Bayesian tensor factorization for predicting clinical outcomes

Predictive utility of human genetics evidence

Onuralp Soylemez

Global Blood Therapeutics

onuralp@gmail.com

Abstract

- Only 1 in 10 drug candidates receive marketing approval. Majority of drug candidates fail due to safety concerns or lack of efficacy. Increasingly available high dimensional information on targets, drug molecules and indications provides an opportunity for ML methods to integrate multiple data modalities and better predict clinically promising drug targets.

- Notably, drug targets with human genetics evidence are shown to have higher approval success rates. However, a recent tensor factorization-based approach found that additional information on targets and indications might not necessarily improve the predictive accuracy underscoring the importance of feature selection and data curation [1].

- Here we revisit this approach by integrating different types of human genetics evidence collated from publicly available sources to support each target-indication pair.

- We use Bayesian tensor factorization to show that models incorporating all available human genetics evidence (rare disease, gene burden, common disease) modestly improves the clinical outcome prediction over models using single line of genetics evidence.

Data and Methods

We used three lines of human genetics evidence based on disease variant frequency to support the statistical and biological association between human genetic variation in a drug target and their impact on medical outcomes. All the data used in this analysis are publicly available on Open Targets Platform [2]. Detailed information on each data source is available on Github at **cx0/icml-human-genetics**.

Evidence type	Description
Rare disease	List of curated genes with established causal link between gene and disease.
Gene burden	Gene-based rare variant associations in UK Biobank using whole exome sequencing data.
GWAS	Prioritization of causal genes at GWAS locus based on genetic and functional genomics features using locus-to-gene (L2G) model.
Combined evidence	Integrating human genetics evidence from all three types of evidence.

We created rank-3 tensors with each mode referring to drug targets, indications and human genetics evidence, respectively, and used Bayesian probabilistic matrix factorization using MCMC [3] to factorize the binary matrices as implemented in SMURFF, a highly optimized framework for Bayesian tensor factorization (BTF) [4].

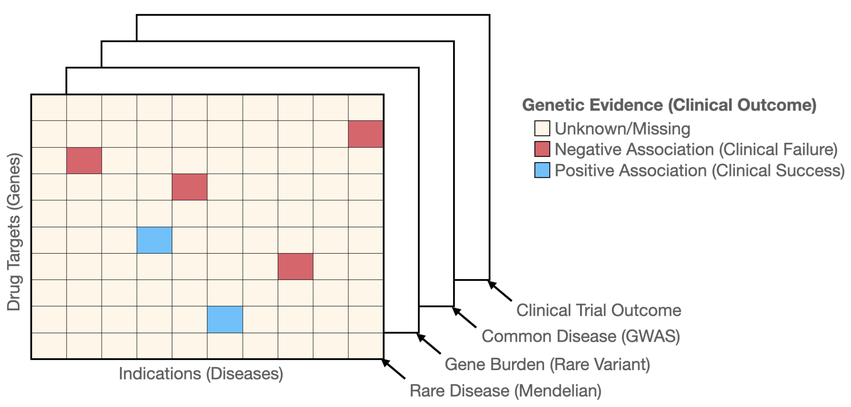

Figure 1: Schematic representation of rank-3 tensor for the 'combined evidence' model.

Results

We evaluated the predictive performance of each model using AUROC in a random 80/20 training-test split, and the model with combined evidence across three lines of human genetics evidence performed slightly better than the other models (see **Table 1**).

NLP-based classification of clinical trial stop reasons yielded a small conservative set of negative outcomes resulting in significant class imbalance between clinical success and failure. To address the class imbalance, we also computed F1 scores for each model.

Model/Evidence	AUROC	F1 score	Imbalance
Rare disease	93.2 $\pm$ 0.3	96.6 $\pm$ 0.2	87.2%
Gene burden	92.6$\pm$ 0.3	81 $\pm$ 0.6	2.5%
GWAS	93.3$\pm$ 0.2	95.4 $\pm$ 0.2	39.4%
Combined	94.5$\pm$ 0.2	98.1$\pm$ 0.1	29.3%

Table 1: Classification accuracies for the models considered in this study. F1 score was calculated using a threshold of 0.5. Class imbalance shows the proportion of positive labels out of total labels for the respective model.

We corroborate the previous finding that target-indication pairs from Phase 3 are enriched for validated or de-risked drug targets and therefore have higher probability of success. clinical trials at later stages are more likely to succeed [1]

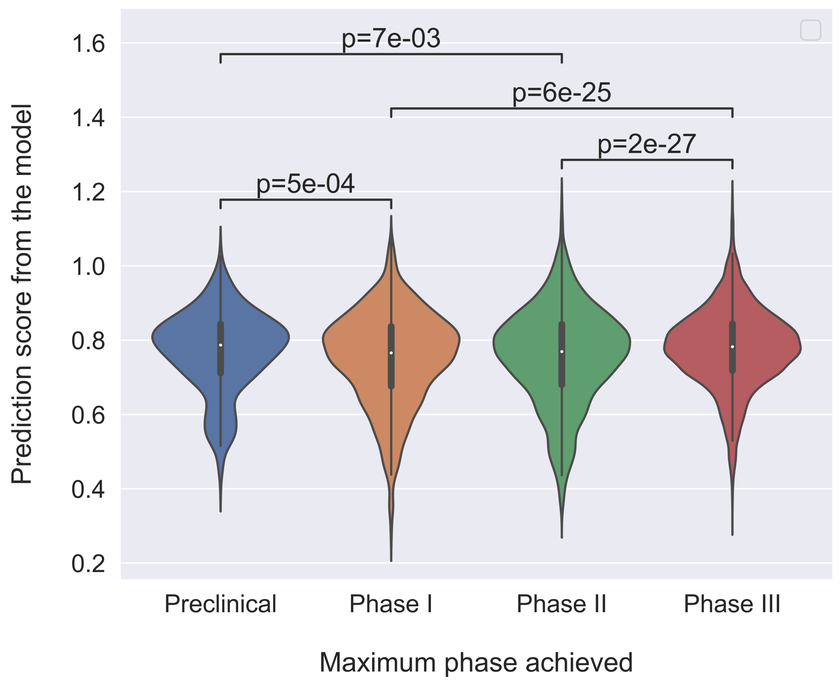

Figure 2: Bayesian tensor factorization model prediction scores from the best performing model ('combined model'). Each target-indication pair was grouped by the maximum clinical phase reached. Preclinical phase refers to research compounds that have not made to Phase I clinical trials yet. P-values were calculated using two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction.

Conclusions and Future Research

- BTF offers a scaleble approach to handling data scarcity when integrating different lines of human genetics evidence. Combined evidence appears to have a very modest improvement on the predictive accuracy when predicting clinically promising drug targets.
- It is conceivable that the poor predictive performance of the burden model is largely due to high class imbalance in this model as well as relatively few available labels. Further research and cross-validation strategies are necessary to probe whether this class of genes with burden evidence biologically represent difficult-to-target genes for therapeutic modulation (e.g., highly selective targeting) or empirical significance thresholds for these genes are too conservative.
- We relied on NLP classification for labeling the clinical trial outcomes, however, it is very likely that text-based classifications do not completely capture the complex nature of a particular trial failure. There is significant need for better metadata standards for clinical trial outcomes to improve the effectiveness of semantic analysis.

References

[1] J. Yao et al. Predicting clinically promising therapeutic hypotheses using tensor factorization. *BMC Bioinformatics*, 20(69), 2019.

[2] D. Ochoa et al. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acid Research*, 49:1302–1310, 2021.

[3] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 880–887, 2008.

[4] Tom Vander Aa, Imen Chakroun, and Thomas J. Ashby. Smurff: a high-performance framework for matrix factorization. In *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 304–308, 2019.

Acknowledgements

We are grateful to the Open Targets team and public/private partner institutions for their commitment to open data sharing.