M2 Bio-Informatique 2025-2026

Rapport du Projet Kaggle :
Parkinson's Disease Progression Prediction

DESVILLES Roxane, DUFOUR Laura, IANCU Nicoleta et NABI Appoline

I. **Introduction** :

La maladie de Parkinson est une pathologie neurodégénérative chronique, caractérisée par une dégradation progressive des neurones dopaminergiques. À ce jour, aucun traitement curatif n'existe, ce qui rend cruciale la capacité à prédire son évolution pour adapter au mieux la prise en charge des patients. Dans ce contexte, ce projet vise à développer un modèle prédictif robuste permettant d'anticiper l'évolution de la maladie, pour un mois de visite donné, à partir de données protéomiques longitudinales.

Dans un premier temps, un travail de prétraitement et d'analyse exploratoire a été réalisé pour mieux comprendre les données protéomiques. Cela a permis de nettoyer les valeurs manquantes, de structurer les variables, et d'examiner les tendances globales entre patients. Suite à ce traitement, deux méthodes d'imputation des données ont été testées en parallèle afin d'être comparées. Des techniques comme l'ACP et UMAP ont été utilisées pour visualiser les profils biologiques et repérer d'éventuels regroupements ou trajectoires temporelles. Par la suite, un modèle de type Multi-Layer Perceptron (MLP), et un de type Random Forest ont été mis en place, pour prédire l'évolution clinique des patients. Ces modèles sont respectivement connus pour leur capacité à modéliser des relations non linéaires entre variables, et à faire face au bruit en capturant des interactions complexes entre variables. L'objectif est de proposer un outil fiable et interprétable, capable d'accompagner la stratification des patients et d'anticiper les formes évolutives de la maladie.

II. **Matériel et méthodes** :

A) **Traitement des fichiers bruts** :

Lors de ce projet, quatre fichiers fournis par le défi Kaggle 'Parkinson's Progression Prediction' ont été utilisés : *train_peptides.csv*, *train_proteins.csv, train_clinical_data.csv* et *supplemental_clinical_data.csv*.

Dans un premier temps, les deux jeux cliniques (*train_clinical_data.csv* et *supplemental_clinical_data.csv*) ont été concaténés verticalement afin de réunir l'ensemble des scores UPDRS disponibles. Puis le tableau à été trié par *patient_id* et *visit_month* pour une meilleur lisibilité. Une vérification d'intégrité a confirmé l'absence de doublons sur la clé de visite (*visit_id*), garantissant l'unicité des enregistrements cliniques par passage. Puis, les tables omiques "longues" (*train_proteins.csv* et *train_peptides.csv*) ont été fusionnées (jointure interne) à la table clinique concaténée sur la clé composite ["*patient_id*", "*visit_id*", "*visit_month*"]. Ce choix assure que chaque mesure d'abondance d'une protéine (NPX) est associée aux scores UPDRS et à l'état médicamenteux correspondant à la même visite. Deux tables labellisées ont ainsi été obtenues et exportées : *train_proteins_labeled.csv* (données protéiques + variables cliniques) et *train_peptides_labeled.csv* (données peptidiques + variables cliniques). Pour améliorer la lisibilité et garantir un ordre reproductible, les deux tables ont été triées par *patient_id*, *visit_month* et *UniProt*. Pour finir, *train_proteins_labeled*, a été pivoté afin d'avoir la correspondance *patient_id* x *visit_month* en ligne et les valeurs NPX des protéines d'UniProt en colonnes. La colonne visit_id à ensuite été supprimée étant donné sa redondante avec les informations présentes en ligne.

On obtient ainsi une table d'expression protéique, par patient et mois de visite, à laquelle est intégré les quatre scores cliniques UPDRS correspondants, en conditions "on/off medication". Les analyses ultérieures, comprenant une visualisation des données suivies d'une imputation hiérarchique, d'une transformation logarithmique et d'un centrage-réduction, décrites dans les sections suivantes, seront appliquées sur la table ainsi obtenue.

B) **Données Patients :**

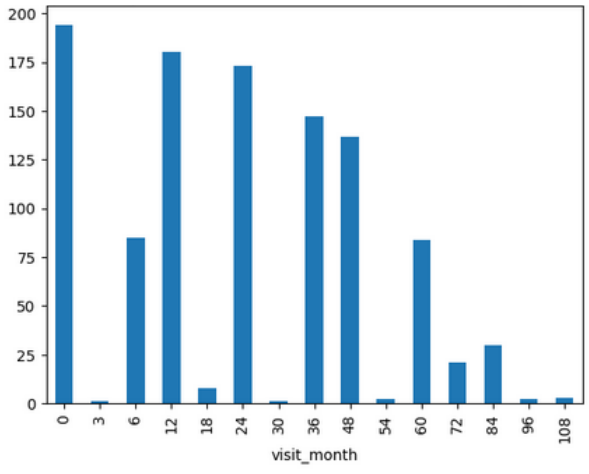


Figure 1 *: Répartition des patients par mois de visite*

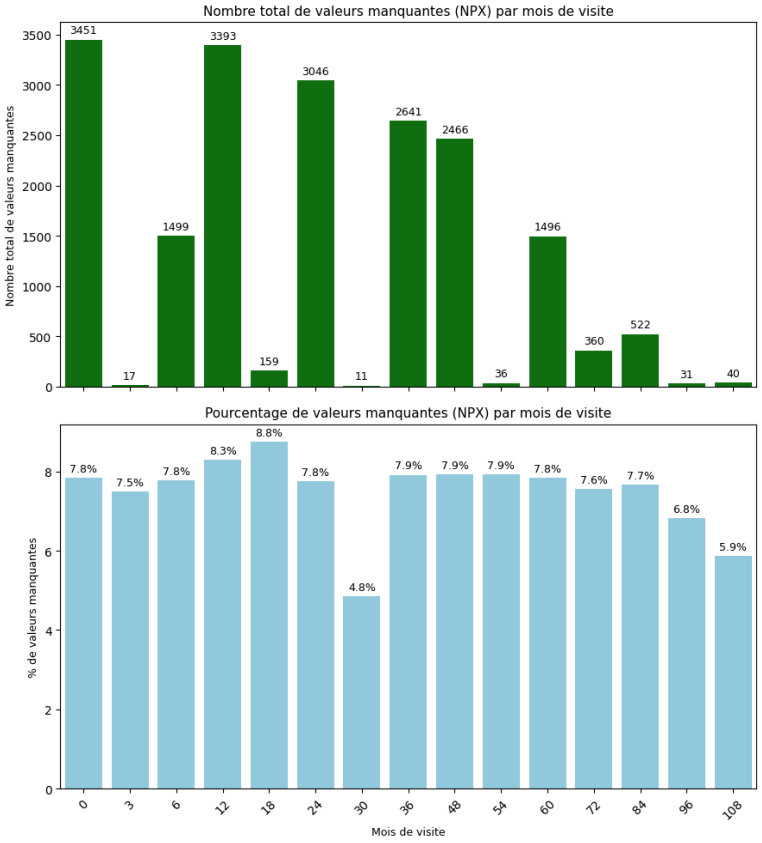
Notre jeu de données traité contient ainsi 248 patients, ayant effectué en moyenne 4 visites sur la période de 108 mois correspondant à l'étude. Ces patients incluent ceux atteints, mais également des individus sains non étiquetés.

La figure 1 met en évidence une répartition des patients très hétérogène au cours des mois de visite, qui pourrait être corrélé à la présence des individus sains.

C) Données protéiques :

Les protéines ont été quantifiées par Olink® Proximity Extension Assay (PEA) couplée à la spectrométrie de masse, permettant une mesure multiplexe sensible sur de faibles volumes d'échantillon. La matrice d'analyse utilisée après transformation des fichiers bruts comprenait 227 protéines mesurées chez 248 patients, chaque ligne représentant les dosages normalisés (NPX) pour une visite par patient. Cette structure permet l'analyse de trajectoires intra-individus d'abondance protéique au cours du suivi de l'étude de 108 mois.

a) Analyse exploratoire de la complétude des données protéiques :



L'exploration descriptive a mis en évidence une proportion modérée et relativement stable de valeurs manquantes dans les mesures NPX protéiques (Figure 2). En effet, la part de données manquantes est comprise entre 4,8 % à 8,8 % selon le mois de visite, avec des maximums observés à 12 et 18 mois, et un minimum à 30 mois, ce qui révèle une stabilité temporelle générale. En effectifs absolus, les visites précoces concentrent logiquement davantage de valeurs manquantes du fait du nombre d'observations plus élevé. Une observation plus détaillée des données révèle cependant l'existence d'un petit sous-ensemble de protéine fortement lacunaire (jusqu'à ≈ 58 % de manquants). La stabilité temporelle des taux de manquants et la limitation de ce sous-ensemble nous a poussé à conserver ces protéines afin d'éviter une perte d'information potentiellement pertinente sur des sous-phénotypes biologiques. Cependant ces valeurs manquantes seront prises en compte grâce à une imputation simple et hiérarchisée, qui exploite d'abord l'information longitudinale (intra-patient), puis l'information transversale (inter-patients à un temps donné).

***Figure 2** : Représentation du nombre de valeurs manquantes par mois de visite, en occurrence (en haut - en vert) et en pourcentage (en bas - en bleu).*

b) Procédure d'imputation des données :

Les 19 168 valeurs manquantes présentes dans notre jeu de données restructuré ont été imputées en totalité selon deux méthodes plus ou moins complexes et réalistes d'un point de vue temporel. D'une part, une imputation hiérarchique complexe en trois étapes à été réalisée, afin de respecter la structure longitudinale et inter-patients:

- **Interpolation intra-patient linéaire (1ère étape)** :

Pour chaque patient, lorsque les valeurs manquantes étaient encadrées par deux visites observées, une imputation par interpolation linéaire suivant la chronologie (*visit_month*) à été réalisée, afin de préserver la continuité des trajectoires individuelles.

- **Médiane inter-patients au même temps de visite (2e étape)** :

Les valeurs restant manquantes ont été remplacées, protéine par protéine, par la médiane calculée parmi les patients disposant d'une mesure valide au même mois de visite. Cette étape garantit une cohérence « verticale » entre individus, au temps *t*.

- **Médiane globale par protéine (3e étape) :**

Les rares valeurs résiduelles ont été imputées par la médiane globale de la protéine (toutes visites et patients confondus), de sorte à éliminer tout NaN résiduel.

D'autre part, une imputation uniquement sur la médiane globale par protéine (3e étape) à été réalisée, afin de déterminer l'apport d'une imputation plus complexe qui tient compte d'une structure longitudinale au cours du temps, sur la prédiction des scores UPDRS.

c) Transformation logarithmique et normalisation des données protéiques

Afin d'harmoniser l'échelle des variables et de stabiliser la variance avant les analyses multivariées, nous avons appliqué une transformation logarithmique suivie d'un centrage-réduction à la matrice imputée.

- Transformation logarithmique

Sur la matrice des protéines imputées, noté X, appartenant à R^(n×p), avec n observations et p=227 protéines, nous appliquons une transformation logarithmique de type log1p :

$$X_{ij}^{(\log)} = \log(1 + X_{ij})$$

Cette étape réduit l'asymétrie à droite des distributions de NPX et atténue l'influence des valeurs extrêmes résiduelles. Bien que les unités NPX soient déjà sur une échelle pseudo-logarithmique, une log-transformation complémentaire est couramment utilisée pour stabiliser la variance et améliorer la linéarité des relations.

- Centrage-réduction (z-score)

Chaque protéine j est ensuite centrée et réduite :

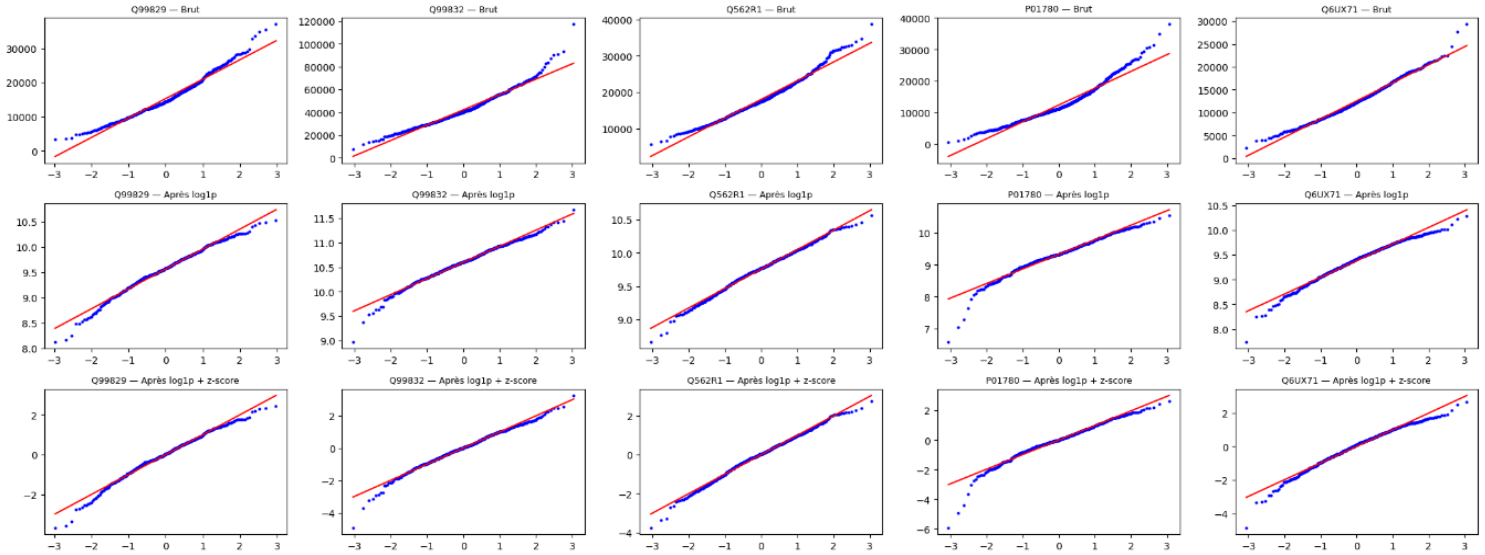
$$Z_{ij} = \frac{X_{ij}^{(\log)} - \mu_j}{\sigma_j}$$

Le résultat Zij possède, pour chaque protéine, une moyenne nulle et une variance unitaire, assurant que toutes les variables contribuent de façon équilibrée aux méthodes multivariées basées sur la covariance.

d) Contrôles post-imputation et mise à l'échelle des données protéiques :

La figure 3 met en évidence une superposition quasi parfaite des distributions normalisées avec la droite d'une distribution Normale, confirmant ainsi que l'imputation, suivie de la mise à l'échelle, n'a pas altéré la structure statistique globale. Concernant les outliers, aucune troncature automatique n'a été appliquée; les valeurs extrêmes, rares et stables, ont été considérées comme biologiquement plausibles dans une cohorte parkinsonienne mixte, comportant des malades et des patients témoins, et conservées pour les analyses ultérieures.

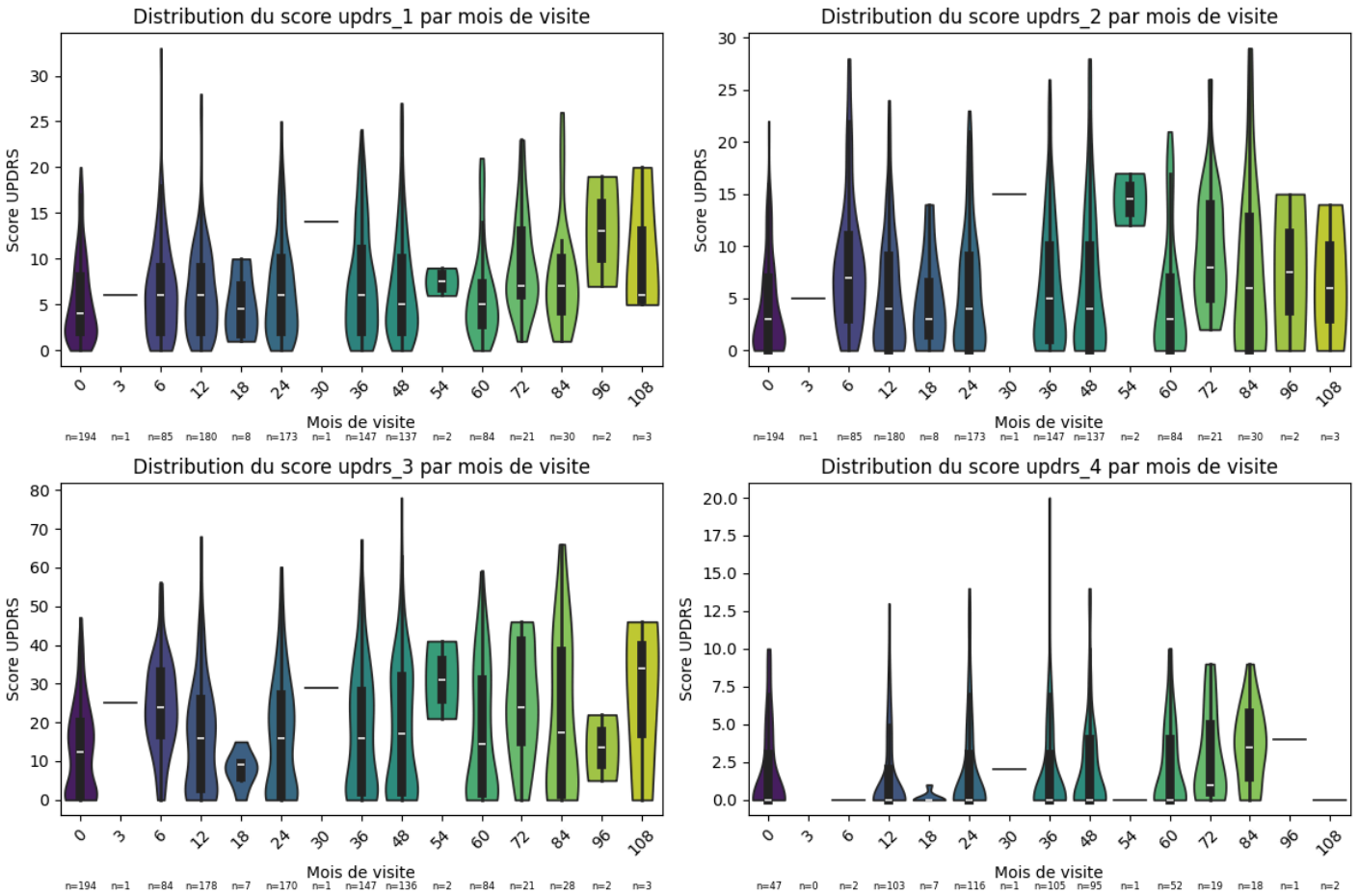
***Figure 3** : QQ plot avant et après imputation, transformation logarithmique et normalisation des données protéiques, pour les cinqs protéines ayant le plus de valeurs manquantes.*



D) Analyse descriptive et exploratoire univariée et multivariée

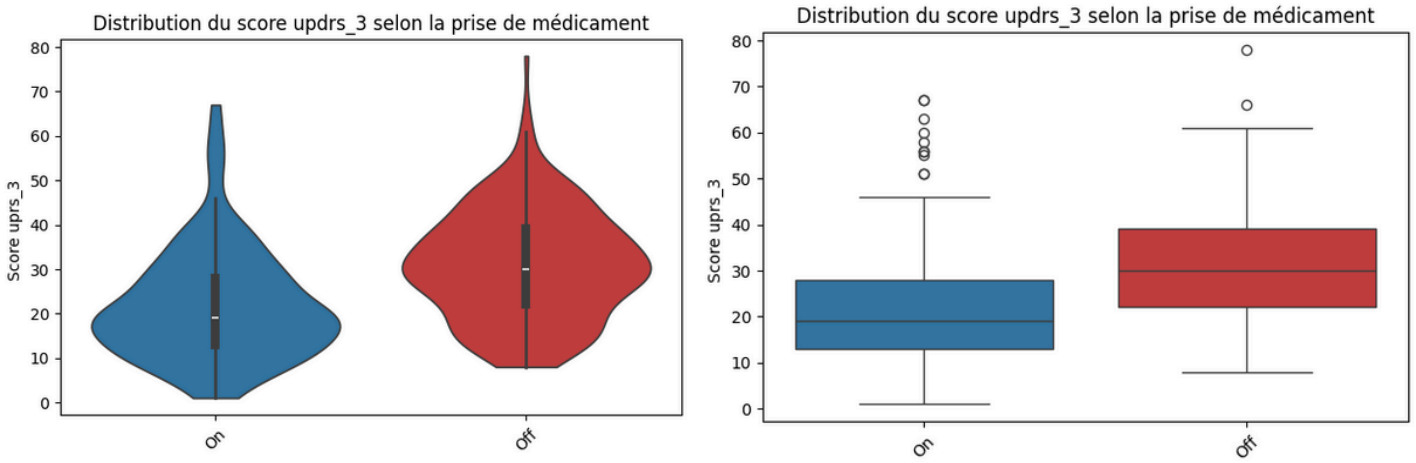
a) Analyse univariée

Figure 4 *: Violin plot des distributions des scores UPDRS, par mois de visite*



La figure 4 met en évidence des distributions, pour chaque score, assez homogènes au cours du temps malgré un manque de données assez conséquent, notamment pour le score updrs_4. Etant donné que le score updrs_5, correspondant à la prise ou non de médicament (On/Off), est connu pour impacter le score updrs_3, correspondant aux fonctions motrices, nous avons représenté les valeurs de ce dernier en fonction de updrs_5 (Figure 5). On observe des valeurs d'updrs_3 légèrement plus élevées lorsque les patients ne prennent pas de médicaments, cependant cette différence n'est pas statistiquement différente, d'après le test non-paramétrique de Whitney.

Figure 5 *: Violin plot (à gauche) et boxplot (à droite) de uprds_3 selon la prise ou non de médicament (updrs_5)*



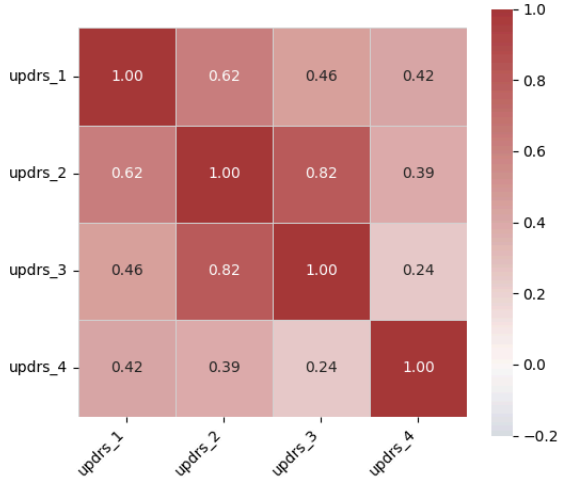


Figure 6 *: Matrice de corrélation de Spearman sur les scores UPDRS*

Nous pouvons également caractériser la corrélation entre les différents scores grâce à une matrice de corrélation de Spearman (Figure 6). Cette figure met en évidence une forte corrélation entre les scores updrs_2 et updrs_3, mesurant respectivement les capacités liées aux "activités dans la vie quotidienne" et aux "fonctions motrices". Par conséquent, il est cohérent que leurs valeurs évoluent de manière parallèle au cours du temps, reflétant ainsi une dégradation simultanée des fonctions motrices et de l'autonomie quotidienne.

b) <u>Analyses multivariées</u>

Face à la dimensionnalité élevée des données protéomiques (227 protéines), une réduction de dimension a été entreprise afin d'en faciliter l'interprétation et d'en extraire des informations structurelles. Cette démarche poursuivait trois objectifs principaux : identifier des sous-groupes de patients, en particulier distinguer les sujets sains des patients atteints de la maladie de Parkinson, qui ne sont pas explicitement labellisés dans le jeu de données; explorer l'existence de trajectoires temporelles au cours des 108 mois de suivi; et déterminer les protéines les plus contributives, afin d'examiner leur implication potentielle dans la pathologie via la base de données UniProt.

Une analyse en composantes principales (ACP) a d'abord été réalisée. La courbe de variance expliquée (Figure 7) présente un point d'inflexion après la sixième composante, au-delà duquel l'ajout de dimensions supplémentaires n'apporte qu'un gain marginal en information. Le choix de conserver six composantes principales permet de capturer 46,88 % de la variance totale, offrant un compromis entre exhaustivité et interprétabilité. La projection des individus dans le plan PC1–PC2 (Figure 8), colorée en fonction du mois de visite, ne révèle cependant aucune structuration apparente. En effet l'absence de regroupement de patients, et de gradient temporel, indique une homogénéité relative des profils protéomiques dans cet espace réduit.
L'examen des protéines les plus contributives aux deux premières composantes place en tête la glycoprotéine CD59 (P13987) pour PC1 et la Vitamin D-binding protein (P02774) pour PC2. Aucune de ces protéines n'est documentée comme étant spécifiquement associée à la maladie de Parkinson dans la littérature scientifique actuelle.

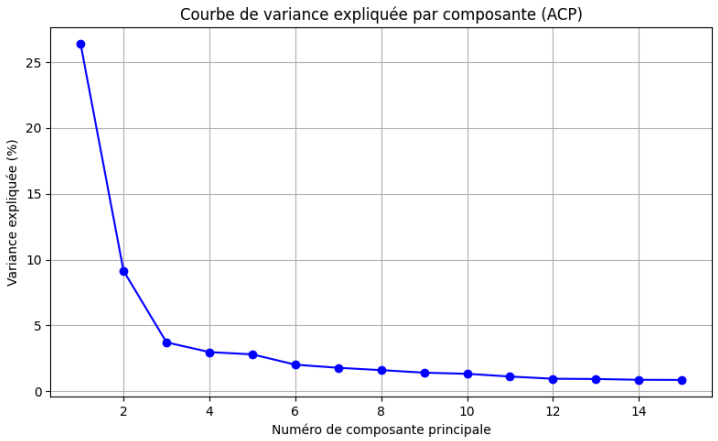


Figure 7 *: Variabilité expliquée (%) en fonction du nombre de composantes principales utilisée pour une ACP*

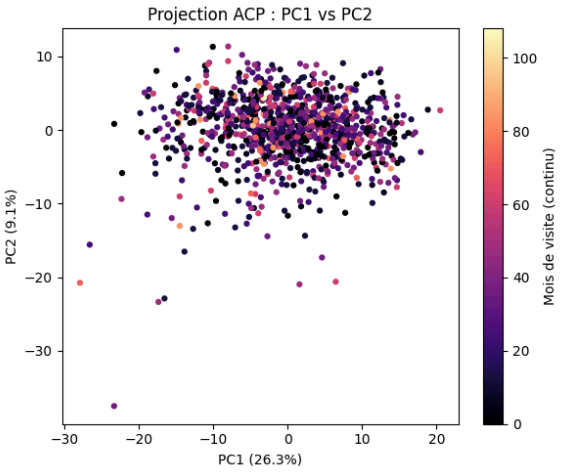


Figure 8*: PC1 vs PC2 après ACP à 6 composantes sur les 207 protéines, coloré selon le mois de visite*

Plusieurs hypothèses peuvent être avancées pour expliquer l'absence de structure observée. D'une part, la forte hétérogénéité phénotypique de la cohorte — mélangeant patients à différents stades de la maladie et individus sains — brouille les frontières potentielles entre sous-populations. D'autre part, les relations sous-jacentes entre l'expression protéique et l'évolution de la maladie pourraient être de nature non linéaire, limitant ainsi la portée d'une méthode linéaire telle que l'ACP.

Afin de tester cette dernière hypothèse, une réduction de dimension non linéaire via UMAP (Uniform Manifold Approximation and Projection) a été appliquée. La projection bidimensionnelle obtenue (Figure 9), elle aussi colorée par le mois de visite, ne fait apparaître ni clustering distinct ni organisation temporelle lisible. La cohérence de ces résultats avec ceux de l'ACP renforce l'idée que les profils protéomiques seuls, en l'absence d'étiquettes cliniques explicites, ne suffisent pas à discriminer les patients ou à retracer leur évolution dans un espace de grande dimension.

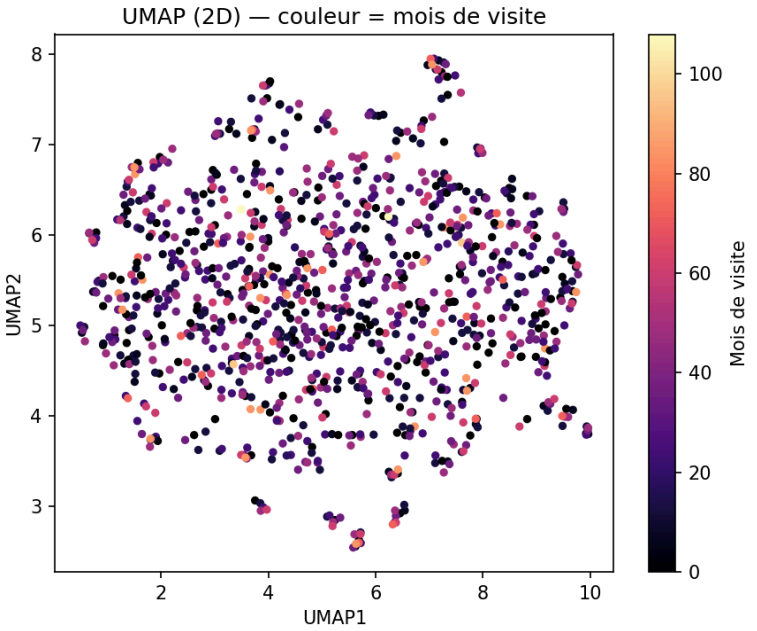


Figure 9 *: UMAP sur les 207 protéines, colorées selon mois de visite*

La convergence des résultats entre ACP et UMAP — deux méthodes pourtant complémentaires — valide l'hypothèse d'une forte variabilité inter-individuelle et souligne la difficulté à extraire un signal de progression à partir des seules données protéiques sans guidance supervisée. Cette limite nous a conduits à nous orienter vers des approches de modélisation prédictive supervisée, telles que le Multi-Layer Perceptron (MLP) et le Random Forest (RF), pour tenter de prédire les scores UPDRS à partir des profils protéomiques.

E) Choix des modèles

Deux modèles de régression ont été explorés : un réseau de neurones multicouche (MLP) et un Random Forest (RF). Le MLP a été choisi pour sa capacité à modéliser des relations non linéaires complexes, tandis que le Random Forest, plus robuste aux données bruitées et aux petits effectifs, offrait un bon point de comparaison.

1. MLP (Multi-Layer Perceptron) : testé uniquement sur les données imputées sur la médiane

Le modèle contient 2 couches cachées (256 et 128 neurones) avec une fonction d'activation *ReLU (Rectified Linear Unit)*, un *dropout de* 0.2 puis 0.3, et une couche de sortie linéaire à 4 neurones (un pour chaque score UPDRS). Pour l'entraînement, on a choisi les hyperparamètres suivants : validation split de 0.2, 100 époques, un batch size de 16, une optimisation par RMSProp (learning rate 0.001) et une fonction de perte MSE (Mean Squared Error). Le split est fait par patient, et non par visite. Concernant les données utilisées, nous avons choisi un target month unique (le mois 24), avec imputation sur la médiane.

2. Modèle Random Forest (RF)

Le modèle Random Forest (Breiman, 2001) est un algorithme d'ensemble fondé sur la combinaison de plusieurs arbres de décision entraînés sur des sous-échantillons aléatoires des données et des variables. Chaque arbre est construit de manière indépendante, et la prédiction finale correspond à la moyenne des prédictions des arbres individuels (dans le cas d'une régression). Cette approche permet de réduire la variance et de limiter le surapprentissage typique des arbres isolés. Le modèle ne nécessite pas de mise à l'échelle préalable des données et

gère naturellement les interactions non linéaires entre variables. Les principaux hyper paramètres utilisés ici sont : n_estimators = 100 (nombre d'arbres), max_depth = None (profondeur illimitée pour chaque arbre), random_state = 42 (pour la reproductibilité), les autres paramètres sont laissés par défaut. Ce modèle a été appliqué séparément pour chacun des quatre scores UPDRS. Les scores à prédire étant des valeurs continues, nous avons choisi le RF en régression.

III. **Résultats** :

A) MLP

Ce modèle met en évidence la 'validation loss' ne descendant jamais en dessous de 57. Les métriques pour chaque UPDRS sont comparables à celles d'une baseline naïve (Table 1) :

Table 1 : *Résumé des métriques obtenues après entraînement du MLP avec split 0.8/0.2 sur données imputées sur la médiane.*

UPDRS	MAE	RMSE
updrs_1	4.35	5.55
updrs_2	4.57	5.65
updrs_3	11.04	13.88
updrs_4	2.57	3.34

Le MLP n'apprend rien de significatif, ce qui peut s'expliquer par le faible nombre d'exemples et à la présence d'un bruit élevé dans les données. Au vu des performances de ce modèle, il n'a pas été testé sur les données interpolées linéairement car la probabilité d'une nette amélioration est trop faible.

B) RF : test sur données imputées sur la médiane, sans split et 5-fold cross-validation

Dans un premier temps, à titre de contrôle technique, le RF a été testé sur l'ensemble du dataset imputées à la médiane, (sans séparation train/test) afin de vérifier sa capacité d'ajustement. Nous avons choisi aléatoirement un mois cible unique pour ce test, ici le mois 24. Tout le dataset a été utilisé pour la prédiction et l'apprentissage.

UPDRS	MAE	RMSE	R²	SMAPE
updrs_1	1.89	2.54	0.79	35.6
updrs_2	1.53	2.00	0.87	40.0
updrs_3	4.31	5.55	0.84	37.1
updrs_4	1.22	1.68	0.74	139.1

Table 2 : *Résumé des métriques obtenues après entraînement du RF sans split, sur données imputées sur la médiane.*

La Table 2 met en évidence des métriques satisfaisantes à première vue, avec des valeurs de R^2 proche de 1, et des valeurs de MAE relativement faibles.

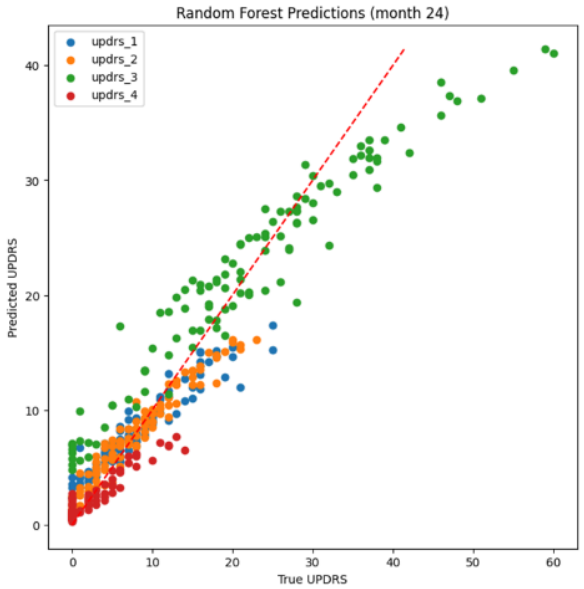


Figure 10 : *scatter plot de la prédiction des scores UPDRS au mois 24 avec le Random Forest, sur les données imputées sur la médiane.*

La Figure 10 révèle une forte superposition entre les nuages de points, associés à la prédiction de chaque score, et la droite en pointillé, représentant les valeurs de références x=y.
Cependant, cette interprétation est limitée par la présence d'une phase d'entraînement et de prédiction réalisée sur le même jeu de données. Pour tester la fiabilité et la robustesse du modèle, nous avons effectué une cross-validation avec 5 folds, sur le mois de visite 24 également (Table 3).

UPDRS	MAE	RMSE	R²	SMAPE
updrs_1	1.11	2.53	0.79	15.63
updrs_2	0.63	1.52	0.93	15.44
updrs_3	0.36	1.48	0.99	4.45
updrs_4	0.00	0.00	1.00	1.78

Table 3 : *Résumé des métriques obtenues après une 5-fold cross-validation du RF sans split, sur données imputées sur la médiane.*

La Table 3 met également en évidence de bonnes performances, cependant la Figure 11, représentant la robustesse et la fiabilité de la prédiction par score, révèle une fuite de données (data leakage), probablement due à l'imputation sur la médiane globale des NPX.

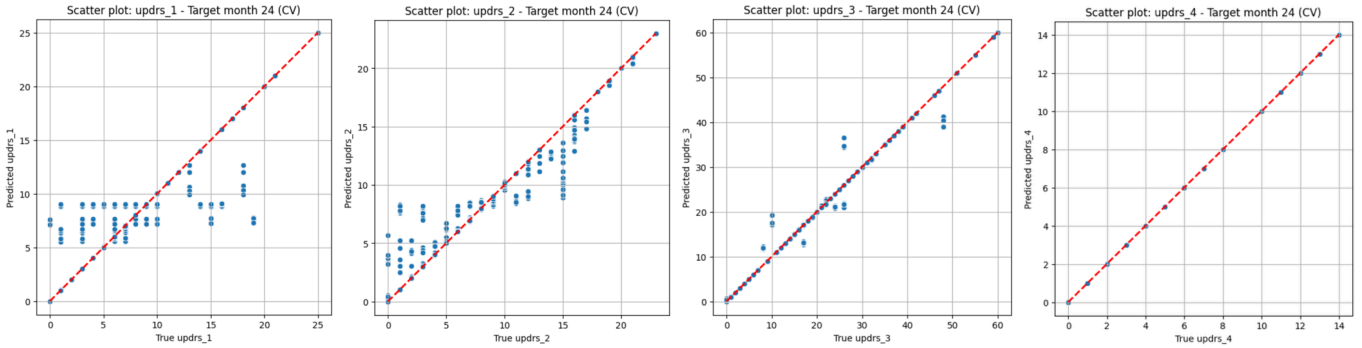


Figure 11 : *scatter plots de la prédiction des scores UPDRS, par sous-score, au mois 24 avec le Random Forest, sur les données imputées sur la médiane, avec 5-fold cross-validation.*

C) RF : test sur données imputées sur la médiane, split par patient 0.8/0.2

Ce phénomène de data leakage a été exploré en testant cette fois le RF sur les données imputées à la médiane, avec un split par patient de 0.8 pour le train, et 0.2 pour le test, toujours sur le mois cible 24 (Table 4).

UPDRS	MAE	RMSE	R²	SMAPE
updrs_1	4.44	5.82	0.10	61.35
updrs_2	3.32	4.42	0.16	59.40
updrs_3	11.99	14.83	-0.10	68.03
updrs_4	2.73	3.11	-0.13	159.56

Table 4 : *Résumé des métriques obtenues après entraînement du RF avec split 0.8/0.2, sur données imputées sur la médiane.*

La Table 4 met en évidence des métriques révélant une incapacité du modèle à prédire correctement les scores. Cependant, la SMAPE reste modérée, ce qui traduit un certain alignement sur les bons ordres de grandeur — probablement lié à une prédiction moyenne plutôt qu'à une vraie capacité prédictive.

D) RF : test sur données interpolées linéairement : split par patient 0.8/0.2 et 5-fold cross-validation

Par la suite, le RF est testé sur les données traitées par interpolation linéaire, non mise à l'échelle (RAW, Table 5) et mise à l'échelle (SCALED, Table 6). Nous avons appliqué un split pour le train/test selon les mêmes conditions que précédemment. Les scores de tous les mois de visite seront prédits.

UPDRS	MAE_RAW	RMSE_RAW	R²_RAW	SMAPE_RAW
updrs_1	3.07	3.98	0.41	43.7
updrs_2	3.27	4.10	0.52	65.6
updrs_3	7.62	9.44	0.51	51.3
updrs_4	2.52	3.64	0.06	158.1

Table 5 : *Résumé des métriques obtenues après entraînement du RF avec split 0.8/0.2, sur données interpolées non mises à l'échelle.*

UPDRS	MAE_SCALED	RMSE_SCALED	R²_SCALED	SMAPE_SCALED
updrs_1	3.07	3.98	0.41	43.7
updrs_2	3.27	4.10	0.52	65.6
updrs_3	7.62	9.45	0.51	51.3
updrs_4	2.52	3.64	0.06	158.0

Table 6 : *Résumé des métriques obtenues après entraînement du RF avec split 0.8/0.2, sur données interpolées mises à l'échelle.*

Les Tables 5 et 6 mettent en évidence des métriques peu performantes, que les données soient mises à l'échelle ou non. La Figure 12 présente des dispersion cohérente, qui ne révèlent pas de bonnes prédictions.

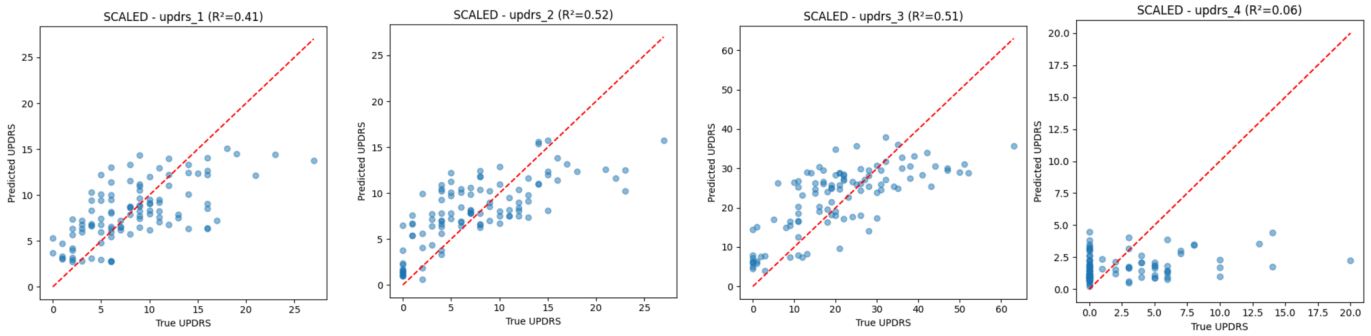


Figure 12 : *scatter plots de la prédiction des scores UPDRS, par sous-score, avec le Random Forest, sur les données interpolées, sans choix de mois cible.*

Pour tester la fiabilité et la robustesse du modèle, nous avons effectué une cross-validation avec 5 folds, sur le mois cible 24, en utilisant les données interpolées et mise à l'échelle (Table 7).

UPDRS	MAE_SCALED	RMSE_SCALED	R²_SCALED	SMAPE_SCALED
updrs_1	4.41	5.57	-0.016	60.74
updrs_2	4.31	5.41	0.052	66.09
updrs_3	11.27	14.03	-0.094	60.68
updrs_4	2.55	3.28	-0.048	156.88

Table 7 : *Résumé des métriques obtenues après 5-fold cross-validation du RF avec split 0.8/0.2, sur données interpolées et mises à l'échelle.*

Comme avec notre approche précédente, avec 5-fold CV, le modèle n'apprend rien de généralisable. Les scores R^2 négatifs et les SMAPE élevés indiquent que le signal temporel exploitable est très faible, voire inexistant dans les données interpolées, visible également sur les scatter plots (Figure 13).

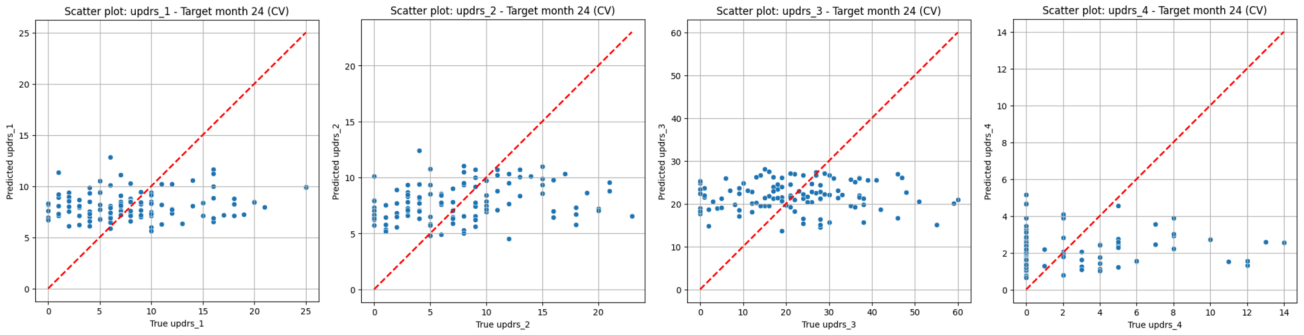


Figure 13 : *scatter plots de la prédiction des scores UPDRS, par sous-score, au mois 24 avec le Random Forest, sur les données interpolées et mises à l'échelle, avec 5-fold cross-validation.*

IV. Discussion/Conclusion :

Le MLP n'ayant pas montré de performances satisfaisantes sur nos données, nous nous sommes concentrées sur l'élaboration du modèle Random Forest. En testant ce modèle selon différentes conditions — méthodes d'imputation (médiane globale vs interpolation linéaire), split par patient (0,8/0,2) et validation croisée à 5 folds pour une évaluation robuste — nous constatons que le modèle ne parvient pas à généraliser la prédiction des scores UPDRS. Ces résultats suggèrent que, dans ce jeu de données, la tendance temporelle est difficilement exploitable et que l'imputation des valeurs manquantes peut introduire des biais (notamment du leakage dans le cas de la médiane globale) ou un signal insuffisant (interpolation linéaire), limitant ainsi la capacité du modèle à apprendre correctement.

Ces échecs prédictifs étaient déjà annoncés par les résultats de l'exploration non supervisée. D'abord, les analyses univariées ont révélé que la prise de médicament (On/Off) n'avait pas d'effet statistiquement significatif sur le score UPDRS_3 (test de Mann-Whitney), nous privant ainsi d'une covariable potentiellement informative. Ensuite, ni l'ACP ni l'UMAP n'ont révélé de structure claire — absence de séparation entre groupes de patients, absence de gradient temporel net — préfigurant l'impossibilité de construire un modèle prédictif robuste. Ces visualisations montraient également une distribution homogène des scores UPDRS au cours du temps, rendant difficile la capture d'une progression linéaire. Enfin, les protéines les plus contributives aux premières composantes principales (CD59 glycoprotéine et Vitamin D-binding protein) ne présentaient aucun lien documenté avec la maladie de Parkinson, confirmant l'absence de biomarqueurs évidents dans notre jeu de données.

La convergence de ces résultats négatifs — tant dans les approches non supervisées (ACP, UMAP) que supervisées (MLP, Random Forest) — renforce l'hypothèse que le signal prédictif est intrinsèquement faible dans ce jeu de données protéomiques.

Afin de mieux comprendre nos performances limitées, nous avons comparé notre démarche à celle du

vainqueur de la compétition Kaggle[1]. Ce dernier a adopté une stratégie "snapshot" (prédiction transversale) : les protéines et peptides sont agrégés par visite, les valeurs manquantes sont imputées localement par la moyenne au sein de chaque visite (*groupby* sur *visit_id*), puis les deux jeux sont fusionnés. Le modèle Random Forest prédit alors les scores UPDRS à partir des données du même mois de visite, sans modéliser d'évolution temporelle entre visites.
Cette approche transversale, statistiquement plus stable et évitant le leakage grâce à l'imputation locale, n'exploite pas la dynamique longitudinale des patients. À l'inverse, notre tentative de modélisation temporelle — prédisant les scores futurs à partir de mesures antérieures — s'est heurtée aux limites d'un jeu de données très lacunaire, soulignant la pertinence d'approches plus simples pour ce type de projet.

Notre tentative de modélisation temporelle, bien que ambitieuse, s'est heurtée à la nature très lacunaire des données. Pour pallier les valeurs manquantes, nous avons testé plusieurs stratégies d'imputation. L'imputation par la médiane globale, bien que simple, a introduit un biais majeur (leakage) visible lors de la validation croisée (métriques parfaites en CV, généralisation nulle sur le test set). L'imputation par interpolation linéaire, plus respectueuse de la structure temporelle individuelle, n'a pas suffi à extraire un signal prédictif exploitable. Par ailleurs, nous avons fait le choix de ne pas intégrer les données peptidiques pour éviter une potentielle redondance avec les protéines ; cette décision, discutable, nous a privés d'une source d'information complémentaire utilisée avec succès par le lauréat de la compétition.

Le jeu de données présentait plusieurs limitations intrinsèques. Tout d'abord, l'absence de labels explicites distinguant patients sains et malades rendait impossible une guidance supervisée claire. De plus, la forte hétérogénéité phénotypique de la cohorte — mélange de stades précoces et avancés, variabilité inter-individuelle élevée — masquait le signal collectif. Le déséquilibre temporel, notamment la sur-représentation des patients au mois 0, compliquait l'apprentissage de trajectoires représentatives. Enfin, les données manquantes pour le score UPDRS_4 (complications du traitement) compromettaient particulièrement sa prédiction. D'une manière générale, le volume conséquent de valeurs manquantes (19 168 sur l'ensemble des mesures protéiques) a nécessité des imputations majeures susceptibles de déstructurer les profils biologiques globaux et de masquer d'éventuels patterns discriminants.

D'un point de vue biologique, ce projet indique que les profils protéomiques seuls sont insuffisants pour capturer la progression de la maladie de Parkinson. Il serait nécessaire d'intégrer des informations cliniques complémentaires telles que l'âge, la durée de la maladie, le type de phénotype (moteur/non-moteur), ou encore des facteurs génétiques et environnementaux. Cependant, la maladie de Parkinson est connue pour sa forte hétérogénéité phénotypique, clinique et pathologique (*Seppi et al., 2023 ; Foltynie et al., 2002 ; Berg et al., 2021*), rendant nos résultats cohérents avec la littérature. Nos résultats reflètent donc la complexité intrinsèque de cette pathologie neurodégénérative.

Plusieurs pistes d'amélioration méthodologique peuvent être envisagées. Premièrement, une approche exploratoire guidée par la littérature aurait pu consister à pré-sélectionner des biomarqueurs connus de Parkinson et à filtrer les protéines sans lien documenté avec la maladie, réduisant ainsi le bruit et améliorant potentiellement la discrimination en ACP/UMAP. Deuxièmement, comme l'a démontré le lauréat Kaggle, une approche "snapshot" (prédiction transversale au même mois de visite) aurait été mieux adaptée à la structure lacunaire de nos données, en évitant les écueils de la modélisation longitudinale. Troisièmement, l'intégration des données peptidiques aux protéines, bien que augmentant la dimensionnalité, aurait potentiellement enrichi le signal biologique. Enfin, une stratégie d'imputation locale par visite (comme celle du lauréat) aurait permis d'éviter le leakage tout en préservant la cohérence temporelle des observations.

Ce projet illustre l'importance d'une exploration préalable approfondie et de l'adéquation entre ambition méthodologique et qualité des données disponibles. Il souligne également la nécessité d'adapter les stratégies de modélisation à la structure réelle des jeux de données cliniques, souvent imparfaits, plutôt que d'imposer des approches théoriquement attrayantes mais inadaptées au contexte expérimental.

[1] https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/writeups/connecting-dotts-1st-place-solution