



# Prediction of ADRs using NMF and Weighted NMFs

Arshad Ahamed Shajahan, Sai Shasank Dandu

<sup>1</sup> School of Mathematical and Statistical Sciences, University of Galway

## Background and Objectives

Adverse drug reactions (ADRs) are a common problem in clinical and pharmacovigilance research and can lead to serious patient harm and biased conclusions if modelled poorly. Sparse, noisy, and highly imbalanced drug ADR data often cause standard machine learning methods to perform no better than naïve frequency-based approaches, unless appropriate low-rank and kernel-based methods are used with clear assumptions [1].

## Objectives

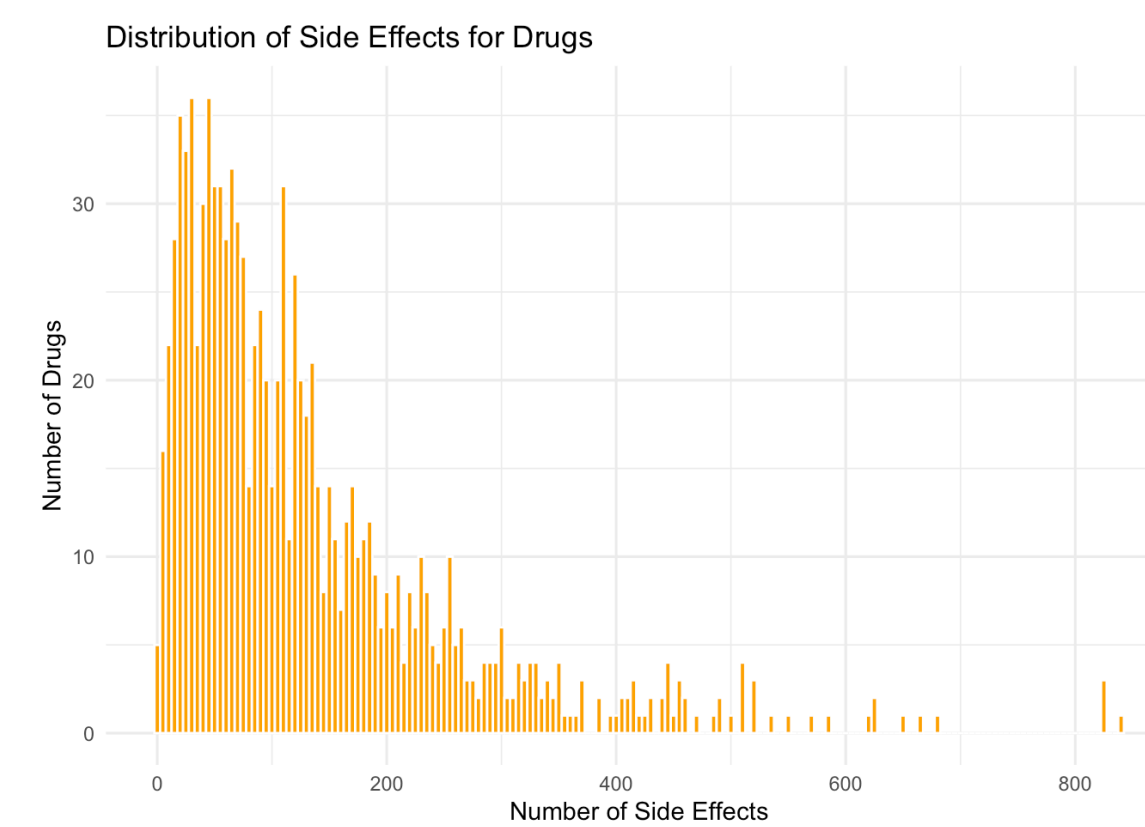
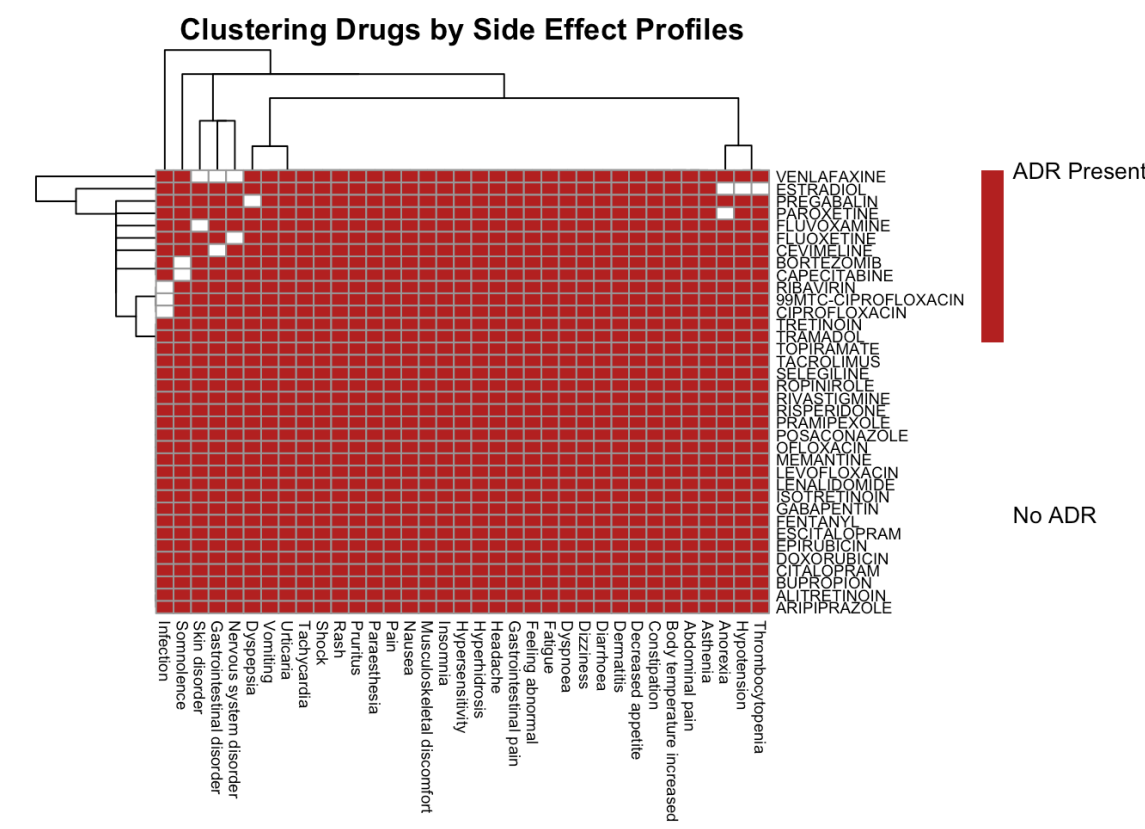
1. Explain the ADR profile prediction problem and its statistical challenges in imbalanced, noisy health data.
2. Explore low-rank and kernel-based methods (e.g. NMF+kernel regression and alternatives) for ADR prediction.

## Datasets

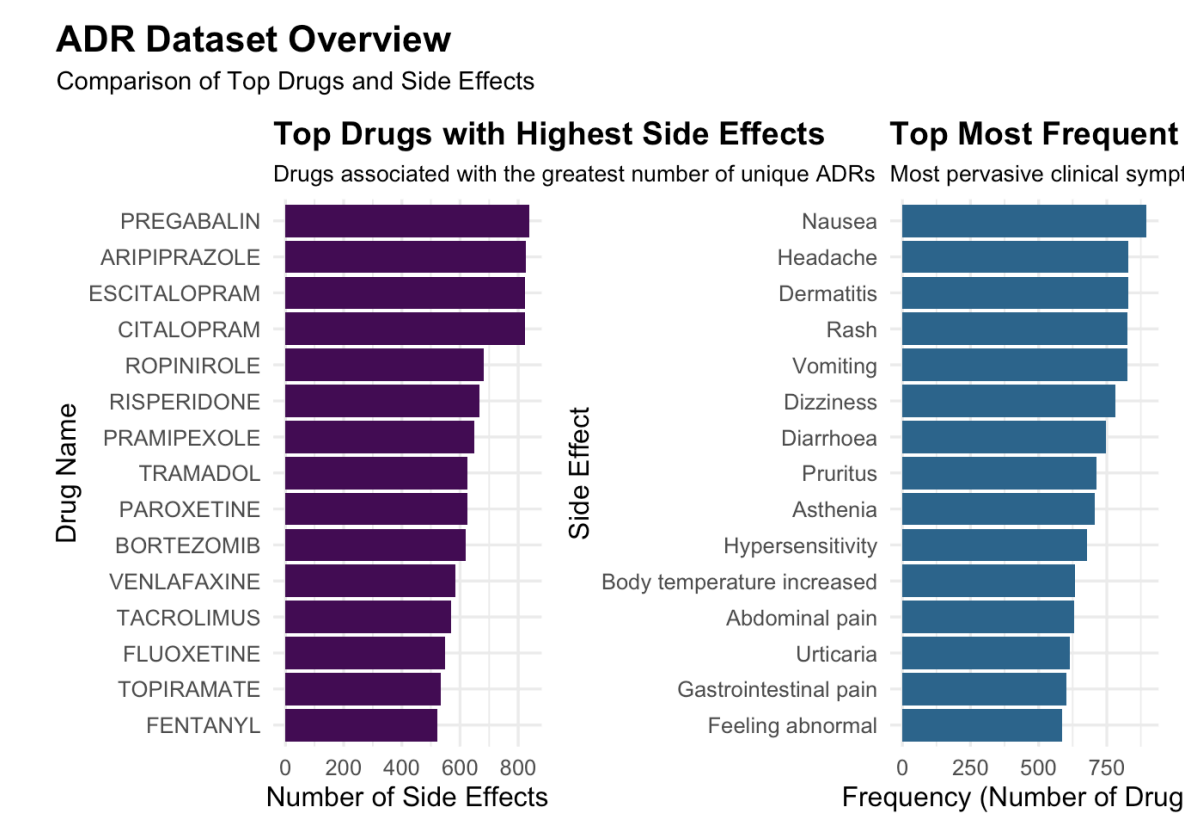
The primary source of datasets are DGIdb 4.0, SIDER 4.1 and PubChem database.

1. Drug-gene interaction pair: Intersection of drugs from DGIdb 4.0 and SIDER 4.1, generated to the binary matrix form.
2. Chemical fingerprints: Data from PubChem database, generated to the binary matrix form.
3. Drug & side effects: Drug along with it's side effects are extracted from SIDER 4.1 database.

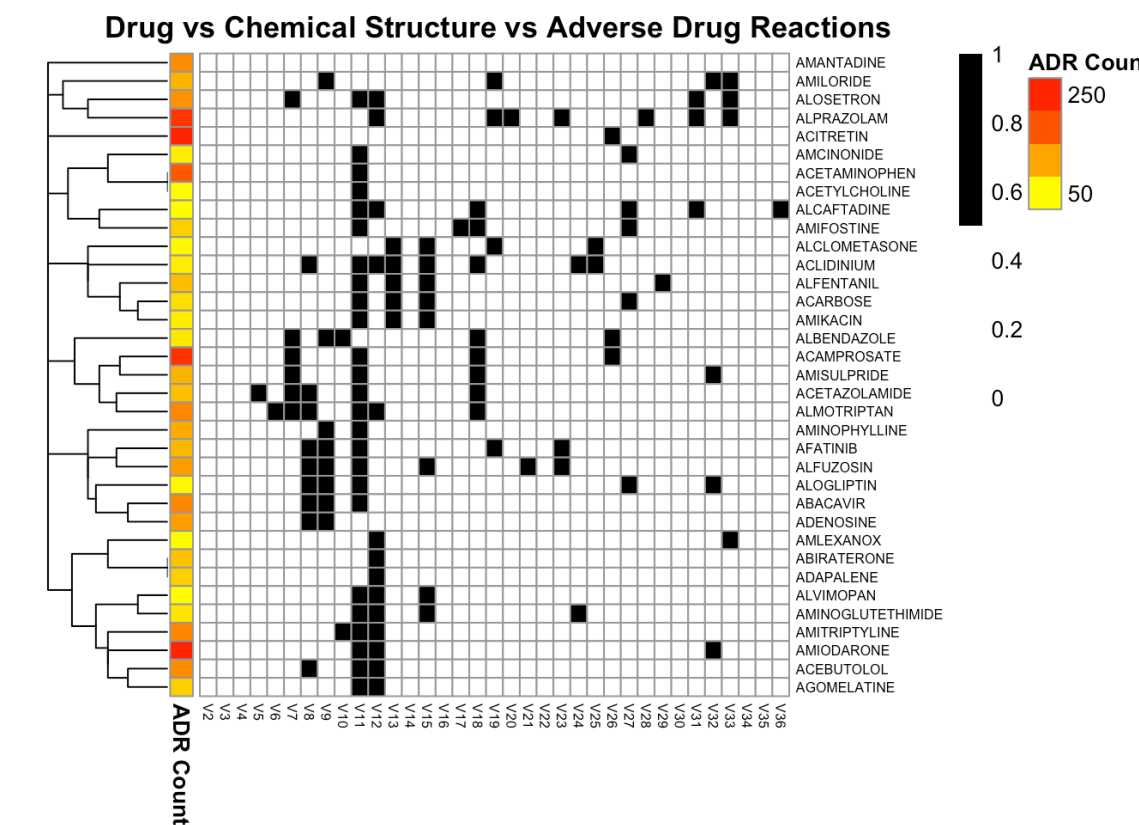
## Descriptive Statistics



The above graph mentions the drugs majority of the drugs have more than 200 side effects, emphasising the importance of this study.



From the dataset, these are the drugs which possess the most side effects displayed alongside the most common side effects.

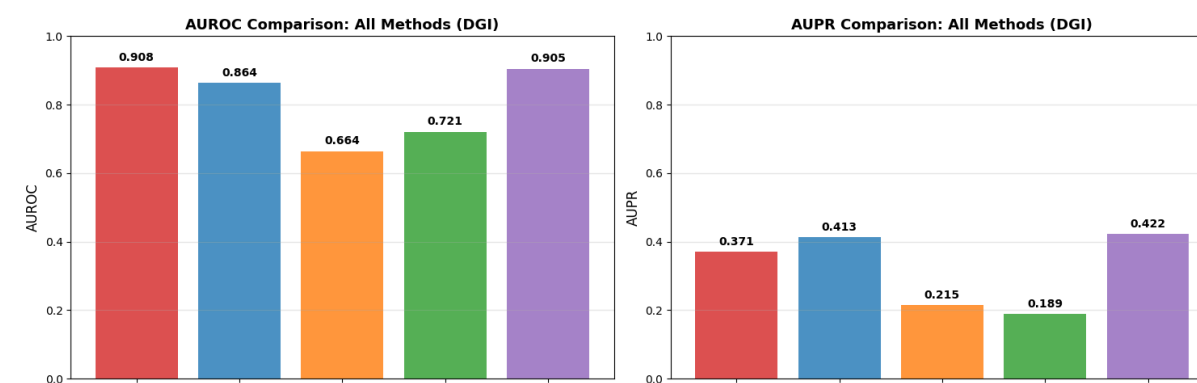


## Methods

1.We implemented five ADR profile prediction methods using DGI features only: Naïve frequency model (ADR prevalence per column), kernel regression (KR), linear SVM, RBF-kernel SVM, and VKR (NMF + kernel ridge regression).

## Early Results

Preliminary analysis on the DGI dataset shows that the Naïve baseline and VKR achieve the highest AUROC ( $\approx 0.91$ ), while KR and VKR achieve the best AUPR ( $\approx 0.41$ – $0.42$ ), clearly outperforming SVM variants on both metrics. VKR therefore provides the best overall trade-off between discrimination (AUROC) and rare ADR detection (AUPR), motivating its use as the main reference method in further experiments.



## Future Work

We plan to conduct further analysis using:

- Extended Data set with latest drug available from SIDER 4.1 & DGIdb database <sup>1</sup>
- New methods using weighted NMF, SVD and weighted SVD <sup>2</sup>.

We will use the **PYTHON** Programming for this.

## GitHub

The code and datasets for this project can be viewed at our GitHub repository here: <https://github.com/arshad4387/ADR-Prediction.git>

## Acknowledgement

We would like to thank Dr.Yezhao Zhong, Cathal Seoighe , Dr.Haixuan Yang for their work in ADR prediction and sharing the code and data through the github page.

## References

[1] Zhong, Y., Seoighe, C., & Yang, H. (2024). Non-Negative matrix factorization combined with kernel regression for the prediction of adverse drug reaction profiles. Bioinformatics Advances, 4(1), vbae009.

1. <https://doi.org/10.1093/nar/gkv1075>
2. <https://doi.org/10.3390/rs13020268>