Commorbidity of Alzheimer's and Parkinson's Diseases: A Protein-Level Bioinformatics Analysis

Biological Datasets for Computational Physics

Miguel Avilés Moreno

September 12, 2025

Contents

1 Introduction

1.1 Brief description of the biological problem, knowns and unknowns

Alzheimer's disease (AD) is a chronic, progressive neurodegenerative disorder characterized by cognitive decline, memory loss, and impaired reasoning. It represents the leading cause of dementia in older adults. Pathological hallmarks include extracellular deposition of amyloid beta (Aβ) plaques, intracellular neurofibrillary tangles of phosphorylated tau (pTau), and widespread neuronal damage. These changes severely impact daily functioning and quality of life, and no curative therapies are currently available [Safiri et al., 2024].

Parkinson's disease (PD) is another major neurodegenerative disorder of the central nervous system, clinically manifesting as tremors, rigidity, and bradykinesia. The disease predominantly affects the elderly, with men being more frequently diagnosed than women. The global burden of PD has more than doubled in the past two decades, reflecting its increasing impact on public health. Pathologically, PD is characterized by dopaminergic neuron loss and the presence of intracellular inclusions rich in α-synuclein (Lewy bodies) [Sayyaed et al., 2023].

The comorbidity of AD and PD further complicates their clinical and biological understanding. Substantial overlap in symptoms, particularly in advanced disease stages, often leads to misdiagnosis or late recognition. Pathological studies have revealed that co-occurrence of AD and PD hallmarks is not uncommon, with Aβ, pTau, and α-syn frequently observed together in affected brains. This suggests potential molecular interactions between these proteins, raising the possibility of shared mechanisms underlying disease onset and progression [Zhang et al., 2025]. However, the precise molecular basis for AD–PD comorbidity remains poorly understood, representing both a diagnostic challenge and an opportunity for mechanistic insight.

Proteins often rely on recognizing and binding to other molecules in order to perform their biological functions. While traditionally this was understood as interactions between well-folded protein domains, it has become clear that many crucial interactions are mediated instead by intrinsically disordered regions (IDRs). These regions lack a stable structure in isolation but can undergo a disorder-to-order transition when binding their partners. Short linear motifs (SLiMs) represent a complementary concept: short stretches of amino acids that act as recognition signals, typically located in flexible and accessible parts of proteins [Mészáros et al., 2012].

1.2 Defining the question

Given the clinical and pathological overlap, a central question arises: *Do AD and PD share molecular features at the protein sequence level—specifically in terms of intrinsically disordered regions (IDRs), short linear motifs (SLiMs), and genetic variants—that could help explain their observed comorbidity?*

1.3 Defining the plan

To approach this question, we designed a computational workflow that translates clinical observations into protein-level analyses (sequence based predictions). The plan begins by identifying disease-associated proteins and extracting their amino acid sequences. These sequences are then examined for structural and functional features that are frequently implicated in neurodegeneration:

- Quantification of IDRs, which are polypeptide segments that do not fold into a stable tertiary structure [Li and Babu, 2018].
- Evaluation of disease overlap using similarity measures, Jaccard Index, to quantify the extent of molecular commorbidity between AD and PD.
- Computation of amino acids frequencies within the IDRs of AD, PD and human proteome.

- Identification of single variants in AD and PD proteins in IDRs.
- Identification of SLiMs embedded within IDRs, representing functional microdomains critical for protein interactions and regulation.

This conceptual framework allows us to translate the comorbidity of AD and PD into computational terms. By focusing on disorder, genetic variation and motifs at the sequence level, the study seeks to uncover molecular signatures that may contribute to the shared pathology of these two major neurodegenerative diseases.

2 Methods

2.1 Dataset assembly

All analyses were performed in Python 3.12.3 using Jupyter Notebook. The workflow relied on widely used libraries for data handling and analysis, including `pandas` for tabular manipulation, `numpy` for numerical operations, `requests` for programmatic access to web Applications of Programming Interface (APIs), `matplotlib` for visualization, and `scipy.stats` for statistical testing.

To construct the disease datasets, AD and PD were first defined using the National Center for Biotechnology Information (NCBI) web [**?**], including OMIM and MeSH. Gene–disease associations were retrieved from DisGeNET [Piñero et al., 2020], yielding two curated lists of genes ordered by disease–gene association score. These gene sets served as the entry point for downstream analyses.

Protein sequences corresponding to these genes and reported sequence variants, including annotated positions and amino acid substitutions, were retrieved via the UniProt REST API [The UniProt Consortium, 2025]. To obtain functional and structural features, additional programmatic resources were queried: (i) MobiDB API [**?**] for the annotation of IDRs, (ii) the ELM (Eukaryotic Linear Motif) API [Kumar et al., 2024] for the identification of SLiMs.

As a reference background, we sampled a set of 100 proteins from the human proteome using UniProt. These proteins were processed with the same pipeline as AD and PD datasets, providing a control distribution for comparison purposes.

2.2 Methods used for analysis

The computational workflow focused on translating disease-associated protein sets into measurable sequence-level features. For each protein, IDR content was quantified, variants were positioned along the primary sequence and SLiMs were mapped. Comparative analyses were then conducted at different levels:

- **IDR characterization:** global fraction of residues in IDRs, and amino acid frequency distributions within IDRs, for AD, PD, and control proteins.
- **Variant localization:** proportion of reported variants falling inside IDRs.
- **SLiM mapping:** occurrence and distribution of motifs, with particular attention to those embedded within disordered regions.
- **Disease overlap:** similarity between AD and PD associated gene sets was quantified using the Jaccard Index as a measure of comorbidity at the molecular level.

This design ensured that all results could be interpreted at the protein sequence level, while enabling comparisons between the two neurodegenerative diseases.

2.3 Statistical analysis

To assess whether proteins associated with AD and PD exhibit comparable molecular properties, we applied a set of inferential statistical tests across different analyses. In each case, the null hypothesis was that the distributions of the investigated features are comparable between the two disease-associated protein sets.

For the comparison of overall disorder content between AD and PD proteins, we used an independent two-sample t-test. This test assumes approximate normality of the distributions and is commonly employed to evaluate whether two independent samples differ significantly in their means.

For features where the assumption of normality did not hold, such as the distribution of longest intrinsically disordered region (IDR) lengths, we applied the Mann–Whitney U test. This non-parametric test assesses whether one distribution tends to have larger values than the other, without requiring normality. To evaluate differences in the fraction of genetic variants localized within IDRs, we again used the Mann–Whitney U test.

For completeness, additional similarity measures, such as the Jaccard Index and Euclidean distance between amino acid composition vectors, were also computed. These measures were used descriptively rather than inferentially.

3 Results

3.1 Gene–Disease Associations and Intrinsically Disordered Regions Characterization

The first question we addressed was whether proteins genetically associated with AD and PD display specific signatures of intrinsic disorder in their amino acid sequences. The goal was to establish the datasets of disease-related genes and their corresponding proteins, and to characterize their IDRs.

To construct the datasets, we defined the two diseases using OMIM and MeSH controlled identifiers from the NCBI webpage. We then retrieved the lists of genes associated with each disease from DisGeNET. Separate dataframes were generated for AD and PD, ordered by the DisGeNET association score. Crucially, each entry included the UniProt accession (UniProtID), which allowed mapping between gene identifiers and protein sequences. Because DisGeNET does not provide protein sequences, we used the UniProt REST API to query the sequence information directly. A custom function was implemented to perform this mapping, retrieving protein sequences for every UniProtID in the dataset.

Following sequence retrieval, we characterized intrinsic disorder using the MobiDB API (v6.0 OAS 3.0). For consistency, all proteins were annotated using the same consensus predictor, *prediction-disorder-disHL*, which identifies regions of intrinsic disorder along the amino acid chain. The workflow extracted IDR regions, total IDR length, and the fraction of disordered residues (IDR content). Entries with missing sequences were filtered out before analysis.

The final curated datasets included 196 proteins associated with AD and 147 proteins associated with PD. Among these, well-known disease drivers were correctly retrieved. In AD, the dataset includes APP (amyloid precursor protein, UniProtID P05067) and MAPT (microtubule-associated protein tau, UniProtID P10636), both of which are central to AD pathology [Safiri et al., 2024]. Indeed, APP cleavage generates β-amyloid peptides forming extracellular plaques, while hyperphosphorylated MAPT forms neurofibrillary tangles (NFTs) inside neurons. A slice of the dataset illustrates the structure of the final dataframe in Table 1.

For PD, the retrieved dataset included SNCA (alpha-synuclein, UniProtID P37840), PRKN (parkin, UniProtID O60260), and LRRK2 (leucine-rich repeat serine/threonine, UniProtID Q5S007) , all recognized as major contributors to PD development and progression [Sayyaed et al., 2023]. Example entries are displayed in Table 2.

genep_str_ids	gene_symbol	disease	seq_length	idr_content	idr_length	region_interval
P05067	APP	Alzheimer	770	0.2935	226	1–2, 25–25, 49–57...
P10636	MAPT	Alzheimer	758	0.6108	463	1–4, 6–7, 10–10, 13–26...

Table 1: *Example of Alzheimer's-associated proteins annotated with sequence and IDR information.*

genep_str_ids	gene_symbol	disease	seq_length	idr_content	idr_length	region_interval
P37840	SNCA	Parkinson	140	0.6643	93	1–6, 9–11, 13–16...
O60260	PRKN	Parkinson	465	0.3400	156	1–2, 7–23, 31–39...

Table 2: *Example of Parkinson's-associated proteins annotated with sequence and IDR information.*

To compare overall disorder tendencies, we computed the mean IDR content across all proteins for each disease. The results indicate disorder fractions of 0.29 for AD-associated proteins and 0.31 for PD-associated proteins. To assess whether this difference is statistically significant, we performed an independent two-sample t-test, assuming that IDR content distributions follow a normal distribution shown in Figure 1. The test yielded a t-statistic of -1.10 and a p-value of 0.27. As the p-value is greater than the conventional threshold of 0.05, we conclude that there is no statistically significant difference in mean IDR content between AD and PD protein sets. This suggests that neurodegenerative-disease proteins, regardless of etiology, share a comparable tendency to contain IDRs.

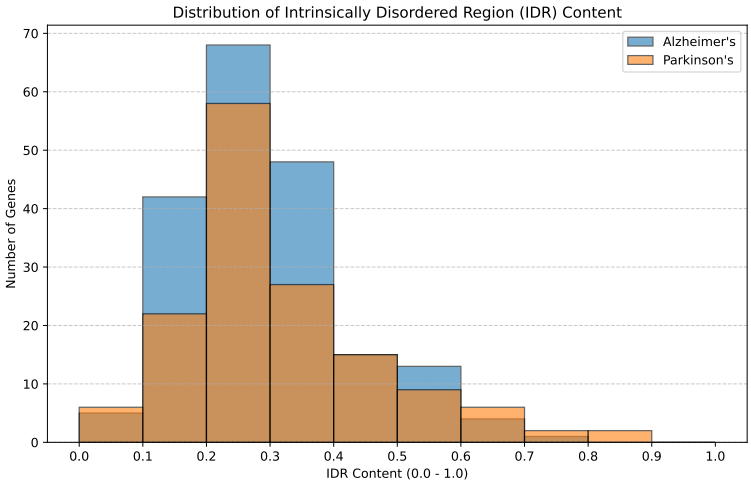

Figure 1: *Distribution of IDR content across proteins associated with Alzheimer's and Parkinson's disease. Bars represent the number of proteins with a given IDR fraction (x-axis bins from 0 to 1).*

In addition, we investigated the longest continuous IDR segment per protein. The frequency distribution of longest IDR lengths (Figure 2) shows that approximately 10% of proteins in both datasets contain disordered regions of 40–50 amino acids, and about 5% contain regions of 50–60 amino acids. To assess statistical differences between groups, we applied the Mann–Whitney U test, as the distributions of longest IDR lengths deviated significantly from normality. The test yielded a p-value of 0.50, indicating no significant difference between AD- and PD-associated proteins in terms of their longest disordered region length. These findings corroborate prior observations that neurodegenerative-disease proteins are highly enriched in disordered regions, at levels comparable to signaling and cancer-related proteins [Uversky, 2009]. The presence of long IDRs in central disease drivers such as MAPT and SNCA is consistent with their pathological

aggregation tendencies.

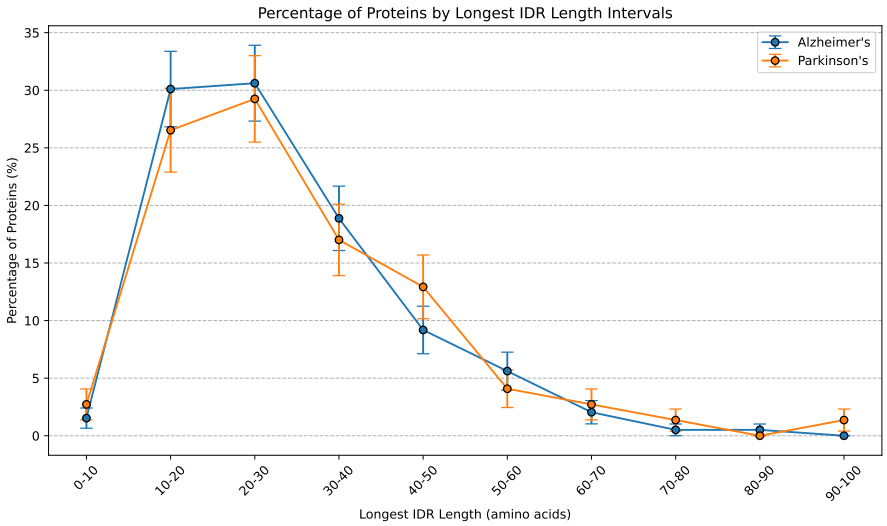

Figure 2: *Distribution of longest intrinsically disordered region (IDR) length in proteins associated with Alzheimer's and Parkinson's disease. The x-axis shows disorder region length in amino acids (10-aa bins), and the y-axis the percentage of proteins in each bin.*

Taken together, these results demonstrate that proteins genetically associated with AD and PD show intrinsic disorder. The similarity of mean disorder content between both datasets and the presence of long disordered regions in key proteins suggest that intrinsic disorder is a common structural feature of neurodegeneration-associated proteomes, potentially contributing to protein aggregation and functional dysregulation.

3.2 Jaccard Index, Shared Proteins and Aminoacid Composition of IDRs

In this section, we aimed to investigate whether AD and PD show a true comorbidity relationship by testing if they share a significant fraction of genes, and whether their amino acid composition within IDRs is more similar to each other than to the general human proteome.

For this analysis, we used three datasets: (i) AD associated proteins, (ii) PD associated proteins, and (iii) a reference dataset of human proteins sampled from UniProt. UniProt provides a downloadable reviewed human proteome containing approximately 20,000 proteins with UniProt accessions, gene symbols, and sequence length. To build the reference dataset, we filtered for Swiss-Prot entries with 5-star annotation accuracy and sequence lengths comparable to the GDA datasets (mean sequence length AD: 605.63, PD: 634.95). From this filtered dataset we randomly sampled 100 proteins, renaming columns for compatibility with the GDA structure, ensuring the same functions for IDR characterization could be applied. This allowed a consistent comparison among the three datasets: AD, PD, and Human proteome.

To quantify gene overlap, we computed the Jaccard Index (JI) [Travieso et al., 2024], defined as:

$$JI(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

We calculated the JI for three pairs: AD–PD, AD–Human proteome, and PD–Human proteome. The results are shown in Table 3.

The AD–PD Jaccard Index (0.125) demonstrates a strong overlap compared to negligible overlap with the general proteome (0.004 and 0.000), supporting a significant comorbidity relationship.

Comparison	Jaccard Index
AD–PD	0.125
AD–Human Proteome	0.004
PD–Human Proteome	0.000

Table 3: *Jaccard Index values for gene set overlaps.*

We found 38 shared proteins between AD and PD, listed in Table 4. These include highly relevant genes frequently cited in the literature as associated with neurodegeneration and comorbidity, such as *MAPT*, *SNCA*, *HFE*, *IL1B*, and *TNF*.

Shared Proteins (N=38)
MAPT, NECTIN2, EIF2AK2, INS, ALDH2, BDNF, IL1B, GSTO1, MTHFR, INSR, IGF2, GSTP1, IGF1R, GSTO2, CYP2D6, HLA-DRB1, HMGCR, HFE, S100B, MAOB, HMOX1, CASP3, UCHL1, SNCA, TNF, EPO, ND1, IGF2R, PARP1, NOS1, MAOA, ESR2, A2M, TFAM, KLK6, SOD2, GSTT1, ND2

Table 4: *Shared proteins between Alzheimer's disease and Parkinson's disease (N=38). Highlighted proteins are widely reported in comorbidity.*

The presence of *MAPT* and *SNCA* is particularly notable, as they represent hallmark genes for AD and PD, respectively. The overlap supports the hypothesis that these diseases share pathogenic mechanisms at the protein level.

We next analyzed amino acid frequencies within the IDRs of the three datasets. Figure 3 shows the distributions. The results are consistent with published findings that IDRs are enriched in disorder-promoting residues (Q, S, R and P) and depleted in order-promoting residues (C, W, I, Y, F, V and N) [Uversky, 2010]. In our case, both AD and PD associated proteins followed these patterns.

To quantify similarities, we computed pairwise Euclidean distances between amino acid frequency vectors and converted them into similarity values as:

$$Sim(A, B) = \frac{1}{1 + d(A, B)},$$

where $d(A, B)$ is the Euclidean distance. Results are shown in Table 5.

Comparison	Similarity
AD–PD	0.985
AD–Human Proteome	0.979
PD–Human Proteome	0.982

Table 5: *Pairwise similarities of amino acid frequency distributions in IDRs. Higher values indicate stronger similarity.*

The amino acid composition analysis shows that AD and PD have nearly identical IDR frequency profiles (0.985), stronger than their similarity to the general human proteome. This strongly supports the notion that comorbidity between AD and PD is not only genetic but also rooted in the biochemical properties of their disordered protein regions.

Taken together, these results demonstrate a consistent genetic and biochemical overlap between AD and PD. The shared proteins and the strong similarity in amino acid composition of IDRs support a true comorbidity relationship between these neurodegenerative disorders, rather than random overlap.

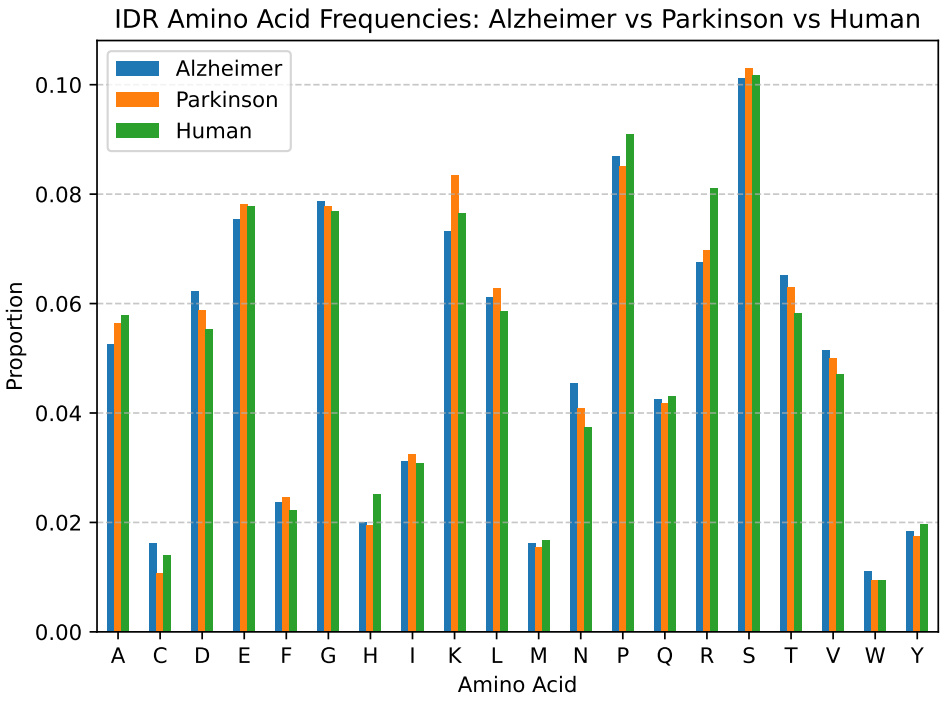

Figure 3: *Amino acid frequency distributions within IDRs for Alzheimer's disease, Parkinson's disease, and Human proteome. Both AD and PD datasets show enrichment in disorder-promoting residues and depletion in order-promoting residues.*

3.3 Variants in Intrinsically Disordered Regions

The goal of this section was to investigate whether variants in proteins associated with AD and PD preferentially fall within IDRs, and to assess whether the distribution of such variants corroborates previous findings in the literature [Vacic et al., 2012]. As described in the Methods, we queried UniProt through its API to extract all known single amino acid variants for proteins in the GDA lists of AD and PD, independently of whether the variants were directly related to disease. For each variant, we retrieved its UniProt accession, gene symbol, dbSNP identifier, position, reference and alternative amino acids, and determined whether the substitution fell within an annotated IDR.

Table 6 and Table 7 provide representative slices of the resulting variant datasets for AD and PD, respectively.

variant_id	uniprot_acc	gene_symbol	dbSNP	pos	ref_aa	alt_aa	in_idr
VAR_022315	P05067	APP	rs45588932	501	E	K	False
VAR_006413	P49768	PSEN1	rs63749824	79	A	V	False

Table 6: *Example of AD-associated protein variants.*

variant_id	uniprot_acc	gene_symbol	dbSNP	pos	ref_aa	alt_aa	in_idr
VAR_007957	P37840	SNCA	rs104893878	30	A	P	True
VAR_024931	Q5S007	LRRK2	rs2256408	50	R	H	False

Table 7: *Example of PD-associated protein variants.*

Consistent with the study of [Vacic et al., 2012], which reported that approximately 21.7% of annotated disease mutations fall within IDRs, we found that 28.1% of AD variants and 29.6% of PD variants occur in IDRs. A Mann–Whitney U test comparing the distributions of variant

percentages between AD and PD proteins yielded $U = 13453.5$ and $p = 0.288$, indicating no significant difference between the two diseases. This agreement strengthens the evidence that mutations within disordered regions are not rare, but instead represent a substantial fraction of the mutational landscape in neurodegenerative disease proteins.

We further analyzed the spectrum of amino acid substitutions within IDRs. Figure 4 shows the 15 most frequent substitution types observed in disordered regions across AD and PD proteins. The distribution of substitution types closely matches previous observations that disorder-to-order transition mutations are restricted to a limited repertoire. In particular, [Vacic et al., 2012] showed that only a handful of substitutions (RRW, RRC, ERK, RRH, RRQ) account for 44% of deleterious transitions. Our results corroborate this finding, as the most enriched substitution classes in our dataset overlapped substantially with this restricted repertoire, highlighting the reproducibility of the phenomenon across different datasets.

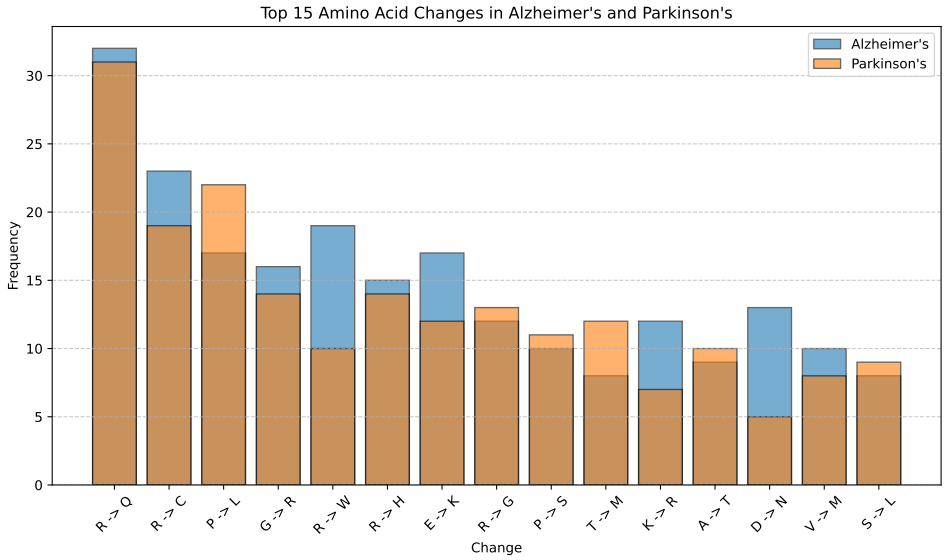

Figure 4: *Top 15 most frequent amino acid substitutions occurring within intrinsically disordered regions (IDRs) across AD and PD protein datasets.*

Taken together, these results demonstrate that variants located within IDRs are not only prevalent in AD and PD associated proteins, but also display a substitution repertoire highly consistent with previously reported disorder-to-order transition mutations. This convergence strongly supports the notion that IDR variants contribute mechanistically to the pathological aggregation and dysregulation processes underlying both neurodegenerative diseases.

3.4 Short Linear Motifs

The goal of this section is to investigate the presence and distribution of SLiMs in proteins associated with AD and PD. SLiMs are short peptide sequences that mediate protein–protein interactions, regulate cellular processes, and may reveal shared molecular mechanisms underlying comorbidity between these two neurodegenerative disorders.

We accessed the ELM database through its API. For each UniProt entry in the AD and PD datasets, we retrieved SLiMs and their annotated regions, filtering only those motifs that passed the ELM quality filters. Among the 353 classes available, we focused on seven biologically relevant motif types: 14-3-3, SH2, SH3, CDK, DEG, PDZ, and PKA. These motifs were selected due to their strong connections to protein regulation, phosphorylation, degradation, and signaling pathways implicated in AD and PD.

For each protein, we constructed a table containing the gene symbol, UniProt identifier,

motif type and motif region. Additionally, we implemented a function to evaluate whether each SLiM falls within an IDR. This will be useful for further analysis, since SLiMs are frequently enriched in disordered regions. An example slice of the results is shown in Tables 8 and 9.

gene_symbol	genep_str_ids	type_ELM	region	inside_idr
APP	P05067	DEG	[767, 770]	False
APP	P05067	SH2	[728, 731]	True

Table 8: *Example of SLiMs retrieved for Alzheimer's Disease-associated proteins.*

gene_symbol	genep_str_ids	type_ELM	region	inside_idr
SNCA	P37840	DEG	[1, 3]	True
GBA1	P04062	14-3-3	[170, 175]	False

Table 9: *Example of SLiMs retrieved for Parkinson's Disease-associated proteins.*

To better understand the global distribution of SLiMs, we plotted the counts of the selected motif types for both diseases (Figure 5).

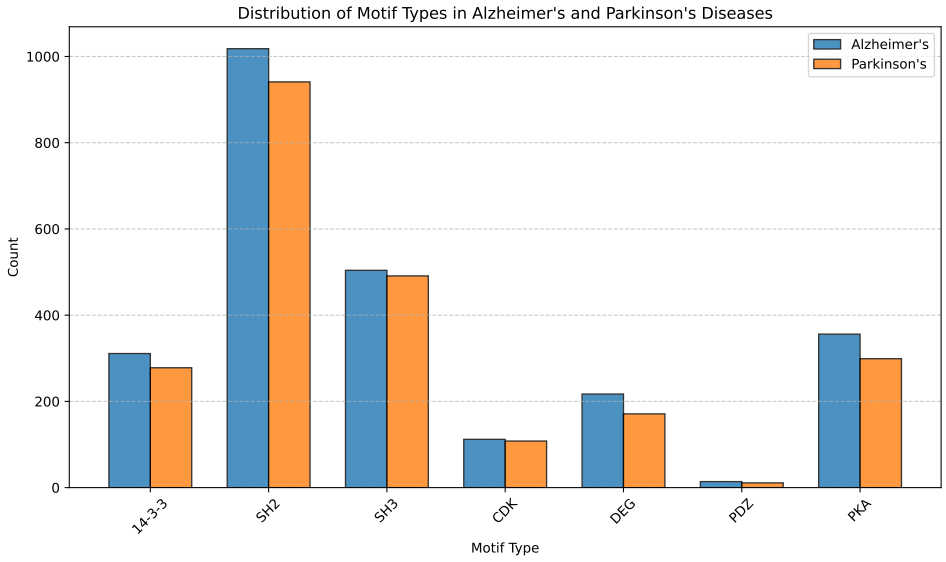

Figure 5: *Distribution of selected SLiM types (14-3-3, SH2, SH3, CDK, DEG, PDZ, PKA) across AD and PD proteins.*

The results indicate that AD and PD proteins show a remarkably similar distribution of motif types, being SH2 and SH3 the most abundant SLIMs, characteristics of signaling and transforming proteins [Koch et al., 1991]. This observation suggests that proteins in both diseases may engage in overlapping regulatory mechanisms. This observation suggest that a mutation affecting a motif could simultaneously disrupt multiple processes, explaining comorbidity. SLiMs often appear in hub proteins, and disruption of the same hub types may contribute to both diseases.

Taken together, the results consistently show that AD and PD proteins harbor overlapping sets of SLiMs, supporting the hypothesis that comorbidity may arise from dysregulation of shared molecular pathways.

3.4.1 Relationship between Short Linear Motifs and Intrinsically Disordered Regions

IDRs and SLiMs describe molecular interactions that are short, transient, and highly adaptable, making them central to regulation and signaling processes. Importantly, many SLiMs are located within IDRs, reflecting the shared principles of flexibility, accessibility, and weak but specific binding [Mészáros et al., 2012].

To investigate this relationship, we quantified the fraction of SLiMs located inside IDRs in our two disease-associated protein datasets. The results show that in AD-related proteins, 23.18% of the identified SLiMs fall within disordered regions. For PD proteins, the corresponding percentage is 21.01%.

These results indicate that a substantial proportion of SLiMs are embedded in IDRs for both diseases, consistent with the idea that intrinsic disorder provides the structural context for motif-mediated interactions.

4 Discussion and Conclusion

4.1 Discussion

In this study, we addressed the biological question of whether AD and PD share common molecular features that may underlie their frequent comorbidity. Specifically, we examined whether proteins associated with both diseases display comparable levels of intrinsic disorder, overlapping genetic associations, similar amino acid composition in disordered regions, and similar variants and SLiMs distributions.

Our results consistently support the presence of shared molecular signatures. Both AD and PD associated proteins exhibited nearly identical mean intrinsic disorder content (0.29 vs. 0.31) and contained long disordered regions in key drivers such as *MAPT* and *SNCA*, consistent with their roles in pathological aggregation. Furthermore, we identified 38 proteins shared between AD and PD, including hallmark genes *MAPT* and *SNCA*, as well as inflammatory regulators like *IL1B* and *TNF*, strongly supporting the hypothesis of a genetic overlap. The amino acid composition of IDRs revealed nearly identical frequency profiles for AD and PD (similarity score = 0.985), reinforcing that their biochemical properties converge. Finally, both datasets showed highly similar SLiM repertoires, particularly in motifs related to phosphorylation, degradation, and regulatory interactions, suggesting that dysregulation of shared signaling pathways contributes to disease progression.

These findings are in strong agreement with previous reports. [Uversky, 2009] demonstrated that neurodegeneration-related proteins are among the most enriched in IDRs, comparable to proteins in cancer and signaling. Our results directly support this observation by showing broad distributions of disorder across both datasets. Moreover, [Sayyaed et al., 2023, Safiri et al., 2024] identifies *SNCA*, *MAPT*, and *APP* as central drivers of PD and AD pathology, respectively, consistent with their identification in our curated datasets and with the observed enrichment of disordered regions that favor aggregation. The overlap in genetic and biochemical features we report adds novel evidence that comorbidity between AD and PD is not incidental but mechanistically rooted in shared protein disorder and regulatory motifs.

Taken together, we propose a model in which AD and PD proteins converge on a "disorder-driven" mechanism: their enrichment in IDRs, coupled with a high density of SLiMs within these regions, facilitates aberrant protein–protein interactions, aggregation, and dysregulation of signaling pathways. The shared presence of inflammatory regulators among the overlapping proteins further suggests that chronic neuroinflammation may be a unifying consequence of this molecular convergence [Heneka and van der Flier, 2025].

4.2 Conclusion

In conclusion, our study demonstrates that AD and PD share strong molecular parallels at the levels of protein disorder, genetic overlap, amino acid composition, single variants and SLiM signatures. These findings emphasize that the comorbidity between AD and PD likely arises from shared disorder-driven mechanisms of protein aggregation and dysregulation. From a broader perspective, these insights highlight the importance of targeting intrinsically disordered proteins and their interaction motifs in future therapeutic strategies. For further studies, one promising direction would be to obtain and analyze the variants specifically associated with the diseases, in order to disentangle disease-driving mutations from neutral background variation. Importantly, the framework we have developed is general: we created a Jupyter Notebook that allows analysis of any GDA dataset downloaded from DisGeNET, since the implemented functions are disease-agnostic. We tested this approach on Hypertension and Obesity, and it worked seamlessly without the need for any code adaptation, supporting its robustness and broad applicability.

References

[Heneka and van der Flier, 2025] Heneka, M. T. and van der Flier (2025). Neuroinflammation in Alzheimer disease. *Nature Reviews Immunology*, 25(5):321–352.

[Koch et al., 1991] Koch, C. A., Anderson, D., Moran, M. F., Ellis, C., and Pawson, T. (1991). SH2 and SH3 domains: Elements that control interactions of cytoplasmic signaling proteins. *Science (New York, N.Y.)*, 252(5006):668–674.

[Kumar et al., 2024] Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J. M., Kalman, Z. E., Pascarelli, S., Palopoli, N., Dobson, L., Suarez, C. F., Van Roey, K., Krystkowiak, I., Griffin, J. E., Nagpal, A., Bhardwaj, R., Diella, F., Mészáros, B., Dean, K., Davey, N. E., Pancsa, R., Chemes, L. B., and Gibson, T. J. (2024). ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research*, 52(D1):D442–D455.

[Li and Babu, 2018] Li, X.-H. and Babu, M. M. (2018). Human Diseases from Gain-of-Function Mutations in Disordered Protein Regions. *Cell*, 175(1):40–42.

[Mészáros et al., 2012] Mészáros, B., Dosztányi, Z., and Simon, I. (2012). Disordered binding regions and linear motifs–bridging the gap between two models of molecular recognition. *PloS One*, 7(10):e46829.

[Piñero et al., 2020] Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855.

[Safiri et al., 2024] Safiri, S., Ghaffari Jolfayi, A., Fazlollahi, A., Morsali, S., Sarkesh, A., Daei Sorkhabi, A., Golabi, B., Aletaha, R., Motlagh Asghari, K., Hamidi, S., Mousavi, S. E., Jamalkhani, S., Karamzad, N., Shamekh, A., Mohammadinasab, R., Sullman, M. J. M., Şahin, F., and Kolahi, A.-A. (2024). Alzheimer's disease: A comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. *Frontiers in Medicine*, 11.

[Sayyaed et al., 2023] Sayyaed, A., Saraswat, N., Vyawahare, N., and Kulkarni, A. (2023). A detailed review of pathophysiology, epidemiology, cellular and molecular pathways involved in the development and prognosis of Parkinson's disease with insights into screening models. *Bulletin of the National Research Centre*, 47(1):70.

[The UniProt Consortium, 2025] The UniProt Consortium (2025). UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617.

[Travieso et al., 2024] Travieso, G., Benatti, A., and Costa, L. d. F. (2024). An Analytical Approach to the Jaccard Similarity Index.

[Uversky, 2009] Uversky, V. N. (2009). Intrinsic disorder in proteins associated with neurodegenerative diseases. *Frontiers in Bioscience (Landmark Edition)*, 14(14):5188–5238.

[Uversky, 2010] Uversky, V. N. (2010). Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: Another illustration of the D2 concept. *Expert Review of Proteomics*, 7(4):543–564.

[Vacic et al., 2012] Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., and Iakoucheva, L. M. (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS computational biology*, 8(10):e1002709.

[Zhang et al., 2025] Zhang, Q.-L., Liu, Y., Tu, T., and Yan, X.-X. (2025). The Comorbidity of AD and PD: Exploring Clinical, Pathological, and Biomarker Interactions. *Aging and Disease*.