ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fgene.2021.744557

Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status

Justin L. Cope[1], *Hannes A. Baukmann*[1], *Jörn E. Klinger*[1], *Charles N. J. Ravarani*[1], *Erwin P. Böttinger*[2], *Stefan Konigorski*[2] and *Marco F. Schmidt*[1]*

[1]*biotx.ai GmbH, Potsdam, Germany,* [2]*Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of Potsdam, Potsdam, Germany*

Polygenic risk scores (PRS) aggregating results from genome-wide association studies are the state of the art in the prediction of susceptibility to complex traits or diseases, yet their predictive performance is limited for various reasons, not least of which is their failure to incorporate the effects of gene-gene interactions. Novel machine learning algorithms that use large amounts of data promise to find gene-gene interactions in order to build models with better predictive performance than PRS. Here, we present a data preprocessing step by using data-mining of contextual information to reduce the number of features, enabling machine learning algorithms to identify gene-gene interactions. We applied our approach to the Parkinson's Progression Markers Initiative (PPMI) dataset, an observational clinical study of 471 genotyped subjects (368 cases and 152 controls). With an AUC of 0.85 (95% CI = [0.72; 0.96]), the interaction-based prediction model outperforms the PRS (AUC of 0.58 (95% CI = [0.42; 0.81])). Furthermore, feature importance analysis of the model provided insights into the mechanism of Parkinson's disease. For instance, the model revealed an interaction of previously described drug target candidate genes *TMEM175* and *GAPDHP25*. These results demonstrate that interaction-based machine learning models can improve genetic prediction models and might provide an answer to the missing heritability problem.

Keywords: epistasis, machine learning, feature selection, parkinson's disease, PPMI (parkinson's progression markers initiative)

OPEN ACCESS

Edited by:
Jiangning Song,
Monash University, Australia

Reviewed by:
Jiancheng Zhong,
Hunan Normal University, China
Zhen Chen,
Qingdao University, China

****Correspondence:***
Marco F. Schmidt
ms@biotx.ai

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *20 July 2021*
Accepted: *27 September 2021*
Published: *20 October 2021*

Citation:
Cope JL, Baukmann HA, Klinger JE, Ravarani CNJ, Böttinger EP, Konigorski S and Schmidt MF (2021) Interaction-Based Feature Selection Algorithm Outperforms Polygenic Risk Score in Predicting Parkinson's Disease Status. Front. Genet. 12:744557. doi: 10.3389/fgene.2021.744557

INTRODUCTION

The need to understand how to predict phenotypes from genetic data is becoming ever-more important for the prediction of disease risk for individuals and for plant and animal breeding as well as for genome editing. Polygenic risk scores (PRS), simple additive models, are the state of the art in the investigation of the genetic architecture of complex traits or diseases, and, more importantly, in the prediction of disease susceptibility. (Wray et al., 2007; Evans et al., 2009; International Schizophrenia Consortium et al., 2009). A Polygenic Risk Score is calculated for a given individual as the weighted sum of the number of risk allele single nucleotide polymorphisms (SNP) for which the individual was tested. The weights used in this calculation are the regression coefficients from a prior genome-wide association study (GWAS).

Importantly, PRS models are not optimized for predictive performance. (Chatterjee et al., 2013; Dudbridge, 2013). There are three reasons for this:

(1) Due to the current limited sample size of discovery GWAS datasets (<1,000,000 individuals), biologically relevant rare variants with small effect sizes cannot be detected. Additionally, the limited sample sizes of discovery GWAS can lead to biased PRS models that might not perform well in populations with ancestry different to that of the discovery dataset. (Reisberg et al., 2017; Duncan et al., 2019).
(2) It has been shown that statistically significant outcome-associated SNPs are not automatically good predictors of that outcome. (Lo et al., 2015).
(3) It has been reported that genetic effects discovered in genome-wide association studies do not sum to the estimate of the heritability of the trait derived from twin studies. (Yang et al., 2010). This has been called the *missing heritability problem* in GWAS. (Manolio et al., 2009). Besides potentially missing relevant rare variants and suboptimal SNP selection based on *p*-values, classical PRS models ignore complex gene-gene interactions, also known as *epistasis*, of the trait or disease due to their simple additive structure.

The concept of epistasis was first described more than 100 years ago. (Bateson, 1906). Statistical epistasis, as observed in genome-wide association studies, is genetic variance that can be attributed to gene interaction and is defined as a function of the allele frequencies in a population. Detection of epistasis in discovery GWAS and modeling its impact is challenging because of linkage disequilibrium (LD), replication of identified gene-gene interactions in validation datasets, model complexity, and high dimensionality. (Wei et al., 2014).

Machine learning algorithms that improve automatically through the use of data represent an opportunity to find gene-gene interactions in order to build models with better predictive performance than PRS. Nevertheless, in a recent study, a PRS model outperformed five machine learning algorithms (Naïve Bayes classifier, regularized regression, random forest, gradient boosting, and support vector machine) that were used to build predictive models for coronary artery disease status. (Gola et al., 2020).

Here we revisit the potential of machine learning algorithms to predict disease status compared to a PRS model. For this purpose, we adopt the Parkinson's Progression Markers Initiative (PPMI) dataset (Marek et al., 2011, 2018) (https://www.ppmi-info.org) as this dataset has been intensively analyzed and is broadly available for replication studies. We explore two machine learning approaches in particular, which complement those applied by Gola et al.: deep learning and interaction-based feature selection. The first approach, deep learning, employs artificial neural networks to discover automatically from raw data the representations needed for classification. Despite not being widely used in the field of genomics, there is work on applying deep learning to GWAS: Romero et al., 2016 predict genetic ancestry by introducing a multi-task architecture including a parameter prediction network, thereby considerably reducing the feature space under consideration. The second approach, interaction-based feature selection, also drastically reduces the feature space–in this case, by leveraging contextual information obtained via data mining, allowing for the testing of a small set of complex hypotheses containing interactions of multiple variants. Further details concerning these approaches are described in the Methods section, following a presentation of the results of our investigation below.

RESULTS

Data Preparation

For all 471 subjects in the PPMI database (368 cases and 152 controls) subject genotyping information was collected from two complementary genotyping chips (NeuroX and ImmunoChip). After careful quality control and harmonization, we merged that information into a single dataset with 369,036 variants and 436 individuals (296 cases and 140 controls). The data was then split into three disjoint sets: a training set (n = 367) for training predictive models; a validation set (n = 33) for so-called *hyperparameter tuning*, and a test set (n = 36) for model evaluation. Training and validation are described in further detail below for each approach as appropriate. In all cases, evaluation metrics were calculated on the basis of bootstrap resampling with 10^4 iterations.

Genome-Wide Association Study

A genome-wide association (GWA) analysis was performed on the training data. The Manhattan plot of the *p*-values resulting from the analysis is shown in **Figure 1**. Seven single nucleotide polymorphisms (SNPs) showed *p*-values less than 10^{-4} (**Table 1**).

Polygenic Risk Score

To calculate the PRS, seven different *p*-value thresholds (0.001, 0.05, 0.1, 0.2, 0.4, and 0.5) for the subjects in the training, validation and test set were used. The PRS of the subjects in the training set were then used to train a separate logistic regression classifier for each *p*-value threshold. Receiver operating characteristics (ROC) curves were used to evaluate classifier performance relative to the validation data. The classifier with the highest mean area under the curve (AUC) was that which had been trained on the PRS resulting from the 0.05 *p*-value threshold, comprising the weighted sum of 57 different SNPs. This classifier was finally evaluated relative to the test data set, where the mean AUC was 0.58 with a 95% confidence interval from 0.42 to 0.81 **Figure 3**. **Table 2** presents these results, along with the estimates of accuracy, sensitivity, and specificity corresponding to the optimal Youden's index of 0.21.

Deep Learning

We applied Romero *et al.*'s approach to the PPMI dataset, again using the training data set to train competing networks with distinct hyperparameter settings and the validation data set to select between these networks. When evaluated relative to the test data set, the mean AUC of the final deep learning model was 0.67 (95% CI = [0.47; 0.83]) and the optimal Youden index

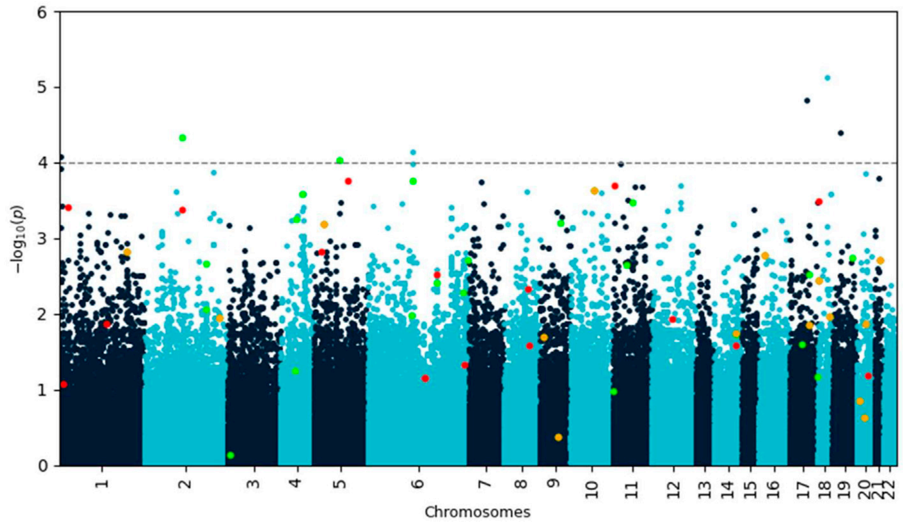

FIGURE 1 | Manhattan plot of negative decadic logarithm of *p*-values for SNPs as determined by SAIGE analysis. Variants identified by Lasso with feature selection are highlighted in red and green if they increase or decrease disease risk, respectively. Variants highlighted in orange occur in both protective and risk-enhancing groups of SNPs, depending on their genotype. Most of these biologically meaningful variants would have been missed by using a simple *p*-value cutoff.

TABLE 1 | PPMI GWAS results identified seven SNPs with a *p*-value < 10^{-4}. Positions and rs IDs according to Human Genome Reference hg19 (GRCh37).

Chr	Pos	SNP Id	rs Id	Gene	*p*-value
1	173,266,578	imm_1_171,533,201	rs4916319	*TNFSF4* (upstream)	0.000083
2	209,087,335	exm2261159	rs4675743		0.000046
5	156,376,703	exm498917	rs6873053	*TIMD4* (downstream)	0.000092
6	133,716,974	rs212805	rs212805	*EYA4*	0.000074
17	25,895,033	imm_17_22,919,160	rs4795747		0.000015
18	5,479,093	rs7238186	rs7238186	*EPB41L3* (downstream)	0.000007
19	57,909,872	exm1513284	rs4801478	*ZNF548*	0.000040

TABLE 2 | Performance comparison of all models.

Method	AUC [95% CI]	Accuracy	Sensitivity	Specificity	Youden's index
PRS	0.56 [0.42; 0.81]	0.60	0.62	0.56	0.21
Deep learning	0.67 [0.47; 0.83]	0.60	0.42	0.88	0.29
LASSO w/feature selection	0.85 [0.72; 0.96]	0.81	0.81	0.80	0.61
LASSO w/o feature selection	0.51 [0.39; 0.63]	0.62	0.87	0.09	0.12

corresponding to the accuracy, sensitivity, and specificity measures reported in **Table 2** was 0.29.

Feature Selection and LASSO Regression

A set of less than 100 polygenic hypotheses were generated using the interaction-based feature selection approach applied to the training data, as described in the Methods section below. (See also an overview of our approach in **Figure 3**.). These hypotheses were summarized in a term that was used to build a LASSO regression model on the basis of the validation data. (Tibshirani, 1996). The predictive performance of this model, based on 47 SNPs in several different interaction terms, was then evaluated relative to the test set **Figure 4**. The mean area under the curve (AUC) for the LASSO model with prior feature selection was 0.85 [95% CI = (0.72; 0.96)] and the optimal Youden index corresponding to the accuracy, sensitivity, and specificity measures reported in **Table 2** was 0.61. A LASSO model without prior feature selection that was built for comparison was evaluated in the same manner but did not deliver outcomes that were significantly better than chance (**Table 2**), in line with Gola et al. (2020) s results for regularized regression.

Exploring the feature selection based model with its interactive terms provides insights about the genes associated with Parkinson's disease. An annotation of all 47 SNPs in our

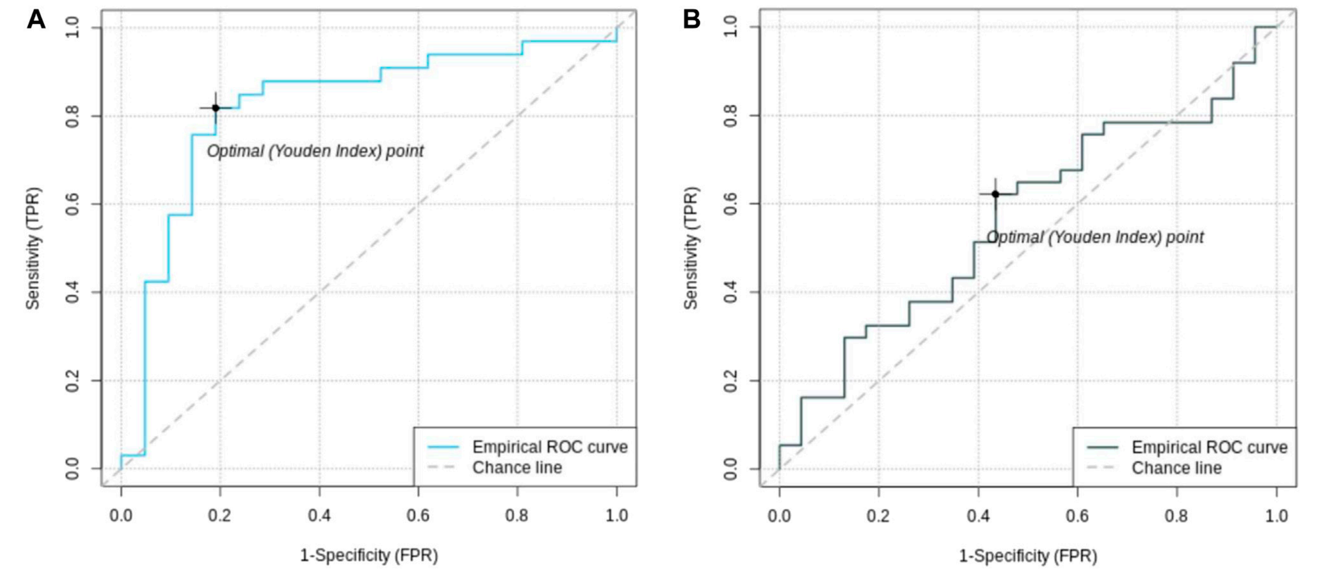

FIGURE 2 | Receiver operating characteristic (ROC) curves of feature selected machine learning model **(A)** and polygenic risk score **(B)**. The AUC of the feature selected model with 0.85 [95% CI = (0.72; 0.96)] is better than the AUC of the PRS with 0.56 [95% CI = (0.42; 0.81)].

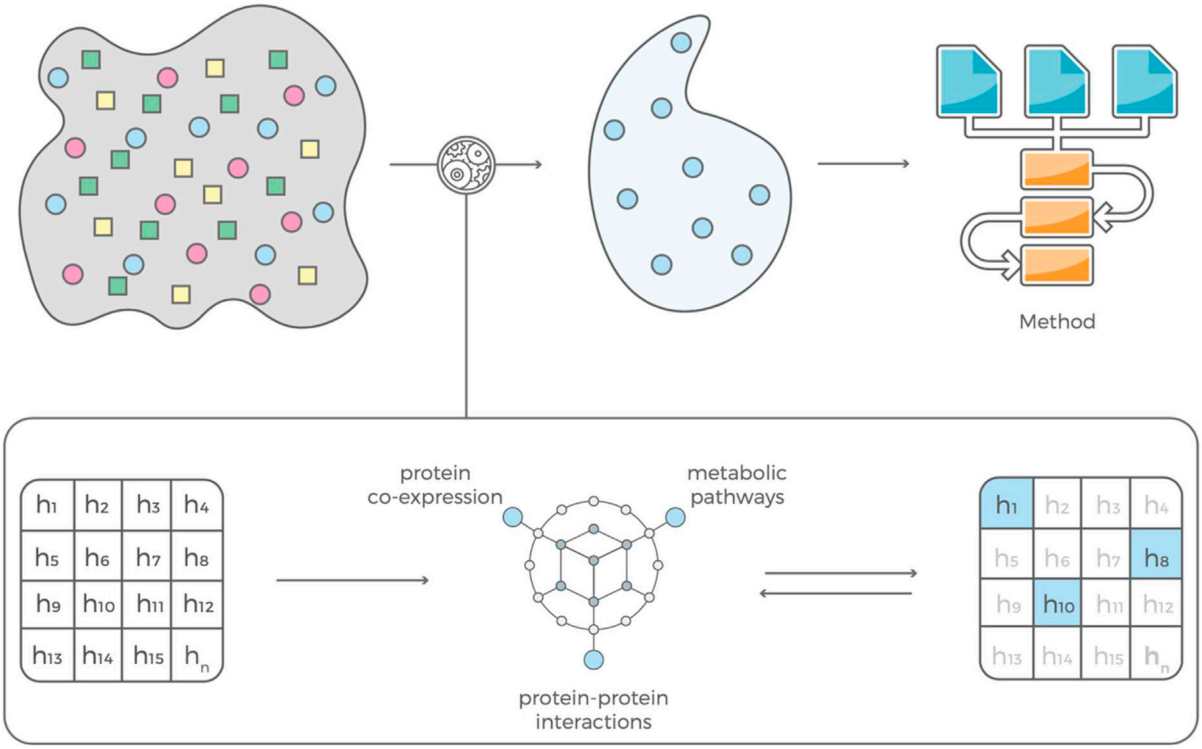

FIGURE 3 | Our feature selection consists of two complementary modules that are in feedback with each other. The contextual module uses information mined from the scientific literature, pathway libraries and protein co-expression data and an evaluation module that estimates predictive power of a feature based on that contextual information. The selected features can be used to build prediction models with standard machine learning algorithms.

model can be found in the Supplementary Information. An exciting result from this analysis of the PPMI dataset is the statistical interaction of variants rs3822019 on chromosome four in gene *TMEM175*, coding for a potassium channel in late endosomes, and rs17022,452 on chromosome 2, close to the coding region of *GAPDHP25*, glyceraldehyde-3 phosphate

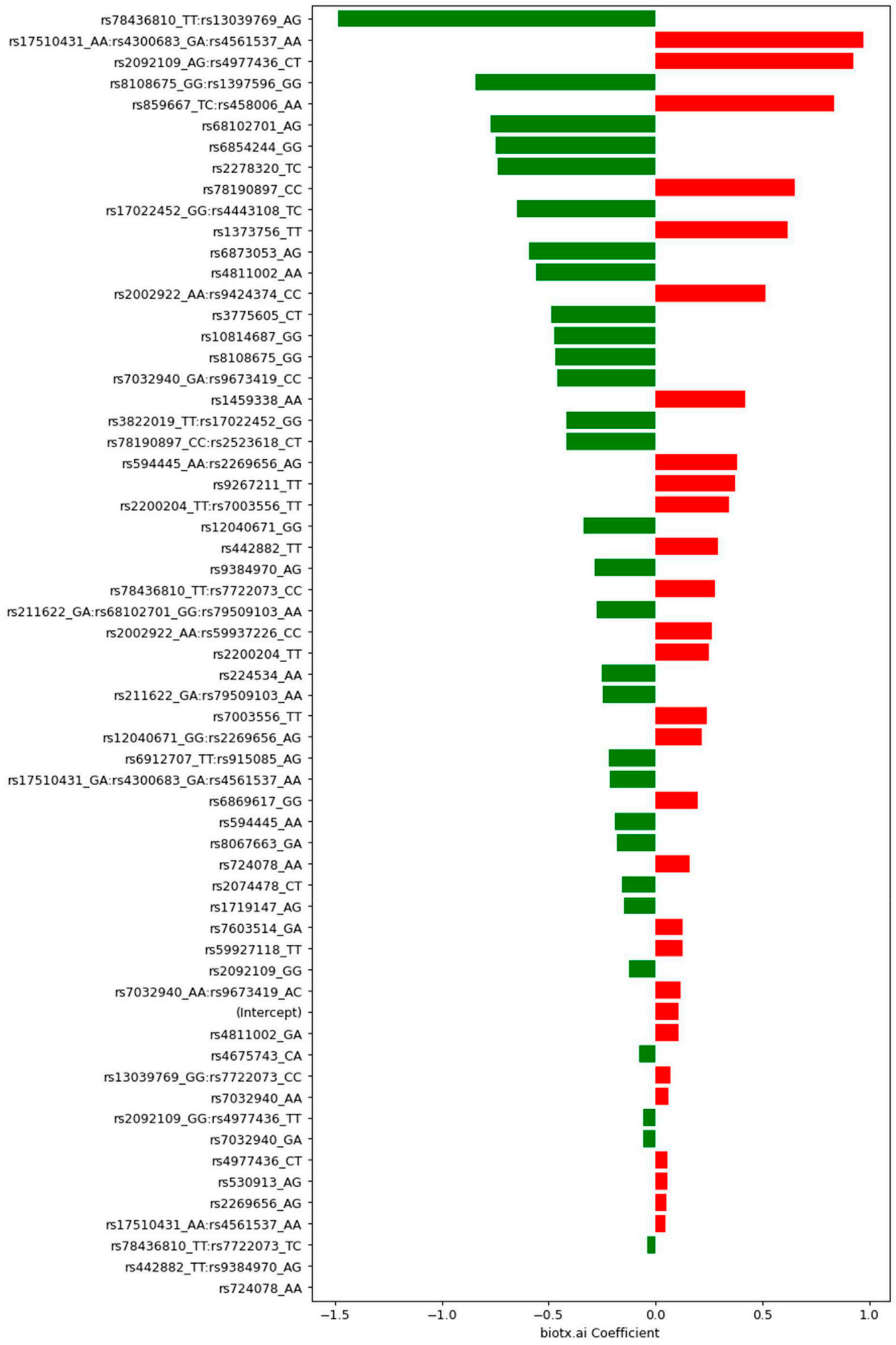

FIGURE 4 | Coefficients determined by Lasso with feature selection for SNPs and groups of SNPs. Negative values (green) indicate protective (combinations of) variants, positive values (red) mark risk variants. The respective genotypes of each variant are indicated by one-letter codes of the bases, where the first letter corresponds to the reference allele, and the second corresponds to the observed, alternative allele.

TABLE 3 | PD cases and controls among bearers of the respective genotype combinations of the identified variants rs3822019 and rs17022,452.

Genotype combination	Cases	Controls
rs3822019_TT/rs17022,452_GG	0	0
rs3822019_TT/rs17022,452_GA	6/100%	0
rs3822019_TC/rs17022,452_GG	2/50%	2/50%
rs3822019_TT/-	7/100%	0
-/rs17022,452_GG	7/87.5%	1/12.5%
rs3822019_TC/rs17022,452_GA	27/87.1%	4/12.9%
rs3822019_TC/-	68/73.9%	24/26.1%
-/rs17022,452_GA	66/75%	22/25%
-/-	113/56.5%	87/43.5%

dehydrogenase pseudogene 25. rs3822019 is an intron variant that has been linked to Parkinson's disease. (Nalls et al., 2014).

DISCUSSION

We analyzed the PPMI dataset and built predictive models using polygenic risk scores, a deep learning algorithm for genomic data (Romero et al., 2016), and LASSO regression with and without interaction-based feature selection to reduce the hypothesis space. The PRS model comprises 57 SNPs and showed an AUC of 0.58 whereas the deep learning model had an AUC of 0.67. Notably, the deep learning model consists of abstract embeddings instead of single SNPs like the PRS. Therefore, identification of disease-associated SNPs and further insights into the disease mechanism are not possible here. The LASSO regression model built on interactions containing only 47 SNPs that were discovered via the use of contextual information outperformed the other predictive models with an AUC of 0.85. Beyond that, the approach was able to associate new variants with the disease that would not have shown up under an additive approach such as PRS.

We investigated how the combinations of the relevant genotypes rs3822019_TT (*TMEM175*) and rs17022,452_GG (*GAPDHP25*) split the individuals into cases and controls (**Table 3**). All subjects that are homozygous for rs3822019_TT are affected by PD. Furthermore, most individuals heterozygous for this variant (rs3822019_TT) or homozygous for rs17022,452_GG are cases (76.4 and 75.0%, respectively). These results support the relevance of the association between these variants and PD status.

The *TMEM175/GAK/DGKQ* locus was the third strongest risk locus in a GWA study of Parkinson's disease (Krohn et al., 2020) and has been described as a potential drug target. (Diogo et al., 2018; Jinn et al., 2019). Deficiency in the potassium channel TMEM175 results in unstable lysosomal pH, which leads to decreased lysosomal catalytic activity and increased α-synuclein aggregation, among other effects. As a potassium channel, TMEM175 has a high potential as a druggable target and a tractable therapeutic strategy has been proposed. (Jinn et al., 2017).

GAPDH has been targeted with the investigational drug Omigapil for prevention of PD, ALS, congenital muscular dystrophy and myopathy. The drug has been shown to protect against behavioural abnormalities and neuro-degeneration in animal models of Parkinson's disease. However, PD development has been terminated due to lack of benefit. (Olanow et al., 2006).

There seem to be various causes of Parkinson's disease, yet the pathogenesis of this disease appears to be converging on common themes—oxidative stress, mitochondrial dysfunction, and protein aggregation—all of which are tightly linked to autophagy. (Lynch-Day et al., 2012). Both TMEM175 (Jinn et al., 2019) and GAPDH (Butera et al., 2019) regulate autophagy. Disturbed expression of autophagy genes in blood of PD patients. (Lynch-Day et al., 2012).

To summarize, we here present an approach to apply machine learning algorithms to high-dimensional genomic data using a contextual knowledge based feature selection. PRS models require a large set of SNPs, which leads to overfitting and limits their use in clinical practice. We generated more parsimonious models overcoming these limitations–with only 47, partly interacting SNPs, our model was able to outperform a PRS model based on 57 SNPs for Parkinson's disease. Analysis of feature importance of our model identified a gene-gene interaction of *TMEM175* and *GAPDHP25*. TMEM175 has been described as a potential drug target and further information on its mechanism of action could be invaluable. A recently discovered interaction with pseudogene *GAPDHP25* could provide helpful insights. In conclusion, applying machine learning algorithms to feature-selected genomic data led to an interaction-based model with better predictive performance than PRS and has paved the way for the generation of new insights into disease mechanisms.

METHODS

Parkinson's Progression Marker Initiative Dataset

The Parkinson's Progression Marker Initiative (PPMI) dataset (https://www.ppmi-info.org) contains 471 subjects (368 cases and 152 controls), and for each subject, genotyping information collected from two complementary chips (NeuroX and ImmunoChip) is available. (Marek et al., 2011). After careful quality control and harmonization (e.g., genome build conversion, strand alignment) as described in the literature (Marees et al., 2018), we merged that information into a single dataset with 380,939 variants in total.

After this initial data harmonization, an additional set of quality control steps were performed on variants and individuals that aimed to remove biases that could affect the downstream analysis. First, SNPs and individuals were filtered based on their missingness in the dataset. This ensured the exclusion of SNPs that had a high proportion of subjects where genotyping information was unavailable or of poor quality. Similarly, individuals where a large proportion of SNPs could not be measured were excluded. This step was achieved by setting the missing call rate threshold to 0.02 (i.e., >2%); as a result, 6,084 variants and 22 people were removed. SNP filtering was performed before individual filtering.

With high missing call rates filtered, all variants not on autosomal chromosomes were removed (5,731 variants). This was followed by the identification and removal of variants deviating from Hardy-Weinberg equilibrium, which can indicate genotyping errors. These variants were identified in a two-stage process whereby we first applied a threshold of 1e-6 exclusively to controls, followed by a threshold of 1e-10 applied to all samples, leading to the removal of 0 and 202 variants, respectively.

Next, individuals were filtered based on their heterozygosity rates, which can indicate sample contamination. Individuals deviating by more than 3 standard deviations from the mean of the rate of all samples (13 individuals) were removed. To assess the heterozygosity rate per sample, variants in linkage disequilibrium were first extracted, scanning the genome at a window size of 50 variants, a step size of 5, and a pairwise correlation threshold of 0.2.

Finally, relatedness between individuals was ascertained through the calculation and assessment of their respective identity by descent coefficients (IBD). Only one individual in a related pair would be kept, although in this case, no related individuals were identified and so none were removed.

The final quality-controlled dataset contained 369,036 variants and 436 individuals passing the various filters.

GENOME-WIDE ASSOCIATION STUDY

As a preliminary step, a genome-wide association (GWA) analysis was performed with the R package SAIGE (Zhou et al., 2018) to test individual variants for their association with Parkinson's disease.

Polygenic Risk Score

The PRS was constructed by using PLINK (Purcell et al., 2007) following the guidelines provided by Choi et al. (Choi et al., 2020) and the accompanying tutorial (https://choishingwan.github.io/PRS-Tutorial/plink/). The clumping cut-off of r^2 was 0.1. For all subjects in the training, validation and test sets, seven distinct risk scores were calculated, corresponding to seven potential p-value thresholds (0.001, 0.05, 0.1, 0.2, 0.4, 0. 5). The seven risk scores for the subjects in the training set were then used to train seven separate logistic regression classifiers (one for each p-value threshold) using the *glm* function in R (www.R-project.org). These classifiers were evaluated relative to the validation data set, leading to the selection of the classifier based on the PRS calculated using the p-value threshold of 0.05. The predictions of this final classifier were then evaluated relative to the test set.

Deep Learning

The deep learning prediction model was built using a Diet Network according to the procedure described by Romero et al. (Romero et al., 2016) The model is composed of three networks: one basic and two auxiliary networks. After a basic discriminative network with optional reconstruction path, follows a network that predicts the input fat layer parameters, and finally, a network that predicts the reconstruction fat layer parameters. The official code can be found here: https://github.com/adri-romsor/DietNetworks.

Feature Selection

The interaction-based feature selection approach that we adopt organizes data mined from journal articles, pathway libraries, protein co-expression libraries, and drug candidate libraries (e.g., dbSNP, ClinVar, OMIM, Reactome, STRING database) into a hierarchical knowledge graph, which generates disease-specific hypotheses based on interactions of genetic variants (**Figure 1**). Each interaction's predictive power is determined using the training data set and the *glm* function in R (www.R-project.org). If an interaction predicts disease status well, the graph is incentivized to 'fine-tune' the hypothesis by comparing a set of very similar hypotheses. If a hypothesis has little or no predictive power, the graph is not incentivized to explore it or similar hypotheses further and will instead propose hypotheses containing different variants. (Klinger et al., 2021). This learning process is driven by gradient descent, meaning that it converges when the average performance of the new multi-variant hypothesis does not increase. After convergence, the selected features are used to build prediction models with standard machine learning algorithms, such as LASSO regression (Friedman et al., 2010).

LASSO Regression

LASSO (least absolute shrinkage and selection operator) regression models were computed by using the glmnet package (https://glmnet.stanford.edu/index.html) for R (www.R-project.org) and its function cv.glmnet with five-fold cross-validation in order to avoid overfitting. (Friedman et al., 2010).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ppmi-info.org/.

AUTHOR CONTRIBUTIONS

JK, CR, EB, SK, and MS contributed to conception and design of the study. JC, HB, JK, and CR organized the database and performed the statistical analysis. MS wrote the first draft of the manuscript. JC, HB, JK, CR, and MS wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The research work was supported by the Investitionsbank des Landes Brandenburg (ILB), the European Regional Development Fund (ERDF), and the European Social Fund+ (ESF+).

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI–a public-private partnership–is funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners. List of full names of all the PPMI funding partners can be found at www.ppmi-info.org/fundingpartners. We also thank the program digital solutions made in Brandenburg (digisolBB) for its continued support.

REFERENCES

Bateson, W. (1906). The Progress of Genetics since the Rediscovery of Mendel's Papers. *Prog. Rei Bot.* 1, 368.

Butera, G., Mullappilly, N., Masetto, F., Palmieri, M., Scupoli, M. T., Pacchiana, R., et al. (2019). Regulation of Autophagy by Nuclear GAPDH and its Aggregates in Cancer and Neurodegenerative Disorders. *Int. J. Mol. Sci.* 20, 2062. doi:10.3390/ijms20092062

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J. H. (2013). Projecting the Performance of Risk Prediction Based on Polygenic Analyses of Genome-Wide Association Studies. *Nat. Genet.* 45 (4), 400–405. doi:10.1038/ng.2579

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: A Guide to Performing Polygenic Risk Score Analyses. *Nat. Protoc.* 15, 2759–2772. doi:10.1038/s41596-020-0353-1

Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C. A., et al. (2018). Phenome-Wide Association Studies across Large Population Cohorts Support Drug Target Validation. *Nat. Commun.* 9, 4285. doi:10.1038/s41467-018-06540-3

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *Plos Genet.* 9, e1003348. doi:10.1371/journal.pgen.1003348

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., et al. (2019). Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations. *Nat. Commun.* 10, 3328. doi:10.1038/s41467-019-11112-0

Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the Information Contained within Genome-wide Association Studies to Improve Individual Prediction of Complex Disease Risk. *Hum. Mol. Genet.* 18, 3525–3531. doi:10.1093/hmg/ddp295

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I. R. (2020). Polygenic Risk Scores Outperform Machine Learning Methods in Predicting Coronary Artery Disease Status. *Genet. Epidemiol.* 44, 125–138. doi:10.1002/gepi.22279

International Schizophrenia ConsortiumPurcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., et al. (2009). Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature* 460, 748–752. doi:10.1038/nature08185

Jinn, S., Drolet, R. E., Cramer, P. E., Wong, A. H.-K., Toolan, D. M., Gretzula, C. A., et al. (2017). TMEM175 Deficiency Impairs Lysosomal and Mitochondrial Function and Increases α-synuclein Aggregation. *Proc. Natl. Acad. Sci. USA* 114, 2389–2394. doi:10.1073/pnas.1616332114

Jinn, S., Blauwendraat, C., Toolan, D., Gretzula, C. A., Drolet, R. E., Smith, S., et al. (2019). Functionalization of the TMEM175 p.M393T Variant as a Risk Factor for Parkinson Disease. *Hum. Mol. Genet.* 28, 3244–3254. doi:10.1093/hmg/ddz136

Klinger, J., Ravarani, C., Bannard, C., Lamparter, M., Schwinges, A., Cope, J., et al. (2021). *Critically Ill COVID-19 Status Associated Trait Genetics Reveals CDK6 Inhibitors as Potential Treatment*. doi:10.21203/rs.3.rs-568366/v1 Available at https://www.medrxiv.org/content/10.1101/2021.05.18.21256584v2

Krohn, L., Öztürk, T. N., Vanderperre, B., Ouled Amar Bencheikh, B., Ruskey, J. A., Laurent, S. B., et al. (2020). Genetic, Structural, and Functional Evidence Link TMEM175 to Synucleinopathies. *Ann. Neurol.* 87, 139–153. doi:10.1002/ana.25629

Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why Significant Variables Aren't Automatically Good Predictors. *Proc. Natl. Acad. Sci. USA* 112, 13892–13897. doi:10.1073/pnas.1518285112

Lynch-Day, M. A., Mao, K., Wang, K., Zhao, M., and Klionsky, D. J. (2012). The Role of Autophagy in Parkinson's Disease. *Cold Spring Harbor Perspect. Med.* 2, a009357. doi:10.1101/cshperspect.a009357

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the Missing Heritability of Complex Diseases. *Nature* 461, 747–753. doi:10.1038/nature08494

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et al. (2018). A Tutorial on Conducting Genome-wide Association Studies: Quality Control and Statistical Analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608. doi:10.1002/mpr.1608

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* 95, 629–635. doi:10.1016/j.pneurobio.2011.09.005

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., et al. (2018). The Parkinson's Progression Markers Initiative (PPMI) - Establishing a PD Biomarker Cohort. *Ann. Clin. Transl Neurol.* 5, 1460–1477. doi:10.1002/acn3.644

Nalls, M. A., Pankratz, N., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., et al. (2014). Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease. *Nat. Genet.* 46, 989–993. doi:10.1038/ng.3043

Olanow, C. W., Schapira, A. H., LeWitt, P. A., Kieburtz, K., Sauer, D., Olivieri, G., et al. (2006). TCH346 as a Neuroprotective Drug in Parkinson's Disease: A Double-Blind, Randomised, Controlled Trial. *Lancet Neurol.* 5, 1013–1020. doi:10.1016/s1474-4422(06)70602-0

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., and Vilo, J. (2017). Comparing Distributions of Polygenic Risk Scores of Type 2 Diabetes and Coronary Heart Disease within Different Populations. *PLoS One* 12, e0179238. doi:10.1371/journal.pone.0179238

Romero, A., Carrier, P. L., Erraqabi, A., Sylvain, T., Auvolat, A., Dejoie, E., et al. (2016). Diet Networks: Thin Parameters for Fat Genomics. arXiv [cs.LG]. Available at: http://arxiv.org/abs/1611.09340.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting Epistasis in Human Complex Traits. *Nat. Rev. Genet.* 15, 722–733. doi:10.1038/nrg3747

Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of Individual Genetic Risk to Disease from Genome-wide Association Studies. *Genome Res.* 17, 1520–1528. doi:10.1101/gr.6665407

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs Explain a Large Proportion of the Heritability for Human Height. *Nat. Genet.* 42, 565–569. doi:10.1038/ng.608

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently Controlling for Case-Control Imbalance and Sample

Relatedness in Large-Scale Genetic Association Studies. *Nat. Genet.* 50, 1335–1341. doi:10.1038/s41588-018-0184-y

Conflicts of Interest: JC, HB, JK, CR, and MS are employed by biotx.ai GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.