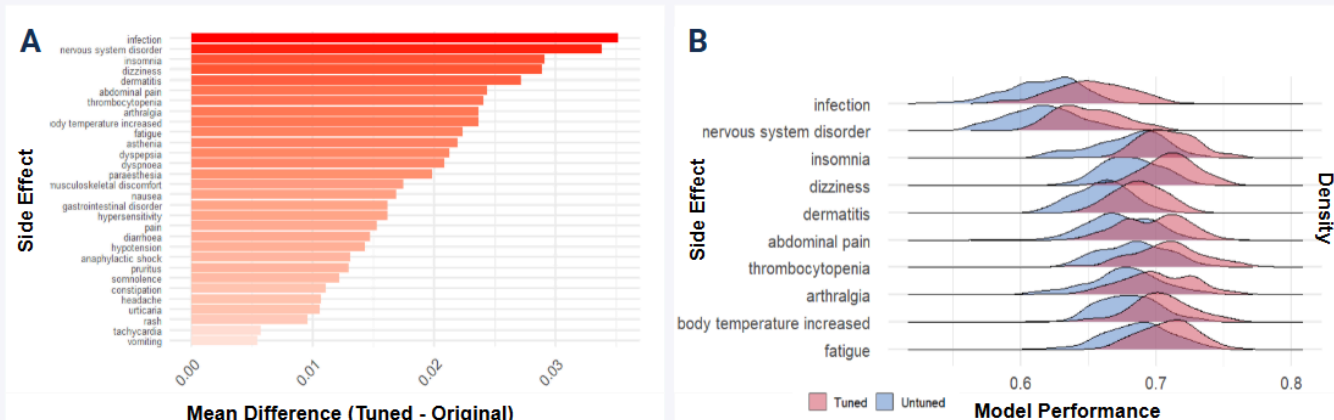


Figure 5: Performance Comparison of Tuned (Maximized J and Optimal C value) Logistic Regression (LR) Model and Original Model for Drug Targets (DT) only. (A) Mean difference plot (tuned - original) shows fine tuned model outperforms original model for all top 30 side effects up to around 3%. (B) Density distribution of top 10 side effects (mean difference) for tuned v. untuned model.

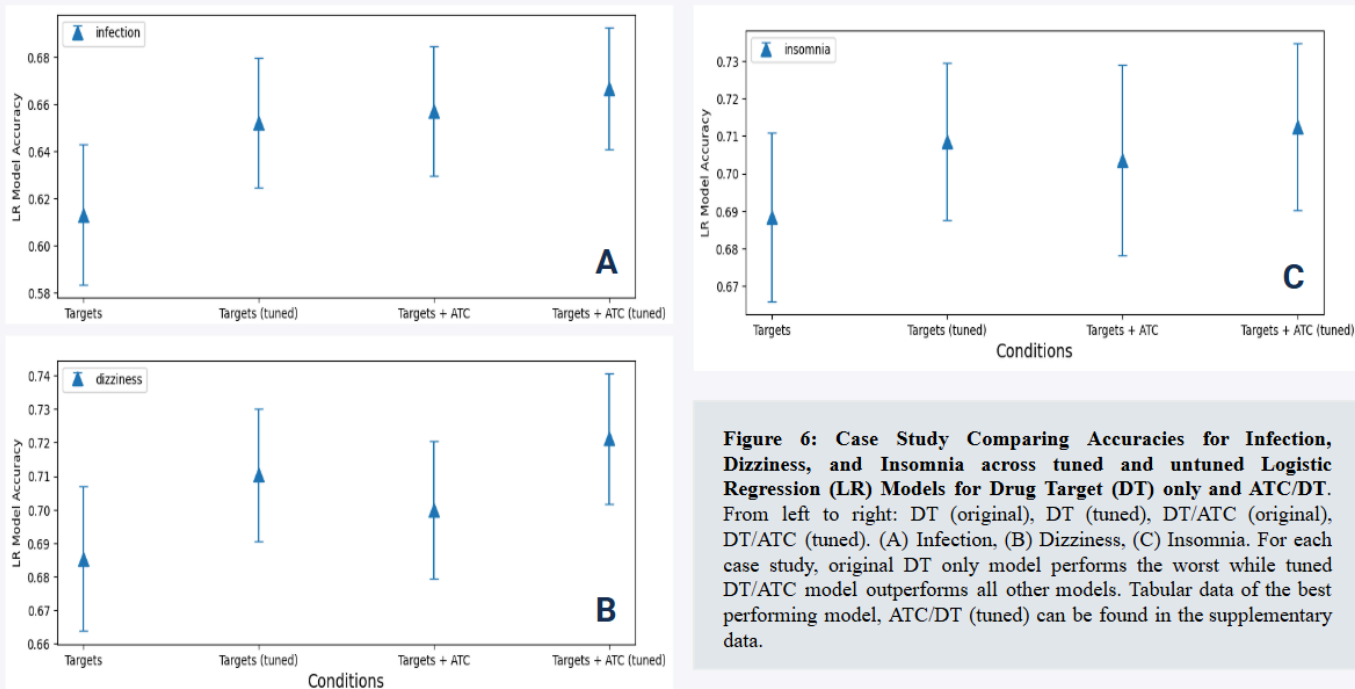


Figure 6: Case Study Comparing Accuracies for Infection, Dizziness, and Insomnia across tuned and untuned Logistic Regression (LR) Models for Drug Target (DT) only and ATC/DT. From left to right: DT (original), DT (tuned), DT/ATC (original), DT/ATC (tuned). (A) Infection, (B) Dizziness, (C) Insomnia. For each case study, original DT only model performs the worst while tuned DT/ATC model outperforms all other models. Tabular data of the best performing model, ATC/DT (tuned) can be found in the supplementary data.

achieved a mean accuracy of 0.7016, with the highest accuracies observed for thrombocytopenia (0.7316), urticaria (0.7263), and somnolence (0.7262). The full

results are shown in the supplementary data. The mean decision threshold used was 0.46, indicating that a slightly more stringent decision threshold optimized model

performance. The distribution of C values favored lower values, suggesting that higher regularization was beneficial to prevent overfitting.

Looking across three specific side effects—infection, dizziness, and insomnia—even the tuned DT-only model outperformed the original DT/ATC model (previously identified as the highest-performing model in the original study) for dizziness and insomnia (**Figure 6**). This indicates that fine-tuning the LR model alone has the potential to improve accuracy beyond any domain feature combination tested in the study.

DISCUSSION AND CONCLUSION

Supervised machine learning models for side effect prediction have gained increasing attention in recent years as a promising alternative to clinical or animal testing. Beyond broad safety evaluations, they also hold potential in personalized side effect risk analysis using patient-specific data [19]. Our project extended the existing study of predicting drug side effects and focused specifically on the predictive power of ATC codes and fine-tuning the logistic regression model. Through extensive analysis and model evaluation, we found several notable trends that highlight the tradeoffs required when focusing on specificity and granularity.

One of the primary research questions was to evaluate whether increasing the specificity of ATC codes would improve the predictive power of ML models. Our results showed that while ATC levels 2, 3, and 4 performed comparably, level 5 showed significant underperformance due to overfitting. This is likely due to level 5 codes being too specific, leading to a lack of sufficient data points for certain drug classifications. As a result, the model may begin learning noise or irrelevant details

specific to the training set rather than being able to generalize patterns that can apply to data it doesn't see. This phenomenon, known as overfitting, is a common issue in machine learning when feature granularity increases but the amount of available training data is limited [20]. As a result, it is vital to carefully identify and include only the optimal factors rather than every single available one, as redundant factors can potentially negatively affect the model as seen with our results for level 5 ATC codes.

Our detailed comparisons between level 2 and level 4 ATC codes showed that the performance difference was not statistically significant. This suggests that while more detailed ATC classifications can provide additional pharmacological context, they may not necessarily translate into improved predictive power. This finding highlights the importance of balancing feature specificity and generalizability in ML models when working with hierarchical classification systems like ATC codes. The difference was still not statistically significant even after combining DT features with ATC codes, reinforcing the idea that the relative contributions of domain knowledge features depend on the complexity of the dataset and the structure of the model.

The second major aspect of the study was to fine-tune the parameters of the LR model, namely through optimizing decision thresholds and adjusting the regularization strength. Through modifying these factors, we achieved an improvement in model accuracy across all side effects, with certain side effects having especially substantial gains (*e.g.*, infection, nervous system disorder, and insomnia).

This improvement observed in sensitivity-specificity trade-offs suggests that threshold optimization is a critical step

in refining ML models for drug safety prediction. The default threshold, 0.5, may not be ideal for every side effect, particularly for those that are rare or highly specific. Maximizing Youden's J statistic provided a more balanced approach, reducing both false positives and false negatives and offering the greatest diagnostic accuracy [21]. Similarly, tuning the regularization strength (C) showed that smaller values, or higher regularization, led to more stable models with a reduced risk of overfitting. Tuning the LR model led to improvements even without introducing additional domain knowledge features, indicating that model optimization can be as impactful as feature selection in enhancing predictive strength. Ultimately, combining both our feature selection and fine-tuning efforts led to the best-performing model (tuned DT/ATC level 4).

This study had several limitations. Performing analysis on different ATC levels and adjusting parameters of the LR model were computationally expensive, expending significant time while yielding small improvements. Further fine-tuning the model could yield higher accuracies, but these improvements will likely be marginal when also considering the model's runtime [22]. This also makes it difficult to explore the predictive accuracy of side effects beyond those tested in this project. Additionally, a more systematic approach to this project could lead to improved performance. Namely, fine-tuning the LR model as the first extension to get the best-performing model could have then been used to identify optimal performance for the higher level ATC code extension. In addition, since the model has not yet been tested on less common or unknown side effects, it may not be generalizable, as model accuracy may differ when dealing with some of the rarer side effects. Future

studies could explore alternative feature engineering approaches or machine learning models, as well as integrate more datasets to further refine drug safety prediction models.

While the findings provide insights into predictive modeling, real-world validation is still needed to assess how well the model performs in clinical settings [23]. Collaborating with pharmacological experts to test model predictions against real patient data would be a valuable next step. By integrating data-driven insights with clinical expertise, ML has potential to enhance drug safety assessments and ultimately reduce the risks associated with adverse reactions.

ACKNOWLEDGEMENTS

We would like to thank our professor Dr. Brian Nadel for his guidance and support throughout developing our project in our class C&S BIO 185 at UCLA. We also would like to express our gratitude towards the original study by Liu and Wilson, which provided the foundational framework for our research. Finally, we are grateful to UCLA Library Services for providing access to essential databases such as PubMed and NCBI, which were crucial for finding scholarly literature resources to support and guide our project.

CODE AND DATA AVAILABILITY

The code and data from the original study we built upon can be found at the following GitHub link:

https://github.com/jenwilson521/ML_At看_D_T_PFX

The code that we developed built upon the original study, and we modified it to analyze higher level ATC codes and also utilized different parameters for fine tuning. Our code and data can be found below:

https://github.com/abeni8/ML_At看_DT_PF_X

REFERENCES

1. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*. 2011;12(1):169. doi:10.1186/1471-2105-12-169
2. Santaella RM, Fraunfelder FW. Ocular adverse effects associated with systemic medications. *Drugs*. 2007;67(1):75-93. doi:10.2165/00003495-200767010-00006
3. Kremer I, Gatton DD, David M, Gatton E, Shapiro A. Toxic effects of systemic retinoids on meibomian glands. *Ophthalmic Res*. 2009;26(2):124-128. doi:10.1159/000267402
4. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507
5. Winters ZE, Griffin C, Horne R, Bidad N, McCulloch P. Barriers to accrue to clinical trials and possible solutions. *Br J Cancer*. 2014;111(4):637-639. doi:10.1038/bjc.2014.318
6. Amorim AMB, Piochi LF, Gaspar AT, Preto AJ, Rosário-Ferreira N, Moreira IS. Advancing drug safety in drug development: bridging computational predictions for enhanced toxicity prediction. *Chem Res Toxicol*. 2024;37(6):827-849. doi:10.1021/acs.chemrestox.3c00352
7. Galeano D, Paccanaro A. Machine learning prediction of side effects for drugs in clinical trials. *Cell Rep Methods*. 2022;2(12):100358. doi:10.1016/j.crmeth.2022.100358
8. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. 2016;44(D1):D1075-D1079. doi:10.1093/nar/gkv1075
9. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34(Database issue):D668-D672. doi:10.1093/nar/gkj067
10. Knox C, Wilson M, Klinger CM, *et al*. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. doi:10.1093/nar/gkad976
11. Liang H, Chen L, Zhao X, Zhang X. Prediction of drug side effects with a refined negative sample selection strategy. *Comput Math Methods Med*. 2020;2020:1573543. doi:10.1155/2020/1573543
12. Liu HJ, Wilson JL. Drug target, class level, and PathFX pathway information share utility for machine learning prediction of common drug-induced side effects. *Front Drug Saf Regul*. 2023;3:1287535. doi:10.3389/fdsfr.2023.1287535
13. Guo W, Liu J, Dong F, *et al*. Review of machine learning and deep learning models for toxicity prediction. *Exp Biol Med*. 2023;248(21):1952-1973. doi:10.1177/15353702231209421

14. Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. *Sci Transl Med*. 2011;3(114):114ra127. doi:10.1126/scitranslmed.3002774
15. Chen JJ, Tsai CA, Moon H, Ahn H, Young JJ, Chen CH. Decision threshold adjustment in class prediction. *SAR QSAR Environ Res*. 2006;17(3):337-352. doi:10.1080/10659360600787700
16. Njieutcheu Tassi CR, Börner A, Triebel R. Regularization strength impact on neural network ensembles. In: *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*. Association for Computing Machinery; 2023:1-9. doi:10.1145/3579654.3579661
17. Zhao X, Chen L, Lu J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math Biosci*. 2018;306:136-144. doi:10.1016/j.mbs.2018.09.010
18. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64. doi:10.1186/s13054-019-2351-7
19. Das P, Mazumder DH. An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. *Artif Intell Rev*. 2023;56(9):9809-9836. doi:10.1007/s10462-023-10413-7
20. Cook JA, Ranstam J. Overfitting. *Br J Surg*. 2016;103(13):1814. doi:10.1002/bjs.10244
21. Yin J, Tian L. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. *Stat Med*. 2014;33(8):1426-1440. doi:10.1002/sim.6046
22. Baughman M, Chakubaji N, Truong HL, Kreics K, Chard K, Foster I. Measuring, quantifying, and predicting the cost-accuracy tradeoff. In: *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*. IEEE; 2019:3616-3622. doi:10.1109/BigData47090.2019.9006370
23. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4(125):125ra31. doi:10.1126/scitranslmed.3003377

GLOSSARY

1. *Side Effect Resource (SIDER)* – A database that compiles information on drug side effects from various sources.
2. *DrugBank* – A comprehensive database containing detailed information on drugs, including their chemical properties, targets, and interactions.
3. *Anatomical Therapeutic Chemical (ATC) Classification* – A system used to classify drugs based on their therapeutic use and chemical properties.
4. *Drug Targets (DT)* – Specific proteins or molecules in the body that a drug interacts with to produce a therapeutic effect.
5. *Logistic Regression (LR)* – A statistical model used for binary classification tasks, like in our case, predicting the occurrence of side effects.
6. *Protein-Protein Interaction (PPI) Networks* – Complex networks of proteins that interact with each other, used in drug side effect prediction.
7. *PathFX* – A pathway modeling tool that predicts drug-induced phenotypic effects based on protein-protein interactions.
8. *Overfitting* – A situation where a model learns patterns from the training data too well, resulting in poor performance on new and unseen data.
9. *Regularization* – A technique used in machine learning to prevent overfitting by penalizing large model coefficients.
10. *Grid Search* – A method used in machine learning to systematically test different hyperparameter values to find the best-performing model.
11. *Decision Threshold* – A value that determines the cutoff point for classifying predictions in a machine learning model.
12. *C Value* – A regularization parameter in logistic regression and support vector machines that controls the trade-off between achieving a low error on training data and maintaining model generalization (higher C means less regularization and lower C increases regularization).
13. *Youden's J Statistic* – A metric used to find the optimal decision threshold for classification models by maximizing the difference between true positive and false positive rates.

SUPPLEMENTARY DATA

Table 1: Overview of First 100 Entries in drugbank.tsv (Mapped and Processed Data)

DrugBank ID	Name	Type	Group(s)	ATC Codes
DB00001	Lepirudin	biotech	approved	B01AE02
DB00002	Cetuximab	biotech	approved	L01XC06
DB00003	Dornase alfa	biotech	approved	R05CB13
DB00004	Denileukin diftitox	biotech	approved investigational	L01XX29
DB00005	Etanercept	biotech	approved investigational	L04AB01
DB00006	Bivalirudin	small molecule	approved investigational	B01AE06
DB00007	Leuprolide	small molecule	approved investigational	L02AE51 L02AE02
DB00008	Peginterferon alfa-2a	biotech	approved investigational	L03AB11 L03AB61
DB00009	Alteplase	biotech	approved	B01AD02 S01XA13
DB00010	Sermorelin	biotech	approved withdrawn	V04CD03 H01AC04
DB00011	Interferon alfa-n1	biotech	approved investigational	L03AB06
DB00012	Darbepoetin alfa	biotech	approved investigational	B03XA02
DB00013	Urokinase	biotech	approved investigational withdrawn	B01AD04
DB00014	Goserelin	small molecule	approved	L02AE03
DB00015	Reteplase	biotech	approved investigational	B01AD07
DB00016	Erythropoietin	biotech	approved	B03XA01
DB00017	Salmon calcitonin	biotech	approved investigational	H05BA01
DB00018	Interferon alfa-n3	biotech	approved investigational	
DB00019	Pegfilgrastim	biotech	approved	L03AA13
DB00020	Sargramostim	biotech	approved investigational	L03AA09
DB00022	Peginterferon alfa-2b	biotech	approved	L03AB10 L03AB60
DB00023	Asparaginase Escherichia coli	biotech	approved investigational	L01XX02
DB00024	Thyrotropin alfa	biotech	approved vet_approved	H01AB01
DB00025	Antihemophilic factor, human recombinant	biotech	approved investigational	B02BD02
DB00026	Anakinra	biotech	approved	L04AC03
DB00027	Gramicidin D	small molecule	approved	R02AB30
DB00028	Human immunoglobulin G	biotech	approved investigational	
DB00029	Anistreplase	biotech	approved	B01AD03
DB00030	Insulin human	biotech	approved investigational	A10AC01 A10AE01
DB00031	Tenecteplase	biotech	approved	B01AD11
DB00032	Menotropins	biotech	approved	G03GA02
DB00033	Interferon gamma-1b	biotech	approved investigational	
DB00034	Interferon alfa-2a, Recombinant	biotech	approved investigational	L03AB04
DB00035	Desmopressin	small molecule	approved	H01BA02
DB00036	Coagulation factor VIIa Recombinant Human	biotech	approved	B02BD08
DB00038	Oprelvekin	biotech	approved investigational	L03AC02
DB00039	Palifermin	biotech	approved	V03AF08
DB00040	Glucagon	biotech	approved	H04AA01
DB00041	Aldesleukin	biotech	approved	L03AC01
DB00042	Botulinum Toxin Type B	biotech	approved investigational	M03AX01
DB00043	Omalizumab	biotech	approved investigational	R03DX05
DB00044	Lutropin alfa	biotech	approved	G03GA07
DB00045	Lyme disease vaccine (recombinant OspA)	biotech	approved withdrawn	
DB00046	Insulin lispro	biotech	approved	A10AC04 A10AB04 A10AD04
DB00047	Insulin glargine	biotech	approved	A10AE54 A10AE04
DB00048	Collagenase clostridium histolyticum	biotech	approved investigational	D03BA02 M09AB02 D03BA52
DB00049	Rasburicase	biotech	approved investigational	V03AF07 M04AX01
DB00050	Cetorelix	small molecule	approved investigational	H01CC02
DB00051	Adalimumab	biotech	approved	L04AB04
DB00052	Somatotropin	biotech	approved investigational	H01AC01
DB00053	Imiglucerase	biotech	approved	A16AB02
DB00054	Abciximab	biotech	approved	B01AC13
DB00055	Drotrecogin alfa	biotech	approved investigational withdrawn	B01AD10
DB00056	Gemtuzumab ozogamicin	biotech	approved investigational	L01XC05
DB00057	Indium In-111 satumomab pendetide	biotech	experimental withdrawn	V09IB02
DB00058	Alpha-1-proteinase inhibitor	biotech	approved	B02AB02
DB00059	Pegaspargase	biotech	approved investigational	L01XX24
DB00060	Interferon beta-1a	biotech	approved investigational	L03AB07

DB00061	Pegademase	biotech	approved	L03AX04
DB00062	Albumin human	biotech	approved	
DB00063	Eptifibatide	biotech	approved investigational	B01AC16
DB00065	Infliximab	biotech	approved	L04AB02
DB00066	Follitropin	biotech	approved	G03GA06 G03GA10 G03GA05
DB00067	Vasopressin	small molecule	approved	
DB00068	Interferon beta-1b	biotech	approved	L03AB08
DB00069	Interferon alfacon-1	biotech	approved investigational	L03AB09
DB00070	Hyaluronidase (ovine)	biotech	approved investigational	
DB00071	Insulin pork	biotech	approved	A10AC03 A10AB03
DB00072	Trastuzumab	biotech	approved investigational	L01XC03
DB00073	Rituximab	biotech	approved	L01XC02
DB00074	Basiliximab	biotech	approved investigational	L04AC02
DB00075	Muromonab	biotech	approved investigational	L04AA02
DB00076	Digoxin Immune Fab (Ovine)	biotech	approved	V03AB24
DB00078	Ibritumomab tiuxetan	biotech	approved investigational	V10XX02
DB00080	Daptomycin	small molecule	approved investigational	J01XX09
DB00081	Tositumomab	biotech	approved investigational	V10XA53
DB00082	Pegvisomant	biotech	approved	H01AX01
DB00083	Botulinum toxin type A	biotech	approved investigational	M03AX01
DB00085	Pancrelipase	biotech	approved investigational	A09AA02
DB00086	Streptokinase	biotech	approved investigational	B01AD01 B06AA55
DB00087	Alemtuzumab	biotech	approved investigational	L04AA34
DB00088	Alglucerase	biotech	approved investigational	A16AB01
DB00089	Capromab pendetide	biotech	approved	V09IB04
DB00090	Laronidase	biotech	approved	A16AB05
DB00091	Cyclosporine	small molecule	approved investigational vet_approved	L04AD01 S01XA18
DB00092	Alefacept	biotech	approved investigational withdrawn	L04AA15
DB00093	Felypressin	small molecule	experimental	
DB00094	Urofollitropin	biotech	approved vet_approved	G03GA04
DB00095	Efalutuzumab	biotech	approved investigational	L04AA21
DB00097	Choriogonadotropin alfa	biotech	approved	G03GA08
DB00098	Antithymocyte immunoglobulin (rabbit)	biotech	approved	L04AA04
DB00099	Filgrastim	biotech	approved	L03AA02
DB00100	Coagulation Factor IX (Recombinant)	biotech	approved investigational	

Table 2: LR Model Performance Across ATC Code Levels 2-5 (only) for Top 30 Side Effects.

Side Effect	LR: ATC Level 2	LR: ATC Level 3	LR: ATC Level 4	LR: ATC Level 5	F-Value	P-value
dizziness	0.675898305	0.66779661	0.679627119	0.539118644	1168.720528	4.73E-164
nausea	0.683143713	0.668892216	0.683263473	0.556317365	1232.911303	3.09E-167
headache	0.665696203	0.680221519	0.684778481	0.570506329	728.4721486	1.50E-136
rash	0.664735099	0.670496689	0.674503311	0.565496689	798.0560486	9.38E-142
vomiting	0.696447368	0.673815789	0.663223684	0.526282895	1532.945956	2.46E-180
asthenia	0.707677903	0.711123596	0.708202247	0.546816479	1499.301749	5.42E-179
diarrhea	0.690797101	0.684710145	0.681086957	0.544202899	1168.512581	4.84E-164
pruritus	0.700074627	0.710335821	0.703320896	0.586679104	747.6217069	5.04E-138
hypersensitivity	0.68328	0.67948	0.6614	0.582	511.3902923	6.17E-117
abdominal pain	0.698893617	0.698978723	0.689744681	0.582212766	488.4736446	1.81E-114
urticaria	0.677719298	0.692105263	0.698815789	0.588201754	540.6800074	5.88E-120
body temperature increased	0.712366071	0.687633929	0.697678571	0.567276786	859.9684322	4.68E-146
dermatitis	0.66947541	0.679770492	0.654590164	0.534852459	1108.05689	6.86E-161
pain	0.656165049	0.663300971	0.674902913	0.559271845	423.9395871	5.70E-107
gastrointestinal disorder	0.639086758	0.63739726	0.663150685	0.553881279	398.027845	1.07E-103
fatigue	0.699315068	0.712191781	0.694109589	0.555159817	657.9036026	8.38E-131
dyspepsia	0.657607656	0.669282297	0.6815311	0.554497608	596.8888118	2.19E-125
constipation	0.708651163	0.698	0.697395349	0.550232558	1036.848993	5.71E-157
dyspnoea	0.697980296	0.691773399	0.708916256	0.531674877	1290.015007	6.07E-170
hypotension	0.7084375	0.715677083	0.68265625	0.534739583	993.1681834	1.93E-154
insomnia	0.685048544	0.692815534	0.69092233	0.562669903	722.7689693	4.20E-136
anaphylactic shock	0.606235294	0.597117647	0.635764706	0.527294118	318.0105783	2.19E-92
paraesthesia	0.656020408	0.638520408	0.64744898	0.53	759.0456664	6.89E-139
somnolence	0.733349515	0.753932039	0.737864078	0.579320388	1065.862391	1.35E-158
nervous system disorder	0.609840426	0.610531915	0.615212766	0.530744681	294.7019984	1.11E-88
thrombocytopenia	0.716054054	0.696756757	0.700324324	0.519513514	1837.22993	2.33E-191
tachycardia	0.663693182	0.660284091	0.626590909	0.494488636	994.9981469	1.50E-154
arthralgia	0.657267442	0.675988372	0.680116279	0.540988372	641.4836479	2.17E-129
infection	0.642259887	0.641694915	0.673785311	0.495875706	1027.24426	2.01E-156
musculoskeletal discomfort	0.688965517	0.677142857	0.728669951	0.54862069	1203.050782	8.95E-166
Average	0.6784060825	0.6779256039	0.6806532383	0.5486312581		

Table 3: Best-performing LR Model (ATC Level 4/DT) Accuracy for Top 30 Side Effects: Tuned v. Untuned, Sorted by Mean Difference

Side Effect	LR: Untuned	LR: Tuned	Mean Difference (Tuned - Untuned)
asthenia	0.647903	0.705169	0.057266
infection	0.616836	0.66661	0.049774
thrombocytopenia	0.684378	0.731568	0.04719
nervous system disorder	0.616223	0.662021	0.045798
dyspnoea	0.661379	0.705025	0.043646
dizziness	0.681458	0.721186	0.039728
arthralgia	0.676395	0.715698	0.039303
musculoskeletal discomfort	0.678768	0.716404	0.037636
nausea	0.65485	0.691198	0.036348
urticaria	0.690219	0.726316	0.036097
fatigue	0.686804	0.72137	0.034566
diarrhea	0.681667	0.715797	0.03413
hypotension	0.678177	0.711771	0.033594
insomnia	0.67932	0.712476	0.033156
paraesthesia	0.660969	0.692959	0.03199
abdominal pain	0.673745	0.702383	0.028638
dermatitis	0.660295	0.688787	0.028492
pain	0.65534	0.683544	0.028204
body temperature	0.679509	0.707634	0.028125
dyspepsia	0.679809	0.707751	0.027942
gastrointestinal disorder	0.65274	0.680457	0.027717
somnolence	0.699029	0.726214	0.027185
rash	0.652483	0.67755	0.025067
pruritus	0.655224	0.678619	0.023395
vomiting	0.696316	0.718158	0.021842
constipation	0.699581	0.718186	0.018605
hypersensitivity	0.67596	0.69336	0.0174
anaphylactic shock	0.691176	0.708176	0.017
headache	0.677627	0.693639	0.016012
tachycardia	0.658977	0.668977	0.01
Average	0.670105	0.701633	0.031528