Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Literature-Based Knowledge Graph of Histone Deacetylase 6 in Alzheimer's Disease Using Large Language Models

Aneesh Mazumder
Harvard College
Anirudh Mazumder
Stanford University
Prasenjit Mondal
Massachusetts General Hospital, Harvard Medical School
Licong Cui
The University of Texas Health Science Center at Houston
Claire Suen
Stanford University
Rudolph Tanzi
Massachusetts General Hospital, Harvard Medical School
Can Zhang
Massachusetts General Hospital, Harvard Medical School
Haoqi Sun
hsun3@bidmc.harvard.edu

Beth Israel Deaconess Medical Center

Article

Keywords: Histone Deacetylase 6, Alzheimer's Disease, Large Language Model, Knowledge Graph

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11453216/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Alzheimer's disease (AD) is a leading cause of dementia worldwide. Current therapies primarily target amyloid pathology and offer modest clinical benefit. Histone deacetylase 6 (HDAC6) has emerged as a potential therapeutic target in neurodegeneration through its roles in cytoskeletal dynamics, protein quality control, autophagy, and inflammatory signaling. However, the molecular landscape of HDAC6 in AD remains fragmented across a rapidly expanding literature. Our hypothesis is that a large language model (LLM)-enabled pipeline can systematically extract and organize molecular relationships from the HDAC6-AD literature into a biologically coherent knowledge graph (KG). We processed 106 full-text PubMed articles (2008 to 2025) with GPT-4.1-mini to extract molecular interactions and standardized entities using UniProt. The resulting KG contained 508 unique molecular entities and 511 edges from 760 extracted relationships. The graph was sparse and organized around HDAC6 as the principal hub and bridge node, with small isolated islands suggesting research gaps. We identified mechanistic modules for tau regulation, epigenetic signaling, and pharmacologic modulation through pathway analysis. This KG validates known relationships, highlights underexplored connections relevant to AD pathogenesis and therapy, and supports HDAC6 as a key regulator in AD. This framework may extend to other pathways and neurodegenerative disorders.

INTRODUCTION

Alzheimer's disease (AD) affects millions of individuals worldwide and represents one of the most pressing challenges in modern medicine [1]. Despite decades of research, therapeutic options remain limited, with recently approved monoclonal antibodies targeting early amyloid pathology providing modest slowing of disease progression and carrying side effects [2, 3]. There is a gap in identifying alternative therapeutic targets that address the broader molecular complexity of AD.

Histone deacetylase 6 (HDAC6), originally investigated in cancer biology and epigenetic regulation, has emerged as a promising candidate for therapeutic repurposing in neurodegeneration [4, 5]. Unlike many other histone deacetylases, HDAC6 is primarily cytoplasmic and regulates diverse cellular processes, including microtubule stability, protein quality control, autophagy, and stress responses [5]. These functions place HDAC6 at the intersection of several pathological mechanisms implicated in AD, including tau pathology, impaired axonal transport, and disrupted proteostasis [5]. Increasing experimental evidence suggests that pharmacologic inhibition of HDAC6 may restore neuronal function and improve cognitive outcomes in preclinical AD models [5, 6, 7]. Despite growing interest in HDAC6 as a therapeutic target, the exact regulatory role of HDAC6 in the nervous system remains incompletely understood, and systematic mapping of its interaction network has only recently been undertaken [8, 9]. HDAC6 interacts with numerous proteins and signaling pathways, and its effects appear to vary with experimental context, cell type, and disease stage [5]. As a result, the current knowledge of HDAC6 biology in AD is scattered across publications and difficult to synthesize using traditional manual methods. A systematic and scalable approach is needed to summarize the knowledge into a structured format.

Here, we developed a large language model (LLM)-enabled knowledge graph (KG) by preprocessing, extracting, standardizing, and integrating molecular relationships from the AD and HDAC6 literature into a structured knowledge graph (KG) (Fig. 1). We quantified graph metrics and identified hubs and modules in the resulting KG, providing a literature-driven systematic understanding of HDAC6's roles in AD and the potential gaps.

METHODS

Literature Acquisition and Corpus Development

A comprehensive literature corpus was assembled using the Biopython Entrez module to query PubMed. The search query was designed to capture publications at the intersection of Alzheimer's disease and HDAC6 research: ((Alzheimer's disease[Title/Abstract] OR Alzheimer's[Title/Abstract]) AND (HDAC6[Title/Abstract] OR histone deacetylase 6[Title/Abstract])). The query date was July 21, 2025. The query was further refined to include only English-language publications with full text. We did not restrict the paper types or species. For each retrieved article, metadata including title, authors, publication date, and PubMed ID were extracted. We manually downloaded all full texts in PDF format. Papers that were not open access were excluded from the analysis.

Text Extraction and Preprocessing

Full-text PDFs were converted to machine-readable text using the Mistral optical character recognition (OCR) API (model mistral-ocr-2505). For API checkpointing, the corpus was processed in five batches of 20, 20, 20, 20, and 26 papers. The OCR process included base64-encoding images and returned structured Markdown output. To optimize downstream processing, all embedded images, figures, and diagrams were removed, retaining only the textual content. The raw OCR output underwent a multi-stage cleaning pipeline using GPT-4.1-mini (see section "LLM Performance Comparison" below) to ensure high-quality input for triplet extraction (see Pre-processing prompts in the Supplementary Methods).

The first cleaning stage removed OCR-specific artifacts and non-sentence elements, including figure and table captions, section headers and subheadings, in-text citations (parenthetical and bracketed references), superscripts and subscripts, footnotes and page numbers, and equation numbers. The prompt instructed the model to preserve only complete, readable sentences from the body of the paper while retaining the article title for context.

The second preprocessing stage expanded all abbreviations to their full forms as defined within the paper. For example, "HDAC6" was consistently replaced with "histone deacetylase 6" throughout the text. This expansion was critical for ensuring that the downstream triplet extraction process captured complete entity names without requiring resolution of abbreviation-definition pairs.

The third preprocessing stage performed coreference resolution, which resolved anaphoric references and vague pronouns (e.g., "it", "they", "this", "these") to their corresponding antecedent noun phrases. This transformation made each sentence interpretable independently, removing the need for contextual information from preceding sentences during triplet extraction. For example, "It phosphorylates tau protein" would be transformed to "Histone deacetylase 6 phosphorylates tau protein." Each stage was applied paragraph by paragraph to maintain coherence.

Finally, the cleaned full text was segmented into individual sentences using the pre-trained Punkt sentence tokenizer in NLTK 3.9.1 (sent_tokenize with the English punkt_tab model), which employs unsupervised machine learning to detect sentence boundaries.

Triplet Extraction and Multi-Stage Refinement

For each sentence, GPT-4.1-mini was queried to extract all subject-predicate-object triplets representing distinct molecular relationships. Each triplet had: Subject (the entity initiating the interaction), Predicate (the relationship or action verb), and Object (the entity being acted upon). Each refinement step operated on individual triplets, using independent LLM API calls to ensure fine-grained quality control (see the extraction prompts in the Supplementary Methods). All GPT-4.1-mini calls used the gpt-4.1-mini-2025-04-14 snapshot through the OpenAI Responses API with default sampling parameters (temperature 1.0, top_p 1.0) and JSON-object output for the triplet stages.

After extraction, triplets underwent a multi-stage refinement to ensure accuracy, biological relevance, and compatibility with the KG. See refinement prompts in the Supplementary Methods. The 1st refinement stage filtered triplets to retain only those in which both the subject and object represented molecular entities (proteins, enzymes, small molecules, or complexes), excluding cellular processes, phenotypes, or abstract concepts that do not represent nodes in a molecular interaction network. The 2nd stage decomposed sentences containing multiple independent relationships into separate triplets. For example, "Protein A activates Protein B and inhibits Protein C" yielded two distinct triplets: (Protein A, activates, Protein B) and (Protein A, inhibits, Protein C). The 3rd stage further decomposed triplets with compound subjects or objects (e.g., "Proteins A and B") into individual binary relationships, ensuring that each triplet represented a single pairwise interaction suitable for graph representation. The 4th refinement stage mapped each predicate verb to standardized relationship categories: regulatory interactions (Activate, Amplify, Inhibit, Neutralize, Terminate), binding interactions (Bind, Form Complex, Dissociate), post-translational modifications (Phosphorylate, Dephosphorylate), structural changes (Cleave, Assemble, Reconfigure, Unfold, Aggregate), localization (Transport, Anchor, Recruit), signaling (Transmit, Trigger), and degradation (Degrade, Hydrolyze, Expel, Digest). If a predicate did not fit existing categories, the model was prompted to return "I don't know" along with a suggested new category for manual review. The 5th stage excluded triplets representing implicit or inferred relationships not explicitly stated in the original sentence, reducing LLM-induced hallucinations. The 6th refinement stage removed triplets derived from hypothetical statements, author speculations, or proposed mechanisms, retaining only experimentally validated or previously established relationships. The final 7th stage

excluded sentences containing uncertainty markers (e.g., "may", "might", "could"), questions, negations, or contradictory statements, ensuring that only confident, affirmative relationships were included in the KG.

Post-processing and Entity Standardization

For each refined triplet, GPT-4.1-mini was employed to identify and extract all proteins, enzymes, and small molecules mentioned in the subject and object fields (see post-processing prompts in the Supplementary Methods). The extraction prompt distinguished between proteins (including receptors, kinases, phosphatases, and transcription factors) and small molecules (including metabolites, drugs, and chemical compounds), providing specific guidelines for extracting core entity names while avoiding generic terms. Small molecule names were standardized at two levels. At the graph level, all entity names, including small molecules, went through the same Unicode normalization pipeline of NFKD normalization, lowercasing, stripping of non-alphanumeric characters, and whitespace collapsing, which collapsed variant spellings and punctuation differences to a single canonical form (e.g., "Trichostatin A" and "trichostatin A" map to the same node). At the annotation level, each extracted small molecule name was queried against PubChem and ChEMBL to retrieve structural identifiers. Each extracted protein was queried against the UniProt REST API to retrieve the standardized UniProt accession number and metadata. The search employed a two-strategy approach: an exact phrase search was performed first for maximum precision, followed by a fuzzy search if the exact search yielded no results. Results were preferentially filtered for Homo sapiens entries when appropriate, and Swiss-Prot (reviewed) entries were prioritized over TrEMBL (unreviewed) entries. For matched entries, the following metadata fields were extracted and stored: UniProt accession number, entry name, recommended and alternative protein names, gene names, organism, sequence length, and functional annotations. When multiple UniProt entries matched a protein name, GPT-4.1-mini was employed for context-aware disambiguation. The model was provided with the original triplet context (subject, predicate, object), all candidate UniProt entries with full metadata including truncated functional descriptions, and the complete sentence from which the triplet originated. The selection prompt instructed the model to evaluate candidates based on name similarity to the target entity, biological relevance to the triplet context, species relevance (preferring human when applicable), and functional relevance. This GPT-mediated disambiguation proved particularly valuable for proteins with multiple isoforms or homologs, where simple string matching would be insufficient.

Manual Quality Control and Knowledge Graph Finalization

Following automated post-processing and entity standardization, GPT-5-mini screened the 5,377 post-processed triplets and retained 1,593 (see the screening prompt in the Supplementary Methods). A human reviewer then manually reviewed all 1,593 screened triplets to ensure biological accuracy and KG quality. The reviewer evaluated each triplet for correct entity identification and appropriate predicate categorization. Triplets identified as containing errors in entity extraction, mapping, or relationship

interpretation were corrected or removed from the final dataset. This manual curation step ensured the final KG contained only high-confidence, biologically meaningful relationships suitable for downstream analyses and hypothesis generation. To better characterize performance across the full corpus, we also quantified the effect of human review on the complete set of extracted triplets. The original post-processed KG contained 5,377 triplets, of which 1,593 were retained after GPT-5-mini screening and 760 after manual curation. This corresponds to 4,617 removed triplets, or an 85.9% reduction, with 14.1% of extracted triplets retained in the final KG.

LLM Performance Comparison

To ensure optimal triplet extraction quality, the process was performed in parallel using three non-reasoning LLMs: GPT-4.1-mini (gpt-4.1-mini-2025-04-14), DeepSeek (deepseek-chat, DeepSeek-V3-0324), and Gemini 2.5-flash (gemini-2.5-flash with thinking disabled). Two independent human reviewers (AnM and ArM; see Author contributions) evaluated the outputs from all three models on a random subset of 10 papers. The reviewers assessed each model's performance across multiple criteria, including the accuracy of relationship identification, the completeness of entity extraction, consistency with the source text, and the overall biological plausibility of extracted triplets. The results are shown in Supplementary Table S2. GPT-4.1-mini was selected as the optimal model for full corpus processing, based on its superior performance in accurately capturing complex biological relationships while minimizing false-positive extractions. Therefore, all formal extraction and refinement steps were performed exclusively using GPT-4.1-mini; a separate GPT-5-mini pass screened the post-processed triplets before manual curation (see Manual Quality Control and Knowledge Graph Finalization). This is consistent with recent biomedical natural language processing (NLP) literature showing that LLMs can support knowledge graph construction and information extraction, but that performance varies across tasks and models and therefore benefits from benchmarking and human validation [10, 11, 12].

Graph Construction and Analysis

The knowledge graph was constructed, and all graph metrics were computed, in Python 3.13 with NetworkX 3.6.1; degree summaries were computed with NumPy. Community detection used the Louvain method as implemented in NetworkX (unweighted, resolution 1.0, modularity threshold 1e-7, random seed 42), run on the largest connected component. Network figures were drawn with Plotly 7.1.0 (https://plotly.com) using the Fruchterman–Reingold spring layout (seed 42).

Pathway Analysis

To characterize the biological functions represented within each community, we performed functional enrichment analysis on the three largest modules using the Database for Annotation, Visualization, and Integrated Discovery (DAVID). For each module, we curated UniProt accession numbers of its constituent protein nodes and submitted them as the input gene list, using DAVID's default background.

Enrichment was assessed across Gene Ontology (biological process, molecular function, and cellular component), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [13], and Reactome pathway [14] annotation categories, and p-values were corrected for multiple testing using the Benjamini-Hochberg procedure; enriched terms were reported at a false discovery rate (FDR) < 0.05. Analyses were performed using DAVID Knowledgebase v2026_1 on June 18, 2026 [15].

RESULTS

Corpus Characteristics

As shown in Supplementary Table S1, the corpus comprised 106 papers published between 2008 and 2025. Their journal impact factors ranged from 1.6 to 17.5, reflecting a broad distribution of publication venues from more specialized journals to higher-impact journals.

Knowledge Graph Characteristics

We extracted 760 molecular relationship triplets from the processed full-text corpus and assembled a literature-based KG after entity normalization. Node normalization reduced the pre-normalization total from 572 nodes to 508 nodes, indicating that 64 nodes were merged through canonical naming and synonym resolution. The final KG contained 508 unique nodes and 511 edges, as shown in Fig. 2.

We first characterized the overall graph sparsity to understand how sparse the connected entities were in the literature. The graph density was 0.004. The mean degree was 2.01, the median degree was 1.00, the minimum degree was 1, the maximum degree was 165, and the standard deviation was 7.58. For directed relationships, the mean in-degree was 1.04 (median 1, min 0, max 105, std 4.91) and the mean out-degree was 1.04 (median 1, min 0, max 73, std 3.54). The high standard deviation relative to the mean indicates a strongly right-skewed distribution where a small number of hub entities accumulate the vast majority of connections, while most entities appear in only one extracted relationship.

We then examined connectivity to determine whether the extracted relationships formed a single coherent network or multiple disconnected islands. The graph was not fully connected, with 58 disconnected islands. The largest connected island contained 367 nodes, while the remaining components were much smaller. We further quantified fragmentation by counting small components. There were 56 such small islands, indicating that many extracted relationships remained isolated from the main network because they lacked shared entities with other triplets in the corpus.

We then assessed local organization and global navigability to understand whether the graph contained densely interconnected neighborhoods and whether long or short chains of relationships separated entities in the main island. The average clustering coefficient was 0.03, indicating limited triangle structure and local redundancy among neighbors. The average shortest path length within the connected components of the graph was 3.5, indicating that entities in the core network were separated by only a few steps on average, despite the overall sparsity.

Knowledge Graph Hubs

We next identified hubs to determine which entities dominated the literature-derived molecular interaction landscape and which entities linked distinct regions of the graph. As shown in Fig. 3, HDAC6 was the dominant hub as expected, with a degree of 165 and a degree centrality of 0.33. HDAC6 also had the highest betweenness centrality, 0.46, indicating that it served as a principal connector on many shortest paths in the graph. In addition to HDAC6, Tau had a degree of 30 (Fig. 4). HDAC1 had a degree of 17. Amyloid β had a degree of 17. HDAC3 had a degree of 16. Pharmacologic entities also appeared among the most connected nodes, including CKD-504 (degree 14) and valproic acid (degree 14).

Knowledge Graph Modules

We next performed community detection to determine how the extracted relationships were organized into modules. We identified 27 modules, showing that the graph separates into multiple discrete groups of entities and interactions rather than collapsing into a single continuous structure. The top 3 largest modules are colored in Fig. 2.

The largest module contains 122 nodes and 121 edges and is organized almost entirely around HDAC6, which has a degree of 121 within the module and connects to nearly every other node. The density of 0.016 and the average degree of 2.0 indicate that this module is dominated by one-to-one links between HDAC6 and its partners, with relatively few connections among the surrounding nodes. That structure is informative because it shows that HDAC6 occupies a central coordinating role across multiple biological processes in this corpus. Its linked nodes include cytoskeletal proteins such as α-tubulin and β1-tubulin, which reflect HDAC6's known role in microtubule regulation and intracellular transport; signaling proteins such as G protein-coupled receptor kinase 2, which suggest broader effects on cellular response pathways; and autophagy-associated proteins such as SQSTM1 (also known as p62), which point to its role in protein quality control and cargo handling. The presence of inhibitors such as phenylhydroxamic acid and CKD 504 adds a translational dimension to the same module by showing that pharmacologic targeting is embedded directly within the core HDAC6 interaction landscape. Rather than representing a single narrow pathway, this module positions HDAC6 as a molecular hub linking structural maintenance, stress response, and therapeutic intervention.

The second largest module contains 37 nodes and 39 edges and focuses on Tau-related protein homeostasis. The dominant hub nodes are Tau (degree 19) and tau protein (degree 10), indicating that this module is organized around the regulation of Tau state and fate. Several of the surrounding nodes make this cluster especially interesting from a mechanistic perspective. The presence of Sirtuin 1, another deacetylation-related epigenetic regulator in the same neighborhood, suggests that Tau deacetylation is not controlled solely by HDAC6 but rather reflects a broader balance among deacetylase systems. Heat Shock Protein 70 and heat shock cognate 70 link the module to chaperone-based surveillance, indicating that Tau handling depends not only on its modification state but also on protein-folding and triage machinery. HECT, UBA and WWE domain-containing E3 ubiquitin protein ligase 1

(HUWE1) and ubiquitin-conjugating enzyme E2 O extend this logic further by linking Tau to ubiquitin-dependent processing, thereby connecting the acetylation state to downstream degradation decisions. The density of 0.059 is higher than that of the largest module, indicating that this cluster contains more internal mechanistic coupling and fewer isolated islands. This module covers several steps in Tau regulation: modification, recognition, trafficking, and degradation. In the context of AD, that organization highlights a specific region of the graph where HDAC6-related biology converges directly with pathogenic protein handling.

The third largest module contains 33 nodes and 34 edges and places HDAC6 within a broader epigenetic regulatory network. The leading hub nodes within the module are HDAC2 (degree 7), HDAC4 (degree 6), and suberoylanilide hydroxamic acid (SAHA; degree 5; Fig. 5 shows the full neighbourhoods of these three hubs in the whole graph), indicating that this module is structured around relationships among multiple HDAC family members and HDAC-targeted compounds rather than around HDAC6 alone. This is useful for interpretation because it shows that HDAC6-related mechanisms are situated within a broader deacetylase landscape that includes overlapping regulatory programs and competing therapeutic targets. Several nodes make that point more specific. Tubastatin A introduces selective HDAC6 inhibition into a module that also contains SAHA, allowing selective and pan-HDAC targeting within the same regulatory neighborhood. Yin Yang 1 (YY1) and lysine-specific demethylase 1 connect the cluster to transcriptional control and chromatin remodeling, broadening the module beyond deacetylase biology to epigenetic coordination more generally. ApoE3 and ApoE4 are particularly notable because they link this regulatory network to genotype-dependent AD biology, particularly pathways involving lipid metabolism, neuronal vulnerability, and transcriptional dysregulation. The density of 0.064 is the highest among the three largest modules, indicating the greatest degree of internal connectivity and reinforcing that this cluster captures a tightly related regulatory system. This module, therefore, provides critical context for interpreting HDAC6 by showing its position relative to other HDAC family members, epigenetic modifiers, and AD-relevant risk-associated pathways.

Pathway Analysis

To characterize the biological functions represented within each community, we performed functional pathway enrichment analysis on the protein nodes of the three largest modules, which yielded 253, 25, and 88 significantly enriched terms (FDR < 0.05) for Modules 1, 2, and 3, respectively (Supplementary Tables S3–S5). Enriched Gene Ontology terms surviving correction are shown in Supplementary Figs. S3–S5.

For Module 1 (HDAC6-centered), enrichment spanned localization, protein-degradation, cytoskeletal, and stress-response functions. The most significant cellular-component terms were cytoplasm (n = 43, fold enrichment [FE] 2.8) and cytosol (n = 41, FE 3.2), followed by more specific localizations including the centrosome (n = 14, FE 9.0) and the microtubule cytoskeleton (all FDR < 0.001). At the molecular-function level, the module was enriched for ubiquitin protein ligase binding (n = 15, FE 20.6) and protein-folding chaperone binding (n = 7, FE 26.7; both FDR < 0.001). Enriched biological-process and pathway terms

included microtubule cytoskeleton organization (n = 7, FE 18.2), aggresome assembly (n = 4, FE 277.9), cellular response to oxidative stress (n = 7, FE 25.4), Selective autophagy (n = 11, FE 32.7), Macroautophagy (n = 11, FE 20.4), and Aggrephagy (n = 8, FE 47.0; all FDR < 0.001). The KEGG term Pathways of neurodegeneration – multiple diseases (n = 14, FE 6.8) and Protein processing in endoplasmic reticulum (n = 8, FE 10.8) were also enriched (FDR < 0.001). These terms correspond to HDAC6's established functions in microtubule regulation, aggresome and autophagic clearance of misfolded proteins, and the oxidative stress response, consistent with the cytoskeletal, autophagy-associated, and quality-control nodes that define this module.

For Module 2 (Tau-centered), the enriched terms were fewer, reflecting the module's smaller size, and were centered on chaperone activity and protein stability control. The strongest signals were the cellular-component terms cytosol (n = 16, FE 3.3) and nucleoplasm (n = 14, FE 3.8), the molecular-function term enzyme binding (n = 7, FE 19.1), and the biological-process term regulation of cellular response to heat (n = 4, FE 300.7; all FDR < 0.001). Additional significant terms included Hsp90 protein binding (n = 3, FE 74.1), regulation of protein stability (n = 4, FE 28.6), protein ubiquitination (n = 6, FE 12.1), the axon compartment (n = 5, FE 13.5), and the Reactome pathways Cellular response to heat stress (n = 5, FE 39.4) and Regulation of HSF1-mediated heat shock response (n = 4, FE 40.3; FDR < 0.05). This heat-shock and chaperone-related profile matches the Heat Shock Protein 70, heat shock cognate 70, and ubiquitin-processing nodes present in the module, supporting its interpretation as a cluster organized around chaperone-based surveillance and degradation of Tau. Because several of these terms rest on small gene counts (n = 3–6), the associated fold-enrichment values should be read as directional rather than precise.

For Module 3 (HDAC2- and HDAC4-centered), enrichment was dominated by deacetylase catalytic activity and transcriptional regulation. The most significant molecular-function terms were protein lysine deacetylase activity (n = 7, FE 420.0), histone deacetylase activity (n = 7, FE 326.6), histone deacetylase binding (n = 8, FE 49.4), and DNA-binding transcription factor binding (n = 9, FE 41.8), accompanied by the cellular-component term histone deacetylase complex (n = 7, FE 143.7; all FDR < 0.001). At the biological-process level, the module was enriched for epigenetic regulation of gene expression (n = 5, FE 61.3) and negative regulation of transcription by RNA polymerase II (n = 12, FE 9.2; FDR < 0.001), with the chromatin compartment also enriched (n = 9, FE 6.6, FDR < 0.01). The Reactome Notch-HLH transcription pathway (n = 7, FE 119.2, FDR < 0.001) was among the enriched pathway terms. Together, these deacetylase-, chromatin-, and transcription-associated terms confirm the module's identity as a broader HDAC-family and epigenetic regulatory cluster, distinct from the HDAC6-specific interactions concentrated in Module 1.

Subgraphs

Finally, we generated three subgraph views to improve interpretability and test whether key structural patterns persisted across different evidence-focused views. First, a high-impact-factor filtered subgraph that included only relationships drawn from journals with an impact factor of ≥ 10 contained 108 nodes and 97 edges (Supplementary Fig. S1). This subgraph preserved a compact core centered on HDAC6

while reducing peripheral complexity. Second, a hub-focused subgraph that included HDAC6 and other high-degree nodes contained 254 nodes and 304 edges (Supplementary Fig. S2). This view focused on the graph's densest portion and highlighted how a small set of hubs organized much of the reachable interaction landscape. Third, we generated a movie (Supplementary Video S1) showing the graph's evolution as a function of publication year, showing the growth of both molecules connected to HDAC6 and those peripheral to HDAC6.

DISCUSSION

This study transformed the HDAC6-AD literature into a KG of molecular interactions using an LLM, providing a structured framework for understanding HDAC6 in the context of AD and systematically discovering underexplored molecular connections relevant to AD pathogenesis and therapeutic development.

Biological Insights Confirming Existing Findings

HDAC6 biology in AD exhibits complex and context-dependent behavior across multiple molecular pathways [5, 16, 17]. Previous work has shown that HDAC6 can participate in protective chaperone complexes that target aberrantly modified tau proteins for degradation, suggesting a role in maintaining protein quality control [18]. In contrast, other studies demonstrate that pharmacologic inhibition of HDAC6 improves cognitive outcomes and reduces pathological tau accumulation in experimental AD models [5, 6, 7, 19]. These apparently contradictory observations indicate that HDAC6 may play both protective and detrimental roles depending on disease stage and cellular environment.

The broad range of substrates of HDAC6 further contributes to this complexity. In addition to tau and tubulin, HDAC6 interacts with numerous proteins involved in cytoskeletal organization, stress responses, and inflammatory signaling [5, 20, 21]. These interactions link HDAC6 to multiple biological processes, including autophagy, oxidative stress regulation, and intracellular transport [5, 22, 23]. The modular structure of the KG reflects this biological diversity. Distinct communities corresponding to cytoskeletal regulation, epigenetic signaling, and pharmacologic modulation emerged from the graph analysis. These modules align with major mechanistic themes described in the HDAC6 literature and support the biological face validity of the extracted relationships.

Biological Insights that Require More Studies: Potential New Knowledge

First, several proteins identified in recent studies appeared in peripheral regions of the graph with relatively few connections. For example, a recent case report described a patient with a C9orf72 expansion who presented with an AD-like clinical, biomarker, and imaging profile, including progressive memory and language deficits, amyloid positivity, and AD-like temporoparietal hypometabolism [24]. Because this evidence comes from a single case, it is better understood as an isolated observation than as a broadly established pattern. The peripheral placement of proteins or pathways linked to

observations such as these likely reflects their only weak integration into the broader HDAC6 literature and insufficient evidence to be centrally embedded within the graph.

Second, the graph revealed connections between HDAC6 and pathways that have been studied in other neurodegenerative and systemic diseases. For example, interactions involving α-synuclein suggest potential overlap between HDAC6-mediated mechanisms in AD and those implicated in Parkinson's disease [25]. Similarly, several nodes associated with cancer signaling pathways appeared in the graph, reflecting the historical focus of HDAC6 research in oncology. These observations support the possibility of drug repurposing strategies, where compounds originally developed for cancer or other neurodegenerative disorders could be investigated for AD therapy.

Third, the 56 small islands in the KG indicate that many molecular relationships remained isolated, suggesting the need for further research. These 56 disconnected components comprise 41 isolated pairs, 11 triplets, 3 groups of four nodes, and 1 group of five nodes. Their biological content reveals several recurring patterns. Several pairs represent molecular variants or post-translational states of entities that appear in the main network under different canonical names—for example, hyper-phosphorylated tau paired with a plant-derived inhibitor, and monomeric amyloid beta 1–42 paired with phosphorylated tau at threonine 181, both suggesting niche interaction contexts not yet linked back to the central network. Other isolated pairs capture single-paper observations from pathways adjacent to but not yet wired into the main HDAC6 interaction landscape, including cytoskeletal pairs such as alpha-tubulin acetyltransferase 1 with vimentin, receptor pairs such as C-X-C chemokine receptor type 2 with its cognate ligand, and autophagy-related pairs such as active lysosome-associated membrane protein 2A (LAMP2A) with lysosomal heat shock cognate 70 (HSC70). Among the larger isolated components, the triplet containing FUS and TARDBP alongside RNA processing machinery reflects disease biology from amyotrophic lateral sclerosis (ALS)-associated papers whose entities do not overlap with HDAC6 network entities, and the triplet containing nuclear respiratory factor 1, nuclear respiratory factor 2, and peroxisome proliferator-activated receptor gamma coactivator 1-alpha (PGC-1 alpha) captures a mitochondrial biogenesis cluster that appears in a single study context. The presence of these islands reflects both the precision of the triplet extraction pipeline, which does not fabricate connections, and the inherent sparsity of the current 106-paper corpus, where many interactions are reported only once without overlapping entities to bridge them into the main component.

Methodological Implications

The primary methodological contribution of this work is the demonstration that LLMs can support scalable construction of literature-derived molecular KGs [10, 12]. The extracted KG was sufficiently large to support systems-level interpretation yet remained sparse. The low graph density and low median degree indicated that most molecular entities were supported by only a few explicit relationships in the processed corpus. This pattern reflects a common structure in the biomedical literature, in which a small set of well-studied entities dominates the interaction landscape. Connectivity analysis revealed both a dominant connected component and many smaller isolated components. These small islands likely reflect two factors: the inherent modularity of biological systems and incomplete coverage of bridging

relationships within the current corpus. Expanding the corpus and relaxing certain filtering constraints may reduce this fragmentation in future iterations. Centrality analysis identified HDAC6 as both the primary hub and the primary bridge node in the KG. Tau and amyloid β also emerged as highly connected nodes, reinforcing the biological relevance of the extracted relationships. Additional hubs, such as HDAC1, HDAC3, CKD-504, and valproic acid, further demonstrate that both molecular regulators and pharmacologic agents play central roles in the literature-derived interaction landscape. Clustering and path length analyses provided additional insight into the organization of the KG. The low clustering coefficient indicated limited redundancy among neighboring nodes, while the relatively short path length within the main component suggested that major mechanistic themes remain interconnected through a small number of bridging entities.

Implications for Drug Discovery

The HDAC6-AD KG provides a systematic framework for organizing existing molecular evidence and identifying candidate interactions and pathways for subsequent drug-discovery studies. By integrating relationships that are otherwise dispersed across the literature, the graph can highlight well-supported mechanisms, underexplored connections, and disconnected regions that may warrant additional investigation. This role is consistent with recent recommendations that AI applications in drug discovery should be driven by specific biological and translational questions and should ultimately support improved scientific decision-making rather than model performance alone [26]. In this way, the KG can help define which HDAC6-associated mechanisms, interacting proteins, and pathways should be prioritized for further computational or experimental study.

Future development of the KG could extend this framework from literature organization toward predictive analysis. Incorporating contextual information such as cell type, species, disease stage, experimental system, and direction of molecular effects would allow individual relationships to be interpreted within the biological conditions in which they occur. Graph-based modeling could then be used to identify or simulate pathways through which perturbation of HDAC6 and connected molecular nodes may lead to reductions in tau- and amyloid-β-associated pathology. These predicted pathways could prioritize candidate mechanisms and molecular targets for experimental validation and, subsequently, guide the identification or development of therapeutic compounds. Thus, the current KG establishes a foundation for a more systematic progression from literature-derived molecular knowledge to computational hypothesis generation and ultimately to HDAC6-directed drug discovery.

Limitations and Future Directions

The graph lacks contextual information such as species, cell type, experiment, disease stage, or perturbation direction. Incorporating these attributes would allow context-dependent relationships to be represented more accurately. For example, HDAC6-AD KGs in mice and humans may share important similarities and differences that suggest drug transferability. Another limitation is the imperfect standardization of molecule names, so that entities referring to the same biological target appear in different forms and are represented as separate nodes. Future iterations should improve molecule name

standardization, reduce redundancy, and enable more robust linkage to external databases. Additionally, although each retained relationship was manually curated for biological plausibility by human review, the extracted edges were not systematically cross-referenced against curated interaction databases such as STRING or BioGRID; such benchmarking is complicated by the literature-specific and potentially novel nature of the relationships captured here, and programmatic cross-referencing represents a clear direction for future work [27, 28]. There are also potential improvements to the KG, such as edge weighting based on literature frequency, which could help distinguish widely supported interactions from single-report findings. Lastly and importantly, we only used papers with full text available, which introduces a selection bias that warrants further investigation.

A particularly promising future direction is link prediction and structural inference within the KG. Because the current KG already contains a connected core with short path lengths, graph-based algorithms (such as graph neural networks) can propose statistically plausible but currently unknown interactions. These predicted links could guide targeted literature searches and experimental validation. Integration with AlphaFold structural predictions and experimental validation studies may further strengthen this approach [29]. Over time, combining automated literature mining, computational KG analysis, and experimental feedback could create a closed-loop system to continuously refine molecular knowledge graphs for complex diseases such as AD.

Declarations

Competing interests

The authors declare no competing interests.

Funding

Aneesh Mazumder discloses support for the research of this work from the Yun Family Research Fund and the Herchel Smith Undergraduate Science Research Program at Harvard University.

Author Contribution

AnM, Aneesh Mazumder; ArM, Anirudh Mazumder.Conceptualization: HS, CZ;Data curation: AnM, ArM;Formal analysis: AnM, ArM;Funding acquisition: AnM, CZ, HS;Writing – original draft: AnM, ArM, LC, HS;Writing – review & editing: All authors.

Acknowledgement

Aneesh Mazumder would like to thank the Office of Undergraduate Research and Fellowships (URAF) at Harvard University for its support of this research through the Yun Family Research Fund and the Herchel Smith Undergraduate Science Research Program.

Data Availability

The code and data generated in this study are available in the GitHub repository at https://github.com/anirudhmazumder/AI-Park.

References

1. World Health Organization. Dementia. (2025). https://www.who.int/news-room/fact-sheets/detail/dementia
2. van Dyck, C. H. et al. Lecanemab in early Alzheimer's disease. *N Engl. J. Med.* **388**, 9–21 (2023).
3. Sims, J. R. et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA* **330**, 512–527 (2023).
4. Qureshi, T. & Chinnathambi, S. Histone deacetylase-6 modulates Tau function in Alzheimer's disease. *Biochim. Biophys. Acta Mol. Cell. Res.* **1869**, 119275 (2022).
5. Mondal, P. et al. Structure-based discovery of a small molecule inhibitor of histone deacetylase 6 (HDAC6) that significantly reduces Alzheimer's disease neuropathology. *Adv. Sci.* **11**, 2304545 (2024).
6. Govindarajan, N. et al. Reducing HDAC6 ameliorates cognitive deficits in a mouse model for Alzheimer's disease. *EMBO Mol. Med.* **5**, 52–63 (2013).
7. Onishi, T. et al. A novel orally active HDAC6 inhibitor T-518 shows a therapeutic potential for Alzheimer's disease and tauopathy in mice. *Sci. Rep.* **11**, 15423 (2021).
8. Bahram Sangani, N. et al. A novel insight into neurological disorders through HDAC6 protein–protein interactions. *Sci. Rep.* **14**, 14666 (2024).
9. Zhang, W., Zhang, Y., Wang, Y., Wang, C. & Zhang, C. Advancing histone deacetylase 6 (HDAC6) as a promising therapeutic target for Alzheimer's disease: from molecular insights to clinical prospects. *Front. Drug Discov*. **5**, 1662691 (2025).
10. Arsenyan, V., Bughdaryan, S., Shaya, F., Small, K. W. & Shahnazaryan, D. Large language models for biomedical knowledge graph construction: information extraction from EMR notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*295–317Association for Computational Linguistics, (2024).
11. Jahan, I., Laskar, M. T. R., Peng, C. & Huang, J. X. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Comput. Biol. Med.* **171**, 108189 (2024).
12. Chen, Q. et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat. Commun.* **16**, 3280 (2025).

13. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677. https://doi.org/10.1093/nar/gkae909 (2025).
14. Ragueneau, E. et al. The Reactome Knowledgebase 2026. *Nucleic Acids Res.* **54**, D673–D681. https://doi.org/10.1093/nar/gkaf1223 (2026).
15. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221. https://doi.org/10.1093/nar/gkac194 (2022).
16. Kim, J. Y. et al. HDAC6 inhibitors rescued the defective axonal mitochondrial movement in motor neurons derived from the induced pluripotent stem cells of peripheral neuropathy patients with HSPB1 mutation. *Stem Cells Int.* 9475981 (2016). (2016). https://doi.org/10.1155/2016/9475981
17. Dompierre, J. P. et al. Histone deacetylase 6 inhibition compensates for the transport deficit in Huntington's disease by increasing tubulin acetylation. *J. Neurosci.* **27**, 3571–3583. https://doi.org/10.1523/JNEUROSCI.0037-07.2007 (2007).
18. Trzeciakiewicz, H. et al. An HDAC6-dependent surveillance mechanism suppresses tau-mediated neurodegeneration and cognitive decline. *Nat. Commun.* **11**, 5522 (2020).
19. Liang, T. et al. Multimechanism biological profiling of tetrahydro-β-carboline analogues as selective HDAC6 inhibitors for the treatment of Alzheimer's disease. *Eur. J. Med. Chem.* **275**, 116624. https://doi.org/10.1016/j.ejmech.2024.116624 (2024).
20. Wang, L. et al. HDAC6/aggresome processing pathway importance for inflammasome formation is context-dependent. *J. Biol. Chem.* **300**, 105638. https://doi.org/10.1016/j.jbc.2024.105638 (2024).
21. Jagaselvan, E., Mondal, P. & Zhang, C. Modulating stress granules via HDAC6 inhibition: a promising strategy against Alzheimer's disease. *ACS Chem. Neurosci.* **17**, 2771–2774. https://doi.org/10.1021/acschemneuro.6c00264 (2026).
22. Ganai, S. A. Small-molecule modulation of HDAC6 activity: the propitious therapeutic strategy to vanquish neurodegenerative disorders. *Curr. Med. Chem.* **24**, 4104–4120. https://doi.org/10.2174/0929867324666170209104030 (2017).
23. Li, Y., Lin, S., Gu, Z., Chen, L. & He, B. Zinc-dependent deacetylases (HDACs) as potential targets for treating Alzheimer's disease. *Bioorg. Med. Chem. Lett.* **76**, 129015. https://doi.org/10.1016/j.bmcl.2022.129015 (2022).
24. Garrou, F. et al. Challenging the boundaries: c9orf72 mutation presenting as Alzheimer’s disease. *Amyotroph. Lateral Scler. Frontotemporal Degener*. **27**, 215–218 (2026).
25. Crowe, A. et al. HDAC6 inhibition reduces seeded tau and α-synuclein pathologies in primary neuron cultures and wild-type mice. *J. Neurosci.* **46**, e1092252025 (2026).
26. Bender, A. et al. Artificial intelligence in drug discovery — what it is, where we stand and the path forward. *Nat. Rev. Drug Discov*. https://doi.org/10.1038/s41573-026-01496-2 (2026).
27. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–

D646. https://doi.org/10.1093/nar/gkac1000 (2023).

28. Oughtred, R. et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200. https://doi.org/10.1002/pro.3978 (2021).
29. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Figures

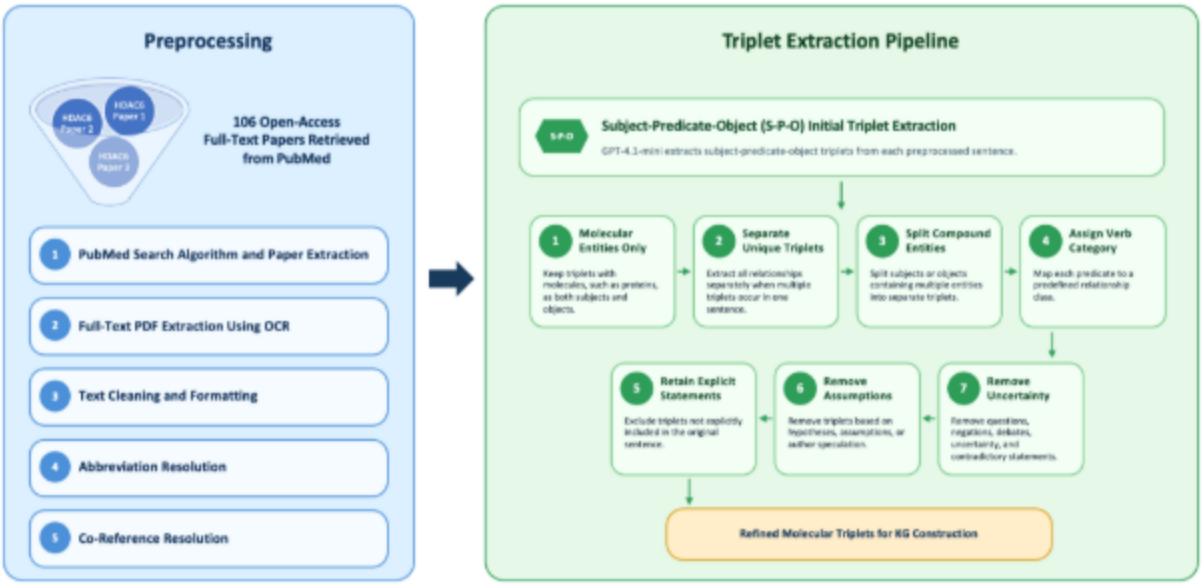


Figure 1

Overview of the LLM-based extraction method, including initial triplet extraction followed by seven refinement steps.

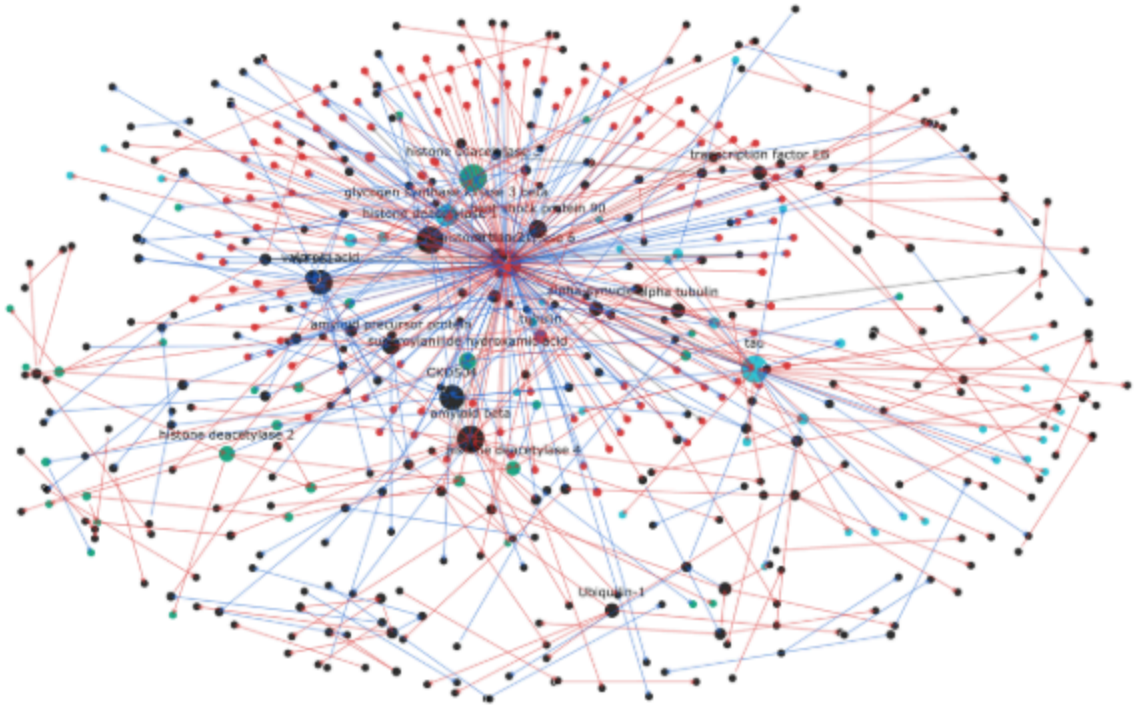


Figure 2

Whole-graph visualization of the literature-derived molecular interaction network. Nodes represent normalized molecular entities and are colored by community module: red nodes indicate Module 1, which is HDAC6-centered; cyan nodes indicate Module 2, which is Tau-centered; green nodes indicate Module 3, which is HDAC2- and HDAC4-centered; and black nodes indicate other or isolated entities. Node size scales with degree. Red edges denote activating or binding relationships, blue edges denote inhibiting or degrading relationships, and black edges denote triplets for which the large language model could not confidently classify the predicate.

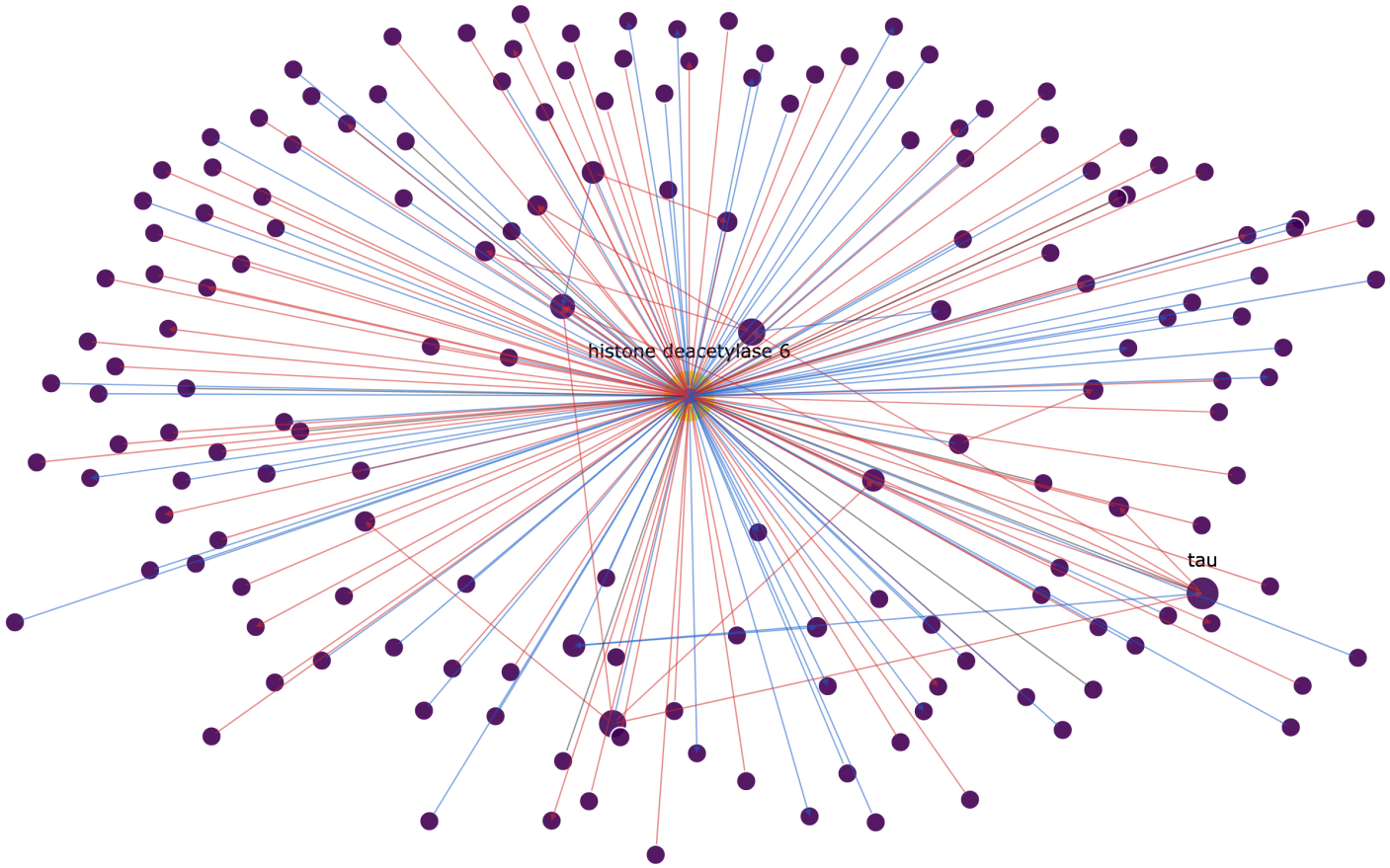


Figure 3

HDAC6-centered subgraph of the final literature-derived molecular interaction network. The subgraph shows all nodes directly connected to histone deacetylase 6. Nodes represent normalized molecular entities, edges represent extracted molecular relationships, and node size scales with degree.

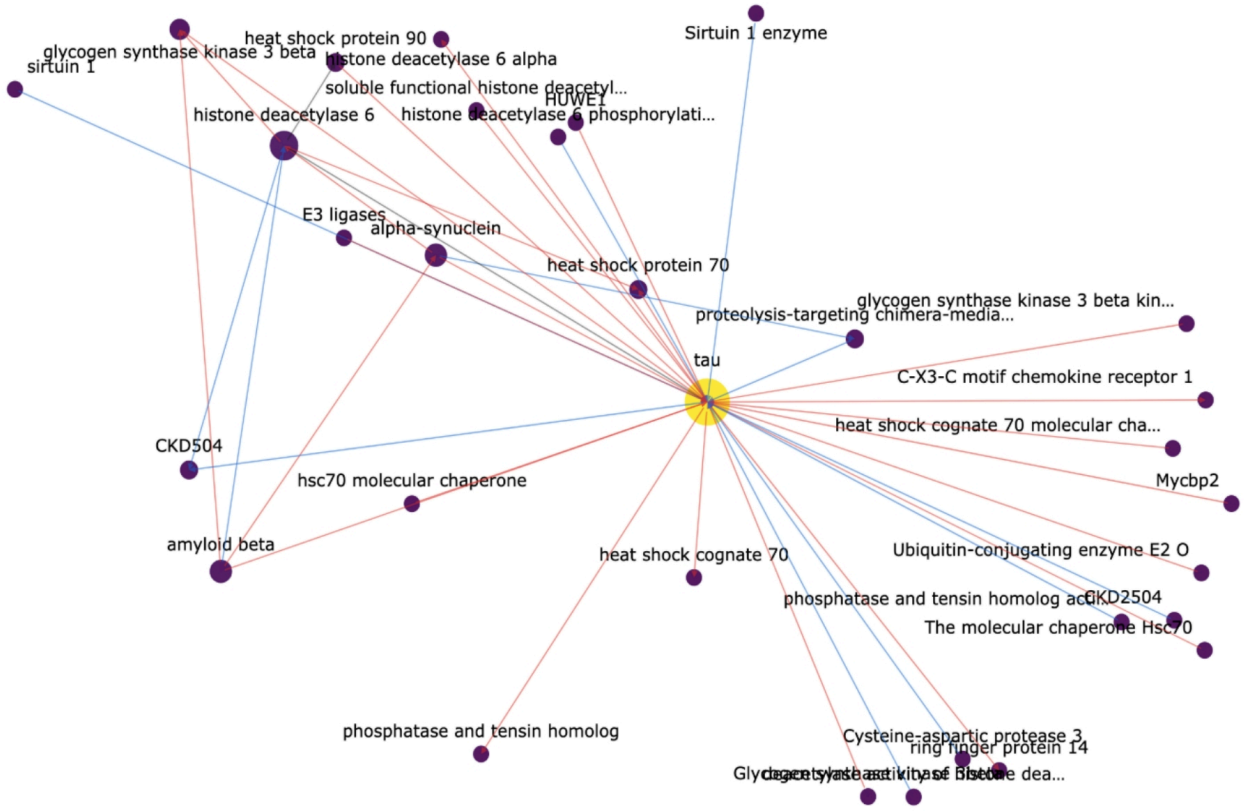


Figure 4

Tau-centered subgraph of the literature-derived molecular interaction network. The subgraph shows all nodes directly connected to Tau. Nodes represent normalized molecular entities, edges represent extracted pairwise relationships, and node size scales with degree.

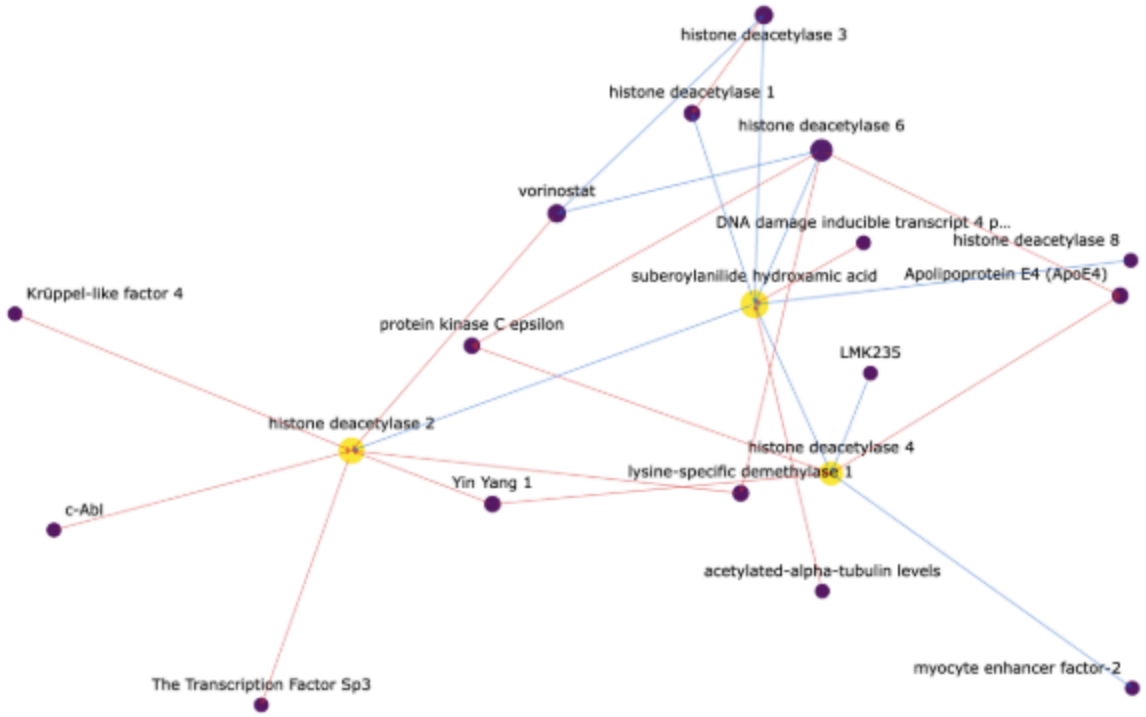


Figure 5

HDAC2, HDAC4, and SAHA-centered subgraph of the literature-derived molecular interaction network. The subgraph shows all nodes directly connected to HDAC2, HDAC4, and SAHA. Nodes represent normalized molecular entities, edges represent extracted pairwise relationships, and node size scales with degree.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HDAC6ADSupplementaryInformationFinal.docx