scientific reports

OPEN

Multi-omics integration of scRNA-seq time series data predicts new intervention points for Parkinson's disease

Katarina Mihajlović[1], Gaia Ceddia[1], Noël Malod-Dognin[1], Gabriela Novak[3,4], Dimitrios Kyriakis[3], Alexander Skupin[3,4,5] & Nataša Pržulj[1,2,6]✉

Parkinson's disease (PD) is a complex neurodegenerative disorder without a cure. The onset of PD symptoms corresponds to 50% loss of midbrain dopaminergic (mDA) neurons, limiting early-stage understanding of PD. To shed light on early PD development, we study time series scRNA-seq datasets of mDA neurons obtained from patient-derived induced pluripotent stem cell differentiation. We develop a new data integration method based on Non-negative Matrix Tri-Factorization that integrates these datasets with molecular interaction networks, producing condition-specific "gene embeddings". By mining these embeddings, we predict 193 PD-related genes that are largely supported (49.7%) in the literature and are specific to the investigated *PINK1* mutation. Enrichment analysis in Kyoto Encyclopedia of Genes and Genomes pathways highlights 10 PD-related molecular mechanisms perturbed during early PD development. Finally, investigating the top 20 prioritized genes reveals 12 previously unrecognized genes associated with PD that represent interesting drug targets.

Parkinson's disease (PD) is a complex multifactorial disease and the second most prevalent neurodegenerative disorder affecting about 2–3% of the population over the age of 65[1]. Due to an ageing society, PD will continue to increase its burden on social systems and economy. In the United States alone, it is projected that by 2037 PD will impact more than 1.6 million individuals, surpassing the economic burden of $79 billion[2]. PD's exact cause is unknown, with only 5–10% of cases being linked to identified genetic mutations, while the remainder are categorized as idiopathic, lacking a known cause[3]. Even for PD cases with identified genetic causes, PD exhibits clinical and genetic heterogeneity with PD causing mutations including, e.g., *PINK1*, *SNCA*, *LRRK2* and *PARK2*. The current intervention strategies are limited to treating symptoms, and no curative treatment is available[3]. PD is characterized by the intracellular accumulation of misfolded α-synuclein proteins called Lewy bodies and the subsequent death of midbrain dopaminergic (mDA) neurons in the substantia nigra part of the brain[3]. Furthermore, increasing evidence implicates multiple molecular mechanisms in the disease, including disrupted mitochondrial function, calcium and protein homeostasis as well as oxidative and endoplasmic reticulum stress[1,3].

One of the main challenges in studying PD is the availability of tissue samples, as 60% of the mDA neurons have already died by the time of the diagnosis and 90% at the later stages of the disease[4]. This issue is limiting our understanding of the early stage of PD development. The recent technology of cellular reprogramming provides an alternative way of obtaining mDA neurons by converting the somatic cells of PD patients carrying disease-associated mutations into induced pluripotent stem cells (iPSCs) and differentiating them into mDA neurons[5–7]. This technique provides a practically unlimited source of mDA neurons that can be studied to uncover the molecular mechanisms driving PD. Recently, we used mDA neurons differentiated from iPSCs in the early stages of neural development (i.e., before PD is established) to investigate PD mechanisms by applying single-cell (SC) RNA sequencing (scRNA-seq)[7].

The emergence of SC sequencing techniques has led to an explosion of high-throughput measurements that can investigate cellular heterogeneity, offering the opportunity to study individual mDA neurons primed for

[1]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain. [2]Department of Computer Science, University College London, WC1E 6BT London, UK. [3]The Integrative Cell Signalling Group, Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg. [4]Luxembourg Institute of Health (LIH), Esch-sur-Alzette, Luxembourg. [5]University of California San Diego, La Jolla, CA 92093, USA. [6]ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain. ✉email: natasha@bsc.es

degeneration. Different SC sequencing techniques have been applied to study the cellular response of heterogeneous cell types and further understand the molecular mechanisms underlying PD pathology and other diseases[5–7]. However, SC measurements are associated with high levels of noise that impede the subsequent data analysis. This challenge is often addressed with data integration strategies combining multi-omics data sets (as seen in the review of Luecken et al.[8]). Such integrated data is often subjected to cell-level downstream analysis, including cell clustering, identifying cell types, and trajectory inference[9]. On the other hand, gene-level analysis (e.g., identifying disease gene markers) is typically based on identifying differentially expressed genes (DEGs)[7,9,10]. However, DEG analyses cannot uncover, for example, disease-related genes whose expressed proteins do not have altered expression but have undergone post-translational modifications, leading to disease pathogenesis[11]. This prompts the need to develop novel non-DEG-based integration methods to discover new disease markers by analyzing SC time-series data. To jointly analyze time-series scRNA-seq data and identify different cell types, Jung et al.[12] suggested a Non-Negative Matrix Factorization (NMF)-based approach.

Matrix factorization techniques (such as NMF and its extension Non-Negative Matrix Tri-Factorization (NMTF)) are popular co-clustering, dimensionality reduction and inference methods recently gaining attention for data integration. They project the original highly-dimensional data into lower-dimensional embedding (latent) spaces that are easier to handle and analyze[13]. In contrast to other artificial intelligence algorithms, these methods are linear in nature and define latent spaces with natural biological interpretations[14,15], offering traceability of their predicted values, which are crucial properties for biological data mining. These inherent properties of NMF and NMTF have been widely used for bulk data analysis to study, for instance, molecular networks (i.e. networks that capture relevant information about cellular functions and pathways) to suggest novel disease-related genes[16,17], protein functions[18] and drug-repurposing options[19]. NMF-based approaches have also shown promising results in dealing with sparse SC samples[10,12,14,15]. Furthermore, NMF-based methods have been used to jointly integrate SC data with one molecular interaction network to identify types of SCs[20], to discover interpretable gene programs[21] and to generate protein representations within various cellular contexts to identify therapeutic targets and nominate cell type contexts for rheumatoid arthritis and inflammatory bowel diseases[22]. Using matrix factorization approaches to integrate SC data with molecular networks (i.e., prior knowledge) allows us to benefit from the biologically relevant information in molecular networks and simultaneously minimize the inherent noisiness of SC data. However, current SC integration methods are only considering one molecular network to guide the organization of gene embeddings, often using PPI or COEX networks as prior knowledge. Because PD is considered a metabolic disease[23], also including the MI network during the integration process would allow for investigating the alterations in metabolic pathways between control and disease states. Despite the advances in SC data analysis, no existing method is designed to uncover novel PD gene markers while fully exploiting time-series SC data and the information in prior knowledge contained in multiple molecular interaction networks.

To go beyond the limitations of these methods, here, we propose a new NMTF-based method, NetSC-NMTF, which simultaneously decomposes a time point-specific scRNA-seq dataset of a cell line harbouring a PD-associated mutation in the *PINK1* gene (I368N mutation), or a control one, with prior knowledge from multiple molecular networks – protein-protein interaction (PPI; obtained from BioGRID[24]), gene co-expression (COEX; obtained from CoexpressDB[25]), metabolic interaction (MI; obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG)[26]), and genetic interaction (GI; obtained from BioGRID) networks (for details on the data see Section "Datasets"). NetSC-NMTF produces gene embedding vectors (i.e., "gene embeddings") that are biologically relevant, as shown by clustering and enrichment analysis in biological annotations from Gene Ontology (GO)[27], KEGG pathways (KP)[26] and Reactome pathways (RP)[28]. Then, we introduce a 2-step downstream method that mines the "gene embeddings" across all cell conditions, identifying 193 PD-related gene predictions, of which 49.7% are associated with PD in the literature. While our methodological framework does not directly incorporate time-series data during the integration step, as each time point is studied individually, our 2-step downstream analysis approach predicts and prioritizes genes considering all time points collectively, handling time-series single-cell data similar to DEG-based approaches. However, in contrast to previous studies on SC data, our workflow reveals PD-associated genes beyond the standard DEG analysis[7,10]. As the literature indicates that PD is a metabolic disease[23], we relate our 193 gene predictions to metabolic pathways by performing an enrichment analysis in KPs[26]. We highlight 10 significantly enriched KPs whose impairments in PD are supported by the literature, shedding light on the metabolic mechanisms that drive the progression of PD. Moreover, we show that incorporating MI during integration allows us to reveal more molecular mechanisms that are involved in PD pathogenesis. Then, we manually validate the top 20 highest-scoring predictions to propose 12 new and promising PD-associated genes that include seven known and two potential new drug targets, representing potential candidates for developing novel treatments for PD. Finally, we demonstrate that the predictions are not only associated with PD, but are specific to the *PINK1* mutation. The methodological pipeline presented here is a flexible framework that could be extended to incorporate other types of SC, or bulk data and applied to other complex diseases.

Results and discussion

We apply our NetSC-NMTF data integration framework (see Section "NetSC-NMTF data integration model") to scRNA-seq time-series data and molecular networks to obtain "gene embeddings", which we investigate with our 2-step downstream mining method (see Section "Predicting novel PD-associated genes: A 2-step downstream method") to uncover impaired PD pathways, novel PD-associated genes and suggest drug-repurposing options.

DisGeNet PD genes have specific properties in the embedding spaces of genes and single cells

Integrating SC expression data with prior knowledge in all four molecular networks produces functionally meaningful "gene embeddings", as measured by the clustering and enrichment analysis in GO, KP and RP annotations (Supplementary Section "Integrating single-cell expression data with molecular networks captures the functional organization of cell conditions" and Supplementary Fig. 12). We obtain gene clusters by applying the k-means clustering algorithm[29] (a widely used algorithm for getting gene clusters) on the "gene embeddings" of each time point-specific cell condition. Here, we cluster the genes in the number of clusters corresponding to the k_1 dimension of G_1, but demonstrate that our model is robust to the choice of the number of clusters in Supplementary Section "Robustness of NetSC-NMTF gene embeddings to the number of clusters and dimension". Consequently, we see that genes embedded close to each other are functionally related, leading to the interpretation that such genes participate in the same molecular pathways. As PD is characterized by perturbation of many molecular mechanisms, we assume that *DisGeNet PD genes* participate in the same molecular pathways and investigate if *DisGeNet PD genes* are also embedded close to each other. Therefore, we perform an enrichment analysis (detailed in Supplementary Section "Enrichment analysis") in *DisGeNet PD genes* of the clusters described above, measuring the percentage of clusters significantly enriched in *DisGeNet PD genes* (*p-value* $\leq$ 5%). We observe around 18% of significantly enriched clusters across all cell conditions (see Supplementary Fig. 6), where approximately half of the genes in these clusters are *DisGeNet PD genes* (average fold enrichment is 2.06). To confirm the hypothesis that genes that group with *DisGeNet PD Genes* could indeed be used to uncover new PD genes, we perform a 5-fold cross-validation with *DisGeNet PD Genes* and observe that test *DisGeNet PD Genes* co-occur with train *DisGeNet PD Genes* more than background (see Supplementary Section "Cross-fold validation using *DisGeNet PD genes*" and Supplementary Fig. 13) These results indicate that *DisGeNet PD genes* are not interspersed throughout the gene embedding spaces but rather group together and could be analyzed further to extract novel PD-associated genes.

Furthermore, we hypothesize that *DisGeNet PD genes* participate in molecular pathways that are altered more than the pathways characterized by other expressed genes between control and PD cell conditions at individual time points. To measure this alteration, we apply the method described in Section "Definition of the gene movement" to compute the distributions of the changes in the relative position (i.e., "gene movement") of *DisGeNet PD genes* and non-*DisGeNet PD genes* (background) between the embedding spaces of each PD cell condition and its time point-matching control. We compare the two "gene movement' distributions at each time point by performing a one-sided Mann-Whitney U (MWU) test (with a significance level of 5%). The *DisGeNet PD genes* "movement" distributions across all time points are statistically significantly larger than the one of background genes, with *p-values* $\leq 1.65e^{-05}$ (see Supplementary Fig. 7).

In conclusion, we observe that: (1) *DisGeNet PD genes* cluster together in the gene embedding spaces of individual cell conditions; (2) the "movement" of *DisGeNet PD genes* is statistically significantly larger than that of background genes for each time point-specific control and PD pairwise comparison. The following sections build upon these two observations to predict and validate novel PD-associated genes.

Obtaining *Core PD predictions*

To uncover novel PD-associated genes, we mine the "gene embeddings" obtained with our NetSC-NMTF framework with our 2-step downstream method (see Section "Predicting novel PD-associated genes: A 2-step downstream method"). In the first step, we obtain PD predictions for a given time point (i.e., ***Stage-specific PD predictions***), by extracting non-*DisGeNet PD genes* from the clusters of PD cell conditions (obtained above) significantly enriched in *DisGeNet PD genes*. We show that relevant biological information can be gained from analyzing SC data at individual time points by observing that gene predictions associated with all four time points of cell development (i.e., ***Stage-specific PD predictions***) are significantly associated with PD (see Supplementary Section "Stage-specific PD predictions", Supplementary Figs 9 and 10).

To account for the noisiness of the scRNA-seq measurements and to focus on the genes involved in the PD progression across all time points, we adopt a consensus approach used in other studies[7,12] that makes a final list of predictions based on all available time points. Thus, we apply the second step of our 2-step downstream method to define a final set of predictions by intersecting all sets of ***Stage-specific PD predictions*** and ranking the genes in the overlap according to the average "movement" across all time points so that the genes with the largest average "movement" are ranked the highest (Section "Predicting novel PD-associated genes: A 2-step downstream method"). This results in 193 PD predictions that we name ***Core PD predictions*** (Supplementary File 1).

To assess if such a consensus approach is applicable to extract novel PD-associated genes, we test if the four sets of ***Stage-specific PD predictions*** significantly overlap by applying a sampling with replacement technique (Supplementary Section "Sampling with replacement"). Over 10,000 repetitions, the overlap of ***Stage-specific PD prediction*** sets is always larger compared to random (*p-value* $\leq 1e^{-05}$). This significant overlap shows that a consensus approach is possible and can be used to uncover a set of new genes predicted to be related to PD across all time points.

Core PD predictions are relevant for PD

To determine the PD-relevance of the ***Core PD predictions***, we follow the validation procedures described in Section "Validating predictions". We obtain the co-occurrence distributions of the predictions and the background genes with the term "Parkinson's disease" in PubMed publications, which we compare with a one-sided MWU test (with a significance level of 5%). We find that the co-occurrence distribution of ***Core PD predictions*** is significantly greater than the one of background (*p-value* $= 4.96e^{-12}$) (see Supplementary Fig. 11). Furthermore, we observe that our ***Core PD predictions*** are significantly enriched in the PD-related genes (*p-value* $= 1.48e^{-10}$), with 49.7% of our ***Core PD predictions*** belonging to the set of PD-related genes. These results demonstrate that the 193 ***Core PD predictions*** are significantly associated with PD.

We also examine whether our methodology predicts PD-related genes that can not be uncovered through conventional DEG analysis. Thus, we compare our ***Core PD predictions*** with the 232 protein-coding DEG-based predictions obtained by Novak et al.[7] (Section "Biological annotations, PD genes and DEGs") (note that the same scRNA-seq dataset is used in our study), by checking the overlap between the two sets of genes. We uncover eight genes (*GOLT1B*, *PDIA6*, *RPN2*, *PFKP*, *FOS*, *EGLN3*, *GNAS* and *LMAN1*) (Fig. 1a), all of which have been linked with PD (Supplementary Table 4), but whose exact roles in the pathogenesis of PD are not fully understood. Identifying these genes through two independent analyses suggests their importance in PD and warrants more extensive studies to elucidate their involvement in PD and investigate them as potential intervention opportunities. Additionally, by integrating scRNA-seq data with prior knowledge, we reduce noise characteristic for the scRNA-seq measurements, which allows us to predict and prioritize a set of PD-related genes beyond standard DEG analysis, offering new insights into PD and further proving the value of our data integration model and the 2-step downstream analysis.

In conclusion, we prove the PD-relevance of ***Core PD predictions*** and show that the consensus approach allows for uncovering novel PD-associated genes. Furthermore, in Supplementary Section "Comparison with other methods" and Supplementary Table 66, we show that our 2-step method outperforms other approaches based on: (1) DEGs (LIGER[10], a state-of-the-art method for integrating SC data), (2) distance of non-*DisGeNet PD genes* to *DisGeNet PD genes* in the embedding spaces of PD cell conditions, or (3) "movement" of genes between control and PD time point-specific cell conditions. Our method leads to more PD-relevant predictions, which we measure by validating the predictions obtained with other methods through an automated PubMed search, and performing an enrichment analysis in a PD-related set of genes (Section "Biological annotations, PD genes and DEGs") and KPs.

KEGG pathways enriched in our *Core PD predictions* are associated with PD

Because PD is characterized by disruption of many molecular pathways (suggesting a PD is a metabolic disorder[23]), we believe that adding the MI network in our gene embedding methodology will allow us to better uncover PD molecular mechanisms (i.e. pathways) when comparing the embeddings of genes between disease and control. To test this, we investigate the PD-relevance of the pathways uncovered by our methodology, with and without including the MI network during integration.

First, we determine the metabolic functions associated with our 193 ***Core PD predictions*** obtained by including MI during intergation and investigate if they agree with the literature. To that aim, we perform enrichment analysis (Supplementary Section "Enrichment analysis") in KPs, identifying 37 significantly enriched ones (Supplementary File 2). Here, we present the top 10 most significantly enriched KPs of the ***Core PD predictions*** (see Fig. 1b), which we rank according to the most significant *p-value* (smallest first). Most importantly, *Parkinson's disease* is one of the most enriched pathways (rank 8), further showing that our predictions are related to this disease. *Protein processing in the endoplasmic reticulum* (ER) is the first-ranked pathway. It is indeed relevant for

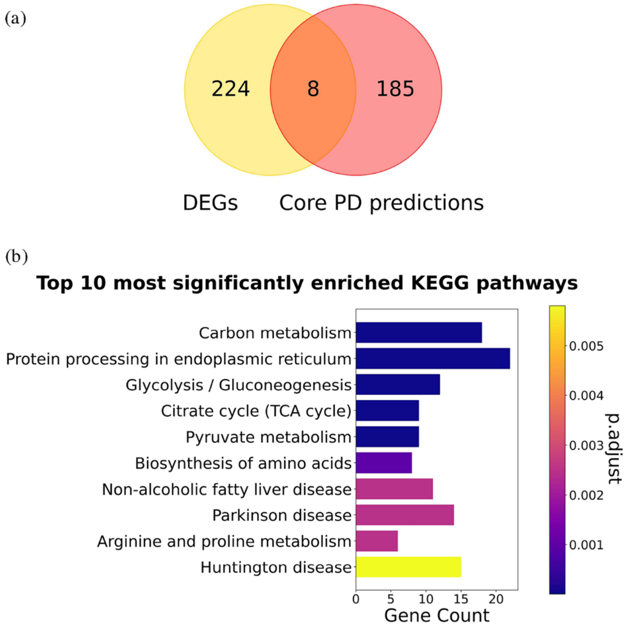

Figure 1. **(a)** Overlap between DEGs from Novak et al.[7] and ***Core PD predictions***. **(b)** Top 10 KPs significantly enriched in ***Core PD predictions***. p.adjust represents adjusted *p-values* (obtained from enrichment analysis) for multiple hypothesis testing using a method from Benjamini and Hochberg[35]. Gene Count are the number of ***Core PD predictions*** that participate in a KP.

PD as the accumulation of α-synuclein proteins in PD induces ER stress by inhibiting the ER-Golgi trafficking and leading to the dysregulation of protein processing in ER and eventually cell death[30]. *Carbon metabolism* (rank 2) is altered in PD, leading to a decrease in glucose metabolism and abnormally elevated levels of pyruvate (*pyruvate metabolism*, rank 5)[23]. Furthermore, *glycolysis/gluconeogenesis* (rank 3) and *citrate cycle* (rank 4) contribute to the overarching *carbon metabolism pathway*[26], emphasizing the contribution of the altered carbon metabolism in the progression of PD. Thus, protein processing in the ER and carbon metabolism undoubtedly play a vital role in the progression of PD and should be further investigated. PD is also characterized by the dysregulation of the *biosynthesis of amino acids* (rank 6)[31], with evidence pointing to a lower abundance of arginine and unaltered amounts of proline (*arginine and proline metabolism*, rank 7)[32], highlighting the need to study these mechanisms in more detail and uncover their exact role in PD. Although the association between the *non-alcoholic fatty liver disease pathway* (rank 9) and PD is not well understood, evidence suggests that it plays a role in PD[33]. Finally, it is not surprising that Huntington's disease (rank 10) is one of the most enriched pathways as it shares many disrupted mechanisms with PD, including protein processing in ER[34]. In Supplementary Section "Enriched KEGG pathways shared between DisGeNet PD genes and ***Core PD predictions*** are relevant for PD", we also perform an enrichment analysis in KPs of *DisGeNet PD genes* and compare the results with those of ***Core PD predictions***. Notably, we find that *DisGeNet PD genes* are enriched in 144 KPs, sharing 14 common pathways with the ***Core PD predictions*** that most importantly include *Parkinson's disease* (rank 8). Overall, we show that the top 10 KPs associated with our ***Core PD predictions*** are all relevant for PD development at the early stages of the disease.

In contrast, when not including MI during integration, we observe that the resulting set of 90 ***Core PD predictions*** are enriched in only three PD-related metabolic pathways (see Supplementary Section "Enriched KEGG pathways in predictions obtained by not including MI network during integration" and Supplementary Fig. 14a), unlike 37 pathways enriched for the 193 ***Core PD predictions***, 20 of which have been further investigated and associated with PD above and in Supplementary Section "Enriched KEGG pathways shared between DisGeNet PD genes and ***Core PD predictions*** are relevant for PD".

To better position our findings with respect to conventional DEG analysis, we perform enrichment analysis in KPs of the total set of 232 DEGs from Novak et al.[7], the set of 224 unique DEGs (i.e., only found by Novak et al.) and the 185 unique ***Core PD predictions*** (i.e., only found by our analysis). We reveal that the 232 DEGs are not enriched in any pathway, that the 224 unique DEGs are enriched in three KEGG terms not indicative of PD (*Prion disease*, *Antigen processing and presentation* and *Legionellosis*) (Supplementary Fig. 14b), and that the 185 unique ***Core PD predictions*** are enriched in 35 KPs (20 of which have been investigated and associated with PD, as discussed above and in Supplementary Section "Enriched KEGG pathways shared between *DisGeNet PD genes* and ***Core PD predictions*** are relevant for PD"), sharing the same top 10 most significantly enriched KPs as the full set of the 193 ***Core PD predictions*** (Supplementary Fig. 14c). This further shows that our methodology uncovers PD-related genes that complement the findings obtained by the traditional DEG approach.

Thus, we suggest that future research endeavors aim to further understand the relationship of the discussed pathways with PD, as they might significantly contribute to the pathogenesis of this disease and provide opportunities for drug discovery efforts.

Literature validation of the top 20 *Core PD predictions*

The results above show that our ***Core PD predictions*** are globally associated with PD, suggesting that the remaining non-validated ***Core PD predictions*** are also relevant for PD. Thus, we focus on the top 20 prioritized ***Core PD predictions*** to better characterize their relationship to PD and their potential role in drug-repurposing strategies. We manually assess if the top 20 ***Core PD predictions*** are related to PD in the literature and find that eight genes (40%) have a known association. We find literature evidence for the remaining 12 genes that could explain their potential role in the disease (Supplementary Table 5). Additionally, we identify seven (out of 12) druggable genes that represent candidates for future drug-repurposing investigations and suggest two other genes for drug discovery studies, providing potential novel therapeutic opportunities for PD (Supplementary Table 4). The 12 gene predictions could play a role in PD based on the metabolic pathways they participate in, or their involvement in other neurodegenerative diseases. Here, we discuss seven of those predictions and their druggability.

The mutation of *PFN1* (rank one) leads to the development of a neurodegenerative disease called amyotrophic lateral sclerosis[36]. *PFN1* also regulates the dynamics of the actin cytoskeleton, whose dysregulation has been implicated in multiple neurodegenerative diseases such as PD and Alzheimer's. *PFN1* is also a target of Artenimol, a drug originally used to treat malaria[37]. Gao et al.[38] proved that Artenimol could be used for treating neuroinflammatory diseases by inactivating the PI3K/AKT and NF-κB signalling pathways, two pathways that are dysregulated in PD[39,40], suggesting that the Artenimol-*PFN1* drug-target interaction could be exploited for treating PD. We highlight six gene predictions (*APLP2*, *RRBP1*, *RCN1*, *SEC63*, *KDELR1*, *SSR4*) for their role in maintaining the proper functioning of ER (see Supplementary Table 4). *APLP2* (rank 4) is a target of zinc and some of its compounds. It could be exploited for maintaining the optimal levels of zinc, whose alterations have been implicated in the pathophysiology of PD[41]. We also find that *RRBP1* (rank 5) is druggable by Radezolid[42], which has been used in trials to treat skin diseases and might be repurposed for PD. *RCN1* (rank 8) and *SSR4* (rank 15) are affected by calcium[43], providing opportunities to maintain Ca^{2+} homeostasis and reverse the toxicity of the misfolded α-synuclein proteins, thereby preventing ER stress. Given that four of the six genes under investigation are established drug targets, the remaining two, *SEC63* and *KDELR1*, warrant consideration as prospective candidates for future PD drug discovery studies.

Here, we have shown evidence that the top prioritized ***Core PD predictions*** are associated with PD, motivating us to propose 12 novel PD-associated genes. Seven of the 12 uncovered genes are known drug-targets and candidates for future drug-repurposing investigations, while two genes represent possible intervention points for drug discovery studies.

Core PD predictions are associated with PD subtype carrying a *PINK1* mutation

To determine if ***Core PD predictions*** are relevant to the *PINK1* subtype of PD considered in this study, we investigate if our gene predictions are closely related to the *PINK1* gene by studying the subgraph that the predictions and *PINK1* form in the PPI network. The intuition behind this approach is that proteins in the same neighbourhood in a PPI network are likely to participate in the same functional modules, such as protein complexes, metabolic pathways or signal transduction systems. We generate the subgraph of the PPI network obtained from Biogrid (defined in Section "Datasets") induced by the genes expressed in at least one of our cell conditions, making it more relevant for our data. In this data-specific PPI network, we measure the shortest path of ***Core PD predictions*** and background (genes in the PPI subgraph that are not ***Core PD predictions***) to the *PINK1* gene and compare the two shortest paths distributions using a one-sided MWU test (with a significance level of 5%). We observe that ***Core PD predictions*** are statically significantly closer ($p\text{-}value = 6.91e^{-17}$) to the *PINK1* gene than the background (see Fig. 2a), with the average shortest path length of 1.92 of our predictions and 2.26 of the background. To determine if our predictions are more specific to the *PINK1* subtype of PD than the 232 DEGs from Novak et al.[7], we perform the same experiment with this set of genes and observe that the DEGs are also statistically significantly closer ($p\text{-}value = 2.06e^{-04}$) to the *PINK1* gene than the background (genes in the PPI subgraph that are not DEGs). However, as the average shortest path of DEGs to *PINK1* is larger than that of our ***Core PD predictions*** (2.13 compared to 1.92), we conclude that our method allows us to find genes that are more specific to the *PINK1* subtype of PD.

Having shown that ***Core PD predictions*** are close to *PINK1* in the PPI network, we hypothesize that they participate in the same PD-related pathways with *PINK1*. We further assume that the genes closest to *PINK1* in the PPI network are the ones that experience the effects of *PINK1*'s mutation first, subsequently leading to the

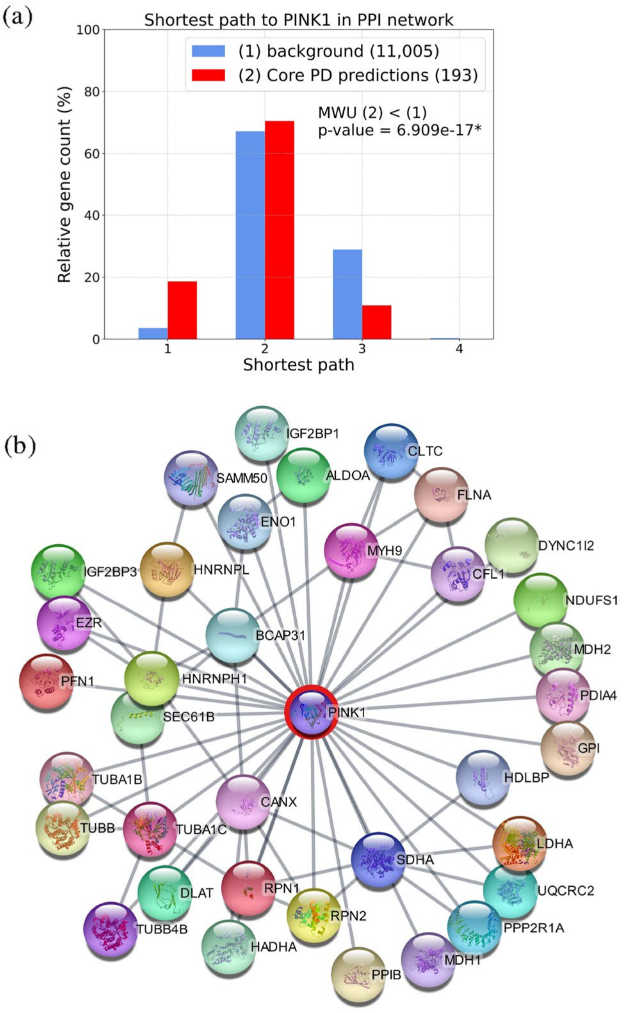

Figure 2. **(a)** Shortest path length distribution of genes to *PINK1*gene in the PPI network. We represent the background gene set as group (1) (blue) and ***Core PD predictions*** set as group (2) (red), indicating the number of genes in each set in the brackets on the right. MWU (2) < (1) indicates that we perform a one-sided MWU (with a significance level of 0.05) to test if the shortest path distribution of ***Core PD predictions*** is significantly smaller than the one of background genes (with $p\text{-}value < 0.05$ indicated by *). **(b)** PPI network of *PINK1*and its 36 first neighbours from ***Core PD predictions***. Network visualization was done using Cytoscape 3.10.0[50].

impairment of the biological mechanisms these genes participate in, thereby contributing to PD development. Therefore, we focus on the 36 ***Core PD predictions*** that are *PINK1*'s first network neighbours (see Fig. 2b). To test if these genes indeed participate in PD-associated biological mechanisms and how they relate with *PINK1*, we perform an enrichment analysis in PD terms (i.e., pathways) obtained from PD map[44]; we exclude *Scrapbook* and *Parkinsons UK Gene Ontology genes* from the terms collected from PD map, as they do not represent biological pathways. We find that our 36 ***Core PD predictions*** are statistically significantly enriched in *glycolysis* (*p-value* $= 5.68e^{-03}$), *actin filament organization* (*p-value* $= 9.07e^{-03}$), *mitochondrial and ROS metabolism* pathway (*p-value* $= 4.71e^{-02}$), and *axonal remodeling and CDK5 signaling* (*p-value* $= 3.10e^{-02}$). A recent study[45] showed that glycolysis is elevated in a PD model harbouring a *PINK1* I368N mutation, a mutation also investigated in this paper. Additionally, altered actin dynamics were observed in *PINK1* knockdown dopamine neuronal cells[46]. Mitochondrial dysfunction and increased mitochondrial ROS are known as hallmarks of the PD subtype carrying *PINK1* mutations[47]. Interestingly, there is no direct evidence linking *PINK1* with *axonal remodelling and CDK5 signaling* mechanism. However, *PINK1* mutations influence *LLRK2* levels[48] (a commonly mutated protein in PD)[49], which in turn disrupts *axonal remodelling and CDK5 signaling* in PD. Therefore, *PINK1* mutations may contribute to PD development by disrupting *axonal remodelling and CDK5 signaling* through their relationships with *LLRK2*. Our findings emphasize the need to investigate *axonal remodelling and CDK5 signaling* as a new and important pathway in PD pathogenesis associated with *PINK1* mutations.

These results demonstrate the power of our methodological pipeline to predict genes pertinent to a PD subtype characterized by a *PINK1* mutation. Additionally, the PPI subnetwork from Fig. 2b) reveals the molecular interactions that connect the *PINK1* mutations with the four PD pathways mentioned above. To further understand how a mutation in *PINK1* leads to PD development, we recommend future research to be directed at studying the 36 ***Core PD predictions*** that are *PINK1*'s first PPI network neighbours and their related molecular mechanisms, such as *axonal remodelling and CDK5 signaling*.

Conclusion

The complexity of PD requires new integration methods capable of exploiting multi-omics data. To address the open challenge of integrating time-series scRNA-seq data with molecular networks to uncover novel PD-associated genes beyond DEGs, we propose our NetSC-NMTF framework (Section "NetSC-NMTF data integration model") and the 2-step downstream method (Section "Predicting novel PD-associated genes: A 2-step downstream method"). Using prior knowledge hidden in the molecular networks, our framework effectively minimizes noise intrinsic to the scRNA-seq measurements, thereby prioritizing genes and pathways pertinent to PD.

We apply our methodology to integrate and analyze four condition-specific molecular interaction networks and scRNA-seq data of a cell line harbouring a PD-associated mutation in the *PINK1* gene (I368N), or a control cell line, at four time points of cell development (eight cell conditions). We identify 193 PD-related predictions that we call ***Core PD predictions***, out of which 49.7% have been previously associated with PD in the literature. We also show that DEG-based approaches cannot uncover our genes and that our method predicts genes that are more PD relevant than DEG-based methods such as LIGER[10] and the method presented in Novak et al.[7] (Section "Core PD predictions are relevant for PD", Fig. 1a). To shed light on biologically relevant PD mechanisms, we discuss the top 10 most enriched KPs of the ***Core PD predictions*** (see Fig. 1b). We perform a manual literature validation on the top 20 predictions to suggest 12 novel PD-associated genes implicated in PD based on the metabolic pathways they participate in, or their role in other neurodegenerative disorders. Seven out of 12 novel PD-associated genes are druggable, so we recommend future drug-repurposing directions that could represent new therapeutic options for treating PD. Additionally, we predict two new PD-associated genes that are not known drug targets but represent potential intervention points that should be considered in future drug discovery studies on PD. Furthermore, we demonstrate that the ***Core PD predictions*** are specific for the PD subtype associated with the *PINK1* gene mutation. To our knowledge, this is the only non-DEG-based approach that integrates scRNA-seq time-series data of Parkinson's disease and control samples with molecular networks and exploits the information of the "gene embeddings" to: (1) uncover novel disease-related genes, (2) reveal critical metabolic pathways and (3) propose new drug-repurposing and drug discovery options.

Here, we analyze iPSC-derived data from a mutation-specific neuronal cell line against the control. While the iPSC-derived neuronal models represent the gold standard for analyzing PD in vitro, significantly contributing to understanding this disease[6,7], it suffers from several biological drawbacks. For example, ageing-related effects and epigenetic influence are lost during reprogramming to iPSCs. Additionally, as iPSC is a two-dimensional cell culture model, it does not fully recapitulate the cell-cell/cell-matrix interactions and cell morphology present in vivo[51]. However, our methodology is generic and could be modified to accommodate data from tissue samples and other data types, such as metabolomics, proteomics, and epigenetics (including bulk datasets). New data could complement the information in the expression data and molecular networks, allowing for uncovering novel biological knowledge. In this study, we consider all input matrices equally and investigate how different combinations of molecular networks and SC expression data influence the functional organization of gene embeddings by performing an ablation study. In general, the different input data could have been differently weighted to highlight different disease aspects, such as metabolic impairment, for which MI could be integrated with a higher importance. Additionally, performing wet lab experiments could provide stronger evidence for the involvement of our predicted genes in PD pathogenesis, which we leave for future work. Although we apply our integration method to PD data associated with *PINK1* mutation, our framework is generic and versatile and could be used to investigate PD heterogeneity resulting from other PD-causing mutations to uncover genes and mechanisms common to various PD subtypes, potentially determining disease causality. Additionally, the comparison across PD subtypes could not only identify common PD mechanisms shared by all subtypes but also subtype-specific mechanisms that could help further PD treatment strategies in the context of personalized medicine. Finally, our framework could also be applied to other diseases, or processes where analysis of time series expression data is key, e.g., gender aging differences, or cell response to drug treatment.

Methods

Datasets

Expression matrices and molecular networks

From Novak et al.[7], we obtain the SC dataset that contains normalized scRNA-seq data of mDA neurons of two cell lines: a Parkinson's disease cell line obtained from a 64-year-old male with a homozygous ILE368ASN mutation (P.I368N/P.I368N) in the *PINK1* gene and an age- and sex-matched control cell line, both at four time points (stages) (day 0, 6, 15, and 21), corresponding to the initial phase of the development of the disease. The scRNA-seq data is also available through the Gene Expression Omnibus (GEO), accession number GSE183248. In this study, we call a cell line at a specific time point a cell condition, leading to eight cell conditions and use a convention *cell line*$_{day}$ (e.g., Control$_{\mathrm{D0}}$ for control cell line at day 0; PD$_{\mathrm{D0}}$ for PD cell line at day 0) to refer to a particular cell condition (see Supplementary Table 1). We model the expression data of each cell condition by a matrix E in which rows represent genes, columns represent cells, and an entry E_{ij} is the normalized read count of gene i in cell j.

To integrate the data with prior knowledge, we collect four molecular networks for *Homo sapiens*. To create the PPI network, we collect all physical interactions between proteins from BioGRID 4.3.195[24], captured by at least one of the following experiments: Two-hybrid, Affinity Capture-Luminescence, Affinity Capture-MS, Affinity Capture-RNA, Affinity Capture-Western. To make the GI network, we fetch genetic interactions reported in BioGRID 4.3.195[24]. We create the COEX network by collecting the top 1% strongest correlations between genes from CoexpressDB v.7.3[25]. Finally, we construct the MI network by connecting genes participating in the same metabolic pathways in KEGG. We retrieve the pathways that are annotated by at least one of the following metabolism-related keywords in KEGG 2021/01[26]: metabolism, metabolic, glycolysis, TCA, oxidative phosphorylation, fatty acid, pentose, degradation, or biosynthesis.

We filter the SC expression data for each cell condition to keep only protein-coding genes with at least one PPI in BioGRID, as PPIs are the most direct evidence that two proteins interact. Similarly to what was done in Malod-Dognin et al.[17], we construct condition-specific PPI, GI, COEX and MI networks by considering protein-coding genes expressed in a cell condition (as measured by scRNA-seq). An edge connects nodes in the networks if the corresponding genes (or, equivalently, their protein products) interact in the molecular interaction networks obtained from the databases (detailed above) (see Supplementary Table 2).

Biological annotations, PD genes and DEGs

To assess if our integration framework produces biologically coherent "gene embeddings", we obtain biological annotations from Gene Ontology (GO)[27], KEGG pathways (KP)[26] and Reactome pathways (RP)[28] (all annotations were collected on 10 March 2021). We also use KPs to assess the biological relevance of our predicted set of genes and identify metabolism mechanisms perturbed during early PD development. Importantly, we only use the interactions from KEGG to create the MI network, used in the integration process, and the KP annotations of genes to evaluate gene embeddings, which are obtained from the network topology created from the interactions and not from annotations.

Additionally, we collect genes from DisGeNet associated with Parkinson's Disease (Concept Unique Identifier: C0030567) (collected on 14 May 2021) and consider them our ground-truth PD genes, terming them *DisGeNet PD genes*. We keep only those *DisGeNet PD genes* expressed in our transcriptomics data, resulting in 1,378 genes. To examine whether our methodology predicts PD-related genes that cannot be uncovered through conventional DEG analysis, we obtain the 232 protein-coding DEGs from the original study of the SC analysis[7] and investigate their overlap with our ***Core PD predictions***.

NetSC-NMTF data integration model

To integrate a condition-specific single-cell expression matrix, E, with molecular interaction networks, we extend an NMTF-based method, iCell[17], to our new framework NetSC-NMTF (see Fig. 3).

Molecular interaction networks are represented by their adjacency matrices, $A_{i\in\{1,\ldots,4\}}$, where each A_i is a symmetric matrix with $A_i[v][w]$ entry value of one if genes v and w interact with each other, and zero otherwise. All input matrices are simultaneously decomposed into products of three matrix factors so that $A_i \approx G_1 S_i G_1^T$ and $E \approx G_1 S_5 G_2^T$, where $G_1 \in \mathbb{R}^{n\times k_1}$, $G_2 \in \mathbb{R}^{m\times k_2}$, $S_{i\in\{1,\ldots,4\}} \in \mathbb{R}^{k_1\times k_1}$ and $S_5 \in \mathbb{R}^{k_1\times k_2}$, with n and m being the number of genes and SCs, respectively. k_1 and k_2 represent the optimal dimensions of the latent embedding spaces, which we obtain by computing the maximum dispersion coefficient[52] (see Supplementary Section "Choosing the number of dimensions", Supplementary Table 3 and Supplementary Figs. 2 and 3). Additionally, in Supplementary Section "Robustness of NetSC-NMTF gene embeddings to the number of clusters and dimensions", we demonstrate that our integration model is robust to the choice of k_1 and k_2 parameters. Note that matrix factor G_1 is shared across all decompositions, facilitating the information flow and learning from all data.

According to the embedding interpretation of NMTF, the set of rows of matrix G_1 defines the set of embedding vectors of the genes (also called "gene embeddings"), and the set of rows of matrix G_2 defines the set of embedding vectors of the SCs. To emphasize the contribution of a biological condition of SCs, we transform "gene embeddings" from G_1 to the space spanned by G_2 by using the transformation matrix S_5 to compute $U = G_1 * S_5$ (detailed in Supplementary Section "Movement of DisGeNet PD genes projected in the SC embedding spaces"). In the rest of the paper, the rows of matrix U (that we call u_i) are referred to as "gene embeddings". On the other hand, NMTF also has a co-clustering interpretation where G_1 and G_2 are interpreted as cluster indicator matrices of genes and SCs, grouping genes and SCs into k_1 and k_2 clusters, respectively. $S_{i\in\{1,\ldots,4\}}$ matrices are interpreted

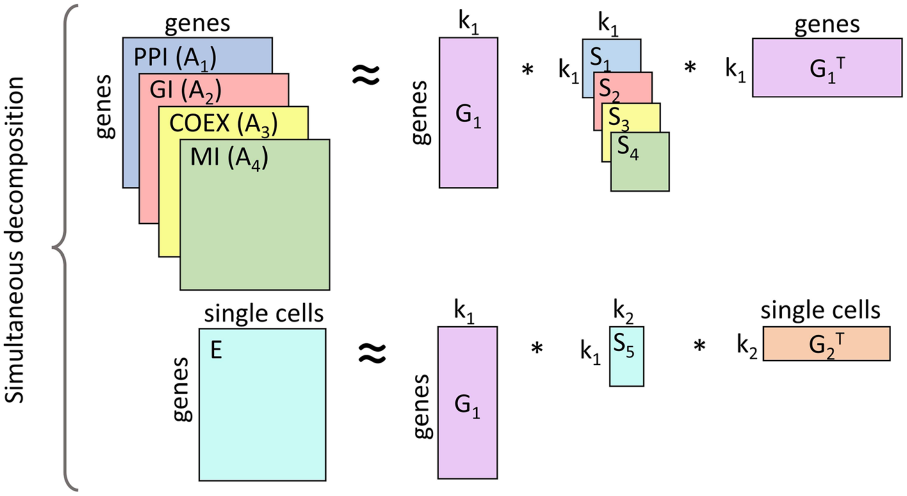

Figure 3. NetSC-NMTF model**.** Non-negative Matrix Tri-Factorization (NMTF)-based model used for integrating SC expression matrix, E, with molecular interaction networks PPI, GI, COEX, and MI (represented by their adjacency matrices, A_1, A_2, A_3, and A_4, respectively). The matrix factor G_1 is shared across decompositions to allow learning from all input matrices. The parameters k_1 and k_2 indicate the reduced dimensions of the embedding spaces of human genes and single cells, respectively.

as the compressed representations of molecular networks, and S_5 is the compressed representation of the SC expression matrix. Based on this interpretation, we use G_1 when performing clustering, as it leads to higher enrichments in biological annotations (Supplementary Fig. 4) than the clusters of matrix U (Supplementary Fig. 5). The intuition behind this observation is that gene embeddings in G_1 matrices are mostly influenced by biological networks, and we expect these embeddings to have high enrichments in biological annotations. On the other hand, gene embeddings that we map to the space of single cells (U matrices) are more suited for capturing PD biological signal caused by the differences between SC phenotypes of disease and control cell conditions at individual time points as measured by the higher "movement" of *DisGeNet PD genes* in U than in G_1 (Supplementary Section "Movement of DisGeNet PD genes projected in the SC embedding spaces"). Hence, G_1 embeddings capture more generic biological processes, while U embeddings, which are driven by phenotype-associated data, highlight PD-related biology.

Solving NMTF is an NP-hard continuous optimization problem[53]. Thus, we obtain all matrix factors by applying a heuristic fixed-point solver based on multiplicative update rules (MURs)[54] (see Supplementary Section "Multiplicative update rules") to solve the following optimization problem:

$$\min_{G_1,G_2,S_i,S_5} \left[\sum_{i=1}^{4}(\| \| A_i - G_1 S_i G_1^T \| \|_F^2) + \| \| E - G_1 S_5 G_2^T \| \|_F^2\right] \tag{1}$$

where $G_1, G_2 \geq 0$ and $\| \|_F$ denotes the Frobenius norm. Starting from an initial solution, the solver iteratively uses MURs to converge towards a locally optimal solution. To initialize the matrix factors, we apply singular value decomposition (SVD) on the original input matrices because it reduces the number of iterations needed to achieve convergence and makes the solver deterministic[55]. To comply with the non-negativity constraint of NMTF, we generate the initial solution by taking the absolute values of the entries of the resulting SVD matrices. The iterative process stops when the objective function converges (e.g., Supplementary Fig. 1), which we measure every 10 iterations with $\left|\frac{\mathscr{F}_{i-10} - \mathscr{F}_i}{\mathscr{F}_i}\right| \leq 10^{-3}$, where $\mathscr{F}_i$ is the value of the objective function at the current iteration and $\mathscr{F}^{i-10}$ is the value computed at iteration $i - 10$.

Definition of the "gene movement"

To determine how different biological states/conditions of SCs alter "gene embeddings", we introduce the so-called *Gene Mapping Matrix* (*GMM*). *GMM* is a symmetric distance matrix that captures the relative positions between the "gene embeddings" in the embedding space of SCs of one cell condition. We compute a *GMM* of a cell condition with $GMM[u_i][u_j] = d(u_i, u_j)/\| U \|_F$, where each entry corresponds to the norm-scaled Euclidean distance between the "gene embeddings" u_i and u_j of two genes i and j in the matrix U and $\| \|_F$ denotes the Frobenius norm. Since *GMMs* encode such relative and normalized gene positions, we can directly compare the position of one gene in one *GMM*, g_i, with its position in another *GMM*, g_i'. To do this, we compute the Euclidean distance between these positions with $GM_i = d(g_i, g_i')$, where GM_i is a scalar that we call a "movement" of a gene *i* between two cell conditions (i.e., "gene movement"). In other words, "gene movement" represents how the relative position of a gene (i.e., relative to all other genes in one cell condition) changes between the embedding spaces of two cell conditions. While "gene movement" can be computed for genes between any two cell conditions, we calculate the "gene movement" of genes between time point-matching PD and control cell conditions in our study. This allows us to investigate how PD affects the spatial organization of embedding spaces of genes compared to healthy controls at different stages of PD development. The "gene movement" is either a positive value and indicates to what extent PD alters the relative position of a gene compared to a corresponding control, or zero if there is no such change.

Predicting novel PD-associated genes: A 2-step downstream method

We uncover novel PD-associated genes by applying our NetSC-NMTF framework to obtain "gene embeddings" of time point-specific data, which we mine using the following 2-step downstream method. In summary, we select the top PD-related genes (i.e., *Core PD predictions*) by applying our 2-step downstream analysis method. For each time point, we define the ***Stage-specific PD predictions*** as the genes that are statistically significantly associated with *DisGeNet PD Genes*, i.e., genes that appear in the clusters that are significantly enriched in *DisGeNet PD Genes* (*p-value enrichment* $\leq$ 5%, a threshold that is standardly used to determine the significance of enrichments). In Section "*DisGeNet PD genes* have specific properties in the embedding spaces of genes and single cells", we experimentally validate that these genes are indeed likely to be PD-related. Then, we define the ***Core PD predictions*** as genes that are statistically significantly associated to *DisGeNet PD Genes* at all time points, which we obtain by intersecting all sets of ***Stage-specific PD predictions***. Finally, we prioritize the ***Core PD predictions*** by computing their average "movement" across all time points, ranking the ones with the highest "movement" on top.

In the first step, we obtain PD predictions for a given time point (i.e., ***Stage-specific PD predictions***). This step is based on the hypothesis that *DisGeNet PD genes* (defined in Section "Biological annotations, PD genes and DEGs") cluster in the gene embedding spaces of PD cell conditions (see Section "DisGeNet PD genes have specific properties in the embedding spaces of genes and single cells") and that genes that group with *DisGeNet PD genes* are also relevant for PD. A similar approach has been successfully applied by Gligorijevic et al.[56] to find new cancer driver genes. Therefore, we apply a k-means clustering algorithm (Pedregosa et al.[29]; a widely used algorithm for getting gene clusters) on G_1 matrix of a time point-specific PD cell condition, clustering the genes in the number of clusters corresponding to the k_1 dimension of G_1. Next, we perform an enrichment analysis (see Supplementary Section "Enrichment analysis") in *DisGeNet PD genes* of these clusters and keep the significantly enriched ones. Finally, from significantly enriched clusters, we retain those genes that are not labelled as *DisGeNet PD genes* and are also expressed in the control cells of the corresponding time point, calling this set of genes ***Stage-specific PD predictions***. We apply the first step for each of the four time points, resulting in four sets of ***Stage-specific PD predictions***, which we validate in the literature in Supplementary Section "Stage-specific PD predictions", Supplementary Figs 9 and 10.

In the second step, we intersect the four ***Stage-specific PD predictions*** obtained from step 1 to define our ***Core PD predictions***. By focusing on the intersection of ***Stage-specific PD predictions***, we remove the potential stage-specific noise and hypothesize that we uncover genes that drive PD across all time points caused by the *PINK1* mutation. Furthermore, previous studies on time-dependent SC data have shown that a significant signal could be detected when SC data across all time points are exploited, stressing the importance of such approaches[7,12]. To prioritize the ***Core PD predictions***, we compute their average "movement" across all time points (Section "Definition of the gene movement") and rank the predictions according to their average "movement", the largest first. This approach is based on the observation that the relative positions of *DisGeNet PD genes* are more altered in PD versus control conditions compared to other genes (see Section "DisGeNet PD genes have specific properties in the embedding spaces of genes and single cells"). Therefore, we hypothesize that the more the relative position of a gene is altered in PD versus control (at a given time point, or across all time points), the more relevant it is for PD (which we confirm in Supplementary Section "Stage-specific PD predictions" and Supplementary Fig. 8).

Validating predictions

We validate our gene predictions and quantify their association with PD in the literature by using an automated PubMed publication search to count the co-occurrence of each gene from the set of predictions and the background set (defined below) with the term "Parkinson's disease" in PubMed publications. To measure if our predicted genes are significantly more co-occurring with PD in the literature, we perform a one-sided Mann-Whitney U (MWU) test (with a significance level of 5%) between the co-occurrence distributions of the set of predictions and the corresponding background. For validating each set of ***Stage-specific PD predictions***, we define the background set of genes as genes that are expressed at a particular time point and remove *DisGeNet PD genes* and the set of ***Stage-specific PD predictions***. For validating the ***Core PD predictions***, we define the background set of genes as genes that are expressed across all time points and do not belong to the *DisGeNet PD genes* and the ***Core PD predictions***.

For a complementary validation of our predictions, we do an additional validation experiment evaluating if our sets of gene predictions are statistically significantly related to PD. We perform enrichment analysis (Supplementary Section "Enrichment analysis") (significance threshold of 0.05) in a set of PD-related genes less validated in the literature than *DisGeNet PD genes* on each set of our gene predictions against the background (defined above). The set of PD-related genes are genes that: 1) co-occur with the term "Parkinson's disease" in at least one PubMed study, in an automatic literature search of PubMed (2,031 genes) (search was performed on 13 May 2022), or 2) are in the Gene4PD database (a database containing PD associated genes extracted from more than 60 genomic data sources) (2,121 genes)[57] (collected on 07 October 2021). The overlap between the two sets is 517 genes, resulting in 3,635 unique PD-related genes.

Data availability

This paper analyzes existing, publicly available data. The data used in this work are available at 10.5281/zenodo.10391382[58].

Code availablity

All original code and reproducibility materials have been deposited at 10.5281/zenodo.10391382[58] and is publicly available as of the publication date. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Received: 30 January 2024; Accepted: 10 May 2024
Published online: 14 May 2024

References

1. Poewe, W. *et al.* Parkinson disease. *Nat. Rev. Dis. Primers* **3**, 1–21 (2017).
2. Yang, W. *et al.* Current and projected future economic burden of Parkinson's disease in the us. *NPJ Parkinson's Dis.* **6**, 1–9 (2020).
3. Balestrino, R. & Schapira, A. Parkinson disease. *Eur. J. Neurol.* **27**, 27–42 (2020).
4. Giguère, N., Burke Nanni, S. & Trudeau, L. E. On cell loss and selective vulnerability of neuronal populations in Parkinson's disease. *Front. Neurol.* **9**, 383041 (2018).
5. Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: Technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 (2020).
6. Smajić, S. *et al.* Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. *Brain* **145**, 964–978 (2022).
7. Novak, G. *et al.* Single-cell transcriptomics of human IPSC differentiation dynamics reveal a core molecular network of Parkinson's disease. *Commun. Biol.* **5**, 1–19 (2022).
8. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
9. Luecken, M. D. & Theis, F. J. Current best practices in single-cell rna-seq analysis: A tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
10. Welch, J. D. *et al.* Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
11. Thygesen, C., Boll, I., Finsen, B., Modzel, M. & Larsen, M. R. Characterizing disease-associated changes in post-translational modifications by mass spectrometry. *Expert Rev. Proteom.* **15**, 245–258 (2018).
12. Jung, I., Choi, J. & Chae, H. A non-negative matrix factorization-based framework for the analysis of multi-class time-series single-cell rna-seq data. *IEEE Access* **8**, 42342–42348 (2020).
13. Yang, J., Yang, S., Fu, Y., Li, X. & Huang, T. Non-negative graph embedding. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8 (IEEE, 2008).
14. Huizing, G.-J., Deutschmann, I. M., Peyré, G. & Cantini, L. Paired single-cell multi-omics data integration with Mowgli. *Nat. Commun.* **14**(1), 7711 (2023).
15. Argelaguet, R. *et al.* Mofa+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 1–17 (2020).
16. Hwang, T. *et al.* Co-clustering phenome-genome for phenotype classification and disease gene discovery. *Nucl. Acids Res.* **40**, e146–e146 (2012).
17. Malod-Dognin, N. *et al.* Towards a data-integrated cell. *Nat. Commun.* **10**, 1–13 (2019).
18. Peng, W., Li, L., Dai, W., Du, J. & Lan, W. Predicting protein functions through non-negative matrix factorization regularized by protein-protein interaction network and gene functional information. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*, 86–89 (IEEE, 2019).
19. Tang, X. *et al.* Indicator regularized non-negative matrix factorization method-based drug repurposing for covid-19. *Front. Immunol.* **11**, 3824 (2021).
20. Elyanow, R., Dumitrascu, B., Engelhardt, B. E. & Raphael, B. J. NETNMF-SC: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res.* **30**, 195–204 (2020).
21. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'r, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* **21**, 1–12 (2023).
22. Li, M. M. *et al.* Contextualizing protein representations using deep learning on protein networks and single-cell data. *bioRxiv* (2023).
23. Anandhan, A. *et al.* Metabolic dysfunction in Parkinson's disease: Bioenergetics, redox homeostasis and central carbon metabolism. *Brain Res. Bull.* **133**, 12–30 (2017).
24. Oughtred, R. *et al.* The biogrid interaction database: 2019 update. *Nucl. Acids Res.* **47**, D529–D541 (2019).
25. Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. & Kinoshita, K. Coxpresdb v7: A gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucl. Acids Res.* **47**, D55–D62 (2019).
26. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucl. Acids Res.* **45**, D353–D361 (2017).
27. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
28. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucl. Acids Res.* **48**, D498–D503 (2020).
29. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
30. Ghemrawi, R. & Khair, M. Endoplasmic reticulum stress and unfolded protein response in neurodegenerative diseases. *Int. J. Mol. Sci.* **21**, 6127 (2020).
31. Dong, C. *et al.* Plasma metabolite signature classifies male lrrk2 Parkinson's disease patients. *Metabolites* **12**, 149 (2022).
32. Figura, M. *et al.* Serum amino acid profile in patients with Parkinson's disease. *PLoS One.* **13**(1), e0191670 (2018).
33. Chi, J. *et al.* Integrated analysis and identification of novel biomarkers in Parkinson's disease. *Front. Aging Neurosci.* **18**(10), 178 (2018).
34. Roussel, B. D. *et al.* Endoplasmic reticulum dysfunction in neurological disease. *Lancet Neurol.* **12**, 105–118 (2013).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
36. Teyssou, E. *et al.* The amyotrophic lateral sclerosis m114t pfn1 mutation deregulates alternative autophagy pathways and mitochondrial homeostasis. *Int. J. Mol. Sci.* **23**, 5694 (2022).
37. Organization, W.H. *Guidelines for the treatment of malaria* (World Health Organization, 2015).
38. Gao, Y. *et al.* Dihydroartemisinin ameliorates LPS-induced neuroinflammation by inhibiting the pi3k/akt pathway. *Metab. Brain Dis.* **35**, 661–672 (2020).
39. Rai, S. N. *et al.* The role of pi3k/akt and ERK in neurodegenerative disorders. *Neurotox. Res.* **35**, 775–795 (2019).

40. Singh, S. S. *et al.* Nf-κb-mediated neuroinflammation in Parkinson's disease and potential therapeutic effect of polyphenols. *Neurotox. Res.* **37**, 491–507 (2020).
41. Sikora, J. & Ouagazzal, A.-M. Synaptic zinc: An emerging player in Parkinson's disease. *Int. J. Mol. Sci.* **22**, 4724 (2021).
42. Wishart, D. S. *et al.* Drugbank 5.0: a major update to the Drugbank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
43. Wishart, D. S. *et al.* Hmdb 5.0: The human metabolome database for 2022. *Nucl. Acids Res.* **50**, D622–D631 (2022).
44. Fujita, K. A. *et al.* Integrating pathways of Parkinson's disease in a molecular interaction map. *Mol. Neurobiol.* **49**, 88–102 (2014).
45. Travaglio, M. *et al.* Increased cysteine metabolism in pink1 models of Parkinson's disease. *Dis. Models Mech.* **16**, 049727 (2023).
46. Kim, K.-H. & Son, J. H. Pink1 gene knockdown leads to increased binding of parkin with actin filament. *Neurosci. Lett.* **468**, 272–276 (2010).
47. Requejo-Aguilar, R. *et al.* Pink1 deficiency sustains cell proliferation by reprogramming glucose metabolism through hif1. *Nat. Commun.* **5**, 4514 (2014).
48. Azkona, G. *et al.* Lrrk2 expression is deregulated in fibroblasts and neurons from Parkinson patients with mutations in pink1. *Mol. Neurobiol.* **55**, 506–516 (2018).
49. Shah, K. & Rossie, S. Tale of the good and the bad cdk5: Remodeling of the actin cytoskeleton in the brain. *Mol. Neurobiol.* **55**, 3426–3438 (2018).
50. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
51. Lopes, F. M., Bristot, I. J., Da Motta, L. L., Parsons, R. B. & Klamt, F. Mimicking Parkinson's disease in a dish: Merits and pitfalls of the most commonly used dopaminergic in vitro models. *NeuroMol. Med.* **19**, 241–255 (2017).
52. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**, 4164–4169 (2004).
53. Vavasis, S. A. On the complexity of nonnegative matrix factorization. *SIAM J. Optim.* **20**, 1364–1377 (2010).
54. Ding, C., Li, T., Peng, W. & Park, H. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 126–135 (2006).
55. Qiao, H. New svd based initialization strategy for non-negative matrix factorization. *Pattern Recogn. Lett.* **63**, 71–77 (2015).
56. Gligorijević, V., Malod-Dognin, N. & Pržulj, N. Patient-specific data fusion for cancer stratification and personalised treatment. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, 321–332 (World Scientific, 2016).
57. Li, B. *et al.* Gene4PD: A comprehensive genetic database of Parkinson's disease. *Front. Neurosci.* **15**, 679568 (2021).
58. Mihajlovic, K. KatarinaMihajlovic/NetSCNMTF-2stepmining: NetSCNMTF-2stepmining_DOI, https://doi.org/10.5281/zenodo.10391383, https://doi.org/10.5281/zenodo.10391383 (2023).

Acknowledgements

This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2019-105500GB-I00 / AEI / 10.13039/501100011033 and grant PID2022-141920NB-I00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536. DK was supported by the PRIDE program of the Luxembourg National Research Fund through PRIDE17/12244779/PARK-QC.

Author contributions

Conceptualization, N.P. and N.M.D.; Methodology, N.P., N.M.D., G.C. and K.M.; Validation, K.M.; Formal Analysis, K.M.; Investigation, K.M.; Resources, N.P., A.S., G.N. and D.K.; Writing - Original Draft, G.C. and K.M.; Writing - Review & Editing, N.P., A.S., N.M.D. and K.M.; Visualization, K.M.; Supervision, N.P.; Funding Acquisition, N.P. and A.S. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-61844-3.

Correspondence and requests for materials should be addressed to N.P.

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

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