DrugReX: an explainable drug repurposing system powered by large language models and literature-based knowledge graph

Liang-Chin Huang[1], Hunki Paek[1], Kyeryoung Lee[1], Ediz Calay[1], Deepak Pillai[1], Nneka Ofoegbu[1], Bin Lin[1], Hua Xu[2], and Xiaoyan Wang[1, 3, *]

[1]IMO Health, Rosemont, IL, 60018, USA
[2]Department of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, 06510, USA
[3]Department of Health Policy and Management, Tulane University, New Orleans, LA, 70118, USA
[*]xw108@caa.columbia.edu

Supplementary Figures and Tables

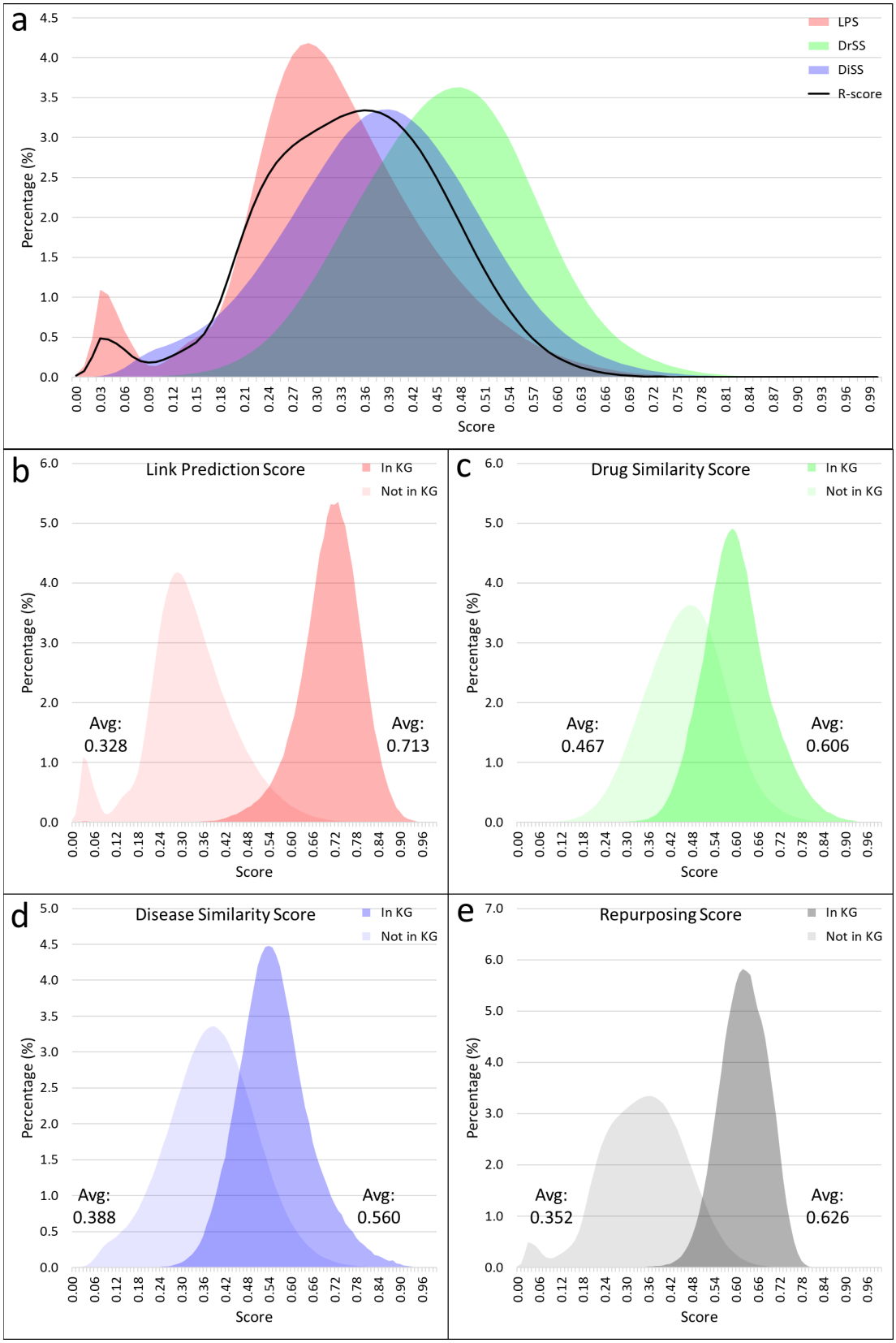

Figure S1. Distributions of different scores in the scoring system. LPS: Link Prediction Score; DrSS: Drug Similarity Score; DiSS: Disease Similarity Score; R-score: Repurposing Score; KG: knowledge graph. The average score for each group is labeled.

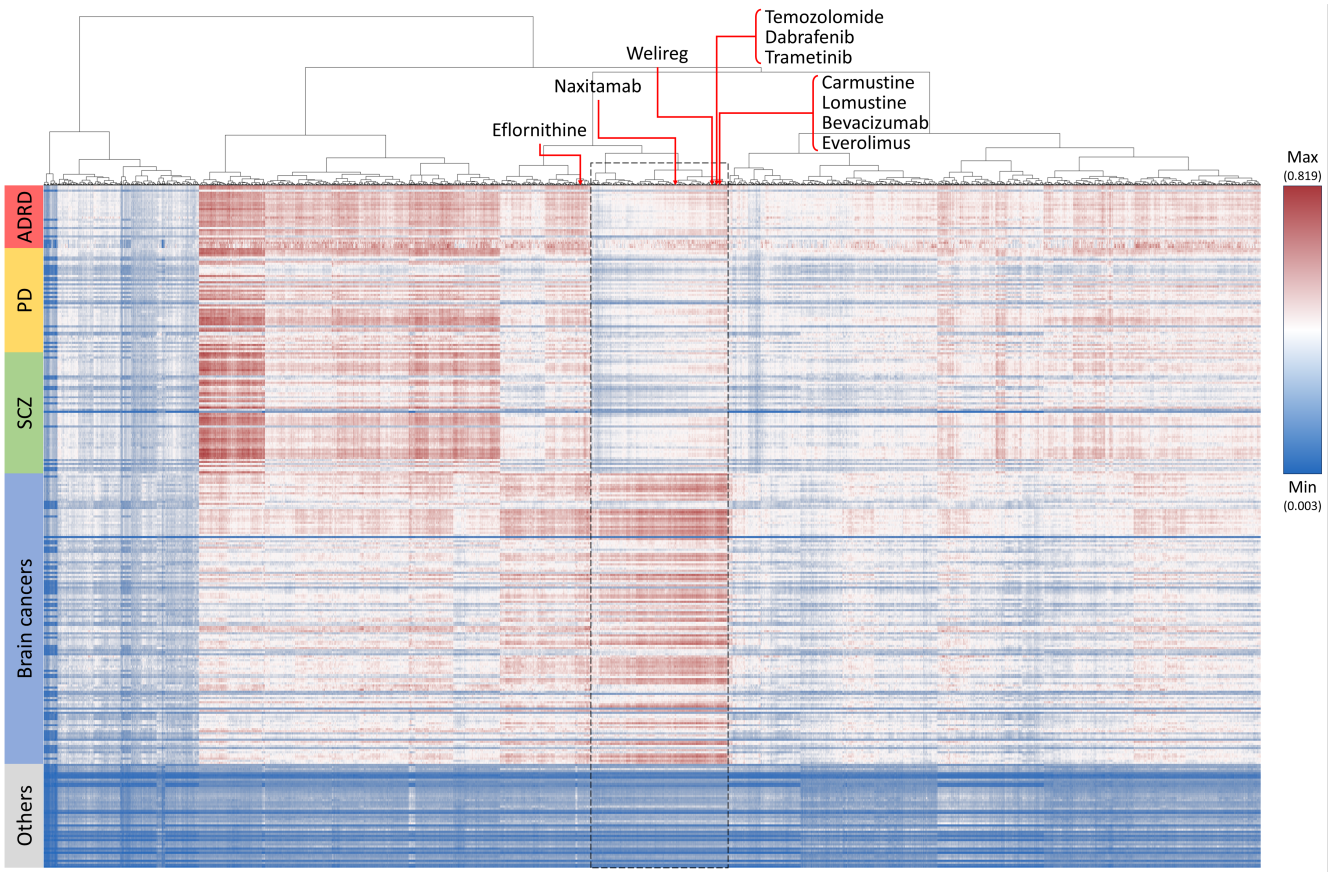

Figure S2. Clustering analysis for FDA-approved brain cancer drugs. Brain cancer drugs are indicated by red arrows. PD: Parkinson's disease; SCZ: schizophrenia.

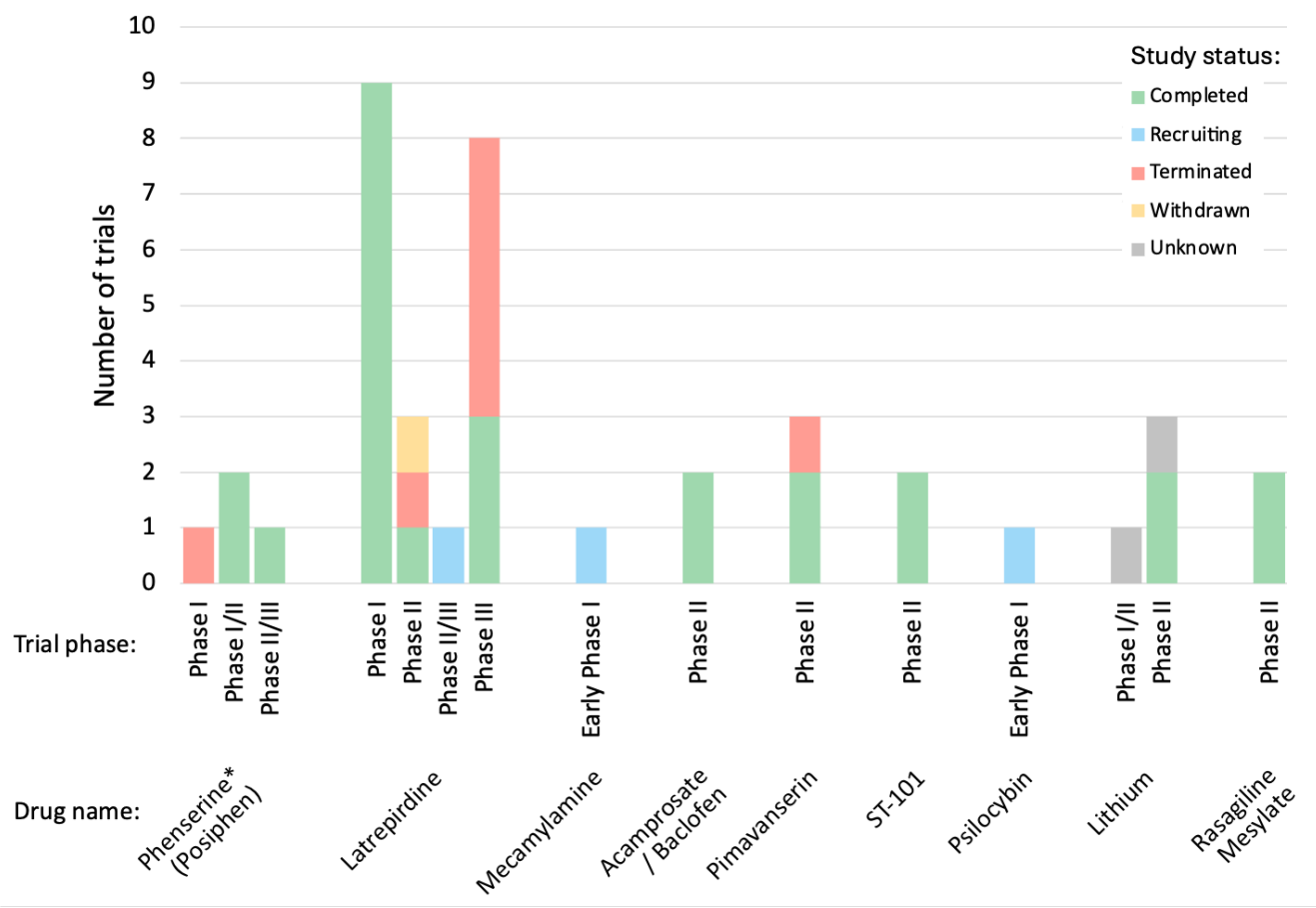

Figure S3. Candidate ADRD drugs with clinical trials for treating ADRD.

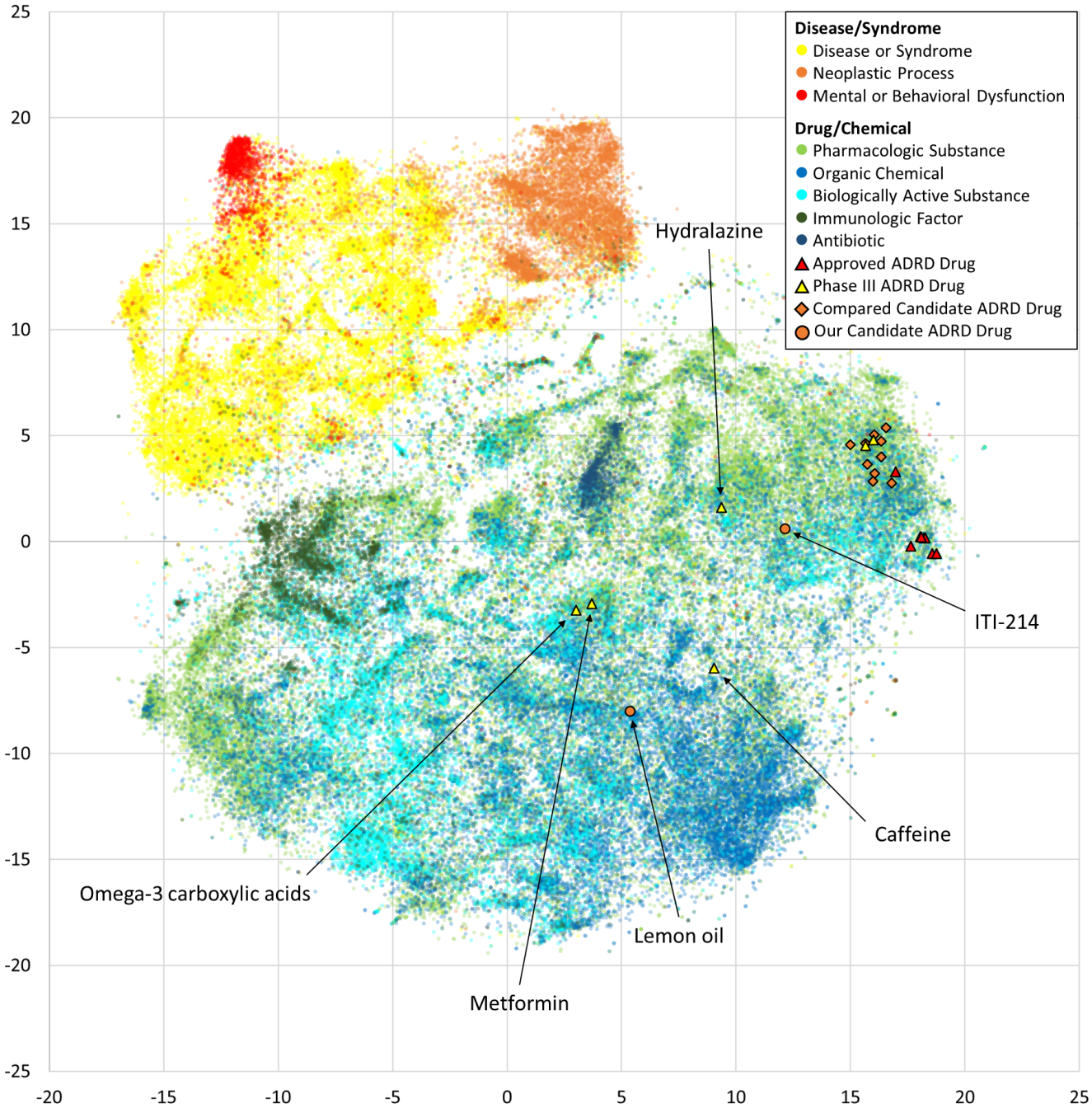

Figure S4. Visualization of compared candidate drugs embedded in a 2D space.

Iteration	BERT$_{cased}$	BERT$_{uncased}$	BioBERT	BioClinicalBERT	BlueBERT	PubMedBERT
1	0.6667	0.6651	0.7625	0.7410	0.7316	0.7919
2	0.6566	0.6451	0.7372	0.7355	0.7308	0.7697
3	0.6724	0.6690	0.7467	0.7425	0.7337	0.7805
4	0.6391	0.6635	0.7302	0.7476	0.7504	0.7788
5	0.6612	0.6651	0.7384	0.7267	0.7352	0.7886
6	0.6599	0.6620	0.7851	0.7428	0.7379	**0.8082**
7	0.6590	0.6675	0.7925	0.7446	0.7061	0.7912
8	0.6737	0.6468	0.7606	0.7524	0.6881	0.7820
9	0.6635	0.6675	0.7540	0.7414	0.7234	0.7610
10	0.6643	0.6667	0.7910	0.7206	0.7197	0.7880

Table S1. Performance (F_1 score) of relation classification using BERT-based models. The best performing model is highlighted in bold.

Entity type	Semantic type defined by UMLS	Count
Drug/Chemical	Pharmacologic Substance	48,461
	Organic Chemical	26,086
	Biologically Active Substance	13,254
	Immunologic Factor	4,171
	Antibiotic	1,774
	Subtotal	93,746
Gene/Protein	Gene or Genome	41,800
	Enzyme	8,702
	Amino Acid, Peptide, or Protein	7,640
	Nucleic Acid, Nucleoside, or Nucleotide	1,528
	Subtotal	59,670
Disease/Syndrome	Disease or Syndrome	25,972
	Neoplastic Process	7,898
	Mental or Behavioral Dysfunction	1,746
	Subtotal	35,616
Pathway/Function	Molecular Function	7,369
	Cell Function	3,559
	Genetic Function	1,140
	Subtotal	12,068
Others	Finding	21,276
	Therapeutic or Preventive Procedure	11,115
	Body Part, Organ, or Organ Component	4,998
	Laboratory Procedure	457
	Others	43,125
	Subtotal	80,971
Total		282,071

Table S2. Number of entities in the LitKG.

Relation type	# Relations	# Sentences	Example sentence
ASSOCIATED_WITH	6,297,142	19,261,066	Amyloid beta (Aβ) peptides are characterized as the major factors **associated with** neuron death in Alzheimer's disease[1]
CAUSES	2,789,328	9,768,918	β-amyloidosis and oxidative stress have been implicated as root **causes** of Alzheimer's disease[2]
COEXISTS_WITH	4,034,508	11,110,829	Because cerebral amyloid angiopathy-related hemorrhage often **co-exists with** AD, we examined the IL-1A polymorphism in cerebral amyloid angiopathy-related hemorrhage[3]
COMPLICATES	92,645	152,195	Alzheimer's disease **complicated** by a terminal salmonella infection[4]
INHIBITS	2,964,981	6,220,215	Berberine (BBR) can improve antioxidative capacity and **inhibit** Abeta protein aggregation and tau protein hyperphosphorylation in AD, and stem cell therapy is also increasingly recognized as a therapy for AD[5]
INTERACTS_WITH	2,970,495	7,237,927	These in vitro **interactions** between tau protein and sulfated glycosaminoglycans reproduced the known characteristics of paired helical filament-tau from Alzheimer's disease brain[6]
MANIFESTATION_OF	66,108	120,821	In its earliest clinical **manifestation**, AD often presents as Mild Cognitive Impairment (MCI), a term used to describe older adults with cognitive decline that is more severe than expected for healthy aging but does not meet standard criteria for dementia[7]
PRODUCES	701,231	2,026,850	Biochemical characterization of the gamma-secretase activity that **produces** beta-amyloid peptides[8]
STIMULATES	2,812,479	6,848,726	In conclusion, aluminum appears to **induce** isolated tau protein to aggregate in a phosphate-independent way, without the formation of fibrils[9]
TREATS	1,854,718	8,733,721	In conclusion, the well-known antidiabetic drug, metformin, could be a promising drug for AD **treatment**[10]
Total (unique)	24,583,635	66,324,693	

Table S3. Number of relations and supporting sentences in the LitKG. Subjects and objects are highlighted by underlining; relations are highlighted in bold.

Algorithm	MR	MRR	HITS@1	HITS@3	HITS@10
ComplEx	1.7391	0.8325	0.7280	**0.9306**	0.9872
DistMult	1.7738	0.8232	0.7144	0.9243	0.9870
RotatE	3.3377	0.5705	0.3793	0.7018	0.9533
TransE	2.0370	0.7710	0.6396	0.8868	0.9817
TransE_l1	**1.6969**	**0.8327**	**0.7295**	0.9276	**0.9896**
TransE_l2	2.0974	0.7676	0.6356	0.8838	0.9786

Table S4. Performance of knowledge graph embedding models. MR: mean rank; MRR: mean reciprocal rank. TransE_l1: TransE regularized by L1 normalization; TransE_l2: TransE regularized by L2 normalization. The best algorithm and performances are highlighted in bold.

Task	Hypothesis
Generating a hypothesis based on supporting sentences (”LitKG-GPT4”)	”Based on your knowledge and the following evidence, generate a hypothesis of how {DRUG} can treat Alzheimer’s disease in a paragraph and provide citations in the format [number, number, ...]. 1. {SENTENCE 1} 2. {SENTENCE 2} ...”
Generating a hypothesis based on supporting sentences and unrelated sentences (”LitKG’-GPT4”)	”Based on your knowledge and the following evidence, generate a hypothesis of how {DRUG} can treat Alzheimer’s disease in a paragraph and provide citations in the format [number, number, ...]. 1. {SENTENCE 1} 2. {UNREL_SENT 2} ...”
Generating a hypothesis based on relation types and unrelated relations (”KG-GPT4”)	”Based on your knowledge and the following evidence, generate a hypothesis of how {DRUG} can treat Alzheimer’s disease in a paragraph and provide citations in the format [number, number, ...]. 1. {SUBJECT 1} {RELATION 1} {OBJECT 1} 2. {UNREL_SUBJECT 2} {RELATION 2} {UNREL_OBJECT 2} ...”
Generating a hypothesis without further information (”GPT4”)	”Based on your knowledge, generate a hypothesis of how {DRUG} can treat Alzheimer’s disease in a paragraph.”
Generating a hypothesis by Consensus without further information (”Consensus”)	”Based on your knowledge, generate a hypothesis of how {DRUG} can treat Alzheimer’s disease in a paragraph.”

Table S5. The prompt templates for explaining the potential of new ADRD treatments. Replace {DRUG}, {SENTENCE n}, {UNREL_SENT 1}, {SUBJECT n}, {UNREL_SUBJECT n}, {RELATION n}, {OBJECT n}, and {UNREL_OBJECT n} with the candidate drug’s name, supporting sentences, unrelated sentences, subject name, unrelated subject name, relation type, object name, and unrelated object name respectively.

Drug name	FDA-approved indication	Closest ADRD drug	Correlation	R-score	P-value	First CT for ADRD	Adj PV^0
Levetiracetam	Seizures	Memantine	0.9928	0.7153	0.0002	09/12/01	0.0004
Clomipramine	Schizophrenia	Brexpiprazole	0.9932	0.6824	0.0007	NA	NA
Duloxetine	Depression	Brexpiprazole	0.9893	0.6691	0.0012	03/12/01	0.0017
Fluoxetine	Depression	Brexpiprazole	0.9953	0.6976	0.0004	NA	NA
Maprotiline	Depression	Brexpiprazole	0.9924	0.6565	0.0019	NA	NA
Armodafinil	Schizophrenia	Brexpiprazole	0.9944	0.6338	0.0040	NA	NA
Sertraline	Depression	Brexpiprazole	0.9921	0.6808	0.0008	NA	NA
Lisdexamfetamine	ADHD	Brexpiprazole	0.9922	0.6463	0.0027	NA	NA
Atomoxetine	ADHD	Brexpiprazole	0.9784	0.7051	0.0003	03/10/01	0.0033
Dextroamphetamine	ADHD	Brexpiprazole	0.9957	0.6965	0.0004	03/10/01	0.0002
Brexpiprazole	ADRD	Brexpiprazole	1.0000	0.6827	0.0007	13/07/11	NA
Caffeine	Apnea	Donepezil	0.9860	0.6535	0.0021	06/06/01	0.0072
Escitalopram	Depression	Brexpiprazole	0.9948	0.7067	0.0003	95/09/01	NA
Guanfacine	ADHD	Brexpiprazole	0.9917	0.7124	0.0002	19/01/04	0.0011
Hydralazine	Hypertension	Donepezil	0.9226	0.6134	0.0074	21/08/02	0.0010
Metformin	Hyperglycemia	Donepezil	0.9076	0.6196	0.0062	08/06/01	0.0057
OM3-CA	HTG	Galantamine	0.9457	0.5639	0.0260	NA	NA

Table S6. Compared candidate drugs' R-scores for ADRD. Candidate drugs proposed by a previous study are highlighted in bold. *: date format: YY/MM/DD. Correlation: the Pearson correlation coefficient determined by comparing the drug and its closest ADRD drug, as per the R-score profile. ADHD: attention-deficit/hyperactivity disorder; OM3-CA: omega-3-carboxylic acids; HTG: hypertriglyceridemia.

Original relation	Normalized relation
ADMINISTERED_TO	TREATS
AFFECTS	ASSOCIATED_WITH
AUGMENTS	STIMULATES
DISRUPTS	INHIBITS
PREVENTS	INHIBITS
METHOD_OF	ISA
USES	ISA
LOCATION_OF	PART_OF
PREDISPOSES	CAUSES

Table S7. Relation normalization.

References

1. Tung, N. T., Derreumaux, P., Vu, V. V., Nam, P. C. & Ngo, S. T. C-terminal plays as the possible nucleation of the self-aggregation of the s-shape aβ11–42 tetramer in solution: Intensive md study. *ACS omega* **4**, 11066–11073 (2019).
2. Prabhakar, R. Computational insights into the development of novel therapeutic strategies for alzheimer's disease. *Futur. medicinal chemistry* **1**, 119–135 (2009).
3. McCarron, M. *et al.* Association between interleukin-1a polymorphism and cerebral amyloid angiopathy–related hemorrhage. *Stroke* **34**, e193–e195 (2003).
4. Himmelhoch, E., Latham, O. & McDONALD, C. Alzheimer's disease complicated by a terminal salmonella infection. *Med. J. Aust.* **1**, 701–703 (1947).
5. Wen, C. *et al.* The secretion from bone marrow mesenchymal stem cells pretreated with berberine rescues neurons with oxidative damage through activation of the keap1-nrf2-ho-1 signaling pathway. *Neurotox. Res.* **38**, 59–73 (2020).
6. Hasegawa, M., Crowther, R. A., Jakes, R. & Goedert, M. Alzheimer-like changes in microtubule-associated protein tau induced by sulfated glycosaminoglycans: inhibition of microtubule binding, stimulation of phosphorylation, and filament assembly depend on the degree of sulfation. *J. biological chemistry* **272**, 33118–33124 (1997).
7. Pirogovsky, E. *et al.* Temporal sequence learning in healthy aging and amnestic mild cognitive impairment. *Exp. aging research* **39**, 371–381 (2013).
8. Zhang, L. *et al.* Biochemical characterization of the γ-secretase activity that produces β-amyloid peptides. *Biochem.* **40**, 5049–5055 (2001).
9. Scott, C. W., Fieles, A., Sygowski, L. A. & Caputo, C. B. Aggregation of tau protein by aluminum. *Brain research* **628**, 77–84 (1993).
10. Lu, X.-Y. *et al.* Metformin ameliorates aβ pathology by insulin-degrading enzyme in a transgenic mouse model of alzheimer's disease. *Oxidative Medicine Cell. Longev.* **2020**, 2315106 (2020).