Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Máster Universitario en Innovación Digital

COMPLEX DATA IN HEALTH

Studying Alzheimer's Disease

Authors:

Emanuele Alberti
Leandro Duarte
Ottavia Biagi

December 2025

Contents

Introduction

For this assignment, we were assigned to study **Alzheimer's Disease (AD)**. The following identifiers were provided: UMLS CUI `C0002395`, NCIt code `C2866`, recommended protein Apolipoprotein E (UniProt: `P02649`), transcriptome dataset `GSE300079`, and image dataset from OASIS (Kaggle).

1 Medical Terminologies and Semantic Datasets

1.1 Disease General Information

We explored biomedical databases to gather standardized information about Alzheimer's Disease.

NCBI MedGen (UID: 1853) provides the following definition: *"A degenerative disease of the brain that causes dementia, which is a gradual loss of memory, judgment, and ability to function. This disorder usually appears in people older than age 65, but less common forms of the disease appear earlier in adulthood."* The disease is classified as *Disease or Syndrome* with concept ID `C0002395`. The directly associated gene is `APP` (21q21.3), with related genes including `APOE`, `PSEN1`, `PSEN2`, `ABCA7`, `MPO`, and `PLAU` [1].

NCBI MeSH (ID: D000544) describes AD as a degenerative brain disease with insidious onset of dementia, impairment of memory, judgment, and attention span, followed by apraxias and global loss of cognitive abilities. Pathologically, it is marked by senile plaques, neurofibrillary tangles, and neuropil threads [2].

ICD-10 classifies AD under code `G30`, with subtypes: `G30.0` (early onset, before age 65), `G30.1` (late onset, after age 65), `G30.8` (other), and `G30.9` (unspecified) [3].

Orphanet (ORPHA:1020) describes Early-onset autosomal dominant Alzheimer disease (EOAD), representing less than 1% of all AD cases, caused by mutations in `PSEN1` (69%), `APP` (13%), or `PSEN2` (2%) [4].

Table 1 summarizes the disease codifications across vocabularies.

Table 1: Alzheimer's Disease codifications across medical vocabularies [1, 2, 3, 5, 6, 7, 8, 4].

Vocabulary	**Code**
UMLS CUI	C0002395
MedGen UID	1853
MeSH	D000544
ICD-10	G30
SNOMED CT	26929004
NCIt	C2866
OMIM	104300, 516000
Orphanet	ORPHA:238616, ORPHA:1020
MONDO	MONDO:0004975
HPO	HP:0002511

1.2 SPARQL Query Results

We queried the NCIt SPARQL endpoint (`https://shared.semantics.cancer.gov/sparql`) to retrieve annotation properties for Alzheimer's Disease (code `C2866`) [6]. The query extracted the preferred label, synonyms, definition, semantic type, and UMLS CUI.

Summary of SPARQL Findings: The query retrieved annotation properties for NCIt code C2866. The disease is classified under semantic type "Mental or Behavioral Dysfunction" and maps to UMLS CUI C0002395. Seven synonyms were identified, including "Alzheimer's Dementia", "Alzheimer Disease", and "Alzheimer dementia". The definition describes AD as a progressive neurodegenerative disease characterized by nerve cell death leading to loss of cognitive function such as memory and language.

2 Bioinformatics

2.1 Disease Genes

We identified three main genetic factors linked to Alzheimer's disease from NCBI MedGen [1]: APP and PSEN1, when mutated, can directly cause rare, early-onset familial forms of the disease. APOE primarily affects an individual's risk for the more common late-onset form.

APP (Amyloid Precursor Protein) – NCBI Gene ID: 351, chromosome 21q21.3. The APP gene provides the blueprint for amyloid precursor protein. This protein is cut into smaller fragments, including amyloid-beta ($A\beta$), which is a key component of the plaques observed in the brains of Alzheimer's patients [9, 10]. Certain APP mutations or extra copies of the gene lead to an overproduction or more "sticky" forms of $A\beta$, which promotes plaque buildup and can directly cause early-onset familial Alzheimer's disease [9].

PSEN1 (Presenilin 1) – NCBI Gene ID: 5663, chromosome 14q24.2. PSEN1 encodes presenilin-1, which is the catalytic core of the γ-secretase enzyme. This enzyme performs the final cut of APP to release $A\beta$. Mutations in PSEN1 typically alter this cutting process, leading to the production of more of the longer, aggregation-prone $A\beta$ peptides. This makes PSEN1 the most common genetic cause of autosomal-dominant early-onset Alzheimer's, with symptoms often beginning before age 65, and sometimes even before 40 [11, 12].

APOE (Apolipoprotein E) – NCBI Gene ID: 348, chromosome 19q13.32. APOE produces apolipoprotein E, a protein that assists in transporting fats and cholesterol within the body and brain [13]. Its common variants ($\varepsilon 2$, $\varepsilon 3$, $\varepsilon 4$) differ in how they manage $A\beta$ and brain lipid metabolism. Carrying the $\varepsilon 4$ form increases the risk of late-onset Alzheimer's in a dose-dependent manner, while the $\varepsilon 2$ variant tends to lower this risk. APOE primarily acts as a risk modifier; individuals with the $\varepsilon 4$ allele may never develop Alzheimer's, and conversely, those without it can still develop the disease [13, 14].

Gene Sequence: Following the assignment recommendation, the APOE mRNA FASTA sequence (RefSeq: `NM_000041.4`) is provided in Appendix A [15].

2.2 Proteins

We analyzed Apolipoprotein E (APOE), the recommended protein for this assignment, using UniProt accession `P02649` [16].

Function: APOE is an apolipoprotein that associates with lipid particles and transports lipids between organs via plasma and interstitial fluids. It is a core component of plasma lipoproteins (chylomicrons, VLDL, IDL, HDL) and participates in their production, conversion, and clearance. APOE binds to cellular receptors (LDLR, LRP1, LRP2, LRP8, VLDLR) that mediate uptake of APOE-containing lipoprotein particles. The protein plays a role in cholesterol homeostasis through reverse cholesterol transport and regulates lipid transport in the central nervous system, affecting neuron survival and sprouting [16].

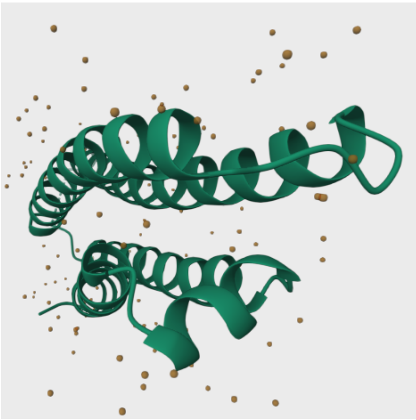
Role in Alzheimer's Disease: The APOE*4 allele is associated with late-onset familial and sporadic Alzheimer disease. Risk for AD increases from 20% to 90% and mean age at onset decreases from 84 to 68 years with increasing number of APOE*4 alleles. [16].

Protein Sequence: The APOE protein sequence (317 amino acids) was retrieved from UniProt and is provided in Appendix B.

Molecular Interactions: APOE does not interact with DNA or RNA. The protein primarily interacts with lipids through its C-terminal domain (residues 222-299), which binds phospholipids and cholesterol in lipoprotein particles. The N-terminal domain (residues 1-191) contains the LDLR-binding region (residues 134-150) that mediates receptor-mediated cellular uptake of lipoproteins. APOE also binds heparin and heparan sulfate proteoglycans on cell surfaces through its N-terminal domain [16].

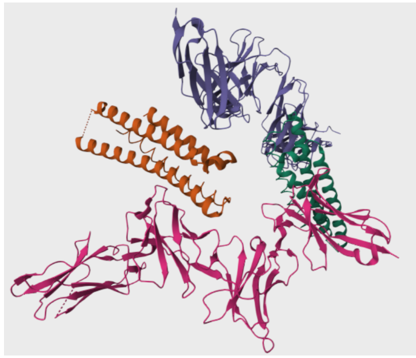
Protein Structures: We obtained protein structures from classical experimental methods and computational prediction.

Figure 1 shows the four structures obtained from classical methods and AlphaFold prediction.

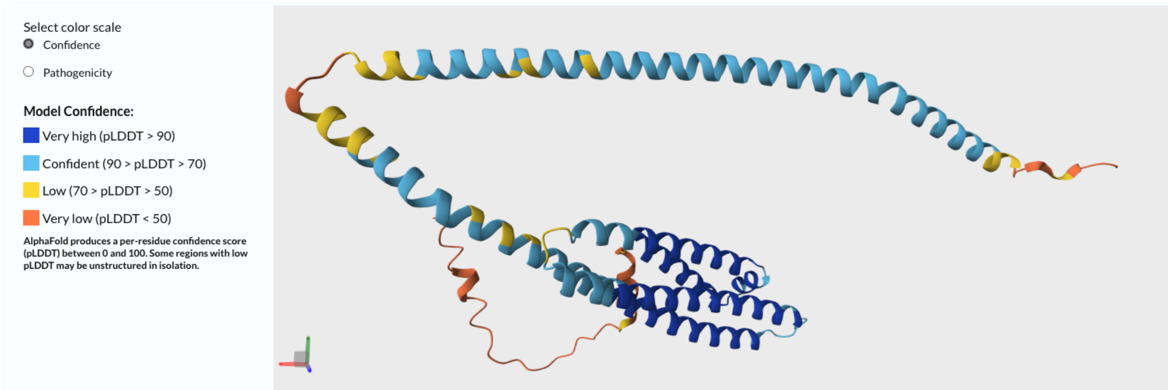


(a) X-ray (1B68)

(b) NMR (1OEF)



(c) Cryo-EM (8GRX)



(d) AlphaFold prediction (AF-P02649-F1)

Figure 1: APOE protein structures from different determination methods [17, 18, 19, 20].

Structural Comparison: The X-ray structure displays a static, well-defined conformation typical of crystallographic methods, capturing residues 19-209. Crystallography freezes the protein in a single state, producing a clean ribbon structure. The NMR structure shows multiple overlapping conformations (appearing as a bundle of strands), representing an ensemble of solution-state structures for residues 281-304. NMR captures protein dynamics in solution, resulting in multiple overlaid structures. The Cryo-EM structure presents a complete assembly (residues 41-180), revealing protein organization in a near-native frozen state. AlphaFold predicts the full-length structure (1-317), displayed as a single ribbon colored by confidence: blue indicates high confidence, cyan for confident regions, yellow for low confidence, and orange for very low confidence. Classical methods provide fragmented views of specific domains, while AlphaFold offers a complete structure prediction across the entire sequence.

Cavity Detection: We used CB-Dock2 [21] to identify potential binding pockets on the AlphaFold predicted structure. CB-Dock2 is a web-based tool for blind protein-ligand docking that combines geometry-based cavity detection with template information. The analysis identified 5 potential binding pockets (Table 2). Cavities are grooves or holes in the protein surface where other molecules (lipids, cholesterol, or drugs) can bind.

Table 2: Binding pockets detected by CB-Dock2 on APOE AlphaFold structure.

Pocket ID	Volume (Å^3)	Center (x, y, z)	Size (x, y, z)
C1	1123	(10, -3, 10)	(12, 16, 16)
C2	287	(48, 27, -33)	(8, 13, 8)
C3	134	(-6, -13, 7)	(11, 6, 5)
C4	97	(20, -2, -15)	(7, 8, 9)
C5	93	(1, 1, -7)	(8, 8, 7)

Cavity C1 has the largest volume (1123 Å^3), which is 4 times larger than the second pocket. This indicates C1 is the primary binding site where lipids or potential therapeutic ligands would bind. The remaining pockets (C2-C5) are smaller and represent secondary binding sites.

2.3 Transcriptome

Gene expression data were retrieved from GEO accession `GSE300079` [22]. This dataset contains single-cell RNA-seq from mouse brain tissue (Mus musculus), examining the effect of switching from the APOE4 risk allele to the APOE2 allele. The analysis code is provided in the attached notebook.

Dataset: samples from whole brain tissue of 4–7 month old female mice. We selected 5 samples for comparison: 2 controls (APOE4) and 3 switched (APOE4→E2).

Method: For each sample, counts were summed across all cells (8,355 to 29,793 cells per sample). Expression values were normalized to CPM and $\log_2$-transformed. 14,945 genes passed filtering (mean CPM > 1). Differential expression was assessed using Welch's t-test.

Results: We identified 63 nominally significant genes ($p < 0.05$, $|\log_2 \text{FC}| > 0.5$). The volcano plot (Figure 2) shows a bias toward downregulation: 59 genes decreased and 4 increased after the switch.

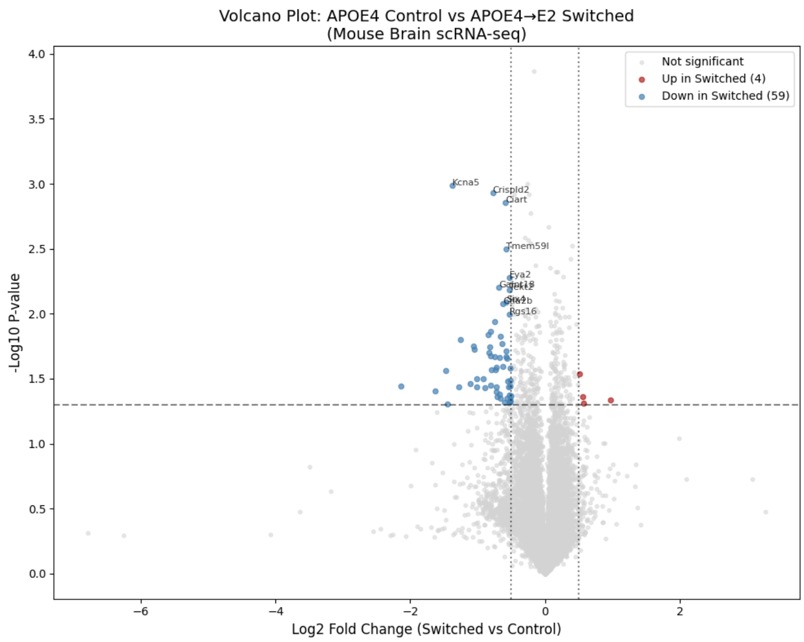


Figure 2: Volcano plot of differential expression. Blue: genes downregulated. Red: genes upregulated.

Functional Enrichment: The 59 downregulated genes were submitted to Enrichr [23] for Gene Ontology enrichment. The top biological process was *Dopamine Metabolic Process* (GO:0042417), shown in Figure 3. This pathway includes gene `Gch1` (GTP Cyclohydrolase 1), which is required for dopamine synthesis. The result suggests that the APOE4 (risk) brain has elevated dopamine turnover, which is reduced after switching to APOE2.

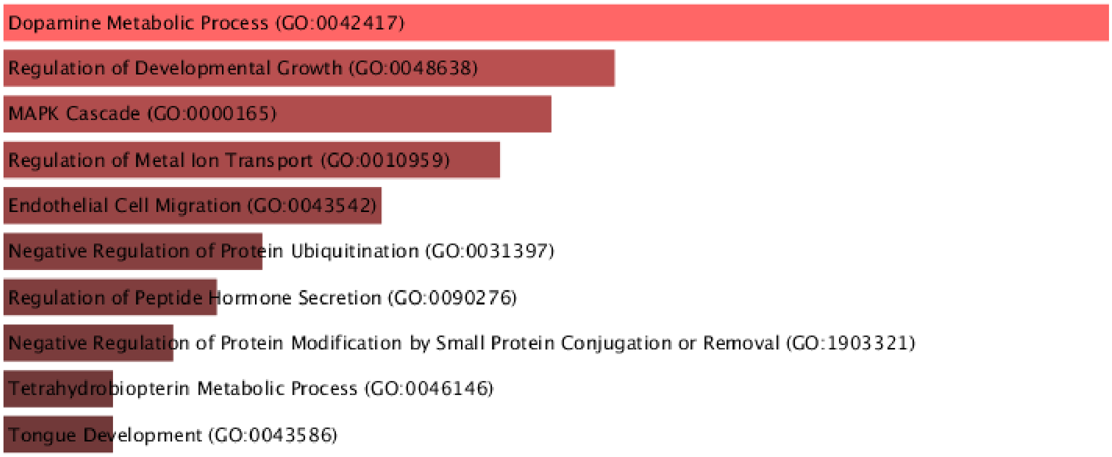


Figure 3: Enrichr GO Biological Process 2025 enrichment for downregulated genes.

3 Network Medicine

3.1 Disease Module

To identify the disease module for Alzheimer's Disease (AD), disease-gene associations were integrated with the human protein-protein interactome (PPI). The interactome contains 18,508 nodes and 332,645 edges. A total of 101 AD-associated genes (UMLS CUI: C0002395) were extracted from the disease-gene dataset. After mapping onto the interactome, 91 genes were present.

The AD-specific subgraph was generated from these 91 genes. This subgraph contained 42 connected components. The Largest Connected Component (LCC) was selected as the disease module. The AD module contains:

- 46 genes (nodes)
- 132 interactions (edges), reduced to 87 after removing 45 self-loops

Statistical Significance: To evaluate whether this module is non-random, the observed LCC size was compared against 1,000 random modules of equal size (91 genes sampled from the interactome). Random LCC sizes followed a distribution with mean = 2.31 and std = 1.41. The AD module LCC size of 46 is far outside the random distribution (Figure 4). This yields a z-score of 30.91 and empirical p-value of 0.0, indicating the module is not due to chance.

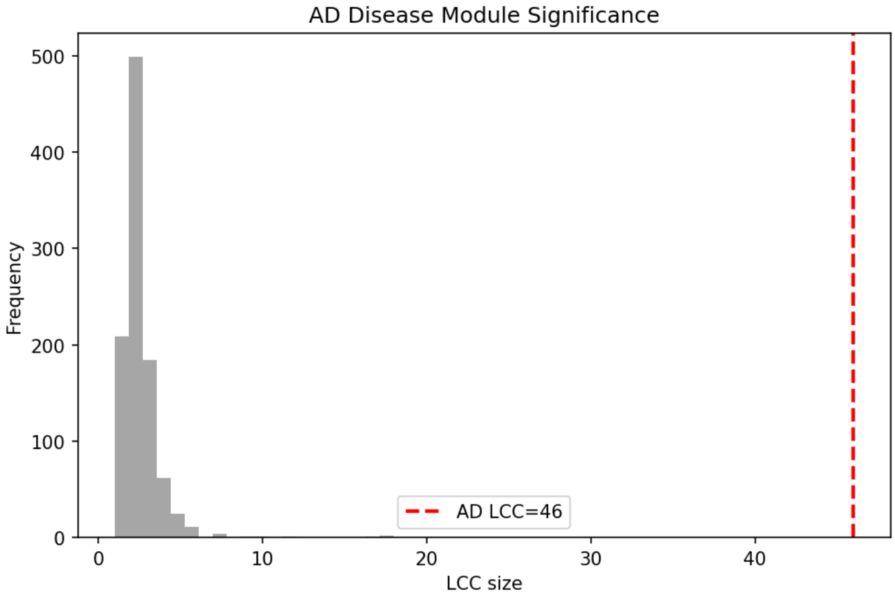


Figure 4: Histogram of random LCC sizes (grey) vs observed AD LCC size (red dashed line at 46).

Visualization: Figure 5 shows the AD module visualized in Gephi. Nodes are sized by degree centrality. Hub genes APP, MAPT, and PSEN1 form a central core. Peripheral nodes connect through short paths, forming communities that correspond to AD-related pathways.

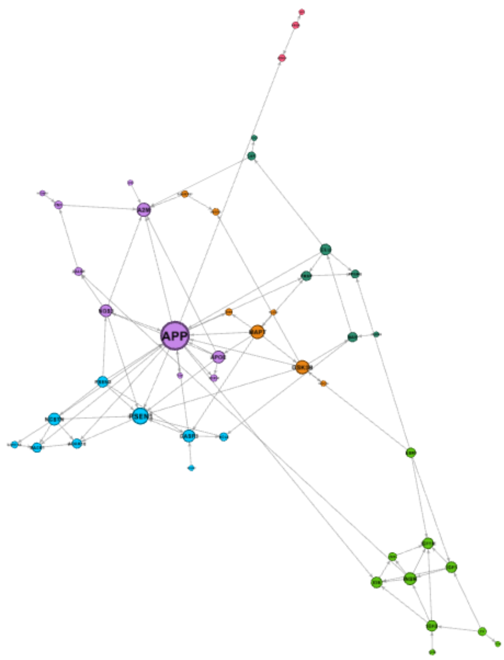


Figure 5: Gephi visualization of the AD disease module. Node size indicates degree centrality.

3.2 Disease Separation

To assess molecular relationships between AD and other conditions, we selected two additional diseases for comparison. We used the 'separation.py' script provided by the professors, which implements the disease separation metric.

Selected Diseases:

- **Close Disease:** presenile dementia (CUI: C0011265). Biologically similar to AD, sharing 89 genes out of 91 AD genes (97.8% overlap).

- **Distant Disease:** fatty Liver (CUI: C0015695). Affects a different organ system, sharing only 6 genes with AD (6.6% overlap).

Separation Results:

- AD vs presenile dementia: $s = 0.0167$
- AD vs fatty Liver: $s = 1.4424$

The separation value of $s = 0.0167$ between AD and presenile dementia indicates a very small positive separation. This value is close to zero, suggesting a high degree of molecular overlap and topological proximity in the interactome. The separation value of $s = 1.4424$ between AD and fatty Liver indicates a large positive separation. This confirms that these diseases are topologically distinct in the interactome, reflecting their different biological pathways and organ systems.

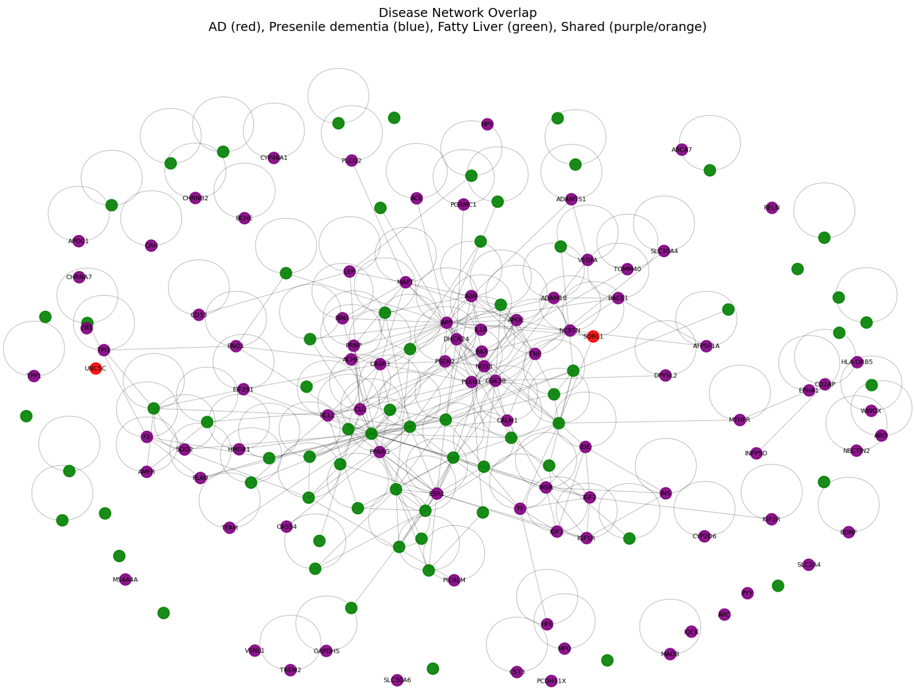


Figure 6: Network overlap between AD (red), presenile dementia (blue), and fatty Liver (green)

3.3 Disease-Drug Proximity

This task explores drug repurposing by assessing the network proximity between drug targets and the AD module. The core idea is that drugs with targets topologically close to the disease module may be effective, even if not directly indicated for AD. Proximity is measured as the average shortest path distance from drug targets to the nearest disease gene. A Z-score quantifies the significance of this proximity compared to random targets.

We processed 22,007 drugs with valid gene targets present in the human protein-protein interactome. Drugs with existing indications for Alzheimer's Disease or dementia were excluded from the repurposing candidates to focus on novel opportunities.

Raw Proximity Distribution: This initial step involved computing the average closest distance from drug targets to the AD module for all drugs. This rapid calculation served as a preliminary screen to identify drugs with potentially close targets before performing the more computationally intensive full significance analysis.

To evaluate if their targets are significantly closer to the AD module than expected by chance, we performed a Z-score analysis for selected drugs. This involved comparing the observed closest distance to a distribution of closest distances obtained from 1000 random simulations, where drug targets of the same size were randomly selected from the interactome.

The method successfully identified known AD drugs as positive controls, confirming its ability to detect significant proximity between drug targets and the AD module. For example, Tacrine showed strong significance ($z = -3.84$, $p = 0.000$ and $z = -4.51$, $p = 0.000$) and VP025 also demonstrated significant proximity ($z = -2.67$, $p = 0.008$).

Several drugs with no prior AD indication demonstrated significant proximity, suggesting repurposing opportunities (Figure 7). Notable examples include Simvastatin ($z = -5.61$, $p = 0.000$), which shows very strong proximity, and ATL1101 ($z = -2.04$, $p = 0.042$).

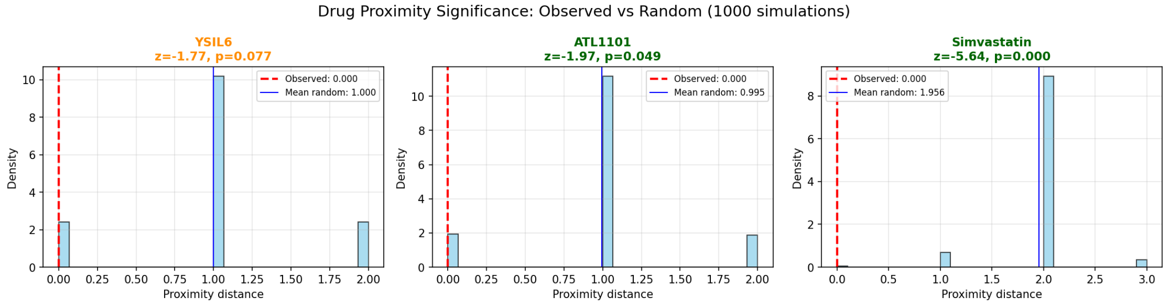


Figure 7: Drug proximity significance for selected repurposing candidates: YSIL6, ATL1101, and Simvastatin. Red dashed line: observed proximity. Blue line: mean random proximity. The plot indicates whether observed proximity is significantly closer to the AD module than random.

A APOE Gene FASTA Sequence

```
>NM_000041.4 Homo sapiens apolipoprotein E (APOE), transcript variant 2, mRNA
CTACTCAGCCCCAGCGGAGGTGAAGGACGTCCTTCCCCAGGAGCCGACTGGCCAATCACAGGCAGGAAGA
TGAAGGTTCTGTGGGCTGCGTTGCTGGTCACATTCCTGGCAGGATGCCAGGCCAAGGTGGAGCAAGCGGT
GGAGACAGAGCCGGAGCCCGAGCTGCGCCAGCAGACCGAGTGGCAGAGCGGCCAGCGCTGGGAACTGGCA
CTGGGTCGCTTTTGGGATTACCTGCGCTGGGTGCAGACACTGTCTGAGCAGGTGCAGGAGGAGCTGCTCA
GCTCCCAGGTCACCCAGGAACTGAGGGCGCTGATGGACGAGACCATGAAGGAGTTGAAGGCCTACAAATC
GGAACTGGAGGAACAACTGACCCCGGTGGCGGAGGAGACGCGGGCACGGCTGTCCAAGGAGCTGCAGGCG
GCGCAGGCCCGGCTGGGCGCGGACATGGAGGACGTGTGCGGCCGCCTGGTGCAGTACCGCGGCGAGGTGC
AGGCCATGCTCGGCCAGAGCACCGAGGAGCTGCGGGTGCGCCTCGCCTCCCACCTGCGCAAGCTGCGTAA
GCGGCTCCTCCGCGATGCCGATGACCTGCAGAAGCGCCTGGCAGTGTACCAGGCCGGGGCCCGCGAGGGC
GCCGAGCGCGGCCTCAGCGCCATCCGCGAGCGCCTGGGGCCCCTGGTGGAACAGGGCCGCGTGCGGGCCG
CCACTGTGGGCTCCCTGGCCGGCCAGCCGCTACAGGAGCGGGCCCAGGCCTGGGGCGAGCGGCTGCGCGC
GCGGATGGAGGAGATGGGCAGCCGGACCCGCGACCGCCTGGACGAGGTGAAGGAGCAGGTGGCGGAGGTG
CGCGCCAAGCTGGAGGAGCAGGCCCAGCAGATACGCCTGCAGGCCGAGGCCTTCCAGGCCCGCCTCAAGA
GCTGGTTCGAGCCCCTGGTGGAAGACATGCAGCGCCAGTGGGCCGGGCTGGTGGAGAAGGTGCAGGCTGC
CGTGGGCACCAGCGCCGCCCCTGTGCCCAGCGACAATCACTGAACGCCGAAGCCTGCAGCCATGCGACCC
CACGCCACCCCGTGCCTCCTGCCTCCGCGCAGCCTGCAGCGGGAGACCCTGTCCCCGCCCCAGCCGTCCT
CCTGGGGTGGACCCTAGTTTAATAAAGATTCACCAAGTTTCACGCA
```

B Apolipoprotein E FASTA Sequence

```
>sp|P02649|APOE_HUMAN Apolipoprotein E OS=Homo sapiens OX=9606 GN=APOE PE=1 SV=1
MKVLWAALLVTFLAGCQAKVEQAVETEPEPELRQQTEWQSGQRWELALGRFWDYLRWVQT
LSEQVQEELLSSQVTQELRALMDETMKELKAYKSELEEQLTPVAEETRARLSKELQAAQA
RLGADMEDVCGRLVQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKRLAVY
QAGAREGAERGLSAIRERLGPLVEQGRVRAATVGSLAGQPLQERAQAWGERLRARMEEMG
SRTRDRLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLKSWFEPLVEDMQRQWAGLVEK
VQAAVGTSAAPVPSDNH
```

References

[1] NCBI. *MedGen UID 1853: Alzheimer's Disease.* https://www.ncbi.nlm.nih.gov/medgen/1853.

[2] NCBI. *MeSH D000544: Alzheimer Disease.* https://www.ncbi.nlm.nih.gov/mesh/68000544.

[3] WHO. *ICD-10 G30: Alzheimer's Disease.* https://icd.who.int/browse10/2019/en#G30.

[4] Orphanet. *Orphanet 1020: Early-onset autosomal dominant Alzheimer disease.* https://www.orpha.net/en/disease/detail/1020.

[5] NCBI. *SNOMED CT 26929004: Alzheimer disease.* https://vsac.nlm.nih.gov/context/cs/codesystem/SNOMEDCT/version/2021-09/code/26929004/info.

[6] NCI. *NCIt C2866: Alzheimer's Disease.* https://evsexplore.semantics.cancer.gov/evsexplore/concept/ncit/C2866.

[7] Monarch Initiative. *MONDO:0004975 Alzheimer disease.* https://monarchinitiative.org/MONDO:0004975.

[8] MSeqDR. *HPO HP:0002511 Dementia.* https://mseqdr.com/hpo_browser.php?2511.

[9] NCBI. *APP amyloid beta precursor protein [Homo sapiens (human)].* https://www.ncbi.nlm.nih.gov/datasets/gene/351/. Gene ID: 351.

[10] Julia TCW and Alison M Goate. "Genetics of -Amyloid Precursor Protein in Alzheimer's Disease". In: *Cold Spring Harbor Perspectives in Medicine* (2017). DOI: 10.1101/cshperspect.a024539.

[11] NCBI. *PSEN1 presenilin 1 [Homo sapiens (human)].* https://www.ncbi.nlm.nih.gov/datasets/gene/5663/. Gene ID: 5663.

[12] Jaya Bagaria, Eva Bagyinszky, and Seong Soo A An. "Genetics, Functions, and Clinical Impact of Presenilin-1 (PSEN1) Gene". In: *International Journal of Molecular Sciences* (2022). DOI: 10.3390/ijms231810970.

[13] NCBI. *APOE apolipoprotein E [Homo sapiens (human)].* https://www.ncbi.nlm.nih.gov/datasets/gene/348/. Gene ID: 348.

[14] Amy C Raulin, Symone V Doss, Zachary A Trottier, Tadafumi C Ikezu, Guojun Bu, and Chia-Chen Liu. "APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches". In: *Molecular Neurodegeneration* (2022). DOI: 10.1186/s13024-022-00574-4.

[15] NCBI. *NM_000041.4 Homo sapiens apolipoprotein E (APOE), transcript variant 2, mRNA.* https://www.ncbi.nlm.nih.gov/nuccore/NM_000041.4?report=fasta.

[16] UniProt. *P02649 · APOE_HUMAN - Apolipoprotein E.* https://www.uniprot.org/uniprotkb/P02649/entry.

[17] RCSB PDB. *PDB 1B68 - Apolipoprotein E4, 22K Fragment.* https://www.rcsb.org/structure/1B68. X-ray, 2.00 Å, residues 19-209.

[18] RCSB PDB. *PDB 1OEF - Apolipoprotein E C-terminal domain.* https://www.rcsb.org/structure/1OEF. NMR, residues 281-304.

[19] RCSB PDB. *PDB 8GRX - Apolipoprotein E structure.* https://www.rcsb.org/structure/8GRX. Cryo-EM, 3.00 Å, residues 41-180.

[20] AlphaFold DB. *AlphaFold Structure Prediction for P02649.* https://alphafold.ebi.ac.uk/entry/P02649. Predicted structure, residues 1-317.

[21] Yang Liu, Xiaocong Yang, Jianhong Gan, Shuo Chen, Zhixiang Xiao, and Yang Cao. *CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting.* https://cadd.labshare.cn/cb-dock2/index.php. 2022.

[22] Lesley R Golden, Dahlia S Siano, Isaiah O Stephens, Steven M MacLean, Kai Saito, Georgia L Nolt, Jessica L Funnell, Akhil V Pallerla, Sangderk Lee, Cathryn Smith, Jing Chen, Haining Zhu, Clairity Voy, Callie M Whitus, Gabriela Hernandez, Brandon C Farmer, Kumar Pandya, Dale O Cowley, Shannon L Macauley, Scott M Gordon, Josh M Morganti, and Lance A Johnson. *GSE300079: APOE4 to APOE2 allelic switching in mice improves Alzheimer's disease-related metabolic signatures, neuropathology and cognition.* https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE300079. 2025.

[23] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gustavo V Meirelles, Neil R Clark, and Avi Ma'ayan. *Enrichr: Gene Set Enrichment Analysis Web Server.* https://maayanlab.cloud/Enrichr/. Ma'ayan Laboratory, Icahn School of Medicine at Mount Sinai.