Title

Prediction of Neurodevelopmental Disorders Based on De Novo Coding Variation

Journal

Journal of Autism and Developmental Disorders, 53(3)

ISSN

0162-3257

Authors

Chow, Julie C

Hormozdiari, Fereydoun

Publication Date

2023-03-01

DOI

10.1007/s10803-022-05586-z

Peer reviewed

Journal of Autism and Developmental Disorders (2023) 53:963–976
https://doi.org/10.1007/s10803-022-05586-z

ORIGINAL PAPER

Prediction of Neurodevelopmental Disorders Based on De Novo Coding Variation

Julie C. Chow[1] · **Fereydoun Hormozdiari**[1,2,3]

Accepted: 21 April 2022 / Published online: 20 May 2022

Abstract

The early detection of neurodevelopmental disorders (NDDs) can significantly improve patient outcomes. The differential burden of non-synonymous de novo mutation among NDD cases and controls indicates that de novo coding variation can be used to identify a subset of samples that will likely display an NDD phenotype. Thus, we have developed an approach for the accurate prediction of NDDs with very low false positive rate (FPR) using de novo coding variation for a small subset of cases. We use a shallow neural network that integrates de novo likely gene-disruptive and missense variants, measures of gene constraint, and conservation information to predict a small subset of NDD cases at very low FPR and prioritizes NDD risk genes for future clinical study.

Keywords De novo mutation · Early prediction · Neural network · Likely gene-disruptive · Missense

Abbreviations

ASD	Autism spectrum disorder
AUC	Area under the curve
CI	Confidence interval
DD	Developmental disability
FPR	False positive rate
ID	Intellectual disability
LGD	Likely gene-disruptive
LOEUF	Loss-of-function observed/expected upper bound fraction
NDD	Neurodevelopmental disorder
pLI	Probability of loss-of-function intolerance
PR-AUC	Precision recall-area under the curve
ROC-AUC	Receiver operating characteristic-area under the curve
RVIS	Residual Variation Intolerance
SSC	Simons Simplex Collection
SNN	Shallow neural network
SVM	Support-vector machine
TPR	True positive rate

✉ Julie C. Chow
jcchow@ucdavis.edu

✉ Fereydoun Hormozdiari
fhormozd@ucdavis.edu

[1] UC Davis Genome Center, University of California, Davis, CA 95616, USA

[2] MIND Institute, University of California, Davis 95817, USA

[3] Biochemistry and Molecular Medicine, University of California, Davis 95616, USA

Introduction

Neurodevelopmental disorders (NDDs), such as autism spectrum disorder (ASD), epilepsy, intellectual disability (ID), and developmental disability (DD) are complex disorders characterized by impairment in cognition, learning, and motor skills. From twin and family studies, it has become apparent that NDDs possess a strong genetic component (Freitag, 2007; Gejman et al., 2010). Estimates of heritability for various NDDs have ranged from 0.3 to 0.9, with heritability estimated to be greater than 0.5 for both ASD and ID (Flint, 2001; Kaufman et al., 2010; Tick et al., 2016). The evident contribution of genetic factors to NDD diagnoses has provided reason for routine prenatal whole exome or genome sequencing to identify potentially deleterious genetic variations (Soden et al., 2014; Tărlungeanu & Novarino, 2018). In particular, whole exome sequencing has proved a useful tool to identify, at a low cost, coding variants in genes that are highly intolerant to mutation and play important roles in typical neurodevelopment (Srivastava et al., 2019).

The identification and prioritization of NDD risk genes is important for the discovery of underlying biological mechanisms that are perturbed in NDDs (Cardoso et al., 2019).

Previous studies have identified many monogenic forms of NDDs and revealed the multifactorial and polygenic nature of most NDD diagnoses (De Felice et al., 2015; Niemi et al., 2018; Sztainberg & Zoghbi, 2016). In particular, rare de novo mutations that are observed in genes in NDD cases at a significantly higher rate than expected relative to unaffected controls have pinpointed many candidate NDD genes, with more than one thousand genes estimated to be NDD risk genes (De Rubeis et al., 2014; Heyne et al., 2018; Iossifov et al., 2012; Kaplanis et al., 2020; McRae et al., 2017; O'Roak et al., 2012a, b; Sanders et al., 2012; Satterstrom et al., 2020; Wilfert et al., 2017).

De novo mutations are a class of rare genetic variation in which variants, that are not observed in parental genomes, exist in offspring due to mutagenesis in germ cells or errors in replication or recombination (Acuna-Hidalgo et al., 2016). De novo mutations may exist as single nucleotide variants, insertions and deletions (indels), and copy number variants. Because de novo mutations are not inherited, highly penetrant mutations can arise in genes that are critical to neurodevelopment and likely under purifying selection (Iossifov et al., 2012; Uddin et al., 2014). In fact, individuals affected by NDDs experience a greater burden of non-synonymous de novo mutation compared to unaffected controls (Coe et al., 2019; Wilfert et al., 2017). Study of ASD simplex families from the Simons Simplex Collection (SSC) has found that de novo likely gene disruptive (LGD) mutations occur at a nearly twofold increased rate in affected cases (0.21) relative to controls (0.12), as well as displaying an increased rate of missense mutation (Iossifov et al., 2014). Furthermore, the study of genetic modules impacted by these de novo mutations has pinpointed several biological processes relevant to NDD etiology, such as chromatin remodeling, the Wnt pathway, synaptic transmission, and the long-term potentiation pathway (Chow et al., 2019; Kwan et al., 2016; O'Roak et al., 2012a, b; Wilfert et al., 2017).

The benefits associated with successful early prediction of NDDs include the improved ability of parents to make informed decisions about potential early application of treatments (Boivin et al., 2015; Cioni et al., 2016; Corsello, 2005). It is important to note that most NDDs cases cannot be predicted using de novo coding variation alone; the exome constitutes 1–2% of the human genome and the majority of NDD-associated variants are likely to reside in non-coding regions involved in the regulation of gene expression (Short et al., 2018; Turner & Eichler, 2019). Currently, it is estimated that only ~ 10% of ASD cases and ~ 20–30% of ID/DD cases have de novo LGD variants, and the rate of such variants in the general population is significantly lower (Wang et al., 2021). The genetically and phenotypically heterogeneous nature of NDDs indicates that many factors, including common and non-coding genetic variants and non-genetic factors, account for a large fraction of diagnoses, further complicating our ability for the early prediction of these disorders. However, it is possible to confidently predict a subset of individuals who will likely develop NDDs due to de novo coding variation in the form of non-synonymous de novo mutations. Despite the polygenic nature of NDDs and the multitude of potential genetic or environmental causes, focusing specifically on un-inherited, de novo mutations that disrupt protein coding sequence permits early prediction for a small fraction of cases with very low false positive rates.

The early prediction of NDDs requires a very low false positive rate (FPR) due to potential negative consequences, such as the costs associated with early intervention treatments, that may result from false positive prediction. The minimization of the FPR is clinically most relevant in genetic counseling settings for parents with suspected or confirmed familial risk for NDDs and to aid in the decision to begin early intervention treatments in young children. Early diagnosis of NDDs via a combination of behavioral and motor assessments, imaging, and genetic testing followed by early prediction methods can greatly benefit patient outcomes and lead to timely, appropriate treatment (Hadders-Algra, 2021). Previously, a method for the early prediction of complex disorders, Odin, used de novo LGD variants observed in cases and controls and co-expression data to identify cases at very low FPR (Huynh & Hormozdiari, 2018). The shallow neural net (SNN) with novel objective function introduced here incorporates LGD de novo mutation, constraint, and conservation data to achieve a higher (>0.30129) true positive rate (TPR) at very low FPR (<0.01) in comparison to traditional classification models such as random forest (RF), support-vector machine (SVM), and logistic regression (LR). Furthermore, the proposed SNN model achieves similar PR-AUC and ROC-AUC to other machine learning approaches. An ensemble model that averages predictions among the SNN, RF, SVM, and LR models is able to achieve a slightly increased TPR at FPR <0.01 and comparable PR-AUC. Additionally, the SNN is able to rank genes according to their relative importance in NDDs given LGD or missense de novo variation, prioritizing candidate NDD genes.

Methods

Objective

The main objective is to investigate the potential of using machine learning approaches for early prediction of NDDs using de novo coding genetic variants in a subset of cases. More formally, we are interested in utilizing de novo coding variants in maximizing the fraction of affected NDD cases accurately predicted when limiting the FPR to virtually zero.

Data Collection and Preprocessing

To distinguish NDD cases from unaffected controls using de novo coding variation, de novo likely gene-disruptive (LGD) and missense variants were retrieved from denovo-db (version 1.6.1) (Turner et al., 2017, p.). These data consisted of 9962 individuals with primary phenotypes of autism spectrum disorder (ASD), intellectual disability, and developmental disability and 2245 controls, of which 6509 cases and 1251 controls possess non-synonymous coding de novo mutation (Supplementary Table 1). In total, the 7760 samples possessed 1974 LGD (cases: 1715; controls: 259) and 10,777 (cases: 9073; controls: 1704) missense de novo coding mutations. *PrimateAI* scores were used to quantify the pathogenicity of missense variants, in which position-specific scores were calculated for each missense variant while incorporating conservation, solvent accessibility, and secondary structure data to yield predictions of deleteriousness (Sundaram et al., 2018). Probability of loss-of-function intolerance (pLI) and loss-of-function observed/expected upper bound fraction (LOEUF) scores from the gnomAD browser (v2.1.1), Residual Variation Intolerance (RVIS) scores based on ExAC v2 release 2.0 (March 15, 2017 version), and phastCons element scores were also used as features (Karczewski et al., 2020; Petrovski et al., 2013; Siepel et al., 2005).

LGD-specific and missense-specific feature matrices were generated, in which rows represent individuals with LGD or missense variation from denovo-db and columns represent genes containing de novo mutations (Fig. 1A, Additional File 1).

Model Architecture Development and Hyperparameter Tuning

Separate models were trained for de novo LGD variation and missense variation, referred to as SNN LGD-specific and missense-specific models. Each variation-specific SNN consists of two phases, a hyperparameter optimization phase and a prediction phase. After splitting all samples into training (75%) and testing (25%) sets, the hyperparameter optimization phase is applied to the training set, choosing optimal hyperparameters within a selected search space (Fig. 1B, Supplementary Table 2, Additional File 1). The purpose of the hyperparameter optimization phase for the SNN is to select a set of hyperparameters that yield the largest TPR at very low FPR on the training set to use during the prediction phase. Similarly, RF (sklearn.ensemble.RandomForestClassifier), SVM (sklearn.svm.LinearSVC), and LR (sklearn.linear_model.LogisticRegression) classifiers, hereon referred to as baseline models, are individually subjected to hyperparameter optimization and prediction phases. To allow direct comparison of each model's performance, identical training/testing splits are provided to SNN and baseline models. The performance of SNN and baseline models are additionally compared to the TPR and FPR of the following heuristics, in which an individual is classified as a case if the individual has an LGD mutation in: (1) any gene with a (i) SFARI score of 1 (high confidence ASD gene) or (ii) SFARI score of 1 or 2 (strong candidate ASD gene) (https://gene.sfari.org/database/gene-scoring/), (2) any gene identified by SPARK as a (i) prioritized or (ii) risk gene, and (3) any gene with i) pLI $>=0.90$ or (ii) LOEUF <0.35 (Additional File 1).

In the hyperparameter optimization and prediction phases (Fig. 1C),

$$loss = 1 - (TP - (\lambda_1 * FP)) \quad (1)$$

is used as a custom loss function (Eq. 1) for the SNNs, in which the objective is to minimize the product of the number of false positives (FP) and the hyperparameter λ_1 subtracted from the true positives (TP). The value of λ_1 is selected during the hyperparameter optimization phase. The SNN architecture consists of an input layer, a hidden layer with ReLU activation and an optimized number of neurons, and an output layer with sigmoid activation and L2 regularization with an optimized regularization parameter λ_2. The SNN uses the Adam optimization algorithm.

To return a prediction that incorporates both LGD and missense variation for individuals who possess both types of variants simultaneously (referred to as a 'combined' prediction), predictions are retrieved from the separately trained LGD- and missense-specific models for SNN and baseline models. For a given sample with both LGD and missense variation, the maximum predicted probability from the two separately trained variation-specific models is returned as the predicted probability of being a case primarily due to de novo coding variation (Fig. 1D). By using the maximum predicted probability, the model is trained to learn the class of an individual given their de novo mutation that is predicted to have the largest deleterious effect.

The average performance of a model over 100 independent training and testing splits is measured by determining the average TPR at FPR <0.01, ROC-AUC, and PR-AUC for LGD-specific, missense-specific, and combined predictions for the SNN approach using the custom loss function, three baseline models, an ensemble model, and an ensemble model excluding SNN predictions (Additional File 1). To demonstrate the importance of gene score features and PrimateAI scores to increased TPR at FPR <0.01, SNN and baseline models were trivially trained on one-hot encoded feature matrices indicating only the presence or absence (denoted as 1 or 0, respectively) of de novo LGD or missense mutation, and performance metrics were returned. To additionally assess the performance of the missense-specific model using only deleterious missense

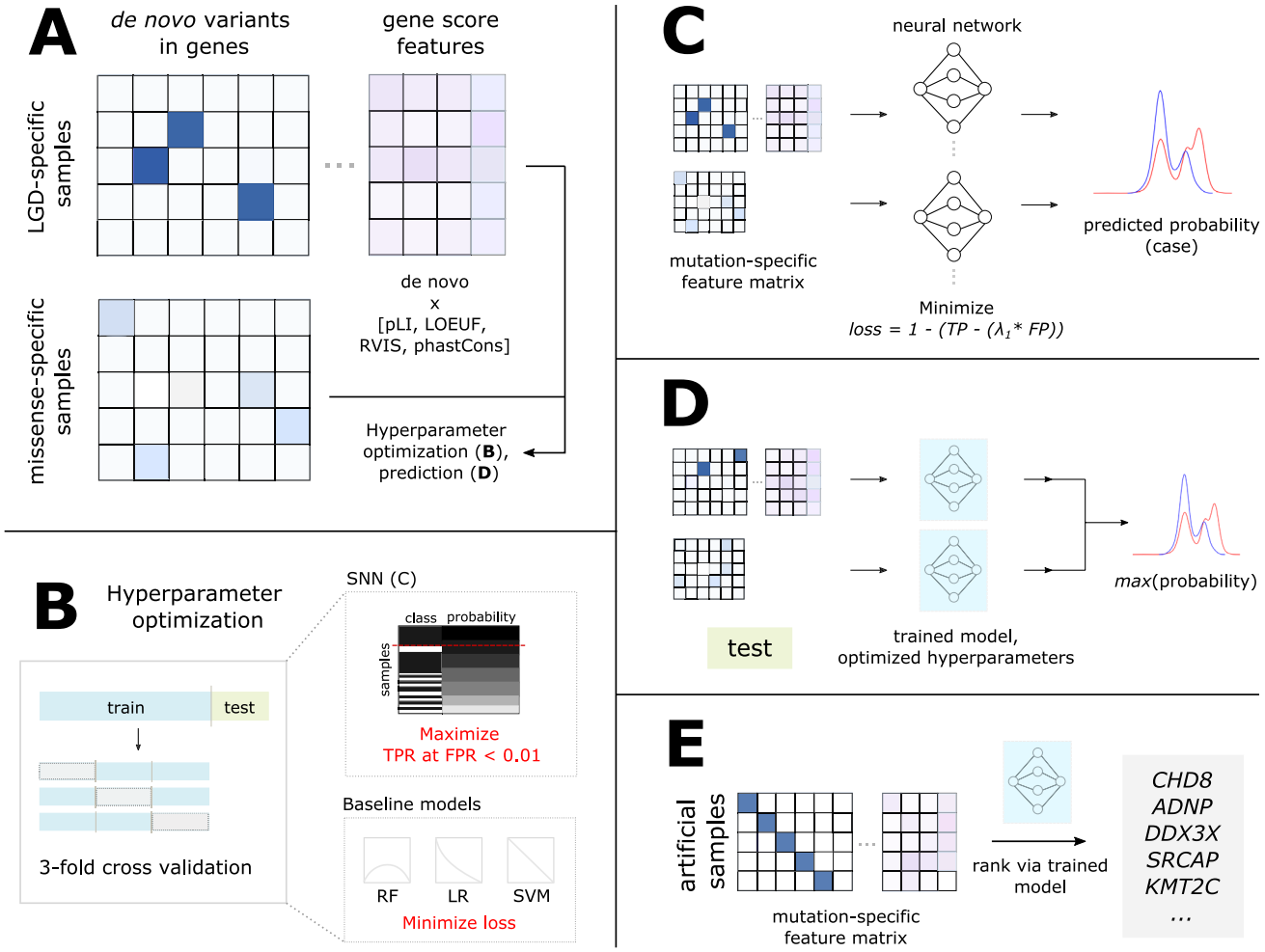


Fig. 1 Methods overview. **A** De novo LGD and missense variants from probands with NDDs and controls were retrieved from denovo-db and arranged into feature matrices. Constraint and conservation information, in the forms of pLI, LOEUF, RVIS, and average phastCons element conservation scores were incorporated as gene score features (Karczewski et al., 2020; Petrovski et al., 2013; Siepel et al., 2005) (Additional File 1). **B** To perform hyperparameter optimization and model training, samples were divided into training (75%) and testing (25%) sets. Hyperparameter optimization occurs via threefold cross validation on the partitioned training set. For the SNN model (**C**), performance is measured as the TPR at FPR < 0.01, which is calculated by determining the number of cases (class: black) with predicted probability greater than that of any control (class: white) in the validation fold. For baseline models, consisting of the RF, LR, and SVM classifiers, respective loss functions are minimized. **C** The SNN consists of a single hidden layer and a loss function that seeks to minimize the product of predicted FP and a parameter λ_1, subtracted from the TP. **D** During the prediction phase, using the model trained with optimized hyperparameters, a prediction is made on the withheld testing set. For samples that simultaneously have both LGD and missense variation, two separate probabilities are retrieved from LGD- and missense-specific models for a given individual, and the maximum predicted probability is returned per individual. **E** For ranking genes based on their importance to NDDs, artificial samples are generated such that each artificial sample has a single de novo variant in a unique gene, using either LGD or missense variation, separately. Application of the prediction phase (**D**) on artificial samples yielded a ranking of the relative importance of a gene to NDDs determined via de novo coding variation

variation with PrimateAI scores ≥ 0.803 (as described in Sundaram et al., 2018), the missense-specific model (i) was trained using only samples with deleterious missense variation (PrimateAI ≥ 0.803) without discarding any samples, or (ii) was executed while excluding samples without deleterious missense mutation from training and testing sets.

NDD Gene Ranking

To rank genes according to their relative importance to NDDs using de novo coding variation in the form of de novo LGD mutations or missense mutations, two sets of artificial samples (LGD- and missense-specific) were created. The artificial samples each contain a single LGD (or missense)

variant in a unique gene in the human genome (Fig. 1E). The probability of being a case is predicted for each of these artificial samples using the previously trained SNN LGD- or missense-specific models. For every artificial sample and its corresponding gene containing a de novo LGD (or missense) variant, the predicted probability indicates the relative importance of the gene to NDD risk from de novo coding variation. Enrichment of de novo LGD and missense mutation in NDD cases relative to controls was assessed (Additional File 1).

Results

To identify, at very low FPRs, a subset of affected NDD cases using rare coding variation consisting of de novo LGD and missense variation, LGD- and missense-specific feature matrices indicating the presence of de novo variation within genes were constructed (Fig. 1A). Additional features incorporating constraint and conservation data were used to improve classification of NDD cases using LGD variation. The ability of SNNs (Fig. 1C) to classify NDD cases at very low FPRs were compared to various classifiers, including RF, SVM, and LR (baseline models), in addition to three heuristics.

De Novo LGD Mutations Distinguish a Subset of NDD Cases from Controls with Low FPR

At very low FPRs (FPR < 0.01), an SNN trained on an LGD-specific feature matrix captures 30.1% of NDD cases possessing any de novo LGD coding variation. In comparison to baseline models, the SNN trained on an LGD-specific feature matrix is able to identify 5.29–10.25% [95% confidence interval (CI)] more NDD cases at FPR < 0.01 than the RF classifier, and more than 5.73–17.26% (95% CI) NDD cases than SVM or LR models (Fig. 2, Table 1, Supplementary Fig. 1). To measure the extent to which the SNN and other models achieve increased TPR at FPR < 0.01 compared to a randomized model, a z-score was also calculated (Additional File 1, Table 1).

For the SNN, ROC-AUC and PR-AUC values of 0.72785 (0.7227–0.7326, 95% CI) and 0.9505 (0.9490 to 0.9519, 95% CI), respectively, were observed (Table 1). Observed PR-AUC values were comparable among the SNN and baseline models in their deviance from the randomized model, displaying similar z-scores. Note that due to the large number of cases in proportion to controls in available datasets, PR-AUC values for SNN and baseline models are significantly inflated; the random assignment model has an PR-AUC of over 0.85.

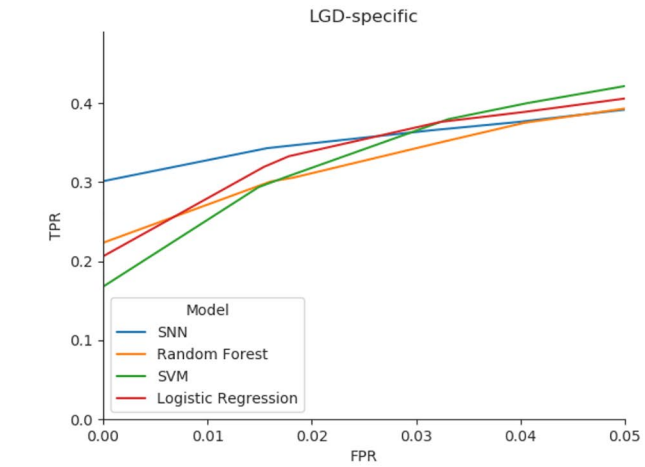
The inclusion of gene score features derived from pLI, LOEUF, RVIS, and phastCons element scores improves upon an SNN trivially trained only on LGD mutations (TPR at FPR < 0.01 = 0.24532, ROC-AUC = 0.66696, PR-AUC = 0.93597) (Supplementary Table 3, Supplementary Fig. 2). In addition, baseline and SNN models yield similar performance metrics when trivially trained on only LGD mutations, indicating. that the inclusion of gene constraint and conservation information is important to accurate classification of NDD cases using de novo LGD mutations (Supplementary Table 3).

Compared to the TPR and FPR of the three previously described heuristics, in which a sample was classified as a case if the sample possessed an LGD mutation in a set of prioritized genes, decreased TPR at low FPR thresholds in comparison to the SNN was observed for each heuristic (Supplementary Table 4, Supplementary Fig. 3). No heuristic achieved similar TPR values greater than 0.30 at FPR less than 0.01.

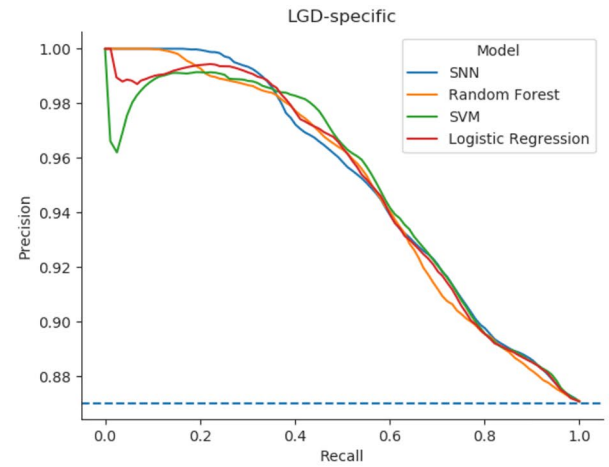
Integration of Missense and LGD-Specific Models Improves Prediction on Individuals with Both De Novo Missense and LGD Variants

To assess the ability of de novo missense mutations to distinguish NDD cases from unaffected controls, de novo missense variants from individuals with at least one missense variant were retrieved, consisting of 6947 samples possessing a total of 10,777 missense mutations. SNN and baseline models trained on missense variation capture less than 2.6% of NDD cases at FPR < 0.01 (Fig. 2, Table 1), indicating that accurate prediction of NDDs using only missense de novo variants is an extremely challenging problem. Slightly increased TPR at FPR < 0.01 is observed when the missense-specific model is trained only on deleterious missense variation without removing any samples from training and testing; excluding samples without deleterious missense variation from training and testing yields 2242 samples (2257 cases; 248 controls) with 2,505 deleterious missense variants and increased TPR at FPR < 0.01 (Supplementary Table 5).

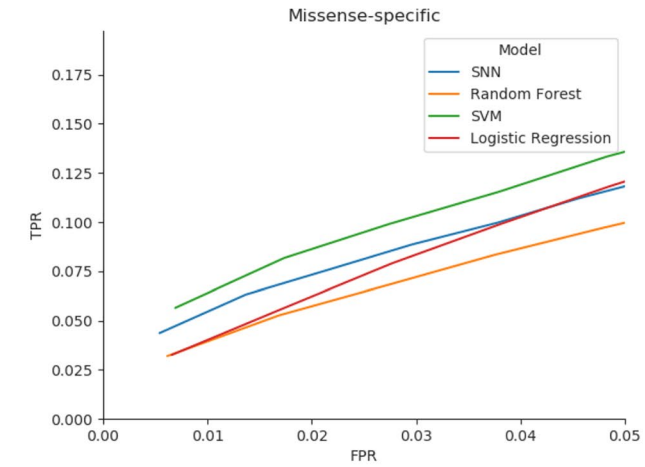
For samples possessing both de novo missense and LGD variants, accurate prediction of NDD cases at low FPR can be improved by taking the maximum predicted probability from two models trained separately on only missense or LGD variation, referred to as a 'combined' prediction (Fig. 2, Table 1). Combined prediction on samples with both missense and LGD variation captures an increased fraction of cases. For example, compared to the LGD-specific SNN, an SNN using combined prediction is able to detect at most 4.22% more affected cases at FPR < 0.01 (95% CI). TPR at FPR < 0.01—associated z-scores for the SNN are greater by 1.41- 2.51 than values observed for baseline models using combined predictions.


LGD-specific
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

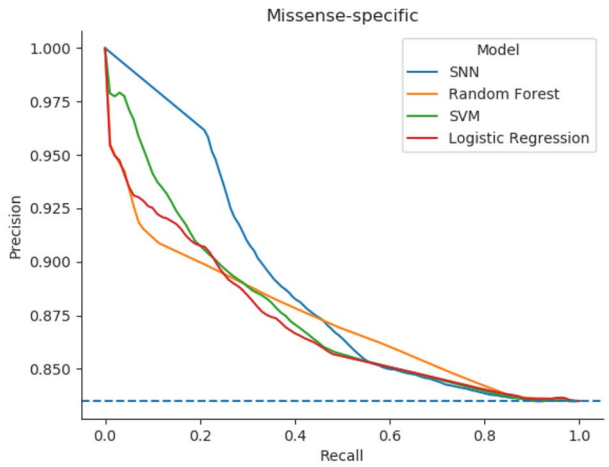
A


LGD-specific
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

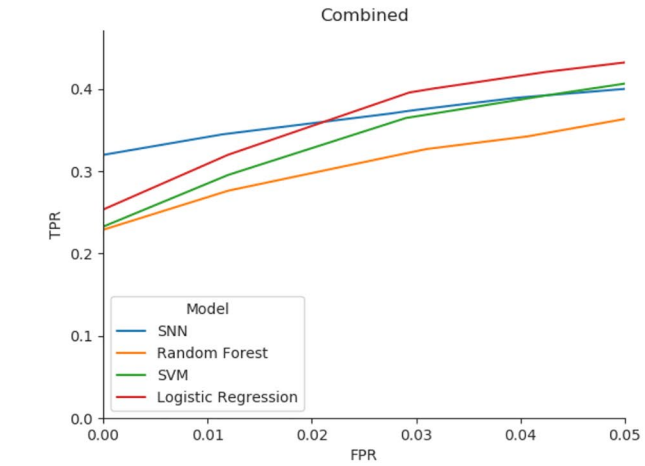
B


Missense-specific
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

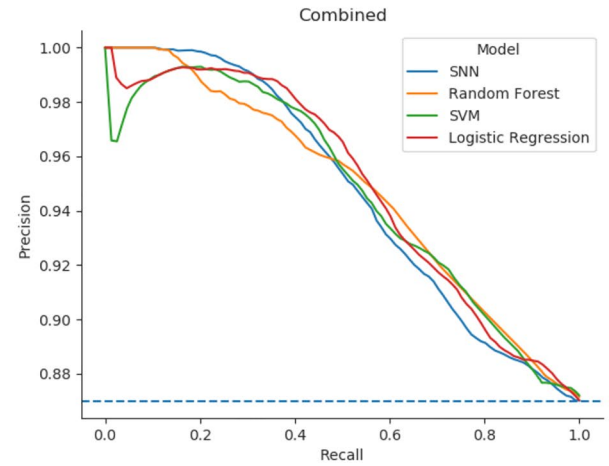
C


Missense-specific
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

D


Combined
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

E


Combined
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

F

◂**Fig. 2** Receiver operating characteristic and PR curves for LGD- and missense-specific and combined predictions for SNN and baseline (RF, SVM, and LR) models. Random classification is displayed as a dashed blue line in all PR curves. Models trained on LGD-specific variation feature matrices additionally use constraint and conservation gene score information, whereas models provided with missense-specific feature matrices do not use gene score information. For LGD-specific features, the SNN achieves greater TPR at low FPR < 0.01 compared to baseline models, a trend which is evident even at FPR < 0.05 (**A**), and the SNN achieves comparable precision at lower recall compared to baseline models (**B**). Models trained on missense-specific variation alone are poor predictors of NDD status; SNN and baseline models show similar TPR at FPR < 0.05, with the SNN achieving slightly higher rates at low FPR (**C**). The SNN displays comparable precision at low recall thresholds when trained on missense-specific variation (**D**). **E** For combined prediction for samples with both missense and LGD variation, the proportion of cases captured at FPR < 0.01 is largest for the SNN, and similar precision at low recall is observed for the SNN compared to baseline models (**F**)

Ensemble Prediction Yields Increased TPR at Very Low FPRs Compared to Separately Trained SNN and Baseline Models

An ensemble prediction was generated by returning the average predicted probability from the SNN, RF, SVM, and LR models for a given sample in the testing set. Compared to SNN and baseline models for LGD-specific, missense-specific, and combined models, the ensemble model consistently yields a larger TPR at low FPR values (Supplementary Fig. 4, Table 1). The predictive contribution of the SNN to the ensemble model is more substantial than that of the baseline models. For example, the TPR at FPR < 0.01 is greater for LGD-specific and combined prediction SNNs than ensemble models that exclude SNN predictions, referred to as 'Ensemble—SNN' (Table 1, Supplementary Fig. 4). Additionally, for LGD-specific and combined predictions, there is no overlap of 95% CIs between SNN and Ensemble—SNN models. From the ensemble prediction's constituent models, the SNN performs most similarly to the full ensemble method, differing by 0.586% and 1.582% in TPR at FPR < 0.01 given LGD-specific and combined predictions, respectively. In addition, the ensemble model achieves a slightly higher average PR-AUC, as evidenced by an increased z-score, than any individual SNN or baseline model for corresponding LGD-specific (0.95176) or combined predictions (0.95215) (Table 1).

Integration of Constraint, Conservation, and De Novo Mutation Data Permit NDD Gene Prioritization

Training of SNNs (Fig. 1C) on variation-specific feature matrices enables NDD risk gene ranking according to the effect of de novo missense and LGD mutations within specific genes (Fig. 1E). For example, using only LGD variants during SNN training reveals genes that are sensitive to LGD mutations and play important roles in typical neurodevelopment. Gene rankings and associated SFARI Gene scores are displayed in Supplementary Table 6 in descending order according to their relative importance to NDD risk.

For artificial LGD samples (that each possess a single LGD variant in a unique gene), an increased enrichment of LGD variants is observed in NDD cases relative to unaffected controls at increasing predicted probabilities (Fig. 3A), and a slight increased enrichment of missense variants is also observed in NDD cases for genes ranked according to a trained LGD-specific SNN (Fig. 3B). The difference in enrichment (E_{diff}) of LGD or missense mutation in cases relative to controls per gene is calculated by Eq. 2 (Additional File 1). Significant correlation exists between pLI ($p\text{-}value < 2.25e-79$) and LOEUF ($p\text{-}value < 1.09e-63$) values with predicted probability ranks for artificial LGD samples (Fig. 3C, D). For gene rankings produced by a missense-specific SNN, similar trends in enrichment of de novo coding variation in NDD cases relative to controls are observed, although the range of probabilities predicted by the missense-specific SNN narrows compared to the LGD-specific SNN, and the strength of correlation amongst pLI and LOEUF values with predicted probabilities is reduced (Supplementary Fig. 5).

For the LGD-specific SNN model, inclusion of gene score features generated from pLI, LOEUF, RVIS, and PhastCons produces rankings with greater enrichment of LGD variation in cases relative to controls at higher probabilities than an LGD-specific SNN model trivially trained on one-hot encoded mutation information that excludes gene score features (Supplementary Fig. 6).

Discussion

To distinguish NDD cases from unaffected controls at extremely low FPRs using de novo coding variation and measures of gene constraint and conservation, we developed a SNN with a customized objective function to maximize TP while simultaneously minimizing FP (Fig. 1). Although most cases of NDDs arise from a variety of classes of genetic variation, particularly common, non-coding, and structural variants, focusing specifically on de novo coding variation of relatively large effect size is a tradeoff to obtain significantly reduced FPR on a small but significant subset of samples. Compared to traditional machine learning techniques, such as RF, support vector machine (SVM), and LR (referred to as 'baseline' models), the constructed SNN is able to achieve greater TPR at FPR less than 0.01 given LGD-specific variation (Table 1, Fig. 2). The ability of the SNN to capture

Table 1 Average TPR at FPR < 0.01, ROC-AUC, and PR-AUC for LGD-specific, missense-specific, and combined SNN, baseline, ensemble models, and randomized predictions

Input features	Model	TPR at FPR < 0.01 (95% CI); z-score	ROC-AUC (95% CI); z-score	PR-AUC (95% CI); z-score
LGD-specific	SNN	0.30129 (0.2906, 0.3124); 4.93244	0.72785 (0.7227, 0.7326); 4.01329	0.95050 (0.949, 0.9519); 5.86600
	Random forest	0.22342 (0.2099, 0.2377); 2.83170	0.71997 (0.7154, 0.7244); 3.95991	0.94866 (0.9472, 0.95); 5.81660
	SVM	0.16790 (0.1398, 0.1962); 1.04685	**0.73199** (0.7278, 0.7365); 4.18017	0.94825 (0.9463, 0.9498); 5.33855
	Logistic regression	0.20632 (0.18, 0.2333); 1.34869	0.72695 (0.7222, 0.7317); 4.06566	0.94877 (0.9471, 0.9504); 5.58760
	Ensemble	**0.30715** (0.2965, 0.3174); 5.08163	0.73037 (0.7261, 0.7347); 4.14049	**0.95176** (0.9504, 0.953); 6.08741
	Ensemble—SNN	0.23347 (0.2213, 0.2453); 3.33032	0.72823 (0.724, 0.7325); 4.10213	0.95023 (0.9488, 0.9515)); 6.00325
	Randomized	0.01660 (0.0135, 0.0202)	0.50627 (0.4963, 0.5164)	0.8698
Missense-specific	SNN	0.02334 (0.0199, 0.0267); 1.09477	0.54378 (0.5391, 0.5483); 1.23832	**0.88139** (0.878, 0.885); 2.40309
	Random forest	0.01279 (0.0109, 0.0151); 0.78867	0.53086 (0.5287, 0.533); 1.17197	0.87220 (0.8705, 0.8738); 3.97519
	SVM	**0.02610** (0.022, 0.0301); 1.09631	0.55910 (0.5564, 0.5618); 2.22556	0.87486 (0.8737, 0.876); 5.88837
	Logistic regression	0.01214 (0.0101, 0.0144); 0.72456	0.55810 (0.5551, 0.5609); 2.13551	0.87071 (0.8694, 0.872); 4.82097
	Ensemble	0.02530 (0.022, 0.0288); 1.18239	**0.56006** (0.5571, 0.5631); 2.18983	0.87374 (0.8726, 0.8749); 5.71154
	Ensemble—SNN	0.02386 (0.0205, 0.0272); 1.13687	0.55915 (0.5564, 0.5619); 2.18614	0.87383 (0.8726, 0.8751); 5.69270
	Randomized	0.00406 (0.0033, 0.0048)	0.50304 (0.4991, 0.5071)	0.8350
Combined	SNN	0.31985 (0.3038, 0.3348); 3.55285	0.71422 (0.7071, 0.7215); 2.93749	0.94685 (0.9445, 0.949); 3.95676
	Random forest	0.22892 (0.2129, 0.2456); 2.39793	0.71830 (0.7121, 0.7246); 3.15223	0.94740 (0.9453, 0.9494); 4.02305
	SVM	0.23267 (0.2058, 0.2598); 1.41386	0.72803 (0.7211, 0.7346); 3.21778	0.94620 (0.9437, 0.9486); 3.67270
	Logistic regression	0.25347 (0.226, 0.2837); 1.48639	0.73280 (0.7269, 0.7389); 3.34153	0.94874 (0.9466, 0.951); 4.05063
	Ensemble	**0.33567** (0.3216, 0.3508); 3.87773	**0.74128** (0.7345, 0.7481); 3.42116	**0.95215** (0.9501, 0.9541); 4.35673
	Ensemble—SNN	0.23961 (0.2249, 0.2549); 2.63914	0.73737 (0.7302, 0.7447); 3.30974	0.94899 (0.9468, 0.951); 4.09443
	Randomized	0.02898 (0.0224, 0.036)	0.53177 (0.5218, 0.5409)	0.8701

An ensemble model generated only from the predictions of baseline models while excluding SNN predictions is referred to as 'Ensemble—SNN'. To generate randomized predictions, probabilities drawn from a uniform distribution were randomly assigned to samples. Average performance metrics are measured over 100 independent iterations of randomized training/testing splits on the testing set, in which the same training/testing partition is provided to all models at each iteration. Confidence intervals (95% CI) are indicated in parentheses, followed by a z-score quantifying the deviance from the mean performance metric of a certain model and the randomized model (Additional File 1). The PR-AUC values associated with randomized predictions were calculated by dividing the number of cases in a testing set by the total number of samples within the testing set. The largest average TPR at FPR < 0.01, ROC-AUC, and PR-AUC values are bolded for LGD-specific, missense-specific, and combined models

more than 30% of cases at FPR < 0.01, corresponding to at least 5.29% more cases than any baseline model (Table 1), indicates that the use of a SNN with the custom loss function (Eq. 1) is beneficial in classifying NDD cases at very low FPR. Note that it is estimated that LGD variants have been observed in roughly 10% of ASD cases and up to 30% of DD cases (Wang et al., 2021). Thus, our results indicate that the proposed SNN should be able to identify > 3% of ASD

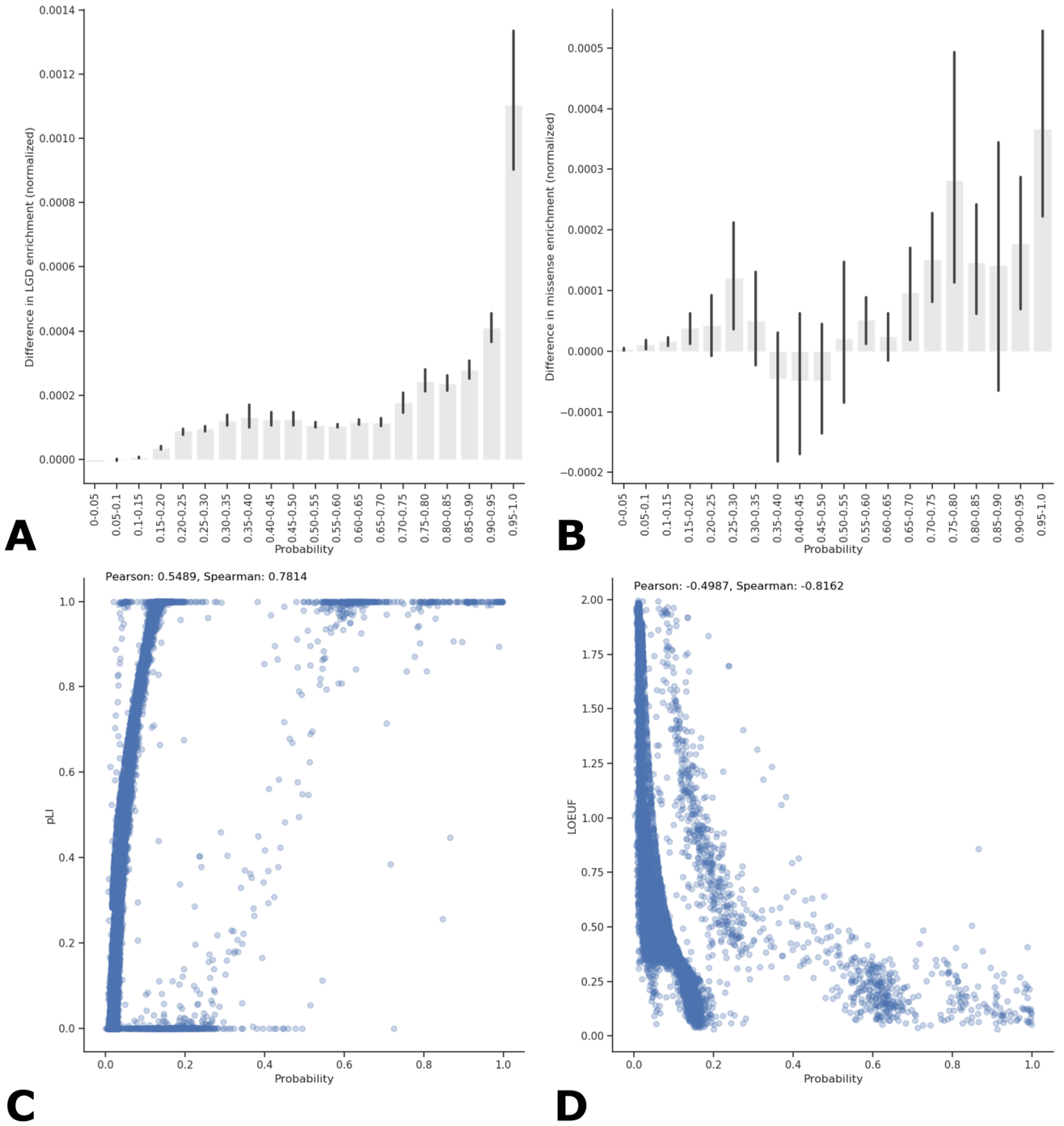


Fig. 3 Increased enrichment of de novo LGD and missense mutation in NDD cases relative to unaffected controls in highly ranked NDD genes according to an SNN trained on an LGD-specific feature matrix. Applying a trained SNN on artificial samples containing a single unique LGD variant allows the SNN to rank genes according to their relative importance to NDD risk with respect to LGD coding variation. The difference in enrichment in NDD cases versus controls per ranked gene is calculated by Eq. 2 (Additional File 1) and displayed on the y-axes. Increasing probability (x-axes) indicates increasing importance to NDD risk. The average predicted probability was determined for each artificial sample over 100 independent iterations, and 95% CI are shown. At increasing probabilities for artificial samples with LGD variants, a steady, increased enrichment of LGD in cases (A) is observed, and a slight enrichment of missense variation (B) in cases relative to controls is also observed at increasing probabilities. The probability (ranks) assigned to genes is significantly correlated with both pLI (C) and LOEUF (D) values retrieved from gnomAD (v2.1.1). pLI values range from 0 to 1, where values above 0.9 suggest intolerance to LGD mutation, whereas LOEUF values represent a ratio of observed over expected LGD mutations and values less than 0.35 suggest intolerance to LGD mutation

and > 10% of all DD cases while having an FPR of virtually zero simply by considering de novo LGD variants.

To demonstrate that gene scores related to constraint and conservation, including pLI, LOEUF, RVIS, and phastCons, were useful and necessary for the SNN to yield elevated TPR at FPR < 0.01 compared to baseline models given LGD-specific variation, the performance of trivially trained SNN and baseline models were measured (Supplementary Table 3, Supplementary Fig. 2). During trivial training, *only* a feature matrix of one-hot encoded values (1 or 0) denoting the presence or absence of a de novo coding variation within a gene were provided as input features to models. We note that most de novo mutations retrieved from denovo-db were identified via simplex studies that facilitate the identification of de novo variants, thus potentially introducing biased prediction in favor of variants identified via simplex rather than multiplex studies. We would also like to note that multiplex NDD cases will have a potentially lower chance of being caused by de novo variants and thus reduce the ability of our model's accurate prediction of these cases. Similar TPR at FPR < 0.01 values are reported for trivially trained and trivially trained baseline models, indicating that the inclusion of gene score features greatly contributes to the SNN's improved ability to classify NDD cases at very low FPR.

In addition, a simple ensemble method that uses the average predicted probability from SNN and baseline model predictions is able to identify NDD cases at greater TPR at FPR < 0.01 and slightly increased precision at lowered recall than any of its constituent models (Table 1, Supplementary Fig. 4). Excluding SNN predictions from the ensemble model reveals that the SNN, compared to baseline models, contributes substantially to the ensemble model's ability to accurately classify NDD cases at low FPR values. In fact, for LGD-specific variation, an ensemble method that excludes SNN predictions produces decreased TPR at FPR < 0.01 metrics compared to the SNN alone (Table 1).

The ability of SNN and baseline models to use only missense variation to identify NDD cases is relatively poor. However, the incorporation of both missense and LGD-specific predictions during 'combined' prediction for samples containing both LGD and missense variation, in which the maximum predicted probability from two separately trained missense- and LGD-specific models are returned, increases average TPR at FPR < 0.01 compared to using only probabilities predicted by an LGD-specific model (Table 1, Fig. 2). The improved performance of combined predictions indicates that certain samples possessing very deleterious missense variation (in addition to LGD variation) are correctly classified as cases when the predicted probability associated with the missense-specific model, rather than the LGD-specific model, is retrieved.

SNNs trained on LGD- and missense-specific feature matrices containing de novo coding variation from NDD cases and controls are able to rank genes according to their relative importance to NDD risk when applied to artificial samples which each contain a single type of de novo variant in a single gene (Supplementary Table 6). An increased enrichment of de novo LGD and missense mutation in NDD cases relative to controls is observed in highly ranked genes (those with higher predicted probability of being a case) using LGD-specific variation (Fig. 3). Significant, strong correlation exists between predicted probability for artificial samples for both the pLI and LOEUF constraint metrics, showing that the ranking via LGD-specific variation can accurately detect most high risk NDD genes. Among the 50 most highly ranked genes using LGD-specific variation, a total of 47 out of 50 genes are classified as high confidence (39 genes with score 1), strong candidate (6 genes with score 2), and suggestive evidence (2 genes with score 3) autism spectrum disorder (ASD) risk genes, including genes relevant to syndromes, according to SFARI Gene and OMIM (Supplementary Table 6). Among genes with predicted probabilities greater than 0.90 (ranks 1–55), four genes (*WDR45*, *CLTC*, *BRPF1*, and *GATAD2B*) do not possess SFARI annotations, but have been associated with neurodegeneration and intellectual disability according to OMIM annotations. Highly ranked genes lacking both SFARI Gene scores and OMIM annotations suggest candidate NDD genes susceptible to de novo LGD variation. Evidence of association with NDDs [*ZFHX3* (Fuller et al., 2018), *CHD5* (Parenti et al., 2021), *UBR3* (Murcia Pienkowski et al., 2020)) or enrichment of de novo LGD mutation in NDD cases (*ANP32A, SKIDA1* (Coe et al., 2019)), neurodegeneration (*ANP32A* (Podvin et al., 2020), *HECTD1* (Schmidt et al., 2021)), gliomas (*LARP4B* (Koso et al., 2016)), synapses and neuronal formation (*LMTK3* (Takahashi et al., 2020), *DOT1L* (Franz et al., 2019)] have been studied in model organisms, cell lines, and families for these candidate NDD genes.

Weaker correlation is observed for missense-specific rankings with pLI and LOEUF values, and enrichment of de novo non-synonymous mutation is also present in NDD cases relative to controls, although to a lesser extent compared to LGD-specific rankings (Supplementary Fig. 5). The missense-specific rankings are distinct from LGD rankings in their ability to identify genes potentially sensitive to missense variation (Supplementary Table 6). Among highly ranked genes lacking SFARI Gene scores and OMIM annotations, previous studies suggest association with NDDs and schizophrenia [*OBSCN* (Hashimoto et al., 2016), *PLEC* (Dincer et al., 2015), *RYR2* (Lieve et al., 2019), *ZSWIM8* (Tischfield et al., 2017)], cortical formation and thickness [*LAMA5* (Omar et al., 2017), *GOLGA3* (Kim et al., 2017)), and neurodegenerative diseases (*PKHD1* (Santos-Laso et al., 2020), *DNAH*1 (Thonberg et al., 2017)].

Our results indicate that we can accurately predict a small, yet significant fraction of NDD cases using de novo coding variants. Currently, whole-exome or whole-genome sequencing of trios is not common practice. However, to make the early prediction of these disorders a reality, such sequencing should become common practice. Furthermore, our approach only covers a small fraction of affected patients and additional methods that use other types of biomolecular signatures, such as common variants, rare non-coding variants, and epigenomic markers, are needed to increase the reach of early prediction to a larger fraction of cases.

Conclusions

In summary, the described SNN identifies NDD cases at higher TPR while having very low FPR in comparison to traditional machine learning methods. Several factors contribute to the improved performance of the proposed approach, namely: the use of gene constraint and conservation features in LGD-specific prediction and a custom loss function that specifically seeks to maximize the TPR while minimizing the FPR. An ensemble method, aggregated from SNN and baseline model predictions, is able to correctly classify a greater proportion of cases at FPR < 0.01 compared to any individual model. The SNN itself is a major contributor to increased TPR at FPR < 0.01 observed in the ensemble model. Although de novo missense mutation alone is a poor predictor of case status relative to LGD mutation, missense-specific predictions are useful during combined prediction for identifying additional cases that possess highly deleterious missense mutation in addition to LGD mutation. Fully trained SNNs on LGD- or missense-specific variation are also useful in NDD risk gene prioritization, revealing candidate NDD genes enriched in de novo non-synonymous mutations in NDD cases relative to controls.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10803-022-05586-z.

Acknowledgements We would like to acknowledge Farhad Hormozdiari for help with discussing the paper and providing helpful comments.

Author Contributions JC wrote the SNN and associated scripts and performed experiments devised by FH and JC. JCand FH wrote the manuscript. All authors read and approved the final manuscript.

Funding This work has also been supported in part by NSF award DBI-2042518 to FH.

Data Availability Codes associated with the SNN and analysis are available at https://github.com/jchow32/EarlyPredictionSNN.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology, 17*(1), 241. https://doi.org/10.1186/s13059-016-1110-1

Boivin, M. J., Kakooza, A. M., Warf, B. C., Davidson, L. L., & Grigorenko, E. L. (2015). Reducing neurodevelopmental disorders and disability through research and interventions. *Nature, 527*(7578), S155–S160. https://doi.org/10.1038/nature16029

Cardoso, A. R., Lopes-Marques, M., Silva, R. M., Serrano, C., Amorim, A., Prata, M. J., & Azevedo, L. (2019). Essential genetic findings in neurodevelopmental disorders. *Human Genomics*. https://doi.org/10.1186/s40246-019-0216-4

Chow, J., Jensen, M., Amini, H., Hormozdiari, F., Penn, O., Shifman, S., Girirajan, S., & Hormozdiari, F. (2019). Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. *Genome Medicine, 11*(1), 65. https://doi.org/10.1186/s13073-019-0678-y

Cioni, G., Inguaggiato, E., & Sgandurra, G. (2016). Early intervention in neurodevelopmental disorders: Underlying neural mechanisms. *Developmental Medicine and Child Neurology, 58*(Suppl 4), 61–66. https://doi.org/10.1111/dmcn.13050

Coe, B. P., Stessman, H. A. F., Sulovari, A., Geisheker, M. R., Bakken, T. E., Lake, A. M., Dougherty, J. D., Lein, E. S., Hormozdiari, F., Bernier, R. A., & Eichler, E. E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature Genetics, 51*(1), 106–116. https://doi.org/10.1038/s41588-018-0288-4

Corsello, C. M. (2005). Early Intervention in Autism. *Infants & Young Children*, *18*(2), 74–85. https://journals.lww.com/iycjournal/fulltext/2005/04000/early_intervention_in_autism.2.aspx. Accessed 17 November 2020

De Felice, A., Ricceri, L., Venerosi, A., Chiarotti, F., & Calamandrei, G. (2015). Multifactorial origin of neurodevelopmental disorders:

Approaches to understanding complex etiologies. *Toxics, 3*(1), 89–129. https://doi.org/10.3390/toxics3010089

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., ErcumentCicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Fu, S.-C., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., … Buxbaum, J. D. (2014). Synaptic, transcriptional, and chromatin genes disrupted in autism. *Nature, 515*(7526), 209–215. https://doi.org/10.1038/nature13772

Dincer, A., Gavin, D. P., Xu, K., Zhang, B., Dudley, J. T., Schadt, E. E., & Akbarian, S. (2015). Deciphering H3K4me3 broad domains associated with gene-regulatory networks and conserved epigenomic landscapes in the human brain. *Translational Psychiatry, 5*(11), e679–e679. https://doi.org/10.1038/tp.2015.169

Flint, J. (2001). Genetic basis of cognitive disability. *Dialogues in Clinical Neuroscience*, *3*(1), 37–46. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181642/. Accessed 12 November 2020

Franz, H., Villarreal, A., Heidrich, S., Videm, P., Kilpert, F., Mestres, I., Calegari, F., Backofen, R., Manke, T., & Vogel, T. (2019). DOT1L promotes progenitor proliferation and primes neuronal layer identity in the developing cerebral cortex. *Nucleic Acids Research, 47*(1), 168–183. https://doi.org/10.1093/nar/gky953

Freitag, C. M. (2007). The genetics of autistic disorders and its clinical relevance: A review of the literature. *Molecular Psychiatry, 12*(1), 2–22. https://doi.org/10.1038/sj.mp.4001896

Fuller, T. D., Westfall, T. A., Das, T., Dawson, D. V., & Slusarski, D. C. (2018). High-throughput behavioral assay to investigate seizure sensitivity in zebrafish implicates ZFHX3 in epilepsy. *Journal of Neurogenetics, 32*(2), 92–105. https://doi.org/10.1080/01677063.2018.1445247

Gejman, P., Sanders, A., & Duan, J. (2010). The role of genetics in the etiology of schizophrenia. *The Psychiatric Clinics of North America, 33*(1), 35–66. https://doi.org/10.1016/j.psc.2009.12.003

Hadders-Algra, M. (2021). Early diagnostics and early intervention in neurodevelopmental disorders—age-dependent challenges and opportunities. *Journal of Clinical Medicine, 10*(4), 861. https://doi.org/10.3390/jcm10040861

Hashimoto, R., Nakazawa, T., Tsurusaki, Y., Yasuda, Y., Nagayasu, K., Matsumura, K., Kawashima, H., Yamamori, H., Fujimoto, M., Ohi, K., Umeda-Yano, S., Fukunaga, M., Fujino, H., Kasai, A., Hayata-Takano, A., Shintani, N., Takeda, M., Matsumoto, N., … Hashimoto, H. (2016). Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *Journal of Human Genetics, 61*(3), 199–206. https://doi.org/10.1038/jhg.2015.141

Heyne, H. O., Singh, T., Stamberger, H., AbouJamra, R., Caglayan, H., Craiu, D., De Jonghe, P., Guerrini, R., Helbig, K. L., Koeleman, B. P. C., Kosmicki, J. A., Linnankivi, T., May, P., Muhle, H., Møller, R. S., Neubauer, B. A., Palotie, A., Pendziwiat, M., Striano, P., … Lemke, J. R. (2018). De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics, 50*(7), 1048–1053. https://doi.org/10.1038/s41588-018-0143-7

Huynh, L., & Hormozdiari, F. (2018). Combinatorial approach for complex disorder prediction: Case study of neurodevelopmental disorders. *Genetics, 210*(4), 1483–1495. https://doi.org/10.1534/genetics.118.301280

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., … Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature, 515*(7526), 216–221. https://doi.org/10.1038/nature13908

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h, Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., … Wigler, M. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron, 74*(2), 285–299. https://doi.org/10.1016/j.neuron.2012.04.009

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., Gallone, G., Lelieveld, S. H., Martin, H. C., McRae, J. F., Short, P. J., Torene, R. I., de Boer, E., Danecek, P., Gardner, E. J., Huang, N., Lord, J., Martincorena, I., Pfundt, R., … Retterer, K. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature, 586*(7831), 757–762. https://doi.org/10.1038/s41586-020-2832-5

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kaufman, L., Ayub, M., & Vincent, J. B. (2010). The genetic basis of non-syndromic intellectual disability: A review. *Journal of Neurodevelopmental Disorders, 2*(4), 182–209. https://doi.org/10.1007/s11689-010-9055-2

Kim, D., Basile, A. O., Bang, L., Horgusluoglu, E., Lee, S., Ritchie, M. D., Saykin, A. J., & Nho, K. (2017). Knowledge-driven binning approach for rare variant association analysis: Application to neuroimaging biomarkers in Alzheimer's disease. *BMC Medical Informatics and Decision Making, 17*(Suppl 1), 61. https://doi.org/10.1186/s12911-017-0454-0

Koso, H., Yi, H., Sheridan, P., Miyano, S., Ino, Y., Todo, T., & Watanabe, S. (2016). Identification of RNA-binding protein LARP4B as a tumor suppressor in glioma. *Cancer Research, 76*(8), 2254–2264.

Kwan, V., Unda, B. K., & Singh, K. K. (2016). Wnt signaling networks in autism spectrum disorder and intellectual disability. *Journal of Neurodevelopmental Disorders, 8*, 45. https://doi.org/10.1186/s11689-016-9176-3

Kyle Satterstrom, F., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Walters, R. K. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell, 180*(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Lieve, K. V. V., Verhagen, J. M. A., Wei, J., MartijnBos, J., van der Werf, C., RosésiNoguer, F., Mancini, G. M. S., Guo, W., Wang, R., van den Heuvel, F., Frohn-Mulder, I. M. E., Shimizu, W., Nogami, A., Horigome, H., Roberts, J. D., Leenhardt, A., Crijns, H. J. G., Blank, A. C., Aiba, T., … Wilde, A. A. M. (2019). Linking the heart and the brain: Neurodevelopmental disorders in patients with catecholaminergic polymorphic ventricular tachycardia. *Heart Rhythm, 16*(2), 220–228. https://doi.org/10.1016/j.hrthm.2018.08.025

McRae, J. F., Clayton, S., Fitzgerald, T. W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., Ambridge, K., Barrett, D. M., Bayzetinova, T., Jones, P., Jones, W. D., King, D., Krishnappa, N., Mason, L. E., Singh, T., … Hurles, M. E. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature, 542*(7642), 433–438. https://doi.org/10.1038/nature21062

Murcia Pienkowski, V., Kucharczyk, M., Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., Stawiński, P., Pollak, A., Kosińska, J., Wojciechowska, K., Lejman, M., Cieślikowska, A., Wicher, D., Stembalska, A., Matuszewska, K., Materna-Kiryluk, A., Gambin, A., Chrzanowska, K., Krajewska-Walasek, M., & Płoski, R. (2020). Breakpoint mapping of symptomatic balanced translocations links the EPHA6, KLF13 and UBR3 genes to novel disease

phenotype. *Journal of Clinical Medicine, 9*(5), 1245. https://doi.org/10.3390/jcm9051245

Niemi, M. E. K., Martin, H. C., Rice, D. L., Gallone, G., Gordon, S., Kelemen, M., McAloney, K., McRae, J., Radford, E. J., Yu, S., Gecz, J., Martin, N. G., Wright, C. F., Fitzpatrick, D. R., Firth, H. V., Hurles, M. E., & Barrett, J. C. (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature, 562*(7726), 268–271. https://doi.org/10.1038/s41586-018-0566-4

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., Smith, J. D., Turner, E. H., Stanaway, I. B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J. M., Borenstein, E., Rieder, M. J., … Eichler, E. E. (2012b). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature, 485*(7397), 246–250. https://doi.org/10.1038/nature10989

Omar, M. H., Campbell, M. K., Xiao, X., Zhong, Q., Brunken, W. J., Miner, J. H., Greer, C. A., & Koleske, A. J. (2017). CNS neurons deposit laminin α5 to stabilize synapses. *Cell Reports, 21*(5), 1281–1292. https://doi.org/10.1016/j.celrep.2017.10.028

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., O'Day, D. R., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., … Shendure, J. (2012a). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science (new York, NY), 338*(6114), 1619–1622. https://doi.org/10.1126/science.1227764

Parenti, I., Lehalle, D., Nava, C., Torti, E., Leitão, E., Person, R., Mizuguchi, T., Matsumoto, N., Kato, M., Nakamura, K., de Man, S. A., Cope, H., Shashi, V., Undiagnosed Diseases Network, Friedman, J., Joset, P., Steindl, K., Rauch, A., Muffels, I., … Mignot, C. (2021). Missense and truncating variants in CHD5 in a dominant neurodevelopmental disorder with intellectual disability, behavioral disturbances, and epilepsy. *Human Genetics, 140*(7), 1109–1120. https://doi.org/10.1007/s00439-021-02283-2

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genetics, 9*(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

Podvin, S., Jones, A., Liu, Q., Aulston, B., Ransom, L., Ames, J., Shen, G., Lietz, C. B., Jiang, Z., O'Donoghue, A. J., Winston, C., Ikezu, T., Rissman, R. A., Yuan, S., & Hook, V. (2020). Dysregulation of exosome cargo by mutant tau expressed in human-induced Pluripotent Stem Cell (iPSC) neurons revealed by proteomics analyses. *Molecular & Cellular Proteomics: MCP, 19*(6), 1017–1034. https://doi.org/10.1074/mcp.RA120.002079

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Jeremy Willsey, A., Gulhan Ercan-Sencicek, A., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., … State, M. W. (2012). De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature, 485*(7397), 237–241. https://doi.org/10.1038/nature10945

Santos-Laso, A., Izquierdo-Sanchez, L., Rodrigues, P. M., Huang, B. Q., Azkargorta, M., Lapitz, A., Munoz-Garrido, P., Arbelaiz, A., Caballero-Camino, F. J., Fernández-Barrena, M. G., Jimenez-Agüero, R., Argemi, J., Aragon, T., Elortza, F., Marzioni, M., Drenth, J. P. H., LaRusso, N. F., Bujanda, L., Perugorria, M. J., & Banales, J. M. (2020). Proteostasis disturbances and endoplasmic reticulum stress contribute to polycystic liver disease: new therapeutic targets. *Liver International: Official Journal of the International Association for the Study of the Liver, 40*(7), 1670–1685. https://doi.org/10.1111/liv.14485

Schmidt, M. F., Gan, Z. Y., Komander, D., & Dewson, G. (2021). Ubiquitin signalling in neurodegeneration: Mechanisms and therapeutic opportunities. *Cell Death & Differentiation, 28*(2), 570–590. https://doi.org/10.1038/s41418-020-00706-7

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., Wright, C. F., Firth, H. V., FitzPatrick, D. R., Barrett, J. C., & Hurles, M. E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature, 555*(7698), 611–616. https://doi.org/10.1038/nature25983

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., JamesKent, W., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., LePichon, J.-B., Miller, N. A., Thiffault, I., Dinwiddie, D. L., Twist, G., Noll, A., Heese, B. A., Zellmer, L., Atherton, A. M., Abdelmoity, A. T., Safina, N., Nyp, S. S., Zuccarelli, B., … Kingsmore, S. F. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science Translational Medicine, 6*(265), 265ra168. https://doi.org/10.1126/scitranslmed.3010076

Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., Firth, H. V., Frazier, T., Hansen, R. L., Prock, L., Brunner, H., Hoang, N., Scherer, S. W., Sahin, M., & Miller, D. T. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine, 21*(11), 2413–2421. https://doi.org/10.1038/s41436-019-0554-6

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., Xu, J., Batzoglou, S., Li, X., & Kai-How Farh, K. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics, 50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Sztainberg, Y., & Zoghbi, H. Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. *Nature Neuroscience, 19*(11), 1408–1417. https://doi.org/10.1038/nn.4420

Takahashi, M., Sugiyama, A., Wei, R., Kobayashi, S., Fukuda, K., Nishino, H., Takahashi, R., Tsutsumi, K., Kita, I., Ando, K., Manabe, T., Kamiguchi, H., Tomomura, M., & Hisanaga, S.-I. (2020). Hyperactive and impulsive behaviors of LMTK1 knockout mice. *Scientific Reports, 10*(1), 15461. https://doi.org/10.1038/s41598-020-72304-z

Tărlungeanu, D. C., & Novarino, G. (2018). Genomics in neurodevelopmental disorders: An avenue to personalized medicine. *Experimental & Molecular Medicine, 50*(8), 1–7. https://doi.org/10.1038/s12276-018-0129-7

Thonberg, H., Chiang, H.-H., Lilius, L., Forsell, C., Lindström, A.-K., Johansson, C., Björkström, J., Thordardottir, S., Sleegers, K., Broeckhove, C., Rönnbäck, A., & Graff, C. (2017). Identification and description of three families with familial Alzheimer disease that segregate variants in the SORL1 gene. *Acta Neuropathologica Communications, 5*, 43. https://doi.org/10.1186/s40478-017-0441-9

Tick, B., Bolton, P., Happé, F., Rutter, M., & Rijsdijk, F. (2016). Heritability of autism spectrum disorders: A meta-analysis of twin studies. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 57*(5), 585–595. https://doi.org/10.1111/jcpp.12499

Tischfield, D. J., Saraswat, D. K., Furash, A., Fowler, S. C., Fuccillo, M. V., & Anderson, S. A. (2017). Loss of the neurodevelopmental gene Zswim6 alters striatal morphology and motor regulation. *Neurobiology of Disease, 103*, 174–183. https://doi.org/10.1016/j.nbd.2017.04.013

Turner, T. N., & Eichler, E. E. (2019). The role of de novo noncoding regulatory mutations in neurodevelopmental disorders. *Trends in Neurosciences, 42*(2), 115–127. https://doi.org/10.1016/j.tins.2018.11.002

Turner, T. N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., Stessman, H. A. F., Doebley, A.-L., Bernier, R. A., Nickerson, D. A., & Eichler, E. E. (2017). denovo-db: A compendium of human de novo variants. *Nucleic Acids Research, 45*, D804–D811. https://doi.org/10.1093/nar/gkw865

Uddin, M., Tammimies, K., Pellecchia, G., Alipanahi, B., Hu, P., Wang, Z., Pinto, D., Lau, L., Nalpathamkalam, T., Marshall, C. R., Blencowe, B. J., Frey, B. J., Merico, D., Yuen, R. K. C., & Scherer, S. W. (2014). Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nature Genetics, 46*(7), 742–747. https://doi.org/10.1038/ng.2980

Wang, T., Kim, C., Bakken, T. E., Gillentine, M. A., Henning, B., Mao, Y., Gilissen, C., Nowakowski, T. J., & Eichler, E. E. (2021). Integrated gene analyses of de novo mutations from 46,612 trios with autism and developmental disorders. *bioRxiv*. https://doi.org/10.1101/2021.09.15.460398

Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P., & Eichler, E. E. (2017). Recurrent de novo mutations in neurodevelopmental disorders: Properties and clinical implications. *Genome Medicine, 9*(1), 101. https://doi.org/10.1186/s13073-017-0498-x