A machine learning approach for gene prioritisation in Parkinson's

Aymeric Lanore, Aymeric Basset, Suzanne Lesage

SCIENTIFIC COMMENTARY

A machine learning approach for gene prioritisation in Parkinson's disease

This scientific commentary refers to 'Machine learning nominates the inositol pathway and novel genes in Parkinson's disease' by Yu *et al.* (https://doi.org/10.1093/brain/awad345).

Recent genome-wide association studies (GWAS) have identified 90 independent risk variants for idiopathic Parkinson's disease at 78 loci scattered across the genome.[1] Each of these variants individually confers a low risk of developing Parkinson's disease, rendering the disease polygenic. Single-nucleotide polymorphisms (SNPs) may be pathogenic through their effects on the protein, due to the disruption of normal protein function, or via effects on gene expression.[2] By tracing variants back to genes and then to pathways, we should be able to identify the genetic basis of biological susceptibilities and improve our understanding of the underlying mechanisms. However, it is difficult to prove the causality of variants in disease due to linkage disequilibrium, the systematic association of alleles at different loci on the same chromosome[3]; indeed, significant SNPs identified in GWAS may be genuinely causal or merely correlated with causal SNPs within loci. The presence within the newly identified GWAS loci of genes with well-established causal roles in monogenic Parkinson's disease is of particular interest. In this issue of *Brain*, Yu and colleagues[4] propose two hypotheses: firstly, that these genes mediate causality at the loci within which they are found; and secondly, that this causality can be determined using genomic information, potentially enabling the discovery of causal genes at other loci.

The objective of the study was to prioritise the genes at each locus according to how likely they are to be involved in Parkinson's disease, using a new machine learning approach. The

authors first defined 78 loci based on the 90 independent risk variants previously described[1], including all protein-coding genes within 1 Mb of the risk variants. They excluded SNPs identified as non-causal by 'echolocatoR',[5] and then created a classification model from the remaining SNPs.[6] The training task was the classification of seven well-established Parkinson's disease genes (*GBA1*, *LRRK2*, *SNCA*, *GCH1*, *MAPT*, *TMEM175*, *VPS13C;* positively labelled) relative to 205 other genes at the same loci which were thus unlikely to be driving the association with Parkinson's disease (negative labelling).

The features tested were 1) distance from the top-ranking SNPs, 2) the predicted consequences of variants for protein structure and function according to the bioinformatics tools 'Variant Effect Predictor' (VEP) and Polyphen-2; 3) molecular quantitative trait loci (QTL), including expression QTL (eQTL) and splicing QTL (sQTL); and 4) gene expression in different tissues and cell types. The authors used SHAP values[7] to identify the features of the model making the largest contributions to prediction.[7] They found that the following features were the most useful for predicting causality: distance from the top-ranked Parkinson's disease-associated SNP in the locus to the transcription start site or start of the gene; the severity of the consequences of the variant predicted by VEP; and mRNA levels within specific dopaminergic neuron subtypes. Epigenetic features were not found to have significant predictive value. Once trained, the model was used to prioritise genes not seen during the training phase within new loci. The model identified 76 genes at 78 loci, two of which (*MAPT* and *TOX3*) were identified twice in neighbouring loci (Fig. 1).

In addition to missense variants of *GBA1*, *LRRK2* and *GCH1*, which are known to contribute to Parkinson's disease, the authors identified missense SNPs that contributed to the scores of two candidate genes, *SPNS1* and *MLX* (p.L512M, rs7140 and p.Q139R, rs665268, respectively) prioritised by the model within loci with possible functional consequences. Interestingly, both of the SNPs identified are also eQTLs/sQTLs for *SPNS1* and *MLX* in

several Parkinson's disease-relevant tissues in the Genotype-Tissue Expression (GTEx) database. The authors investigated the potential consequences of these variants by performing *in silico* structural analyses of the proteins encoded by these genes. *SPNS1* encodes a lysophospholipid transporter with several alternative isoforms. The p.L512M variant is found in an unstructured region of the C-terminus of one of the isoforms of this membrane-bound protein. It is located in the lumen of the lysosome, but the effect of this variant on SPNS1 is unknown. *MLX* encodes a Max-like protein X from a family of transcription factors involved in energy metabolism that interact with other related proteins, such as the MAD family of transcriptional repressors and the MONDO family of transcriptional activators. The p.Q139R variant may affect the dimerization of MLX with its transcriptional repressors or activators. This variant has been suggested to increase oxidative stress and to suppress autophagy in immune cells and has been shown to be associated with Takayasu's arteritis.

Yu and co-workers performed gene set enrichment analysis to identify the specific pathways and mechanisms via which the genes they identified could be involved in Parkinson's disease. Enrichment was found for eight gene ontology (GO) biological processes and eight GO cellular components, after correction for the false discovery rate. The authors focused on two novel pathways among the GO biological processes found to display significant enrichment: inositol phosphate biosynthesis and polyol biosynthesis. Importantly, inositol is associated with four candidate genes: *ITPKB*, *IP6K2*, *PPIP5K2*, and *INPP5F*. The authors investigated the association between these putative novel Parkinson's disease pathways and Parkinson's disease status by calculating pathway-specific polygenic risk scores (PRS) with PRSet.[8] We assume that the authors used 14,207 Parkinson's disease cases and 12,981 controls from six cohorts, all of European ancestry, after excluding one cohort (the Vance cohort) from the meta-analysis due to significant heterogeneity. In three analyses of PRS, heterogeneity remained significant, making it necessary to consider a random effects

model. The associations detected were significant, but neither of the two new pathways identified was a strong risk factor, with odds ratios of 1.15 for inositol phosphate biosynthesis and 1.20 for polyol biosynthesis. These associations remained significant even after the exclusion of the top-ranking genes identified in the analysis.

In their analysis of single-cell RNA data for dopaminergic neurons from two datasets, the authors detected differential expression for many genes, including, in particular, *INPP5F*, which is involved in the inositol pathway. By contrast, no significant differential expression was observed in the bulk RNAseq analysis. Finally, the authors performed meta-analyses of rare variants in genes identified by their model, based on whole-exome sequencing data for 3,202 patients with Parkinson's disease, 6,284 proxy patients and 143,884 controls from two cohorts. They identified rare variants significantly associated with idiopathic Parkinson's disease in genes implicated in the monogenic form of the disease (*GBA1, LRRK2* and *GCH1*), and in two new genes, *KCNIP3* and *LSM7*. Rare variant burden tests also revealed an association of the polyol/inositol biosynthetic pathways with Parkinson's disease.

The machine learning model used in this study has several limitations. First, the authors assumed that the behaviour of features depends on specific loci, an assumption that could lead to bias. Outliers at these loci could mask significant signals or give undue weight to irrelevant traits. Secondly, the assumption during training that all genes identified at each locus are negative, bar the seven well-established genes, could lead to underperformance of the model, particularly if this negative status is not demonstrated. Another possible approach would involve assuming that other genes have an indeterminate status, thereby bringing the task into the domain of semi-supervised learning.[9] The use of an advanced approach, such as XGBoost, on a dataset for only 212 genes, seven of which are positive, is audacious. It might be more prudent, as a first approach, to use simpler models, evaluate their performance, and then compare the results obtained with those of more complex and harder-to-optimise algorithms,

to mitigate the risk of overfitting. In addition, the size of the dataset and the lack of a designated test set greatly increase the probability of overfitting due to feature selection and hyperparameter tuning being performed on the same set. The number of features retained in the model, 78, also seems quite high, given the risk of overfitting. A more stringent feature selection step would further improve the model. The SHAP analysis provides information about the model, but it should be borne in mind that its relevance depends on the performance of the underlying model.

In terms of validation, comparison with the other six pathways identified by analysis of GO biological processes might provide a better understanding of the importance of the inositol phosphate and polyol biosynthetic pathways. For the analysis of PRS pathways, a comparison with the global PRS and the other PRS pathways would make it possible to weight the pathways studied. The persistence of heterogeneity in the meta-analysis of PRS may necessitate the elimination of additional cohorts, but the odds ratios appeared low for both random- and fixed-effects models.

In conclusion, 83% of the top-ranking genes identified by Yu and co-workers were the closest genes in terms of distance from the top-ranking SNPs in the GWAS already identified by simple proximity annotation. The method used is innovative and could be improved by taking into account the points discussed above. The involvement of the inositol phosphate and polyol biosynthetic processes in Parkinson's disease should be confirmed by further analysis, particularly in populations of non-European ancestry, and by functional analysis. The authors also identified missense SNPs in *SPNS1* and *MLX* and obtained significant burden test results for variants of the *KCNIP3* and *LSM7* genes, for which further functional investigation would be valuable, to shed light on their role in Parkinson's disease.

Aymeric Lanore[1,2]*, Aymeric Basset*[1]*, Suzanne Lesage*[1]

1. Sorbonne Université, Institut du Cerveau – Paris Brain Institute – ICM, Inserm, CNRS, Paris, France

2. Assistance Publique Hôpitaux de Paris, Département de Neurologie, CIC Neurosciences, Hôpital Pitié-Salpêtrière, Paris, France

E-mail: suzanne.lesage@upmc.fr

Conflicts of interest

The authors report no conflicts of interest.

Funding

S.L. has received grants from *Fondation pour la Recherche Médicale* (FRM, MND202004011718).

Figure legend

Figure 1 Schematic representation of the machine-learning analysis and results. The authors applied a classification model to prioritise genes most likely to be involved in Parkinson's disease within loci identified in recent GWAS, based on features such as distance from the top-ranking SNPs, variant consequences, molecular QTL, and gene expression. They identified 76 genes in 78 loci with the trained model, with *MAPT* and *TOX3* identified twice. The authors then explored the genes identified, demonstrating differential expression of many in single-cell RNA data from dopaminergic neurons. In addition, they detected missense SNPs in *SPNS1* and *MLX* with potential consequences for protein function, identified

previously unknown rare variants of *KCNIP3* and *LSM7*, and obtained evidence suggesting involvement of the inositol phosphate and polyol biosynthetic pathways in Parkinson's disease.

GWAS, genome-wide association study; QTL, quantitative trait loci; RNAseq, RNA sequencing; scRNA, single-cell RNA; SNPs, single-nucleotide polymorphisms

References

1. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019;18(12):1091-1102. doi:10.1016/S1474-4422(19)30320-5

2. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Primers*. 2021;1(1):1-21. doi:10.1038/s43586-021-00056-9

3. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. 2008;9(6):477-485. doi:10.1038/nrg2361

4. Yu E, Larivière R, Thomas RA, et al. Machine learning nominates the inositol pathway and novel genes in Parkinson's disease. *Brain*. Published online October 6, 2023:awad345. doi:10.1093/brain/awad345

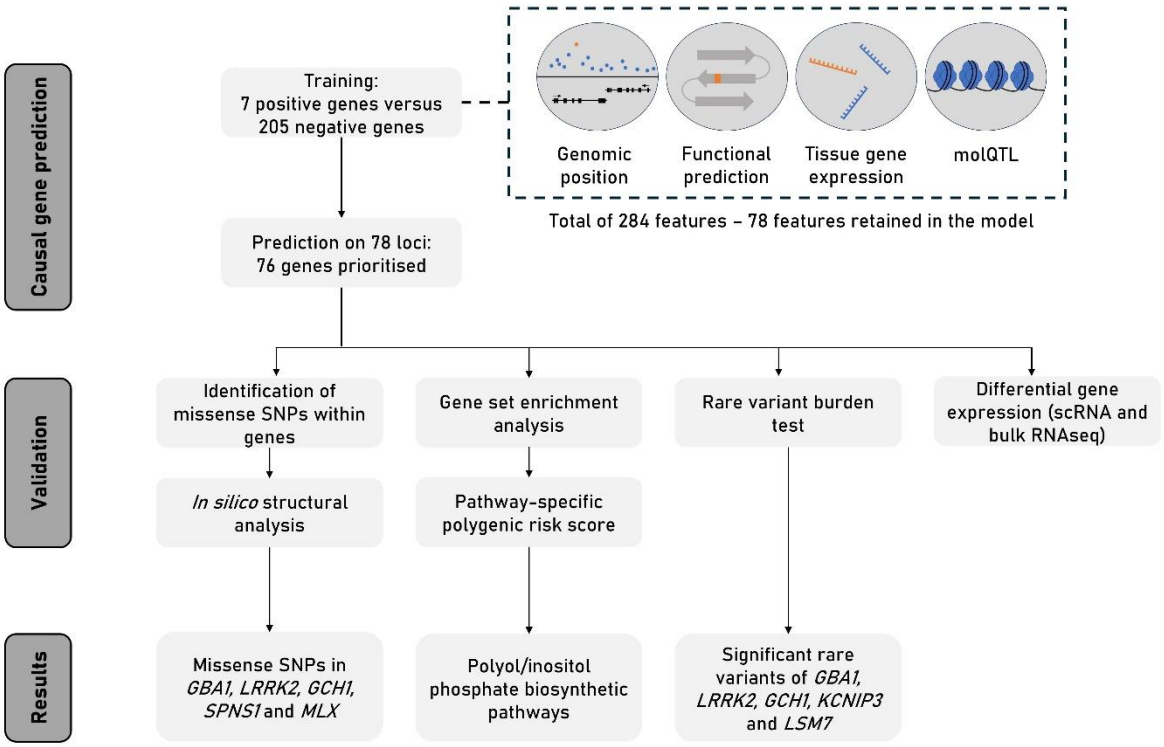
5. Schilder BM, Humphrey J, Raj T. echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline. *Bioinformatics*. 2021;38(2):536-539. doi:10.1093/bioinformatics/btab658

6. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ; 2016:785-794. doi:10.1145/2939672.2939785

7. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed January 9, 2024. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

8. Choi SW, García-González J, Ruan Y, et al. PRSet: Pathway-based polygenic risk score analyses and software. *PLOS Genetics*. 2023;19(2):e1010624. doi:10.1371/journal.pgen.1010624

9. Chapelle O, Schölkopf B, Zien A, eds. *Semi-Supervised Learning*. MIT Press; 2006.


Causal gene prediction
Validation
Results
Training:
7 positive genes versus
205 negative genes
Genomic position
Functional prediction
Tissue gene expression
molQTL
Total of 284 features – 78 features retained in the model
Prediction on 78 loci:
76 genes prioritised
Identification of missense SNPs within genes
Gene set enrichment analysis
Rare variant burden test
Differential gene expression (scRNA and bulk RNAseq)
In silico structural analysis
Pathway-specific polygenic risk score
Missense SNPs in GBA1, LRRK2, GCH1, SPNS1 and MLX
Polyol/inositol phosphate biosynthetic pathways
Significant rare variants of GBA1, LRRK2, GCH1, KCNIP3 and LSM7