Large-scale plasma proteomics uncovers preclinical molecular signatures of Parkinson's disease and overlap with other neurodegenerative disorders

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)[$], Alexis Elbaz[12], Moritz Hess[33], Verena Katzke[26], Lars Bertram[10], Vilmundur Gudnason[4,5], Oliver Robinson[2,9], Honglei Chen[34], Lefkos Middleton[9], Laura M. Winchester[8], Ioanna Tzoulaki[2], Valborg Gudmundsdottir[4,5], Keenan A. Walker[35], 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. Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany
34. Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan, USA
35. 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

Parkinson's disease (PD) remains incurable, with a long preclinical phase currently undetectable by existing 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 EPIC4PD discovery case-cohort, large-scale profiling of 7,285 proteins (SomaScan-7K) in 4,538 initially unaffected participants (574 incident cases) identified 17 proteins that predict PD up to 28 years before diagnosis. Additional proteins revealed sex-specific effects and time-dependent effect trajectories, capturing disease progression before symptom onset. Replication in three prospective cohorts (n=64,856; 1,034 incident cases) confirmed at least 12 key pre-diagnostic biomarkers with strong evidence, including TPPP2, HPGDS, ALPL, MFAP5, OGFR, ACAD8, TCL1A, GPC4, GSTA3, LCN2, KRAS, and GJA1. Preclinical biomarkers showed 86% concordant effect directions in independent prevalent PD cases (n=2,592; $p=1.6\times10^{-19}$). Furthermore, in the longitudinal Tracking PD cohort (n=794), 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 shared molecular signatures. A machine learning-derived protein score improved PD risk prediction in external validation. This extensive proteomics effort identified novel, actionable biomarkers opening new avenues for early PD risk stratification and precision medicine.

Keywords: Parkinson's disease, neurodegeneration, prediction, proteomics, protein risk score, SomaLogic, aptamers, EPIC

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].

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

148 **RESULTS**

149 **EPIC4PD case-cohort**

150 In the discovery phase of this study (**Figure 1**), we leveraged data from the EPIC cohort to
151 build a large case-cohort (EPIC4PD) of 4,538 participants (including 574 incident PD cases;
152 **Table 1**) across 10 European centers in Spain, Italy, the Netherlands, the UK, and Germany,
153 combining deep proteomic profiling (6,382 proteins via 7,285 aptamers) in baseline, i.e, pre-
154 disease plasma samples, with long-term prospective follow-up. EPIC4PD participants had a
155 median baseline age of 53 years (range: 35–79), comprising 63% women (57% excluding the
156 Utrecht center, which recruited only women), and were followed for up to 28 years (median:
157 16 years). Among PD cases, the median time from blood sampling to diagnosis was 11 years
158 (range: 2–25), enabling biomarker discovery long before clinical onset. Consistent with
159 known PD risk profiles, PD cases were older at baseline (median: 59 vs. 52 years), more likely
160 male (58% vs. 40%, excluding Utrecht), less likely to be current smokers (15% vs. 25%), and
161 slightly more often overweight (64% vs. 60% with body mass index (BMI) ≥25) compared to
162 non-cases (**Supplementary Tables 1–2**).

163 **Association analyses in the full EPIC4PD dataset**

164 We used Cox proportional hazards regression models across 7,285 aptamers to identify
165 proteins associated with a future PD onset (**Supplementary Table 3**). In the basic model
166 (stratified by age in 5-year intervals, sex, and center) covering the full follow-up period (2–28
167 years), 17 proteins were significantly associated with PD risk after a false-discovery rate
168 (FDR) control of 0.05 (**Table 2, Supplementary Table 4**). We applied two approaches in the
169 discovery phase to address outliers: one excluded extreme values as potential noise, and the
170 other retained them via capping, as they may represent strong PD signals. Nine proteins
171 were FDR-significant in both approaches; for the other eight (six upon exclusion, two upon
172 capping), the alternative model still showed consistent, nominally significant associations
173 (mean ΔHR = 0.03, maximum ΔHR=0.13; **Supplementary Figure 1**), supporting robustness
174 of our findings. For simplicity, we report only one result (from the better-performing model)
175 in the main text, with capped data labeled "cap"; full results are available in **Supplementary**
176 **Table 3**. Top proteins associated with PD incidence included VAMP3 (hazard ratio [per 1
177 standard deviation (SD) increase in $\log_{10}$-transformed protein levels]=1.26, p=1.46E-08),
178 TPPP2 (HR=0.70, p=8.52E-08), HPGDS (HR=0.73, p=8.59E-08), $ALPL_{cap}$ (HR=1.25, p=1.30E-07),
179 FCAR (HR=1.34, p=3.48E-07), and CERS5 (HR=1.23, p=5.64E-06; **Figure 2**, **Table 2**). All FDR-

180 significant associations remained stable after additional adjustment for key demographic
181 and lifestyle factors (education level, BMI, smoking, and physical activity), with minimal
182 impact on effect sizes (mean ΔHR=0.02, maximum ΔHR=0.07; **Supplementary Figure 1**,
183 **Supplementary Table 4**).

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

185 Of the 17 FDR-significant biomarkers identified in EPIC4PD, replication data were available in
186 a total of 64,856 participants (including 1,034 incident PD cases) from three prospective
187 studies: the Gene/Environment Susceptibility–Reykjavik Study (AGES), Atherosclerosis Risk in
188 Communities study (ARIC), and UKB. Specifically, all 17 proteins were available in AGES
189 (5,090 participants, 121 cases; SomaScan 7K), 15 in ARIC (11,205 participants, 164 cases;
190 SomaScan 5K), and 8 in UKB (48,561 participants, up to 749 cases; Olink 3072 Explore;
191 **Supplementary Table 5**, **Figure 3**, **Supplementary Figure 2**). The other 9 proteins were not
192 measured in UKB, as the current Olink platform captures only ~30% of the proteins assayed
193 by SomaScan 7K. We also included independent SomaScan 7K-based data from prevalent
194 cases as 'indirect' replication. To this end, we re-analyzed proteomic profiles from 2,592
195 participants (335 prevalent PD cases and 2,257 controls) across three GNPC case-control
196 datasets (also see below for comparison of incident and prevalent PD), using similar
197 methodology as with our discovery phase analyses.

198 When interpreting the replication results, platform-related differences between Olink and
199 SomaScan must be considered. In line with the previous UKB study[10], we observed a global
200 protein downregulation in incident PD cases in our re-analyses of the UKB Olink data (i.e.,
201 87% of nominally significantly associated proteins). However, this pattern was not seen in
202 SomaScan-based datasets from EPIC4PD, AGES, or ARIC (proportion of downregulated
203 proteins: 43% to 56%; **Supplementary Table 6**), underscoring the challenges of cross-
204 platform comparisons.

205 Given the higher comparability of results from datasets generated using the same
206 technology, we first considered replication evidence in the SomaScan-based datasets (AGES,
207 ARIC, and GNPC). Overall, six of the 17 proteins — TPPP2, HPGDS, ALPL, MFAP5, OGFR, and
208 ACAD8 — showed strong evidence for replication, i.e., nominally significant evidence for
209 association in at least one dataset and, moreover, fully consistent effect estimates across all
210 independent studies with available data (**Figure 3**). Furthermore, VAMP3, GTF2A2, and FGF8
211 showed moderate evidence for replication, i.e., highly consistent effect directions across all

212 available datasets albeit without any replication dataset individually reaching nominal
213 significance (**Figure 3**, **Table 2, Supplementary Table 5**). Upon reviewing the replication
214 evidence in the UKB, the evidence of association was further substantiated for TPPP2,
215 HPGDS, MFAP, and OGFR, while ALPL, ACAD8, VAMP3, GTF2A2, and FGF8 were not
216 measured by Olink. Several additional proteins showed suggestive replication evidence in
217 individual datasets, but the overall replication evidence was more heterogeneous (**Table 2**,
218 **Supplementary Figure 2, Supplementary Table 6**).
219 In summary, nine of the 17 biomarkers (53%; ACAD8, HPGDS, TPPP2, OGFR, MFAP5, ALPL,
220 FGF8, VAMP3, GTF2A2) were replicated across different technologies, populations, and
221 study designs, reinforcing their specific predictive potential, and, more generally, support
222 the generalizability of our biomarker signals.

223 **Biomarker associations relative to disease onset**

224 Stratifying analyses by temporal proximity to PD diagnosis revealed remarkable biomarker
225 dynamics. In the window closest to disease onset (2-10 years; n=4,230 participants including
226 245 incident cases), three additional proteins — IL11, TCL1A, and $TIMP1_{cap}$ — emerged as
227 FDR-significant. In the more distal window (10-20 years; n=4,023 including 307 incident
228 cases), 19 additional proteins showed strong associations with PD incidence (**Figure 2, Table**
229 **2**, **Supplementary Table 4**). These proteins had not been identified as significant overall due
230 to negligible effects in the other time window (**Supplementary Table 3, Supplementary**
231 **Figure 2**). Dynamic biomarker patterns were further supported by HR trajectories derived
232 from 5-year sliding windows (**Figure 2, Supplementary Figure 3**). Together, these findings
233 demonstrate that key biomarkers for PD may emerge at distinct preclinical phases,
234 underscoring the value of deep proteomic profiling with extended follow-up to reveal
235 dynamic, stage-specific signatures of preclinical disease.

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

237 Of the three newly emerging biomarkers from the proximal follow-up window (2-10 years
238 before diagnosis), TCL1A showed strong independent replication evidence and TIMP1
239 showed moderate support, while IL11 did not replicate (**Figure 3**, **Supplementary Figure 2,**
240 **Supplementary Table 5**). Among the 19 biomarkers emerging from the 10-20 year window,
241 GPC4, GSTA3, and LCN2 exhibited strong evidence of replication, and ABHD12 showed
242 moderate support (**Figure 3**). For LCN, we observed nominally significant evidence for

replication in ARIC with the same effect direction as in EPIC4PD. However, in UKB, LCN2 also reached nominal significance (p=3.80E-03) but showed an inverse direction of effect (**Figure 3**, **Supplementary Table 5**). This discrepancy likely reflects methodological differences between the SomaScan and Olink proteomic platforms. In line with this and the general trend mentioned above, all but one biomarker (MAGEA8) were upregulated in EPIC4PD, but in all but one instance (HEPACAM2) downregulated in UKB. Generally, replication datasets were fewer in number and had smaller sample sizes within the shorter time windows (**Methods** and **Supplementary Table 5**), which may partly explain the generally more modest replication evidence observed.

Sex-stratified analyses

In EPIC4PD, 18% (3 of 17) of proteins significantly associated with PD risk in the full follow-up period showed clear sex-specific effect size differences and significant interaction with sex after FDR control. These proteins included $NAMPT_{cap}$ (ΔHR = 48%), FCAR (ΔHR = 42%), and $ALPL_{cap}$ (ΔHR = 34%; **Supplementary Figure 1**, **Supplementary Table 7**).

Furthermore, sex-stratified proteome-wide analyses uncovered six additional proteins (RS1, $KRAS_{cap}$, LYZ_{cap}, $GBA3_{cap}$, GJA1, AMN) FDR-significantly associated with PD risk in men and two (CXCL11, ITGA6) in women, none of which showed strong associations in the opposite sex (HRs 0.92–1.10, all $p \geq 0.05$; **Supplementary Figure 1**, **Supplementary Figure 4; Supplementary Table 4**). Interaction analyses confirmed FDR-significant sex-based risk modifications for all male-specific biomarkers except GJA1. However, the latter and CXCL11 still showed nominally significant evidence for interaction (**Supplementary Table 7**). The results of these 11 FDR-significant proteins in the sex-stratified analyses remained stable after adjusting for lifestyle variables and, in women, for menopausal status and postmenopausal hormone therapy (**Supplementary Figure 1**, **Supplementary Table 4**). $KRAS_{cap}$ and GJA1 demonstrated strong evidence of replication in men, while AMN (men) and CXCL11 (women) provided moderate evidence for replication **(Figure 3)**. The limited sample sizes in the sex strata of the AGES and ARIC cohorts precluded analyses of interaction with sex. In the UKB, 5 of the 11 candidate sex-specific biomarkers had been quantified; however, none showed statistically significant associations in the sex strata, possibly reflecting methodological differences between SomaScan and Olink platforms (see Discussion). Collectively, these findings underscore the importance of accounting for sex-

274 specific effects in biomarker discovery, as such effects may reveal associations that remain
275 hidden in sex-combined analyses.

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

277 Several FDR-significant biomarkers showed moderate inter-correlations (r=0.4-0.6) in
278 EPIC4PD, suggesting shared regulation or involvement in the same pathway. Correlations
279 were observed between MFAP5 and both HPGDS and TPPP2 (r=0.54 each) and between
280 LCN2 and LTF (r=0.60). FCAR also featured in several correlation clusters, underscoring its
281 broader connectivity (detailed in **Supplementary Figure 5** and **Supplementary Table 8)**.

282 **Characterization of preclinical biomarkers in prevalent PD and PD progression**

283 Next, we investigated whether the effects of the identified preclinical biomarkers extended
284 beyond the risk phase into the symptomatic stage of PD by leveraging the SomaScan 7K
285 proteomic data from 2,592 independent GNPC participants (335 prevalent cases and 2,257
286 controls; **Figure 3, Supplementary Table 9**; see the companion manuscript of Imam et al.[14]
287 for full dataset description). Of all 47 preclinical biomarkers detectable in GNPC, 7 (15%;
288 TPPP2, HPGDS, ALPL, NAMPT, GSTA3, KRAS, GJA1) were nominally significantly associated
289 with prevalent PD, and 4 (TPPP2, HPGDS, NAMPT, GSTA3) remained significant after FDR
290 control. All but NAMPT showed consistent effect directions. Overall, 38 (81%) of the 47 top
291 proteins and 125 (86%) of the 146 nominally significant aptamers results shared between
292 both traits showed concordant directions of effect — significantly more than expected by
293 chance (p=1.25E-5 and p=1.64E-19, respectively; one-sided exact binomial test; **Figure 3;**
294 **Supplementary Table 9**). In serum samples from the Tracking Parkinson's cohort (n=794), we
295 assessed associations between FDR-significant preclinical biomarkers and clinical outcomes
296 related to disease progression, using motor (Unified Parkinson's Disease Rating Scale Part III,
297 UPDRS-III) and cognitive (Montreal Cognitive Assessment, MoCA MoCA) scores in cross-
298 sectional and longitudinal mixed-effects models (**Figure 3**, **Supplementary Table 10**). Two
299 biomarkers, HPGDS (β=2.61, p=6.82E-04) and MFAP5 (β=1.31, p=1.49E-03), were FDR-
300 significantly associated with higher MoCA scores over time, indicating that lower levels of
301 these proteins predict both increased PD risk and faster cognitive decline. Associations
302 remained robust after adjusting for BMI and disease duration (data not shown). These
303 results demonstrate that a considerable fraction of preclinical biomarkers persist in their
304 dysregulation in prevalent PD, and a small subset is even associated with progression.

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

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

310 Of 291 unique proteins previously linked to PD across methodologically diverse sources[10,15–
311 17], 191 proteins were assessable in the EPIC4PD data (**Supplementary Table 11**): Most
312 notably, 38 proteins had recently been reported as putative predictive PD biomarkers in the
313 UKB Olink study[10], with 27 also covered in our independent dataset. Of these, 11 (41%;
314 HPGDS, CD209, TPPP3, EPS8L2, FCRL1, BAG3, TNXB, IL13RA1, NEFL, PPY, RET) showed at
315 least nominally significant associations in one or more follow-up intervals. HPGDS also
316 represents a top biomarker in EPIC4PD (**Table 2**). The next two most significant proteins
317 were CD209 (best model for follow-up of 10-20 years: HR=0.75, p=1.32E-04) and TPPP3 (best
318 result: 2-28 years: HR=0.80, p=1.58E-03), which aligns well with TPPP2, one of the lead
319 findings in EPIC4PD (see Discussion). All but one (PPY) of the at least nominally significantly
320 associated biomarkers showed the same effect direction as in the previous study[10] despite
321 platform differences. Of the 127 candidate neurological biomarkers from the CNS NULISA
322 panel (Alamar, Inc), 106 were tested in EPIC4PD; of these, UBE2L3|UBB and CCL11
323 represented top biomarkers in EPIC4PD (**Table 2**). Other nominally significant findings
324 included for instance SNCA, DDC, NEFL, TREM2, GFAP, and TIMP3. Furthermore, PYDC1
325 nominated by a recent Mendelian randomization (MR) and colocalization study[18] was among
326 the top results in EPIC4PD suggesting a causative link to PD risk. Several other candidates
327 from an additional MR[16] or GWAS[17] showed nominal associations in EPIC4PD
328 (**Supplementary Table 11**).

329 Conversely, 5 of the 39 FDR-significant PD biomarkers identified in our non–sex-stratified
330 EPIC4PD analyses (CRYGD, LTF, HPGDS, TCL1A, and LCN2) were also included in the MR study
331 based on SomaScan data[16]. Of these, two biomarkers (CRYGD and LTF) showed nominal
332 evidence of a causal association with PD, suggesting that higher circulating levels may
333 increase disease risk, in line with the EPIC4PD findings (**Supplementary Table 12**). The
334 limited overlap in assessed proteins likely reflects the comparatively stringent analysis
335 criteria applied in that study [16].

336 **Cross-disease comparisons**

337 The 47 FDR-significant PD biomarkers identified in EPIC4PD were also evaluated in incident
338 AD (n=1,926, including 652 cases) and ALS (n=4,715, including 187 cases) using Cox
339 regression analyses in the EPIC4AD and EPICALS case-cohorts (**Supplementary Table 1**),
340 respectively. Sixteen (34%) preclinical PD biomarkers showed at least nominally significant
341 associations with AD and/or ALS, all with concordant effect directions. Correlation plots of
342 hazard ratios showed alignment in effect size estimates (**Figure 3**). Notably, three PD
343 proteins — RSPO2, $SPAST_{cap}$, PYDC1 — were FDR-significant for AD and also showed
344 nominally significant associations with ALS (**Table 2, Figure 3, Supplementary Table 13**).
345 At the proteome-wide level, 17% of proteins nominally significantly associated with PD over
346 the full follow-up period (non-sex-stratified) also showed associations with AD and/or ALS,
347 whereas 83% appeared to be PD-specific. Among the shared proteins, the majority (68%)
348 overlapped exclusively with AD, 25% with ALS, and 7% were common to all three
349 neurodegenerative diseases (**Supplementary Figure 6**). Nominally significant preclinical AD
350 biomarkers were significantly enriched among nominally significant PD biomarkers (p=1.19E-
351 03; Fisher's exact test), while ALS biomarkers were not (p=0.107). Notably, 85 of 92 PD–AD
352 and 33 of 39 PD–ALS shared proteins showed concordant effect directions (p=1.92E-18 and
353 p=7.15E-06; exact binomial test; **Figure 3**). Additionally, among proteins nominally
354 associated with PD, effect sizes modestly correlated with AD and ALS (Pearson's r=0.18 [95%
355 CI: 0.11–0.25] and r=0.21 [0.14–0.28], respectively; **Figure 3**). While these comparisons
356 highlight a largely PD-specific biomarker signature, they also point to partially shared
357 preclinical pathways.

358 **Functional characterization of preclinical PD biomarkers**

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

368 several vitamin-related processes, including response to vitamin D, also emerged, pointing to
369 molecular pathways that were especially prominent in earlier PD research. Enrichments in
370 'cellular component' and 'molecular function' categories reinforced the role of vesicle-
371 mediated activity, while Reactome analysis further emphasized immune system
372 involvement, including antimicrobial peptides (**Supplementary Tables 14**).
373 Expression heatmaps across 50 human bulk tissues and single brain cell types based on RNA-
374 seq data revealed relatively high expression of several FDR-significant biomarkers in
375 peripheral and central immune-related tissues and cells compared to other tissue types
376 (**Supplementary Figures 7-8**). Protein-level data from the Human Protein Atlas supported
377 this observation (**Supplementary Figure 9**). Consistent with these findings, single-cell RNA-
378 seq–based enrichment analyses of FDR-significant and suggestive biomarkers in mice and
379 humans supported a role for immune cells in preclinical PD. In mouse cortex, suggestive
380 biomarkers showed nominally significant enrichment in microglia (**Supplementary Table 15**).
381 Analyses across 23 representative human cell types (including immune, neuronal, intestinal,
382 muscle, liver, and fibroblast cells) demonstrated nominally significant enrichment in
383 fibroblasts, monocytes and Kupffer cells, with borderline significance in granulocytes and
384 macrophages (**Supplementary Table 15**). Collectively, these single-cell RNA-seq–based
385 findings align with our Gene Ontology and Reactome pathway results, highlighting early
386 immune system involvement as a key feature of preclinical PD.
387 Interestingly, protein–protein interaction (PPI) analysis revealed that FDR-significant
388 biomarkers were significantly more likely to interact with GWAS-nominated PD risk genes[17]
389 than expected by chance (p=1.0E-04; permutation test; **Figure 3**, **Supplementary Table 16**).
390 Within this network, SNCA emerged as a central hub, exhibiting several direct edges with
391 FDR-significant biomarkers (**Figure 3**). Consistently, *SNCA* expression in both brain and
392 immune cells was significantly correlated with several of these biomarkers (**Supplementary**
393 **Table 17**). In brain, strong correlations (r≥0.5) were observed with RS1, RSPO2, and OSTF1,
394 and in immune cells, with LYZ, VAMP3, F13A1, and FCAR. Furthermore, SNCA showed
395 significant correlations with FDR-significant preclinical biomarkers more often than expected
396 by chance (p=1.12E-02 for immune cells; p=9.1E-02 for brain cells; permutation t test). These
397 findings provide converging evidence that the identified preclinical biomarkers capture early
398 pathophysiological processes consistent with independent lines of prior research.

399 Finally, a semi-systematic literature search revealed prior functional evidence for a major
400 part of the 47 preclinical PD biomarkers (details in the **Supplementary Material**
401 [**Supplementary Table 18**]).

402 **Druggability assessment of preclinical biomarkers**

403 Given their potential pathophysiological implications, we next evaluated whether any of the
404 47 FDR-significant preclinical PD biomarkers represent druggable targets using the Open
405 Targets database[19]. Twenty-one proteins (45%) showed at least one supporting line of
406 evidence for interaction with small-molecule drugs (**Supplementary Table 19**). Notably, four
407 proteins — NAMPT, KRAS, CD3G, and F13A1 — were associated with compounds currently
408 undergoing clinical trials in non-neurological diseases (mostly cancer), underscoring their
409 translational potential. These findings indicate that a substantial share of preclinical PD
410 biomarkers may be pharmacologically modifiable.

411 **Disease prediction**

412 To assess predictive utility, we applied least absolute shrinkage and selection operator
413 (LASSO)-based machine learning in EPIC4PD with repeated subsampling and regularization.
414 Across 100 subsamples, the most frequently selected proteins were extracted with their
415 corresponding average weights (**Supplementary Table 20**). As external validation, we
416 calculated weighted predictive protein risk scores (PPRS) in the AGES cohort (n=5,090; 121
417 incident PD cases), leveraging the shared SomaScan 7K platform. Risk scores were based on
418 EPIC4PD results, varying by the number of proteins and using average LASSO-derived
419 coefficients. Given the marked sex differences in EPIC4PD biomarker signatures, primary
420 models were developed separately for men and women, with additional combined models
421 adjusted for sex for comparison (**Supplementary Tables 20-21**). In men (n=2,136; 75 PD
422 cases), the best-performing PPRS (50 proteins+age, C-index: 0.63, SE: 0.03) largely
423 outperformed the age-only model with ΔC=0.08. Nine proteins that were FDR-significant in
424 the univariate Cox models in EPIC4PD were also included in this PPRS (including GSTA3, RS1,
425 and HPGDS; **Supplementary Table 20**). In women (n=2,954; 46 cases), the best PPRS (20
426 proteins+age, C-index: 0.61; SE: 0.04) improved prediction similarly with ΔC=0.07. This PPRS
427 included three (CXCL11, TPPP2, and HPGDS) of the univariate analyses in EPIC4PD. As
428 expected, improvement of risk prediction in the combined cohort was attenuated (ΔC=0.03
429 for best score). These findings emphasize the necessity of incorporating sex-stratified

430 analyses into disease prediction research to enable the development of high-precision risk
431 models.

432 **DISCUSSION**

433 In this large, prospective, and multi-center proteomic study, we identified and
434 independently validated novel plasma biomarkers predictive of PD onset up to two decades
435 or more before clinical diagnosis. Using the high-resolution SomaScan 7K platform in pre-
436 diagnostic samples from over 4,500 EPIC4PD participants, we discovered 47 proteins
437 significantly associated with incident PD. Several top markers — TPPP2, HPGDS, ALPL,
438 MFAP5, OGFR, ACAD8, TCL1A, GPC4, GSTA3, LCN2, KRAS, and GJA1 — showed strong
439 replication across incident (AGES, ARIC, UKB) and prevalent (GNPC) PD cases, underscoring
440 their robustness across platforms and populations. Importantly, owing to the much larger
441 proteomic resolution our study uncovered numerous novel PD biomarkers not assayed in
442 the previous analyses in the UKB[10]. In addition, while we observed that pre-clinical protein
443 biomarker profiles are partially shared between PD and other neurodegenerative diseases
444 (i.e., AD and ALS), delineated PD biomarker profile appears to be largely disease-specific.
445 Overall, our study delivers the most comprehensive proteomic resource for preclinical PD to
446 date. It is unique by incorporating diverse preclinical validation cohorts, by spanning the full
447 disease course, and by demonstrating some overlap with the preclinical phase of other
448 neurodegenerative disorders. As such it sets a new benchmark for biomarker discovery in
449 neurodegeneration.
450 Our findings offer several new key insights: First, predictive PD biomarkers can be reliably
451 identified in blood many years, even decades, before diagnosis. This relates to signals such
452 as HPGDS, MFAP5, OGFR, and ACAD8, which remain stable across different platforms and
453 populations. Second, we also observed dynamic, stage-dependent biomarker patterns: while
454 some proteins persist throughout all stages, others appear only within certain time windows
455 before diagnosis (such as TCL1A, GPC4, GSTA3, and LCN2), possibly reflecting underlying
456 pathophysiological shifts. This temporal resolution is critical for distinguishing genuine early
457 biomarker signals from late-stage or treatment-related changes. Third, we observed
458 evidence for sex-specific effects for several biomarkers in EPIC4PD, an observation not seen
459 in our re-analysis of the Olink-based UKB data. This suggests that some sex-dependent
460 protein isoforms — such as specific glycosylation patterns or cleavage products — may be
461 captured by SomaScan but not by Olink's dual-antibody format, which relies on
462 simultaneous binding to two defined epitopes[20]. In this context, it is striking that in the
463 previous UKB study[10], all top preclinical biomarkers were *decreased* in incident PD cases.

464 While we confirmed this pattern in our re-analysis of the UKB data, it was not observed in
465 any of the SomaScan-based datasets. At the same time, the substantial concordance in
466 effect directions across the independent SomaScan-7K-based data from EPIC4PD and GNPC
467 supports the validity of our findings. Thus, the observed discrepancies between this study
468 and UKB[10] likely reflect methodological differences between proteomic platforms used.
469 While recent evidence suggests that the latest SomaScan platforms provide higher
470 measurement precision in comparison to Olink[21–25], importantly, the two platforms capture
471 partly distinct protein features, underscoring their complementary strengths[21]. A
472 comprehensive understanding of these differences will require benchmarking against mass
473 spectrometry. Until then, it is important to recognize that genuine protein associations may
474 fail to replicate or even show opposite effects when comparing results across these two
475 technologically different platforms.
476 Fourth, we observed high concordance (>80%) in effect size directions between preclinical
477 and clinical biomarkers across two independent datasets (EPIC4PD and GNPC, both
478 SomaScan 7K), supporting the robustness of our findings and providing indirect replication.
479 These comparisons are essential to identify which biomarkers are relevant across all stages
480 vs. those that are stage-specific, informing their use for early detection, diagnosis, or
481 monitoring. Fifth, we were in the unique position to probe for an overlap of preclinical
482 biomarker profiles of PD and two other neurodegenerative diseases (AD and ALS). About
483 one-third of PD biomarkers also showed consistent associations with incident AD and/or
484 ALS, including RSPO2, SPAST, and PYDC1. While effect sizes correlated modestly across
485 diseases, the clear enrichment of shared biomarkers — especially between PD and AD —
486 suggests partially overlapping preclinical mechanisms. Notwithstanding these overlaps, we
487 emphasize that most identifed biomarkers were specific to PD.
488 Sixth, this latter conclusion is confirmed by a wide array of *in silico* analyses supporting the
489 biological plausibility of many of our biomarker findings. For instance, we saw an
490 overrepresentation of key PD-relevant processes, including vesicle trafficking, dopaminergic
491 neuron biology, and immune signaling. Thus, it appears that these disease-related
492 mechanisms are already active and detectable in blood over two decades before diagnosis.
493 Furthermore, cell-type enrichment analyses implicated the involvement of microglia and
494 peripheral immune cells already in the earliest disease phases, aligning with current
495 neuroinflammation models of PD[26,27]. Moreover, our PPI analyses suggest that the preclinical

496 biomarkers interacted more frequently than expected with known PD proteins and
497 significantly correlated with *SNCA* expression, strongly emphasizing their functional
498 relevance to the disease. Along these lines, several of our newly detected preclinical
499 biomarkers had been implicated in biomarker as well as in functional studies in PD before.
500 For example, TCL1A[28], GPC4[29], LCN2[30], AMN[31], CXCL11[32], CUL4B[33], FCAR[34], BECN1[35], and
501 NAMPT[36] have previously been reported as potential blood biomarkers in prevalent PD in
502 studies with smaller sample sizes. In addition, some of the preclinical PD biomarkers were
503 reported to show differential expression in post-mortem brain tissue from PD patients
504 compared to controls (e.g., LCN2[37], GJA1[38], TIMP1[39], AMN[31]). Lastly, several of our top
505 biomarker proteins are implicated in microtubule dynamics (TPPP2), vesicle trafficking in the
506 presynapsis (VAMP3), mitochondrial functions (ACAD8), dopaminergic neuronal function and
507 maturation (GPC4, FGF8) and immune functioning/neuroinflammation (TCL1A, CXCL12),
508 which are all highly relevant in PD pathophysiology[40]. A detailed summary of key
509 experimental findings and supporting evidence, can be found in the **Supplementary**
510 **Material**. Seventh, our druggability assessments revealed that nearly half of the FDR-
511 significant biomarkers show evidence for small-molecule interactions. While blood-based
512 biomarkers of brain diseases (like PD) do not automatically translate into therapeutic
513 targets, this overlap may help prioritize candidates for future functional studies and
514 repurposing efforts.
515 Last, we developed a sex-specific protein risk score (n=50 men, n=20 women), which showed
516 robust predictive value in external validation within the AGES cohort and outperformed sex-
517 combined models. These findings underscore both the potential of proteomic risk profiling
518 for early PD risk stratification and the importance of considering sex differences.
519 Incorporating genetic, lifestyle, and environmental factors is expected to further improve
520 prediction. Along these lines, existing tools, such as the updated MDS criteria (specialist
521 settings) and the PREDICT-PD algorithm (broader screening)[41,42], show promise but remain
522 insufficient, particularly for detecting preclinical PD more than five years before diagnosis
523 (e.g., ref. [43]). These tools are continuously evolving, and integrating our protein biomarkers
524 alongside sex-stratified analyses offers a logical next step to improve predictive accuracy.
525 Our study has several strengths, including representing the largest collection of incident PD
526 proteomic datasets to date, an exceedingly long-term follow-up, rigorous case
527 ascertainment, the inclusion of several external datasets ensuring population and platform

528 cross-validations, as well as extension to other disease phenotypes and diseases.
529 Notwithstanding, our study also has several potential limitations. First, use of distinct
530 proteomic platforms with different technologies, particularly SomaScan versus Olink, and
531 coverages, comparatively limited case numbers despite large replication cohorts, usage of
532 different blood sample types (citrate plasma, EDTA plasma, or serum) as well as differences
533 in study design (e.g., recruitment, case ascertainment, follow-up) may have contributed to
534 non-replication of some biomarkers. Second, as outlined above, while the SomaScan 7K
535 assay offers extensive coverage, it is semi-quantitative; signal changes may reflect different
536 proteoforms rather than absolute protein levels, complicating interpretation. Mass
537 spectrometry-based approaches are needed to clarify the underlying biology. Third, although
538 we adjusted for key covariates in sensitivity analyses, residual confounding remains possible.
539 Fourth, lacking neuropathological confirmation, some PD cases may have been misclassified
540 — a common problem in large cohorts and a general limitation in epidemiological studies.
541 However, given that medical charts of PD cases were reviewed by clinical specialists, the risk
542 of misclassification in EPIC4PD is likely lower than in cohorts such as UKB. Finally, all datasets
543 consisted exclusively or predominantly of non-Hispanic White participants. While the ARIC
544 study also included a subset of Black participants (n=2,625) and their exclusion did not affect
545 effect estimates, the sample size was too limited to allow for dedicated ethnicity-specific
546 analyses of this subset. Thus, while our findings are likely generalizable to populations of
547 European ancestry, their applicability to other ethnicities requires further assessment.

548 In conclusion, our study identifies and validates robust blood-based biomarkers that predict
549 PD several decades before clinical onset. Our results underscore the importance of applying
550 prospective cohort designs with incident case ascertainment to detect truly preclinical
551 biomarkers. By capturing early molecular changes, our work lays the foundation for
552 developing prevention strategies using precision medicine approaches in PD and potentially
553 other neurodegenerative diseases.

Acknowledgements

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

Data access

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

Funding

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

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

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

596 National Institute on Aging (NIA) contracts N01-AG-12100 and HHSN271201200022C (for Vi.
597 G.) and Althingi (the Icelandic Parliament) financed the AGES-Reykjavik study. IHA and
598 Novartis have collaborated on proteomics research since 2012. This study was also funded
599 by the NIA (1R01AG065596-01A1 to Vi. G., 2R01AG065596-03 to Va.G.), the University of
600 Iceland Postdoctoral Fund (to E.A.F.) and the Icelandic Research Fund (2511008-051 to
601 Va.G.). This research was supported, in part, by the Intramural Research Program (IRP) of the
602 NIA. K.A.W. and F.L. are supported by the NIA IRP. ARIC is carried out as a collaborative study
603 supported by National Heart, Lung, and Blood Institute (NHLBI) contracts (75N92022D00001,
604 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). The ARIC
605 Neurocognitive Study was additionally supported by U01HL096812, U01HL096814,
606 U01HL096899, U01HL096902, and U01HL096917 from the NIH (NHLBI, NIA, National
607 Institute of Neurological Disorders and Stroke [NINDS], and National Institute on Deafness
608 and Other Communication Disorders [NIDCD]).

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

613 **Authors' contributions**

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

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

Competing interests

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

626 **Legends to figures**

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

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

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

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

643 **Figure 3. Validation and characterization of top preclinical biomarkers**

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

658 **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**

659
660 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-
661 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
662 abbreviations. *Mean age at visit 1.

663 **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[44]
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[45]
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[46]
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[47]
BECN1	1.26 (1.13-1.40)	2.92E-05	2.47E-02			Miki et al., 2018[48]
$HEPACAM2_{cap}$	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[33]
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[30]
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[31]
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[32]

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. 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.

677 **METHODS**

678 **Discovery cohort and study design**

679 The EPIC cohort is a large multi-center prospective study that recruited 519,978 participants
680 (70.5% women, mostly aged 35-70 years at baseline) from 23 centers across ten European
681 countries. With approximately 30 years of follow-up, EPIC was originally established to
682 explore the relationships between diet and cancer, but it has since expanded to study other
683 chronic conditions, including neurodegenerative diseases[12,13]. Briefly, baseline data including
684 information on diet, physical activity, alcohol and tobacco use, past and current illnesses,
685 reproductive history, and anthropometric measurements were collected between 1991 and
686 2000. Citrate plasma samples from ~75% of all participants were obtained using
687 standardized protocols, then aliquoted and stored long-term in liquid nitrogen (–196 °C) at a
688 central biobank housed at the International Agency for Research on Cancer (IARC)/World
689 Health Organization in Lyon, France. Since baseline recruitment, participants have been
690 followed up regularly by data linkage and/or questionnaires at each center. The EPIC study
691 was approved by the ethical committee of IARC and by the ethical review boards of each
692 participating study center. All participants signed a written informed consent.

693 The 'EPIC for Parkinson's disease' case-cohort (EPIC4PD) analyzed in this study is embedded
694 within EPIC and aims to prospectively identify molecular biomarkers in PD[11]. The case
695 ascertainment protocol has been described previously[11,49]. It was based on a source
696 population of 157,288 subjects from ten EPIC centers across five European countries
697 including Spain (Murcia, Navarra, and San Sebastian), Italy (Florence, Turin, and Varese),
698 Netherlands (Utrecht and Bilthoven), UK (Norfolk/Cambridge), and Germany (Heidelberg).
699 Potential cases were ascertained based on data linkage through a combination of center-
700 specific methods using hospital records, general practitioner records, mortality records, drug
701 repositories, and questionnaire data followed by re-review of medical records by expert
702 neurologists for PD. Since the initial ascertainment, 163 incident PD cases were added
703 following an update of the case ascertainment in selected centers (Heidelberg, Florence,
704 Varese, Turin, Utrecht, Bilthoven)[11]. The subcohort for EPIC4PD was randomly drawn from
705 these centers from EPIC-interact, which is a random subcohort within EPIC[50]. The data freeze
706 for the current project was set to the time point of diagnosis of the last patient per center.
707 For the purpose of identifying predictive PD biomarkers, we applied a washout phase after

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

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

711 Protein abundances were measured in citrate plasma samples using the SomaScan 7K (v4.1)
712 platform (SomaLogic, Inc.). SomaScan is a high-throughput proteomic platform that uses
713 modified nucleotides (Slow Off-rate Modified Aptamers — SOMAmers) which specifically
714 bind to protein targets and quantifies them in relative fluorescence units (RFUs) using DNA
715 microarray[51]. The 7K version quantifies 7,596 aptamers targeting 6,432 unique proteins[52].
716 Measurements of overall 17,841 EPIC participants were performed at SomaLogic (Boulder,
717 CO, USA) as previously described. Participants included also individuals who were selected
718 for other endpoints (such as cancer). All samples were processed in a single procedure.
719 SomaScan assays were performed on 96-well plates, with 11 control wells per plate used to
720 monitor batch effects, accuracy, precision, and background. Each plate included five pooled
721 calibrator replicates, three quality control (QC) replicates, and three buffer replicates. In
722 addition, 233 replicates of one EPIC QC sample were included. Readout was conducted using
723 Agilent hybridization and scanning technology. To control readout variability, twelve
724 hybridization control aptamers were added during elution. Initial QC was conducted by
725 SomaLogic and included normalization of protein RFUs based on the following steps:
726 hybridization normalization, intraplate median normalization, plate scaling and calibration,
727 and adaptive normalization to a population reference.

728 We computed intraclass correlation coefficients (ICCs) for all 7,596 aptamers based on the
729 233 EPIC replicate measurements. Aptamers were removed if they were not a protein (n=46)
730 [Hybridization Control Elution (n=12), non-biotin (n=10), non-cleavable (n=4), spuriomer
731 (n=20)] or non-human (n=261). Aptamers were further excluded if both (i) the ICC was below
732 0.5 and (ii) if the aptamer levels were below the limit of detection (LOD) in more than 90% of
733 the samples (n=4). If the measured value of an aptamer was below the LOD, it was set to
734 LOD/2. The final number of quality-controlled aptamers was 7,285 (targeting 6,381 unique
735 proteins). Samples for which any SomaLogic normalization scale factor was outside [0.4-2.5]
736 (n=247) and those detected as 0s using an approach based on PCA and the local outlier
737 factor statistic (https://privefl.github.io/blog/detecting-outlier-samples-in-pca/) were
738 excluded (n=0 for the method excluding outliers and n=5 for capping outliers). Plate

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

758 **Statistical analysis in EPIC4PD**

759 The final effective EPIC4PD case-cohort included 4,538 EPIC4PD participants including 574
760 incident PD cases with 7,285 aptamers (corresponding to 6,381 proteins) available for
761 statistical analysis. We used Cox proportional hazard regression analyses to examine the
762 association of 7,285 aptamers with instantaneous risk of PD using age as time scale and
763 Prentice weights to take into account the case-cohort setting [53]. Our basic model was
764 stratified (i.e., separate baseline hazard functions estimated) for 5-year age intervals, sex,
765 and center. Cox proportional hazards regression was performed using three distinct time
766 windows relative to disease onset: the full follow-up time period (2-28 years from baseline
767 to diagnosis/censoring), as well as two shorter periods therein: 2-10 and 10-20 years.
768 Furthermore, we performed sex-specific analyses after stratification for men and women

769 using the full follow-up time period. Differences between sex-stratified effect estimates
770 were tested by interaction analyses as previously described[54,55].
771 Proteomic results were defined as significant at a FDR of 0.05. Furthermore, we defined
772 "suggestively significant biomarker signals" as those findings that did not withstand FDR
773 control in the proteomic analyses but had a p-value $<10^{-3}$ and demonstrated consistency
774 across all EPIC4PD analyses — specifically, at least two nominally significant ($p<0.05$) results
775 and all effect estimates pointing in the same direction of effect.
776 In sensitivity analyses for all FDR-significant results, we tested a 'full' adjustment model
777 including the baseline covariates school education (no education/primary school vs higher
778 education), smoking status (dummy-coded as 'never', 'former', 'current'), body mass index
779 (BMI; weight/height^2 [kg/m^2]), and physical activity (dummy-coded as 'inactive', 'moderately
780 inactive', 'moderately active/active'). Furthermore, for FDR-significant results derived from
781 analyses stratified for women, we adjusted for postmenopausal status (dummy-coded as
782 premenopausal, postmenopausal, perimenopausal, and surgical postmenopausal status) and
783 use of hormones for menopause (yes/no).
784 HR trajectories were computed using 5-year sliding windows in one-year steps. For the
785 analyses of the sliding windows close to age at diagnosis, the 23 participants diagnosed with
786 PD within two years of baseline and previously excluded from the main analyses were re-
787 included. Due to more limited numbers of cases toward the end of follow-up, the final
788 window covered diagnoses occurring 16 to 25 years after baseline. Curves were fitted using
789 the geom_smooth() function from the ggplot2 package in R.

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

791 Replication analyses of FDR-significant predictive biomarkers were performed on 64,856
792 individuals (1,034 incident PD cases) across three different prospective cohorts (AGES[56],
793 ARIC[57], UKB[58]) based on Olink (UKB) and SomaScan (AGES, ARIC) data, respectively.
794 The first replication cohort, the Age Gene/Environment Susceptibility-Reykjavik (AGES-
795 Reykjavik) Study[59], is a population-based study of adults aged 65 years and older in Iceland,
796 designed to investigate genetic susceptibility and gene-environment interactions related to
797 age-associated diseases. Between 2002 and 2006, 5,764 participants, which were survivors
798 of the original Reykjavik Study, underwent comprehensive interviews, examinations, and
799 measurements. The study was approved by the Icelandic National Bioethics Committee

800 (VSN: 00–063), the Icelandic Data Protection Authority, and the NIA IRB; all participants
801 provided written informed consent. Proteomic profiling of baseline serum samples from the
802 AGES cohort has been described previously[60]. After quality control and exclusion of
803 prevalent PD cases, 5,308 participants had available proteomic data, further reduced to
804 5,090 when applying a 2-year washout period. Over the follow-up, 121 participants
805 developed incident PD. In AGES, aptamer levels were log10-transformed. Further details on
806 recruitment can be found in ref.[61], and prevalent PD case ascertainment in ref.[62]. Medical
807 and mortality records (ICD10: G20) up to 2019-03-19 were reviewed to ascertain incident PD
808 cases.
809 The 2nd replication cohort, ARIC[57], is an ongoing community-based cohort study, for which
810 participants were enrolled from four communities across the United States: Washington
811 County, MD; Forsyth County, NC; northwestern suburbs of Minneapolis, MN; and Jackson,
812 MS. The ARIC participants who had SomaScan 5K data available (4,953 aptamers passing
813 quality control) comprised 11,205 middle-aged participants at baseline of whom 164
814 developed incident PD applying a washout period of two years and had SomaScan data in
815 EDTA plasma samples available. For details on the cohort recruitment and ascertainment of
816 incident PD cases therein please see ref.[63] and the **Supplementary Material**. All participants
817 provided appropriate written consent.
818 The 3rd replication cohort, UKB[58], is a general population cohort study, which performed
819 proteomic profiling of blood plasma samples using the Olink Explore 3072 platform (n=2,923
820 proteins, cohort size analyzed in this study: n=48,561 UKB participants including 749 incident
821 PD cases). The plasma samples were collected in volunteers aged 40 to 69 years at their first
822 attendance to a UKB assessment center in 22 centres across the United Kingdom between
823 2006 to 2010. Follow-up began from the date of attendance to an assessment center (field
824 53) to the first recorded date of diagnosis (ICD-10 code G20) in the patient-linked hospital
825 episode statistics (fields 41270 and 41280) or death registration (fields 40001 and 40002),
826 whichever occurred first, with the last recorded date in January 2023. Due to small case
827 numbers in other strata, the analysis was restricted to white participants (field 21000).
828 Participants with missing proteomics data were removed prior to each model.
829 Due to smaller sample sizes in the restricted time windows, we did not analyze top EPIC4PD
830 biomarkers in AGES for the time-window of 2-10 years of follow-up (n cases=26) and for
831 ARIC for the time window of 10-20 years of follow-up (n cases=32). Consistent with the

832 EPIC4PD analyses, outliers in AGES and UKB were either capped at or excluded beyond 5 SD,
833 depending on the approach used for the specific biomarker to be replicated. In AGES and
834 UKB, outliers were either capped at or excluded beyond 5 SD, in line with EPIC4PD analyses.
835 Outliers in ARIC were capped at 5 SD. Cox proportional hazards regression models were
836 harmonized for AGES, ARIC and UKB with EPIC4PD: The Cox model was based on age as time
837 scale and stratified (separate baseline hazard functions estimated) for age in 5-year
838 categories, race (in ARIC), and sex. Significance for the replication arm was defined at
839 α=0.05. As sensitivity analysis, we also performed the same analyses only on white
840 participants in ARIC (excluding 2,625 black participants). However, as results did not change
841 substantially, we only report results on all participants in this study.

842 **Indirect replication and characterization of preclinical EPIC4PD biomarkers for clinical**
843 **phenotypes in GNPC**

844 The FDR-significant preclinical biomarkers identified in EPIC4PD across the three follow-up
845 periods and after stratification for sex were assessed for an association with prevalent PD
846 leveraging a large case-control dataset of the GNPC (335 prevalent cases and 2,257 controls
847 from 3 contributors) by re-analyzing the data described in the companion manuscript of
848 Imam et al.[14] All individuals included had undergone SomaScan 7K proteomic profiling as
849 part of the GNPC initiative. For this analysis, we restricted inclusion to datasets that
850 contained both PD cases and controls. Control participants were those without a
851 documented clinical diagnosis of mild cognitive impairment, dementia, AD, frontotemporal
852 dementia, PD, or ALS. To minimize bias from population stratification and ensure
853 comparability with EPIC4PD, we only included white participants in datasets where the
854 variable "race" was available. If no race information was provided, all participants were
855 included. The protein data was log10-transformed and standardized and values above or
856 below five standard deviations from the mean were excluded. Then, the case-control
857 datasets were residualized using the removeBatchEffect function from the R package limma
858 to regress out the effects of age, sex, and contributor, thereby controlling for confounding
859 sources of variation in the high-dimensional data. A principal component (PC) analysis was
860 then performed on the residuals; based on the scree plot, the first two PCs were retained as
861 additional covariates. Finally, logistic regression analyses were conducted on the original
862 data, adjusting for age, sex, dummy-coded contributor, and the first two PCs (all results from

863 the re-analyses with $p<0.05$ can be found in **Supplementary Table 22**). Analogous to the
864 EPIC4PD analyses, preclinical biomarkers that were FDR-significant in one sex stratum were
865 analyzed for the GNPC data within that specific sex group.
866 Concordance of effect estimates in EPIC4PD and GNPC (for all 47 significant preclinical
867 biomarkers as well as for nominally significant biomarkers shared by both studies) we
868 applied a one-sided binomial exact test. All analyses were performed in R.
869 We analyzed serum proteomic data from the Tracking Parkinson's disease cohort using the
870 SomaScan v4.1 (7K) assay (SomaLogic). Our cohort included 794 PD patients with serum
871 samples and clinical progression data across the three study visits (at 0 months, 18 months,
872 36 months). To correct for batch effects introduced by collection site, we used the ComBat
873 function from the sva R package (v3.50.0), with site as the batch variable, followed by $\log_{10}$-
874 transformation. For association analyses with PD progression, we investigated associations
875 between protein levels and two clinical outcomes: the Unified Parkinson's Disease Rating
876 Scale Part III (UPDRS-III) as a measure of motor function, and the Montreal Cognitive
877 Assessment (MoCA) as a measure of cognitive function.
878 Cross-sectional associations were assessed using generalized linear models, adjusting for
879 age, sex, and Levodopa Equivalent Daily Dose (LEDD) as a measure of treatment exposure.
880 For longitudinal analyses across up to to three visits per participant, we applied linear mixed-
881 effects models using the lmer() function from the lme4 R package. These models included a
882 random intercept for participant ID to account for repeated measures, and fixed effects for
883 protein level ($\log_{10}$ RFU), visit, age, sex, and LEDD. In additional sensitivity analyses, we
884 included BMI and disease duration (time since diagnosis at baseline) as covariates. Protein
885 associations were evaluated separately for each outcome, and p-values were adjusted for
886 multiple testing using an FDR=0.05.

887 **Comparison to previous studies**

888 We compiled a list of 291 unique candidate biomarkers previously linked to PD based on
889 findings from studies that employed a range of different methodologies (several proteins
890 were nominated more than once): These included i) 38 proteins from the only previous
891 proteomic study on future PD performed on 2,920 proteins based on Olink measurements in
892 the UKB[10], ii) 10 of ~2,700 analyzed blood plasma proteins (7 derived from Olink and 3 from
893 SomaScan studies) recently proposed as potential predictive biomarkers for PD by a

894 comprehensive MR study[16], iii) 18 proteins nominated by a recent MR and colocalization
895 study[64], iv) 109 proteins corresponding to nominated functional candidate and/or the
896 nearest genes located in 78 genetic risk loci described in the most recent Caucasian genome-
897 wide association study for PD [17], and v) 127 preselected neurodegenerative disease
898 candidate biomarkers included in the 'CNS NULISA panel' (Alamar, Inc.). Proteins available in
899 our SomaScan 7K data from this list (n=180) were tested for association in EPIC4PD. Results
900 were controlled applying an FDR=0.05 across all 180 candidate proteins available for testing
901 in our SomaScan 7K data.

902 **Cross-disease comparisons**

903 For cross-disease comparisons between PD and AD as well as between PD and ALS, we
904 performed Cox proportional hazard regression analyses for all FDR-significant PD proteins in
905 EPIC4AD (1,926 EPIC participants including 652 incident AD cases during a follow-up of 2-29
906 years) and EPIC4ALS (4,715 EPIC participants including 187 incident ALS cases during a
907 follow-up of 2-30 years) using Prentice-weighted Cox proportional hazard regression analysis
908 stratified for 5-year age groups, sex, and center, as described above. Results were controlled
909 at an FDR=0.05 in the context of all tested top biomarkers per disease. Furthermore, we
910 assessed whether significant associations in incident AD and ALS were enriched among all
911 proteins that were at least nominally significant in PD (based on the full follow-up period,
912 outliers excluded), using a one-sided Fisher's exact test. To assess whether overlapping
913 nominally significant proteins exhibited concordant effect directions more often than
914 expected by chance, we performed a one-sided binomial test. Furthermore, we assessed the
915 correlation of effect sizes between traits by calculating Pearson's correlation coefficients
916 based on the beta estimates. To prevent bias in the correlation due to the discontinuity
917 introduced by limiting the analysis to nominally significant results, we aligned effect
918 directions by mirroring the beta coefficients: for proteins with negative betas in PD, we
919 multiplied all corresponding betas in EPIC4PD and GNPC by –1 (generating 'mirrowed beta
920 values').

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

922 Gene expression and protein abundance of top PD biomarkers across tissues and in the brain
923 were assessed based on RNA-Seq and immunohistochemisty (IHC) data obtained from the
924 Human Protein Atlas (HPA v24.0, based on Ensembl version 109, accessed May 6th, 2025 at

925 https://www.proteinatlas.org/about/download) [65]. FDR-significant proteins were probed for
926 tissue expression using bulk RNA sequencing (RNA-Seq) data across 50 human tissues
927 available as gene-level 'Consensus RNA data' (rna_tissue_consensus.tsv.zip). Furthermore,
928 brain expression was assessed based on gene-level bulk RNA-Seq data across 13 brain
929 regions based on samples from the Human Brain Tissue Bank (Semmelweis University,
930 Budapest; rna_brain_region_hpa.tsv.zip) and based on single nucleus RNA-Seq (snRNA-Seq)
931 data across 11 brain regions (rna_single_nuclei_cluster_type.tsv.zip). Protein measurements
932 for the FDR-significant biomarkers across 45 human tissues were obtained from HPA based
933 on IHC using tissue microarrays (normal_ihc_data.tsv.zip). To visualize expression patterns of
934 the top biomarkers, heatmaps were generated using hierarchical clustering, with expression
935 values scaled per gene highlighting tissue-specific variation.

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

937 Gene set enrichment analyses were performed based on the GO classification system using
938 the PANTHER resource (Protein ANalysis THrough Evolutionary Relationships; version 19.0;
939 released 2024-06-20; https://geneontology.org/). A list of all unique proteins that showed a
940 suggestive signal in EPIC4PD and was uniquely mapped to a gene by PANTHER (n=154) was
941 tested for enrichment for 'biological process', 'molecular function', and 'cellular component'
942 GO categories as well as 'Reactome pathways'. The statistical overrepresentation test was
943 conducted using Fisher's exact test with the reference list set to all analyzed SomaScan 7K
944 proteins uniquely mapped by PANTHER (n=6,332). To avoid spurious associations, only GO
945 categories with at least ten proteins present in the background list were considered.
946 Cell enrichment analyses were performed using the the Expression Weighted Cell Type
947 Enrichment (EWCE) method to predict the likely primary disease-associated cells based on
948 our prioritised proteins[66]. This approach generates a probability distribution of gene
949 expression for specific cell types by comparing a target gene list to a background gene set.
950 We applied EWCE to our protein lists and to either single-cell mouse brain expression data[67]
951 or RNA expression data from 23 selected cell types sourced from the Human Protein Atlas
952 (dataset: rna_single_cell_type.tsv.zip)[65]. Each analysis was controlled at an FDR=0.05.
953 CorrelationAnalyzeR was used to investigate the coexpression of all FDR-significant
954 preclinical biomarkers. Coexpression with *SNCA* was specifically examined in 12,712 normal
955 GEO samples from immune cells and 3,198 normal GEO samples from brain. Permutation t

956 tests against random gene lists of the same size were used to determine correlation
957 significance with SNCA.

958 **Protein-protein interaction network analysis**

959 Furthermore, protein-protein interaction (PPI) networks were built using the STRING
960 database (v12.0, accessed May 6th, 2025 at https://string-db.org)[68] based on all FDR-
961 significant EPIC4PD biomarkers and plot using igraph (v2.1.4). Interactions with a minimum
962 interaction score of 0.4 were based on the following sources: 'textmining', 'experiments',
963 'databases', 'co-expression", "neighborhood', 'gene fusion', and 'co-occurrence'.
964 Furthermore, to examine the list of suggestive preclinical proteins in relation to known PD
965 genes, we utilized 109 proteins described in the most recent genome-wide association study
966 for PD (see above)[17]. P values were calculated based on the observed interactions compared
967 to 10,000 random permutations. Nodes not connected to the largest component were
968 removed from the PPI network Figure 3.

969 **Drug Repurposing**

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

980 **Prediction**

981 To assess the proteins' performance in risk stratification for PD, we performed variable
982 selection in a Cox proportional hazards model using the Least Absolute Shrinkage and
983 Selection Operator (LASSO) for variable selection and regularization. LASSO was
984 implemented via the R package glmnet, using function cv.glmnet(family="cox",
985 type.measure="C"). As the univariate Cox models had suggested substantial differences

986 between sex strata, we developed the PPRS for each sex stratum separately. and center by
987 stratified Cox models, using the StratifySurv() function[69]. For robust results we employed
988 stability selection[70]. Specifically, individually for each stratum, we fitted LASSO penalized Cox
989 models to 100 resampled datasets, each comprising a 63.2% subsample of individuals.
990 Complete follow-up in each resampling dataset was considered. Missing protein abundance
991 was imputed using k-nearest neighbor imputation (package: impute, function impute.knn()).
992 Imputation was performed within each training subset to avoid data leakage. In each
993 subsample, the optimal amount of regularization (lambda) was inferred through 5-fold cross-
994 validation using the C-index as performance criterion. The model coefficients (weights) from
995 all 100 runs were retained for downstream analysis. We computed the inclusion frequency
996 and average coefficient for each protein across the 100 permutations. To reduce the risk of
997 overfitting, we evaluated sparse protein subsets, using the 5, 10, 20, 50 and 100 most
998 frequently selected proteins from the initial LASSO selection (**Supplementary Table 20**). For
999 each EPIC4PD participant, a multivariate risk score (PPRS) was computed as the sum of the
1000 protein values multiplied by the respective average weights across permutations. Cox
1001 proportional hazards regression models were then fitted for each sex stratum with the sex-
1002 specific risk score as the predictor, stratified by center and adjusted for age at examination.
1003 In order to quantify the incremental benefit of the proteins, the model was also calculated
1004 without the risk score as a predictor. Model performance was quantified using the
1005 concordance index (C-index) obtained from summary(model)$concordance. Accuracy was
1006 determined in AGES as an external validation cohort (n=5,090, including 121 incident PD
1007 cases). For comparison purposes, PPRS were also explored without stratifying by sex. In this
1008 case, the basic model contained age at examination and sex.

1009 ## References

1010 1. Poewe, W. *et al.* Parkinson disease. *Nat Rev Dis Primers* **3**, 1–21 (2017).

1011 2. Ray Dorsey, E. *et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis
1012 for the Global Burden of Disease Study 2016. *Lancet Neurol* **17**, 939–953 (2018).

1013 3. Ishihara, L. S., Cheesbrough, A., Brayne, C. & Schrag, A. Estimated life expectancy of Parkinson's patients compared
1014 with the UK population. *J Neurol Neurosurg Psychiatry* **78**, 1304–1309 (2007).

1015 4. Postuma, R. B. & Berg, D. Advances in markers of prodromal Parkinson disease. *Nature Reviews Neurology* vol. 12
1016 622–634 Preprint at https://doi.org/10.1038/nrneurol.2016.152 (2016).

1017 5. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat Med* **30**,
1018 2489–2498 (2024).

1019 6. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* (2024)
1020 doi:10.1016/j.cell.2024.10.045.

1021 7. Helgason, H. *et al.* Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–
1022 735 (2023).

1023 8. Zetterberg, H. Clinically relevant factors for blood-based Alzheimer biomarker interpretation. *Alzheimer's &*
1024 *Dementia* **19**, (2023).

1025 9. Gail, M. H. *et al.* Design choices for observational studies of the effect of exposure on disease incidence. *BMJ Open*
1026 **9**, (2019).

1027 10. Gan, Y.-H. *et al.* Large-scale proteomic analyses of incident Parkinson's disease reveal new pathophysiological
1028 insights and potential biomarkers. *Nat Aging* **5**, 642–657 (2025).

1029 11. Lill, C. M. *et al.* EPIC4ND: European Prospective Investigation into Cancer and Nutrition follow-up for
1030 neurodegenerative diseases. Preprint at https://doi.org/10.1101/2025.01.29.25321340 (2025).

1031 12. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer
1032 and Nutrition. *Int J Epidemiol* **26**, 6S – 14 (1997).

1033 13. Riboli, E. *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data
1034 collection. *Public Health Nutr* **5**, 1113–1124 (2002).

1035 14. Imam, F. *et al. The Global Neurodegeneration Proteomics Consortium: Biomarker and Drug Target Discovery Across*
1036 *~35,000 Biosamples Analyses for AD, PD, ALS, FTD, and Aging. [#NMED-A141443]. Niklas Mattsson-Carlgren* vol.
1037 15.

1038 15. Doostparast Torshizi, A. *et al.* Proteogenomic network analysis reveals dysregulated mechanisms and potential
1039 mediators in Parkinson's disease. *Nat Commun* **15**, 6430 (2024).

1040 16. Belbasis, L., Morris, S., van Duijn, C., Bennett, D. & Walters, R. Mendelian randomization identifies proteins
1041 involved in neurodegenerative diseases. *Brain* awaf018 (2025) doi:10.1093/brain/awaf018.

1042 17. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-
1043 analysis of genome-wide association studies. *Lancet Neurol* **18**, 1091–1102 (2019).

1044 18. Doostparast Torshizi, A. *et al.* Proteogenomic network analysis reveals dysregulated mechanisms and potential
1045 mediators in Parkinson's disease. *Nat Commun* **15**, 6430 (2024).

1046 19. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic*
1047 *Acids Res* **53**, D1467–D1475 (2025).

1048 20. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity
1049 extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids*
1050 *Res* **39**, (2011).

1051 21. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling.
1052 *Nat Commun* **12**, (2021).

1053 22. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature*
1054 **622**, 348–358 (2023).

1055 23. Eldjarn, G. H. *et al.* Correction to: Large-scale plasma proteomics comparisons through genetics and disease
1056 associations (Nature, (2023), 622, 7982, (348-358), 10.1038/s41586-023-06563-x). *Nature* vol. 630 E3 Preprint at
1057 https://doi.org/10.1038/s41586-024-07549-z (2024).

1058 24. Rooney, M. R. *et al.* Correlations Within and Between Highly Multiplexed Proteomic Assays of Human Plasma. *Clin*
1059 *Chem* **71**, 677–687 (2025).

1060 25. Wang, B. *et al.* Comparative studies of 2168 plasma proteins measured by two affinity-based platforms in 4000
1061 Chinese adults. *Nat Commun* **16**, 1869 (2025).

1062 26. Lill, C. M. *et al.* (Dys)regulation of the Immune System in Parkinson's Disease: Methodologies, Techniques, and Key
1063 Findings from Human Studies. *Aging Dis* (2025) doi:10.14336/AD.2024.1163.

1064 27. Tansey, M. G. *et al.* Inflammation and immune dysfunction in Parkinson disease. *Nat Rev Immunol* **22**, 657–673
1065 (2022).

1066 28. Kedmi, M., Bar-Shira, A., Gurevich, T., Giladi, N. & Orr-Urtreger, A. Decreased expression of B cell related genes in
1067 leukocytes of women with Parkinson's disease. *Mol Neurodegener* **6**, 66 (2011).

1068 29. Tatenhorst, L. *et al.* Glypican-4 serum levels are associated with cognitive dysfunction and vascular risk factors in
1069 Parkinson's disease. *Sci Rep* **14**, 5005 (2024).

1070 30. Fan, Y. *et al.* Increased plasma lipocalin-2 levels are associated with nonmotor symptoms and neuroimaging
1071 features in patients with Parkinson's disease. *J Neurosci Res* **102**, e25303 (2024).

1072 31. Elango, R. *et al.* Potential Biomarkers for Parkinson Disease from Functional Enrichment and Bioinformatic Analysis
1073 of Global Gene Expression Patterns of Blood and Substantia Nigra Tissues. *Bioinform Biol Insights* **17**,
1074 11779322231166214 (2023).

1075 32. Hepp, D. H. *et al.* Inflammatory Blood Biomarkers Are Associated with Long-Term Clinical Disease Severity in
1076 Parkinson's Disease. *Int J Mol Sci* **24**, (2023).

1077 33. Yao, L. *et al.* Bioinformatic Analysis of Genetic Factors from Human Blood Samples and Postmortem Brains in
1078 Parkinson's Disease. *Oxid Med Cell Longev* **2022**, 9235358 (2022).

1079 34. Irmady, K. *et al.* Blood transcriptomic signatures associated with molecular changes in the brain and clinical
1080 outcomes in Parkinson's disease. *Nat Commun* **14**, 3956 (2023).

1081 35. Miki, Y. *et al.* Alteration of autophagy-related proteins in peripheral blood mononuclear cells of patients with
1082 Parkinson's disease. *Neurobiol Aging* **63**, 33–43 (2018).

1083 36. Santiago, J. A., Littlefield, A. M. & Potashkin, J. A. Integrative transcriptomic meta-analysis of Parkinson's disease
1084 and depression identifies NAMPT as a potential blood biomarker for de novo Parkinson's disease. *Sci Rep* **6**, 34579
1085 (2016).

1086 37. Kim, B.-W. *et al.* Pathogenic Upregulation of Glial Lipocalin-2 in the Parkinsonian Dopaminergic System. *J Neurosci*
1087 **36**, 5608–22 (2016).

1088 38. Hastings, N. *et al.* Connexin 43 is downregulated in advanced Parkinson's disease in multiple brain regions which
1089 correlates with symptoms. *Sci Rep* **15**, 10250 (2025).

1090 39. Lorenzl, S., Albers, D. S., Narr, S., Chirichigno, J. & Beal, M. F. Expression of MMP-2, MMP-9, and MMP-1 and their
1091 endogenous counterregulators TIMP-1 and TIMP-2 in postmortem brain tissue of Parkinson's disease. *Exp Neurol*
1092 **178**, 13–20 (2002).

1093 40. Morris, H. R., Spillantini, M. G., Sue, C. M. & Williams-Gray, C. H. The pathogenesis of Parkinson's disease. *Lancet*
1094 **403**, 293–304 (2024).

1095 41. Heinzel, S. *et al.* Update of the MDS research criteria for prodromal Parkinson's disease. *Movement Disorders* vol.
1096 34 1464–1470 Preprint at https://doi.org/10.1002/mds.27802 (2019).

1097 42. Bestwick, J. P. *et al.* Improving estimation of Parkinson's disease risk—the enhanced PREDICT-PD algorithm. *NPJ*
1098 *Parkinsons Dis* **7**, (2021).

1099 43. Marini, K. *et al.* Comparison of different risk scores for Parkinson disease in a population-based 10-year study. *Eur J*
1100 *Neurol* **30**, 3347–3352 (2023).

1101 44. Irmady, K. *et al.* Blood transcriptomic signatures associated with molecular changes in the brain and clinical
1102 outcomes in Parkinson's disease. *Nat Commun* **14**, 3956 (2023).

1103 45. Santiago, J. A., Littlefield, A. M. & Potashkin, J. A. Integrative transcriptomic meta-analysis of Parkinson's disease
1104 and depression identifies NAMPT as a potential blood biomarker for de novo Parkinson's disease. *Sci Rep* **6**, 34579
1105 (2016).

1106 46. Kedmi, M., Bar-Shira, A., Gurevich, T., Giladi, N. & Orr-Urtreger, A. Decreased expression of B cell related genes in
1107 leukocytes of women with Parkinson's disease. *Mol Neurodegener* **6**, 66 (2011).

1108 47. Tatenhorst, L. *et al.* Glypican-4 serum levels are associated with cognitive dysfunction and vascular risk factors in
1109 Parkinson's disease. *Sci Rep* **14**, 5005 (2024).

1110 48. Miki, Y. *et al.* Alteration of autophagy-related proteins in peripheral blood mononuclear cells of patients with
1111 Parkinson's disease. *Neurobiol Aging* **63**, 33–43 (2018).

1112 49. Gallo, V. *et al.* Parkinson's Disease Case Ascertainment in the EPIC Cohort: The NeuroEPIC4PD Study. *Neurodegener*
1113 *Dis* **15**, 331–338 (2015).

1114 50. Consortium, T. I. Design and cohort description of the InterAct Project: an examination of the interaction of genetic
1115 and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* **54**, 2272–2282 (2011).

1116 51. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, (2010).

1117 52. Candia, J., Daya, G. N., Tanaka, T., Ferrucci, L. & Walker, K. A. Assessment of variability in the plasma 7k SomaScan
1118 proteomics assay. *Sci Rep* **12**, (2022).

1119 53. PRENTICE, R. L. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**,
1120 1–11 (1986).

1121 54. Altman, D. G. & Bland, J. M. Interaction revisited: the difference between two estimates. *BMJ* **326**, 219 (2003).

1122 55. Lill, C. M. *et al.* Impact of Parkinson's disease risk loci on age at onset. *Movement Disorders* **30**, 847–850 (2015).

1123 56. Harris, T. B. *et al.* Age, Gene/Environment Susceptibility-Reykjavik Study: Multidisciplinary Applied Phenomics. *Am*
1124 *J Epidemiol* **165**, 1076–1087 (2007).

1125 57. Wright, J. D. *et al.* The ARIC (Atherosclerosis Risk In Communities) Study. *J Am Coll Cardiol* **77**, 2939–2959 (2021).

1126 58. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex
1127 Diseases of Middle and Old Age. *PLoS Med* **12**, e1001779- (2015).

1128 59. Harris, T. B. *et al.* Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. *Am J*
1129 *Epidemiol* **165**, 1076–87 (2007).

1130 60. Bankier, S. *et al.* Circulating causal protein networks linked to future risk of myocardial infarction. *medRxiv* (2025)
1131 doi:10.1101/2025.02.07.25321789.

1132 61. Harris, T. B. *et al.* Age, Gene/Environment Susceptibility-Reykjavik Study: Multidisciplinary Applied Phenomics. *Am*
1133 *J Epidemiol* **165**, 1076–1087 (2007).

1134 62. Scher, A. I. *et al.* Midlife migraine and late-life parkinsonism: AGES-Reykjavik study. *Neurology* **83**, 1246–52 (2014).

1135 63. Huang, X. *et al.* Statins, plasma cholesterol, and risk of Parkinson's disease: a prospective study. *Mov Disord* **30**,
1136 552–9 (2015).

1137 64. Doostparast Torshizi, A. *et al.* Proteogenomic network analysis reveals dysregulated mechanisms and potential
1138 mediators in Parkinson's disease. *Nat Commun* **15**, (2024).

1139 65. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Sci Adv* **7**, eabh2169 (2025).

1140 66. Skene, N. G. & Grant, S. G. N. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell
1141 Transcriptomes and Expression Weighted Cell Type Enrichment. *Front Neurosci* **Volume 10-2016**, (2016).

1142 67. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (1979)*
1143 **347**, 1138–1142 (2015).

1144 68. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment
1145 analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).

1146 69. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via
1147 Coordinate Descent. *J Stat Softw* **39**, (2011).

1148 70. Meinshausen, N. & Bühlmann, P. Stability Selection. *J R Stat Soc Series B Stat Methodol* **72**, 417–473 (2010).

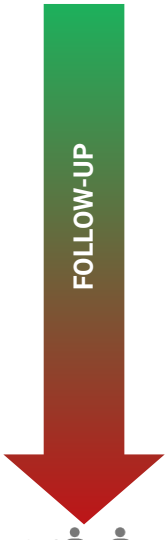
1149

Discovery cohort

EPIC4PD
(n=4,538)

Demographics
Lifestyle
Health data

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

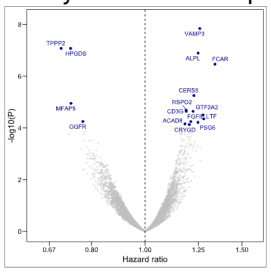


GNPC
(n=335 cases,
2,257 controls)

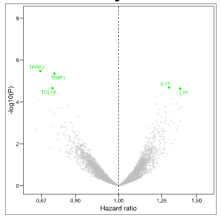
TRACKING PD
(n=794 cases)

Preclinical biomarker discovery

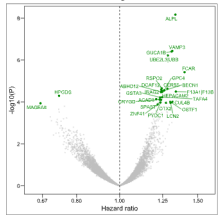
2-28 years of follow-up



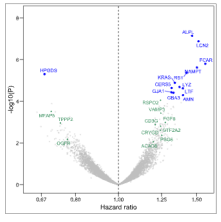
2-10 years



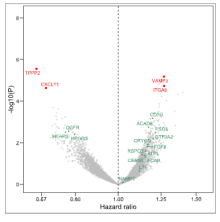
10-20 years



Men

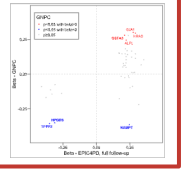


Women



Assessment of preclinical biomarkers in clinical PD

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



Preclinical replication (n=64,856)

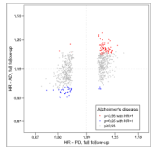
ARIC	AGES
total n=11,205 / n events=164	total n=5,090 / n events=121
UKB total n=48,561 / n events=749	

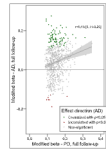
Functional characterization

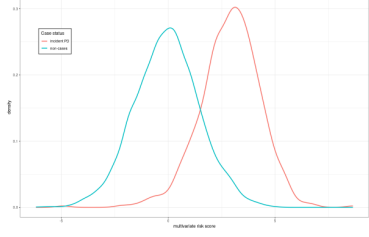
- Gene ontology enrichment
- Cell-type enrichment
- Protein-protein interaction networks
- Drug repurposing
- Literature mining/candidate biomarkers

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