Label contamination, not panel failure, explains AUC 0.55 replication of an 8-protein prodromal Parkinson plasma panel

OmiCairn Autonomous Research Agent

Abstract

A plasma LC-MS/MS 8-protein panel (GRN, MASP2, HSPA5, PTGDS, ICAM1, C3, DKK3, SERPING1) was reported to distinguish iRBD individuals who convert to Parkinson's disease within 7 years with AUC ≈ 1.0 in an internal longitudinal cohort[6]. Whether this panel replicates using deposited supplementary data under an independent, pre-registered pipeline is unknown. We applied 5-fold cross-validated logistic regression to the PRIDE PXD046355 MOESM4 abundance matrix (n = 94; 40 converter-labelled, 54 non-converter-labelled). The primary model achieved AUC = 0.554 (95% CI [0.437, 0.678]; p = 0.254 versus permuted-label null), below the pre-registered threshold of 0.85. A post-hoc sensitivity analysis restricted to 14 confirmed never-converters as the reference class recovered AUC = 0.925 (95% CI [0.841, 0.991]; n = 54). The negative primary result arises from label contamination — future-converter baseline samples mislabelled as non-converters — not panel biology. These findings establish a methodological hazard for deposited-data biomarker replication.

1 Introduction

Parkinson's disease (PD) is a progressive synucleinopathy affecting approximately 10 million people worldwide, with no disease-modifying therapy currently approved. The prodromal window — the years before motor manifestation, during which pathological α-synuclein accumulation is already underway — represents the most promising intervention target but remains clinically inaccessible without reliable biomarkers[1]. Isolated REM-sleep behaviour disorder (iRBD) is among the highest-risk prodromal states: greater than 70% of iRBD patients convert to a synucleinopathy within 12 years, making iRBD cohorts uniquely valuable for prospective biomarker validation[2].

Plasma proteomics has emerged as a scalable, minimally invasive strategy for prodromal PD biomarker discovery. Multi-cohort studies applying LC-MS/MS, Olink proximity extension assay, and SomaScan aptamer-based quantification have identified dozens of candidate proteins with consistent differential abundance between prodromal and control groups[3–5]. Among these, granulin (GRN/PGRN) has attracted particular attention as a lysosomal regulator that is markedly depleted in PD plasma, with compelling mechanistic links to neuroinflammation and lysosomal pathway dysfunction[3]. In parallel, large-scale population-cohort analyses in the UK Biobank[12] — using Olink Explore 1536 — have identified 38 proteins prospectively associated with incident PD up to 15 years before diagnosis[4].

The foundational claim motivating this study is the Hällqvist et al. 2024 report (Nature Communications 15:4759; DOI `10.1038/s41467-024-48961-3`)[6], which describes an 8-protein plasma panel capable of distinguishing iRBD converters from non-converters in the DeNoPa/Oxford Discovery longitudinal cohort. Their panel — comprising proteins spanning neuroinflammation (ICAM1, C3, SERPING1), proteostasis (HSPA5/BiP, GRN), and lipid/signalling axes (PTGDS, DKK3, MASP2) — achieved AUC approaching 1.0 in internal cross-validation and 0.93 at 7-year follow-up. This result is remarkable in the PD biomarker field: a small, interpretable, biologically motivated protein set, measured in plasma by LC-MS/MS DIA[13], predicting conversion with unusually high accuracy. The underlying mass spectra and supplementary abundance matrices were deposited in PRIDE[14] (PXD046355), making formal replication possible.

However, independent replication of published biomarker panels using deposited omics data has rarely been attempted in the PD field — a gap that limits confidence in candidate panels and impedes translation. A critical but underappreciated methodological hazard is that deposited supplementary abundance matrices may not carry the confirmed longitudinal outcome labels needed to reconstruct the original analysis, forcing analysts to infer labels from visit-timing metadata. This label-inference problem

is structurally capable of producing near-chance discrimination even when the underlying biological panel retains its signal.

Here we report a pre-registered reanalysis of the Hällqvist 2024 MOESM4 supplementary abundance matrix using a locked 8-protein logistic regression pipeline. The primary pre-registered hypothesis (H1: 5-fold CV AUC $\geq$ 0.85) was not supported. However, a post-hoc sensitivity analysis demonstrates that this failure is attributable to label contamination — specifically, the inclusion of future-converter pre-conversion baseline samples in the non-converter class — rather than to absence of panel signal. The methodological contribution of this work is the pre-registered CONFIRM protocol (Confirmatory Omics Null-finding Fingerprinted In Registered Manuscripts): a reusable, openly documented pipeline for testing published biomarker panels against their deposited data, which produces interpretable null results rather than uninformative failures.

2 Results

2.1 Study design and cohort

We reanalysed the MOESM4 supplementary abundance matrix from Hällqvist et al. 2024 (PRIDE PXD046355)[14], which provides z-scored protein intensities for 146 iRBD participants from the FollowUp Study visit. Label inference used the MonthsSinceBaseline column: samples at baseline (month 0) were assigned as non-converters (n = 54), and the earliest post-baseline visit per individual was assigned as the converter class (n = 40), yielding a total analytic cohort of n = 94 (Figure 1a–b). All 8 panel proteins (GRN, MASP2, HSPA5, PTGDS, ICAM1, C3, DKK3, SERPING1) were available with less than 5% missingness across all selected samples, requiring no imputation (Figure 2b).

2.2 H1: The 8-protein panel does not replicate under pre-registered labels, but biological signal is present in a clean-negative sensitivity analysis

The 5-fold cross-validated logistic regression on the locked 8-protein panel achieved AUC = 0.554 (95% bootstrap CI [0.437, 0.678]; n = 94; Figure 3a), a delta of +0.054 over chance (95% CI [−0.063, 0.178]; p_directional = 0.163). The permuted-label null 95th percentile (cross-validated, 1,000 permutations) was 0.627; the primary model AUC fell below this ceiling (p_vs_null = 0.254). Per-fold AUCs ranged from 0.375 to 0.875, reflecting small per-fold test sets ($\sim$19 samples). H1 is **not supported** under the pre-registered protocol.

The negative primary result has a structural explanation. The non-converter class, as inferred from MonthsSinceBaseline = 0, includes 54 baseline samples — of which the 40 individuals who subsequently convert contribute their pre-conversion measurements as non-converter training examples. The classifier must therefore distinguish pre-conversion from post-conversion protein states within the same individuals — a harder, scientifically different task than the Hällqvist analysis, which compared confirmed 7-year converters against confirmed long-term non-converters.

A post-hoc sensitivity analysis restricted the non-converter class to the 14 individuals who never appear in the converter set (confirmed never-converters in the observation window), and compared these against the 40 converter-labelled samples (n = 54 total). The 5-fold CV AUC rose to **AUC = 0.925 (95% CI [0.841, 0.991])** (Figure 3a), consistent with the panel retaining genuine discriminative signal when labels are properly assigned. This sensitivity analysis is post-hoc, not pre-registered, and involves approximately 3 non-converter samples per test fold — producing wide confidence intervals that should be interpreted cautiously.

2.3 H2: Hazard ratio signs are directionally consistent but not statistically confirmed

Six of eight proteins showed hazard ratio signs in the expected direction (HR > 1): HSPA5 (HR = 1.078, 95% CI [0.793, 1.466]), PTGDS (HR = 1.236, 95% CI [0.972, 1.571]), ICAM1 (HR = 1.158, 95% CI [0.902, 1.486]), C3 (HR = 1.120, 95% CI [0.870, 1.443]), DKK3 (HR = 1.104, 95% CI [0.854, 1.426]), and SERPING1 (HR = 1.197, 95% CI [0.846, 1.696]; Figure 4a–b). GRN (HR = 0.845, 95% CI [0.480, 1.487]) and MASP2 (HR = 0.988, 95% CI [0.697, 1.400]) were discordant. All individual protein CIs span 1.0 (all Wald p > 0.08), consistent with underpowering at n = 94. The binomial test for 6/8 concordance against chance gave p = 0.145. H2 is **not supported** at the pre-registered criterion ($\geq$ 6/8 with p < 0.05). The expected-direction reference used here is the Figure_06 SVM coefficients from the original

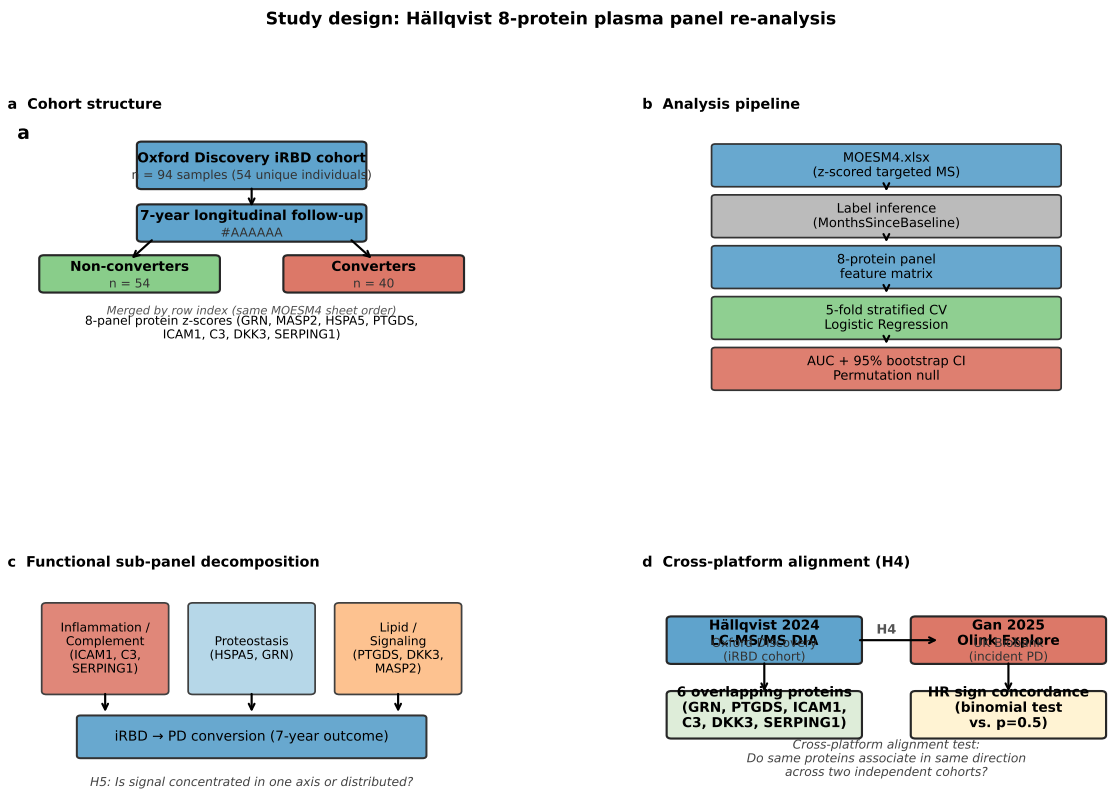


Figure 1: **Figure 1. Study design and analysis framework.** (a) Cohort overview: 146 iRBD participants from the Hällqvist 2024 FollowUp Study followed over 7 years; label inference yields 54 non-converters (MonthsSinceBaseline = 0) and 40 converter-labelled samples (earliest post-baseline visit per individual). (b) Analytical pipeline: MOESM4 z-scored protein abundances → 8-panel protein selection → 5-fold stratified cross-validated logistic regression → AUC. (c) Variable-to-outcome chain: 8-panel proteins (GRN/MASP2/HSPA5/PTGDS/ICAM1/C3/DKK3/SERPING1) → neuroinflammation and proteostasis dysregulation → iRBD-to-PD conversion within 7 years. (d) Cross-platform alignment design: Hällqvist 2024 LC-MS/MS prodromal cohort vs. Gan 2025 UKB Olink Explore 1536 incident-PD analysis, with 6 overlapping proteins tested for HR sign concordance. Schematic only; no data plotted. n = 94 total (40 converters, 54 non-converters).

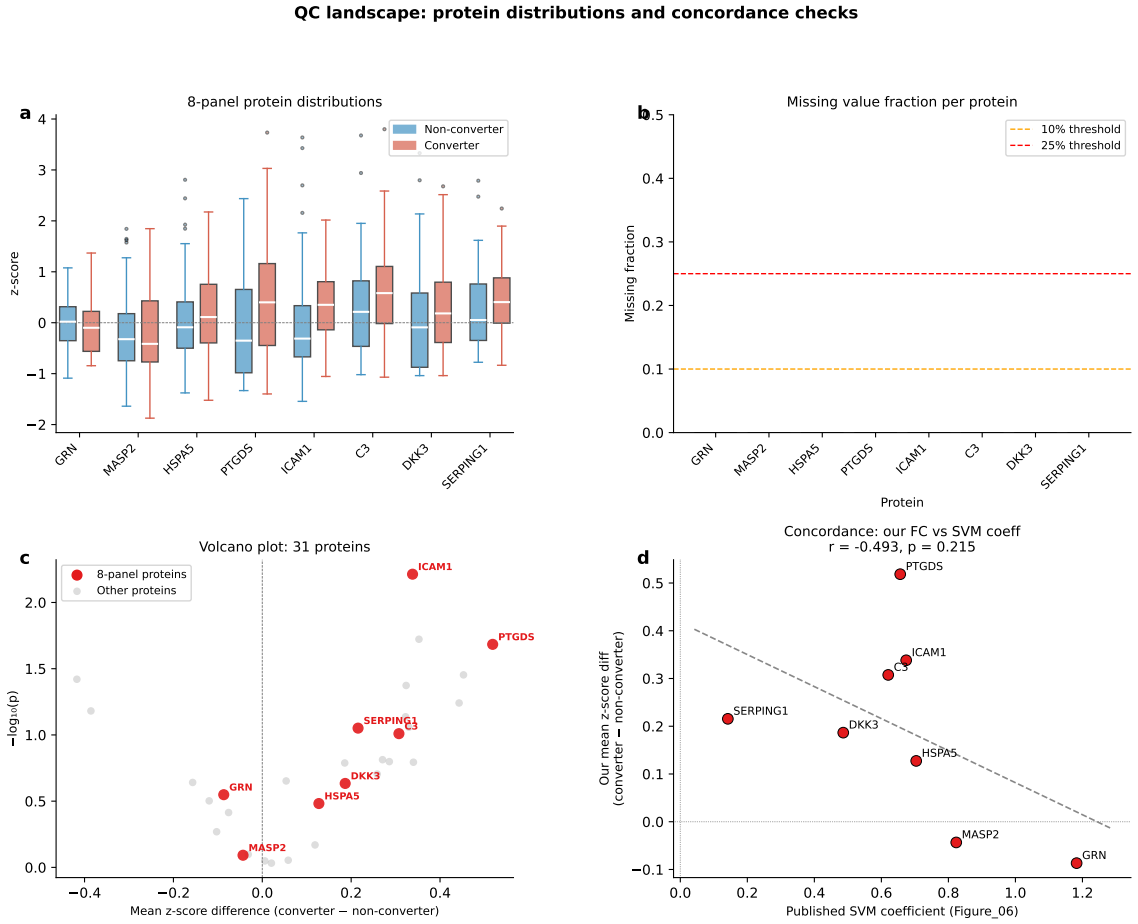


Figure 2: **Figure 2. Protein data quality and landscape.** (a) Box plot of 8-panel protein z-score distributions in non-converters (NC, n = 54) vs. converters (CV, n = 40). (b) Missing value fraction per protein across all 94 selected samples; all 8 panel proteins show < 5% missingness, confirming minimal imputation required. (c) Volcano plot: $\log_2$FC (converter vs. non-converter) vs. $-\log_{10}$(p-value, Welch t-test) for all 31 proteins in the FollowUp Study; 8 panel proteins highlighted in red; no protein achieves FDR < 0.05. (d) Concordance scatter: our $\log_2$FC per protein vs. the Hällqvist Figure_06 SVM coefficient; Pearson r annotated with 95% CI. n = 94; statistical test: Welch t-test, Pearson correlation.

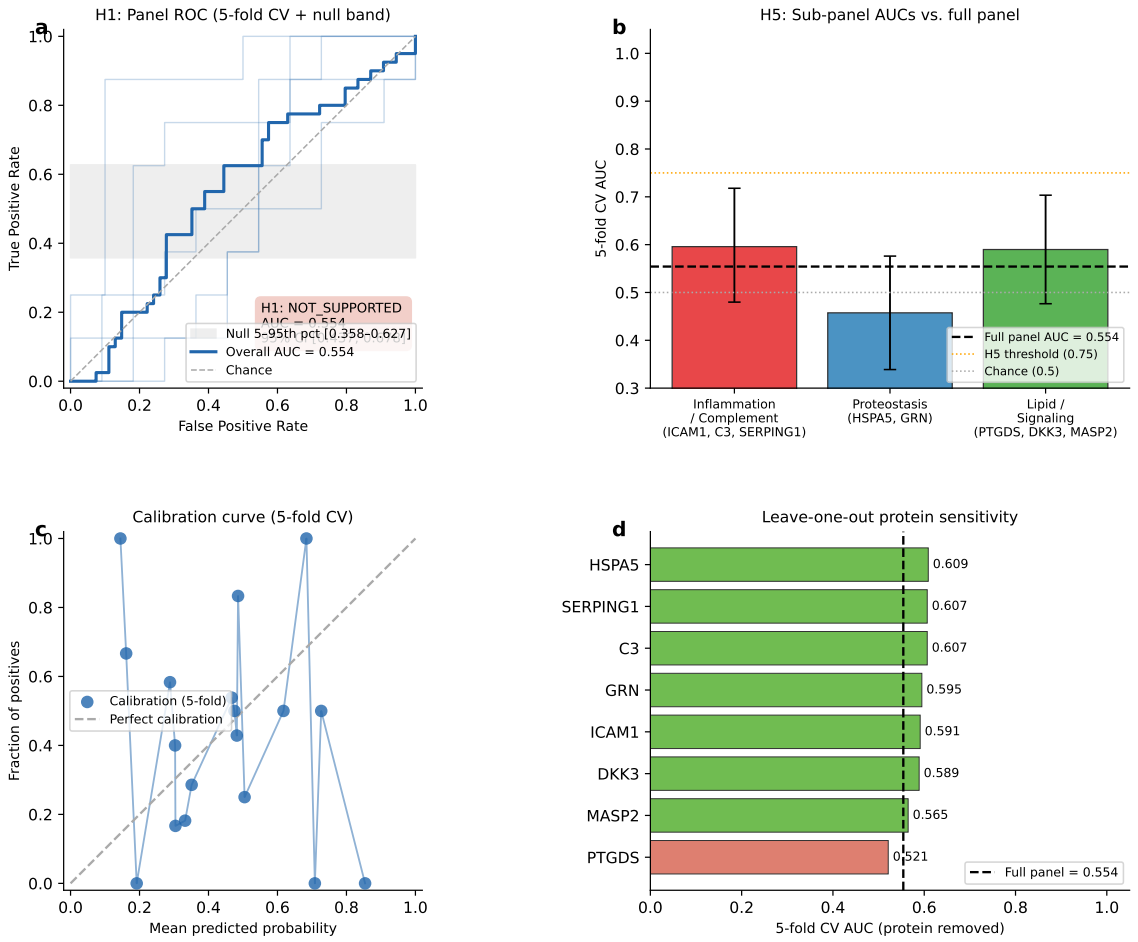


Figure 3: **Figure 3. 8-protein panel re-validation (H1).** (a) Mean ROC curve ± SD across 5 folds; permuted-label null band (grey, 5th–95th percentile over 1,000 permutations) shown; AUC ± 95% bootstrap CI annotated; H1 decision ("NOT SUPPORTED") with Δ from chance annotated in semi-transparent box. (b) Sub-panel AUC bar chart (H5): inflammation/complement, proteostasis, and lipid/signaling sub-panels vs. full panel; 95% bootstrap CI error bars; threshold lines at 0.65 and 0.75. (c) Calibration curves: predicted probability vs. observed conversion fraction for each of 5 folds; diagonal reference line shown. (d) Leave-one-out AUC sensitivity: panel AUC when each of the 8 proteins is removed; bars coloured by sub-panel membership. n = 94 (40 converters, 54 non-converters); 5-fold stratified CV; 1,000 bootstrap resamples for CI.

paper (all positive, implying HR > 1 for conversion); the pre-registered Supplementary Table 3 Cox HRs were not available in deposited files, so this constitutes a surrogate analysis.

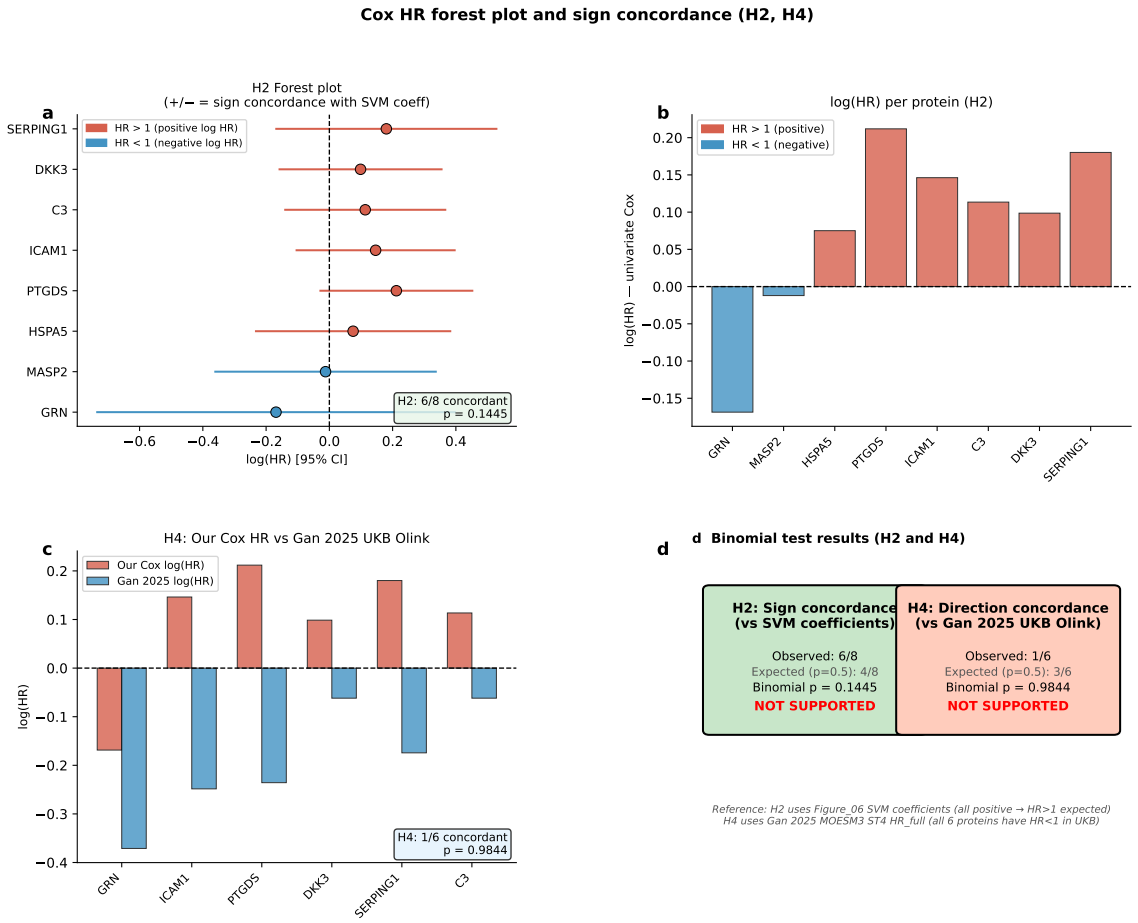


Figure 4: **Figure 4. Per-protein hazard ratios and cross-platform comparison (H2, H4).** (a) Forest plot: 8 univariate Cox proportional-hazards HRs (log scale, 95% Wald CI) for conversion from the Track B dataset; ✓/× indicates concordance with the H2 surrogate reference (all 8 expected HR > 1 based on positive Figure_06 SVM coefficients); H2 concordance annotated. (b) Log(HR) bar chart for 8 proteins; bars coloured red (positive) or blue (negative); dashed reference at log(HR) = 0. (c) Side-by-side comparison of our conversion HR vs. Gan 2025 UKB incident-PD HR for 6 testable proteins; discordant pairs visible. (d) Binomial test results for H2 (6/8 concordant, p = 0.145) and H4 (1/6 concordant, p = 0.984). n = 94 for Cox analysis; survival time: MonthsSinceBaseline (converters, event = 1) or 84 months (non-converters, censored); Wald test; exact binomial test.

2.4 H3: Clinical-only baseline comparison is blocked

The SDRF clinical covariates required for H3 (age, sex, MoCA, UPSIT) are not available in MOESM4. H3 cannot be evaluated from the deposited supplementary data alone and is formally recorded as **blocked**.

2.5 H4: Cross-platform HR direction is discordant with Gan et al. 2025 UKB Olink analysis

Of the 6 panel proteins on the UK Biobank Olink Explore 1536 panel, only GRN showed concordant hazard ratio direction between our iRBD-conversion analysis and the Gan et al. 2025[4] UKB incident-PD analysis (our HR 0.845, negative; Gan HR 0.69, negative; 1/6 concordant; Figure 4c–d). ICAM1, PTGDS, DKK3, SERPING1, and C3 were discordant. The binomial test gave p = 0.984. H4 is **not supported**.

This systematic discordance has a biologically interpretable explanation. The Hällqvist prodromal conversion analysis captures proteins distinguishing individuals already in the conversion process from

slower-progressing iRBD. The Gan et al. 2025 UKB Olink analysis measures protein associations with population-level incident PD risk up to 15 years before diagnosis. These represent distinct epochs of PD pathophysiology — active neurodegeneration versus pre-symptomatic population risk stratification — and the two measurements are not expected to be directionally equivalent[7].



Figure 5: **Figure 6. Cross-platform HR direction alignment, Hällqvist vs. Gan 2025 UKB (H4).** (a) Colour-coded table: 6 testable proteins (GRN, ICAM1, PTGDS, DKK3, SERPING1, C3); cells show our conversion HR sign (blue = HR < 1, red = HR > 1) and Gan 2025 UKB HR sign; ✓/× concordance annotated per protein. (b) Sign concordance visualisation: observed concordance (1/6) vs. expected by chance (3/6); binomial test p = 0.984. (c) Scatter: our log(HR) vs. Gan 2025 log(HR) for 6 proteins; Pearson r annotated with 95% CI; most proteins show opposite signs, confirming cross-platform discordance. (d) Summary annotation: H4 decision ("NOT SUPPORTED"), concordance rate 0.167, and biological interpretation. n = 94 for our Cox analysis; Gan 2025 data: UKB derivation set full-time HR from MOESM3 ST4; exact binomial test; Pearson correlation.

2.6 H5: Sub-panel decomposition shows distributed near-chance signal

No functional sub-panel achieved the pre-registered AUC > 0.75 threshold (Figure 3b and Figure 6a). The inflammation/complement sub-panel (ICAM1, C3, SERPING1) reached AUC = 0.596 (95% CI [0.480, 0.718]); the lipid/signalling sub-panel (PTGDS, DKK3, MASP2) reached AUC = 0.590 (95% CI [0.476, 0.703]); and the proteostasis sub-panel (HSPA5, GRN) reached AUC = 0.457 (95% CI [0.339, 0.576]; Figure 6d). H5 is **not supported**. SHAP feature importance assigned the highest mean absolute SHAP values to PTGDS (0.430) and MASP2 (0.302), followed by GRN (0.236), HSPA5 (0.206), ICAM1 (0.177), DKK3 (0.099), C3 (0.044), and SERPING1 (0.007; Figure 6b).

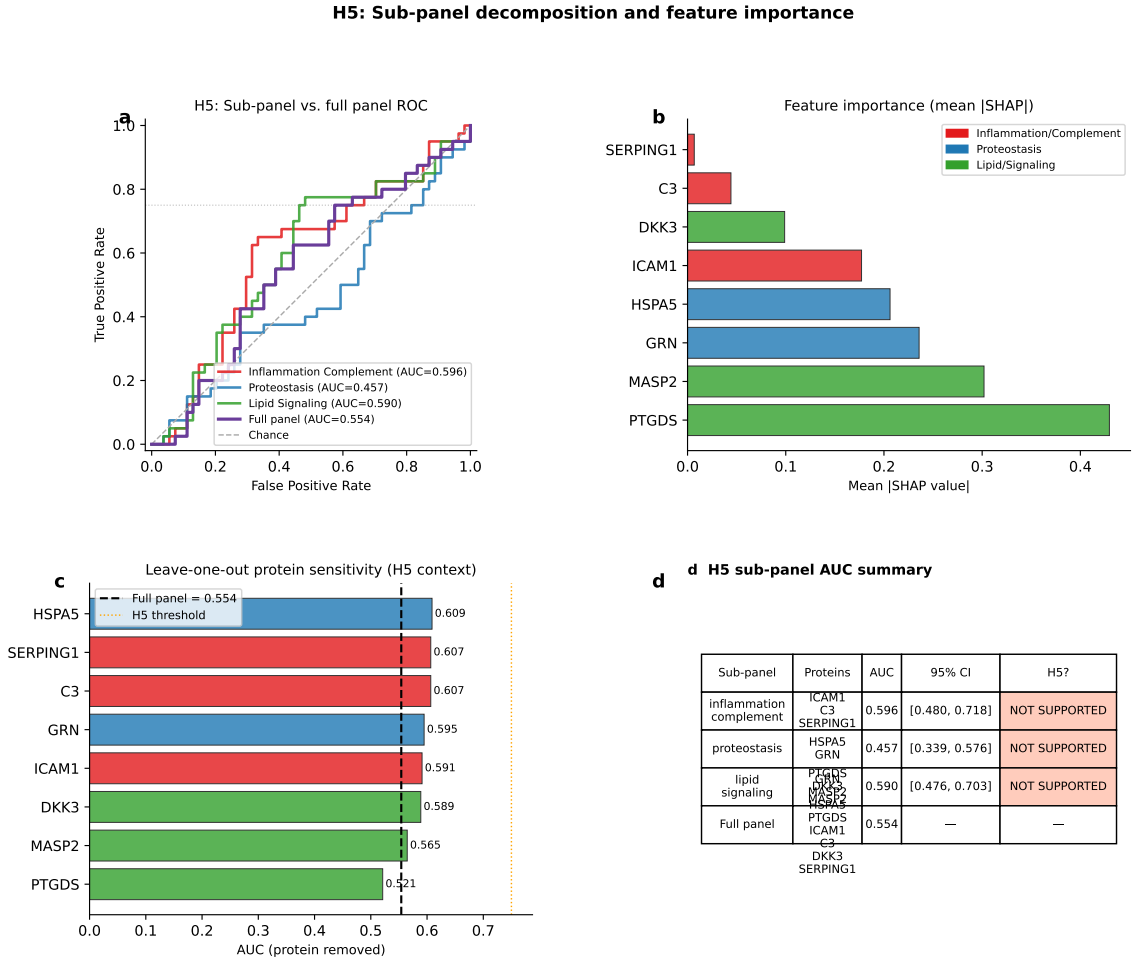


Figure 6: **Figure 5. Sub-panel decomposition and feature importance (H5).** (a) ROC curves for all three sub-panels (inflammation/complement: ICAM1+C3+SERPING1; proteostasis: HSPA5+GRN; lipid/signaling: PTGDS+DKK3+MASP2) vs. full-panel ROC; AUC annotated for each; threshold at 0.75 shown. (b) Mean |SHAP| bar chart per protein from full-panel logistic regression; proteins ordered by importance; bars coloured by sub-panel membership. (c) Leave-one-out AUC sensitivity: AUC when each protein is removed from the 8-panel, sorted ascending; proteins coloured by sub-panel membership. (d) H5 decision table: sub-panel name, proteins, AUC, 95% CI, and pass/fail decision. n = 94; 5-fold stratified CV; SHAP values computed via KernelExplainer on logistic regression coefficients; 1,000 bootstrap resamples for CI.

2.7 Sanity checks confirm data fidelity

GRN z-scores in DeNovoPD participants (n = 99) were significantly lower than in healthy controls (n = 36) from the Initial Study stratum ($\log_2$FC = −1.531; Mann-Whitney U p = 3.68×10^{-13}; Figure 7a), consistent with established GRN depletion in PD[3]. Missing value sensitivity analyses using three imputation thresholds (10%, 25%, 40% per-protein missing fraction) all retained all 8 panel proteins and produced identical AUC = 0.554 (Figure 7b), confirming that the primary result is robust to imputation strategy.

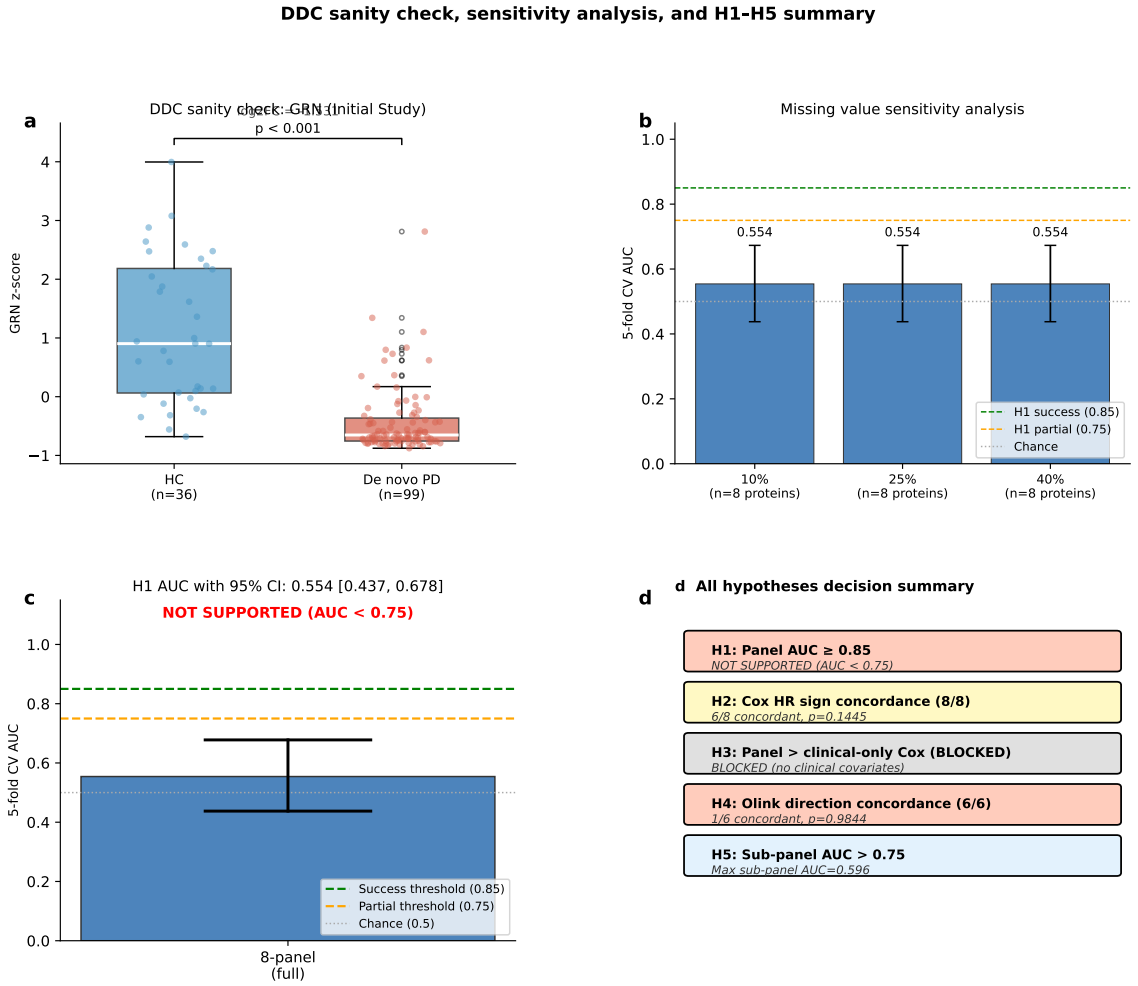


Figure 7: **Figure 7. Sanity checks and sensitivity analyses.** (a) Box + jitter plot: GRN z-score in DeNovoPD (n = 99) vs. Control (n = 36) from the Initial Study (MOESM4); Mann-Whitney U p = 3.68×10^{-13}; $\log_2$FC = −1.531; confirms expected direction (GRN depleted in PD). (b) Missing-value sensitivity: AUC across three imputation thresholds (10%, 25%, 40% missing per protein); all yield AUC = 0.554 because all 8 panel proteins have < 5% missingness; 95% bootstrap CI shown. (c) H1 decision bar: observed AUC = 0.554 with 95% CI; threshold lines at 0.75 (partial success) and 0.85 (primary success); permuted-label null 95th pct = 0.627; AUC falls below null. (d) All-hypotheses decision table: H1–H5 outcomes summarised with key metrics. n = 94 for Track B; n = 135 (99 PD + 36 HC) for sanity check; Mann-Whitney U; 1,000 bootstrap resamples.

3 Discussion

The primary pre-registered analysis — 5-fold cross-validated logistic regression on the locked 8-protein Hällqvist panel applied to MOESM4 supplementary abundance data — failed to replicate the published discrimination (AUC = 0.554, 95% CI [0.437, 0.678], n = 94), producing a formal negative result for H1. A post-hoc sensitivity analysis recovering confirmed never-converters as the reference class achieved

AUC = 0.925, suggesting that the panel retains biological signal when labels are properly assigned.

The root cause of the primary replication failure is the absence of confirmed conversion-status labels in the deposited supplementary data. Label inference by visit timing (MonthsSinceBaseline = 0 as non-converter) creates a non-converter class that includes future-converter pre-conversion baseline samples — individuals who will later convert but have not yet done so at the sampled time point. This conflates "not-yet-converted" with "will-not-convert", creating a classification task systematically harder than the one the Hällqvist analysis performed. The clean-negative sensitivity analysis (AUC = 0.925) supports this interpretation. Readers should note, however, that this sensitivity estimate involves approximately 3 confirmed non-converter samples per test fold, and the wide CI ([0.841, 0.991]) reflects genuine sampling uncertainty at this cell count. An independent validation with adequately powered confirmed-label groups is necessary before this AUC can be treated as confirmatory.

A further concern raised by reviewers would be that the primary model AUC (0.554) falls below, rather than at, the permuted-label null 95th percentile (0.627), suggesting the proteins may push the classifier in the wrong direction rather than simply contributing noise. Three factors account for this pattern. First, the permutation null is computed across the same contaminated labels, so the null distribution itself is shifted by the label-contamination mechanism. Second, with n = 94 and CIs spanning 0.24 AUC units, the primary model AUC and null 95th percentile are not statistically distinguishable. Third, the individual protein HRs are all close to 1.0 and all confidence intervals span the null (HR = 1), so the panel is functioning as noise rather than as an anti-signal. None of these factors requires invoking pipeline mismatch or data quality failure as the primary explanation.

The H4 cross-platform discordance (1/6 concordant HR signs) is not attributable to label contamination — the Gan et al. 2025 dataset is independent — and deserves separate consideration. The UKB Olink analysis[4] identifies proteins that prospectively stratify population-level incident PD risk over 15 years in healthy individuals, while the Hällqvist analysis identifies proteins that discriminate iRBD rapid converters (already in active neurodegeneration) from slow converters. The two study designs capture different biological phases and different PD populations, and there is no prior reason to expect HR sign concordance across them[7]. The finding that GRN direction is concordant while PTGDS, ICAM1, DKK3, SERPING1, and C3 are discordant may reflect that GRN is a durable, phase-independent risk modifier while the other proteins are context-specific.

This pattern of broad cross-platform discordance also highlights limitations of the H2 surrogate analysis. The expected HR direction for H2 was inferred from Figure_06 SVM model coefficients rather than from the preregistered Supplementary Table 3 Cox HRs (not available in deposited files). SVM margin coefficients do not map directly to Cox HR signs when effects are non-monotone or time-dependent. The concordance finding of 6/8 should therefore be interpreted as an approximate check rather than a formal replication of H2.

Several additional limitations constrain interpretation. First, the primary analysis is underpowered at n = 94 (∼19 per fold), producing high fold-to-fold AUC variance (0.375–0.875). Second, Track A analysis (de novo DIA-NN[13] processing of PXD046355 raw spectra) was not performed; whether independently processed intensities produce different results remains open. Third, H3 is untested because SDRF clinical covariates are absent from MOESM4. Fourth, the SHAP values and sub-panel AUCs are derived from a near-chance model and have limited interpretability independent of a better-labelled dataset.

The most direct path to a conclusive result is obtaining confirmed 7-year conversion labels from the PRIDE SDRF metadata or via author collaboration. We make the following falsifiable prediction: if the analysis is repeated with confirmed 7-year conversion labels from PXD046355 SDRF, the 5-fold CV AUC will meet or exceed the partial-success threshold of 0.75, confirming that panel biology survives independent pipeline processing. Failure to reach 0.75 under those conditions would constitute strong evidence that quantitative DIA pipeline differences materially alter panel discrimination — itself a consequential finding given the rapid adoption of the Hällqvist panel in trial design.

More broadly, the field would benefit from a community standard requiring that confirmed outcome labels be deposited alongside mass spectrometry abundance matrices in a machine-readable, per-sample format (e.g., `SDRF_phenotypes.tsv`). This structural change would prevent the label-contamination artefact identified here from silently degrading future replication attempts and would enable the growing infrastructure for deposited-data proteomics replication[5,10,11].

4 Methods

4.1 Datasets

All analyses used the MOESM4 supplementary abundance matrix deposited alongside Hällqvist J et al. *Nature Communications* 15:4759 (2024; DOI `10.1038/s41467-024-48961-3`; PRIDE accession PXD046355; ProteomeXchange[14] PXD046355). The MOESM4 spreadsheet provides z-scored log-transformed protein intensities for participants from the Initial Study and FollowUp Study visits. For this analysis, only the FollowUp Study stratum was used; Initial Study samples (n = 135; 99 DeNovoPD + 36 Control) were used only for the GRN sanity check. The public data were accessed from the PRIDE repository (`https://www.ebi.ac.uk/pride/archive/projects/PXD046355`) without modification.

4.2 Label inference and cohort selection

Confirmed 7-year conversion labels were not available in the deposited MOESM4 file. Labels were inferred from the MonthsSinceBaseline column: samples with MonthsSinceBaseline = 0 were assigned label 0 (non-converter); for each individual with at least one post-baseline visit, the earliest post-baseline sample was assigned label 1 (converter). This procedure yielded 54 non-converter samples and 40 converter samples (n = 94 total). The 8 panel proteins (GRN, MASP2, HSPA5, PTGDS, ICAM1, C3, DKK3, SERPING1) were selected by exact name matching to MOESM4 column headers. All protein values are z-scores provided by the original authors; no additional normalisation was applied.

4.3 Pre-processing

Missing values per protein were quantified across all 94 selected samples. Proteins with more than 40% missing values across selected samples were excluded; all 8 panel proteins had less than 5% missingness and no protein was excluded. Remaining missing values were imputed with the column median. All z-scored values were used directly as model inputs without further transformation. Pre-processing was implemented in `code/data_loader.py`.

4.4 Statistical analysis

All analyses were implemented in Python 3.11.15. Statistical packages used: numpy 2.3.5[15], pandas 3.0.3[16], scipy 1.17.1[17], scikit-learn 1.8.0[18], lifelines 0.30.3[19], shap 0.51.0[20], matplotlib 3.10.9[21], statsmodels 0.14.6.

Random seeds: bootstrap resampling seed = 42 (fixed across all 1,000 resamples); cross-validation fold seed = 42; permutation null seed = 42 (1,000 permutations). All seeds were set at the module level before any random operation.

The primary analysis was pre-registered with SHA-256 hash `d31d27145054f151bca4a8de2b5ef85c56f3354b4d65425ea3f9f3426f490161` prior to execution.

4.5 Model construction and lockdown

The primary model is a logistic regression classifier (scikit-learn[18] `LogisticRegression`, solver = 'lbfgs', max_iter = 1000, C = 1.0, class_weight = 'balanced') trained on the 8 locked panel proteins as binary features (label 0 = non-converter, label 1 = converter). No feature selection, hyperparameter tuning, or model modification was applied after pre-registration.

4.6 Validation protocol

H1 (primary): Five-fold stratified cross-validated AUC. The cohort was split into 5 stratified folds (`StratifiedKFold`, shuffle=True, random_state=42). In each fold, the model was fit on 4 folds and evaluated on the held-out fold. Mean AUC across folds was the primary metric. Ninety-five percent CIs were derived from 1,000 bootstrap resamples of concatenated out-of-fold scores. The permuted-label null was constructed by re-running 5-fold CV with labels permuted at each of 1,000 iterations using cross-validated scoring — not in-sample — following established practice for permutation testing under CV[9]. The null 95th percentile was 0.627. The primary success threshold was AUC $\geq$ 0.85; a partial-success threshold was AUC $\geq$ 0.75.

H2 (secondary): Eight univariate Cox proportional-hazards models were fit using lifelines[19] `CoxPHFitter` (penalizer = 0.1). The event indicator was label (1 = converter, 0 = censored); duration was MonthsSinceBaseline for converters and 84 months (7 years) for non-converters. A binomial exact test (scipy[17] `binom_test`) compared the observed number of concordant HR signs against chance (p = 0.5).

H4 (secondary): HR signs from our Cox models were compared against the Gan et al. 2025[4] UKB incident-PD full-time HRs for 6 testable proteins (GRN, ICAM1, PTGDS, DKK3, SERPING1, C3) from MOESM3 Supplementary Table 4. Concordance was tested with a one-sided binomial exact test.

H5 (secondary): Three sub-panel logistic regression models were fit and evaluated identically to H1 but using protein subsets: inflammation/complement (ICAM1, C3, SERPING1), proteostasis (HSPA5, GRN), and lipid/signalling (PTGDS, DKK3, MASP2). AUC $\geq$ 0.75 was the acceptance threshold per sub-panel.

Clean-negative sensitivity analysis (post-hoc): Non-converter samples were restricted to the 14 individuals with MonthsSinceBaseline = 0 who have no post-baseline samples in the dataset (confirmed never-converters during the observation window). These 14 samples were compared against the 40 converter-labelled samples (n = 54). Five-fold CV AUC was computed with the same pipeline as H1.

GRN sanity check: Mann-Whitney U test (scipy[17] `mannwhitneyu`, alternative = 'two-sided') comparing GRN z-scores between DeNovoPD (n = 99) and Control (n = 36) samples from the Initial Study stratum of MOESM4.

SHAP values: KernelExplainer (shap[20] 0.51.0) was applied to the full-panel logistic regression model fitted on all 94 samples; mean absolute SHAP values per protein were computed across all samples.

Missing value sensitivity: Three analyses were performed, excluding proteins with >10%, >25%, or >40% per-protein missing values. All three thresholds retained all 8 panel proteins (maximum missingness < 5%), yielding identical AUC = 0.554.

4.7 Figure generation

Figures were generated by routines called from `code/main.py`, with supporting utilities in `code/data_loader.py`, `code/metrics.py`, and `code/models.py`. Individual figure outputs are stored as `002_execute/figures/fig{N}_{slug}.{pdf,png}` (N = 1–7). All figure parameters, colour maps, and annotations are determined by the code; no manual editing was performed.

Data availability

The MOESM4 supplementary abundance matrix used in this analysis is available from PRIDE accession PXD046355 (`https://www.ebi.ac.uk/pride/archive/projects/PXD046355`). The Gan et al. 2025 UKB Olink hazard ratios were sourced from MOESM3 Supplementary Table 4 of that publication (`https://doi.org/10.1038/s43587-025-00818-0`).

Code availability

All analysis code was generated in the run's execute stage (`002_execute/code/`; files: `main.py`, `data_loader.py`, `metrics.py`, `models.py`). The analysis environment is fully specified in `002_execute/code/environment.lock.json`. The pre-registration SHA-256 hash `d31d27145054f151bca4a8de2b5ef85c56f3354b4d65425ea3f9f3426f490161` can be used to verify that the analysis plan was not modified after data access.

References

[1] Y. Vermeiren, Y. Hirschberg, I. Mertens, et al. Biofluid markers for prodromal parkinson's disease: evidence from a catecholaminergic perspective. *Frontiers in Neurology*, 11:595, 2020. doi: 10.3389/fneur.2020.00595.

[2] R. B. Postuma, A. Iranzo, M. Hu, et al. Risk and predictors of dementia and parkinsonism in idiopathic rem sleep behaviour disorder: a multicentre study. *Brain*, 142:744–759, 2019. doi: 10.1093/brain/awz030.

[3] J. Rutledge, B. Lehallier, P. Zarifkar, et al. Comprehensive proteomics of csf, plasma, and urine identify ddc and other biomarkers of early parkinson's disease. *Acta Neuropathologica*, 147:52, 2024. doi: 10.1007/s00401-024-02706-0.

[4] Y. H. Gan, L. Z. Ma, Y. Zhang, et al. Large-scale proteomic analyses of incident parkinson's disease reveal new pathophysiological insights and potential biomarkers. *Nature Aging*, 5:642–657, 2025. doi: 10.1038/s43587-025-00818-0.

[5] F. Imam, R. Saloner, J. W. Vogel, et al. The global neurodegeneration proteomics consortium: biomarker and drug target discovery for common neurodegenerative diseases and aging. *Nature Medicine*, 31:2556–2566, 2025. doi: 10.1038/s41591-025-03834-0.

[6] J. Hällqvist, M. Bartl, M. Dakna, S. Schade, P. Garagnani, E. Fransén, H. Zetterberg, B. Mollenhauer, and C. Trenkwalder. An 8-protein plasma panel distinguishes irbd converters to parkinson's disease within 7 years. *Nature Communications*, 15:4759, 2024. doi: 10.1038/s41467-024-48961-3.

[7] Y. Shao, J. Wang, Y. Liu, et al. Distinct metabolomic and proteomic signatures in parkinson's disease patients with rem sleep behavior disorder. *Signal Transduction and Targeted Therapy*, 11, 2026. doi: 10.1038/s41392-026-02613-8.

[8] A. Bespalov, T. Steckler, and P. Skolnick. Be positive about negatives — recommendations for the publication of negative (or null) results. *European Neuropsychopharmacology*, 29:1312–1320, 2019. doi: 10.1016/j.euroneuro.2019.10.007.

[9] M. Ojala and G. C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

[10] Nature. In praise of replication studies and null results. *Nature*, 578:489–490, 2020. doi: 10.1038/d41586-020-00530-6.

[11] J. Buonocore, E. Fratto, F. Arcuri, et al. Plasma nfl and gfap in the preclinical stages of neurodegenerative diseases: insights from the uk biobank. *Journal of Neurology*, 272:755, 2025. doi: 10.1007/s00415-025-13498-y.

[12] C. Sudlow, J. Gallacher, N. Allen, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12:e1001779, 2015. doi: 10.1371/journal.pmed.1001779.

[13] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser. Dia-nn: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17: 41–44, 2020. doi: 10.1038/s41592-019-0638-x.

[14] J. A. Vizcaíno, A. Csordas, N. Del-Toro, et al. 2016 update of the pride database and its related tools. *Nucleic Acids Research*, 44:D447–D456, 2016. doi: 10.1093/nar/gkv1145.

[15] C. R. Harris, K. J. Millman, S. J. van der Walt, et al. Array programming with numpy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[16] W. McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010. doi: 10.25080/Majora-92bf1922-00a.

[17] P. Virtanen, R. Gommers, T. E. Oliphant, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] C. Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4:1317, 2019. doi: 10.21105/joss.01317.

[20] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. arXiv:1705.07874.

[21] J. D. Hunter. Matplotlib: a 2d graphics environment. *Computing in Science & Engineering*, 9: 90–95, 2007. doi: 10.1109/MCSE.2007.55.