Statistical integration of multi-omics data identifies potential therapeutic targets for synucleinopathies

Said el Bouhaddani[a,*], Matthias Höllerhage[b], Hae-Won Uh[a], Günter U. Höglinger[b,**], Jeanine J. Houwing-Duistermaat[c,**]

[a] *Dept. of Data Science and Biostatistics, Division Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX, Utrecht, Netherlands*

[b] *Department of Neurology, Hannover Medical School, Carl-Neuberg-Str. 1, D-30625, Hannover, Germany*

[c] *Dept. of Statistical Sciences, University of Bologna, Via Zamboni 33, 40126, Bologna, Italy*

Abstract

Parkinson's disease and multiple system atrophy are characterized by aggregation of the α-synuclein protein in the brain. The causes and consequences of this aggregation are currently unclear, hampering the development of an effective therapy against these synucleinopathies. To gain understanding in the neurobiological events underlying α-synuclein aggregation, LUHMES cell lines have been utilized to measure transcriptomics and proteomics data and perform differential regulation analyses. The LUHMES cells have also been used to screen 1600 FDA-approved drugs for their ability to reduce α-synuclein-induced toxicity.

We implemented a probabilistic data integration method, POPLS-DA, to model α-synuclein overexpression in terms of the transcriptomics and proteomics data jointly. Using POPLS-DA, we selected 200 relevant genes and proteins and integrated this gene list with the FDA drugs that were validated to be protective against α-synuclein-induced toxicity. Using interactome and functional enrichment analyses, we identified clusters of synaptic and lysosome-related genes and proteins that were targeted by the protective drugs.

We found that *HSPA5*, a member of the heat shock protein 70 family, was one of the most targeted genes by the validated drugs. Our integrative approach results provide new directions for therapeutic targets for synucleinopathies.

Keywords: Alpha-synuclein, cell lines, latent variable modeling, probabilistic data integration, interactome

*Corresponding author. Email: `s.elbouhaddani@umcutrecht.nl`

**Equal contributions.

1. Introduction

Synucleinopathies comprise neurodegenerative disorders, the common denominator of which is the pathological aggregation of α-synuclein, encoded by the *SNCA* gene. The most common synucleinopathy is Parkinson's disease (PD), characterized by the presence of α-synuclein aggregates in the form of Lewy bodies in neuronal cells. Clinically, PD is characterized by the presence of bradykinesia in combination with rigidity and/or tremor at rest. Another neurodegenerative synucleinopathiy is multiple system atrophy (MSA) that affects the nigrostriatal system, the cerebellum, and the autonomic nervous system, with a typical survival time of 6 to 10 years. Clinically, MSA is characterized by the presence of autonomic dysfunction, as well as parkinsonism or cerebellar symptoms, which determine the parkinsonian (MSA-P) and cerebellar subtype (MSA-C). In MSA, proteinaceous inclusions of α-synuclein can be found in oligodendrocytes, glial cytoplasmic inclusions, and neurons. In general, aggregated forms of α-synuclein are assumed to be toxic, leading to cell death and neurodegeneration. The mechanisms of how α-synuclein confers toxicity are not yet fully understood. Current therapies are purely symptomatic, and anti-Parkinsonian drugs are much less effective than for PD. A neuroprotective or disease modifying therapy for synucleinopathies is yet to be found. Omics research can provide insight in biological processes and molecular functions and interactions. Its ultimate goal is to identify biomarkers and drug targets.

LUHMES cell lines, which are derived from human embryonic dopaminergic cells [1], are a rich source to study the neurobiological processes underlying neurodegeneration. A recent study used LUHMES cells overexpressing α-synuclein and GFP as a control protein, respectively, for omics research [2]. Transcriptomic and proteomic differential expression analyses have been performed. In both analyses, statistically significant results were found, but only little overlap between the two datasets appeared to be present. Specifically, enrichment analysis of the transcriptomic data identified genes involved in cell death and apoptosis, showing regulation of cell death-associated genes in the affected cells as a consequence of α-synuclein overexpression. Enrichment analyses of the proteomics data identified gene ontology terms "vesicular trafficking", "synapses" and "lysosomes". These terms have been linked to PD. Since the analyses were performed on each omics dataset separately, a holistic overview of gene deregulation in α-synuclein overexpressing cells is lacking. Such an overview can be obtained by using an multi-omics approach [3], and provides a unified list of the most relevant features. Moreover, this list can comprise features that were not found with single omics analyses, because of a lack of statistical evidence when analyzed separately. For example, genes whose transcripts and proteins were close to being significantly differentially expressed in two separate analyses might become significant in an integrated multi-omics analysis.

To identify drugs that protect against α-synuclein-induced cell death, LUHMES cells were used to screen 1600 FDA-approved drug compounds [4]. The screening revealed 41 drugs that significantly reduced cell death in LUHMES cells overexpressing α-synuclein. Dipyridamole, a non-specific phosphodiesterase inhibitor, was found most promising to reduce α-synuclein toxicity. However, this drug does not pass the blood-brain-barrier and therefore cannot be directly used as a treatment. The screening results can be

combined with results of a multi-omics data analysis by integrating genes and proteins targeted by the protective drug compounds with the prioritized genes and proteins from the omics analyses.

For the analysis of multiple omics data in cell lines, the multivariate integrative (MINT) method can be used. MINT is an extension of partial least squares discriminant analysis (PLS-DA) [5] for multiple omics levels. It estimates linear combinations of features that maximally discriminate between experimental groups across all omics datasets. However, this method does not model omics-specific variation. For heterogeneous datasets, several studies have shown that modeling omics-specific variation yields better results [6, 7]. To integrate transciptomics and proteomics, including data specific components can be beneficial as these data are different in, e.g., scale and measurement platform. For clinical and epidemiological studies, the multi-omics integration approaches, algorithmic and probabilistic two-way orthogonal partial least squares (O2PLS [8] resp. PO2PLS [9]), model the systematic difference between datasets. For small datasets, algorithmic approaches tend to overfit and a probabilistic approach is preferred [7]. However, these methods cannot directly be applied to data from cell lines, because they require that the datasets are measured on the same samples. We will adapt the probabilistic approach to multiple omics data measured in cell lines by linking the datasets via genes instead of samples.

We present an integrated analysis of transcriptomics, proteomics, and FDA-approved drugs data using a novel probabilistic approach named POPLS-DA. We applied POPLS-DA to integrate both transcriptomic and proteomic datasets and identified relevant variables that are associated with the two experimental conditions (α-synuclein versus control protein overexpression). Next, based on the genes that correspond to these relevant variables, a protein-protein interaction (PPI) network was constructed and the genes that are targeted by the drugs were identified. We used String-DB, where a link between genes indicates evidence for a biologically plausible relation.

2. Methods

2.1. Processing samples and measuring data

The performance of the screening of 1600 FDA-approved drugs and the acquisition of proteome and transcriptome data are described in detail elsewhere[4, 2] and summarized here in brief.

High-throughput drug screening and validation. For the drug screening, LUHMES cells were plated in flasks coated with poly-L-lysine (0.1 mg/ml) and fibronectin (5 μg/ml) in growth medium (DMEM/F12 with 1% N2 supplement and 0.04 μl/ml basic fibroblast growth factor). After 24h, the medium was changed to differentiation medium (DMEM/F12 with 1% N2-supplement, 1 μg/ml tetracycline, 0.49 μg/ml N6,2'-O-dibutyryladenosine 3',5'-cyclic monophosphate, 2 ng/ml glial derived neurotrophic factor. After another 24 hours, the cells were transduced with adenoviral vectors leading to overexpression of human wild-type α-synuclein. On the day after transduction, viral vectors were washed away and the cells were replated on 384-well multi-well plates and

treated with the different drugs solved in dimethyl sulfoxide. In the primary screening, all drugs were tested in 3 μM and 10 μM. For validation, dose-response analyses were performed in concentrations between 0.6 to 80 μM, depending on the effect at 20 μM. Compounds that already showed a plateau protection or toxicity at 20 μM were not tested in higher concentrations.

Transcriptome and proteome data acquisition and processing. For the transcriptome and proteome data, cells were plated directly in differentiation medium and transduced with adenoviral vectors leading to overexpression of either human wild-type α-synuclein or GFP, repectively. After 24 hours, remaining viral vectors were removed. Samples were collected four days after transduction. Transcriptome data were measured with Illumina HumanHT-12_V3 bead chips (Illumina, San Diego, CA, USA). Proteome data were measured by liquid chromatography mass spectrometry (LC-MS/MS). The transcription probes were filtered based on the Illumina detection p-value, where probes were removed when all samples had a p-value above 0.05. A variance filter was applied where probes were removed if the inter-quantile range (IQR) of a probe didn't exceed the 0.5-th quantile of all IQRs. Probes that couldn't be mapped to an ENTREZ identifier were removed.

After pre-processing and filtering, the transcriptomics and proteomics datasets contained 15660 transcription probes and 2577 proteins, respectively, that could be mapped to Entrez identifiers. 2292 of these identifiers were present in both datasets. We excluded the *SNCA* measurements from the data analysis. For the rest of this paper, we refer to 'genes' as the term for the transcripts and proteins with the same Entrez identifier.

For the transcriptomics data, we have three samples from LUHMES cells overexpressing α-synuclein, and three samples from LUHMES cells expressing GFP as a control protein. For proteomics, the numbers are equal to nine. The correlation of the transcripts with the α-synuclein overexpressing cells versus the control cells was taken as a reference; if this correlation was negative, the corresponding protein feature was multiplied by minus one. The measurements for both datasets were scaled to have zero mean and unit variance.

2.2. Probabilistic data integration with POPLS-DA

For the joint analysis of the transcriptomics and proteomics datasets, a novel statistical data integration method, probabilistic orthogonal partial least squares discriminant analysis (POPLS-DA) was applied. Details can be found in the online supplement, Section S2.

Briefly, let x_t and x_p be random vectors representing the 2292 transcriptomics and proteomics data with the same Entrez identifier, respectively. Similarly, let y_t (for transcriptomics) and y_p (for proteomics) be univariate random variables encoding the two experimental conditions (one for cells overexpressing α-synuclein, and zero for the controls). In the POPLS-DA model, the group status and omics data are linked via latent variables u_t and u_p of size r, with r much smaller than the data dimensions. These latent variables represent joint parts of the two omics datasets that are correlated with the group status. To model omics variables that contribute to the variation in the omics

dataset, but do not play a role in distinguishing between the two experimental groups, specific components v_t and v_p were added to the model. The numbers of components here are denoted by r_t and r_p, respectively. The number of joint and specific components, r, r_t, and r_p have to be specified and can be based on scree plots [10] of the eigenvalues of the stacked datasets. Residual variation was modeled by including noise terms e_t and e_p for the transcriptomics and proteomics datasets, and ϵ_t and ϵ_p for the random variable representing the groups. All random variables were assumed to be normally distributed with zero mean. The mathematical model for POPLS-DA can be written as

$$\begin{aligned} x_t &= u_t W^{\mathrm{T}} + v_t P_t^{\mathrm{T}} + e_t, & x_p &= u_p W^{\mathrm{T}} + v_p P_p^{\mathrm{T}} + e_p, \\ y_t &= u_t \beta + \epsilon_t, & y_p &= u_p \beta + \epsilon_p. \end{aligned} \tag{1}$$

The matrix W contains the joint loadings, and the matrices P_t and P_p contain the specific loadings for each gene in each component. The vector β contains the regression coefficients of y_t on the joint components u_t. Note that these regression coefficients are the same for y_p on the joint components u_p. The vector $W\beta$ represents effect sizes for how well each gene, represented by its transcript and protein, discriminates between the two experimental conditions.

The parameters of POPLS-DA were estimated with maximum likelihood. See the online supplement (Section S3) for more details. To select a set of relevant genes, the obtained effect sizes $W\beta$ were ranked and plotted. The elbow criterion [10] was used to determine the threshold between relevant and irrelevant genes.

2.3. FDA drug compound analysis

The relevant genes identified by POPLS-DA were further analyzed using databases on protein-protein interactions (PPI) and drug targets, as well as gene ontology and gene-disease association databases.

The PPI network of the relevant genes (in short, αSyn-PPI) was estimated using the String-DB [11] website and the `STRINGdb` package in R [12]. In the String-DB analysis, text mining evidence was excluded to base the network only on curated, computational, and experimental data sources. A medium evidence threshold (0.4) was used to define an interaction (edge in the network).

To integrate information on FDA drug compounds with the obtained αSyn-PPI network, the lists of genes targeted by the compounds that significantly reduced toxicity of α-synuclein aggregation were obtained from the DrugBank database [13]. By using the `string_db$get_neighbors` function, the direct neighbors of these targeted genes in the String-DB database were added to the gene lists for each compound. Then, for each gene in the αSyn-PPI network, the number of compounds targeting that gene was determined. Genes in the αSyn-PPI network were colored according to this number using the "payload" function on the String-DB website.

To identify sub-clusters of genes in the αSyn-PPI network, the Markov cluster algorithm (MCL) [14] on the String-DB website was used. The inflation parameter was set to two.

Finally, GO pathway and disease enrichment analyses were performed. For GO, the `goana` function from the `limma` package [15] in R was used, with the whole human genome as background. For disease enrichment analysis, DisGeNet [16], a disease-gene association database implemented in the `disgenet2r` R package, was used. Here, the "BEFREE" text mining database was used for the disease-gene enrichment score as it contains more annotations regarding MSA.

The methodology is presented as a workflow in Figure 1.

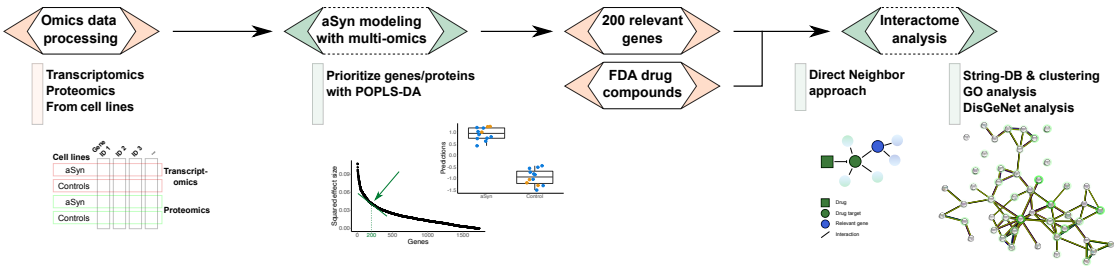

Figure 1: **Workflow of the multi-omics integration approach.** (1) To link the two datasets, the transcripts and proteins were mapped to their Entrez identifier, referred to as genes. (2) To identify the relevant genes that discriminate between the two experimental groups (α-synuclein overexpression vs control), POPLS-DA was applied. (3) The prioritized genes were integrated with the FDA drug compound data using a direct neighbor approach. (4) Bioinformatics analyses: interactome, GO and DisGeNet enrichment analyses analyses. Protein-protein interaction networks were built using String-DB where each node is a gene. The genes that were direct neighbors of a drug target were highlighted with a green halo, where its color intensity is proportional to the number of direct neighbors. The abbreviations aSyn, GO and FDA stand for α-synuclein, gene ontology, and Food and Drug Administration, respectively.

3. Results

POPLS-DA was used to identify relevant genes by analyzing the pre-processed transcriptomics (x_t) and proteomics (x_p) datasets. Based on a scree plot of the eigenvalues of the row-wise concatenated x_t and x_p, we choose two joint and two specific components (Supplementary Fig. S1). Based on a scree plot of $W\beta$, presented in Figure 2, 200 genes were retained for further analysis. A list of all genes and their weights are shown in Supplementary Table S1.

Of the total variance in the transcriptomics and proteomics data, 21 and 18 percent, respectively, were explained by the integrative parts. Omics-specific variation accounted for around 27 percent of the total variation. The selected relevant genes appeared to be able to distinguish between the two experimental groups (Figure 2 and Supplementary Fig. S2).

3.1. Omics and FDA data integration with String-DB

Figure 3 shows the protein-protein interaction network (αSyn-PPI) constructed using String-DB. The network had significantly more interactions than expected from an arbitrary subset of 200 genes. The network enrichment p-value was smaller than 10^{-15}.

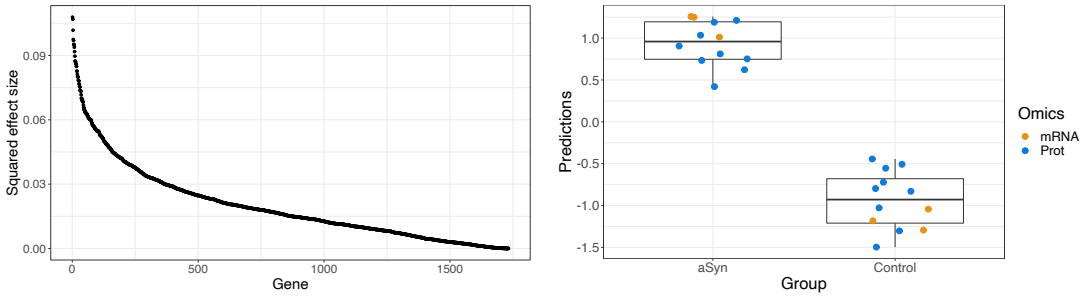

Figure 2: **Selection of relevant genes and their ability to discriminate.** The *left panel* shows the sorted squared effect size per gene given by the elements of $W\beta$. The 200 relevant genes correspond to the 'elbow', which is visually determined. The *right panel* shows boxplots of score predictions based on the selected 200 relevant genes. The dots, representing individual samples, are added with a horizontal 'jitter' to reduce overlap. The transcriptomics and proteomics samples are colored in orange and blue, respectively. The abbreviations aSyn, miRNA, and Prot stand for α-synuclein, transcriptome, and proteome, respectively.

From the 41 drugs that were validated to reduce α-synuclein-induced toxicity [4], we found 27 targeting at least one gene in the αSyn-PPI network.

The *HMGCR*, *NQO1*, *HSPA5*, and *YBX1* genes were most targeted by the validated drugs, namely by 20, 18, 17 resp. 17 compounds.

Table 3.1 shows for each compound the number of direct neighbors it targets in the network. Dexibuprofen, a non-steroidal anti-inflammatory drug (NSAID), targeted the highest number of direct neighbors (N=65), followed by imatinib (N=55), which is acting on the tyrosine kinase enzyme and is involved in apoptosis.

Drug name	**Direct neighbors**	**Drug name**	**Direct neighbors**
Dexibuprofen	60	Flunarizine	14
Imatinib	47	Quinacrine	13
Amiodarone	44	Pentoxyverine	10
Risperidone	41	Dicyclomine	8
Bepridil	36	Doxazosin	8
Astemizole	26	Ajmaline	7
Telmisartan	24	Ifenprodil	7
Amlodipine	21	Lomerizine	7
Benazepril	18	Pentamidine	6
Reserpine	18	Guanfacine	5
Dipyridamole	16	Dyclonine	4
Triflupromazine	16	Tropisetron	3
Nefazodone	15	Clemastine	1
Trimipramine	15		

Table 1: **FDA-approved drugs protective against α-synuclein-induced toxicity with direct neighbors of the top 200 genes.** For each drug, the number of genes from the top 200 is shown for which any of its target has at least one direct interactor.

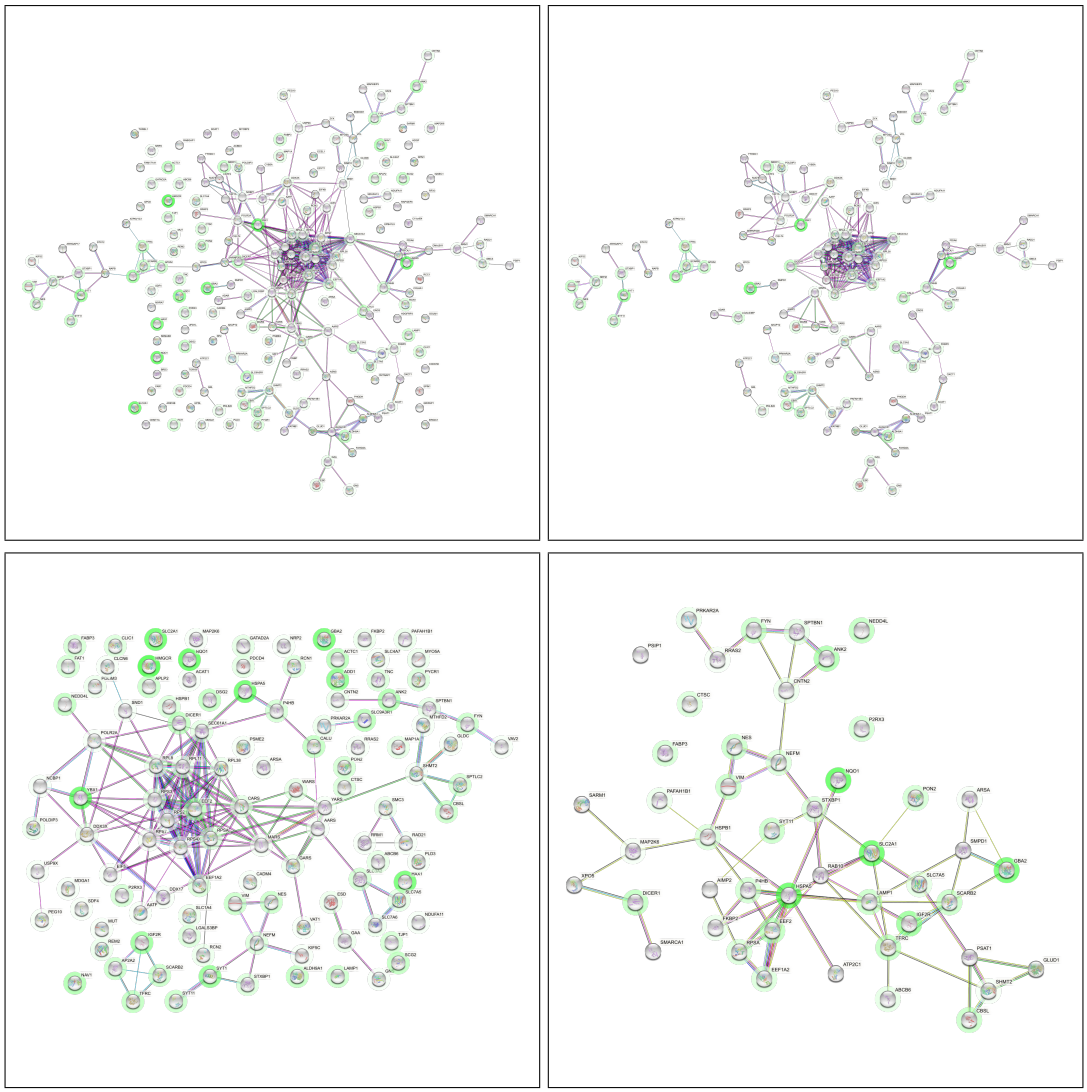

Figure 3: **String-DB and clustering analyses of the top 200 genes.** Figures are numbered from left to right, from top to bottom. In panel (a) a network of interactions between top 200 genes (estimated with POPLS-DA) was constructed using String-DB. Each node is a gene, and a connection between genes indicates evidence for a biologically plausible link. Text-mining was excluded as an evidence source, and a medium confidence threshold was used. For genes that were (indirectly) targeted by a drug compound, a green 'halo' is drawn. The intensity of the green color is proportional to the number of drug compounds for which the gene was a (indirect) target. In panel (b) the interaction network from (a) was clustered using the MCL clustering algorithm from the String-DB website. The edges between the clusters are removed for visual aid. In panel (c) an interaction network is shown for a druggable subset of the top 200 genes, consisting of 116 genes that were (indirectly) targeted by an FDA-approved drug compound. In panel (d) an interaction network of top genes in the "Parkinson's disease" DisGeNet term was constructed using String-DB. Text mining as evidence was included here.

Using the MCL clustering algorithm on the αSyn-PPI network, the network was split into sub-clusters, see Figure 3. Several clusters consist of genes that are involved in neurogenerative disorders. The largest cluster is a highly connected hub of 22 genes, mostly ribosomal proteins. Ribosomal proteins have been reported to mediate neurodegeneration in PD [17]. Another cluster of five genes includes the top two genes *MTHFD2* and *CBS*. The *CBS* (cystathionine-beta-synthase) gene was protective in a PD toxin model in rats [18]. The network also contains clusters with genes involved in neurological development, e.g. *TFRC* (transferrin receptor), and neuronal adhesion and growth, e.g. *TNC* (tenascin C). Further, four genes encoding intermediate filaments are clustered together, such as *NES* (nestin). They are primarily responsible for cell adhesion [19]. Furthermore, a cluster of three genes from the solute carrier family was found (*SLC3A2*, *SLC7A5*, and *SLC7A6*). These genes may play a role in the transport across the blood-brain barrier.

3.2. Enrichment analyses with GO and DisGeNet

In Table 3.2, the most significant pathways enriched in the top 200 genes are shown. The GO terms "intracellular" and "cytoplasm" were most significant. Furthermore, enrichment was found for members of the extracellular region and space terms (extracellular exosome, vesicle, and organelle). Genes with these annotations are involved in vesicular transport and signaling [20]. Nine of the ten terms are from the "cellular compound" GO category. The term 'cellular protein localization' is from the 'biological process' category.

Term	**Ontology**	**Genes in set**	**FDR**
cytoplasm	CC	178	1.28e-19
extracellular exosome	CC	71	4.67e-17
extracellular vesicle	CC	71	5.51e-16
extracellular organelle	CC	71	5.51e-16
intracellular anatomical structure	CC	190	7.62e-12
cell junction	CC	61	7.62e-12
vesicle	CC	88	2.30e-11
extracellular space	CC	81	7.09e-11
protein binding	CC	179	7.65e-11
cytosol	MF	100	1.48e-10

Table 2: **Gene ontology enrichment analysis of the top 200 genes.** Out of 1732 genes, the 200 genes with the highest weight (estimated using POPLS-DA) were used to perform GO enrichment analysis. All three GO ontologies were considered: cellular compound (CC), biological process (BP), and molecular function. For each term, a p-value was computed using Fisher's exact test with the whole genome as background and corrected for multiple testing. The ten most significant terms are shown, ranked by their GO enrichment FDR-adjusted p-value. The column "Genes in set" shows how many of the top 200 genes were annotated with the respective term.

Table 3.2 shows the five diseases for which the most significant enrichment was found in the top 200 genes. The second term is "Parkinson's disease" and contained 48 genes.

Figure 3 shows the PPI network for genes in this disease term. Note that the *HSPA5* gene, which was one of the most targeted by the FDA drug compounds, is in this network.

Term	**Genes in set**	**FDR**
Neoplasms	142	2.84e-07
Parkinson's Disease	48	1.94e-06
Carcinogenesis	100	3.36e-06
Malignant neoplasm of prostate	75	2.54e-05
Neoplasm Metastasis	98	2.89e-05

Table 3: **Top five disease enrichment clusters based on DisGeNet.** The top 200 relevant genes were analyzed using DisGeNet, a disease-gene database. The BEFREE database was used. For each term, a p-value was calculated using a Fisher exact test and FDR adjusted. The terms are ranked by this adjusted p-value.

Finally, enrichment analyses were performed for the subset of 116 genes that were direct neighbors of genes targeted by at least one of the drugs. Figure 3 shows the protein-protein interaction network for this subset. The corresponding network enrichment p-value was smaller than 10^{-10}. Also, the most significant GO terms agreed with the GO analysis of the total set of 200 genes (Supplementary Table S2).

4. Discussion

In this study, we integrated transcriptomics and proteomics data, as well as data on the protection of FDA-approved drugs against α-synuclein-induced toxicity from LUHMES cells overexpressing α-synuclein and those overexpressing GFP as a control protein. After mapping all variables to common gene identifiers, we applied POPLS-DA to the two omics datasets. We found 200 relevant genes based on an elbow plot of the squared effect size of each gene (Figure 2). A protein-protein interaction network was constructed in which the FDA drug compound data was integrated by identifying direct neighbors of their drug targets in the network. These genes appeared to perfectly discriminate between the two experimental groups. Interactome analyses revealed several genes and pathways linked to PD, the most common synucleinopathy.

The drugs that targeted the highest number of direct neighbors were Dexibuprofen and Imatinib. Both drugs were previously found to lower the risk of PD in epidemiological studies [21, 22].

The genes that were targeted by the majority of drugs are *HMGCR*, *NQO1*, *HSPA5*, and *YBX1*. Note that *HMGCR* (3-Hydroxy-3-Methylglutaryl-CoA reductase) is the main target of statins. *NQO1* (NAD(P)H quinone dehydrogenase 1) has multiple functions and has been linked to Alzheimer's disease [23]. The *HSPA5* gene (heat shock protein family A member 5) is a heat shock protein within the *HSPA* (*HSP70*) family localized to the endoplasmic reticulum [**?**]. It is found to be involved in maintaining correct folding behavior of proteins [24]. Most notably, the HSP70 family has shown to be protective against α-synuclein-induced cell death [**?**]. The *YBX1* gene (Y-box binding protein 1) is a cold shock domain protein, predominantly localized to neurons and

essential for brain development [25]. Furthermore, several clusters were observed in the network involving ribosomal proteins and intermediate filaments. Enrichment analyses identified several terms involving "extracellular" and "vesicles".

Using DisGeNet, the second most significantly enriched disease term was "Parkinson's disease", with 48 genes from our list of relevant genes. The other disease terms involved neoplasms and carcinogenesis. Interestingly, previous studies already suggested shared molecular mechanisms between PD and cancer [26, 27, 28].

In a previous study on these data [2], the authors identified significant deregulation of synaptic proteins, in particular *STXBP1*, *STX1B*, *SYT1*, and *FYN*. In our POPLS-DA integration, we found three of these genes: *STXBP1*, *SYT1*, and *FYN*, ranked 16, 18, and 72, respectively. We identified two additional synaptic genes in the top 200, namely *STX12* as well as *SYT11* which is linked to PD [29]. Additionally, they reported deregulation of lysosome related proteins, *SCARB2*, *CTSB*, *SMPD1*. Our list of top genes contained 22 genes annotated with the lysosome GO term. From these 22 genes are the previously found *SCARB2*, with position 7 in our top genes, as well as *SLC7A5* and *SLC3A2* who had position 5 and 6, respectively. The authors pointed to lysosome membrane protein 2 expression (encoded by *SCARB2*) as a therapeutic target to increase lysosomal glucocerebrosidase 458 activity to promote *alpha*-synuclein clearance. Note that one of our clusters in Figure 3 consists of genes encoding lysosome related proteins (e.g., *SCARB2*, *AP2A2*, *ATP6V1C1*, *IGF2R*, and *TFRC*) that are also direct neighbors of the drug compounds.

The data used in this study are measured on cell lines, where each dataset is measured on a different sample. Most of the data integration methods require multiple datasets measured on the same samples to calculate correlations across omics datasets from which joint components can be derived. These methods are sometimes referred to as horizontal integration approaches, where the data are laid out next to each other with the samples across the rows. In this paper, we adapted a framework for horizontal integration, namely PO2PLS, to vertical integration, where different groups of samples are laid down below each other with the same (type of) variables across the columns. A drawback of such an approach is that only genes with observations in both datasets can be analyzed.

We utilized four bioinformatics databases in our analysis: String-DB, GeneBank, GO, and DisGeNet. Several databases are available with information on interaction between genes and other molecular variables (interactomes) [30]. We chose String-DB since it performs well with regard to disease prediction [31]. However, Mosca et al [31] showed that other databases might comprise complementary information. Currently, there is no method that can utilize all information across different interactomes. Information bias is another unsolved problem when using these databases: less studied diseases have less coverage in these databases [32]. Since MSA is rare and a less studied disease, the bioinformatics interpretation of our results are likely to be biased towards well-studied disease. Indeed, in DisGeNet, MSA has only 19 annotated genes, while neoplasms have more than 10000 genes. This explains that our enrichment analysis based on DisGeNet (Table 3.2) showed many genes annotated with "neoplasms".

The POPLS-DA method can integrate more than two datasets to distinguish between

experimental conditions in cell lines if the variables in these datasets can be mapped to the same unit, i.e. genes. For example, SNPs from a GWAS study can be integrated by mapping them to genes. Here the distance of a SNP to a gene or its transcription factor binding site [33] can be used. For methylation, CpG sites are commonly mapped to genes based on their distance to a transcription start site [34]. Since typically, multiple SNPs and CpG sites are mapped to the same gene, an aggregation step is needed to obtain one measurement per gene. For example, the first principal component of the SNPs or CpG sites corresponding to a gene can be taken as the final aggregated measurement.

The FDA drug compound data were integrated and projected onto the interaction network of the 200 relevant genes obtained from the transcriptomics and proteomics data analyses. An alternative approach is to augment the data on transcriptomics and proteomics with the screening data on the compounds. It has been shown that including such information via penalties into data integration analysis yields robust results [35]. Such a joint model can pinpoint genes that can discriminate the two groups while targeted by the compounds, improving the clinical relevance of the list of top genes.

Conclusion. We validated 41 FDA-approved drugs to be protective against α-Synuclein-induced toxicity. We applied a novel data integration approach to combine these interventional data with experimental multi-omics data. We identified a set of 200 relevant genes that discriminated between samples overexpressing α-Synuclein and controls, as well as the drugs they are targeted by. These findings can potentially be used to develop therapeutic targets for PD, MSA, and other synucleinopathies.

Sources of funding

All authors were funded by ERA-Net E-Rare JTC 2018 (MSA-omics) [40-44000-98-2006/ 90030376507]. JD was supported by European Union's Horizon 2020 (IMforFUTURE) [721815]. GH was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (EXC 2145 SyNergy – ID 390857198) and within the Hannover Cluster RESIST (EXC 2155 – - project number 39087428), Deutsche Forschungsgemeinschaft (DFG, HO2402/18-1 MSAomics), the Bavarian Ministry for Education, Culture, Science and Art (Grant 8810001412 ForIPS), Niedersächsisches Ministerium für Wissenschaft und Kunst (MWK, ZN3440.TP): REBIRTH – Forschungszentrum für translationale regenerative Medizin, VolkswagenStiftung (Niedersächsisches Vorab), Petermax-Müller Foundation (Etiology and Therapy of Synucleinopathies and Tauopathies).

References

[1] J. Lotharius, Progressive Degeneration of Human Mesencephalic Neuron-Derived Cells Triggered by Dopamine-Dependent Oxidative Stress Is Dependent on the Mixed-Lineage Kinase Pathway, J. Neurosci. 25 (27) (2005) 6329–6342. doi:10.1523/JNEUROSCI.1746-05.2005.

[2] M. Höllerhage, M. Stepath, M. Kohl, K. Pfeiffer, O. W. ho Chua, L. Duan, F. Hopfner, M. Eisenacher, K. Marcus, G. U. Höglinger, Transcriptome and Proteome Analysis in LUHMES Cells Overexpressing Alpha-Synuclein, Front. Neurol. 13 (April) (2022) 1–19. doi:10.3389/fneur.2022.787059.

[3] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype–phenotype interactions, Nat. Rev. Genet. 16 (2) (2015) 85–97. doi:10.1038/nrg3868.

[4] M. Höllerhage, C. Moebius, J. Melms, W.-H. Chiu, J. N. Goebel, T. Chakroun, T. Koeglsperger, W. H. Oertel, T. W. Rösler, M. Bickle, G. U. Höglinger, Protective efficacy of phosphodiesterase-1 inhibition against alpha-synuclein toxicity revealed by compound screening in LUHMES cells, Sci. Rep. 7 (1) (2017) 11469. doi:10.1038/s41598-017-11664-5.

[5] L. Ståhle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study, J. Chemom. 1 (3) (1987) 185–196. doi:10.1002/cem.1180010306.

[6] J. Trygg, S. Wold, O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter, J. Chemom. 17 (1) (2003) 53–64. doi:10.1002/cem.775.

[7] S. el Bouhaddani, H.-W. Uh, G. Jongbloed, C. Hayward, L. Klarić, S. M. Kiełbasa, J. Houwing-Duistermaat, Integrating omics datasets with the OmicsPLS package, BMC Bioinformatics 19 (1) (2018) 371. doi:10.1186/s12859-018-2371-3.

[8] S. el Bouhaddani, J. Houwing-Duistermaat, P. Salo, M. Perola, G. Jongbloed, H.-W. Uh, Evaluation of O2PLS in Omics data integration, BMC Bioinformatics 17 (S2) (2016) S11. doi:10.1186/s12859-015-0854-z.

[9] S. el Bouhaddani, H.-W. Uh, G. Jongbloed, J. Houwing-Duistermaat, Statistical Integration of Heterogeneous Data with PO2PLS (mar 2021). arXiv:2103.13490.

[10] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate analysis, Academic Press, 1979.

[11] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. von Mering, STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res. 47 (D1) (2019) D607–D613. doi:10.1093/nar/gky1131.

[12] R Core Team, R: A Language and Environment for Statistical Computing (2014).

[13] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration., Nucleic Acids Res. 34 (Database issue) (2006) D668–D672. doi:10.1093/nar/gkj067.

[14] A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families., Nucleic Acids Res. 30 (7) (2002) 1575–84. doi:10.1093/nar/30.7.1575.

[15] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47–e47. doi:10.1093/nar/gkv007.

[16] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, L. I. Furlong, The DisGeNET knowledge platform for disease genomics: 2019 update, Nucleic Acids Res. 48 (D1) (2020) D845–D855. doi:10.1093/nar/gkz1021.

[17] I. Martin, J. W. Kim, B. D. Lee, H. C. Kang, J.-C. Xu, H. Jia, J. Stankowski, M.-S. Kim, J. Zhong, M. Kumar, S. A. Andrabi, Y. Xiong, D. W. Dickson, Z. K. Wszolek, A. Pandey, T. M. Dawson, V. L. Dawson, Ribosomal Protein s15 Phosphorylation Mediates LRRK2 Neurodegeneration in Parkinson's Disease, Cell 157 (2) (2014) 472–485. doi:10.1016/j.cell.2014.01.064.

[18] W.-l. Yin, W.-g. Yin, B.-s. Huang, L.-x. Wu, Neuroprotective effects of lentivirus-mediated cystathionine-beta-synthase overexpression against 6-OHDA-induced parkinson's disease rats, Neurosci. Lett. 657 (2017) 45–52. doi:10.1016/j.neulet.2017.07.019.

[19] R. Perrot, J. Eyer, Intermediate Filaments in Neurodegenerative Diseases, in: Neurodegener. Dis., InTech, 2013. doi:10.5772/54676.

[20] C.-Y. Kao, E. T. Papoutsakis, Extracellular vesicles: exosomes, microparticles, their parts, and their targets to enable their biomanufacturing and clinical applications, Curr. Opin. Biotechnol. 60 (2019) 89–98. doi:10.1016/j.copbio.2019.01.005.

[21] K. Rees, R. Stowe, S. Patel, N. Ives, K. Breen, C. E. Clarke, Y. Ben-Shlomo, Non-steroidal anti-inflammatory drugs as disease-modifying agents for Parkinson's disease: evidence from observational studies, Cochrane Database Syst. Rev. (11) (nov 2011). doi:10.1002/14651858.CD008454.pub2.

[22] M. Kumar, R. Kulshrestha, N. Singh, A. S. Jaggi, Expanding spectrum of anticancer drug, imatinib, in the disorders affecting brain and spinal cord, Pharmacol. Res. 143 (2019) 86–96. doi:10.1016/j.phrs.2019.03.014.

[23] N. Simidjievski, C. Bodnar, I. Tariq, P. Scherer, H. Andres Terre, Z. Shams, M. Jamnik, P. Liò, Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice, Front. Genet. 10 (2019) 1205. doi:10.3389/fgene.2019.01205.

[24] M. S. Lyon, C. Milligan, Extracellular heat shock proteins in neurodegenerative diseases: New perspectives, Neurosci. Lett. 711 (2019) 134462. doi:10.1016/j.neulet.2019.134462.

[25] N. V. Bobkova, D. N. Lyabin, N. I. Medvinskaya, A. N. Samokhin, P. V. Nekrasov, I. V. Nesterova, I. Y. Aleksandrova, O. G. Tatarnikova, A. G. Bobylev, I. M. Vikhlyantsev, M. S. Kukharsky, A. A. Ustyugov, D. N. Polyakov, I. A. Eliseeva, D. A. Kretov, S. G. Guryanov, L. P. Ovchinnikov, The Y-Box Binding Protein 1 Suppresses Alzheimer's Disease Progression in Two Animal Models, PLoS One 10 (9) (2015) e0138867. doi:10.1371/journal.pone.0138867.

[26] P. Tacik, S. Curry, S. Fujioka, A. Strongosky, R. J. Uitti, J. A. van Gerpen, N. N. Diehl, M. G. Heckman, Z. K. Wszolek, Cancer in Parkinson's disease, Parkinsonism Relat. Disord. 31 (2016) 28–33. doi:10.1016/j.parkreldis.2016.06.014.

[27] M. Ejma, N. Madetko, A. Brzecka, K. Guranski, P. Alster, M. Misiuk-Hojło, S. G. Somasundaram, C. E. Kirkland, G. Aliev, The Links between Parkinson's Disease and Cancer, Biomedicines 8 (10) (2020) 416. doi:10.3390/biomedicines8100416.

[28] P. S. Filippou, T. F. Outeiro, Cancer and Parkinson's Disease: Common Targets, Emerging Hopes, Mov. Disord. 36 (2) (2021) 340–346. doi:10.1002/mds.28425.

[29] C. Wang, X. Kang, L. Zhou, Z. Chai, Q. Wu, R. Huang, H. Xu, M. Hu, X. Sun, S. Sun, J. Li, R. Jiao, P. Zuo, L. Zheng, Z. Yue, Z. Zhou, Synaptotagmin-11 is a critical mediator of parkin-linked neurotoxicity and Parkinson's disease-like pathology, Nat. Commun. 9 (1) (2018) 81. doi:10.1038/s41467-017-02593-y.

[30] D. J. Rigden, X. M. Fernández, The 2022 Nucleic Acids Research database issue and the online molecular biology database collection, Nucleic Acids Res. 50 (D1) (2022) D1–D10. doi:10.1093/nar/gkab1195.

[31] E. Mosca, M. Bersanelli, T. Matteuzzi, N. Di Nanni, G. Castellani, L. Milanesi, D. Remondini, Characterization and comparison of gene-centered human interactomes, Brief. Bioinform. 00 (October 2020) (2021) 1–16. doi:10.1093/bib/bbab153.

[32] W. A. Haynes, A. Tomczak, P. Khatri, Gene annotation bias impedes biomedical research, Sci. Rep. 8 (1) (2018) 1362. doi:10.1038/s41598-018-19333-x.

[33] X. Hu, H. Kim, E. Stahl, R. Plenge, M. Daly, S. Raychaudhuri, Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets, Am. J. Hum. Genet. 89 (4) (2011) 496–506. doi:10.1016/j.ajhg.2011.09.002.

[34] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J.-B. B. Fan, R. Shen, High density DNA methylation array with single CpG site resolution, Genomics 98 (4) (2011) 288–295. doi:10.1016/j.ygeno.2011.07.007.

[35] Z. Gu, S. el Bouhaddani, J. Pei, J. Houwing-Duistermaat, H.-W. Uh, Statistical integration of two omics datasets using GO2PLS, BMC Bioinformatics 22 (1) (2021) 131. doi:10.1186/s12859-021-03958-3.