Plasma proteomics in the prediagnostic phase of Parkinson's disease: a multi-cohort study of 74,000 participants

Jan Homann[1], Alexander G. Smith[2,*], Sarah Morgan[3,*], Elisabet A. Frick[4,5*], Fangyu Liu[6,*], Vivian Viallon[7,*], Rashmi Maurya[8], Roxanna Korologou-Linden[9], Valerija Dobricic[10], Olena Ohlei[1], Laura Deecke[1], Fatema Hajizadah[3], Yujia Zhao[11], Fanny Artaud[12], Karl Smith-Byrne[13], P. Martijn Kolijn[11], Jose Maria Huerta[14,15], Nils Winter[16], Marcela Guevara[15,17,18], Ana Jimenez-Zabala[19,20], María José Sánchez[15,21,22], Camino Trobajo-Sanmartín[15,17,18], Natalia Cabrera-Castro[14,15], Ana Vinagre[19,23,], Dafina Petrova[15,21,22,24], Sabina Sieri[25], Tim J. Key[13], Nick Wareham[26], Rudolph Kaaks[27], Ruth C. Travis[13], Tim Hahn[16], Susan Baker[9], Sean M. Kelly[28], Roel Vermeulen[11,29], Susan Peters[11], Giovanna Masala[30], Carlotta Sacerdote[31], Nancy Finkel[32], The Global Neurodegeneration Proteomics Consortium (GNPC)[$], Adam S Butterworth[33,34,35,36], Alexis Elbaz[12], Moritz Hess[37], Verena Katzke[27], Lars Bertram[10], Vilmundur Gudnason[4,5], Oliver Robinson[2,9], Honglei Chen[38], Lefkos Middleton[9], Laura M. Winchester[8], Ioanna Tzoulaki[2], Valborg Gudmundsdottir[4,5], Keenan A. Walker[39], Pietro Ferrari[7], Elio Riboli[2], Marc J. Gunter[2], Christina M. Lill[1,9,#]

Affiliations:

1. Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany
2. Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
3. Centre for Neuroscience, Surgery and Trauma, Blizard Institute, Queen Mary University of London, London, UK
4. Icelandic Heart Association, Kopavogur, Iceland
5. Faculty of Medicine, Department of Health Sciences, University of Iceland, Reykjavik, Iceland
6. Laboratory of Behavioral Neuroscience, National Institute on Aging, Baltimore, Maryland, USA
7. International Agency for Research on Cancer (IARC/WHO), Nutrition and Metabolism Branch, Lyon, France
8. Department of Psychiatry, University of Oxford, Oxford, UK
9. Ageing and Epidemiology Unit (AGE), School of Public Health, Imperial College London, London, UK
10. Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), University of Lübeck, Lübeck, Germany
11. Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands
12. Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France
13. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
14. Department of Epidemiology, Murcia Regional Health Council, Biomedical Research Institute of Murcia-IMIB, Murcia, Spain
15. Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
16. Institute of Translational Psychiatry, University of Münster, Münster, Germany
17. Instituto de Salud Pública y Laboral de Navarra, Pamplona, Spain
18. Navarre Institute for Health Research (IdiSNA), Pamplona, Spain
19. Biogipuzkoa Health Research Institute, San Sebastian, Gipuzkoa, Spain
20. Ministry of Health of the Basque Government, Sub-Directorate for Public Health and Addictions of Gipuzkoa, San Sebastián, Gipuzkoa, Spain
21. Escuela Andaluza de Salud Pública (EASP), Granada, Spain
22. Instituto de Investigación Biosanitaria ibs.GRANADA, Granada, Spain

23. Osakidetza Basque Health Service, Donostia University Hospital, Neurology Department, San Sebastián, Gipuzkoa, Spain
24. Servicio de Oncología Médica. Hospital Universitario Virgen de las Nieves, Granada, Spain
25. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale Dei Tumori Di Milano Via Venezian, Milan, Italy
26. Medical Research Council Epidemiology Unit, University of Cambridge, Cambridge, UK
27. Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany
28. NYU Langone Health Department of Neurology
29. University Medical Centre Utrecht, Utrecht, The Netherlands
30. Clinical Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network (ISPRO), Florence, Italy
31. Department of Health Sciences, University of Eastern Piedmont, Novara, Italy
32. Novartis, Cambridge, MA, USA
33. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
34. Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, Cambridge, UK.
35. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK.
36. NIHR Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, UK.
37. Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany
38. Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, USA
39. Laboratory of Behavioral Neuroscience, National Institute on Aging, Baltimore, Maryland, USA
*These authors contributed equally as shared second authors.
[$]For the index GNPC author list, see Supplementary Information.

[#]**Corresponding Author:**
Dr. Christina Lill, MD, MSc, Translational Epidemiology Unit, Institute of Epidemiology and Social Medicine, University of Münster, Domagkstr. 3, 48149 Münster, Germany, telephone: +49 251 83 57101, email: christina.lill@uni-muenster.de and clill@ic.ac.uk

ABSTRACT

Background: Parkinson's disease (PD) remains incurable, with a long prediagnostic phase currently undetectable by existing methods. Identifying individuals at high risk of PD would enhance our understanding of the underlying pathophysiology and will open up new preventive or early therapeutic avenues.

Methods: In the largest proteomic study in neurodegenerative diseases to date, we analyzed blood samples from ~74,000 individuals across discovery and validation cohorts. In the discovery phase, we leveraged data from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort with up to 28 years of follow-up to identify incident PD cases across five European countries. Prediagnostic plasma samples from initially healthy participants underwent high-throughput proteomic profiling (7,285 protein markers) using the SomaScan platform. Cox proportional hazards models based on 4,538 participants (including 574 incident PD cases) were used to identify protein biomarkers associated with a future PD diagnosis. In the validation phase, we tested three cohorts with incident PD cases (AGES, ARIC, UK Biobank, (n=64,856; 1,034 incident cases), a case-control dataset with prevalent cases (GNPC, n=2,592), and the longitudinal Tracking PD cohort with data on PD progression (n=794).

Findings: In the EPIC4PD discovery case-cohort, 17 proteins that predict PD up to 28 years before diagnosis were detected (FDR=0.05). Additional proteins revealed sex-specific and time-varying effects, capturing disease progression before symptom onset. Replication in three prospective cohorts confirmed at least 12 key prediagnostic biomarkers with strong evidence, including TPPP2, HPGDS, ALPL, MFAP5, OGFR, ACAD8, TCL1A, GPC4, GSTA3, LCN2, KRAS, and GJA1. Furthermore, in the longitudinal Tracking PD cohort, HPGDS and MFAP5 also predicted cognitive decline. Notably, several of the identified PD biomarkers overlapped with those for incident Alzheimer's disease and amyotrophic lateral sclerosis, indicating partially shared molecular signatures. A machine learning-derived PD protein risk score improved PD but not Alzheimer's disease risk prediction in the independent AGES cohort.

Conclusion: Our extensive proteomics effort examining PD from the prediagnostic to the progression phase identified novel, actionable biomarkers opening new avenues for early PD risk stratification and precision medicine.

Funding: Michael J Fox Foundation, Cure Alzheimer's Fund, Clinical Research in ALS and Related Disorders for Therapeutic Development (CreATe) Consortium, Michael J Fox Foundation, Interdisciplinary Centre for Clinical Research, University Münster

INTRODUCTION

Parkinson's disease (PD) is the most common movement disorder and the second most common neurodegenerative disease after Alzheimer's disease (AD)[1]. It is characterized by progressive, disabling motor and non-motor symptoms, substantially reducing quality of life and life expectancy[2,3]. By the time PD is clinically diagnosed, significant neuronal loss has already occurred, and available treatments are only symptomatic, with diminishing benefits over time[1]. Consequently, the disease poses a significant burden not only on patients but also on their families and caregivers. Furthermore, its prevalence is projected to rise sharply, particularly in aging societies, leading to increasing financial and social costs[2].

Importantly, non-specific symptoms may begin up to 20 years before diagnosis, suggesting that PD pathophysiology starts even earlier[4]. This time course points to a critical window for preventive or early therapeutic interventions that could delay or halt disease onset. However, identifying at-risk but clinically unaffected individuals remains a significant challenge. Developing a reliable PD risk score - such as one based on molecular biomarkers - will be essential to address this gap. Such a score would be especially valuable for risk stratification in vulnerable populations, including workers in rural areas with pesticide exposure or individuals with a family history of PD. By identifying those at highest risk, targeted monitoring and the initiation of early clinical and preventive trials could be implemented more effectively. In this context, high-throughput proteomic platforms such as SomaScan® (SomaLogic, Inc.) and Olink® (Olink Proteomics AB) offer new exciting opportunities to identify novel biological signatures of PD development. These methods can quantify thousands of proteins simultaneously and hold strong potential for identifying biomarkers that can detect diseases in their earliest or even preclinical stages[5–7]. In neurodegenerative diseases interest in blood-based biomarkers has grown, owing to their minimally invasive collection, easier clinical integration, and strong performance of candidate blood biomarkers in AD[8]. In PD, such markers could enhance early diagnosis, enable risk stratification in clinical trials, and support preventive strategies. However, given PD's comparatively low prevalence (2-3% in individuals aged 65+)[1], prospective cohorts with long-term follow-up are rare. Even more rare are cohorts with baseline blood samples - collected before PD onset - which are required to identify predictive biomarkers. This is why most prior biomarker studies have compared prevalent PD patients to controls, risking reverse causation and limiting predictive value[9].

142 To our knowledge, only one prospective cohort, i.e., the UK Biobank (UKB), has hitherto been
143 used to discover predictive PD biomarkers. Based on this resource, a recent study analyzed
144 2,920 plasma proteins measured with Olink technology in up to 859 incident PD patients[10].
145 Although 38 candidate biomarkers were identified, the findings of that study remain
146 unvalidated in independent prospective cohorts. In addition to closing this gap, our study
147 represents a major advance by nearly tripling the number of protein markers (n=7,285;
148 SomaScan 7K array) analyzed in prediagnostic blood samples from ~4,600 participants in a
149 large, well-characterized case-cohort ('European Investigation into Cancer and Nutrition for
150 Parkinson's disease', EPIC4PD)[11] within the EPIC study[12,13]. Overall, EPIC4PD features up to
151 28 years of follow-up time and 574 specialist-confirmed incident PD diagnoses. Crucially, we
152 externally validated our top biomarkers in ~65,000 individuals, including >1,000 incident PD
153 cases, across three independent prospective cohorts. Overall, this resulted in the largest and
154 most comprehensive investigation of predictive biomarkers for PD to date, both in terms of
155 sample size and proteomic resolution. To assess these markers across the disease timeline -
156 including diagnosis and progression - we also incorporated independent SomaScan 7K data
157 from the Global Neurodegenerative Proteomics Consortium (GNPC; ~300 prevalent cases
158 and ~2,300 controls)[14] and the Tracking PD cohort (~800 cases with progression data). Using
159 the top results from these analysis arms, we also performed enrichment and network
160 analyses to gain insights into biological pathways, and evaluated PD biomarker profiles in AD
161 and amyotrophic lateral sclerosis (ALS) to assess their potential as shared markers of
162 neurodegeneration (**Figure 1**). Finally, using machine-leaning we constructed and externally
163 validated a PD protein risk score (PPRS) in EPIC4PD. This score improved risk prediction
164 decades before clinical PD onset. Combined with traditional risk factors and markers from
165 other molecular domains, the biomarkers identified in this study have the potential to
166 substantially advance ultra-early PD risk stratification and diagnosis paving the way for early
167 treatment and prevention.

RESULTS

EPIC4PD case-cohort

In the discovery phase of this study (**Figure 1**), we leveraged data from the EPIC cohort to build a large case-cohort (EPIC4PD) of 4,538 participants (including 574 incident PD cases; **Table 1**) across 10 European centers in Spain, Italy, the Netherlands, the UK, and Germany, combining deep proteomic profiling (6,382 proteins via 7,285 aptamers) in baseline, i.e, pre-disease plasma samples, with long-term prospective follow-up. EPIC4PD participants had a median baseline age of 53 years (range: 35–79), comprising 63% women (57% excluding the Utrecht center, which recruited only women), and were followed for up to 28 years (median: 16 years). Among PD cases, the median time from blood sampling to diagnosis was 11 years (range: 2-25), enabling biomarker discovery long before clinical onset. The age distribution at diagnosis in women and men was highly similar (median diagnosis age in each group: 71, interquartile range 65-76). As expected and consistent with known PD risk profiles, PD cases were older at baseline (median: 59 vs. 52 years), more likely male (58% vs. 40%, excluding Utrecht), less likely to be current smokers (15% vs. 25%), and slightly more often overweight (64% vs. 60% with body mass index (BMI) ≥25) compared to non-cases (**Supplementary Tables 1–2**).

Association analyses in the full EPIC4PD dataset

We used Cox proportional hazards regression models across 7,285 aptamers to identify proteins associated with a future PD onset (**Supplementary Table 3**). In the basic model (stratified by age in 5-year intervals, sex, and center) covering the full follow-up period (2–28 years), 17 proteins were significantly associated with PD risk after a false-discovery rate (FDR) control of 0.05 (**Table 2, Supplementary Table 4**). We applied two approaches in the discovery phase to address outliers: one excluded extreme values as potential noise, and the other retained them via capping, as they may represent strong PD signals. Nine proteins were FDR-significant in both approaches; for the other eight (six upon exclusion, two upon capping), the alternative model still showed consistent, nominally significant associations with smiliar effect size estimates (mean ΔHR = 0.03, maximum ΔHR=0.13; **Supplementary Figure 1A**), supporting robustness of our findings. For simplicity, we report only one result (from the better-performing model) in the main text, with capped data labeled "cap"; full results are available in **Supplementary Table 3**. Top proteins associated with PD incidence included VAMP3 (hazard ratio [per 1 standard deviation (SD) increase in $\log_{10}$-transformed

200 protein levels]=1.26, p=1.46E-08), TPPP2 (HR=0.70, p=8.52E-08), HPGDS (HR=0.73, p=8.59E-
201 08), $ALPL_{cap}$ (HR=1.25, p=1.30E-07), FCAR (HR=1.34, p=3.48E-07), and CERS5 (HR=1.23,
202 p=5.64E-06; **Figure 2A+D**, **Table 2**). All FDR-significant associations remained stable after
203 additional adjustment for key demographic and lifestyle factors (education level, BMI,
204 smoking, and physical activity), with minimal impact on effect sizes (mean ΔHR=0.02,
205 maximum ΔHR=0.07; **Supplementary Figure 1**, **Supplementary Table 4**).

206 **External replication of predictive biomarkers in the full follow-up period**

207 Of the 17 FDR-significant biomarkers identified in EPIC4PD, replication data were available in
208 a total of 64,856 participants (including 1,034 incident PD cases) from three prospective
209 studies: the Gene/Environment Susceptibility–Reykjavik Study (AGES), Atherosclerosis Risk in
210 Communities study (ARIC), and UKB. Specifically, all 17 proteins were available in AGES
211 (5,090 participants, 121 cases; SomaScan 7K), 15 in ARIC (11,205 participants, 164 cases;
212 SomaScan 5K), and 8 in UKB (48,561 participants, up to 749 cases; Olink 3072 Explore;
213 **Supplementary Table 5**, **Supplementary Figure 2**). The other 9 proteins were not measured
214 in UKB, as the current Olink platform captures only ~30% of the proteins assayed by
215 SomaScan 7K. We also included independent SomaScan 7K-based data from prevalent cases
216 as 'indirect' replication. To this end, we re-analyzed proteomic profiles from 2,592
217 participants (335 prevalent PD cases and 2,257 controls) across three GNPC case-control
218 datasets (also see below for comparison of incident and prevalent PD), using similar
219 methodology as with our discovery phase analyses.

220 Given the higher comparability of results from datasets generated using the same
221 technology, we first considered replication evidence in the SomaScan-based datasets (AGES,
222 ARIC, and GNPC). Overall, nine of the 17 proteins - TPPP2, HPGDS, ALPL, MFAP5, OGFR,
223 ACAD8, VAMP3, GTF2A2, and FGF8 showed strong or moderate evidence for replication, i.e.,
224 nominally significant evidence for association in at least one dataset and/or fully consistent
225 effect estimates across all independent studies with available data (**Figure 3A**, **Table 2,**
226 **Supplementary Table 5**). Upon reviewing the replication evidence in the UKB, the evidence
227 of association was further substantiated for TPPP2, HPGDS, MFAP, and OGFR, while ALPL,
228 ACAD8, VAMP3, GTF2A2, and FGF8 were not measured by Olink. Several additional proteins
229 showed suggestive replication evidence in individual datasets, but the overall replication
230 evidence was more heterogeneous (**Table 2**, **Supplementary Figure 2, Supplementary Table**
231 **5**). Overall, platform-related differences between Olink and SomaScan must be considered.

232 In line with the previous UKB study[10], we observed largely decreased protein signals in
233 incident PD cases in our re-analyses of the UKB Olink data (i.e., 87% of nominally significantly
234 associated proteins) – a pattern was not seen in SomaScan-based datasets from EPIC4PD,
235 AGES, or ARIC (proportion of proteins with a decreased protein signal for incident PD: 43% to
236 56%; **Supplementary Table 6**), underscoring the challenges of comparisons across different
237 protein platforms (see **Discussion**).
238 In summary, nine of the 17 biomarkers (53%; ACAD8, HPGDS, TPPP2, OGFR, MFAP5, ALPL,
239 FGF8, VAMP3, GTF2A2) were replicated across different technologies, populations, and
240 study designs, reinforcing their specific predictive potential, and, more generally, support
241 the generalizability of our biomarker signals.

242 **Biomarker associations relative to disease onset**

243 Stratifying analyses by temporal proximity to PD diagnosis revealed remarkable time-varying
244 biomarker effects. In the window closest to disease onset (2-10 years; n=4,230 participants
245 including 245 incident cases), three additional proteins - IL11, TCL1A, and $TIMP1_{cap}$ -
246 emerged as FDR-significant. In the more distal window (10-20 years; n=4,023 including 307
247 incident cases), 19 additional proteins showed strong associations with PD incidence (**Figure**
248 **2B+C, Table 2**, **Supplementary Table 4**). These proteins had not been identified as significant
249 overall due to negligible effects in the other time window (**Supplementary Table 3**). Time-
250 varying biomarker signals were further supported by HR pseudotrajectories derived from 5-
251 year sliding windows (**Figure 2B+C, Supplementary Figure 3**). Consequently, we performed
252 additional exploratory statistical analyses, including interaction analyses with time-to-event
253 and spline regressions to assess non-linear association effects. Importantly, a substantial
254 number of biomarkers (all biomarkers from the proximal window and nine biomarkers from
255 the distal window) demonstrated nominally significant interaction results and/or evidence
256 from spline regressions for non-linear associations, providing enhanced statistical support
257 for time-varying effects (**Supplementary Table 7, Supplementary Figure 4**). Together, these
258 findings demonstrate that key biomarkers for PD may emerge at distinct prediagnostic
259 phases, underscoring the value of deep proteomic profiling with extended follow-up to
260 reveal dynamic, stage-specific signatures of prediagnostic disease.

261 **External replication of the temporal analysis results**

262 Of the three newly emerging biomarkers from the proximal follow-up window (2-10 years
263 before diagnosis), TCL1A showed strong independent replication evidence and TIMP1
264 showed moderate support, while IL11 did not replicate (**Figure 3A**, **Supplementary Figure 2,**
265 **Supplementary Table 5**). Generally, the amount of replication data for the distal time
266 window (10-20 year window) was more restricted due to shorter follow-up durations in
267 some of the replication cohorts (**Methods** and **Supplementary Table 5**). Among the
268 biomarkers emerging from the 10-20 year window, GPC4, GSTA3, and LCN2 exhibited strong
269 evidence of replication, and ABHD12 showed moderate support (**Table 2, Figure 3A**). For
270 LCN, we observed nominally significant evidence for replication in ARIC with the same effect
271 direction as in EPIC4PD. In UKB, LCN2 also reached nominal significance (p=3.80E-03) but
272 showed an inverse direction of effect (**Figure 3A**, **Supplementary Table 5**). This discrepancy
273 likely reflects methodological differences between the SomaScan and Olink proteomic
274 platforms (see **Discussion**).

275 **Sex-stratified analyses**

276 In EPIC4PD, 18% (3 of 17) of proteins significantly associated with PD risk in the full follow-up
277 period showed clear sex-specific effect size differences and significant interaction with sex
278 after FDR control. These proteins included $NAMPT_{cap}$ (ΔHR=48%), FCAR (ΔHR=42%), and
279 $ALPL_{cap}$ (ΔHR=34%; **Supplementary Figure 1G**, **Supplementary Table 8**).
280 Furthermore, sex-stratified proteome-wide analyses uncovered six additional proteins (RS1,
281 $KRAS_{cap}$, LYZ_{cap}, $GBA3_{cap}$, GJA1, AMN) FDR-significantly associated with PD risk in men and
282 two (CXCL11, ITGA6) in women, none of which showed strong associations in the opposite
283 sex (HRs 0.92–1.10, all $p \geq 0.05$; **Supplementary Figure 1G**, **Supplementary Figure 5;**
284 **Supplementary Table 4**). Interaction analyses confirmed FDR-significant sex-based risk
285 modifications for all male-specific biomarkers except GJA1. However, the latter and CXCL11
286 still showed nominally significant evidence for interaction (**Supplementary Table 8**). The
287 results of these 11 FDR-significant proteins in the sex-stratified analyses remained stable
288 after adjusting for lifestyle variables and, in women, for menopausal status and
289 postmenopausal hormone therapy (**Supplementary Figure 1D+E**, **Supplementary Table 4**).
290 While the amount of replication data was more limited due to smaller strata in the
291 replication cohorts, $KRAS_{cap}$ and GJA1 demonstrated strong evidence of replication in men,
292 while AMN (men) and CXCL11 (women) provided moderate evidence for replication **(Figure**
293 **3A)**. Collectively, these findings underscore the importance of accounting for sex-specific

294 effects in biomarker discovery, as such effects may reveal associations that remain hidden in
295 sex-combined analyses.

296 **Inter-correlations of predictive biomarkers in EPIC4PD**

297 Several FDR-significant biomarkers showed moderate inter-correlations (r=0.4-0.6) in
298 EPIC4PD, suggesting shared regulation or involvement in the same pathway. Correlations
299 were observed between MFAP5 and both HPGDS and TPPP2 (r=0.54 each) and between
300 LCN2 and LTF (r=0.60). FCAR also featured in several correlation clusters, underscoring its
301 broader connectivity (detailed in **Supplementary Figure 6** and **Supplementary Table 9)**.

302 **Characterization of prediagnostic biomarkers in prevalent PD and PD progression**

303 Next, we investigated whether the effects of the identified prediagnostic biomarkers
304 extended beyond the risk phase into the symptomatic stage of PD by leveraging the
305 SomaScan 7K proteomic data from 2,592 independent GNPC participants (335 prevalent
306 cases and 2,257 controls; see the ref. [14] for full GNPC dataset description). Of all 47
307 prediagnostic biomarkers detectable in GNPC, 7 (15%; TPPP2, HPGDS, ALPL, NAMPT, GSTA3,
308 KRAS, GJA1) were nominally significantly associated with prevalent PD, and 4 (TPPP2,
309 HPGDS, NAMPT, GSTA3) remained significant after FDR control. All but NAMPT showed
310 consistent effect directions. Overall, 38 (81%) of the 47 top proteins and 125 (86%) of the
311 146 nominally significant aptamers results shared between both traits showed concordant
312 directions of effect - significantly more than expected by chance (p=1.25E-5 and p=1.64E-19,
313 respectively; one-sided exact binomial test; **Figure 3B, Supplementary Table 10**). In serum
314 samples from the Tracking Parkinson's cohort (n=794), we assessed associations between
315 FDR-significant prediagnostic biomarkers and clinical outcomes related to disease
316 progression, using motor (Unified Parkinson's Disease Rating Scale Part III, UPDRS-III) and
317 cognitive (Montreal Cognitive Assessment, MoCA) scores in cross-sectional and longitudinal
318 mixed-effects models. Two biomarkers, HPGDS (β=2.61, p=6.82E-04) and MFAP5 (β=1.31,
319 p=1.49E-03), were FDR-significantly associated with higher MoCA scores over time,
320 indicating that lower levels of these proteins predict both increased PD risk and faster
321 cognitive decline (**Figure 3E**; **Supplementary Table 11**). Associations remained robust after
322 adjusting for BMI and disease duration (data not shown). These results demonstrate that a
323 considerable fraction of prediagnostic biomarkers persist in their dysregulation in prevalent
324 PD, and a small subset is even associated with progression. Comparing incident and

325 prevalent disease settings, such as between EPIC4PD and GNPC, is essential for determining
326 which biomarkers are robust across the disease continuum and which are specific to
327 particular stages. This distinction enables identification of biomarkers suitable for early
328 detection versus those better suited for diagnosis or disease monitoring.

329 **Comprehensive assessment of previously described biomarker candidates**

330 Of 291 unique proteins previously linked to PD across methodologically diverse sources[10,15–
331 17], 191 proteins were assessable in the EPIC4PD data (**Supplementary Table 12**): Most
332 notably, 38 proteins had recently been reported as putative predictive PD biomarkers in the
333 UKB Olink study[10], with 27 also covered in our independent dataset. Of these, 11 (41%;
334 HPGDS, CD209, TPPP3, EPS8L2, FCRL1, BAG3, TNXB, IL13RA1, NEFL, PPY, RET) showed at
335 least nominally significant associations in one or more follow-up intervals in EPIC4PD. HPGDS
336 also represents a top biomarker in EPIC4PD (**Table 2**). The next two most significant proteins
337 were CD209 (best model for follow-up of 10-20 years: HR=0.75, p=1.32E-04) and TPPP3 (best
338 result: 2-28 years: HR=0.80, p=1.58E-03), which aligns well with TPPP2, one of the lead
339 findings in EPIC4PD. All but one (PPY) of the at least nominally significantly associated
340 biomarkers showed the same effect direction as in the previous study[10] despite platform
341 differences. Of the 127 candidate neurological biomarkers from the CNS NULISA panel
342 (Alamar, Inc), 106 were tested in EPIC4PD; of these, UBE2L3|UBB and CCL11 represented
343 top biomarkers in EPIC4PD (**Table 2**). Other nominally significant findings included for
344 instance SNCA, DDC, NEFL, TREM2, GFAP, and TIMP3. Furthermore, PYDC1 nominated by a
345 recent Mendelian randomization (MR) and colocalization study[18] was among the top results
346 in EPIC4PD suggesting a causative link to PD risk. Several other candidates from an additional
347 MR[16] or GWAS[17] showed nominal associations in EPIC4PD (**Supplementary Table 12**).
348 Conversely, 5 of the 39 FDR-significant PD biomarkers identified in our non–sex-stratified
349 EPIC4PD analyses (CRYGD, LTF, HPGDS, TCL1A, and LCN2) were also included in the MR study
350 based on SomaScan data[16]. Of these, two biomarkers (CRYGD and LTF) showed nominal
351 evidence of a causal association with PD, suggesting that higher circulating levels may
352 increase disease risk, in line with the EPIC4PD findings (**Supplementary Table 13**). The
353 limited overlap in assessed proteins likely reflects the comparatively stringent analysis
354 criteria applied in that study [16].

355 **Cross-disease comparisons**

356 The 47 FDR-significant PD biomarkers identified in EPIC4PD were also evaluated in incident
357 AD (n=1,926, including 652 cases) and ALS (n=4,715, including 187 cases) using Cox
358 regression analyses in the EPIC4AD and EPICALS case-cohorts (**Supplementary Table 1**),
359 respectively. Sixteen (34%) prediagnostic PD biomarkers showed at least nominally
360 significant associations with AD and/or ALS, all with concordant effect directions. Correlation
361 plots of hazard ratios showed alignment in effect size estimates. Notably, three PD proteins -
362 RSPO2, $SPAST_{cap}$, PYDC1 - were FDR-significant for AD and also showed nominally significant
363 associations with ALS (**Table 2, Figure 3C, Supplementary Table 14**).
364 At the proteome-wide level, 17% of proteins nominally significantly associated with PD over
365 the full follow-up period (non-sex-stratified) also showed associations with AD and/or ALS,
366 whereas 83% appeared to be PD-specific. Among the shared proteins, the majority (68%)
367 overlapped exclusively with AD, 25% with ALS, and 7% were common to all three
368 neurodegenerative diseases (**Supplementary Figure 7**). Nominally significant prediagnostic
369 AD biomarkers were significantly enriched among nominally significant PD biomarkers
370 (p=1.19E-03; Fisher's exact test), while ALS biomarkers were not (p=0.107). Notably, 85 of 92
371 PD–AD and 33 of 39 PD–ALS shared proteins showed concordant effect directions (p=1.92E-
372 18 and p=7.15E-06; exact binomial test; **Figure 3C**). Additionally, among proteins nominally
373 associated with PD, effect sizes modestly correlated with AD and ALS (Pearson's r=0.18 [95%
374 CI: 0.11–0.25] and r=0.21 [0.14–0.28], respectively; **Figure 3C**). While these comparisons
375 highlight a largely PD-specific biomarker signature, they also point to partially shared
376 prediagnostic pathways.

377 **Functional characterization of prediagnostic PD biomarkers**

378 Gene ontology (GO) enrichment analysis of suggestive PD biomarkers (n=154 proteins with
379 p<1.0E-03 and consistent effect directions; see Methods) did not yield FDR-significant results
380 but revealed 68 gene sets with strong nominal enrichment (p<1.0E-02), highlighting
381 numerous biologically plausible and likely disease-relevant pathways based on prior
382 evidence. Strikingly, despite the long time between blood draw and clinical diagnosis, we
383 observed several well-established mechanisms in PD among the top 'biological processes'
384 categories, including vesicle trafficking (e.g., endoplasmatic-reticulum-to-Golgi transport,
385 vesicle docking), dopaminergic neuron differentiation, microtubule dynamics, and immune-
386 related responses (e.g., to interleukin-6, cytokines, neutrophil-mediated immunity). Lastly,
387 several vitamin-related processes, including response to vitamin D, also emerged, pointing to

388 molecular pathways that were especially prominent in earlier PD research. Enrichments in
389 'cellular component' and 'molecular function' categories reinforced the role of vesicle-
390 mediated activity, while Reactome pathway analyses further emphasized immune system
391 involvement, including antimicrobial peptides (**Supplementary Tables 15**).
392 Expression heatmaps across 50 human bulk tissues and single brain cell types based on RNA-
393 seq data revealed relatively high expression of several FDR-significant biomarkers in
394 peripheral and central immune-related tissues and cells compared to other tissue types
395 (**Supplementary Figures 8-9**). Protein-level data from the Human Protein Atlas supported
396 this observation (**Supplementary Figure 10**). Consistent with these findings, single-cell RNA-
397 seq–based enrichment analyses of FDR-significant and suggestive biomarkers in mice and
398 humans supported a role for immune cells in prediagnostic PD. In mouse cortex, suggestive
399 biomarkers showed nominally significant enrichment in microglia (**Supplementary Table 16**).
400 Analyses across 23 representative human cell types (including immune, neuronal, intestinal,
401 muscle, liver, and fibroblast cells) demonstrated nominally significant enrichment in
402 fibroblasts, monocytes and Kupffer cells, with borderline significance in granulocytes and
403 macrophages (**Supplementary Table 16**). Collectively, these single-cell RNA-seq–based
404 findings align with our Gene Ontology and Reactome pathway results, highlighting early
405 immune system involvement as a key feature of prediagnostic PD.
406 Interestingly, protein–protein interaction (PPI) analysis revealed that FDR-significant
407 biomarkers were significantly more likely to interact with GWAS-nominated PD risk genes[17]
408 than expected by chance (p=1.0E-04; permutation test). Within this network, SNCA emerged
409 as a central hub, exhibiting several direct edges with FDR-significant biomarkers (**Figure 3D**,
410 **Supplementary Table 17**). Consistently, *SNCA* expression in both brain and immune cells was
411 correlated with several of these biomarkers (**Supplementary Table 18**). In brain, strong
412 correlations (r≥0.5) were observed with RS1, RSPO2, and OSTF1, and in immune cells, with
413 LYZ, VAMP3, F13A1, and FCAR. In paired samples of healthy participants from the GNPC,
414 several proteins (GTF2A2, LCN2, LYZ, TIMP1, OGFR, NAMPT, and CULB4B) exhibited
415 moderate to strong correlations between blood and CSF (**Supplementary Table 19**). These
416 findings provide converging evidence that the identified prediagnostic biomarkers capture
417 early pathophysiological processes consistent with independent lines of prior research.

418 Finally, a semi-systematic literature search revealed prior functional evidence for a major
419 part of the 47 prediagnostic PD biomarkers (details in the **Supplementary Material**
420 [**Supplementary Table 20**]).

421 **Druggability assessment of prediagnostic biomarkers**

422 Given their potential pathophysiological implications, we next evaluated whether any of the
423 47 FDR-significant prediagnostic PD biomarkers represent druggable targets using the Open
424 Targets database[19]. Twenty-one proteins (45%) showed at least one supporting line of
425 evidence for interaction with small-molecule drugs (**Supplementary Table 21**). Notably, four
426 proteins - NAMPT, KRAS, CD3G, and F13A1 - were associated with compounds currently
427 undergoing clinical trials in non-neurological diseases (mostly cancer), underscoring their
428 translational potential. These findings indicate that a substantial share of prediagnostic PD
429 biomarkers may be pharmacologically modifiable.

430 **Disease prediction**

431 To assess predictive utility, we fitted penalized Cox models based on least absolute shrinkage
432 and selection operator (LASSO)-based machine learning in EPIC4PD, excluding the
433 Heidelberg center as training and using it as test dataset. The external validation of the best
434 performing PD protein risk score (PPRS) from the test data (as determined based on the
435 concordance [C-]index) in the independent AGES cohort (n=5,090; 121 incident PD cases),
436 which also used the SomaScan 7K platform, yielded an improvement of the C-index of 0.07,
437 i.e., from 0.65 (baseline based on age and sex only) to 0.72 (baseline + 8 proteins) in the
438 overall sample. This improvement is statistically highly significant (p=4.66E-09, likelihood
439 ratio test). Three of the eight proteins from the PPRS (i.e., TPPP2, HPGDS, VAMP3) also
440 appear as top results in the single-protein analyses (**Supplementary Tables 22-23**). In AGES,
441 the PPRS translates to a nearly 6-fold increase of PD risk in the top compared to the bottom
442 PPRS quintile during the AGES follow-up of up to 17 years prior to disease diagnosis
443 (HR=5.70, p= 6.62E-07). Importantly, for incident AD in AGES, the increase in C-index using
444 the eight-protein PPRS was negligible (ΔC-index=0.0019), suggesting specificity of the score
445 for PD despite overall concordant effect directions across the two neurodegenerative
446 endpoints (**Supplementary Table 23**).

447 **DISCUSSION**

448 In this large, prospective, and multi-center proteomic study, we identified and
449 independently validated novel plasma biomarkers predictive of PD onset up to 28 years
450 before clinical diagnosis. Using the high-resolution SomaScan 7K platform in prediagnostic
451 samples from over 4,500 EPIC4PD participants, we discovered 47 proteins significantly
452 associated with incident PD. Several top markers - TPPP2, HPGDS, ALPL, MFAP5, OGFR,
453 ACAD8, TCL1A, GPC4, GSTA3, LCN2, KRAS, and GJA1 - showed strong replication across
454 incident (AGES, ARIC, UKB) and prevalent (GNPC) PD cases, underscoring their robustness
455 across platforms and populations. Importantly, owing to the much larger proteomic
456 resolution our study uncovered numerous novel PD biomarkers not assayed in the previous
457 UKB analyses[10]. In addition, while we observed that prediagnostic protein biomarker profiles
458 are partially shared between PD and other neurodegenerative diseases (i.e., AD and ALS),
459 the delineated PD risk score appears to be disease-specific. Overall, our study delivers the
460 most comprehensive proteomic resource for prediagnostic PD to date. It is unique by
461 incorporating diverse prediagnostic validation cohorts, by spanning the full disease course,
462 and by demonstrating some overlap with the prediagnostic phase of other
463 neurodegenerative disorders. As such it sets a new benchmark for biomarker discovery in
464 neurodegeneration.

465 Our findings offer several new key insights: First, predictive PD biomarkers can be reliably
466 identified in blood many years, even decades, before diagnosis. In this context, signals such
467 as HPGDS, MFAP5, OGFR, and ACAD8 appear to remain stable across different populations
468 and proteomic platforms. Furthermore, we also observed dynamic, stage-dependent
469 biomarker association patterns: while some protein signals persist throughout all stages,
470 others appear only within certain time windows before diagnosis (such as IL11, TIM1, and
471 TCL1A in the 2-10 year window, and several others in the 10-20 year window prior to
472 diagnosis), possibly reflecting underlying pathophysiological shifts. Interestingly, all three
473 biomarkers from the proximal time window are linked to immune processes and
474 inflammation (as detailed in the **Supplementary Material**). Furthermore, it is notable that
475 we observed a high degree of agreement (>80%) in effect size directions between
476 prediagnostic and clinical biomarkers across the independent EPIC4PD and GNPC datasets
477 (both SomaScan 7K). Additionally, two of the prediagnostic biomarkers (HPGDS and MFAP5)

478 also predicted longitudinal cognitive decline in the course of PD. These comparisons - which
479 uniquely span the broad spectrum from preclinical PD to disease progression - enable the
480 identification of biomarkers relevant across all stages versus those that are stage-specific,
481 thereby informing their optimal use for early detection, diagnosis, and disease monitoring.
482 Second, we observed evidence for sex-specific effects for several biomarkers in EPIC4PD,
483 with predominance in men, which were not driven by other relevant covariates such as
484 smoking or BMI. This is interesting in light of the increased prevalence of PD in men and
485 warrants detailed follow-up in future investigations.
486 Third, the substantial concordance of findings across different independent SomaScan-based
487 datasets supports the robustness of our findings. In this context, we note that the previous
488 UKB study[10] reported a consistent decrease of top prediagnostic biomarkers in incident PD
489 cases when compared to non-cases. While we were able to confirm this pattern in our re-
490 analysis of the publicly available UKB data, we did not observe this in any of the four
491 SomaScan-based datasets. Likely, this discrepancy reflects methodological differences
492 between the currently available proteomic platforms: While recent evidence suggests that
493 the latest SomaScan platforms provide higher measurement precision when compared to
494 Olink[20–24], it is important to note that both platforms frequently capture distinct protein
495 features, underscoring their complementary strengths and limitations[20] (see **Supplementary**
496 **Text** for a more detailed discussion on this). A better understanding of these differences will
497 require large paired SomaScan and Olink datasets as well as benchmarking against mass
498 spectrometry. Until such data are available, it is important to recognize that – although
499 being genuine – certain protein associations may fail to replicate or show opposite effects
500 when compared across these two technologically different platforms. Thus, while orthogonal
501 replication is confirmatory, lack of replication across datasets using these two different
502 platforms does not invalidate the observed signals.
503 Fourth, we were in the unique position to probe for an overlap of prediagnostic biomarker
504 profiles of PD and two other neurodegenerative diseases (AD and ALS) owing to the design
505 of the EPIC4ND study. About one-third of PD biomarkers also showed consistent associations
506 with incident AD and/or ALS, including RSPO2, SPAST, and PYDC1. Even though effect sizes
507 correlated only modestly across diseases, the enrichment of shared biomarkers - especially
508 between PD and AD - suggests partially overlapping prediagnostic mechanisms. While a
509 limited biomarker overlap has recently also been reported by the GNPC for the symptomatic

510 phase of these two diseases[25,26]; our results are the first to demonstrate that this is already
511 detectable in the prediagnostic phase. Notwithstanding these overlaps, we emphasize that
512 most identifed biomarkers were specific to PD, and the PD-specific protein prediction score
513 was only predictive for PD but not for AD as demonstrated in the independent AGES cohort.
514 Fifth, it appears that disease-related mechanisms are already active and detectable in blood
515 more than two decades before diagnosis. A wide array of *in silico* analyses supporting the
516 biological plausibility of many of our biomarker findings showed, for instance, an
517 overrepresentation of key PD-relevant processes, including vesicle trafficking, dopaminergic
518 neuron biology, and immune signaling. Furthermore, cell-type enrichment analyses
519 implicated the involvement of microglia and peripheral immune cells already in the earliest
520 disease phases, aligning with current neuroinflammation models of PD[27,28]. Moreover, our
521 PPI analyses suggest that the prediagnostic biomarkers interacted more frequently than
522 expected by chance with known PD proteins and correlated with *SNCA* expression, strongly
523 emphasizing their functional relevance to the disease. Along these lines, several of our newly
524 detected prediagnostic biomarkers had been implicated in other studies in PD before. For
525 example, TCL1A[29], GPC4[30], LCN2[31], AMN[32], CXCL11[33], CUL4B[34], FCAR[35], BECN1[36], and
526 NAMPT[37] have previously been reported as potential blood biomarkers in (prevalent) PD in
527 studies with smaller sample sizes. In addition, some of the prediagnostic PD biomarkers were
528 reported to show differential expression in post-mortem brain tissue from PD patients
529 compared to controls (e.g., LCN2[38], GJA1[39], TIMP1[40], AMN[32]). Lastly, several of our top
530 biomarker proteins are implicated in microtubule dynamics (TPPP2), vesicle trafficking in the
531 presynapsis (VAMP3), mitochondrial functions (ACAD8), dopaminergic neuronal function and
532 maturation (GPC4, FGF8) and immune functioning/neuroinflammation (CXCL12, IL11, TIM1,
533 and TCL1A), which are all highly relevant in PD pathophysiology[41]. A detailed summary of key
534 experimental findings and supporting evidence, can be found in the **Supplementary**
535 **Material**.
536 Sixth, our druggability assessments revealed that nearly half of the FDR-significant
537 biomarkers show evidence for small-molecule interactions. While blood-based biomarkers of
538 brain diseases (like PD) do not automatically translate into therapeutic targets, this
539 assessment may help prioritize candidates for future functional studies and repurposing
540 efforts. The list of biomarkers we identified as correlating between blood and CSF may also
541 aid in prioritization efforts.

542 Lastly, using machine-learning, we developed a PD-specific protein risk score comprising
543 eight proteins, which showed robust predictive value in external validation within the AGES
544 cohort within a follow-up of up to 17 years. Along these lines, existing tools, such as the
545 updated MDS criteria (specialist settings) and the PREDICT-PD algorithm (broader
546 screening)[42,43] based on combinations of lifestyle, genetic, and/or clinical variables show
547 promise but remain insufficient, particularly for detecting prediagnostic PD more than five
548 years prior to diagnosis (e.g., ref. [44]). These tools are continuously evolving, and integrating
549 our PD-specific protein risk score offers an immediate translational next step to improve the
550 classification of prodromal PD.

551 Our study has several strengths, including representing the largest collection of incident PD
552 proteomic datasets to date, an exceedingly long-term follow-up, rigorous case
553 ascertainment, the inclusion of several external datasets ensuring population and platform
554 cross-validations, as well as extension to other disease phenotypes and diseases.
555 Notwithstanding, the results of our study must be interpreted with the following potential
556 limitations in mind. First, use of distinct proteomic platforms with different technologies,
557 particularly SomaScan versus Olink, and coverages, comparatively limited case numbers
558 despite large replication cohorts, usage of different blood sample types (citrate plasma,
559 EDTA plasma, or serum) as well as differences in study design (e.g., recruitment, case
560 ascertainment, follow-up times) may have contributed to non-replication of some
561 biomarkers. At the same time, the biomarkers that did show replication across the datasets
562 analyzed can be considered particularly robust. Second, while the SomaScan 7K assay offers
563 extensive coverage, it is semi-quantitative. This means that signal changes may reflect
564 different proteoforms rather than absolute protein levels. Mass spectrometry-based
565 approaches are needed to clarify the underlying biology. Third, although we carefully
566 adjusted for key covariates in numerous sensitivity analyses, residual confounding remains
567 possible. Fourth, lacking neuropathological confirmation, some PD cases may have been
568 misclassified - a common problem in large cohorts and a general limitation in
569 epidemiological studies. However, given that medical charts of PD cases were reviewed by
570 clinical specialists, the risk of misclassification in EPIC4PD is likely lower than in cohorts such
571 as UKB. Finally, all datasets consisted exclusively or predominantly of non-Hispanic White
572 participants. While the ARIC study also included a subset of Black participants (n=2,625) and
573 their exclusion did not affect effect estimates, the sample size was too limited to allow for

574 dedicated ethnicity-specific analyses of this subset. Thus, while our findings are likely
575 generalizable to populations of European ancestry, their applicability to other ethnic groups
576 requires further assessment.

577 In conclusion, our study newly identifies and validates a number of robust blood-based
578 biomarkers that predict PD several decades before clinical onset. Our results underscore the
579 importance of applying prospective cohort designs with incident case ascertainment to
580 detect genuinely *pre*diagnostic biomarkers. By capturing early molecular changes, our work
581 lays the foundation for developing novel prevention strategies applying precision medicine
582 to PD and potentially other neurodegenerative diseases.

583 **Acknowledgements**

584 The authors thank all EPIC participants and the countless scientists who have contributed to
585 the EPIC cohort over the past 30 years. We thank the National Institute for Public Health and
586 the Environment (RIVM), Bilthoven, the Netherlands, for contributing cases and ongoing
587 support to the EPIC study. We sincerely thank Bertrand Hemon for his assistance with data
588 management, and Prof. Paolo Vineis, Prof. Beate Ritz, Prof. Martin Wolkewitz, Prof. Klaus
589 Berger, Prof. Valentina Gallo, Dr. Mark Frasier, and Dr. Lazaros Belbasis for helpful
590 discussions.

591 **Data access**

592 Data can be accessed by external researchers after approval by the EPIC4ND working group
593 and the EPIC steering committee. Potential collaborators can contact the EPIC4ND working
594 group chair, Christina Lill (christina.lill@uni-muenster.de and clill@ic.ac.uk) and EPIC
595 administrator Sherry Morris (epicadmin@imperial.ac.uk).

596 **Funding**

597 Updates of the PD case ascertainments, generation and processing of proteomic data were
598 funded by the Michael J Fox Foundation (#008994 to C.M.L. and E.R.). Part of the proteomic
599 data were also generated with funding from the Cure Alzheimer's Fund (to C.M.L. and L.B.),
600 and the 'CReATe- Clinical Research in ALS and Related Disorders for Therapeutic
601 Development' Consortium (to C.M.L. and L.B.). CReATe (U54 NS092091) is part of the Rare
602 Diseases Clinical Research Network (RDCRN), an initiative of the Office of Rare Diseases
603 Research (ORDR), NCATS and funded through collaborations between NCATS, NINDS, and
604 the ALS Association. Intramural funding was provided by the Interdisciplinary Centre for
605 Clinical Research, University Münster (to C.M.L.). Additional funding was supplied by the
606 Interdisciplinary Centre for Clinical Research, University Münster (to C.M.L., Lil3/001/25).
607 C.M. Lill was supported by the Heisenberg program of the DFG (DFG; LI 2654/4-1).

608 The coordination of EPIC-Europe is financially supported by International Agency for
609 Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics,
610 School of Public Health, Imperial College London which has additional infrastructure support
611 provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are
612 supported by Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Italian Ministry of
613 Health, Italian Ministry of University and Research (MUR), Compagnia di San Paolo (Italy);

614 Dutch Ministry of Public Health, Welfare and Sports (VWS), the Netherlands Organisation for
615 Health Research and Development (ZonMW), World Cancer Research Fund (WCRF), (The
616 Netherlands); Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía,
617 Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO
618 (Spain); Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford),
619 Medical Research Council (MR/N003284/1, MC-UU_12015/1 and MC_UU_00006/1 to EPIC-
620 Norfolk; MR/Y013662/1 to EPIC-Oxford) (United Kingdom). Previous support has come from
621 "Europe against Cancer" Programme of the European Commission (DG SANCO). SomaScan®
622 data were generated under Master Research Agreement, 14th December 2021, between
623 Imperial College London and SomaLogic Inc. SomaLogic were not involved in analyzing or
624 interpreting the data; or in writing or submitting the manuscript for publication.

625 National Institute on Aging (NIA) contracts N01-AG-12100 and HHSN271201200022C (for Vi.
626 G.) and Althingi (the Icelandic Parliament) financed the AGES-Reykjavik study. IHA and
627 Novartis have collaborated on proteomics research since 2012. This study was also funded
628 by the NIA (1R01AG065596-01A1 to Vi. G., 2R01AG065596-03 to Va.G.), the University of
629 Iceland Postdoctoral Fund (to E.A.F.) and the Icelandic Research Fund (2511008-051 to
630 Va.G.). This research was supported, in part, by the Intramural Research Program (IRP) of the
631 NIA. K.A.W. and F.L. are supported by the NIA IRP. ARIC is carried out as a collaborative study
632 supported by National Heart, Lung, and Blood Institute (NHLBI) contracts (75N92022D00001,
633 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). The ARIC
634 Neurocognitive Study was additionally supported by U01HL096812, U01HL096814,
635 U01HL096899, U01HL096902, and U01HL096917 from the NIH (NHLBI, NIA, National
636 Institute of Neurological Disorders and Stroke [NINDS], and National Institute on Deafness
637 and Other Communication Disorders [NIDCD]).

638 Disclaimer: Where authors are identified as personnel of the International Agency for
639 Research on Cancer / World Health Organization, the authors alone are responsible for the
640 views expressed in this article and they do not necessarily represent the decisions, policy or
641 views of the International Agency for Research on Cancer / World Health Organization.

642 **Authors' contributions**

643 Study concept and supervision: LB, PF, MJG, ER, CML; acquisition/contribution of data: SM,
644 JMH, MG, AJ-Z, MJS, CT-S, NC-C, AV, DP, SS, TK, NW, RK, RT, RV, SP, GM, CS, NF, GNPC, VK,

645 ViG, HG, LM, LMW, IT, VaG, KAW, PF, ER, MJG, CML, data analysis and interpretation: JH, AS,
646 SM, EAF, FL, VV, RM, RK-L, VD, OO, LD, FH, YZ, FA, KSB, PMK, JMH, NW,TH, SB, AE, MH, VaG,
647 LB, OR, LMW, IT, ViG, KAW, PF, ER, MJG, CML; drafting the manuscript: JH, SM, CML. Critical
648 revision of the manuscript for content: all co-authors

649 **Competing interests**

650 N.F. is an employee and stockholder of Novartis. K.A.W. is an Associate Editor at Alzheimer's
651 & Dementia, a member of the Editorial Board of Annals of Clinical and Translational
652 Neurology, and on the Board of Directors of the National Academy of Neuropsychology.
653 K.A.W. and I.W. have given unpaid seminars on behalf of SomaLogic. None of the other
654 authors reports any competing interest.

655 **Legends to figures**

656 **Figure 1. Visual summary of study design.**

657 Legend. This study leveraged plasma proteomics from 4,538 individuals in the 'European Prospective Investigation into
658 Cancer and Nutrition for Parkinson's disease' case-cohort (EPIC4PD) to discover prediagnostic protein biomarkers of
659 Parkinson's disease (PD). Using the SomaScan 7K platform, 7,285 aptamers were analyzed across multiple time windows.
660 Identified biomarkers were externally validated in three large cohorts ('Age, Gene/Environment Susceptibility–Reykjavik
661 Study' [AGES], 'Atherosclerosis Risk in Communities' study [ARIC], and 'United Kingdom Biobank' [UKB]) and assessed in
662 clinical PD cases from the 'Global Neurodegenerative Proteomics Cohort' [GNPC] and the longitudinal 'Tracking Parkinson's
663 disease' cohort. Functional characterization included pathway and cell-type enrichment, protein-protein networks, cross-
664 trait analyses with Alzheimer's disease (AD) and amyotrophic lateral sclerosis (ALS). Predictive protein risk scores were
665 calculated in EPIC4PD and externally validated in AGES.

666 **Figure 2. Cox proportional hazard regression analyses across different time-windows in**
667 **EPIC4PD.**

668 Legend. Association results of prediagnostic protein biomarkers with incident Parkinson's disease across different follow-up
669 windows in EPIC4PD. Panels **A–C**: effect estimates for three time windows: full follow-up (2–28 years), proximal (2–10
670 years), and distal (10–20 years). **(D)** summary of associations of significant biomarkers (false-discovery rate of 0.05 in at
671 least one time-window) across all follow-up periods.

672 **Figure 3. Validation and characterization of top prediagnostic biomarkers**

673 Legend. (**A**) Forest plot of results of prediagnostic biomarkers for incident PD in the EPIC4PD case-cohort, the 'Age,
674 Gene/Environment Susceptibility–Reykjavik Study' (AGES), the 'Atherosclerosis Risk in Communities' study (ARIC), and
675 'United Kingdom Biobank' [UKB] cohorts. **(B)** Comparison of top prediagnostic PD biomarkers with their associations in
676 prevalent PD within the 'Global Neurodegenerative Proteomics Cohort' (GNPC); left: false-discovery rate (FDR)-significant
677 proteins from EPIC4PD; right: nominally significant proteins. Color-coding refers to risk (red) or protective (blue) effects in
678 GNPC. **(C)** Cross-disease analysis of prediagnostic PD-associated proteins compared to incident Alzheimer's disease (AD) and
679 amyotrophic lateral sclerosis (ALS) in EPIC4AD and EPIC4ALS. The top panels display all FDR-significant prediagnostic PD
680 proteins from the full follow-up analysis, while the bottom panels show proteins with nominal significance ($p<0.05$). Protein
681 names in orange indicate nominal significance in AD (top left) or ALS (top right). Orange bolded names indicate FDR-
682 significant proteins. For each disease, the bottom right plot illustrates the correlation of effect estimates using 'mirrored'
683 beta values (see Methods). **(D)** Protein-protein interaction network of FDR-significant predictive proteins (black border) and
684 GWAS-identified hits (no border). Eight clusters were identified using igraph's greedy optimisation of modularity. Node size
685 increases for nodes with 10 or more edges, proportional to the number of edges. **(E)** Integrated overview of datasets for
686 validation and characterization of FDR-significant biomarkers across different proteomic and transcriptomic layers.

687 **Table 1. Overview of cohort datasets included in this study**

Cohort	Population	Phenotype	Platform	Cases			Non-cases		N total
				Number (% women)	AAE (range)	AAD (range)	Number (% women)	AAE (range)	
EPIC4PD	Europe	incident PD	SomaScan 7K	574 (45%)	59 (38-77)	71 (45-89)	3,964 (66%)	52 (35-79)	**4,538**
ARIC	USA	incident PD	SomaScan 5K	164 (41%)	59 (48-68)	75 (56-87)	11,041 (56%)	57 (46-70)	**11,205**
AGES	Iceland	incident PD	SomaScan 7K	121 (38%)	78 (69-88)	84 (72-93)	4,969 (58%)	78 (68-98)	**5,090**
GNPC	Europe, USA	prevalent PD	SomaScan 7K	335 (37%)	75 (49-90)	-	2,257 (60%)	71.5 (27-90)	**2,592**
TRACKING PD	UK	PD progression	SomaScan 7K	794 (34%)	66 (38-85)*	66 (38-85)	-	-	**794**
EPIC4AD	Europe	incident AD	SomaScan 7K	652 (71%)	59 (36-69)	78 (53-93)	1,274 (63%)	49 (35-69)	**1,926**
EPIC4ALS	Europe	incident ALS	SomaScan 7K	187 (57%)	58 (38-76)	73 (48-95)	4,528 (66%)	52 (35-79)	**4,715**
UK Biobank	UK	incident PD	Olink Explore 3072	749 (37%)	64 (41-70)	71 (44-83)	47,812 (54%)	58 (40-70)	**48,561**
TOTAL				**3,572**			**70,523**		**74,095**

688
689 Legend. This table presents the cohort datasets included in the study. AAE = median age at examination (baseline), AAD = median age at diagnosis. Note that the individual EPIC-derived case-
690 cohorts (EPIC4PD, EPIC4AD, EPIC4ALS) the vast majority of non-cases, thus the total number of non-cases is not the sum of non-cases of the individual datasets. See Methods for full cohort name
691 abbreviations. *Mean age at visit 1.

692 **Table 2. Plasma protein biomarkers significantly associated with future onset of Parkinson's disease**

Protein	HR (95% CI)	p	q	Replication	Cross-disease	Prior evidence
2-28 years of follow-up						
VAMP3	1.26 (1.16-1.36)	1.46E-08	1.07E-04	moderate		
TPPP2	0.70 (0.61-0.80)	7.06E-08	3.15E-04	**strong**		
HPGDS	0.73 (0.65-0.82)	8.59E-08	2.09E-04	**strong**	AD	
$ALPL_{cap}$	1.25 (1.15-1.36)	1.30E-07	3.16E-04	**strong**		
FCAR	1.34 (1.20-1.50)	3.48E-07	6.34E-04	suggestive		Irmady et al., 2023[45]
CERS5	1.23 (1.12-1.34)	5.64E-06	6.85E-03	suggestive	ALS	
$MFAP5_{cap}$	0.73 (0.64-0.84)	1.14E-05	1.38E-02	**strong**		
RSPO2	1.19 (1.10-1.29)	2.09E-05	2.10E-02	suggestive	**AD**, ALS	
CD3G	1.19 (1.10-1.29)	2.30E-05	2.10E-02	suggestive		
$GTF2A2_{cap}$	1.22 (1.11-1.34)	2.30E-05	2.40E-02	moderate		
LTF	1.28 (1.14-1.43)	3.21E-05	2.34E-02		AD	
$NAMPT_{cap}$	1.28 (1.14-1.44)	4.51E-05	3.65E-02		ALS	Santiago et al., 2016[46]
OGFR	0.77 (0.68-0.87)	5.04E-05	3.67E-02	**strong**	AD	
FGF8	1.21 (1.10-1.33)	5.75E-05	3.46E-02	moderate		
PSG6	1.25 (1.12-1.39)	6.17E-05	3.46E-02			
ACAD8	1.18 (1.09-1.28)	7.05E-05	3.65E-02	**strong**	**AD**	
CRYGD	1.20 (1.10-1.32)	7.52E-05	3.65E-02			
2-10 years of follow-up						
$TIMP1_{cap}$	0.72 (0.62-0.83)	4.52E-06	1.65E-02	moderate		
IL11	1.30 (1.15-1.47)	2.04E-05	4.23E-02			
TCL1A	0.71 (0.61-0.83)	2.23E-05	4.23E-02	**strong**		Kedmi et al., 2011[47]
10-20 years of follow-up						
$GUCA1B_{cap}$	1.31 (1.18-1.45)	3.95E-07	1.44E-03			
$UBE2L3\|UBB_{cap}$	1.29 (1.16-1.42)	6.08E-07	1.48E-03			

DCAF12	1.24 (1.12-1.37)	2.63E-05	2.47E-02			
$GPC4_{cap}$	1.26 (1.13-1.40)	2.80E-05	3.33E-02	**strong**		Tatenhorst et al., 2024[48]
BECN1	1.26 (1.13-1.40)	2.92E-05	2.47E-02			Miki et al., 2018[49]
$HEPACAM2_{cap}$	1.25 (1.12-1.39)	3.05E-05	3.33E-02	suggestive		
$F13A1\|F13B_{cap}$	1.34 (1.17-1.54)	3.20E-05	3.33E-02			
ABHD12	1.25 (1.12-1.38)	3.37E-05	2.47E-02	moderate		
IRAG2	1.24 (1.12-1.37)	3.38E-05	2.47E-02		AD	
$TAFA4_{cap}$	1.25 (1.12-1.39)	5.93E-05	3.93E-02			
GSTA3	1.23 (1.11-1.37)	6.08E-05	3.69E-02	**strong**	AD	
$SPAST_{cap}$	1.24 (1.11-1.38)	7.94E-05	4.45E-02		**AD**, ALS	
CUL4B	1.32 (1.15-1.51)	9.85E-05	4.22E-02			Yao et al., 2022[34]
OSTF1	1.30 (1.14-1.49)	9.92E-05	4.22E-02			
PYDC1	1.24 (1.11-1.38)	1.07E-04	4.22E-02		**AD**, ALS	D. Torshizi et al., 2024[18]
LCN2	1.30 (1.14-1.49)	1.10E-04	4.22E-02	**strong**		Fan et al., 2024[31]
OTX2	1.27 (1.13-1.44)	1.10E-04	4.22E-02		AD	
MAGEA8	0.66 (0.54-0.82)	1.16E-04	4.22E-02			
ZNF41	1.22 (1.10-1.35)	1.32E-04	4.57E-02			
Men						
RS1	1.42 (1.22-1.65)	4.34E-06	9.10E-03		**AD**	
$KRAS_{cap}$	1.34 (1.17-1.53)	1.30E-05	1.89E-02	**strong**		
LYZ_{cap}	1.37 (1.19-1.59)	2.06E-05	2.50E-02			
$GBA3_{cap}$	1.33 (1.16-1.52)	3.91E-05	3.17E-02		**AD**	
GJA1	1.32 (1.16-1.50)	3.70E-05	3.37E-02	**strong**	**AD**	
AMN	1.40 (1.19-1.64)	5.07E-05	3.69E-02	moderate	ALS	Elango et al., 2023[32]
Women						
ITGA6	1.27 (1.14-1.42)	1.87E-05	4.29E-02	suggestive		
CXCL11	0.69 (0.58-0.82)	2.34E-05	5.67E-02	moderate		Hepp et al., 2023[33]

Legend. This table presents all plasma proteins showing significant (false discovery rate [FDR]=0.05) association with incident Parkinson's disease (PD) in Cox proportional hazard regressions in the EPIC4PD case-cohort stratified for center, sex (where non-sex-stratified) and 5-year age categories. Analyses were conducted for the entire follow-up period (2–28 years), as well as for proximal (2–10 years) and distal (10–20 years) time windows before disease onset or censoring, and additionally analyzed per sex stratum over the full follow-up. Please note that for the distal time window, the amount of available replication data was limited due to limited follow-up in some cohorts; in these cases, replication evidence was scored down. Strong evidence for replication: At least one nominally significant replication (for SomaScan in the same effect direction; for Olink, direction not required), and all additional SomaScan-based datasets showing effect estimates in the same direction. Moderate evidence for replication: Effects observed in the same direction in all (at least two) SomaScan datasets datasets, but non-significant results, or alternatively, one nominally significant SomaScan-based replication result (same direction), another SomaScan dataset showing the same and one showing an opposite direction. Suggestive evidence for replication: One nominally significant replication result (same direction for SomaScan, for Olink, direction not required) but heterogeneous effect directions among remaining datasets, or only two SomaScan datasets with consistent (but non-significant) effects. Cross-disease=at least nominally significant (uncorrected $p<0.05$) associations with instantaneous risk of Alzheimer's disease (AD) and amyotrophic lateral sclerosis (ALS) based on EPIC4AD and EPIC4ALS, respectively. Prior evidence: publications describing differential protein and/or mRNA levels in blood in (always prevalent) PD patients compared to controls. Note that these were always in line with effect directions in EPIC4PD. HR=hazard ratio; CI=95% confidence interval; q=FDR-adjusted p-value (FDR=0.05). Bolded words for 'cross-disease' indicate significance at FDR=0.05. Note that from both outlier handling approaches, i.e., excluding outliers beyond 5 standard deviations from the log-mean or capping these outliers at the value of 5 standard deviations (indicated as "cap"), only the more significant result (smaller q) is shown. However, in all instances, at least nominally significant evidence was also observed for the alternative outlier removal method.

707 **METHODS**

708 **Discovery cohort and study design**

709 The EPIC cohort is a large multi-center prospective study that recruited 519,978 participants
710 (70.5% women, mostly aged 35-70 years at baseline) from 23 centers across ten European
711 countries. With approximately 30 years of follow-up, EPIC was originally established to
712 explore the relationships between diet and cancer, but it has since expanded to study other
713 chronic conditions, including neurodegenerative diseases[12,13]. Briefly, baseline data including
714 information on diet, physical activity, alcohol and tobacco use, past and current illnesses,
715 reproductive history, and anthropometric measurements were collected between 1991 and
716 2000. Citrate plasma samples from ~75% of all participants were obtained using
717 standardized protocols, then aliquoted and stored long-term in liquid nitrogen (–196 °C) at a
718 central biobank housed at the International Agency for Research on Cancer (IARC)/World
719 Health Organization in Lyon, France. Since baseline recruitment, participants have been
720 followed up regularly by data linkage and/or questionnaires at each center. The EPIC study
721 was approved by the ethical committee of IARC and by the ethical review boards of each
722 participating study center. All participants signed a written informed consent.

723 The 'EPIC for Parkinson's disease' case-cohort (EPIC4PD) analyzed in this study is embedded
724 within EPIC and aims to prospectively identify molecular biomarkers in PD[11]. The case
725 ascertainment protocol has been described previously[11,50]. It was based on a source
726 population of 157,288 subjects from ten EPIC centers across five European countries
727 including Spain (Murcia, Navarra, and San Sebastian), Italy (Florence, Turin, and Varese),
728 Netherlands (Utrecht and Bilthoven), UK (Norfolk/Cambridge), and Germany (Heidelberg).
729 Potential cases were ascertained based on data linkage through a combination of center-
730 specific methods using hospital records, general practitioner records, mortality records, drug
731 repositories, and questionnaire data followed by re-review of medical records by expert
732 neurologists for PD. Since the initial ascertainment, 163 incident PD cases were added
733 following an update of the case ascertainment in selected centers (Heidelberg, Florence,
734 Varese, Turin, Utrecht, Bilthoven)[11]. The subcohort for EPIC4PD was randomly drawn from
735 these centers from EPIC-interact, which is a random subcohort within EPIC[51]. The data freeze
736 for the current project was set to the time point of diagnosis of the last patient per center.
737 For the purpose of identifying predictive PD biomarkers, we applied a washout phase after

738 blood draw of two years prior to diagnosis to avoid reverse causation resulting in an
739 exclusion of 23 PD patients.

740 **Generation and quality control of proteomic data in EPIC4PD**

741 Protein abundances were measured in citrate plasma samples using the SomaScan 7K (v4.1)
742 platform (SomaLogic, Inc.). SomaScan is a high-throughput proteomic platform that uses
743 modified nucleotides (Slow Off-rate Modified Aptamers - SOMAmers) which specifically bind
744 to protein targets and quantifies them in relative fluorescence units (RFUs) using DNA
745 microarray[52]. The 7K version quantifies 7,596 aptamers targeting 6,432 unique proteins[53].
746 Measurements of overall 17,841 EPIC participants were performed at SomaLogic (Boulder,
747 CO, USA) as previously described. Participants included also individuals who were selected
748 for other endpoints (such as cancer). All samples were processed in a single procedure.
749 SomaScan assays were performed on 96-well plates, with 11 control wells per plate used to
750 monitor batch effects, accuracy, precision, and background. Each plate included five pooled
751 calibrator replicates, three quality control (QC) replicates, and three buffer replicates. In
752 addition, 233 replicates of one EPIC QC sample were included. Readout was conducted using
753 Agilent hybridization and scanning technology. To control readout variability, twelve
754 hybridization control aptamers were added during elution. Initial QC was conducted by
755 SomaLogic and included normalization of protein RFUs based on the following steps:
756 hybridization normalization, intraplate median normalization, plate scaling and calibration,
757 and adaptive normalization to a population reference.

758 We computed intraclass correlation coefficients (ICCs) for all 7,596 aptamers based on the
759 233 EPIC replicate measurements. Aptamers were removed if they were not a protein (n=46)
760 [Hybridization Control Elution (n=12), non-biotin (n=10), non-cleavable (n=4), spuriomer
761 (n=20)] or non-human (n=261). Aptamers were further excluded if both (i) the ICC was below
762 0.5 and (ii) if the aptamer levels were below the limit of detection (LOD) in more than 90% of
763 the samples (n=4). If the measured value of an aptamer was below the LOD, it was set to
764 LOD/2. The final number of quality-controlled aptamers was 7,285 (targeting 6,381 unique
765 proteins). Samples for which any SomaLogic normalization scale factor was outside [0.4-2.5]
766 (n=247) and those detected as 0s using an approach based on PCA and the local outlier
767 factor statistic (https://privefl.github.io/blog/detecting-outlier-samples-in-pca/) were
768 excluded (n=0 for the method excluding outliers and n=5 for capping outliers). Plate

769 correction was further applied through a residual approach: for each SOMAmer, its
770 measurements were corrected for plate effect estimated in linear mixed effect models
771 adjusted for center, age, sex, BMI, smoking status, and incidence of multiple cancer types,
772 CVD, T2D, neurodegenerative diseases and death, to preserve possible biological variation
773 due to these factors. Finally, only the samples belonging to the EPIC4ND case-cohort were
774 retained (n=6,543 for both outlier handling methods).
775 Measurements of each SOMAmer were log_{10}-transformed and further centered and scaled
776 so that their mean was 0 and their standard deviation was 1 in the subcohort.
777 Measurements deviating more than 5 SD from the mean of the normalized log_{10}-
778 transformed data were either excluded or 'capped' i.e., set to the value of 5 SD from the
779 normalized log_{10}-transformed mean. The first approach is based on the assumption that
780 outliers represent technical or biological variation unrelated to PD (e.g., outliers due to
781 blood type or sex or genetic variation not associated with PD) obscuring any association
782 signal nearer to the mean. However, as we also acknowledge that outliers may represent
783 strong PD biomarker signals, in the discovery phase, we equally ran all statistical analyses by
784 capping these outliers. For all statistically significant findings, log_{10}-transformed values were
785 visually inspected (**Supplementary Figure 11**) to verify the appropriateness of the data
786 transformation and the cutoff for the outlier removal. In the subsequent text and main
787 tables, protein names based on results using capped data are labeled with an 'cap'.

788 **Statistical analysis in EPIC4PD**

789 The final effective EPIC4PD case-cohort included 4,538 EPIC4PD participants including 574
790 incident PD cases with 7,285 aptamers (corresponding to 6,381 proteins) available for
791 statistical analysis. We used Cox proportional hazard regression analyses to examine the
792 association of 7,285 aptamers with instantaneous risk of PD using age as time scale and
793 Prentice weights to take into account the case-cohort setting [54]. Our basic model was
794 stratified (i.e., separate baseline hazard functions estimated) for 5-year age intervals, sex,
795 and center. Cox proportional hazards regression was performed using three distinct time
796 windows relative to disease onset: the full follow-up time period (2-28 years from baseline
797 to diagnosis/censoring), as well as two shorter periods therein: 2-10 and 10-20 years.
798 Furthermore, we performed sex-specific analyses after stratification for men and women

799 using the full follow-up time period. Differences between sex-stratified effect estimates
800 were tested by interaction analyses as previously described[55,56].
801 Proteomic results were defined as significant at a FDR of 0.05. Furthermore, we defined
802 "suggestively significant biomarker signals" as those findings that did not withstand FDR
803 control in the proteomic analyses but had a p-value $<10^{-3}$ and demonstrated consistency
804 across all EPIC4PD analyses - specifically, at least two nominally significant ($p<0.05$) results
805 and all effect estimates pointing in the same direction of effect.
806 In sensitivity analyses for all FDR-significant results, we tested a 'full' adjustment model
807 including the baseline covariates school education (no education/primary school vs higher
808 education), smoking status (dummy-coded as 'never', 'former', 'current'), body mass index
809 (BMI; weight/height^2 [kg/m^2]), and physical activity (dummy-coded as 'inactive', 'moderately
810 inactive', 'moderately active/active'). Furthermore, for FDR-significant results derived from
811 analyses stratified for women, we adjusted for postmenopausal status (dummy-coded as
812 premenopausal, postmenopausal, perimenopausal, and surgical postmenopausal status) and
813 use of hormones for menopause (yes/no).
814 Time-varying effects of PD-associated proteins were explored using different approaches
815 across the EPIC4PD dataset with 2-30 year follow-up. To visualize **time-varying effects** of top
816 results in different time windows, HR pseudotrajectories were computed using 5-year sliding
817 windows in one-year steps. For the analyses of the sliding windows close to age at diagnosis,
818 the 23 participants diagnosed with PD within two years of baseline and previously excluded
819 from the main analyses were re-included. Due to more limited numbers of cases toward the
820 end of follow-up, the final window covered diagnoses occurring 16 to 25 years after
821 baseline. Curves were fitted using the geom_smooth() function from the ggplot2 package in
822 R.
823 Furthermore, exploratory analyses of time-varying effects of ALS-associated proteins
824 showing FDR-significant association in either the proximal or distal time windows (see
825 above) were performed by analyzing Cox models (Prentice-weighted, stratified for age, sex,
826 and center) with a 'protein level × log time-to-event' interaction term, and and by fitting Cox
827 models incorporating natural splines to assess non-linear time-dependent effects. For all
828 these analyses, we used time since blood collection instead of age as the time scale. For the
829 splines, we split the follow-up time into fine intervals (using the survSplit() function) and
830 applied a time-varying coefficient (TVC) approach for each protein using spline interactions

831 (protein × spline(time-to-event)). The Prentice-weighted Cox models were used adjusting for
832 age (as a time-varying variable) and stratifying by sex and center. The nsk() function from the
833 R package splines was used with 2 degrees of freedom. We evaluated the statistical evidence
834 for time-varying effects using Wald tests.

835 **Independent replication of predictive biomarkers in prospective cohorts**

836 Replication analyses of FDR-significant predictive biomarkers were performed on 64,856
837 individuals (1,034 incident PD cases) across three different prospective cohorts (AGES[57],
838 ARIC[58], UKB[59]) based on Olink (UKB) and SomaScan (AGES, ARIC) data, respectively.
839 The first replication cohort, the Age Gene/Environment Susceptibility-Reykjavik (AGES-
840 Reykjavik) Study[60], is a population-based study of adults aged 65 years and older in Iceland,
841 designed to investigate genetic susceptibility and gene-environment interactions related to
842 age-associated diseases. Between 2002 and 2006, 5,764 participants, which were survivors
843 of the original Reykjavik Study, underwent comprehensive interviews, examinations, and
844 measurements. The study was approved by the Icelandic National Bioethics Committee
845 (VSN: 00–063), the Icelandic Data Protection Authority, and the NIA IRB; all participants
846 provided written informed consent. Proteomic profiling of baseline serum samples from the
847 AGES cohort has been described previously[61]. After quality control and exclusion of
848 prevalent PD cases, 5,308 participants had available proteomic data, further reduced to
849 5,090 when applying a 2-year washout period. Over the follow-up, 121 participants
850 developed incident PD. In AGES, aptamer levels were log10-transformed. Further details on
851 recruitment can be found in ref.[62], and prevalent PD case ascertainment in ref.[63]. Medical
852 and mortality records (ICD10: G20) up to 2019-03-19 were reviewed to ascertain incident PD
853 cases.
854 The 2nd replication cohort, ARIC[58], is an ongoing community-based cohort study, for which
855 participants were enrolled from four communities across the United States: Washington
856 County, MD; Forsyth County, NC; northwestern suburbs of Minneapolis, MN; and Jackson,
857 MS. The ARIC participants who had SomaScan 5K data available (4,953 aptamers passing
858 quality control) comprised 11,205 middle-aged participants at baseline of whom 164
859 developed incident PD applying a washout period of two years and had SomaScan data in
860 EDTA plasma samples available. For details on the cohort recruitment and ascertainment of

861 incident PD cases therein please see ref.[64] and the **Supplementary Material**. All participants
862 provided appropriate written consent.
863 The 3rd replication cohort, UKB[59], is a general population cohort study, which performed
864 proteomic profiling of blood plasma samples using the Olink Explore 3072 platform (n=2,923
865 proteins, cohort size analyzed in this study: n=48,561 UKB participants including 749 incident
866 PD cases). The plasma samples were collected in volunteers aged 40 to 69 years at their first
867 attendance to a UKB assessment center in 22 centres across the United Kingdom between
868 2006 to 2010. Follow-up began from the date of attendance to an assessment center (field
869 53) to the first recorded date of diagnosis (ICD-10 code G20) in the patient-linked hospital
870 episode statistics (fields 41270 and 41280) or death registration (fields 40001 and 40002),
871 whichever occurred first, with the last recorded date in January 2023. Due to small case
872 numbers in other strata, the analysis was restricted to white participants (field 21000).
873 Participants with missing proteomics data were removed prior to each model.
874 Due to smaller sample sizes in the restricted time windows, we did not analyze top EPIC4PD
875 biomarkers in AGES for the time-window of 2-10 years of follow-up (n cases=26) and for
876 ARIC for the time window of 10-20 years of follow-up (n cases=32).

877 Consistent with the EPIC4PD analyses, outliers in AGES and UKB were either capped at or
878 excluded beyond 5 SD, depending on the approach used for the specific biomarker to be
879 replicated. In AGES and UKB, outliers were either capped at or excluded beyond 5 SD, in line
880 with EPIC4PD analyses. Outliers in ARIC were capped at 5 SD. Cox proportional hazards
881 regression models were harmonized for AGES, ARIC and UKB with EPIC4PD: The Cox model
882 was based on age as time scale and stratified (separate baseline hazard functions estimated)
883 for age in 5-year categories, race (in ARIC), and sex. Significance for the replication arm was
884 defined at α=0.05. As sensitivity analysis, we also performed the same analyses only on
885 white participants in ARIC (excluding 2,625 black participants). However, as results did not
886 change substantially, we only report results on all participants in this study.
887 We explicitely note that the smaller amount of replication data of distal biomarkers (i.e.
888 falling into the 10-20-year pre-diagnosis window) is a a result of shorter follow-up periods in
889 some replication cohorts, reducing the number of available replication datasets (excluding
890 ARIC, and, in many instances, UKB), as well as limited case numbers (AGES) after time-
891 window stratification. Specifically, while all time-window independent biomarkers were

892 assessed in 3-4 independent datasets, for 14 out of 19 biomarkers from the 10-20 year
893 window, only 1-2 replication datasets were available. Consequently, we have intentionally
894 exercised caution and avoided highlighting confirmatory (but limited) external replication
895 evidence to prevent overinterpretation.

896 **Indirect replication and characterization of prediagnostic EPIC4PD biomarkers for clinical**
897 **phenotypes in GNPC**

898 The FDR-significant prediagnostic biomarkers identified in EPIC4PD across the three follow-
899 up periods and after stratification for sex were assessed for an association with prevalent PD
900 leveraging a large case-control dataset of the GNPC (335 prevalent cases and 2,257 controls
901 from 3 contributors) by re-analyzing the data described in the companion manuscript of
902 Imam et al.[14] All individuals included had undergone SomaScan 7K proteomic profiling as
903 part of the GNPC initiative. For this analysis, we restricted inclusion to datasets that
904 contained both PD cases and controls. Control participants were those without a
905 documented clinical diagnosis of mild cognitive impairment, dementia, AD, frontotemporal
906 dementia, PD, or ALS. To minimize bias from population stratification and ensure
907 comparability with EPIC4PD, we only included white participants in datasets where the
908 variable "race" was available. If no race information was provided, all participants were
909 included. The protein data was log10-transformed and standardized and values above or
910 below five standard deviations from the mean were excluded. Then, the case-control
911 datasets were residualized using the removeBatchEffect function from the R package limma
912 to regress out the effects of age, sex, and contributor, thereby controlling for confounding
913 sources of variation in the high-dimensional data. A principal component (PC) analysis was
914 then performed on the residuals (i.e., after having removed the effects of age, sex, dummy-
915 coded contributor); based on the scree plot, the respective first two PCs were retained as
916 additional covariates. Finally, logistic regression analyses were conducted on the original
917 data, adjusting for age, sex, dummy-coded contributor, and the first two PCs (all results from
918 the re-analyses with $p<0.05$ can be found in **Supplementary Table 24**). Analogous to the
919 EPIC4PD analyses, prediagnostic biomarkers that were FDR-significant in one sex stratum
920 were analyzed for the GNPC data within that specific sex group.

921 Concordance of effect estimates in EPIC4PD and GNPC (for all 47 significant prediagnostic
922 biomarkers as well as for nominally significant biomarkers shared by both studies) we
923 applied a one-sided binomial exact test. All analyses were performed in R.
924 We analyzed serum proteomic data from the Tracking Parkinson's disease cohort using the
925 SomaScan v4.1 (7K) assay (SomaLogic). Our cohort included 794 PD patients with serum
926 samples and clinical progression data across the three study visits (at 0 months, 18 months,
927 36 months). To correct for batch effects introduced by collection site, we used the ComBat
928 function from the sva R package (v3.50.0), with site as the batch variable, followed by log_{10}-
929 transformation. For association analyses with PD progression, we investigated associations
930 between protein levels and two clinical outcomes: the Unified Parkinson's Disease Rating
931 Scale Part III (UPDRS-III) as a measure of motor function, and the Montreal Cognitive
932 Assessment (MoCA) as a measure of cognitive function.
933 Cross-sectional associations were assessed using generalized linear models, adjusting for
934 age, sex, and Levodopa Equivalent Daily Dose (LEDD) as a measure of treatment exposure.
935 For longitudinal analyses across up to to three visits per participant, we applied linear mixed-
936 effects models using the lmer() function from the lme4 R package. These models included a
937 random intercept for participant ID to account for repeated measures, and fixed effects for
938 protein level (log_{10} RFU), visit, age, sex, and LEDD. In additional sensitivity analyses, we
939 included BMI and disease duration (time since diagnosis at baseline) as covariates. Protein
940 associations were evaluated separately for each outcome, and p-values were adjusted for
941 multiple testing using an FDR=0.05.

942 **Comparison to previous studies**

943 We compiled a list of 291 unique candidate biomarkers previously linked to PD based on
944 findings from studies that employed a range of different methodologies (several proteins
945 were nominated more than once): These included i) 38 proteins from the only previous
946 proteomic study on future PD performed on 2,920 proteins based on Olink measurements in
947 the UKB[10], ii) 10 of ~2,700 analyzed blood plasma proteins (7 derived from Olink and 3 from
948 SomaScan studies) recently proposed as potential predictive biomarkers for PD by a
949 comprehensive MR study[16], iii) 18 proteins nominated by a recent MR and colocalization
950 study[65], iv) 109 proteins corresponding to nominated functional candidate and/or the
951 nearest genes located in 78 genetic risk loci described in the most recent Caucasian genome-

952 wide association study for PD [17], and v) 127 preselected neurodegenerative disease
953 candidate biomarkers included in the 'CNS NULISA panel' (Alamar, Inc.). Proteins available in
954 our SomaScan 7K data from this list (n=180) were tested for association in EPIC4PD. Results
955 were controlled applying an FDR=0.05 across all 180 candidate proteins available for testing
956 in our SomaScan 7K data.

957 **Cross-disease comparisons**

958 For cross-disease comparisons between PD and AD as well as between PD and ALS, we
959 performed Cox proportional hazard regression analyses for all FDR-significant PD proteins in
960 EPIC4AD (1,926 EPIC participants including 652 incident AD cases during a follow-up of 2-29
961 years) and EPIC4ALS (4,715 EPIC participants including 187 incident ALS cases during a
962 follow-up of 2-30 years) using Prentice-weighted Cox proportional hazard regression analysis
963 stratified for 5-year age groups, sex, and center, as described above. Results were controlled
964 at an FDR=0.05 in the context of all tested top biomarkers per disease. Furthermore, we
965 assessed whether significant associations in incident AD and ALS were enriched among all
966 proteins that were at least nominally significant in PD (based on the full follow-up period,
967 outliers excluded), using a one-sided Fisher's exact test. To assess whether overlapping
968 nominally significant proteins exhibited concordant effect directions more often than
969 expected by chance, we performed a one-sided binomial test. Furthermore, we assessed the
970 correlation of effect sizes between traits by calculating Pearson's correlation coefficients
971 based on the beta estimates. To prevent bias in the correlation due to the discontinuity
972 introduced by limiting the analysis to nominally significant results, we aligned effect
973 directions by mirroring the beta coefficients: for proteins with negative betas in PD, we
974 multiplied all corresponding betas in EPIC4PD and GNPC by –1 (generating 'mirrowed beta
975 values').

976 **Brain expression data and protein abundance of top biomarkers**

977 Gene expression and protein abundance of top PD biomarkers across tissues and in the brain
978 were assessed based on RNA-Seq and immunohistochemisty (IHC) data obtained from the
979 Human Protein Atlas (HPA v24.0, based on Ensembl version 109, accessed May 6th, 2025 at
980 https://www.proteinatlas.org/about/download) [66]. FDR-significant proteins were probed for
981 tissue expression using bulk RNA sequencing (RNA-Seq) data across 50 human tissues
982 available as gene-level 'Consensus RNA data' (rna_tissue_consensus.tsv.zip). Furthermore,

983 brain expression was assessed based on gene-level bulk RNA-Seq data across 13 brain
984 regions based on samples from the Human Brain Tissue Bank (Semmelweis University,
985 Budapest; rna_brain_region_hpa.tsv.zip) and based on single nucleus RNA-Seq (snRNA-Seq)
986 data across 11 brain regions (rna_single_nuclei_cluster_type.tsv.zip). Protein measurements
987 for the FDR-significant biomarkers across 45 human tissues were obtained from HPA based
988 on IHC using tissue microarrays (normal_ihc_data.tsv.zip). To visualize expression patterns of
989 the top biomarkers, heatmaps were generated using hierarchical clustering, with expression
990 values scaled per gene highlighting tissue-specific variation.

991 **Gene set, cell type enrichment and network analyses**

992 Gene set enrichment analyses were performed based on the GO classification system using
993 the PANTHER resource (Protein ANalysis THrough Evolutionary Relationships; version 19.0;
994 released 2024-06-20; https://geneontology.org/). A list of all unique proteins that showed a
995 suggestive signal in EPIC4PD and was uniquely mapped to a gene by PANTHER (n=154) was
996 tested for enrichment for 'biological process', 'molecular function', and 'cellular component'
997 GO categories as well as 'Reactome pathways'. The statistical overrepresentation test was
998 conducted using Fisher's exact test with the reference list set to all analyzed SomaScan 7K
999 proteins uniquely mapped by PANTHER (n=6,332). To avoid spurious associations, only GO
1000 categories with at least ten proteins present in the background list were considered.
1001 Cell enrichment analyses were performed using the the Expression Weighted Cell Type
1002 Enrichment (EWCE) method to predict the likely primary disease-associated cells based on
1003 our prioritised proteins[67]. This approach generates a probability distribution of gene
1004 expression for specific cell types by comparing a target gene list to a background gene set.
1005 We applied EWCE to our protein lists and to either single-cell mouse brain expression data[68]
1006 or RNA expression data from 23 selected cell types sourced from the Human Protein Atlas
1007 (dataset: rna_single_cell_type.tsv.zip)[66]. Each analysis was controlled at an FDR=0.05.
1008 CorrelationAnalyzeR was used to investigate the coexpression of all FDR-significant
1009 prediagnostic biomarkers. Coexpression with *SNCA* was specifically examined in 12,712
1010 normal GEO samples from immune cells and 3,198 normal GEO samples from brain.

1011 **Protein-protein interaction network analysis**

1012 Furthermore, protein-protein interaction (PPI) networks were built using the STRING
1013 database (v12.0, accessed May 6th, 2025 at https://string-db.org)[69] based on all FDR-

1014 significant EPIC4PD biomarkers and plot using igraph (v2.1.4). Interactions with a minimum
1015 interaction score of 0.4 were based on the following sources: 'textmining', 'experiments',
1016 'databases', 'co-expression", "neighborhood', 'gene fusion', and 'co-occurrence'.
1017 Furthermore, to examine the list of suggestive prediagnostic proteins in relation to known
1018 PD genes, we utilized 109 proteins described in the most recent genome-wide association
1019 study for PD (see above)[17]. P values were calculated based on the observed interactions
1020 compared to 10,000 random permutations. Nodes not connected to the largest component
1021 were removed from the PPI network Figure 3.

1022 **Drug Repurposing**

1023 We leveraged the Open Targets platform[19] to explore druggability and drug-repurposing
1024 potential for all FDR-significant, putative prediagnostic PD biomarkers. Open Targets
1025 integrates multiple data sources to systematically link genes and proteins with drug
1026 molecules and disease associations. We queried the database focusing on small-molecule
1027 evidence, excluding other therapeutic modalities such as biologics or gene therapies.
1028 Specifically, approximately 17,000 small-molecule drugs and 21,087 protein targets were
1029 screened for evidence of interaction with the 47 FDR-significant biomarkers. Evidence of
1030 druggability was defined by reported direct binding or predicted high-quality
1031 pharmacological modulation supported by curated experimental, 3D protein structure or
1032 clinical data within the Open Targets resource.

1033 **Prediction**

1034 To assess the proteins' performance in risk stratification for PD, we performed variable
1035 selection in a Cox proportional hazards model using the Least Absolute Shrinkage and
1036 Selection Operator (LASSO) for variable selection and regularization. Specifically, EPIC4PD
1037 (excluding the Heidelberg center) served as the training set. Incident PD cases were
1038 Prentice-weighted to account for the case-cohort sampling and class imbalance, and
1039 penalized Cox models were fitted using LASSO as implemented in the R package glmnet with
1040 cv.glmnet(family = "cox", type.measure = "C"). The amount of penalization (λ) was tuned via
1041 5-fold cross-validation using the C-index as the performance criterion. For model
1042 selection, each candidate λ from the cross-validation grid was evaluated in a Cox model
1043 restricted to participants from the Heidelberg center, and the λ yielding the highest C-index
1044 in Heidelberg defined the final set of proteins and their regression coefficients. These

1045 coefficients were used to compute the PD proteomic risk score (PPRS), which was then
1046 evaluated in the independent external AGES cohort in Cox models adjusted for age and sex.
1047 To this end, the improvement of the C-index from the baseline model (comprising age and
1048 sex only) to the baseline model + the top PPRS was quantified. This approach controls for
1049 age effects and avoids selecting aging markers as PD predictors. To assess whether the PPRS
1050 captures general neurodegenerative process, we also quantified the ΔC-index for AD in AGES
1051 with and without inclusion of *APOE* epsilon 4 genotypes.

1052 **References**

1053 1. Poewe, W. *et al.* Parkinson disease. *Nat. Rev. Dis. Primers* **3**, 1–21 (2017).

1054 2. Ray Dorsey, E. *et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis
1055 for the Global Burden of Disease Study 2016. *Lancet Neurol.* **17**, 939–953 (2018).

1056 3. Ishihara, L. S., Cheesbrough, A., Brayne, C. & Schrag, A. Estimated life expectancy of Parkinson's patients compared
1057 with the UK population. *J. Neurol. Neurosurg. Psychiatry* **78**, 1304–1309 (2007).

1058 4. Postuma, R. B. & Berg, D. Advances in markers of prodromal Parkinson disease. *Nature Reviews Neurology* vol. 12
1059 622–634 Preprint at https://doi.org/10.1038/nrneurol.2016.152 (2016).

1060 5. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**,
1061 2489–2498 (2024).

1062 6. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell*
1063 https://doi.org/10.1016/j.cell.2024.10.045 (2024) doi:10.1016/j.cell.2024.10.045.

1064 7. Helgason, H. *et al.* Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–
1065 735 (2023).

1066 8. Zetterberg, H. Clinically relevant factors for blood-based Alzheimer biomarker interpretation. *Alzheimer's &*
1067 *Dementia* **19**, (2023).

1068 9. Gail, M. H. *et al.* Design choices for observational studies of the effect of exposure on disease incidence. *BMJ Open*
1069 **9**, (2019).

1070 10. Gan, Y.-H. *et al.* Large-scale proteomic analyses of incident Parkinson's disease reveal new pathophysiological
1071 insights and potential biomarkers. *Nat. Aging* **5**, 642–657 (2025).

1072 11. Lill, C. M. *et al.* EPIC4ND: European Prospective Investigation into Cancer and Nutrition follow-up for
1073 neurodegenerative diseases. Preprint at https://doi.org/10.1101/2025.01.29.25321340 (2025).

1074 12. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer
1075 and Nutrition. *Int. J. Epidemiol.* **26**, 6S – 14 (1997).

1076 13. Riboli, E. *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data
1077 collection. *Public Health Nutr.* **5**, 1113–1124 (2002).

1078 14. Imam, F. *et al.* The Global Neurodegeneration Proteomics Consortium: biomarker and drug target discovery for
1079 common neurodegenerative diseases and aging. *Nat. Med.* **31**, 2556–2566 (2025).

1080 15. Doostparast Torshizi, A. *et al.* Proteogenomic network analysis reveals dysregulated mechanisms and potential
1081 mediators in Parkinson's disease. *Nat. Commun.* **15**, 6430 (2024).

1082 16. Belbasis, L., Morris, S., van Duijn, C., Bennett, D. & Walters, R. Mendelian randomization identifies proteins
1083 involved in neurodegenerative diseases. *Brain* awaf018 (2025) doi:10.1093/brain/awaf018.

1084 17. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-
1085 analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

1086 18. Doostparast Torshizi, A. *et al.* Proteogenomic network analysis reveals dysregulated mechanisms and potential
1087 mediators in Parkinson's disease. *Nat. Commun.* **15**, 6430 (2024).

1088 19. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic*
1089 *Acids Res.* **53**, D1467–D1475 (2025).

1090 20. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling.
1091 *Nat. Commun.* **12**, (2021).

1092 21. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature*
1093 **622**, 348–358 (2023).

1094 22. Eldjarn, G. H. *et al.* Correction to: Large-scale plasma proteomics comparisons through genetics and disease
1095 associations (Nature, (2023), 622, 7982, (348-358), 10.1038/s41586-023-06563-x). *Nature* vol. 630 E3 Preprint at
1096 https://doi.org/10.1038/s41586-024-07549-z (2024).

1097 23. Rooney, M. R. *et al.* Correlations Within and Between Highly Multiplexed Proteomic Assays of Human Plasma. *Clin.*
1098 *Chem.* **71**, 677–687 (2025).

1099 24. Wang, B. *et al.* Comparative studies of 2168 plasma proteins measured by two affinity-based platforms in 4000
1100 Chinese adults. *Nat. Commun.* **16**, 1869 (2025).

1101 25. Ali, M. *et al.* Shared and disease-specific pathways in frontotemporal dementia and Alzheimer's and Parkinson's
1102 diseases. *Nat. Med.* **31**, 2567–2577 (2025).

1103 26. Heo, G. *et al.* Large-scale plasma proteomic profiling unveils diagnostic biomarkers and pathways for Alzheimer's
1104 disease. *Nat. Aging* **5**, 1114–1131 (2025).

1105 27. Lill, C. M. *et al.* (Dys)regulation of the Immune System in Parkinson's Disease: Methodologies, Techniques, and Key
1106 Findings from Human Studies. *Aging Dis.* https://doi.org/10.14336/AD.2024.1163 (2025)
1107 doi:10.14336/AD.2024.1163.

1108 28. Tansey, M. G. *et al.* Inflammation and immune dysfunction in Parkinson disease. *Nat. Rev. Immunol.* **22**, 657–673
1109 (2022).

1110 29. Kedmi, M., Bar-Shira, A., Gurevich, T., Giladi, N. & Orr-Urtreger, A. Decreased expression of B cell related genes in
1111 leukocytes of women with Parkinson's disease. *Mol. Neurodegener.* **6**, 66 (2011).

1112 30. Tatenhorst, L. *et al.* Glypican-4 serum levels are associated with cognitive dysfunction and vascular risk factors in
1113 Parkinson's disease. *Sci. Rep.* **14**, 5005 (2024).

1114 31. Fan, Y. *et al.* Increased plasma lipocalin-2 levels are associated with nonmotor symptoms and neuroimaging
1115 features in patients with Parkinson's disease. *J. Neurosci. Res.* **102**, e25303 (2024).

1116 32. Elango, R. *et al.* Potential Biomarkers for Parkinson Disease from Functional Enrichment and Bioinformatic Analysis
1117 of Global Gene Expression Patterns of Blood and Substantia Nigra Tissues. *Bioinform. Biol. Insights* **17**,
1118 11779322231166214 (2023).

1119 33. Hepp, D. H. *et al.* Inflammatory Blood Biomarkers Are Associated with Long-Term Clinical Disease Severity in
1120 Parkinson's Disease. *Int. J. Mol. Sci.* **24**, (2023).

1121 34. Yao, L. *et al.* Bioinformatic Analysis of Genetic Factors from Human Blood Samples and Postmortem Brains in
1122 Parkinson's Disease. *Oxid. Med. Cell. Longev.* **2022**, 9235358 (2022).

1123 35. Irmady, K. *et al.* Blood transcriptomic signatures associated with molecular changes in the brain and clinical
1124 outcomes in Parkinson's disease. *Nat. Commun.* **14**, 3956 (2023).

1125 36. Miki, Y. *et al.* Alteration of autophagy-related proteins in peripheral blood mononuclear cells of patients with
1126 Parkinson's disease. *Neurobiol. Aging* **63**, 33–43 (2018).

1127 37. Santiago, J. A., Littlefield, A. M. & Potashkin, J. A. Integrative transcriptomic meta-analysis of Parkinson's disease
1128 and depression identifies NAMPT as a potential blood biomarker for de novo Parkinson's disease. *Sci. Rep.* **6**, 34579
1129 (2016).

1130 38. Kim, B.-W. *et al.* Pathogenic Upregulation of Glial Lipocalin-2 in the Parkinsonian Dopaminergic System. *J. Neurosci.*
1131 **36**, 5608–22 (2016).

1132 39. Hastings, N. *et al.* Connexin 43 is downregulated in advanced Parkinson's disease in multiple brain regions which
1133 correlates with symptoms. *Sci. Rep.* **15**, 10250 (2025).

1134 40. Lorenzl, S., Albers, D. S., Narr, S., Chirichigno, J. & Beal, M. F. Expression of MMP-2, MMP-9, and MMP-1 and their
1135 endogenous counterregulators TIMP-1 and TIMP-2 in postmortem brain tissue of Parkinson's disease. *Exp. Neurol.*
1136 **178**, 13–20 (2002).

1137 41. Morris, H. R., Spillantini, M. G., Sue, C. M. & Williams-Gray, C. H. The pathogenesis of Parkinson's disease. *Lancet*
1138 **403**, 293–304 (2024).

1139 42. Heinzel, S. *et al.* Update of the MDS research criteria for prodromal Parkinson's disease. *Movement Disorders* vol.
1140 34 1464–1470 Preprint at https://doi.org/10.1002/mds.27802 (2019).

1141 43. Bestwick, J. P. *et al.* Improving estimation of Parkinson's disease risk—the enhanced PREDICT-PD algorithm. *NPJ*
1142 *Parkinsons Dis.* **7**, (2021).

1143 44. Marini, K. *et al.* Comparison of different risk scores for Parkinson disease in a population-based 10-year study. *Eur.*
1144 *J. Neurol.* **30**, 3347–3352 (2023).

1145 45. Irmady, K. *et al.* Blood transcriptomic signatures associated with molecular changes in the brain and clinical
1146 outcomes in Parkinson's disease. *Nat. Commun.* **14**, 3956 (2023).

1147 46. Santiago, J. A., Littlefield, A. M. & Potashkin, J. A. Integrative transcriptomic meta-analysis of Parkinson's disease
1148 and depression identifies NAMPT as a potential blood biomarker for de novo Parkinson's disease. *Sci. Rep.* **6**, 34579
1149 (2016).

1150 47. Kedmi, M., Bar-Shira, A., Gurevich, T., Giladi, N. & Orr-Urtreger, A. Decreased expression of B cell related genes in
1151 leukocytes of women with Parkinson's disease. *Mol. Neurodegener.* **6**, 66 (2011).

1152 48. Tatenhorst, L. *et al*. Glypican-4 serum levels are associated with cognitive dysfunction and vascular risk factors in
1153 Parkinson's disease. *Sci. Rep.* **14**, 5005 (2024).

1154 49. Miki, Y. *et al*. Alteration of autophagy-related proteins in peripheral blood mononuclear cells of patients with
1155 Parkinson's disease. *Neurobiol. Aging* **63**, 33–43 (2018).

1156 50. Gallo, V. *et al*. Parkinson's Disease Case Ascertainment in the EPIC Cohort: The NeuroEPIC4PD Study.
1157 *Neurodegener. Dis.* **15**, 331–338 (2015).

1158 51. Consortium, T. I. Design and cohort description of the InterAct Project: an examination of the interaction of genetic
1159 and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* **54**, 2272–2282 (2011).

1160 52. Gold, L. *et al*. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, (2010).

1161 53. Candia, J., Daya, G. N., Tanaka, T., Ferrucci, L. & Walker, K. A. Assessment of variability in the plasma 7k SomaScan
1162 proteomics assay. *Sci. Rep.* **12**, (2022).

1163 54. PRENTICE, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**,
1164 1–11 (1986).

1165 55. Altman, D. G. & Bland, J. M. Interaction revisited: the difference between two estimates. *BMJ* **326**, 219 (2003).

1166 56. Lill, C. M. *et al*. Impact of Parkinson's disease risk loci on age at onset. *Movement Disorders* **30**, 847–850 (2015).

1167 57. Harris, T. B. *et al*. Age, Gene/Environment Susceptibility-Reykjavik Study: Multidisciplinary Applied Phenomics. *Am.*
1168 *J. Epidemiol.* **165**, 1076–1087 (2007).

1169 58. Wright, J. D. *et al*. The ARIC (Atherosclerosis Risk In Communities) Study. *J. Am. Coll. Cardiol.* **77**, 2939–2959 (2021).

1170 59. Sudlow, C. *et al*. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex
1171 Diseases of Middle and Old Age. *PLoS Med.* **12**, e1001779- (2015).

1172 60. Harris, T. B. *et al*. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am.*
1173 *J. Epidemiol.* **165**, 1076–87 (2007).

1174 61. Bankier, S. *et al*. Circulating causal protein networks linked to future risk of myocardial infarction. *medRxiv*
1175 https://doi.org/10.1101/2025.02.07.25321789 (2025) doi:10.1101/2025.02.07.25321789.

1176 62. Harris, T. B. *et al*. Age, Gene/Environment Susceptibility-Reykjavik Study: Multidisciplinary Applied Phenomics. *Am.*
1177 *J. Epidemiol.* **165**, 1076–1087 (2007).

1178 63. Scher, A. I. *et al*. Midlife migraine and late-life parkinsonism: AGES-Reykjavik study. *Neurology* **83**, 1246–52 (2014).

1179 64. Huang, X. *et al*. Statins, plasma cholesterol, and risk of Parkinson's disease: a prospective study. *Mov. Disord.* **30**,
1180 552–9 (2015).

1181 65. Doostparast Torshizi, A. *et al*. Proteogenomic network analysis reveals dysregulated mechanisms and potential
1182 mediators in Parkinson's disease. *Nat. Commun.* **15**, (2024).

1183 66. Karlsson, M. *et al*. A single–cell type transcriptomics map of human tissues. *Sci. Adv.* **7**, eabh2169 (2025).

1184 67. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell
1185 Transcriptomes and Expression Weighted Cell Type Enrichment. *Front. Neurosci.* **Volume 10-2016**, (2016).

1186 68. Zeisel, A. *et al*. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (1979).*
1187 **347**, 1138–1142 (2015).

1188 69. Szklarczyk, D. *et al*. The STRING database in 2023: protein–protein association networks and functional enrichment
1189 analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

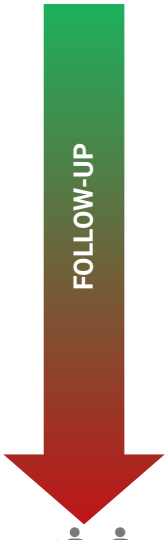
1190

Discovery cohort

EPIC4PD
(n=4,538)

Demographics
Lifestyle
Health data

Plasma proteomics
(SomaScan 7K,
7,285 aptamers)

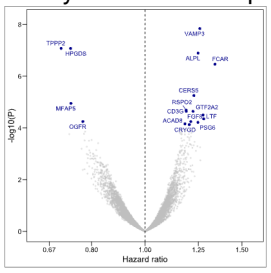


GNPC
(n=335 cases,
2,257 controls)

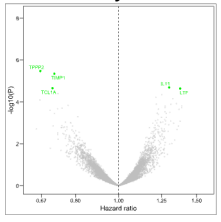
TRACKING PD
(n=794 cases)

Preclinical biomarker discovery

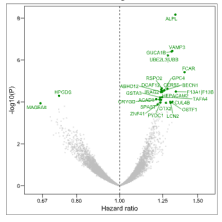
2-28 years of follow-up



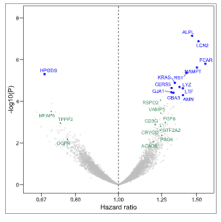
2-10 years



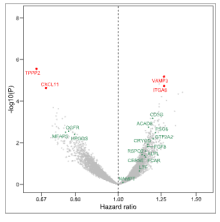
10-20 years



Men

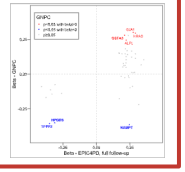


Women



Assessment of preclinical biomarkers in clinical PD

- 'Indirect' replication
- Cross-trait enrichment analysis
- Correlation analyses



Preclinical replication (n=64,856)

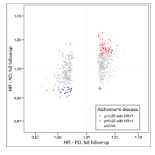
ARIC	AGES
total n=11,205 / n events=164	total n=5,090 / n events=121
UKB total n=48,561 / n events=749	

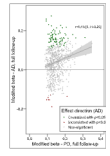
Functional characterization

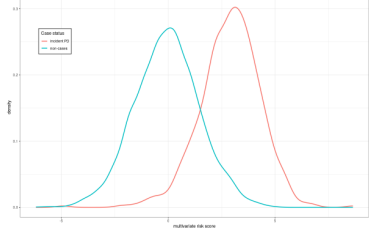
- Gene ontology enrichment
- Cell-type enrichment
- Protein-protein interaction networks
- Drug repurposing
- Literature mining/candidate biomarkers

Cross-phenotype comparisons

- Comparison with AD and ALS
- Cross-trait enrichment analysis





Prediction

- Variable selection by LASSO and construction of protein risk score
- External validation in AGES