AI Math Agents

LLMs, Information Systems Biology, Precision Health, and Alzheimer's Genomics (arXiv: 2307.02502)

Melanie Swan,* DIYgenomics, University College London (Research Associate CBT); Takashi Kido, Advanced Comprehensive Research Organization, Teikyo University; Eric Roland, RedBud AI, LLC; Renato P. dos Santos, Centre for Generative AI in Cognition and Education, Lutheran University of Brazil

Background

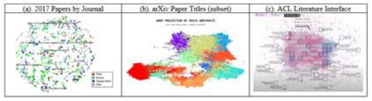
A contemporary challenge is deploying generative AI technologies to facilitate scientific advance. Generative AI mainly refers to Large Language Models (LLMs), computerized language models with billions of parameters (weighting nodes) pre-trained on large data corpora, e.g. GPT-4 (OpenAI), LaMDA (Google), and LLaMA (Meta AI). LLMs are transformer neural networks, an advance allowing an entire data corpus to be processed simultaneously to analyze connections between data elements. LLMs could be crucial to genomic medicine to process large volumes of multiscalar data and find the relevant level of interrelation in the complex domain [1,2].

The important result of LLMs is that they are a language-based interface to the computational infrastructure. This means via natural language for human-AI chat (e.g. ChatGPT), but more extensively, via formal languages such as programmatic code, mathematics, and physics for AI-directed computational interaction. The entirety of data corpora, not only word-based knowledge bases but also software code and mathematics, are being digitized and mobilized as available easy-to-use tools. Into this trajectory, the current work formulates an AI-based mathematics approach to Alzheimer's genomics, introducing Math Agents, the mathematical embedding, and equation clusters as tools for the autonomous evaluation and analysis of mathematical ecologies/mathscapes (sets of equations).

Method

AI-enabled mathematics tools are introduced and demonstrated in the case of Alzheimer's disease and aging as information systems biology problems. A theoretical model is elaborated, applying multiscalar physics mathematics (identifying near-far entropic correlations in systems) to disease mathematics and whole-human genomic data for two precision medicine participants. Vector embedding as a standard machine learning method is employed with mathematical equations and genomic data as the input.

Figure 1. Embedding Visualization examples with Academic Papers as the Data Corpus



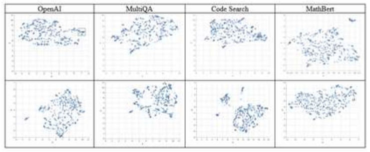
Vector embedding is the algorithmic processing of data into character strings for high dimensional analysis with results then translated back into low dimensional (2D) output for interpretation. Used routinely in machine learning, the method primarily targets traditional content types (text, images, sound). Word embeddings are used in LLMs to process news, social media, and other online data corpora such as Wikipedia. Embeddings are starting to be developed as a routine big data tool for the analysis of scientific content, but are still targeted to words (Figure 1). Visualizations of embeddings for three projects, each of which uses the entirety of an academic literature as the data corpus, are presented: (a) all papers published in 2017 by journal [3], (b) all arXiv paper titles and abstracts (2.3 million) [4], and (c) an interactive visualization of the ACL Anthology (Association of Computational Linguistics) of 85,000 papers [5]. The interpretation is that clustering provides a relevant signal, with zoom-explorer functionality as visibility into the data set.

Mathematical Embedding

The mathematical embedding is novel as an equation encoded in the form of a character-based vector string for input to high dimensional analysis in machine learning systems. The research aim is implementing mathematics as an approach to solving problems in data-intensive areas such as gene regulation, aging, and Alzheimer's disease. Causality is difficult to trace across tiers of complexity in biosystems. Hence, mathematics for modeling multiscalar physical systems is suggested, for example, renormalization mathematics which formalize a system-wide factor such as symmetry (in the universe) or free energy (in biological systems) conserved across scale tiers. Renormalization methods AdS/CFT correspondence and Chern-Simons theory are utilized.

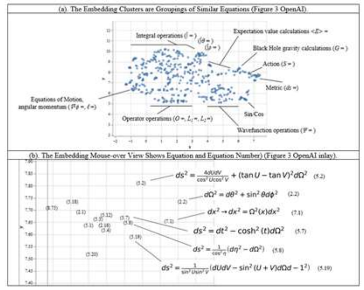
The mathematical embedding visualization of a 476-equation AdS/CFT mathematical ecology [6] is rendered with four standard embedding models (OpenAI, MultiQA, Code Search, MathBert) in LaTeX and symbolic Python (Figure 2). The abstracted view of mathematics as embeddings has a distinct structure. Clustering is relevant to interpretation (vs x,y values), indicating the grouping of similar kinds of equations, irrespective of order of appearance in the linear progression of the paper. The result is a new and higher level means of interacting with a set of equations, using Code Interpreter to "chat with a paper" or "chat with the equations in a paper," graph equations, generate data sets to fit mathematics, and write computer code to further vector-embed and evaluate systems of equations and model-fit with real-life data.

Figure 2. AdS/CFT Equation Clusters in Embedding Visualization (LaTeX and SymPy)



Annotated views illustrate (a) how similar groups of equations are grouped in the embedding method and (b) the mouse-over view of equation images by equation number (Figure 3 inlay). Equations are converted to LaTeX first, then symbolic Python as a readily mobilized computational form conducive to automated evaluation.

Figure 3. Annotated AdS/CFT Equation Clusters



Abbreviations

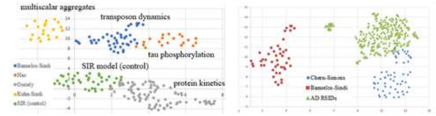
AD: Alzheimer's disease; AdS/CFT: anti-de Sitter space/conformal field theory; CSF: cerebrospinal fluid
eQTLs: expression quantitative trait loci; GWAS-EWAS: genome-wide and epigenome-wide association studies
SNP: single nucleotide polymorphisms; UMAP: uniform manifold approximation and projection (dimensionality reduction)

Results: Alzheimer's Genomics

Alzheimer's genomics is a whole-human genome-based approach to Alzheimer's disease involving GWAS SNPs, EWAS SNPs, eQTL transcriptomics (expressed RNA), and transposon indels. Beyond the ApoE profile, the current understanding of Alzheimer's disease genomic risk assessment includes multiple factors in a multiscalar analysis [7]. First, GWAS SNPs may suggest overall risk propensity for the disease. Second, EWAS SNPs may indicate which disease genes are actually expressed, as confirmed with transcriptome and biomarker assays (Aβ42/40 blood test, CSF). Third, the relation of GWAS-EWAS SNPs on a cis-trans (near-far) correlative basis may be relevant in terms of how different parts of the genome control which genes are expressed. Fourth, transposable elements may be related in activating Alzheimer's disease through genomic insertion-deletion events produced by viruses and other factors.

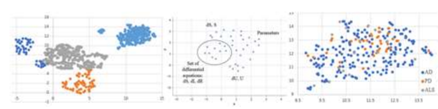
The benefit of the mathematical embedding for descriptive and interventional mathematics in Alzheimer's genomics is that the entirety of a mathscape (set of equations) can be seen in one at-a-glance abstracted and consolidated view. The visualization provides a first-pass view of the mathematics in a paper in the form of equation clusters and mouse-over images of equations. The aggregate view also allows related mathematical ecologies to be compared, and mathematics and data to be investigated in one view as two (ideally corresponding) representations of a system.

Figure 4. Mathematical Ecologies (a) Alzheimer's + SIR Model (control math); (b) Chern-Simons + AD SNPs



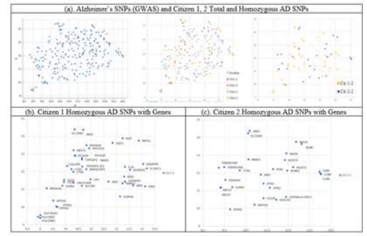
Biophysics renormalization mathematics are applied to Alzheimer's disease mathematics by extending the mathematical embedding to Alzheimer's disease models and genomic data, seen in four embedding visualizations of Alzheimer's disease mathscapes, with the SIR model as a mathematics control example [8] (Figure 4a). Low overlap is not surprising as the research programs target different aspects of Alzheimer's disease and have different bodies of mathematics: transposon dynamics, multiscalar aggregates, tau phosphorylation, and protein kinetics [9-12]. The SIR model and the protein kinetics embeddings are closer as both mathematical ecologies have differential equations as a focal point. The right figure examines mathematics and data together (Figure 4b), suggesting a better model-fit between the Chern-Simons mathematics and the Alzheimer's SNP data than with the transposon dynamics mathematics. Ecosystem analysis is implicated in modeling host-virus interactions in transposon indel activations of Alzheimer's disease. One next step is multigenic disease risk analysis as pathologies share pathways (Figure 5).

Figure 5. (a) AdS/CFT Mathematical Ecology (b) SIR Mathematics (c) Multi-disease Genomic view: AD, PD, ALS



Alzheimer's Genomics and SIR Precision Health

Figure 6. Embeddings Visualization of Data: Alzheimer's SNPs applied to Citizen 1, Citizen 2 Precision Health



Alzheimer's disease genomic risk is analyzed for two precision health participants with whole-genome sequencing data (Figure 6). An embedding visualization is performed for all GWAS-linked Alzheimer's disease SNPs and presented for Citizen 1 and Citizen 2's heterozygous (one alternative allele) and homozygous (two alternative alleles) SNPs (a). A gene-level zoom for homozygosity is shown to identify possible pathway-related interventional starting points. Notably, each individual is homozygous for different subsets of genes, with Citizen 1 at higher risk (confirmed empirically by an Aβ42/40 blood test). Citizen 1 has immune system related SNPs (CD33, HLA-DRB1), and the clathrin binder (PICALM) implicated in Alzheimer's disease. Citizen 2 is homozygous for cancer-upregulated membrane proteins (TREM) and cytokine-dependent hematopoietic cell linkers (CLNK). Both are homozygous for the solute carrier protein (SLC24A4) and the intracellular trafficking protein nexin (SNX1). Near-far genomic correlation analysis is enabled by AdS/CFT multiscalar modeling and deployed via the UMAP dimensionality-reduction embedding technique.

Conclusion

There is an opportunity to deploy AI-based tools to mobilize mathematics as a high-validation data corpus towards broadly humanity-benefiting use cases in global disease-preventing healthy well-being. Future work could elaborate Math Agent use cases in an SIR (sustaining, intervening, recovering) model for the societal realization of Precision Health based on two tiers of ongoing informational and interventional cycling of the population. Other mathematical discovery with AI Math Agents could include synthetic data generation to solve complex math ecologies and address causality. AI Math Agents with episodic memory (per file time-stamping) could be deployed in blockchain health ledgers and longitudinal personal health dossiers to identify the foundations of pathogenesis as dysregulatory genes are activated. In the short term, genomic variant and eQTL expression data is indicated for practical application to the unresolved challenge of Alzheimer's disease as the only top-five human killer with no survivors.

References

[1] Nguyen, E., Poli, M., Faizi, F. et al. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. arXiv:2306.15794v1.
[2] Batzoglou, S. (2023). Large Language Models in Molecular Biology: Deciphering the language of biology, from DNA to cells to human health. Towards Data Science. 2 June 20213. https://towardsdatascience.com/large-language-models-in-molecular-biology-9eb6b65d8a30.
[3] Meijer, H.J., Truong, J. & Karimi, R. (2021). Document Embedding for Scientific Articles: Efficacy of Word Embeddings vs TFIDF. arXiv:2107.05151v1.
[4] Depue, W. (2023). Embeddings for every research paper on the arXiv. Twitter. 25 May 2023. https://twitter.com/willdepue/status/1661781355452325889?lang=en. arXiv Title and Abstract embeddings available at: https://alex.macrocosm.so/download.
[5] Wang, Z.J., Hohman, F. & Chau, D.H. (2023). WIZMAP: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. arXiv:2306.09328v1. https://github.com/poloclub/wizmap.
[6] Kaplan, J. (2016). Lectures on AdS/CFT from the bottom up. Johns Hopkins Lecture Course.
[7] Sun, N., Akay, L.A., Murdock, M.H. et al. (2023.) Single-nucleus multi-region transcriptomic analysis of brain vasculature in Alzheimer's disease. Nat Neurosci. 26:970–982. https://doi.org/10.1038/s41593-023-01334-3.
[8] Wyss, A. & Hidalgo, A. (2023). Modeling COVID-19 Using a Modified SVIR Compartmental Model and LSTM-Estimated Parameters. Mathematics. 11:1436. https://doi.org/10.3390/math11061436.
[9] Banuelos, M. & Sindi, S. (2018). Modeling transposable element dynamics with fragmentation equations. Mathematical Biosciences. 302:46–66. https://doi.org/10.1016/j.mbs.2018.05.009.
[10] Banwarth-Kuhn, M. & Sindi, S. (2019). Multi-Scale Mathematical Modeling of Prion Aggregate Dynamics and Phenotypes in Yeast Colonies. Biomathematics. 1-30. http://dx.doi.org/10.5772/intechopen.88575.
[11] Hao, W. & Friedman, A. (2016). Mathematical model on Alzheimer's disease. BMC Syst Biol 10(108). https://doi.org/10.1186/s12918-016-0348-2.
[12] Fornari, S., Schafer, A., Kuhl, E. & Goriely, A. (2020). Spatially-extended nucleation-aggregation-fragmentation models for the dynamics of prion-like neurodegenerative protein-spreading in the brain and its connectome. Journal of Theoretical Biology. 486:110102. https://doi.org/10.1016/j.jtbi.2019.110102.