Identifying Neglected Hypotheses in Neurodegenerative Disease with Large Language Models

Spencer Phillips Hey
Prism Analytic Technologies
Cambridge, MA 02142
spencer@prism.bio

Darren Angle
Prism Analytic Technologies
Cambridge, MA 02142
darren@prism.bio

Chris Chatham
Genentech
South San Francisco, CA 94080
chatham.christopher@gene.com

Abstract

Neurodegenerative diseases remain a medical challenge, with existing treatments for many such diseases yielding limited benefits. Yet, research into diseases like Alzheimer's often focuses on a narrow set of hypotheses, potentially overlooking promising research avenues. We devised a workflow to curate scientific publications, extract central hypotheses using gpt3.5-turbo, convert these hypotheses into high-dimensional vectors, and cluster them hierarchically. Employing a secondary agglomerative clustering on the "noise" subset, followed by GPT-4 analysis, we identified signals of neglected hypotheses. This methodology unveiled several notable neglected hypotheses including treatment with coenzyme Q10, CPAP treatment to slow cognitive decline, and lithium treatment in Alzheimer's. We believe this methodology offers a novel and scalable approach to identifying overlooked hypotheses and broadening the neurodegenerative disease research landscape.

1 Introduction

Neurodegenerative diseases remain an area of high unmet medical need. For example, despite recent regulatory approvals for therapies targeting amyloid-β (Aβ), treatments for Alzheimer's Disease (AD) have still shown only modest clinical benefits [1]. One explanation for the lack of significant, disease-modifying therapeutics in AD is the scientific community's decades-long focus on a limited subset of "dominant" hypotheses. For example, it has been widely argued that the so-called "Amyloid Hypothesis" accumulated a disproportionate share of the research funding and activities in AD [2]. The small set of (perceived) viable hypotheses, combined with the high cost and the high failure rate of development programs in AD, has led to a reduced investment from pharmaceutical manufacturers in the neurodegenerative disease space over the past 20 years [3].

However, the existence of dominant scientific hypotheses, which crowd out research into other ideas, suggests that there is a complementary set of "neglected hypotheses," which are underfunded and underexplored. From the perspective of a drug developer, systematically identifying neglected hypotheses is potentially of high value: It may help to identify promising new ideas for research

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

and development; it can aid company strategy by providing a more comprehensive understanding of the total scientific landscape; it may uncover evidence that conflicts with the dominant hypotheses, suggestive of revisions or refinements needed to the reigning scientific paradigms.

In what follows, we present a proof-of-concept workflow—leveraging large language models (LLMs) to extract, generate, and cluster scientific hypothesis from a body of scientific literature—intended to systematically identify such neglected hypotheses in the space of neurodegenerative diseases.

2 Methods

Recognizing that valuable, neglected hypotheses may be latent in scientific literature that cuts across domains [4], we chose to explore a broad body of literature for our proof-of-concept. To wit: We conducted a keyword search of the National Library of Medicine's PubMed database for "neurodegenerative diseases," retrieving all records published between 2002 and 2022, filtering by "clinical trial" as an article type. For each record, we used a Python script and PubMed's API (through Biopython Entrez) to extract additional metadata on keywords and substance/compound terms that were not included in the original data download.

To extract the central hypothesis from each record abstract, we used OpenAI's gpt3.5turbo large language model (LLM). Since explicit hypothesis statements are not always present in abstracts, we wrote a prompt that provided the LLM with each record's title, abstract, keywords, and substance terms and instructed it to return 1- or 2-sentence hypothesis. In the prompt, we provided the LLM with a persona ("You are an AI that has been specifically trained to identify and report hypotheses for scientific papers. Accuracy in your response is paramount and may lead to saving lives."), context ("Your answer will be stored in a database, and used to compare semantic similarity between this paper's hypothesis and the hypothesis of other papers."), and restrictions ("You do not need to include the instructions, or include any social speech like 'sure' or summarize the task given to you.").

We then embedded each hypothesis into 1536-dimensional vectors using OpenAI's text-embedding-ada-002 model.[1] To reduce the dimensionality of the embeddings, we utilized the Uniform Manifold Approximation and Projection (UMAP) method with neighbors set to 15 [5]. We then applied the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm to segregate our data points into distinct clusters [6].

The points that HDBSCAN labeled as "noise" from this clustering were prima facie of interest as potentially neglected hypotheses. Thus, we performed a second pass on the noise using agglomerative clustering.[2] Then as the final step, to summarize the meaning of the "neglected" hypotheses resulting from the second clustering, we provided gpt4.0 with the list of hypotheses for each of the 25 largest "noise" clusters and instructed it to extract the "central hypothesis".

3 Results

Our PubMed search identified 8,749 records. Each record was then processed through gpt3.5turbo with the hypothesis generation prompt. This data pipeline failed and returned no hypothesis for sixty-five records (0.7%). The remaining 8,684 records had a central hypothesis statement generated and underwent semantic embedding. Table 1 shows a representative transformation of a the original PubMed article metadata and the output of the hypothesis extraction algorithm. The complete dataset of metadata and LLM-generated hypotheses is provided in the supplementary materials.

Distinct from the pipeline failures, in 14 instances the LLM correctly generated a "null result" hypothesis for articles that lacked sufficient information in the abstract or other metadata to generate a meaningful hypothesis. In 2 instances, the LLM correctly recognized that an article had been retracted—however, in only 1 of these 2 did it successfully capture the retraction in the content of the "hypothesis" (in the other, it returned "N/A"). These successful classifications of "low quality data" are encouraging signs for the automation and scalability of this approach.

[1] We chose to embed only the extracted, central hypothesis, rather than the complete abstract, because we were not interested in clustering results by the methods, results, or conclusions. Our aim was to try and uncover neglected *scientific hypotheses*, not neglected papers more generally.

[2] For agglomerative clustering, we brute force tested the cluster size parameter from 15–1000 items per cluster to find the optimal number by silhouette score. This lead to 505 clusters being optimal.

Table 1: Example Hypothesis Extraction

PMID	34193504
Title	Repurposing bromocriptine for Aβ metabolism in Alzheimer's disease (REBRAnD) study: randomised placebo-controlled double-blind comparative trial and open-label extension trial to investigate the safety and efficacy of bromocriptine in Alzheimer's disease with presenilin 1 (PSEN1) mutations.
Abstract	Alzheimer's disease (AD) is one of the most common causes of dementia. Pathogenic variants in the presenilin 1 (PSEN1) gene are the most frequent cause of early-onset AD. Medications for patients with AD bearing PSEN1 mutation (PSEN1-AD) are limited to symptomatic therapies and no established radical treatments are available. Induced pluripotent stem cell (iPSC)-based drug repurposing identified bromocriptine as a therapeutic candidate for PSEN1-AD. In this study, we used an enrichment strategy with iPSCs to select the study population, and we will investigate the safety and efficacy of an orally administered dose of bromocriptine in patients with PSEN1-AD. This is a multicentre, randomised, placebo-controlled trial. AD patients with PSEN1 mutations and a Mini Mental State Examination-Japanese score of $\leq$25 will be randomly assigned, at a 2:1 ratio, to the trial drug or placebo group ($\geq$4 patients in TW-012R and $\geq$2 patients in placebo). This clinical trial consists of a screening period, double-blind phase (9 months) and extension phase (3 months). The double-blind phase for evaluating the efficacy and safety is composed of the low-dose maintenance period (10 mg/day), high-dose maintenance period (22.5 mg/day) and tapering period of the trial drug. Additionally, there is an open-labelled active drug extension period for evaluating long-term safety. Primary outcomes are safety and efficacy in cognitive and psychological function. Also, exploratory investigations for the efficacy of bromocriptine by neurological scores and biomarkers will be conducted.
Substance Terms	bromocriptine, psen1 protein human, presenilin1
Keyword Terms	dementia, neurogenetics, clinical trials, neurology, neurobiology
Hypothesis	The hypothesis of this paper is that orally administered bromocriptine will be safe and effective in improving cognitive and psychological function in patients with Alzheimer's disease bearing PSEN1 mutations. Additionally, exploratory investigations will be conducted to evaluate the efficacy of bromocriptine by neurological scores and biomarkers.

Figure 1 shows the 3-D UMAP representation of the hypothesis embeddings. Exploration of this visualization supports the face validity of this method, since it is readily able to identify clusters related to Aβ, donepezil, deep brain stimulation, psycho-social interventions, and several other well-known (i.e., "dominant") hypotheses in the domain of neurodegenerative diseases.

The results of re-analyzing, re-clustering, and then summarizing what was classified as "noise" revealed several interesting hypotheses. We provide the complete list of 25 neglected hypothesis in the supplementary materials, but will enumerate three of what (in our view) are the most interesting of these here:

1. Coenzyme Q10 plays a therapeutic role in neurodegenerative diseases like Parkinson's, Alzheimer's, and Huntington's by modulating oxidative damage, improving cerebral energy metabolism, and potentially serving as a biomarker of antioxidant status.
2. Continuous positive airway pressure (CPAP) treatment for obstructive sleep apnea (OSA) may lead to improved cognitive function and slowed cognitive decline in patients with neurodegenerative diseases such as Alzheimer's and Parkinson's, potentially linked to changes in Aβ burden and sleep patterns in specific brain regions.
3. Lithium treatment in Alzheimer's disease patients may exert neuroprotective effects, potentially moderating cognitive decline and influencing Alzheimer's-related biomarkers such as tau phosphorylation, GDNF, and BDNF levels.

All three of these hypotheses have some measure of scientific validity (i.e., a plausible mechanistic rationale for the relationship between intervention and disease modification). Supplementary literature searches on these hypotheses as topics (e.g., searching "CPAP AND Alzheimer's disease" in PubMed) also confirms that these are (as yet) little explored, but potentially burgeoning areas.

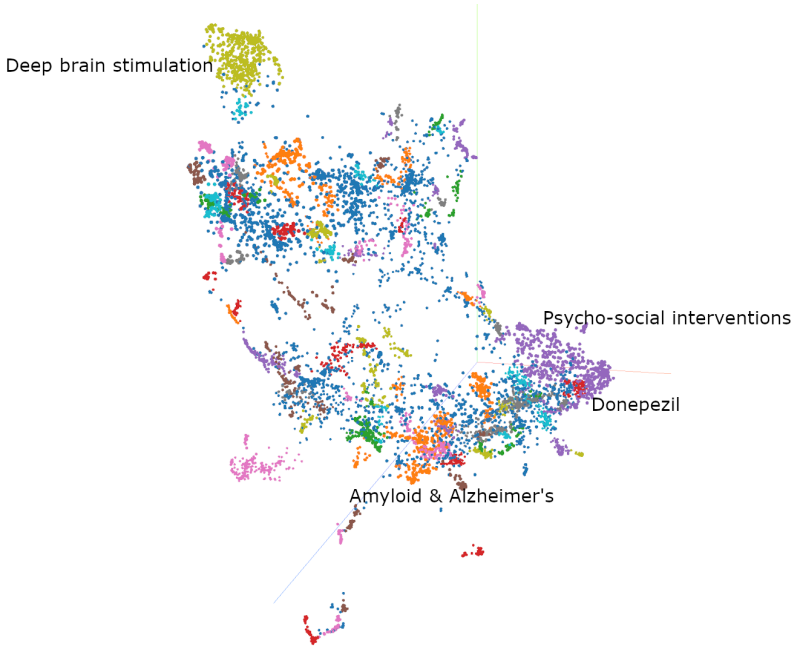

Figure 1: 3-D UMAP representation of the 8,689 hypotheses obtained from our PubMed search on neurodegeneration. The colors in the figure correspond to the major clusters identified by HDBSCAN. Deep brain stimulation is the chartreuse cluster in the upper left; psycho-social interventions are purple in the lower right; donepezil is dark red in the lower right; Aβ targetting agents and related imaging technologies for AD are orange in the lower middle. The dark blue points scattered throughout are the noise that was re-analyzed to identify neglected hypotheses.

However, there were also several hypotheses among our 25 "neglected" that were similar to dominant hypotheses from the larger set. In particular, our top 25 neglected hypotheses included hypotheses about donepezil and levodopa treatment in Parkinson's diseases, as well as Aβ, β-secretase, and other well-known and well-tested mechanistic hypotheses in AD.

4 Discussion

We have described a largely automated method, leveraging LLMs, high-dimensional embeddings, and dimensionality reduction techniques to visually represent, explore, and classify clusters of scientific hypotheses. The face validity of this method is demonstrated when observing the visual clusters corresponding to well-known hypotheses and research programs after the first pass of HDBSCAN. For organizations with interests in venturing into alternate mechanistic pathways or targets, the second step of re-analyzing and re-clustering the "noise" represents a promising advance. By employing our workflow, these organizations can systematically and confidently omit dominant clusters to explore the lesser-traversed territories.

Our approach is similar to Sourati and Evans' recent work in which they used semantic embeddings of paper abstracts and authors to conduct a "similarity walk" of the scientific literature [7]. Their aims were to (a) predict hypotheses that were likely to appear in future publications, and (b) identify hypotheses that, despite scientific plausibility, were very unlikely to appear in the literature. The latter goal, in particular, aligns closely with our aims here.

However, one outstanding challenge for Sourati and Evans, as well as our approach, is distinguishing between neglected (or unlikely) hypotheses that are worth further study and those that are neglected for good reasons (e.g., because the hypothesis has been thoroughly tested and refuted). The well-documented phenomenon of publication bias will amplify this challenge, since negative tests of highly novel or otherwise "neglected" ideas are much less likely to appear in print [8, 9].

But bearing that limitation in mind, this kind of computational investigation of hypotheses latent in the "noise" of the scientific literature represents a pragmatic, and relatively low-cost strategy to gain visibility into the landscape of overlooked or emerging ideas. Indeed, we are encouraged to find that our method required essentially no subject-matter expertise to generate the output. In this workflow, an experts time and attention is only required at the first and final steps—that is, to specify the search query and to review the output of neglected hypotheses. Although we focused on hypotheses related to drug development in neurodegenerative diseases, there is nothing in the workflow (apart from the initial literature search) that depends on these parameters. This makes our approach quite generalizable and potentially suitable for any domain of science.

For the specific use-case of a large pharmaceutical company's research and development portfolio, this kind of visibility into the landscape of scientific hypotheses could be valuable to help combat "cabal-like" dynamics, by surfacing alternative hypotheses. It would also be valuable as a highly-scalable approach to help a company establish an optimal, data-driven mix of "safe bets" on dominant hypotheses and "risky bets" on newer or neglected hypotheses.

The workflow piloted here motivates several lines of further meta-scientific inquiry. The use of full-text embeddings may yield additional neglected hypotheses (e.g., discussion of future directions that are shared across papers not otherwise belonging to the same cluster), potentially at the expense of less informative noise clusters (representing idiosyncratic differences between papers). Second, beginning from a broader or narrower search of the literature (e.g., beyond clinical trials or neurodegenerative disease; or limiting the results to early stage trials) may each offer advantages for identification of latent hypotheses that might otherwise be dismissed as noise. This along with other parameters may be optimized in future work with various objective functions (e.g., reducing the number of papers for which hypotheses cannot be extracted, using temporal dynamics in the evolution of hypothesis clusters as an additional self-supervised signal, and excluding apparently "dominant" hypotheses from the set of neglected hypotheses).

In sum, elaboration of this core workflow may accelerate the conception of novel or otherwise overlooked ideas and direct scientists to more efficiently evaluate the full set of hypotheses relevant to the development of breakthrough therapies.

Supplementary Material

You can download all the code and data necessary to replicate our analysis from the zip at the following link:

`https://drive.google.com/file/d/1URbMFUOxL1Vs96tmjlGW_FXzAOLCX-GJ/view?usp=sharing`

References

[1] Erik S Musiek, Teresa Gomez-Isla, David M Holtzman, et al. Aducanumab for alzheimer disease: the amyloid hypothesis moves from bench to bedside. *The Journal of Clinical Investigation*, 131(20), 2021.

[2] Sharon Begley. The maddening saga of how an alzheimer's 'cabal'thwarted progress toward a cure for decades. *STAT*, 2019.

[3] Jeffrey L Cummings, Dana P Goldman, Nicholas R Simmons-Stern, and Eric Ponton. The costs of developing treatments for alzheimer's disease: A retrospective exploration. *Alzheimer's & Dementia*, 18(3): 469–477, 2022.

[4] Feng Shi and James Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641, 2023.

[5] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[6] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[7] Jamshid Sourati and James A Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, pages 1–15, 2023.

[8] Colin B Begg. Publication bias. *The handbook of research synthesis*, 25:299–409, 1994.

[9] Phillipa J Easterbrook, Ramana Gopalan, JA Berlin, and David R Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.