Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease

J. Diana Zhang,[1,2] Chonghua Xue,[2] Vijaya B. Kolachalama,[2,3,*] and William A. Donald[1,*]

[1] *School of Chemistry, University of New South Wales, Sydney, Australia*

[2] *Department of Medicine, Boston University School of Medicine, Boston, MA, USA*

[3] *Department of Computer Science and Faculty of Computing & Data Sciences, Boston University, Boston, MA, USA*

Correspondence to:

vkola@bu.edu

w.donald@unsw.edu.au

Abstract

The use of machine learning (ML) with metabolomics provides opportunities for the early diagnosis of disease. However, the accuracy and extent of information obtained from ML and metabolomics can be limited owing to challenges associated with interpreting disease prediction models and analysing many chemical features with abundances that are correlated and ‘noisy’. Here, we report an interpretable neural network (NN) framework to accurately predict disease and identify significant biomarkers using whole metabolomics datasets without feature selection. The performance of the NN approach for predicting Parkinson’s disease (PD) from blood plasma metabolomics data was significantly higher than classical ML methods with a mean area under the curve of > 0.995. PD-specific markers that contribute significantly to early disease prediction were identified including an exogenous polyfluoroalkyl substance. It is anticipated that this accurate and interpretable NN-based approach can improve diagnostic performance for many other diseases using metabolomics and other untargeted ‘omics methods.

Introduction

The rate of Parkinson's Disease (PD) is growing more rapidly than any other neurological disease.[1] PD is typically diagnosed according to a clinical criteria of motor symptoms which include bradykinesia (slowness of movement), a resting tremor, and rigidity.[2] However, the onset of atypical non-motor symptoms such as sleep disorder, constipation, apathy, and loss of smell can predate clinically relevant symptoms by several years to decades.[3-5] In addition, for patients who present with Parkinson-like symptoms, the current process for identifying PD can often be inconclusive. For example, according to a meta-analysis by Rizzo et al.,[6] the overall diagnostic accuracy for PD based on an initial clinical assessment is 81% and the error rate for misclassifying PD is up to 20%. Accurate identification of PD using biomarker signatures rather than relying primarily on physical symptoms would be highly beneficial.

Biomarkers associated with metabolic processes are used extensively for understanding, diagnosing and monitoring diseases.[7, 8] Such metabolites are typically sampled from well-established matrices such as blood plasma and serum for trace-level analysis of tens to hundreds of metabolites using mass spectrometry (MS).[9] Additional matrices of emerging interest for biomarker discovery and disease diagnosis applications include the rapid and non-invasive sampling of skin sebum and breath.[10-12] Using MS, differences in the metabolite profiles in the blood plasma of pre-PD subjects were identified up to 15 years prior to a clinical diagnosis when compared to healthy controls who did not develop PD.[13] These results suggest that PD may potentially be diagnosed using metabolite biomarkers significantly earlier than in current practice, particularly if analysing such metabolites can result in high diagnostic accuracy and be validated on large-scale cohort studies.

To develop accurate prediction models for disease diagnosis using large metabolomics datasets, machine learning (ML) approaches are widely used. However, the use of whole metabolomics datasets to build prediction models is rarely used because these datasets often contain many highly correlated and 'noisy' chemical features which can risk model overtraining and reduce diagnostic performance.[14] As a result, models are typically based on a smaller subset of features which are determined by conventional statistical methods (e.g., based on *p*-values and fold-changes of individual features). For example, Gonzalez-Riano et al.[13] reported a classical ML model using 20 pre-selected biomarkers to diagnose pre-PD vs healthy controls from blood plasma samples. Similarly, Sinclair et al.[15] used a classical ML model with 15 and 26 pre-selected

biomarkers to diagnose drug naïve PD and medicated PD vs healthy control, respectively, from skin sebum samples. However, given that the abundances of metabolites are often correlated and can depend nonlinearly on the abundances of other metabolites, such classical ML approaches may potentially 'hide' some key features in metabolomics datasets.

Advanced ML approaches such as neural networks (NN) are particularly well-suited for processing large volumes of data interconnectivity and building models for datasets that contain nonlinear effects.[16] However, a fundamental issue in using methods such as NN for classifying complex mixtures based on metabolomics data is that the resulting predictive models are generally considered as uninterpretable 'black boxes,' which cannot be readily used to reveal mechanistic information.[17, 18] Recently, a new approach entitled Shapley Additive exPlanations (SHAP) was developed to 'interpret' ML models by retrospectively calculating the contribution of individual features to the accurate predictive performance of a model.[19] However, SHAP has not been used in the analysis of metabolomics datasets given that methods for interpreting ML models have only recently been developed and using all chemical features risks overtraining prediction models. Ideally, whole metabolomics datasets should be included in the ML model for SHAP to identify key metabolites that drive model prediction.

Here, we report an interpretable and computationally efficient neural network-based framework for analysing datasets generated by untargeted mass spectrometry-based methods (Figure 1) entitled, 'CRANK-MS' (Classification and Ranking Analysis using Neural network generates Knowledge from Mass Spectrometry). CRANK-MS has several built-in features including: (i) integrated model parameters that allow the high dimensionality of metabolomics datasets to be analysed without the need for pre-selecting chemical features; (ii) SHAP to retrospectively 'mine' key chemical features that contribute the most to an accurate model prediction; and (iii) benchmark testing with five alternative ML methods to compare diagnostic performance and further verify significant chemical features. Using CRANK-MS, we report the highest diagnostic performance to date for binary classification of PD *vs* healthy control with a mean area under the curve of > 0.995 Additionally, NN-driven predictions were used to reveal new PD-specific chemical features which were not previously identified. The program for implementing this approach is freely available online at https://github.com/CRANK-MS.

Methods

Data

Datasets from two cross-sectional PD metabolomics studies[13, 15] were used throughout. The Spanish European Prospective Study on Nutrition and Cancer (EPIC) study[13] involved metabolomics data from blood plasma samples taken from subjects who later developed PD up to 15 years later, and those who did not develop PD (total number of participants, n = 78; Table 1). The blood plasma samples from the EPIC study were analysed using four different instrumental methods (gas chromatography-MS, GC-MS; capillary electrophoresis-MS, CE-MS; and liquid chromatography-MS, LC-MS, in positive (+) and negative (-) ionisation modes). The NHS study[15] involved the LC-MS (+) analysis of skin sebum sampled from drug-naïve and medicated PD patients, and healthy controls (n = 274). The composite dataset from the EPIC study was prepared using the data from all four methods (i.e., GC-MS, CE-MS, LC-MS (+) and LC-MS (-)). Six participants were excluded in the composite dataset as data from one or more of the four methods were missing. The number of reported molecular features in the metabolomics datasets ranged from 60 to 6502 (Table 1). Binary classification (i.e., PD *vs* healthy) was conducted on each dataset using ML algorithms (see below).

Machine learning algorithms

The six machine learning algorithms and SHAP analysis were implemented in Python (v. 3.8). Random forest (RF), extreme gradient boosting (XGB), linear discriminant analysis (LDA), logistic regression (LR), and support vector machine (SVM) were written using scikit-learn packages (v 1.0.2). NN with multilayer perceptron was written using PyTorch (v. 1.10.2). Additional Python libraries used to support data analysis and visualisation include pandas (v. 1.4.2), numpy (v. 1.21.5), and matplotlib (v. 3.5.1). Full details of the open access code are available at https://github.com/CRANK-MS.

Hyperparameter tuning

For each ML model, hyperparameters correspond to specific model parameters that an algorithm uses to train on a given dataset. To determine the optimal hyperparameters, the composite dataset from the EPIC study was used (see above). Hyperparameter tuning for each ML model was optimised using the *GridSearchCV* package in scikit-learn. For non-linear ML models, the number of possible permutations resulting from different parameters ranged from 120-158. For LDA, the number of permutations used was 12. For each permutation, a bootstrap model was used in which the dataset was split randomly 100 times into 60%

training data and 40% validation data (i.e., 100 'bootstraps'). The optimised hyperparameters were determined based on the combination of hyperparameters that gave the highest Matthews correlation coefficient (MCC) after 100 iterations per permutation. These hyperparameters were then applied to all datasets in the study.

Performance metrics

To calculate diagnostic performance, the 100 randomly selected testing datasets from the bootstrapping process were used to calculate the overall diagnostic performance. The use of a bootstrap model with more replicates can lower absolute error compared to other sampling methods such as cross-validation and is considered useful for relatively small sample sizes.[20, 21] The diagnostic performance for each ML model was calculated based on the mean of the 100 bootstrap measurements and error was calculated as one standard deviation of the mean. For each ML model, accuracy, precision, sensitivity/recall, specificity, *F*1 score, and MCC score were calculated. Receiver operating characteristic (ROC) and precision-recall (PR) curves were generated to calculate area-under-curve (AUC). Briefly, AUC (ROC) is a plot of the true positive rate (i.e., how many PD patients were correctly predicted) *vs* the true negative rate (i.e., how many healthy controls were correctly predicted). In contrast, AUC (PR) plots the precision rate (i.e., how many PD predictions were correct) *vs* recall or sensitivity rate (i.e., how many PD patients were correctly predicted).

Annotation of chemical features

For each chemical feature in the metabolomics datasets, a SHAP score was calculated based on the absolute average from the 100 bootstraps. The greater the SHAP scores, the greater the contribution to the overall prediction of PD across all 100 bootstraps and all participants. Metabolites were annotated based on the highest mass accuracies that were obtained by comparing the measured monoisotopic neutral masses (accounting for protonation, sodiation and potential loss of a water molecule) to those from the Human Metabolome Database (https://hmdb.ca/) using a threshold of ± 20 ppm. One top scoring chemical feature (*m*/*z* 942.9824) had a relatively large negative mass defect which is indicative of an exogenous, synthetic compound. Thus, this ion was annotated using the PubChem (http://www.cheminfo.org/) database.

Results and Discussion

Highest diagnostic accuracy to date for PD using metabolomics: NN outperforms five other ML algorithms

The overall diagnostic performance for all six ML algorithms was assessed using a composite dataset featuring metabolites from blood plasma that were detected using four analytical methods as reported in the EPIC PD study.[13] The diagnostic performance of NN was higher than the other five ML algorithms across all metrics investigated (Figure 2, Table S2). Specifically, the binary classification of PD *vs* healthy using NN resulted in AUCs of 0.994 and 0.995 for ROC and PR, respectively. Extreme gradient boosting and logistic regression performed similarly with AUC (ROC) and AUC (PR) of 0.967 and 0.968 for extreme gradient boosting, and 0.968 and 0.969 for logistic regression, respectively. In contrast, the performance of the RF, SVM and LDA classifiers were relatively low with AUC (ROC) and AUC (PR) values of 0.829 and 0.836 for the RF classifier, 0.647 and 0.661 for the SVM classifier, and 0.681 and 0.634 for the LDA classifier, respectively. Recently, Chicco et al.[22] has shown that the MCC score is a more informative and reliable metric for evaluating binary classification accuracy as it considers all four values in the confusion matrix (i.e., true positive, false positive, true negative, and false negative). Thus, MCC can be considered as less biased towards datasets with imbalanced cohorts. Based on the MCC score, NN performed significantly higher with 0.918 compared to 0.815, 0.787, 0.433, 0.272, and 0.213 for the LR, XGB, RF, LDA, and SVM classifiers, respectively.

SHAP analysis was used to identify the metabolites and the corresponding mass spectrometry-based method that contributed the most to the accurate prediction of PD using the composite metabolomics dataset for blood plasma. Five of the top six metabolites were detected using LC-MS (+) (i.e., *m*/*z* 942.9824, 467.3822, 393.3454, 379.3289, and 613.4767) (Figure 2). To further validate the contribution of these chemical features in predicting PD, all six ML algorithms were applied to the LC-MS (+) dataset without including the LC-MS (-), GC-MS and CE-MS datasets. Similar to that for the composite dataset, binary classification of PD using NN and the LC-MS (+) data was highest across all performance metrics (Figure 3, Table S2). For example, the AUC (ROC), AUC (PR), and MCC score for NN was 0.983, 0.984, and 0.894, respectively. Both XGB and LR classifiers performed about the same or slightly lower with AUC (ROC), AUC (PR) and MCC values of 0.968, 0.966, 0.805, and 0.972, 0.976, 0.869, respectively. RF, SVM, and LDA classifiers were substantially lower with AUC (ROC), AUC (PR), and MCC scores of 0.894, 0.911, 0.589 for the RF classifier; 0.856, 0.869, 0.582 for the SVM classifier; and 0.878, 0.877, and 0.626 for the LDA classifier. Standard deviation errors across all performance metrics were pooled to calculate the average relative standard deviation (RSD). RSD

was lowest for NN (5.17%) followed by LR (5.19%), XGB (7.70%), LDA (12.8%), SVM (14.3%), and RF (17.2%). Based on a SHAP analysis, five of the six top-scoring chemical features in the LC-MS (+) dataset were also in the top six highest scoring features for the composite dataset (see above, Figure 3). These results indicate that the diagnostic performance for using the LC-MS (+) data outperformed the three other mass spectrometry-based methods across all six ML methods. The use of NN resulted in higher diagnostic accuracy than the other five ML methods for the LC-MS (+) dataset which has significantly fewer chemical features (509) than in the composite dataset (1430). Given these results and the low RSD of NN compared to other ML methods, any overtraining effects are minimal under these conditions.

Overall, NN resulted in the highest diagnostic performance and the lowest %RSD in predicting PD from blood plasma using either the composite or LC-MS (+) datasets compared to the other five ML algorithms. Disease classification using NN involving 100 bootstraps and 1430 total metabolites required < 1 min on a consumer laptop computer (Surface Laptop 3, Microsoft) with a 1.2 GHz processor (Intel i5-core). The diagnostic performance of NN was at least 10% higher than previously reported by Gonzalez-Riano et al when a support vector machine model was used.[13] In addition, the diagnostic performance obtained using NN is the highest reported to date for any PD diagnosis regardless of the sample matrix including blood plasma,[23, 24] blood serum,[25] and skin sebum.[10, 11, 15, 26, 27]

Diagnostic performance is uncompromised by including the whole metabolomics dataset

The NN model required no pre-processing steps as all metabolites or features from the dataset were used as inputs for the model rather than using models with pre-selected chemical features. For example, in Gonzalez-Riano et al.,[13] biomarkers for PD were first screened for significance and a small subset of these biomarkers (up to 20) were used in the final diagnostic model. In this previous study, the highest ROC (AUC) value obtained was 0.919 for the composite dataset using a 20-feature linear SVM model.[13] In the current study, a similar linear SVM model was applied to the composite dataset without feature selection which resulted in a ROC (AUC) of 0.647. In contrast, using NN on all features in the composite dataset resulted in a ROC (AUC) of 0.994. These results are consistent with classical ML models having relatively low predictive performance when incorporating large datasets that contain many 'noisy' features.

A feature-selected NN model was developed using the ten chemical features that had the highest SHAP scores using the LC-MS (+) dataset. The AUC ROC and PR values for the feature-selected NN-model were

comparable or slightly higher (0.997 ± 0.006 and 0.997 ± 0.006) than that obtained using all chemical features in the dataset (0.983 ± 0.022 and 0.984 ± 0.022). Given that the diagnostic performance of both models was comparable, these data indicate that NN can be highly tolerant of many chemical features (> 1500) that do not contribute substantially to accurate disease prediction; i.e., diagnostic performance is essentially uncompromised by including the whole metabolomics dataset without feature selection. In addition, the relatively high diagnostic accuracy further verifies the use of SHAP to accurately identify chemical features that contribute significantly to disease classification.

Revealing new metabolite biomarkers for PD by retrospective analysis

The metabolites that contributed the most significantly to the accurate prediction of PD are more likely to be more basic and readily ionised by cation adduction, rather than acidic, given that higher diagnostic performance for PD was obtained using LC-MS (+) that LC-MS (-) and that similar numbers of metabolites (~510 to 530) were measured using each method, consistent with previous reports.[23, 25] SHAP analysis on LC-MS (+) data for blood plasma revealed that five of the top six highest scoring metabolites were consistent across all six ML algorithms (Figure 3). The detected metabolites were different compared to those determined using a linear-based classical ML model[13] (Figure S2), which can fundamentally be attributed to the difference between projecting data in a linear *vs* non-linear space.

The five metabolites that contributed the most to a PD prediction could serve as potential indicators for disease status and were annotated (Table 2). The five annotated ions corresponded to polyfluoroalkyl substance (PFAS), triterpenoids, cholestane steroids, diacylglycerol and vitamin D steroids of either endogenous or exogenous origins, which have all been linked to PD in the literature previously (Table 2). For example, the ion with an *m*/*z* value of 942.9824 had the highest SHAP value and likely corresponds to the sodiated PFAS [3-(2,2,3,3,4,4,4-heptafluorobutanoyloxy)-2,2-bis(2,2,3,3,4,4,4-heptafluorobutanoyloxymethyl)propyl] 2,2,3,3,4,4,4-heptafluorobutanoate (DTXSID70325550). Ions corresponding to this PFAS and its oxidation ([3-(2,2,3,3,4,4,4-heptafluorobutanoyloxy)-2,2-bis(2,2,3,3,4,4,4-heptafluorobutanoyloxymethyl)-1-hydroxypropyl] 2,2,3,3,4,4,4-heptafluorobutanoate) and hydrolysis (i.e., sodiated 2,2-bis(hydroxymethyl)propane-1,3-diol) products were all higher in PD participants than health controls. The presence of PFAS compounds are ubiquitous in the environment and human blood given their propensity to bioaccumulate, chemical longevity and widespread use in industrial and consumer

products such as plastics, non-stick cookware, and food packaging.[28, 29] For example, in the U.S. population, PFAS was detected in the blood serum of over 98% of Americans that were sampled during 2003-2004 (n = 2,094).[30] DTXSID70325550 is a PFAS compound of interest that is currently listed under the U.S. Environmental Protection Agency CompTox Chemicals Database,[31] and appears pre-organised for the non-covalent, multi-dentate binding of Na^+, K^+, Ca^{2+}, Cu^{2+} and Zn^{2+}. Thus, such a compound could potentially disrupt neuronal activity by affecting intracellular ion homeostasis.[32, 33] A potential mechanism proposed for PFAS-induced neurotoxicity involves the increase of intracellular Ca^{2+} which is implicated in impacting neuronal cell processing, signalling, and function.[28, 33] Although further *in vitro* and *in vivo* studies are needed to investigate the effects of DTXSID70325550 on neuronal cell function, these data suggest that elevated levels of specific PFAS compounds in blood plasma may be an early indicator of PD. Overall, these results further support that SHAP analysis can be useful in identifying potential biomarkers for PD (Table 2) that were not initially found using classical statistical approaches.[13]

NN resulted in higher performance for diagnosing PD than five alternative ML methods using a larger metabolomics dataset from sebum samples

The performance of CRANK-MS was assessed on a larger sample size with more chemical features. The six ML algorithms were used to analyse the data from the NHS study[15] to predict PD patients that were drug naïve or medicated from healthy controls using LC-MS (+) metabolomics data from skin swab samples of sebum. Across all performance metrics, NN performed comparatively better on average compared to the five other ML algorithms (Figure 5, Table S2). Binary classification of drug-naïve PD *vs* healthy control using NN resulted in AUC (ROC), AUC (PR), MCC, and $F1$ scores of 0.843, 0.896, 0.530, and 0.800 respectively. In addition, average %RSD was lowest for NN at 9.51% followed by SVM (9.63%), LR and LDA (10.0%), XGB (11.7%), and RF classifiers (14.0%). For the medicated PD *vs* healthy control dataset, XGB and NN performing better overall compared to the other ML algorithms (Table S2).

The diagnostic performance obtained using NN on skin sebum was more than 8% higher than previously reported by Sinclair et al.[15] which used a multivariate principle least squares-discriminant analysis model based on 15 pre-selected features. Specifically, the AUC (ROC) using multivariate principle least squares-discriminant analysis was 0.779 compared to 0.843 using NN.[15] The NN approach required no pre-selection of features and took 10 min to obtain the key diagnostic performance metrics (i.e., accuracy,

sensitivity, specificity, precision, *F*1 score, MCC score, AUC ROC, AUC PR) using all 6,502 chemical features in the dataset.

Conclusions

In this study, a 'lightweight' and optimised NN framework with interpretable feature analysis, entitled CRANK-MS, is reported that can be used to establish accurate disease prediction models using whole MS datasets without pre-selecting features. The NN is highly tolerant of 'noisy' metabolomics data that can contain thousands of metabolites which do not contribute significantly to model prediction. Using CRANK-MS, we report the highest diagnostic performance to date for predicting PD using blood plasma metabolomics data (> 0.995 AUC) when benchmarked with classical ML algorithms. Diagnostic accuracy in predicting PD using skin sebum metabolomics data was also enhanced using NN compared to alternative, widely used ML approaches. PD-specific biomarkers in blood plasma were identified including triterpenoids, diacylglycerols and a polyfluoroalkyl substance that contribute significantly to ML model predictions and are potential early indicators for PD. These results are consistent with specific food diets (such as the Mediterranean diet[34]) for PD prevention and that exposure to some exogenous chemicals (such as PFASs that can disrupt neuronal activity via changes to intracellular ion homeostasis[32, 33]) may contribute to the development of PD.

Given the improved diagnostic performance of CRANK-MS, it is anticipated that this neural network-based framework can be a powerful tool to build accurate prediction models for other diseases using metabolomics data. Interpretable ML methods can also be used to retrospectively 'mine' metabolomics datasets to identify early 'lead' compounds within the biomarker discovery pipeline. Biomarkers identified using this approach can be further validated using high-resolution tandem mass spectrometry and NMR for complete structure elucidation and targeted quantification using clinical MS-based methods, in addition to using *in-vitro* cell-based assays and *in-vivo* disease models. For example, there are over 800 publicly available metabolomics studies in the Metabolomics Workbench data repository in which prediction models could be used for the binary classification of diseases such as diabetes, fatty liver disease, heart disease, chronic obstructive pulmonary disease, and COVID-19. Using advanced ML methods, retrospectively 'mining' such databases for biomarkers that contribute significantly to the prediction of these diseases could reveal novel mechanistic information that may not necessarily be apparent using traditional linear approaches. In addition,

CRANK-MS could be adapted to support clinical workflows and improve confidence in diagnosis particularly where disease stratification is important. For example, CRANK-MS can be used with metabolomics data in conjunction with alternative 2-dimensional clinical information such as medical history and neurological examination scores, as well as 3-dimensional brain structural imaging scans to further differentiate clinical PD from other types of Parkinsonian-like diseases. Ultimately, the use of CRANK-MS should enhance the accuracy of disease prediction models based on metabolomics and many other types of '-omics' experiments, and facilitate biomarker discovery.

Acknowledgements

JDZ completed this research while undertaking a Fulbright Future Scholarship funded by the Kinghorn Foundation. WAD acknowledges funding from the Australian Research Council (FT200100798). VBK acknowledges support from the Karen Toffler Charitable Trust, a Strategically Focused Research Network (SFRN) Center Grant (20SFRN35460031) from the American Heart Association, a subaward (32307-93) from the NIDDK Diabetic Complications Consortium grant (U24-DK115255), and a Hariri Research Award from the Hariri Institute for Computing and Computational Science & Engineering at Boston University, and NIH grants (R01-AG062109, R01-HL159620, R21-CA253498, and R43-DK134273). Additional support was provided by Boston University's Affinity Research Collaboratives program, and the Boston University Alzheimer's Disease Center (P30-AG013846). We thank Dr. Shangran Qiu (Boston University), and Professors Jonathon E. Beves and Thanh Vinh Nguyen (UNSW Sydney) for helpful discussions.

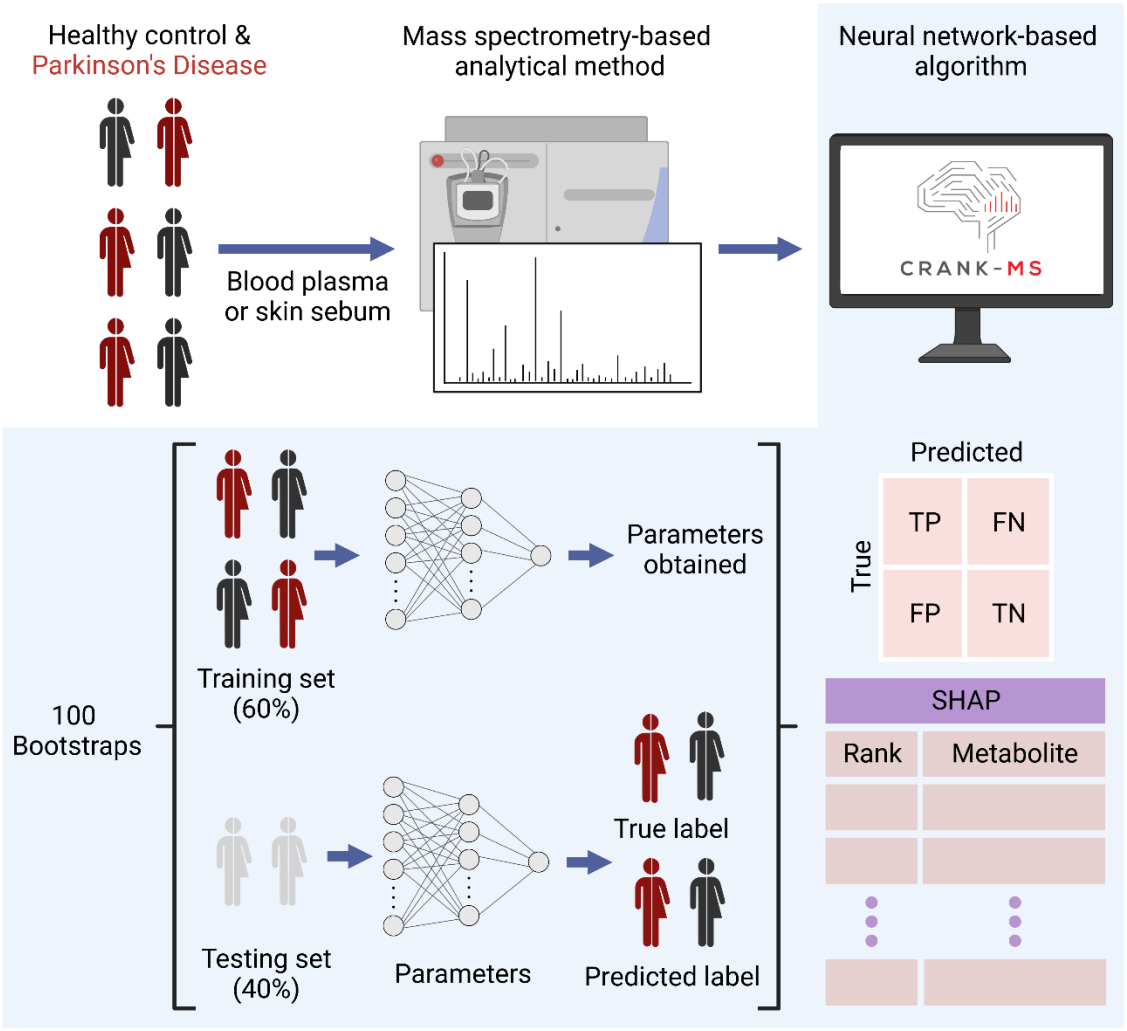

Figure 1. Neural network (NN) framework for predicting Parkinson's disease using large mass spectrometry-based metabolomics data. Whole metabolomics datasets without feature selection can be analysed directly by NN for the binary classification of Parkinson's disease. Using a 100-iteration bootstrap model, 60% of the data was randomly distributed for training and 40% for testing. Diagnostic performance for each bootstrap was calculated based on the absolute values obtained for true positive (TP), false negative (FN), false positive (FP), and true negative (TN). Shapely additive explanations (SHAP) analysis of NN was used to rank chemical features based on the extent of their contribution to a positive Parkinson's disease prediction.

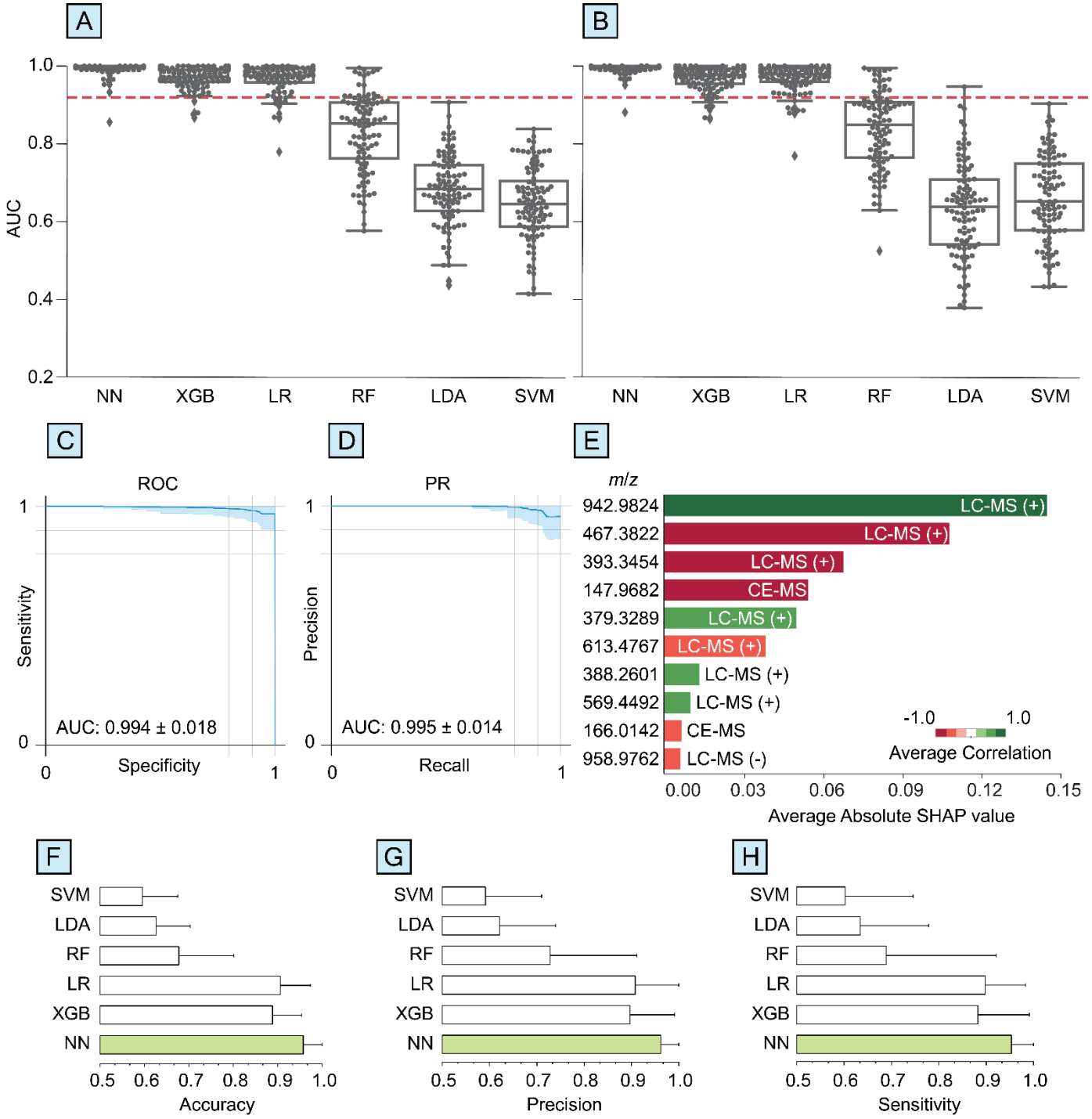

Figure 2. The use of neural networks (NN) can outperform other machine learning algorithms for the early diagnosis of Parkinson's disease using blood plasma metabolomics data. A composite dataset from the EPIC PD study was used involving metabolomics data from liquid chromatography-mass spectrometry (LC-MS) in positive and negative ionisation modes, capillary electrophoresis-mass spectrometry (CE-MS), and gas chromatography-MS (GC-MS) without any feature selection.[13] Box-swarm plots of area under curves (AUC) of (A) receiver-operating curve (ROC) and (B) precision-recall (PR) for NN, extreme gradient boosting (XGB), logistic regression (LR), random forest (RF), linear discriminant analysis (LDA), and support vector machine (SVM) classifiers. The red line corresponds to the AUC previously reported[13] based on a feature-selected classical ML model. Overall (C) ROC and (D) PR plots are shown for NN. (E) Shapely additive explanations (SHAP) values for NN for the top ions (*m*/*z*) that had the highest contribution to a positive PD prediction and the corresponding analytical method. The average correlation corresponds to whether the feature is up- (green) or down- (red) regulated. For each algorithm, (F) accuracy, (G) precision, and (H) sensitivity are shown (green bars correspond to the highest performance).

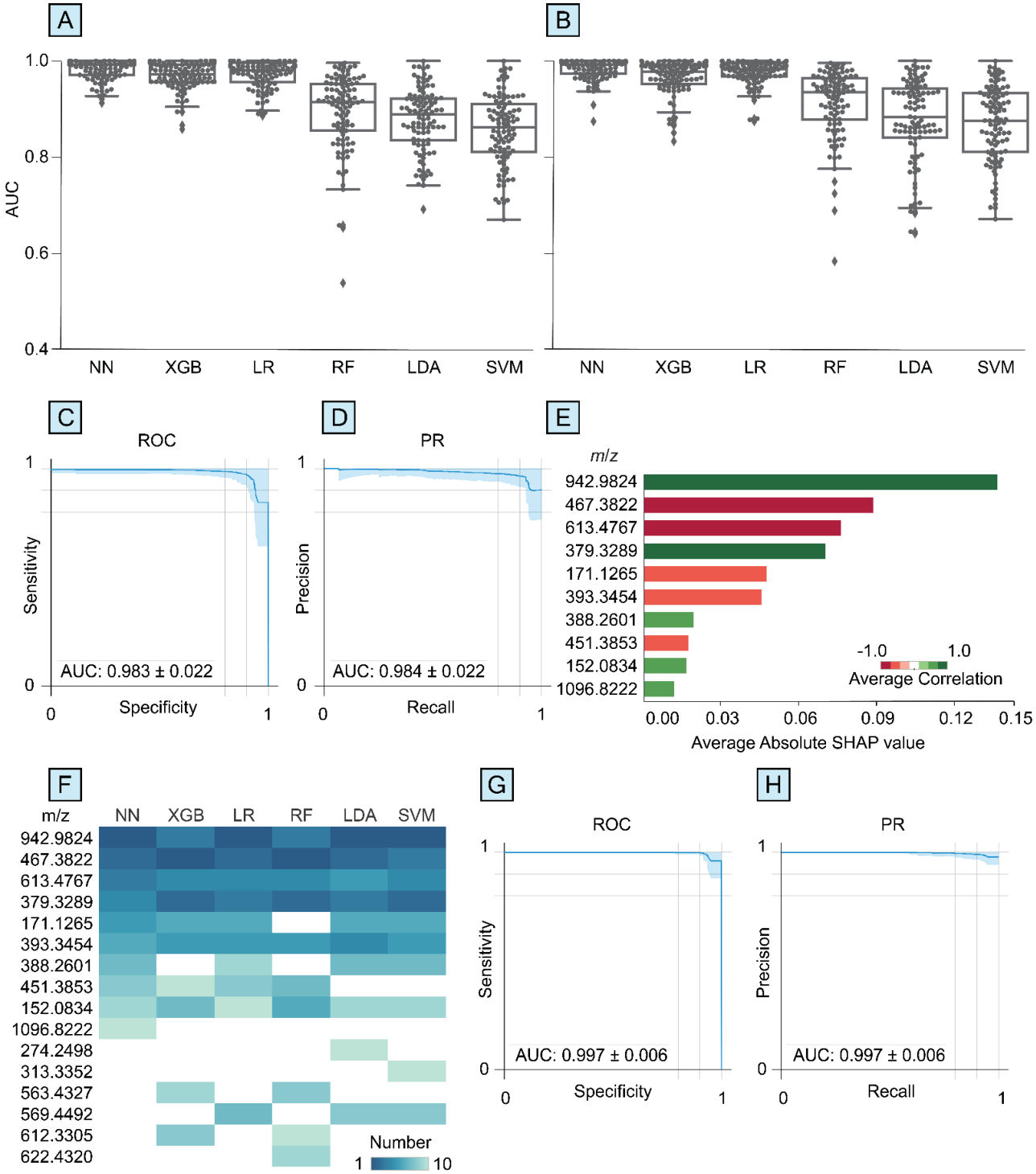

Figure 3. The use of neural networks (NN) resulted in the highest performance for diagnosing Parkinson's disease from blood plasma metabolomics data (EPIC PD study[13]) obtained by liquid chromatography-mass spectrometry in positive mode (LC-MS (+)) without any pre-selection of chemical features. Box-swarm plots of area under curves (AUC) for (A) receiver-operating curve (ROC) and (B) precision-recall (PR) for NN, extreme gradient boosting (XGB), logistic regression (LR), random forest (RF), linear discriminant analysis (LDA), and support vector machine (SVM) classifiers. Overall (C) ROC and (D) PR plots are shown for NN. (E) Shapely additive explanations (SHAP) values for NN shown for the top ten ions (*m/z*) using LC-MS (+) that had the highest contribution to a positive PD prediction. The average correlation corresponds to whether the feature is up- (green) or down- (red) regulated. (F) Comparative SHAP values and relative rankings for the top ten metabolites for all six ML algorithms. (G) ROC and (H) PR plots are shown for a feature-selected NN-model using the top ten metabolites identified from SHAP.

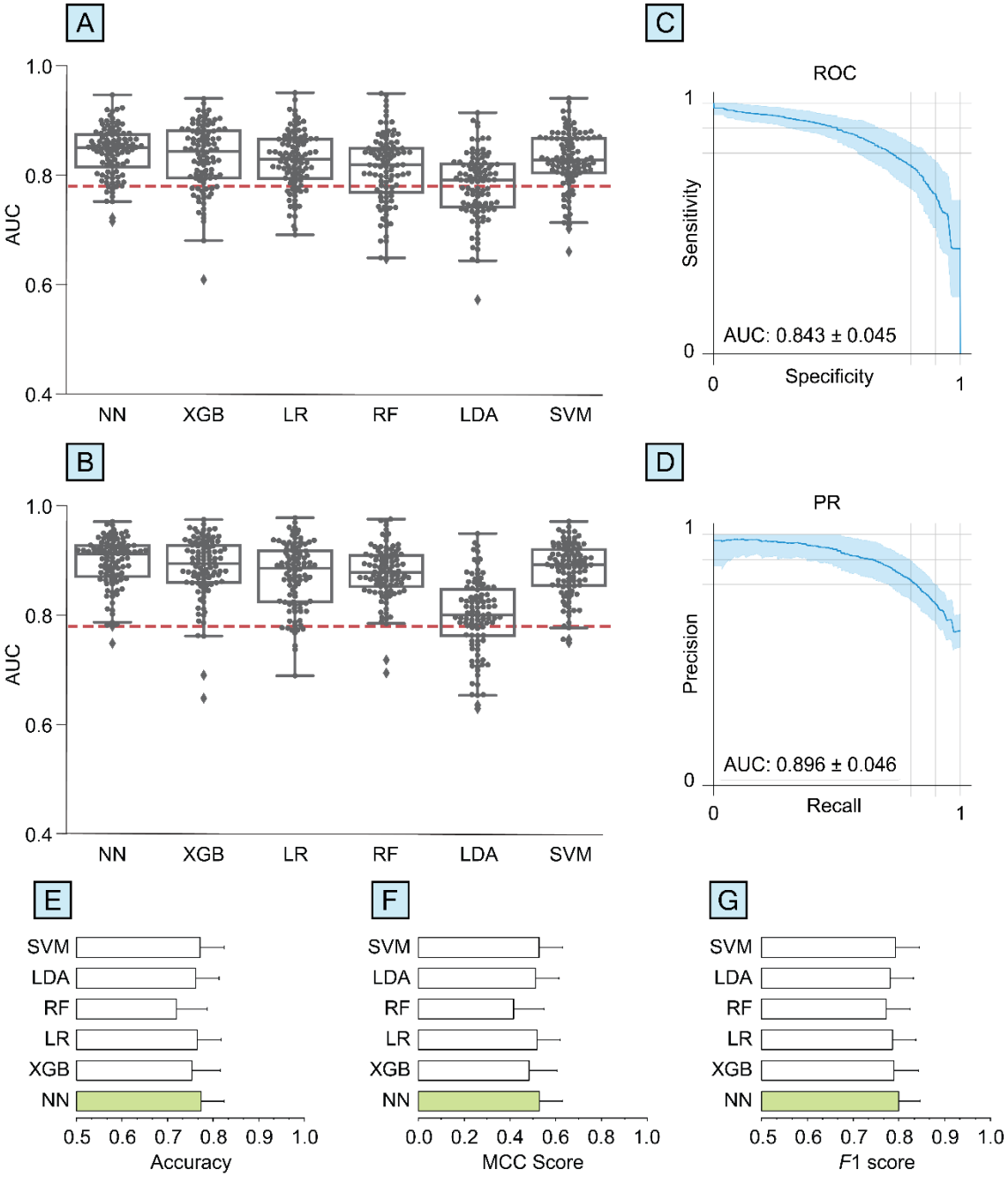

Figure 4. The use of neural networks (NN) can result in higher overall diagnostic performance for drug-naïve PD *vs* healthy control from metabolomics data (NHS PD study[15]) of skin sebum samples without any selection of > 6500 chemical features. Box-swarm plots of area under curves (AUC) of (A) receiver-operating curve (ROC) and (B) precision-recall (PR) for NN, extreme gradient boosting (XGB), logistic regression (LR), random forest (RF), linear discriminant analysis (LDA), and support vector machine (SVM) classifiers. The red line corresponds to the AUC previously reported[15] based on a feature-selected classical ML model. ROC and PR plots are shown for NN in (C) and (D), respectively. For each algorithm, (E) accuracy, (F) MCC score, and (G) *F*1 score are shown (green bars correspond to the highest performance).

Table 1. Summary of demographic and chemical feature information for each metabolomics dataset

Cohort	Dataset	N = Control	N = PD	Number of features	Sample type
EPIC[13]	GC-MS	39	36	60	Plasma
	CE-MS	39	39	329	Plasma
	LC-MS (+)	39	39	509	Plasma
	LC-MS (-)	34	39	532	Plasma
	Composite (GC-MS, CE, LC-MS (+), LC-MS (-)	37	35	1430	Plasma
NHS[15]	LC-MS (+)	56	80[a]	6502	Sebum
	LC-MS (+)	56	138[b]	6502	Sebum

[a] Healthy control *vs* drug-naïve PD.

[b] Healthy control *vs* medicated PD.

Table 2. Summary of five annotated metabolites that contributed most to a Parkinson's disease prediction.

m/z	Compound class	Annotation	Chemical formula	Ion	Up/Down regulated	*p*-value	Link to PD
942.9824	Polyfluorinated alkyl substance	[3-(2,2,3,3,4,4,4-heptafluorobutanoyloxy)-2,2-bis(2,2,3,3,4,4,4-heptafluorobutanoyloxymethyl)propyl] 2,2,3,3,4,4,4-heptafluorobutanoate	$C_{21}H_8F_{28}O_8$	$[M+Na]^+$	Up	5.4×10^{-11}	A proposed mechanism for PFAS-induced neurotoxicity involves the increase of intracellular Ca^{2+} which is implicated in impacting neuronal cell processing, signalling, and function.[28, 33] Non-covalent binding of metal ions by this PFAS could disrupt neuronal activity by affecting intracellular ion homeostasis.[32, 33]
467.3822	Triterpenoid	Dammarenediol II [a]	$C_{30}H_{52}O_2$	$[M+Na]^+$	Down	3.1×10^{-09}	Triterpenoids have been linked to the activation of the nuclear factor-E2-related factor-2 (Nrf2)/antioxidant response element (ARE) signalling pathway which regulates oxidative stress.[35-37] Oxidative stress is a leading factor in the pathogenesis of PD which includes dopaminergic cell death, mitochondrial dysfunction, and inflammation.[38] Triterpenoids can be consumed through food sources including apple, olive, tomato, and soybean.[36, 39, 40]

613.4767	Diacylglycerol	1,2-diacylglycerol (34:3) isomers [b]	$C_{37}H_{66}O_5$	$[M+Na]^+$	Down	3.2×10^{-06}	Diacylglycerols are naturally found in vegetable oils such as olive oil,[41] where consumption of unsaturated lipids is an important component in a Mediterranean diet.[34] A recent study by Barbalace et al.[42] reported that extra virgin olive oil extract can significantly increase the brain-derived neurotrophic factor, which is a key signalling pathway for neuronal survival, regulation, and regeneration.[43, 44]
379.3289	Steroid	Vitamin D2 [c]	$C_{28}H_{44}O$	$[M+H-H_2O]^+$	Up	4.5×10^{-07}	The presence of Vitamin D has previously been implicated as biomarkers in PD.[13, 45]
393.3454	Cholestane steroid	Cholest-5-ene	$C_{27}H_{46}$	$[M+Na]^+$	Down	1.1×10^{-06}	Cholestane derivatives have been shown to have neuroprotectant properties.[46] For example, using animal models, Hu et al.[47] reported that an endogenous cholestane derivative could directly block NMDA receptors where overactivation of these receptors are typically observed in PD.[48]

[a] See Table S3 for other potential triterpenoid isomers listed in the HMDB that have not been detected in blood, unlike Dammarenediol II.

[b] 1,3-diacylglycerol (34:3) isomers were also listed in the HMDB.

[c] Vitamin D2 agrees with the annotation by Gonzalez-Riano et al.[13] based on the EPIC cohort. See Table S3 for other potential steroid isomers listed in the HMDB.

References

1. Feigin, V. L., et al., Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Neurology* **2017,** *16* (11), 877-897.
2. Armstrong, M. J., et al., Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA* **2020,** *323* (6), 548-560.
3. Chaudhuri, K. R., et al., Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment. *The Lancet Neurology* **2009,** *8* (5), 464-474.
4. Poewe, W., et al., Parkinson disease. *Nature Reviews Disease Primers* **2017,** *3* (1), 17013.
5. Hawkes, C. H., et al., A timeline for Parkinson's disease. *Parkinsonism & Related Disorders* **2010,** *16* (2), 79-84.
6. Rizzo, G., et al., Accuracy of clinical diagnosis of Parkinson disease. *Neurology* **2016,** *86* (6), 566.
7. Gowda, G. A. N., et al., Metabolomics-based methods for early disease diagnostics. *Expert Review of Molecular Diagnostics* **2008,** *8* (5), 617-633.
8. Griffiths, W. J., et al., Targeted Metabolomics for Biomarker Discovery. *Angewandte Chemie International Edition* **2010,** *49* (32), 5426-5445.
9. Dunn, W. B., et al., Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews* **2011,** *40* (1), 387-426.
10. Sinclair, E., et al., Validating Differential Volatilome Profiles in Parkinson's Disease. *ACS Central Science* **2021,** *7* (2), 300-306.
11. Trivedi, D. K., et al., Discovery of Volatile Biomarkers of Parkinson's Disease from Sebum. *ACS Central Science* **2019,** *5* (4), 599-606.
12. Hanna, G. B., et al., Accuracy and Methodologic Challenges of Volatile Organic Compound–Based Exhaled Breath Tests for Cancer Diagnosis: A Systematic Review and Meta-analysis. *JAMA Oncology* **2019,** *5* (1), e182815-e182815.
13. Gonzalez-Riano, C., et al., Prognostic biomarkers of Parkinson's disease in the Spanish EPIC cohort: a multiplatform metabolomics approach. *npj Parkinson's Disease* **2021,** *7* (1), 73.
14. Worley, B., et al., Multivariate Analysis in Metabolomics. *Current Metabolomics* **2013,** *1* (1), 92-107.
15. Sinclair, E., et al., Metabolomics of sebum reveals lipid dysregulation in Parkinson's disease. *Nature Communications* **2021,** *12* (1), 1592.
16. Gardner, M. W., et al., Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* **1998,** *32* (14), 2627-2636.
17. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019,** *1* (5), 206-215.
18. Watson, D. S., et al., Clinical applications of machine learning algorithms: beyond the black box. *BMJ* **2019,** *364*, l886.
19. Qiu, S., et al., Multimodal deep learning for Alzheimer's disease dementia assessment. *Nature Communications* **2022,** *13* (1), 3404.
20. Smith, G. C. S., et al., Correcting for Optimistic Prediction in Small Data Sets. *American Journal of Epidemiology* **2014,** *180* (3), 318-324.
21. Molinaro, A. M., et al., Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **2005,** *21* (15), 3301-3307.
22. Chicco, D., et al., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020,** *21* (1), 6.
23. Zhao, H., et al., Potential biomarkers of Parkinson's disease revealed by plasma metabolic profiling. *Journal of Chromatography B* **2018,** *1081-1082*, 101-108.
24. Shao, Y., et al., Comprehensive metabolic profiling of Parkinson's disease by liquid chromatography-mass spectrometry. *Molecular Neurodegeneration* **2021,** *16* (1), 4.
25. Pereira, P. A. B., et al., Multiomics implicate gut microbiota in altered lipid and energy metabolism in Parkinson's disease. *npj Parkinson's Disease* **2022,** *8* (1), 39.
26. Uehara, Y., et al., Non-invasive diagnostic tool for Parkinson's disease by sebum RNA profile with machine learning. *Scientific Reports* **2021,** *11* (1), 18550.

27. Fu, W., et al., Artificial Intelligent Olfactory System for the Diagnosis of Parkinson's Disease. *ACS Omega* **2022,** *7* (5), 4001-4010.
28. Cao, Y., et al., Absorption, distribution, and toxicity of per- and polyfluoroalkyl substances (PFAS) in the brain: a review. *Environmental Science: Processes & Impacts* **2021,** *23* (11), 1623-1640.
29. Foguth, R., et al., Per- and Polyfluoroalkyl Substances (PFAS) Neurotoxicity in Sentinel and Non-Traditional Laboratory Model Systems: Potential Utility in Predicting Adverse Outcomes in Human Health. *Toxics* **2020,** *8* (2), 42.
30. Calafat Antonia, M., et al., Polyfluoroalkyl Chemicals in the U.S. Population: Data from the National Health and Nutrition Examination Survey (NHANES) 2003–2004 and Comparisons with NHANES 1999–2000. *Environmental Health Perspectives* **2007,** *115* (11), 1596-1602.
31. US Environmental Protection Agency: PFAS structures in DSSTox. https://comptox.epa.gov/dashboard/chemical-lists/PFASSTRUCTV5 (accessed November 2, 2022).
32. Harik, S. I., Blood--brain barrier sodium/potassium pump: modulation by central noradrenergic innervation. *Proceedings of the National Academy of Sciences* **1986,** *83* (11), 4067-4070.
33. Starnes, H. M., et al., A Critical Review and Meta-Analysis of Impacts of Per- and Polyfluorinated Substances on the Brain and Behavior. *Frontiers in Toxicology* **2022,** *4*, 881584.
34. Trichopoulou, A., et al., Modified Mediterranean diet and survival: EPIC-elderly prospective cohort study. *BMJ* **2005,** *330* (7498), 991.
35. Yang, L., et al., Neuroprotective Effects of the Triterpenoid, CDDO Methyl Amide, a Potent Inducer of Nrf2-Mediated Transcription. *PLOS ONE* **2009,** *4* (6), e5757.
36. Szakiel, A., et al., Fruit cuticular waxes as a source of biologically active triterpenoids. *Phytochemistry Reviews* **2012,** *11* (2), 263-284.
37. Ma, Q., Role of Nrf2 in oxidative stress and toxicity. *Annual Review of Pharmacology and Toxicology* **2013,** *53*, 401-426.
38. Jenner, P., Oxidative stress in Parkinson's disease. *Annals of Neurology* **2003,** *53* (S3), S26-S38.
39. Ma, C. M., et al., The cytotoxic activity of ursolic acid derivatives. *European Journal of Medicinal Chemistry* **2005,** *40* (6), 582-589.
40. Krishnamurthy, P., et al., High-Throughput Screening and Characterization of a High-Density Soybean Mutant Library Elucidate the Biosynthesis Pathway of Triterpenoid Saponins. *Plant and Cell Physiology* **2019,** *60* (5), 1082-1097.
41. Lee, Y.-Y., et al., Production, safety, health effects and applications of diacylglycerol functional oil in food systems: a review. *Critical Reviews in Food Science and Nutrition* **2020,** *60* (15), 2509-2525.
42. Barbalace, M. C., et al., Antioxidant and Neuroprotective Activity of Extra Virgin Olive Oil Extracts Obtained from Quercetano Cultivar Trees Grown in Different Areas of the Tuscany Region (Italy). *Antioxidants* **2021,** *10* (3), 421.
43. Jiang, L., et al., Serum level of brain-derived neurotrophic factor in Parkinson's disease: a meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **2019,** *88*, 168-174.
44. Chmielarz, P., et al., Neurotrophic factors for disease-modifying treatments of Parkinson's disease: gaps between basic science and clinical studies. *Pharmacological Reports* **2020,** *72* (5), 1195-1217.
45. Evatt, M. L., et al., Prevalence of Vitamin D Insufficiency in Patients With Parkinson Disease and Alzheimer Disease. *Archives of Neurology* **2008,** *65* (10), 1348-1352.
46. Bansal, R., et al., Exploring the potential of natural and synthetic neuroprotective steroids against neurodegenerative disorders: A literature review. *Medicinal Research Reviews* **2018,** *38* (4), 1126-1158.
47. Hu, H., et al., The major cholesterol metabolite cholestane-3β,5α,6β-triol functions as an endogenous neuroprotectant. *The Journal of Neuroscience* **2014,** *34* (34), 11426-11438.
48. Yan, M., et al., Characterization of a Synthetic Steroid 24-keto-cholest-5-en-3β, 19-diol as a Neuroprotectant. *CNS Neuroscience & Therapeutics* **2015,** *21* (6), 486-495.