Bayesian tensor factorization for predicting clinical outcomes using integrated human genetics evidence

Onuralp Soylemez [1]

Abstract

The approval success rate of drug candidates is very low with the majority of failure due to safety and 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 better odds to succeed. However, a recent tensor factorization-based approach found that additional information on targets and indications might not necessarily improve the predictive accuracy. 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. We provide additional insight into the relative predictive power of different types of human genetics evidence for predicting the success of clinical outcomes.

1. Motivation

The approval success rate of drug candidates is less than 10% with most of clinical trial failures attributed to safety concerns and a lack of clinical efficacy (Cook et al., 2014). Increasingly available high dimensional information on targets, drug molecules and indications provides an opportunity for ML methods to better predict clinically promising drug targets. Notably, drug targets with human genetics evidence are twice more likely to be approved [(Nelson et al., 2015), (King et al., 2019)] and the most recent drug approvals from FDA corroborate this strong trend (Ochoa et al., 2022).

However, a recent tensor factorization-based approach from (Yao et al., 2019) found that additional information on targets and indications might not necessarily improve the predictive accuracy underscoring the importance of feature selection and evidence data quality. Here we revisit this approach by integrating different lines of human genetics evidence collated from publicly available sources and assess the relative predictive performance of models incorporating different types of human genetics evidence.

2. Data curation

Open Targets Platform curates and maintains target-disease evidence by harmonizing information on genetic diseases, genetic variation, clinical trial outcomes, gene expression data and biomedical literature (Ochoa et al., 2021). 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 (see Table 1).

2.1. Rare genetic diseases

We used a list of curated genes with reasonably well-established causal link with a disease. This class of genetics evidence is enriched for genes associated with rare Mendelian diseases whereby very rare variants with large effect and high penetrance are the causative genetic alteration underlying the disease or clinical manifestation. We leveraged expert manual curation of diagnostic-grade gene-disease relationships from ClinGen (Strande et al., 2017) and Genomics England (Martin et al., 2019) to annotate target-indication (gene-disease) pairs.

2.2. Gene-level burden association

Gene-level collapsing methods combine information from genetic variants found at appreciable frequencies in the general population and assess the statistical strength of association between the aggregate variation and health outcomes. We collated a list of significant gene burden associations across thousands of target-disease pairs from UK Biobank.

[1]Global Blood Therapeutics, South San Francisco, CA.. Correspondence to: Onuralp Soylemez <onuralp@gmail.com>.

The 2022 ICML Workshop on Computational Biology.

Public-private partner institutes analyzed whole-exome sequencing data from more than 400,000 UK Biobank participants to identify genes with coding variants that are collectively enriched in individuals with a selected set of medical outcomes [(Backman et al., 2021), (Wang et al., 2021), (Karczewski et al., 2022)]. Any target-indication pair that did not reach the empirical significance threshold in the respective study was labeled as negative association rather than missing.

2.3. Evidence from GWAS

Genome-wide association studies (GWAS) test the statistical association between common genetic variants and diseases, and have identified many disease-variant associations that recapitulate known disease biology as well as nominate novel therapeutic hypotheses. While GWAS have become an indispensable tool in drug discovery for novel drug target identification and validation, prioritization of causal genes among dozens of candidate genes with equally compelling biological explanation remains a significant challenge.

We leveraged a recently developed prioritization model, 'locus-to-gene' (L2G) model, that integrates human genetics from GWAS Catalog and UK Biobank, known target and disease biology and multi-omic datasets (Ghoussaini et al., 2020), (Mountjoy et al., 2021). We used the scoring predictions from the L2G model from Open Targets Platform to annotate drug targets that are predicted to be the causal genetics factor for a disease or phenotype.

2.4. Clinical trial outcomes

For clinical trial outcome data, we followed a similar label annotation procedure as employed in (Yao et al., 2019). Specifically, we labeled every drug target - indication pair as "approved" (positive label) if there is at least one drug molecule that has been approved for the corresponding indication. For the remaining target-indication pairs, a "failure" status (negative label) was assigned if there is at least one clinical trial for the pair that was either terminated or suspended. Additionally, we leveraged the data from Open Target Platform's NLP-based classification of clinical trials to annotate the reason for trial failure. Clinical trials that were inferred to be unfavorable due to safety concerns or efficacy were also assigned negative labels.

2.5. Disease ontology

To facilitate the integration of indication data from multiple sources, we mapped the EFO (experimental factor ontology) IDs for each indication to MeSH IDs using EBI-EMBL ontology cross-reference database (OxO) (Malone et al., 2010).

Table 1. Description of the three lines of human genetics evidence used in this analysis.

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.

3. Model description

Given a binary matrix of drug targets (genes) and clinical outcomes (success, failure, unknown), our goal is to impute the unknown cells or missing entries using interrelationships among targets and indications. We considered four models incorporating human genetics evidence either individually or altogether. Specifically, 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 (Salakhutdinov & Mnih, 2008) to factorize the binary matrices as implemented in SMURFF, a highly optimized framework for Bayesian tensor factorization (Aa et al., 2019).

For each tensor factorization, we built a model with 32 latent dimensions and used burn-in of 500 samples for the Gibbs sampler. We collected 3500 samples from the model, and kept every 350th sample and averaged the predictions from these samples for the final prediction.

4. Results

We evaluated the predictive performance of each model using AUROC, and the model with combined evidence across three lines of human genetics evidence performed slightly better than the other models (see Table 2). 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. In particular, the discrepancy between AUROC and F1 scores for the gene burden model highlights the

Table 2. 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.

MODEL/EVIDENCE	AUROC	F1 SCORE	IMBALANCE
RARE DISEASE	93.2 ± 0.3	96.6 ± 0.2	87.2%
GENE BURDEN	92.6± 0.3	81 ± 0.6	2.5%
GWAS	93.3± 0.2	95.4 ± 0.2	39.4%
COMBINED	94.5± 0.2	98.1± 0.1	29.3%

dramatic class imbalance for this model. It is very likely that a non-trivial fraction of target-indication pairs may reach statistical significance when larger sample sizes and more refined definition of indications are considered. Alternatively, more nuanced set of rare and common variants with overlapping burden signal can be considered (Weiner et al., 2022).

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 (Yao et al., 2019) (see Figure 1).

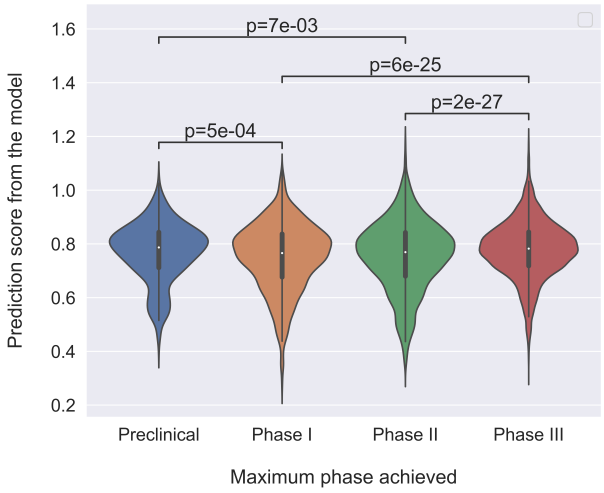

Figure 1. Bayesian tensor factorization model prediction scores from the best performing model ('combined model'). Each target-indication pair was grouped by the maximum clinical phase reached. Preclinical phase refers to research compounds that have not made to Phase I clinical trials yet. P-values were calculated using two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction.

Interestingly, we also find that registered trials at preclinical stage (research compounds) appear to have better odds of success than trials that have progressed (i.e., Phase 1 and Phase 2) suggesting that sponsors consider validated drug-indication pairs increasingly more often in their early drug discovery research programs.

5. Discussion

Here we used publicly available data on approved drug indications and human genetics evidence for the corresponding drug targets to predict the outcome of ongoing clinical trials. Notably, we used Bayesian tensor factorization with additional information from different lines of human genetics evidence to support these targets. Our results show that the model with combined evidence (rare disease, gene burden, common disease) modestly improves the accuracy of predicting clinically promising drug targets when compared to alternative models with single line of evidence. While this finding is encouraging for the increasing appreciation for the role of human genetics in drug target discovery and validation efforts, substantial class imbalance due to necessity of expert manual curation poses a significant challenge for model comparison and establishing benchmarks for further method development.

While the approved drug indications may be considered as true positive labels, there are numerous reasons why the outcome of a clinical trial was not favorable. Even when a clinical trial meets its primary objectives, trial sponsors may choose not to move forward with the trial due to business reasons, rapidly changing standard of care or anticipation of difficulty to enroll eligible patients. Here we relied on NLP classification for labeling the clinical trial outcome data, however, it is very likely that text-based classifications may not completely capture the nature of a particular trial failure. There is significant need for better documentation and structure of clinical trial data to improve the effectiveness of text-based classification and semantic analysis.

Choosing the most appropriate strategy to integrate different lines of human genetics evidence remains to be an active area of research in drug discovery (e.g., (Stacey et al., 2019), (Dornbos et al., 2022)). Historically, Mendelian genetics evidence has proven to be a convenient source of strong causal evidence between drug target and indication where the evidence implicates a single mutation or gene as the molecular cause underlying the disease. However, as DNA sequencing has become increasingly inexpensive, genomics studies in much larger populations and more complex medical conditions have begun to yield human genetics evidence for gene-disease associations in the form of hundreds of mutations with individually marginal effects on health outcomes and complex traits. While there are significant advances in statistical and computational approaches to combining these effects for patient stratification (see (Khera et al., 2018)), integrating

evidence from genetics associations from rare and common diseases has proven to be difficult.

Understanding the relevant importance of different sources of human genetics evidence for predicting clinically promising drug targets will significantly help develop safe and effective therapies. In the experimental setup presented in this study, we built multiple models to consider each source of human genetics evidence separately and combined to compare the relative predictive performance of each model. Notably, the burden model performed the poorest. It is conceivable that the poor predictive performance is largely due to high class imbalance in this model as well as relatively few available labels. Further research is necessary to probe whether this class of genes with burden evidence biologically represent difficult drug targets for therapeutic modulation (e.g., highly selective targeting) or empirical significance thresholds for these genes are too conservative.

Integration of matrix (or tensor) factorization approaches with neural networks has proven to be very useful for predicting gene expression from highly structured data modalities such as genomic and epigenomic data (Schreiber et al., 2020). In case that there are substantial non-linear relationship between different sources of human genetics data, this approach can be useful for predicting trial outcome data using informative latent representation of genetics evidence.

Data Availability

All the data used in this analysis are publicly available on Open Targets Platform (Ochoa et al., 2021): https://platform.opentargets.org/downloads. Data on human genetics evidence and clinical trial outcomes were downloaded from the latest release of the platform (v22.06). Detailed information on each data source is available at https://github.com/cx0/icml-human-genetics.

Acknowledgements

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

References

Aa, T. V., Chakroun, I., Ashby, T. J., Simm, J., Arany, A., Moreau, Y., Van, T. L., Dzib, J. F. G., Wegner, J., Chupakhin, V., Ceulemans, H., Wuyts, R., and Verachtert, W. Smurff: a high-performance framework for matrix factorization. 2019. URL https://arxiv.org/abs/1904.02514.

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., and Center, R. G. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599:628–634, 2021.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., and Pangalos, M. N. Lessons learned from the fate of astrazeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13:419–431, 2014.

Dornbos, P., Singh, P., Jang, D.-K., Mahajan, A., Biddinger, S. B., Rotter, J. I., McCarthy, M. I., and Flannick, J. Evaluating human genetic support for hypothesized metabolic disease genes. *Cell Metabolism*, 34(5):661–666, 2022.

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M., Koscielny, G., Machlitt-Northen, S., Malangone, C., Pendlington, Z. M., Roncaglia, P., Suveges, D., Wright, D., Vrousgou, O., Papa, E., Parkinson, H., MacArthur, J. A. L., Todd, J., Barrett, J. C., Schwartzentruber, J., Hulcoop, D., Ochoa, D., McDonagh, E. M., and Dunham, I. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49: 1311–1320, 2020.

Karczewski, K. J., Solomonson, M., Chao, K. R., Goodrich, J. K., Tiao, G., Lu, W., Riley-Gillis, B. M., Tsai, E. A., Kim, H. I., Zheng, X., Rahimov, F., Esmaeeli, S., Grundstad, A. J., Reppell, M., Waring, J., Jacob, H., Sexton, D., Bronson, P. G., Chen, X., Hu, X., Goldstein, J. I., King, D., Vittal, C., Poterba, T., Palmer, D. S., Churchhouse, C., Howrigan, D. P., Zhou, W., Watts, N. A., Nguyen, K., Nguyen, H., Mason, C., Farnham, C., Tolonen, C., Gauthier, L. D., Gupta, N., MacArthur, D. G., Rehm, H. L., Seed, C., Philippakis, A. A., Daly, M. J., Davis, J. W., Runz, H., Miller, M. R., and Neale, B. M. Systematic single-variant and gene-based association testing of thousands of phenotypes in 426,370 uk biobank exomes. *medRxiv*, 2022.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., and Kathiresan, S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224, 2018.

King, E., Davis, W., and Degner, J. Are drug targets with genetic support twice as likely to be approved? revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genetics*, 15(12), 2019.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., and Parkinson, H. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

Martin, A. R., Williams, E., Foulger, R. E., Leigh, S., Daugherty, L. C., Niblock, O., Leong, I. U., Smith, K. R., Gerasimenko, O., Haraldsdottir, E., et al. Panelapp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature genetics*, 51(11):1560–1565, 2019.

Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M. A., et al. An open approach to systematically prioritize causal variants and genes at all published human gwas trait-associated loci. *Nature Genetics*, 53(11):1527–1533, 2021.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., Cardon, L. R., Whittaker, J. C., and Sanseau, P. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47:856–860, 2015.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., Baker, J., Ferrer, J., Raies, A., Razuvayevskaya, O., Faulconbridge, A., Petsalaki, E., Mutowo, P., Machlitt-Northen, S., Peat, G., McAuley, E., Ong, C. K., Mountjoy, E., Ghoussaini, M., Pierleoni, A., Papa, E., Pignatelli, M., Koscielny, G., Karim, M., Schwartzentruber, J., Hulcoop, D. G., Dunham, I., and McDonagh, E. M. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acid Research*, 49:1302–1310, 2021.

Ochoa, D., Karim, M., Ghoussaini, M., Hulcoop, D. G., McDonagh, E. M., and Dunham, I. Human genetics evidence supports two-thirds of the 2021 fda-approved drugs. *Nature reviews. Drug discovery*, 2022.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pp. 880–887, 2008. URL https://doi.org/10.1145/1390156.1390267.

Schreiber, J., Durham, T., Bilmes, J., and Noble, W. S. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome biology*, 21(1):1–18, 2020.

Stacey, D., Fauman, E., Ziemek, D., Sun, B., Harshfield, E., and Wood, A. Progem: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acid Research*, 47, 2019.

Strande, N. T., Riggs, E. R., Buchanan, A. H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S. S., Goldstein, J., Ghosh, R., Seifert, B. A., Sneddon, T. P., et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *The American Journal of Human Genetics*, 100(6):895–906, 2017.

Wang, Q., Dhindsa, R., Carss, K., Harper, A., and Initiative, A. G. Rare variant contribution to human disease in 281,104 uk biobank exomes. *Nature*, 597:527–532, 2021.

Weiner, D. J., Nadig, A., Jagadeesh, K. A., Dey, K. K., Neale, B. M., Robinson, E. B., Karczewski, K. J., and O'Connor, L. J. Polygenic architecture of rare coding variation across 400,000 exomes. *medRxiv*, 2022. URL https://www.medrxiv.org/content/early/2022/07/07/2022.07.06.22277335.

Yao, J., Hurle, M. R., Nelson, M. R., and Agarwal, P. Predicting clinically promising therapeutic hypotheses using tensor factorization. *BMC Bioinformatics*, 20(69), 2019.