

# Supplementary Information for Graph AI generates hypotheses validated across neurological systems

Ayush Noori<sup>1,2,3,4,5,6,7</sup> , Joaquín Polonuer<sup>1</sup> , Katharina Meyer<sup>2,5,8</sup> , Bogdan Budnik<sup>2,5</sup> , Shad Morton<sup>2,5</sup> , Xinyuan Wang<sup>6,9</sup> , Sumaiya Nazeen<sup>6,9</sup> , Yingnan He<sup>3</sup> , Iñaki Arango<sup>1</sup> , Lucas Vittor<sup>1</sup> , Matthew Woodworth<sup>2,5,8</sup> , Richard C. Krolewski<sup>6,9</sup> , Michelle M. Li<sup>1,6</sup> , Ninning Liu<sup>2,5</sup> , Tushar Kamath<sup>10</sup>, Evan Macosko<sup>10</sup> , Dylan Ritter<sup>6,11</sup> , Jalwa Afroz<sup>6,11</sup> , Alexander B. H. Henderson<sup>3,6</sup>, Lorenz Studer<sup>6,11</sup> , Samuel G. Rodrigues<sup>12</sup> , Andrew White<sup>12</sup> , Noa Dagan<sup>7,13,14</sup> , David A. Clifton<sup>4,15</sup> , George M. Church<sup>2,5,8</sup> , Sudeshna Das<sup>3,†</sup> , Jenny M. Tam<sup>2,5,8,†</sup> , Vikram Khurana<sup>6,9,10,16,†</sup> , Marinka Zitnik<sup>1,6,7,10,17,18,†</sup> 

<sup>1</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>2</sup>Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, MA, USA

<sup>3</sup>Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

<sup>4</sup>Department of Engineering Science, University of Oxford, Oxford, UK

<sup>5</sup>BD<sup>2</sup>: Breakthrough Discoveries for thriving with Bipolar Disorder, Santa Monica, CA, USA

<sup>6</sup>Aligning Science Across Parkinson's (ASAP) Collaborative Research Network, Chevy Chase, MD, USA

<sup>7</sup>The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at

Harvard Medical School and Clalit Research Institute, Boston, MA, USA

<sup>8</sup>Department of Genetics, Harvard Medical School, Boston, MA, USA

<sup>9</sup>Department of Neurology, Brigham and Women's Hospital, Boston, MA, USA

<sup>10</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA

<sup>11</sup>The Center for Stem Cell Biology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

<sup>12</sup>FutureHouse Inc., San Francisco, CA, USA

<sup>13</sup>Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat-Gan, Israel

<sup>14</sup>Faculty of Computer and Information Science, Ben Gurion University of the Negev, Be'er Sheva, Israel

<sup>15</sup>Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou, Jiangsu, China

<sup>16</sup>Harvard Stem Cell Institute, Cambridge, MA, USA

<sup>17</sup>Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA

<sup>18</sup>Harvard Data Science Initiative, Cambridge, MA, USA

<sup>†</sup>Correspondence: [sdas5@mgh.harvard.edu](mailto:sdas5@mgh.harvard.edu), [jenny.tam@wyss.harvard.edu](mailto:jenny.tam@wyss.harvard.edu), [vkhurana@bwh.harvard.edu](mailto:vkhurana@bwh.harvard.edu), [marinka@hms.harvard.edu](mailto:marinka@hms.harvard.edu) (lead contact)

PROTON website: <https://protonmodel.ai>

PROTON code: <http://github.com/mims-harvard/PROTON>

PROTON model: <https://huggingface.co/mims-harvard/PROTON>

<sup>1</sup> This PDF file includes:

<sup>2</sup> Supplementary Notes 1 to 5

<sup>3</sup> Supplementary Figures 1 to 8

<sup>4</sup> Supplementary Tables 1 to 5

5 **Supplementary Note 1: PROTON embeddings are biomedically organized**

6 The learned embedding space of PROTON was visualized using the uniform manifold approximation  
7 and projection algorithm (UMAP) for non-linear dimensionality reduction [1, 2] implemented in the  
8 scikit-learn machine learning library [3]. We used a cosine similarity distance function, a local  
9 neighborhood size of 15, and a minimum distance of 0.8; however, UMAP results were largely invariant to  
10 these hyperparameter choices. Clustering was observed by node type identity – that is, nodes with the same  
11 biomedical identity occupied similar regions of the embedding space – suggesting that PROTON effectively  
12 projects NEUROKG into a meaningful learned latent space (Supplementary Figure 1).

13 **1.1 Related diseases cluster together in the PROTON latent space**

14 To further evaluate PROTON embeddings, we examined disease nodes. MONDO disease terms were grouped  
15 into four categories based on their hierarchical position in the MONDO disease ontology [4, 5]: neurodegen-  
16 erative diseases (descendants of [MONDO:0005559](#),  $n = 772$ ), cancers (descendants of [MONDO:000499](#),  
17  $n = 2,208$ ), infectious diseases (descendants of [MONDO:0005550](#),  $n = 1,060$ ), and an “other” category  
18 for all remaining diseases. Diseases classified under multiple parent categories were assigned to “other,”  
19 resulting in 750 neurodegenerative diseases, 2,198 cancers, and 1,028 infectious diseases. The disease nodes  
20 in NEUROKG corresponding to these MONDO disease terms formed four subgraphs: 693 neurodegenerative  
21 diseases, 982 infectious diseases, 2,117 cancers, and 18,409 other diseases.

22 The PROTON disease embedding space was visualized using cosine-similarity-based UMAP (with the  
23 same parameters as above), which revealed that neurodegenerative diseases segregate from cancers and  
24 infectious diseases in the PROTON latent space (Supplementary Figure 2a). To quantitatively assess this  
25 separation based on distance in the high-dimensional embedding space – rather than in the low-dimensional  
26 UMAP projection, which may not preserve distances or structure [6] – we computed the silhouette coefficient  
27 for each disease node:

$$S(u) = \frac{b(u) - a(u)}{\max(a(u), b(u))}$$

28 where  $a$  is the mean intra-cluster distance and  $b$  is the mean nearest-cluster distance:

$$a(u) = \frac{1}{|C_u| - 1} \sum_{v \in C_u, v \neq u} \text{dist}(u, v)$$
$$b(u) = \min_{C_k \neq C_u} \left( \frac{1}{|C_k|} \sum_{v \in C_k} \text{dist}(u, v) \right)$$

29 Here,  $C_u$  represents the set of diseases in the same cluster as  $u$ ,  $|C_u|$  is its cardinality, and  $\text{dist}(u, v)$  is the  
30 cosine-similarity-based distance between diseases  $u$  and  $v$ , given by:

$$\text{dist}(u, v) = 1 - \frac{\mathbf{x}_u \cdot \mathbf{x}_v}{|\mathbf{x}_u||\mathbf{x}_v|}$$

31 which is scaled to the  $[0, 1]$  range using min-max normalization. The term  $b(u)$  is computed by taking the  
32 minimum average distance over all clusters  $C_k$  to which  $u$  does not belong. Finally, the silhouette coefficient  
33 for each cluster was determined by averaging over all constituent diseases.

34 The silhouette coefficient ranges from  $[-1, 1]$ , where 1 indicates well-separated clusters (e.g., perfect  
35 clusters in synthetic data), 0 suggests overlapping clusters, and negative values imply potential misclassi-  
36 fications since a disease is closer to another cluster than its assigned one. As shown in Supplementary

37 Figure 2b, neurodegenerative diseases possess the highest silhouette score (0.1847), followed by cancers  
38 (0.1331) and then infectious diseases (0.0255). Diseases not assigned to a specific category have a negative  
39 score (-0.1951), suggesting they lack clear separation relative to the three selected categories. These results  
40 demonstrate that PROTON capably organizes diseases into broad medical categories in its learned embedding  
41 space.

## 42 1.2 Structurally-similar drugs cluster in the PROTON latent space

43 Next, we examined drug representations learned by PROTON. We first sought to identify groupings of  
44 drugs in NEUROKG by structural similarity. For each drug node in NEUROKG, the corresponding chemical  
45 structure was retrieved in the simplified molecular-input line-entry system (SMILES) format [7]. Isomeric  
46 SMILES structures with isotopic and chiral specifications were obtained by querying the PubChem [8]  
47 application programming interface (API) with DrugBank identifiers [9]. Using RDKit [10], an open-source  
48 cheminformatics toolkit, the SMILES representations of all drugs were converted to molecular graphs, where  
49 nodes correspond to atoms and edges correspond to chemical bonds. Based on the molecular graphs, we  
50 then created molecular fingerprints, or vector representations of the structural properties of each molecule,  
51 which encode the local chemical environment of a molecule by iteratively applying a hashing function to  
52 molecular substructures [11]. In particular, we use the Morgan fingerprint, a 2048-bit vector where bits are  
53 assigned based on the presence of circular substructures around each atom in a molecule [12]. We use a  
54 radius of 2; therefore, the hashing function is applied to all substructures within 2 bonds of each atom in  
55 the molecule. Of the 8,160 drugs in NEUROKG, SMILES structures were available, molecular graphs were  
56 created, and Morgan fingerprints were computed for 6,943 drugs (85.09%).

57 Next, for all drugs with Morgan fingerprints, the pairwise Tanimoto similarity was computed and  
58 converted to a distance metric with  $\text{dist}(d_1, d_2) = 1 - \text{Tanimoto}(d_1, d_2)$ . Finally, drugs were grouped  
59 into structurally similar clusters by agglomerative hierarchical clustering with the average distance linkage  
60 criterion and a distance threshold of 0.8. A total of 847 clusters were identified, from which the 20 largest  
61 clusters containing at least 55 molecules were analyzed. Altogether, these 20 clusters encompassed 1,844  
62 drugs; the remaining 5,099 drugs were assigned a cluster of “other.” The pharmacological similarities  
63 uniting drug clusters were interpreted by querying GPT-4o [13] with the constituent drugs of each cluster,  
64 along with the prompt, “What do these drugs have in common?” Based on model responses, selected clusters  
65 of interest were manually assigned human-readable labels.

66 The PROTON drug embedding space was visualized using cosine-similarity-based UMAP (with the  
67 same parameters as above), which revealed that drugs segregate by structural similarity in the PROTON learned  
68 embedding space (Supplementary Figure 3). The embedding-based silhouette scores of the 20 largest clusters  
69 relative to each other and to the outgroup of 5,099 uncategorized drugs are shown in Supplementary Figure 3b.  
70 Selected clusters of interest are highlighted in the drug embedding UMAP in Supplementary Figure 3a and  
71 with colored bars in Supplementary Figure 3b.

72 The largest cluster, Cluster 1 (with a cluster-wide mean silhouette score of 0.2359), contained 250  
73 nucleoside or nucleotide analogs, including reverse transcriptase inhibitors used to treat HIV/AIDS, herpes  
74 viruses, and hepatitis B virus; antimetabolite chemotherapeutic agents; and nucleotide derivatives involved  
75 in cellular metabolism. Cluster 4 (0.2076) contained 126 steroids, including glucocorticoids, mineralo-  
76 corticoids, progestins, androgens, antiandrogens, and anabolic steroids. Cluster 9 (0.0902) contained 79  
77 additional steroid hormones and hormone analogs, including estrogens, progestins, and androgens. Given  
78 that clusters 4 and 9 both contain steroid hormones, the overlap between these clusters in the latent space  
79 is encouraging (Supplementary Figure 3a). Cluster 12 (0.1303) was composed of phosphate-containing  
80 molecules primarily involved in metabolism, nucleotide biosynthesis, and cellular signaling, including

81 phosphorylated sugars and carbohydrates involved in glycolysis, gluconeogenesis, and pentose phosphate  
82 pathways; phosphate-linked nucleosides; inositol phosphates; phospholipids; and organophosphates. Cluster  
83 16 (0.2191) contained 61 benzimidazole-based compounds that impede tubulin polymerization with antipa-  
84 rasitic, antifungal, or microtubule-disrupting properties. Cluster 17 (0.3344) contained 58 cephalosporins and  
85 related  $\beta$ -lactam antibiotics that inhibit bacterial cell wall synthesis. Cluster 18 (0.1211) contained 57 drugs  
86 that modulate neurotransmitter systems, including anticholinergics, opioids and opioid-like compounds, and  
87 antihistamines. Finally, Cluster 20 (0.0905) contained 55 drugs that share a sulfonamide functional group,  
88 including selective cyclooxygenase-2 inhibitors, sulfonamide-based antiepileptics, and antibacterial sulfon-  
89 amides. By contrast, the group of 5,099 uncategorized drugs had a silhouette score of -0.1586, suggesting  
90 that they lack clear separation relative to the top 20 largest drug categories.

91 Of note, PROTON was never provided with molecular structure during pre-training; rather, PROTON was  
92 trained on the interactome of each drug represented in NEUROKG. Nonetheless, this analysis indicates that  
93 PROTON can recover the structural organization of drugs purely from their interaction networks. Given that  
94 structurally similar drugs often share similar pharmacological properties, these findings suggest that PROTON  
95 captures functionally relevant features beyond what is explicitly provided in the training data.

## 96 **Supplementary Note 2: An AI agent validates disease edges in NEUROKG**

97 We sought to evaluate whether edges in NEUROKG are high-quality and supported by evidence in the  
98 scientific literature. To do so, we used PaperQA2, a frontier language-based generative AI agent that  
99 retrieves and reasons on information from the scientific literature [14, 15]. PaperQA2 answers user questions  
100 by performing multi-step literature-grounded retrieval-augmented generation (RAG) [16] using four tools:  
101 “Paper Search,” which identifies, parses, and embeds candidate scientific papers from keywords in the user  
102 query; “Gather Evidence,” which injects summaries of relevant papers identified by top- $k$  dense vector  
103 retrieval, LLM reranking, and contextual summarization steps; “Citations Traversal,” which adds additional  
104 sources from the citation graph; and “Generate Answer,” which produces a final answer. Among the databases  
105 that PaperQA2 can retrieve papers from include arXiv, medRxiv, bioRxiv, ChemRxiv, PubMed, PubMed  
106 Central, and open-access papers.

107 We selected six neurological diseases and, for each disease, queried NEUROKG for all edges of type  
108 (disease, disease), (disease, gene/protein), (disease, drug), or (disease,  
109 exposure) in the neighborhood of that disease. There were 433 edges incident on amyotrophic lateral  
110 sclerosis, 648 edges incident on epilepsy, 1,033 edges incident on multiple sclerosis, 1,072 edges  
111 incident on Parkinson’s disease, 685 edges incident on bipolar disorder, and 326 edges incident on major  
112 depressive disorder, for a total of 4,197 edges. For each edge, we tasked PaperQA2 with evaluating whether  
113 that association is supported by evidence in the scientific or medical literature. Specifically, PaperQA2 was  
114 prompted to assign an integer score from 1 to 5 for each edge, where 1 indicates no supporting evidence  
115 found, 2 indicates weak evidence, with 1-2 papers in support; 3 indicates moderate evidence, with 3-4 papers  
116 in support; 4 indicates strong evidence, with 5-6 papers in support; and 5 indicates very strong evidence,  
117 with over 6 supporting papers or substantial experimental evidence. For each response, PaperQA2 also  
118 provided reasoning behind each rating and referenced scientific or medical sources to support its assessment,  
119 including peer-reviewed studies, clinical guidelines, or experimental data retrieved via API calls to the Open  
120 Targets platform. As a negative control, PaperQA2 was also tasked with evaluating 100 random non-edges  
121 for each disease (*i.e.*, associations between each disease and 100 randomly selected nodes of type disease,  
122 gene/protein, drug, or exposure), using the same prompt and instructions. PaperQA2 generated  
123 structured responses in eXtensible Markup Language (XML).

124 The LLM used for answer generation and contextual summarization was Claude 3.5 Sonnet (version  
125 20241022) from Anthropic, while the `text-embedding-3-large` model from OpenAI was used to  
126 generate embedding vectors of parsed paper sections. PaperQA2 was queried using the `pqapi` package  
127 (version 7.2.0) in Python (version 3.11.10). From a total of 4,797 queries to PaperQA2, the AI agent could  
128 not generate an answer for 13 responses, six of which were negative control queries. Among the successful  
129 experimental queries, for each query and response, PaperQA2 parsed and reviewed an average of  $14.50 \pm 5.75$   
130 (mean  $\pm$  SD,  $n = 4,197$ ) papers, cited average  $5.43 \pm 1.79$  references in the final answer, and required average  
131  $140.48 \pm 81.99$  seconds to generate a response at an average cost of  $\$0.26 \pm 0.09$  per query.

132 PaperQA2 identified scientific evidence to support 83.87% of edges ( $n = 3,514$ ) (Supplementary  
133 Figure 6). Of those edges, 35.37% were weakly supported (*i.e.*, 1-2 papers,  $n = 1,243$ ), 28.03% were  
134 moderately supported (*i.e.*, 3-4 papers,  $n = 985$ ), 12.64% were strongly supported (*i.e.*, 5-6 papers,  $n = 444$ ),  
135 and 23.96% were very strongly supported (*i.e.*, greater than 6 papers,  $n = 842$ ). Only 16.13% of edges  
136 were not supported by any evidence ( $n = 676$ ); of these, the majority were gene-disease associations, which  
137 is expected, as many gene-disease associations were derived directly from experimental evidence from the  
138 Open Targets Platform or other sources and may not yet be supported by published papers in the scientific  
139 literature. By contrast, for the negative control of random non-edges, 75.54% ( $n = 448$ ) of non-edges were  
140 not supported by any evidence, and only 1.85% ( $n = 11$ ) of edges were very strongly supported. These  
141 results suggest that, across several neurological diseases, associations in NEUROKG are high-quality, factual,  
142 and can be supported by the scientific and medical literature.

### 143 **Supplementary Note 3: NEUROKG encompasses nervous system entities**

144 NEUROKG offers broad coverage of biomedical entities in the human nervous system across health and  
145 disease. For example, according to the Human Protein Atlas (HPA), 16,465 human protein-coding genes are  
146 expressed in the brain, based on the highest expression value of the gene detected across 13 brain regions, the  
147 spinal cord, and the corpus callosum [17]. Of those, 16,089 genes, or 97.72%, are represented in NEUROKG.  
148 Similarly, of the 3,195 neuroanatomical entities included within the neurological (*i.e.*, “nervous-minimal”)  
149 subset of the Uberon integrated anatomical ontology [18, 19], 2,938 entities, or 91.96%, are represented in  
150 NEUROKG. Finally, of the 5,554 neurological diseases that are descendants of the “nervous system disorder”  
151 term ([MONDO:0005071](#)) in the MONDO disease ontology [4, 5], 5,130 diseases, or 92.37%, are represented  
152 in NEUROKG. Altogether, the coverage of neuroanatomical features, brain-expressed genes, and neurological  
153 diseases suggests that NEUROKG may be a valuable resource for research into the human nervous system.

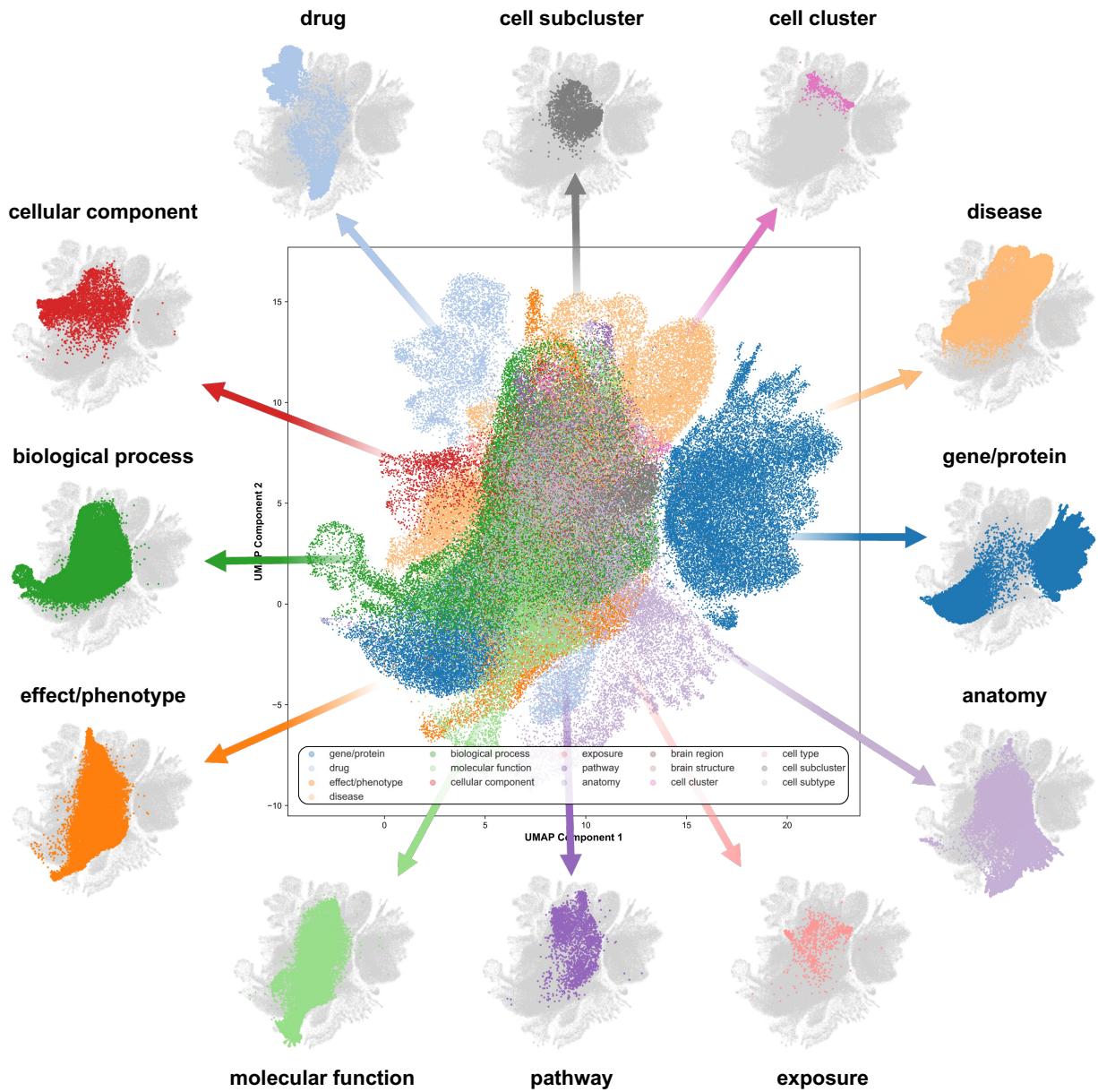
### 154 **Supplementary Note 4: NEUROKG is more comprehensive than current KGs**

155 Even with the stringent filtering steps taken during its construction, NEUROKG is more comprehensive than  
156 other state-of-the-art biomedical KGs. For example, PrimeKG has 129,375 nodes and 4,050,249 edges across  
157 10 node types and 30 edge types [20]; [ogbl-biokg](#) has 93,773 nodes and 5,088,434 edges [21]; Hetionet  
158 features only 47,031 nodes and 2,250,197 edges across 11 node types and 24 edge types [22]; PharmKG has  
159 7,603 nodes and 500,958 edges [23]; DRKG has 97,238 nodes and 5,874,261 edges; and AlzKB has 118,902  
160 nodes and 1,309,527 edges [24]. By the numbers of nodes, edges, and databases considered, these graphs all  
161 contain less information than NEUROKG. For instance, when compared to PrimeKG, which shares a similar  
162 graph construction methodology as NEUROKG, 8,038 genes exist in NEUROKG but not in PrimeKG, while  
163 only 511 genes exist in PrimeKG but not in NEUROKG. These differences are driven by increased coverage of

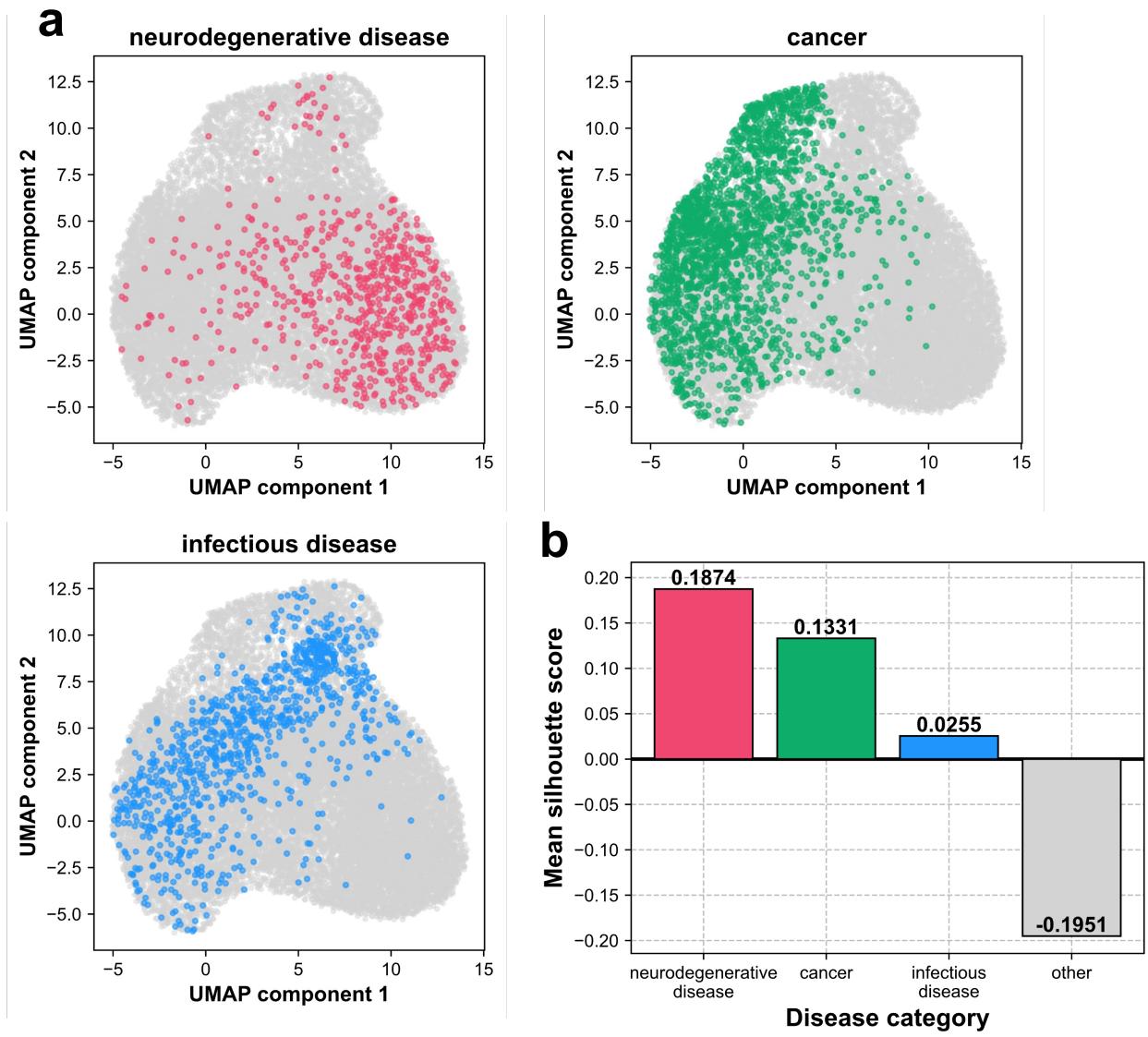
<sup>164</sup> both coding and non-coding genes: NEUROKG features 1,913 microRNAs, 2,063 long intergenic non-protein  
<sup>165</sup> coding RNAs, and various pseudogenes. Further, NEUROKG is unique in integrating scRNA-seq data at  
<sup>166</sup> subcluster resolution, enabling patient single-cell data analyses.

<sup>167</sup> **Supplementary Note 5: PROTON makes consistent predictions across related  
168 diseases**

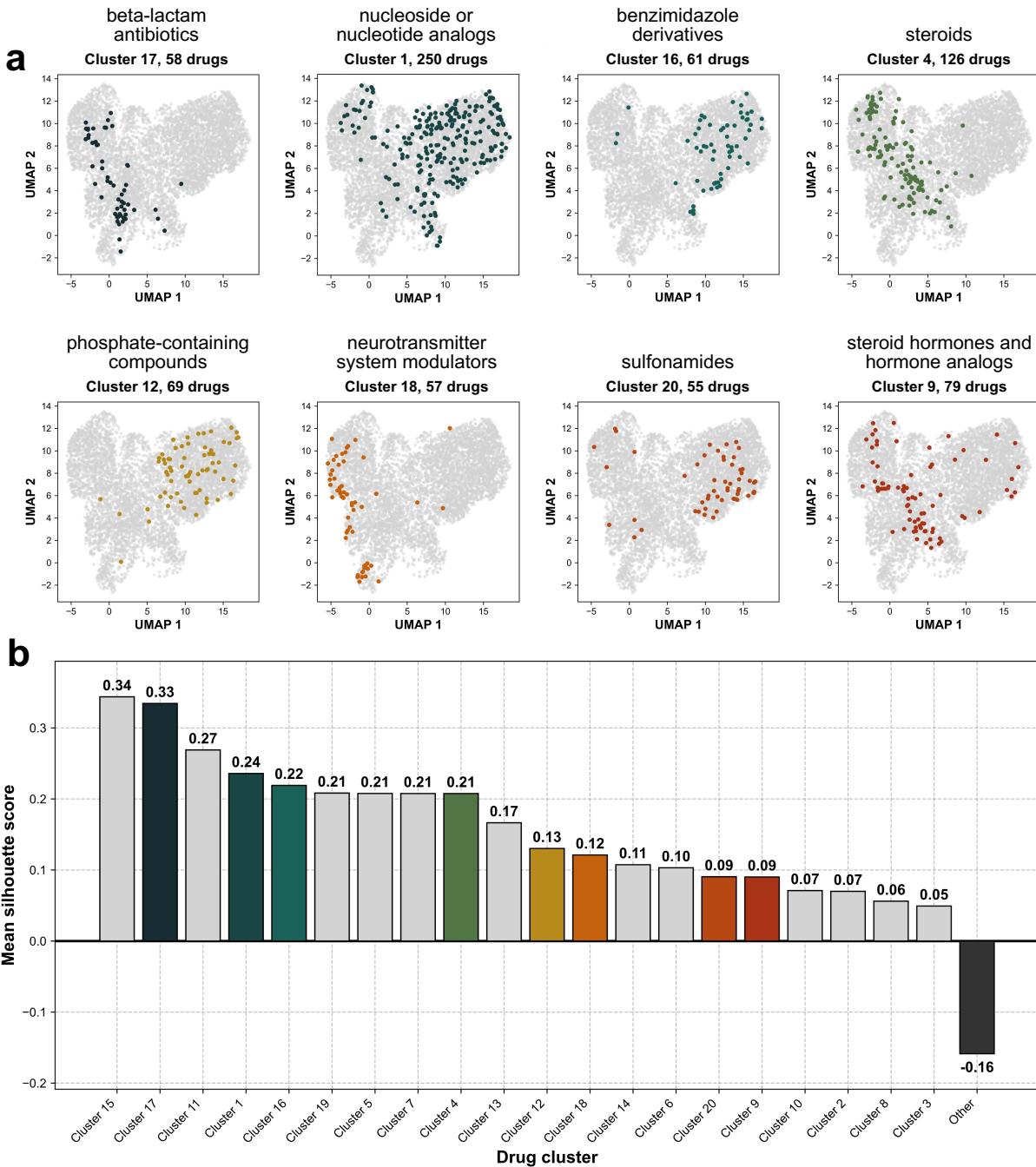
<sup>169</sup> We also assessed whether PROTON’s predictions are consistent across clinically related diseases. We reasoned  
<sup>170</sup> that a drug relevant to one disease should be similarly prioritized for related conditions. To evaluate this,  
<sup>171</sup> we used PROTON (trained on the full NEUROKG graph) to generate therapeutic predictions across 25 neuro-  
<sup>172</sup> logical diseases and all 8,160 drugs; then, we computed pairwise Spearman rank correlations across disease  
<sup>173</sup> pairs. To exclude consistently low-ranked drugs that may be poorly annotated or not relevant to neurological  
<sup>174</sup> disease, we focused on the 3,071 drugs that appear at rank  $\leq 2,000$  in at least one of the 25 evaluation  
<sup>175</sup> diseases. Diseases were grouped into five categories: neurodegenerative, neuropsychiatric, neuroinfectious,  
<sup>176</sup> neurological cancer, and other. Encouragingly, PROTON indeed makes similar predictions for related diseases  
<sup>177</sup> (Supplementary Figure 8a). Brain cancers – including oligodendrogloma, medulloblastoma, astrocytoma,  
<sup>178</sup> ependymoma, craniopharyngioma, and meningioma – had strong intragroup and low intergroup correla-  
<sup>179</sup> tion. Similarly, neuroinfectious diseases, including meningitis, neurosyphilis, and encephalitis, are clustered  
<sup>180</sup> tightly. Interestingly, neuropsychiatric conditions – such as anxiety, autism spectrum disorder, attention  
<sup>181</sup> deficit hyperactivity disorder (ADHD), schizophrenia, and bipolar disorder – clustered alongside neurode-  
<sup>182</sup> generative diseases, suggesting overlapping pharmacological profiles captured by PROTON. This is consistent  
<sup>183</sup> with clinical observations: behavioral and neuropsychiatric symptoms are prevalent in neurodegenerative  
<sup>184</sup> disorders like Alzheimer’s disease [25] and Parkinson’s [26] disease. Further, several drugs originally de-  
<sup>185</sup> veloped for neuropsychiatric disease – including brexpiprazole (approved for schizophrenia), escitalopram  
<sup>186</sup> (approved for major depression), methylphenidate (approved for ADHD), and lithium (approved for BD) –  
<sup>187</sup> are in phase II or phase III clinical trials for Alzheimer’s disease [27].



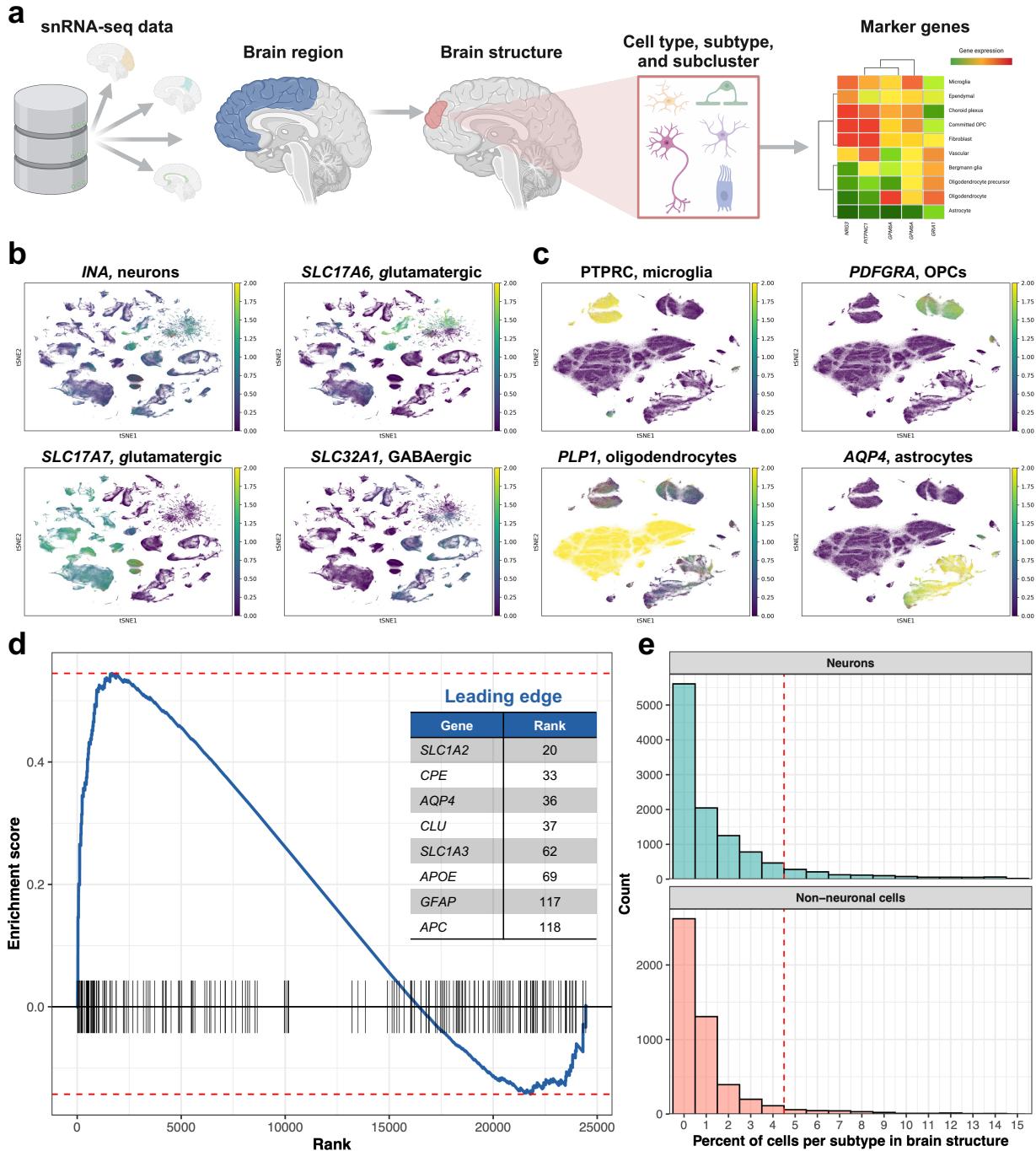
**Supplementary Figure 1: PROTON embeddings are organized by biomedical identities and relationships.** The learned embedding space of PROTON was visualized using the UMAP algorithm for non-linear dimensionality reduction. Nodes clustered based on their biomedical identity, suggesting that PROTON successfully maps NEUROKG into a structured and meaningful latent representation.



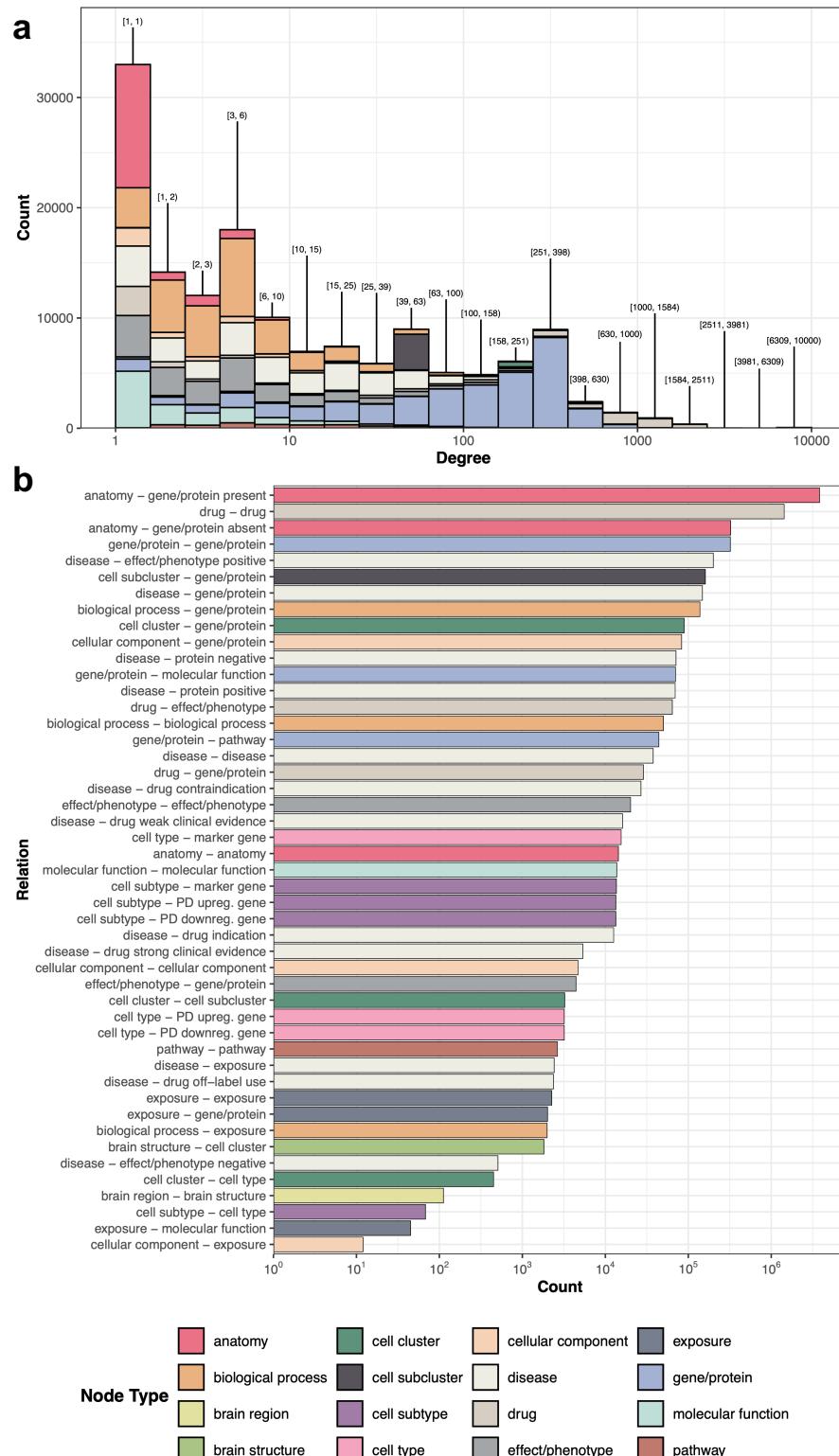
**Supplementary Figure 2: Related diseases cluster together in the learned PROTON latent space.** (a) A UMAP projection of disease embeddings based on cosine similarity shows that neurodegenerative diseases (red) segregate from cancers (green) and infectious diseases (blue) relative to all other diseases (gray). (b) Silhouette scores quantify the degree of separation between disease categories in the high-dimensional embedding space. Note that the separation between the three disease groups is relative to the fourth group of uncategorized diseases.



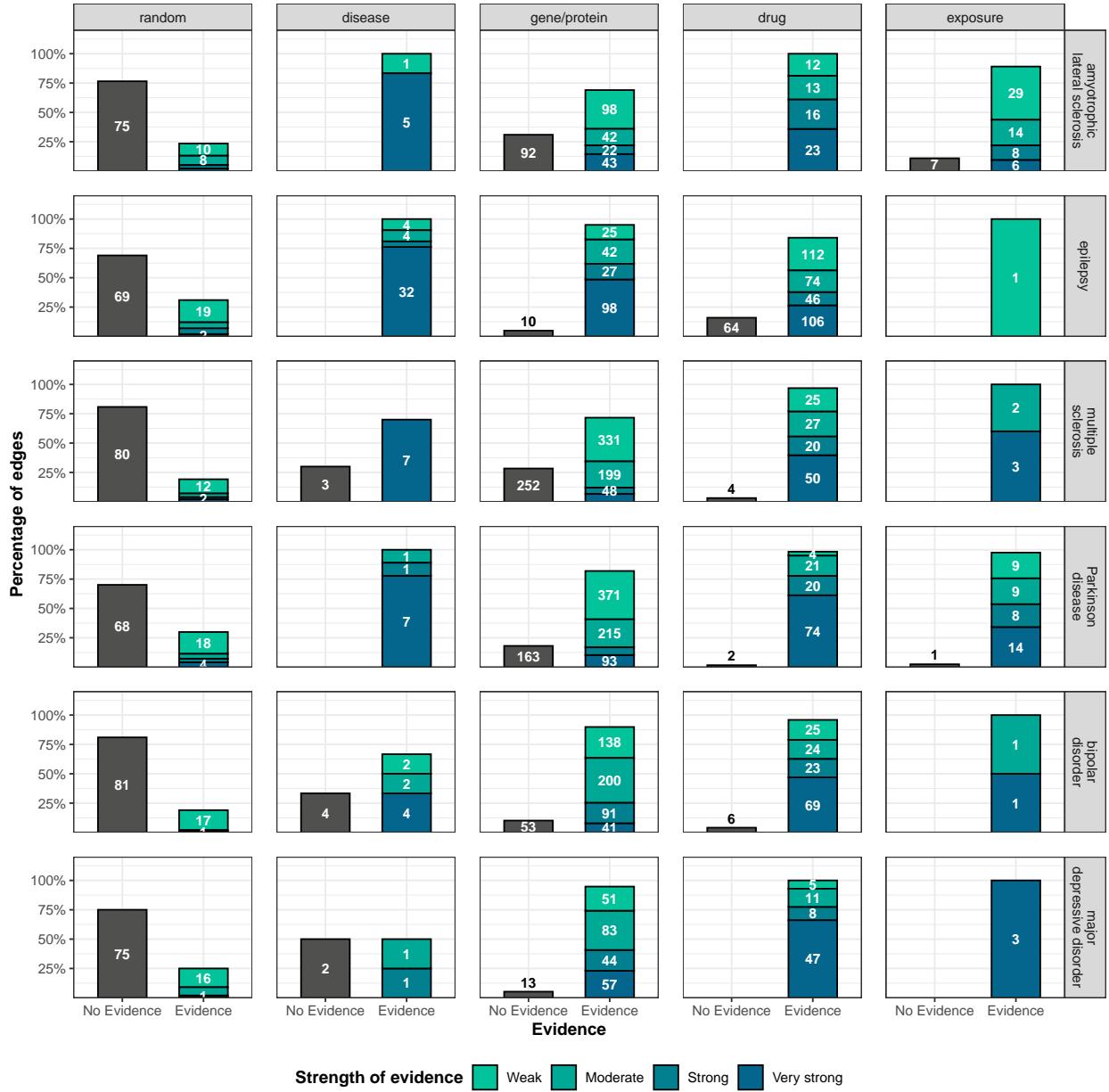
**Supplementary Figure 3: Structurally-similar drugs cluster together in the learned PROTON latent space.** (a) A UMAP projection of drug embeddings based on cosine similarity shows that drugs are grouped by structural and functional similarity. Colored points belong to the selected cluster, while gray points do not. (b) Silhouette scores quantify the degree of separation between drug categories in the high-dimensional embedding space. Note that the separation between the 20 largest clusters (containing a total of 1,844 drugs) is relative to the outgroup of 5,099 uncategorized drugs. Colored bars correspond to selected clusters of interest highlighted in panel (a).



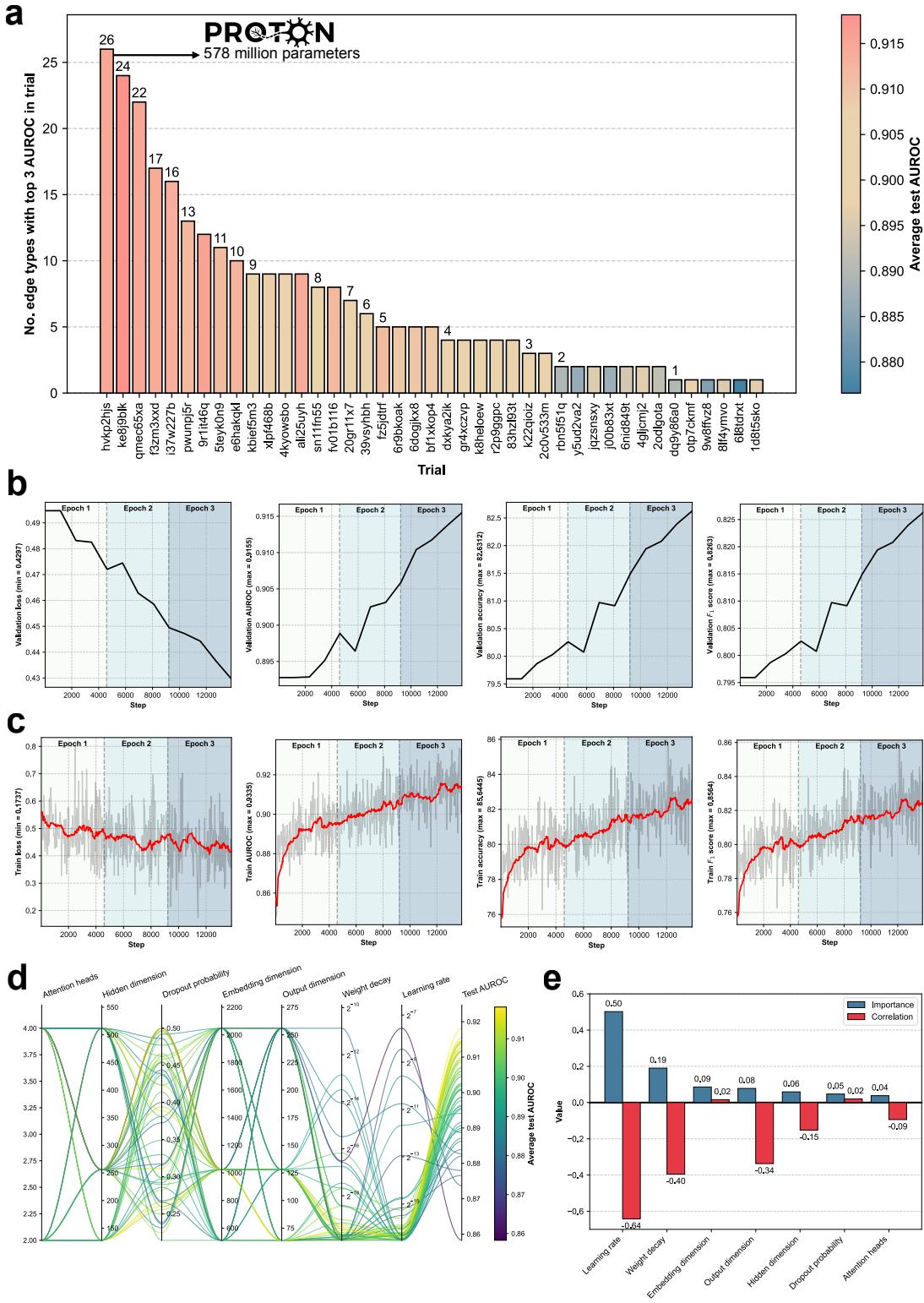
**Supplementary Figure 4: Analysis of 2,480,956 neurons and 888,263 non-neuronal cells integrated into NEUROKG.** (a) Structure of snRNA-seq derived edges in NEUROKG. Edges were constructed between brain regions, brain structures, cell types, cell subtypes, cell subclusters, and marker genes. (b) Neuronal and (c) non-neuronal cell types identified by Siletti *et al.* express known cell type-specific markers. See Figure 1 of Siletti *et al.* [28] for reference. (d) Astrocyte marker genes are significantly enriched for known immunohistochemical markers of astrocytes under gene set enrichment analysis ( $p = 1.612 \times 10^{-4}$ , ES = 0.545, NES = 1.623). (e) Proportions of cells of each subtype present per each brain structure. Most brain structures contain less than 5% of each subtype.



**Supplementary Figure 5: NEUROKG is a knowledge graph with 147,020 nodes and 7,366,745 edges.**  
**(a)** Distribution of degree by node type, see also Supplementary Table 2. **(b)** Number of edges per edge type. Bars are colored by the type of the tail node. Note that the x-axes of both panels are in logarithmic scale.

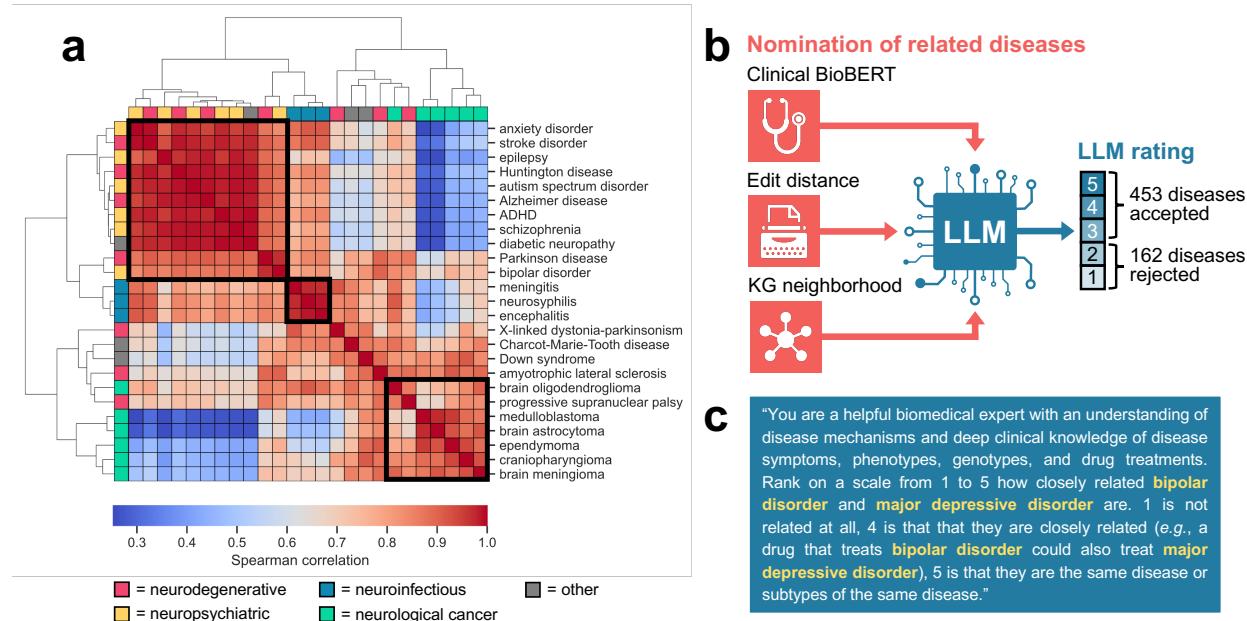


**Supplementary Figure 6: Disease edges in NEUROKG are supported by literature evidence retrieved by an AI agent.** PaperQA2, a frontier language-based generative AI agent that retrieves and reasons on information from the scientific literature [14], was used to assess the presence or absence of evidence supporting edges of type (disease, disease), (disease, gene/protein), (disease, drug), or (disease, exposure) incident on six neurological diseases in NEUROKG. The PaperQA2 AI agent identified evidence in the scientific literature to support 83.87% of edges ( $n = 3,514$ ). By contrast, 75.54% ( $n = 448$ ) of randomly sampled non-edges were not supported by any evidence (leftmost column). The height of bars represents group-wise percentages, while numeric labels represent edge count.



**Supplementary Figure 7: Performance across 55 hyperparameter optimization trials to design the PROTON architecture.** (a) For each trial, the number of edge types in which that trial is among the top 3 best-performing trials (based on edge-type-specific test AUROC) is shown. (cont.)

**Supplementary Figure 7:** (cont.) PROTON (trial ID: hvkp2hjs) is best performing in 26 edge types. **(b)** Validation loss, AUROC, accuracy, and F<sub>1</sub> score across three epochs of pre-training for a representative trial. **(c)** Pre-training loss, AUROC, accuracy, and F<sub>1</sub> score for a representative trial. Values are sampled for visualization. The rolling window average with a window size of 25 is shown in red. **(d)** Parallel coordinate plot of hyperparameter settings across trials. Trials are colored by test AUROC. **(e)** Importance of each hyperparameter with respect to test AUROC, as well as correlation with test AUROC.



**Supplementary Figure 8: PROTON makes consistent predictions across related diseases.** **(a)** PROTON makes similar predictions for related diseases based on Spearman correlation of the 3,071 drugs that appear at rank  $\leq 2,000$  in at least one of the 25 diseases. Diseases are grouped into neurodegenerative, neuropsychiatric, neuroinfectious, neurological cancer, and other categories. **(b)** The process of identifying diseases for inclusion in the disease-centric split is depicted. For each of 17 neurological diseases, related diseases were nominated for inclusion using Clinical BioBERT embedding similarity [29], Levenshtein edit distance between disease names, or one-hop neighborhood overlap in NEUROKG. Candidate diseases were reviewed and rated by GPT-4o, with diseases scoring  $\geq 3$  included in the split. **(c)** An example of the prompt used for disease similarity rating; in this case, comparing bipolar disorder and major depressive disorder.

Cell type	Total cells	Subtypes	Subclusters
<b>Neurons</b>			
Upper-layer intratelencephalic	452,715	16	139
Splatter	273,703	90	1135
CGE interneuron	226,007	21	180
MGE interneuron	219,640	26	211
Deep-layer intratelencephalic	209,135	14	127
Medium spiny neuron	150,318	18	143
Upper rhombic lip	136,377	6	21
Midbrain-derived inhibitory	124,243	12	64
Amygdala excitatory	106,706	18	138
Deep-layer corticothalamic and 6b	77,365	17	116
Thalamic excitatory	75,529	17	109
Hippocampal dentate gyrus	67,245	7	26
Hippocampal CA1-3	52,990	14	73
Lower rhombic lip	49,523	6	25
LAMP5-LHX6 and Chandelier	44,519	12	55
Eccentric medium spiny neuron	39,628	14	96
Miscellaneous	24,422	22	102
Deep-layer near-projecting	18,736	13	72
Mammillary body	14,205	9	37
Cerebellar inhibitory	13,974	10	38
Hippocampal CA4	10,575	9	40
<b>Non-neuronal cells</b>			
Oligodendrocyte	489,306	8	24
Astrocyte	154,679	13	71
Oligodendrocyte precursor	105,640	5	19
Microglia	91,432	9	30
Vascular	9,792	11	41
Fibroblast	9,036	8	25
Bergmann glia	8,039	1	1
Choroid plexus	7,404	7	22
Ependymal	5,432	10	43
Committed oligodendrocyte precursor	4,538	7	21

**Supplementary Table 1: Neuronal and non-neuronal cell types in Siletti *et al.*** Cell type counts are shown after pre-processing and quality control analyses.

<b>Node type</b>	<b>Number of nodes</b>	<b>Average degree</b>
gene/protein	35,198	$163.93 \pm 157.85$
biological process	27,668	$8.70 \pm 27.20$
disease	22,201	$28.42 \pm 145.41$
effect/phenotype	16,711	$18.55 \pm 81.27$
anatomy	14,384	$290.93 \pm 2197.55$
molecular function	11,228	$8.71 \pm 128.94$
drug	8,160	$370.45 \pm 537.77$
cellular component	4,054	$22.81 \pm 202.48$
cell subcluster	3,244	$50.52 \pm 3.92$
pathway	2,629	$18.79 \pm 29.59$
exposure	860	$12.80 \pm 34.45$
cell cluster	450	$210.67 \pm 11.14$
brain structure	112	$17.25 \pm 11.59$
cell subtype	68	$595.12 \pm 48.51$
cell type	39	$575.69 \pm 325.95$
brain region	14	$8.00 \pm 9.27$

**Supplementary Table 2: Node types and statistics in NEUROKG.** Average degree is reported as mean  $\pm$  SD.

<b>ID</b>	<b>Line</b>	<b>Cohort or study ID</b>	<b>Age</b>	<b>Sex</b>	<b>Diagnosis</b>
CM2	PSC-01-020	Yankner/Tsai	38	M	Control
CM3	MH0185983	NRGR Study ID 163	29	M	Control
CM4	MH0185932	NRGR Study ID 163	19	M	Control
CF4	MH0185863	NRGR Study ID 163	36	F	Control
BF1	BD-220-5	Yankner/Tsai	49	F	BD
BF3	BD-12-33	Yankner/Tsai	35	F	BD
BF4	BD-193-3	Yankner/Tsai	32	F	BD
BF5	BD-PSC-01-002	Yankner/Tsai	46	F	BD
BM4	MH0185869	NRGR Study ID 163	26	M	BD

**Supplementary Table 3: iPSC lines used to generate cerebral brain organoids.** iPSC lines in this study were obtained from Meyer *et al.* [30].

<b>ICD-10 code</b>	<b>Description</b>
<b>Alzheimer's disease and related dementias (ADRD)</b>	
F01.50	Vascular dementia without behavioral disturbance
F01.51	Vascular dementia with behavioral disturbance
F02.80	Dementia in other diseases classified elsewhere without behavioral disturbance
F02.81	Dementia in other diseases classified elsewhere with behavioral disturbance
F03.90	Unspecified dementia, unspecified severity, without behavioral disturbance, psychotic disturbance, mood disturbance, and anxiety
F03.91	Unspecified dementia, unspecified severity, with behavioral disturbance
F04	Amnestic disorder
G30.0	Alzheimer's disease with early onset
G30.1	Alzheimer's disease with late onset
G30.8	Other Alzheimer's disease
G30.9	Alzheimer's disease, unspecified
G31.0	Frontotemporal dementia
G31.01	Pick's disease
G31.09	Other frontotemporal neurocognitive disorder
G31.1	Senile degeneration of brain, not elsewhere classified
G31.8	Other specified degenerative disease of nervous system
G31.83	Neurocognitive disorder with Lewy bodies
G31.84	Mild cognitive impairment of uncertain or unknown etiology
G31.85	Corticobasal degeneration
G31.89	Other specified degenerative disease of nervous system
G31.9	Degenerative disease of nervous system, unspecified
R41.81	Age-related cognitive decline
<b>Type 2 diabetes mellitus</b>	
E11*	Type 2 diabetes mellitus, including all subcategories
<b>Wet age-related macular degeneration</b>	
H35.32	Exudative age-related macular degeneration
<b>Hypercholesterolemia</b>	
E78.00	Pure hypercholesterolemia, unspecified
<b>Hyperlipidemia</b>	
E78*	Disorders of lipoprotein metabolism and other lipidemias*
<b>Hypertension</b>	
I10*	Essential hypertension*

**Supplementary Table 4: ICD-10 diagnosis codes.** Diagnosis codes were used to define dementia and non-dementia indications for emulated target trials. For hyperlipidemia and hypertension, ICD codes were further refined using SQL LIKE clauses to those containing the phrases “hyperlipidemia” or “hypertension,” respectively. This step improved precision in cohort definitions (*e.g.*, by excluding secondary hypertensive disorders from the hypertension cohort).

Sweep	AUROC	Accuracy	AP	F <sub>1</sub> score	Loss
<b>Training metrics</b>					
ke8j9blk	0.9217	0.8271	0.9158	0.8271	0.6093
9r1it46q	0.9133	0.8252	0.9059	0.8252	0.6645
qmec65xa	0.9189	0.8350	0.9104	0.8350	0.3970
<b>hvkp2hjs</b>	0.9129	0.8232	0.9064	0.8232	0.4939
ali25uyh	0.9103	0.8223	0.9068	0.8223	0.5563
f3zm3xxd	0.9115	0.8232	0.9049	0.8232	0.4369
i37w227b	0.9229	0.8311	0.9165	0.8311	0.2383
e6hakqkl	0.9089	0.8223	0.9001	0.8223	0.5751
pwunpj5r	0.9125	0.8223	0.9088	0.8223	0.3639
fv01b116	0.9272	0.8350	0.9213	0.8350	0.3425
<b>Validation metrics</b>					
ke8j9blk	0.9179	0.8289	0.9115	0.8289	0.4291
9r1it46q	0.9155	0.8263	0.9092	0.8263	0.4297
qmec65xa	0.9143	0.8244	0.9076	0.8244	0.4345
<b>hvkp2hjs</b>	0.9148	0.8228	0.9087	0.8228	0.4354
ali25uyh	0.9145	0.8225	0.9086	0.8225	0.4465
f3zm3xxd	0.9144	0.8237	0.9082	0.8237	0.4438
i37w227b	0.9136	0.8219	0.9072	0.8219	0.4370
e6hakqkl	0.9127	0.8214	0.9065	0.8214	0.4571
pwunpj5r	0.9119	0.8209	0.9051	0.8209	0.4329
fv01b116	0.9116	0.8203	0.9056	0.8203	0.4519
<b>Test metrics</b>					
ke8j9blk	0.9182	0.8293	0.9118	0.8293	0.4291
9r1it46q	0.9155	0.8265	0.9091	0.8265	0.4329
qmec65xa	0.9145	0.8246	0.9081	0.8246	0.4357
<b>hvkp2hjs</b>	0.9145	0.8223	0.9085	0.8223	0.4356
ali25uyh	0.9142	0.8222	0.9083	0.8222	0.4512
f3zm3xxd	0.9140	0.8235	0.9077	0.8235	0.4466
i37w227b	0.9136	0.8220	0.9073	0.8220	0.4348
e6hakqkl	0.9127	0.8215	0.9066	0.8215	0.4546
pwunpj5r	0.9119	0.8206	0.9051	0.8206	0.4395
fv01b116	0.9118	0.8206	0.9059	0.8206	0.4505

**Supplementary Table 5: Pre-training metrics of top 10 hyperparameter optimization trials ranked by test AUROC.** The trial configuration with the final selected hyperparameters is shown in bold typeface.

188 **References**

- 189 1. Healy, J. & McInnes, L. Uniform manifold approximation and projection. *Nature Reviews Methods Primers* **4**, 1–15. doi:[10.1038/s43586-024-00363-x](https://doi.org/10.1038/s43586-024-00363-x) (2024).
- 190 2. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861. doi:[10.21105/joss.00861](https://doi.org/10.21105/joss.00861) (2018).
- 191 3. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- 192 4. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world 2022. doi:[10.1101/2022.04.13.22273750](https://doi.org/10.1101/2022.04.13.22273750).
- 193 5. Shefczek, K. A. *et al.* The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research* **48**, D704–D715. doi:[10.1093/nar/gkz997](https://doi.org/10.1093/nar/gkz997) (2020).
- 194 6. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLOS Computational Biology* **19**, e1011288. doi:[10.1371/journal.pcbi.1011288](https://doi.org/10.1371/journal.pcbi.1011288) (2023).
- 195 7. Weininger, D. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences* **28**, 31–36. doi:[10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005) (1988).
- 196 8. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Research* **53**, D1516–D1525. doi:[10.1093/nar/gkae1059](https://doi.org/10.1093/nar/gkae1059) (2025).
- 197 9. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082. doi:[10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037) (2018).
- 198 10. Landrum, G. *et al.* RDKit: Open-source cheminformatics. 2022. doi:[10.5281/zenodo.6483170](https://doi.org/10.5281/zenodo.6483170).
- 199 11. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S. & Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods. Virtual Screening* **71**, 58–63. doi:[10.1016/j.ymeth.2014.08.005](https://doi.org/10.1016/j.ymeth.2014.08.005) (2015).
- 200 12. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-a Technique Developed at Chemical Abstracts Service. *Journal of chemical documentation* **5**, 107–113 (1965).
- 201 13. OpenAI *et al.* GPT-4o System Card 2024. doi:[10.48550/arXiv.2410.21276](https://doi.org/10.48550/arXiv.2410.21276).
- 202 14. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge 2024. doi:[10.48550/arXiv.2409.13740](https://doi.org/10.48550/arXiv.2409.13740).
- 203 15. Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodrigues, S. G. & White, A. D. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research 2023. doi:[10.48550/arXiv.2312.07559](https://doi.org/10.48550/arXiv.2312.07559).
- 204 16. Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks in Advances in Neural Information Processing Systems **33** (Curran Associates, Inc., 2020), 9459–9474.
- 205 17. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947. doi:[10.1126/science.aay5947](https://doi.org/10.1126/science.aay5947) (2020).

- 227 18. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an  
228 integrative multi-species anatomy ontology. *Genome Biology* **13**, R5. doi:[10.1186/gb-2012-13-1-r5](https://doi.org/10.1186/gb-2012-13-1-r5) (2012).
- 230 19. Haendel, M. A. *et al.* Unification of multi-species vertebrate anatomy ontologies for comparative  
231 biology in Uberon. *Journal of Biomedical Semantics* **5**, 21. doi:[10.1186/2041-1480-5-21](https://doi.org/10.1186/2041-1480-5-21)  
232 (2014).
- 233 20. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine.  
234 *Scientific Data* **10**, 67. doi:[10.1038/s41597-023-01960-3](https://doi.org/10.1038/s41597-023-01960-3) (2023).
- 235 21. Hu, W. *et al.* Open graph benchmark: datasets for machine learning on graphs in *Proceedings  
236 of the 34th International Conference on Neural Information Processing Systems* (Curran  
237 Associates Inc., Red Hook, NY, USA, 2020), 22118–22133.
- 238 22. Himmelstein, D. S. *et al.* Systematic integration of biomedical knowledge prioritizes drugs  
239 for repurposing. *eLife* **6** (ed Valencia, A.) e26726. doi:[10.7554/eLife.26726](https://doi.org/10.7554/eLife.26726) (2017).
- 240 23. Zheng, S. *et al.* PharmKG: a dedicated knowledge graph benchmark for biomedical data  
241 mining. *Briefings in Bioinformatics* **22**, bbaa344. doi:[10.1093/bib/bbaa344](https://doi.org/10.1093/bib/bbaa344) (2021).
- 242 24. Romano, J. D. *et al.* The Alzheimer’s Knowledge Base: A Knowledge Graph for Alzheimer  
243 Disease Research. *Journal of Medical Internet Research* **26**, e46777 (2024).
- 244 25. Lyketsos, C. G. *et al.* Neuropsychiatric symptoms in Alzheimer’s disease. *Alzheimer’s &  
245 Dementia* **7**, 532–539. doi:[10.1016/j.jalz.2011.05.2410](https://doi.org/10.1016/j.jalz.2011.05.2410) (2011).
- 246 26. Aarsland, D., Marsh, L. & Schrag, A. Neuropsychiatric symptoms in Parkinson’s disease.  
247 *Movement Disorders* **24**, 2175–2186. doi:[10.1002/mds.22589](https://doi.org/10.1002/mds.22589) (2009).
- 248 27. Ihara, M. & Saito, S. Drug Repositioning for Alzheimer’s Disease: Finding Hidden Clues in  
249 Old Drugs. *Journal of Alzheimer’s Disease* **74**, 1013–1028. doi:[10.3233/JAD-200049](https://doi.org/10.3233/JAD-200049) (2020).
- 250 28. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science*  
251 **382**, eadd7046. doi:[10.1126/science.add7046](https://doi.org/10.1126/science.add7046) (2023).
- 252 29. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings in *Proceedings of the 2nd  
253 Clinical Natural Language Processing Workshop* (Association for Computational Linguistics,  
254 Minneapolis, Minnesota, USA, 2019), 72–78.
- 255 30. Meyer, K. *et al.* Impaired neural stress resistance and loss of REST in bipolar disorder.  
256 *Molecular Psychiatry* **29**, 153–164. doi:[10.1038/s41380-023-02313-7](https://doi.org/10.1038/s41380-023-02313-7) (2024).