

1 **Interaction-based feature selection algorithm outperforms**  
2 **polygenic risk score in predicting Parkinson's Disease**  
3 **status**

4

5 Jörn E. Klinger<sup>1</sup>, Charles N. J. Ravarani<sup>1</sup>, Hannes A. Baukmann<sup>1</sup>, Justin L. Cope<sup>1</sup>,  
6 Erwin P. Böttinger<sup>2</sup>, Stefan Konigorski<sup>2</sup>, Marco F. Schmidt<sup>1\*</sup>

7

8 <sup>1</sup>biotx.ai GmbH, Am Mühlenberg 11, 14476 Potsdam, Germany

9 <sup>2</sup>Digital Health Center, Hasso Plattner Institute for Digital Engineering, University of  
10 Potsdam, Rudolf-Breitscheid-Straße 187, 14482 Potsdam, Germany

11

12 \*Corresponding author

13 E-mail: [ms@biotx.ai](mailto:ms@biotx.ai)

## 14 **Abstract**

15 Polygenic risk scores (PRS) aggregating results from genome-wide association  
16 studies are state of the art to predict the susceptibility to complex traits or diseases.  
17 Novel machine learning algorithms that use large amounts of data promise to find  
18 gene-gene interactions in order to build models with better predictive performance than  
19 PRS. Here, we present a data preprocessing step by using data-mining of contextual  
20 information to reduce the number of features, enabling machine learning algorithms to  
21 identify gene-gene interactions. We applied our approach to the Parkinson's  
22 Progression Markers Initiative (PPMI) dataset, an observational clinical study of 471  
23 genotyped subjects (368 cases and 152 controls). With an AUC of 0.85 (95% CI =  
24 [0.72; 0.96]), the interaction-based prediction model outperforms the PRS (AUC of 0.58  
25 (95% CI = [0.42; 0.81])). Furthermore, feature importance analysis of the model  
26 provided insights into the mechanism of Parkinson's Disease. For instance, the model  
27 revealed an interaction of previously described drug target candidate genes *TMEM175*  
28 and *GAPDHP25*. These results demonstrate that interaction-based machine learning  
29 models can improve genetic prediction models and might provide an answer to the  
30 missing heritability problem.

31

32

## 33 **Introduction**

34 The need to understand how to predict phenotypes from genetic data becomes ever-  
35 more important for individual's disease risk prediction, animal and plant breeding as  
36 well as genome editing. Polygenic risk scores (PRS), simple additive models, are state  
37 of the art to investigate the genetic architecture and, more importantly, to predict the  
38 susceptibility of complex traits or diseases. (Wray et al., 2007; Evans et al., 2009;  
39 International Schizophrenia Consortium et al., 2009) For each individual a score is  
40 calculated as a weighted sum of the number of risk allele single nucleotide  
41 polymorphisms (SNP) an individual was tested for. The used weights are regression  
42 coefficients from previous genome-wide association studies (GWAS).

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

45 **1.** Due to the current limited sample size of discovery GWAS datasets (< 1,000,000  
46 individuals), biologically relevant rare variants with small effect sizes cannot be  
47 detected. Additionally, the limited sample sizes of discovery GWAS can lead to biased  
48 PRS models that might not perform well in populations with ancestry different to that  
49 of the discovery dataset. (Reisberg et al., 2017; Duncan et al., 2019)

50 **2.** It has been shown that statistically associated SNPs are not automatically good  
51 predictors. (Lo et al., 2015)

52 **3.** It has been reported that genetic effects discovered in genome-wide association  
53 studies do not sum to the estimate of the heritability of the trait compared to twin  
54 studies. (Yang et al., 2010) This has been called the missing heritability problem in  
55 GWAS. (Manolio et al., 2009) Beside potentially missing relevant rare variants and  
56 suboptimal SNP selection based on  $p$  values, classical PRS models ignore complex  
57 gene-gene interactions, also known as epistasis, of the trait or disease due to their  
58 simple additive structure.

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

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

73 Here, we explore the potential of using contextual information obtained via data mining  
74 to strongly reduce the hypothesis space, which, in turn, allows for testing a small set  
75 of complex hypotheses, containing interaction of multiple variants. This approach  
76 organizes data mined from journal articles, pathway libraries, protein co-expression  
77 libraries and drug candidate libraries into a hierarchical graph, which generates  
78 disease-specific hypotheses based on interactions of genetic variants. Each  
79 interaction's predictive power is determined using the training data set. If an interaction  
80 predicts disease status well, the graph is incentivized to 'fine-tune' the hypothesis by  
81 comparing a set of very similar hypotheses. If a hypothesis has little or no predictive  
82 power, the graph is not incentivized to explore it, or similar hypotheses further and will  
83 instead propose hypotheses containing different variants. This learning process is  
84 driven by gradient descent, meaning that it converges when the average performance

85 of the new multi-variant hypothesis does not increase. After convergence, the selected  
86 features are used to build prediction models with standard machine learning  
87 algorithms, such as LASSO (least absolute shrinkage and selection operator)  
88 regression (Friedman et al., 2010). An overview of our approach is given in Figure 1.  
89 The Parkinson's Progression Markers Initiative (PPMI) dataset (Marek et al., 2011,  
90 2018) (<https://www.ppmi-info.org>) was selected for the comparison as the dataset has  
91 been intensively analyzed and, from there, a broad audience is able to reproduce our  
92 results.

93

94

## 95 **Results**

### 96 **GWAS**

97 In a preliminary proof-of-concept step, a genome-wide association (GWA) analysis  
98 was performed. For all 471 subjects in the PPMI database, 368 cases and 152 controls,  
99 subject genotyping information was collected from two complementary genotyping  
100 chips (NeuroX and ImmunoChip). After careful quality control and harmonization, we  
101 merged that information into a single dataset with 369,036 variants and 436 individuals.  
102 The Manhattan plot of the  $p$  values resulting from SAIGE analysis is shown in Fig. 2.  
103 Seven single nucleotide polymorphisms (SNPs) showed smaller  $p$  values than  $10^{-4}$   
104 (Table 1).

105

### 106 **Polygenic risk score**

107 For all analyses described in the following, the data was split into training, validation  
108 and test sets. The same sets were used for constructing a polygenic risk score (PRS)  
109 and machine learning prediction models with and without feature selection. To  
110 calculate the PRS, 7 different  $p$  value thresholds (0.001, 0.05, 0.1, 0.2, 0.4, 0.5) for the

111 subjects in the training, validation and test set were used. The PRS of the subjects in  
112 the training set were then used to train a separate logistic regression classifier for each  
113  $p$  value threshold. The validation data set was used to determine which of these  
114 thresholds produces the best classifier, which was then used to predict the test set.  
115 This classifier is based on the PRS of 57 different SNPs.  
116 The receiver operating characteristics (ROC) curve was used to evaluate the predictive  
117 power of the PRS. The area under the curve (AUC) was 0.58 (95% CI = [0.42; 0.81])  
118 and the Youden's index was 0.21 (Tab. 2).

119

## 120 **Deep learning**

121 Deep learning is a machine learning technique based on artificial neural networks with  
122 representation learning that allows a system to automatically discover the  
123 representations needed for feature detection or classification from raw data. Despite  
124 not being widely used in the field of genomics, there is work on applying Deep Learning  
125 to GWAS, such as Romero *et al.* (Romero *et al.*, 2016). Romero *et al.* use a Diet  
126 Network, a neural network parameterization, which considerably reduces the number  
127 of free parameters. The model is composed of 3 networks, one basic and two auxiliary  
128 networks. After a basic discriminative network with optional reconstruction path, a  
129 network that predicts the input fat layer parameters, and finally, a network that predicts  
130 the reconstruction fat layer parameters. We applied their approach to the PPMI  
131 dataset. The area under the curve (AUC) of the deep learning model was 0.67 (95%  
132 CI = [0.47; 0.83]) and the Youden index was 0.29 (Tab. 2). Notably, the deep learning  
133 model contains abstract embeddings instead of concrete SNPs as in the PRS.

134

## 135 **Feature selection and LASSO regression**

136 The hierarchical graph, as well as the training set of the GWAS data were used to  
137 select a set of less than 100 polygenic hypotheses using our approach as described in  
138 the introduction above. The remaining hypotheses were summarized in a term that was  
139 used to train a LASSO regression model on the validation data. (Tibshirani, 1996) This  
140 model, based on 47 SNPs in several different interaction terms, then predicted the test  
141 set. The area under the curve (AUC) for the LASSO model with prior feature selection  
142 was 0.85 (95% CI = [0.72; 0.96]) and the Youden index was 0.61. A LASSO model  
143 without prior feature selection that was built for comparison did not deliver outcomes  
144 that were significantly better than chance (Tab. 2).

145 Exploring the feature selection based model with its interactive terms, provides insights  
146 about the genes associated with the disease. An annotation of all 47 SNPs in our model  
147 can be found in the Supplementary Information. An exciting result from this analysis of  
148 the PPMI dataset is the statistical interaction of variants rs3822019 on chromosome 4  
149 in gene *TMEM175*, coding for a potassium channel in late endosomes, and  
150 rs17022452 on chromosome 2, close to the coding region of *GAPDHP25*,  
151 glyceraldehyde-3 phosphate dehydrogenase pseudogene 25. rs3822019 is an intron  
152 variant that has been linked to Parkinson's Disease. (Nalls et al., 2014)

153

154

## 155 **Discussion**

156 We analyzed the PPMI dataset and built predictive models using PLINK for a polygenic  
157 risk score, a diet-net deep learning algorithm for genomic data (Romero et al., 2016),  
158 and LASSO regression with and without the above proposed approach of using  
159 contextual data to reduce the hypothesis space. The PRS model comprises 16,135  
160 SNPs and showed an AUC of 0.56 whereas the deep learning model had an AUC of

161 0.52. Notably, the deep learning model consists of abstract embeddings instead of  
162 single SNPs like the PRS. Therefore, identification of disease-associated SNPs and  
163 further insights into the disease mechanism are not possible here. The LASSO  
164 regression model built on interactions containing only 47 SNPs that were discovered  
165 via the use of contextual information outperformed the other predictive models with an  
166 AUC of 0.82. Beyond that, the approach was able to associate new variants with the  
167 disease that would have not shown up under an additive approach such as PRS.

168 We investigated how the combinations of the relevant genotypes rs3822019\_TT  
169 (*TMEM175*) and rs17022452\_GG (*GAPDHP25*) split the individuals into cases and  
170 controls (Tab. 3). All subjects that are homozygous for rs3822019\_TT are affected by  
171 PD. Furthermore, most individuals heterozygous for this variant (rs3822019\_TT) or  
172 homozygous for rs17022452\_GG are cases (76.4% and 75.0%, respectively). These  
173 results support the relevance of the association between these variants and PD status.

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

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



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

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

205

206

## 207 **Methods**

### 208 **Parkinson’s progression marker initiative dataset**

209 The Parkinson’s progression marker initiative (PPMI) dataset (<https://www.ppmi->  
210 [info.org](https://www.ppmi-info.org)) contains 471 subjects, 368 cases and 152 controls, for each subject  
211 genotyping information from two complementary chips (NeuroX and ImmunoChip) was

212 collected. (Marek et al., 2011) After careful quality control and harmonization (e.g.  
213 genome build, strand alignment) as described in the literature (Marees et al., 2018) we  
214 merged that information into a single dataset with 380,939 variants in total. An  
215 additional set of quality control steps were performed on variants and individuals that  
216 aimed to remove biases that could affect the downstream analysis.

217 In more detail, in a first stage SNPs and individuals were filtered based on their  
218 missingness in the dataset. This ensures that SNPs are excluded that have a high  
219 proportion of subjects where genotyping information was unavailable or of poor quality.  
220 Similarly, individuals where a large proportion of SNPs could not be measured were  
221 excluded. This step was achieved by setting a threshold of 0.02 (i.e. >2%; 6,084  
222 variants and 22 people were removed). SNP filtering was performed before individuals  
223 were filtered.

224 With high missing rates filtered, all variants not on autosomal chromosomes were  
225 removed (5,731 variants were removed). This was followed by the identification and  
226 removal of variants that deviate from the Hardy-Weinberg equilibrium. These variants  
227 were identified in a two-stage process whereby we first applied a threshold of  $1e-6$   
228 exclusively to controls, followed by a threshold of  $1e-10$  applied to all samples (0 and  
229 202 variants were removed). This is a common indicator of genotyping errors.

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

236 Finally, related individuals were removed, which was achieved through the assessment  
237 of their respective identity by descent coefficients (IBD) that were calculated. Only one  
238 individual in a related pair would be kept (0 individuals were removed).

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

241

## 242 **GWAS**

243 In a preliminary proof-of-concept step, a genome-wide association (GWA) analysis  
244 was performed with the R package SAIGE (Zhou et al., 2018) to test individual variants  
245 for their association with Parkinson's Disease.

246

## 247 **Polygenic risk score**

248 The data ( $n = 436$ ) was then split into training ( $n = 367$ ), validation ( $n = 33$ ) and test  
249 sets ( $n = 36$ ). The same sets were used for constructing a polygenic risk score (PRS)  
250 and machine learning prediction models with and without feature selection. To  
251 calculate the PRS, different  $p$  value thresholds for the subjects in the training, validation  
252 and test set were used. The PRS was constructed by using PLINK(Purcell et al., 2007)  
253 following the guidelines provided by Choi et al. (Choi et al., 2020) and the  
254 accompanying tutorial (<https://choishingwan.github.io/PRS-Tutorial/plink/>.) The  
255 clumping cut-off of  $r^2$  was 0.1. The  $p$ -value threshold was 0.05 for the subjects in the  
256 training, validation, and test sets. The PRS of the subjects in the training set were then  
257 used to train a separate logistic regression classifier using the *glm* function in R  
258 ([www.R-project.org](http://www.R-project.org)) for each  $p$  value threshold. The validation data set was used to  
259 determine which of these thresholds produces the best classifier, which was then used  
260 to predict the test set.

261

## 262 **Deep learning**

263 The deep learning prediction model was built using a Diet Network according to the  
264 procedure described by Romero et al.. (Romero et al., 2016) The official code can be  
265 found here: <https://github.com/adri-romsor/DietNetworks>.

266

## 267 **Feature selection**

268 A knowledge graph containing contextual information mined from public databases  
269 such as e.g. dbSNP, ClinVar, OMIM, Reactome, STRING database proposes 100  
270 SNPs and SNP-SNP interactions. The proposed SNPs and SNP-SNP interactions are  
271 evaluated in the training data by drop one model comparison procedure using the *glm*  
272 function in R ([www.R-project.org](http://www.R-project.org)). (Klinger et al.) A gradient descent algorithm directs  
273 the search across the graph based on whether proposed hypotheses are correlated  
274 with the disease status or not and corrects accordingly the SNP and SNP-SNP  
275 interactions list until all members of the list show strong correlation with the disease  
276 status.

277

## 278 **LASSO regression**

279 LASSO regression models were computed by using the glmnet package  
280 (<https://glmnet.stanford.edu/index.html> for R ([www.R-project.org](http://www.R-project.org)) and its function  
281 cv.glmnet. (Friedman et al., 2010)

282

283

## 284 **Acknowledgements**

285 Data used in the preparation of this article were obtained from the Parkinson's  
286 Progression Markers Initiative (PPMI) database ([www.ppmi-info.org/data](http://www.ppmi-info.org/data)). For up-to-  
287 date information on the study, visit [www.ppmi-info.org](http://www.ppmi-info.org). PPMI – a public-private

288 partnership – is funded by The Michael J. Fox Foundation for Parkinson’s Research  
289 and funding partners. List the full names of all of the PPMI funding partners can be  
290 found at [www.ppmi-info.org/fundingpartners](http://www.ppmi-info.org/fundingpartners). The research work was supported by the  
291 Investitionsbank des Landes Brandenburg (ILB), the European Regional Development  
292 Fund (ERDF), and the European Social Fund+ (ESF+). We also thank the program  
293 digital solutions made in Brandenburg (digisolBB) for its continued support.

294

295

## 296 **References**

- 297 Bateson, and William (1906). The progress of genetics since the rediscovery of  
298 Mendel's papers. *Progress. Rei Bot.* 1, 368.
- 299 Butera, G., Mullappilly, N., Masetto, F., Palmieri, M., Scupoli, M. T., Pacchiana, R., et  
300 al. (2019). Regulation of Autophagy by Nuclear GAPDH and Its Aggregates in  
301 Cancer and Neurodegenerative Disorders. *Int. J. Mol. Sci.* 20.  
302 doi:10.3390/ijms20092062.
- 303 Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H.  
304 (2013). Projecting the performance of risk prediction based on polygenic analyses  
305 of genome-wide association studies. *Nat. Genet.* 45, 400–5, 405e1–3.
- 306 Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing  
307 polygenic risk score analyses. *Nat. Protoc.* 15, 2759–2772.
- 308 Diogo, D., Tian, C., Franklin, C. S., Alanne-Kinnunen, M., March, M., Spencer, C. C.  
309 A., et al. (2018). Phenome-wide association studies across large population  
310 cohorts support drug target validation. *Nat. Commun.* 9, 4285.
- 311 Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS*  
312 *Genet.* 9, e1003348.
- 313 Duncan, L., Shen, H., Gelaye, B., Meijssen, J., Ressler, K., Feldman, M., et al. (2019).  
314 Analysis of polygenic risk score usage and performance in diverse human  
315 populations. *Nat. Commun.* 10, 3328.
- 316 Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information  
317 contained within genome-wide association studies to improve individual prediction  
318 of complex disease risk. *Hum. Mol. Genet.* 18, 3525–3531.

- 319 Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for  
320 Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22.
- 321 Gola, D., Erdmann, J., Müller-Myhsok, B., Schunkert, H., and König, I. R. (2020).  
322 Polygenic risk scores outperform machine learning methods in predicting coronary  
323 artery disease status. *Genet. Epidemiol.* 44, 125–138.
- 324 International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L.,  
325 Visscher, P. M., O'Donovan, M. C., et al. (2009). Common polygenic variation  
326 contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752.
- 327 Jinn, S., Blauwendraat, C., Toolan, D., Gretzula, C. A., Drolet, R. E., Smith, S., et al.  
328 (2019). Functionalization of the TMEM175 p.M393T variant as a risk factor for  
329 Parkinson disease. *Hum. Mol. Genet.* 28, 3244–3254.
- 330 Jinn, S., Drolet, R. E., Cramer, P. E., Wong, A. H.-K., Toolan, D. M., Gretzula, C. A.,  
331 et al. (2017). TMEM175 deficiency impairs lysosomal and mitochondrial function  
332 and increases  $\alpha$ -synuclein aggregation. *Proc. Natl. Acad. Sci. U. S. A.* 114, 2389–  
333 2394.
- 334 Klinger, J., Ravarani, C., Bannard, C., Lamparter, M., Schwinges, A., Cope, J., et al.  
335 Critically ill COVID-19 status associated trait genetics reveals CDK6 inhibitors as  
336 potential treatment. doi:10.21203/rs.3.rs-568366/v1.
- 337 Krohn, L., Öztürk, T. N., Vanderperre, B., Ouled Amar Bencheikh, B., Ruskey, J. A.,  
338 Laurent, S. B., et al. (2020). Genetic, Structural, and Functional Evidence Link  
339 TMEM175 to Synucleinopathies. *Ann. Neurol.* 87, 139–153.
- 340 Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why significant variables aren't  
341 automatically good predictors. *Proc. Natl. Acad. Sci. U. S. A.* 112, 13892–13897.

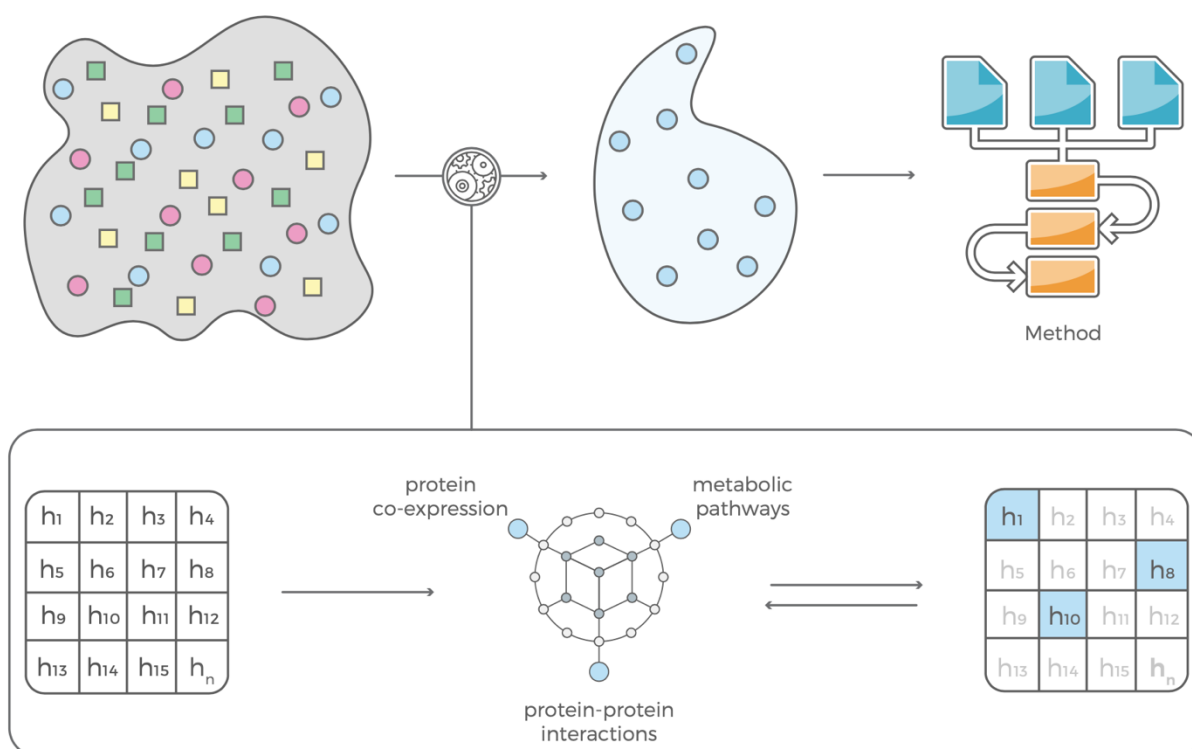
- 342 Lynch-Day, M. A., Mao, K., Wang, K., Zhao, M., and Klionsky, D. J. (2012). The role of  
343 autophagy in Parkinson's disease. *Cold Spring Harb. Perspect. Med.* 2, a009357.
- 344 Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J.,  
345 et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461,  
346 747–753.
- 347 Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., et  
348 al. (2018). A tutorial on conducting genome-wide association studies: Quality  
349 control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608.
- 350 Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C.,  
351 et al. (2018). The Parkinson's progression markers initiative (PPMI) - establishing  
352 a PD biomarker cohort. *Ann Clin Transl Neurol* 5, 1460–1477.
- 353 Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011).  
354 The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol.* 95, 629–  
355 635.
- 356 Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., et al.  
357 (2014). Large-scale meta-analysis of genome-wide association data identifies six  
358 new risk loci for Parkinson's disease. *Nat. Genet.* 46, 989–993.
- 359 Olanow, C. W., Schapira, A. H. V., LeWitt, P. A., Kieburtz, K., Sauer, D., Olivieri, G., et  
360 al. (2006). TCH346 as a neuroprotective drug in Parkinson's disease: a double-  
361 blind, randomised, controlled trial. *Lancet Neurol.* 5, 1013–1020.
- 362 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et  
363 al. (2007). PLINK: a tool set for whole-genome association and population-based  
364 linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.



- 365 Reisberg, S., Ijjasenko, T., Läll, K., Fischer, K., and Vilo, J. (2017). Comparing  
366 distributions of polygenic risk scores of type 2 diabetes and coronary heart disease  
367 within different populations. *PLoS One* 12, e0179238.
- 368 Romero, A., Carrier, P. L., Erraqabi, A., Sylvain, T., Auvolat, A., Dejoie, E., et al. (2016).  
369 Diet Networks: Thin Parameters for Fat Genomics. *arXiv [cs.LG]*. Available at:  
370 <http://arxiv.org/abs/1611.09340>.
- 371 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*  
372 58, 267–288.
- 373 Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human  
374 complex traits. *Nat. Rev. Genet.* 15, 722–733.
- 375 Wray, N. R., Goddard, M. E., and Visscher, P. M. (2007). Prediction of individual  
376 genetic risk to disease from genome-wide association studies. *Genome Res.* 17,  
377 1520–1528.
- 378 Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et  
379 al. (2010). Common SNPs explain a large proportion of the heritability for human  
380 height. *Nat. Genet.* 42, 565–569.
- 381 Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et  
382 al. (2018). Efficiently controlling for case-control imbalance and sample  
383 relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341.

384 **Figures and Tables**

385 **Figure 1**

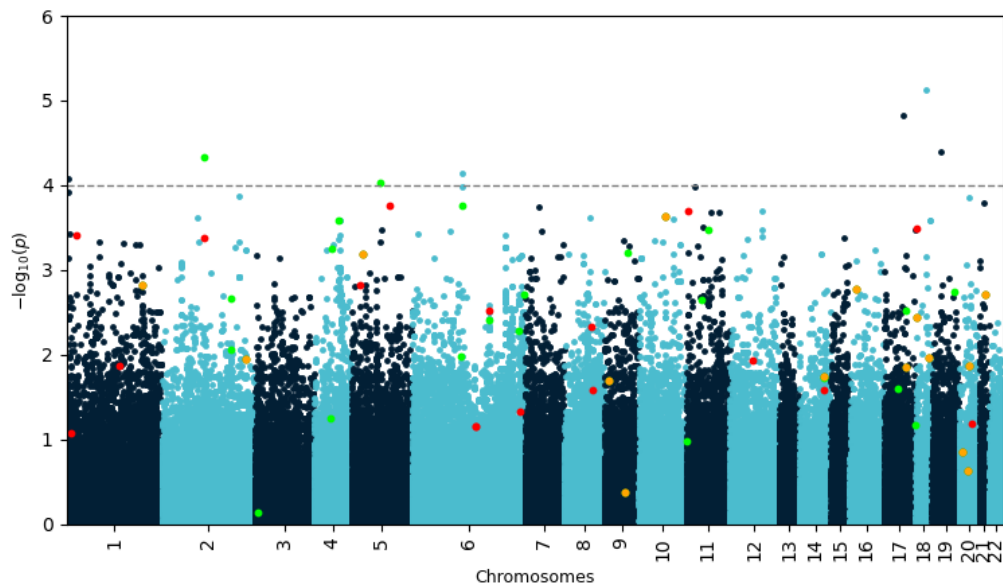


386

387 **Fig. 1.** Our feature selection consists of two complementary modules that are in  
388 feedback with each other. The contextual module uses information mined from the  
389 scientific literature, pathway libraries and protein co-expression data and an evaluation  
390 module that estimates predictive power of a feature based on that contextual  
391 information. The selected features can be used to build prediction models with  
392 standard machine learning algorithms.

393

394 **Figure 2**

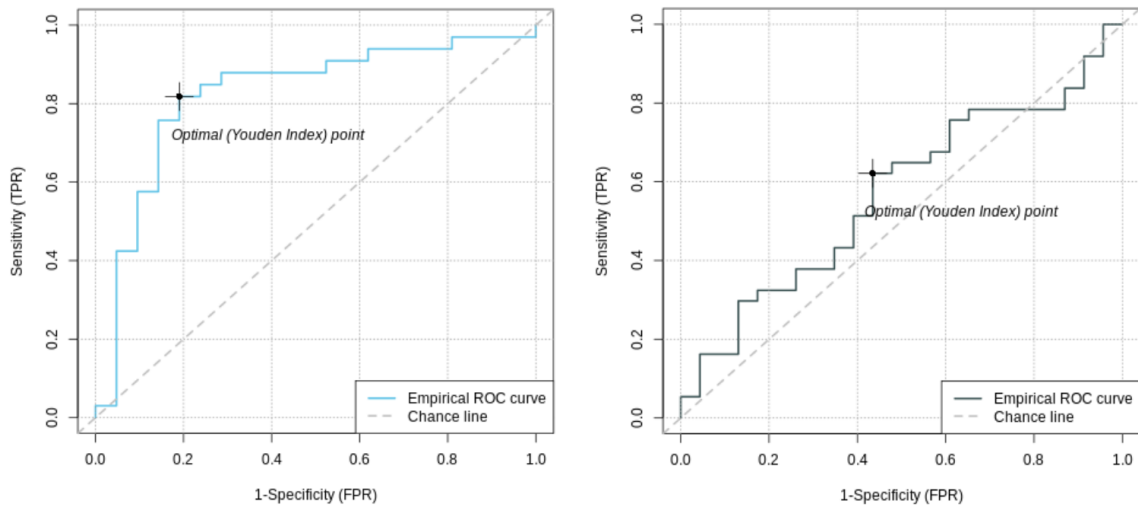


395

396 **Fig. 2.** Manhattan plot of negative decadic logarithm of  $p$  values for SNPs as  
397 determined by SAIGE analysis. Variants identified by our biotx.ai model are highlighted  
398 in red and green if they increase or decrease disease risk, respectively. Variants  
399 highlighted in orange occur in both protective and risk-enhancing groups of SNPs,  
400 depending on their genotype. Most of these biologically meaningful variants would  
401 have been missed by using a simple  $p$  value cutoff.

402

403 **Figure 3**



404

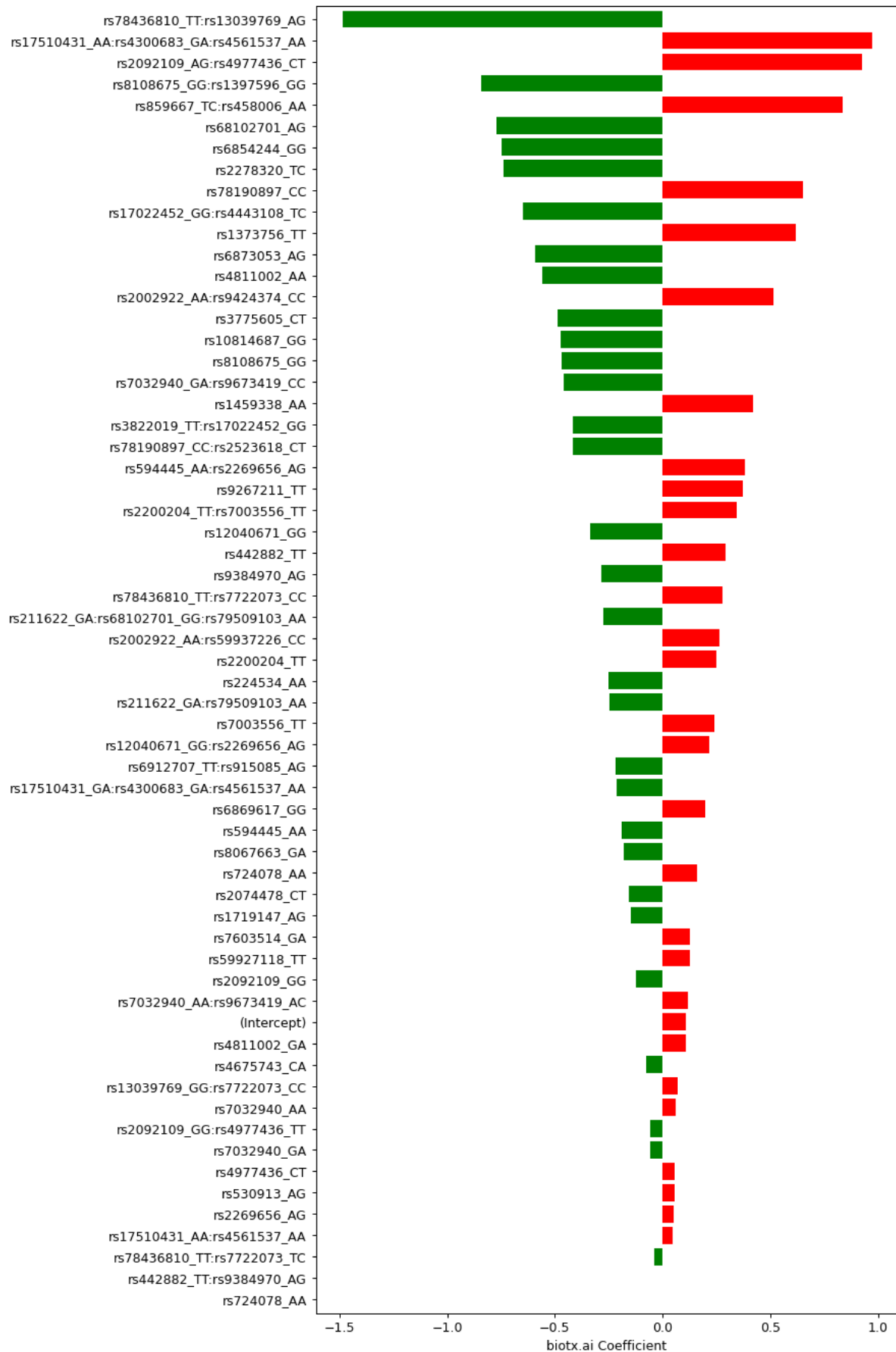
405 **Fig. 3.** Receiver operating characteristic (ROC) curves of feature selected machine  
406 learning model (left) and polygenic risk score (right). The AUC of the feature selected  
407 model with 0.85 (95% CI = [0.72; 0.96]) is better than the AUC of the PRS with 0.56  
408 (95% CI = [0.42; 0.81]).

409

410

411

412 **Figure 4**



414 **Fig. 4.** Coefficients determined by our biotx.ai model for SNPs and groups of SNPs.  
415 Negative values (green) indicate protective (combinations of) variants, positive values  
416 (red) mark risk variants. The respective genotypes of each variant are indicated by  
417 one-letter codes of the bases, where the first letter corresponds to the reference allele,  
418 and the second corresponds to the observed, alternative allele.

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439 **Table 1**

440 **Tab 1.** PPMI GWAS results identified 7 SNPs with a p value < 10<sup>-4</sup>. Positions and rs  
441 IDs according to Human Genome Reference hg19 (GRCh37).

Ch r	Pos	SNP ID	rs ID	Genic Context	p value
1	173266578	imm_1_17 1533201	rs4916319	<i>TNFSF4</i> (upstream)	0.000083
2	209087335	exm22611 59	rs4675743		0.000046
5	156376703	exm49891 7	rs6873053	<i>TIMD4</i> (downstream)	0.000092
6	133716974	rs212805	rs212805	<i>EYA4</i>	0.000074
17	25895033	imm_17_2 2919160	rs4795747		0.000015
18	5479093	rs7238186	rs7238186	<i>EPB41L3</i> (downstream)	0.000007
19	57909872	exm15132 84	rs4801478	<i>ZNF548</i>	0.000040

442

443

444

445

446

447

448

449

450

451

452

453

454 **Table 2**

455 **Tab. 2.** Performance comparison of all models.

<b>Method</b>	<b>AUC [95% CI]</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>Youden' s Index</b>
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

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472



473 **Table 3**

474 **Tab. 3.** PD cases and controls among bearers of the respective genotype combinations  
 475 of the identified variants rs3822019 and rs17022452.

<b>Genotype combination</b>	<b>cases</b>	<b>controls</b>
rs3822019_TT / rs17022452_GG	0	0
rs3822019_TT / rs17022452_GA	6 / 100%	0
rs3822019_TC / rs17022452_GG	2 / 50%	2 / 50%
rs3822019_TT / -	7 / 100%	0
- / rs17022452_GG	7 / 87.5%	1 / 12.5%
rs3822019_TC / rs17022452_GA	27 / 87.1%	4 / 12.9%
rs3822019_TC / -	68 / 73.9%	24 / 26.1%
- / rs17022452_GA	66 / 75%	22 / 25%
- / -	113 / 56.5%	87 / 43.5%

476

477