

Functional Gene Embeddings for Quantitative Genetics

Ege Kurt¹

¹TUM School of Computation, Information and Technology (CIT), Technical University Munich

Abstract

The yeast "*Saccharomyces cerevisiae*" is broadly utilized as a model organism to study genetic mechanisms. Its extensive genomic resources and compact genome, compared to more complex organisms, make it an excellent basis for developing machine learning models that can generalize to higher-order biological systems. The first part of this project focuses on various prediction tasks within the yeast genome by integrating multiple data sources and machine learning models.

Building on this baseline, the second phase extends functional gene embedding approaches to human genomics by integrating RNA embeddings derived from the Orthrus study. Within the scope of human genomics, we aim to predict disease-associated genes with a primary focus on those related to cancer. By integrating multiple data sources into gene embeddings, we convert composite biological information into numerical vector representations that can be efficiently used in machine learning applications.

Introduction

Comprehending gene function and genetic interactions remains a major challenge in modern genomics. Genes act within complex networks that determine cellular behavior, adaptation, function, and other processes such as disease susceptibility.

Saccharomyces cerevisiae, commonly known as baker's yeast, is one of the most studied eukaryotic organisms. It acts as an excellent model organism for studying these networks due to its compact genome, well-characterized genetic landscape, and suitability for high-throughput experimentation [1]. These properties facilitate the development of computational models that can generalize to more complex systems, including humans.

For predictive models, we use gene embeddings, which represent biological entities as numerical vectors that capture functional and relational information across multiple data sources. They are known as an efficient way to integrate biological data into machine learning models, enhancing their ability to interpret gene function and relationships [2].

The first part of this project focuses on yeast genomics, where diverse data modalities, including promoter and UTR sequence embeddings from DNALM [3], yeast functional interaction networks from YeastNet [4], gene expression vectors from Yeast Phenome [5], and orthologous groups from STRING [6], were combined to generate gene embeddings. These embeddings were then used to predict genetic interactions, double mutant fitness scores, and gene essentiality classification based on data from the landmark

genetic interaction map established by Costanzo et al.

In the second part of the project, the focus shifts from yeast to human genomics, not by directly applying yeast-derived techniques, but by building upon recent advances in RNA-based gene embeddings introduced in the Orthrus study [7]. Particularly, it integrates an existing benchmarking framework for functional gene embeddings [2] with the Orthrus RNA foundation model.

The embeddings were extracted from the Orthrus RNA language model, which was trained using contrastive self-supervision on RNA sequences to capture functional and evolutionary relationships across transcripts. These embeddings are used both as standalone predictors and in combination with different multi-omic datasets to improve disease-gene prediction.

Predicting Genetic Interactions in *Saccharomyces Cerevisiae* Using Gene Embeddings

Understanding how genes interact to control cellular functions is one of the central challenges in modern genomics. While individual gene effects can often be measured directly, many biological traits and phenotypes arise from the combined action of multiple genes. In *Saccharomyces cerevisiae*, large-scale genetic interaction maps, such as those established by Costanzo et al.[8], provides a rich foundation for

studying these complex relationships. However, the vast number of possible gene combinations and the incomplete coverage of experimental datasets make it difficult to fully characterize the genetic interaction network.

The goal of this part of the project is to explore whether machine learning models can learn patterns within the yeast genome and predict gene–gene interactions based on existing knowledge. First evaluating the predictive structure of the interaction data itself, and later incorporating gene embeddings derived from multiple biological data sources, this project aims to assess how well numerical representations can capture functional dependencies among genes.

Datasources

Costanzo et al. (2016) – Global genetic interaction network Genetic interactions reveal functional relationships between genes. The Costanzo et al. (2016) dataset [8] represents the most comprehensive map of genetic interactions in *Saccharomyces cerevisiae* currently available. In the study, the genes were grouped based on their essentiality status. Essential genes are those whose deletion results in loss of viability, whereas knocking out nonessential genes does not prevent survival ability.

Using systematic Synthetic Genetic Array (SGA) analysis, the study quantifies pairwise gene–gene interactions, covering almost all essential and nonessential genes in yeast. Each interaction is represented by an epsilon (ϵ) score that measures the deviation of the observed double-mutant fitness from the expected fitness based on the two single mutants. Negative scores indicate synthetic sick or lethal interactions, whereas positive scores represent epistatic suppression or alleviating effects.

We utilized this genetic interaction network data, along with additional complementary data sources, to develop and evaluate several predictive tasks, including:

1. Predicting genetic interactions from existing patterns within the network
2. Predicting double mutant fitness scores
3. Predicting interaction scores
4. Classifying interaction types

Genomic sequences The first datasource consists of DNA sequence data extracted from a large, multi-species genomic dataset by Karoliuss, Gagneur, and colleagues, which was initially used for DNA language models [3]. The dataset includes extensive annotations for protein-coding genes, as well as non-coding regions, which play an essential role in gene regulation.

The upstream sequences contain promoters and transcription factor binding sites, that regulate gene transcription, while the downstream regions include untranslated regions (UTRs) and RNA-binding protein motifs, which influence mRNA stability, localization, and translation efficiency [9].

YeastNet YeastNet¹ is a probabilistic network describing functional relationships among coding genes in *Saccharomyces cerevisiae*. The network models similarity to neighboring genes in the network, and modulatory roles across different cellular states [4]. Furthermore, it integrates diverse data sources to predict functional associations between genes.

Gene embeddings derived from the YeastNet network were generated using the node2vec algorithm, implemented through the *pecanpy* framework by Liu and Krishnan. [10]

StringDB The *STRING*² systematically integrates protein-protein interactions from diverse sources, including experimental data, computational predictions, curated pathways data, and text mining of scientific literature. It provides a vast platform for protein-protein interaction networks and functional enrichment analyses, which are broadly utilized in genomics research. [6] The STRING database was incorporated using one-hot encodings to represent orthologous relationships between genes within the network.

YeastPhenome The Yeast Phenome³ dataset is resource for studying the phenotypic effects of gene knockouts in *Saccharomyces cerevisiae*. It aggregates data from approximately diverse knockout screens, encompassing causal gene-to-phenotype links [11]. In this project, solely gene expression data from the Yeast Phenome dataset was employed to construct embeddings.

Methods

In this phase of the project, we will utilize multiple machine learning methods for the analysis. For numerical predictions, the methods PCA combined with Linear Regression, Ridge Regression, Random Forests, Gradient Boost Regression, and Neural Networks are used. For the classification task, Logistic Regression and Random Forest Classifier were used.

Linear Regression combined with PCA provides a simple and interpretable baseline for predicting continuous outcomes. Ridge Regression extends this

¹ <https://www.inetbio.org/yeastnet/>, accessed on 2025-10-30

² STRING: functional protein association networks, <https://string-db.org/> database, accessed on 2025-10-30.

³ Yeast Phenome: A comprehensive resource for genome-wide knockout phenotypic screens, <https://www.yeastphenome.org/>, accessed on 2025-01-22.

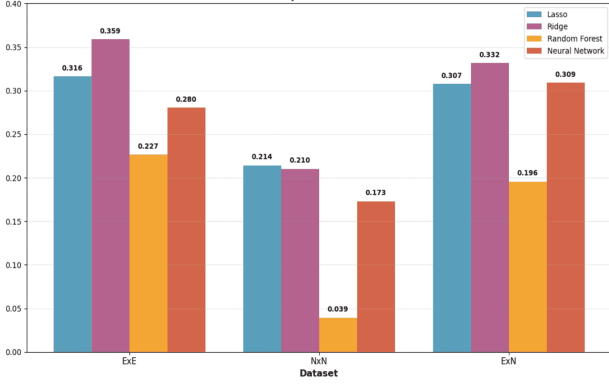


Figure 1: R^2 score comparison across Models and Interaction Classes.

by adding regularization. It stabilizes predictions by preventing large coefficient values [12]. Random Forest Regression improves robustness by bagging multiple decision trees, whereas Gradient Boosting Regression optimizes predictive performance by sequentially correcting previous trees by bagging one new tree at a time [13]. Neural Networks, while requiring careful hyperparameter tuning, are effective in learning complex, high-dimensional patterns, and are powerful for intricate data structures [14].

For classification, we employ a similar approach, utilizing a basic and interpretable baseline model, Logistic Regression, and more complex models, Random Forest Classifier and XGBoost Classifier, which use bagging to improve predictions in the presence of complex data patterns.

An 80–20 train–test split was applied to all datasets in this part of the project. Further methodological decisions were made for the prediction tasks, but because each subtask differs in scope, the corresponding settings and model choices are discussed individually in the respective sections.

Results

Predicting Interactions from Interactions

The first task focused on predicting interactions directly from the genetic interaction matrix itself, before moving on to any embedding-based prediction. The motivation behind the first step was to determine whether the interaction data contained intrinsic patterns that could be learned and predicted from within itself.

The interaction types were analyzed separately in essential–essential (ExE), nonessential–nonessential (NxN), and essential–nonessential (ExN) combinations. The reason for that is better modality, as well as the variant availability and distribution of data between these groups. As input and target features, genetic interaction scores derived from large-scale

yeast screens by Constanzo et. al. [8] were used. Genetic interaction scores assess the fitness effects observed in double mutants, where each value represents the interaction strength between a pair of genes. For each category, models were trained using an 80–20 train–test split, and parameter tuning was performed via grid search. Individual results were evaluated using leave-one-out cross validation approach, where each gene was iteratively held out as the test instance while the remaining genes were used for training.

As demonstrated in Figure 1: among the models used, Ridge Regression performed best overall, achieving stable predictions for ExE, ExN, and NxN. It suggests that the relationships between gene interaction profiles are mostly linear. However, regularization is necessary for handling potential collinearity and noise in the data. In contrast to Ridge Regression, Lasso Regression enforces stronger sparsity by driving less informative coefficients of features to zero, which in our case resulted in slightly lower performance. This suggests that a broader set of weakly contributing features jointly explain gene–gene interaction patterns better than feature elimination. Random Forest models underperformed compared to both linear approaches. As tree-based models exhibit overfitting tendencies despite hyperparameter optimization, in our case, this happened most likely due to the high dimensionality of the data. Neural Network models achieved competitive performance but showed higher variability across datasets, likely because their training is more sensitive to parameter initialization and data size.

Higher scores observed for ExE and ExN pairs suggest that essential genes demonstrate more consistent interaction patterns compared to nonessential genes, which tend to be more variable. Finally, we can conclude that genes with similar interaction profiles exhibit predictable relationships to some extent. This provides a proper baseline for following embedding-based predictions, which aim to integrate multimodal data sources for capturing deep biological insights.

Predicting Double Mutant Score From Embeddings

The second task was to predict double mutant fitness (DMF) scores directly from gene embeddings. The DMF score represents the observed fitness of a double knockout relative to the expected value based on the fitness of the individual single mutants [8]. To perform this prediction, gene embeddings derived from the combination of the data sources DNALM, YeastNet, and Yeast Phenome were used as input features, and DMF scores as target features. Alongside multiple regression models, different combinations of

data sources (such as DNALM, DNALM + YeastNet, and DNALM + Yeast Phenome) were evaluated to compare their predictive strengths and to scrutinize consistency between models and data sources.

The predictive models used in this task were Ridge Regression, Random Forest, and XGBoost, with hyperparameter optimization performed through grid search. Additionally, two different evaluations were done using different data splitting approaches. The first was a standard split, in which interactions were randomly divided into training and test sets. The second was a more stringent gene holdout split, where all interactions involving a selected subset of genes (accounting for 20% of the total interactions) were completely excluded from the training data. The goal of this setup was to evaluate our embeddings' ability to generalize to completely unseen genes and to investigate whether consistent patterns could still be identified based on similarities in their interaction profiles.

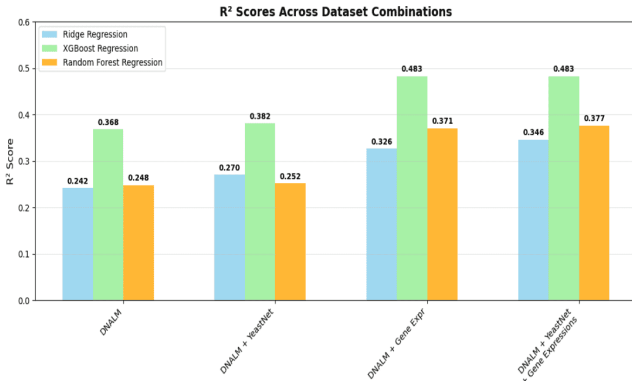


Figure 2: R^2 Score Distributions of Data Sources For Standard split

Under the standard split, embedding-based models achieved valid predictive performance, and among the evaluated models, Ridge Regression and XGBoost produced the best results. These results show that the embeddings capture a meaningful portion of the variance in double mutant fitness. In particular, embeddings generated from gene expression and DNA sequence features contributed most to prediction accuracy when they were added, as can be seen from Figure 2.

In the more challenging and interesting gene holdout setting, predictive performance decreased. The drop in accuracy was expected, as the model had no prior exposure to test genes and their interactions. Nevertheless, it is noteworthy that the models trained on embedding features were still able to predict interactions between genes that they had never seen during training. The results are demonstrated in Figure 3

Despite declining performance when generalizing

to unseen genes, the relative performance of the models and the dataset combinations remained consistent. With Ridge Regression and XGBoost outperforming other methods, and the dataset combination of DNALM, YeastNet, and Gene expression outperforming others. This consistency implies that the learned representations encode generalizable biological information rather than dataset-specific artifacts.



Figure 3: R^2 Score Distributions of Data Sources For Holdout split

Taken together, the results suggest that the information derived from multimodal biological data sources, when converted to numerical gene embeddings, encodes functional relationships relevant to genetic interactions.

Predicting Gene Interaction Score From Embeddings

In this segment, we focus on predicting the gene interaction scores between two genes, which reflect the deviation in double-mutant fitness relative to the expected additive effects of single mutants [8]. For this task, again, we used Ridge Regression, Lasso, Random Forest, and XGBoost models.

All models yielded R^2 values very close to zero, meaning that the models' predictions were not able to explain the variability in the true data. Figure 4 showcases this as true versus predicted values appear randomly distributed and do not show any correlation. The R^2 results are provided in the appendix.

Interestingly, in contrast to this, the double mutant fitness scores were predictable using the same approach. The inability to predict interaction scores, despite moderate performance on DMF prediction, raises questions about the internal consistency of the dataset or the stability of the derived interaction score metric. It might indicate that the transformation from double and single mutant fitness scores to interaction scores introduce either additional noise or nonlinearities that cannot be captured in the embedding space.

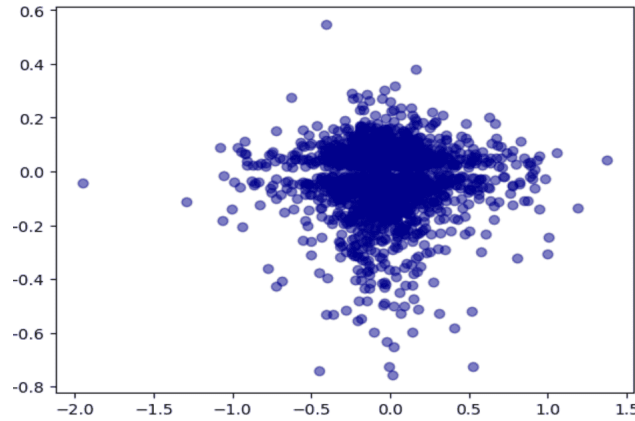


Figure 4: *Cloud-like Distributions of True and Predicted scores.*

Predicting Interaction classifications From Embeddings

To assess whether embeddings could still distinguish interaction types, the task was reformulated as a classification problem. Interactions were categorized as negative, neutral, or positive, using the thresholds from Costanzo et al [8]. With $e > 0.08$ being positive, $e < -0.08$ being negative, and everything else being neutral.

If we look at the distribution of the dataset, which is available in the appendix as a graph, we observe a quantitative dominance of neutral interactions. Out of a total of 933,148 interactions, 567,151 were classified as neutral, 233,621 as negative, and only 132,868 as positive interactions. This strong class imbalance meant that the models were exposed to far more neutral examples during training, and might lead them to favor neutral predictions over positive or negative ones.

For this task, XGBoost and Random Forest classifiers were trained with an 80–20 split and evaluated on precision, recall, and F1 scores. As expected, the Random Forest model tended to predict nearly all interactions as neutral, failing to correctly classify positive and negative interactions. XGBoost was better at distinguishing between positive and negative interactions but made more errors when predicting neutral ones. Nevertheless, XGBoost achieved a better overall performance across all classes whilst considered as a whole. The results are demonstrated in Figure 5.

Discussion

While the models performed well in several cases, successfully predicting interactions from the interaction matrix itself and double mutant fitness scores using embeddings as input features, the results for interaction classification were less successful in com-

parison. Moreover, predicting quantitative interaction scores directly from gene embedding features did not achieve meaningful results.

Our primary motivation was to understand phenotypes arising from genetic interactions. However, our application scope has an inescapable limitation; we focused only on interactions of genes in pairs, even though, in reality, most phenotypes are shaped by the coordinated action of many genes. In such cases, numerous genes have cumulative effects, making it challenging to point out individual contributions. As the number of interacting genes increases in a phenotype, the number of potential combinations grows exponentially. Such scenarios would be computationally infeasible, but this is not an immediate concern, as we are already limited by data availability. Currently, datasets capturing higher-order interactions are not available, which constrains our analysis to pairwise interactions.

Furthermore, additional biological factors need to be considered. Even mutations in nonessential genes may not directly cause lethality, but can produce unpredictable downstream effects that disrupt other parts of the genome, which can then lead to fatal consequences over an extended period. Such indirect consequences illustrate the complexity of genetic networks once again and highlight the importance of considering indirect effects in future analyses.

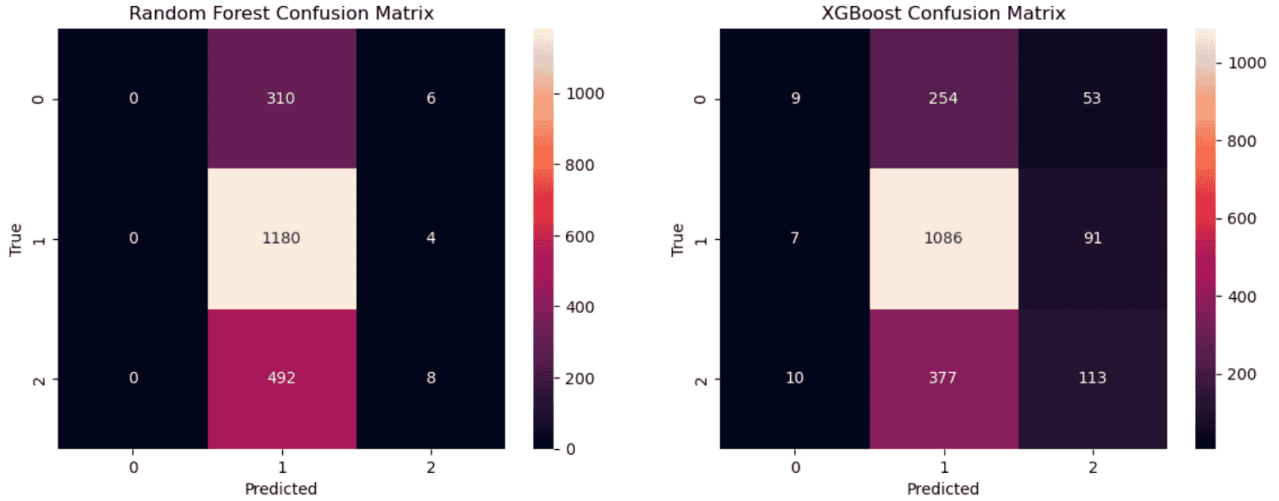


Figure 5: Confusion Matrix Comparison of Random Forest and XGBoost Models

Predicting Disease Genes in *Homo Sapiens* Using Gene Embeddings

Complex diseases such as cancer originate from disruptions in multiple biological processes; therefore, it is crucial to understand gene function beyond single experimental measurements. But most traditional approaches rely on manually curated knowledge or single data modalities.

Functional Gene Embeddings addresses this by transforming diverse data modalities such as gene expression, protein-protein interactions, and literature evidence into numerical embeddings while being easy to integrate with modern machine learning approaches [2]. Nevertheless, current methods and datasets are often biased toward well-studied genes and fail to generalize to unexplored regions of the genome.

RNA language models are a new, state-of-the-art approach beyond prior gene- or protein-level embeddings; they capture transcript-level diversity, splicing variation, and evolutionary conservation across species. Building on this outline, this part of the project integrates RNA-based gene embeddings derived from the Orthrus foundation model. This approach aims to provide fresh predictive power for disease-associated and cancer-related genes, especially those that are understudied or novel.

Orthrus

The Orthrus model is a large-scale RNA foundation model designed to learn functional and evolutionary relationships across transcriptomes. It is pretrained on splicing isoforms and orthologous transcripts from over 400 mammalian species, supplying

a broad phylogenetic scope [7]. Unlike conventional sequence models that rely on token prediction, Orthrus is trained by comparing pairs of RNA transcripts to learn their relative functional similarity. It uses a contrastive learning framework that positions functionally and evolutionarily related transcripts close together in the embedding space.

Through this structure, Orthrus can capture information on RNA structure, evolutionary constraints, and regulatory patterns. Each transcript is encoded as a six-track input, consisting of a 4-base one-hot nucleotide base sequence (A, C, G, U) along with splice-site and codon position tracks. The model is trained on a massive scale, comprising 49 million transcripts and 870 million contrastive pairs.

Originally, Orthrus embeddings were developed and benchmarked for several RNA-level predictive tasks, including RNA half-life prediction (transcript stability), mean ribosome load (translation efficiency), gene ontology, molecular function classification, protein subcellular localization, and transcript structural feature prediction, such as UTR lengths, exon counts, and coding region length.

Datasources

To benchmark the predictive performance of Orthrus RNA embeddings, we used multiple diverse datasets for comparison and combined them with. For this purpose, we adopted the same data sources used in the benchmark evaluations from Brechtmann et al. [2].

The predicted genes stem from different sources. While cancer-associated genes are derived from a large and comprehensive study, genes related to other diseases are sourced from a variety of independent resources. Even though all diseases except cancer

are being evaluated within a unified benchmarking workflow. In the following, a concise description of each data source will be provided.

OMICS Embeddings The Omics embedding dataset is a collection of experimental data, comprising gene activity, essentiality, and protein-level information. It combines GTEx: human tissue-wide gene expression profiles [15], DepMap: CRISPR-Cas9 essentiality screens across hundreds of human cancer cell lines [16], ProtT5: Protein sequence embeddings derived from a large protein language model, containing structural and functional properties of proteins [17].

StringDB As explained in the first phase of the project, the STRING database provides a comprehensive view of protein–protein interactions through both experimental and knowledge-based associations⁴. In addition to data for *Saccharomyces cerevisiae*, it also contains information on *Homo sapiens* genes.

In the evaluation pipeline, there are two versions of this database: STRING Full, which combines all available evidence, and STRING Experimental, which includes only experimentally validated interactions.

PoPS (Polygenic Priority Score) Features PoPS is a gene-level feature representation datasource integrating diverse functional evidence. It provides information relevant to disease heritability and the interpretation of GWAS results [18]. Again, as in the STRING datasource, there are two versions of this database: PoPS Full, which includes all available data, and PoPS Experimental, which restricts features to experimental and expression-based data.

EMOGI Cancer Gene Predictions The EMOGI (Explainable Multi-Omics Graph Integration) provides a reference dataset for pan-cancer gene prediction. It integrates four omics data modalities: Single-nucleotide variants, Copy-number alterations, DNA methylation studies, and Gene expression [19]. EMOGI uses the structure of protein–protein interaction networks to find cancer-related genes that are affected by pathway disruptions.

Multiple Disease Driver Gene Datasources The cancer-associated genes used in this study were from the Pan-Cancer Analysis of Whole Genomes project [20]. This large-scale study analyzes 2,658 whole cancer genomes spanning 38 tumor types.

The non-cancer disease gene sets were collected from multiple studies covering a diverse range of disorders. Datasets originate from published sources such as Schlieben et al.(2020) [21] for mitochondrial

disorders, Frésard et al.(2019) [22] for ophthalmological and neurological conditions, Rapaport et al.(2021) [23] for inborn errors of immunity, Gonorazky et al.(2019) [24] for neuromuscular diseases, and Wang et al.(2017) [25] for epilepsy.

Methods

The first methodological choice was about the model used. Two tree-based methods were considered: Random Forest and XGBoost. Both are ensemble learning algorithms that are known for performing well on structured, tabular biological datasets. Random Forests are generally easier to use, as it requires minimal hyperparameter tuning and are robust against overfitting. It can perform reliably across various types of data; however, it cannot directly handle missing values. This often results in larger model sizes due to the high number of trees, and may lead to biased feature importance scores. XGBoost is an algorithm that can achieve higher predictive accuracy by sequentially minimizing model bias and variance when properly tuned. It can handle missing values natively and typically yields better generalization performance. But XGBoost introduces greater complexity: it involves many hyperparameters, is more computationally expensive, and carries a higher risk of overfitting.

In each prediction task, both models were tested, and XGBoost consistently outperformed Random Forest in every experiment, with further details provided in the appendix. Therefore, it was selected as the primary model, and all subsequent predictions and evaluations were conducted using XGBoost. Just as in the referenced benchmarks [2], hyperparameters of XGBoost were optimized through nested cross-validation, where the L1 and L2 regularization parameters in the XGBoost loss function were tuned over a log-uniform grid ranging from 10^{-2} to 10^{-5} for both parameters.

In the cancer gene prediction task, the predictor datasets included Omics, STRING (Full and Experimental), EMOGI, and Orthrus embeddings, resulting in 12,874 common genes after intersecting all datasets (764 cancer and 12,110 non-cancer genes). For the prediction of other disease-associated genes, the same workflow was applied using Omics, STRING (Full and Experimental), PoPS (Full and Experimental), and Orthrus embeddings, yielding 17,208 common genes across all sources.

⁴ STRING: functional protein association networks, <https://string-db.org/> database, accessed on 2025-10-30.

Results

Cancer Genes Prediction

The first task focused on predicting cancer genes. As described in the Methods section, we performed nested cross validation with an 80–20 training–test split.

Models based on STRING achieved the highest performances with AuPRC values of 0.424 and 0.443, respectively, followed by the experimental-only version of STRING and EMOGI predictors. When evaluated individually, Orthrus embeddings reached an AuPRC of 0.134, which, although noticeably lower than the other feature sources, still performed above the random baseline. This demonstrates that Orthrus captures biologically meaningful RNA-level information even when used alone. However, combining Orthrus embeddings with other datasets yields only minor improvements. The results are demonstrated in Figure 6.

These findings indicate that, although Orthrus embeddings encode valuable RNA-level information, their contribution to cancer gene prediction is limited within the current approach.

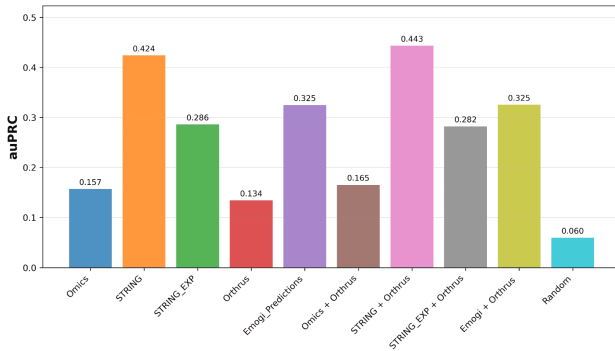


Figure 6: AuPRC Comparison of Embeddings in Cancer Gene Predictions

Other Disease Predictions

The second task was closely associated with the first one and focused on predicting genes related to other diseases. As explained in the methods section, a very similar methodological framework was applied for this task.

Across all disease categories, STRING-based features again achieved the strongest overall performance, outperforming all other datasets in most cases, except for inborn errors of immunity category. PoPS predictors followed closely in overall performance. The highest average performances were observed for mitochondrial and neuromuscular disease gene predictions, with AuPRC values exceeding 0.5,

indicating that these datasets capture clear functional signals. The results are demonstrated in Figure 7.

Here, our results are highly consistent with those reported in the referenced benchmarks [2], with only small deviations that are likely caused by differences in data splitting during cross-validation. For comparison, the corresponding plot from the original study is provided in the appendix.

Here, it is generally important to emphasize that the dataset is highly imbalanced with rare positive cases, which may introduce bias into the predictors. The detailed proportions of positive and negative gene counts for each disease category are provided in the appendix.

When evaluated individually, Orthrus embeddings generally performed lower than other feature sources, with some minor exceptions, as can be seen in Figure 7. Similar to the cancer gene results, combining Orthrus with other datasets led to only marginal or no improvements. A detailed visualization of all combinations is provided in the appendix.

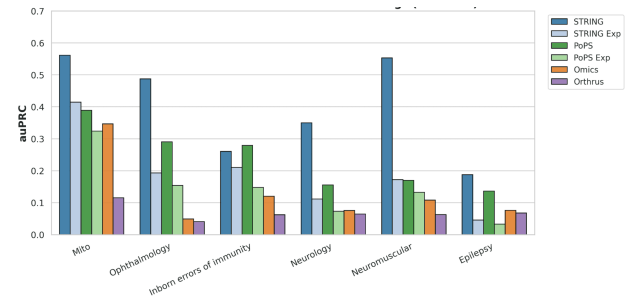


Figure 7: AuPRC Comparison of Embeddings in Multiple Disease Gene Predictions

Discussion

In this project, Orthrus RNA embeddings were evaluated for their ability to predict disease-associated genes, including both cancer and other complex disorders.

While Orthrus embeddings captured biologically meaningful signals, as they outperformed random baselines when used individually, their overall predictive power was limited compared to other data modalities such as STRING, PoPS, and EMOGI. Moreover, integrating Orthrus embeddings with other datasets did not yield significant improvements either.

Despite these limitations, the study established an extensible and ready pipeline that enables the integration and benchmarking of further diverse data sources. This project will facilitate future research and the evaluation of new potential embeddings.

In future work, it would be interesting to explore fine-tuning Orthrus embeddings at the gene level

or for phenotype-specific tasks, or to combine them with tissue-specific expression and regulatory data to better use the RNA-level information they encode. As RNA foundation models are a state-of-the-art approach and continue to evolve, their integration with functional embeddings may provide deeper insights into the molecular mechanisms underlying complex diseases or other prediction tasks.

References

- [1] Maria Parapouli et al. “Saccharomyces cerevisiae and its industrial applications”. In: *AIMS Microbiology* 6.1 (2020), pp. 1–31. DOI: 10.3934/microbiol.2020001. URL: <http://www.aimspress.com/journal/microbiology>.
- [2] Felix Brechtmann et al. “Evaluation of input data modality choices on functional gene embeddings”. In: *NAR Genomics and Bioinformatics* 5.4 (2023), p. 13. DOI: 10.1093/nargab/lqad095. URL: <https://doi.org/10.1093/nargab/lqad095>.
- [3] Alexander Karoliuss et al. “Species-aware DNA language models capture regulatory elements and their evolution”. In: *Genome Biology* 25.83 (2024). DOI: 10.1186/s13059-024-03221-x. URL: <https://doi.org/10.1186/s13059-024-03221-x>.
- [4] Hanhae Kim et al. “YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*”. In: *Nucleic Acids Research* 42.D1 (Oct. 2013), pp. D731–D736. ISSN: 0305-1048. DOI: 10.1093/nar/gkt981. eprint: <https://academic.oup.com/nar/article-pdf/42/D1/D731/16803898/gkt981.pdf>. URL: <https://doi.org/10.1093/nar/gkt981>.
- [5] Gina Turco et al. “Global Analysis of the Yeast Knockout Phenome”. In: *Science Advances* 9 (2023). DOI: 10.1126/sciadv.adg5702.
- [6] Damian Szklarczyk et al. “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D638–D646.
- [7] Philip Fradkin et al. “Orthrus: Towards Evolutionary and Functional RNA Foundation Models”. In: *bioRxiv* (2024). DOI: 10.1101/2024.10.10.617658. URL: <https://www.biorxiv.org/content/10.1101/2024.10.10.617658v1>.
- [8] Michael Costanzo et al. “A global genetic interaction network maps a wiring diagram of cellular function”. In: *Science* 353.6306 (2016), aaf1420. DOI: 10.1126/science.aaf1420. eprint: <https://www.science.org/doi/pdf/10.1126/science.aaf1420>. URL: <https://www.science.org/doi/abs/10.1126/science.aaf1420>.
- [9] Flavio Mignone et al. “Untranslated regions of mRNAs”. In: *Genome Biology* 3.3 (2002), reviews0004.1–reviews0004.10. DOI: 10.1186/gb-2002-3-3-reviews0004. URL: <http://genomebiology.com/2002/3/3/reviews/0004>.
- [10] Renming Liu and Arjun Krishnan. “PecanPy: a fast, efficient and parallelized Python implementation of node2vec”. In: *Bioinformatics* 37.19 (Mar. 2021), pp. 3377–3379. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab202. eprint: <https://academic.oup.com/bioinformatics/article-pdf/37/19/3377/50338217/btab202.pdf>. URL: <https://doi.org/10.1093/bioinformatics/btab202>.
- [11] Gina Turco et al. “Global Analysis of the Yeast Knockout Phenome”. In: *Science Advances* 9 (2023). DOI: 10.1126/sciadv.adg5702.
- [12] Christopher M. *Pattern Recognition and Machine Learning*. en. 1st ed. Information Science and Statistics. New York, NY: Springer, Aug. 2006.
- [13] William W. Hsieh. “Decision Trees, Random Forests and Boosting”. In: *Introduction to Environmental Data Science*. Cambridge University Press, 2023, pp. 473–493.
- [14] Jürgen Schmidhuber. “Deep Learning in Neural Networks: An Overview”. In: *Technical Report IDSIA-03-14/ arXiv:1404.7828* (2014). URL: <https://arxiv.org/abs/1404.7828>.
- [15] GTEx Consortium. “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509 (2020), pp. 1318–1330. DOI: 10.1126/science.aaz1776.
- [16] Robin M Meyers et al. “Computational correction of copy-number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells”. In: *Nature Genetics* 49.12 (2017), pp. 1779–1784. DOI: 10.1038/ng.3984.
- [17] Ahmed Elnaggar et al. “ProtTrans: Toward Understanding the Language of Life’s Code Through Self-Supervised Deep Learning and High-Performance Computing”. In: *arXiv arXiv:2007.06225* (2021). Preprint.

- [18] Elle M. Weeks et al. "Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases". In: *Nature Genetics* 55.8 (2023), pp. 1267–1276. doi: 10.1038/s41588-023-01443-6.
- [19] Rieke Schulte-Sasse et al. "Integration of multi-omics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms". In: *Nature Machine Intelligence* 3 (2021), pp. 513–526. doi: 10.1038/s42256-021-00342-6.
- [20] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. "Pan-cancer analysis of whole genomes". In: *Nature* 578 (2020), pp. 82–93. doi: 10.1038/s41586-020-1969-6.
- [21] Linda D. Schlieben and Holger Prokisch. "The dimensions of primary mitochondrial disorders". In: *Frontiers in Cell and Developmental Biology* 8 (2020), p. 600079. doi: 10.3389/fcell.2020.600079.
- [22] Laure Frésard et al. "Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts". In: *Nature Medicine* 25 (2019), pp. 911–919. doi: 10.1038/s41591-019-0433-3.
- [23] Florence Rapaport et al. "Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance". In: *Proceedings of the National Academy of Sciences of the United States of America* 118 (2021), e2001248118. doi: 10.1073/pnas.2001248118.
- [24] H. David Gonorazky et al. "Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease". In: *American Journal of Human Genetics* 104 (2019), pp. 466–483. doi: 10.1016/j.ajhg.2019.01.012.
- [25] Jia Wang et al. "Epilepsy-associated genes". In: *Seizure* 44 (2017), pp. 11–20. doi: 10.1016/j.seizure.2016.11.030.

Appendix

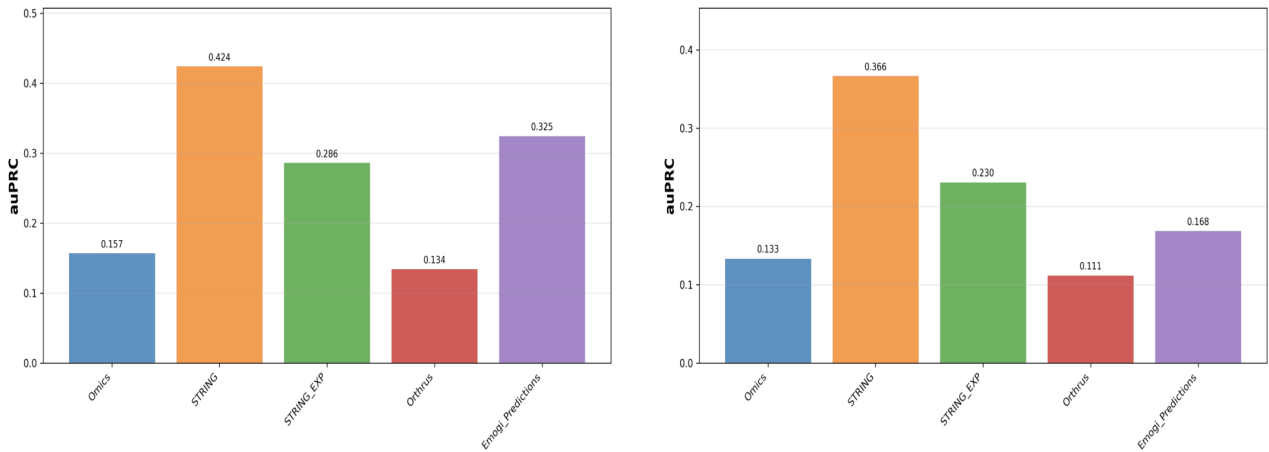


Figure 8: AuPRC Comparison of Embeddings in Cancer Gene Predictions XGBoost (left) vs Random Forests (right)

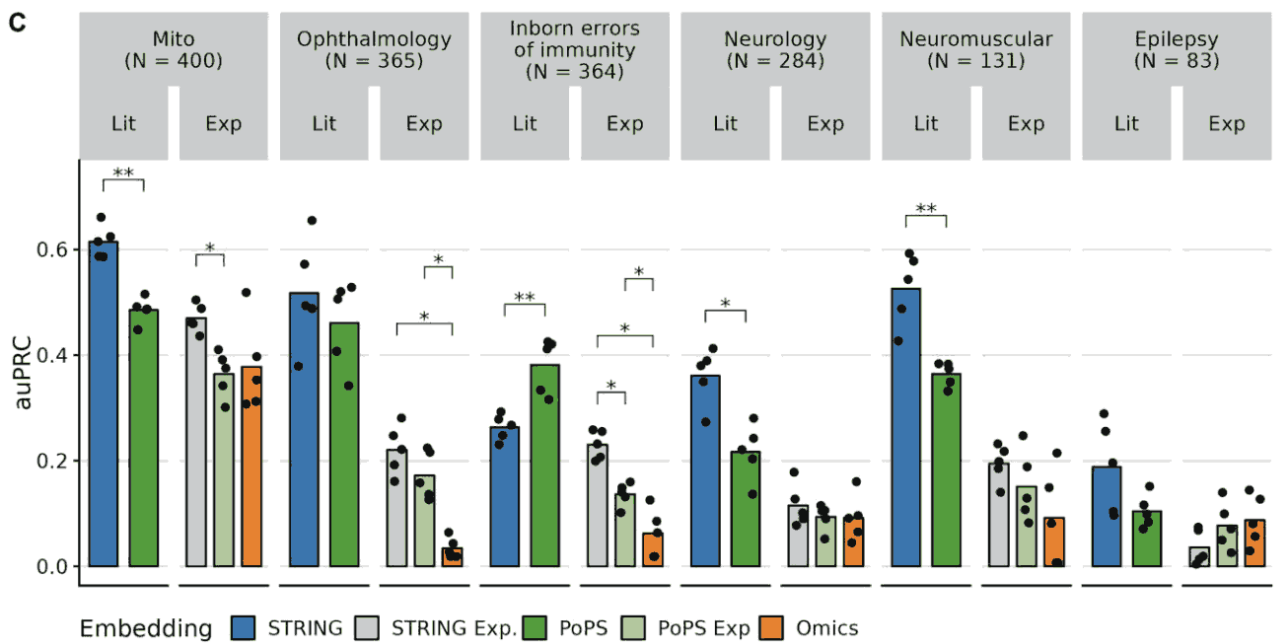


Figure 9: AuPRC Comparison of Embeddings in Multiple Disease Gene Predictions from [2]

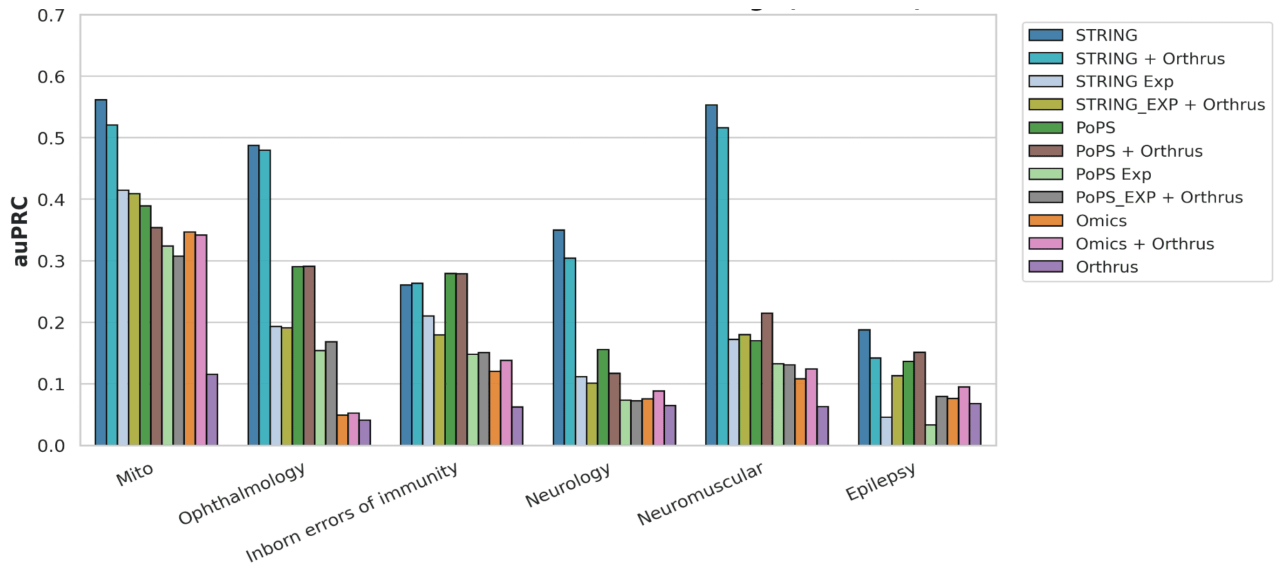


Figure 10: *AuPRC Comparison of Individual Embeddings and Combination With Orthrus in Multiple Disease Gene Predictions*

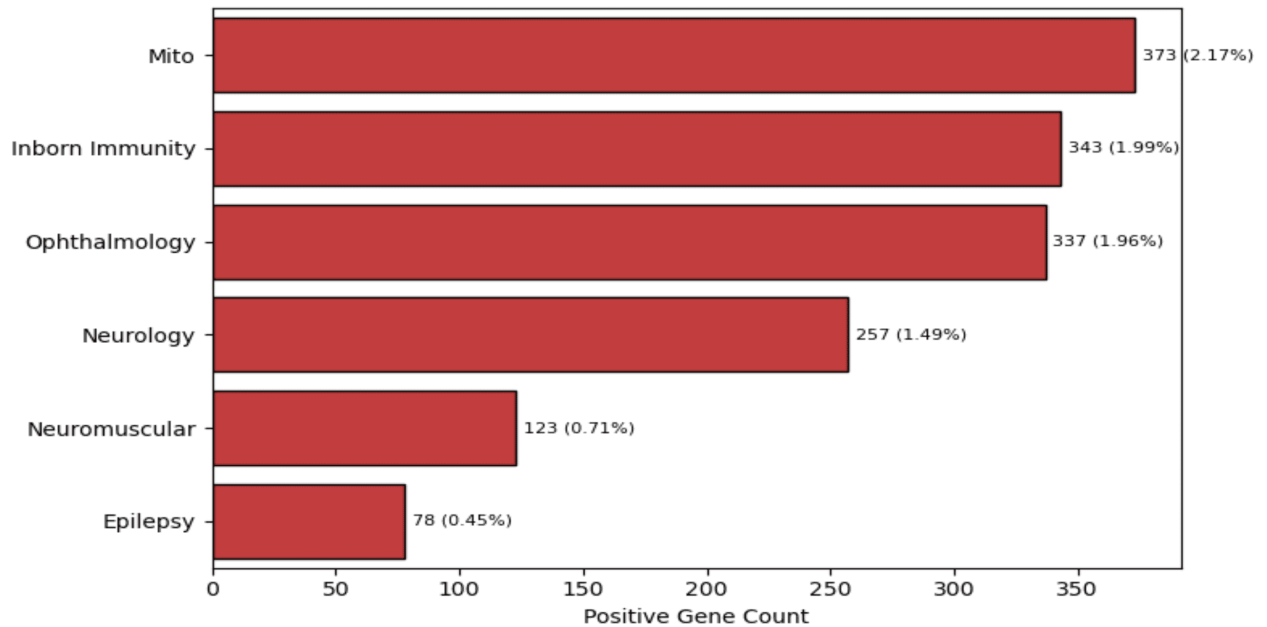


Figure 11: *Number and Proportion of Positive Disease-associated Genes Across Disease Categories*

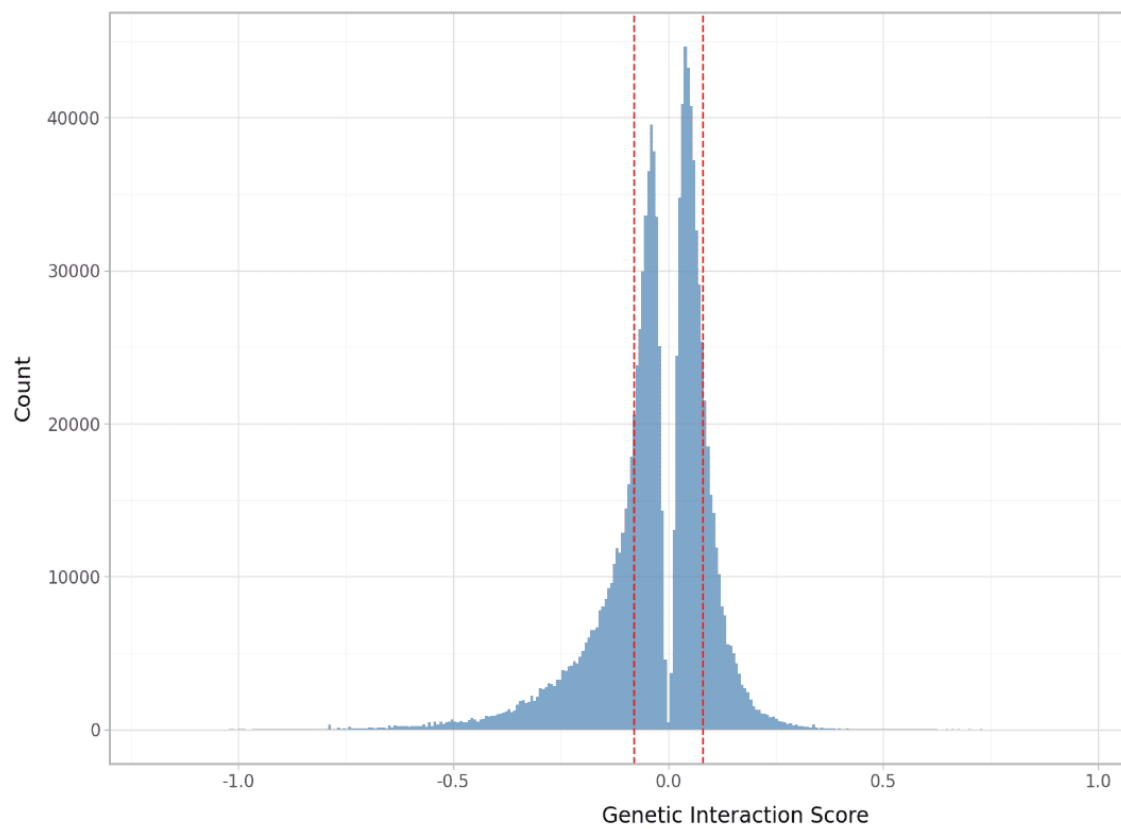


Figure 12: Histogram of Genetic Interaction Scores with Red Vertical Lines Marking the Cutoff Values for Positive and Negative Interactions