UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Prediction of genes associated with Autism Spectrum Disorder using sequence and graph embedding methods

João Pedro da Silva Inácio

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Doutor Hugo Filipe de Mesquita Costa Martiniano
Professora Doutora Astride Carolina Lentz de Moura Vicente

2024

Acknowledgements

This master's thesis is the culmination of a year's worth of hard work and dedication. While there have been a few hardships, I'm extremely happy with the end result and what I was able to accomplish. This achievement was made possible thanks to all the people who helped me along the way, I offer this as an homage to their support. First, I would like to thank my family for all their support, patience, and care this past year. Thank you for helping me cross the finish line and not give up. Although finishing a master's thesis is extremely fulfilling for someone who believes you can always improve further it is also a challenge, thank you for being there when I needed it the most. I would like to thank my adviser Dr. Hugo Martiniano for all the help and guidance since day one. Thank you for all the explanations simple to complex on those longer than expected meetings and for always being available. Theory crafting and finding ways to solve the problems that appeared along the process was extremely gratifying and discussing it with someone passionate about this field gave me the motivation to press on. I also want to thank my advisor Dra. Astrid Vicente for giving me the opportunity to work at INSA with the DPS team and for all the guidance throughout this year of hard work. I am grateful as well to all the members of the INSA DPS for making me feel at home and for their help, from new ideas to expert interpretations of results. I would like to thank my friends new and old, for all their support, honorable mention to my bioinformatics group not only for the good times at minicampos but for also being able to count on each other and help one another. Saving the best for last, I want to give a special thank you to my girlfriend Catarina. This past year was filled with highs and lows, and you were always there for me. Thank you for not letting me get too absorbed in my own world and future problems, you made me enjoy the present more. Thank you for listening to all my scrutinizing rants and for pushing me to face my insecurities. Thank you for being there and for being my girlfriend and best friend.

Abstract

Neurodevelopmental disorders impose a significant social and economic burden on individuals with these conditions and their families. Given that all neurodevelopmental disorders have a genetic component, identifying the risk genes for these disorders enhances our understanding of their etiology and can aid in the development of future screening methods and targeted therapies. Autism Spectrum Disorder (ASD) is a prototypical complex neurodevelopmental disorder characterized by high heritability and a heterogeneous genetic architecture and phenotypic presentation. This thesis presents a Machine Learning (ML) approach that improves upon state-of-the-art methods for ASD risk gene prediction, this thesis presents a Machine Learning (ML) approach capable of improving state-of-the-art methods for ASD risk gene prediction. To achieve this goal, a novel approach is created using publicly available ASD-associated genes and graph and sequence gene embeddings with supervised ML classifiers. Using a 5-fold nested stratified cross-validation, the pipeline achieved an AUC of 0.90, F1 of 0.82, and MCC of 0.77.

Additionally, the top decile of the ranked list of predicted risk genes generated by the model was significantly enriched for ASD phenotypes but not other brain-specific disorders. The proposed pipeline improved state-of-the-art approaches in predicting genes targeted by LOF mutations in the MSSNG and SCC studies. A functional network characterization of the top decile identified four distinct communities significantly enriched for biological pathways associated with ASD. Of the 50 top predicted genes by the pipeline, 37 were already present in ASD risk gene databases, while 13 were not yet linked to ASD. The 13 genes were significantly enriched in the cerebral cortex, and the telencephalon cell migration processes critical for brain development and linked to neurodevelopmental disorders. This thesis provides an accurate comparison of embedding methods for risk gene discovery and improves existing ASD risk gene predictions, taking a step closer to a better understanding of this complex genetic disorder.

Keywords: Autism Spectrum Disorder, Machine Learning, Graphs, Transformers, Embeddings.

Resumo

As perturbações do neurodesenvolvimento colocam um fardo social e económico significativo nos indivíduos afetados tal como nas suas famílias. Embora se tenham registado avanços significativos nas tecnologias de sequenciação de genes, ainda não foram descobertos todos os aspetos da arquitetura genética destas perturbações. Tendo em conta que perturbações do neurodesenvolvimento apresentam uma componente genética, é de elevada importância identificar os genes de risco responsáveis. Determinar os genes de risco não só aprofunda o nosso conhecimento sobre a etiologia de cada perturbação, como também contribui para o desenvolvimento de métodos de rastreio e terapias direcionadas. A Perturbação do Espectro do Autismo (PEA) é uma perturbação do neurodesenvolvimento complexa, altamente hereditária e heterogénea. Esta heterogeneidade cria desafios significativos na identificação de todos os genes de risco associados ao PEA. Métodos convencionais para identificar genes de risco do PEA, como o *Transmission and De Novo Association* (TADA), dependem de dados clínicos de pacientes com PEA, como os obtidos por *Whole Exome Sequencing* (WES). Embora eficazes, estes estudos são dispendiosos e os dados gerados são restritos devido às leis de proteção de dados, o que limita a partilha de dados e impossibilita a reprodução dos estudos. Dada a complexidade da PEA e as limitações dos métodos mais convencionais para a descoberta de genes de risco, esta tese adota uma abordagem de aprendizagem automática (AA).

O principal objetivo desta tese é melhorar os métodos *state-of-the-art* para a previsão de genes de risco do PEA. Para alcançar este objetivo, é criada uma abordagem que utiliza genes associados ao PEA disponíveis em bases de dados públicas com *embeddings* de genes, utilizando classificadores de aprendizagem automática supervisionada. Num problema de aprendizagem automática supervisionada é necessário definir um conjunto de dados de treino positivos e negativos. Neste caso os dados positivos vão ser genes que foram previamente associados com PEA e os dados negativos vão ser genes o mais distantes possível dos genes positivos. Os genes positivos usados nesta tese foram os genes da *Simons Foundation Autism Research Initiative* (SFARI) gene dataset. Esta base de dados é constituída por um conjunto de genes com evidencia bibliográfica de associação com PEA. Cada gene presente nesta base de dados é classificado de um a três. Os genes que têm a categoria um são os genes para qual existe uma maior evidencia biliografica da associação com PEA. Os de categoria três por sua vez são os que demostram ter menor evidencia. Seguindo a metodologia de abordagens *state-of-the-art*, nesta tese, os genes negativos usados foram os do artigo de *Krishnan et al.* de 2016. De modo a conferir que os genes negativos eram relevantes, foi realizada uma verificação utilizando os códigos de *ICD10* com uma base

de dados proteína-doença. Devido a novos estudos funcionais alguns dos genes negativos do artigo de 2016 estavam agora associados a genes do neurodesenvolvimento e como o objetivo dos genes negativos é ser o mais distante possível de genes do PEA esses genes foram removidos. Um dos maiores desafios nesta tese foi em encontrar a melhor maneira de representar genes numa forma compatível com os modelos de AA. Foram utilizadas duas abordagens, grafos e sequencias para criar as representações, ou *embeddings*. Para a abordagem de grafos foi utilizada a base de dados de interação proteína-proteína *STRING*, onde cada nó desse grafo corresponde a uma proteína. Foi criado para cada gene Embeddings utilizando os transcritos canónicos e diversos modelos de *embeddings*. Os transcritos canónicos são os transcritos que são considerados os principais de um gene por serem os mais expressos, serem os mais conservados, melhor caracterizados e por codificarem para a proteína funcional principal. No caso das representações de sequências, tal como as representações de grafos utilizei as sequencias de DNA e de aminoácidos de transcritos canónicos. Os *embeddings* de sequências de DNA foram gerados com o modelo *BERT*, *DNABERT-2*, e os *embeddings* de sequências de aminoácidos foram criados utilizando o modelo de aprendizagem profunda *PortT5*. Para fazer a previsão binária se um gene é ou não um gene de risco da PEA, nesta tese utilizei no total 6 modelos de AA de diferentes tipos. Os modelos utilizados foram modelos lineares como regressão logística, Support Vector Machines (SVM), modelos baseados em vizinhança como K-Nearest Neighbors (KNN), e modelos de ensemble como Random Forest, LightGBM e XGBoost. De modo a garantir previsões mais precisas dos modelos de AA testei diferentes permutações dos datasets de treino. Por exemplo, em vez de utilizar o dataset completo da SFARI como os genes positivos, utilizei apenas os genes da categoria um, os genes das categorias um e dois, entre outras combinações. A ideia é encontrar o melhor equilíbrio entre um dataset com genes mais específicos e um dataset não enviesado. Como o dataset de treino era limitado, nesta tese utilizei uma validação *nested stratified cross fold*. Métodos de *oversampling* por vezes são utilizados quando o dataset está enviesado, com mais elementos de uma categoria que a outra, contudo como neste caso como cada gene é altamente específico utilizar métodos de *oversampling* pode introduzir viés, que põe em causa a relevância das previsões do modelo. Deste modo, utilizar uma validação *nested stratified cross fold* não só possibilita o uso da totalidade dos dados para treinar e validar como assegura que o mesmo rácio de dados positivos e negativos são utilizados. Os resultados gerados pela *pipeline* bioinformática proposta demonstram a eficacidade e capacidade de esta *pipeline* em classificar com sucesso os genes de risco do PEA. Ao utilizar uma validação *nested stratified cross fold* foi possivel de obter um *AUC* de 0,90, *F1* de 0,82 e *MCC* de 0,77. O modelo que demostrou os melhores resultados foi a regressão logística com os *embeddings* de grafos *Deepwalk CBOW*. Usando esta combinação de modelo e embedding, a pipeline foi aplicada a todos os genes humanos disponíveis, gerando uma lista com cerca de 18000 genes. Cada gene tinha uma percentagem associada que representava a confiança do modelo em classificar o gene como de risco para PEA. Esta lista de genes foi ordenada pela percentagem de confiança, do maior para o menor valor, e posteriormente dividida em decis. O primeiro decil da lista ordenada de genes de risco previstos pelo modelo demostrou estar significativamente enriquecido para fenótipos de PEA mas não para outros Transtornos específicos do cérebro, deste modo demostrando a precisão da previsão. A pipeline proposta melhorou as abordagens *state-of-the-art* na previsão de genes associados a mutações de perda de função (LOF) presentes nos estudos MSSNG e SCC. Ao fazer uma caracterização funcional

de rede do primeiro decil foi possível identificar quatro comunidades distintas que demostraram estar significativamente enriquecidas para o desenvolvimento do sistema nervoso, organização e remodelação da cromatina, transdução de sinal intercelular e modificação de proteínas, processos diretamente ligados ao PEA. A identificação de comunidades foi feita utilizando o algoritmo de Leiden. Dos 50 genes de topo do primeiro decil da lista gerada pela pipeline, 37 genes já tinham sido previamente associados ao PEA, enquanto 13 genes ainda não tinham sido associados no momento de escrita desta tese. Ao fazer análise de enriquecimento funcional, treze dos genes demostraram estar significativamente enriquecidos para processos de migração no córtex cerebral e no telencéfalo, críticos no desenvolvimento cerebral e ligados a distúrbios do neurodesenvolvimento. Em conclusão, esta tese fornece uma comparação de métodos de embeddings para a previsão de genes de risco e melhora as abordagens existentes de previsão de genes de risco do PEA usando métodos de AA, contribuindo para uma melhor compreensão desta complexa perturbação.

Palavras Chave: Perturbação do Espectro do Autismo, Aprendizagem Automática, Grafos, Transformers, Embeddings.

Contents

List of Figures

List of Tables

Acronyms

PBBC Bioinformatics and Computational Biology Project

MBBC Master in Bioinformatics and Computational Biology

ASD Autism Spectrum Disorder

ML Machine Learning

TADA Transmission and De Novo Association

AUC Area Under the Curve

MCC Matthews Correlation Coefficient

LOF Loss of Function

SFARI Simons Foundation Autism Research Initiative

DSM-5-TR Diagnostic and Statistical Manual of Mental Disorders, Text Revision

WES Whole Exome Sequencing

BERT Bidirectional Encoder Representations from Transformers

PCA Principal Component Analysis

SVD Singular Value Decomposition

SVM Support Vector Machine

KNN K-Nearest Neighbors

LLM Large Language Model

RNN Recurrent Neural Network

CNN Convolutional Neural Network

NPL Natural Programming Language

STRING Search Tool for the Retrieval of Interacting Genes/Proteins

API Application Programming Interface

Chapter 1

Introduction

1.1 Motivation

Advancements in genome sequencing technologies have provided a large volume of information on the genetic causes of neurodevelopmental disorders. Understanding these disorders, particularly identifying associated risk genes, is crucial for improving our knowledge of their etiology and helping the development of future screening methods and targeted therapies. Among these disorders, Autism Spectrum Disorder (ASD) is a major focus in neurodevelopmental disorder research.

ASD is a complex, prototypical neurodevelopmental disorder characterized by high heritability and a heterogeneous genetic architecture and phenotypic presentation. [1]. Individuals with ASD have atypical cognitive profiles, such as impaired social cognition and social perception, executive dysfunction, and atypical perceptual and information processing [2]. ASD is a common disease, and the prevalence of ASD in the world population, according to a systematic review, is 1% [3]. In Portugal, the overall prevalence is 0.5%. [4].

According to the Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-5-TR), there is no physical test to diagnose ASD. Despite significant advances in our understanding of the genetic causes of ASD, there remains a lack of consensus on specific genes and pathways that contribute to this disorder. This thesis aims to tackle this issue by developing a machine learning approach to identify the genetic factors that contribute to ASD and pave the way for more personalized approaches to diagnosis and treatment [5].

Although conventional genetic association studies such as Transmission and De Novo Association (TADA) using Whole Exome Sequencing (WES) data have been the most successful approach for discovering risk genes for ASD [6], they face limitations due to the expense and challenges associated with acquiring and reusing WES data. In TADA studies, a Bayesian gene-based likelihood model is used that includes per-gene mutation rates, allele frequencies, and relative risks of particular classes of sequence change. In contrast, ML classification methods offer a faster, more cost-effective alternative for predicting ASD risk genes from genetic data, showing promising results when combined with techniques such

as embeddings or graph embedding [7], [8].

1.2 Problem

Autism Spectrum Disorder (ASD) is a complex heterogeneous neurodevelopmental disorder influenced by genetic and environmental factors [9]. While ASD has a strong genetic component, uncovering the full genetic architecture of ASD continues to be a major challenge. Although there have been several studies advancing our understanding of ASD not all the biological pathways and processes underlying ASD have been fully characterized. To better understand this complex disease, several machine learning approaches aimed at the detection of risk genes for ASD were developed [7; 8; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25].

A central challenge of this thesis is effectively representing gene features for machine learning models. To facilitate the development of an ML-based bioinformatic pipeline for predicting ASD risk genes, the effective transformation of biological data into an ML-usable format is essential.

Working with genetic data is often hampered by data access limitations. This limited data availability makes it challenging to assemble sufficiently large and unbiased datasets. Approaches such as TADA analysis [26] are recognized as effective for identifying ASD risk genes. Despite their effectiveness, TADA and similar methods rely on datasets with limited accessibility due to privacy concerns. Furthermore, the inherent complexity of ASD, with its multitude of interacting biological pathways, adds to the challenge of achieving biologically meaningful and translatable findings.

ASD is a prototypical complex genetic disorder that involves numerous pathways and biological processes. Due to its complexity, achieving biologically relevant results poses a significant challenge.

To address these challenges, this thesis proposes a ML approach that leverages computational methods to analyze large datasets and enhances the accuracy and reliability of gene identification, without relying on restricted genetic data.

1.3 Objectives

The primary goal of this thesis is to create a bioinformatics pipeline that enhances existing state-of-the-art ML methods to predict ASD risk genes. While simultaneously achieving the following specific goals:

- Finding the most effective gene representation approach.
- Identifying the optimal training datasets.
- Creating a ranked list of ASD risk genes.
- Identifying new possible ASD risk genes.

As mentioned in the problem section, finding a way to represent genes in a format compatible with ML models that can be used to generate significant biological predictions presents a significant challenge. In this thesis, two gene representation approaches are used to find which of the two can produce better gene representations. In this thesis, I also create subsets of state-of-the-art ASD risk gene datasets to find the optimal balance between highly specific and balanced datasets. Using the created, validated, and trained pipeline, in this thesis a ranked list of possible ASD risk gene candidates is provided. This ranked list can be used as a starting point for future ASD risk gene research, ultimately contributing to a better understanding of this disorder.

1.4 Workflow

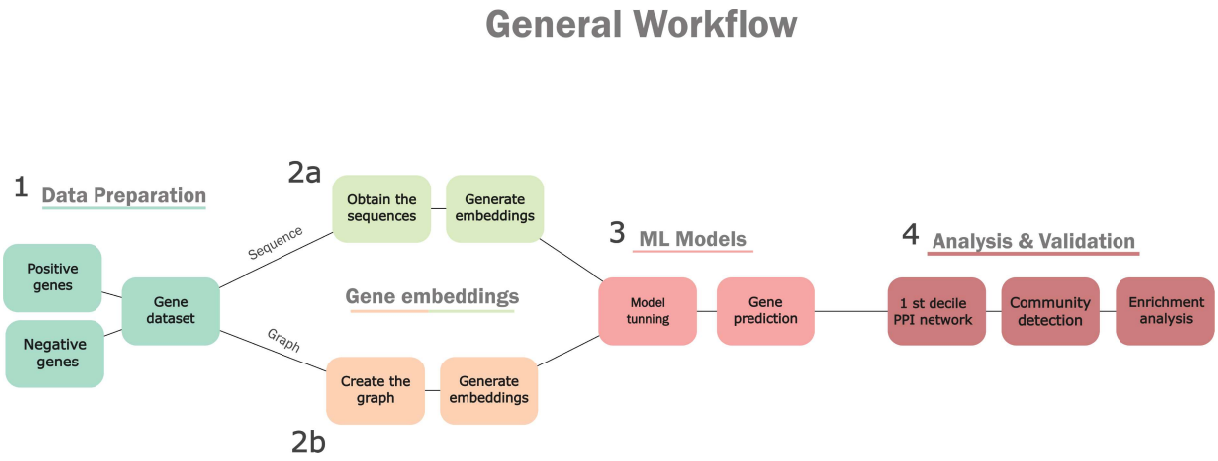

Figure 1.1: Representation of the general workflow of this thesis.

The first step of this thesis involves data preparation, which includes choosing, extracting, and cleaning the positive and negative ASD risk gene datasets. The list of positive genes consists of genes proven

to be associated with ASD through experimental evidence. The negative list consists of genes that are as different from the positive genes as possible. This step is challenging due to the limited availability of confirmed positive ASD risk genes and the evolving nature of negative gene associations, where some previously classified negative genes have been newly linked to neurodevelopmental disorders.

Once the positive and negative genes have been curated, the next step involves generating embeddings for all available genes. This thesis compares two distinct embedding methods: graph embeddings, which have shown efficacy in previous research, and sequence embeddings, a novel approach that demonstrates significant potential when applied to text modeling.

After creating the embeddings, a set of ML models is applied to determine which model performs best with each embedding method. This step includes hyperparameter tunning and ML model validation. Once the optimal model is identified, it is trained using the positive and negative gene datasets and applied to the entire gene set, resulting in a ranked list of predicted ASD risk genes.

The ranked list is subsequently divided into deciles, following methodologies such as those used by [7] and [18]. The first decile is characterized through enrichment analysis for several pathologies and integrated with various available proband ASD data to compare with other published ML methods. In the last step, communities within the top decile are identified and characterized using enrichment analysis functional annotations.

1.5 Thesis Structure

This thesis is divided into four chapters, which are the following:

- **Chapter** 1 **- Introduction:** This chapter presents the goals of the thesis as well as the motivation, the problems associated with this subject and the workflow used to achieve the goals.
- **Chapter** 2 **- Related work:** The related work chapter introduces all of the technical concepts needed to understand the methods applied in this thesis. The concepts include an introduction to graphs, machine learning, embeddings, and metrics used to evaluate said models.
- **Chapter** 3 **- Prediction of ASD risk genes:** The third chapter features the methodology used and the results obtained.
- **Chapter** 4 **- Discussion:** The last chapter provides a review of all the key findings uncovered in this thesis.

Chapter 2

Related Work

2.1 Background

2.1.1 Graphs

Graphs are organized representations of entities and their respective relations. In a graph, objects are represented as nodes (or vertices), and their relationships are defined as edges. Graphs play an important role in biology and medicine by representing complex biological organization and biomedical knowledge in a more straightforward and efficient manner. For example, if we consider a network of protein interactions, each node is a protein, and the edges represent the interactions between these proteins. A key principle in network biology, and highly relevant to risk gene discovery, is that mutations affecting proteins that interact within the same biological pathways can often contribute to related disease phenotypes. Graph representations are also able to display how similar entities group together. For instance, cellular components linked to the same phenotype generally group within the same network neighborhood, and essential genes are frequently found at the hubs of a molecular network [27]. Arranging genes in a graph better represents all the intricate interactions and relations between each of them, facilitating the use of ML models as well.

2.1.2 Machine learning

Arthur Samuel, a pioneer in artificial intelligence and computer gaming, defined the term machine learning as ”the field of study that gives computers the ability to learn without being explicitly programmed” [28]. Extracting meaningful insights from large datasets has become increasingly necessary with the exponential growth of information generated by our civilization. Machine learning emerges as a subset of artificial intelligence that involves creating algorithms that allow computers to learn from data to make predictions and find relations. There are four types of machine learning algorithms: supervised, unsupervised, semi-supervised, and reinforcement. In this thesis, only supervised and unsupervised models

were used.

2.1.2.1 Supervised learning

Supervised learning is an approach in machine learning in which models are trained on labeled datasets to make predictions. A labeled dataset comprises a series of input-output pairs, where each input is associated with its corresponding correct output. This approach enables the model to learn the relationships between features and target variables [29]]. There are two main categories of supervised learning algorithms:

- Classification: Classification algorithms predict a discrete label or category for a given input. Classification can be binary (two classes) or multiclass (more than two). For instance, a weather forecast model might predict windy, sunny, rainy, or stormy (multiclass) outcomes. At the same time, a spam filter would classify emails as either spam or not spam (binary) [29].

- Regression: Regression algorithms predict a continuous target numeric value for a given input. Opposite to classification, in regression, the predicted value is a numerical value within a specific range. For example, a regression model might predict car prices based on features such as mileage, age, and brand [29]. Another application could be forecasting the average glucose levels of an individual, given his age, address, job, and sex.

2.1.2.2 Unsupervised learning

Unsupervised learning is a branch of machine learning where algorithms are tasked with finding patterns or structures in unlabeled data. Unlike supervised learning, there are no predefined output labels, and the model must discover inherent relationships within the data on its own [29]. Unsupervised learning is beneficial for exploratory data analysis, feature learning, and finding hidden patterns. The main categories of unsupervised learning algorithms include the following:

- Clustering: Clustering algorithms group similar data points together based on their inherent characteristics or distances between them. These algorithms aim to maximize intra-cluster similarity while minimizing inter-cluster similarity. Common clustering techniques include K-means, hierarchical clustering, and DBSCAN. For example, clustering can be used in customer segmentation for targeted marketing or image segmentation for computer vision tasks.

- Dimensionality Reduction: These algorithms aim to reduce the number of input features while preserving the most essential information in the data. This is useful for visualization, noise reduction, and improving computational efficiency. PCA and SVD are popular dimensionality reduction techniques.

- Association Rule Learning: This algorithm aims to discover interesting relationships or associations among variables in large datasets. It is commonly used in market basket analysis to find items frequently bought together. The Apriori algorithm is a classic example of association rule learning, often applied in retail for product placement and recommendation systems.
- Anomaly Detection: These algorithms identify rare items, events, or observations that differ significantly from most data. Anomaly detection is crucial in various domains, such as fraud detection in financial transactions and intrusion detection in cybersecurity.

The main goal of this thesis is to create a bioinformatics pipeline capable of predicting whether a gene is an ASD risk gene. Similarly to other ML ASD risk prediction studies, a binary classification approach was used. To predict the risk genes, various classification models were used, ranging from simpler models such as logistic regression to more complex models such as XGBoost.

2.1.2.3 Logistic regression

Logistic regression, also named the log-linear classifier, is a binary classification model commonly used to estimate the probability that an instance belongs to a particular class. If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class, and otherwise, it predicts that it does not. It does this by computing the input features' weighted sum while adding a bias term [29].

This is one of the simplest models for training or making predictions, but it still presents significant results when used.

2.1.2.4 SVM

Support Vector Machines SVM are effective and versatile machine learning models used for classification and regression tasks, although they are most commonly applied in classification problems. In classification tasks, SVMs aim to find the optimal hyperplane that separates different classes in a high-dimensional space. This hyperplane is chosen to have the largest margin between the two classes, where the margin is defined as the width of the separation between the hyperplane and the nearest data points from each class, known as support vectors. In risk gene discovery, SVMs have been shown to produce excellent results, as shown in the Krishnan et al. [7] article.

While SVMs can be computationally intensive for large-scale problems, they remain popular due to their Accuracy and versatility, especially in cases where the number of dimensions exceeds the number of samples [29].

2.1.2.5 KNN

Nearest neighbor-based classification is a form of instance-based or non-generalizing learning. KNN does not require a training phase in the traditional sense, as it does not create an explicit model. Instead,

it retains training data instances and makes predictions based on the closest training samples to a new data point.

The idea behind the model is to find the number of training samples closest in distance (typically Euclidean) to a new point and predict the corresponding label. Instead of having a trained model, KNN retains instances of the training data. The standard nearest-neighbor classification uses uniform weights, meaning a straightforward majority vote of the nearest neighbors determines the value assigned to a query point. While a simple model, KNN has been shown to be effective in many classification and regression problems. Making it still a relevant model today, in part due to techniques such as Ball Tree or KD Tree for efficient neighbor searches in higher-dimensional spaces

2.1.2.6 Random Forrest

Random Forest is an ensemble learning method used for classification and regression tasks. Developed in 2001 by Breiman [30], this model combines two techniques: a bootstrapping method with decision trees. Bootstrapping consists of sampling at random from a training dataset with replacement. This means that individual data points may be chosen for each sample more than once. After creating several data samples, a decision tree is trained independently for each data sample. In classification issues, the final prediction of a Random Forest model will be the majority vote of all the trees [29].

A decision tree is a model that continually splits the data into subsets, creating a hierarchical tree structure. Each split is made to find the optimal splits that best separate the data to minimize the impurity. In other words, each split aims to make the resulting subsets as homogeneous as possible regarding the target variable. Decision trees use a greedy approach, where each split is determined sequentially without considering the global optimal separation, sometimes finding the best overall hierarchical separation is impossible [29].

By training various decision trees with Bootstrap, we can get a more generalized model, reducing the risk of overfitting, a common issue with decision trees. This makes Random Forest one of the most used classification and regression models.

2.1.2.7 Gradient-boosted trees

Similar to the Random forest model, a boosting model is also an ensemble model. However, instead of sampling the data and running a set of models simultaneously, a boosting model works by training predictors sequentially, with each model trying to correct its predecessor. In the case of the gradient-boosted models, each new predictor is trained with the residual errors made by the previous predictor. These models are highly efficient and currently present state-of-the-art predictions. In this thesis, two gradient-boosted models, XGBOOST[1] and LightGBM[2], were used. Both models use a decision tree as the base model, and both have been shown to have state-of-the-art performance.

[1] `https://xgboost.readthedocs.io/en/stable/`

[2] `https://lightgbm.readthedocs.io/en/stable/`

2.1.3 Transformers

Transformers are neural networks with a specific architecture, enabling them to extract contextual cues and relationships in sequential data. They were introduced in a 2017 paper by Google researchers titled ”Attention Is All You Need” [31], and have since revolutionized the AI landscape due to their use in Large Language Model (LLM). Many AI models currently used daily, such as ChatGPT, Gemini, and GitHub Copilot, are based on the Transformer architecture.

The most essential feature created in the Transformer models was the self-attention mechanism, which allows the model to dynamically weigh the importance of different segments of a sequential input. Transformer models are also able to process long-range dependencies more effectively than previous deep learning models such as Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN) due to their positional encodings.[31]. Due to their architecture, transformers are not influenced by the vanishing-gradient problem commonly faced by RNNs. They also reduce the sequential computation bottleneck by allowing for parallel processing of input data.

One significant drawback of transformer models is their quadratic time and memory complexity regarding sequence length, which can pose challenges when dealing with long sequences. Additionally, while Transformers excel in many areas, they lack an inherent sequential inductive bias, relying instead on positional encodings to represent the order of input data. This can sometimes lead to inefficiencies in tasks where sequential relationships are crucial. One of the most popular transformer-based models is the Bidirectional Encoder Representations from Transformers (BERT). [32]. BERT is designed to pretrain deep bidirectional representations from unlabeled text by joint conditioning on both left and right contexts in all layers. In the original transformer architecture, every token could only attend to previous tokens in the self-attention layers, so it only worked left to right. This showed issues when applied to sentence-level tasks, such as answering questions where it is essential to take both directions’ contexts into account [32].

As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

2.1.4 Embeddings

Embeddings represent real-world objects as numeric vectors in a high-dimensional space. Most ML models are unable to make predictions with complex data because they cannot grasp all of its intricate relationships and inherent properties. Embeddings bridge that gap by representing each object as a vector in a dimensional space, following the homophily theory, which states that objects that share similar functions or features would be represented closer together.

For example, if a model was tasked with predicting the function of a protein using its amino acid sequence, it would be difficult for the model to make predictions based on just the amino acid sequence. However, by representing the sequence as a vector, the model can more easily identify patterns and make

accurate predictions.

Embeddings play a vital role in dimensionality reduction with models such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). As well as in Natural language Processing Natural Programming Language (NPL) with the model word2vec, where each word is represented as a vector [33]. This thesis explores two different embedding approaches. The first is graph embeddings, which use knowledge graphs as the data. The other is sequence embeddings, where the DNA or amino acid sequence is used as the primary source of information.

2.1.5 Graph embeddings

As previously mentioned, embeddings are a vector representation of an object. In the case of graph embeddings, each vector will represent a node of a graph while preserving the graph's relational and structural data. Graph embeddings are extremely useful when dealing with large and complex graphs since these representations significantly reduce the computation time of ML models while still preserving the graph's features. The most popular graph embedding models are **DeepWalk** and **Node2Vec**. Both models take inspiration from the word2vec while also implementing random walks, which consist of using uniformly distributed probabilities to randomly select nodes sequentially in a graph. The critical difference between these two models is the length of the walk. In the Deepwalk algorithm, the walks always have the same length; in the Node2Vec case, the length of the walk is defined based on the distance between the nodes while also factoring in the return parameter and the in-out parameter [34; 35].

2.1.6 Sequence embeddings

Sequence embeddings have become increasingly popular due to the advances in NPL models. While models such as Word2Vec showed great promise, their representation of a word was limited. Word2Vec maps individual tokens into a single vector, not considering the different meanings of a word depending on its content [33]. Consider the phrases: "I am going to be the lead in a school play" and "I used to play basketball in college.". In this example, the word "play" has different meanings, but using word2vec, only one will be represented as an embedding. A BERT model improves on this by generating a different representation depending on the context of the input phrase [32].

Suppose we apply these principles to genomics, but instead of using a word sequence, we use a DNA or protein sequence. In that case, we can leverage sequence embeddings to capture the contextual information of genetic sequences. Just like in NLP, where the meaning of a word depends on the context of the sentence, the function or significance of a DNA sequence depends on its neighboring nucleotides and the larger genetic context in which it appears.

DNABERT-2 [3] is a state-of-the-art BERT model trained on large-scale multi-species genomes. Each word is clearly separated by a space in sentences or sequences of words. In the case of DNA sequences,

[3] `https://github.com/MAGICS-LAB/DNABERT_2`

each genome is an extensive sequence of As, Ts, Cs, and Gs without any spaces or commas. One of the most used methods to separate the DNA sequence into smaller and more readable chunks was the k-mer tokenization [36]. The k-mer tokenization consists in breaking a DNA sequence into fixed-length permutations. DNABERT-2 uses a different method to create tokens of a sequence. Instead of k-mer tokenization, they use the Byte Pair Encoding (BPE). BPE is a statistics-based data compression algorithm that creates each token interactively by merging the most frequent co-occurring genome segment. This implementation, paired with techniques such as Attention with Linear Biases (ALiBi) and Low-Rank Adaptation (LoRA), makes DNABERT-2 a more efficient state-of-the-art model [36].

PROTTRANS [4] ProtT5, similar to DNABERT-2, uses a BERT architecture to create state-of-the-art protein embeddings using their amino acid sequences. Unlike DNABERT-2, where each DNA sequence is tokenized in different sizes, ProtT5 processes each amino acid separately to generate its embedding. The model is pre-trained using 393 billion amino acids and without using evolutionary information [37].

2.1.7 Evaluation metrics

In a binary classification task, there needs to be a clear set of evaluation metrics to evaluate the performance of each ML model and embedding method. In this thesis, the model aims to predict whether or not a gene is linked to ASD. If the model correctly predicts a gene's association with ASD, that prediction is considered a True Positive (TP). If the model incorrectly predicts a gene's association with ASD, it is a False Positive (FP). The same principle applies to the negative predictions. If the model correctly predicts no association, it is a True Negative (TN). However, if it incorrectly predicts no association, it is a False Negative (FN) [29].

- **Accuracy** is a metric that measures the amount of correctly made predictions by the ML model out of the total number of predictions. It is calculated as the sum of the True Positives (TP) and the True Negatives (TN) divided by the sum of all the other predictions (which is the sum of True Positives, True Negatives, False Positives (FP), and False Negatives (FN)). The Accuracy of a model is highly dependent on how much the dataset is skewed. If, in this thesis, a dataset with ninety genes not associated with ASD and ten that were was used. The model could have 90% Accuracy by predicting that a gene was not associated every time. Discarding the need to predict an association.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** is used to determine how often the model correctly predicts the positive class. In other words, Precision shows how frequently the positive predictions are correct. To calculate the Precision, we divide the True Positives (TP) by the True Positives and the False Positives sum.

[4] `https://github.com/agemagician/ProtTrans`

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** or True Positive Rate (TPR) is a metric used to determine the proportion of all the true positives correctly classified as positives. In this case, a Recall close to 1 means that most of the genes predicted by the model as associated are true positives. Recall is calculated as the sum of the True positives (TP) divided by the sum of the True positives (TP) and False negatives (FN).

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Specificity** or True Negative Rate (TNR) is a metric used to determine the proportion of actual negatives correctly identified by the model. A high specificity indicates that the model is effective in identifying negative cases. Specificity is calculated as the number of True Negatives (TN) divided by the sum of True Negatives (TN) and False Positives (FP).

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- The **F1** score is the harmonic mean of recall and Precision. This metric provides a balanced score of a model due to being a harmonic mean. This means that, for a model to have a high F1 score, it also needs to have a high recall and Precision.

$$\text{F1} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

- **ROC AUC** or Receiver Operating Characteristic - Area Under Curve is the area under the ROC curve. The ROC curve illustrates the trade-off between TPR and FPR, with the Area Under the Curve (AUC) providing a single value that summarizes the model's performance across all thresholds. A higher AUC value indicates a better model performance.

- **Matthew Correlation Coefficient (MCC)**, also known as phi coefficient, is a metric that provides a balanced measure of the classification model's performance by taking into account true and false positives and negatives. It is beneficial for imbalanced datasets.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

2.1.8 Cross-Validation

In ML the simplest way to test a model is to separate the data into a training set and a testing set, also called the train-test split. Both the training and testing set can be divided into two classes, the features

(X) and the label [29]. The features are what the model will use to make predictions, the label is the value that is going to be predicted. The train-test split works well in most simple ML tasks but is limited when working with unbalanced datasets and complex data. As an alternative, this thesis utilizes a stratified K-fold cross-validation approach to ensure the training phase is as unbiased as possible [29].

This method validates the ML models with the training data and reduces possible biases with unbalanced datasets. The dataset is equally divided into folds (parts) with the same size and proportion of positive and negative samples. The idea behind K-fold cross-validation is to divide a dataset, for example, in 5 folds, and train and test an ML model 5 times, changing the fold used for testing each time and using the remaining folds as training data, as seen in Figure 2.1. (Figure created with the inspiration of this figure [5]). This way all of the data is used for both testing and training, which is important for this thesis since the dataset used is relatively small.

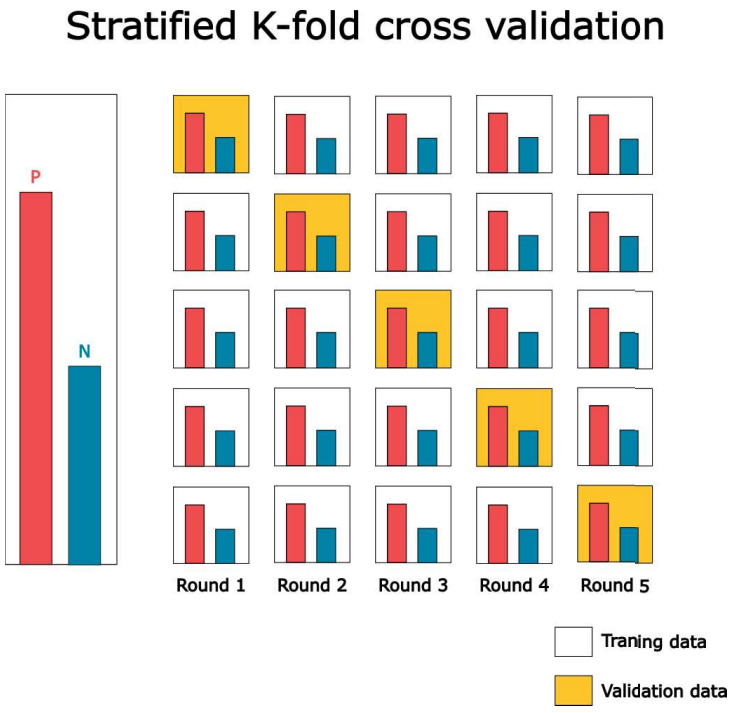

Figure 2.1: Diagram showcasing the workflow of K-fold stratified cross-validation. The red bar represents the positive genes, and the blue bar represents the negative genes.

[5] https://medium.com/@ompramod9921/cross-validation-623620ff84c2

2.2 Systematic literature review

2.2.1 Identification

A comprehensive systematic literature review was conducted to better understand the methodologies and data used in ASD-related research. The first step in this literature review is to define keywords to search for relevant papers in different databases. To achieve this, a set of relevant keywords were drafted. The keywords were aimed at narrowing the search to include only pertinent studies for this thesis. For this review, the keywords drafted were:

- 'ASD' OR 'Autism': These keywords specify the target disorder of the papers.
- "Candidate gene*" OR "Risk gene*" OR "Disease gene*": These keywords aim to restrict the search to studies that focus on risk genes.
- "Machine learning" OR "Deep Learning": These keywords specify the methodological approach used in the studies.

The selected keywords were used to search for papers from three databases: PubMed, Scopus, and Web of Science. The search produced 36, 47, and 37 papers, respectively. After intersecting the Digital Object Identifiers (DOIs) of each paper, 67 unique papers remained from the initial total of 122. To not exclude any possible relevant articles, the keywords were used to search throughout the entire article, not just titles and abstracts.

2.2.2 Screening

In the screening phase, the screening questions were defined to evaluate the relevance of each paper. These questions were designed to ensure that only studies relevant to predicting ASD risk genes using machine learning methods were included. The screening phase was necessary since this thesis aimed to predict ASD risk genes without using patient sequence data. In most ASD studies of ASD risk gene prediction, WES data was used.

The screening questions were the following:

1. Were risk genes predicted using machine learning methods?
2. Were the predictions made for ASD risk genes?

Each paper was thoroughly read and evaluated based on these criteria. Articles that did not meet both requirements were excluded from further analysis. Due to the small number of articles that passed the screening phase, articles published before 2018 were considered relevant to this thesis. After this screening phase, 17 articles were selected from the initial 67. After carefully reading each of the 17 articles, two were excluded. The first was the Bern et al. [38] study published in 2019. This article was

excluded from the systematic literature review because ASD risk genes were only predicted for a specific phenotype. This thesis aims to obtain a more global understanding of ASD, not restricted to a particular phenotype.

The other excluded study was the Huang et al. [39] published in 2021. Similarly to the other study, in the Huang et al. [39] paper, the ASD risk genes were predicted in conjunction with toxic chemicals, making it a multilabel classification. Restricting the predicted ASD risk genes as well.

2.2.3 Results

Table 2.1: Positive and Negative Genes Used in Models

Study	**Positive Genes**	**Negative Genes**
Fan et al. (2023)	1005 (SFARI dataset)	1590 (source unspecified)
Zhang et al. (2020)	732 (SFARI), 28 (OMIM)	1102 (Duda et al., 2018)
Dang et al. (2019)	From Dang Hung Tran et al. (2014)	From Dang Hung Tran et al. (2014)
J. Anitha et al. (2021)	366 (Cogill and Wang et al.)	From Cogill and Wang et al. (2014)
Lin et al. (2021)	336 (SFARI, AutismKB, De Rubeis)	366 (randomly extracted from Cogill and Wang et al. (2014)
Asif et al. (2018)	990 (SFARI)	1140 (Krishnan et al.)
Krishan et al. (2016)	894 (SFARI, OMIM, GAD,HUGE, DGA, Gene2Mesh)	1189 (non-mental illnesses)
Martiniano et al. (2020)	739 (SFARI)	1132 (Krishnan et al., 2016)
Beyreli et al. (2022)	594 (SFARI)	1074 (Krishnan et al.)
Duda et al. (2018)	143 (SFARI)	1176 (Krishnan et al.)
Brueggeman et al. (2020)	76 (SFARI high-confidence)	1000 (randomly selected)
Suratanee et al. (2023)	SFARI , AutismKB	Randomly sampled to match positive genes
Gök et al. (2018)	366 (BrainSpan)	All genes excluding ASDgenes
Pastorino et al. (2019)	1013 (SFARI)	13,568 (all other genes)
Ying Lin et al. (2020)	121 (filtered from Duda et al., 2018)	963 (Krishnan et al.)

Of these 15 papers, six used only the SFARI database as the source of the positive genes. In comparison, the other 11 papers used a mixture of SFARI genes with different databases, such as OMIM,

AutismKB, De Rubeis, DGA, and Gene2Mesh, as seen in Table 2.1.

For the negative genes, a total of 7 articles use some or all of the negative genes used in the Krishnan et al. [7] study. At the same time, 2 articles use all other genes.

Each study presented ML metrics of their model's performance, such as AUC, Accuracy, Matthews Correlation Coefficient (MCC), and others. Each study shows a high variability of the chosen negative and positive genes. For example, in the Krishnan et al. [7] study, 894 positive genes from different databases are used, while in the Duda et al. [18], only 143 positive genes are used (Table 2.1). Since the gene lists are so diverse, comparing each study's performance metrics becomes senseless. In this thesis, the positive genes used were extracted from the SFARI gene dataset, and the negative genes were obtained from the Krishnan et al. [7] study.

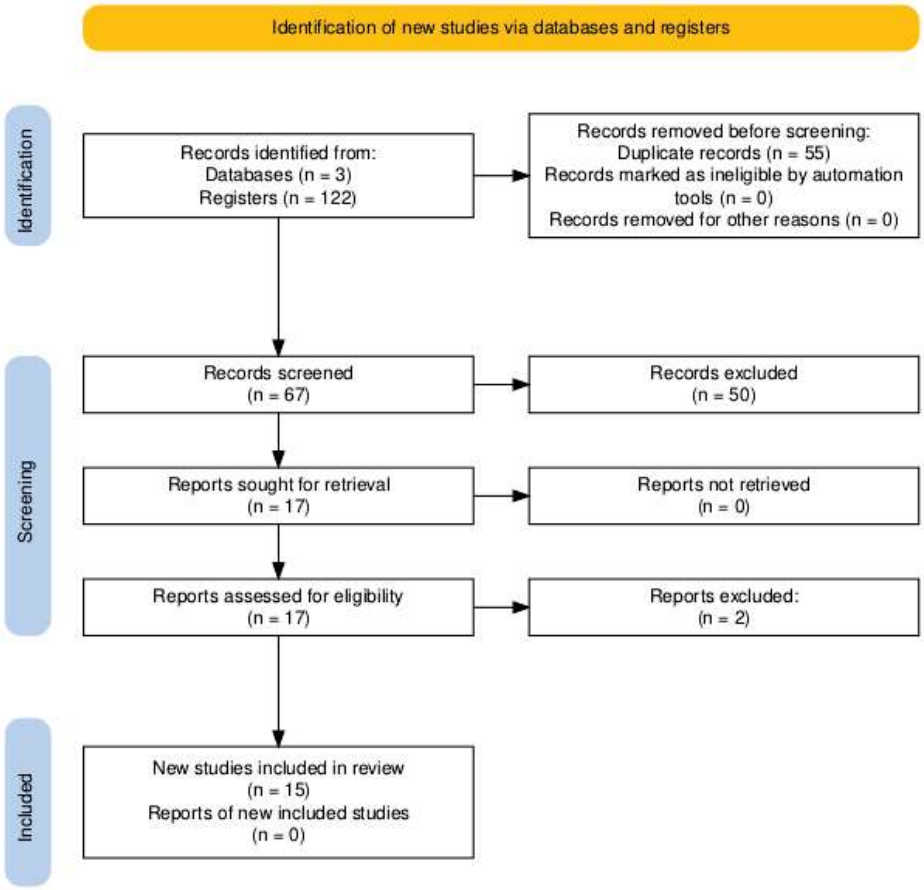

Figure 2.2: PRISMA diagram of the systematic literature review.

2.3 Data resources

2.3.1 Positive genes

Positive genes are genes that have been experimentally linked to ASD through various studies. In this thesis, I used the Simons Foundation Autism Research Initiative (SFARI) gene database [6] to extract the positive genes dataset. The SFARI gene is a comprehensive database of risk genes for ASD. For a gene to be included, it needs to have at least one reported de novo likely-gene-disrupting mutation and evidence from a significant association study[7].

To further characterize the evidence of the association of a gene with ASD, a ranking system named Gene Score was created. The gene score provides the user with the ability to get an estimate of the strength of the evidence of association of each gene with ASD. The Gene Score system classifies genes on a scale from 1 to 3, where genes with a score of 1 are genes with more bibliographic support than those with a score of 3:

- Score 1: High confidence
- Score 2: Strong candidate
- Score 3: Suggestive evidence

In addition, the SFARI gene database identifies syndromic genes. These are genes associated with syndromes that include ASD as one of their symptoms. For example, genes involved in conditions such as Rett syndrome and Fragile X syndrome are classified as syndromic due to their established connections to ASD. Following the methodologies used in 13[7; 8; 10; 11; 12; 13; 15; 16; 17; 18; 19; 21; 24] out of the 15 papers present in the systematic review, in this thesis the SFARI gene dataset was used as the positive genes.

2.3.2 Negative genes

Similarly to 7 other articles present in the systematic review, in this thesis the negative genes used were the negative genes of the Krishnan et al. [7] study. Since the article was published in 2016 the negative gene list was also manually checked to see if some of its genes had been recently associated with neurological disorders.

2.3.3 Graph data

In this thesis, I used the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) [8] protein-protein interaction network database as the features for the graph embedding models [40]. STRING is

[6] `https://gene.sfari.org/database/human-gene/`
[7] `https://gene.sfari.org/about-gene-scoring/`
[8] `https://string-db.org/`

an extensive database that gathers and integrates protein-protein interactions, including both physical and functional associations. It sources data from several origins: automated text mining of scientific literature, computational predictions from co-expression and conserved genomic context, databases of experimental interactions, and curated collections of known complexes and pathways. Each interaction is assessed and scored.

2.3.4 Sequence data

The complete genomic DNA and protein sequence files were manually downloaded as FASTA files from the Ensembl website in the FTP download section [9]. The canonical transcripts were obtained using the pybiomart library, in Python. The pybiomart library was created to enable the use of the Ensembl biomart tool directly in Python. This library provides an easier and simpler way to manage sequence data.

Using the Dataset class from the pybiomart library, for the homo sapiens species and with the host of *http://www.ensembl.org.*, the canonical sequences were extracted.

2.3.5 Software

The bioinformatics pipeline developed in this thesis was created using the programming language Python. Python is a staple in bioinformatics and computational biology because of its simplicity and compatibility with most biological data Application Programming Interface (API). The models used in this thesis, specifically linear regression, random forest, support vector machines, and K-nearest neighbors, were all scikit-learn implementations. Scikit-learn is an open-source machine-learning Python library that provides supervised and unsupervised ML models and other ML tools such as scalers and feature reduction models [41]. The code developed for this thesis, along with the gene lists utilized, can be found on my GitHub page[10].

[9] "https://ftp.ensembl.org/pub/release-110/fasta/homo_sapiens/"

[10] https://github.com/a59490/Tese_ASD_Gene_Pred

Chapter 3

Prediction of risk genes associated with ASD

3.1 Methods

In this chapter, a supervised ML bioinformatic approach to predict ASD risk genes is presented. To train a machine learning model to classify if a gene is an ASD risk gene correctly, it is necessary to use a list of positive and negative genes. Positive genes are genes previously associated with ASD, and negative genes should be as distant as possible from ASD and other positive genes.

The SFARI gene dataset was used as the positive gene list for this binary classification problem. A filtered version of the negative list presented in the Krishnan et al. [7] study was used for the negative set. Sequence and graph embedding models were used to create gene representations of each gene. By comparing two types of embedding approaches and seven different ML models in this thesis, I provide a comparison between state-of-the-art techniques while also providing relevant results.

After training an ML model with the appropriate training data, we would provide a gene with a corresponding Ensembl ID and an embedding, either a sequence or graph embedding. The model will then output a score from 0 to 1, with one corresponding to high confidence in association and 0 to no association. After applying this methodology to all available genes, the final result will be a ranked list of genes. This ranked list had the genes with the highest association score, provided by the model, as the first entries.

3.1.1 Data collection and preprocessing

The SFARI dataset used in this thesis was the updated version 1-16-2024. Ten entries on this dataset were missing an Ensembl gene ID. Each gene with the missing Ensembl IDs was cross-checked with the protein symbol, the chromosome, and the gene name on the official Ensembl site. The final positive gene

list had 1161 genes associated with ASD, as shown in Table 3.1.

Category	Gene amount
Category 1	232
Category 2	705
Category 3	131
Syndromic	94
Total	1162

Table 3.1: SFARI gene category distribution

The negative genes used in this thesis were those proposed in the Krishnan et al. [7] study. This negative list comprised 1189 genes associated with non-neurodevelopmental disorders. Considering the literature review, it was possible to verify that some authors used a filtered version of the Krishnan et al. [7] negative gene list (Table 2.1). Due to the advances in gene association studies, some of the negative genes that were previously not associated with neurodevelopmental disorders are now associated with ASD.

To validate the negative genes obtained from Krishnan et al. a comparison of gene-disease annotations from an updated dataset was made. The first step was to extract all of the codes regarding neurodevelopmental disorders from the DSM-5-TR [42]. Using a list of all the ICD-10 neurodevelopmental codes a dataset containing a list of genes associated with neurodevelopmental disorders was obtained. Comparing the neurodevelopmental disorders gene list with the Krishnan gene list we found that only 825 of the 1089 genes in the Krishnan et al. [7] study were not associated with neurodevelopmental disorders, as shown in Figure 3.1. The disgenet [1] API was used to obtain the neurodevelopmental disorders gene list. Using the ICD-10 codes as a query, the API returned a list of genes associated with neurodevelopmental disorders that should not be present in the negative gene list.

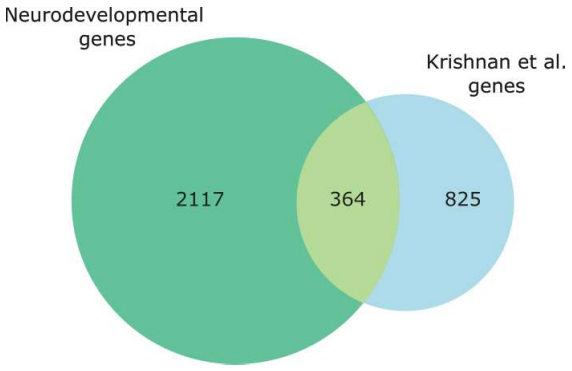

Figure 3.1: Venn diagram of the neurodevelopmental genes with the genes in Krishnan et al. study.

[1] https://www.disgenet.com/

To ensure that the negative genes were as distinct as possible from the positive genes, the ASD gene Ensemble IDs were matched with the negative gene Ensembl IDs. Of the 825 negative genes, 35 were also present in the SFARI gene dataset. The final complete gene list consisted of 1160 positive genes and 794 negative genes, for a total of 1954 genes (Table 3.2).

Gene type	Gene amount
Positives	1160
Negatives	794
Total	1954

Table 3.2: Final gene list, after validating the negative genes

3.1.2 Gene representation

In this study, two different embedding methods were used. The first was using the STRING database, a protein-to-protein network, to create the graph embeddings. The second was using a gene's DNA or protein sequence with a BERT transformer model. DNABERT 2 for the DNA sequences and ProtT5 for the protein sequences.

When working with gene data, a significant challenge is selecting the correct cDNA sequence. Alternative splicing and post-translation processing can result in one gene being translated into different proteins, known as isoforms. Due to the limitations of protein sequencing studies and the nature of each gene, some genes have annotated more than six cDNA sequences while others only have one. In this thesis, the canonical sequences were used to address this issue.

The canonical sequence, also known as the consensus sequence, is a protein or nucleic acid sequence that represents the most expressed sequence of a particular gene. These sequences are chosen to represent the most typical sequence for a gene or protein. The pybiomart Python package, specifically the Dataset class, was used to obtain the canonical sequences for each gene. Using the methodology presented on the pybiomart documentation page, the name of the dataset was *"hsapiensgene_ensembl"*, and the host was *"http://www.ensembl.org."* After creating the dataset instance, we specify the query attributes to *"ensembl_gene_id"*, *"external_gene_name"*, and *"ensembl_peptide_id"*. The result is a dataset of a gene name and ID with their respective canonical gene ID.

3.1.3 Graph approach

To create the graph embeddings, the first step is to download the protein-to-protein interaction network. The network was directly downloaded on the STRINGdb website [2], specifying the organism to Homo sapiens. The downloaded dataset consists of 13277326 rows and three columns. Each row of the STRING

[2] `https://string-db.org/cgi/download?sessionId=bsyF5t3lO2DX&species_text=Homo+sapiens`

dataset represents a link between a pair of proteins and a score relative to how confident the STRING model judges an interaction to be accurate, given the available evidence. Since the STRING database is a protein graph, the protein IDs were changed to Ensembl gene IDs using the previously mentioned pybiomart library, and all duplicates were eliminated.

After editing the STRING dataset, the next step is to recreate the graph. This was done using the Python library Grape [3] (Graph Representation leArning, Predictions, and Evaluation). Grape is a fast, high-performance processing and embedding library developed by AnacletoLAB (Dept. of Computer Science of the University of Milan). This library, written in Rust and Python, was created to optimize the creation of big graphs and efficient embedding techniques.

The graph representing each gene and the respective connections was created using Grape. It had 19.18 thousand nodes and 6.64 million edges, with a minimum edge weight of 150 and a maximum edge weight of 999. The Grape library gives access to different types of embedding models to create graph node embeddings. This thesis used 23 different graph embedding models, with each model generating embeddings with dimensions of 1000, 500, 200, and 100 rows.

3.1.4 Sequence approach

Two BERT embedding models were used to create sequence embeddings. The first was DNABERT-2, which uses DNA sequences, and the second was ProtT5, which uses protein sequences. Contrary to the previously mentioned graph embedding approach, BERT models create embeddings using sequences instead of a graph. In the case of these two models, DNABERT-2, as the name implies, uses DNA sequences to produce embeddings, while ProtT5 uses aminoacid sequences. Following the same methodology used in the graph embeddings, canonical sequences were also used for the sequence embeddings.

As previously mentioned, Fasta files containing all the protein and cDNA sequences were downloaded from the Ensembl website's FTP download section. The DNA embeddings were created following the methodologies presented in the DNABERT-2 GitHub page. They were created using an iterable Python script where each gene's canonical sequence was read and transformed by the BERT model. When the DNABERT-2 model is applied, it creates embeddings for each of the tokens created out of a gene's sequence. To obtain a single embedding of a sequence, we calculate the mean values of all the token embeddings of the sequence. The final DNA sequence embedding file comprised 28000 genes and their respective DNA embeddings.

The GitHub page methodologies for the ProtT5 model were used for the protein sequence embeddings. Using a similar Python script to the one used on the DNA embeddings, each protein embedding was created iteratively, resulting in a dataset of 23773 and their respective embeddings.

[3] `https://github.com/AnacletoLAB/grape`

Model	Number of columns	Number of rows
DNABERT-2	768	28000
ProtT5	1023	23773

Table 3.3: Sequence embeddings final datasets

3.1.5 Machine learning models

3.1.5.1 Data unbalance

Once all the embedding files have been created, the next step is to address the class imbalance in the gene list used for training. The original gene list, Table 3.2, has 1954 genes, 1160 positive and 794 negative. In a classification problem, the ideal scenario would be to have the same number of positive and negative genes. As mentioned in the introduction, using imbalanced datasets in this specific task is expected due to the limitations in available data. The Pastorino et al. [21] study used an oversampling technique to increase the data entries of the least occurring class to create a smoother dataset. They achieve this by randomly duplicating the samples of the less occurring class. While oversampling techniques provide a solid solution for tasks with repeating data entries, in the case of genetic data, where each gene has unique properties, creating artificial data may not capture the true complexity and variability of genetic sequences. This could lead to oversimplifications or inaccuracies in the analysis, as each gene's distinct features and interactions might be lost or misrepresented.

This thesis takes a novel approach. Instead of only using a positive and negative dataset, six datasets were used and compared using metrics such as the F1 score and the MCC, which take into account skewed datasets. The six data sets comprised all combinations of categories 1 through 3 of the SFARI gene score and whether syndromic genes were used (Table 3.4).

Category	Positive genes	Total
1	231	1019
1 and Sd	325	1113
1 and 2	928	1716
1, 2 and sd	1022	1810
1, 2 and 3	1058	1846
Complete	1152	1940

Table 3.4: Sequence embedding final datasets, all datasets have 789 negative genes.

While a balanced dataset is essential for classification tasks, it is equally important to consider the significance of each positive gene. Category 1 genes, more frequently observed in gene association studies, are more relevant than genes in other categories. In this classification task, the positive genes represent the most significant genes associated with ASD. Using less important genes might make the models

make worse predictions, ultimately making them more meaningless. By using six different datasets in this thesis, we can find the equilibrium between a balanced dataset and a significant positive genes.

3.1.5.2 Training, validation and prediction

This thesis utilizes a stratified K-fold cross-validation approach to ensure the training phase is as unbiased as possible. Category one genes are considered the most essential ASD risk genes. Therefore, this thesis utilizes a modified stratified k-fold cross-validation approach, using test folds exclusively from category one genes. This method ensures that each permutation of the gene list (Table 3.4) is evaluated using the same subset of genes instead of using test datasets derived from a mix of gene categories.

In this thesis, the scikit_learn implementation StratifiedKFold was used to divide the dataset into five folds with the same negative and positive genes in each fold.

This thesis used linear ML models, logistic regression, support vector machines, and non-linear models such as KNN. In addition, ensemble models such as random forest and boosting models, particularly XGBoost and LightGBM, were used.

Hyperparameter tuning was used to validate these models using GridSearchCV. GridSearchCV is a scikit-learn class used to do a systematic hyperparameter optimization by thoroughly searching all the possible parameter combinations of a specified parameter grid. It performs cross-validation for each combination of parameters, thus identifying the set that produces the best performance according to a predefined scoring metric. Each model was tested using a Python script with a wide range of parameters, as seen in Table 3.5. In this thesis, the F1 score was chosen because it is a score designed to provide less biased results when working with unbalanced data sets.

Given the stratified nature of the cross-validation folds, this is a form of nested cross-validation. Nested cross-validation provides an unbiased estimate of model performance by rooting the hyperparameter tuning with cross-validation (via GridSearchCV) within the cross-validation process. This ensures that the hyperparameter tuning is performed independently for each training fold, thus preventing data leakage and providing a more reliable evaluation of the model's performance.

Model performance was assessed based on several metrics, including the area under the curve (AUC), accuracy, F1 score, recall, precision, average precision, sensitivity, and specificity. Each presented result was a mean of the five stratified folds. The standard deviation was also obtained to correctly compare the different models with the dataset permutations.

All computational experiments were executed on the BIOISI SLURM Cluster, ensuring a high-performance model training and evaluation computing environment. After validating all models and embeddings, the combination that provided the best results was trained again and applied to the complete gene dataset to make a genome-wide prediction. The result was a dataset where each gene was assigned a numeric value from 0 to 1, representing the model's confidence in how likely a gene was an ASD risk gene. By using a nested stratified 5-fold cross-validation once more, the model was also able to make an unbiased prediction of the data used for its training.The gene list dataset was ordered by the model's

Model	Parameters
Logistic regression	**C** : [0.001, 0.1, 1, 10, 100, 1000]
Random forest	**n_estimators**: [100, 200, 300, 400, 500, 1000], **max_features**: [sqrt, log2], **max_depth**: [3, 5, 10, 20, 30, 40, 50], **min_samples_split**: [2, 5, 10], **min_samples_leaf**: [1, 2, 4]
Support vector machines	**C**: [0.1, 1, 10, 100, 1000], **gamma**: [scale,auto,1, 0.1, 0.01, 0.001, 0.0001], **degree**: [3, 4, 5, 6, 7, 8, 9, 10], **kernel**: [rbf, poly, sigmoid,sigmoid]
K-Neighbors	**n_neighbors**: [2, 3, 5, 7, 9, 11], **weights**: [uniform, distance], **algorithm**: [auto, ball_tree, kd_tree]
LightGBM	**n_estimators**: [100, 200, 300, 1000], **learning_rate**: [0.0001, 0.01, 0.05, 0.1, 0.5, 1, 10, 100], **max_depth**: [2, 3, 5, 10, 20, 30], **reg_alpha**: [0, 0.1, 0.5, 1, 2, 5, 10]
XGBoost	**n_estimators**: [100, 200, 300, 1000], **learning_rate**: [0.0001, 0.01, 0.05, 0.1, 0.5, 1, 10, 100], **max_depth**: [2, 3, 5, 10, 20, 30]

Table 3.5: Models with their respective hyper-parameters used with GridsearchCV.

confidence value, making a ranked list of genes associated with ASD, which was then divided into ten equal sublists, also referred to as deciles.

3.1.5.3 Validation of predicted gene ranking

A decile enrichment analysis was performed using Gprofiler to assess and validate the model's predictions. Gprofiler is a tool used to perform enrichment analysis of gene lists to obtain their statistically significant functions across various data sources. In Gprofiler, the statistical significance of enrichment for a given gene list is obtained using the cumulative hypergeometric test with the g:SCS testing correction method, the standard of the Gprofiler tool [43].

In this thesis, three types of decile enrichment tests will be used to assess the significance of the model's prediction. The three types are Phenotypes, Neurodevelopmental disorders, and Psychiatric Disorders unrelated to ASD.

The first set of tests compares the top decile with three gene lists of phenotypes closely related to ASD: *Autistic behavior* HP:0000729, *Abnormal social development* HP:0025732, and *Restricted or repetitive behaviors or interests* HP:0031432. The next set compares disorders that have been shown to be related to ASD due to being comorbidly-related disorders with the top decile: *Attention deficit hyperactivity disorder* HP:0007018, *Intellectual disability* HP:001249, and *Schizophrenia* HP:0100753. The last set is composed of brain-specific disorders not associated with ASD: *Amyotrophic lateral sclerosis* HP:0007354, *Parkinsonism* HP:0001300, and *Dementia* HP:0000726. If the model made relevant predictions, it is

expected that the first decile will be significantly enriched for Autistic behavior, Abnormal social development, and Restricted or repetitive behaviors or interests. At the same time, having less significant enrichment to Attention deficit hyperactivity disorder, Intellectual disability, and Schizophrenia. The last set of enrichment tests should present no significant enrichment to the brain-specific disorders since they do not share a close relation with ASD.

To further validate the predicted rank list, an enrichment analysis was performed to compare with two independent sequencing studies. The first is the Simons Simplex Collection [44] exome-sequencing study of 2517 families where 350 genes with de novo Loss of Function (LOF) mutations in probands were identified. Of the 350 genes, 27 were present in more than one proband, and 174 genes with de novo LOF mutations in unaffected siblings were also identified. The second enrichment analysis was using the MSSNG cohort, which consists of a set of over 5000 samples where 214 genes were identified as Denovo LOF mutations in probands. Out of the 240 genes, 61 received genome-wide significance of association with ASD, with 18 genes having no prior evidence. By comparing the results of these two enrichment analyses with other state-of-the-art approaches we can verify the effectiveness of the proposed pipeline in predicting ASD risk genes.

3.1.5.4 Identification and characterization of communities

Community detection and enrichment analysis were used to further characterize the top decile of the ranked ASD risk gene list. To find the gene communities present in the first decile, a subgraph of the STRINGdb network was created, using only genes from the first decile (n=1810). Community detection was done using the Leiden algorithm. To better characterize each community enrichment analysis for pathways and biological processes was performed.

The igraph and leidenalg python libraries were used for community detection. Enrichment analysis was performed using the Gprofiler Python package using the Reactome and Gene Ontology Biological Processes databases.

3.1.6 Results

Using an embedding-based machine learning approach, a genome-wide ASD risk gene prediction was produced. By comparing all 1224 different combinations of embedding type, model, and dataset used, the dataset consisting of Category 1 and syndromic genes consistently provided superior performance and the best ratio between significant ASD risk genes and a balanced dataset. Due to the sheer amount of results from the different combinations, they were not presented in this thesis. However, they can be viewed on my GitHub[4] page.

Figure 3.2 shows the results of using a support vector machine in combination with deepwalk CBOW embedding on the different datasets. Category 1 and category 1 with syndromic provide more balanced

[4] `https://github.com/a59490/Tese_ASD_Gene_Pred`

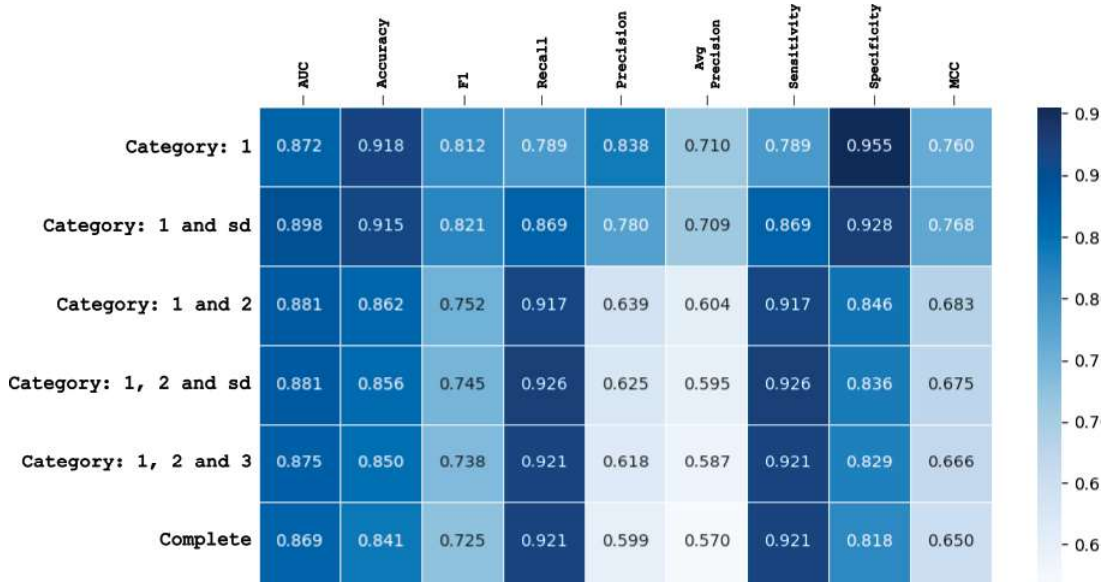

Figure 3.2: Results using support vector machines with deepwalk CBOW embeddings with different gene lists. Sd refers to syndromic genes

results, showing a much higher F1, Precision, and MCC compared to the other gene lists. While category 1 and category 1 with syndromic both provide prominent results, category 1 with syndromic pulls head. This thesis aims to predict ASD risk genes, so a higher Recall is more important than a higher Precision. In this particular case it is better to have more false positives than false negatives, this way no new ASD risk genes are excluded.

3.1.6.1 Comparison of graph and sequence embeddings

Gene lists were generated using embeddings created with ProtT5, DNABERT-2, and STRINGdb features. Due to limitations of the STRINGdb, not all 1985 genes from the initial gene list could be utilized. Only 1915, 1136 positive, and 779 negative, out of the 1985 genes from the original gene list, were successfully used to create the graph embeddings.

Out of the 32 embedding algorithms, deepwalk CBOW produced the best results when implemented in the nested cross-fold validation. Using the best-performing embedding, the deepwalk CBOW, with the category 1 and syndromic gene dataset, we achieved an AUC of 0.9, an F1 score of 0.82, and an MCC of 0.77. Logistic regression is the better-performing model for graph embeddings, as it produced a higher F1 score and MCC while still being the fastest and simplest model. These metrics are vital in unbalanced datasets, which is the case with this one (Table 3.6).

Similarly to the graph embeddings, not all genes in the initial gene list could be used to create sequence embeddings. In total, only 1945 out of the 1989 were used.

Sequence-based embedding methods provided promising results, specifically the Protein sequence embeddings. While still falling short compared with the graph embeddings, the protein sequence embeddings with XGBoost generated an AUC of 0.83, an F1 of 0.77, and an MCC of 0.65. While the DNA sequence embeddings using support vector machines only produced an AUC of 0.79, an F1 of 0.63, and an MCC of 0.52. (Table 3.6)

Model	**Type**	**AUC**	**Accuracy**	**F1**	**Recall**	**Precision**	**Sensitivity**	**Specificity**	**MCC**
KNN	Protein	0.81	0.83	0.68	0.76	0.61	0.76	0.86	0.58
KNN	DNA	0.70	0.79	0.54	0.54	0.54	0.54	0.86	0.40
KNN	Graph	0.81	0.89	0.71	0.67	0.81	0.67	0.95	0.66
LGBM	Protein	0.83	0.85	0.70	0.78	0.64	0.78	0.87	0.61
LGBM	DNA	0.76	0.80	0.61	0.68	0.55	0.68	0.84	0.48
LGBM	Graph	0.85	0.90	0.78	0.76	0.80	0.76	0.94	0.72
LR	Protein	0.82	0.80	0.67	0.86	0.55	0.86	0.79	0.57
LR	DNA	0.78	0.78	0.62	0.79	0.52	0.79	0.79	0.51
LR	Graph	**0.90**	**0.92**	**0.82**	0.87	0.78	0.87	0.93	**0.77**
RF	Protein	0.84	0.84	0.70	0.83	0.62	0.83	0.85	0.61
RF	DNA	0.75	0.80	0.60	0.66	0.55	0.66	0.84	0.48
RF	Graph	0.82	0.89	0.74	0.68	**0.82**	0.68	0.95	0.68
SVM	Protein	0.83	0.84	0.70	0.81	0.63	0.81	0.86	0.61
SVM	DNA	0.79	0.80	0.63	0.75	0.55	0.75	0.82	0.52
SVM	Graph	**0.90**	0.91	0.81	**0.88**	0.75	**0.88**	0.91	0.75
XGB	Protein	0.83	0.88	0.73	0.74	0.73	0.74	0.92	0.65
XGB	DNA	0.77	0.84	0.64	0.63	0.65	0.63	0.90	0.54
XGB	Graph	0.84	0.91	0.78	0.71	0.85	0.71	**0.96**	0.72

Table 3.6: Results for different ML models using Graph embeddings and Protein and DNA sequence embeddings with the category one and syndromic genes. The graph embeddings used were created by the deepwalk CBOW model.

Comparing the generated results, the pipeline that provided the best performance used the graph embedding Deepwalk CBOW with the model Logistic regression coupled with the category one and syndromic gene list.

3.1.6.2 Prediction validation and Functional network characterization

Following the proposed methodology, a genome-wide prediction was made utilizing graph embeddings and logistic regression using all available genes (n=18100). The final result was a ranked ASD prediction table divided into ten sections. The top genes that the model predicted to be closely related to ASD,

excluding genes from the SFARI gene database, are presented in Table 3.7.

Position	Gene name	Ensembl ID
12	*BICRAL*	ENSG00000112624
24	*FAM199X*	ENSG00000123575
29	*RALGAPA1*	ENSG00000174373
30	*Novel Protein*	ENSG00000285708
34	*NPAS3*	ENSG00000151322
35	*PAFAH1B1*	ENSG00000007168
37	*SRGAP2*	ENSG00000266028
44	*REST*	ENSG00000084093
45	*ZDHHC8*	ENSG00000099904
46	*DCX*	ENSG00000077279
47	*IQSEC1*	ENSG00000144711
48	*LARP4B*	ENSG00000107929
50	*CHD6*	ENSG00000124177

Table 3.7: Genes in the top 50 predicted gene list not present in the SFARI gene dataset.

Only 13 genes out of the top 50 predicted genes were not present in the SFARI dataset. Of the 50 genes, 24 were assigned Category 1, and 13 were assigned Category 2. Therefore, 74% of the genes in the top 50 are genes previously associated with ASD. Doing an enrichment analysis using the 13 new genes, we find that they are significantly enriched for genes involved in cerebral cortex radial glia-guided and cell migration, telencephalon cell migration, and forebrain cell migration, Table 3.8.

Process	GO term	p-value
cerebral cortex radial glia-guided migration	GO:0021801	0.000641
telencephalon glial cell migration	GO:0022030	0.000641
cerebral cortex radially oriented cell migration	GO:0021799	0.001515
cerebral cortex cell migration	GO:0021795	0.003578
glial cell migration	GO:0008347	0.006963
telencephalon cell migration	GO:0022029	0.007315
forebrain cell migration	GO:0021885	0.008440
dendrite development	GO:0016358	0.010238
dendritic spine development	GO:0060996	0.029079

Table 3.8: Enrichment analysis of the 13 genes not present in the SFARI gene dataset.

These three biological processes are critical in brain development and are associated with neurodevelopmental disorders including epilepsy, dyslexia, schizophrenia (SCZ), and autism spectrum disorders

(ASD) [45; 46]. The 13 proposed genes can be good contenders for ASD risk genes in future association studies.

To further validate the proposed bioinformatics pipeline, a decile enrichment was made using Gprofiler. By comparing the predictions with different ranges of phenotypes, it is possible to confirm the specificity of the predictions in relation to ASD, neurodevelopmental disorders, and other neurological disorders. In Figure 3.3(a), we can confirm that the proposed pipeline was significantly enriched for all the ASD phenotypes, with a p-value = $1,21\times10^{-73}$ in *Autistic behavior*, $6,75\times10^{-4}$ in *Abnormal social development*, and a p-value = $2,00\times10^{-29}$ in *Restricted or repetitive behaviors or interests*.

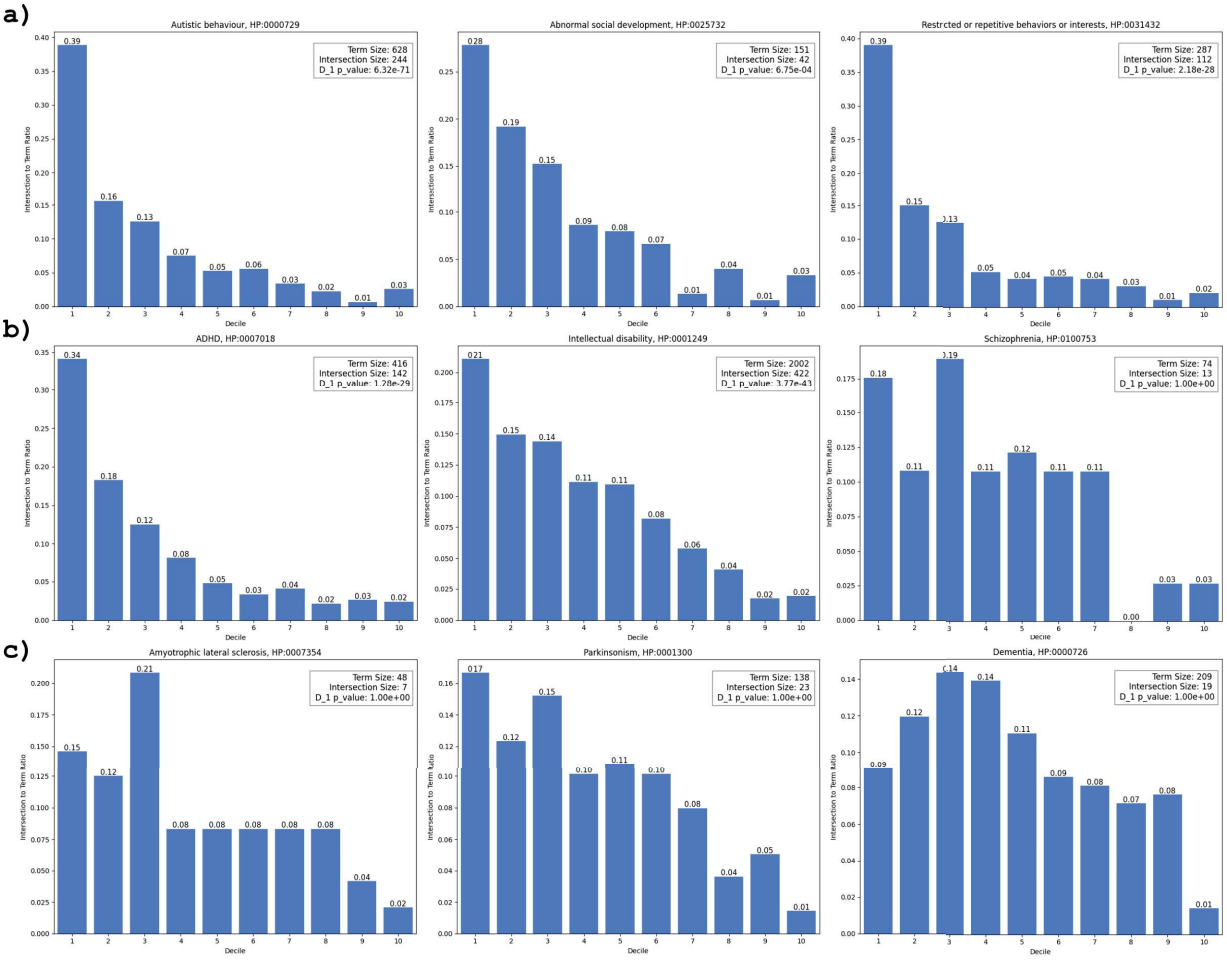

Figure 3.3: Distribution of the predicted deciles when comparing with (*a*) ASD phenotypes, Autistic behaviour, Abnormal social development, Restricted or repetitive behaviors or interests. (*b*) Neurodevelopmental disorders, ADHD, intellectual disability, Schizophrenia. (*c*) Neurological disorders, Amyotrophic Lateral Sclerosis (ALS), Parkinsonism and Dementia. Some percentages don't add to 100% due to the loss of some genes in the embedding process.

Due to the pleiotropic nature of genes linked to neurodevelopmental disorders, it is expected that some of the genes that influence ASD can also be responsible for other neurodevelopmental disorders, which is the case for *ADHD* and *Intellectual disability* [47]. This is in line with the results for enrichment in the first decile genes for these two disorders, with p-values of $7,87\times10^{-37}$ and $2,34\times10^{-60}$, respectively, Figure 3.3(b). This also correlates with the higher comorbidity rate between ASD and ADHD than with ASD and Schizophrenia [48]. Other brain-specific disorders show no significant enrichment, with p-values = 1, which is the foreseen result since they are more distant to ASD Figure 3.3(c).

The predicted ranked gene list was also compared with two other ASD risk gene prediction studies to prove its effectiveness. The first was the SCC [44] study that was also compared to the Krishnan et al. and Duda et al. [7; 18] studies. In this thesis, the proposed pipeline revealed significant enrichment in the top decile when examining proband de novo LOF mutations, with a p-value of 8.29×10^{-11}. Further, there was considerable enrichment for de novo recurrent LOF mutations, with a p-value of 7.96×10^{-10} with 70% of the genes present in the first decile, 3% more than the Duda study and a 10% increase when compared with the Krishnan study, Figure 3.4 (a).

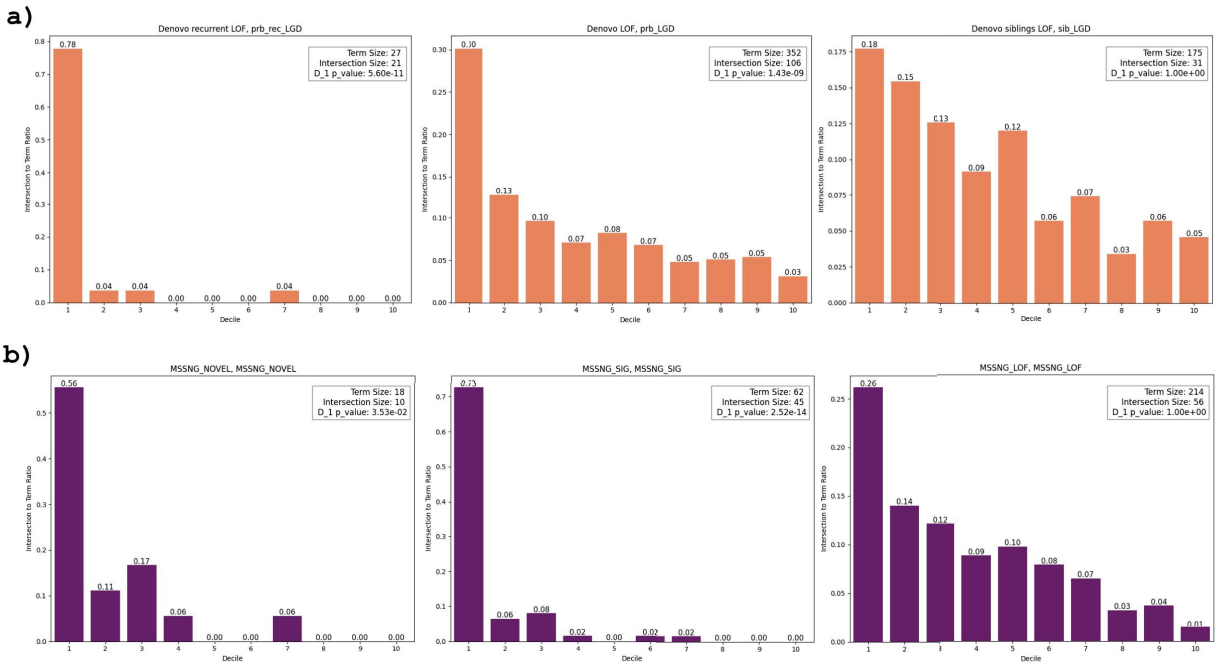

Figure 3.4: Distribution of the predicted deciles when comparing with (**a**) exome-sequencing study SSC and (**b**) with the MSSNG cohort data.

The predicted ranked gene list was also enriched for ASD candidate genes identified in the MSSNG version 5 data (as obtained from Duda et al. [18]). Significant enrichment was observed with a p-value of $8,19\times10^{-4}$ for the 18 novel genes and 1.49×10^{-15} for the 62 associated ASD genes, as shown in Figure 3.4 (b). These results show an increase of 6% and 9% in the amount of novel and associated genes

present in the first decile when compared with the study by Duda et al.[18].

3.1.6.3 Identification and characterization of communities

Autism spectrum disorder (ASD) is a complex disorder with both genetic and environmental roots. Identifying ASD risk genes is important, but understanding their functional relationships within biological networks is crucial for a complete picture of the disorder. This thesis addressed this by performing community detection using the Leiden algorithm on the protein interaction network derived from top-ranked genes, followed by functional characterization of the communities. This analysis revealed four distinct communities (Figure 3.5).

Community 1 exhibited enrichment for the biological processes of *nervous system development*, *synaptic signaling*, *calcium ion binding*, and *glutamate receptor activity*, which are known to be previously associated to ASD. Community 2 is significantly enriched in *Protein ubiquitination*, *Antigen processing*, and *Protein binding*. The biological processes of *Chromatin organization and remodeling*, *DNA-templated transcription*, and *histone-modifying activity* were associated with the community 3. All of these are also closely related to ASD. Community 4 was shown to be enriched for the ASD-associated *MAPK signaling pathway*, *intercellular signal transduction*, and *GTPase regulator activity*, Figure 3.5. This functional analysis not only provides a better understanding of the types of biological processes and pathways the ASD genes share but also clarifies the complex molecular interactions of ASD.

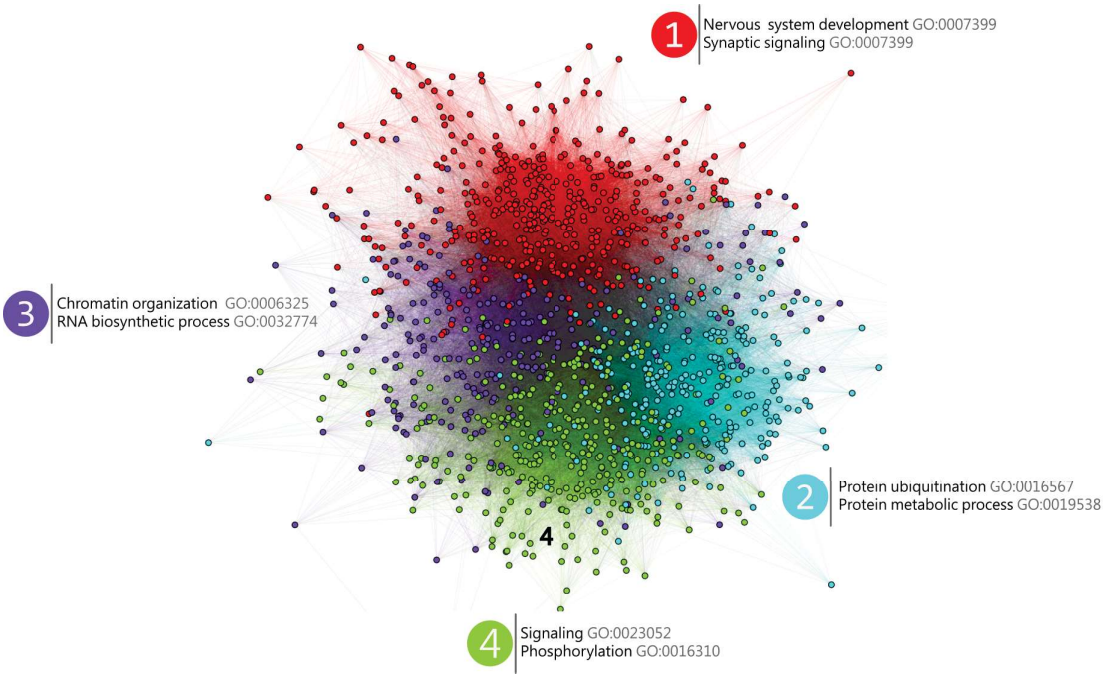

Figure 3.5: Functional communities of the first predicted decile.

Chapter 4

Discussion

ASD and other neurodevelopmental disorders are a common reality in many families around the world. With the increase in ASD prevalence [3] and the advances in gene discovery, understanding the genetic causes of ASD has become an increasingly important topic. Identifying novel risk genes associated with ASD not only provides a better understanding of the disorder as a whole but also enables possible screening methods, paving the way for personalized medicine. Machine learning approaches for ASD risk gene discovery have proven to achieve significant results while not directly dependent on patient genetic data. In this thesis, the proposed bioinformatics pipeline using ML models with embeddings improved upon state-of-the-art approaches to predicting the risk genes associated with ASD.

Genetic patient data, such as Whole Exome Sequencing and Whole Genome Sequencing, is essential for gene association studies. However, due to privacy and security concerns restrict access to genetic data. Machine learning (ML) approaches such as the one described in this thesis offer an alternative by leveraging publicly available data and embedding methods to infer gene information. ML approaches for risk gene prediction often operate under the assumption of ”guilt by association.” This principle posits that genes with associations similar to those of known disease-associated genes (e.g., in gene expression, interactions, or sequence) are also likely to be involved in the disease. However, while intuitively reasonable, this assumption can make it challenging to identify high-impact ASD genes, as predictions are inherently biased towards already discovered ASD genes. Consequently, assigning a specific weight to a gene-based solely on guilt-by-association becomes less straightforward. The SFARI gene dataset and a validated version of the Krishnan et al.[7] were used as this thesis's positive and negative features. The subset of the SFARI gene database composed of category 1 and syndromic genes provided more specificity by using a smaller portion of the positive dataset. By filtering the negative gene list obtained from the Krishnan et al. study [7], I was able to achieve a more specific negative dataset that did not include neurodevelopmental disorder genes. One of the more prominent issues with these approaches is the dataset limitations. In the graph approach, some genes were not represented on the STRINGdb. This lack of representation made the training dataset smaller. The same issue occurred in the sequence embeddings, with some sequences being too large for the model to create the embeddings. This lack of

available data impacts the model's accuracy in making relevant predictions.

Representing complex biological data in a format compatible with ML models is a challenge. In this thesis, sequence, and graph embedding approaches capable of creating these representations are presented. While DNA and protein sequence embeddings showed promising results, the graph embeddings provided a better gene representation, achieving an AUC of 0.9, an F1 score of 0.82, and an MCC of 0.77. However, it is important to note that BERT sequence embeddings are a relatively new approach that can potentially offer better performance than graph embedding models, provided that better training criteria are established. Sequence embeddings also present the benefit of not needing any prior information about a gene only its sequence.

Using the proposed bioinformatic pipeline an ASD risk gene ranked list was created. 13 new possible ASD risk genes were identified. These were shown to be associated with biological processes critical in brain development, such as neuronal migration. The first decile of the ranked list was significantly enriched for the ASD phenotypes *Autistic behavior*, *Abnormal social development*, and *Restricted or repetitive behaviors or interests*, as well as other neurodevelopmental disorders such as *ADHD* and *Intellectual disability*. However, the first decile was revealed not to be enriched for *Schizophrenia*, *Amyotrophic lateral sclerosis*, *Parkinsonism*, and *Dementia*. These findings validate the model's ability to accurately predict genes specific to ASD phenotypes rather than predicting genes related to other brain-specific disorders unrelated to ASD. Enrichment for genes associated with ADHD and Intellectual Disability is consistent with the known pleiotropic nature of genes involved in neurodevelopmental disorders [47] and the comorbidity rates[48]. When compared with similar studies, it is possible to verify that the top decile ASD risk genes predicted using this thesis's bioinformatic pipeline had more mutated genes present. Showing a 3% to a 10% increase of correctly predicted genes in the SCC [44], and MSSNG datasets, in relation to the Krishnan et al.[7] and Duda et al. studies[18].

4.1 Future work

In this thesis, a comparison of sequence and graph embeddings is made, showcasing their effectiveness as features to predict ASD risk genes. Although graph models currently provide better results, it is possible that this might change in the near future. Deep learning models are currently revolutionizing natural language processing with models such as ChatGPT and copilot, as well as discovering protein structures with the AlphaFold model. BERT models for DNA and protein sequences are expanding. As more genes are annotated and more interactions are discovered, better sequence embedding models will be created. The top-ranked genes showcased in this thesis provide a good starting point for future gene association studies. The filtered negative list used in this thesis can be used as an updated version of the Krishnan negative gene list in future ASD studies. Although the objective of this thesis was to identify ASD risk genes, the pipeline can be used for other disorders, provided an appropriate positive and negative list exists. Overall, this thesis improves upon existing ASD risk gene predictions using ML methods, taking a step closer to a better understanding of this complex genetic and environmental disorder.

References

[1] Catherine Lord, Traolach S. Brugha, Tony Charman, James Cusack, Guillaume Dumas, Thomas Frazier, Emily J. H. Jones, Rebecca M. Jones, Andrew Pickles, Matthew W. State, Julie Lounds Taylor, and Jeremy Veenstra-VanderWeele. Autism spectrum disorder. *Nat Rev Dis Primers*, 6(1):5, 2020.

[2] Shuang Qiu, Yingjia Qiu, Yan Li, and Xianling Cong. Genetics of autism spectrum disorder: an umbrella review of systematic reviews and meta-analyses. *Transl Psychiatry*, 12(1):249, 2022.

[3] Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S. Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. Global prevalence of autism: A systematic review update. *Autism Research*, 15(5):778–790, 2022.

[4] Célia Rasga, João Xavier Santos, Cátia Café, Alexandra Oliveira, Frederico Duque, Manuel Posada, Ana Nunes, Guiomar Oliveira, and Astrid Moura Vicente. Prevalence of Autism Spectrum Disorder in the Centro region of Portugal: a population based study of school age children within the ASDEU project. *Front. Psychiatry*, 14:1148184, 2023.

[5] Lauren Rylaarsdam and Alicia Guemez-Gamboa. Genetic Causes and Modifiers of Autism Spectrum Disorder. *Front. Cell. Neurosci.*, 13:385, 2019.

[6] De Rubeis, Homozygosity Mapping Collaborative for Autism, UK10K Consortium, The Autism Sequencing Consortium, Silvia De Rubeis, Xin He, Arthur P. Goldberg, Christopher S. Poultney, Kaitlin Samocha, A. Ercument Cicek, Yan Kou, Li Liu, Menachem Fromer, Susan Walker, Tarjinder Singh, Lambertus Klei, Jack Kosmicki, Shih-Chen Fu, Branko Aleksic, Monica Biscaldi, Patrick F. Bolton, Jessica M. Brownfeld, Jinlu Cai, Nicholas G. Campbell, Angel Carracedo, Maria H. Chahrour, Andreas G. Chiocchetti, Hilary Coon, Emily L. Crawford, Lucy Crooks, Sarah R. Curran, Geraldine Dawson, Eftichia Duketis, Bridget A. Fernandez, Louise Gallagher, Evan Geller, Stephen J. Guter, R. Sean Hill, Iuliana Ionita-Laza, Patricia Jimenez Gonzalez, Helena Kilpinen, Sabine M. Klauck, Alexander Kolevzon, Irene Lee, Jing Lei, Terho Lehtimäki, Chiao-Feng Lin, Avi Ma'ayan, Christian R. Marshall, Alison L. McInnes, Benjamin Neale, Michael J. Owen, Norio Ozaki, Mara Parellada, Jeremy R. Parr, Shaun Purcell, Kaija Puura, Deepthi Rajagopalan, Karola

Rehnström, Abraham Reichenberg, Aniko Sabo, Michael Sachse, Stephan J. Sanders, Chad Schafer, Martin Schulte-Rüther, David Skuse, Christine Stevens, Peter Szatmari, Kristiina Tammimies, Otto Valladares, Annette Voran, Li-San Wang, Lauren A. Weiss, A. Jeremy Willsey, Timothy W. Yu, Ryan K. C. Yuen, Edwin H. Cook, Christine M. Freitag, Michael Gill, Christina M. Hultman, Thomas Lehner, Aarno Palotie, Gerard D. Schellenberg, Pamela Sklar, Matthew W. State, James S. Sutcliffe, Christopher A. Walsh, Stephen W. Scherer, Michael E. Zwick, Jeffrey C. Barrett, David J. Cutler, Kathryn Roeder, Bernie Devlin, Mark J. Daly, and Joseph D. Buxbaum. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215, 2014.

[7] Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra L Theesfeld, Aaron K Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, and Olga G Troyanskaya. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci*, 19(11):1454–1462, 2016.

[8] Muhammad Asif, Hugo F. M. C. M. Martiniano, Astrid M. Vicente, and Francisco M. Couto. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLoS ONE*, 13(12):e0208626, 2018.

[9] Padideh Karimi, Elahe Kamali, SeyyedMohammad Mousavi, and Mojgan Karahmadi. Environmental factors influencing the risk of autism. *J Res Med Sci*, 22(1):27, 2017.

[10] Ying Lin, Shiva Afshar, Anjali M. Rajadhyaksha, James B. Potash, and Shizhong Han. A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates. *Front. Genet.*, 11:500064, 2020.

[11] Leo Brueggeman, Tanner Koomar, and Jacob J. Michaelson. Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Sci Rep*, 10(1):4569, 2020.

[12] Yongxian Fan, Hui Xiong, and Guicong Sun. DeepASDPred: a CNN-LSTM-based deep learning method for autism spectrum disorders risk RNA identification. 24(1):261, 2023.

[13] Yu Zhang, Yuanzhu Chen, and Ting Hu. Panda: Prioritization of autism-genes using network-based deep-learning approach. *Genetic Epidemiology*, 44(4):382–394, 2020.

[14] Dr.J. Anitha. Autism Spectrum Disorder Risk Gene Prediction Using Improved Salp Swarm Algorithm and Enhanced Convolution Neural Network Algorithm. *NeuroQuantology*, 19(9):97–109, 2021.

[15] Yenching Lin, Srinivasulu Yerukala Sathipati, and Shinn-Ying Ho. Predicting the risk genes of autism spectrum disorders. *Frontiers in Genetics*, 12, 2021.

[16] Ilayda Beyreli, Oguzhan Karakahya, and A. Ercument Cicek. DeepND: Deep multitask learning of gene risk for comorbid neurodevelopmental disorders. *Patterns*, 3(7):100524, 2022.

[17] Ying Lin, Shiva Afshar, Anjali M. Rajadhyaksha, James B. Potash, and Shizhong Han. A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. *Frontiers in Genetics*, 11, 2020.

[18] Marlena Duda, Hongjiu Zhang, Hong-Dong Li, Dennis P. Wall, Margit Burmeister, and Yuanfang Guan. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Transl Psychiatry*, 8(1):56, 2018.

[19] Apichat Suratanee and Kitiporn Plaimas. Gene association classification for autism spectrum disorder: Leveraging gene embedding and differential gene expression profiles to identify disease-related genes. *Applied Sciences*, 13(15), 2023.

[20] Murat Gök. A novel machine learning model to predict autism spectrum disorders risk gene. *Neural Computing and Applications*, 31(10):6711–6717, 2019.

[21] Javier Pastorino and Ashis Kumer Biswas. Texanasd: Text analytics for asd risk gene predictions. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1350–1357, 2019.

[22] S. Cogill and L. Wang. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics*, 32(23):3611–3618, 2016.

[23] Dang Hung Tran, Thanh-Phuong Nguyen, Laura Caberlotto, and Corrado Priami. Inference of autism-related genes by integrating protein-protein interactions and mirna-target interactions. In Van Nam Huynh, Thierry Denoeux, Dang Hung Tran, Anh Cuong Le, and Son Bao Pham, editors, *Knowledge and Systems Engineering*, pages 299–311, Cham, 2014. Springer International Publishing.

[24] Hugo F. M. C. Martiniano, Muhammad Asif, Astrid Moura Vicente, and Luís Correia. Network propagation-based semi-supervised identification of genes associated with autism spectrum disorder. In Maria Raposo, Paulo Ribeiro, Susana Sério, Antonino Staiano, and Angelo Ciaramella, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 239–248, Cham, 2020. Springer International Publishing.

[25] Xuan Tho Dang, Duong Hung Bui, Thi Hong Nguyen, Tran Quoc Vinh Nguyen, and Dang Hung Tran. Prediction of autism-related genes using a new clustering-based under-sampling method. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6, 2019.

[26] Xin He, Stephan J. Sanders, Li Liu, Silvia De Rubeis, Elaine T. Lim, James S. Sutcliffe, Gerard D. Schellenberg, Richard A. Gibbs, Mark J. Daly, Joseph D. Buxbaum, Matthew W. State, Bernie

Devlin, and Kathryn Roeder. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLOS Genetics*, 9(8):1–12, 2013.

[27] Michelle M. Li, Kexin Huang, and Marinka Zitnik. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369, 2022.

[28] Arthur Samuel. Some studies in machine learning using the game of checkers. 1967.

[29] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media, 2nd edition, 2019.

[30] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. 2017.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[34] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks, 2016. arXiv:1607.00653 [cs, stat].

[35] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations, 2014.

[36] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome, 2023.

[37] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[38] Miriam Bern, Alexander King, Derek A. Applewhite, and Anna Ritz. Network-based prediction of polygenic disease genes involved in cell motility. *BMC Bioinformatics*, 20(S12):313, 2019.

[39] Zhi-An Huang, Jia Zhang, Zexuan Zhu, Edmond Q. Wu, and Kay Chen Tan. Identification of autistic risk candidate genes and toxic chemicals via multilabel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):3971–3984, 2021.

[40] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2022.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[42] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association, 5th ed edition.

[43] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023.

[44] Ivan Iossifov, Brian J. O'Roak, Stephan J. Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A. Stessman, Kali T. Witherspoon, Laura Vives, Karynne E. Patterson, Joshua D. Smith, Bryan Paeper, Deborah A. Nickerson, Jeanselle Dea, Shan Dong, Luis E. Gonzalez, Jeffrey D. Mandell, Shrikant M. Mane, Michael T. Murtha, Catherine A. Sullivan, Michael F. Walker, Zainulabedin Waqar, Liping Wei, A. Jeremy Willsey, Boris Yamrom, Yoon-ha Lee, Ewa Grabowska, Ertugrul Dalkic, Zihua Wang, Steven Marks, Peter Andrews, Anthony Leotta, Jude Kendall, Inessa Hakker, Julie Rosenbaum, Beicong Ma, Linda Rodgers, Jennifer Troge, Giuseppe Narzisi, Seungtai Yoon, Michael C. Schatz, Kenny Ye, W. Richard McCombie, Jay Shendure, Evan E. Eichler, Matthew W. State, and Michael Wigler. The contribution of de novo coding mutations to autism spectrum disorder. 2014.

[45] Yi-Hsuan Pan, Nan Wu, and Xiao-Bing Yuan. Toward a better understanding of neuronal migration deficits in autism spectrum disorders. 7:205, 2019.

[46] Orly Reiner, Eyal Karzbrun, Aditya Kshirsagar, and Kozo Kaibuchi. Regulation of neuronal migration, an emerging topic in autism spectrum disorders. *Journal of Neurochemistry*, 136(3):440–456, 2016.

[47] Anbo Zhou, Xiaolong Cao, Vaidhyanathan Mahaganapathy, Marco Azaro, Christine Gwin, Sherri Wilson, Steven Buyske, Christopher W. Bartlett, Judy F. Flax, Linda M. Brzustowicz, and Jinchuan Xing. Common genetic risk factors in asd and adhd co-occurring families. *Human Genetics*, 142:217–230, 10 2022.

[48] L Ghirardi, I Brikell, R Kuja-Halkola, C M Freitag, B Franke, P Asherson, P Lichtenstein, and H Larsson. The familial co-aggregation of asd and adhd: a register-based cohort study. *Molecular Psychiatry*, 23:257–262, 02 2017.