

## Subject Section

# Ranking the Predictability of Drug Side Effects Based on Their Chemical Structures

Aseel Ibrahim<sup>1,\*</sup>, Haixuan Yang<sup>2</sup> and Yezhao Zhong<sup>2,\*</sup>

<sup>1</sup>Mathematics, Statistics and Applied Mathematics, University of Galway, Galway, Galway City, Post Code, Ireland and

<sup>2</sup>Mathematics, Statistics and Applied Mathematics, University of Galway, Galway, Galway City, Post Code, Ireland.

\*To whom correspondence should be addressed.

## Abstract

**Motivation:** The prediction of adverse drug reactions (ADRs) is a critical aspect of pharmaceutical development, essential for ensuring patient safety and reducing financial losses associated with drug recalls and adverse event litigations. Traditional methods, such as clinical trials, are often limited by ethical concerns, time consumption, and financial costs, and may not capture all possible side effects due to limited sample sizes and diversity. Motivated by these challenges, our study leverages computational approaches to predict ADRs using logistic regression with Lasso regularization, utilizing data from the SIDER and PubChem databases.

**Results:** Our model achieved high predictive accuracy for several ADRs, including increased potassium levels (AUC: 1.00), red blood cell disorders (AUC: 0.993), and gastrointestinal perforation (AUC: 0.982). These high AUC values are due to specific and measurable clinical markers influenced by certain drugs, such as diuretics affecting potassium levels and NSAIDs impacting gastrointestinal integrity. However, the model showed lower AUC values for multifactorial ADRs like urinary tract disorders, nephropathy, and ventricular tachycardia, which present varied symptoms influenced by broader factors. Our study also identified common significant chemical fingerprints among the top predicted side effects, particularly fingerprint 1, which was influential across multiple ADRs. This identification aids in targeted monitoring and enhances the drug development process.

## Availability:

**Contact:** A.Ibrahim12@universityofgalway.ie

**Supplementary information:** Supplementary data are available at <https://github.com/A1Ibrahim/Ranking-the-Predictability-of-Drug-Side-Effects-Based-on-Their-Chemical-Structures-codes> online.

## 1 Introduction

The prediction of drug side effects is crucial for pharmaceutical companies to overcome financial losses and, more importantly, to ensure patient safety (Zaheer, n.d.; Walsh et al., 2015). Significant efforts have been made in this field to explore side effects before marketing drugs. Traditional methods, such as clinical trials, involve human subjects to determine both the efficacy and side effects of drugs. However, these methods have limitations in terms of ethical concerns, time consumption, and financial costs (DiMasi et al., 2016; Schipper, 2020). Moreover, the sample size and diversity in clinical trials may not be sufficient to capture all possible side effects, leading to less reliable predictions (Greenwood, 2023). In recent years, computational methods have been developed to predict side effects before proceeding with clinical trials. Techniques such as machine

learning have shown promise in identifying potential side effects based on existing drug data (Vilar et al., 2018; Toni et al., 2024). However, these methods also face challenges, including data quality issues, the complexity of biological systems, and the need for extensive computational resources (Kavakiotis et al., 2017). Low-quality data can lead to inaccurate predictions, which may result in undetected adverse effects and pose significant risks to patient safety. High-quality, diverse data is essential for building reliable predictive models that can generalize across different populations and drug interactions. Chemical fingerprints, which are unique digital representations of the molecular structure of a compound, have emerged as a promising tool in this domain. These fingerprints allow for the characterization of a compound's properties and potential interactions, providing a robust basis for predictive modeling (Yang et al., 2022). Databases such as the SIDER database for known side effects and the PubChem database for drug fingerprints are invaluable resources in this

context. By leveraging these databases, we can develop more accurate predictive models. This study aims to advance the prediction of drug side effects by utilizing chemical fingerprints of drugs to predict side effects even before manufacturing. By building a predictive model using logistic regression with Lasso regularization, we intend to improve the accuracy and reliability of side effect predictions. Our approach focuses on ranking side effects by their predictability, which could significantly enhance the drug development process. This research has the potential to lead to safer drugs and more efficient use of resources, contributing to both economic savings and improved patient outcomes. Notably, our model demonstrated high accuracy in predicting the side effect of increased potassium levels, though it performed less well in predicting ventricular tachycardia, indicating areas for further refinement and improvement in our predictive methodology.

## 2 Methods

### 2.1 Data Collection and Preparation

There are two databases utilized in this study: the SIDER database, which contains side effects presented as binary data (1 for the presence of each side effect with each drug and 0 for the absence)(Kuhn et al., 2010, 2016), and the PubChem database, which contains fingerprints also presented as binary data (1 for the presence of each fingerprint with each drug and 0 for the absence). To prepare the data, we converted the datasets to dataframes and then merged the two dataframes based on the common drug names to create a unified dataset containing both fingerprints and side effects. We then cleaned the missing values or incomplete data to ensure data integrity. Next, the cleaned data was split into training and testing sets using an 80/20 split with a fixed random seed.

### 2.2 Predictive Modeling

For each side effect, a logistic regression model with Lasso regularization was trained using the training data. Logistic regression is a statistical method for modeling the probability of a binary outcome (Seifeddine et al., 2020). The logistic regression model can be expressed as follows:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_m x_{im})}} \quad (1)$$

(Akalin, n.d.)

where  $P_i$  is the probability of the side effect being present,  $\beta_0$  is the intercept,  $\beta_i$  are the coefficients, and  $x_i$  are the features (chemical fingerprints) for the  $i$ -th (drug) observation.

We used lasso regularization (Least Absolute Shrinkage and Selection Operator) to overcome over-fitting (which is model performs very well on the training data but poorly on new, unseen data (testing data) (AWS, 2024)). Lasso adds a penalty based on the absolute value of the coefficients' magnitudes to the loss function. This penalty causes the coefficients of less important features to shrink to zero, thereby effectively performing feature selection (Pillai, 2023).

firstly, the log-likelihood for the logistic regression model is maximized as follows:

$$\max \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log (1 - p_i)] \quad (2)$$

(Akalin, n.d.)

then, the negative log-likelihood, which we aim to minimize it and Combining the logistic loss with the Lasso penalty term:

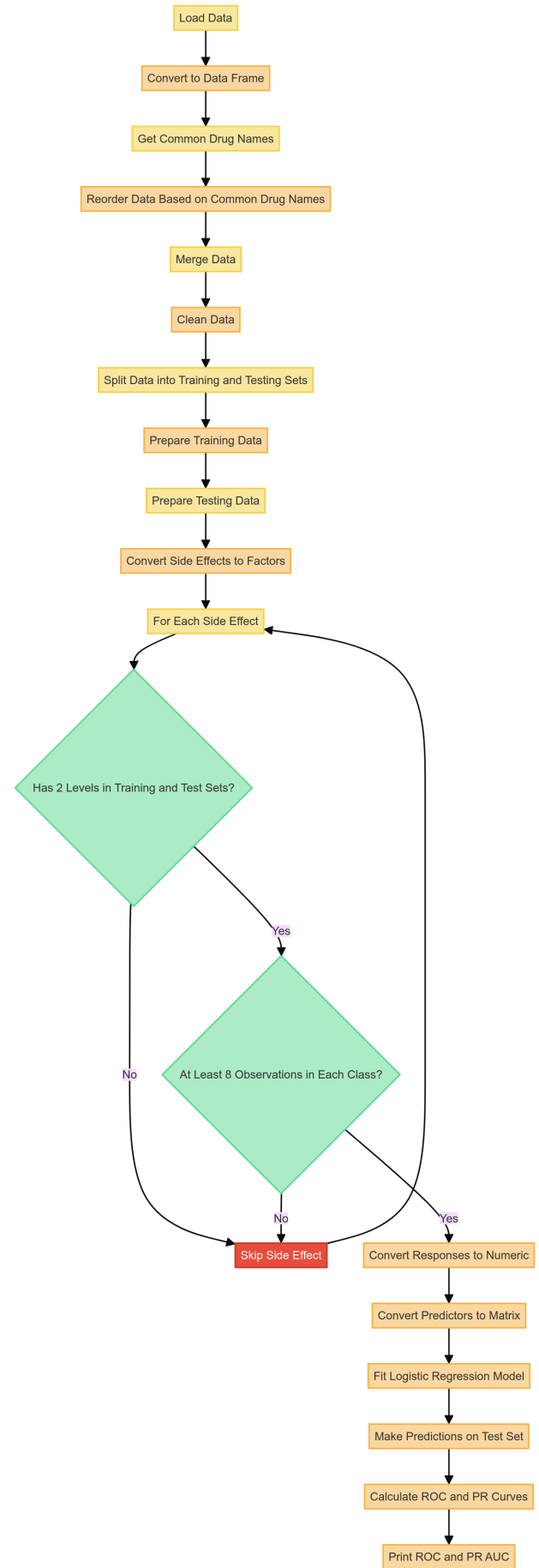


Fig. 1. FLOWCHART OF METHOD

$$\min \left( - \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log (1 - p_i)] + \lambda \sum_{j=1}^p |\beta_j| \right) \quad (3)$$

(Akalin, n.d.)

where  $P_i$  is the logistic regression which is the predicted probability for the  $i$ -th (drugs) observation.  $y_i$  is the ground truth which is the actual binary outcome for the  $i$ -th (drugs) observation.  $N$  is the number of observations (drugs).  $B_j$  are the coefficients of the model.  $\lambda$  is the regularization parameter.

We employed 10-fold cross-validation, a statistical method used to evaluate and compare learning algorithms by dividing the data into training and validation sets (Refaeilzadeh et al., 2009). In this technique, the training and validation sets are alternated in successive iterations, ensuring that each data point is used for validation at least once. This approach was used to select the optimal  $\lambda$  value for the model to ensure robustness and reliability. Subsequently, predictions were made on the test data, and the model's performance was evaluated using ROC AUC and PR AUC metrics to assess its discriminative ability.

### 2.3 Addressing Class Imbalance

Class imbalance (CI) in classification issues occurs when the number of observations in one class is significantly lower than in another class (Khan et al., 2023). To address class imbalance issues, we ensured that each side effect had at least two levels in both the training and test sets, and each class in the training data had at least eight observations. This helps to make sure that model don't end up being trained only on negative labels. So, the model will at least see some positive instances.

### 2.4 Evaluation

The performance of our predictive models was evaluated using the following metrics:

#### 2.4.1 ROC AUC (Receiver Operating Characteristic Area Under the Curve)

The ROC AUC score is a single value that encapsulates a classifier's performance across every possible classification threshold. It is obtained by calculating the area under the ROC curve. The ROC AUC score measures the classifier's ability to differentiate between positive and negative classes, ranging from 0 to 1. A higher ROC AUC score signifies better performance, with a perfect model scoring 1 and a random model scoring 0.5. (Evidently AI Team, n.d.) (fig.2,A).

#### 2.4.2 Precision-Recall AUC

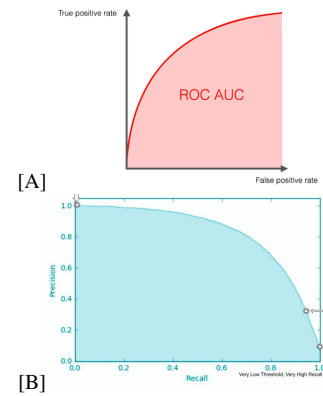
Precision indicates the confidence that a predicted positive instance is indeed positive, while recall measures the model's ability to capture all actual positive instances in the dataset. The positive class refers to the category of interest in the classification task (Chugh, n.d.):

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$TP$  (True Positive) represents cases where the model exactly predicts positive values,  $TN$  (True Negative) indicates instances where the model exactly predicts negative values,  $FP$  (False Positive) occurs when the model wrongly predicts positive values, and  $FN$  (False Negative) arises when the model wrongly predicts negative values (Seo et al., 2020).

This metric is particularly useful for imbalanced datasets as it focuses on the performance of the model for the positive class (fig.2,B).



**Fig. 2.** Combined ROC AUC and PR Curve

The outcomes demonstrated high ROC AUC in predicting the side effect of increased potassium levels, though there was less ROC AUC in predicting ventricular tachycardia, indicating areas for further refinement and improvement in our predictive methodology. By concentrating on these predictive methods and ranking side effects by their predictability, this research could significantly enhance the drug development process, leading to safer drugs and more efficient use of resources.

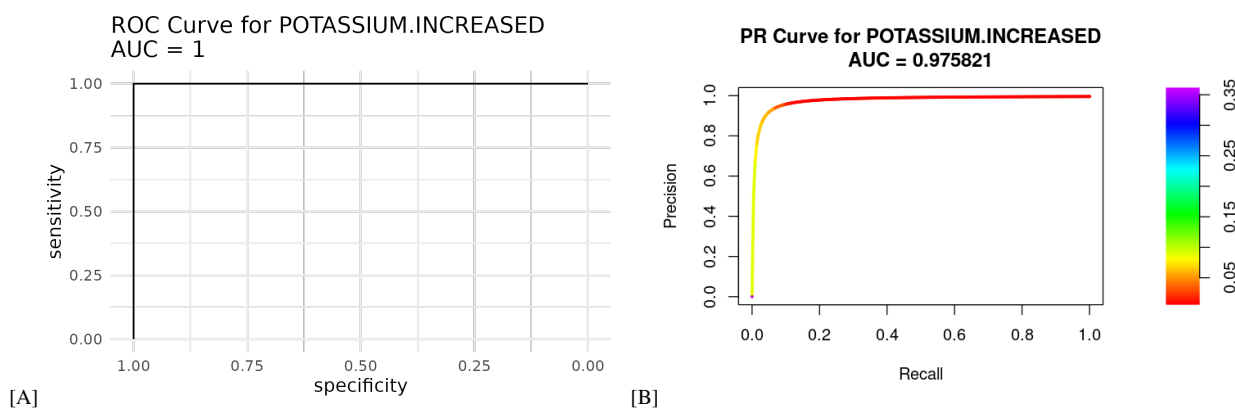
## 3 Result

Our study on predicting ADRs using logistic regression with Lasso regularization, leveraging data from the SIDER database and PubChem, demonstrated high ROC AUC for several ADRs as seen in table 1, notably potassium increased (AUC: 1.00), red blood cell disorders (AUC: 0.993), and gastrointestinal perforation (AUC: 0.982). The high AUC values for these ADRs can be attributed to the specific and measurable clinical markers influenced by certain drugs. Blood glucose levels, easily monitored and influenced by the drugs, yielded an AUC of 0.9688, while aneurysms, with distinct physiological indicators detectable via imaging and correlated with drugs affecting blood pressure and vessel integrity, showed an AUC of 0.9263.

Hypotension symptomatic, nephrotoxicity, lactic acidosis, breast tenderness, and extrapyramidal symptoms also demonstrated high predictability due to their clear clinical presentations and strong associations with specific medications.

In contrast, lower AUC values were observed (as seen in table 2) for ADRs like urinary tract disorder (AUC: 0.616), nephropathy (AUC: 0.615), and ventricular tachycardia (AUC: 0.611), reflecting the multifactorial nature and varied symptoms of these conditions, which complicate precise predictions based solely on drug interactions.

The PR curve for the prediction of drugs causing potassium increased side effect demonstrates an AUC of 0.975821 (fig.3). This high AUC value indicates that the model exhibits excellent performance in distinguishing between drugs that are likely to cause increased potassium levels and those that are not. The curve shows high precision across various levels of recall, implying that the model maintains a high accuracy in identifying true positives while minimizing false positives. This result underscores the model's effectiveness in predicting the risk of hyperkalemia associated with certain medications.



**Fig. 3.** Performance curves for POTASSIUM.INCREASED: (A) Receiver Operating Characteristic (ROC) curve, (B) Precision-Recall (PR) curve

Side Effects	ROC AUC score	PR AUC score
POTASSIUM.INCREASED	1.0000000	1.0000000
RED.BLOOD.CELL.DISORDERS	0.9933036	0.250000000
GASTROINTESTINAL.PERFORATION	0.9820628	0.183100192
GLUCOSE.INCREASED	0.9687500	0.06528025
ANEURYSM	0.9263393	0.0279597
HYPOTENSION.SYMPTOMATIC	0.9226457	0.5206606
NEPHROTOXICITY	0.9219939	0.492440087
LACTIC.ACIDOSIS	0.9215247	0.5204233
BREAST.TENDERNESS	0.9189498	0.55597696
EXTRAPYRAMIDAL.SYMPTOMS	0.8938727	0.6119965

Table 1. Top 10 in ROC AUC score

Side Effects	ROC AUC score	PR AUC score
VENTRICULAR.TACHYCARDIA	0.6107000	0.145420188
RENAL.FAILURE	0.6116000	0.349775712
URETHRAL.DISORDER	0.6117811	0.3401693
INFLUENZA	0.6120095	0.26662258
PROTHROMBIN.LEVEL.INCREASED	0.6123272	0.2506872
DIPLOPIA	0.6125073	0.2130609
PROCTOCOLITIS	0.6141141	0.02279677
DEMENTIA	0.6154545	0.4149633
NEPHROPATHY	0.6154835	0.07202773
URINARY.TRACT.DISORDER	0.6163033	0.3570276

Table 2. LEAST 10 in ROC AUC score

### 3.1 Identifying Common and Significant Fingerprints in Top Side Effects

By identifying and analyzing the most influential chemical fingerprints for the top 5 side effects based on ROC AUC values, we aimed to understand which specific chemical structures are strongly associated with these side effects. This understanding can help pharmaceutical companies prioritize their focus on the most significant risks, thereby improving patient safety and achieving economic savings.

Top Fingerprints for Each Side Effect Here are the fingerprints that were identified as most influential for the top 5 side effects (fig.6) (see supplementary file for more information):

1- POTASSIUM.INCREASED: Fingerprints 1, 155, 230, 292, 372, 375, 402, 700, 861

2- RED.BLOOD.CELL.DISORDERS: Fingerprints 1, 93

3- GASTROINTESTINAL.PERFORATION: Fingerprints 1, 85, 133, 193, 230, 493, 621, 668, 701, 902, 923, 992

4- GLUCOSE.INCREASED: Fingerprints 1, 8, 97, 331, 402, 458, 636, 666, 756, 768, 785, 882, 936, 946, 1014

5- ANEURYSM: Fingerprints 1, 437, 579, 780, 797, 868

## 4 Discussion

Our logistic regression model with Lasso regularization has proven highly effective in predicting certain ADRs, especially those with specific and measurable clinical markers. For instance, potassium increased achieved a perfect AUC of 1.0000, underscoring the clear clinical markers influenced by drugs such as diuretics (Bandari et al., 2018), which directly impact potassium levels. This high ROC AUC is critical for preventing hyperkalemia, a potentially life-threatening condition.






































TOP 5 SIDE EFFECTS	TOP FINGERPRINTS IN THEIR SIDE EFFECTS									
POTASSIUM INCREASED										
RED BLOOD CELL DISORDERS										
GASTROINTESTINAL PERFORATION										
GLUCOSE INCREASED										
ANEURYSM										

Fig. 4. TOP FINGERPRINTS IN TOP 5 SIDE EFFECTS

red blood cell disorder and gastrointestinal perforation also showed high ROC AUC scores of 0.9933 and 0.9821, respectively. These ADRs can be closely monitored using clinical markers influenced by specific drugs, such as NSAIDs' impact on gastrointestinal health (Wallace, 2004). However, the lower PR AUC scores for these ADRs suggest that while the model is good at distinguishing between affected and unaffected cases, it may produce a higher number of false positives, necessitating further refinement.

Glucose increased (ROC AUC: 0.9688) benefits from the easy monitoring of blood glucose levels, which can be significantly affected by drugs like corticosteroids and antipsychotics (Saag, 2021). "aneurysm" (ROC AUC: 0.9263) is detectable via imaging and is associated with drugs affecting blood pressure and vessel integrity.

Other ADRs, such as hypotension symptomatic, nephrotoxicity, lactic acidosis, breast tenderness, and extrapyramidal symptoms, showed high predictability due to their clear clinical presentations and strong associations with specific medications. These ADRs' relatively high PR AUC scores indicate a better precision-recall balance, reflecting reliable predictions.

In contrast, lower AUC values for other ADRs, such as urinary tract disorder (AUC: 0.616), nephropathy (AUC: 0.615), and ventricular tachycardia (AUC: 0.611), highlight the multifactorial nature and varied symptoms of these conditions, which complicate precise predictions based solely on drug interactions. These findings suggest that additional factors beyond drug interactions need to be considered to improve the prediction accuracy for these complex ADRs.

Overall, our model demonstrates high ROC AUC and reliability for predicting several ADRs with clear clinical markers, providing valuable insights for enhancing patient safety and treatment outcomes. Further refinement and inclusion of additional variables may improve the model's performance for ADRs with lower AUC scores. The results of our study underscore the efficacy of using logistic regression with Lasso regularization to predict ADRs leveraging data from the SIDER and PubChem databases. The high AUC values for several ADRs demonstrate the potential of this computational approach in accurately identifying drug side effects, which is crucial for ensuring patient safety and mitigating financial losses in the pharmaceutical industry.

We can categorize the side effects into three groups based on our model's performance:

**High ROC AUC for Specific ADRs:** The model exhibited high predictive accuracy for ADRs such as increased potassium levels, red

blood cell disorders, and gastrointestinal perforation. This success can be attributed to the availability of specific and measurable clinical markers for these conditions, which are strongly influenced by certain drugs. For instance, diuretics can significantly affect potassium levels, and NSAIDs are known to impact gastrointestinal integrity, providing clear indicators for these ADRs.

**Moderate to High AUC for Other ADRs:** Conditions like hemolytic anemia, acute renal failure, ileus, delusion, and prolonged QT electrocardiogram also demonstrated high predictability. These conditions have clear clinical presentations and are strongly associated with specific medications, aiding in the model's accurate predictions.

**Challenges with Multifactorial ADRs:** The model's lower AUC values for ADRs such as urinary tract disorders, nephropathy, and ventricular tachycardia highlight the challenges in predicting conditions that are multifactorial and present varied symptoms. These conditions are influenced by a broader range of factors, complicating precise predictions based solely on drug interactions.

#### 4.1 Common and Merged Fingerprints

The fingerprints 1, 230, and 402 are associated with multiple ADRs, underscoring their significance in various drug interactions. Fingerprint 1 (fig.4 red circle), characterized by the substructure containing 8 hydrogen atoms (PubChem Substructure Fingerprint V1.3, n.d.), appears in the contexts of increased potassium levels, red blood cell disorders, gastrointestinal perforation, increased glucose levels, and aneurysms. This fingerprint is linked to biochemical and physiological processes influenced by diuretics, NSAIDs, corticosteroids, and antipsychotics. Notably, fingerprint 1 is present in a wide range of drugs (as seen in supplementary material), which explains its involvement in a broad spectrum of side effects. Fingerprint 230 (fig.4 green circle), characterized by a saturated or aromatic heteroatom-containing ring of size 8 (PubChem Substructure Fingerprint V1.3, n.d.), is present in both potassium increased and gastrointestinal perforation, indicating its role in potassium regulation and gastrointestinal health, particularly in response to diuretics and NSAIDs. Fingerprint 402 (fig.4 blue circle), characterized by the substructure N(O)(O) (PubChem Substructure Fingerprint V1.3, n.d.), is associated with potassium increased and glucose increased, highlighting its involvement in potassium homeostasis and glucose metabolism, influenced by diuretics, corticosteroids, and antipsychotics. These overlapping markers illustrate the complex interactions between

drugs and various physiological systems, emphasizing the need for comprehensive monitoring in clinical settings to manage these ADRs effectively.

Overall, our logistic regression model with Lasso regularization demonstrates high ROC AUC in predicting certain ADRs with clear clinical markers. However, a few limitations must be highlighted. ADRs can be influenced by a multitude of factors beyond drug interactions, such as genetic predispositions, environmental influences, and patient comorbidities. Addressing class imbalance remains a challenge, especially for ADRs that are rare.

## 5 Conclusion

This study highlights the potential of using logistic regression with Lasso regularization for predicting ADRs based on chemical fingerprints. By leveraging comprehensive datasets from the SIDER and PubChem databases, we developed a model that demonstrates high ROC AUC in predicting certain ADRs with clear clinical markers, such as increased potassium levels, red blood cell disorders, and gastrointestinal perforation. These findings underscore the importance of high-quality data and robust computational methods in improving the predictability of drug side effects.

The identification of key fingerprints, particularly fingerprints 1, 230, and 402, and their association with multiple ADRs, emphasizes the complex interactions between drugs and physiological systems. These fingerprints serve as crucial indicators for monitoring potential adverse effects in clinical settings, thereby enhancing patient safety.

Despite the model's success in predicting several ADRs, challenges remain, particularly for multifactorial conditions with varied symptoms. This indicates the need for further refinement and the inclusion of additional variables to improve the prediction accuracy for these complex ADRs.

Overall, this research provides valuable insights into the use of chemical fingerprints for predicting drug side effects, offering a promising approach for safer drug development and more efficient use of resources. By ranking side effects by their predictability, this study contributes to a more systematic and reliable evaluation of drug safety, potentially leading to better patient outcomes and significant economic savings in the pharmaceutical industry.

## SUPPLEMENTAL INFORMATION

The code and more supplementary information can be found online at: <https://github.com/AIbrahim/Ranking-the-Predictability-of-Drug-Side-Effects-Based-on-Their-Chemical-Structures-codes->

## Acknowledgements

I would like to express my heartfelt thanks to all those who have supported me in this research, both mentally and emotionally.

## References

Zaheer A. Top 5 Medicine Producing Countries in the World. *InsiderMonkey*. Available: <https://www.insidermonkey.com/blog/top-5-medicine-producing-countries-in-the-world-1162493/> [Accessed: 2024-07-28].

Walsh, D., Lavan, A., Cushen, A.-M., & Williams, D. (2015). Adverse drug reactions as a cause of admission to a Dublin-based university teaching hospital. *Ir J Med Sci*, 184(2), 441-447. doi: 10.1007/s11845-014-1140-1.

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*, 47, 20-33. doi: 10.1016/j.jhealeco.2016.01.012.

Schipper, I. (2020). The ethics of clinical trials in times of corona. *SOMO*. Available: <https://www.somo.nl/the-ethics-of-clinical-trials-in-times-of-corona/> [Accessed: 2024-07-28].

Greenwood, M.Sc., Michael. (2023). Side Effects in Clinical Trials. *News-Medical*. Available: <https://www.news-medical.net/life-sciences/Side-Effects-in-Clinical-Trials.aspx> [Accessed: 2024-07-28].

Vilar, S., Friedman, C., & Hripcsak, G. (2018). Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. *Brief Bioinform*, 19(5), 863-877. doi: 10.1093/bib/bbx010.

Toni, E., Ayatollahi, H., Abbaszadeh, R., & Fotuhi Siahpirani, A. (2024). Machine Learning Techniques for Predicting Drug-Related Side Effects: A Scoping Review. *Pharmaceuticals*, 17(6), 795. doi: 10.3390/ph17060795.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15, 104-116. doi: 10.1016/j.csbj.2016.12.005.

Yang, J., Cai, Y., Zhao, K., Xie, H., & Chen, X. (2022). Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov Today*, 27(11), 103356. doi: 10.1016/j.drudis.2022.103356.

Seifeddine, M., Bradai, A., Bukhari, S., Pham Tran Anh, Q., Ben Ahmed, O., & Atri, M. (2020). A survey on machine learning in Internet of Things: Algorithms, strategies, and applications. *Internet of Things*, November 2020. doi: 10.1016/j.iot.2020.100314.

Khan, A. A., Chaudhari, O., & Chandra, R. (2023). A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. Available: <http://arxiv.org/abs/2304.02858> [Accessed: 2024-07-28].

Evidently AI Team. (n.d.). How to explain the ROC AUC score and ROC curve? *Evidently AI*. Available: <https://www.evidentlyai.com/classification-metrics/explain-roc-curve>.

Chugh, V. (n.d.). Precision-Recall Curve in Python Tutorial. *DataCamp*. Available: <https://www.datacamp.com/tutorial/precision-recall-curve-tutorial>.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. *Encyclopedia of Database Systems*, 532-538. doi: 10.1007/978-0-387-39940-9\_65.

Akalin, A. Computational Genomics with R. Available: <https://compmgenomr.github.io/book/> [Accessed: 2024-07-28].

Bandari, T. F., Venkatesh, P. K., Dryden, L. M., & Smith, N. L. (2018). Thiazide diuretics alone or in combination with a potassium-sparing diuretic on blood pressure-lowering in patients with primary hypertension: protocol for a systematic review and network meta-analysis. *Systematic Reviews*. Available: <https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-018-0737-5> [Accessed: 2024-07-28].

Wallace, M. B. (2004). Review: NSAIDs and gastrointestinal complications. *Postgraduate Medical Journal*. Available: <https://academic.oup.com/pmj/article/80/942/201/7033888> [Accessed: 2024-07-28].

Saag, P. (2021). Impact of corticosteroids on blood glucose levels. *Postgraduate Medical Journal*. Available: <https://academic.oup.com/pmj/article/98/1160/477/6958929>.

PubChem Substructure Fingerprint V1.3. *PubChem*. Available: [https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem\\_fingerprints.pdf](https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf). 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. Available: <https://doi.org/10.1093/nar/gkv1075>

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1), 343. Available: <https://doi.org/10.1038/msb.2009.98>

Amazon Web Services (2024). Underfitting vs. Overfitting. *Amazon Machine Learning Documentation*. Available: <https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html>

Pillai, G. (2023). Feature Selection: Enhancing Model Accuracy and Interpretability. *Medium*. Available: <https://medium.com/@pillagreesh16/feature-selection-enhancing-model-accuracy-and-interpretability-19c119cd4bb5>

Seo, S., Lee, T., Kim, M., & Yoon, Y. (2020). Prediction of Side Effects Using Comprehensive Similarity Measures. *BioMed Research International*.