npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01175-9

Identification of Parkinson's disease PACE subtypes and repurposing treatments through integrative analyses of multimodal data

Chang Su[1,2], Yu Hou[3], Jielin Xu[4], Zhenxing Xu[1,2], Manqi Zhou[2,5], Alison Ke[2,5], Haoyang Li[1,2], Jie Xu[6], Matthew Brendel[7], Jacqueline R. M. A. Maasch[2,8], Zilong Bai[1,2], Haotan Zhang[9], Yingying Zhu[10], Molly C. Cincotta[11], Xinghua Shi[12], Claire Henchcliffe[13], James B. Leverenz[14], Jeffrey Cummings[15], Michael S. Okun[16], Jiang Bian[6], Feixiong Cheng[4,17] & Fei Wang[1,2]

Parkinson's disease (PD) is a serious neurodegenerative disorder marked by significant clinical and progression heterogeneity. This study aimed at addressing heterogeneity of PD through integrative analysis of various data modalities. We analyzed clinical progression data (≥5 years) of individuals with de novo PD using machine learning and deep learning, to characterize individuals' phenotypic progression trajectories for PD subtyping. We discovered three pace subtypes of PD exhibiting distinct progression patterns: the Inching Pace subtype (PD-I) with mild baseline severity and mild progression speed; the Moderate Pace subtype (PD-M) with mild baseline severity but advancing at a moderate progression rate; and the Rapid Pace subtype (PD-R) with the most rapid symptom progression rate. We found cerebrospinal fluid P-tau/α-synuclein ratio and atrophy in certain brain regions as potential markers of these subtypes. Analyses of genetic and transcriptomic profiles with network-based approaches identified molecular modules associated with each subtype. For instance, the PD-R-specific module suggested *STAT3*, *FYN*, *BECN1*, *APOA1*, *NEDD4*, and *GATA2* as potential driver genes of PD-R. It also suggested neuroinflammation, oxidative stress, metabolism, PI3K/AKT, and angiogenesis pathways as potential drivers for rapid PD progression (i.e., PD-R). Moreover, we identified repurposable drug candidates by targeting these subtype-specific molecular modules using network-based approach and cell line drug-gene signature data. We further estimated their treatment effects using two large-scale real-world patient databases; the real-world evidence we gained highlighted the potential of metformin in ameliorating PD progression. In conclusion, this work helps better understand clinical and pathophysiological complexity of PD progression and accelerate precision medicine.

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by changes in both motor and non-motor functions and involves degeneration of multiple basal ganglia and cortical related circuits. PD is the second most prevalent neurodegenerative disorder, impacting approximately 2–3% of individuals aged over 65[1,2]. The prevalence of the disease increases with advancing age, and it has emerged as a prominent health concern for the aging population[1,2]. PD is characterized by the loss of dopamine-producing (dopaminergic) neurons in the substantia nigra and the accumulation of α-synuclein aggregates across multiple brain circuits and regions[1,2]. However, the precise etiological and pathological mechanisms underlying PD remain elusive. Consequently, as of now, there are no approved disease-modifying treatments known to slow, prevent, or reverse the progression of PD[3].

A full list of affiliations appears at the end of the paper. e-mail: few2001@med.cornell.edu

In the past decades, there has been increasing recognition that "Parkinson's disease" is not a single entity, but rather multiple sub-disorders classified under the "Parkinson's disease" term with multiple overlapping etiologies[4–6], leading to distinct progression trajectories during the PD course. Past clinical trials have struggled to account for the considerable heterogeneity in symptoms and progression and the pathophysiology underlying this clinical heterogeneity[7]. This clinical and progression heterogeneity may have contributed to PD clinical trial failures[7,8]. Administering the same medication to patients who share a PD clinical diagnosis but possess diverse underlying pathobiological processes, is unlikely to ameliorate disease progression. Moreover, without proper characterization of phenotypic variations, evaluating treatment response in clinical trials for PD is challenging. In this context, segmenting the heterogeneous population of a disease into relatively pathologically and biologically homogeneous subgroups, i.e., so-called subtypes, has shown significant promise for precision medicine and drug development. Recently, PD research has begun to shift its focus towards identifying PD subtypes[8–11]. Previous studies have utilized various approaches for PD subtyping, such as determining subtypes based on PD risk genotypes, categorizing subtypes based on motor (e.g., postural instability and gait difficulty [PIGD]) or non-motor manifestations (e.g., mild cognitive decline subtype), or by employing emerging data-driven methods[10,11]. However, despite the progress made, there is yet no consensus on a single subtyping method that effectively aids in PD therapy[10] and accounts for disease progression[12–14]. One potential reason has been the insufficient appreciation and consideration given to PD progression heterogeneity. Approaches that utilize longitudinal data, beginning from a specific disease duration or disease milestone (e.g., early stage), could offer new insights in addressing PD heterogeneity[15].

Increasing efforts have been dedicated to exploring the complex pathological and biological underpinnings of PD and its progression[1,2,16,17]. These efforts have uncovered crucial mechanisms involved in PD, such as the aggregation of α-synuclein, neuroinflammation, mitochondrial dysfunction, and oxidative stress[16–18]. Additionally, associations have been established between specific genes and the disease manifestations although notably there has been substantial heterogeneity even in monogenetic causes of PD[19–21]. Most data have focused predominantly on the disease state of PD, offering limited insight into the disease's progression. A few recent genetic studies have identified risk loci associated with motor and non-motor progression in PD[22,23]. However, our understanding of the pathophysiological mechanisms that drive the heterogeneous progression of PD remains incomplete. This knowledge gap is an obstacle to discovering effective disease-modifying treatments that can slow, halt, or reverse PD progression. Further studies aimed at better understanding the complex interplay of factors influencing PD progression are needed to achieve the goal of developing efficacious therapeutic strategies.

In this study, our goal was to disentangle the clinical and progression heterogeneity of PD to accelerate precision medicine. To achieve this, we established an integrated data-driven framework that combines machine learning, deep learning, network medicine, and statistical approaches, enabling a multifaceted analysis of diverse data types (see Fig. 1). These included individual-level clinical records, biospecimens, neuroimaging, genetic and transcriptomic information, publicly available protein-protein interactome (PPI) and transcriptomics-based drug-gene signature data[24], as well as patient-level real-world data (RWD). First, we characterized individual's high-dimensional phenotypic progression data to uncover PD subtypes. This led to the identification of three pace subtypes of PD that exhibited distinct phenotypic progression patterns and were stable over time. Next, we identified indicative cerebrospinal fluid (CSF) and neuroimaging markers of the subtypes we discovered. Furthermore, by interrogating genetic and transcriptomic data with network medicine approaches, we identified subtype-specific molecular modules, revealing potential pathophysiological underpinnings driving the subtypes with distinct progression trends. Finally, we predicted potential therapeutic candidates by targeting subtype-specific molecular modules and estimated treatment effects of the candidates using real-world evidence (RWE) via analysis of two large-scale RWD databases. Our data suggested metformin as a potential candidate in mitigating PD progression, as (1) it could counteract molecular alterations triggered in the rapid pace subtype, and (2) it was associated with an improved PD progression based on RWE.

Results

Study cohorts

In this investigation, we adopted data of participants in the Parkinson's Progression Markers Initiative (PPMI) study, an international observational PD study that systematically collected clinical, biospecimen, multi-omics, and brain imaging data of participants[25]. Our analysis included 406 de novo PD participants (PD diagnosis within the last 2 years and untreated at enrollment) in the PPMI cohort, comprising 140 (34.5%) women and 266 (65.5%) men, with an average age of 59.6 ± 10.0 years at PD onset; 188 healthy control (HC) volunteers, comprising 67 (35.6%) women and 121 (64.4%) men; and 61 participants who had dopamine transporter scans without evidence of dopaminergic deficit (SWEDD), comprising 23 (37.7%) and 38 (62.3%) men (see Supplementary Table 1). Specifically, we developed a deep learning model to capture PD phenotypic progression trends using over 5-year longitudinal clinical assessments of the de novo PD, HC, and SWEDD participants. Clustering analysis was conducted based on the learned progression profiles among the de novo PD participants to derive subtypes. We further examined individual's neuroimaging, CSF, genetic, and transcriptomic data to identify subtype-specific biomarkers and molecular modules. To demonstrate robustness of our method in capturing PD progression trends to identify subtypes, we replicated our deep learning model among participants in the Parkinson Disease Biomarkers Program (PDBP)[26]. More details of participants included in this study can be found in the "Methods" and Supplementary Table 1.

Discovery of three pace subtypes of de novo PD

We used five-year longitudinal records of individuals in over 140 items of diverse motor and non-motor assessments (see Supplementary Table 2). We built a deep learning model, termed deep phenotypic progression embedding (DPPE), for holistically modeling such multidimensional, longitudinal progression data of the participants (see Fig. 1g). The DPPE extended our previous work[27], integrating a Long Short-Term Memory (LSTM)[8,13] network with an autoencoder architecture. The LSTM is a specialized deep neural network designed for multivariate time sequence data analysis[28,29]. Specifically, DPPE had two components: (1) an encoder that used a LSTM to receive the longitudinal clinical data of each participant as input, capturing his/her phenotypic progression profile, and map it into a compact embedding vector; and (2) conversely, a decoder that employed another LSTM to unpack this compact embedding vector to reconstruct the individual's raw input data. In this way, the DPPE was trained in an unsupervised manner by minimizing the difference between the input raw data and the reconstructed output. For training DPPE, we leveraged data of de novo PDs, HCs, and SWEDD individuals in the PPMI (see Supplementary Table 1). The trained model can generate a machine-readable embedding vector for each individual, encoding his/her phenotypic progression profile. More details of the DPPE model can be found in Fig. 1g and "Methods" section.

Next, we conducted cluster analysis based on the DPPE-learned embedding vectors of individuals to identify PD subtypes. We used the agglomerative hierarchical clustering (AHC) algorithm with Euclidean distance calculation and Ward linkage criterion, because it has demonstrated to be robust to different types of data distributions[30]. An issue of cluster analysis is how to determine the optimal cluster number. To this end, we considered (1) cluster separation in the dendrogram produced by the AHC, (2) 18 cluster structure measurements using the 'NbClust' software[31], (3) cluster separation in the 2-dimensional (2D) space based on the t-distributed stochastic neighbor embedding (t-SNE) algorithm[32], and (4) clinical interpretations of clusters. Using our method, three distinct subtypes were identified (see Supplementary Fig. 1 and Supplementary Note 1). The subtypes' demographics and clinical characteristics at baseline and follow-

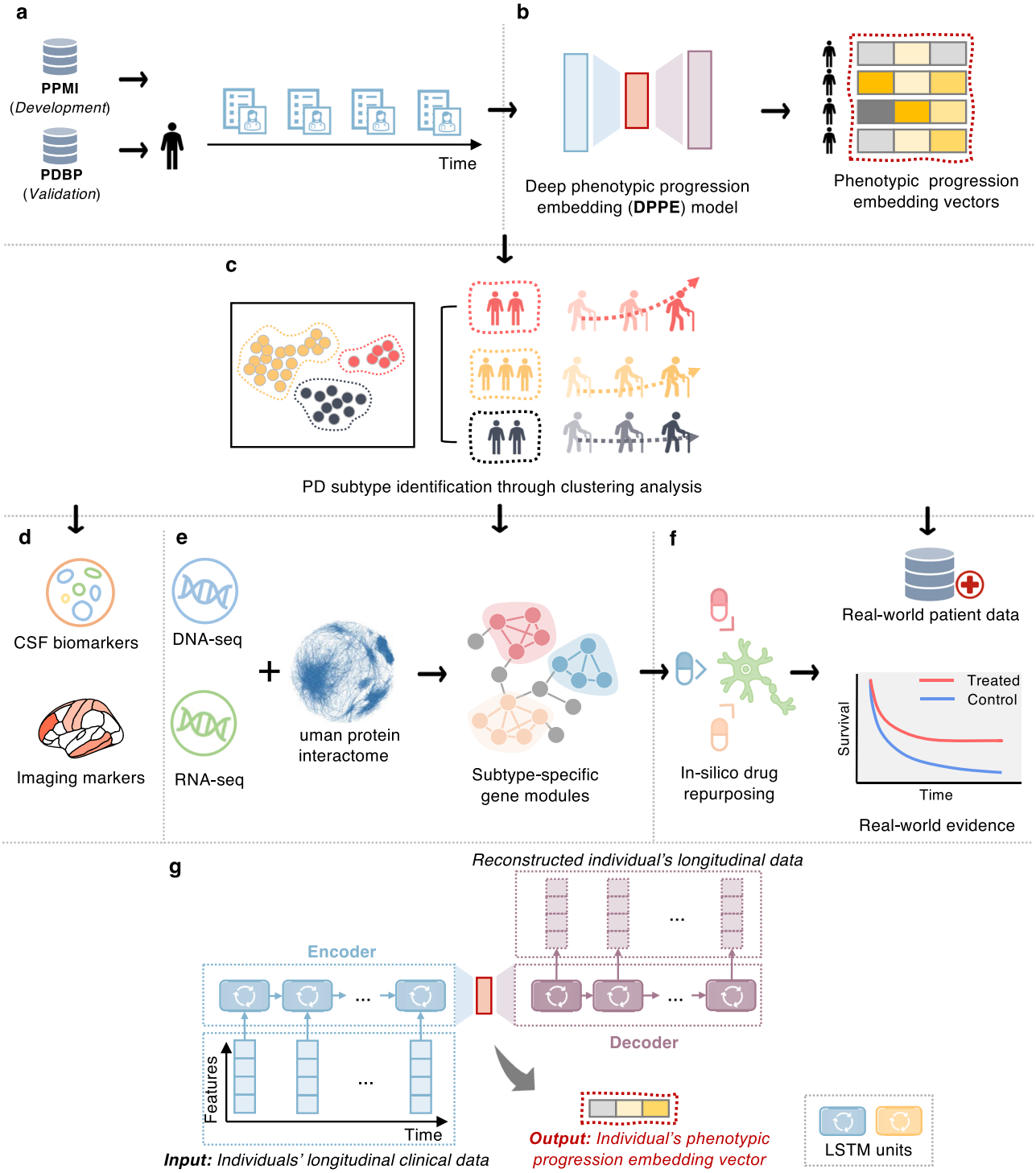

Fig. 1 | A diagram illustrating the present analysis. a Collecting longitudinal clinical data from the Parkinson's Progression Markers Initiative (PPMI) and Parkinson's Disease Biomarkers Program (PDBP) cohorts and conducting necessary data cleaning and preprocessing. **b** Development of a deep phenotypic progression embedding (DPPE) model to learn a progression embedding vector for each individual, which encodes his/her PD symptom progression trajectory. **c** Cluster analysis with the learned embedding vectors to identify PD subtypes, each of which reveal a unique PD progression pattern. **d** Identifying CSF biomarkers and imaging markers the discovered PD subtypes. **e** Construction of PD subtype-specific molecular modules based on genetic and transcriptomic data, along with human protein-protein interactome (PPI) network analyses, using network medicine approaches. **f** In silico drug repurposing based on subtype-specific molecular profiles and validation of drug candidates' treatment efficiency based on analysis of large-scale real-world patient databases, i.e., the INSIGHT and OneFlorida + . **g** Architecture of the DPPE model. Specifically, DPPE engaged two Long-Short Term Memory (LSTM) units—one as encoder receiving an individual's longitudinal clinical records and compacting them into a low-dimensional embedding space; while another taking the individual's embedding vector to reconstruct the original clinical records. DPPE was trained by minimizing the reconstruction difference.

Table 1 | Demographics and baseline clinical characteristics by subtypes within the PPMI cohort

Variables	Subtype PD-I (inching pace)	Subtype PD-M (moderate pace)	Subtype PD-R (rapid pace)	*P* value[a]	Post hoc[b]	*P* value adjusted[c]
# of participants	**145**	**207**	**54**	–	–	–
Age at onset, year, mean (SD)	58.1 (9.9)	59.4 (10.0)	64.4 (8.6)	<0.001**	III vs. rest	–
Sex male, N (%)	67 (46.2)	155 (74.9)	44 (81.5)	<0.001**	–	–
Race white, N (%)	140 (96.6)	196 (94.7)	50 (92.6)	0.422	–	–
Symptom duration, month, mean (SD)	0.593 (0.682)	0.546 (0.742)	0.667 (0.752)	0.526	–	–
Genetic risk score[d], mean (SD)	−0.009 (0.004)	−0.009 (0.004)	−0.01 (0.004)	0.615	–	0.879
Family history, N (%)	140 (96.6)	196 (94.7)	50 (92.6)	0.711	–	–
Education history group, N (%)					–	–
<12 years	10 (6.9)	9 (4.4)	7 (13.0)	0.102	–	–
12–16 years	90 (62.1)	132 (63.8)	26 (48.2)			
>16 years	45 (31.0)	66 (31.9)	21 (38.9)			
Motor manifestations						
MDS-UPDRS Part II, mean (SD)	5.2 (3.7)	5.7 (4.2)	7.9 (4.4)	<0.001**	III vs. rest	<0.001**
MDS-UPDRS Part III, mean (SD)	19.5 (8.8)	20.7 (8.4)	24.8 (9.6)	<0.001**	III vs. rest	0.005*
H&Y Stage, mean (SD)	1.55 (0.51)	1.54 (0.49)	1.70 (0.49)	0.093	–	0.378
Schwab and England score, mean (SD)	93.9 (5.9)	93.2 (5.6)	90.9 (5.7)	0.005*	III vs. rest	0.004*
Tremor score, mean (SD)	0.46 (0.32)	0.50 (0.31)	0.55 (0.34)	0.14	–	0.252
PIGD score, mean (SD)	0.22 (0.25)	0.20 (0.19)	0.34 (0.26)	<0.001***	III vs. rest	<0.001**
Motor phenotype, N (%)						
Tremor	98 (67.6)	154 (74.4)	33 (61.1)	0.235	–	–
Indeterminate	18 (12.4)	22 (10.6)	6 (11.1)			
PIGD	29 (20.0)	31 (15.0)	15 (27.8)			
Non-motor manifestations						
MDS-UPDRS Part I, mean (SD)	5.2 (4.2)	5.4 (3.8)	6.6 (4.4)	0.076	–	0.034
Hallucination, mean (SD)	0.01 (0.12)	0.03 (0.18)	0.04 (0.19)	0.467	–	0.304
Apathy, mean (SD)	0.18 (0.51)	0.21 (0.48)	0.20 (0.49)	0.82	–	0.657
Pain, mean (SD)	0.67 (0.83)	0.72 (0.83)	0.65 (0.84)	0.806	–	0.426
Fatigue, mean (SD)	0.56 (0.71)	0.67 (0.79)	0.74 (0.84)	0.272	–	0.058
Sleep, mean (SD)						
Epworth sleepiness score	4.9 (3.4)	6.0 (3.1)	6.4 (3.8)	0.001**	I vs. rest	0.002*
REM sleep behavior disorder score	3.4 (2.4)	4.2 (2.7)	5.1 (2.9)	<0.001**	1 vs. rest	<0.001**
Sleep phenotype, missing = 1, N (%)						
REM sleep behavior disorder positive	38 (26.2)	84 (40.8)	26 (48.2)	0.003	–	–
REM sleep behavior disorder negative	107 (73.4)	122 (59.2)	28 (51.8)			
QUIP (Impulse control disorders)	0.27 (0.62)	0.31 (0.64)	0.17 (0.46)	0.284	–	0.360
Geriatric depression scale	2.26 (2.79)	2.23 (2.801)	2.87 (2.41)	0.212	–	0.078
Depression phenotype, missing = 1, N (%)						
Normal	125 (86.2)	181 (87.9)	42 (77.8)	0.295	–	–
Mild	9 (6.2)	17 (8.3)	8 (14.8)			
Moderate	9 (6.2)	7 (3.4)	4 (7.4)			
Severe	2 (1.4)	1 (0.5)	0 (0)			
State trait anxiety index, mean (SD)						
State subscore	32.3 (10.8)	33.2 (9.9)	33.2 (9.5)	0.716	–	0.318
Trait subscore	31.8 (10.3)	32.7 (8.8)	32.4 (9.7)	0.713	–	0.107
SCOPA autonomic questionnaire, mean (SD)						
Gastrointestinal (up+down)	1.81 (1.93)	2.1 (1.95)	3.13 (2.35)	<0.001**	III vs. rest	<0.001*
Urinary	3.88 (2.63)	4.21 (2.89)	5.41 (7.45)	0.040	I vs. III	0.198

Table 1 (continued) | Demographics and baseline clinical characteristics by subtypes within the PPMI cohort

Variables	Subtype PD-I (inching pace)	Subtype PD-M (moderate pace)	Subtype PD-R (rapid pace)	*P* value[a]	Post hoc[b]	*P* value adjusted[c]
Cardiovascular	0.39 (0.68)	0.51 (0.86)	0.52 (0.66)	0.309	–	0.316
Thermoregulatory	0.46 (0.87)	0.47 (0.84)	0.24 (0.72)	0.19	–	0.340
Pupillomotor	0.35 (0.58)	0.42 (0.64)	0.50 (0.79)	0.289	–	0.182
Skin	0.69 (0.88)	0.68 (0.91)	0.83 (0.96)	0.53	–	0.243
Sexual	4.47 (6.48)	3.43 (5.68)	4.48 (6.71)	0.233	–	0.629
Total (sum all)	12.04 (8.39)	11.81 (8.32)	15.11 (11.16)	0.044	II vs. III	0.020
Cognitive function, mean (SD)						
MoCA-visuospatial	4.6 (0.8)	4.5 (0.8)	4.4 (0.8)	0.23	–	0.271
MoCA-naming	3.0 (0.2)	2.9 (0.3)	2.9 (0.4)	0.043	I vs. III	0.135
MoCA-attention	5.8 (0.5)	5.8 (0.6)	5.7 (0.7)	0.532	–	0.393
MoCA-language	2.7 (0.6)	2.5 (0.7)	2.6 (0.7)	0.013*	I vs. II	0.011*
MoCA-delayed recall	3.7 (1.3)	3.3 (1.4)	2.9 (1.6)	<0.001**	I vs. rest	0.043
MoCA total score	27.9 (2.0)	27.0 (2.4)	26.7 (2.4)	<0.001**	I vs. rest	0.004*
Benton judgment of line orientation	13.0 (2.1)	12.9 (2.1)	12.1 (2.2)	0.036	III vs. rest	0.020
HVLT-total recall	25.5 (4.6)	24.3 (5.1)	22.0 (4.7)	<0.001***	III vs. rest	0.023
HVLT-delayed recall	8.8 (2.4)	8.3 (2.5)	7.2 (2.5)	<0.001**	III vs. rest	0.016
HVLT-discrimination recognition	10.1 (2.3)	9.7 (2.5)	8.5 (2.9)	<0.001**	III vs. rest	0.572
HVLT-retention	0.9 (0.2)	0.8 (0.2)	0.8 (0.2)	0.072	–	0.073
LNS	11.4 (2.7)	10.3 (2.4)	9.5 (2.8)	<0.001***	I vs. rest	<0.001**
Semantic fluency	52.1 (12.1)	48.3 (10.5)	41.3 (11.4)	<0.001***	All comparisons	<0.001**
Symbol digit test	47.0 (8.6)	44.6 (9.0)	40.7 (9.9)	<0.001***	All comparisons	<0.001**
Cognitive phenotype, missing = 7, N (%)						
Normal	135 (97.8)	200 (96.6)	41 (75.9)	<0.001**	–	–
MCI	2 (1.4)	4 (1.9)	8 (14.8)			
Dementia	1 (0.7)	3 (1.5)	5 (9.3)			

HVLT Hopkins Verbal Learning Test, *LNS* letter-number sequencing, *MCI* mild cognitive impairment, *MDS-UPDRS* Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale, *MoCA* Montreal Cognitive Assessment, *PIGD* postural instability and gait disorder, *PPMI* the Parkinson's Progression Markers Initiative, *SCOPA* Scales for Outcomes in Parkinson's Disease.

[a]*P* values were calculated using ANOVA (for continuous variables) and χ^2 test (for categorical variables) where appropriate.

[b]Post hoc analysis was performed using the Tukey HSD test when the ANOVA *P* value < 0.05.

[c]ANCOVA was used to calculate *p* values (for continuous variables) adjusting for age, sex, and levodopa equivalent daily dose.

[d]Genetic risk scores were calculated based on 90 PD-related loci reported in the latest Genome wide association study[33].

Multiple correction was conducted by controlling false discovery rate (FDR). * FDR adjusted *P* value < 0.05; ** FDR adjusted *P* value < 0.01; *** FDR adjusted *P* value < 0.001.

up were detailed in Table 1 and Supplementary Tables 3 and 4, respectively. Subtype-specific averaged symptom progression trajectories and annual progression rates in terms of each single clinical assessment – as estimated by linear mixed effect models (adjusting for age, sex, and levodopa equivalent dose (LED) usage at visit)—were illustrated in Fig. 2a and summarized in Table 2, respectively. The specific characteristics of each subtype were elaborated upon below.

The Rapid Pace subtype (PD-R), marked by rapid symptom progression, consisted of 54 (13.3%) individuals (see Fig. 2a and Tables 1, 2). Compared to other subtypes, PD-R had more males ($N = 44$ [81.5%]) with the highest average age at PD onset (64.4 ± 8.6 years). Individuals of PD-R experienced more severe motor symptoms at baseline, as indicated by MDS-UPDRS Parts II and III and the Schwab and England scale, as well as more non-motor problems, especially cognitive impairment (see Table 1). Remarkably, PD-R was associated with the most rapid annual progression rates in most motor and non-motor symptoms among the three subtypes. For instance, compared to other subtypes, PD-R exhibited greater annual progression rates in motor assessments including MDS-UPDRS Part II (1.98/year, 95% CI [1.52, 2.44], $P < 0.001$) and Part III (3.08/year, 95% CI [2.12, 4.03], $P < 0.001$), and Schwab and England score (−3.89/year, 95% CI [−4.93, −2.85], $P < 0.001$). PD-R also showed the greatest annual increase rate regarding the overall non-motor function, measured by MDS-UPDRS Part I (1.70/year, 95% CI [1.53, 1.88], $P < 0.001$). Rapid progression rates were also observed in specific non-motor functions in PD-R such as sleep (e.g., Epworth sleepiness score and REM sleep behavior disorder), mood (e.g., State trait anxiety index and Geriatric depression scale), autonomic problem (Scales for Outcomes in Parkinson's Disease autonomic questionnaire), and cognitive performance (e.g., MoCA) (see Fig. 2a and Table 2).

The Inching Pace subtype (PD-I), characterized by mild baseline symptoms and mild symptom progression, encompassed 145 participants (35.7%) (see Fig. 2a and Tables 1 and 2). Compared to others, PD-I had a relatively lower proportion of men (46.2%, $N = 67$) and a younger age at PD onset (58.1 ± 9.9 years). In addition, individuals of PD-I exhibited milder motor and non-motor symptoms at baseline. This was substantiated by their lower scores on the Movement Disorders Society—Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Parts II and III, and their less severe sleep and cognitive impairments (see Table 1). Furthermore, PD-I demonstrated the most gradual PD progression among all subtypes, indicated by the lowest annual progression rates in most PD symptoms estimated by the linear mixed effect models (see Fig. 2a and Table 2). Specifically, the progression rates of overall motor symptoms were low in MDS-UPDRS Part II (0.42/year, 95% CI [0.24, 0.60], $P < 0.001$) and Part III (1.03/year, 95% CI [0.62, 1.43], $P < 0.001$), and Schwab and England score

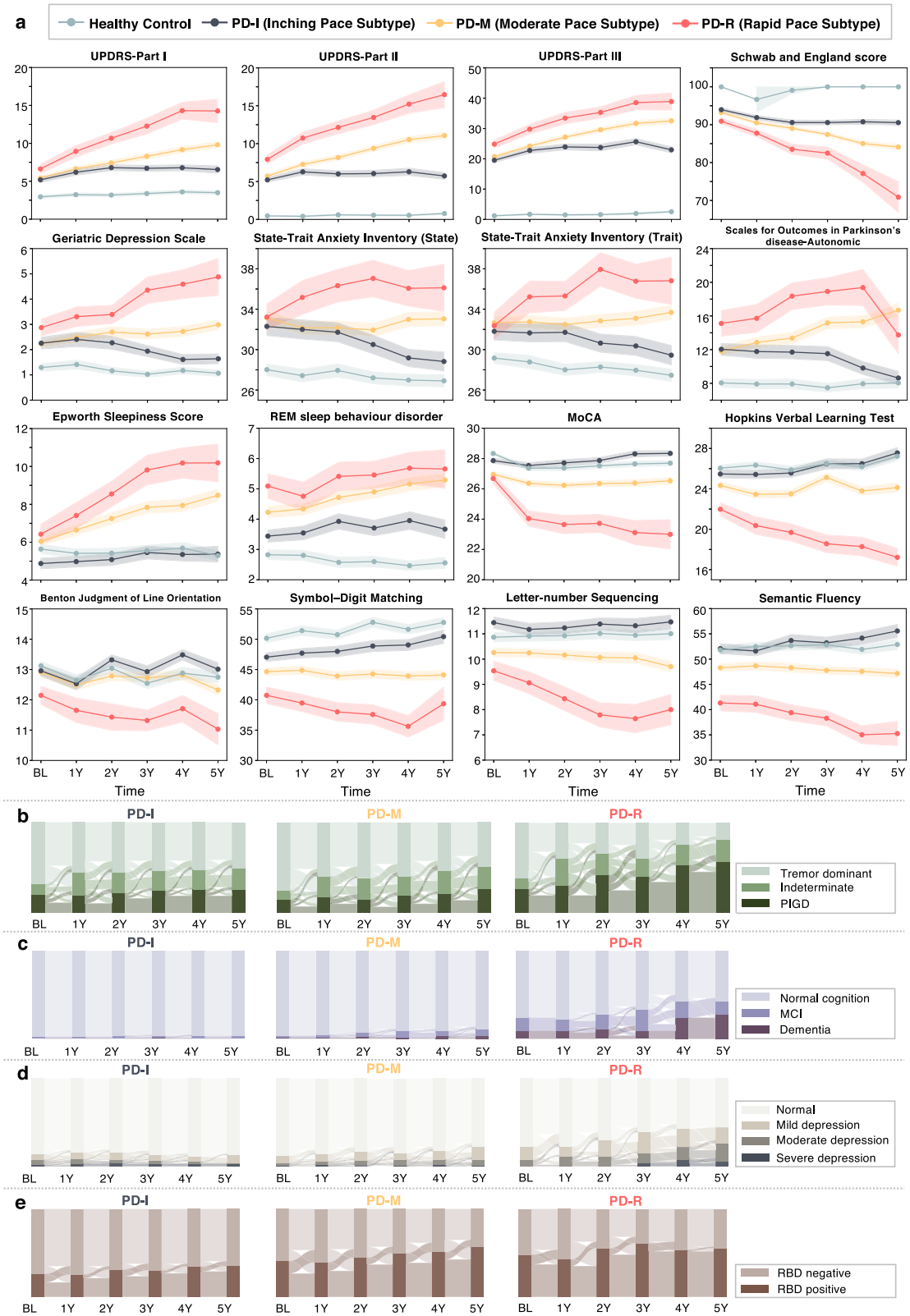

Fig. 2 | Progression patterns of the three PD subtypes within the PPMI cohort. **a** Averaged progression trajectories in clinical manifestations by subtypes, with shading indicating standard error of the mean (SEM). **b** Sankey diagrams showing evolution patterns of motor phenotypes (tremor dominant, indeterminate, and PIGD) by subtypes. **c** Sankey diagrams showing evolution patterns of cognition phenotypes (normal cognition, MCI, and dementia) by subtypes. **d** Sankey diagrams showing evolution patterns of mood phenotypes (normal, mild depression, moderate depression, and severe depression) by subtypes. **e** Sankey diagrams showing evolution patterns of sleep phenotypes (REM sleep behavior disorder [RBD] negative and positive) by subtypes.

Table 2 | Annual progression rates in clinical manifestations and CSF biomarkers by subtypes assessed by linear mixed effects models within the PPMI cohort

Variable	Subtype PD-I (inching pace)		Subtype PD-M (moderate pace)		Subtype PD-R (rapid pace)	
	β	*P* value	*β*	*P* value	*β*	*P* value
Motor manifestations						
UPDRS-Part II	0.42 (0.24, 0.60)	<0.001***	1.02 (0.88, 1.16)	<0.001***	1.98 (1.52, 2.44)	<0.001***
UPDRS-Part III	1.03 (0.62, 1.43)	<0.001***	2.07 (1.74, 2.40)	<0.001***	3.08 (2.12, 4.03)	<0.001***
H&Y Stage	0.07 (0.05, 0.09)	<0.001***	0.09 (0.08, 0.10)	<0.001***	0.15 (0.10, 0.20)	<0.001***
Schwab and England score	−0.98 (−1.26, −0.69)	<0.001***	−1.73 (−1.98, −1.48)	<0.001***	−3.89 (−4.93, −2.85)	<0.001***
Tremor score	0 (−0.02, 0.01)	0.917	0.03 (0.02, 0.04)	<0.001***	0 (−0.03, 0.02)	0.842
PIGD score	0.04 (0.02, 0.05)	<0.001***	0.06 (0.04, 0.07)	<0.001***	0.19 (0.13, 0.24)	<0.001***
Non-motor manifestations						
UPDRS-Part I	0.54 (0.45, 0.62)	<0.001***	0.88 (0.80, 0.96)	<0.001***	1.70 (1.53, 1.88)	<0.001***
Hallucination	0.02 (0.01, 0.03)	<0.001***	0.03 (0.02, 0.04)	<0.001***	0.11 (0.06, 0.16)	<0.001***
Apathy	0.01 (−0.01, 0.03)	0.219	0.05 (0.03, 0.07)	<0.001***	0.12 (0.06, 0.18)	<0.001***
Pain	0.03 (0, 0.06)	0.026*	0.06 (0.04, 0.09)	<0.001***	0.14 (0.10, 0.18)	<0.001***
Fatigue	0.05 (0.02, 0.07)	<0.001***	0.08 (0.06, 0.11)	<0.001***	0.18 (0.13, 0.23)	<0.001***
Sleep						
Epworth sleepiness score	0.17 (0.05, 0.30)	0.008*	0.49 (0.37, 0.60)	<0.001***	1.07 (0.79, 1.35)	<0.001***
REM sleep behavior disorder	0.18 (0.09, 0.28)	<0.001***	0.27 (0.19, 0.34)	<0.001***	0.14 (−0.01, 0.32)	0.076
QUIP (Impulse control disorders)	0.02 (−0.01, 0.04)	0.232	0.06 (0.03, 0.08)	<0.001***	0.02 (−0.02, 0.06)	0.332
Geriatric depression scale	−0.12 (−0.21, −0.02)	0.014*	0.11 (0.04, 0.19)	0.002**	0.47 (0.25, 0.69)	<0.001***
State trait anxiety index						
State subscore	−0.53 (−0.83, −0.24)	<0.001**	−0.02 (−0.27, 0.22)	0.851	0.81 (0.24, 1.40)	0.008*
Trait subscore	−0.32 (−0.58, −0.02)	0.020*	0.16 (−0.07, 0.39)	0.172	1.09 (0.50, 1.69)	<0.001**
SCOPA autonomic questionnaire						
Gastrointestinal	0.28 (0.19, 0.36)	<0.001***	0.31 (0.25, 0.37)	<0.001***	0.39 (0.25, 0.54)	<0.001***
Urinary	0.12 (0.04, 0.20)	0.004*	0.28 (0.17, 0.38)	<0.001***	0.34 (−0.07, 0.76)	0.111
Cardiovascular	0.05 (0.02, 0.09)	0.002**	0.07 (0.04, 0.09)	<0.001***	0.21 (0.13, 0.29)	<0.001***
Thermoregulatory	0.02 (−0.02, 0.06)	0.254	0.06 (0.03, 0.09)	<0.001***	0.07 (0.02, 0.12)	0.017*
Pupillomotor	0.02 (0, 0.05)	0.057	0.04 (0.02, 0.06)	<0.001***	0.07 (0.02, 0.13)	0.006
Skin	0.03 (0.00, 0.07)	0.051	0.10 (0.07, 0.13)	<0.001***	0.14 (0.06, 0.23)	0.002**
Sexual	0.14 (−0.08, 0.37)	0.216	0.37 (0.21, 0.53)	<0.001***	0.69 (0.30, 1.07)	0.001**
Total	0.68 (0.42, 0.94)	<0.001***	1.23 (0.97, 1.47)	<0.001***	1.88 (1.15, 2.61)	<0.001***
Cognitive function						
MoCA	0.05 (−0.02, 0.13)	0.170	−0.09 (−0.17, −0.02)	0.018*	−0.99 (−1.32, −0.67)	<0.001***
Benton judgment of line orientation	0.02 (−0.04, 0.08)	0.548	−0.06 (−0.12, −0.01)	0.026*	−0.27 (−0.41, −0.13)	<0.001**
HVLT-total recall	0.22 (0.05, 0.38)	0.006*	0.03 (−0.10, 0.17)	0.674	−0.98 (−1.24, −0.72)	<0.001***
HVLT-delayed recall	0.11 (0.04, 0.19)	0.005*	0.01 (−0.06, 0.08)	0.735	−0.55 (−0.74, −0.34)	<0.001***
HVLT-discrimination recognition	0.23 (0.14, 0.32)	<0.001***	0.16 (0.09, 0.23)	<0.001***	0.05 (−0.14, 0.25)	0.602
HVLT-retention	0 (0, 0.01)	0.206	0 (0, 0)	0.575	−0.05 (−0.07, −0.03)	<0.001***
LNS	−0.10 (−0.17, −0.02)	0.017*	−0.10 (−0.17, −0.04)	0.002**	−0.48 (−0.68, −0.28)	<0.001***
Semantic fluency	0.25 (−0.11, 0.61)	0.183	−0.23 (−0.47, 0.01)	0.058	−1.46 (−2.07, −0.84)	<0.001***
Symbol digit test	0.34 (−0.04, 0.71)	0.081	−0.12 (−0.37, 0.12)	0.335	−1.23 (−1.90, −0.56)	<0.001**

For each variable, we fitted a linear mixed effect model for each PD subtype, specifying time (year) as the explanatory variable of interest. For all models, individual variation was included as a random effect, and age, sex, and levodopa equivalent daily dose were included as covariates. The annual progression rate of a clinical assessment was obtained as the estimated as *β* (95% CI) of time.
Multiple correction was conducted by controlling false discovery rate (FDR). * FDR adjusted *P* value < 0.05; ** FDR adjusted *P* value < 0.01; *** FDR adjusted *P* value < 0.001.
HVLT Hopkins Verbal Learning Test, *LNS* Letter-number sequencing, *MDS-UPDRS* Movement Disorders Society–Unified Parkinson's Disease Rating Scale, *MoCA* Montreal Cognitive Assessment, *PIGD* postural instability and gait disorder, *PPMI* the Parkinson's Progression Markers Initiative, *SCOPA* Scales for Outcomes in Parkinson's Disease.

(−0.98 /year, 95% CI [−1.26, −0.69], $P < 0.001$). Non-motor symptoms also progressed at a moderate rate, for example, MDS-UPDRS Part I of PD-I increased at a low rate of 0.54 (95% CI [0.45, 0.62], $P < 0.001$).

The Moderate Pace subtype (PD-M), characterized by mild baseline symptoms and moderate symptom progression, included 207 (50.9%) individuals (see Fig. 2a and Tables 1, 2). This subtype had a higher proportion of men ($N = 155$ [74.9%]) compared to PD-I, and these individuals presented with an average age of 59.4 ± 10.0 years at PD onset. Although individuals of PD-M exhibited mild motor and non-motor symptoms at enrollment, mixed with PD-I, they demonstrated worse clinical symptoms

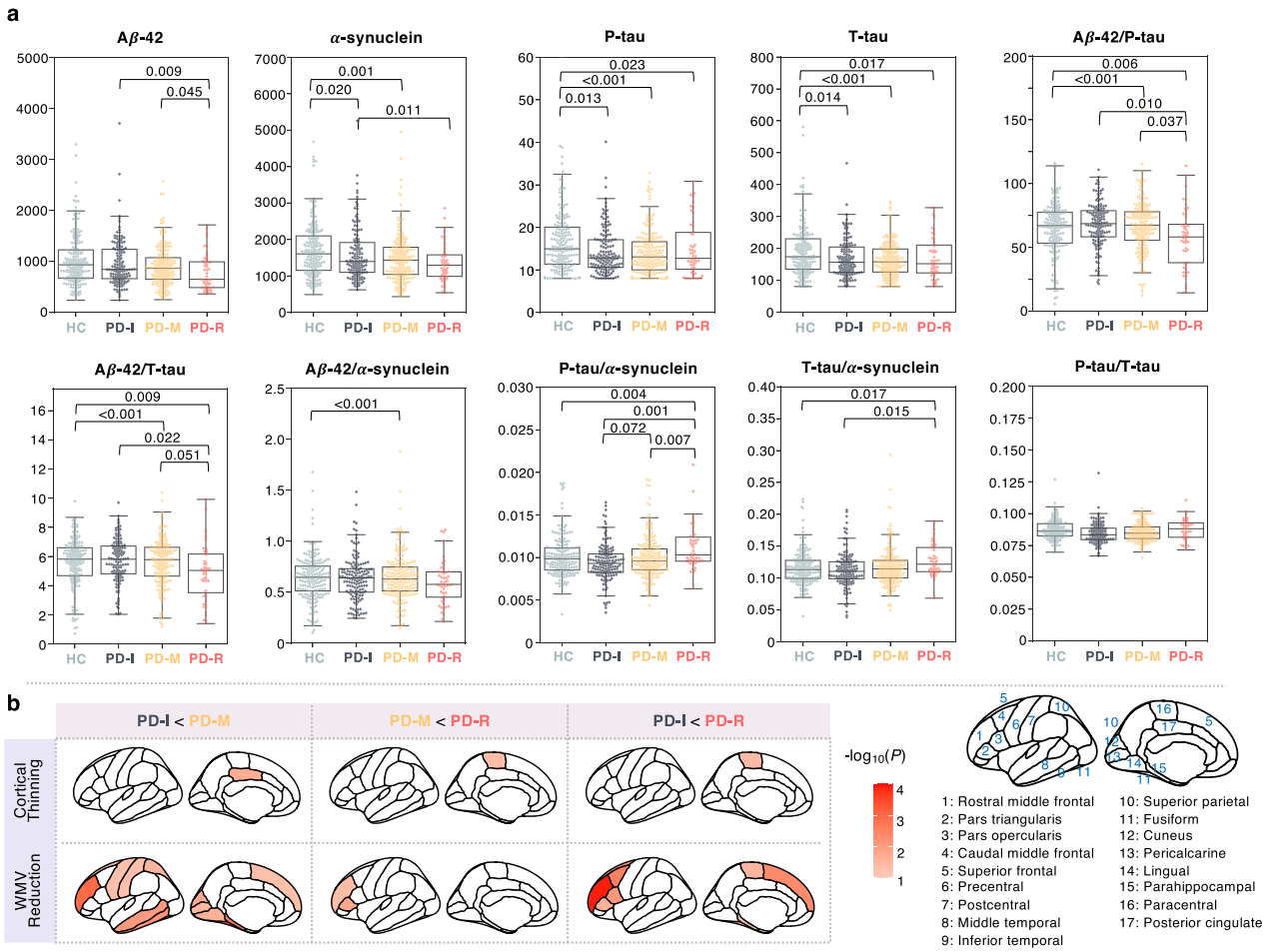

Fig. 3 | CSF biomarkers and neuroimaging markers of the identified subtypes. **a** CSF biomarkers by PD subtypes. On each box plot, the central mark indicates the median value and the bottom and top edges of the box indicate the interquartile range (IQR) with whiskers covering the most extreme values within 1.5 × IQR. **b** Regions showing significant signals in 1-year brain atrophy between a pair of subtypes. 1-year brain atrophy was measured by cortical thickness and white matter volume from 34 region of interests (ROIs), defined by the Desikan-Killiany atlas (averaged over the left and right hemispheres). Color density denotes significance in terms of -$\log_{10}(P)$.

since the second year of follow-up when compared to PD-I (see Supplementary Tables 3, 4). Notably, compared to the PD-I, PD-M displayed faster (generally moderate level) progression rates in most clinical manifestations (see Fig. 2a and Table 2). In terms of overall motor symptoms, the MDS-UPDRS Part II score showed an annual increase of 1.02 (95% CI [0.88, 1.16], $P < 0.001$), the MDS-UPDRS Part III score had an annual increase of 2.07 (95% CI [1.74, 2.40], $P < 0.001$), and the Schwab and England score showed an annual change of −1.73 (95% CI [−1.98, −1.48], $P < 0.001$). For non-motor symptoms, for instance, MDS-UPDRS Part I exhibited an estimated annual increase of 0.88 (95% CI [0.80, 0.96], $P < 0.001$).

We further investigated PD clinical phenotype alterations over time across our identified subtypes. Here, we examined motor phenotypes (tremor dominant [TD], indeterminate, and PIGD) and non-motor phenotypes including cognition (normal cognition/mild cognitive impairment/dementia), REM sleep behavior disorder [RBD] (positive/negative), and depression (non/mild/moderate/severe depression) using Sankey diagrams (see Fig. 2b–e). Overall, the individuals of PD-R had an increasing risk of developing more advanced phenotypes. These included PIGD (Fig. 2b), cognitive dysfunction (mild cognition impairment [MCI] and dementia) (Fig. 2c), and moderate-to-severe depression (Fig. 2d). Individual of PD-I and PD-M were likely to had stable phenotypes throughout the PD course.

The subtypes were reproducible in the PDBP cohort. Following the same procedure above based on clinical variables shared with PPMI, we also obtained the three-cluster structure using clinical progression data of early PD individuals in the PDBP cohort (see Supplementary Note 1). Clinical characteristics at baseline and follow-up, and PD symptom progression profiles of these re-identified subtypes closely mirrored those uncovered in our primary analysis using the PPMI cohort (see Supplementary Figs. 2, 3 and Tables 5–8). These demonstrated the reproducibility and robustness of our subtyping approach and validated the pace subtypes we identified in the PPMI cohort.

Discovery of CSF biomarkers of the PD pace subtypes

We assessed cerebrospinal fluid (CSF) biomarkers among the identified subtypes at baseline and identified potential indicators of the pace subtypes (see Fig. 3a and Supplementary Table 9). While the phosphorylated tau (P-tau) and total tau (T-tau) levels differentiated each subtype from HCs, these biomarkers were not effective at differentiating among the subtypes. The P-tau/α-synuclein ratio might be the most efficacious biomarker for distinguishing among the three subtypes at baseline. Furthermore, the amyloid beta-42 (Aβ-42)/P-tau and Aβ-42/T-tau ratios showed potential in differentiating PD-R (rapid progression) from the other two subtypes. However, they faced challenges when attempting to distinguish PD-M against PD-I. This echoed the similarity in clinical manifestations of the two subtypes at baseline (see Fig. 2 and Tables 1 and 2).

Discovery of imaging biomarkers of the PD pace subtypes

We examined brain atrophy measured by alterations of cortical thickness and white matter volume (WMV) across 34 brain regions of interest (ROIs) defined by the Desikan-Killiany atlas (averaged over

the left and right hemispheres) within the first year after baseline. In Fig. 3b, we marked in red the ROIs where atrophy measures significantly differentiated each pair of PD subtypes ($P < 0.05$), with color intensity reflecting the significance level, i.e., $-\log_{10}(P)$. We observed that reduction of WMV may be more useful than that of cortical thickness in differentiating the subtypes. Specifically, compared to PD-I, PD-M had significantly higher WVM reduction in certain ROIs, such as rostral middle frontal, superior frontal, precentral, postcentral, middle and inferior temporal, fusiform, lingual, parahippocampal gyri, superior parietal, cuneus, and pericalcarine, and cortical thickness atrophy in the posterior cingulate. Compared to PD-M, PD-R had more WVM reduction in rostral middle frontal gyrus and pars triangularis as well as cortical thickness reduction in paracentral gyrus. Compared to PD-I, PD-R had higher WVM reduction in multiple ROIs including rostral middle frontal, pars triangularis, pars opercularis, caudal middle frontal, superior frontal, paracentral, fusiform, and parahippocampal gyri, as well as cortical thickness of the paracentral gyrus. Together, these results indicated that increased WVM atrophy at earlier stage could serve as potential markers for the pace subtypes with distinct progression patterns.

Discovery of genetic components of the PD pace subtypes

We conducted genetic data analyses to identify genetic factors associated to each PD subtype. Given the limited sample size, which makes a genome-wide association study (GWAS) unfeasible, we focused on 90 PD-related single-nucleotide polymorphisms (SNPs) reported in the latest GWAS study[33] and apolipoprotein E (*APOE*) alleles, a known risk factor of AD. Similar to a previous study[34], we utilized the hypergeometric tests to identify variants enriched in each subtype (see "Methods"). Our data suggested certain SNPs that could potentially be associated with each PD subtype (see Supplementary Fig. 4). Since impact of single genetic factors (i.e., SNPs) is low, we further conducted network-based analysis to amplify the genetic signals. We first mapped the subtype-associated SNPs to potential causal genes, leading to a list of genetic associated genes for each subtype. Next, in a human protein-protein interactome (PPI) network we built (see "Methods"), we located these genetic associated genes along with their connected PD contextual genes to construct a sub-network as the genetic molecular module for each subtype. The PD contextual genes were obtained through single nucleus RNA sequence (snRNA-seq) data from dopamine neurons. More details can be found in the "Methods" section.

We successfully identified genetic molecular module for each subtype (see Fig. 4a and Supplementary Fig. 5). For instance, 17 genes were identified as potential genetic associated genes of PD-R. Among these, 14 (*CNTNAP1*, *TRIM40*, *RETREG3*, *FYN*, *MBNL2*, *ATP6V0A1*, *STAT3*, *BECN1*, *BAG6*, *HLA-DRB1*, *HLA-DRB5*, *HLA-DQB1*, *SIPA1L2*, and *PRRC2A*) directly connected to snRNA-seq-derived PD contextual genes in the PPI network, thus constructing a genetic molecular module of PD-R, as depicted in Fig. 4a. Among the genes identified in the module was the signal transducer and activator of transcription 3 (*STAT3*), which transmits signals for the maturation of immune system cells. In microglia, *STAT3* is known to influence the expression of inflammatory genes[35]. *FYN* is a member of the protein tyrosine kinase oncogene family, which plays a critical role in cell communication and signaling pathways, particularly in the nervous system. *FYN* can promote *STAT3* signaling as part of proinflammatory priming of microglia[36]. *BECN1* (beclin 1) regulates autophagy, which clears toxic proteins such as α-synuclein aggregates in brain. *LRRK2* was found as a hub node in the modules of PD-I and PD-M. This was supported by previous evidence that mutated *LRRK2* is likely to be associated with lower progression rate of PD[37]. Finally, pathway enrichment analyses pinpointed prominent pathways enriched in the three subtypes (see Fig. 4b and Supplementary Fig. 5b, d). For instance, some notable pathways that were associated with the PD-R include neuroinflammation (immune system-related pathways), oxidative stress, metabolism (insulin receptor signaling and Type I diabetes pathways), and the Alzheimer's disease (AD) pathway. Top-ranked pathways for each subtype can be found in Fig. 4b and Supplementary Fig. 5b, d.

Discovery of transcriptomic profiles of the PD pace subtypes

We next investigated the transcriptomic changes associated with each PD subtype using the whole blood bulk RNA-seq data in the PPMI cohort (see "Methods"). Differential gene expression analyses, comparing each subtype with HCs, identified 2176, 2376, and 2305 differentially expressed genes (DEG, adjusted $P < 0.05$) for the PD-I, PD-M, and PD-R subtypes, respectively (see Supplementary Fig. 6).

Drawing on the DEGs of each subtype and PPI network we built as input, we utilized our Genome-wide Positioning Systems network (GPSnet) algorithm[38] to identify transcriptomic molecular modules specific to each subtype (see "Methods"). Our analysis relied on the notion that a subtype-specific molecular module is a set of genes which are (1) densely interconnected within the PPI network and (2) differentially expressed in the specific subtype. We identified three distinct molecular modules for PD-I, PD-M, and PD-R, consisting of 211, 213, and 240 genes, respectively (see Supplementary Figs 7–9). Taking PD-R as an example, we found several genes of interest within the module of the subtype (see Supplementary Fig. 9). For instance, the module gene *E2F1*, a member of the *E2F* family of transcription factors, has been previously linked to neuronal damage and death[39,40]. The module gene apolipoprotein A1 (*APOA1*) produces the *APOA1* protein, a major protein component of high-density lipoprotein (HDL) in the blood. A previous PPMI cohort study[41] has revealed the association between plasma *APOA1* level with age at PD onset and motor symptom severity. Another module gene, *NEDD4*, promotes removal of α-synuclein via a lysosomal process in human cells which may help resist PD[42]. Furthermore, the *GATA* binding protein 2 (*GATA2*) has demonstrated its crucial role in neuronal development, specifically influencing the cell fate determination of catecholaminergic sympathetic neurons and modulating the expression of endogenous neuronal α-synuclein[43]. Pathway enrichment analyses were conducted based on the subtype-specific modules and suggested pathways enriched in each subtype (see Fig. 4c-d and Supplementary Fig. 10). Again, taking PD-R as an example, Fig. 4c illustrated a sub-network of its transcriptomic molecular module. The phosphoinositide-3-kinase/protein kinase B (PI3K/AKT), angiogenesis, T cell receptor signaling, and senescence pathways may play a role in rapid PD progression, i.e., PD-R. Top-ranked pathways in each subtype can be found in the Fig. 4c, d and Supplementary Fig. 10.

A classification model for distinguishing the PD pace subtypes early

The identified PD pace subtypes demonstrated clear progression trends within the PD course (over 5 years). To gain more prognostic insights, we built a classification model for separating the pace subtypes at early stage. To build the model, we used individual-level demographics, genetic data, as well as clinical assessments, CSF biomarkers, and brain atrophy measures collected within the first year after baseline. We leveraged a cascade framework[44] consisting of two base random forest classifiers: one separating PD-R from the others and another distinguishing PD-I and PD-M (see Supplementary Fig. 11). As shown in the figure, the model was very effective in separating PD-R from the other two subtypes, attaining an area under the receiver operating characteristics curve (AUC) of 0.87 ± 0.05. While distinguishing PD-I and PD-M was challenging due to their similar clinical characteristics at baseline, the model still achieved acceptable performance with an AUC of 0.74 ± 0.07.

Discovery of PD pace subtype-based repurposable drug candidates

We next sought to identify potential treatments to slow or halt PD progression by targeting the subtype-specific molecular bases. To achieve this, we used transcriptomics-based drug-gene signature data in human cell lines from the Connectivity Map (CMap) database[24]. We evaluated each drug's

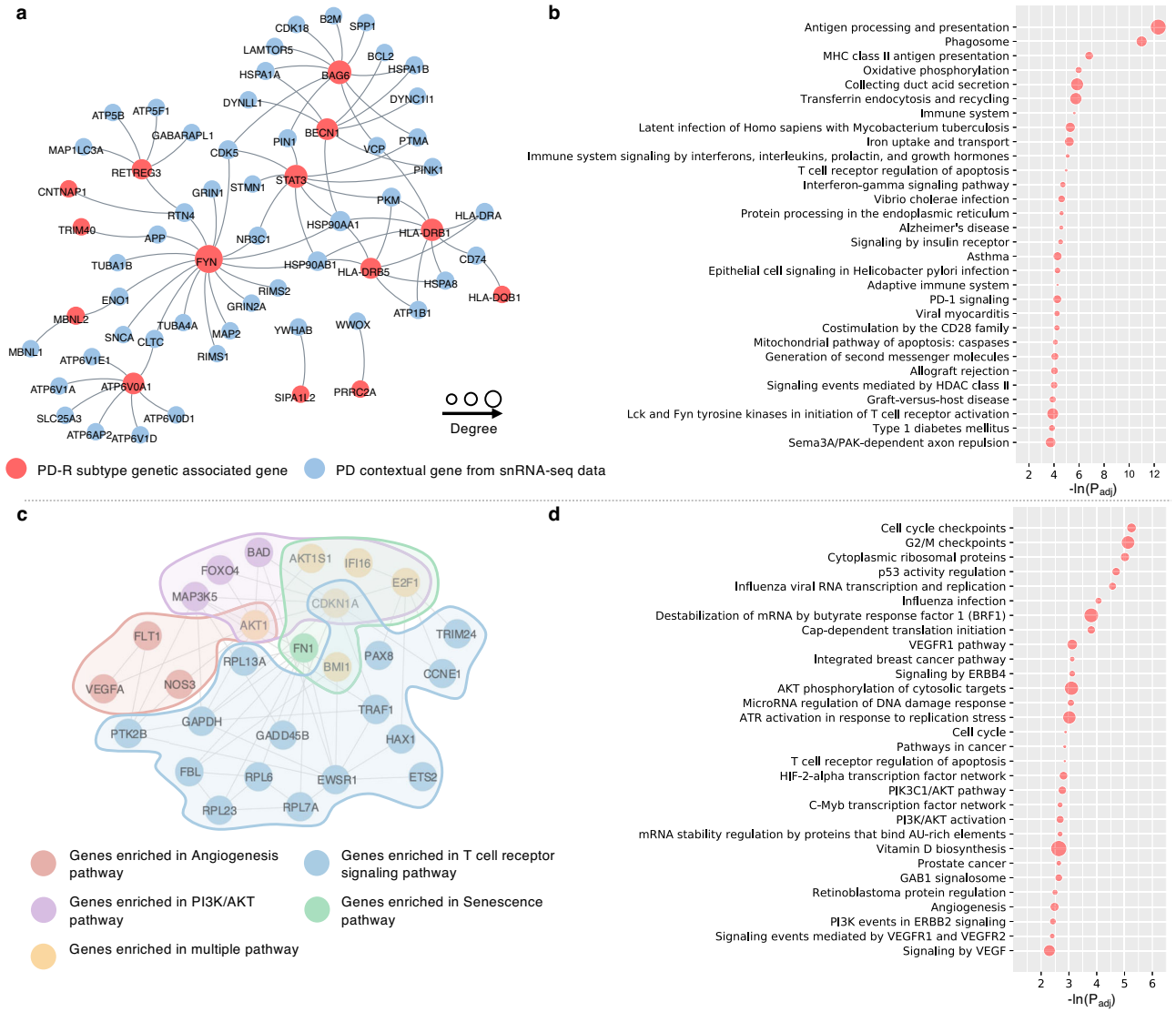

Fig. 4 | PD-R subtype-specific molecular modules revealing potential biological mechanisms of rapid PD progression. a Genetic molecular module of PD-R. **b** Pathways enriched based on genetic molecular module of PD-R. **c** A sub-network of transcriptomic molecular module of PD-R. The entire transcriptomic molecular module of PD-R can be in the Supplementary Fig. 9. **d** Pathways enriched based on transcriptomic molecular module of PD-R.

potential therapeutic effects for preventing PD progression by assessing its ability to reverse dysregulated gene expression levels of the subtype-specific molecular modules (see "Methods"). More specifically, we used gene set enrichment analysis (GSEA) to compute an enrichment score (ES) with permutation tests for each tested drug[45,46]. We chose ES > 3 and $Q < 0.05$ as the cutoffs to prioritize potential drug candidates (see "Methods" for more details). In total, we investigated 1,309 drugs from the CMap, resulting in 49, 65, and 207 candidates (ES > 3 and $Q < 0.05$) based on molecular modules of the PD-I, PD-M, and PD-R, respectively.

Particularly, drug candidates predicted by targeting PD-R-specific gene module fell into fourteen pharmacological categories (according to Anatomical Therapeutic Chemical [ATC] code), including agents for the nervous and cardiovascular systems, immunomodulating agents, etc. The top-ranked drug candidates for PD-R were highlighted in Fig. 5a. Of note, our predictions included three US Food and Drug Administration (FDA)-approved drugs for PD treatment, including biperiden, amantadine, and levodopa. This demonstrated that our method can predict effective medication for PD. In addition, our data suggested some potential repurposable drug candidates. For instance, ambroxol[47], an FDA-approved drug for respiratory disorders was predicted to be potentially useful for PD-R. Ambroxol has shown promise as a potential disease-modifying treatment for PD in a phase II, single-center, double-blind, randomized placebo-controlled trial[48]. Guanabenz, another drug prioritized for PD-R, is traditionally used for hypertension treatment. Guanabenz has been found to reduce 6-hydroxydopamine (6-OHDA)-induced cell death in ventral midbrain dopaminergic neurons in culture, and in dopaminergic neurons in the substantia nigra of mice[49].

Real-world evidence (RWE) highlighted metformin as a promising candidate to mitigate PD progression

RWD, such as the patient's clinical records collected from regular healthcare procedures, has been an important resource for gathering clinical evidence, i.e., RWE, about the usage and potential benefits or risks of medical treatment. RWE can provide insights into how a treatment works in a broader patient population in regular healthcare circumstances[50]. To gain RWE for validating treatment effects of the identified drug candidates, we used two large-scale, de-identified, patient-level RWD databases: the INSIGHT[51], containing clinical information of ~21.6 million patients in the New York City and Houston as well as the OneFlorida+[52], covering ~17 million patients in Florida, ~2 million in Georgia, and ~1 million in Alabama. We focused on five drugs based on subject matter expertise and a combination of multiple factors including: (1) the strength of network-based prediction

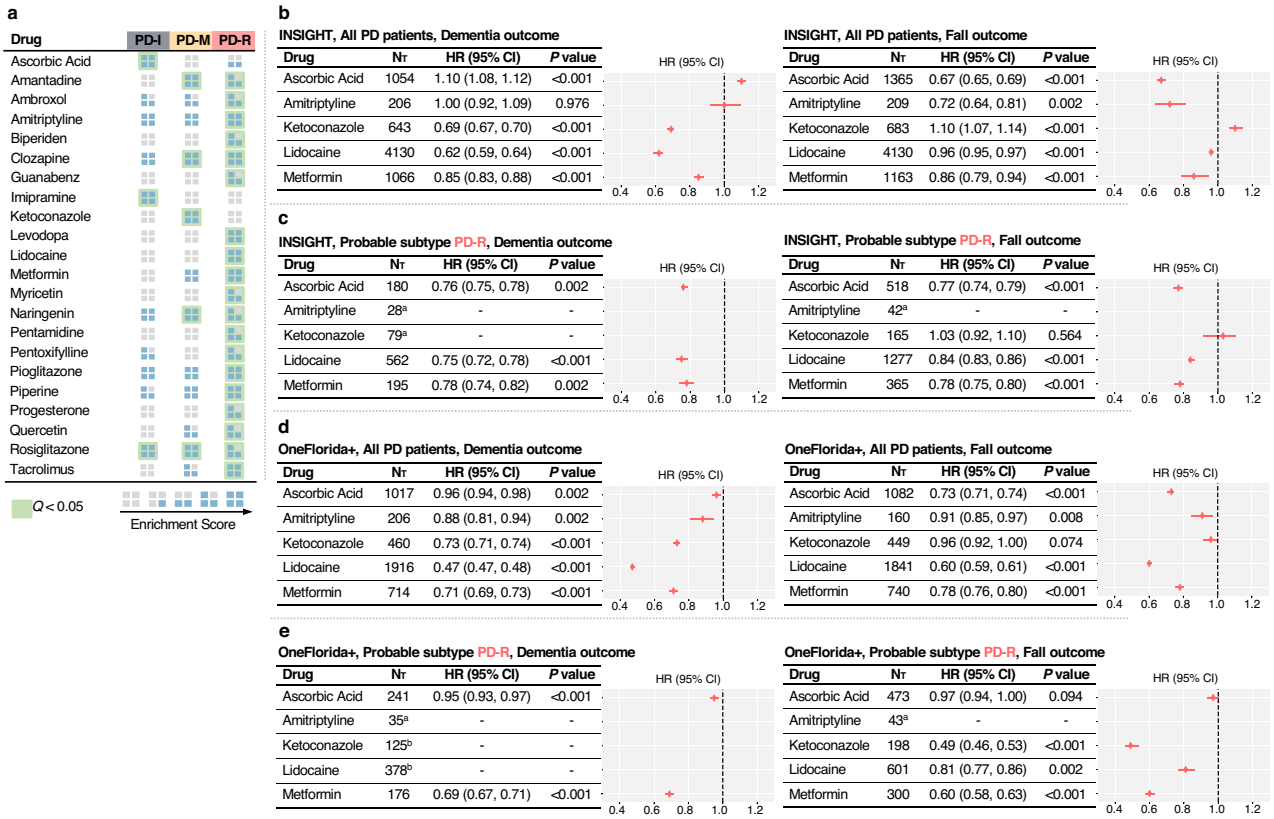

Fig. 5 | Identified repurposable drug candidates for preventing PD progression by targeting subtype-specific molecular changes. a Gene set enrichment analysis (GSEA) based on subtype-specific gene modules with bulk RNA-seq data of individuals and transcriptomics-based drug-gene signature data in human cell lines identified repurposable drug candidates for different PD pace subtypes. Treatment effect estimation using the INSIGHT data within the broad PD population (**b**) and probable PD-R population (**c**). Treatment effect estimation using the OneFlorida+ data within the broad PD population (**d**) and probable PD-R population (**e**). [a]The drug doesn't have sufficient patient data (<100) for analysis. [b]The drug does not have sufficient balanced emulated trials (<10). N_T indicates the number of eligible PD patients who received the tested drug after PD initiation.

(ES score >3 and *Q* value < 0.05); (2) sufficient data for the target of interests (we required the tested drug to be taken by more than 100 patients in the RWD database, see "Methods"); (3) pharmacokinetics profiles; and (4) knowledge of functional data from the literature (for instance, we paid attention to drugs that have been reported to interact with PD-related pathways).

To evaluate treatment effects of the tested drugs, we conducted trial emulation based on our computational trial emulation framework[53] (see "Methods" and Supplementary Fig. 12). Specifically, for each emulated trial, we built its treated group as the set of eligible patients who received a specific tested drug after PD initiation. Subsequently, we can formulate a control group as the set eligible patients who received alternative medications (i.e., those falling under the same ATC level 2 class as the tested drug, excluding the tested drug itself), through 1:1 nearest-neighbor matching. A propensity score method was used to learn empirical treatment assignment given the baseline covariates and an inverse probability of treatment weighting was used to balance the treated and control groups[54]. We created 100 emulated trials for each tested drug and excluded drugs which had less than 10 balanced trials. We finally estimated drug treatment effects for the balanced trials using the Cox proportional hazard models[55]. Following a previous study[56], treatment efficiency was measured as the reduced risk to develop PD-related outcomes, including falls (a proxy of advanced motor impairment and dyskinesia relevant to PD progression) and dementia. More details of this analysis can be found in the "Methods" section and Supplementary Fig. 12.

Using the eligible PD patient population within both INSIGHT and OneFlorida+ data, we were able to create sufficient (≥10) successfully balanced emulated trials for the five tested drugs. Metformin, an FDA-approved anti-diabetic medication, emerged as the most promising candidate for potentially preventing PD progression (see Fig. 5). Specifically, within the INSIGHT data, the usage of metformin was associated with 15% reduced likelihood of developing dementia (hazard ratio [HR] = 0.85, 95% CI [0.83, 0.88], *P* value < 0.001) as well as 14% reduced likelihood of onset of falls (HR = 0.86, 95% CI [0.79, 0.94], *P* value < 0.001), compared to the usage of alternative drugs (see Fig. 5b). Similarly, in the OneFlorida+ data, metformin usage was associated with 29% reduced likelihood of developing dementia (HR = 0.71, 95% CI [0.69, 0.73], *P* value < 0.001) as well as 22% reduced likelihood of onset of falls (HR = 0.78, 95% CI [0.76, 0.80], *P* value < 0.001) compared to the alternative medication usage (see Fig. 5d). Of note, we noticed a discrepancy in the therapeutic signals (i.e., HRs) of metformin between the INSIGHT and OneFlorida+ data. This may be due to the variations in healthcare settings and population diversity within the two databases. However, what stands out more significantly is that our data, drawn from the two independent real-world healthcare databases, consistently indicate that the use of metformin post-PD diagnosis may play a role in mitigating the progression of both motor and cognitive manifestations throughout the course of PD.

PD patients with early cognitive impairment are likely a subpopulation of PD-R (see Fig. 2c). We further examined the drug candidates' treatment effects in this probable PD-R population. Specifically, we defined the probable PD-R as (1) patients diagnosed with PD, and (2) with any cognitive deficit diagnosis within the first year following PD initiation but prior to initiation of the tested drug. Under such settings, we repeated the trial emulation. We found that metformin could still be a good candidate for preventing PD progression within this population, and its treatment efficiency increased compared to that within the general PD population (see

Clinical subtypes of Parkinson's disease	Method	Data used for subtyping	Baseline clinical manifestations	Progression
Classical motor subtypes used by many authors • Tremor dominant • Akinetic-rigid (AR) • Postural instability and gait disorder (PIGD)	Hypothesis-based (clinician observation)	• Single-domain (motor only) • Usually applied to a single point in time	• PIGD/AG has less tremor, more gait and balance challenges and longer symptom duration.	• **Unstable subtype over time**. Using this subtyping the phenotypes tend over time to shift from TD subtype to others.
Data-driven subtypes introduced by Fereshtehnejad et al.) • Mild motor-predominant (MMP) • Intermediate • Diffuse malignant (DM)	Data-driven	• Multiple-domains including motor and non-motor symptoms. • Data usually collected at one time point.	• Early PDs at baseline • There can be mixed symptom duration at baseline. • The subtypes manifest distinct clinical severities at baseline • MMP had the mildest symptom severity • DM had the most severe symptom severity and more PIGD	• **Modest** differences in progression rates among the three subtypes. • Correlations between subtypes using motor and non-motor phenotype dynamics were not explored.
PD pace subtypes (this work) • PD-I (Inching Pace) – mild baseline severity and mild progression • PD-M (Moderate Pace) – mild baseline severity and moderate progression • PD-R (Rapid Pace) – worse baseline severity and rapid progression	Data-driven using ML and DL	• Multiple-domain motor and non-motor symptoms used • Requires 2-5 years of clinical symptom progression data	• Early PDs at baseline • Mixed symptom durations at baseline in the cohort. • PD-R had worse motor and cognitive symptom severity at baseline, compared to PD-I and PD-M. • PD-I and MD-M had mixed, mild clinical manifestations at baseline.	• There were **significantly distinct annual progression rates** in diverse clinical assessments among the three subtypes, adjusting for age, sex, levodopa equivalent dosage at clinical visits. • PD-R had an increasing risk when compared to the classical phenotypes to morph into PIGD and to develop mild cognitive impairment, dementia, and/or depression.

Fig. 6 | Comparisons of the identified pace subtypes with conventional motor subtypes and prior data-driven subtypes. Notably, our subtyping algorithm was completely data-driven and hypothesis-free. In addition, since our method modeled individuals' phenotypic progression profile for PD subtyping, the identified subtypes demonstrated unique progression patterns and, importantly, were stable over time.

Fig. 5b–e). Within the INSIGHT data, the usage of metformin was observed to be associated with 22% reduced likelihood of dementia (HR = 0.78, 95% CI [0.74, 0.82], P value = 0.002) as well as associated with 22% reduced likelihood of fall events (HR = 0.78, 95% CI [0.75, 0.80], P value < 0.001) compared to the non-metformin alternative drug usage (see Fig. 5c). Similarly, within the OneFlorida+ data, the usage of metformin was associated with 31% reduced likelihood of dementia (HR = 0.69, 95% CI [0.67, 0.71], P value < 0.001) and with 40% reduced likelihood of fall events (HR = 0.60, 95% CI [0. 58, 0. 63], P value < 0.001) compared to the alternative medication usage (see Fig. 5e)

Discussion

The heterogeneous nature of PD progression has been widely recognized, yet it remains incompletely deciphered[6–8]. In this investigation, we conducted comprehensive analyses of various types of data to study PD progression, including longitudinal clinical data, CSF biospecimen, neuroimaging, genetic data, and transcriptomic data of individuals, publicly available PPI data and transcriptomics-based drug-gene signature data in cell lines (CMap), and two large-scale patient-level RWD databases (INSIGHT and OneFlorida +). To this end, we employed machine learning, deep learning, network medicine, and RWD-based trial emulation techniques.

Overall, our findings in this study—not only the discovery of pace subtypes with distinct progression patterns, but also the identification of differentiated CSF biomarker levels, atrophy in different brain regions, and distinct molecular modules associated with these subtypes—provided evidence supporting the existence of different pathophysiologic mechanisms driving different PD subtypes, leading to differentiated PD progression trajectories. This highlighted the necessity of treating PD subtypes as unique sub-disorders within clinical practice, where our pace subtypes could inform patient stratification and management. Additionally, we have built a classification model that can effectively distinguish the pace subtypes early, based on demographic, genetic, and clinical data gathered within the first-year post-baseline (see Supplementary Fig. 11). This model could serve as a valuable tool in clinical settings, enabling the swift identification of subtypes for newly diagnosed patients, thereby facilitating prompt and precise patient stratification and management. Lastly, this study, through subtype-specific in silico drug repurposing, has identified potential therapeutic candidates (e.g., metformin), underscoring the importance of developing subtype-specific drug treatments for PD. This tailored approach could revolutionize how PD is treated, offering more personalized and potentially more effective therapeutic strategies.

Conventional PD subtyping algorithms typically rely on a priori hypotheses focusing on single-domain information (e.g., tremor) and have resulted in clinical subtypes of PD[10]. Notable examples included the motor subtypes such as tremor dominant (TD), akinetic-rigid (AR), and postural instability and gait disorder (PIGD). Recent data-driven methods have become increasingly popular in subtyping complex diseases such as PD, due to their hypothesis-free nature that allows unbiased investigation of vast and complicated clinical data to identify subtypes. Compared to the previous PD subtyping systems, our PD subtyping approach has remarkable advantages (see Fig. 6). First, unlike the conventional approaches that may be biased due to researcher's knowledge, our subtyping algorithm was completely data-driven and hypothesis-free. Second, while the importance of subtype analysis in studying PD heterogeneity has been acknowledged[8–11], the debate persists on whether the subtypes represent truly distinct entities or different stages within the same progression trajectory of the disease[10,57], particularly as recent cohort studies have demonstrated that the conventional clinical subtypes were unstable over time (see Fig. 6)[12,13]. Additionally, prior data-driven methods were mainly based on individual's data collected at a specific time point (e.g., the baseline visit). This limited their capacities to accurately capture the evolving patterns of PD progression to address subtype instability over time. In contrast, our approach leveraged deep learning to model high-dimensional, longitudinal (over 5 years) clinical progression data of individuals, capturing the intrinsic phenotypic progression profiles of individuals to detect subtypes. Therefore, individuals' subtype memberships were inherently stable over time, with each subtype reflecting a distinct PD progression pattern (in Fig. 2 and Table 2).

A remarkable finding is the PD-R subtype, which had worse symptom severities at baseline and, more importantly, experienced the most rapid motor and non-motor symptom progression paces throughout the PD course. A recent data-driven study[58] has reported a "diffuse malignant" clinical subtype of PD, which also exhibited poor baseline clinical manifestations and faster progression during follow-up. However, the progression speed of this "diffuse malignant" subtype was only marginally faster than the others and was noted in only a few clinical assessments, such as the Schwab and England score, MoCA, and global composite outcome. In contrast, PD-R demonstrated the greatest progression rates in a diverse spectrum of motor and non-motor manifestations among the PD population, estimated by the linear mixed effects models that considered multiple

covariates (age, sex, levodopa equivalent daily dose usage at visits) (see Table 2 and Supplementary Table 8). Additionally, when compared to other subtypes, PD-R showed a more consistent trend to shift from TD toward PIGD (see Fig. 2b). This demonstrated that PD-R can progress faster than the others, but also indicated that the TD and PIGD motor phenotypes are more likely to represent different stages of PD. Similarly, PD-R was more likely to develop cognitive and behavioral abnormalities, including MCI and/or dementia as well as depression (see Fig. 2c-d).

Another two subtypes, i.e., PD-I and PD-M, exhibited comparable symptom severities at baseline (see Fig. 2a and Table 1). Nevertheless, importantly, substantial differences in clinical manifestations, both motor and non-motor, became apparent after two years from baseline (see Fig. 2a, Table 1, and Supplementary Tables 3 and 4). PD-M manifested a much faster annual progression in most clinical manifestations compared to PD-I (see Table 2). The findings suggested that even when individuals with PD initially presented with similar symptom severities at baseline, their progression can vary in pace, leading to distinct trajectories and prognoses. This also supported the hypothesis that subtypes defined relying on baseline data can display fluctuations in symptom severity and/or outcomes during follow-up[12–14]. Our findings underscore the importance of thoroughly modeling individual's symptom progression trajectories when attempting to identify PD subtypes.

Cognitive impairment-related pathology may play a role in driving rapid PD progression, i.e., PD-R. Firstly, compared to the other subtypes, PD-R had worse cognitive performances in numerous clinical assessments and demonstrated an 8 to 10-fold higher prevalence of cognitive impairment at baseline (see Table 1 and Fig. 2c). Secondly, the studied PD participants had no AD diagnosis at baseline and follow-up visits, yet our analysis of baseline CSF biomarkers suggested that several AD-type biomarkers could serve as potential early indicators for the PD pace subtypes we identified (see Fig. 3a). More specifically, the CSF P-tau/α-synuclein ratio appeared to be a promising biomarker for differentiating the three PD subtypes from one another at baseline, although it had limited capacity to distinguish these subtypes against HC participants. Aside from the P-tau/α-synuclein ratio, we didn't identify any additional CSF biomarkers that could effectively separate the PD-I and PD-M at baseline, which was consistent with the mixed symptom severities of the two subtypes at baseline. Other AD-related biomarkers including CSF Aβ-42, Aβ-42/P-tau ratio, and Aβ-42/T-tau ratio distinguished the PD-R from the PD-I and PD-M subtypes. Our findings are supported by the prior study[59], which suggested that although the AD biomarkers alone are not helpful in diagnosing PD, they could improve prognostic evaluation. This emphasizes the critical role that the primary hallmarks of AD, the amyloid and tau pathologies, may have in the progression of PD. Last, we also observed that early (within the first year after baseline) structural atrophy in certain brain regions responsible for cognitive functions may be potential markers of the PD pace subtypes (see Fig. 3b). Some notable markers we found included frontal lobe regions including rostral middle frontal, pars triangularis, caudal middle frontal, superior frontal gyri, as well as fusiform and parahippocampal gyri.

Analyses of individuals' genetic and transcriptomic profiles armed with network medicine approaches led to the identification of subtype-specific molecular modules and biological pathways. Despite the commonalities in certain pathways across subtypes, each subtype was associated with unique pathways. This suggested that there might be unique pathophysiological underpinnings that drive the distinct progression patterns (i.e., the pace subtypes) observed in PD (see Supplementary Figs 5e and 10c). In particular, the molecular modules of PD-R suggested the potential roles of neuroinflammation, oxidative stress, metabolism, and AD pathways, along with PI3K/AKT and angiogenesis pathways in rapid PD progression (see Fig. 4). Neuroinflammation, a crucial factor in PD pathogenesis, involves the chronic inflammatory response in the brain, particularly the activation of microglia[60]. Upon activation, microglia release pro-inflammatory cytokines and reactive oxygen species (ROS). Neuroinflammation-related ROS production subsequently leads to oxidative stress and neuronal damage, thereby playing a significant role in PD's pathogenesis and progression[60,61]. The PI3K/AKT pathway is fundamental to various physiological processes, such as cell survival, growth, proliferation, and metabolism. It also plays a crucial role in modulating the mammalian target of rapamycin (mTOR), a pivotal regulator of autophagy—an essential cellular process involved in degrading and recycling damaged or unnecessary cell components[62,63]. Previous evidence has indicated that disruptions in the PI3K/AKT pathway impairs autophagy, leading to ineffective elimination of abnormal and toxic protein aggregates, contributing to neuronal death[57,58]. In addition, angiogenesis, the process of forming new blood vessels from existing ones, has been implicated in PD pathologies[64]. Aberrant angiogenesis has been associated with blood-brain barrier (BBB) dysfunction. This dysfunction may lead to changes in BBB permeability, potentially permitting neurotoxic substances to enter the brain[65]. It could also trigger neuroinflammation and abnormal immune responses, both contributing to neurodegeneration[65].

Using in silico drug repurposing approaches, we prioritized drug candidates that could target specific PD subtypes (see Fig. 5). Metformin, a traditional type 2 diabetes (T2D) medication, stood out as a potentially promising candidate for repurposing. Our data demonstrated the potential effect of metformin in counteracting molecular changes triggered in PD-R (ES = 4.44, P value < 0.001). Furthermore, RWE gained from large-scale individual-level RWD substantiated our findings, showing that metformin usage was associated with a decreased risk of advanced PD outcomes including dementia (cognitive impairment) and falls (advanced motor dysfunction). Notably, a stronger treatment effect was observed within the surrogate PD-R subtype population (INSIGHT data: dementia, HR = 0.78, 95% CI [0.74, 0.82], P = 0.002; falls, HR = 0.78, 95% CI [0.75, 0.80], P value < 0.001; OneFlorida +: dementia, HR = 0.69, 95% CI [0.67, 0.71], P value < 0.001; falls, HR = 0.60, 95% CI [0.58, 0.63], P value < 0.001) compared to that within the broader PD population (INSIGHT data: dementia, HR = 0.85, 95% CI [0.83, 0.88], P value < 0.001; falls, HR = 0.86, 95% CI [0.79, 0.94], P value < 0.001; OneFlorida +: dementia, HR = 0.71, 95% CI [0.69, 0.73], P value < 0.001; falls, HR = 0.78, 95% CI [0. 76, 0. 80], P value < 0.001). Metformin, known for its efficacy in reducing hepatic glucose release and augmenting the body's insulin sensitivity, has recently also been highlighted for its potential neuroprotective role in neurodegenerative disorders including PD. This is supported by a growing body of in vitro and in vivo studies of PD[66–68]. First, known as a strong AMPK (5'-adenosine mono-phosphate-activated protein kinase) activator, metformin's neuroprotective effects may be primarily attributable to its role in activating the AMPK signaling pathway[69,70]. Past research found that AMPK can suppress microglial activation, thus potentially mitigating neuroinflammation[69]. Recent investigations employing animal models revealed metformin's ability to reduce microglial cell numbers, modify microglial activation pathways, and attenuate both inflammasome activation and the accumulation of ROS in microglia[71,72]. In addition, metformin has been found to modulate PI3K/AKT activity[73,74]. Enhanced activations of AMPK and PI3K/AKT can amplify autophagy, which aids in the removal of toxic proteins like α-synuclein, thereby preventing dopaminergic neuronal death[73,74]. Finally, animal model studies have found that metformin may induce both angiogenesis and neurogenesis in the brain under different experimental conditions[75–77]. In conclusion, the neuroprotective effect of metformin against the PD-R subtype (rapid PD progression) could be attributed to its ability to inhibit neuroinflammation, enhance PI3K/AKT activity for efficient α-synuclein clearance, and stimulate angiogenesis. Our data further supported the potential of metformin in treating PD, especially the rapid pace PD subtype patients, but also indicated the potential of our method in accelerating precision medicine approaches towards identifying therapeutics that reduce PD progression.

Limitations of this work should also be acknowledged. First of all, the PPMI and PDBP cohorts collected abundant individual-level phenotypic, molecular, and imaging data, enabling comprehensive multi-modal, longitudinal analysis of PD; however, the majority of participants in the two

cohorts are white with high education level (see Supplementary Table 1). Such a skewed population may weaken the generalizability of our findings to the entire PD population.

We modeled longitudinal phenotypic data to capture the heterogeneous progression procedure of PD. We acknowledge, however, that the variability in individuals' health conditions at baseline could influence the modeling of PD progression, and this factor was not fully considered in the present analysis. Therefore, there is an unmet need to perform comparative analyses with data from individuals with more advanced stages of PD at baseline, such as those diagnosed for over three years in PDBP. This may help us better understand how PD evolves from its early to later stages and refine our progression models accordingly. Data incompleteness may also impact PD progression modeling and subtype identification. To address this issue, we excluded individuals who had insufficient longitudinal information and used a combination of the last observation carried forward (LOCF) and next observation carried backward (NOCB) methods[78] to impute missing values while retaining longitudinal consistency. We also imputed the variables missing all values along the timeline of an individual based on median values of the population (see "Methods"). Such procedures help mitigate the missing value issue but might also include unexpected noise and bias in analysis.

Our omics data analyses, aiming to explore potential biological mechanisms driving the subtypes, may have limitations. First, instead of exhaustively examining the genetic factors whole genome-wide, our genetic examination was selective, focusing on known PD-related variants reported in the latest GWAS study[33] and *APOE* alleles. Such a targeted strategy could mitigate the issues caused by the small impact sizes of single genetic variants and limited sample size, which become particularly evident in subtype-specific investigations. Additionally, utilization of a network medicine strategy with the PPI network allowed us to extend our genetic and transcriptomic insights on the genome-wide scale. However, it remains possible that key genes pivotal to specific PD subtypes may have been missed. Second, our study was potentially limited to the analysis of whole-blood transcriptomics. Exploring transcriptomic data from other tissues (especially the brain tissues) and CSF, as well as integrating other omics approaches like proteomics and single-cell RNA-seq data, might yield more comprehensive insights into the molecular underpinnings driving PD progression heterogeneity. Third, potential confounding factors, such as genetic-environmental interactions, comorbidities, and socioeconomic status, were not adequately accounted for in our current study. These variables require further examination in future research to clarify their influence on PD progression. Lastly, our molecular findings, derived based on the PPMI dataset, lack extensive validation in independent cohorts or through functional analyses. Future studies should endeavor to corroborate our results in molecular data from separate validation cohorts, such as the PDBP, and employ functional analyses to further authenticate the molecular observations.

Translating subtype findings from research cohorts (e.g., PPMI and PDBP) to RWD to gather real-world evidence is crucial but poses significant challenges due to data discrepancy, skewed population, and data quality. We defined surrogate PD-R in RWD based on its unique clinical feature of early cognitive impairment, which may not be accurate. Future research employing more sophisticated techniques like transfer learning, which facilitates the transfer of knowledge across cohorts with partially overlapping features, is needed. In addition, RWD analysis may not be suitable for testing drugs that have too few patient data. We were able to perform population-based validation for metformin, as it had a large amount of patient data available. For drugs with fewer patient data, replication of the associations using multiple large population-based cohorts is suggested. Moreover, our RWD analysis cannot build causal relationships between use of a specific medication (e.g., metformin) and beneficial clinical response of PD or PD-R. Causal methods, such as Mendelian randomization studies[79,80], should be utilized in the future. Last, we didn't examine the safety of the predicted drug candidates and it must be rigorously evaluated through dedicated studies.

Methods

Study cohorts for PD subtyping

The present study included two longitudinal PD cohorts for identifying PD subtypes: the Parkinson's Progression Markers Initiative (PPMI, http://www.ppmi-info.org)[25] and the Parkinson Disease Biomarkers Program (PDBP, https://pdbp.ninds.nih.gov)[26].

PPMI, launched in 2010 and sponsored by the Michael J. Fox Foundation, is an international and multi-center observational study dedicated to identifying biomarkers indicative of PD progression[25]. The present study included the following participants in the PPMI cohorts: de novo PD participants (diagnosed with PD within the last 2 years and untreated at enrollment), HCs, and individuals with SWEDD (scans without evidence of dopaminergic deficit). More information about PPMI participants has been described elsewhere[25]. The PPMI study protocol was approved by the institutional review board of the University of Rochester (NY, USA), as well as from each PPMI participating site. Data used in the preparation of this article were obtained on Jun 25, 2020 from the PPMI database (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431. For up-to-date information on the study, visit www.ppmi-info.org. This analysis used data openly available from PPMI as well as whole exome sequencing data, whole blood RNA-seq data, and high resolution T1-weighted 3 T Magnetic Resonance Imaging (MRI) data, obtained from PPMI upon request after approval by the PPMI Data Access Committee.

PDBP, established in 2012 and funded by the National Institute of Neurological Disorders and Stroke (NINDS), is an observational study for advancing comprehensive PD biomarker research[26]. The present study included the participants with PD and HCs in the PDBP cohort. More information about the PDBP participants has been described elsewhere[26]. The study protocol for each PDBP site was approved by institutional review board of each participating site. Data of the PDBP cohort were obtained on Jan 26, 2021 via the Accelerating Medicines Partnership Parkinson's disease (AMP-PD) Knowledge Platform (http://amp-pd.org) under AMP-PD Data Use Agreement.

We used the PPMI cohort as the development cohort. Participants who had less than 1-year historical records were excluded as lack of longitudinal information. Participants' longitudinal clinical data were collected for modeling PD symptom progression trajectories to produce a progression embedding vector for each participant, using deep learning. Subtypes were identified using the learned progression embedding vectors of participants with PD. We used the PDBP cohort as the validation cohort. Participants who had less than 1-year historical records were excluded. In the PDBP, only the early PDs (symptom duration < 3 years at enrollment) were used to re-identify the PD subtypes. More details of the studied cohorts and data utilization were illustrated in the Supplementary Table 1. All study participants provided written informed consent for their participation in both studies.

PD progression modeling for subtype identification

Clinical variables. We used participant's longitudinal data in diverse clinical assessments. Specifically, we used motor manifestation data including Movement Disorders Society–revised Unified Parkinson's Disease Rating Scale (MDS-UPDRS) Parts II and III[81] and Schwab-England activities of daily living score, as well as non-motor manifestation data including MDS-UPDRS Part I[81], Scales for Outcomes in Parkinson's disease-Autonomic (SCOPA-AUT)[82], Geriatric depression scale (GDS)[83], Questionnaire for Impulsive-Compulsive Disorders in Parkinson's disease (QUIP)[84], State-Trait Anxiety Inventory (STAI)[85], Benton Judgment of Line Orientation (JOLO)[86], Hopkins Verbal Learning Test (HVLT)[87], Letter-number sequencing (LNS), Montreal Cognitive Assessment (MoCA)[88], Semantic verbal-language fluency test[89], Symbol–Digit Matching (SDM)[90], Epworth Sleepiness Score (ESS)[91], REM sleep behavior disorder (RBD)[92], and Cranial Nerve Examination. Usage of dopaminergic medication was transformed into levodopa equivalent daily dose usage[93].

For the PDBP cohort, we used the clinical variables shared with the PPMI cohort. More details of the clinical variables used for PD subtyping can be found in the Supplementary Table 2.

Data preparation. In both the PPMI and PDBP cohorts, we extracted clinical data at baseline and follow-up visits for each participant. In this way, each participant was associated with a multivariate clinical time sequence. Missing values were imputed using the combination of the last observation carried forward (LOCF) and next observation carried backward (NOCB) strategies which have demonstrated effectiveness in our previous work[78]. In a last observation carried forward (LOCF) procedure, a missing follow-up visit value is replaced by (imputed as) that subject's previously observed value; similarly, in a next observation carried backward (NOCB) procedure, a missing previous visit value is replaced by (imputed as) that subject's follow-up observed value. The variables missing all values along the timeline of a participant were imputed by median values of the population. To eliminate the effects of value magnitude, all variables were scaled based on z-score, i.e., $\hat{x} = (x - \mu)/\sigma$, where, $\hat{x}$ is the z-scored value, x is the original value, and μ and σ are the mean value and standard deviation of the data.

Learning individuals' phenotypic progression embedding vectors using deep learning. Our goal was to identify PD subtypes, each of which can reflect a unique PD symptom progression pattern within the course of PD. To this end, there is the need of fully considering longitudinal data of individuals to derive PD subtypes. Here, we developed a deep learning model, termed deep phenotypic progression embedding (DPPE), which took the multivariate clinical time sequence data of each participant as input to learn a machine-readable vector representation, encoding his/her PD phenotypic progression trajectory over time (see Fig. 1g). Specifically, the DPPE model was based on the Long-Short Term Memory (LSTM) units[28,29], a deep learning model designed for time sequence data modeling. The LSTM unit has a "memory cell" that stores historical information for extended periods, making it an excellent choice for modeling disease progression using longitudinal clinical data[94]. The DPPE engaged an autoencoder architecture that consisted of two components, an encoder and a decoder, each of which is a LSTM model: (1) the encoder took the longitudinal clinical data of each individual (i.e., multivariate time sequence) as input and learned a progression embedding vector that encoded his/her PD symptom progression trajectory; and (2) the decoder, with reversed architecture of the encoder, tried to reconstruct the input time sequence of each individual based on the embedding vector. The DPPE model was trained by minimizing difference between the input multivariate clinical time sequences and the reconstructed ones. To enhance model training, we used data of PDs, HCs, and SWEDDs in the PPMI cohorts. After training of the DPPE model, we obtained a learned progression embedding vector for each participant, encoding his/her PD progression profile.

We deployed DPPE using PyTorch (https://pytorch.org) with Python 3.6. Specifically, we used one-layer LSMT in both the encoder and decoder in DPPE. We set the embedding size as 16. In model training, to take full advantage of patient data, we set batch size as 1. We trained the model using the "Adam" optimizer in PyTorch.

Cluster analysis for subtype identification in the PPMI cohort. The agglomerative hierarchical clustering (AHC) with Euclidean distance calculation and Ward linkage criterion[30] was applied to the individuals' embedding vectors learned by the DPPE model. We used AHC because that, (1) unlike other clustering methods like k-means clustering (which requires a sphere-like distribution of the data), AHC is usually robust as it's not sensitive to the distribution of the data and doesn't require an initialization procedure (e.g., k-means) that may incorporate uncertainty; (2) AHC can produce a tree diagram known as a dendrogram, visually interpreting how the data points are agglomerated together in a hierarchical manner and also illustrating the distances between the clusters at different layers in the hierarchy, providing visible guidance in determining the optimal cluster number. AHC has previously shown promise in identifying underlying patterns from clinical profiles for disease subtyping[95–97]. We implemented AHC using scikit-learn (https://scikit-learn.org/stable/) with Python 3.6.

A crucial issue of cluster analysis is the determination of cluster number in data. To address this issue, we considered multiple criteria to determine the optimal cluster (i.e., subtype) number in cluster analysis based on AHC: (1) Clusters should be clearly separated in the dendrogram produced by the AHC. (2) We selected optimal cluster numbers suggested based on clustering measurements calculated by the 'NbClust' software[31], an R package for assisting a clustering method to determine the optimal cluster number of the data. Here, we used 18 indices provided by 'NbClust' to evaluate cluster structure of the agglomerative hierarchical clustering model with Ward criterion. The optimal cluster number was determined by the optimal value of each index. The used indices included: Scott index, Marriot index, TrCovW index, TraceW index, Friedman index, Rubin index, DB index, Silhouette index, Duda index, Pseudot2 index, Beale index, Ratkowsky index, Ball index, Ptbiserial index, Frey index, McClain index, Dunn index, SDindex. More details of these indices were introduced elsewhere[31]. (3) We calculated the 2-dimensional (2D) representation for each individual based on his/her progression embedding vector using the t-distributed stochastic neighbor embedding (t-SNE) algorithm[32]. We then visualized individuals' subtype memberships in the 2D t-SNE space. We expected that the clusters, i.e., PD subtypes, could be clearly separated in the 2D t-SNE space. (4) We also considered clinical interpretations of the subtype results to determine the optimal cluster number.

Subtype validation in the PDBP cohort. To enhance reproducibility of the identified subtypes, we validated them using the PDBP cohort. Specifically, we repeated the whole procedure above in the PPMI cohort to re-identify the subtypes in the PDBP cohort. We re-trained the DPPE model to calculate individuals' progression embedding vectors. We used data of HCs and all participants with PD in the PDBP cohort for training the DPPE model. PD participants in the PDBP cohort had a broad distribution of PD duration[26]. To align with our primary analysis in the PPMI cohort, we used the individuals with early PD (whose PD duration < 3 years) and performed AHC based on their learned embedding vectors to re-derive subtypes.

Exploring clinical characteristics of the identified subtypes

We characterized the identified subtypes in two ways. First, we characterized the subtypes by evaluating their differences in demographics and clinical assessments at baseline as well as 2- and 5-year follow-up. For group comparisons, we performed analysis of variance (ANOVA) for continuous data and χ^2 test for categorical data. Analysis of covariance (ANCOVA) was also applied, adjusting for age, sex, and levodopa equivalent daily dose (LEDD) usage. A two-tailed P value < 0.05 were considered as the threshold for statistical significance.

Second, we estimated annual progression rates in terms of each clinical assessment for each subtype. To accomplish this, for each variable, we fitted a linear mixed effect model for each PD subtype, specifying time (year) as the explanatory variable of interest. For all models, individual variation was included as a random effect, and age, sex, and LEDD at visits were included as the covariates. For each model, we reported the coefficient β (95% CI) of time as annual progression rate of the specific clinical assessment, along with the corresponding P value. A P value < 0.05 was considered for statistical significance. We further utilized Sankey diagrams to visualize the transition trends of motor phenotypes (tremor dominant and postural instability and gait disorder [PIGD]) as well as non-motor phenotypes including cognition (normal cognition, mild cognition impairment [MCI], and dementia), REM sleep behavior disorder (RBD negative or positive), and depression (normal, as well as mild, moderate, and severe depression).

For above statistical analyses, multiple correction was conducted by controlling false discovery rate (FDR).

The statistical analyses were conducted using R 4.3. Linear mixed effect models were built using lme4 (https://github.com/lme4/lme4/) in R.

CSF biomarker analysis

CSF biospecimen data were obtained from the PPMI cohort. We used baseline α-synuclein measured by enzyme-linked immunosorbent assay[98], amyloid-beta1–42 ($A\beta$-42), phosphorylated Tau protein at threonine 181 (P-tau), and total tau protein (T-tau) measured by INNO-BIA AlzBio3 immunoassay. Following the previous studies[99,100], we also evaluated $A\beta$-42/P-tau, $A\beta$-42/T-tau, $A\beta$-42/α-synuclein, P-tau/α-synuclein, T-tau/α-synuclein, and P-tau/T-tau levels. We conducted two-group comparisons (subtype vs. subtype and subtype vs. HCs) for each biomarker using linear mixed effect models, specifying individuals' HC or PD subtype membership as the explanatory variable of interest, adjusting for baseline age and sex as covariates. A P value < 0.05 was considered for statistical significance. Additionally, boxplots were engaged to visualize data distributions.

Neuroimaging biomarker analysis

High resolution T1-weighted 3 T MRI data of participants were available at baseline and follow-up in the PPMI cohort. For each individual, we calculated 1-year brain atrophy measured by cortical thickness and white matter volume in 34 brain regions of interests (ROIs), defined by the Desikan-Killiany atlas (averaged over the left and right hemispheres). We evaluated those measures in separating each pair of PD subtypes. The student's t-tests were used for two-group comparisons. The 'ggseg'[101] in R was used to visualize the neuroimaging biomarkers under the Desikan-Killiany atlas.

Construction of human protein-protein interactome network

We assembled commonly used human protein-protein interactome (PPI) databases with experimental evidence and in-house systematic human PPIs to build a comprehensive human PPI network. PPI databases we used included: (i) kinase-substrate interactions via literature-derived low-throughput and high-throughput experiments from Human Protein Resource Database (HPRD)[102], Phospho.ELM[103], KinomeNetworkX[104], PhosphoNetworks[105], PhosphositePlus[106], and DbPTM 3.0[107]; (ii) binary PPIs from 3D protein structures from Instruct[108]; (iii) binary PPIs assessed by high-throughput yeast-two-hybrid (Y2H) experiments[109]; (iv) protein complex data (~56,000 candidate interactions) identified by a robust affinity purification-mass spectrometry collected from BioPlex V2.0[110]; (v) signaling networks by literature-derived low-throughput experiments from the SignaLink2.0[111]; and (vi) literature-curated PPIs identified by affinity purification followed by mass spectrometry from BioGRID[112], HPRD[113], InnateDB[113], IntAct[114], MINT[115], and PINA[116]. In total, 351,444 PPIs connecting 17,706 unique proteins are now freely available at https://alzgps.lerner.ccf.org[117]. In this study, we utilized the largest connected component of this dataset, including 17,456 proteins and 336,549 PPIs. The PPI network was used for network analyses of genetic and transcriptomic data for subtype-specific molecular module identification, which were detailed below.

Genetic data analysis

To explore genetic components of the PD subtypes, we analyzed genetic data in the PPMI cohort. APOE ε2 and ε4 genotypes and genotypes of 90 PD-related risk loci[33] were collected for analysis. The hypergeometric tests were used to identify SNPs that were enriched in each subtype within the PD population. P value from the hypergeometric analysis indicates association of a SNP to the specific subtype[34]. A P value < 0.05 was considered for statistical significance.

Single nucleus RNA-seq analysis for identifying PD contextual genes

We utilized one set of human brain single nucleus RNA-sequencing data collected from 12 control donors with two brain regions: cortex and substantia nigra (SN) which included approximately 17,000 nuclei. It is available from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) database with accession number GSE140231. We performed the bioinformatics analyses according to the processes described in the original manuscript[118]. Each brain region was analyzed individually, and the whole analyses were implemented with Seurat (4.0.6)[119]. Nuclei expressed with ≤ 500 genes, with ≥ 5% mitochondrial genes and ≥ 5% ribosomal genes were removed. Then the raw count was log-normalized and the top 2000 most variable genes were detected by function *FindVariableFeatures* with *selection.method = 'vst'*. Next, all samples were integrated by functions *FindIntegrationAnchors* using canonical correlation analysis (CCA) and IntegrateData with *dims = 1:42* and *1:25* for cortex and substantia nigra, separately. We then scaled the data and regressed out heterogeneity related with ribosomal, mitochondrial content, and number of UMIs. Clustering was performed with resolution 0.4 and 0.6 for cortex and substantia nigra, separately. We identified dopaminergic neuron with marker genes (*TH*, *LMX1B*, *KCNJ6*, *NR4A2* and *SLC6A3*) provided by the original manuscript[118] for substantia nigra only. DEGs for dopaminergic neuron were calculated against other cell types with MAST R package[120] regarding brain region substantia nigra, and were considered as PD contextual genes.

Construction of genetic molecular modules of the identified subtypes

Network analysis was conducted based on the PPI network we built to expand the genetic signals to identify subtype-specific molecular module. For each subtype, we first selected the SNPs that were associated with the subtype (P value < 0.05) and obtained their nearest genes and/or possible causal genes identified through the PD GWAS Locus Browser[121]. The PD GWAS Locus Browser identified causal genes for each PD risk loci, as protein coding genes within 1 Mb up and downstream of it, via integration of diverse data resources, such as gene expression and expression quantitative trait locus (eQTL) data from different tissues as well as literature. This resulted in a list of genetic associated genes for each subtype. Then, for each subtype, we linked its genetic associated genes with the PD contextual genes in the PPI network we built. Genetic associated genes that cannot link to any contextual genes were removed. In this way, these linked genetic associated genes and their linked contextual genes constructed the genetic molecular module specific to the subtype.

Transcriptomic data analysis

We performed gene expression analysis using whole blood bulk RNA-seq data of participants within the PPMI cohort[122]. Genes were annotated using UniProt[123], and only protein coding genes were included for analysis. Genes with low expression levels across all samples were excluded. Differential gene expression analyses were performed for each subtype compared to healthy controls using DEseq2[124] in R. Age, sex, and LEDD usage were included as covariates. An adjusted P value < 0.05 was considered for statistical significance.

Construction of transcriptomic molecular modules of the identified subtypes using GPSnet algorithm

Our GPSnet (Genome-wide Positioning Systems network)[38] demanded two inputs: the node (gene) scores and a background PPI network we built above. GPSnet first set an initial gene score $z(i)$ for each gene i in the PPI network: for differentially expressed genes (DEGs) with Q value ≤ 0.05, the gene scores were initialized as $z(i) = |\log_2 \mathrm{FC}|$, where FC indicates fold change; for the remaining non-DEGs, $z(i) = 0$. After that, a random walk with restart process was applied to smoothen the scores for all genes in the PPI network. Next, GPSnet utilized the following procedures to build a gene module M: It first initialized module M only containing a randomly selected seed gene and then expanded the module by involving genes one by one, while (1) enhancing gene connectivity within the module and (2) improving module level gene score. Specifically, a gene $i \in \Gamma_M$ will be included into M (Γ_M is a set of genes that interact with genes in module M), if (1)

$P(i) \leq 0.01$ (calculated with Eq. (1), indicating genes within M densely connect to each other after including i) and (2) $S(M \cup \{i\}) > S(M)$ (calculated with Eq. (2), indicating module-level risk score increased after involving gene i).

$$P(i) = \sum_{d=d_n}^{d_i} \frac{\binom{m}{d}\binom{N-m}{d_i-d}}{\binom{N}{d_i}} \quad (1)$$

$$S(M) = \frac{\sum_{j \in M}(z(j) - \omega)}{\sqrt{n}} \quad (2)$$

where, m denotes the number of genes in module M, ω denotes the average score of all genes in the PPI network, and $S(M)$ denotes the module-level gene score of the module M.

We stopped module expansion when $S(M)$ could or be increased or $P(i) \geq 0.01$ by involving new genes. In this study, we repeated above procedures multiple (~100,000) times and obtained a collection of raw modules. All raw modules were ranked in a descending order based on module score. We generated the final gene modules by assembling the top ranked modules. For each subtype, we ran GPSnet based on the DEGs of a specific subtype to gain the transcriptomic molecular module of the subtype.

Pathway enrichment analysis

Pathway enrichment analyses were conducted based on using BioPlanet[125] from Enrichr[126]. For each subtype, we identified pathways according to the genetic and transcriptomic molecular modules of the subtype, respectively. The combined score defined in Enrichr[126] equaled the product of log of P value from the Fisher's test and z-score which characterized the departure from the expected rank.

Building classification model of subtypes

To gain more prognostic insights of the subtypes, we built a prognostic model of subtypes using information collected within the first year after enrollment. Specifically, we used candidate predictors including demographics, 10 principal components derived from neuroimaging biomarkers (1-year brain atrophy measured by reduction of cortical thickness and white matter volume in 34 brain ROIs), 90 PD-related SNPs, and clinical variables at baseline and 1-year follow-up. To improve model practicability in separating multiple subtypes, we built the model using a cascade framework[44], which was proposed to address multi-label classification tasks. Specifically, it contained a sequential of multiple basic classifiers, such that in each step, it predicted a subtype from the remaining subjects that will be sent to the successor basic classifier. We used random forest as the basic classifier and the model was trained using a fivefold cross-validation strategy. We used the receiver operating characteristics curve (ROC) and area under ROC curve (AUC) to measure prediction performance of our model.

Subtype-specific in silico drug repurposing

Connectivity Map (CMap) database. We downloaded transcriptomics-based drug-gene signature data in human cell lines from the Connectivity Map (Cmap) database[24]. The Cmap data used in this study contained 6100 expression profiles relating 1309 compounds[24]. The CMap provided a measure of the extent of differential expression for a given probe set. The amplitude α was defined as Eq. (3) as follows:

$$\alpha = \frac{t - c}{(t + c)/2} \quad (3)$$

where t is the scaled and threshold average difference value for the drug treatment group and c is the threshold average difference value for the control group. Therefore, an $\alpha = 0$ indicates no expression change after drug treatment, while an $\alpha > 0$ indicates elevated expression level after drug treatment and vice versa.

Gene set enrichment analysis (GSEA) for drug repurposing. We applied GSEA algorithm for predicting repurposable drug candidates for each PD subtype. The GSEA algorithm demanded two inputs: the CMap data and a list of module genes. For each subtype, we combined the genetic and transcriptomic molecular module to obtain a list of module genes for this subtype. Detailed descriptions of GSEA have been illustrated in elsewhere[38,46]. Then, for each drug in the CMap database, we calculated an enrichment score (ES) based on genes within each subtype-specific molecular module as Eq. (4). ES represents drug potential capability to reverse the expression of the input molecular network:

$$ES = \begin{cases} ES_{up} - ES_{down}, \mathrm{sgn}(\mathrm{ES_{up}}) \neq \mathrm{sgn}(\mathrm{ES_{down}}) \\ 0, else \end{cases} \quad (4)$$

where ES_{up} and ES_{down} were calculated separately for up- and down-regulated genes from the subtype-specific gene module. We computed $a_{up/down}$ and $b_{up/down}$ as

$$a = \max_{1 \leq j \leq s}\left(\frac{j}{s} - \frac{V(j)}{r}\right) \quad (5)$$

$$b = \max_{1 \leq j \leq s}\left(\frac{V(j)}{r} - \frac{j-1}{s}\right) \quad (6)$$

where $j = 1, 2, 3, \ldots, s$ are the genes within the subtype-specific module sorted in an ascending order by their rank in the gene expression profiles of the tested drug. The rank of gene j was defined as $V(j)$, such that $0 \leq V(j) \leq r$ with r being the number of module genes in drug profile. Then $ES_{up/down}$ was set to $a_{up/down}$ if $a_{up/down} > b_{up/down}$ and was set to $-b_{up/down}$ if $a_{up/down} < b_{up/down}$. Permutation tests repeated 100 times using randomly selected gene lists consisting of the same numbers of up- and down-regulated genes as the input subtype-specific gene module were performed to calculate the significance of the computed ES value. Therefore, drugs with large positive ES values and P values ≤ 0.05 were selected.

Pharmacoepidemiologic validation with large-scale real-world patient data

We estimated treatment effects of the identified repurposable drug candidates. Overall workflow of the real-world patient data analysis can be found in the Supplementary Fig. 12. The detailed procedures were introduced as below.

Real-world patient data. In this study, we used two independent large-scale real-world patient databases.

- *INSIGHT Clinical Research Network (CRN).* The INSIGHT CRN[51] was founded with and continues to be supported by Patient-Centered Outcomes Research Institute (PCORI). The INSIGHT CRN brings together top academic medical centers located in New York City (NYC), including Albert Einstein School of Medicine/Montefiore Medical Center, Columbia University and Weill Cornell Medicine/New York-Presbyterian Hospital, Icahn School of Medicine/Mount Sinai Health System, and New York University School of Medicine/Langone Medical Center. This study used de-identified patient real-world data (RWD) from the INSIGHT CRN, which contained longitudinal clinical data of over 15 million patients in the NYC metropolitan area. The use of the INSIGHT data was approved by the Institutional Review Board (IRB) of Weill Cornell Medicine under protocol 21-07023759.
- *OneFlorida+ Clinical Research Consortium.* OneFlorida+ Clinical Research Consortium[52] is another CRN supported by PCORI, which includes 12 healthcare organizations and contains longitudinal and

linked patient-level data. This study used de-identified, robust linked patient-level RWD of 17 million patients in Florida, 2.1 million in Georgia (via Emory), and 1.1 million in Alabama (via UAB Medicine) since 2012 and covering a wide range of patient characteristics including demographics, diagnoses, medications, procedures, vital signs, and lab tests. The use of the OneFlorida+ approved by the IRB of University of Florida under protocol IRB202300639.

Eligibility criteria. Patient eligibility criteria for analysis included:

- Patients should have at least one PD diagnosis according to International Classification of Diseases 9th and 10th revision (ICD-9/10) codes, including 332.0 (ICD-9) and G20 (ICD-10).
- Patient's age was > = 50 years old at the first PD diagnosis.
- Patients who had neurodegenerative disease diagnoses before his/her first PD diagnosis was excluded.

PD outcomes. We considered PD related outcomes including dementia and falls (indicating advanced motor impairment and dyskinesia), which were defined by ICD-9/10 diagnosis codes. Drug treatment efficiency was defined by reducing the risk to develop the PD outcomes.

Follow-up. Each patient was followed from his/her baseline until the day of the first PD outcome event, or loss to follow-up (censoring), whichever happened first.

Trial emulation. We first obtained DrugBank ID of each tested drug and translated it to RxNorm and NDC codes using the RxNav API (application programming interface). Drugs which were used by less than 100 patients were excluded for analysis. Following Ozery-Flato et al.[56], we defined the PD initiation date of each patient as six months prior to his/her first recorded PD diagnosis event. This accounted for the likelihood that PD may be latently present before formal diagnosis. We defined the index date as the beginning time of treatment of a tested drug candidate or its alternative treatment. We also defined the baseline period as the time interval between the PD initiation date and the index date for each patient. We required that:

- The index date was later than the PD initiation date.
- Onset of PD outcomes were later than the index date.

Then, for each tested drug, we built an emulated trial using the following procedures:

1. We built its treated group as the eligible PD patients who took the tested drug after PD initiation;
2. We built a control group as:
    1. Patients who received alternative treatment of the tested drug, i.e., drugs from a same Anatomical Therapeutic Chemical level 2 [ATC-L2] classification of the tested drug, excluding the drug itself.
    2. To control confounding factors, we performed the propensity score matching as introduced below.

Propensity score matching (PSM). We collected three types of covariates at the baseline period for each patient: (1) We included 64 comorbidities including comorbidities from Chronic Conditions Data Warehouse and other risk factors that were selected by experts[53]. The comorbidities were defined by a set of ICD-9/10 codes. (2) We considered usage of 200 most prevalent prescribed drug ingredients as covariates in this analysis. These drugs were coded using RxNorm and grouped into major active ingredients using Unified Medical Language System. (3) We also included other covariates, including age, gender, race, and the time from the PD initiation date to the drug index date. In total, we included 267 covariates for analysis.

For each emulated trial, we used a propensity score framework to learn the empirical treatment assignment given the baseline covariates and used an inverse probability of treatment weighting to balance the treated and control groups[54]. For each trial, a 1:1 nearest-neighbor matching was performed to build the matched control group[54]. The covariate balance after propensity score matching was assessed using the absolute standardized mean difference (SMD)[127]. For each covariate, it was considered balanced if its SMD≤0.2, and the treated and control group were balanced if only no more than 2% covariates were not balanced[128]. To enhance robustness of the analysis, we created 100 emulated trials for each tested drug. Tested drugs that had <10 successfully balanced trials were excluded for analysis.

Treatment effect estimation. For each tested drug, we estimated drug treatment effect for each balanced trial by calculating the hazard ratio (HR) using a Cox proportional hazard model[55], comparing the risk to develop a specific outcome between the treated and control groups. We reported the median HR with 95% confidence intervals (CI) obtained by bootstrapping[129]. A HR < 1 indicated the tested drug can reduce risk to develop a specific outcome and a P value < 0.05 was considered as statistically significant.

The trial emulation pipeline for treatment effect estimation was implemented using Python packages psmpy[130] for propensity score framework and lifelines[131] for the Cox proportional hazard model.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This study used data from PPMI (www.ppmi-info.org/access-data-specimens/download-data, RRID:SCR 006431) and PDBP (https://pdbp.ninds.nih.gov). Data from the PDBP were download from AMP-PD Knowledge Platform (http://amp-pd.org). Information of the INSIGHT database is available at https://insightcrn.org. Request of INSIGHT data can be sent via: https://nyc-cdrn.atlassian.net/servicedesk/customer/portal/2/group/6/create/16. Information of the OneFlorida+ data is available at: https://onefloridaconsortium.org/. Request of OneFlorida+ data can be sent via: https://onefloridaconsortium.org/front-door/prep-to-research-data-query/.

Code availability

Computer codes for this study are available at https://github.com/changsu10/Parkinson-Progression-Subtyping/tree/main.

Received: 24 December 2023; Accepted: 21 June 2024;
Published online: 09 July 2024

References

1. Poewe, W. et al. Parkinson disease. *Nat. Rev. Dis. Prim.* **3**, 17013 (2017).
2. Balestrino, R. & Schapira, A. H. V. Parkinson disease. *Eur. J. Neurol.* **27**, 27–42 (2020).
3. Mizuno, Y. Where do we stand in the treatment of Parkinson's disease? *J. Neurol.* **254**, 13–18 (2007).
4. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet* **397**, 2284–2303 (2021).
5. Weiner, W. J. There is no Parkinson disease. *Arch. Neurol.* **65**, 705–708 (2008).
6. Farrow, S. L., Cooper, A. A. & O'Sullivan, J. M. Redefining the hypotheses driving Parkinson's diseases research. *npj Parkinson's Dis.* **8**, 45 (2022).
7. Lang, A. E. & Espay, A. J. Disease modification in Parkinson's disease: Current approaches, challenges, and future considerations. *Mov. Disord.* **33**, 660–677 (2018).
8. Espay, A. J., Brundin, P. & Lang, A. E. Precision medicine for disease modification in Parkinson disease. *Nat. Rev. Neurol.* **13**, 119–126 (2017).
9. Sieber, B.-A. et al. Prioritized research recommendations from the national institute of neurological disorders and stroke Parkinson's disease 2014 conference. *Ann. Neurol.* **76**, 469–472 (2014).

10. Fereshtehnejad, S.-M. & Postuma, R. B. Subtypes of Parkinson's disease: what do they tell us about disease progression? *Curr. Neurol. Neurosci. Rep.* **17**, 34 (2017).
11. van Rooden, S. M. et al. The identification of Parkinson's disease subtypes using cluster analysis: A systematic review. *Mov. Disord.* **25**, 969–978 (2010).
12. Simuni, T. et al. How stable are Parkinson's disease subtypes in de novo patients: Analysis of the PPMI cohort? *Parkinsonism Relat. Disord.* **28**, 62–67 (2016).
13. Erro, R. et al. Comparing postural instability and gait disorder and akinetic-rigid subtyping of Parkinson disease and their stability over time. *Eur. J. Neurol.* **26**, 1212–1218 (2019).
14. Eisinger, R. S. et al. Motor subtype changes in early Parkinson's disease. *Parkinsonism Relat. Disord.* **43**, 67–72 (2017).
15. Mestre, T. A. et al. Parkinson's disease subtypes: Critical appraisal and recommendations. *J. Parkinson's Dis.* **11**, 395–404 (2021).
16. Dexter, D. T. & Jenner, P. Parkinson disease: From pathology to molecular disease mechanisms. *Free Radic. Biol. Med.* **62**, 132–144 (2013).
17. Dickson, D. W. Neuropathology of Parkinson disease. *Parkinsonism Relat. Disord.* **46**, S30–S33 (2018).
18. Siderowf, A. et al. Assessment of heterogeneity among participants in the Parkinson's Progression Markers Initiative cohort using α-synuclein seed amplification: a cross-sectional study. *Lancet Neurol.* **22**, 407–417 (2023).
19. Trinh, J. & Farrer, M. Advances in the genetics of Parkinson disease. *Nat. Rev. Neurol.* **9**, 445–454 (2013).
20. Blauwendraat, C., Nalls, M. A. & Singleton, A. B. The genetic architecture of Parkinson's disease. *Lancet Neurol.* **19**, 170–178 (2020).
21. Ron, S. et al. Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676 (2017).
22. Tan, M. M. X. et al. Genome-wide association studies of cognitive and motor progression in Parkinson's disease. *Mov. Disord.* **36**, 424–433 (2021).
23. Hirotaka, I. et al. Genetic risk of Parkinson disease and progression. *Neurol. Genet.* **5**, e348 (2019).
24. Lamb, J. et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
25. Marek, K. et al. The Parkinson progression marker initiative (PPMI). *Prog. Neurobiol.* **95**, 629–635 (2011).
26. Rosenthal, L. S. et al. The NINDS Parkinson's disease biomarkers program. *Mov. Disord.* **31**, 915–923 (2016).
27. Zhang, X. et al. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: A cohort study. *Sci. Rep.* **9**, 797 (2019).
28. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
29. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
30. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *J. Classification* **31**, 274–295 (2014).
31. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R package for determining the relevant number of clusters in a data set. *2014* **61**, 36 (2014).
32. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 11 (2008).
33. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
34. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
35. Przanowski, P. et al. The signal transducers Stat1 and Stat3 and their novel target Jmjd3 drive the expression of inflammatory genes in microglia. *J. Mol. Med.* **92**, 239–254 (2014).
36. Vavougios, G. D., Breza, M., Mavridis, T. & Krogfelt, K. A. FYN, SARS-CoV-2, and IFITM3 in the neurobiology of Alzheimer's disease. *Brain Disord.* **3**, 100022 (2021).
37. Saunders-Pullman, R. et al. Progression in the LRRK2-associated Parkinson disease population. *JAMA Neurol.* **75**, 312–319 (2018).
38. Cheng, F. et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat. Commun.* **10**, 3476 (2019).
39. Jiang, S. X., Sheldrick, M., Desbois, A., Slinn, J. & Hou, S. T. Neuropilin-1 is a direct target of the transcription factor E2F1 during cerebral ischemia-induced neuronal death in vivo. *Mol. Cell. Biol.* **27**, 1696–1705 (2007).
40. Smith, R. A., Walker, T., Xie, X. & Hou, S. T. Involvement of the transcription factor E2F1/Rb in kainic acid-induced death of murine cerebellar granule cells. *Mol. Brain Res.* **116**, 70–79 (2003).
41. Christine, S. et al. Plasma ApoA1 associates with age at onset and motor severity in early Parkinson disease patients (P6.068). *Neurology* **84**, P6.068 (2015).
42. Tofaris, G. K. et al. Ubiquitin ligase Nedd4 promotes α-synuclein degradation by the endosomal–lysosomal pathway. *Proc. Natl. Acad. Sci.* **108**, 17004–17009 (2011).
43. Scherzer, C. R. et al. GATA transcription factors directly regulate the Parkinson's disease-linked gene α-synuclein. *Proc. Natl. Acad. Sci.* **105**, 10907–10912 (2008).
44. Gama, J. & Brazdil, P. Cascade generalization. *Mach. Learn.* **41**, 315–343 (2000).
45. Fang, J. et al. Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease. *Nat. Aging* **1**, 1175–1188 (2021).
46. Zhou, Y. et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* **6**, 14 (2020).
47. Mullin, S. et al. Ambroxol for the treatment of patients with Parkinson disease with and without glucocerebrosidase gene mutations: A nonrandomized, noncontrolled trial. *JAMA Neurol.* **77**, 427–434 (2020).
48. Silveira, C. R. A. et al. Ambroxol as a novel disease-modifying treatment for Parkinson's disease dementia: Protocol for a single-centre, randomized, double-blind, placebo-controlled trial. *BMC Neurol.* **19**, 20 (2019).
49. Sun, X. et al. Guanabenz promotes neuronal survival via enhancement of ATF4 and parkin expression in models of Parkinson disease. *Exp. Neurol.* **303**, 95–107 (2018).
50. Sherman, R. E. et al. Real-world evidence — What is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
51. Kaushal, R. et al. Changing the research landscape: The New York City clinical data research network. *J. Am. Med. Inform. Assoc.* **21**, 587–590 (2014).
52. Shenkman, E. et al. OneFlorida clinical research consortium: Linking a clinical and translational science institute with a community-based distributive medical education model. *Acad. Med*. **93**, 451–455 (2018).
53. Zang, C. et al. High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. *Nat. Commun.* **14**, 8180 (2023).
54. Allan, V. et al. Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *J. Comp. Effectiveness Res.* **9**, 603–614 (2020).
55. Lin, D. Y. & Wei, L. J. The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.* **84**, 1074–1078 (1989).

56. Ozery-Flato, M., Goldschmidt, Y., Shaham, O., Ravid, S. & Yanover, C. Framework for identifying drug repurposing candidates from observational healthcare data. *JAMIA Open* **3**, 536–544 (2020).
57. Nutt, J. G. Motor subtype in Parkinson's disease: Different disorders or different stages of disease? *Mov. Disord.* **31**, 957–961 (2016).
58. Fereshtehnejad, S.-M., Zeighami, Y., Dagher, A. & Postuma, R. B. Clinical criteria for subtyping Parkinson's disease: biomarkers and longitudinal progression. *Brain* **140**, 1959–1976 (2017).
59. Parnetti, L. et al. CSF and blood biomarkers for Parkinson's disease. *Lancet Neurol.* **18**, 573–586 (2019).
60. Hirsch, E. C. & Hunot, S. Neuroinflammation in Parkinson's disease: A target for neuroprotection? *Lancet Neurol.* **8**, 382–397 (2009).
61. Picca, A. et al. Mitochondrial dysfunction, oxidative stress, and neuroinflammation: Intertwined roads to neurodegeneration. *Antioxidants* **9**, 647 (2020).
62. Long, H.-Z. et al. PI3K/AKT signal pathway: a target of natural products in the prevention and treatment of Alzheimer's disease and Parkinson's disease. *Front. Pharmacol.* **12**, 648636 (2021).
63. Heras-Sandoval, D., Pérez-Rojas, J. M., Hernández-Damián, J. & Pedraza-Chaverri, J. The role of PI3K/AKT/mTOR pathway in the modulation of autophagy and the clearance of protein aggregates in neurodegeneration. *Cell. Signal.* **26**, 2694–2701 (2014).
64. Zacchigna, S., Lambrechts, D. & Carmeliet, P. Neurovascular signalling defects in neurodegeneration. *Nat. Rev. Neurosci.* **9**, 169–181 (2008).
65. Sweeney, M. D., Sagare, A. P. & Zlokovic, B. V. Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. *Nat. Rev. Neurol.* **14**, 133–150 (2018).
66. Cao, G. et al. Mechanism of metformin regulation in central nervous system: Progression and future perspectives. *Biomed. Pharmacother.* **156**, 113686 (2022).
67. Li, N., Zhou, T. & Fei, E. Actions of metformin in the brain: A new perspective of metformin treatments in related neurological disorders. *Int. J. Mol. Sci.* **23**, 8281 (2022).
68. Agostini, F., Masato, A., Bubacco, L. & Bisaglia, M. Metformin repurposing for Parkinson disease therapy: Opportunities and challenges. *Int. J. Mol. Sci.* **23**, 398 (2022).
69. Paudel, Y. N., Angelopoulou, E., Piperi, C., Shaikh, M. F. & Othman, I. Emerging neuroprotective effect of metformin in Parkinson's disease: A molecular crosstalk. *Pharmacol. Res.* **152**, 104593 (2020).
70. Katila, N. et al. Metformin lowers α-synuclein phosphorylation and upregulates neurotrophic factor in the MPTP mouse model of Parkinson's disease. *Neuropharmacology* **125**, 396–407 (2017).
71. Ismaiel, A. A. K. et al. Metformin, besides exhibiting strong in vivo anti-inflammatory properties, increases mptp-induced damage to the nigrostriatal dopaminergic system. *Toxicol. Appl. Pharmacol.* **298**, 19–30 (2016).
72. Tayara, K. et al. Divergent effects of metformin on an inflammatory model of Parkinson's disease. *Front. Cell. Neurosci.* **12**, 440 (2018).
73. Ge, X.-H. et al. Metformin protects the brain against ischemia/reperfusion injury through PI3K/Akt1/JNK3 signaling pathways in rats. *Physiol. Behav.* **170**, 115–123 (2017).
74. Ruan, C. et al. Neuroprotective effects of metformin on cerebral ischemia-reperfusion injury by regulating PI3K/Akt pathway. *Brain Behav.* **11**, e2335 (2021).
75. Liu, Y., Tang, G., Zhang, Z., Wang, Y. & Yang, G.-Y. Metformin promotes focal angiogenesis and neurogenesis in mice following middle cerebral artery occlusion. *Neurosci. Lett.* **579**, 46–51 (2014).
76. El-Ghaiesh, S. H. et al. Metformin protects from rotenone–induced nigrostriatal neuronal death in adult mice by activating AMPK-FOXO3 signaling and mitigation of angiogenesis. *Front. Mol. Neurosci.* **13**, 84 (2020).
77. Zhu, X. et al. Metformin improves cognition of aged mice by promoting cerebral angiogenesis and neurogenesis. *Aging (Albany NY)* **12**, 17845–17862 (2020).
78. Engels, J. M. & Diehr, P. Imputation of missing longitudinal data: a comparison of methods. *J. Clin. Epidemiol.* **56**, 968–976 (2003).
79. Andrews, S. J. & Goate, A. Mendelian randomization indicates that TNF is not causally associated with Alzheimer's disease. *Neurobiol. Aging* **84**, 241.e241–241.e243 (2019).
80. Williams, D. M., Finan, C., Schmidt, A. F., Burgess, S. & Hingorani, A. D. Lipid lowering and Alzheimer disease risk: A mendelian randomization study. *Ann. Neurol.* **87**, 30–39 (2020).
81. Goetz, C. G. et al. Movement disorder society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord.* **23**, 2129–2170 (2008).
82. Visser, M., Marinus, J., Stiggelbout, A. M. & Van Hilten, J. J. Assessment of autonomic dysfunction in Parkinson's disease: The SCOPA-AUT. *Mov. Disord.* **19**, 1306–1312 (2004).
83. Yesavage, J. A. & Sheikh, J. I. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. *Clin Gerontol.* **5**, 165–173 (1986).
84. Weintraub, D. et al. Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's disease. *Mov. Disord.* **24**, 1461–1467 (2009).
85. Spielberger, C. D. State-Trait Anxiety Inventory. In *The Corsini Encyclopedia of Psychology* 1-1.
86. Benton, A. L., Varney, N. R. & Hamsher, K. D. Visuospatial judgment: A clinical test. *Arch. Neurol.* **35**, 364–367 (1978).
87. Shapiro, A. M., Benedict, R. H. B., Schretlen, D. & Brandt, J. Construct and concurrent validity of the hopkins verbal learning test – revised. *Clin. Neuropsychologist* **13**, 348–358 (1999).
88. Nasreddine, Z. S. et al. The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *J. Am. Geriatrics Soc.* **53**, 695–699 (2005).
89. Gladsjo, J. A. et al. Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment* **6**, 147–178 (1999).
90. Smith, A. *Symbol digit modalities test.* (Western Psychological Services, Los Angeles, 1973).
91. Johns, M. W. A new method for measuring daytime sleepiness: The Epworth sleepiness scale. *Sleep* **14**, 540–545 (1991).
92. Stiasny-Kolster, K. et al. The REM sleep behavior disorder screening questionnaire—A new diagnostic instrument. *Mov. Disord.* **22**, 2386–2393 (2007).
93. Tomlinson, C. L. et al. Systematic review of levodopa dose equivalency reporting in Parkinson's disease. *Mov. Disord.* **25**, 2649–2653 (2010).
94. Ayala Solares, J. R. et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* **101**, 103337 (2020).
95. Brendel, M., Su, C., Hou, Y., Henchcliffe, C. & Wang, F. Comprehensive subtyping of Parkinson's disease patients with similarity fusion: a case study with BioFIND data. *npj Parkinson's Dis.* **7**, 83 (2021).
96. Su, C. et al. Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *npj Digital Med.* **4**, 110 (2021).
97. Su, C. et al. Identifying organ dysfunction trajectory-based subphenotypes in critically ill patients with COVID-19. *Sci. Rep.* **11**, 15872 (2021).
98. Kang, J.-H. et al. Association of cerebrospinal fluid β-amyloid 1-42, T-tau, P-tau181, and α-synuclein levels with clinical features of drug-naive patients with early Parkinson disease. *JAMA Neurol.* **70**, 1277–1287 (2013).
99. Kang, J.-H. et al. CSF biomarkers associated with disease heterogeneity in early Parkinson's disease: The Parkinson's Progression Markers Initiative study. *Acta Neuropathologica* **131**, 935–949 (2016).

100. Mollenhauer, B. et al. Longitudinal CSF biomarkers in patients with early Parkinson disease and healthy controls. *Neurology* **89**, 1959–1969 (2017).
101. Mowinckel, A. M. & Vidal-Piñeiro, D. Visualization of brain statistics With R packages ggseg and ggseg3d. *Adv. Methods Pract. Psychological Sci.* **3**, 466–483 (2020).
102. Peri, S. et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* **32**, D497–D501 (2004).
103. Dinkel, H. et al. Phospho.ELM: a database of phosphorylation sites—update 2011. *Nucleic Acids Res.* **39**, D261–D267 (2011).
104. Cheng, F., Jia, P., Wang, Q. & Zhao, Z. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* **5**, 3697 (2014).
105. Hu, J. et al. PhosphoNetworks: A database for human phosphorylation networks. *Bioinformatics* **30**, 141–142 (2014).
106. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
107. Lu, C.-T. et al. dbPTM 3.0: An informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **41**, D295–D305 (2013).
108. Meyer, M. J., Das, J., Wang, X. & Yu, H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* **29**, 1577–1579 (2013).
109. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
110. Huttlin, E. L. et al. The BioPlex network: A systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
111. Fazekas, D. et al. SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC Syst. Biol.* **7**, 7 (2013).
112. Chatr-aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–D478 (2015).
113. Goel, R., Harsha, H., Pandey, A. & Prasad, T. K. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst.* **8**, 453–463 (2012).
114. Orchard, S. et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
115. Licata, L. et al. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
116. Cowley, M. J. et al. PINA v2.0: Mining interactome modules. *Nucleic Acids Res.* **40**, D862–D865 (2012).
117. Zhou, Y. et al. AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery. *Alzheimer's Res. Ther.* **13**, 24 (2021).
118. Agarwal, D. et al. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat. Commun.* **11**, 4183 (2020).
119. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
120. Finak, G. et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
121. Grenn, F. P. et al. The Parkinson's disease genome-wide association study locus browser. *Mov. Disord.* **35**, 2056–2067 (2020).
122. Craig, D. W. et al. RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nat. Aging* **1**, 734–747 (2021).
123. The UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
124. Love, M., Anders, S. & Huber, W. Differential analysis of count data–the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).
125. Huang, R. et al. The NCATS BioPlanet–an integrated platform for exploring the universe of cellular signaling pathways for toxicology, systems biology, and chemical genomics. *Front. Pharmacol.* **10**, 445 (2019).
126. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).
127. Austin, P. C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28**, 3083–3107 (2009).
128. Liu, R., Wei, L. & Zhang, P. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat. Mach. Intell.* **3**, 68–75 (2021).
129. Thomas, J. D. & Bradley, E. Bootstrap confidence intervals. *Stat. Sci.* **11**, 189–228 (1996).
130. Kline, A. & Luo, Y. PsmPy: A package for retrospective cohort matching in python. *Proc. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 1354-1357 (2022).
131. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).

Acknowledgements

This study used data from the PPMI (obtained from PPMI database at www.ppmi-info.org/access-data-specimens/download-data) and PDBP (obtained through AMP-PD Knowledge Platform at https://www.amp-pd.org). PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including 4D Pharma, Abbvie, AcureX, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL, BioArctic, Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Capsida Biotherapeutics, Celgene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GlaxoSmithKline plc (GSK), Golub Capital, Handl Therapeutics, Insitro, Janssen Neuroscience, Jazz Pharmaceuticals, Lundbeck, Merck, Meso Scale Discovery, Mission Therapeutics, Neurocrine Biosciences, Neuropore, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi, Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voyager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics. For up-to-date information on the study, visit www.ppmi-info.org. PDBP is supported by the National Institute of Neurological Disorders and Stroke at the National Institutes of Health. Investigators include: Roger Albin, Roy Alcalay, Alberto Ascherio, Thomas Beach, Sarah Berman, Bradley Boeve, F. DuBois Bowman, Shu Chen, Alice Chen-Plotkin, William Dauer, Ted Dawson, Paula Desplats, Richard Dewey, Ray Dorsey, Jori Fleisher, Kirk Frey, Douglas Galasko, James Galvin, Dwight German, Lawrence Honig, Xuemei Huang, David Irwin, Kejal Kantarci, Anumantha Kanthasamy, Daniel Kaufer, James Leverenz, Carol Lippa, Irene Litvan, Oscar Lopez, Jian Ma, Lara Mangravite, Karen Marder, Laurie Orzelius, Vladislav Petyuk, Judith Potashkin, Liana Rosenthal, Rachel Saunders-Pullman, Clemens Scherzer, Michael Schwarzschild, Tanya Simuni, Andrew Singleton, David Standaert, Debby Tsuang, David Vaillancourt, David Walt, Andrew West, Cyrus Zabetian, Jing Zhang, and Wenquan Zou. The PDBP Investigators have not participated in reviewing the data analysis or content of the manuscript. For up-to-date information on the study, visit: https://pdbp.ninds.nih.gov. The AMP-PD program is a public-private partnership managed by the Foundation for the National Institutes of Health and funded by the National Institute of Neurological Disorders and Stroke in partnership with the Aligning Science Across Parkinson's initiative, Celgene, GSK, the Michael J. Fox Foundation for Parkinson's Research, Pfizer, AbbVie, Sanofi, and Verily Life Sciences. For up-to-date information on the study, visit https://www.amp-pd.org. This work was supported by following: the Michael J. Fox Foundation and the National Institute on Aging RF1AG072449 to F.W.; National Institute on Aging R01AG080991 and R01AG076234 to F.W. and J.B.; National Institute on Aging R01AG076448 to F.W. and F.C.; National

Institute on Aging U01AG073323, R21AG083003, R01AG066707, R01AG082118, RF1AG082211, RF1NS133812, 3R01AG066707-01S1, 3R01AG066707-02S1, and R56AG074001 and Alzheimer's Association (ALZDISCOVERY-1051936) to F.C.; National Institute of General Medical Sciences P20GM109025, National Institute of Neurological Disorders and Stroke U01NS093334, National Institute on Aging R01AG053798, P30AG072959, R35AG71476, and AG083721-01, Alzheimer's Disease Drug Discovery Foundation (ADDF), Ted and Maria Quirk Endowment, and Joy Chambers-Grundy Endowment to J.C..

Author contributions

F.W. and C.S. conceived the study. F.W. secured funding and supervised the research. C.S. preprocessed clinical data and implemented subtyping model for primary analysis. Y.H. and H.L. conducted subtype validation analysis. C.S. and Y.H. built subtype prediction model. C.S. and Y.H. conducted genetic data analysis. C.S., J.R.M., Z.B., and H.Z. conducted bulk transcriptomic analysis. X.S. advised genetic and transcriptomic analyses. Y.Z. advised imaging processing. C.S., and M.B. conducted imaging analysis. F.C. designed network medicine method and supervised network analysis and in-silico drug repurposing. JL.X., M.Z., and A.K. conducted network analysis and in-silico drug repurposing. J.B. supervised real-world data collection and analysis. Z.X. and J.X. conducted real-world data analysis. M.C.C., C.H., J.B.L., J.C., and M.S.O. provided clinical expertise. C.S. drafted manuscript. All authors critically revised the manuscript for important intellectual content. All authors provided comments and approved the paper.

Competing interests

J.C. has provided consultation to Acadia, Actinogen, Acumen, AlphaCognition, ALZpath, Aprinoia, AriBio, Artery, Biogen, Biohaven, BioVie, BioXcel, Bristol-Myers Squib, Cassava, Cerecin, Diadem, Eisai, GAP Foundation, GemVax, Janssen, Jocasta, Karuna, Lighthouse, Lilly, Lundbeck, LSP/eqt, Merck, NervGen, New Amsterdam, Novo Nordisk, Oligomerix, Optoceutics, Ono, Otsuka, Oxford Brain Diagnostics, Prothena, ReMYND, Roche, Sage Therapeutics, Signant Health, Simcere, sinaptica, Suven, TrueBinding, Vaxxinity, and Wren pharmaceutical, assessment, and investment companies. The other authors declare no Competing Financial or Non-Financial Interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01175-9.

Correspondence and requests for materials should be addressed to Fei Wang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY, USA. [2]Institute of Artificial Intelligence for Digital Health, Weill Cornell Medicine, Cornell University, New York, NY, USA. [3]Department of Surgery, University of Minnesota, Minneapolis, MN, USA. [4]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [5]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [6]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA. [7]Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [8]Department of Computer Science, Cornell Tech, Cornell University, New York, NY, USA. [9]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [10]Department of Computer Science, University of Texas at Arlington, Arlington, TX, USA. [11]Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA. [12]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [13]Department of Neurology, University of California Irvine, Irvine, CA, USA. [14]Lou Ruvo Center for Brain Health, Neurological Institute, Cleveland Clinic, Cleveland, OH, USA. [15]Chambers-Grundy Center for Transformative Neuroscience, Pam Quirk Brain Health and Biomarker Laboratory, Department of Brain Health, School of Integrated Health Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA. [16]Department of Neurology, Fixel Institute for Neurological Diseases, University of Florida, Gainesville, FL, USA. [17]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA.
✉e-mail: few2001@med.cornell.edu