8-2022

Development Of Graphical Models And Statistical Physics Motivated Approaches To Genomic Investigations

Yashwanth Lagisetty

Development of Graphical Models and Statistical Physics Motivated Approaches to Genomic Investigations

by

Yashwanth Lagisetty, B.S.

APPROVED:

_____
Edgar T. Walters, Ph.D.
Advisory Professor

_____
Olivier Lichtarge, M.D., Ph.D.
Advisory Professor

_____
Prahlad Ram, Ph.D.

_____
Anil Korkut, Ph.D.

_____
Marsal Sanches, M.D., Ph.D.

APPROVED:

_____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Science

Development of Graphical Models and Statistical Physics Motivated Approaches to Genomic Investigation

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Yashwanth Lagisetty, B.S.
Houston, Texas

August, 2022

Acknowledgements

The journey towards a PhD can be a trying experience. There are often short bursts of success followed by long draughts of failed experiments. Over the course of this wearing process, I have had the privilege of meeting and being mentored by several wonderful scientists. First, of course, I would like to thank my advisor, Dr. Olivier Lichtarge for rolling the dice and taking me on in his lab. Dr. Lichtarge provided me the opportunity and freedom to pursue my research interests regardless of how unorthodox they may have been. His lab fostered a collaborative and teamwork driven environment that served only to encourage and reward curiosity. Most of all I am thankful to have spent the last several years in his lab learning how to think critically and ask questions like a scientist.

I would like to thank Dr. Terry Walters for his unwavering support in all of my many, very unorthodox, ventures throughout my time in the MD/PhD program. His advocacy for what is best for students remarkable and a quality that I will strive to replicate in any mentorship role I may take up in the future. I would like to offer my sincere gratitude to the members of my advisory committee - Dr. Prahlad Ram, Dr. Anil Korkut, and Dr. Marsal Sanches. I would not have gotten far without all of the guidance, mentorship, and research advice you have provided over the course of my research training.

I would be lost without all the support I've had from my lab mates! I would like to thank Danny Lee and Thomas Bourquard for answering, with phenomenal patience, the thousands of tedious questions I've asked over the years. Without their incredible mentorship, I very much would have been dead in the water. They are an inspiration and I hope to pay forward the incredible mentorship I have received from them. I thank Jenn Asmussen, Maryam Samieinasab, Dillon Shapiro, and the rest of the lab for their support, friendships, and general shenanigans that have made these last few years enjoyable.

These acknowledgements would be incontrovertibly incomplete without recognizing the immense support I have received from the students, staff, and leadership of the MD/PhD program. First, I would like to thank the directors and co-directors – Dr. Dianna Milewicz, Dr. Terry Walters, Dr. Ruth Heidelberger, Dr. Wendy Woodward, Dr. George Calin, and Dr. Raghu Kalluri – for their support and encouragement in my rather unconventional track through the program thus far. I'd like to thank Rachel Van Drunen, Michael Longmire, Nancy Wearing, Pahul Hanjra, Amanda Lanier, Kareem Wahid, and all of the other GSBS and MD/PhD students with whom I've come to form close friendships.

Lastly, I would like to thank my family including parents, Sai and Vani, and my sister, Ihika, for being a great support system during my PhD training. Above all others, I would like to thank my cats Oppenheimer and Watermelon for always providing me with emotional support, good laughs with their general tomfoolery, and correcting my math.

Development of Graphical Models and Statistical Physics Motivated Approaches to Genomic Investigation

Yashwanth Lagisetty, B.S.

Advisory Professor: Olivier Lichtarge, M.D., Ph.D., Edgar T. Walters, Ph.D

Abstract

Identifying genes involved in disease pathology has been a goal of genomic research since the early days of the field. However, as technology improves and the body of research grows, we are faced with more questions than answers. Among these is the pressing matter of our incomplete understanding of the genetic underpinnings of complex diseases. Many hypotheses offer explanations as to why direct and independent analyses of variants, as done in genome-wide association studies (GWAS), may not fully elucidate disease genetics. These range from pointing out flaws in statistical testing to invoking the complex dynamics of epigenetic processes. In the studies outlined here, however, we focus on the hypothesis that interactions between genes may be a potential culprit. To probe this hypothesis, we begin by developing an algorithm, GeneEMBED, to model the total effect of protein coding variants in various genes across a molecular network of genetic interactions. Given a population of disease and healthy individuals, GeneEMBED systematically evaluates the relative contribution of a gene to disease. The associations are quantified by examining the patterns of differential perturbations in the gene's interactions throughout a biological network. As a proof-of-concept, we applied GeneEMBED to two late-onset Alzheimer's disease (AD) cohorts of 5,169 exomes and 969 genomes. We identified 143 candidate disease-associated genes across the two cohorts and three biological networks. These candidate genes were differentially expressed in both bulk and single-cell RNA expression data from post-mortem AD brains. Knockouts of these candidates in mice were known to lead to abnormal

neurological phenotypes. Lastly, in vivo drosophila assays of candidates showed they modified neurodegenerative phenotypes. Next, we focus on the discrepancies between the functional impact of mutations across different genes. While tools to predict the degree of functional impact a given coding mutation will have on the encoded protein are widely successful, they often make predictions relative to the given gene. To this effect, we extend principles of statistical mechanics to biology to measure any given gene's relative mutational intolerance. Importantly, these mutational intolerance scores can distinguish essential genes from non-essential genes in E.coli. In humans, they can segregate genes that cause autosomal dominant Mendelian diseases from non-disease genes. Similarly, highly mutationally intolerant genes were enriched in core and conserved biological processes across three different species. Conversely, mutationally tolerant genes were involved in adaptive processes, again across three different species. Most notably, we found that mutational intolerance scores highly correlated with experimentally measured fitness effects of gene knockdowns. Together, these efforts provide new tools with which to investigate disease-gene associations and provide insights into the biological dynamics of gene networks.

Table of Contents

Approval Page.....i

Title Page.....ii

Acknowledgements iii

Abstract v

Table of Contents vii

Table of Figures x

List of Tables xi

Chapter 1: Introduction 1

Chapter 1.1: Disease-Gene Association Analyses 1

Chapter 1.1.1: Genetic Interactions 2

Chapter 1.1.2: Graph Based Approaches 3

Chapter 1.2: Graph Learning in Genomics 5

Chapter 1.2.1: What are graphs? 5

Chapter 1.2.2: Types of Graph Learning 6

Chapter 1.2.2.1: Node Level 8

Chapter 1.2.2.1.1: Graph Neural Network Architecture 9

Chapter 1.2.2.1.1.1: Aggregation methods 10

Chapter 1.2.2.1.1.2: Update mechanisms 11

Chapter 1.2.2.1.2: Matrix Factorization 12

Chapter 1.2.2.1.3: Random Walk Approaches 13

Chapter 1.2.2.1.4: Structural Embeddings 14

Chapter 1.2.2.1.5: Node Level applications in large scale -omic analyses 15

Chapter 1.2.2.2: Edge Level 16

Chapter 1.2.2.2.1: Local Similarity Indices 16

Chapter 1.2.2.2.2: Probabilistic models 17

Chapter 1.2.2.2.3: Matrix Completion 17

Chapter 1.2.2.2.4: Embedding based approaches 18

Chapter 1.2.2.3: Graph Level 19

Chapter 1.2.2.3.1: Graph Kernels 19

Chapter 1.2.2.3.2: Graph Embedding via Node Embeddings 20

Chapter 1.3: Quantification of Effects of Coding Mutations 21

Chapter 1.4: Dissertation Objectives 23
Chapter 2: Gene Embedding Provides Novel Insights into Disease Mechanisms 25
Chapter 2.1: ABSTRACT 26
Chapter 2.2: Introduction 27
Chapter 2.3: RESULTS 29
Chapter 2.3.1: GeneEMBED identifies genes that are perturbed in AD 29
Chapter 2.3.2: GeneEMBED CANDIDATES ARE ROBUSTLY CONNECTED AND RELEVANT TO AD 34
Chapter 2.3.3: GeneEMBED candidates are functionally connected and enriched for *in vivo* modulators of neuronal dysfunction triggered by tau and β-amyloid 41
Chapter 2.3.4: GeneEMBED shows robustness across various cohort sizes 48
Chapter 2.3.5: Characterization of performance of PCA vs full embedding distances 52
Chapter 2.3.5: Characterization of alternative edge weighting schemes 53
Chapter 2.3.6 Characterization of sensitivity of GeneEMBED to false negative and false positive edges 54
Chapter 2.3.7: Characterization of GeneEMBED performance in unbiased networks 57
Chapter 2.3.8: Characterization of GeneEMBED in the presence of uninformative mutational data 58
Chapter 2.4: DISCUSSION 59
Chapter 2.5: MATERIALS AND METHODS 66
Chapter 2.5.1:Whole Exome/Genome Sequencing Data 66
Chapter 2.5.2: Variant Scoring Methods 67
Chapter 2.5.3: GeneEMBED 68
Chapter 2.5.4: Downsampling analyses 73
Chapter 2.5.5: Negative control experiment 73
Chapter 2.5.6: MAGMA analyses 73
Chapter 2.5.7: Recall of known AD genes 74
Chapter 2.5.8: Network Analyses 74
Chapter 2.5.9: RNA sequencing Analysis 75
Chapter 2.5.10: Pathway Enrichment Analysis 75
Chapter 2.5.11: Mouse Phenotype Analysis 76
Chapter 2.5.12: Drug interaction Analysis 76
Chapter 2.5.13: Drosophila strains and neuronal dysfunction assay 77
Chapter 3: Thermodynamics inform mutational intolerance 79
Chapter 3.1: Introduction 79

Chapter 3.2: Results 81
Chapter 3.2.1: Energetics of Coding Mutations 81
Chapter 3.2.2: Equipartition Theorem Informs Mutational Intolerance 86
Chapter 3.2.3: Biological Relevance of Mutational Intolerance 87
Chapter 3.3: Discussion 102
Chapter 3.4: Materials and Methods 103
Chapter 3.4.1: Whole Exome/Genome Sequencing Data 103
Chapter 3.4.2: Measurement of goodness of fit to Boltzmann 104
Chapter 3.4.3: Calculation of $\boldsymbol{\mu}$: 106
Chapter 3.4.4: Correlation of $\boldsymbol{\mu}$ with network centralities: 106
Chapter 3.4.5: Pathway Enrichment Analyses: 106
Chapter 3.4.6: Characterization of $\boldsymbol{\mu}$ differences in Mendelian Diseases 107
Chapter 3.4.7: Characterization of $\boldsymbol{\mu}$ differences in Essential, conditionally essential, and auxotrophic genes in *E. Coli*: 107
Chapter 3.4.8: Direct correlation between $\boldsymbol{\mu}$ and gene fitness effect: 108
Chapter 3.4.9: Rank Bias of Olfactory Pathway genes: 108
Chapter 4: Discussion and Future Directions 110
GeneEMBED and exomic analysis 111
Limitations and Future Directions 113
Equipartition and Mutational Intolerance 114
Limitations and Future Directions 116
Appendix of Tables 118
REFERENCES 141
VITA 178

Table of Figures

Figure 1 6
Figure 2 7
Figure 3 31
Figure 4 35
Figure 5 37
Figure 6 39
Figure 7 41
Figure 8 45
Figure 9 46
Figure 10 50
Figure 11 55
Figure 12 63
Figure 13 82
Figure 14 84
Figure 15 87
Figure 16 89
Figure 17 90
Figure 18 91
Figure 19 93
Figure 20 94
Figure 21 96
Figure 22 99
Figure 23 100
Figure 24 105

List of Tables

Table 1 119
Table 2 123
Table 3 124
Table 4 124
Table 5 125
Table 6 127
Table 7 128
Table 8 128
Table 9 129
Figure 10 129
Figure 11 129
Figure 12 134
Figure 13 137
Figure 14 138
Figure 15 138
Figure 16 138
Figure 17 139
Figure 18 139
Figure 19 139
Figure 20 140

Chapter 1: Introduction

The success of the Human Genome Sequencing project has sparked an ongoing revolution in genetic and genomic research. Among the many scientific advances and discoveries being made in this transformative stage in the field, there is one that has arguably the most impact on human health and disease. This is the understanding that nearly all diseases have genetic components [1]. These genetic influences can be small or large, encompassing a single gene (monogenic) or multiple genes (polygenic). In the case of monogenic diseases, clear inheritance patterns have been instrumental in pinpointing causative genetic changes. The genetic cause for cystic fibrosis, for example, was identified through a combination of genetic and pedigree analyses in families [2,3]. However, identifying similar causative variations in polygenic diseases, where there is a complex interplay between many genetic variations, is still an open and challenging problem. One way of identifying so-called 'candidate genes' is through disease-gene association studies. While disease-gene association tools have been fruitful in producing drug targets and prognostic biomarkers [4], they have several methodological limitations. Much of the research presented here aims to address some of these obstacles.

Chapter 1.1: Disease-Gene Association Analyses

Genome Wide Association Studies (GWAS) have been at the forefront of identifying disease susceptibility genes for many years. GWAS candidate genes have been utilized in translational disease research in a variety of ways. From biomarkers to drug development, GWAS candidate genes have even led to discoveries of novel biological mechanisms [4]. For example, the role of autophagy in Crohn's disease was unknown until the identification of autophagy related genes through GWAS [5]. Despite its impressive portfolio of discoveries, GWAS alone is unlikely to give the full genetic picture of diseases [6]. Indeed, in late onset

Alzheimer's disease (AD), the estimated genetic heritability is 60-80% [7,8]. Though more than 40 AD loci have been identified, they account for only a fraction (~33%) of the expected heritability [9,10]. This problem of an incomplete or "missing genetic component" is prevalent in most complex diseases [6,10].

One explanation for this "missing genetic component" stems directly from the burden of multiple testing in GWAS. Due to stringent multiple testing corrections, many loci with small but disease relevant effects do not attain significant q-values [4,11,12]. Intuitively, this can be circumvented by simply increasing sample sizes of analyses. While this approach has been adopted and successful in several diseases due to the recent development of large sequencing projects and consortia [13,14], it is a costly and sometimes infeasible solution [4]. An alternative approach is to consider gene- or gene-set-based associations as they can improve statistical power by reducing the number of tests [4,12]. This approach has motivated the invention of several gene prioritization methods including commonly used tools like SKAT and MAGMA. SKAT (sequence kernel association test) performs multiple regression analysis between phenotype and genetic variants in a specified region (e.g. gene) while considering covariates [15]. Similarly, MAGMA (Multi-marker Analysis of GenoMic Annotations) performs gene- or gene-set-based analysis by using a multiple linear regression strategy between the phenotypes and genotype data. Specifically, it uses the principal components of gene's variants as predictors [16]. These type of gene-level or gene-set-level analyses have been fruitful in identifying risk factors for disease, such as nicotinamide metabolism in colon cancer [17]. However, despite their success, a complete understanding of the genetic background of complex diseases remains elusive.

Chapter 1.1.1: Genetic Interactions

Another, attractive hypothesis suggests that genetic interactions may be complicit in the "missing genetic component" problem. Genetic interactions are functional interactions

observed among variants of a gene where the resulting phenotype differs from the phenotypes of each variant [18]. Thus, otherwise unremarkable variants can combine to generate complex phenotypes [18,19]. Indeed, experiments in yeast demonstrate that genetic interactions greatly influence complex genetic traits [20]. In humans, specific genetic interactions have been associated with complex traits and diseases [21–23]. In one study, Martin et al [22] found that in individuals infected with human immunodeficiency virus type 1 activating KIR KIR3DS1 allele in conjunction with HLA-B Bw4-80Ile allele would delay the progression to AIDS. However, when the HLA-B allele was active without the KIR allele, no effect on the AIDS outcome was observed. Furthermore, they found that activation of KIR allele without HLA-B allele resulted in rapid progression to AIDS. In another example, Leggio et al [23] studied the genetic interaction between dysbindin-1 (Dys) and dopamine D3 receptor variants. The authors found that in both schizophrenia patients and mouse models, the presence of mutations in both Dys and D3 were associated with improved cognitive function. Moreover, they found that genetic interactions between Dys and D3 lead to an increased D2/D3 ratio in the prefrontal cortex of the brain, but this was not seen in the stratum. These results suggest that genetic interactions may have varying effects even within the same organ system. While such interactions have been identified on small scales, genome-wide discovery of pairwise genetic interactions presents major challenges. Even under reasonable assumptions, theoretical analyses suggest that nearly 500,000 subjects would be necessary to identify statistically significant genetic interactions [18].

Chapter 1.1.2: Graph Based Approaches

To combat these constraints, creative solutions using prior knowledge in the form of protein-protein interaction (PPI) networks have been proposed as way to identify candidate disease genes. Tools like GeneWanderer [24], MaxLink [25], and GUILD [26] follow this strategy. GeneWanderer employs a variety of network-based distance metrics to score

distances of genes of interest to all known disease-related genes. Genes of interest are then ranked based on score and prioritized accordingly [24]. MaxLink uses a guilt-by-association network search algorithm which identifies and ranks new candidate genes based on their connectivity to a set of known disease genes [25]. Similarly, GUILD uses multiple algorithms to measure the relatedness of a given gene to a set of previously established disease related genes [26]. Overall, this class of tools aim to characterize the topological traits of potential disease driving genes. However, integration of human cohort data is likely necessary to better understand the role of genetic interactions in disease. To this effect, methods such as PINTA [27] have incorporated disease specific expression data with network-based prioritization algorithms to identify candidate genes. Similarly, HIT'nDRIVE [28] combines patient-specific sequence-altered data with patient-specific expression information. By proposing to find the smallest set of sequence-altered genes which describes the largest portion of expression outliers through network-based algorithms, HIT'nDRIVE identifies candidate disease-genes. While these methods have been successfully applied to identify cancer drivers and predict drug efficacy on cancer cell lines, their dependence on expression data can be a limiting factor in their translation to other complex diseases. HotNet2 [29], and its variations, provide a promising method of identifying significantly mutated subnetworks of genes based on disease-relevant mutation data coupled with network algorithms. Using a PPI network, HotNet algorithms assign nodes a 'heat' based on single nucleotide variants (SNV) and copy number alterations (CNA) data from cancer cohorts. The 'heat' is diffused across the network via diffusion-based methods (HotNet) or random walk algorithms (HotNet2, Heirarchical HotNet [30]), and nodes that send and receive significant heat are reported. While these methods have been successful in identifying cancer drivers and risk factors, they rely on somatic mutational data and are not amenable to the traditional case-control study designs of germline genome-wide association studies.

Chapter 1.2: Graph Learning in Genomics

Advancements in non-Euclidean deep learning have opened the door to new ways to analyze genomic data in the context of biological networks. Indeed, recent years have seen a sharp increase in the number of applications of graph learning applied to bioinformatic questions [31]. The utility of graph learning in large-scale bioinformatic analyses has been exemplified in their success in a number of different applications. For example, Zitnik et al [32] used graph learning architectures to predict the polypharmacy side effects of drugs using protein-protein, drug-protein, and drug-drug interaction networks. Wang et al [33] utilized graph convolutional networks on multi-omic data to successfully classify patients across three different diseases. Similarly, Chereda et al [34] used graph neural network architectures to predict patient-specific cancer metastatic events using personal gene expression data, and, moreover, point to the specific subnetworks responsible for classification. To appreciate the versatility of graph learning and its nuances, it is important to understand the mechanisms by which it is made possible.

Chapter 1.2.1: What are graphs?

A graph is a mathematical structure in which a set of objects are organized by their personal relationship to one another. Consider, as an example, three cities where each can be easily accessed by a major highway from any of the other cities. This accessibility relationship can be described by the graph in **Figure 1A**. Nodes represent the object of interest, in the previous example nodes would represent cities. Edges represent some relationship of interest that exists between two nodes. In our previous example, edges would represent the highways connecting two cities. Edges can also be forced to hold directional information. If, for example, the highway connecting cities 1 and 2 were to be only one-way, then the relationship would be described like **Figure 1B**. Graphs which contain only the

relationships are *undirected*, while graphs which contain both relationship and directional information are *directed*.

Formally, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is given by the set of nodes $\mathcal{V}$ and set of edges $\mathcal{E}$. Edges are denoted as $(u, v) \in \mathcal{E}$ for any edge between $u \in \mathcal{V}$ and $v \in \mathcal{V}$. The overall connectivity of a graph can be represented by its *adjacency* and *degree* matrices. The adjacency matrix, $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, describes the connections between nodes. Typically, the presence of an edge is represented in the matrix as $\boldsymbol{A}[u, v] = 1$, while the absence of an edge is $\boldsymbol{A}[u, v] = 0$. If edges have attributes such as edge weights, then they can be reflected in the adjacency matrix entries. The degree matrix $\boldsymbol{D} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, is a diagonal matrix whose entries describe the number of edges attached to each node.

Chapter 1.2.2: Types of Graph Learning

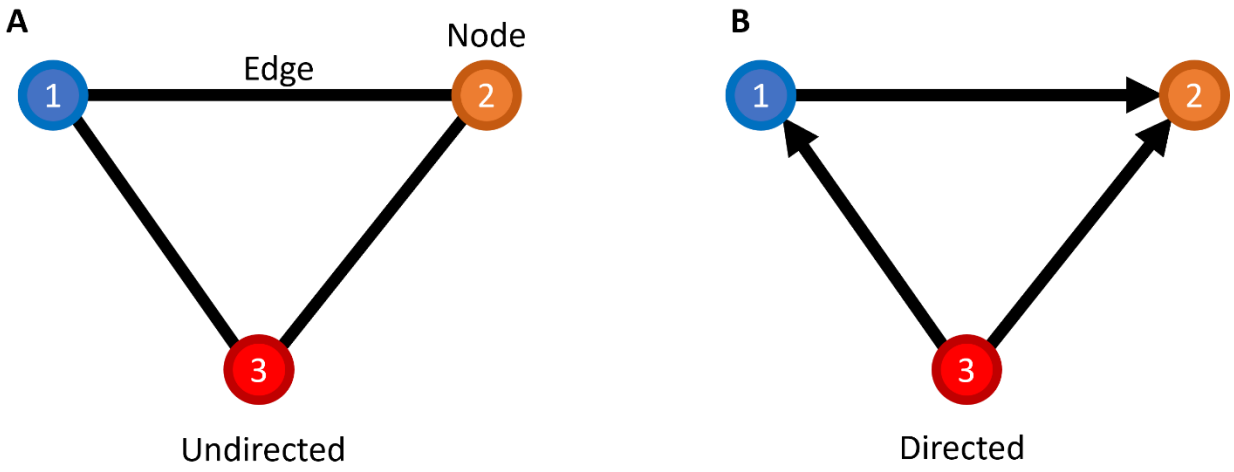


Figure 1: Graphs are mathematical structures which represent physical or abstract relationships between objects. (A) shows an undirected graph wherein nodes are connected if there is a relationship between them. For example, if nodes 1 and 2 are proteins which interact, there exists an edge between them. Undirected graphs do not contain directional information. (B) shows a directed graph. Like (A), edges in (B) represent relationships between nodes. However, in directed graphs, edges have orientations which specify the direction of interactions.

Traditionally, *supervised* and *unsupervised* learning offer broad categorization of machine learning solutions. Supervised learning offers a framework for the task of predicting an output given a set of input data. Unsupervised learning focuses on the task of inferring patterns among input data in order to usefully cluster or categorize the data. Machine learning on graphs is fundamentally no different from classical machine learning, where we seek specific solutions to the questions at hand. However, these typical categorizations are not particularly useful in organizing the full diversity of graph learning species. Instead, it is more helpful to look at the component of the graph with which the graph learning is concerned: *node* level, *edge* level, or *graph* level, as shown in **figure 2**.

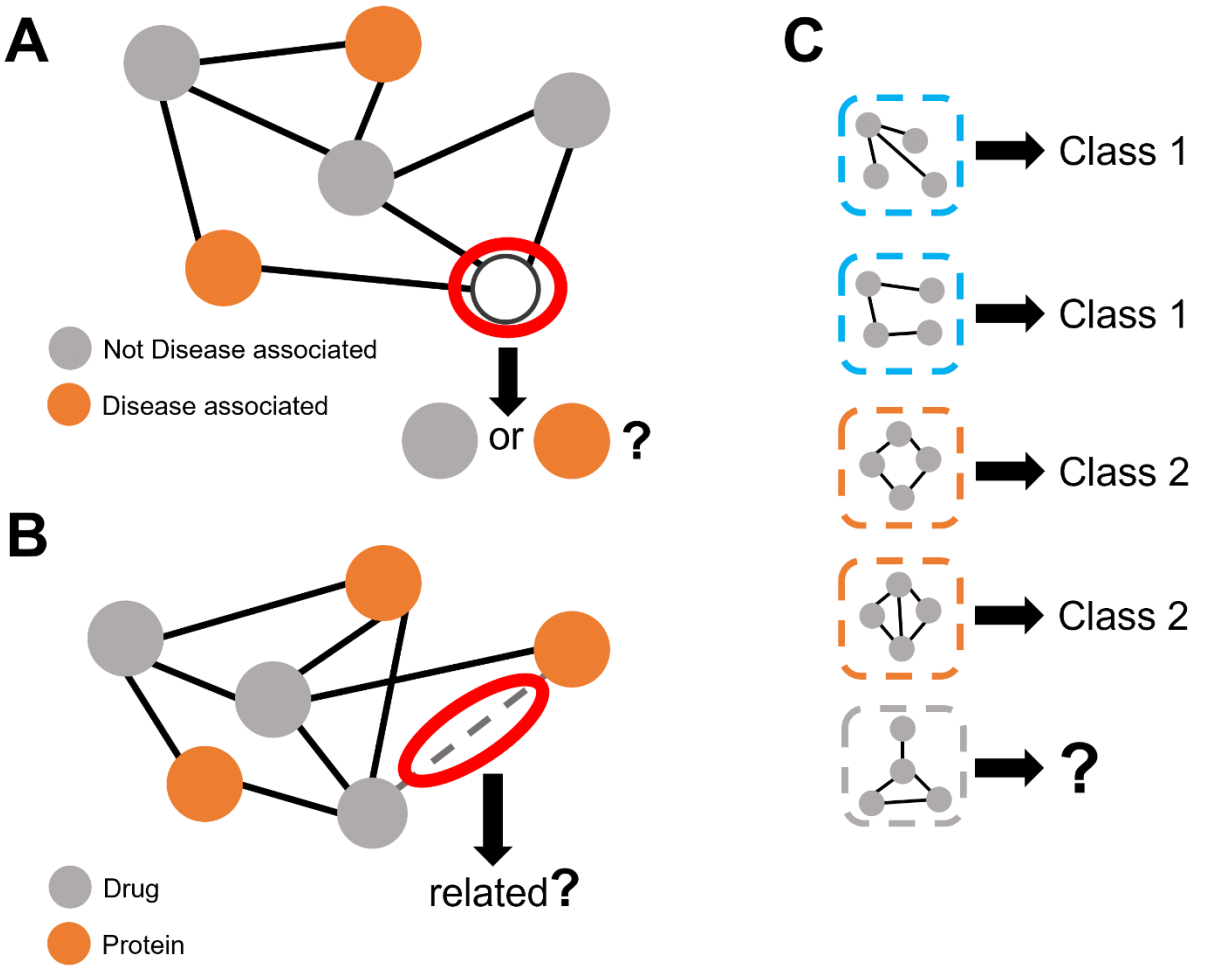


Figure 1: Types of graph learning. (A) Node level tasks often focus on analyzing individual nodes to classify (or regress) them based on specific labels. (B) Edge level tasks focus on predicting specific properties of edges, usually whether or not an edge exists between two nodes. (C) Graph level tasks focus on classifying whole graphs based on specific labels or regressing over them to predict continuous value properties.

Chapter 1.2.2.1: Node Level

An example of a node level task is given in **Figure 2A**. Here we consider a network of genes, some of which are associated with a disease (orange), while others are not associated with the disease (gray). The classic node classification task then asks, given an unclassified node, can we predict whether it is disease associated or not? The ability to classify critically

rests on the construction of *node embeddings*. These are alternative representations onto which a node is mapped, such that its semantic and structural properties are preserved. Thus, in this embedding space, any node which is close to another must have had similar qualitative (topological) and quantitative (features) properties. These embeddings can be constructed in a variety of ways.

Chapter 1.2.2.1.1: Graph Neural Network Architecture

The goal of the graph neural network (GNN) architecture is to extend the success of deep neural networks into non-Euclidean, graph domain. The central tenet being that we would like to define some mapping or encoding function ($f: u \rightarrow \mathbb{R}^d$) which generates representations for nodes that reflect their structural properties and their individual feature information. The groundwork for the basic GNN model is laid through a generalization of convolution functions to discrete non-Euclidean domains [35] (though there have been many alternative derivations leading to the base model). The fundamental mechanism used by the GNN is a form of a *message passing* scheme, wherein vectors of information are traded among connected nodes and at each trade, updated by a neural network. The vector message passed along at each iteration is called a *hidden embedding* $\boldsymbol{h}_u^k$, where u is the node and k is the layer or iteration. In order to generate a hidden embedding, first the hidden embeddings of all nodes v in the neighborhood of u, $\mathcal{N}(u)$, are aggregated. This neighborhood level hidden embedding is then concatenated with the current hidden embedding of node u. Finally, the hidden embedding for $k+1$th iteration is given by an update function on the concatenated embedding vector:

$$\mathbf{h}_u^{k+1} = \mathrm{UPDATE}^k\left(\mathbf{h}_u^k, \mathrm{AGGREGATE}^k\left(\{\mathbf{h}_v^k, \forall v \in \mathcal{N}(u)\}\right)\right) \quad (1)$$

where UPDATE and AGGREGATE are differentiable but arbitrarily defined functions. However, the basic GNN model defines these functions by elementwise nonlinearities commonly used in deep learning (e.g. ReLU or eLU), and affine transformations, respectively:

$$\mathbf{h}_u^{k+1} = \sigma\left(\boldsymbol{W}_{self}^{k+1}\mathbf{h}_u^k + \boldsymbol{W}_{\mathcal{N}}^{k+1}\sum_{v\in\mathcal{N}(u)}\mathbf{h}_v^k + \boldsymbol{b}^{k+1}\right) \quad (2)$$

Where $\boldsymbol{W}_{self}^{k+1}$, $\boldsymbol{W}_{\mathcal{N}}^{k+1} \in \mathbb{R}^{d^{k+1}\times d^k}$ are trainable weight matrices and the σ is the nonlinearity.

Chapter 1.2.2.1.1.1: Aggregation methods

While the simplest way to aggregate the hidden embedding messages of the neighboring nodes is a linear transformation, there are a variety of more complex and generalized methods. One class of alternative aggregation methods normalizes the incoming hidden embeddings by their neighborhood size (degree). This is motivated by the fact that repeated iteration over a neighborhood aggregation seen in equation 2 could lead to explosive gradients or could be highly sensitive to the size of a node's neighborhood. One commonly used aggregation scheme to avoid this invokes a symmetric normalization [36], defining updates as:

$$\mathbf{H}^{k+1} = \sigma\left(\boldsymbol{D}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{A}\boldsymbol{D}^{-\frac{\mathbf{1}}{\mathbf{2}}}\boldsymbol{H}^{\boldsymbol{k}}\boldsymbol{W}^{k+1}\right) \quad (3)$$

where **H** is the matrix of hidden embeddings **h**, ***D*** is the degree matrix, ***A*** is the adjacency matrix and **W** is the trainable weights. The neighborhood aggregation scheme can be seen more clearly in this form:

$$\text{AGGREGATE}\big(\mathcal{N}(u)\big) = \sum_{v\in\mathcal{N}(u)}\frac{\mathbf{h}_v}{\sqrt{|\mathcal{N}(u)||\mathcal{N}(v)|}} \quad (4)$$

Normalized aggregation schemes such as this are attractive for their ability to stabilize message passing and gradient propagation. Interestingly, however, Xu et al [37] demonstrates that this sort of normalized or mean aggregation scheme has less provable expressive power than the simple case of summation in equation 2.

Another popular technique for message aggregation in GNNs is to use attention. Here, each neighbor $v \in \mathcal{N}(u)$ is given an attention weight which describes its relative importance to u compared to all others in the neighborhood. These attention or importance weights are trainable parameters which are optimized over the course of learning. This attention-based aggregation scheme was first introduced by Velickovic et al [38], and is formulated as follows:

$$\text{AGGREGATE}\big(\mathcal{N}(u)\big) = \sum_{v \in \mathcal{N}(u)} \alpha_{u,v} \mathbf{h}_v \tag{5}$$

where $\alpha_{u,v}$ is the attention weight for a neighbor $v \in \mathcal{N}(u)$, and is given by:

$$\alpha_{u,v} = \text{softmax}_v(\mathbf{a}^T[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_v]) = \frac{\exp(\mathbf{a}^T[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_v])}{\sum_{j \in \mathcal{N}(u)} \exp\big(\mathbf{a}^T\big[\mathbf{W}\mathbf{h}_u || \mathbf{W}\mathbf{h}_j\big]\big)} \tag{6}$$

where $\mathbf{a}$ is a trainable attention vector, $\mathbf{W}$ is the trainable weight matrix, and || denotes a concatenation operation. The attention mechanism shown above and variations of it have been useful in a variety of applications. For example, Ingraham [39] employed a variation of the graph attention network introduced in Velickovic et al [38] in order to predict protein sequences given the 3D structures represented as graphs.

Chapter 1.2.2.1.1.2: Update mechanisms

The expressive power of GNNs, much like deep neural networks, relies heavily on their architecture. Specifically, the decision UPDATE and AGGREGATE functions can influence the overall expressiveness of the GNN. In many commonly used GNNs, the UPDATE function is comprised of only single nonlinearity units, e.g. ReLU. However, this type

of use of nonlinearity can lead to linear transformation-like behaviors (Lemma 7 in [37]). As a result, structurally distinct nodes may be embedded similarly. In their theoretical exploration of GNNs, Xu et al [37] demonstrated that an UPDATE function given by a multi-layer perceptron (MLP) provides much greater expressive power than the traditional single nonlinearity units, producing an architecture given by:

$$\mathbf{h}_u^{k+1} = \mathrm{MLP}^{k+1}\left(\left(1+\epsilon^{k+1}\right)\mathbf{h}_u^k + \sum_{v\in\mathcal{N}(u)} \mathbf{h}_v^k\right) \tag{7}$$

Chapter 1.2.2.1.2: Matrix Factorization

While graph neural network strategies have been favored in recent years, owing to their remarkable success, it is worth understanding a few of the other classes of node embedding strategies that exist. Matrix factorization-based approaches follow the principle that we can construct instructive low-dimensional representations using some variation of a node similarity matrix. Of the many matrix factorization methods, *Laplacian eigenmaps* are among the most popular and widely used. In their seminal paper, Belkin et al [40] introduced the idea of Laplacian eigenmaps by considering the node embedding problem as a problem of mapping a graph onto a line such that the points stay as close together as permissible. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, this is to say we would like to find some $y_i \in \mathbb{R}$ which minimizes:

$$\sum_i \sum_j \left(y_i - y_j\right)^2 \tilde{A}_{ij} \tag{8}$$

where $\tilde{A}$ is the weighted adjacency matrix. We can also note that for any *y*, we have that:

$$\sum_i \sum_j \left(y_i - y_j\right)^2 \tilde{A}_{ij} = 2\mathbf{y}^T L\boldsymbol{y} \tag{9}$$

where *L* is the *graph Laplacian*, an extension of the traditional continuous multivariate Laplacian on to discrete spaces. The derivation of this analog is made apparent when we consider that each node $v \in \mathcal{V}$ is akin to a point in Euclidean space and the permissible directions in which 'flow' can occur is delineated by the set of edges $\mathcal{E}$. Laplace eigenmaps therefore state that low-dimensional representations for each node can be acquired by taking the *k*-smallest nontrivial eigenvalues to the eigenvector problem: $L\mathbf{y} = \lambda D\mathbf{y}$.

Chapter 1.2.2.1.3: Random Walk Approaches

Another class of embedding approaches relies on the use of random walk statistics. Among these, *DeepWalk* [41] has had a large influence, it used deep learning for the first time to learn embeddings of nodes by applying natural language processing (NLP) models to random walks along a graph. Specifically, DeepWalk first uses a random walk generator on a graph *G* to sample *N* random walks, each randomly initialized at some node $v \in \mathcal{V}$ with uniform probability across all nodes. Each random walk is permitted to continue for *t* steps. These random walks are then passed into a *SkipGram* language model which maximizes co-occurrence probability between words in a sentence. The key idea being that if nodes co-occur on sufficiently short random walks, they are likely in the same neighborhood and may share properties and therefore should have high similarity. This idea is shared by *node2vec*, another influential and popular random walk-based embedding strategy [42]. The key difference between the two strategies is their definition of random walks. *Node2vec* does not sample random walks in an unbiased manner, in fact it introduces two hyperparameters which adjust the tradeoff between breadth-first-sampling (BFS) and depth-first-sampling (DFS). BFS prioritizes exploration of the broad neighborhood around the initiated node, whereas DFS prioritizes a deep exploration of one path through related nodes.

Chapter 1.2.2.1.4: Structural Embeddings

Algorithms of the structural embedding class concern themselves with learning representations that preferentially capture the structural role of nodes rather than similarities in their global positions in the graph. This is motivated by the idea that in real world networks, nodes tend to have specific functions relative to the system. These functions can be determined by a variety of attributes, but structural embedding algorithms argue that these functional roles are to a large extent hardcoded into the network structure. For example, consider a communication network within a university wherein nodes represent individuals, and an edge represents an email or memo exchanged between two individuals. The topological structure of a node belonging to an esteemed PI is likely very different from that of a lowly graduate student. *Struc2vec* is an example of one such structural embedding algorithm [43]. The *struc2vec* algorithm is made up of four overarching steps in which: (i) the algorithm creates a structural similarity matrix measuring the similarity between each node pair in the graph at varying neighborhood sizes, inducing a hierarchical similarity measurement; (ii) weighted multilayer graphs are created wherein all nodes are represented in each layer but edge weights between nodes at each layer are inversely proportional to their similarity score for the corresponding neighborhood size; (iii) biased random walks are performed on the multilayer graphs to generate node sequences; (iv) random walks are then passed through a standard *skipgram* model to generate embedded representations of the nodes.

Another algorithm for learning structural embeddings is *GraphWave* [44]. This approach learns a continuous vector-valued structural embedding for each node through an unsupervised learning framework aided primarily by heat kernels adjoined with spectral graph wavelet decompositions. Briefly, let $\mathbf{L}$ be the Laplacian of a graph, then, *GraphWave* considers the heat kernelized spectrum of the Laplacian to be informative:

$$\mathbf{L} = \mathbf{U}\mathbf{G}(\mathbf{\Lambda})\mathbf{U}^T \quad (10)$$

where **U** is the matrix of eigenvectors of the Laplacian, **Λ** is the diagonal matrix of eigenvalues of the Laplacian, and **G(Λ)** is the diagonal matrix of heat kernelized eigenvalues, $\mathbf{G}(\mathbf{\Lambda}) = diag(g(\lambda_1), \ldots, g(\lambda_n))$, where $g(\lambda) = e^{-\lambda s}$. Specifically, the embeddings are characterized by the row or column vector of the Laplacian: $\psi_{v_i} = \mathbf{U}\mathbf{G}(\mathbf{\Lambda})\mathbf{U}^T\boldsymbol{v_i}$. Particularly compelling is GraphWave's mathematical guarantees on providing structure preserving embeddings. Consider a pair of nodes *a* and *b* which have identical *K*-hop neighborhoods, that is to say, there exists an injective and surjective mapping π between $\mathcal{N}(a)$ and $\mathcal{N}(b)$ $\forall v, v' \in \mathcal{N}(a), \mathcal{N}(b)$. Then, GraphWave guarantees that each coefficient of ψ_a is within 2ϵ of the corresponding coefficient of ψ_b: $|\psi_{ma} - \psi_{\pi(m)b}| \leq 2\epsilon$, where ϵ is some arbitrarily small error. The authors further extend this finding to structurally *similar* nodes, noting that

$$|\psi_{ma} - \tilde{\psi}_{ma}| \leq \left(\sum_k |\alpha_k| + 1 + C\right)\epsilon \quad (11)$$

where α_k are coefficients of the Stone-Weierstrass polynomial approximating the kernel g_s restricted to the interval $[0, \lambda_n]$, *C* is a constant which bounds the perturbed residual between $g_s(\lambda)$ and $polynomial(\lambda)$. Combined, these suggest that GraphWave can produce identical or nearly identical embeddings for nodes which are structurally identical or nearly identical, respectively.

Chapter 1.2.2.1.5: Node Level applications in large scale -omic analyses

Node embedding techniques have seen a variety of successes in biomedical applications. In their study, Shulte-Sasse et al [45] developed a graph convolutional network model to assess the cancer association of genes by combining various multi-omic pan-cancer data including sequence variations, copy number variations, methylation, and gene expression. They further used a layer-wise propagation mechanism to assess the main

contribution to a gene's prediction, e.g., whether a gene's association to cancer is driven primarily by network information or by omic data. In doing this, they identify a set of candidate genes which, interestingly, are not necessarily enriched for mutations but instead interact with core cancer genes. Turning to singe-cell -omic applications, in their study, Ravindra et al [46] made use of a graph attention network architecture to classify single cells as belonging to healthy individuals or multiple sclerosis patients. To do this, the authors created a k-nearest neighbor graph of distances computed in the PCA space of single cells, where each cell was featurized by its 22k gene expression values.

Chapter 1.2.2.2: Edge Level

An example of an edge level task is given in **Figure 2B.** In this type of task, the goal is to predict whether an interaction between two nodes exists. For example, given a network of known drug – protein interactions, we may want to predict additional, unknown interactions. Like the node level task, there are a variety of strategies to tackle this problem

Chapter 1.2.2.2.1: Local Similarity Indices

Local similarity algorithms measure a similarity metric for each pair of nodes that do not have an observed link [47]. Pairs of nodes which have the highest scores are then predicted to have an unobserved or future link. A wide variety of similarity metrics have been explored. One simple similarity metric examines the sharing of network neighbors between two nodes *x* and *y* [48]:

$$S_{xy}^{CN} = |\mathcal{N}(x) \cap \mathcal{N}(y)| \tag{12}$$

where $\mathcal{N}(\cdot)$ is the set of neighboring nodes. As an extension of this, other similarity metrics aim to look at local communities supposing that two nodes are more likely to be linked if their shared network neighbors are densely linked. To this end, Cannistraci et al [49] propose to weight the shared neighbor metric above with a measure of connectivity density:

$$S_{xy} = S_{xy}^{CN} \cdot \sum_{z \in \mathcal{N}(x) \cap \mathcal{N}(y)} \frac{\left|\mathcal{N}(z) \cap \left(\mathcal{N}(x) \cap \mathcal{N}(y)\right)\right|}{2} \quad (13)$$

These local community based metrics have been successfully used in a variety of settings including bipartite graphs [50].

Chapter 1.2.2.2.2: Probabilistic models

Probabilistic methods aim to develop a probability distribution over the graph. The models consider the relational schema for the graph and the probabilistic dependencies over the domain to estimate the probability of edges not already known to exist [51]. One example comes from Zhao et al [52], wherein the authors use a Bayesian probabilistic approach which builds a model over various node attributes which are encoded in a binary node feature vector. Another example, focused on time varying networks, comes from the work of Das et al [53]. Here, authors leverage Markov processes to model the time varying probabilities of potential edges over multiple time scales.

Chapter 1.2.2.2.3: Matrix Completion

Graph edges are naturally amenable to matrix representation, particularly by an adjacency matrix. As a result, a popular approach is to consider the task of link prediction to be closely related to matrix completion or factorization. Matrix completion approaches often deal with the supervised learning framework of link prediction by adopting an optimization scheme:

$$\min_{\theta} \frac{1}{|E|} \sum_{i,j \in E} \ell\left(G_{ij}, \hat{G}_{ij}(\theta)\right) + \Omega(\theta) \quad (14)$$

where *E* is the set of known edges, θ is the prediction model parameters, ℓ is the loss function, $\hat{G}$ is the model's prediction for edge i,j and Ω is a regularization term. Matrix completion

supposes that G can be factorized as $G \approx L(U\Lambda U^T)$ for some arbitrary $U \in \mathbb{R}^{Vxk}, \Lambda \in \mathbb{R}^{kxk}$, and link function L. The i-th row vector of U corresponding to node i is considered a latent representation of node i. Thus $\hat{G}_{ij}$ can be approximated with $L(u_i \Lambda u_j)$. Various studies use different link functions or regularization schemes to varying degrees of success. In their study Gonen et al [54] implement a Bayesian approach which combines matrix factorization with dimensionality reduction and show that are able to successfully predict drug-target interactions. In a similar study, Cobangolu et al [55] implement a probabilistic matrix factorization approach to perform drug-target interaction prediction on bipartite graphs.

Chapter 1.2.2.2.4: Embedding based approaches

Embedding based link prediction approaches leverage node embedding tools to first define a latent or low-dimensional representation of individual nodes. These approaches then measure the probability of a non observed edge existing between two nodes based on various measures such as cosine similarity. As a result, these approaches are largely no different than the methods covered in the previous section regarding node embeddings. For example, node2vec is often used to generate link prediction strategies in various settings such as predicting the probability that links disappear in the future [56] or dynamic link prediction [57].

Graph neural networks have also proven to be quite useful in link prediction tasks as well. For example, in their study Wu et al [58], the authors use a graph neural network architecture to analyze an interaction network of drugs, proteins, and virtual nodes called bridge nodes used to ensure connectivity of the graph. In a similar task, Zitnik et al [32] try to predict the side effects of drug-drug interactions by using a multimodal graph of protein-protein, drug-protein, and drug-drug interactions. Interactions between drugs and proteins or proteins and proteins were binary, however drug-drug interactions were annotated with a vector encoding the type of side effects seen in patients taking the drug pair. Their

methodology, Decagon, was built in two stages; (i) node embeddings for each of the drugs in the network were built using a graph convolutional network, (ii) a simple feed forward neural network took node embeddings a pair of drugs and predicted the polypharmacy side effect profile.

Chapter 1.2.2.3: Graph Level

A common graph level task is whole graph classification (**Figure 2C**). In this task, the goal is to accurately classify a full graph as belonging to one of two or more classes. Similar to node or edge level tasks, there are a number of different strategies that can be employed to classify graphs.

Chapter 1.2.2.3.1: Graph Kernels

Kernels have become a major tool in machine learning. They measure similarities between data points and certain landmarks (or other data points), thereby allowing the formation of complex and nonlinear decision boundaries. Extending this to graphs, we can define some kernel $k(\mathcal{G}_i, \mathcal{G}_j)$ to measure the similarity between two graphs $\mathcal{G}_i$ and $\mathcal{G}_j$. A commonly used kernel type are random walk kernels [59], which measure similarity as simply the number of walks that are common to two graphs. Another class of kernels are the Weisfeiler-Lehman (WL) kernels which are based on the heuristic WL algorithm for assessing graph isomorphism. These types of graph kernels assess graph similarities based on neighborhood aggregation schemes which count the shared subtrees between the graphs [60]. Another class of kernels use graphlets which are induced subgraphs of the full graph which are non-isomorphic to each other. The graphs $\mathcal{G}_i$ and $\mathcal{G}_j$ are then represented by vectors of graphlet occurrence frequencies $f_{\mathcal{G}_i}$ and $f_{\mathcal{G}_j}$. The graphlet kernel can then be defined by the inner product of the two frequency vectors. Graph kernels have been successful in a number of bioinformatic applications. For example, in their study, Mautner et al [61] developed a methodology that

combines graph kernels and machine learning to predict the secondary structure of mRNA from sequence information. In another study, Tepeli et al [62] develop a graph kernel based clustering approach to cluster patients using multiomic data and pathway knowledgebases. In their clustering analyses they find that their methodology clusters patients such that they have significantly different survival times.

Chapter 1.2.2.3.2: Graph Embedding via Node Embeddings

Recently, many successful studies have adopted the use of graph embeddings. Much like node embeddings, the goal of graph embedding is to identify a low-dimensional representation which preserves the information and properties of the original graph. These embeddings are then used in downstream analyses to learn and perform classification. In fact, graph embeddings are often a composite of their constituent node embeddings. As a result, all the node embedding approaches discussed previously can be directly extended to perform graph classification. This includes matrix factorization tools like Laplacian eigenmaps or node proximity matrices, random walk based node embeddings, or even deep learning and graph learning based node embeddings. Extension of these node embeddings to perform graph embedding is commonly done by simply modifying the standard loss function to:

$$\mathcal{L} = \sum_{\mathcal{G}_i \in \mathbb{D}} \left\| \mathbf{MLP}\left(\boldsymbol{z}_{\mathcal{G}_i}\right) - y_{\mathcal{G}_i} \right\|_2^2 \tag{15}$$

where **MLP** is some multilayer perceptron (or other function) which combines all of the node embeddings $\boldsymbol{z}$ in a graph $\mathcal{G}_i$ into a label prediction, and $y_{\mathcal{G}_i}$ is the ground truth label. These GNN based graph classification approaches have been successful in a variety of bioinformatic applications. For example, Duvenaud et al [63] used GNNs to develop an end-to-end architecture to learn and predict properties of molecules such as their solubility or likely drug efficacy. In another study, Li et al [64] propose a variant of graph convolutional neural

networks to predict properties of molecular graphs. To do this, they develop adaptive graph convolutional architectures by using an 'adaptive' Laplacian given by the general graph Laplacian and a residual Laplacian calculated on a set of 'virtual edges'. They further demonstrate that this adaptive GCN architecture is effective at predicting properties of molecule structures such as solubility and hydration-free energy.

Chapter 1.3: Quantification of Effects of Coding Mutations

While there is a great redundancy in the genetic code which permits biological organisms to tolerate point (missense) mutations to some degree, these mutations can also lead to perturbations in protein shape and function. In fact, these functional perturbations caused by missense mutations can often cause disease. As a result, understanding and predicting the degree to which critical protein functions are disrupted by mutations has been a focus of much research. Many tools have been developed as potential solutions to this problem. Some approaches make predictions based on altered protein stability while others rely on machine learning trained over various features such as phylogenetic sequence conservation, physiochemical properties, homology, population frequency, among many others [65–69]. Another popular strategy is ensemble or consensus methods wherein machine learning techniques are used to weight predictions from various other tools to combine into a 'meta-score' of protein functional perturbation[70,71].

Evolutionary Action (EA) introduced first by Katsonis and Lichtarge in 2014 [69], offers an alternative approach to the variant impact prediction problem by defining a formal genotype-phenotype equation. EA expresses that the genotype-phenotype relationship can be written as $f(\gamma) = \phi$, where evolutionary fitness function (f) maps genotype (γ) onto fitness landscape (ϕ). EA theory then considers any single nucleotide sequence variation (SNV) to be infinitesimally small perturbation in the full genome sequence $(d\gamma)$, which causes an equally small perturbation to the fitness phenotype $(d\phi)$. This relationship is described by: $\nabla f \cdot$

$d\gamma = d\phi$. Assuming a mutual independence of all residues, then for any SNV at amino acid residue r_j, $d\gamma \approx \Delta r_j$, all components of the ∇f will go to zero except $\frac{\partial f}{\partial \gamma_{r_j}}$, leaving: $\frac{df}{d\gamma_{r_j}} \cdot d\gamma_{r_j} \approx d\phi$. The authors demonstrate that all components on the left-hand side of this equation are calculable using the Evolutionary Trace [72,73] algorithm and amino acid substitution log-odds ratios. The resulting estimated perturbation in fitness phenotype is the EA score of a mutation. EA scores are normalized to a range of 0-1 where 0 indicates likely no effect to protein function and 1 indicates a likely total loss of protein function. Interestingly, this formulation of the genotype-phenotype relationship harbors an interesting interpretation of mutations from the perspective of statistical thermodynamics. This idea will be further expanded upon in this thesis.

PolyPhen2 (PPh2) is a well-known tool [67] which uses a naïve-bayes classifier to predict variant impacts. The algorithm is trained on thirty-two predictive features over which the authors implemented a greedy algorithm to iteratively identify the most useful features for impact prediction. Training was done using the HumVar and HumDiv databases. The final trained algorithm uses a set of eight sequence-based features and three structure-based features, with most selected features involving a comparison of some property of wild-type allele to mutant allele.

Combined Annotation-Dependent Depletion (CADD) is another well-known and commonly used tool which uses the consensus or ensemble method strategy [70]. CADD integrates several different annotation and impact scoring systems such as conservation metrics, regulatory information, transcript information, and protein-level scores such as PPh2. Authors then trained a support vector machine with a linear kernel on all 8.6 billion possible SNVs in the human reference genome to produce CADD scores.

Sorting Intolerant from Tolerant (SIFT) algorithm uses sequence homology to estimate the likelihood that any amino acid substitution due to missense mutation will have a detrimental effect on protein function [68]. Methodologically, the algorithm first searches for proteins with similar sequences to a query protein. It then chooses sequences that may share similar functions to the query protein by searching for sequences that have > 90% similarity. Multiple sequence alignments are then performed on the selected sequences and finally, normalized probabilities are calculated for all possible substitutions at any given position along the alignment.

Chapter 1.4: Dissertation Objectives

Advancements in graph learning coupled with the ingenuity of quantitative variant impact predictions have motivated the objective of this dissertation to **develop tools to help gauge the involvement of genes in disease in a manner inclusive and cognizant of their interactive environment, mutational functional perturbation, and relative mutational intolerance.**

To this end, we will begin by leveraging a guiding principle in the study of node embeddings that embeddings should be subject to the restraint that their relative positions in an embedding space must be reflective of the node's original properties. Moreover, a properly defined metric on the embedding space should accurately reflect similarities and differences in the original nodes. If two nodes are close together in the embedding space, it should be because the original nodes are very similar. If two nodes are far apart in an embedding space, it should be because the original nodes are very different. **We can then hypothesize, from this principle, that differences between a gene's embedding in a healthy, wild-type biological network and its embedding in a diseased network should be reflective of its role in disease biology**. If the two embeddings are highly similar, then the gene's function may not be perturbed in disease. If the gene's embedding in disease and healthy are

dissimilar, then the gene may have altered function in the disease, potentially contributing to disease pathology.

Next, we will consider a limitation of many successful variant impact methods: functional perturbations are often estimated relative to the protein rather than the organism. To illustrate this, consider two genes *GATA2* and an olfactory receptor, both with truncating mutations in the main bodies of the genes. Most variant impact predictors will estimate total loss of function for the encoded proteins of both genes, however the resulting perturbation to the overall biology of the system could be very different. *GATA2* is a haploinsufficient gene in which loss of function (LoF) mutations lead to the immunodeficiency syndrome MonoMAC [74]. Conversely, LoF mutations are enriched in olfactory receptor genes suggesting that they are unlikely to significantly perturb the overall organism [75]. Compelled by these disparate effects of genes at the organism level, **we will leverage aspects of statistical thermodynamics to quantify the relative mutational intolerance of all genes across the genome.**

Taken together, these efforts will provide valuable insights into the biological dynamics of gene networks, new tools with which to investigate disease-gene associations and offer alternative ways to think about and probe genetic systems.

Chapter 2: Gene Embedding Provides Novel Insights into Disease Mechanisms

*This chapter is based upon and consists of text written in the following paper, which is published in **Cell Genomics**:*

Identification of Risk Genes for Alzheimer's Disease by Gene Embedding.

Yashwanth Lagisetty[1,2], Thomas Bourquard[2], Ismael Al-Ramahi[2,3,4], Carl Grant Mangleburg[2], Samantha Mota[2], Shirin Soleimani[2], Joshua M. Shulman[2,3,4,5,6], Juan Botas[2,3,4], Kwanghyuk Lee[2], Olivier Lichtarge* [2,4,7]

[1]Department of Biology and Pharmacology, UTHealth McGovern Medical School, Houston, TX
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030
[3]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX 77030
[4]Center for Alzheimer's and Neurodegenerative Diseases, Baylor College of Medicine, Houston, TX, 77030
[5]Department of Neurology, Baylor College of Medicine, Houston, TX 77030
[6]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030
[7]Computational & Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX 77030

Lagisetty Y., Bourquard T., Al-Ramahi I., Mangleburg C.G., Mota S., Soleimani S., Shulman J.M., Botas J., Lee K., Lichtarge O., Identification of Risk genes for Alzheimer's disease by gene embedding, Cell Genomics (2022), https://doi.org/10.1016/j.xgen.2022.100162

Chapter 2.1: ABSTRACT

Most standard disease-gene association methods do not account for gene-gene interactions. This is crucial for complex polygenic diseases like Alzheimer's disease (AD), where our understanding of the genetics and mechanisms remains incomplete. To discover new disease genes whose interactions may contribute to pathology we developed GeneEMBED, a general approach applicable to exome and genome sequences, that models the combined functional impact of coding variants in different genes across a biological network. In this manner, GeneEMBED assesses the mutational perturbation of a gene's neighborhood in disease compared to healthy individuals. In two independent AD cohorts of 5,169 exomes and 969 genomes, in addition to known AD genes, GeneEMBED identifies many novel candidates that are differentially expressed in post-mortem AD brains, whose modulation in mice causes neurological phenotypes and that are interactors of known AD genes. Four genes stood out among the novel candidates since they are upregulated in AD brains and modified neurodegeneration *in vivo*: *PLEC, UTRN, TP53* and *POLD1.* Importantly, *TP53* and *POLD1* are involved in DNA break repair and are targeted by available pharmacological inhibitors. While these data show proof-of-concept in AD, GeneEMBED remains a general approach that can identify genes relevant to risk mechanisms and therapy of other complex diseases.

Chapter 2.2: Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder characterized by progressive memory loss, language deficits, and behavioral abnormalities [76]. An estimated 6 million individuals in the US are afflicted with AD and this number is projected to double by 2050 [77]. The polygenic nature of AD presents an obstacle to early diagnosis and risk prediction. In late onset Alzheimer's disease (LOAD), the estimated genetic heritability is 60-80% [7,8]. Though genome-wide association studies (GWAS) have identified > 40 LOAD loci [78–82], they account for only a fraction (~33%) of the heritability [9,10]. While there are many explanations for this "missing heritability" problem [11,12,83], which is seen across complex diseases [6], an attractive hypothesis suggests that genetic interactions may be a culprit [18]. Genetic interactions are functional interactions observed among gene variants where the resulting phenotype differs from the independent phenotype of each variant [18,19]. Thus, relatively benign mutations may combine to generate complex phenotypes. Indeed, such non-additive genetic interactions have been observed in disease [21–23] and have improved current models of the genotype-phenotype relationship [20,84]. However, genome-wide discovery of pairwise genetic interactions presents major challenges. Theoretical analysis suggests that under reasonable assumptions nearly 500,000 samples would be needed to identify statistically significant genetic interactions [18]. The potential use of prior knowledge to compensate for necessary sample size has motivated the development of network informed gene prioritization methods for various diseases [24,26–29,85]. These approaches do not typically use patient specific genetic data. However, when they do, they often rely on expression data (e.g. HIT'nDRIVE) [28], or they are built for somatic mutations (e.g. HotNet2) [29] and are not immediately amenable to the case-control study designs typical of germ line genome-wide association studies.

Advances in graph representation learning open new opportunities to analyze genomes in the context of biological networks. Graph learning techniques have been successful in a variety of biological applications including predicting protein-protein interactions [86–90] and drug responses or side-effects [91–94]. Specifically, node embedding enables machine learning on networks by compacting the qualitative and quantitative properties of a network node in a mathematically suitable framework. For example, Deep Walk [41] and Node2Vec [42] use random walk algorithms to represent nodes as vectors. Alternatively, Graph Convolutional Networks [36] or Graph Attention Networks [38] use graph neural network architectures to construct node representations instead. Regardless of the approach, node embeddings should conserve the relative properties between original graph nodes, meaning that similar nodes should embed similarly. We hypothesize, based on this principle, that differences in a gene's embedding in a disease network compared to its embedding in a healthy network may reflect a role in disease pathology.

This motivated us to develop GeneEMBED (gene embedding based evaluation of disease-gene relevance) to pinpoint genetic risk factors of disease by examining the differential perturbation patterns of gene interactions. The approach takes a predefined molecular network and annotates it with the functional impact of protein coding variants across cases, and separately controls. Importantly, the approach considers all protein coding variants in estimating gene-level perturbed protein function. Machine learning performs embeddings on each network and then finds which genes have the most difference in case versus controls embeddings. Notably, this approach addresses the limitations of standard models by feasibly assessing the contribution of pairwise, and higher order, genetic interactions on disease and doing so with a case-control study design of typical genome-wide studies. While this approach is general and applicable to many complex diseases, we tested this in two LOAD data sets: the Alzheimer Disease Sequencing Project (ADSP) (dbGaP phs000572.v7.p4) Discovery

cohort comprising 2,729 affected (AD+) individuals and 2,440 healthy (AD-) controls, and the Extension cohort with 481 AD+ and 488 AD- individuals (NIGADS NG00067). To assess robustness of GeneEMBED, we used two variant impact scoring methods, Evolutionary Action (EA) [69] and PolyPhen2 (PPh2) [67], and we tested three different molecular interaction networks, STRING [95], HINT [96], and a brain specific network [97,98]. The candidate genes from the Discovery and Extension cohorts were consistent with one another and with known AD genes. The candidates interacted with manually curated AD-associated genes and were dysregulated in AD brains. Functional *in-silico* analysis showed they were involved in pathways relevant to AD, including for cell cycle and DNA replication. *In vivo* perturbation analysis confirmed that GeneEMBED genes were modifiers of tau and β-amyloid induced phenotypes in well-established *Drosophila* AD models [99–101] and their modulation in mice showed abnormal neurological phenotypes, supporting their role in normal neuronal maintenance and function. Importantly, many GeneEMBED candidates are druggable with already approved compounds. Overall, these results point to new targets for therapeutic development in AD, and broadly support a novel and general paradigm to interrogate other complex genetic diseases.

Chapter 2.3: RESULTS

Chapter 2.3.1: GeneEMBED identifies genes that are perturbed in AD

With a view to discover AD genes, GeneEMBED aims to combine the integrative features of network biology with machine learning to find genes with functional interactions perturbed differently among cases and controls, due to mutations. First, GeneEMBED builds a personalized functional impact network by calculating a perturbation score (PS) for each gene of each subject of a cohort. This score reflects all non-synonymous variants in the gene

(*v*); the impact of each variant estimated by either EA [69] or PPh2 [67] (Variant Impact Score^{EA} and VIS^{PPh2}, respectively); and zygosity (*zyg*) (**Figure 3A** and methods). The PS scores are then mapped to a gene network of choice, such as the STRING protein-protein interaction network, by setting the weight of an existing edge between two genes as the sum of their PS score. Finally, the edge weights are averaged across all cases, or separately across all controls, to produce two global cohort networks that compile the aggregate mutational perturbations of protein-protein interactions in cases and in controls. Both networks are then processed with the GraphWave [44] machine learning algorithm which applies an unsupervised diffusion-aided wavelet decomposition to assign a continuous vector-valued embedding to each gene/node. This embedding is based on the topological (geometric distribution of the edges in the node's vicinity) and functional (functional information associated to each edge) properties surrounding the gene in the network. As a result, the vector assigned to each gene represents the integrated functional perturbation of the variants in its network neighborhood. The final step applies principal component analysis (PCA) to identify vectors with significant differences between the case and control networks (FDR<0.01), suggesting distinct perturbation patterns in these genes between AD vs controls.

Next, to test the algorithm and identify genetic factors underlying AD, we applied GeneEMBED to the whole exome (WES) and whole genome sequencing (WGS) data from the ADSP Discovery and Extension cohorts, respectively, using either *VIS*^{EA} or *VIS*^{PPh2} for the variant impact score, and initially the STRING protein-protein interaction network. Additionally, we applied GeneEMBED to healthy control vs healthy control using both *VIS*^{EA} and *VIS*^{PPh2} to identify potential false positive (FP) genes. After removal of FPs, GeneEMBED identified 69 AD-candidates in the Discovery Cohort and 119 candidates in the Extension cohort with *VIS*^{EA}, and 128 candidates in the Discovery Cohort and 120 genes in the Extension cohort (**Table 1**) with *VIS*^{PPh2}. Fourteen genes overlapped between the Discovery and Extension cohorts when

using *VIS*^{EA} (hypergeometric p ~ 1.86e-16). Of these, nine genes had evidence in literature documenting their association with AD (**Figure 3B**, *APOE, CSF1R, ILR4, MAPK6, MAPT, REST, RIPK4, SP3,* and *TRIB3*) [102–110]. Particularly notable were *MAPT* and *APOE.* Neurofibrillary tangles, one of the primary AD biomarkers, are aggregates of hyperphosphorylated *MAPT* gene products [111]. *APOE*, on the other hand, is one of the strongest genetic predictors of AD [111]. Similarly, 16 genes overlapped between Discovery and Extension cohorts when using *VIS*^{PPh2} (hypergeometric p ~4.25e-15), of which six have been previously linked to AD pathology (**Figure 3B**, *CCT5, ERBB2, MAPK6, REST, SYNJ1,* and *TP53*) [105,107,112–115]. GeneEMBED-*VIS*^{PPh2} did not recover *APOE* in the Discovery cohort but did so in the Extension. GeneEMBED also identified well known genes in which rare variants are associated with AD, including *TREM2* [116] and *SORL1* [117], though these genes are recovered only in the Discovery cohort. Comparing *VIS*^{EA} to *VIS*^{PPh2}, 34 genes overlapped in the Discovery cohort (hypergeometric p ~1.46e-53) and 44 genes overlapped in the Extension network (hypergeometric p ~ 2.46e-64), indicating concordance between these two impact scores. Lastly, we found that 4 genes overlapped among all cohort-VIS combinations with a hypergeometric p-value ~ 8.58e-10. These data suggest that GeneEMBED is robust to inter-cohort variability as well as differences in impact scoring systems and can recover several well-characterized, positive control AD genes.

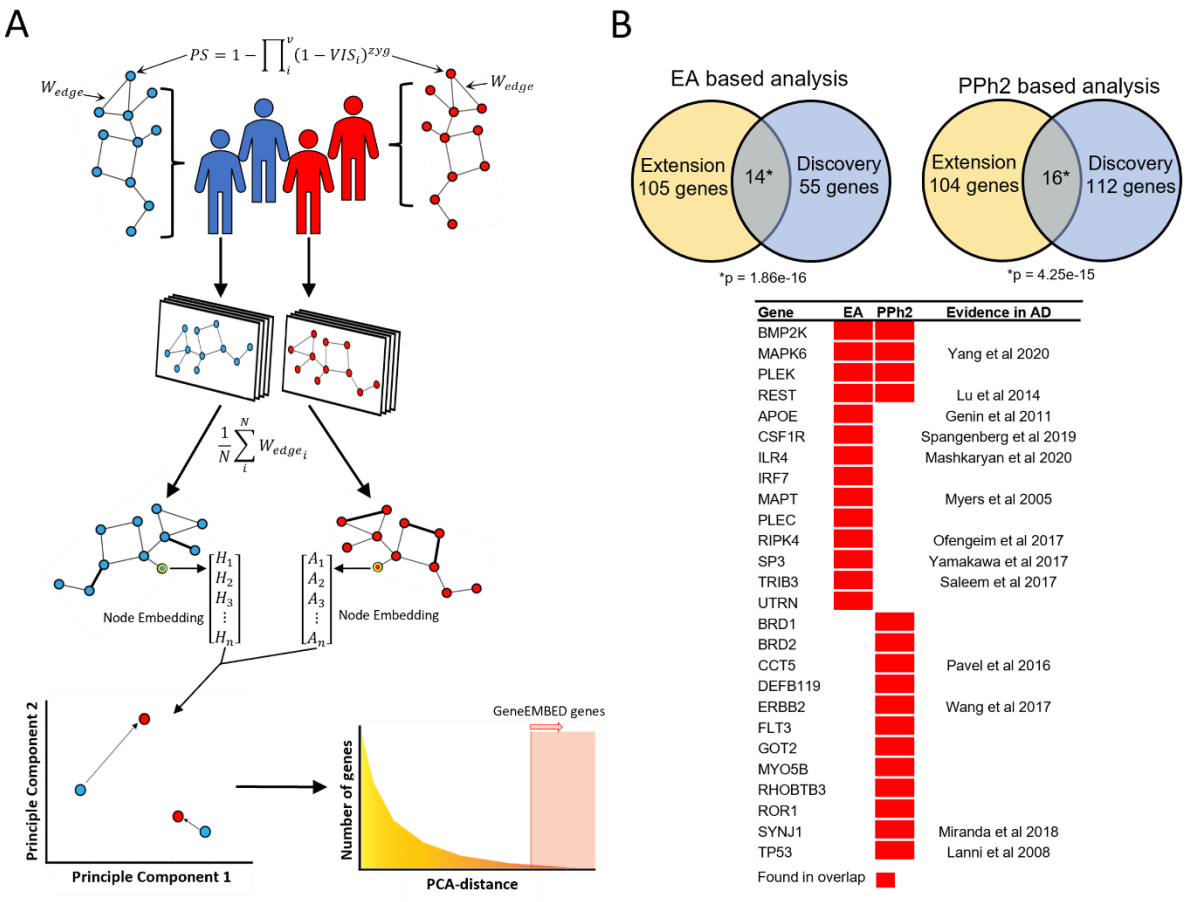


Figure 3: Overview of GeneEMBED and AD candidate genes. (A) GeneEMBED: For an individual, genes are first assigned a perturbation score (PS) consolidating information from all the gene's variants appearing in the individual. The gene PS estimates the total loss of function probability given various combinations of variant level loss of function probabilities. Edge weights for an individual's network are calculated by the sum of the PS of the connected genes. Edge weights are then averaged over to construct one case specific and one control specific graph. Node embedding is performed on the genes in the two networks. Finally, embeddings are projected in a PCA space to measure distances between nodes in case and control networks. (B) GeneEMBED using EA identified 69 candidate genes in Discovery and 119 in Extension with 14 overlapping genes, significant by one-tailed hypergeometric test. In PPh2 analyses, 128 candidate genes were found in Discovery and 120 in Extension with 16 overlapping genes, significant by one-tailed hypergeometric test. A large portion of overlapping genes have been previously implicated in AD biology.

In order to control against a standard method for inferring gene-disease associations, we used MAGMA (Multi-marker Analysis of GenoMic Annotation) which prioritizes genes based on multiple regression analysis. This method can be performed genome-wide, allowing it to be used for gene discovery. [16]. MAGMA identified 31 AD-associated genes in the Discovery cohort ($p < 0.001$) and only 7 in the Extension, with no overlap (**Table 2**). MAGMA in the Discovery cohort shared only *APOE* with both GeneEMBED-VIS^{EA} analyses and GeneEMBED-VIS^{PPh2} in Extension while overlapping with VIS^{PPh2}analysis in Discovery cohort by two genes *SORL1* and *PRIM1*. Similarly, MAGMA in Extension only shared *TPO* with VIS^{PPh2} in Discovery and did not overlap with any other analyses. Of the 31 MAGMA candidates from the Discovery cohort 9 had been previously associated with AD including *APOE* and *TOMM40* [102,118]. This indicates that MAGMA was less effective and less reproducible at this small sample size.

To assess the recovery of GeneEMBED in a systematic manner, we measured hypergeometric overlaps between GeneEMBED candidates and 208 AD-associated genes in the DisGeNet database. DisGeNet compiles gene-disease associations based on genetic, clinical, and animal model curation [119]. We found a significant overlap ($p=0.012 – 5.3e-4$) between GeneEMBED candidates and DisGeNet genes across functional mutational impact methods (VIS^{EA} vs VIS^{PPh2}) and cohorts (Discovery vs Extension, **Table 3**). Additionally, we found significant overlaps between AD-associated genes from the comparative toxicogenomic database (CTD) [120] and GeneEMBED-VIS^{EA} in both the Discovery ($p = 0.047$) and Extension ($p = 3.3e-3$) cohorts. MAGMA candidates in the Discovery cohort recovered similar significant overlaps (**Table 3**, the low number of MAGMA candidates in the Extension cohort prevented a similar analysis). These data suggest that GeneEMBED is able to significantly recover several known AD genes despite large differences in cohort sizes. Moreover, MAGMA was unable to reproducibly retrieve genes between the Discovery and Extension

cohorts while GeneEMBED found significant overlaps. Taken together, these findings demonstrate the robustness of GeneEMBED, compared to MAGMA, to both inter-cohort variability and sample size. Overall, GeneEMBED identifies candidates distinct from MAGMA which are nonetheless enriched for known AD-associated genes, suggesting an identification of disease relevant signal.

Chapter 2.3.2: GeneEMBED CANDIDATES ARE ROBUSTLY CONNECTED AND RELEVANT TO AD

To assess the role of GeneEMBED candidates, we asked if they are implicated in molecular changes related to AD, specifically, dysregulated gene expression as tallied by the Accelerating Medicines Partnership Alzheimer Disease (AMP-AD) RNA-sequencing from seven brain regions [121–126]. To focus on novel genes, we removed GeneEMBED genes that overlapped with any of five curated AD gene sets (DisGeNet, CTD, ClinVar, GWAS Meta 1, GWAS Meta 2 [78,79,119,120,127]). The remainder was significantly dysregulated in the Temporal Cortex of AD patients (TCX, $p<0.05$**, Figure 4A**), independent of both the functional impact method (VIS^{EA} vs VIS^{PPh2}) and the cohort (Discovery vs Extension). However, GeneEMBED-VIS^{EA} candidates were also dysregulated in the Parahippocampal Gyrus (PHG, **Figure 4A**) for both cohorts, and in the cerebellum (CBE), frontal pole (FP), superior temporal gyrus (STG), and dorsolateral prefrontal cortex (DLPFC) ($p<0.05$ **Figure 4A**) for the Extension cohort, whereas that was only true for GeneEMBED-VIS^{PPh2} on the CBE, also in the Extension cohort. MAGMA, in contrast, found no enrichment in dysregulated genes. Secondarily, the number of brain regions with significant dysregulation of candidate genes for GeneEMBED-VIS^{EA} in the Discovery cohort and GeneEMBED-VIS^{PPh2} in the Extension cohort was on par with the number from two AD GWAS meta-analyses (p~ 1.2e-2, p~2.3e-3, pGWAS Meta 1 ~ 2.8e-3 & pGWAS Meta 2 ~4.9e-3, respectively). Remarkably, GeneEMBED-VIS^{EA} applied to the Extension cohort identified candidates significantly dysregulated in 6 brain regions in AD

(p~1.6e-13) (**Figure 4B**). Together, these data indicate a strong link between this group of candidate genes and AD pathology. This link, however, could be either causative or responsive.

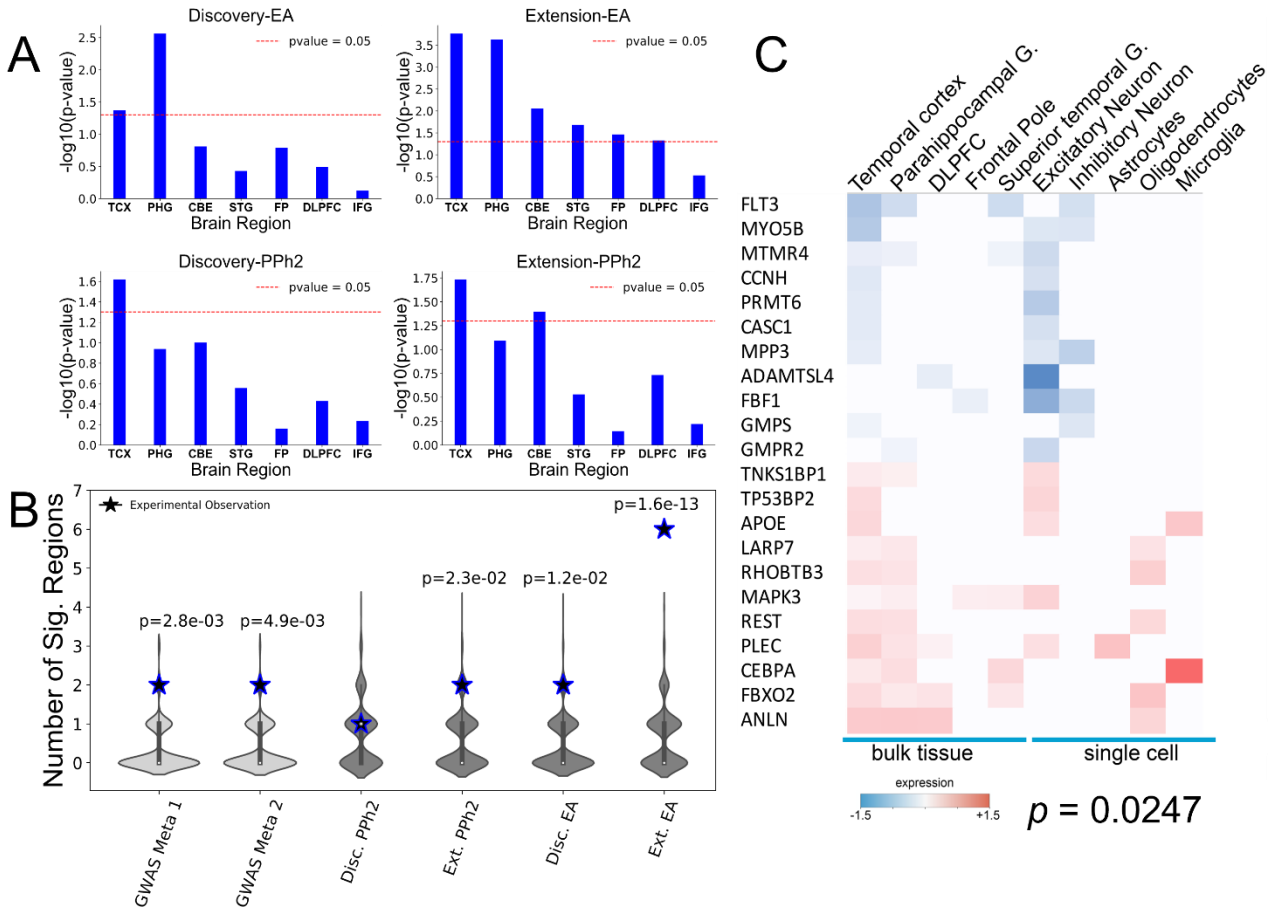


Figure 4: GeneEMBED candidates are differentially expressed in AD brain tissue. (A) One-tailed hypergeometric enrichment of GeneEMBED candidates against differentially expressed genes from seven brain regions: cerebellum (CBE), temporal cortex (TCX), frontal pole (FP), inferior frontal gyrus (IFG), parahippocampal gyrus (PHG), superior temporal cortex (STG), and dorsolateral prefrontal cortex (DLPFC). (B) Comparison of RNA sequencing-based enrichment between known AD gene sets and GeneEMBED candidates. Stars indicates the number of brain regions with significant enrichment in each gene set by permutation testing. Violin plot shows the distribution of expected number of enriched brain regions when using random gene sets. (C) Among the 143 high-confidence genes, a significant number (22, one-tailed Fishers Exact Test p=0.0247) showed differential expression in both bulk tissue from various brain regions and in single cell sequencing of neuronal cell types.

Next, we tested whether novel GeneEMBED candidates were connected to AD-reference gene sets. For this, we measured how well information propagated between them and AD-associated genes in a PPI network [128–130] using the nDiffusion method [131]. Area under a receiver-operator curve (AUROC) measures the strength of their interaction and a *z*-score is calculated for the significance of the observed AUROC compared to a distribution of random gene sets. We used two disease-gene association databases (DisGeNet - 208 genes [119] and CTD - 103 genes [120]) and three variant-based reference gene sets for AD (GWAS Meta 1 - 25 genes [78], GWAS Meta 2 - 38 genes [79] and ClinVar - 21 genes [127]). The GeneEMBED candidates showed statistically significant diffusion (ROC>0.5 + z-score>2) to most selected AD-associated gene sets, regardless of the cohort (**Figure 5, Table 4**, AUROC = 0.63 – 0.84, Z = 2.03 – 5.64). Interestingly, MAGMA candidates also diffused significantly to DisGeNet, CTD, and ClinVar but not to the two GWAS datasets. These data suggest that the GeneEMBED candidates are functionally and significantly connected to previously curated AD-associated genes, further suggesting an identification of disease relevant signal.

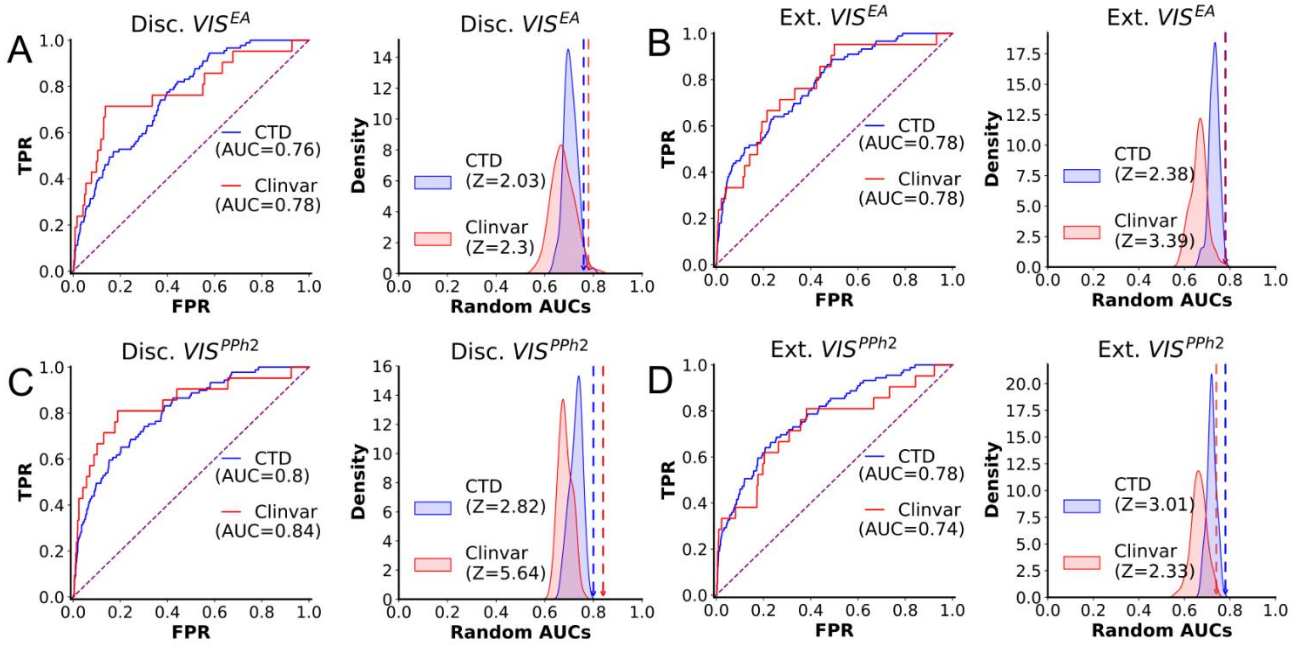


Figure 5: GeneEMBED candidates are significantly related to curated sets of AD genes. (A) Receiver Operator Characteristic curves are shown for Disc. VIS^{EA} for network diffusion to CTD and ClinVar AD gene sets. To determine significance of observed AUC, a permutation testing strategy is used wherein random gene sets of the same size are generated 100 times and analyzed through nDiffusion to create a random distribution of AUCs. Reported z-scores are calculated relative to these backgrounds. Y-axis of the ROC plots are true positive rates (TPR) and x-axis is false positive rate (FPR). Similarly, y-axis of the z-score distribution is probability density and x-axis is the AUROC score of random gene sets. Analogous plots are shown for (B) Ext VIS^{EA} , (C) Disc VIS^{PPh2} , and (D) Ext VIS^{PPh2}

To test GeneEMBED's utility and robustness in alternate protein-protein interaction (PPI) networks, we replicated the experiments from the above sections using the HINT network [96] of curated high-quality PPIs and a second network of physical PPIs specific to brain tissue [97,98]. First, using the HINT network, only VIS^{EA} in Discovery showed significant recall of genes from the CTD, GWAS Meta 1, and GWAS Meta 2 references (p = 0.0014, 0.0058, 0.015) (**Tables 5, 6**). However, nDiffusion found both VIS^{EA} and VIS^{PPh2} in Disc were significantly connected to all curated gene sets except GWAS Meta 1, with AUROCs = 0.62-0.77 (z = 2.31-5.77) and AUROCs = 0.62-0.76 (z = 2.6-3.89) (**Table 7**), respectively. VIS^{EA} and VIS^{PPh2} in Extension also had significant network connectivity with CTD and DisGeNet gene lists with AUROCs = 0.75, 0.7 (z = 3.33, 5.16) and AUROCs = 0.74, 0.67 (z = 3.32, 4.91). Alternately, using the brain specific PPI, both VIS^{EA} and VIS^{PPh2} in Discovery had significant interactions to the curated gene sets, with AUROCs = 0.63-0.78 (z = 2.11-3.91) and AUROCs = 0.64-0.82 (z = 2.43-6.07) (**Table 8-10**). VIS^{EA} in Extension found significant relatedness to CTD and DisGeNet with AUROCs = 0.77, 0.69 (z = 4.64, 5.07). VIS^{PPh2} in Extension did not show any significant links to the curated gene sets. These data show that GeneEMBED robustly identifies genes enriched for functional interactions to curated sets of AD related genes using a variety of alternative PPI networks. Interestingly, a large number of genes were repeatedly identified among two or more GeneEMBED analyses across cohorts, *VIS* systems, and PPI networks (**Figure 6**), suggesting a potential role in AD.

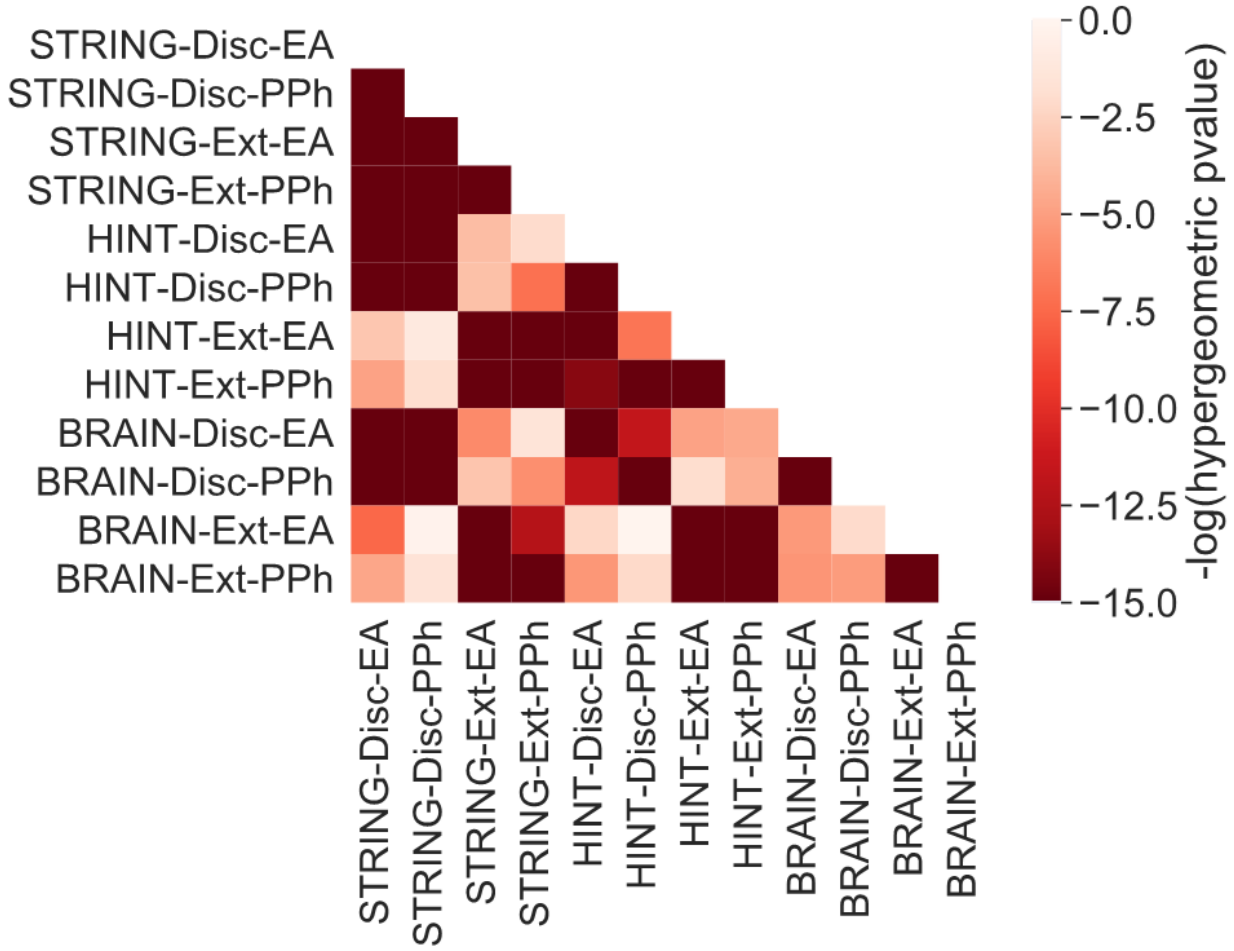


Figure 6: GeneEMBED candidates are consistently identified across various cohorts, networks, and VIS systems. One-tailed hypergeometric overlap tests were done on every pairwise combination of cohort-network-VIS experiments. Among 66 independent pairwise tests, only 11 did not demonstrate statistically significant hypergeometric p-values ($p < 0.05$, $\log(p) < -2.99$).

Chapter 2.3.3: GeneEMBED candidates are functionally connected and enriched for *in vivo* modulators of neuronal dysfunction triggered by tau and β-amyloid

The significant overlap in GeneEMBED candidate genes observed across cohorts and networks (**Figure 6**) indicates that GeneEMBED may be identifying specific pathways where an increased concentration of mutational load modulates AD risk. To investigate this, we performed functional enrichment analysis. We constructed a network in STRING with 143 high confidence hits. These genes were selected using the criteria that they must have been identified at least twice in the same network either across cohorts or across *VIS* methods. Genes were prioritized based on the degree of overlap across networks with more recurrent genes ranking higher, provided that they were never identified in any of the healthy control vs healthy control assays **(Figure 6).** Interestingly, this network showed significant PPI enrichment (*p*-value = 9.56e-07). After clustering with a Louvain algorithm, 127 of the 143 candidate genes mapped to significantly enriched pathways (**Figure 7**), including among others: (1) mechanisms involved in glial biology (Glial cell derived neurotrophic factor receptor) [132,133]; (2) inflammation (Regulation of IP-10 production, positive regulation of TGFβ1 production, chemokine signaling), which is known to be dysregulated in AD [111]; (3) clearance of protein aggregates (regulation of aggrephagy, MTOR signaling); and (4) extracellular signaling cascades. These cascades involved Wnt/β-catenin, G-alpha or ErbB which are dysregulated in AD [134,135] and modulate neurodegeneration in animal models [136], or Syndecan-3, which may play a role in tau and β-amyloid internalization [137]. (5) The largest functional module among the high confidence GeneEMBED candidates is related to DNA-double strand break repair. Interestingly, genes involved in double strand break repair regulation modulate neurodegeneration in animal models [109] and others involved in DNA quality control accumulate in AD brains.

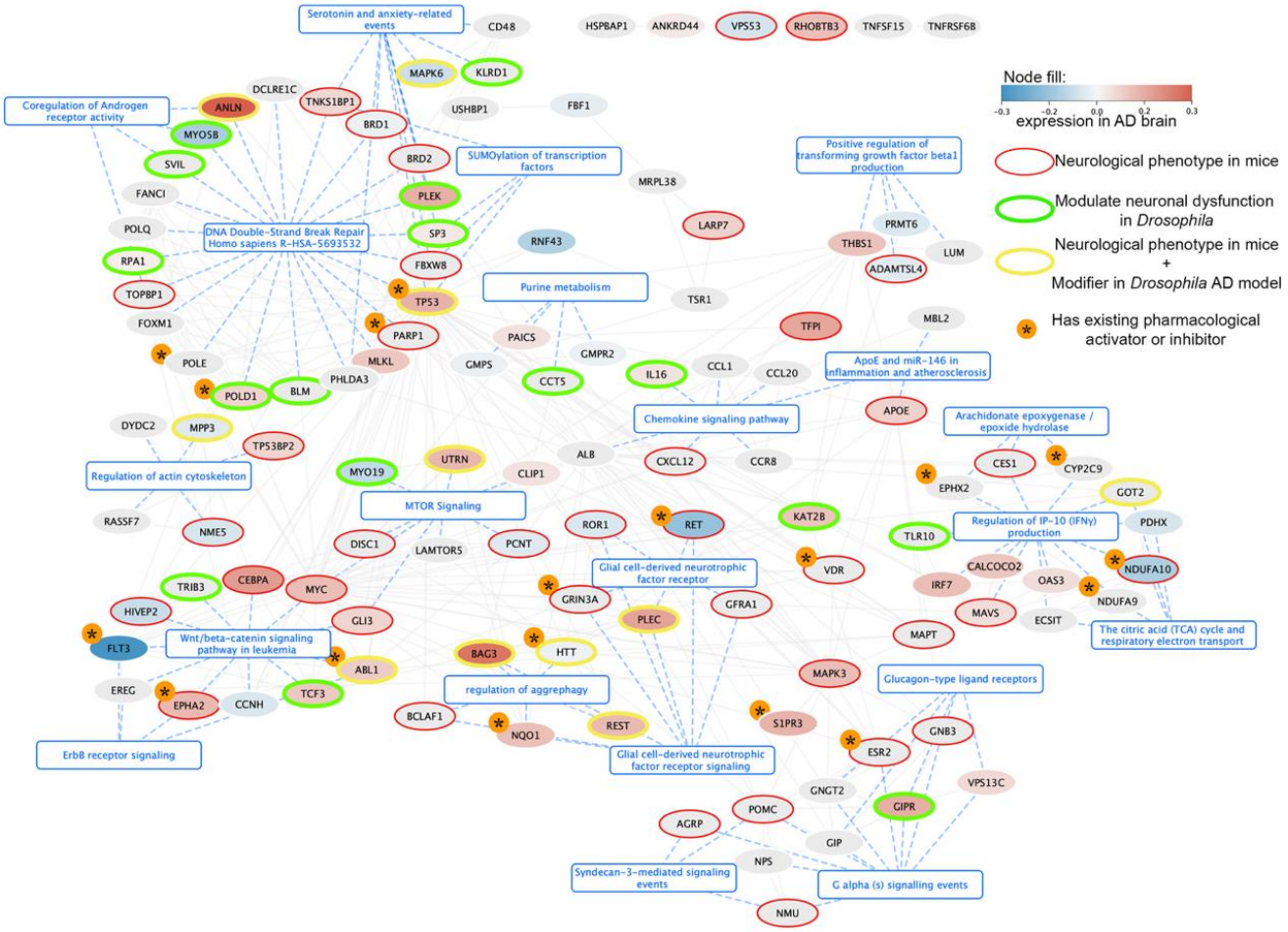


Figure 7: Interaction network among 143 high confidence genes is built using STRING edges. Nodes are colored based on their differential log2 Fold Change expression in AD brains. Red ring around genes indicate that they were reported in MGI to have abnormal neurological phenotype when knocked out. Green rings indicate that the gene was observed to modify AD phenotype in *in vivo* experiments on AD *Drosophila* models. Yellow rings indicate genes which were observed to both modify AD phenotype in *Drosophila* models as well as have reported abnormal neurological phenotype in KO mouse models in MGI. Genes with asterisk next to them are those which have pre-existing FDA approved pharmacological activator or inhibitors, indicating potential targets for drug repurposing studies.

These pathways suggest that modulating GeneEMBED genes may impact neuronal function. This hypothesis is supported by the fact that the 143 high-confidence hits are enriched in differentially expressed genes both in bulk and in single-cell transcriptomic datasets from AD postmortem brains ($p = 0.0247$, **Figure 4B** and **Figure 7**). While many genes have been investigated in AD mouse models to understand their contribution to disease, it is currently impractical to perform this type of analysis with large gene collections. To circumvent this limitation and systematically measure whether GeneEMBED candidates play important roles in CNS, we asked whether modulation of their mouse homologs would cause any neurological phenotypes as tallied in the Mouse Genome Informatics (MGI) database. [138]. This would reveal whether gene candidates are involved in neuronal maintenance and function and whether their loss of function may constitute a risk factor for AD or be a trigger for neurodegeneration. Out of 139 high confidence genes with homologs, 48 (39%) showed abnormal nervous system phenotypes ($p = 0.00024$) when modulated. Notably, among these, a subset of 25 mouse homologs also showed abnormal behavioral/neurological phenotypes ($p = 0.049$). Finally, an additional 11 homologs showed only abnormal behavioral/neurological phenotypes (**Figure 7** shows genes whose modulation causes CNS-associated phenotypes in mice as red or yellow border nodes). Of note, neither the ADSP variant datasets nor the STRING or HINT networks used by GeneEMBED have any bias towards genes expressed in the brain or in neurons. Therefore, the observed enrichment in genes mediating normal neuronal function increases confidence in GeneEMBED and with the potential pathogenic or protective roles of the genes it finds.

To further ascertain the role of GeneEMBED genes in neurodegeneration, we next turned to in vivo experiments. Mouse models recapitulate neuronal dysfunction and neuropathological features of AD; however, they are not amenable for testing a high number of candidates using functional assays. Conversely, cultured cells fail to recapitulate core AD

traits (age dependence, circuit dysfunction, neuron-glia interplay). Therefore, to optimally validate the GeneEMBED candidates in the AD context in vivo we resorted to *Drosophila* AD models, which capture important core AD traits, including age dependence and protein accumulation [139]. This approach is supported by our previous *Drosophila* work in the context of AD and other neurodegeneration disorders where therapeutic targets identified in *Drosophila* have gone on to be validated in mouse or iPSC derived neuronal models [99,100,140–145]. For the GeneEMBED candidates, we modulated the levels of their *Drosophila* homologs in two well-validated *Drosophila* AD models [99–101] to test the effect of each candidate on neuronal dysfunction caused by amyloid (secreted Aβ42) or Tau (2N4R hTau) in the CNS. Expression of secreted *β42* or human *tau* specifically in post-mitotic neurons induces progressive nervous system dysfunction in *Drosophila* which can be monitored by measuring the motor performance of the animals as they age. First, we filtered out high confidence candidate genes that did not have *Drosophila* homologs or available alleles in public repositories. We then tested the resulting 43 genes using both overexpression as well as loss-of-function alleles whenever possible. We found that 28 *Drosophila* genes were modifiers of the β42- and/or tau-induced neuronal dysfunction (**Figure 7**, green and yellow border nodes and **Figure 8, 9**). We further found that of these 28 modifiers, 5 genes (UTRN, REST, PLEC, BAG3, TP53) also showed evidence of dysregulation in human postmortem AD brain transcriptome and abnormal neurological phenotypes in knockout mice. Interestingly, both the MGI hits as well as the *Drosophila* modifiers are very evenly distributed between the different functional clusters (**Figure 7**) indicating that all these pathways may potentially modulate AD pathogenesis. Importantly, some of the *Drosophila* alleles used (inducible overexpression and shRNA lines) were targeted specifically to neurons and therefore likely exerted their effects specifically in neuronal cells. However, other alleles used were classical loss of function or classical rescue constructs (using the endogenous gene promoter) in those cases the effect maybe cell-autonomous or non-cell autonomous for

example through modulation of important functions in glial or muscular cells. Additionally, while some of the modifiers identified may exert their effect through modulating the accumulation of *tau* or *β42*, others may act by protecting or potentiating the predisposition of neurons to degenerate or even by causing certain levels of neurodegeneration themselves. A complete list of the modifier alleles as well as brief description of their putative effect on their target gene is available in **Table 11**.

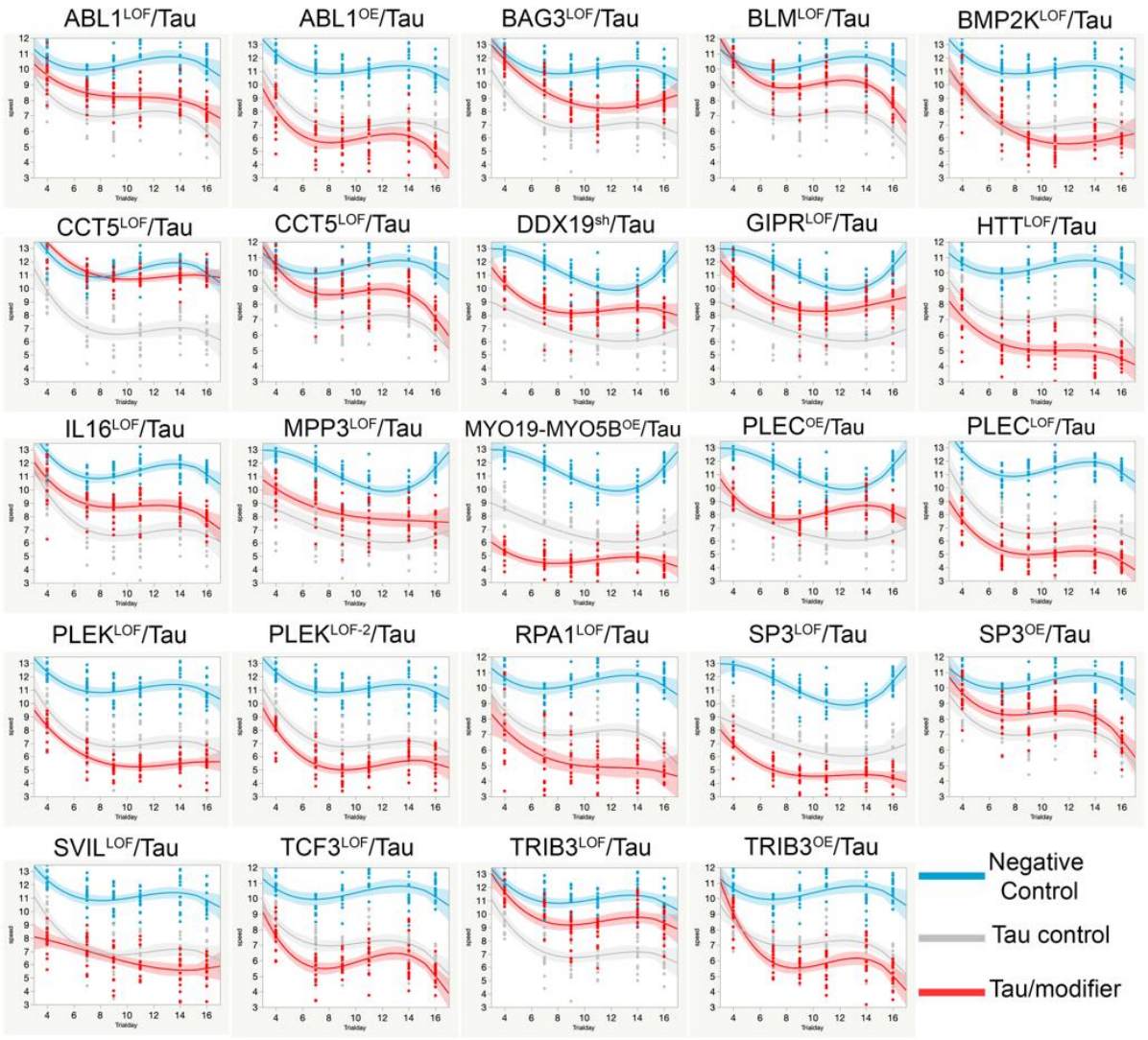


Figure 8: Regressions representing average speed as a function of age in control fruit flies (blue) or flies expressing human wild type *Tau* either alone (grey) or together with the above indicated modifiers (red) on the corresponding Drosophila homolog (see supplementary table 12 for genotype details). Charts show third degree polynomials and confidence intervals. All differential effects were statistically significant ($p<0.01$) following ANOVA analysis on Linear mixed models regression with fitted splines

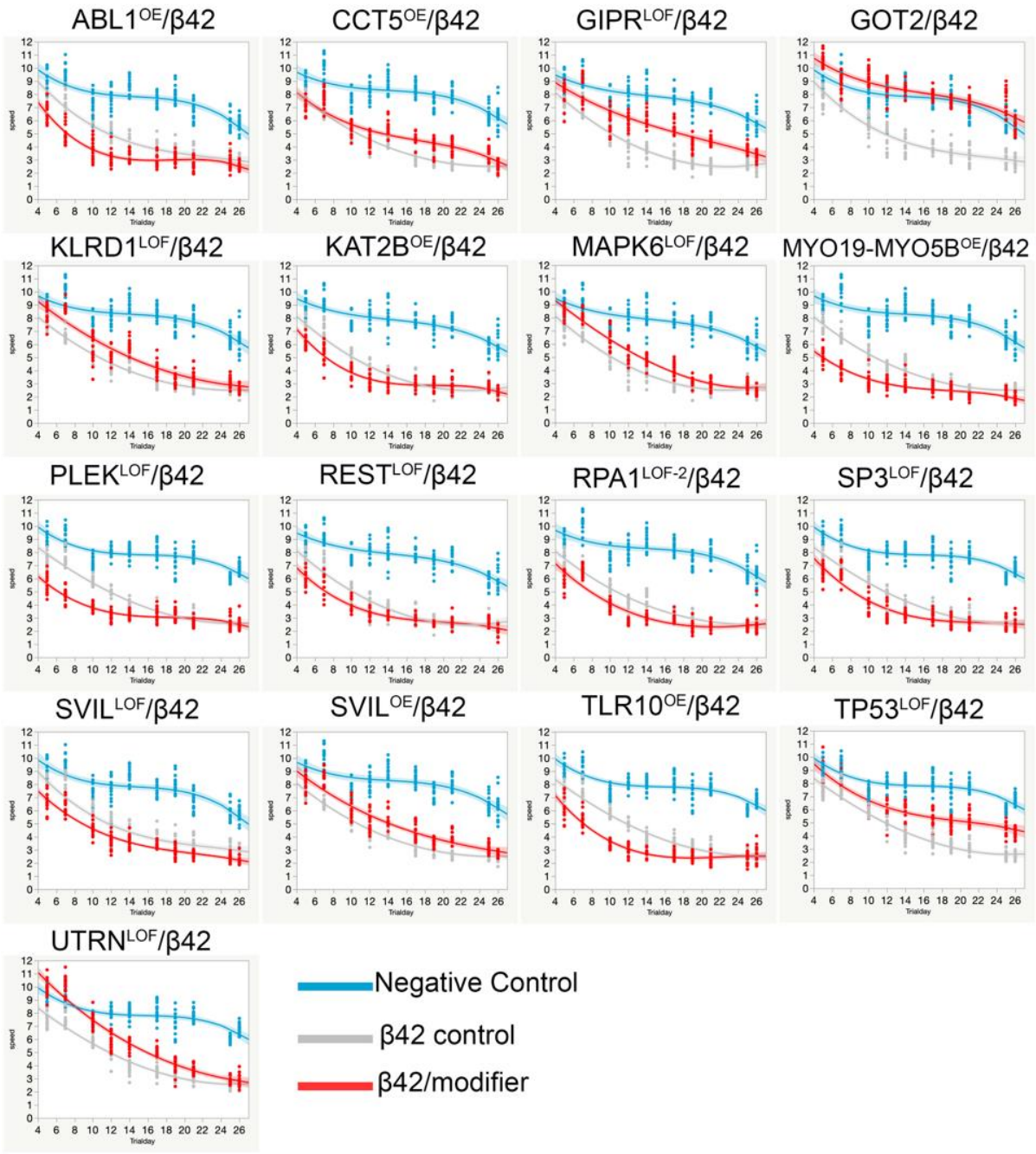


Figure 9: Regressions representing average speed as a function of age in control fruit flies (blue) or flies expressing human wild type *β amyloid* either alone (grey) or together with the above indicated modifiers (red) on the corresponding Drosophila homolog (see supplementary table 12 for genotype details). Charts show third degree polynomials and confidence intervals. All differential effects were statistically significant ($p<0.01$) following ANOVA analysis on Linear mixed models regression with fitted splines

Given the likely neurological role of these high confidence GeneEMBED candidates, we investigated their therapeutic potential. Among the 143 genes, twenty-one have drugs that have been characterized as agonists or antagonists of their function (**Table 12**). Interestingly, of the total 109 compounds activating or inhibiting these genes, 35 have co-mentions with AD in the PubMed database. Noteworthy among these are *EPHA2* and *S1PR3,* both of which were upregulated in AD brains. *EPHA2* has two inhibitors (regorafenib and dasatinib) both of which have shown neuroprotective effects in mouse AD models [146,147]. *S1PR3* has an agonist (fingolimod) which also has therapeutic benefit in mice [148]. Additionally, two genes *FLT3* and *RET* are inhibited by sunitinib, which inhibits cerebrovascular activation to improve cognitive function mouse AD models [149]. Among the genes whose knockdowns ameliorated neurodegeneration in *Drosophila* AD models, three (*ABL1, TP53, POLD1*) have pharmacological agents with previously demonstrated inhibitory effects. While ABL1 inhibition is already being pursued in the context of AD [150,151], TP53 and POLD1 remain to be explored. Together, our results demonstrate that high-confidence GeneEMBED candidates show significant enrichment in modifiers of tau and β-amyloid phenotypes in *Drosophila* models, are differentially expressed in AD brain tissue, and show abnormal neurological phenotypes when modulated in mouse models. These findings highlight the ability of GeneEMBED to successfully identify genes involved in disease pathology, some of which have significant therapeutic potential.

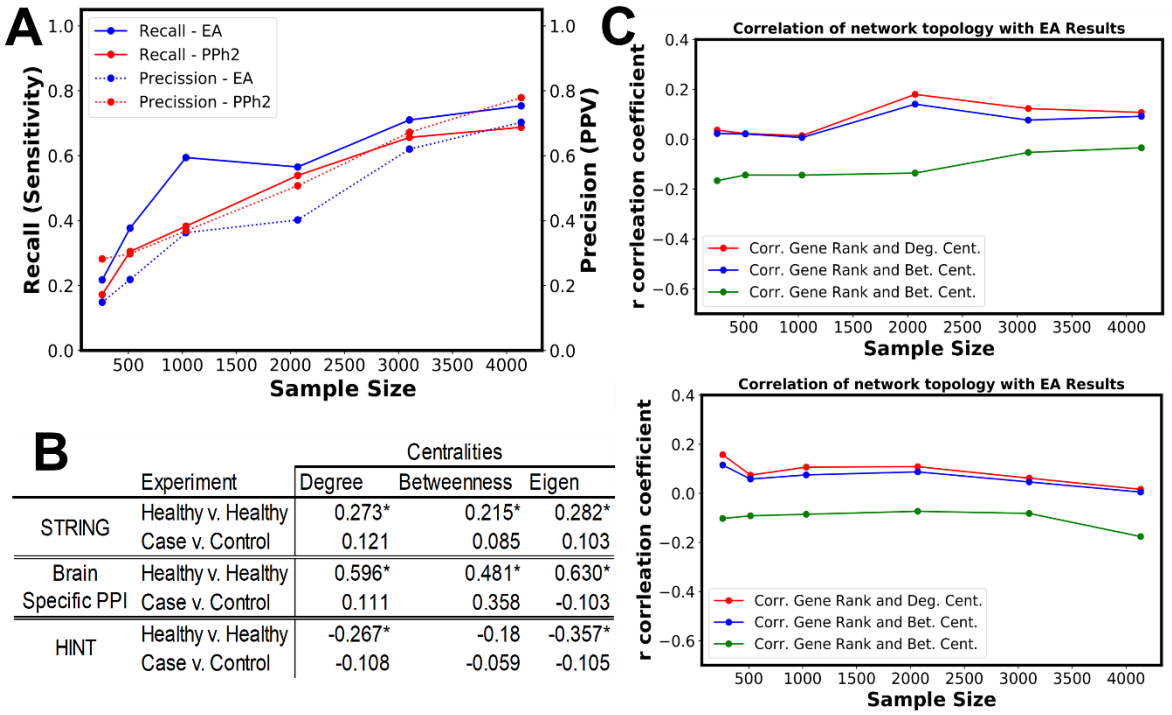
Chapter 2.3.4: GeneEMBED shows robustness across various cohort sizes

While large sequencing cohorts are becoming more commonplace in recent years, for some rarer phenotypes and diseases, it is still a challenge to produce such sample sizes. In order to characterize its performance at various cohort sizes, we performed an iterative downsampling analysis of GeneEMBED on the Discovery cohort. Using the gene set identified with the full cohort as ground truth, we calculated precision and recall of GeneEMBED gene

sets identified at sub-cohort sizes of 80%, 60%, 40%, 20%, 10%, and 5% of the original cohort. Performing this experiment with both EA and PPh2, we found that sub-cohort size could be dropped as far as 40% of the original cohort before recall fell below 0.6 for PPh2 and EA analyses (**Figure 10**). Next, in order to assess the relationship between network connectivity and gene identification by GeneEMBED at various cohort sizes, we correlated the PCA-distances based ranking of identified genes with their ranking using betweenness, degree, and eigen centrality using the Spearman rank-order correlation test. We found that for both EA and PPh2 based analyses, the correlation with these centrality measures was relatively stable regardless of the decrease in cohort size (**Figure 10**). These data suggest that the utility of GeneEMBED is not limited to large cohort sizes but can sufficiently extended to much smaller cohort sizes.

In order to characterize the behavior of GeneEMBED in the absence of disease specific mutational information, we performed the analysis on only healthy controls from the Discovery cohort. We then tested the correlation of ranks of genes which pass FDR threshold with ranks generated by connectivity measures used above. Strikingly, we observed that in GeneEMBED analysis using the STRING network, the PCA-distances correlated with degree, betweenness, and eigenvector centralities at more than twice the rate in healthy vs healthy ($r = 0.273, 0.215, 0.282$) than in case vs control analyses ($r = 0.121, 0.085, 0.103$), and that correlations for healthy vs healthy were all statistically significant while the case vs control analyses were not (**Figure 10**). Even more notable were the disparities of correlations in the Brain Specific PPI network, correlations with degree, betweenness, and eigenvector centralities in the healthy vs healthy analysis ($r = 0.596, 0.481, 0.630$) which were all statistically significant and case vs control analyses ($r = 0.111, 0.358, -0.103$) which were not significant. These findings were again echoed in the HINT network analyses where healthy vs healthy gene set correlated significantly with network centrality measures. The findings

suggest that in the absence of disease-relevant mutational data, GeneEMBED prioritizes genes with large network connectivity as small mutational differences are likely amplified by the gene's network influence.



B

		Centralities		
	Experiment	Degree	Betweenness	Eigen
STRING	Healthy v. Healthy	0.273*	0.215*	0.282*
	Case v. Control	0.121	0.085	0.103
Brain Specific PPI	Healthy v. Healthy	0.596*	0.481*	0.630*
	Case v. Control	0.111	0.358	-0.103
HINT	Healthy v. Healthy	-0.267*	-0.18	-0.357*
	Case v. Control	-0.108	-0.059	-0.105

Figure 10: (A) Plot of precision and recall of GeneEMBED identified genes at decreased sample sizes relative to genes identified using the full Discovery cohort. (B) Spearman rank-order correlation between genes identified using the three brain networks applied to Healthy vs Healthy controls or case vs control experiment. Asterisk indicates statistically significant ($p<0.05$) correlation. When disease relevant information is removed from data, GeneEMBED relies on network topology to rank genes. (C) Spearman rank-order correlation between candidates identified at low cohort sizes.

Chapter 2.3.5: Characterization of performance of PCA vs full embedding distances

In order to assess the effects of PCA on the GeneEMBED methodology, we tested the utility of computing distances based on the full dimensional embedding outputs from the GraphWave algorithm compared to PCA-distances. We ran GeneEMBED with both weighting metrics on the Discovery-VISEA cohort using the STRING network and recovered 82 genes with full embedding distances and 69 with PCA-distances. To test the relevance of the identified gene to AD, we measured their: (i) recovery of AD-associated genes, (ii) connectedness to known AD-associated genes, and (iii) differential expression in postmortem AD brains. We found that distances based on full dimensional embeddings were able to recover statistically significant overlaps with GWAS Meta 1, GWAS Meta 2, and DisGeNet. PCA-distance gene set recovered significant overlaps with DisGeNet and CTD (**Table 16**). Next, we found that genes identified by full dimensional embeddings were significantly connected to GWAS Meta 1, GWAS Meta 3, and CTD gene sets. Comparatively, PCA-distance gene set showed significant connectivity to all five reference gene sets (**Table 17**). Finally, genes identified by full dimensional embeddings showed no statistically significant enrichment for differential expression in post-mortem AD brains, while PCA-distance gene set was enriched in two brain regions with significance ($p = 0.012$) (**Table 18**). These data show that distances based on PCA performs better than full dimensional embeddings. Further examination of the genes identified by both approaches showed that 74% of genes identified by the full dimensional embeddings were also identified in the PCA-distance framework. However, of the 20 genes unique to the full dimensional embeddings' gene set, only 3 were dysregulated in AD brains. Comparatively, of the 7 genes unique to the PCA-distance framework, 4 were dysregulated in AD brains. Overall, this demonstrates the role of principal component analysis in denoising the raw outputs of the GraphWave algorithm.

Chapter 2.3.5: Characterization of alternative edge weighting schemes

In order to assess the effects of PCA on the GeneEMBED methodology, we tested the utility of computing distances based on the full dimensional embedding outputs from the GraphWave algorithm compared to PCA-distances. We ran GeneEMBED with both weighting metrics on the Discovery-VISEA cohort using the STRING network and recovered 82 genes with full embedding distances and 69 with PCA-distances. To test the relevance of the identified gene to AD, we measured their: (i) recovery of AD-associated genes, (ii) connectedness to known AD-associated genes, and (iii) differential expression in postmortem AD brains. We found that distances based on full dimensional embeddings were able to recover statistically significant overlaps with GWAS Meta 1, GWAS Meta 2, and DisGeNet. PCA-distance gene set recovered significant overlaps with DisGeNet and CTD (**Table 16**). Next, we found that genes identified by full dimensional embeddings were significantly connected to GWAS Meta 1, GWAS Meta 3, and CTD gene sets. Comparatively, PCA-distance gene set showed significant connectivity to all five reference gene sets (**Table 17**). Finally, genes identified by full dimensional embeddings showed no statistically significant enrichment for differential expression in post-mortem AD brains, while PCA-distance gene set was enriched in two brain regions with significance ($p = 0.012$) (**Table 18**). These data show that distances based on PCA performs better than full dimensional embeddings. Further examination of the genes identified by both approaches showed that 74% of genes identified by the full dimensional embeddings were also identified in the PCA-distance framework. However, of the 20 genes unique to the full dimensional embeddings' gene set, only 3 were dysregulated in AD brains. Comparatively, of the 7 genes unique to the PCA-distance framework, 4 were dysregulated in AD brains. Overall, this demonstrates the role of principal component analysis in denoising the raw outputs of the GraphWave algorithm.

Chapter 2.3.6 Characterization of sensitivity of GeneEMBED to false negative and false positive edges

While the curation of biological networks has become increasingly more sophisticated, it is important to recognize that even networks built upon stringent curation of experimentally validated edges may be prone to research bias. In order to assess the robustness of GeneEMBED to these potential false positive (FP) or false negative (FN) edges, we performed iterative random deletion or addition of edges to the Brain specific PPI network and applied GeneEMBED to the modified network. Using the genes identified by the Discovery cohort with the VIS-EA as ground truth, we calculated precision and recall of gene sets identified in the modified networks wherein 5% to 70%, with intervals of 5%, of edges were synthetically and randomly deleted (or added). Specifically, we used the Brain specific network for this experiment because its relatively small size of 3.2k nodes and 48k edges allowed for computationally efficient modification and testing. First, testing the relationship between FN edges and GeneEMBED performance, we found that up to 55% of the edges in the original network could be randomly deleted before either recall or precision fell below 0.6 (**Figure 11A**). Moreover, when we restricted random deletion of edges to those involving any of the genes identified in the unaltered Brain specific PPI network, we found that up to 30% of their edges could be deleted before either recall or precision fell below 0.6 (**Figure 11A**). Next, characterizing GeneEMBED performance in the presence of FP edges, we found that we could randomly add edges totaling up to 80% of the original network size (~38.4k edges) before either the precision or recall fell below 0.6 (**Figure 11B**). These data suggest GeneEMBED is highly robust to both false positive and false negative edges. In the case of random deletion of edges (FN edges), it is likely that there are more genes that do not play a role in AD pathobiology than genes that contribute significantly to pathogenesis. Accordingly, there will be more edges that are not associated with AD than edges that are associated with

AD. Therefore, it is possible to randomly delete a large number of edges while maintaining a high recall and precision. However, when there is a bias in the edge deletion process to informative edges, the methodology becomes more sensitive to FN edges. Similar reasoning can be applied to the case of random edge addition (FP edges), as there are likely more edges that are not associated with AD it is possible to have large numbers of FP edges before recall or precision drop below 0.6. Overall, these data show that while there may be potential research bias in curated biological networks, the strategy employed by GeneEMBED allows for its robustness to the presence of false positive and false negative edges.

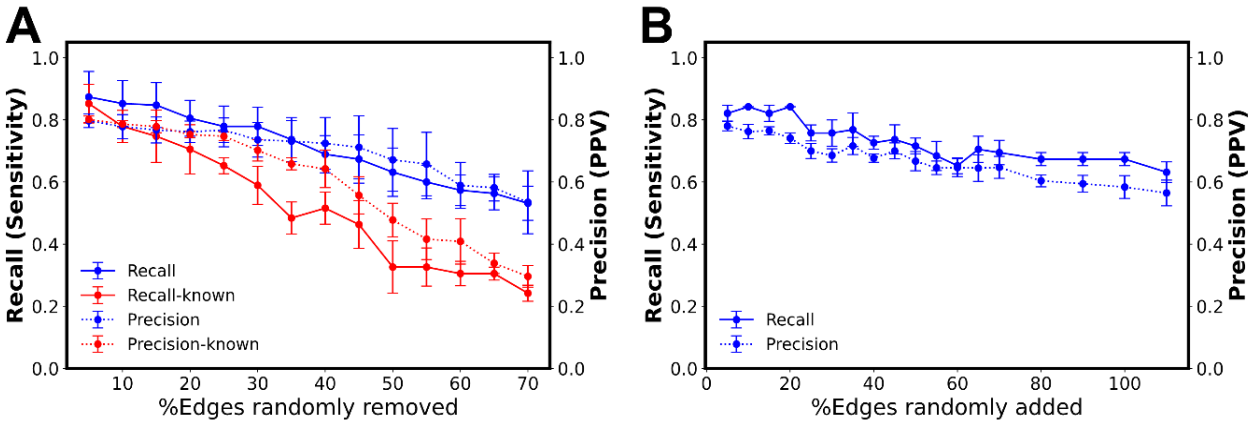


Figure 11: (A) Edges were synthetically and randomly deleted from the Brain network to test sensitivity of GeneEMBED to false negative edges. In blue are plots of precision and recall of GeneEMBED identified genes at various levels of randomly deleted edges. In red are plots of precision and recall of GeneEMBED identified genes when randomly deleted edges are targeted for known (previously identified) genes. (B) Edges were synthetically and randomly added to the Brain network to test sensitivity of GeneEMBED to false positive edges. The plot shows precision and recall of GeneEMBED identified genes at various levels of synthetically added edges. X-axis of '% Edges Added' is relative to the original network size, e.g. at 100%, ~48k edges are randomly added.

Chapter 2.3.7: Characterization of GeneEMBED performance in unbiased networks

In order to benchmark the GeneEMBED strategy with a network without any functional bias or literature curation, we employed the HuRI network [152]. The HuRI network is the largest unbiased interactome map of binary protein-protein interactions. The network contains 8,275 nodes and 52,569 edges generated from an impressive array of nine different 'all-by-all' screens of 17,408 proteins. Using this network as a starting point, we ran GeneEMBED using the Discovery-VISEA cohort and identified a candidate gene set. To test the relevance of the identified genes to AD biology, we examined: (i) direct overlaps with reference gene sets discussed previously, (ii) connectedness between reference gene sets and identified genes, and (iii) dysregulation of identified genes in postmortem AD and non-AD samples from the AMP-AD dataset. Performing these experiments, we found that there was no significant recovery of known AD-associated genes. We also found no significant preferential connectivity between candidate genes and known AD-associated genes (**Tables 13, 14**). We did find an enrichment of the candidate genes for differentially expressed genes in AD vs non-AD brains with marginal significance ($p = 0.06$) (**Table 15**). While these results would seem to suggest that GeneEMBED is unable to perform on such unbiased networks, it is important to consider the HuRI network in the context of AD. Despite being the largest of its kind to date, due to technological limitations, the HuRI network comprises only half the exome. Accordingly, only half or less of the genes in the reference gene sets were present in the HuRI network. Indeed, several genes which are core to AD pathobiology, such as APOE, TREM2, or MAPT, were absent in HuRI. The stringency of the HuRI network's construction suggests that while it has a low FP rate, it may be depleted in protein-protein interactions. Indeed, we have observed that GeneEMBED is more robust to FP edges than FN edges (**Figure 11**). Overall, these data emphasize the importance of appropriately selecting a starting network. While it is

recommended to use an unbiased network when possible, it is also crucial to ensure the network is reflective of the biology of the target disease.

Chapter 2.3.8: Characterization of GeneEMBED in the presence of uninformative mutational data

In the presence of counterproductive mutational data and large influence from network inputs, similar genes will be recovered from various shuffled label experiments leading to inflated overlaps. Indeed, it is likely that due to ambiguous mutational input the identified overlapping genes from randomly shuffled trials are less related to AD than case vs control overlaps. Moreover, the large reliance on network information suggests that identified gene lists are strongly correlated with network connectivity. In order to test these hypotheses, we performed the shuffled labeled experiments in the STRING network for both VIS^{EA} and VIS^{PPh2} using Discovery and Extension cohorts. We found that for VIS^{EA}, an average of 31.7 genes overlapped among gene sets identified using label shuffling for the Discovery and Extension cohorts. Comparatively, an overlap of 14 genes was observed in the original framework after removing potential FPs from control vs control analysis. Importantly, we found that the 14 genes identified in the original analysis showed significant hypergeometric overlap with all five of the reference gene sets of known AD-associated genes ($p = 0.019 – 0.0039$). The overlaps identified by shuffled labeling showed few to no significant overlaps with any of the five reference gene sets (**Table 19**). Next, for each of the 14 genes in the original analysis, we counted the number of publications in the PubMed database co-mentioning the gene with AD in abstracts. We randomly generated 50 gene sets of the same size and counted their co-mentions with AD to build a background distribution. To determine if a gene was related to AD, we used a threshold of at least 5 publications co-mentioning a gene and AD. We found that relative to this background, the original observation of 14 overlapping genes had a z-score of 6.86. Comparatively, the overlaps identified by random shuffling had an average z-

score of 3.43 and stdev of 1.57. Lastly, we found that ranked gene lists derived from random shuffling were significantly correlated with degree centrality (pearson correlation coeff. = 0.2-0.36, p = 0.037 - 1.7e-5), whereas the gene list derived from case vs control analysis was not correlated with pearson correlation coefficient of 0.085 and a pvalue = 0.43. Similarly, in VISPPh2 analysis, an average of 37.5 genes overlapped among gene sets identified by label shuffling. In contrast, 16 genes were found overlapping between Discovery and Extension cohorts using the original framework after removing potential FPs. While no significant overlaps were observed with the reference gene sets (**Table 20**), we found that the PubMed literature curation analysis showed significant association of overlap genes identified from case vs control analysis to AD with a z-score of 4.49. In comparison, overlaps obtained from label shuffling had an average z-score of 0.97 and stdev of 0.82. These data suggest that overlaps observed between Discovery and Extension cohorts in the original analysis are much smaller than expected by label shuffling trials. Despite these large differences in sizes, overlaps from the original framework are more related to AD. Further, they tend to rank genes independently of their pure connectivity, whereas label shuffling leads to a heavy dependence on network information. Overall, these observations demonstrate that during a lack of informative mutational data, GeneEMBED will tend to depend heavily on network information, identifying genes which are less relevant to AD than genes identified through productive (case vs control) mutational data.

Chapter 2.4: DISCUSSION

AD is the leading cause of dementia worldwide. As its prevalence rises, the need to identify therapeutic targets, potential biomarkers, and risk predictive strategies is urgent. These tasks are complicated by the fact that although several AD genes have been discovered, they only partially account for the role of genetics in the disease [9,10]. Here, we

developed GeneEMBED, a new approach to pinpoint genetic risk factors of disease by examining the differential perturbation patterns of gene interactions. Though, in this study, we analyze AD as proof-of-concept, GeneEMBED is a general approach applicable to many complex polygenic diseases.

When applied to the ADSP cohorts, GeneEMBED identified 143 candidate genes that interacted significantly with previously known AD genes (z-score = 2.03 – 6.07), were differentially expressed in bulk tissue and single cells of AD cases (p = 0.0247). While testing such a large collection of genes in AD-related mouse models is currently not possible, we sought to identify experimental links between the GeneEMBED candidates and neuronal biology. We validated candidate genes *in vivo* using two well-characterized *Drosophila* AD models and utilized the MGI database to identify functional links between the GeneEMBED genes and neurological phenotypes. These genes were also linked to known AD pathways and revealed several novel and potentially druggable targets. These pathways included functions related to glial biology, inflammation, protein aggregate clearance, and signaling cascades. While inflammation plays a large role in the pathogenesis of AD, our enrichments draw attention to the regulation of interferon gamma-induced protein 10 (IP-10) production. In AD patients, IP-10 has elevated expression in astrocytes and shows positive correlation between CSF levels and cognitive impairment [153]. In AD transgenic mice it co-localizes with amyloid plaques [153]. Interestingly, among genes responsible for enrichment in this function, three (*NDUFA10, GOT2, TLR10*) show modulation of an abnormal phenotype in animal models (**Figure 7**) while another four (*NDUFA10, NDUFA9, EPHX2, CYP2C9*) have approved pharmacological activators or inhibitors (**Figure 7**). Functions related to glial biology highlighted glial cell-derived neurotrophic factor (*GDNF*) receptor (*GFRa1*) signaling. Studies in transgenic AD mice found that overexpression of *GDNF* induced neuroprotective effects and improved learning and memory [132]. Restoration of *GDNF* effects by introduction of

exogenous *GFRa1* into cortical AD neurons has been shown to alleviate neuronal death [133]. Strikingly, we found that all eight genes (*RET, ROR1, GRIN3A, PLEC, GFRA1, BAG3, NQO1, BCLAF1*) responsible for enrichment in this pathway showed modulation of abnormal neurological phenotypes in mice and *Drosophila* (**Figure 7**). Of these, *RET*, *GRINA3* and *NQO1* all have pharmacological activators or inhibitors which are FDA-approved. *GRINA3*, specifically, interacts with acamprosate which has been associated with decreased incidence of dementia in population studies and has been seen to alleviate cognitive defects in APP transgenic mice [154,155]. Further studies of these gene candidates are needed to disentangle their relationship with AD, however, they present interesting and viable targets for potential therapeutic research.

Several GeneEMBED hits represent novel and unsuspected candidates for AD. Particularly noteworthy were *PLEC* and *UTRN* which, to our knowledge, have not been studied in AD. Both genes were repeatedly identified in multiple GeneEMBED analyses, were significantly upregulated in bulk tissue AD brains, their modulation causes abnormal neurological phenotypes in mouse models [156,157], and they are genetic modifiers of AD-related phenotypes in *Drosophila. PLEC* encodes for plectin, a cytoskeletal protein involved in intermediate filament networks and interacts with actinomycin and microtubules. Mice deficient in *PLEC* isoform *P1c* in neurons demonstrate altered pain sensation, reduced learning, and long-term memory due to increased accumulation of tau proteins with microtubules [156]. Proteomic studies have also associated *PLEC* with AD pathology [158,159]. *UTRN* encodes for utrophin, another component of the cytoskeletal system. Though *UTRN* is downregulated in CA1 neurons containing neurofibrillary tangles [160], its role in the development of tangles is still unclear. The numerous modalities in which *UTRN* and *PLEC* show associations to AD phenotype warrant deeper and more detailed studies to unravel their role in the disease. In a similar vein, we found two additional genes with links to

AD worth highlighting (*TP53 and POLD1*) [161,162] and whose knockdown in *Drosophila* alleviated AD-related phenotypes. Moreover, both of these genes have pre-existing FDA-approved pharmacological inhibitors. We found four compounds (clofarabine, cytarabine, fludarabine, gemcitabine) which inhibit *POLD1*, and one compound (bortezomib) which inhibits *TP53*. Given the distinct effects of these genes in animal models and their druggability, these genes would be priority candidates for further characterization and study in animal models.

GeneEMBED searches for genes that influence disease risk by considering mutational perturbations of function in their molecular interaction network. This is in contrast to variant or gene-based association methods that treat individual genes or variants as independent and isolated risk loci [16,163–166]. To evaluate the functional perturbation of a gene in a disease, GeneEMBED integrates two distinct techniques: variant impact estimators and node embedding algorithms. (**Figure 3A**). Variant impact estimators predict the probable effect of a coding mutation on protein function based on a variety of data. EA is an untrained approach that uses the evolutionary history of sequence variations and phylogenetic divergence to predict the impact of a variant. PPh2 evaluates impacts by applying machine learning tools on sequence and structure features. These estimates are combined across all variants in a gene to predict their total impact on protein function. Node embedding, is a machine learning process which seeks to represent the complex topological properties of a node in an easily manipulatable form. By weighing the interactions of a gene with the sum of its mutational impact and those of its interactors, GeneEMBED uses the perturbed interactions of a gene as learning features rather than their singular mutational burden. Combining these features with node embedding allows GeneEMBED to estimate the differential perturbation of genes in cases versus controls, thereby identifying genes whose disease contribution would not have been apparent in single gene analyses. For example, in AD, the single gene approach

MAGMA did not identify *NQO1* (p_{MAGMA} ~ 0.33) as disease associated despite its links to AD [167–170]. However, its differentially perturbed network interactions between cases and controls allows GeneEMBED to identify *NQO1* with statistical significance (**Figure 12**). This suggests that GeneEMBED identifies genetic processes distinct from those found by standard tools and can offer complementary insights into the factors defining complex diseases.

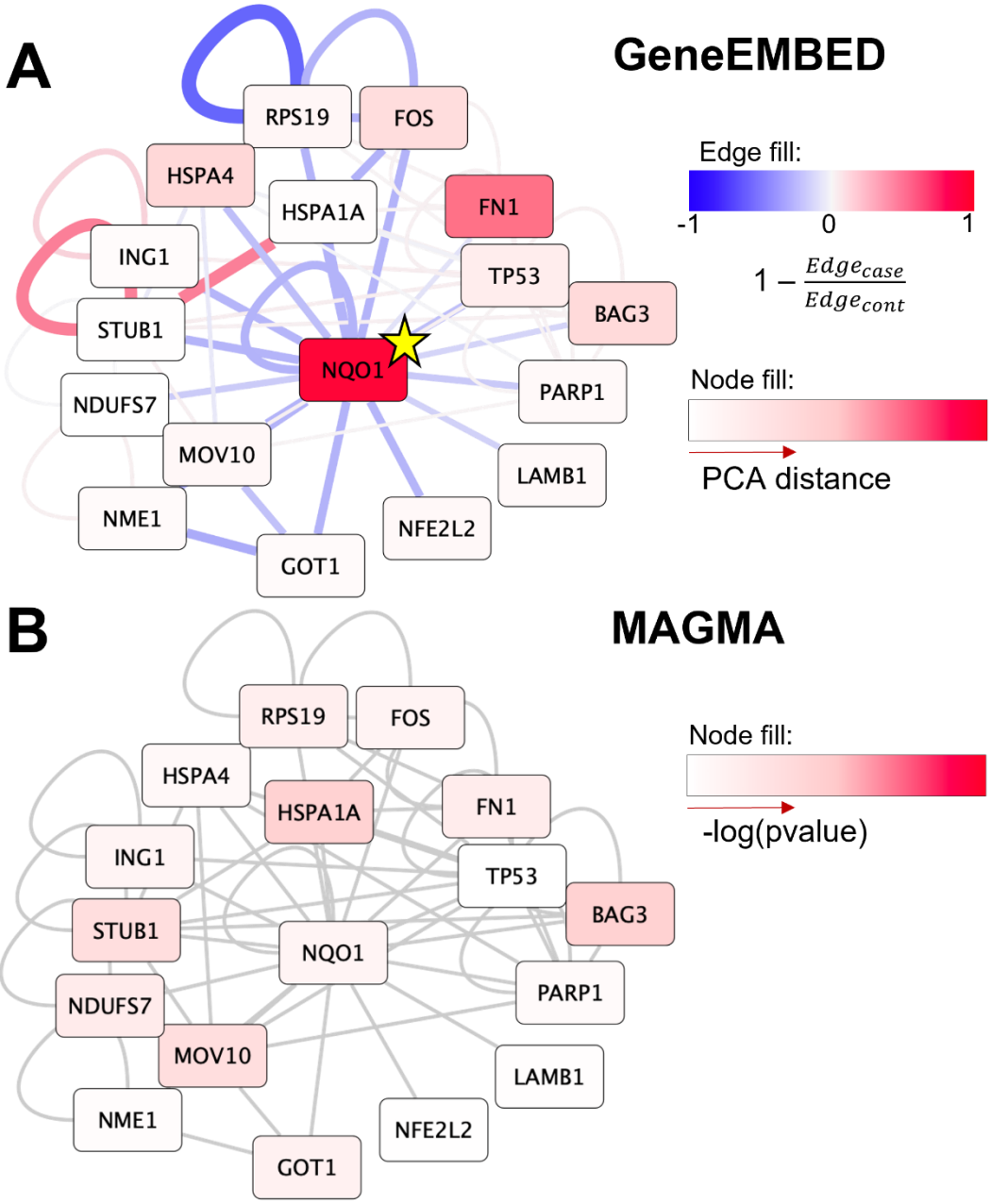


Figure 12: (A) Network of *NQO1* from the Brain network. Edge color represents the zero-centered ratio of mutation edge weight in cases versus controls. Edge width represents the magnitude of this ratio. Node fill is represented by PCA distance from GeneEMBED on the Discovery cohort using EA. The star on *NQO1* indicates that this gene was identified with FDR < 0.01 in GeneEMBED analysis. (B) shows the same network but with node fill corresponding to the -log(pvalue) from MAGMA analysis on the Discovery cohort. Subtle network differences allow GeneEMBED to identify *NQO1* when mutational data alone would not suffice.

The integrative framework of GeneEMBED provides other advantages. First, the integration of network information allows GeneEMBED to be robust to sample sizes. In our analysis of AD, GeneEMBED was able to reliably reproduce findings from the full ADSP Discovery cohort with successively smaller subsampled cohort sizes (**Figure 10**). More than that, GeneEMBED was robust to variations between different cohorts, recovering significant overlaps (p = 1.86e-16, 4.25e-15) in genes identified in the ADSP Discovery and Extension datasets, a challenging task for standard prioritization tools at these sample sizes. Nevertheless, in order to optimally account for the various factors leading to inter-cohort variability and increase robustness of findings, we recommend readers to validate potential candidate gene lists across two or more cohorts. Second, this framework is also flexible in that it is compatible with many different variant impact estimators. Here, we used EA due to its consistently good performance in blind, objective studies [66,171] and overall utility in genomic studies [172,173] in addition to a well-established alternative, PPh2. Despite their differences, we found significant overlap in their predictions (p ~ 2.46e-64 - 1.46e-53), supporting the compatibility of GeneEMBED with multiple impact estimators. The GeneEMBED framework crucially relies on the *PS* metric which is compatible only with estimators that have probabilistic interpretations. While some tools (REVEL, SIFT, MutPred2, or VEST) fit this criterion, many do not have such interpretations or may require further transformations (e.g. CADD or Eigen). The flexibility of the GeneEMBED strategy also applies to different networks. We found that GeneEMBED consistently identified similar genes across the three PPI networks used in this study (p ~ 1.06e-8– 5.78e-28, **Figure 6**), suggesting that usage of any well-constructed and disease relevant network will tend to converge on similar findings. While the use of networks is key in the GeneEMBED strategy, it also introduces a potential source of error. Even stringently curated networks may be prone to research bias. Unbiased networks built with high throughput techniques may provide alternatives. However, they tend to be limited in size due to technical constraints, resulting in an insufficient capture

of disease relevant interactions (**Tables 16-18**). In this regard, the GeneEMBED strategy showed robustness to the presence of both false positive and false negative edges **(Figure 11)**. The flexibility of the framework also provides a channel for improvement in predictive power. Namely, the edge weighting scheme. While other edge weighting approaches were characterized (sup materials, sup tables 13-15), the current framework estimates the perturbation of each interaction independently but considers all edges equally important. However, biological networks are highly robust to mutations due to pathway redundancies [174,175]. Among these, some are dominant while others are auxiliary [176], suggesting that different parts of the network have varying levels of importance. This indicates a potential limitation and area for improvement in the GeneEMBED framework. Potential approaches to address this are to consider alternative methods of node embedding including anisotropic diffusion techniques, which will be the focus of future work.

Chapter 2.5: MATERIALS AND METHODS

Chapter 2.5.1:Whole Exome/Genome Sequencing Data

Whole exome sequencing (WES) data from 5,169 individuals were downloaded from NIH NCBI study ID: phs000572.v8.p459 (ADSP Discovery) and a further 969 whole genome sequences (WGS) were downloaded from National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) dataset: NG0006760 (ADSP Extension). Samples making up the Discovery and Extension cohorts were selected from a set of 24 well characterized cohorts from the Alzheimer's Disease Genetics Consortium. The sample phenotypes were coded as 0 or 1 indicating non-AD and AD, respectively in both cohorts. Only samples of European/White ancestry were used in the analyses of both cohorts. The mean age of AD onset for AD positive Discovery cohort samples was 75.3 years with a

standard deviation of 8.3 years, while the mean age of last exam for control samples was 85.5 years and a standard deviation of 5.1 years. The mean age of AD onset for AD positive Extension cohort samples was 75.5 years with a standard deviation of 7.8 years. Healthy controls of the Extension cohort had a mean age at last exam of 75 years with a standard deviation of 8.3 years. **Quality Control and Annotation of WES and WGS Data**: Although extensive QC procedures were performed on the WES Discovery and WGS extension cohorts by the ADSP and GCAD consortia [177], respectively, we generated QC statistics for Ti/Tv, number of variants, singletons, and missingness for each sample and HWE, genotyping rate (AC/AN) for each variant site across cases and controls. Then, potentially false-positive variants sites and outlier samples were removed. HWE (Hardy Weinberg Equilibrium) exact test [178] was performed on the control samples of each cohort and the variants with HWE violations (HWE p-value < 5E-8) were removed. We also removed the variants that were genotyping rate less than 0.95 in either case and control and in combined case and control samples. Outlier samples including potentially non-whites were identified based on Ti/Tv, total number of variants, singletons, and missingness. To cluster samples with genetic background and identify outliers of clusters, we applied Principal Component Analysis (PCA) method. We identified potentially related samples by estimating genetic relationships between samples with kinship coefficients. We removed outliers that include non-European descendants. To annotate consequences of variants, we used the Annovar133. Then, non-synonymous single nucleotide variants (SNVs) and small indels, which lead to frames-shift, excluding CNVs (copy number variants) were annotated with EA. BCFTOOLS [179], KING [180], and SMARTPCA from Eigenstrat package were used for calculating variant and sample statistics, inferring relationships, and for estimating sample clusters with PCA [181], respectively.

Chapter 2.5.2: Variant Scoring Methods

Two variant scoring methods were used to describe the mutational impact of variant, separately. The first of these two methods are PolyPhen2 (PPh2), which predicts the potential impact of an amino-acid substitution on protein function using a machine learning algorithm trained on sequence and structural information. Here, PPh2 HDIV raw scores were used. PPh2 scores range from 0-1, where increasing value indicates increasing severity of mutation.

The second scoring method we used was Evolutionary Action (EA), which expresses that the genotype-phenotype relationship can be written as $f(\gamma) = \phi$, where evolutionary fitness function (f) maps genotype (γ) onto fitness landscape (ϕ). SNVs are considered small perturbation in the genome $(d\gamma)$ and cause perturbation in fitness $(d\phi)$: $\nabla f \cdot d\gamma = d\phi$. A missense mutation at residue $r_j, d\gamma \approx \Delta r_j$, will cause all components of ∇f to be forced to zero except $\frac{\partial f}{\partial r_j}$, and impact equation simplifies to $\Delta\phi \approx \frac{\partial f}{\partial r_j} \cdot \Delta\gamma$. Evolutionary Trace [72] is used to compute $\frac{\partial f}{\partial r_j}$, and $\Delta\gamma$ can be approximated with amino acid substitution log-odds ratios. EA scores are reported between 0-1 with increasing severity of functional impact, where EA=0 indicates no effect on protein function and EA=1 indicates loss of function. In the EA scoring systems, silent mutations are given a score of EA=0, while frame shift and stop mutations are given a score of EA=1.

Chapter 2.5.3: GeneEMBED

Chapter 2.5.3.1: Network Construction

In the bulk of the work presented here, we use three biological networks for protein-protein interactions including STRING v10 [95], HINT [96], and a brain specific network [97,98]. The STRING network defines edges between genes using many forms of evidence including curated interactions, experimental interactions, protein homology, co-expression, text mining, etc. The HINT network consists of manually and systematically curated edges requiring interactions to have been reported at least twice in literature. The brain specific

network consists of genes who demonstrate tissue specificity per Human Protein Reference database and BRENDA Tissue ontology. Edges in the brain specific network are listed only if there is experimental evidence for an interaction. For in-depth construction details, please see the appropriate publications. For use in this approach, all edge confidence scores in all networks were removed and replaced with a weight of 1, simply indicating an edge exists between two genes.

Networks are first made sample specific by integrating mutational information. First we compile the functional effect of a set of variants in an individual into one gene level score called a perturbation score (*PS*), defined as: $PS_{gene} = 1 - \prod_i^v (1 - VIS)^{zyg}$, where *v* is the number of variants in a gene for the individual, *i* is the index over those variants and $zyg \in \{0,1,2\}$ where 0 denotes wild type, 1 denotes heterozygous, and 2 denotes homozygous for variant *i*, and *VIS* denotes functional impact of variant (EA or PPh2 score). To construct sample specific networks, we calculate edge weights as the sum of the *PS* of the two connected genes: $W_{edge} = |PS_x + PS_y|$. Characterization of alternative edge weighting schemes and their corresponding discussions can be found in **Tables 13-15.** Finally, to construct disease and control specific networks, edge weights are averaged over all cases and controls separately to build a case specific and control specific mutation weighted network.

Chapter 2.5.3.2: Node Embedding Algorithms

In order to assess network perturbations in genes between cases and controls, we use the GraphWave algorithm [44] to generate node embeddings. The GraphWave algorithm has advantages over other embedding algorithms in that it provides rigorous mathematical guarantees on identifying structure preserving embeddings. GraphWave performs

unsupervised node embedding on node structure (i.e. topological patterns of node connectivity). Accordingly, the authors provide proof for the equivalency of embeddings between two structurally identical nodes a and b, which rests on the assumption that there exists a one-to-one mapping between the K-hop neighborhood of the two nodes. We can extend this proof to claim that the embeddings of a node from two identical graphs must also be equivalent since there will exist a mapping between the node neighborhoods. Thus, when comparing disease and healthy graphs wherein the node connectivities are largely unchanged, the descriptive features captured by each dimension in the embedding space are the same, thus allowing for direct comparisons. The GraphWave algorithm is briefly described below.

Let V denote the eigenvectors and λ_n denote the eigenvalues ($\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$) of the graph Laplacian $L = D - A = V\Lambda V^T$, where *D* denotes the degree matrix and *A* denotes the adjacency matrix of the graph. Now consider a low-pass filter kernel $g_s = e^{-\lambda s}$, where *s* is some scaling factor, we may define spectral graph wavelets by modulating the graph Laplacian by kernel g_s:

$$\Psi_a = V\Lambda\big(g_s(\lambda_1), \dots, g_s(\lambda_n)\big)V^T\delta_a \tag{16}$$

Where δ_a is a Dirac signal about node *a*, Ψ_a is an n-dimensional vector representation of the spectral graph wavelet of node *a*, and *s* is a scaling factor corresponding to the radius of the neighborhood around node *a*. GraphWave samples over a set of $s_j \; for \; s_j \in \{s_{min}, s_{max}\}$ where $s_{min}, and \; s_{max}$ are automatically calculated.

We can recover coefficients of the graph spectral wavelet Ψ_a corresponding to a neighbor node *m* by:

$$\Psi_{ma} = \sum_{i=1}^{N} g_s(\lambda_i)V_{mi}V_{ai} \tag{17}$$

Where Ψ_{ma} represents the signal received by node *a* from a neighbor node *m*, and V_{mi}, V_{ai} denote the *i-th* value of the eigenvectors of *m* and *a*. Similarities in node characteristics are carried in Ψ_{ma} coefficients. GraphWave proposes to use the Ψ_{ma} coefficients as components of a characteristic function, which, when sampled at *d* evenly spaced points, allows a 2-d representation of Ψ_a:

$$\phi_a(t_j) = \frac{1}{N}\sum_{m=1}^{N} e^{it\Psi_{\mathrm{ma}}} \tag{18}$$

Where t_j comes from the set of *d* evenly spaced points ($\{t_1, t_2, \ldots, t_d\}$), and *i* is the imaginary unit ($i = \sqrt{-1}$). The final embedding of the node is then collected as:

$$\mathrm{X}_a = \left[Re\left(\phi_a(t_j)\right), Im\left(\phi_a(t_j)\right)\right]_{t_1,..t_d} \tag{19}$$

Chapter 2.5.3.3: Gene Identification

In order to find genes with differentially perturbed network characteristics between case and control, GraphWave is applied to both networks. Each gene will now have two embeddings, one corresponding to the case network and another from the control network. The GeneEMBED hypothesis supposes that genes contributing to disease will have significantly differing case and control embeddings. To prioritize genes accordingly, we perform principal component analysis (PCA) on the node embeddings and measure distances between case and control embeddings in the PCA space. The role of PCA in the methodology is to aid in denoising the full dimensional embeddings retrieved by GraphWave. The full dimensional embeddings produced by the algorithm can encompass signals ranging from immediate neighborhoods to the complete graph. As a result, the full dimensional embedding of a node will be influenced by any change in the edge weight anywhere in the graph. In order to remove some of these noisy influences, we perform a PCA on the full dimensional

embeddings and use the first principal component to measure distances as this component recovers between 78-92% of the variability explained. Characterization of performance between distances computed on full dimensional embedding against distances measured on PCA is shown in **Table 13-15**. By defining distances as the square root of the L2-norm of each gene measured between case and control, we are able to reconstruct a gaussian-like distribution from the positive and negative values. We then compute z-scores and their corresponding p-values for each distance value relative to the full distribution. Then, we perform false discovery rate (FDR) corrections on the p-values using the Benjamini-Hochberg method and genes corresponding to distance values passing FDR < 0.01 are selected as pre-candidate genes. Lastly, the full GeneEMBED process is performed on healthy controls vs healthy controls (details of healthy control selection are discussed below). Genes passing FDR < 0.01 threshold in this control vs control analysis are removed from the list of pre-candidate genes. This is done to filter potential sources of variation which may not be disease specific (false positives, sup materials). The final set of genes passing FDR < 0.01 threshold and not removed by control vs control analysis are considered candidate genes.

Chapter 2.5.3.4: Computational efficiency/requirements.

GeneEMBED offers an analytical framework to appraise all coding genes in the human genome with respect to their attributes in a molecular network. Accordingly, this can be computationally demanding depending on the size of the network being used. In this study we used three different PPI networks, a brain specific network, HINT, and STRING. After annotation and preprocessing of exomic variant calling file (VCF), the computational time required for the brain network consisting of 3.2k nodes and 48k edges was 649 seconds (10.8 minutes). Similarly, for the HINT network consisting of 12.6k nodes and 146k edges, the computational time from annotation of networks with mutational information to identification of candidate genes was 3058 seconds (51 minutes). Lastly, for the STRING network consisting

of 15k nodes and 1.9m edges, the computational time was 6.2 hours. All network analyses were performed on a server with specifications of Intel Xeon Gold 5222 CPU at 3.8GHz with 8 cores and 348gb RAM.

Chapter 2.5.4: Downsampling analyses

To assess sensitivity to cohort size, GeneEMBED was applied to randomly selected sub-cohorts of the ADSP Discovery data set. Sub-cohort sizes were: 80% (4135 samples), 60% (3101), 40% (2068), 20%(1079), 10% (517), 5% (258), 1% (52). Variability of GeneEMBED predictions over decreasing sample sizes was assessed by calculating recall (sensitivity) and precision (positive predictive value) relative to candidate genes identified in the full Discovery cohort. To characterize dependency of GeneEMBED on network information as mutational data decreased, we calculated Pearson correlation between PCA-distances and degree, betweenness, and eigen centralities. Experiments were done using STRING network and both EA and PPh2 scores.

Chapter 2.5.5: Negative control experiment

To analyze the predictions made by GeneEMBED in the absence of meaningful mutational data, we applied it to healthy controls. Healthy samples were defined to be individuals, from the ADSP Discovery cohort, who were homozygous for the APOEε3 variant, and had low BRAAK staging (1 or 2). This filtering resulted in 725 control samples which were randomly split into two groups. GeneEMBED was applied to the two groups and Pearson correlation between PA-distances and degree, betweenness, and eigen centralities were measured for STRING and brain specific networks.

Chapter 2.5.6: MAGMA analyses

We used MAGMA as a methodological control. MAGMA analysis performed on the same vcf files. The variants were annotated with each corresponding NCBI reference genes

of GRCh37 or 38. Next, we calculated each gene's p-values based on the snp-wise Mean model with a '--burden flag' to avoid deteriorating power of extreme rare alleles and the allele frequency threshold '0.1'. A threshold of $p < 0.001$ was used because the FDR thresholds resulted in too few genes for meaningful comparison to GeneEMBED.

Chapter 2.5.7: Recall of known AD genes

To test whether our approach could recover known genes related to AD we assessed direct overlaps. five gene sets were used to define known AD related genes: Comparative Toxicogenomic Database (CTD) gene set of 103 AD related genes [120], a set of 25 genes identified by meta-analysis of large scale GWAS of diagnosed AD (GWAS Meta 1) [78], a set of 38 genes identified by another meta-analysis of AD GWAS studies (GWAS Meta 2) [79], a set of 208 genes with associations to AD from DisGeNET (DGN) [119], and a set of 21 genes acquired from the ClinVar database [127]. Significance of direct overlaps was assessed with hypergeometric tests between sets of known AD genes and candidate gene sets, separately.

Chapter 2.5.8: Network Analyses

nDiffusion [131] was applied to measure how well GeneEMBED candidates were connected to known AD genes (defined above). nDiffusion relies on graph information diffusion methods [128–130] wherein signals are propagated from genes of interest to all genes in a network through their connections. Genes that receive more signal are more connected to genes of interest. Therefore, if known AD genes receive more diffusion signal from GeneEMBED candidates than other genes in the network, they are more connected. nDiffusion also selects random sets of genes with similar degrees of connectivity as genes of interest to produce a background distribution. Two sets of genes are then deemed significantly connected if their area under receiver operating characteristic (AUROC) is > 0.5 and has a z-

score (relative to random) > 2. The nDiffusion webtool was used to perform these analyses with the preset default settings.

Chapter 2.5.9: RNA sequencing Analysis

In order to assess whether expression changes of GeneEMBED candidates in AD brain tissue, we used the AMPAD data sets. Significant differentially expression (DE) was defined, per brain region, as genes which had log2(fold-change) > 0.263 or log2(fold-change) < 0.263 *and* FDR < 0.05. This thresholding provided 1880 DE genes for cerebellum (CBE), 2952 genes for temporal cortex (TCX), 56 genes for frontal pole (FP), 73 genes for inferior frontal gyrus (IFG), 1579 genes for parahippocampal gyrus (PHG), 271 genes for superior temporal gyrus (STG), and 161 genes for dorsolateral prefrontal cortex (DLPRC). AD case versus non-AD control differential expression analysis results from all brain regions listed above are available online (https://doi.org/10.7303/syn9702085). To assess whether GeneEMBED candidates were enriched for DE genes, we performed hypergeometric tests per brain region. These hypergeometric tests were limited only to the set of genes which were present in both the RNA-sequencing data from AMP-AD cohort and the WES data from the ADSP Discovery and Extension cohorts (i.e. only genes sequenced in both data sets were used). We performed these tests over all seven brain regions to identify region specific enrichments. Next, to determine statistical significance of having enrichment in *n* out of seven brain regions, we repeated the above analysis 1000x with randomly selected gene sets of similar size as candidate gene set. P-values were calculated for the observed number of enriched regions in candidate gene set relative to the distribution observed from random gene sets. This analysis was repeated for all GeneEMBED candidate gene sets, GWAS Meta 1 gene set, and GWAS Meta 2, gene set.

Chapter 2.5.10: Pathway Enrichment Analysis

Protein-protein interaction network of high confidence GeneEMBED candidates was built with the Homo sapiens STRING v11[182] using the combined score of all evidence types at a threshold of 0.400. *HiDef-Louvain* algorithm tool in the Community Detection extension algorithm of Cytoscape was used for clustering followed by functional enrichment analysis of each of the 21 main clusters. Gene set enrichment analysis was performed using the *iQuery*, *EnrichR* and *Gprofiler* Community detection interphases. Network was represented using Cytoscape v3.8.2 [183].

Chapter 2.5.11: Mouse Phenotype Analysis

To assess the relationship between high-confidence GeneEMBED genes and mouse phenotypes, we downloaded the files VOC_Mammalian_Phenotype.rpt and HMD_HumanPhenotype.rpt from the Mouse Genome Informatics (MGI) database (downloaded Nov. 2021). Within the downloadable database, we queried our full set of 143 genes and found that only 139 were documented in the database. These 139 genes mapped to 182 mouse homologs/orthologs. We then tallied the number of mouse genes in our candidate set which had annotations for the high level mammalian phenotype of 'Nervous system phenotype'. We then tallied the total number of mouse genes in the downloadable database which had the same high level mammalian phenotype annotation. We then performed a Fisher's Exact Test to determine the statistical significance of our observations. Additionally, we repeated this analysis for high level mammalian phenotype categories of (i) 'Behavioral/Neurological phenotype' and (ii) 'Nervous system phenotype' AND 'Behavioral/Neurological phenotype'.

Chapter 2.5.12: Drug interaction Analysis

To assess whether any of our high confidence candidate genes were potential therapeutic targets, we used the Drug-Gene Interaction database (DGIdb) [184]. The set of high-confidence candidate genes were input into the 'Search Drug-Gene interactions' webtool

on the DGIdb website. We applied preset filters of 'Approved' indicating FDA-approved drugs only. We then filtered the subsequent list of drug-gene interactions for those which were annotated as having a directional (inhibiting or activating) effect. The resulting genes were then queried through PubMed database for co-mentions with 'Alzheimer' in abstracts.

Chapter 2.5.13: Drosophila strains and neuronal dysfunction assay

Genetics and strains: *Drosophila* lines carrying UAS-Tau, and UAS−Aos:β42 have been previously described [185,186] and are available from the Bloomington *Drosophila* Stock Center (BDSC, University of Indiana). For post mitotic pan-neuronal expression we used the elav-GAL4(C155) driver from BDSC. The alleles tested as potential modifiers targeting the *Drosophila* homologs of GeneEMBED candidate genes were obtained from the BDSC. Homologs were identified using BLAST and also the DRSC Integrative Ortholog Prediction Tool (Diopt score) [187,188](**Table 12).** For the neuronal dysfunction tests, we used a highly automated behavioral (motor performance) assay based on the *Drosophila* startle-induced negative geotaxis response as previously described [188,189]. To assess motor performance of fruit flies as a function of age, we used 10 age-matched virgin females per replica per genotype. Four replicates were used per genotype. Flies are collected in a 24-hour period and transferred into a new vial containing 300µl of semi defined media (20g yeast, 20g Tryptone, 30g sucrose, 60g Glucose, 0.5g MgSO4^7H2O, 0.5g CaCl2^2H2O, 80g Inactive Yeast, 1L H2O) every day. Using an automated platform that uses a mechanized arm and clamp (https://nri.texaschildrens.org/core-facilities/high-throughput-behavioral-screening-core), the animals are tapped to the bottom of a plastic vials to trigger their negative geotactic response (climbing response) and are recorded for 7.5 seconds as they climb on the walls of transparent plastic vials. Videos are analyzed using custom software (code available for download on ref [188]) that assigns movement trajectories to each individual animal, assesses their speed

(mm/s) and returns an average per replicate per trial. Three trials per replicate are performed each day shown, and four replicates per genotype are used. A mixed effect model analysis of variance using spline regressions was run on Rstudio, using each four replicates to establish statistical significance across genotypes [189]. Human genes *POLD1* and *ANLN* were identified as modifiers in a separate manuscript currently under revision and were not directly tested here. All shown modifier alleles had a significant effect ($p<0.01$) compared to the disease controls.

Chapter 3: Thermodynamics inform mutational intolerance

Chapter 3.1: Introduction

Understanding the complex relationship between genotypic changes and how they map into phenotypic fitness is arguably the biggest roadblock in our understanding of the evolutionary process [190]. Uncovering such relationships between sequence variations in DNA or proteins and their molecular function would have broad impacts in disease diagnosis, therapeutic development, and even personalized medicine. Predictably, considerable efforts have gone into analyzing these genotype-phenotype maps and have resulted in key realizations that the maps themselves are a source of anisotropic variation which biases the direction of evolution, not always in the most beneficial way [190,191]. Interestingly, in the context of a so-called "weak mutational regime", many, starting with Iwasa in 1988 [192], have demonstrated exact analogies between evolutionary dynamics in equilibrium and statistical mechanics [193–195]. These analogies are driven by the definition of a "free fitness", an evolutionary counterpart to free energy in statistical mechanics. While many other functions and principles of statistical mechanics can be easily derived through this "free fitness" analogue, directly estimating "free fitness" is nontrivial. For this, we turn our attention to the theory of Evolutionary Action (EA) [69].

As species experience genetic drift and undergo adaptation, then their genotypes γ and phenotypes φ will vary. The variation of γ and φ, however, will be coupled to one another through an *evolutionary fitness function* f : $f(\gamma) = \varphi$. Proposing that f exists and is differentiable, EA theory suggests that small perturbations to genotype $d\gamma$, corresponding to missense mutations, will induce variations in global phenotype fitness $d\varphi$ given by:

$$\nabla f \cdot d\gamma = d\varphi \tag{20}$$

where ∇ denotes the gradient operator and $\cdot$ denotes the scalar product. Critically, we note that, as a result of its multivariate nature, the function *f* maps a genomic sequence variation $d\gamma$ to fitness variation $d\varphi$ though two levels: first the change in function of a gene induced by sequence variation $d\gamma$ at position i along the gene, and second the change in global fitness due to change in gene function:

$$\frac{\partial f}{\partial \Gamma_{g,\forall g\in\mathbb{G}}}\frac{\partial \Gamma_g}{\partial \gamma_{i,\forall i\in\Gamma_g}}d\gamma_i = d\varphi \tag{21}$$

where the partial derivative $\frac{\partial f}{\partial \Gamma_{\mathrm{g}}}$ denotes the sensitivity of the fitness function to changes in function Γ of gene *g* for all genes in genome $\mathbb{G}$, and the partial derivative $\frac{\partial \Gamma_g}{\partial \gamma_i}$ denotes the sensitivity of the function Γ of gene *g* to sequence variation $d\gamma$ at position i along the gene. Here, EA proposes that the sensitivity of the fitness function *f* is equal across all genes, $\frac{\partial f}{\partial \Gamma_{\mathrm{g}}} = 1\ \forall g \in \mathbb{G}$, and claim, to first order, the evolutionary gradient lies in $\frac{\partial \Gamma_g}{\partial \gamma_i}$. EA is then defined as the change in phenotype induced by a small change in genotype:

$$EA \triangleq d\varphi \approx \frac{\partial \Gamma_g}{\partial \gamma_i}\cdot d\gamma_i \tag{22}$$

For details on the analytical calculation of EA, we refer the readers to the original publication [69].

Interestingly, we note that this expression is akin to that of work both conceptually and in mathematical form. Work is the energy transferred to an object through the application of a force along a displacement. More specifically, in the absence of changes to velocity or rotation, when an object is displaced in a conservative force field, work is equal to the minus change in potential energy. This work expression can be rewritten as $dW = \frac{dU}{dx}dx$ in the one-dimensional case. While the mathematical similarity between this equation and equation (22)

is apparent, the conceptual similarities are much more striking. Just as *dW* is dependent on *dx*, a small displacement in physical space, EA ($d\varphi$), is dependent on a small displacement in genotype space $d\gamma_i$ along dimension i. Similarly, just as $\frac{dU}{dx}$ informs about the change in potential due to change in physical space, $\frac{\partial \Gamma_g}{\partial \gamma_i}$ dictates the change in fitness potential of gene *g* in response to displacement in genotype space. Taken together, we propose these similarities motivate a reinterpretation of EA as an analog of energy in genotype space. Furthermore, we suggest that if EA can be taken as an energy of mutations (sequence variations), then large systems of mutations must obey the basic tenants of thermodynamics.

Chapter 3.2: Results

Chapter 3.2.1: Energetics of Coding Mutations

In order to assess the proposition equating EA with energy in genome space, we begin by asking whether the distribution of EA in biological systems are Boltzmann in nature, thereby maximizing entropy. We applied the EA calculation to all coding missense variants in the UK Biobank (UKB) [196–198] whole exome sequencing (WES) cohort of 200,000 human individuals. After defining each missense variant to represent a 'particle' or 'bit' of information, we found that the distribution of EAs for all variants in the UKB closely followed a Boltzmann with $R^2 = 0.95$ (**Figure 13A**). Next, to check robustness of this finding across the evolutionary tree, we repeated the experiment using a close relative, *rhesus macaques*, and a much more distant branch, *E. Coli*. We calculated EAs for all coding missense variants from a whole genome sequencing (WGS) cohort of 850 Indian-origin *rhesus macaques* [199] and WGS cohort of 255 environmental *E. Coli* strains from the European Nucleotides Archive [200]. Similar to humans, we observed that the distribution of EAs for coding variants closely followed a Boltzmann distribution in both *macaques* and *E. Coli*, with $R^2 = 0.96$ and 0.97, respectively (**Figure 13A**). We further assessed these entropy maximizing properties of EA

distributions by restricting variants to only those observed within a given individual of a species. Across the 200,000 individuals of the UKB, we found consistently Boltzmann-like distributions of EA with R^2 ranging from 0.94 – 0.98 and mean R^2 = 0.967 (**Figure 13B**). Similarly, in the 855 individual *macaques* we found EA distributions were consistently Boltzmann with an R^2 range of 0.965 – 0.99 and a mean R^2 = 0.984 (**Figure 13B**). Taken together, these data suggest that the statistical mechanical properties of coding variants are universal, permeating all levels of the evolutionary tree. Moreover, these data suggest that regardless of the differences in mutational profile of coding variants between individuals, the overall mutational ensemble is similar. The robustness of these findings across different species further highlights the ubiquitous influence of statistical thermodynamics in biology.

Next, in order to further assess the compliance of EA theory with statistical mechanics, we ask whether the total energy, EA, of individuals in a species is approximately equal. If we consider variants to be equivalent to 'particles' or 'bits' of information, then we can suppose that individuals are simply an ensemble of distinguishable distributions of particles with energies, i.e. microstates. Accordingly, all microstates of a given macrostate must have the same energy. To obtain a total EA for an individual, we simply measured sum of the EAs of each variant seen in the individual. After plotting the distribution of the sum of EAs for all samples in the UKB cohort, we found that the total variant-wide EAs were constrained to a narrow interval with coefficient of variation (CV) = 0.0128 (**Figure 13C**). Strikingly, we observed similar patterns of CV = 0.096 when plotting the distribution of total variant-wide EAs among individuals of the *macaque* cohort (**Figure 13C**). These data demonstrate the adherence of genetics to another principle of thermodynamics, further cementing the governance of biology by statistical mechanics.

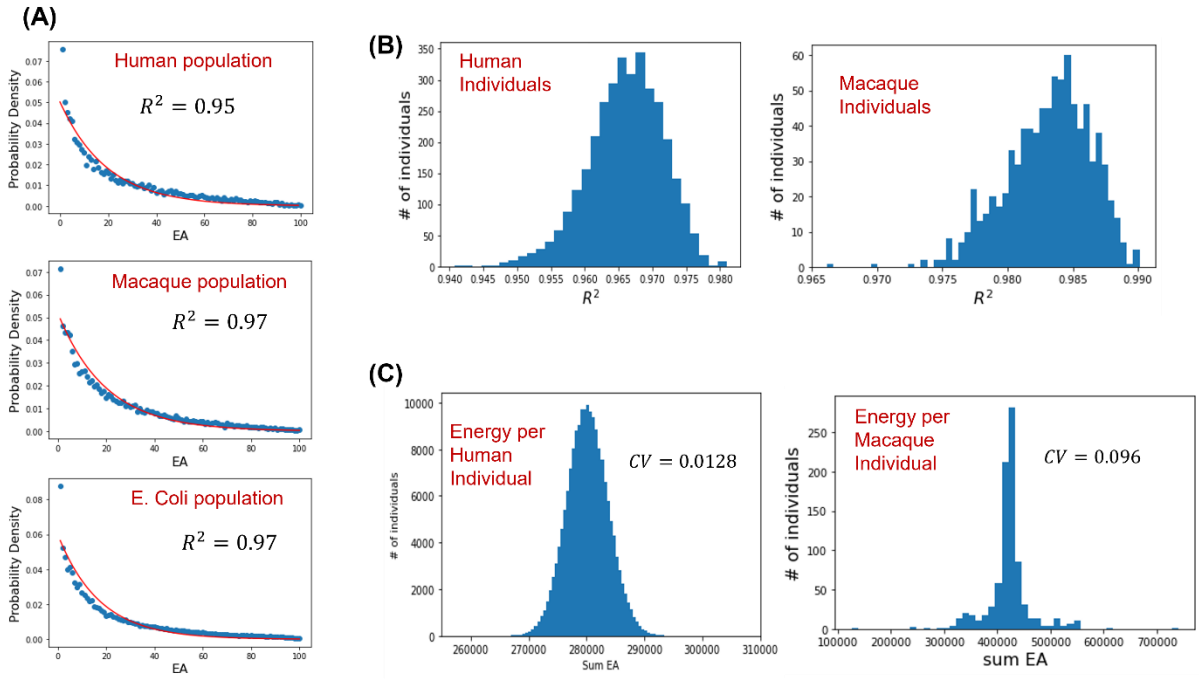


Figure 13: EA obeys thermodynamic expectations. (A) EA distributions of variants across Human, *Macaque*, and *E. Coli* populations follow Boltzmann distribution. (B) Individuals from the Human and *Macaque* population also follow Boltzmann distributions with high R^2 of fit. (C) Distribution of total energies across all variants per individual shows that energies of individuals are narrowly distributed in Humans and *Macaques*.

While the observations above suggest compliance of biological systems to statistical mechanical predictions, these have been primarily under normal, wild-type phenotypes. In order to characterize the properties of ensembles of mutations in non-wild-type settings, we turn to somatic mutations in cancer. Cancers are considered to evolve by accruing large numbers of mutations [201]. Large selective pressures from the host, such as immune response, can help drive tumorigenic mutations [202]. As a result, the distribution of variants and their EAs may break away from the tendencies expected from thermodynamics. To investigate this, we gathered data of EAs calculated for all coding mutations from 9073 patient cancer genomes across 33 cancer types from The Cancer Genome Atlas (TCGA) database [203–206]. We then plotted the distribution of all somatic coding variants for four well-established tumor suppressor genes: *TP53, PTEN, NOTCH1,* and *CDKN2A*. We found that the distribution of variant EAs in these genes diverged from the expected Boltzmann distribution (**Figure 14A**). To further investigate this observation in other biological systems, we examined coding mutations developed by E. Coli populations under the selective pressure of antibiotic treatment from colistin [207]. We gathered EAs of all coding variants for ten well-established colistin antibiotic resistance driver genes [208]: *basS, basR, asmA, ispB, lpxD, lapB, waaQ, ybjX, ynjC,* and *osmE*. We found that mutations acquired by E. Coli strains under the influence of selective pressure from colistin treatment broke away from the expected Boltzman distribution, tending to have higher EAs (**Figure 14B**). Conversely, when the coding variants of the same ten genes were gathered in E. Coli strains which were not exposed to colistin, we found an adherence to Boltzmann distribution (**Figure 14B**). These data confirm our expectation that traits under strong selection are no longer at a steady-state and therefore will not adhere to the Boltzmann distribution. Overall, suggesting that understanding mechanisms of statistical mechanics can be usefully leveraged to identify genes related to specific non-wildtype phenotypes, including disease states.

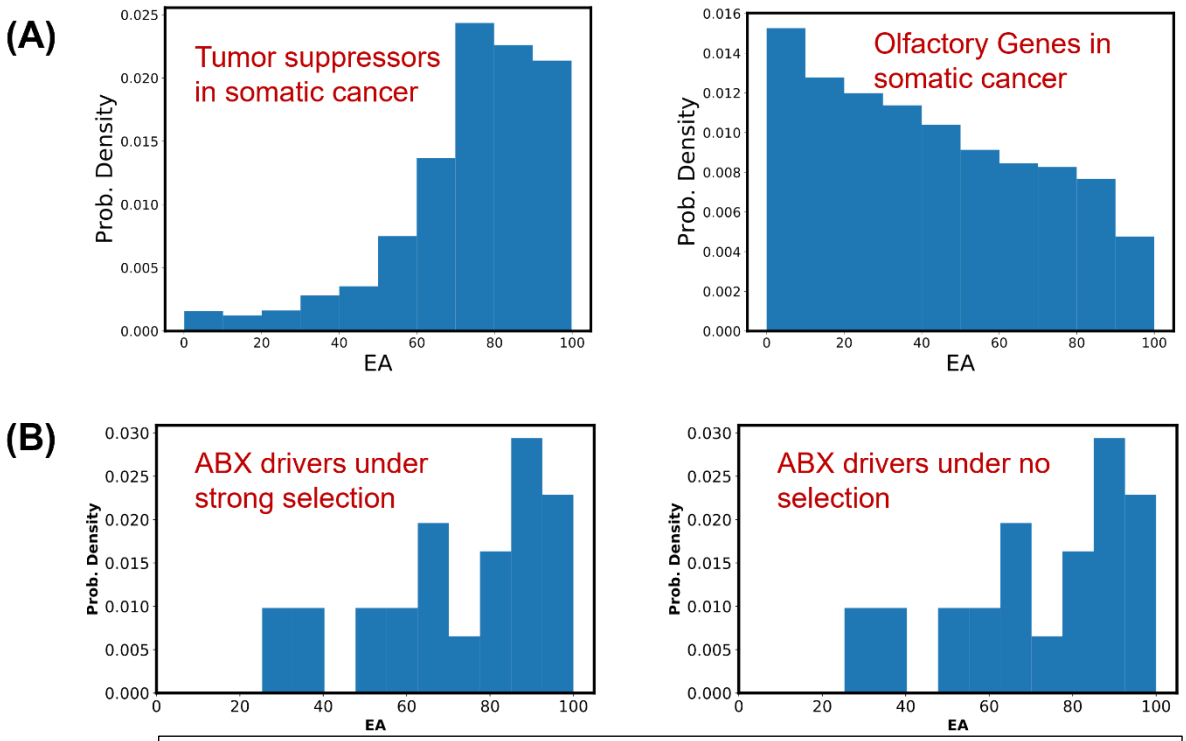


Figure 14: Strong selective pressure induces non-Boltzmann statistics. (A) Distribution of EA scores over well-known cancer driver genes (left) and olfactory receptor genes (right) in somatic sequencing from TCGA. (B) Distribution of EA scores for antibiotic resistance driver genes in *E.Coli* under selective pressures (left) and under no selection pressure (right).

Chapter 3.2.2: Equipartition Theorem Informs Mutational Intolerance

In classical statistical mechanics, the theorem of equipartition states that, if in thermal equilibrium, the energy of a system should be distributed equally among all degrees of freedom. To recast this theorem for biological systems, we first propose that, in the view of coding variants, we may consider independent genes to be degrees of freedom through which populations navigate a fitness landscape. Accordingly, we can extend equipartition to suggest that sum total EA (energy) of all variants in a population (system) should be distributed equally among all genes (degrees of freedom). Interestingly, after computing the sum of all coding variant EAs for each gene across the full population of the UKB, we found that instead of collapsing to one total EA value, genes showed a wide distribution of total EAs. This, however, is not unexpected as our estimation of EA is incomplete as we had considered all genes to contribute equally to phenotypic fitness. As a result, we propose to force populations into equipartition by introducing a new term describing the mutational intolerance or mutational inertia of a gene using the following framework. Let E_T denote the sum total EA of all coding variants in a given population and G denote the number of independent genes in the species' genome $\mathbb{G}$. The expected EA of any gene g according to equipartition is given by the constant $E_G = \frac{E_T}{G}$. The observed EA of a gene g in the population is given by

$$\epsilon_g = \sum_{i\ \forall i \in V_g} EA_i \tag{23}$$

where V_g is the set of all variants observed in g. We then equate the expected EA to the observed EA through an intermediary factor:

$$E_G = \mu_g \cdot \epsilon_g \tag{24}$$

where μ_g is a gene specific coefficient which enforces equipartition. Further examining equation (24), we observe that if $\mu_g > 1$, g is observed to have less mutational energy than

expected. Similarly, if $\mu_g < 1$, g is more mutable than expected. Together, these observations suggest that μ_g is a quantification of a gene's relative intolerance, or resistance, to mutation.

Chapter 3.2.3: Biological Relevance of Mutational Intolerance

In order to test this hypothesis, we applied the above framework to estimate μ for every gene in the genomes of three species including humans, *macaques,* and *E. Coli* looking specifically at rare variants*.* Across the three species, we observed that most μ scores were close to zero with small tail of high μ values (**Figure 15A, B, C**). This suggests a ubiquitous property across species wherein relatively few genes are intolerant to mutations while most have some level of robustness to mutation. Moreover, this suggests that statistical mechanics can offer a universal quantification of the empirical observation that the arrival of mutations in evolution is not random. Next, in order to assess whether μ scores agreed with observations from previous studies that essential genes are more connected in the protein-protein interaction (PPI) network [209–211], we correlated the μ scores with network connectivity. First, we ranked all human genes by decreasing μ and split them into 100 bins, so that bins were ordered relative to their constituent genes. For example, the first bin contains the top 1% of ranked genes and the last bin contains the bottom 1% of ranked genes. We generated a human PPI network by querying the STRING database [182] and calculated the average degree and eigenvector centrality of each bin. Interestingly, we found strong monotonically decreasing relationships between μ ranking and both measures of network connectivity (p_{degree} = 5.4e-56, p_{eigen} = 1.1e-56) (**Figure 15D**). This finding was further replicated in *macaques* and *E. Coli* (**Figure 15E, F**) (p = 4.1e-48 - 1.9e-15). These data are consistent with previous studies [210,211], showing that genes which are intolerant to mutations are more connected and influential in biological networks than genes which are tolerant to mutations. This suggests that genes with high μ are ubiquitously and intimately involved in core processes fundamental to human, *macaque*, and *E. Coli* biology.

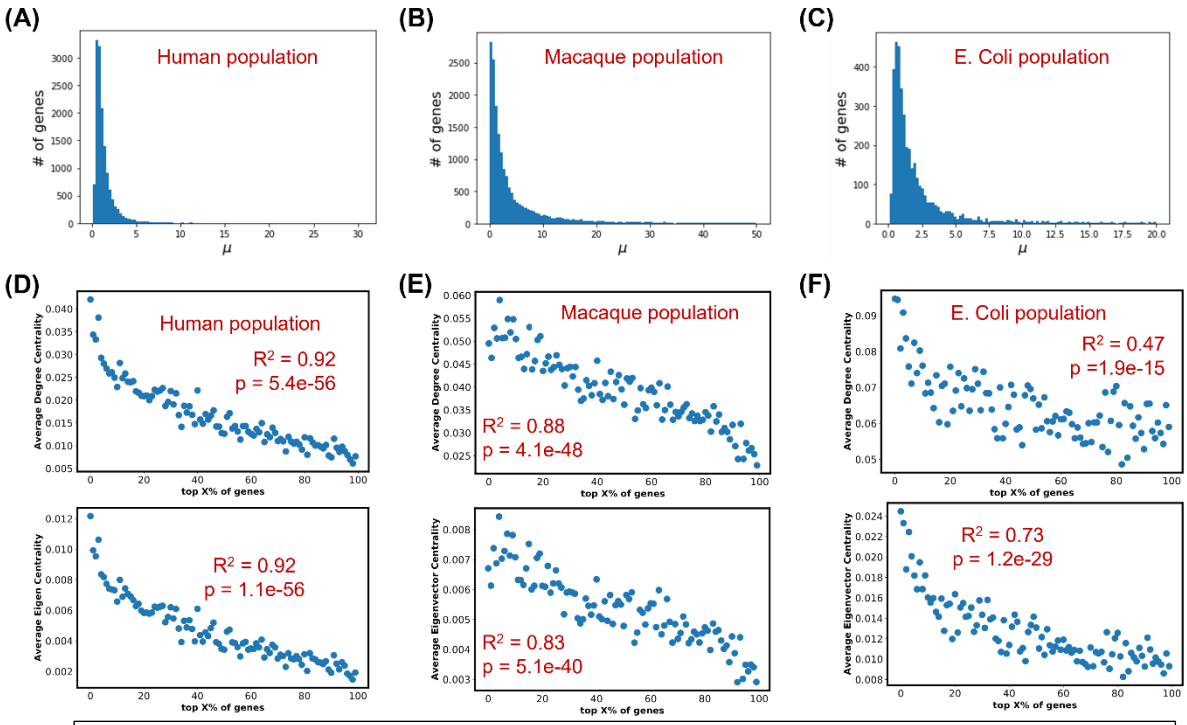


Figure 15: Properties of μ distribution. Distribution of μ scores for (A) Human, (B) *Macaque*, and (C) *E. Coli* populations. Ordered and binned plot of average degree or eigen centrality vs μ ranking (from highest to lowest) for (D) Human, (E) *Macaque*, and (F) *E. Coli* populations.

In order to evaluate this hypothesis, we performed functional enrichment analyses on highly mutationally intolerant genes. We constructed a PPI network using STRING with the top 1% of human genes ranked by μ, corresponding to a threshold of $\mu > 6.25$ (**Figure 16**). Interestingly, we found significant PPI enrichment (p < 1e-16) in this network. After clustering by Markov Clustering algorithm (inflation = 2), we found significant enrichment (FDR < 0.05) in processes including: (1) alternative splicing (spliceosome, regulation of mRNA splicing) which, given that up to 95% of human multi-exomic genes undergo alternative splicing, is central to human biology [212,213]; (2) cell division (oocyte meiosis, mitotic cell cycle processes), and (3) proteosome (ubiquitin mediated proteolysis, regulation of protein catabolic processes) [210,213]. These enriched pathways are all highly conserved across species, some tracing back as far as the last common ancestor of the eukaryotes [214–217]. Next, we repeated this procedure for the top 1% of *macaque* genes by μ, we found significant enrichment (FDR < 0.05) in pathways involving: (1) translation, (2) spliceosome (mRNA splicing, spliceosome complex), and (3) proteosome (complex and regulation) (**Figure 17**). Similarly, for the top 1% of *E. Coli* genes ranked by μ and found significant (p < 1e-16) PPI enrichment (**Figure 18**). We also found significant enrichment (FDR < 0.05) in processes involving: (1) translation (ribosomal assembly, translational fidelity, peptide biosynthesis) and (2) energy metabolism (glycolysis, gluconeogenesis), both of which are highly pervasive and essential throughout the tree of life [218–221]. These data show that, across various branches of the evolutionary tree, genes with high μ tend to form functional groups involved in various core biological processes and are often essential for life. The apparent fundamentality of high μ genes suggest that they may be highly conserved and influence fitness if mutationally perturbed.

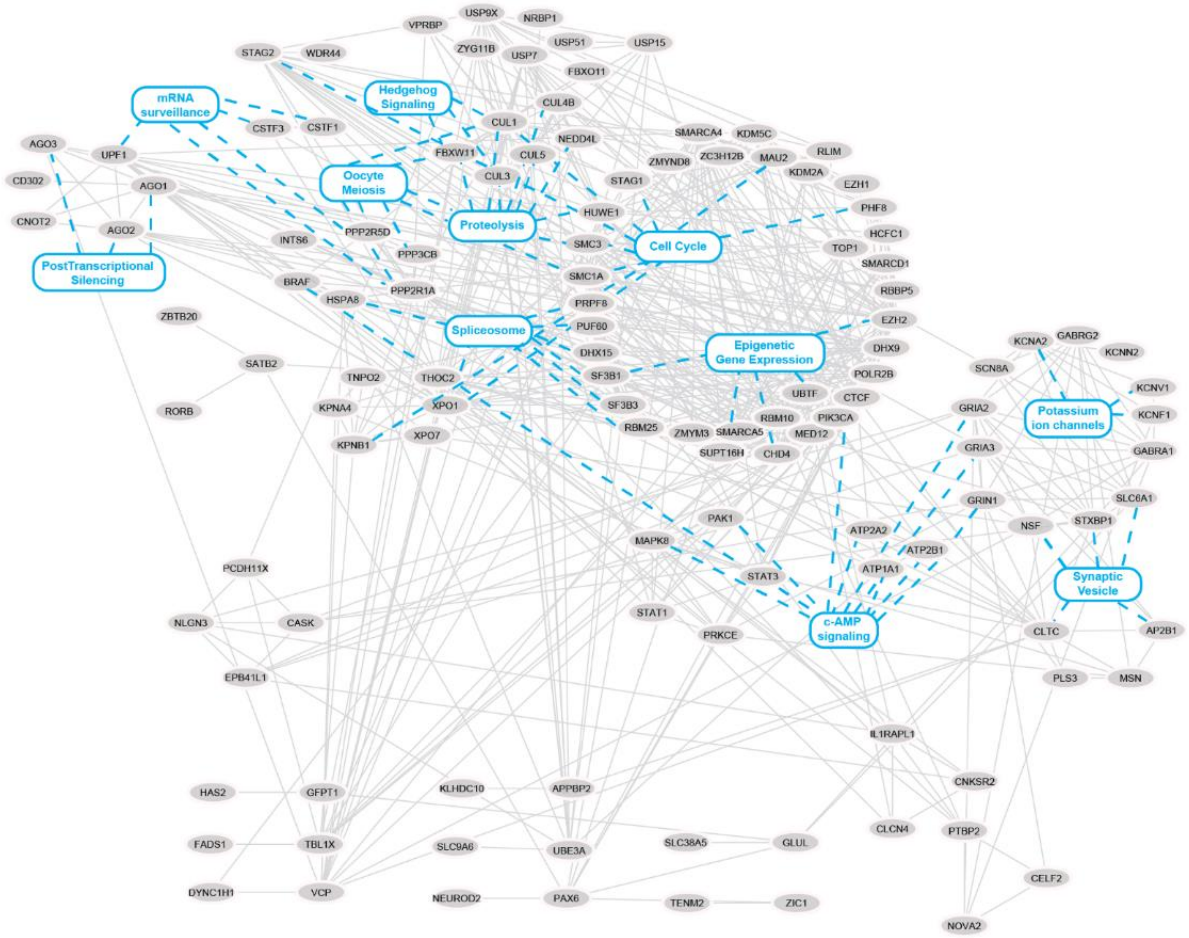


Figure 16: Network of top 1% of Human genes ranked by μ. Blue nodes indicate pathway enrichments and blue dashed edges indicate genes involved in the pathway

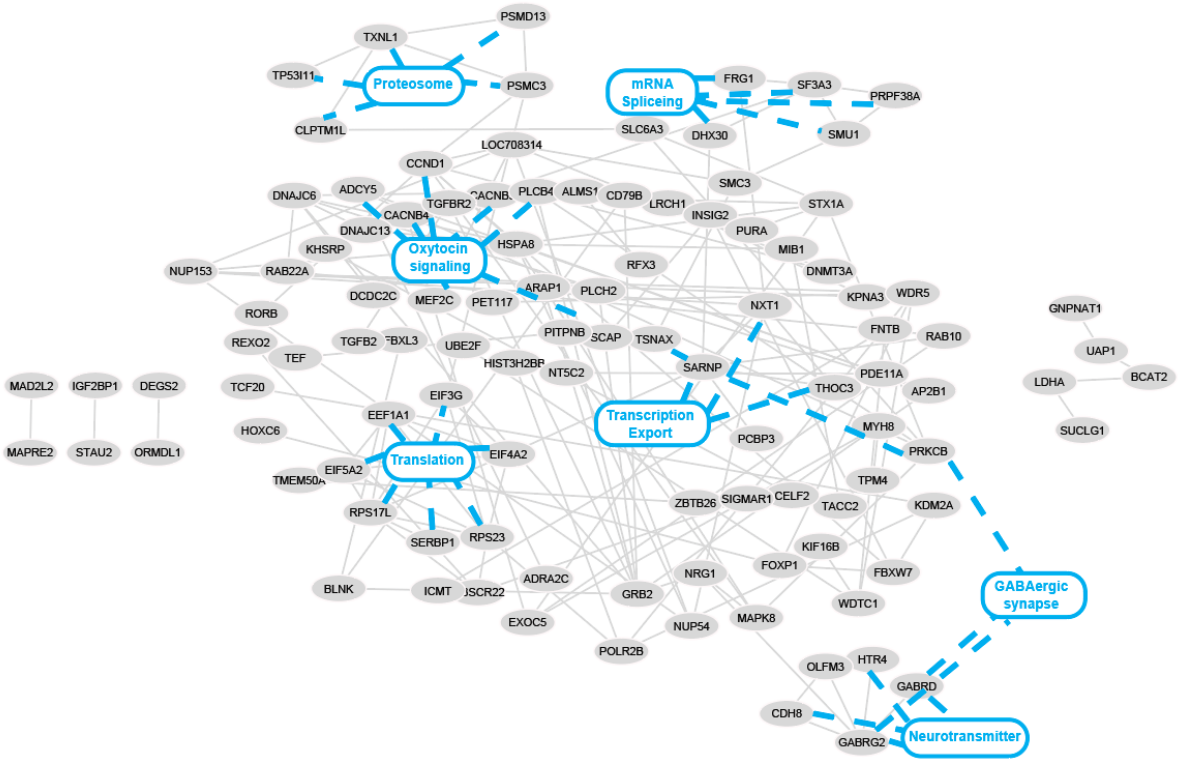


Figure 17: Network of top 1% of *Macaque* genes ranked by μ. Blue nodes indicate pathway enrichments and blue dashed edges indicate genes involved in the pathway

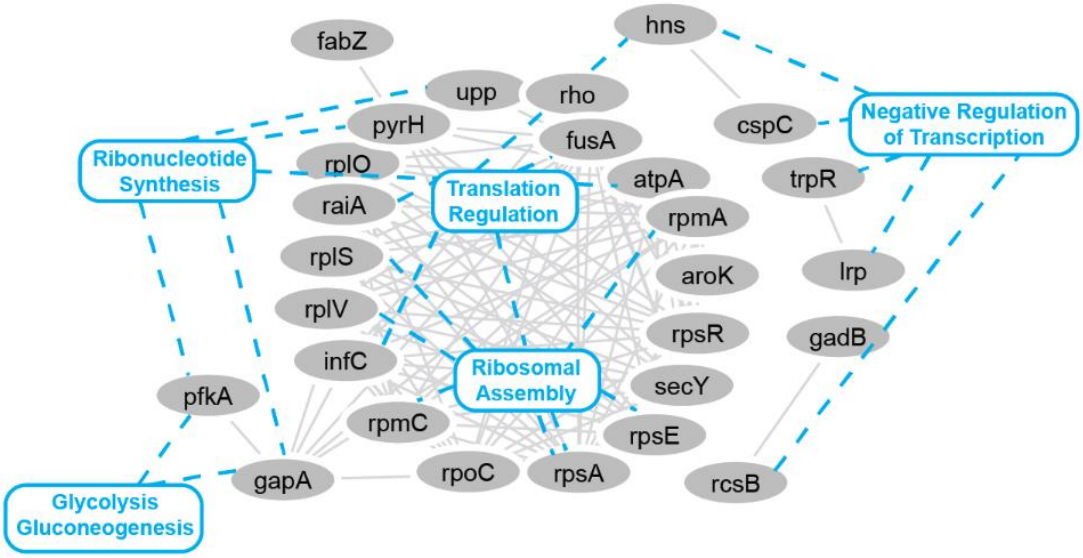


Figure 18: Network of top 1% of *E.Coli* genes ranked by μ. Blue nodes indicate pathway enrichments and blue dashed edges indicate genes involved in the pathway

In order to test this hypothesis, we asked whether μ could segregate genes that are known to cause Mendelian diseases in humans from those which do not cause disease. We curated the Online Mendelian Inheritance in Man (OMIM) database for haploinsufficient, autosomal dominant, and autosomal recessive disease-causing genes [222,223]. Any gene in the human genome that was not curated by OMIM, was defined as non-disease related (non-OMIM). We found that haploinsufficient Mendelian disease genes had significantly higher μ scores than non-OMIM genes (Mann-Whitney u-test p = 9.685e-24) (**Figure 19A**). We also found that μ scores were able to recover autosomal dominant genes with AUROC = 0.XXX. We repeated these experiments using Mendelian disease-causing genes taken from the independent ClinGen database and found that autosomal dominant genes had significantly higher μ scores than non-disease genes (Mann-Whitney u-test p = 3.141e-29) (**Figure 19B**). Next, to extend these observations to *E. Coli*, we compiled a set of 189 'essential' genes from the intersection of three independent analyses of gene essentiality in *E. Coli* [224–227]. All remaining genes in the *E. Coli* genome were considered non-essential. We found that essential genes had significantly higher (Mann-Whitney u-test p = 7.393e-23) μ scores compared to non-essential genes (**Figure 19C**). Moreover, we found that μ scores were able to differentiate between conditionally essential (genes which are essential for growth under particular conditions) [228] and non-essential genes with AUROC = 0.734. A similar segregation was observed between auxotrophic and non-essential genes, AUROC = 0.66 (**Figure 19D**). Together, these data highlight that genes with high μ tend to characterize genes which are essential or liable to induce disease if mutated.

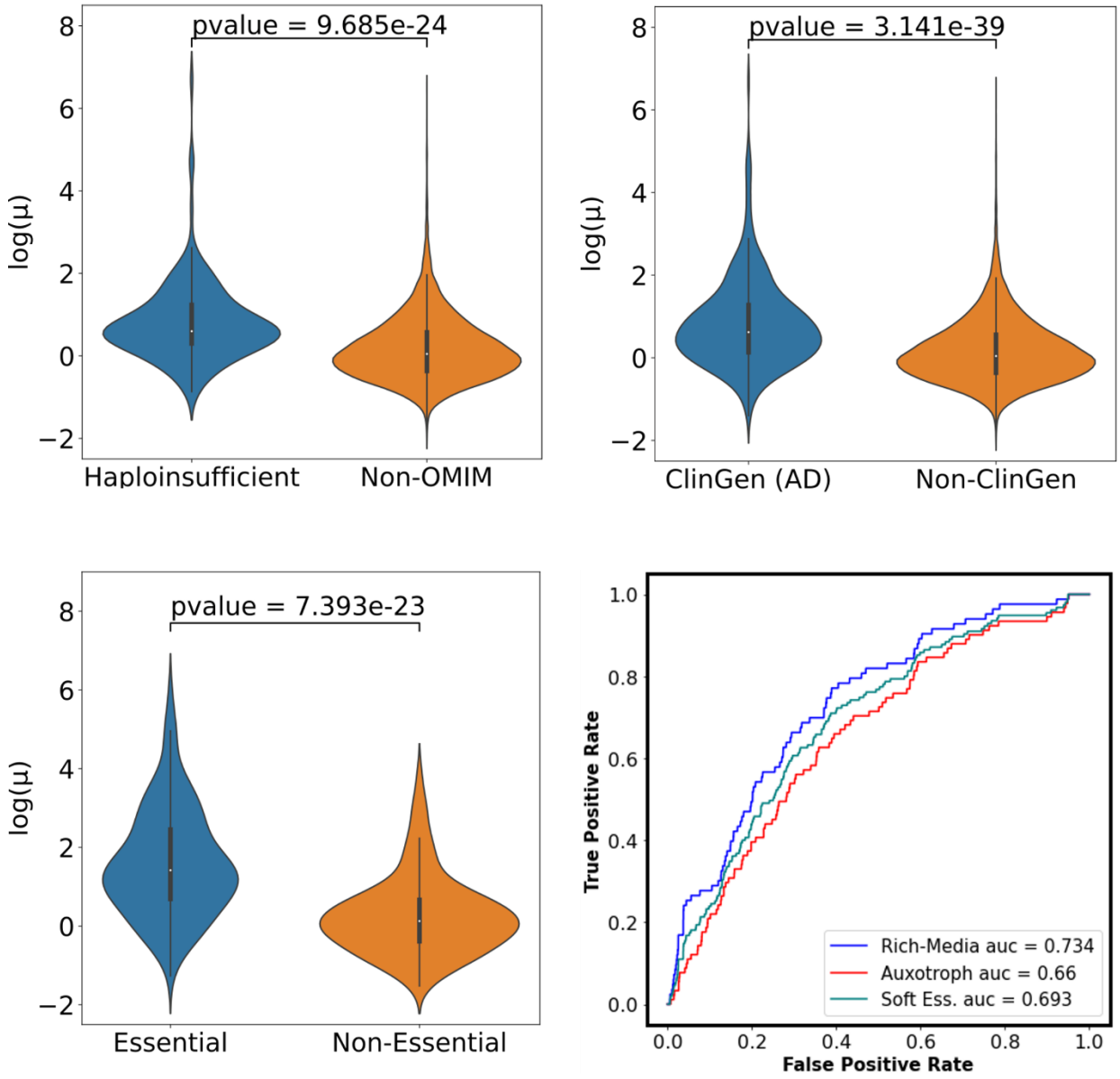


Figure 19: Distribution of μ in essential or disease genes is higher than non-essential genes. (A) distribution of μ scores in OMIM haploinsufficient genes vs non-OMIM. (B) Distribution of μ scores in ClinGen autosomal dominant genes vs non-ClinGen genes. (C) Distribution of μ scores in *E. Coli* essential genes vs non-essential genes. (D) AUROC of *E. Coli* conditionally essential, and auxotrophes.

Next, in order to directly and systematically assess the relationship between μ and the contribution of a gene to overall organismal fitness, we used a set of comprehensive genetic screens in *E. Coli*. Specifically, we downloaded the Rousset et al [229] high-throughput data set of gene repression via CRISPRi, wherein ~3,400 nearly ubiquitous genes in the *E. Coli* genome were systematically "knocked-down". Fitness of the resulting gene repression was measured via the portion change in sgRNA (log2FC) after direct competition. Notably, μ calculations were made on a distinct and independent dataset from the Rousset et al dataset. Strikingly, we found significant correlation (R^2 = 0.722, p = 4.63e-29) in the course-grain comparison of average μ versus average log2FC per percentile (**Figure 20**). These data demonstrate the power and ability of statistical mechanics to accurately and quantitatively characterize a gene's influence in the core biology of a species.

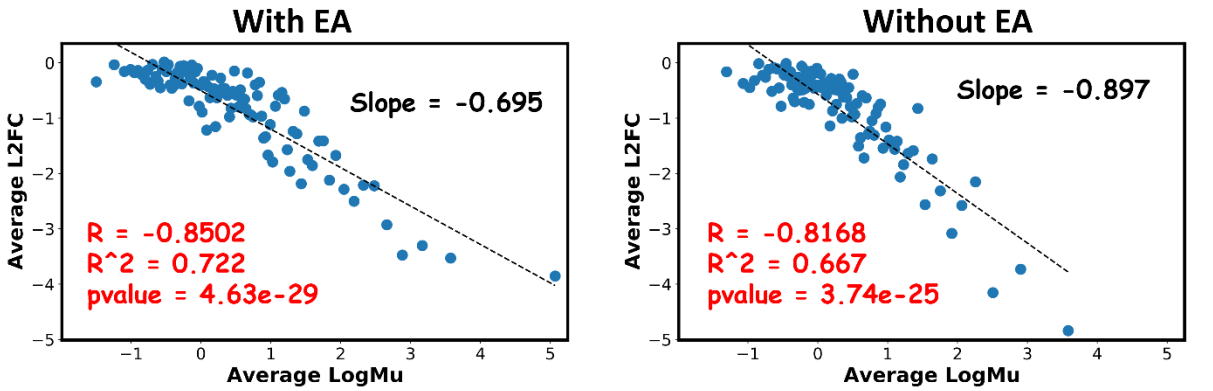
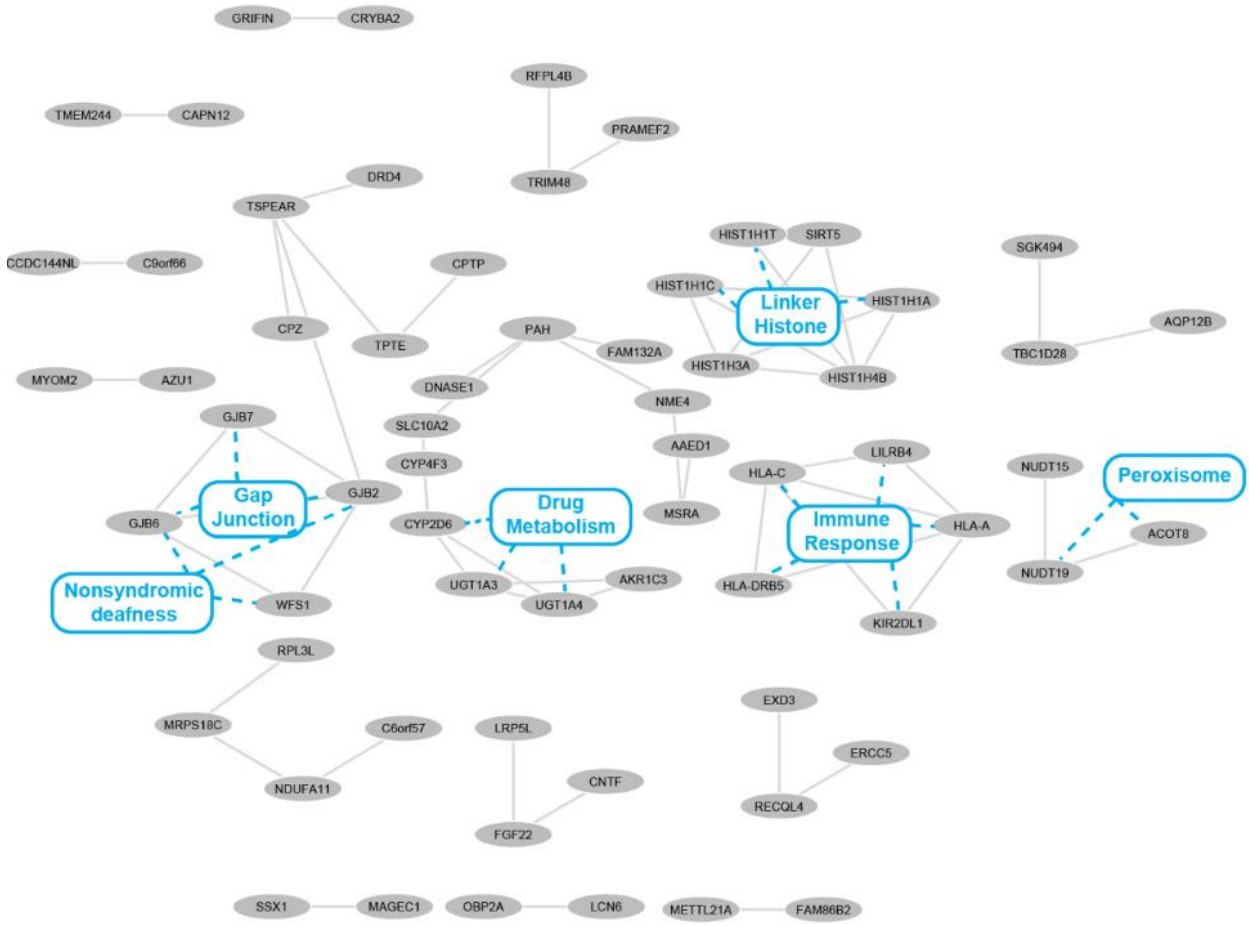


Figure 20: Correlation between binned average μ and average log2FC of sgRNA in genome-wide CRISPRi data of *E. Coli*. (Left) is μ calculated with EA while (Right) is μ calculated without EA

Despite the observation that genes with low μ seem to have low impact on overall fitness of the organism when depleted, the high degree of mutational tolerance among these genes may suggest a role in adaptive responses. To test this in humans, we constructed a PPI network using STRING with the bottom 1% of genes ranked by μ, corresponding to a threshold of $\mu < 0.32$. Here, we found significant PPI enrichment (p = 2.42e-5), suggesting bottom ranking genes are not randomly assorted (**Figure 21**). Following the same procedure as above, we found significant enrichment (fdr < 0.05) in pathways related to drug metabolism, primarily driven by a collection of *CYP* and *UGT* family genes which are known to be highly polymorphic and main contributors to variation in inter-individual drug response [230,231]. We also found enrichment in immune related processes involving a variety of *HLA* genes. The *HLA* system is one of the most polymorphic in the human genome [232]. Indeed, mutations in these systems are widespread and potentially result from evolutionary adaptation [233]. Furthermore, we found that olfactory pathway genes, which are known to be highly mutable and evolutionarily useful in adapting to niche environments [75,210,213,234,235], were systematically downranked by μ across the whole genome (**Figure 21**).



Olfaction pathway

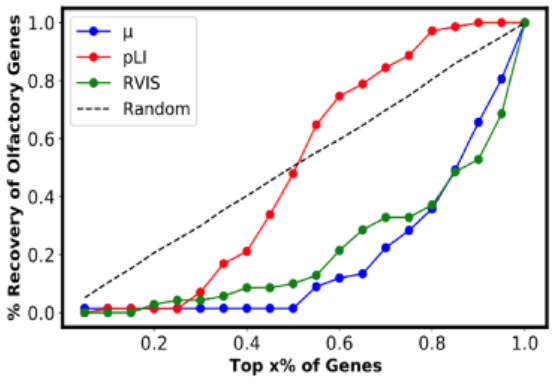


Figure 21: Network of bottom 1% of Human genes ranked by μ. (Top) Shows network of low ranking human genes. (Bottom) shows a plot of % recovery of olfactory genes vs percentile of their ranking

To further validate these observations across species, we investigated the bottom 1% of genes ranked by μ in *macaques*, we found significant PPI enrichment of *p* = 5.56e-5. Significant enrichment (fdr < 0.05) (**Figure 22**) was found in immune related pathways including antigen processing, innate immune response, and MHC class I and II complexes. Much like in humans, *macaques* have wide genetic diversity in immune related genes [236,237]. Similarly in *E. Coli*, corresponding to a threshold of $\mu < 0.22$. We found a significant PPI enrichment of *p* = 2.54 e-5 (**Figure 23**). We also found significant enrichment for the *E. Coli* toxin-antitoxin systems. Toxin-antitoxin systems have a wide variety of functions in bacterial organisms ranging from stress response to protection against bacteriophages [238]. Taken together, these data show that genes with low μ are not randomly segregated, rather they functionally cluster to highlight adaptive pathways, regardless of the vast biological differences between species. This suggest a universal principle that genes with low μ are highly tolerant of mutations owing to involvement in adaptive processes which need to be easily amenable to evolutionary and environmental pressures.

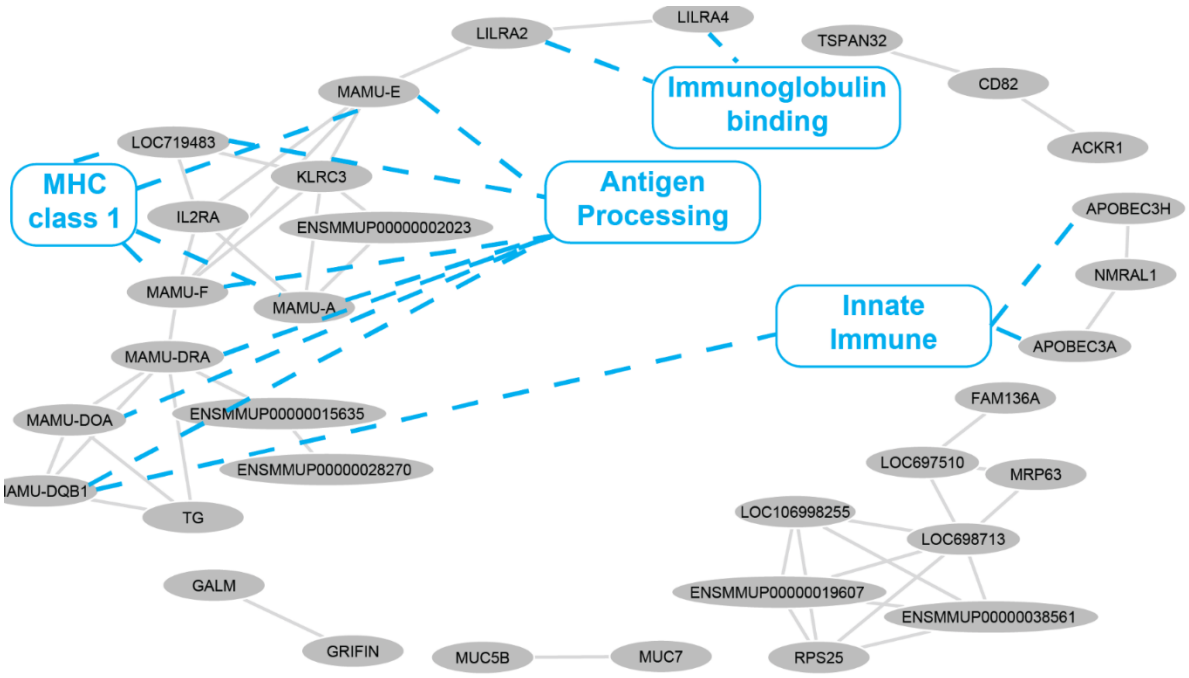


Figure 22: Network of bottom 1% of *Macaque* genes ranked by μ. Blue nodes indicate pathway enrichments and blue dashed edges indicate genes involved in the pathway

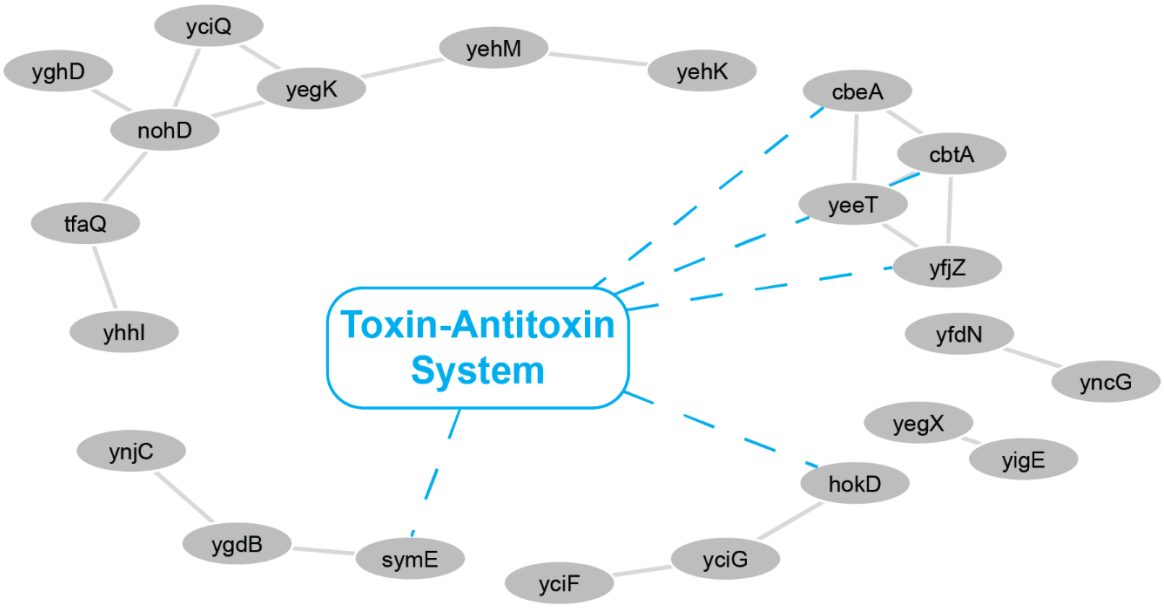


Figure 23: Network of bottom 1% of *E.Coli* genes ranked by μ. Blue nodes indicate pathway enrichments and blue dashed edges indicate genes involved in the pathway

Chapter 3.3: Discussion

Evolutionary processes and biases in the arrival of variation have molded a complex relationship between genotypic change and phenotypic fitness. However, evolutionary processes themselves, can be recast into the theory of learning. As argued by Vanchurin et al [239], under the theory of learning we can imagine any system which increases in complexity over time, from organisms to stars, to do so by predicting how to change to their environment. Such a similarity in the thermodynamics of learning and evolutionary processes have led to many previous studies demonstrating the compatibility of evolution and statistical mechanics [193,239,240].

Here, we follow along the lines of these studies, by starting from the principle of maximum entropy with an alternative measure of fitness effect (an analogue for free energy). We propose that rather than quantifying direct fitness, we may measure the change in fitness due to change in genotype. When applied to WGS/WES cohorts across three different species, we found that distributions of variant EAs fit well to the expected Boltzmann distribution (R^2= 0.95-0.97), seemingly maximizing the entropy of the system. Moreover, individuals in both *macaques* and humans, showed narrowly distributed total EAs, suggesting that individuals of a population may be constrained in the total amount of mutational change allowed relative to the mean. While these findings seem to support the notion of EA as a ΔEnergy and agree with theoretical observations of previous studies [193,239,240], the key contribution of this study is the predictive quantification of the variable contribution of genes to organismal fitness through the extension of the Equipartition theorem. This derivation suggests that μ may be an estimation of the gene contribution component of EA which was previously set equal to 1. The hypothesis that μ represents a gene's mutational inertia, reflecting its central or adaptive role in an organism's biology is supported by the striking observation that μ correlates directly with altered fitness of gene knockdowns in *E. Coli.* This

hypothesis is further reinforced by the findings that genes with high μ are enriched for core and conserved processes like alternative splicing, cell division, and translation while genes with low μ were involved in adaptive processes like immune or environmental response. Together, the findings of this study combined with those of previous thermodynamic studies of evolution suggest that the correspondence between evolution and statistical physics goes beyond analogies and advocate for a reevaluation of the evolutionary process as a thermodynamic system.

Though our findings demonstrate the predictive power of thermodynamics, it should be noted that this study only considered the evolutionary dynamics of protein coding DNA. The theoretical framework itself is not limited to protein coding regions, however, EA can currently only estimate mutational impacts in coding sequences. While future work will be dedicated to extending this to non-coding sequences (e.g. functional non-coding RNA [241]), a potential alternative is to use deep learning approaches to estimate the impact of non-coding mutations [242–244].

Chapter 3.4: Materials and Methods

Chapter 3.4.1: Whole Exome/Genome Sequencing Data

Chapter 3.4.1.1: UKB Human Dataset:

Human WES data was downloaded from the UK Biobank 200K WES release [196–198]. First, following standard GATK procedures, we remove variants with filtered depth (DP) < 10 or GQ < 20 [245–247]. In accordance to guidelines, we also removed variants with an inbreeding coefficient > 0.03 and genotyping rate < 0.95. Lastly, for variant quality control, we removed variants with Hardy-Weinberg Equilibrium exact test p-value < 5e-8 [178]. We also performed quality control on samples by removing genetically related samples through kinship coefficients and identical-by-descent testing as well as thresholding based on variant number,

Ti/Tv ratios, and missingness. Further outliers will be removed by principal component analysis. Calculation and filtering based on these quality control statistics was done through a combination of BCFTOOLS [248], PLINK2 [249], KING [180], and Eigenstrat [250,251]. Variants were then annotated using VEP [252] and EAs were subsequently calculated.

Chapter 3.4.1.2: *Macaque* WGS Dataset:

WGS data for 853 individual *macaques* was acquired from Warren et al [199]. Accordingly, the sample acquisition, sequence assembly, and quality control are all described in detail within the referenced publication [199]. Variants were annotated and mapped to *macaque* proteome using VEP and Ensembl biomart web tool. EAs were calculated given the reference protein and amino-acid sequence variation induced by the sequenced missense mutations.

Chapter 3.4.1.3: E. Coli WGS Dataset:

Environmental E. Coli strain WGS data was downloaded from the European Nucleotide Archive PRJEB232924 and Moradigaravand et a [200]. From this dataset, we extracted WGS data for 255 environmental E. Coli isolates. We then followed the protocol set up by Marciano et al [207]. First, we performed genome assembly on the WGS data using the SPAdes genome assembler [253]. Next, open reading frames were predicted with GeneMarkS-2 [254] and removed. Missense mutations were annotated using the procedure and code library of Marciano et al [207], wherein each missense mutation was mapped to the E. Coli K-12 MG1655 proteome. After removing sequence variations which were unable to be mapped, EA was calculated and annotated for each missense mutation.

Chapter 3.4.2: Measurement of goodness of fit to Boltzmann

In order to assess whether distribution of EAs across all variants in a population fit the predicted Boltzmann distribution, we calculated a histogram over the EA distribution. We then

fit, to the probability density and EA value of each bin, a Boltzmann function which was optimized to best fit over the kT parameter:

$$p_i = \frac{e^{-\frac{\epsilon_i}{kT}}}{\sum_j^M e^{-\frac{\epsilon_j}{kT}}} \tag{25}$$

where p_i is the probability density of bin i, ϵ_i is the EA of bin i, and M is the total number of bins. Given that binning effects the goodness of fit, we binned to the resolution offered by EA (100 bins). In order to further validate that R^2 was not drastically affected by binning, we plotted the R^2 as a function of number of bins **(Figure 24)** and found a plateau in R^2 at bin size 80. This process was repeated for populations of humans, *macaques,* and *E. Coli*, as well as individuals of the human and *macaque* cohort.

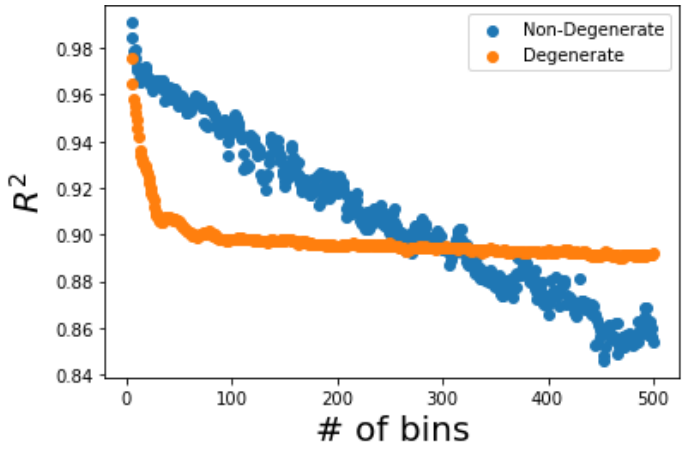


Figure 24: Plot of # of bins used to fit Boltzmann vs R^2 of the resulting Boltzmann fit. Plot shows both degenerate and Non-degenerate fits. This study uses only degenerate distributions

Chapter 3.4.3: Calculation of μ:

In order to evoke equipartition and measure the mutational intolerance of genes μ, we first calculated the expected EA of any given gene as $E_G = \frac{E_T}{G}$, where E_T is the total EA across all variants of all individuals in the population and G is the total number of genes in the species. To further normalize for coding gene length, we modified the denominator to be the total number of amino acids (AA) in the proteome of the species. For all three species used in this study, proteome data was acquired from Ensembl database [255]. Similarly, the observed EA per gene, equation (23), was also normalized by the amino acid length of the encoded protein: $\epsilon_g = \frac{1}{L_g}\sum_{i\forall i\in V_g} EA_i$. μ is then calculated as the ratio of $\frac{E_G}{\epsilon_g}$ for each gene g. For human UKB cohort, we implement an allele frequency threshold of AF < 0.001 to selectively filter out genes that are more common and investigate 'newer' mutations. Due to the vast difference in cohort size, for *macaque* and *E. Coli* data, we investigate variants that pass a less stringent allele frequency threshold of AF < 0.05.

Chapter 3.4.4: Correlation of μ with network centralities:

In order to assess the relationship between a gene's μ and various measures of its network centrality, we downloaded the STRINGv11.5 database for humans, *macaques,* and *E. Coli* [182]. Separately for each species, we ranked genes by their μ in a descending fashion. We then binned the ranked gene list into 100 ordered bins so that the first bin would contain the top 1% of genes ranked by μ and the last bin would contain the bottom 1% of genes. We then measured, for each bin, the average degree and eigenvector centrality of the bin's constituent genes. These average centralities were then plotted against the percentile value of each bin and monotonic decreasing tendency was measured by Spearman correlation test through Python's SciPy library [256].

Chapter 3.4.5: Pathway Enrichment Analyses:

For each species, protein-protein interaction network of the top 1% of genes ranked by μ were generated using the STRINGv11.5 database [182]. Markov Clustering was performed in Cytoscape using ClusterMaker2 library [257–259]. Functional and pathway enrichment analysis of each cluster was then performed by FUMA [260]. Lastly, network representations were created with Cytoscape [257].

Chapter 3.4.6: Characterization of μ differences in Mendelian Diseases

In order to test whether mutational perturbations of genes with high μ in humans lead to disease states, we downloaded the OMIM database and utilized Petrovski et al [223] to categorize genes into categories of haploinsufficient, autosomal dominant, and autosomal recessive. We then defined any gene not annotated in the OMIM database to be non-disease related. We then measured the difference in μ values of genes in OMIM disease categories against non-disease using a Mann-Whitney U test implemented through the SciPy library [256]. Significance of difference in distributions was measured against all non-disease genes and against a randomly selected subset of non-disease genes of equal size to the OMIM disease category tested (**Figure 19**).

Identical experiments were performed for Mendelian disease genes curated from the ClinGen database [261]. Curated genes were filtered for ‘Strong’ or ‘Definitive’ strength of association to Mendelian disease and categorized as autosomal dominant or autosomal recessive based on the mode of inheritance flag.

Chapter 3.4.7: Characterization of μ differences in Essential, conditionally essential, and auxotrophic genes in *E. Coli*:

In order to test essential genes tended to have higher μ scores than non-essential genes, we curated three separate studies of genome-wide *E. Coli* gene essentiality [224–227]. We further defined a gene to be essential if it was identified in each of the three studies,

resulting in 189 essential genes. All other genes in the *E. Coli* genome were designated as non-essential. We then measured the difference in μ distribution of (1) essential genes vs all non-essential genes, and (2) essential genes vs random subset of non-essential genes of the same size (**Figure 19**) using a Mann-Whitney u-test.

We further extended this experiment to "pseudo-essential" genes including conditionally essential and auxotrophic genes. These two gene lists were taken from Nichols et al [228]. Non-essential genes were defined as all other genes in the *E. Coli* genome (bar essential genes defined above). Mann-Whitney u-test was used to measure differences in μ distributions among gene sets.

Chapter 3.4.8: Direct correlation between μ and gene fitness effect:

In order to measure a direct relationship between a gene's μ and its fitness effect, we acquired the Rousset et all dataset of systematic CRISPRi gene 'knock downs' for nearly every ubiquitous gene in the *E. Coli* genome [229]. In their study, Rousset targets each gene with a library of 2-4 sgRNAs per gene. The resulting gene repression's fitness effect was measured by a log2 fold-change (L2FC) in sgRNA after direct competition. In order to generate a gene-specific fitness measure, we averaged over the 2-4 sgRNA L2FCs of each gene. We then measure the spearman correlation between L2FC and μ, which was found to be significant. In order to measure the strength of the relationship between μ and L2FC in a more course-grain manner, we ranked all genes by μ in descending order and organized them into 100 ordered bins. We then measured the spearman correlation between the average μ and average L2FC of each bin.

Chapter 3.4.9: Rank Bias of Olfactory Pathway genes:

To test whether μ downranks highly mutable and adaptive genes, we curated the Reactome database [262] for olfactory pathways. All genes related to these pathways were

defined to be ‘olfactory related genes’. All genes in the human genome were then ranked by μ in descending order. We then measured the percent recovery of ‘olfactory related genes’ at every 5-percentile interval between 0^{th} percentile to 100^{th} percentile. In order to compare the performance of μ to other standard tools, we repeated the experiment with pLI and RVIS [210,223].

Chapter 4: Discussion and Future Directions

Recent years have seen an explosion of genomic research both in the fundamental understanding of the biology and development of methods and tools to probe deeper. The relative ease and cost-effectiveness of modern sequencing techniques have led to large scale sequencing consortia reaching cohort sizes of nearly half a million sequenced exomes [196–198]. Our understanding of the genetic etiology of complex diseases has benefited from such large sequencing efforts. The past decade has seen many genome-wide association studies (GWAS) of complex diseases with impressive sample sizes. For example, the Psychiatric Genome Consortium analyzed the genomes of 150,000 individuals to identify 108 schizophrenia-associated loci [263] . Kunkle et al [78] identified five novel and validated twenty known Alzheimer's disease (AD) associated loci by investigating ~94,000 individuals WGS. An analysis over 41,000 individuals with bipolar disorder and 371,000 controls by Mullins et al [264] identified fifteen disease associated genes. Similarly, the CARDIoGRAMplusC4D consortium identified 64 coronary artery disease related loci over ~190,000 individual genomes [265]. Despite such monumental efforts, our understanding of the genetics of many complex diseases remains incomplete, able to explain only a fraction of expected heritability [PMID: 19812666, 23835440]. Several theories for the "missing heritability" exist including that single nucleotide variants of small effect may be missed because of multiple testing correction stringency [6,266]. However, a particularly interesting hypothesis suggests that our understanding of the genetic underpinnings of complex diseases may be aided by investigating gene-gene interactions [18].

Genetic interactions are interactions between gene variants that result in a phenotype which is significantly different from the individual phenotypes of each variant. In this way, otherwise unremarkable variants with little individual effects can combine to give rise to complex phenotypes [267]. Coupled with the recent wave of advancements around graph-

based learning [31,268], interrogation of genetic interactions presents a potentially fruitful path forward to a better understanding of complex genetic diseases. Proper innovation, however, is necessary to leverage these advances and bridge the two fields. Here, I develop and implement methodologies to do this. First, I motivate and facilitate the translation of node embedding principles to biological settings and architect GeneEMBED to harness this and discover new disease-gene associations. Identification of these associations, however, rests partially on the quantification of functional impact of mutations. Though variant impact predictors have become complex and wide in variety, predictions are often made relative to the affected gene/protein. Next, to attend to this issue, I propose a metric by which to measure the relative contribution of genes to organismal fitness, inspired by recent equivalences established between evolution and thermodynamics [193–195,239,240]. In addition to providing a useful measure, this characterization of genes provokes deeper questions about the nature of evolutionary dynamics and their relation to statistical physics. Together, these efforts are intended to provide valuable new tools with which to probe the genetics of complex diseases and, perhaps more importantly, foster different and unconventional ways to explore genetics.

GeneEMBED and exomic analysis

As described previously, GeneEMBED was designed to aid in the elucidation of the genetic underpinnings of complex diseases. Specifically, GeneEMBED works to pinpoint genetic risk factors of disease by examining the differential perturbation patterns of gene interactions between healthy and affected populations. In our, proof-of-concept demonstration of GeneEMBED, we analyzed Alzheimer's Disease (AD). As the leading cause of dementia worldwide, AD is a neurodegenerative disorder leading to memory loss, language difficulties, and even behavioral issues [76]. AD is currently the sixth leading cause of death in the US and is projected to affect nearly 12 million individuals by 2050 [77]. The rapidly rising

prevalence of AD must be met with effective therapeutic targets, biomarkers, and risk prediction tools. Yet, these efforts are hindered by our overall lack of understanding of the genetic origins of this highly heritable disease [9,10]. To this end, we applied GeneEMBED to two separate AD cohorts taken from the Alzheimer's Disease Sequencing Project (ADSP). Additionally, we tested the flexibility of GeneEMBED with two different variant impact scoring (VIS) systems (EA and PPh2) and three different protein-protein interaction (PPI) networks. Strikingly, we found that GeneEMBED was robust to variations in VIS, PPI, and even cohorts, successfully, and reproducibly, identifying AD related genes. Among the set of repeatedly identified genes in various conditions, we identified a set of 143 high-confidence candidate disease-associated genes. These genes differentially expressed in bulk brain tissue, as well single cells of AD cases. We further found that candidate genes lead to abnormal neurological phenotypes when knocked out in mouse models, suggesting a neuroactive, and potentially, AD related role. To specifically query their role in AD, we investigated candidates *in vivo* using two *Drosophila* AD models and found that 65% of testable genes altered neurodegeneration in flies. Moreover, many of the candidates were targeted with pre-existing, FDA approved activators or inhibitors. Noteworthy among these druggable candidates were *TP53* and *POLD1*, both of which alleviated neurodegeneration phenotype in *drosophila* AD-models when knocked out. Functional inhibition of p53 has been previously suggested as a therapeutic strategy [269] owing to the role of *TP53* in AD pathology [270,271]. In addition, we also identified two genes novelly associated with AD genetics, *PLEC* and *UTRN*. While *PLEC* has been seen to lead to reduced learning and memory as a result of increased *tau* accumulation [156,272] (a key component of AD pathobiology), not much is known about role of *UTRN* in *tau* tangle development.

In light of its robust ability to re-discover known genes associated with AD and identify novel risk factors, we hope that GeneEMBED is widely usable in other case-control studies of

germline contributions to complex genetic diseases. Despite this, we also acknowledge that this study, like others, has limitations. First, all associations identified are computational. *In vivo* experiments and mouse knock-out data suggest a potential role for candidate genes in AD, these data are not sufficient for therapeutic targeting. Further, in-depth, biological characterization will be imperative to explain their roles in neurodegeneration. Second, incorporation of network information, while core to the innovative nature of the approach, presents a potential point of error. Network data contributes greatly to the identification of candidate disease genes. As a result, if a network is utilized which does not reflect key components of the disease process, identified candidates will likely be uninformative. In relatively well studied diseases, these key components may be known. For example, a network suitable for analysis of AD should contain genes such as *APOE, MAPT,* and *TREM2*. However, this point of error is avoidable by using comprehensive PPI networks such as STRING. Though these networks may have higher rates of false positive (FP) edges compared to unbiased networks built through high-throughput screens, GeneEMBED's robustness to FP edges allows them to be used effectively. Lastly, another limitation of the study is that it only considers coding mutations. While coding mutations are vital to disease pathology, a growing body of literature suggests that non-coding variations also contribute greatly to disease. Thus, translating the strategies used in this study to analyze non-coding data would be an interesting direction of future research.

Limitations and Future Directions

The potential for further refinement of the methodologies and the questions they raise are exciting. One interesting direction of future research is the aforementioned extension of GeneEMBED to non-coding domains. There is an increasingly large supply of studies which offer tools to predict the functional impact of non-coding mutations. Many of these methods employ deep learning strategies [242,244], which, in most cases, ensures a probabilistic

interpretation of functional impact prediction. This would be directly compatible with the current GeneEMBED implementation. Moreover, coupled with the availability of molecular networks detailing the interactions of transcription factors, noncoding RNA, and proteins, extension of GeneEMBED to non-coding mutational domain may be readily testable. Another interesting question that arises naturally from this study is the ability of candidate genes to classify individuals by their disease status. While GeneEMBED itself is unable to perform classification, the basic principles can be extended for classification. The approach is to featurize each gene by its overall network neighborhood, annotated with mutations, and compare between cases and controls. Naturally, as this is a supervised learning task, graph neural networks could be used to build a classification algorithm. Graph neural networks with some attention mechanism scheme (e.g. Graph Attention Networks [38]) may be well suited for this task. Though architecture depth will need to be monitored carefully to avoid "oversmoothing" in the case of large, densely connected networks, embeddings of candidate genes could be easily concatenated and used for classification. Another approach may be to adopt a graph information diffusion scheme [29,30]. Here, the mutational burden quantified by some compositive variant impact score could be used as a "signal strength" to be diffused along a PPI network through a heat kernel on the spectrum of the graph. A vector detailing the diffusion profile of the signal across can be calculated and used as a feature vector for genes, which are then fed into a simple statistical machine learning classifier or feed forward neural network. While the prospects of future research are exciting, we also emphasize that we expect the current GeneEMBED implementation to be widely applicable across complex diseases and hope that it will aid in gaining a better understanding of the genetics of disease.

Equipartition and Mutational Intolerance

The use of variant impact prediction tools has become widespread. Many genomic studies, including GeneEMBED, incorporate one or more variant impact prediction tools in

their pipeline. Yet, oftentimes, variant impact prediction tools measure functional impact with respect to the affected protein. For example, both PPh2 and SIFT assess the effect of a missense mutation through measurements of protein structure and function. Under this framework, a loss of function mutation in gene *x* is fundamentally no different from a loss of function in gene *y*. We know, however, that there exist mutational biases within the human genome. For example, olfactory receptor genes are enriched for loss of function mutations compared to the rest of the genome [75]. To help gauge these relative differences, we developed a measure of mutational intolerance and, in doing so, invoked a field of population genetics which seeks to equate laws of evolution to principles of statistical thermodynamics. The argument for such a concordance proposes that "free fitness" may act as a proxy for "free energy" through equilibrium thermodynamics and ideas first presented by Edwin Thompson Jaynes in his seminal two-part paper in 1957 regarding maximum entropy thermodynamics [273,274]. Using the theory of Evolutionary Action (EA) as a starting point, we present a conceptual argument analogizing EA to energy, distinct from the arguments used in thermodynamics-evolution literature. Further supporting the argument, we empirically demonstrate in three different species that EA maximizes entropy and fits a Boltzmann distribution remarkably well. We also show that energies of different macrostates across the population are narrowly distributed, following thermodynamic expectations. Finally, extending equipartition to biological systems allowed us to quantify the relative mutational intolerance (μ) of genes. Strikingly, we found that μ accurately predicted the network influence (degree, betweenness, and eigenvector centralities) of genes, across the three species. More than that, we found that in experimental assays designed to quantify the fitness effect of individual genes by CRISPR repression in *E. Coli*, μ correlated strongly and significantly with fitness effect. Characterizing highly mutationally intolerant genes against mutationally tolerant genes, we found that intolerant genes ($high\ \mu$) were involved in core and highly conserved biological processes ranging from cell division to transcription. Additionally, we found that genes that

caused Mendelian diseases in humans had significantly higher μ than non-disease related genes. Similarly, in bacteria, we found that genes which were essential, conditionally essential, or auxotrophic had significantly higher μ than non-essential genes. Conversely, genes that were highly tolerant to mutations ($low\ \mu$), were involved in adaptive processes including immune pathways and toxin-antitoxin systems. Similarly, genes which are known to be highly mutable and involved in adaptation were preferentially ranked low by μ. Overall, these observations stand as experimental data supporting an equivalence between statistical mechanics and biology. Moreover, they suggest that further analyses of comparisons of statistical mechanics and evolution may lead to informative quantifications of biological systems.

Limitations and Future Directions

Much like GeneEMBED, the extension of equipartition to biology and the resulting quantification of mutational intolerance is limited by the focus on coding mutations. Non-coding DNA, however, plays an important role in biology and may also have conserved regions throughout evolution [275]. While μ is calculated only on exomic mutations, as these are the only ones whose functional impact is quantifiable by EA theory, the theoretical framework is not limited to coding mutations. The framework can be extended to any independent degrees of freedom (e.g. base pairs instead of genes). This presents an interesting direction of future research. One approach may be to utilize deep learning approaches to estimate the impact of non-coding mutations. If any of the existing methods naturally maximize entropy in their estimates of functional impact, they may be a sufficient proxy for estimating non-coding μ. Another interesting direction of future work would be to assess the utility of μ scores in aiding gene discovery and patient classification. This may be done simply by weighting the functional impact of a patient's coding variants by the μ score of the impacted gene. In this way, the frameworks of existing methods (e.g. GeneEMBED, iDEAL

[172], EPIMUTESTR [276]), will not need to be changed. Moreover, the default frameworks will also act as a control against which the improved (or worsened) performance of μ weighted inputs can be measured. Another, potentially more interesting, way to assess the ability of μ to recover disease related genes would be to measure μ separately in cases and controls. Genes which are often mutated and lead to disease state would be found to be "mutationally tolerant" in the case population and "mutationally intolerant" in the control population. Differences in these μ scores could be quantified and used to identify disease-associated genes.

Overall, the objectives of this dissertation were to develop tools to help gauge the involvement of genes in disease in a manner that was inclusive and cognizant of their interactive environment, mutational perturbations, and relative mutational intolerance. In pursuit of these objectives, I hope to have fostered exciting questions and interesting new ways to think about genetic analyses.

Appendix of Tables

Table 1: Candidate genes identified by GeneEMBED on various cohort-VIS combinations (column label) using the **STRING** network. Related to Figure 1.

Discovery - EA	Extension - EA	Discovery - PPh2	Extension - PPh2	Control vs Control - EA	Control vs Control - PPh2
APOE	APOE	RHOBTB3	APOE	PIK3CG	PIK3CG
NQO1	BLM	MYC	BRD2	ATM	CD86
CSF1R	FOXM1	NQO1	TP53	TLR4	MYO5A
MAPT	FLT3	CSF1R	GAK	NRG1	KDR
MYC	MAPK6	LRRC6	MAPK6	BRCA1	NRG1
PAICS	CSF1R	PAICS	BLM	WDHD1	CAD
RHBDF2	PIK3C2B	PRIM1	EPHA2	TTC28	TNS1
EHBP1L1	PDLIM4	TP53	BMP2K	EPHA1	RIPK3
CTSB	TLR2	MAPK6	FOXM1	FOS	TTC28
LRRC6	DCLRE1C	LRRC71	FLT3	FRZB	KNG1
TNNT1	PARP1	ANLN	CSF1R	TLR2	TRAP1
MAPK6	CES1	ERBB2	RET	RAD51C	HERC1
WDHD1	MYO9B	DLG1	TLR2	HERC1	OPRM1
ANLN	GNAS	GNB3	RIPK3	OR10T2	PDGFRA
BMP4	SP3	SCARB1	EPRS	TRAP1	FRZB
PASK	BRCA1	SENP6	DNAJC10	TYK2	IDH1
TRIB3	KIF23	ACTR1B	PNPLA7	SUCLG2	DNAH8
PEG3	BMP2K	MAPT	PARP1	LGALS3	WDHD1
GMPR2	EPHA2	ANKRD44	MAP3K1	CTSB	NDC80
DLG1	POMC	KAT2B	PIK3C2B	RAD51D	EPHA1
HSPBAP1	EPRS	UMPS	CD48	DNAH8	ADCY2
EPHX2	NCOA1	SYNJ1	HSPA4L	EPRS	FGFR4
SLC6A15	GMPS	ZKSCAN3	FGFR4	IFIT2	CHD1
TOPBP1	UTRN	ROR1	TCF3	BCL2A1	RAD51C
ACTR1B	OAS3	ABCC4	UMPS	PEG3	TLR4
GNB3	SERPINA1	RHBDF2	HIVEP2	ACTR1B	ANK3
RHOBTB3	TCF3	CCNH	MX1	GART	MAP3K1
RIPK4	TTC28	FRZB	BDNF	CHD1	TYK2
CCNH	HTT	ZNF35	GMPS	PDGFRA	CD44
KAT2B	ABL1	HERC3	HSPA4	RIPK2	TLR2
PLEC	ACTR1B	RRP12	KLC4	RNASE12	FGFR3
MAVS	RIPK4	ANPEP	VWF	RNASE9	UMPS
SP3	TAP1	HSPA4	ACACB	VCAN	LAMC2

POLQ	PLEC	EPHX2	TANC1	NDC80	SERPINE1
PDHX	ZFPM1	ZRANB3	KIF23	SELL	SHANK2
KDR	CASC1	BMP2K	CES1	PTCH1	BYSL
ICAM3	EPHA10	PDHX	FLII	PDIA6	CACNA1A
MCM8	PLAT	THBS1	ABL1	CD86	EPRS
ABCC4	LGALS3	PCK1	MET	REPIN1	TOP2B
VPS13C	HERC6	MYO5B	UBD	PI3	RBL2
ABCC1	REST	HSPBAP1	PLEK	DPYD	ACTR1B
LPL	RPA1	NUF2	PIK3CD	NFKBIE	BPTF
FN1	TYK2	NOTCH3	RHOBTB3	CALML4	NCOR2
IL4R	GLI3	BRD1	PIKFYVE	HERC6	PTCH1
LRIG1	MPP3	APAF1	ROR1	OAS2	HIF1A
FRZB	MC1R	PIK3CG	CLIP1	HIF1A	LGALS3
HERC3	DPYD	ACACB	GLI3	ACE	ACE
IRF7	PLAU	NACAD	ACTR1B	ANK3	FOS
UTRN	UBD	HADH	PLG	HSPA4	JAG1
TLR4	NUP98	PRKAG3	ACLY	KIF23	ANPEP
ARHGEF3	MLH1	ATM	S1PR3	BYSL	CTSC
EREG	CFTR	CCT5	DCLRE1C	SPAST	CSF1R
SCARB1	RAD51D	CLGN	CP	ADCY2	MYBL2
CYP2B6	SST	DDX60L	DPYD	ERBB2	CMPK1
UMPS	CLIP1	DST	EPHA1	MET	ATM
PLEK	ZNF91	SORL1	MPP3	PEPD	PEG3
THBS1	KIF2C	ZNF3	SMG1	TNFRSF25	GAK
MBL2	TRIB3	NOC2L	ZNF91	SMARCA2	PPP1R12A
ADAR	TLR1	TPO	COL2A1	CENPM	DEFB119
LRRC17	MET	CREBBP	SYNJ1	SPTB	BCL2A1
POLD1	LARS	ZNF224	SMARCA2	CAD	IGF2R
ASB6	CYP2C9	WDHD1	NGF	RAD54L	GPCPD1
TGFB1	GIPR	CDH1	CCT5	MSH2	ATP11B
BAZ1A	GNGT2	BCLAF1	PCNT	HDAC7	ADCY6
IRAK2	LTN1	GMPR2	CR1	CXCL16	SUCLG2
MRPS7	TLR6	FBXW8	ICAM1	EPHA10	MET
PI3	EGF	SRRM1	PTPRJ	ITGA2	OR10T2
BRCA1	RANBP2	LGALS3	ITSN2	NKTR	DNAJC10
CXADR	NPS	SERPINF2	RIPK4	OPRM1	PDGFRB
ANKRD44	KNG1	CASP8	VPS13C	PHLPP1	KIF23
ALB	TNS3	CLSPN	PDGFRA	ZNF680	CASP7
STIL	PMPCB	SERPINA1	GOT2	ZFPM1	ADCY7

BMP2K	PLEK	APPL1	ZNF160	KDR	CYP2B6
ABCC11	IRF7	TMF1	HUNK	KLK3	NNT
HSPA4	COX11	BPTF	XRN1	PPARA	TIE1
C8G	10-Mar	ZNF117	GIPR	BIRC2	CPS1
FBXW8	PGR	ELN	NMU	PIK3C2B	NOS2
EPRS	BRWD1	GHR	MLKL	FANCD2	BLK
SYNJ1	CENPF	BRD2	SON	DDX54	SLC22A1
TBC1D2	CYP4B1	UTRN	POMC	FLT1	ERBB3
DNAH11	MUC5B	GSPT1	FLNC	CACNA1A	ACACB
REST	REPIN1	EREG	CRLF3	MKKS	PHLPP1
POLE	NMU	TRAF3IP2	VDR	PMPCB	PHGDH
DNAH8	DNAJC10	POLQ	PFAS	KNG1	MCM8
PRSS57	ZNF451	CR1	BRD1	TJP1	TNFRSF25
LGALS3	LUM	IL4R	KAT2B	PHGDH	EPHB2
CASP9	S1PR3	PDGFRB	PTPN14	DEFB128	FBXO5
	FANCI	CTNNAL1	NVL	CLSPN	ABCC3
	VDR	NMBR	ASB2	PIK3CA	LCP2
	PCNT	ARFGAP1	CYP2C9	ABCC3	FLT1
	TRPA1	RUNX1	LGALS3	SHANK2	ESR1
	BAG3	SPG7	ACE	TPR	SOS2
	AGXT	DSTYK	BPTF	TBC1D2	MUTYH
	DCAF13	MCM8	TSR1	LAMC2	IRAK4
	FOS	LRRC17	GABBR1	NCOA1	RAD51D
	HPGDS	ABCC1	CAPS2	RPS6KB2	MSH6
	CCT6A	VTN	IRAK3	CDK13	PIK3CD
	RIOK2	FLT3	AKAP13	MYBBP1A	HADH
	CAMKK2	PLEK	CENPF	NCOR2	APP
	CAPS2	INSRR	TFPI	RANBP2	SPTB
	PARP9	KIF9	CASC1	MRPL47	BDNF
	TST	HERC1	REST	FANCC	LRRC8E
	FBXO2	DOCK4	EXOSC3	TTC21A	CR1
	AQP7	ALB	SMTN	GOLGB1	RAD54L
	ANK1	ASPM	MYO5B	HNF4A	CALCA
	CASR	PIK3C2G	ARFGEF1	CACNA1S	MTHFD1
	HIVEP2	MBL2	GNGT2	PLCB3	IFIH1
	CRLF3	DNAJA3	MAPK15	NPHP3	KAT2B
	MAPK3	TLR4	PRKCE	PTPRD	AFP
	ERAP1	JAK2	BYSL	RBL2	RIPK2
	ZNF267	GRK4	ADCY4	NQO1	KITLG

	TSPO	INVS	DMWD	FN1	TJP2
	CLEC7A	FOS	GABPB2	CDKN1A	GATA2
	RIPK3	PPP1R12A	ERBB2	TNS1	CDC25C
	TFPI	OAS3	CARD10	MSH4	VWF
	SUCLG2	EPRS	DEFB108B	ZNF92	BIRC2
	PRELP	NCOR2	MAPK3	INVS	AQP1
	CD48	SETMAR	DEFB119	RAD52	MAP3K5
	CASP7	TPR	LGR4	TIE1	DSC3
	CLSPN	FURIN	ESR2	EGFR	FGFR1
	MLKL	RNASEL	LUM	GPCPD1	UBD
	RET	ABCC3	EBNA1BP2	PGR	HDAC4
	RB1	POLE	DHX58	CPT2	CENPF
	HMMR	GOT2	PKDCC	ATXN3	LRRN1
	NDUFA9	FANK1	FBXO2	SYNJ1	KIF15
	NOS2	GSN	NPS	AXL	CALML4
	LRRK2	LIMK2	KDR	IGF1R	ZAP70
	ANKRD53	MAP3K1	BMP2	PRDM14	TPR
	PLA2G6	POLD1	MSH4	KDM5A	SCARB1
	LRRC40	POLN	GTF3A	PLAUR	MPHOSPH8
	ZNF93	IL16	NCAN	PRKACG	ALDH1B1
	MAPT	TOPBP1	PHLPP1	FAM81B	ATR
	TXNDC16	RUNX3	F5	ALDH1B1	CTSS
	CENPM	GAMT	ANKRD28	SELP	BRCA1
	CTBP2	SKIV2L	ANXA5	MCM8	ANKRD17
	HSPA4	COMP	PFKL		
	ALDH18A1	CAT	TERT		
	ZNF611	DHX36	HERC6		
	EPHB4	IFI16	FANCI		
	RNASE12	POLI	TXNDC16		
	ESR1	MZF1	TRH		
	PTPRZ1	KIT	ATP11B		
	MRPL13	DNAH8	ITPA		
	NRG1	EFCAB3	ACTN1		
	AGRN	HK2	CYP4F2		
	IL4R	REST	HTT		
	GABPB2	POLA2	TNFRSF10B		
		TBL2	NOS2		
		PRDM10	KITLG		
		SOS2	CD86		

		VPS13B	KAT6A		
		PCMTD1	AARS		
		HDC	ACADS		
		VAV1			
		MYB			
		DNAJC10			
		NDUFA10			
		ADCY7			

Table 2: Candidate genes identified by MAGMA in the two AD cohorts. Related to Figure 1.

Discovery Cohort	**Extension Cohort**
APOE	BRINP2
TOMM40	KIAA1614
CNN2	TPO
TCP1	FLT4
SGSM2	CACNA1H
PGLYRP2	GGT5
CFAP74	NKX3-2
MDH1B	
RPA2	
SLC36A1	
DRC3	
PRIM1	
S100PBP	
CBR3	
TNFRSF17	
MICA	
KANSL1	
DHRS4L2	
SAA4	
TRIAP1	
SH3BGRL2	
TGFBR1	
ADRA1A	
CDS1	
RARS1	
PEMT	
ST14	
TREM2	

APOC4	
BPIFA3	
SORL1	

Table 3: One-tailed hypergeometric overlaps between candidate genes identified by GeneEMBED using **STRING** network on various VIS-cohort combinations (column label) and reference gene sets of known AD associated genes (row label). As a comparison to standard approaches, hypergeometric overlaps of MAGMA candidate genes are also shown in the right most column. Related to Figure 1.

	EA Disc. (n=69)		EA Ext. (n=119)		PPh2 Disc. (n=128)		PPh Ext. (n=120)		MAGMA (n=31)	
Gene Set (GS)	ovlp	pval	ovlp	pval	ovlp	pval	ovlp	pval	ovlp	pval
GWAS Meta-analysis 1 (n=25)	1	0.091	1	0.146	1	0.156	1	0.14	3	8.43E-06
GWAS Meta-analysis 2 (n=38)	1	0.142	1	0.214	1	0.227	1	0.216	3	3.30E-06
Comp. Toxico. Database (n=103)	2	0.047	4	3.30E-03	1	0.347	1	0.34	3	3.97E-04
ClinVar (n=21)	1	0.081	2	9.10E-03	0	1	1	0.13	1	0.034
DisGeNet (n=208)	5	1.34E-03	7	5.30E-04	6	3.86E-03	5	0.012	4	4.37E-03

Table 4: nDiffusion analysis AUC and permutation based z-scores between candidate genes identified by GeneEMBED using **STRING** network on various VIS-cohort combinations (column label) and reference gene sets of known AD-associated genes (row labels). As Comparison to standard methods, analysis on MAGMA candidate genes are shown in the right most column. Related to Figure 3.

	EA Disc. (n=69)		EA Ext. (n=119)		PPh2 Disc. (n=128)		PPh Ext. (n=120)		MAGMA (n=31)	
Gene Set (GS)	AUC	z-score	AUC	z-score	AUC	z-score	AUC	z-score	AUC	z-score
Comp. Toxico. Database (n=103)	0.76	2.03	0.78	2.38	0.8	2.82	0.78	3.01	0.79	6.62
GWAS Meta-analysis 1 (n=25)	0.74	3.71	0.71	3.51	0.71	2.85	0.63	0.29	0.66	1.63
GWAS Meta-analysis 2 (n=38)	0.63	2.54	0.61	1.93	0.63	2.6	0.59	1.87	0.56	1.22
ClinVar (n=21)	0.78	2.3	0.78	3.39	0.84	5.64	0.74	2.33	0.77	3.27
DisGeNet (n=208)	0.69	3.26	0.71	3.95	0.71	3.88	0.69	2.3	0.65	2.16

Table 5 : Candidate genes identified by GeneEMBED on various cohort-VIS combinations (column label) using the **HINT** network. Related to Figure 1.

Discovery - EA	**Extension - EA**	**Discovery - PPh2**	**Extension - PPh2**	**Control vs Control - EA**	**Control vs Control - PPh2**
APOE	CEACAM8	ANLN	TCTN3	IL22RA2	ESR1
KCNIP1	CEACAM6	CCR8	CXCL12	NOTUM	IL22RA2
ANLN	IL22RA2	TREM2	IL18RAP	CD69	NINL
KCND3	GFRA2	CCL1	GOLGA2	WNT7A	LRRK2
MYC	CPEB1	MYC	LRRK2	NRG1	NRG1
FCRL4	ESR1	MAPK6	NINL	TMEM190	FAM161A
SLC16A7	SLC16A7	NOTUM	CCR6	SLC16A7	SCN4A
FN1	NRTN	KCNIP1	IL18R1	WDYHV1	WDYHV1
CPEB1	LRRK2	PF4	MAPK6	KCNIP1	CHRNA5
NOTUM	GOLGA2	FNIP2	APOE	CPEB1	CD69
MAPT	TCTN3	FLCN	ESR2	KCND3	ADM
NQO1	APOE	GRIN3A	CCL17	CALY	MCM2
MUC15	PANK3	CCHCR1	CCL20	GFRA2	FN1
MAPK6	GDNF	CDH16	IL22RA2	ZACN	DPYS
MCM2	BLM	TP53BP2	FN1	TNFSF18	WNT7A
CXCL12	ABCB4	TRAF3IP2	ARL14EP	KRT40	MRPL47
MAVS	KCNIP1	CCNJL	LARP7	DRD1	FNIP2
SLC6A15	HTT	G6PC	KLC4	NINL	RIPK3
STIL	MAPK6	TP53	DMWD	FN1	NOTUM
CCL24	CHRNA6	BCLAF1	AGRP	MRPL47	PTCH1
CALCOCO2	CCHCR1	FN1	PTPN14	LGALS3	FLCN
TNNT1	LAMTOR5	ERBB2	TP53	BRCA1	TRAP1
VPS53	PIGY	ECSIT	GFRA1	LRRK2	CCDC155
CCL1	HLA-DPA1	CEBPA	USHBP1	PTCH1	SCN1B
PRKRA	TLR10	ADAMTSL4	MPP3	FAM161A	LGALS3
TNKS1BP1	NME5	NDUFA10	NOTUM	IL22	CAD
PAICS	CEP128	TOLLIP	KLRD1	APOBEC1	PTGDR
CCR8	NOTUM	NQO1	GIPR	GABRB2	ARRDC1
TSSC4	ESR2	MCM2	BATF3	GABRA5	BYSL
NDUFA10	NINL	NINL	TSR1	SUCLG2	CHRNA6
WDHD1	NCOA1	MAPT	GIP	WDHD1	GABRB2
CDS1	DYDC2	EEF2K	SLC45A2	ITGA4	RPGRIP1L
ECSIT	GNAS	VPS53	PLSCR1	PF4	CACNA1A
IDI2	LGALS3	ATXN7	GAK	COA6	DPP4
CCR1	BRCA1	TNIP1	CREB3	XCL1	CCR7
TNFSF18	TLR1	SCN1B	RIPK3	CEACAM8	UBASH3B

CXADR	MPP3	NUF2	MTCH2	EYA4	TLN2
CRELD2	CHRNA4	ZNF614	PTPRO	NRTN	ANKRD13C
CEACAM8	LPIN3	RASSF7	OXSR1	A1CF	PCM1
CHRNA5	CEP250	MRPL47	RNF216	CHRNA6	CCL19
BCLAF1	PRMT6	ARFGAP1	BLM	CHRNA5	GOLGA2
GRIN3A	DSP	WDR18	IL18	ATXN1	APP
TREM2	TUSC1	RRP12	HTT	CEP63	TJP2
KCNIP2	GGPS1	CLGN	MAVS	MYBBP1A	EYA4
CCL15	TSC22D1	GIPR	TCF3	CEACAM6	EWSR1
NME5	AGRP	ABI3	LGALS3	TRAP1	CDS1
MRPL28	IL22	LGALS3	DDX19B	BYSL	TREM2
TRIM25	KRT40	PAICS	SLFN5	GABRB3	SEC16A
DYDC2	COMMD10	NFIX	ABL1	ATM	FNIP1
GFRA2	TCF3	RNF43	SLC25A47	FAM20A	CKAP4
CRHR1	KIF13B	ZC3H3	CCHCR1	ZC3H3	WDHD1
CEACAM6	ABL1	SLC45A2	CISD3	NAGS	LSMEM1
PPHLN1	IDI2	MYOG	FOXM1	IFNL3	HIF1A
RASSF7	OXSM	SVIL	TREM2	ARRDC1	DNASE2B
FXYD7	CISD3	FBF1	MRPL38	HTR3C	TLE2
HBQ1	FOXM1	GIP	DSP	PEPD	TMEM190
LGALS3	MRPL38	TNKS1BP1	TREM1	TTC30A	TICAM1
ZFAND4	PMF1	WNT7A	FBF1	TNFRSF18	DBN1
ANKH	GFRA1	PCM1	GDF5	UCN	SUGP1
DLG1	EVC2	VCAM1	KLRC2	UBASH3B	IDH1
LAMTOR5	CAND1	MTMR4	CHRNA6	STX18	LNX1
EHBP1L1	BAG3	MRM1	CAND1	FCRL4	SH3RF3
TFAP2C	CNTRL	KCND2	EPHA2	CGN	ITSN1
CCL14	LARP7	TNFRSF14	FLNC	CHRNB2	KCNIP2
MAGI1	POMT1	ECE2	PRMT6	HIF1A	MRPS18B
EEF2K	CREB3	ILF3	B3GALNT1	STIL	USP20
DISC1	RMND5B	SCN4A	HSPA4L	EVC2	CHRNB4
ADAMTSL 4	PARP1	ZKSCAN3	ATXN1	PLEKHA7	ATXN7
BYSL	MYBBP1A	KAT2B	SH3RF3	RIF1	CENPO
RNF43	DDX19B	USHBP1	KCNG4	GDNF	SLC45A2
AGAP1	CTBP2	EPN3	BYSL	PLAUR	NCOR2
ABCC1	CCL20	SERPINA10	XRN1	NFKBIE	TTC30A
RFESD	TSR1	MYO19	EXOSC3	CAD	TKT
CEBPA	TLR6	DLG1	DVL2	GOLGB1	RRP7A
MYO19	ITGA4	RHOBTB3	HGD	CHRNA4	PRRC2B
MTMR4	DCLRE1C	NACAD	PYGB	MRPS18B	PAAF1
DDR1	PTPN14	GTF3C1	KRTAP5-9	TBC1D4	MYL2

TP53BP2	NDUFA9	LLGL2	ZZEF1	FOS	HDAC4
RECQL5	UIMC1	LARP7	KRT31	IQCB1	CNOT1
BTNL8	ATXN1	SRRM1	LZTS2	CREB3	LYNX1
FAM171A2	PTPN3	SIRPB1	CHRNA4	EGFR	DSP
BRCA1	SYNPO	DNAJA3	TAX1BP1	RRP7A	PIK3CG
WNT7A	CALCOCO2	DISC1	ESR1	ATXN7	ERBB3
TNFRSF18	RPA1	S100A2	CCR4	TICAM1	DUSP6
NUBP1	HTR1B	KCNIP2	PARP1	AKNAD1	ZNF529
MRPL47	SEC16A	HIPK4	SERPINA10	FANCD2	DENND4C
CHRNB2	SH3D19	SIPA1L1	ITPA	NIN	KCNMA1
ATXN7		DDX19B	MTUS2		USP21
CENPJ		MTA1	NDUFA9		ATXN1
DDX5		AZU1	RET		DSG1
CTSB		GIGYF1	MYO19		MAL
ATL1		EEF1D	BAG3		CDC25C
TNIP1		DST			
CCL16		ACTR1B			
SLC39A4		KLRD1			
CHRNA6		APPL1			
ECH1		LRRC6			
TLR10		FBXW8			
NINL					

Table 6: One-tailed hypergeometric overlaps between candidate genes identified by GeneEMBED using **HINT** network on various VIS-cohort combinations (column label) and reference gene sets of known AD associated genes (row label). Related to Figure 1.

		Disc. -EA (83)		Ext. -EA (67)		Disc. -PPh2 (84)		Ext. -PPh2 (81)	
Gene Set (GS)	GS size	ovlp	pval	ovlp	pval	ovlp	pval	ovlp	pval
Comp. Toxico. Database	103	4	0.0014	1	0.33	2	0.091	1	0.38
GWAS Meta-analysis 1	25	2	0.0058	1	0.086	0	1	1	0.11
DisGeNet	208	3	0.108	1	0.61	1	0.69	3	0.103
ClinVar	21	1	0.11	1	0.095	0	1	1	0.11
GWAS Meta-analysis 2	38	2	0.015	1	0.14	0	1	1	0.17

Table 7: nDiffusion analysis AUC and permutation based z-scores between candidate genes identified by GeneEMBED using **HINT** network on various VIS-cohort combinations (column label) and reference gene sets of known AD-associated genes (row labels). Related to Figure 3.

	Disc. -EA (83)		Ext. -EA (67)		Disc. -PPh2 (84)		Ext. -PPh2 (81)	
Gene Set (GS)	AUC	z-score	AUC	z-score	AUC	z-score	AUC	z-score
Comp. Toxico. Database (n=103)	0.77	4.34	0.75	3.33	0.75	3.34	0.74	3.32
GWAS Meta-analysis 1 (n=25)	0.68	1.86	0.59	0.34	0.66	1.91	0.65	1.12
GWAS Meta-analysis 2 (n=38)	0.62	2.37	0.54	0.36	0.63	2.6	0.59	0.96
ClinVar (n=21)	0.76	3.34	0.66	0.71	0.76	2.71	0.67	1.38
DisGeNet (n=208)	0.7	5.77	0.7	5.16	0.68	3.89	0.67	4.91

Table 8: Candidate genes identified by GeneEMBED on various cohort-VIS combinations (column label) using the **Brain** specific network**.** Related to Figure 1.

Discovery - EA	**Extension - EA**	**Discovery - PPh2**	**Extension - PPh2**	**Control vs Control - EA**	**Control vs Control - PPh2**
APOE	BLM	MYC	APOE	BRCA1	APP
MYC	RPA1	OBSL1	TCF3	ATM	OBSL1
MAPT	BAG3	TRAF3IP2	LRRK2	FBXO6	LRRK2
MAVS	APOE	ATXN7	USP2	OBSL1	RIPK3
NQO1	HTT	NQO1	BAG3	FN1	LNX1
FN1	TCF3	MAPT	BLM	MYBBP1A	DPAGT1
CIDEB	BRCA1	TP53BP2	MAVS	RPS6KB2	FN1
PRKRA	LRRK2	KAT2B	RPA1	FOS	NCOR2
IL16	PARP1	SVIL	FN1	LGALS3	NRG1
DLG1	EIF2B4	PHLDA3	PARP1	ACIN1	HIF1A
OBSL1	MC5R	CREBBP	TP53	APP	FBXO6
MCM2	TNFSF15	TP53	CREB3	ATXN7	RPS6KB2
ATXN7	TESC	IL16	TNFSF15	HIF1A	MCM2
IRF7	GNAS	CTNNAL1	RIPK3	USP8	ATXN7
BRCA1	TNFRSF6B	DLG1	TNFRSF6B	LRRK2	TRAP1
KAT2B	MYBBP1A	FBXW8	PRKCE	DHRS2	NCBP1
DDX5	STX6	CDH1	HTT	NRG1	MYBBP1A
ANKH	DSP	FN1	LGR4	NCOR2	
CALCOCO2	DPP4	RUNX1	TXNDC17	HDAC7	
SVIL	SP3	TNFRSF14	HDAC5	FHOD1	
FBXW8	IRF7	TCF3	LNX1	CREB3	

PHLDA3	SQSTM1	EEF1D	ATXN7	EGFR	
TOPBP1		RHOBTB3		APC2	
		TRAF2			
		CASP8			

Table 9: One-tailed hypergeometric overlaps between candidate genes identified by GeneEMBED using **Brain Specific** network on various VIS-cohort combinations (column label) and reference gene sets of known AD associated genes (row label). Related to Figure 1.

		Disc. -EA (19)		Ext. -EA (19)		Disc. -PPh2 (22)		Ext. -PPh2 (17)	
Gene Set (GS)	GS size	ovlp	pval	ovlp	pval	ovlp	pval	ovlp	pval
Comp. Toxico. Database	103	2	0.024	1	0.22	1	0.25	1	0.18
GWAS Meta-analysis 1	25	1	0.051	1	0.051	0	1	1	0.046
DisGeNet	208	2	0.096	1	0.41	1	0.46	1	0.38
ClinVar	21	1	0.068	1	0.068	0	1	1	0.061
GWAS Meta-analysis 2	38	1	0.057	1	0.057	0	1	1	0.051

Table 10: nDiffusion analysis AUC and permutation based z-scores between candidate genes identified by GeneEMBED using **Brain Specific** network on various VIS-cohort combinations (column label) and reference gene sets of known AD-associated genes (row labels). Related to Figure 3.

	Disc. -EA (19)		Ext. -EA (19)		Disc. -PPh2 (22)		Ext. -PPh2 (17)	
Gene Set (GS)	AUC	z-score	AUC	z-score	AUC	z-score	AUC	z-score
Comp. Toxico. Database (n=103)	0.72	2.11	0.77	4.64	0.75	3.85	0.72	1.36
GWAS Meta-analysis 1 (n=25)	0.71	2.69	0.58	0.2	0.69	2.43	0.54	0.28
GWAS Meta-analysis 2 (n=38)	0.63	2.46	0.58	0.85	0.64	2.89	0.54	0.63
ClinVar (n=21)	0.78	3.91	0.7	1.68	0.82	4.01	0.67	0.62
DisGeNet (n=208)	0.66	2.15	0.69	5.07	0.7	6.07	0.65	1.31

Table 11: Details regarding the specific alleles used to test GeneEMBED high-confidence candidate genes are given below. Related to Figure 4.

Human GeneID	**Drosophila Homolog**	**DIOPT Score (max 15)**	**Allele Class**	**Allele type**	**Specific alleles**
ABL1	Abl	9	LOF (loss of function)	Amorphic (W559term)	Abl[2]/TM6B, Tb[1]

ABL1	Abl	9	OE (Over exrpression)	Myc tagged cDNA inducible	w[*]; P{w[+mC]=UAS-Abl.Myc}attP40
BAG3	stv	9	LOF	Insertion of P-element dysruptin g intron 2 and an alternativ e promoter	P{ry[+t7.2]=PZ}stv[00543] ry[506]/TM3, ry[RK] Sb[1] Ser[1]
BLM	Blm	10	LOF	Imprecise excision of the P{EPgy2 EY03745 .	w[1118]; Blm[N1]/TM3, Sb[1]
BMP2K	Nak	7	LOF	Transpos able element insertion in exon 8	w[1118]; PBac{w[+mC]=WH}Nak[f04720]
CCT5	CCT5	14	LOF	Transpos able element insertion in exon 1	P{ry[+t7.2]=PZ}CCT5[06444] cn[1]/CyO; ry[506]
CCT5	CCT6	2	LOF	Transpos able element insertion in promoter	w[67c23] P{w[+mC]=lacW}CCT6[G0022]/F M7c
DDX19 B	Dbp80	10	RNAi	Inducible RNAi expressio n under UAS	w[*]; P{w[+mC]=UAS-Dbp80.RNAi}19-1
GIPR	Pdfr	3	LOF	Transpos able element insertion in 5'UTR	y[1] Mi{y[+mDint2]=MIC}Pdfr[MI07832] w[*]
GOT2	Got2	13	LOF	Naturally occuring amorphic allele	Got2[nNC]/SM1
HTT	htt	12	LOF	The first three exons of htt have been replaced	TI{RFP[DsRed.3xP3.cUa]=TI}htt[KO]

				with a Disc\RFP 3xP3.cUa marker	
IL16	bbg	3	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}bbg[MI02662]/TM3, Sb[1] Ser[1]
KAT2B	Gcn5	12	LOF	Transposable element insertion in 5'UTR	w[1118]; PBac{w[+mC]=WH}Gcn5[f02830]/TM6B, Tb[1]
KLRD1	rgn	1	LOF	Transposable element insertion in 5'UTR	y[1] w[67c23]; Mi{GFP[E.3xP3]=ET1}rgn[MB01529]
MAPK6	p38c	2	LOF	Transposable element imprecise excision resulting in loss of function	w[*]; p38c[19B1]/TM6
MPP3	metro	11	LOF	Transposable element insertion in 5'UTR	y[1] w[*]; Mi{y[+mDint2]=MIC}metro[MI02273]
MYO19	d	2	OE	UAS inducible overexpression	w[*]; P{w[+mC]=UAS-d.V5}9-F
MYO5B	d	2	OE	UAS inducible overexpression	w[*]; P{w[+mC]=UAS-d.V5}9-F
PLEC	shot	4	OE	UAS inducible overexpression, GFP tagged	w[*]; P{w[+mC]=UAS-shot.L(C)-GFP}3
PLEC	shot	4	LOF	Transposable element carrying a splicing	y[1] w[*]; Mi{y[+mDint2]=MIC}shot[MI03583]/SM6a

				donnor inserted in intronic region	
PLEK	CG32982	1	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}CG32982[MI00088]
PLEK	CG32982	1	LOF	Transposable element inserted in exon 6	y[1] w[*]; Mi{y[+mDint2]=MIC}CG32982[MI00152]
REST	CG9932	2	LOF	Transposable inserted in intronic region	w[1118]; PBac{w[+mC]=PB}CG9932[c00144]/CyO
RPA1	RpA-70	15	LOF	Transposable element insertion in 5'UTR	w[1118]; P{w[+mC]=EP}RpA-70[G5479]/TM6C, Sb[1]
RPA1	RpA-70	15	LOF	Transposable element insertion in promoter	w[1118]; PBac{w[+mC]=PB}RpA-70[c01306]
SP3	Spps	7	LOF	Transposable inserted in exon	w[1118]; P{w[+mC]=EP}Spps[G8810]/TM6C, Sb[1]
SP3	CG3065	2	OE	Rescue construct exoressing GFP tagged CG3065 under the endogenous promoter	w[1118]; PBac{y[+mDint2] w[+mC]=CG3065-GFP.FPTB}VK00033
SVIL	Svil	6	OE	UAS inducible overexpression, GFP tagged	w[*]; P{w[+mC]=UASp-GFP.Svil}attP2

SVIL	Svil	6	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}Svil[MI06800]/TM3, Sb[1] Ser[1]
TCF3	da	12	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}da[MI13697]
TLR10	Tl	3	OE	UAS inducible overexpression, Venus tagged	w[*]; P{w[+mC]=UASp-Tl.Venus}4
TP53	p53	7	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}p53[MI01307]
TRIB3	trbl	9	LOF	Transposable element insertion in promoter	w[1118]; P{w[+mC]=EP}trbl[EP1119]/TM6B, Tb[1]
TRIB3	trbl	9	OE	UAS inducible overexpression	w[1118]; P{w[+mC]=UASp-trbl.M}3
UTRN	Dys	9	LOF	Transposable element carrying a splicing donnor inserted in intronic region	y[1] w[*]; Mi{y[+mDint2]=MIC}Dys[MI01893]/TM3, Sb[1] Ser[1]

Table 12: Drug-Gene interaction information of high-confidence genes is given below. Row labels are candidate genes who had documented interactions with FDA approved drugs in DGIdb. The specific drugs with which they interact is given in the column titled 'drug'. Interaction type is specified in the 'interaction_types' column. Lastly, the sources from which DGIdb acquired the drug-gene interaction is listed in the last column. Related to Figure 4.

gene	drug	interaction_types	sources
S1PR3	SIPONIMOD	agonist	GuideToPharmacology
S1PR3	FINGOLIMOD HYDROCHLORIDE	agonist	ChemblInteractions
VDR	CALCIFEDIOL	agonist	ChemblInteractions|TTD
VDR	ERGOCALCIFEROL	agonist	TdgClinicalTrial|ChemblInteractions|TEND|TTD
VDR	CHOLECALCIFEROL	agonist	DTC|ChemblInteractions|TTD
VDR	PARICALCITOL	agonist	TdgClinicalTrial|ChemblInteractions|TEND|GuideToPharmacology|TTD
VDR	CALCIPOTRIENE	agonist	TdgClinicalTrial|ChemblInteractions
VDR	CALCIFEDIOL	agonist	TTD
VDR	TACALCITOL	agonist	GuideToPharmacology
VDR	DIHYDROTACHYSTEROL	agonist	TdgClinicalTrial|TEND|TTD
ESR2	ESTROGENS, CONJUGATED	agonist	ChemblInteractions
ESR2	ESTROGENS, CONJUGATED SYNTHETIC A	agonist	ChemblInteractions
ESR2	LASOFOXIFENE	agonist	GuideToPharmacology
ESR2	ESTRADIOL	agonist	DTC|TdgClinicalTrial|TEND
ESR2	QUINESTROL	agonist	ChemblInteractions
ESR2	ESTRIOL	agonist	GuideToPharmacology
ESR2	ESTROGENS, ESTERIFIED	agonist	ChemblInteractions
ESR2	SYNTHETIC CONJUGATED ESTROGENS, B	agonist	ChemblInteractions
ESR2	ESTRONE	agonist	GuideToPharmacology
ESR2	DIETHYLSTILBESTROL DIPHOSPHATE	agonist	ChemblInteractions
ESR2	ETHINYL ESTRADIOL	agonist	GuideToPharmacology
ESR2	TAMOXIFEN	agonist|antagonist	DTC|TdgClinicalTrial|TEND|GuideToPharmacology|PharmGKB
VDR	DOXERCALCIFEROL	agonist|suppressor	TdgClinicalTrial|ChemblInteractions|GuideToPharmacology|TTD
ESR2	TRILOSTANE	allosteric modulator	TTD
GRIN3A	ORPHENADRINE CITRATE	antagonist	ChemblInteractions
GRIN3A	ORPHENADRINE	antagonist	TdgClinicalTrial|TEND
GRIN3A	FELBAMATE	antagonist	TdgClinicalTrial|ChemblInteractions|TEND
GRIN3A	ORPHENADRINE HYDROCHLORIDE	antagonist	ChemblInteractions
GRIN3A	AMANTADINE HYDROCHLORIDE	antagonist	ChemblInteractions
GRIN3A	MEMANTINE	antagonist	TdgClinicalTrial|TEND
GRIN3A	KETAMINE	antagonist	TdgClinicalTrial|TEND

GRIN 3A	DEXTROMETHORPHAN	antagonist	TdgClinicalTrial|TEND
GRIN 3A	ACAMPROSATE CALCIUM	antagonist	ChemblInteractions
ESR2	BAZEDOXIFENE	antagonist	TdgClinicalTrial|GuideToPharmacology
ESR2	FULVESTRANT	antagonist	TALC|DTC|ChemblInteractions|GuideToPharmacology|PharmGKB
VDR	CALCITRIOL	antagonist|agonist	DTC|TdgClinicalTrial|ChemblInteractions|NCI|TEND|PharmGKB|TTD
ESR2	RALOXIFENE	antagonist|agonist	DTC|TdgClinicalTrial|TEND|GuideToPharmacology|PharmGKB
PARP 1	NIRAPARIB	antagonist|inhibitor	TALC|MyCancerGenome|TdgClinicalTrial|ClearityFoundationClinicalTrial|ChemblInteractions|GuideToPharmacology
FLT3	SORAFENIB	antagonist|inhibitor	TALC|TdgClinicalTrial|JAX-CKB|TEND|DoCM|COSMIC|CIViC|GuideToPharmacology|PharmGKB|OncoKB
FLT3	MIDOSTAURIN	antagonist|inhibitor	MyCancerGenome|TdgClinicalTrial|JAX-CKB|ChemblInteractions|CGI|DoCM|CIViC|GuideToPharmacology|PharmGKB|TTD|FDA|OncoKB
IL4R	DUPILUMAB	antibody|antagonist	ChemblInteractions|GuideToPharmacology|TTD
PARP 1	NIACINAMIDE	binder	TTD
PARP 1	TALAZOPARIB TOSYLATE	inhibitor	ChemblInteractions
PARP 1	RUCAPARIB CAMSYLATE	inhibitor	ChemblInteractions
PARP 1	OLAPARIB	inhibitor	DTC|MyCancerGenome|ClearityFoundationClinicalTrial|ChemblInteractions|CIViC|GuideToPharmacology
PARP 1	TALAZOPARIB	inhibitor	ChemblInteractions|GuideToPharmacology
HTT	AMITRIPTYLINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	CLOMIPRAMINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	FLUVOXAMINE MALEATE	inhibitor	ChemblInteractions
HTT	ESCITALOPRAM OXALATE	inhibitor	ChemblInteractions
HTT	AMOXAPINE	inhibitor	ChemblInteractions
HTT	DESVENLAFAXINE	inhibitor	ChemblInteractions
HTT	PAROXETINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	DESVENLAFAXINE SUCCINATE	inhibitor	ChemblInteractions
HTT	DULOXETINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	PROTRIPTYLINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	FLUOXETINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	VORTIOXETINE HYDROBROMIDE	inhibitor	ChemblInteractions
HTT	NORTRIPTYLINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	IMIPRAMINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	VILAZODONE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	VENLAFAXINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	MAZINDOL	inhibitor	ChemblInteractions
HTT	TRAZODONE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	SERTRALINE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	NEFAZODONE HYDROCHLORIDE	inhibitor	ChemblInteractions
HTT	PAROXETINE MESYLATE	inhibitor	ChemblInteractions
HTT	CITALOPRAM HYDROBROMIDE	inhibitor	ChemblInteractions
HTT	LEVOMILNACIPRAN HYDROCHLORIDE	inhibitor	ChemblInteractions

TP53	BORTEZOMIB	inhibitor	TALC|CIViC
BMP2K	BARICITINIB	inhibitor	GuideToPharmacology
ABL1	PONATINIB	inhibitor	DTC|MyCancerGenome|TdgClinicalTrial|CGI|DoCM|CIViC|GuideToPharmacology|PharmGKB|TTD|FDA|OncoKB
ABL1	NILOTINIB HYDROCHLORIDE MONOHYDRATE	inhibitor	ChemblInteractions
ABL1	IMATINIB	inhibitor	DTC|MyCancerGenome|TdgClinicalTrial|JAX-CKB|NCI|CGI|TEND|COSMIC|CIViC|GuideToPharmacology|PharmGKB|FDA|OncoKB
ABL1	IMATINIB MESYLATE	inhibitor	ChemblInteractions
ABL1	NILOTINIB	inhibitor	MyCancerGenome|TdgClinicalTrial|JAX-CKB|CGI|TEND|DoCM|COSMIC|CIViC|GuideToPharmacology|PharmGKB|FDA|OncoKB
ABL1	REGORAFENIB	inhibitor	ChemblInteractions
ABL1	PONATINIB HYDROCHLORIDE	inhibitor	ChemblInteractions
ABL1	BOSUTINIB	inhibitor	MyCancerGenome|TdgClinicalTrial|ChemblInteractions|CGI|DoCM|COSMIC|GuideToPharmacology|PharmGKB|TTD|FDA|OncoKB
NQO1	DICUMAROL	inhibitor	NCI|PharmGKB
FLT3	IDARUBICIN	inhibitor	MyCancerGenomeClinicalTrial
FLT3	CERITINIB	inhibitor	GuideToPharmacology
FLT3	PEXIDARTINIB	inhibitor	MyCancerGenome|JAX-CKB|ChemblInteractions|CIViC|GuideToPharmacology|CancerCommons|TTD
FLT3	SUNITINIB	inhibitor	TALC|MyCancerGenome|TdgClinicalTrial|JAX-CKB|TEND|COSMIC|CIViC|GuideToPharmacology|MyCancerGenomeClinicalTrial|PharmGKB
FLT3	AZACITIDINE	inhibitor	JAX-CKB|MyCancerGenomeClinicalTrial
FLT3	GILTERITINIB	inhibitor	JAX-CKB|ChemblInteractions|CIViC|GuideToPharmacology|PharmGKB|TTD|FDA|OncoKB
FLT3	BORTEZOMIB	inhibitor	MyCancerGenomeClinicalTrial
FLT3	CLOFARABINE	inhibitor	MyCancerGenomeClinicalTrial
FLT3	PONATINIB	inhibitor	DTC|MyCancerGenome|JAX-CKB|CGI|DoCM|CIViC|TTD
FLT3	SUNITINIB MALATE	inhibitor	ChemblInteractions
FLT3	SORAFENIB TOSYLATE	inhibitor	ChemblInteractions
FLT3	NINTEDANIB	inhibitor	TTD
FLT3	CABOZANTINIB	inhibitor	MyCancerGenome|JAX-CKB
EPHX2	ZAFIRLUKAST	inhibitor	GuideToPharmacology
EPHX2	OXAPROZIN	inhibitor	GuideToPharmacology
ABCC1	SULFINPYRAZONE	inhibitor	TdgClinicalTrial|TEND|TTD
POLD1	FLUDARABINE PHOSPHATE	inhibitor	ChemblInteractions
POLD1	CLOFARABINE	inhibitor	ChemblInteractions
POLD1	GEMCITABINE HYDROCHLORIDE	inhibitor	ChemblInteractions
POLD1	CYTARABINE	inhibitor	ChemblInteractions
POLE	GEMCITABINE HYDROCHLORIDE	inhibitor	ChemblInteractions
POLE	CYTARABINE	inhibitor	ChemblInteractions
POLE	FLUDARABINE PHOSPHATE	inhibitor	ChemblInteractions
POLE	CLOFARABINE	inhibitor	ChemblInteractions
EPHA2	VANDETANIB	inhibitor	ChemblInteractions
EPHA2	REGORAFENIB	inhibitor	MyCancerGenomeClinicalTrial
CYP2C9	BENZBROMARONE	inhibitor	DTC|PharmGKB

RET	REGORAFENIB	inhibitor	TALC|MyCancerGenome|JAX-CKB|ChemblInteractions|MyCancerGenomeClinicalTrial|TTD
RET	LENVATINIB	inhibitor	TALC|JAX-CKB|CIViC
RET	ALECTINIB HYDROCHLORIDE	inhibitor	ChemblInteractions
RET	PONATINIB	inhibitor	TALC|JAX-CKB|CIViC|GuideToPharmacology|MyCancerGenomeClinicalTrial|TTD
RET	SORAFENIB	inhibitor	DTC|MyCancerGenome|JAX-CKB|CIViC|GuideToPharmacology|PharmGKB
RET	SORAFENIB TOSYLATE	inhibitor	ChemblInteractions
RET	SUNITINIB MALATE	inhibitor	ChemblInteractions
RET	IMATINIB	inhibitor	TdgClinicalTrial|TEND
RET	SUNITINIB	inhibitor	TALC|MyCancerGenome|TdgClinicalTrial|JAX-CKB|NCI|CGI|TEND|DoCM|CIViC|GuideToPharmacology|PharmGKB
RET	VANDETANIB	inhibitor	TALC|DTC|MyCancerGenome|TdgClinicalTrial|ClearityFoundationClinicalTrial|JAX-CKB|ChemblInteractions|CGI|DoCM|CIViC|GuideToPharmacology|MyCancerGenomeClinicalTrial|PharmGKB|TTD|OncoKB
RET	GILTERITINIB	inhibitor	GuideToPharmacology
NDUFA10	METFORMIN HYDROCHLORIDE	inhibitor	ChemblInteractions
NDUFA9	METFORMIN HYDROCHLORIDE	inhibitor	ChemblInteractions
PARP1	RUCAPARIB	inhibitor|antagonist	TALC|MyCancerGenome|ClearityFoundationClinicalTrial|ChemblInteractions|GuideToPharmacology|PharmGKB
EPHA2	DASATINIB	inhibitor|antagonist	DTC|TdgClinicalTrial|ChemblInteractions|CGI|TEND|DoCM|MyCancerGenomeClinicalTrial
RET	CABOZANTINIB	inhibitor|antagonist	TALC|MyCancerGenome|JAX-CKB|CGI|CIViC|GuideToPharmacology|MyCancerGenomeClinicalTrial|PharmGKB|FDA|OncoKB
ESR2	BAZEDOXIFENE ACETATE	modulator	ChemblInteractions
ESR2	TOREMIFENE CITRATE	modulator	ChemblInteractions
ESR2	RALOXIFENE HYDROCHLORIDE	modulator	ChemblInteractions
ESR2	ESTRAMUSTINE PHOSPHATE SODIUM	modulator	ChemblInteractions
ESR2	OSPEMIFENE	modulator	TdgClinicalTrial|ChemblInteractions
ESR2	CHLOROTRIANISENE	modulator	ChemblInteractions
ABL1	DASATINIB	multitarget|inhibitor	MyCancerGenome|TdgClinicalTrial|JAX-CKB|ChemblInteractions|CGI|TEND|DoCM|COSMIC|CIViC|GuideToPharmacology|PharmGKB|FDA|OncoKB

Table 13: One-tailed hypergeometric overlaps between candidate genes identified by GeneEMBED and reference sets of known AD genes. Overlaps are compared across GeneEMBED applied to three curated networks (STRING, HINT, Brain) and an unbiased network (HuRI) all using the Discovery-EA dataset. Related to Figure 1.

Gene Set (GS)	STRING		HINT		Brain		HuRI	
	Ovlp	Pval	Ovlp	Pval	Ovlp	Pval	Ovlp	Pval
GWAS Meta 1 (n = 25)	1	0.091	2	0.005	1	0.051	0	1
GWAS Meta 2 (n = 38)	1	0.142	2	0.015	1	0.057	0	1
Comp. Tox. Database (n = 103)	2	0.047	4	0.001	2	0.024	1	0.34
ClinVar (n = 21)	1	0.081	1	0.11	1	0.068	0	1
DisGeNet (n = 208)	5	1.34E-03	3	0.11	2	0.096	1	0.58

Table 14: nDiffusion AUC and permutation based z-scores between reference sets of known AD-genes and GeneEMBED candidate genes. Comparisons of nDiffusion performance is compared across GeneEMBED applied to three curated networks (STRING, HINT, Brain) and an unbiased network (HuRI) all on the Discovery-EA dataset. Related to Figure 1, 3.

Gene Set (GS)	STRING		HINT		Brain		HuRI	
	AUC	Z-score	AUC	Z-score	AUC	Z-score	AUC	Z-score
GWAS Meta 1 (n = 25)	0.74	3.71	0.68	1.86	0.71	2.69	0.51	-1.59
GWAS Meta 2 (n = 38)	0.63	2.54	0.62	2.37	0.63	2.46	0.54	-0.12
Comp. Tox. Database (n = 103)	0.76	2.03	0.77	4.34	0.72	2.11	0.68	0.48
ClinVar (n = 21)	0.78	2.3	0.76	3.34	0.78	3.91	0.66	0.76
DisGeNet (n = 208)	0.69	3.26	0.7	5.77	0.66	2.15	0.61	-0.41

Table 15: Number of brain regions enriched and the corresponding permutation testing p-value of enrichment is given for GeneEMBED candidiates identified on three curated networks (STRING, HINT, Brain) and an unbiased network (HuRI). Related to Figure 1, 2

	STRING	HINT	Brain	HuRI
Num. of Regions Enriched	2	1	0	1
p-value	0.012	0.048	1	0.06

Table 16: One-tailed hypergeometric overlaps between GeneEMBED candidate genes and reference sets of known AD genes. Comparisons are between the original GeneEMBED implementation, modified implementation which uses full-dimensional embeddings, and modified implementations of max(PS) or thresholded PS for edge weight determination. Related to Figure 1.

Gene Set (GS)	Original Experiment (n=69)		Full Embedding distance (n = 82)		Max PS (n = 73)		PS Threshold (PS < 0.7) (n=72)	
	Ovlp	Pval	Ovlp	Pval	Ovlp	Pval	Ovlp	Pval
GWAS Meta 1 (n = 25)	1	0.091	2	0.006	1	0.101	1	0.1
GWAS Meta 2 (n = 38)	1	0.142	2	0.017	1	0.16	1	0.158
Comp. Tox. Database (n = 103)	2	0.047	2	0.073	2	0.057	2	0.057
ClinVar (n = 21)	1	0.081	1	0.1	1	0.088	1	0.087
DisGeNet (n = 208)	5	1.34 E-03	7	7.40 E-05	5	1.97 E-03	5	1.85 E-03

Table 17: nDiffusion AUC and permutation testing based z-scores between GeneEMBED candidates and reference gene sets are shown below. Comparisons are made between the original implementation of GeneEMBED, modified implementation wherein full-dimensional embeddings are used, and modified edge weighting schemes using Max(PS) or a thresholded PS. Related to Figure 1 and 3.

Gene Set (GS)	Original Experiment (n=69)		Full Embedding distance (n = 82)		Max PS (n = 73)		PS Threshold (PS < 0.7) (n=72)	
	AUC	Z-score	AUC	Z-score	AUC	Z-score	AUC	Z-score
GWAS Meta 1 (n = 25)	0.74	3.71	0.73	2.82	0.73	2.81	0.75	3.47
GWAS Meta 2 (n = 38)	0.63	2.54	0.64	2.05	0.63	1.8	0.65	2.73
Comp. Tox. Database (n = 103)	0.76	2.03	0.78	2.74	0.75	1.27	0.76	1.64
ClinVar (n = 21)	0.78	2.3	0.78	1.96	0.76	2.75	0.77	3.53
DisGeNet (n = 208)	0.69	3.26	0.67	0.97	0.72	1.51	0.7	2.27

Table 18: Number of brain regions enriched and the corresponding permutation testing based p-value of enrichment is given for GeneEMBED candidiates identified using the original framework, modified framework of full-dimensional embeddings, modified edge weighting schemes of max(PS) and a PS threshold. Related to Figure 1, 2.

	Original Experiment (n = 69)	Full Embedding Distances (n = 84)	Max PS (n = 73)	PS Threshold (PS < 0.7) (n = 72)
Num. of Regions Enriched	2	0	1	2
p-value	0.012	1	0.076	0.0014

Table 19: One tailed hypergeometric overlap between reference gene sets (column label) and candidate genes identified by GeneEMBED-VIS-**EA** (row label) using either the observed case vs control experiment or randomly shuffled patient labels. Related to Figure 1

EA Analyses	CTD	GWAS Meta 1	GWAS Meta 2	DGN	ClinVar
observed	0.003991	0.022989	0.0356475	0.015473	0.019344
random 1	0.116209	1	1	0.221469	1
random 2	0.128265	1	1	0.242841	1
random 3	1	1	1	1	1
random 4	0.151891	1	1	0.283848	1
random 5	0.128265	1	1	0.242841	1
random 6	0.134232	0.03429239	0.053003856	0.253307	0.028882

random 7	1	1	1	1	1
random 8	1	1	1	1	1
random 9	0.015581	1	1	0.056689	1
random 10	0.186145	1	1	0.341251	1
random 11	0.016661	0.04705683	0.072467128	0.060349	0.039674
random 12	1	1	1	0.033525	1
random 13	1	1	1	0.263628	1
random 14	1	1	1	1	1
random 15	1	1	1	1	1
random 16	0.151891	0.03909817	0.060348772	0.283848	0.032942
random 17	1	1	1	0.039673	1
random 18	1	1	1	1	1
random 19	1	1	1	0.210559	1
random 20	1	1	1	1	1
random 21	1	1	1	1	1
random 22	1	1	1	1	1
random 23	0.128265	1	1	1	0.027525
random 24	0.180533	1	1	0.332012	1
random 25	0.180533	1	1	0.332012	1

Table 20: One-tailed hypergeometric overlap between reference gene sets (column label) and candidate genes identified by GeneEMBED-VIS-**PPh2** (row label) using either the observed case vs control experiment or randomly shuffled patient labels. Related to Figure 1

PPh2 Analyses	CTD	GWAS Meta 1	GWAS Meta 2	DGN	ClinVar
observed	1	1	1	1	1
random 1	1	1	1	0.36821	1
random 2	1	1	1	0.33201	1
random 3	1	1	1	1	1
random 4	1	1	1	1	1
random 5	1	1	1	0.30351	1
random 6	0.17488	1	1	0.05669	1
random 7	1	1	1	1	1
random 8	1	1	1	1	1
random 9	1	1	1	0.03967	1
random 10	1	1	1	1	1
random 11	1	1	1	1	1
random 12	0.17488	1	1	1	0.03833
random 13	1	1	1	1	1

random 14	1	1	1	1	1
random 15	0.1577	1	1	1	1
random 16	0.15189	1	1	0.28385	1
random 17	0.00978	1	1	0.26363	0.03024
random 18	1	1	1	1	1
random 19	1	1	1	0.30351	1
random 20	1	1	1	0.26363	1
random 21	1	1	1	1	1
random 22	0.16346	1	1	0.30351	0.03564
random 23	1	1	1	1	1
random 24	1	1	1	0.24284	1
random 25	0.01253	1	1	0.29375	1

REFERENCES

1. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. Essays Biochem [Internet]. 2018 Dec 3 [cited 2022 Jun 18];62(5):643–723. Available from: https://pubmed.ncbi.nlm.nih.gov/30509934/

2. Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, Eiberg H, Mohr J, Kennedy D, Plavsic N, Zsiga M, Markiewicz D, Akots G, Brown V, Helms C, Gravius T, Parker C, Rediker K, Donis-Keller H. Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic DNA Marker. Science (1979) [Internet]. 1985 Nov 29 [cited 2022 Jun 23];230(4729):1054–7. Available from: https://www.science.org/doi/10.1126/science.2997931

3. Romeo G, Bianco M, Devoto M, Menozzi P, Mastella G, Giunta AM, Micalizzi C, Antonelli M, Battistini A, Santamaria F. Incidence in Italy, genetic heterogeneity, and segregation analysis of cystic fibrosis. American Journal of Human Genetics [Internet]. 1985 [cited 2022 Jun 23];37(2):338. Available from: /pmc/articles/PMC1684572/?report=abstract

4. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet [Internet]. 2019 Aug 1 [cited 2022 Jun 18];20(8):467–84. Available from: https://pubmed.ncbi.nlm.nih.gov/31068683/

5. Henderson P, Stevens C. The Role of Autophagy in Crohn's Disease. Cells [Internet]. 2012 Aug 3 [cited 2022 Jun 18];1(3):492. Available from: /pmc/articles/PMC3901108/

6. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, MacKay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009.

7. Gatz M, Pedersen NL, Berg S, Johansson B, Johansson K, Mortimer JA, Posner SF, Viitanen M, Winblad B, Ahlbom A. Heritability for Alzheimer's disease: The study of dementia in Swedish twins. Journals of Gerontology - Series A Biological Sciences and Medical Sciences. 1997;

8. Guerreiro R, Escott-Price V, Darwent L, Parkkinen L, Ansorge O, Hernandez DG, Nalls MA, Clark L, Honig L, Marder K, van der Flier W, Holstege H, Louwersheimer E, Lemstra A, Scheltens P, Rogaeva E, St George-Hyslop P, Londos E, Zetterberg H, Ortega-Cubero S, Pastor P, Ferman TJ, Graff-Radford NR, Ross OA, Barber I, Braae A, Brown K, Morgan K, Maetzler W, Berg D, Troakes C, Al-Sarraj S, Lashley T, Compta Y, Revesz T, Lees A, Cairns NJ, Halliday GM, Mann D, Pickering-Brown S, Powell J, Lunnon K, Lupton MK, Dickson D, Hardy J, Singleton A, Bras J. Genome-wide analysis of genetic correlation in dementia with Lewy bodies, Parkinson's and Alzheimer's diseases. Neurobiol Aging [Internet]. 2016 Oct 13 [cited 2021 Dec 12];38:214.e7-214.e10. Available from: https://pubmed.ncbi.nlm.nih.gov/26643944/

9. E G. Missing heritability of complex diseases: case solved? Hum Genet [Internet]. 2020 Jan 1 [cited 2021 Dec 9];139(1):103–13. Available from: https://pubmed.ncbi.nlm.nih.gov/31165258/

10. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Consortium ADG. Alzheimer's Disease: Analyzing the Missing Heritability. PLOS ONE [Internet]. 2013 Nov 7 [cited 2021 Dec 9];8(11):e79771. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079771

11. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008 Nov 6;456(7218):18–21.

12. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet [Internet]. 2004 [cited 2021 Dec 9];75(3):353–62. Available from: https://pubmed.ncbi.nlm.nih.gov/15272419/

13. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, Mattheisen M, Wang Y, Coleman JRI, Gaspar HA, de Leeuw CA, Steinberg S, Pavlides JMW, Trzaskowski M, Byrne EM, Pers TH, Holmans PA, Richards AL, Abbott L, Agerbo E, Akil H, Albani D, Alliey-Rodriguez N, Als TD, Anjorin A, Antilla V, Awasthi S, Badner JA, Bækvad-Hansen M, Barchas JD, Bass N, Bauer M, Belliveau R, Bergen SE, Pedersen CB, Bøen E, Boks MP, Boocock J, Budde M, Bunney W, Burmeister M, Bybjerg-Grauholm J, Byerley W, Casas M, Cerrato F, Cervantes P, Chambert K, Charney AW, Chen D, Churchhouse C, Clarke TK, Coryell W, Craig DW, Cruceanu C, Curtis D, Czerski PM, Dale AM, de Jong S, Degenhardt F, Del-Favero J, DePaulo JR, Djurovic S, Dobbyn AL, Dumont A, Elvsåshagen T, Escott-Price V, Fan CC, Fischer SB, Flickinger M, Foroud TM, Forty L, Frank J, Fraser C, Freimer NB, Frisén L, Gade K, Gage D, Garnham J, Giambartolomei C, Pedersen MG, Goldstein J, Gordon SD, Gordon-Smith K, Green EK, Green MJ, Greenwood TA, Grove J, Guan W, Guzman-Parra J, Hamshere ML, Hautzinger M, Heilbronner U, Herms S, Hipolito M, Hoffmann P, Holland D, Huckins L, Jamain S, Johnson JS, Juréus A, Kandaswamy R, Karlsson R, Kennedy JL, Kittel-Schneider S, Knowles JA, Kogevinas M, Koller AC, Kupka R, Lavebratt C, Lawrence J, Lawson WB, Leber M, Lee PH, Levy SE, Li JZ, Liu C, Lucae S, Maaser A, MacIntyre DJ, Mahon PB, Maier W, Martinsson L, McCarroll S, McGuffin P, McInnis MG, McKay JD, Medeiros H, Medland SE, Meng F, Milani L, Montgomery GW, Morris DW, Mühleisen TW, Mullins N, Nguyen H, Nievergelt CM, Adolfsson AN, Nwulia EA, O'Donovan C, Loohuis LMO, Ori APS, Oruc L, Ösby U, Perlis RH, Perry A, Pfennig A, Potash JB, Purcell SM, Regeer EJ, Reif A, Reinbold CS, Rice JP, Rivas F, Rivera M, Roussos P, Ruderfer DM, Ryu E, Sánchez-Mora C, Schatzberg AF, Scheftner WA, Schork NJ, Shannon Weickert C, Shehktman T, Shilling PD, Sigurdsson E, Slaney C, Smeland OB, Sobell JL, Søholm Hansen C, Spijker AT, St Clair D, Steffens M, Strauss JS, Streit F, Strohmaier J, Szelinger S, Thompson RC, Thorgeirsson TE, Treutlein J, Vedder H, Wang W, Watson SJ, Weickert TW, Witt SH, Xi S, Xu W, Young AH, Zandi P, Zhang P, Zöllner S, Adolfsson R, Agartz I, Alda M, Backlund L, Baune BT, Bellivier F, Berrettini WH, Biernacka JM, Blackwood DHR, Boehnke M, Børglum AD, Corvin A, Craddock N, Daly MJ, Dannlowski U, Esko T, Etain B, Frye M, Fullerton JM, Gershon ES, Gill M, Goes F, Grigoroiu-Serbanescu M, Hauser J, Hougaard DM, Hultman CM, Jones I, Jones LA, Kahn RS, Kirov G, Landén M, Leboyer M, Lewis CM, Li QS, Lissowska J, Martin NG, Mayoral F, McElroy SL, McIntosh AM, McMahon FJ, Melle I, Metspalu A, Mitchell PB, Morken G, Mors O, Mortensen PB, Müller-Myhsok B, Myers RM, Neale BM, Nimgaonkar V, Nordentoft M, Nöthen MM, O'Donovan MC, Oedegaard KJ, Owen MJ, Paciga SA, Pato C, Pato MT, Posthuma D, Ramos-Quiroga JA, Ribasés M, Rietschel M, Rouleau GA, Schalling M, Schofield PR, Schulze TG, Serretti A, Smoller JW, Stefansson H, Stefansson K, Stordal E, Sullivan PF, Turecki G,

Vaaler AE, Vieta E, Vincent JB, Werge T, Nurnberger JI, Wray NR, di Florio A, Edenberg HJ, Cichon S, Ophoff RA, Scott LJ, Andreassen OA, Kelsoe J, Sklar P. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nature Genetics. 2019;

14. Ripke S, Neale BM, Corvin A, Walters JTR, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, Pers TH, Agartz I, Agerbo E, Albus M, Alexander M, Amin F, Bacanu SA, Begemann M, Belliveau RA, Bene J, Bergen SE, Bevilacqua E, Bigdeli TB, Black DW, Bruggeman R, Buccola NG, Buckner RL, Byerley W, Cahn W, Cai G, Campion D, Cantor RM, Carr VJ, Carrera N, Catts S v., Chambert KD, Chan RCK, Chen RYL, Chen EYH, Cheng W, Cheung EFC, Chong SA, Cloninger CR, Cohen D, Cohen N, Cormican P, Craddock N, Crowley JJ, Curtis D, Davidson M, Davis KL, Degenhardt F, del Favero J, Demontis D, Dikeos D, Dinan T, Djurovic S, Donohoe G, Drapeau E, Duan J, Dudbridge F, Durmishi N, Eichhammer P, Eriksson J, Escott-Price V, Essioux L, Fanous AH, Farrell MS, Frank J, Franke L, Freedman R, Freimer NB, Friedl M, Friedman JI, Fromer M, Genovese G, Georgieva L, Giegling I, Giusti-Rodríguez P, Godard S, Goldstein JI, Golimbet V, Gopal S, Gratten J, de Haan L, Hammer C, Hamshere ML, Hansen M, Hansen T, Haroutunian V, Hartmann AM, Henskens FA, Herms S, Hirschhorn JN, Hoffmann P, Hofman A, Hollegaard M v., Hougaard DM, Ikeda M, Joa I, Julià A, Kahn RS, Kalaydjieva L, Karachanak-Yankova S, Karjalainen J, Kavanagh D, Keller MC, Kennedy JL, Khrunin A, Kim Y, Klovins J, Knowles JA, Konte B, Kucinskas V, Kucinskiene ZA, Kuzelova-Ptackova H, Kähler AK, Laurent C, Keong JLC, Lee SH, Legge SE, Lerer B, Li M, Li T, Liang KY, Lieberman J, Limborska S, Loughland CM, Lubinski J, Lönnqvist J, Macek M, Magnusson PKE, Maher BS, Maier W, Mallet J, Marsal S, Mattheisen M, Mattingsdal M, McCarley RW, McDonald C, McIntosh AM, Meier S, Meijer CJ, Melegh B, Melle I, Mesholam-Gately RI, Metspalu A, Michie PT, Milani L, Milanova V, Mokrab Y, Morris DW, Mors O, Murphy KC, Murray RM, Myin-Germeys I, Müller-Myhsok B, Nelis M, Nenadic I, Nertney DA, Nestadt G, Nicodemus KK, Nikitina-Zake L, Nisenbaum L, Nordin A, O'Callaghan E, O'Dushlaine C, O'Neill FA, Oh SY, Olincy A, Olsen L, van Os J, Pantelis C, Papadimitriou GN, Papiol S, Parkhomenko E, Pato MT, Paunio T, Pejovic-Milovancevic M, Perkins DO, Pietiläinen O, Pimm J, Pocklington AJ, Powell J, Price A, Pulver AE, Purcell SM, Quested D, Rasmussen HB, Reichenberg A, Reimers MA, Richards AL, Roffman JL, Roussos P, Ruderfer DM, Salomaa V, Sanders AR, Schall U, Schubert CR, Schulze TG, Schwab SG, Scolnick EM, Scott RJ, Seidman LJ, Shi J, Sigurdsson E, Silagadze T, Silverman JM, Sim K, Slominsky P, Smoller JW, So HC, Spencer CCA, Stahl EA, Stefansson H, Steinberg S, Stogmann E, Straub RE, Strengman E, Strohmaier J, Stroup TS, Subramaniam M, Suvisaari J, Svrakic DM, Szatkiewicz JP, Söderman E, Thirumalai S, Toncheva D, Tosato S, Veijola J, Waddington J, Walsh D, Wang D, Wang Q, Webb BT, Weiser M, Wildenauer DB, Williams NM, Williams S, Witt SH, Wolen AR, Wong EHM, Wormley BK, Xi HS, Zai CC, Zheng X, Zimprich F, Wray NR, Stefansson K, Visscher PM, Adolfsson R, Andreassen OA, Blackwood DHR, Bramon E, Buxbaum JD, Børglum AD, Cichon S, Darvasi A, Domenici E, Ehrenreich H, Esko T, Gejman P v., Gill M, Gurling H, Hultman CM, Iwata N, Jablensky A v., Jönsson EG, Kendler KS, Kirov G, Knight J, Lencz T, Levinson DF, Li QS, Liu J, Malhotra AK, McCarroll SA, McQuillin A, Moran JL, Mortensen PB, Mowry BJ, Nöthen MM, Ophoff RA, Owen MJ, Palotie A, Pato CN, Petryshen TL, Posthuma D, Rietschel M, Riley BP, Rujescu D, Sham PC, Sklar P, St Clair D, Weinberger DR, Wendland JR, Werge T, Daly MJ, Sullivan PF, O'Donovan MC. Biological insights from 108 schizophrenia-associated genetic loci. Nature 2014 511:7510 [Internet]. 2014 Jul 22 [cited

2022 Jul 28];511(7510):421–7. Available from: https://www.nature.com/articles/nature13595

15. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics [Internet]. 2012 Sep [cited 2022 Jun 18];13(4):762–75. Available from: https://pubmed.ncbi.nlm.nih.gov/22699862/

16. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLOS Computational Biology [Internet]. 2015 Apr 1 [cited 2021 Dec 9];11(4):e1004219. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004219

17. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. The American Journal of Human Genetics. 2010 Jun 11;86(6):860–71.

18. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 2012;

19. Cordell HJ. Data mining Machine learning Detecting gene-gene interactions that underlie human diseases. 2009 [cited 2021 Dec 13]; Available from: www.nature.com/reviews/genetics

20. Forsberg SKG, Bloom JS, Sadhu MJ, Kruglyak L, Carlborg Ö. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. Nat Genet [Internet]. 2017 Mar 30 [cited 2021 Dec 9];49(4):497–503. Available from: https://pubmed.ncbi.nlm.nih.gov/28250458/

21. Martínez A, Sánchez E, Valdivia A, Orozco G, López-Nevot MA, Pascual-Salcedo D, Balsa A, Fernández-Gutiérrez B, de La Concha EG, García-Sánchez A, Koeleman BPC, Urcelay E, Martín J. Epistatic interaction between FCRL3 and NFκB1 genes in Spanish patients with rheumatoid arthritis. Annals of the Rheumatic Diseases. 2006;

22. Martin MP, Gao X, Lee JH, Nelson GW, Detels R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, Trowsdale J, Wilson M, O'Brien SJ, Carrington M. Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. Nature Genetics. 2002;

23. Leggio GM, Torrisi SA, Mastrogiacomo R, Mauro D, Chisari M, Devroye C, Scheggia D, Nigro M, Geraci F, Pintori N, Giurdanella G, Costa L, Bucolo C, Ferretti V, Sortino MA, Ciranna L, de Luca MA, Mereu M, Managò F, Salomone S, Drago F, Papaleo F. The epistatic interaction between the dopamine D3 receptor and dysbindin-1 modulates higher-order cognitive functions in mice and humans. Molecular Psychiatry. 2018;

24. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics. 2008 Apr 11;82(4):949–58.

25. Östlund G, Lindskog M, Sonnhammer ELL. Network-based Identification of novel cancer genes. Mol Cell Proteomics [Internet]. 2010 Apr [cited 2022 Jun 18];9(4):648–55. Available from: https://pubmed.ncbi.nlm.nih.gov/19959820/

26. Guney E, Oliva B. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. PLoS One [Internet]. 2012 Sep 21 [cited 2021 Dec 9];7(9). Available from: https://pubmed.ncbi.nlm.nih.gov/23028459/

27. Nitsch D, Tranchevent LC, Gonalves JP, Vogt JK, Madeira SC, Moreau Y. PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Research [Internet]. 2011 Jul 1 [cited 2021 Dec 9];39(suppl_2):W334–8. Available from: https://academic.oup.com/nar/article/39/suppl_2/W334/2505844

28. Shrestha R, Hodzic E, Sauerwald T, Dao P, Wang K, Yeung J, Anderson S, Vandin F, Haffari G, Collins CC, Sahinalp SC. HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. 2017 [cited 2021 Dec 9]; Available from: http://www.genome.org/cgi/doi/10.1101/gr.221218.117.

29. Leiserson MDM, Vandin F, Wu HT, Dobson JR, Eldridge J v., Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L, Raphael BJ. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics 2014 47:2 [Internet]. 2014 Dec 15 [cited 2021 Dec 9];47(2):106–14. Available from: https://www.nature.com/articles/ng.3168

30. Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. Bioinformatics [Internet]. 2018 Sep 1 [cited 2022 Jun 18];34(17):i972–80. Available from: https://pubmed.ncbi.nlm.nih.gov/30423088/

31. Zhang XM, Liang L, Liu L, Tang MJ. Graph Neural Networks and Their Current Applications in Bioinformatics. Front Genet [Internet]. 2021 Jul 29 [cited 2022 Jun 18];12. Available from: https://pubmed.ncbi.nlm.nih.gov/34394185/

32. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics [Internet]. 2018 Jul 1 [cited 2022 Jun 18];34(13):i457–66. Available from: https://pubmed.ncbi.nlm.nih.gov/29949996/

33. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun [Internet]. 2021 Dec 1 [cited 2022 Jun 18];12(1). Available from: https://pubmed.ncbi.nlm.nih.gov/34103512/

34. Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, Auer F, Kramer F, Leha A, Beißbarth T. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. Genome Med [Internet]. 2021 Dec 1 [cited 2022 Jun 18];13(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33706810/

35. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral Networks and Locally Connected Networks on Graphs. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings [Internet]. 2013 Dec 21 [cited 2022 Jun 18]; Available from: https://arxiv.org/abs/1312.6203v3

36. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings [Internet]. 2016 Sep 9 [cited 2021 Dec 9]; Available from: https://arxiv.org/abs/1609.02907v4

37. Xu K, Jegelka S, Hu W, Leskovec J. How powerful are graph neural networks? In: 7th International Conference on Learning Representations, ICLR 2019. 2019.

38. Veličković P, Casanova A, Liò P, Cucurull G, Romero A, Bengio Y. Graph Attention Networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings [Internet]. 2017 Oct 30 [cited 2021 Dec 9]; Available from: https://arxiv.org/abs/1710.10903v3

39. Ingraham J, Garg VK, Barzilay R, Jaakkola T. Generative models for graph-based protein design.

40. Belkin M, Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.

41. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. [cited 2021 Dec 9]; Available from: http://dx.doi.org/10.1145/2623330.2623732

42. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 2016 Jul 3 [cited 2021 Dec 9];13-17-August-2016:855–64. Available from: https://arxiv.org/abs/1607.00653v1

43. Ribeiro LFR, Savarese PHP, Figueiredo DR. struc2vec: Learning Node Representations from Structural Identity. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 2017 Apr 11 [cited 2022 Jun 18];Part F129685:385–94. Available from: http://arxiv.org/abs/1704.03165

44. Donnat C, Zitnik M, Hallac D, Leskovec J. Learning Structural Node Embeddings via Diffusion Wavelets. 2018;1320–9.

45. Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nature Machine Intelligence 2021 3:6 [Internet]. 2021 Apr 12 [cited 2022 Jun 18];3(6):513–26. Available from: https://www.nature.com/articles/s42256-021-00325-y

46. Ravindra N, Sehanobish A, Pappalardo JL, Hafler DA, van Dijk D. Disease State Prediction From Single-Cell Data Using Graph Attention Networks. ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning [Internet]. 2020 Feb 14 [cited 2022 Jun 18];121–30. Available from: https://arxiv.org/abs/2002.07128v2

47. Zhou T. Progresses and challenges in link prediction. iScience. 2021 Nov 19;24(11).

48. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology [Internet]. 2007 May 1 [cited 2022 Jun 22];58(7):1019–31. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.20591

49. Cannistraci CV, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific Reports 2013 3:1 [Internet]. 2013 Apr 8 [cited 2022 Jun 22];3(1):1–14. Available from: https://www.nature.com/articles/srep01613

50. Wu G, Gu C, Yang H, Wolniewicz LM, Berger TA, Huber D, Puccio E, Vassallo P, Piilo J, Daminelli S, Thomas JM, Durán C, Cannistraci CV. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. New Journal of Physics [Internet]. 2015 Nov 16 [cited 2022 Jun 22];17(11):113037. Available from: https://iopscience.iop.org/article/10.1088/1367-2630/17/11/113037

51. Daud NN, Ab Hamid SH, Saadoon M, Sahran F, Anuar NB. Applications of link prediction in social networks: A review. Journal of Network and Computer Applications. 2020 Sep 15;166:102716.

52. Zhao H, Du L, Buntine W. Leveraging Node Attributes for Incomplete Relational Data. International Conference on Machine Learning [Internet]. 2017 [cited 2022 Jun 22]; Available from: https://github.com/

53. Das S, Das SK. A probabilistic link prediction model in time-varying social networks. IEEE International Conference on Communications. 2017 Jul 28;

54. Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics [Internet]. 2012 Sep 15 [cited 2022 Jun 22];28(18):2304–10. Available from: https://academic.oup.com/bioinformatics/article/28/18/2304/241817

55. Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I. Predicting Drug–Target Interactions Using Probabilistic Matrix Factorization. Journal of Chemical Information and Modeling [Internet]. 2013 Dec 12 [cited 2022 Jun 22];53(12):3399. Available from: /pmc/articles/PMC3871285/

56. Li L, Wang W, Yu S, Wan L, Xu Z, Kong X. A Modified Node2vec Method for Disappearing Link Prediction. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). IEEE; 2017. p. 1232–5.

57. de Winter S, Decuypere T, Mitrovic S, Baesens B, de Weerdt J. Combining temporal aspects of dynamic networks with Node2Vec for a more efficient dynamic link prediction. Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018. 2018 Oct 24;1234–41.

58. Wu Y, Gao M, Zeng M, Zhang J, Li M. BridgeDPI: a novel Graph Neural Network for predicting drug–protein interactions. Bioinformatics [Internet]. 2022 Apr 28 [cited 2022 Jun 18];38(9):2571–8. Available from: https://academic.oup.com/bioinformatics/article/38/9/2571/6547049

59. Gärtner T, Flach P, Wrobel S. On graph kernels: Hardness results and efficient alternatives. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science) [Internet]. 2003 [cited 2022 Jun 18];2777:129–43. Available from: https://link.springer.com/chapter/10.1007/978-3-540-45167-9_11

60. Schweitzer PASCAL P, Jan van Leeuwen EJVANLEEUWEN E, Shervashidze N, Schweitzer P, Jan van Leeuwen E, Mehlhorn K, Borgwardt SHERVASHIDZE KM, Leeuwen V. Weisfeiler-Lehman Graph Kernels Nino Shervashidze Kurt Mehlhorn Karsten M. Borgwardt. Journal of Machine Learning Research. 2011;12:2539–61.

61. Mautner S, Montaseri S, Miladi M, Raden M, Costa F, Backofen R. ShaKer: RNA SHAPE prediction using graph kernel. Bioinformatics [Internet]. 2019 Jul 15 [cited 2022 Jun 22];35(14):i354–9. Available from: https://academic.oup.com/bioinformatics/article/35/14/i354/5529262

62. Tepeli YI, Ünal AB, Akdemir FM, Tastan O. PAMOGK: a pathway graph kernel-based multiomics approach for patient clustering. Bioinformatics [Internet]. 2021 Jan 29 [cited 2022 Jun 22];36(21):5237–46. Available from: https://academic.oup.com/bioinformatics/article/36/21/5237/5878954

63. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP. Convolutional Networks on Graphs for Learning Molecular Fingerprints. NeurIPS. 2015;

64. Li R, Wang S, Zhu F, Huang J. Adaptive Graph Convolutional Neural Networks. AAAI [Internet]. 2018 Jan 9 [cited 2022 Jun 22]; Available from: http://arxiv.org/abs/1801.03226

65. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, Park DJ. Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. Human Genomics [Internet]. 2017 Mar 4 [cited 2022 Jun 18];11(1):1–8. Available from: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-017-0104-8

66. Katsonis P, Lichtarge O. Objective assessment of the evolutionary action equation for the fitness effect of missense mutations across CAGI-blinded contests. Hum Mutat [Internet]. 2017 Sep 1 [cited 2022 May 3];38(9):1072–84. Available from: https://pubmed.ncbi.nlm.nih.gov/28544059/

67. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature Methods. 2013.

68. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research [Internet]. 2003 Jul 7 [cited 2022 Jun 19];31(13):3812. Available from: /pmc/articles/PMC168916/

69. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines. 2014;2050–8.

70. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet [Internet]. 2014 [cited 2022 Jun 19];46(3):310–5. Available from: https://pubmed.ncbi.nlm.nih.gov/24487276/

71. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American Journal of Human Genetics [Internet]. 2016 Oct 10 [cited 2022 Jun 23];99(4):877. Available from: /pmc/articles/PMC5065685/

72. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. Journal of Molecular Biology. 1996;

73. Erdin S, Ward RM, Venner E, Lichtarge O. Evolutionary Trace Annotation of Protein Function in the Structural Proteome. Journal of Molecular Biology. 2010;

74. Hsu AP, Sampaio EP, Khan J, Calvo KR, Lemieux JE, Patel SY, Frucht DM, Vinh DC, Auth RD, Freeman AF, Olivier KN, Uzel G, Zerbe CS, Spalding C, Pittaluga S, Raffeld M, Kuhns DB, Ding L, Paulson ML, Marciano BE, Gea-Banacloche JC, Orange JS, Cuellar-Rodriguez J, Hickstein DD, Holland SM. Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. Blood [Internet]. 2011 Sep 8 [cited 2022 Jun 19];118(10):2653–5. Available from: https://pubmed.ncbi.nlm.nih.gov/21670465/

75. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science [Internet]. 2012 Feb 17 [cited 2022 Jun 19];335(6070):823–8. Available from: https://pubmed.ncbi.nlm.nih.gov/22344438/

76. Knopman DS, Amieva H, Petersen RC, Chételat G, Holtzman DM, Hyman BT, Nixon RA, Jones DT. Alzheimer disease. Nat Rev Dis Primers. 2021;7(1):33.

77. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology. 2013 May 7;80(19):1778–83.

78. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, Bellenguez C, Frizatti A, Chouraki V, Martin ER, Sleegers K, Badarinarayan N, Jakobsdottir J, Hamilton-Nelson KL, Moreno-Grau S, Olaso R, Raybould R, Chen Y, Kuzma AB, Hiltunen M, Morgan T, Ahmad S, Vardarajan BN, Epelbaum J, Hoffmann P, Boada M, Beecham GW, Garnier JG, Harold D, Fitzpatrick AL, Valladares O, Moutet ML, Gerrish A, Smith A v., Qu L, Bacq D, Denning N, Jian X, Zhao Y, del Zompo M, Fox NC, Choi SH, Mateo I, Hughes JT, Adams HH, Malamon J, Sanchez-Garcia F, Patel Y, Brody JA, Dombroski BA, Naranjo MCD, Daniilidou M, Eiriksdottir G, Mukherjee S, Wallon D, Uphill J, Aspelund T, Cantwell LB, Garzia F, Galimberti D, Hofer E, Butkiewicz M, Fin B, Scarpini E, Sarnowski C, Bush WS, Meslage S, Kornhuber J, White CC, Song Y, Barber RC, Engelborghs S, Sordon S, Voijnovic D, Adams PM, Vandenberghe R, Mayhaus M, Cupples LA, Albert MS, de Deyn PP, Gu W, Himali JJ, Beekly D, Squassina A, Hartmann AM, Orellana A, Blacker D, Rodriguez-Rodriguez E, Lovestone S, Garcia ME, Doody RS, Munoz-Fernadez C, Sussams R, Lin H, Fairchild TJ, Benito YA, Holmes C, Karamujić-Čomić H, Frosch MP, Thonberg H, Maier W, Roschupkin G, Ghetti B, Giedraitis V, Kawalia A, Li S, Huebinger RM, Kilander L, Moebus S, Hernández I, Kamboh MI, Brundin RM, Turton J, Yang Q, Katz MJ, Concari L, Lord J, Beiser AS, Keene CD, Helisalmi S, Kloszewska I, Kukull WA, Koivisto AM, Lynch A, Tarraga L, Larson EB, Haapasalo A, Lawlor B, Mosley TH, Lipton RB, Solfrizzi V, Gill M, Longstreth WT, Montine TJ, Frisardi V, Diez-Fairen M, Rivadeneira F, Petersen RC, Deramecourt V, Alvarez I, Salani F, Ciaramella A, Boerwinkle E, Reiman EM, Fievet N, Rotter JI, Reisch JS, Hanon O, Cupidi C, Andre Uitterlinden AG, Royall DR, Dufouil C, Maletta RG, de Rojas I, Sano M, Brice A, Cecchetti R, George-Hyslop PS, Ritchie K, Tsolaki M, Tsuang DW, Dubois B, Craig D, Wu CK, Soininen H, Avramidou D, Albin RL, Fratiglioni L, Germanou A, Apostolova LG, Keller L, Koutroumani M, Arnold SE, Panza F, Gkatzima O, Asthana S, Hannequin D, Whitehead P, Atwood CS, Caffarra P, Hampel H, Quintela I, Carracedo Á, Lannfelt L, Rubinsztein DC, Barnes LL, Pasquier F, Frölich L, Barral S, McGuinness B, Beach TG, Johnston JA, Becker JT, Passmore P, Bigio EH, Schott JM, Bird TD, Warren JD, Boeve BF, Lupton MK, Bowen JD, Proitsi P, Boxer A, Powell JF, Burke JR, Kauwe JSK, Burns JM, Mancuso M, Buxbaum JD, Bonuccelli U, Cairns NJ, McQuillin A, Cao C, Livingston G, Carlson CS, Bass NJ, Carlsson CM, Hardy J, Carney RM, Bras J, Carrasquillo MM, Guerreiro R, Allen M, Chui HC, Fisher E, Masullo C, Crocco EA, DeCarli C, Bisceglio G, Dick M, Ma L, Duara R, Graff-Radford NR, Evans DA, Hodges A, Faber KM, Scherer M, Fallon KB, Riemenschneider M, Fardo DW, Heun R, Farlow MR, Kölsch H, Ferris S, Leber M, Foroud TM, Heuser I, Galasko DR, Giegling I, Gearing M, Hüll M, Geschwind DH, Gilbert JR, Morris J, Green RC, Mayo K, Growdon JH, Feulner T, Hamilton RL, Harrell LE, Drichel D, Honig LS, Cushion TD, Huentelman MJ, Hollingworth P, Hulette CM, Hyman BT, Marshall R, Jarvik GP, Meggy A, Abner E, Menzies GE, Jin LW, Leonenko G, Real LM, Jun GR, Baldwin CT, Grozeva D, Karydas A, Russo G, Kaye JA, Kim R, Jessen F, Kowall NW, Vellas B, Kramer JH, Vardy E, LaFerla FM, Jöckel KH, Lah JJ, Dichgans M, Leverenz JB, Mann D, Levey AI, Pickering-Brown S, Lieberman AP, Klopp N, Lunetta KL, Wichmann HE, Lyketsos CG, Morgan K, Marson DC, Brown K, Martiniuk F, Medway C, Mash DC, Nöthen MM, Masliah E, Hooper NM, McCormick WC, Daniele A, McCurry SM, Bayer A, McDavid AN, Gallacher J, McKee AC, van den Bussche H, Mesulam M, Brayne C, Miller BL, Riedel-Heller S, Miller CA, Miller JW, Al-

Chalabi A, Morris JC, Shaw CE, Myers AJ, Wiltfang J, O'Bryant S, Olichney JM, Alvarez V, Parisi JE, Singleton AB, Paulson HL, Collinge J, Perry WR, Mead S, Peskind E, Cribbs DH, Rossor M, Pierce A, Ryan NS, Poon WW, Nacmias B, Potter H, Sorbi S, Quinn JF, Sacchinelli E, Raj A, Spalletta G, Raskind M, Caltagirone C, Bossù P, Orfei MD, Reisberg B, Clarke R, Reitz C, Smith AD, Ringman JM, Warden D, Roberson ED, Wilcock G, Rogaeva E, Bruni AC, Rosen HJ, Gallo M, Rosenberg RN, Ben-Shlomo Y, Sager MA, Mecocci P, Saykin AJ, Pastor P, Cuccaro ML, Vance JM, Schneider JA, Schneider LS, Slifer S, Seeley WW, Smith AG, Sonnen JA, Spina S, Stern RA, Swerdlow RH, Tang M, Tanzi RE, Trojanowski JQ, Troncoso JC, van Deerlin VM, van Eldik LJ, Vinters H v., Vonsattel JP, Weintraub S, Welsh-Bohmer KA, Wilhelmsen KC, Williamson J, Wingo TS, Woltjer RL, Wright CB, Yu CE, Yu L, Saba Y, Pilotto A, Bullido MJ, Peters O, Crane PK, Bennett D, Bosco P, Coto E, Boccardi V, de Jager PL, Lleo A, Warner N, Lopez OL, Ingelsson M, Deloukas P, Cruchaga C, Graff C, Gwilliam R, Fornage M, Goate AM, Sanchez-Juan P, Kehoe PG, Amin N, Ertekin-Taner N, Berr C, Debette S, Love S, Launer LJ, Younkin SG, Dartigues JF, Corcoran C, Ikram MA, Dickson DW, Nicolas G, Campion D, Tschanz JA, Schmidt H, Hakonarson H, Clarimon J, Munger R, Schmidt R, Farrer LA, van Broeckhoven C, C. O'Donovan M, DeStefano AL, Jones L, Haines JL, Deleuze JF, Owen MJ, Gudnason V, Mayeux R, Escott-Price V, Psaty BM, Ramirez A, Wang LS, Ruiz A, van Duijn CM, Holmans PA, Seshadri S, Williams J, Amouyel P, Schellenberg GD, Lambert JC, Pericak-Vance MA. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nature Genetics. 2019;

79. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, Young AMH, Franklin RJM, Johnson T, Estrada K, Gaffney DJ, Beltrao P, Bassett A. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nature Genetics. 2021;

80. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, Salerno WJ, Lancour D, Ma Y, Renton AE, Marcora E, Farrell JJ, Zhao Y, Qu L, Ahmad S, Amin N, Amouyel P, Beecham GW, Below JE, Campion D, Cantwell L, Charbonnier C, Chung J, Crane PK, Cruchaga C, Cupples LA, Dartigues JF, Debette S, Deleuze JF, Fulton L, Gabriel SB, Genin E, Gibbs RA, Goate A, Grenier-Boley B, Gupta N, Haines JL, Havulinna AS, Helisalmi S, Hiltunen M, Howrigan DP, Ikram MA, Kaprio J, Konrad J, Kuzma A, Lander ES, Lathrop M, Lehtimäki T, Lin H, Mattila K, Mayeux R, Muzny DM, Nasser W, Neale B, Nho K, Nicolas G, Patel D, Pericak-Vance MA, Perola M, Psaty BM, Quenez O, Rajabli F, Redon R, Reitz C, Remes AM, Salomaa V, Sarnowski C, Schmidt H, Schmidt M, Schmidt R, Soininen H, Thornton TA, Tosto G, Tzourio C, van der Lee SJ, van Duijn CM, Valladares O, Vardarajan B, Wang LS, Wang W, Wijsman E, Wilson RK, Witten D, Worley KC, Zhang X, Bellenguez C, Lambert JC, Kurki MI, Palotie A, Daly M, Boerwinkle E, Lunetta KL, Destefano AL, Dupuis J, Martin ER, Schellenberg GD, Seshadri S, Naj AC, Fornage M, Farrer LA. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Molecular Psychiatry 2018 25:8 [Internet]. 2018 Aug 14 [cited 2021 Dec 9];25(8):1859–75. Available from: https://www.nature.com/articles/s41380-018-0112-7

81. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, Sealock J, Karlsson IK, Hägg S, Athanasiu L, Voyle N, Proitsi P, Witoelar A, Stringer S, Aarsland D, Almdahl IS, Andersen F, Bergh S, Bettella F, Bjornsson S, Brækhus A, Bråthen G, de Leeuw C, Desikan RS,

Djurovic S, Dumitrescu L, Fladby T, Hohman TJ, Jonsson P v., Kiddle SJ, Rongve A, Saltvedt I, Sando SB, Selbæk G, Shoai M, Skene NG, Snaedal J, Stordal E, Ulstein ID, Wang Y, White LR, Hardy J, Hjerling-Leffler J, Sullivan PF, van der Flier WM, Dobson R, Davis LK, Stefansson H, Stefansson K, Pedersen NL, Ripke S, Andreassen OA, Posthuma D. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nature Genetics 2019 51:3 [Internet]. 2019 Jan 7 [cited 2021 Dec 9];51(3):404–13. Available from: https://www.nature.com/articles/s41588-018-0311-9

82. JC L, CA IV, D H, AC N, R S, C B, AL D, JC B, GW B, B GB, G R, TA TW, N J, AV S, V C, C T, MA I, D Z, BN V, Y K, CF L, A G, H S, B K, ML D, A R, MT B, SH C, C R, F P, C C, D C, N A, C B, OL L, PL DJ, V D, JA J, D E, S L, L L, FJ M, DC R, G E, K S, AM G, N F, MW H, M G, K B, MI K, L K, P BG, B M, EB L, R G, AJ M, C D, S T, D W, S L, E R, J G, P SGH, J C, A L, A B, DW T, L Y, M T, P B, G S, P P, J C, S S, F SG, NC F, J H, MC DN, P B, R C, C B, D G, M M, F M, S M, P M, M DZ, W M, H H, A P, M B, F P, P C, B N, JR G, M M, L L, H H, S P, MM C, M I, D B, V A, F Z, O V, SG Y, E C, KL HN, W G, C R, P P, I M, MJ O, KM F, PV J, O C, MC O, LB C, H S, D B, S M, TH M, DA B, TB H, L F, C H, RF de B, P P, TJ M, K B, JI R, A B, K M, TM F, WA K, D H, JF P, MA N, K R, KL L, JS K, E B, M R, M B, M H, ER M, R S, D R, LS W, JF D, R M, C T, A H, MM N, C G, BM P, L J, JL H, PA H, M L, MA PV, LJ L, LA F, CM van D, C VB, V M, S S, J W, GD S, P A. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet [Internet]. 2013 Dec 1 [cited 2021 Dec 9];45(12):1452–8. Available from: https://pubmed.ncbi.nlm.nih.gov/24162737/

83. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 2019 Aug 1;20(8):467–84.

84. Costanzo M, Kuzmin E, van Leeuwen J, Mair B, Moffat J, Boone C, Andrews B. Global Genetic Networks and the Genotype-to-Phenotype Relationship. Cell [Internet]. 2019 Mar 21 [cited 2021 Dec 9];177(1):85–100. Available from: https://pubmed.ncbi.nlm.nih.gov/30901552/

85. Östlund G, Lindskog M, Sonnhammer ELL. Network-based Identification of novel cancer genes. Mol Cell Proteomics [Internet]. 2010 Apr [cited 2021 Dec 9];9(4):648–55. Available from: https://pubmed.ncbi.nlm.nih.gov/19959820/

86. Fout A, Byrd J, Shariat B, Ben-Hur A. Protein Interface Prediction using Graph Convolutional Networks. NeurIPS. 2017;

87. Yao H, Guan J, Liu T. Denoising Protein-Protein interaction network via variational graph auto-encoder for protein complex detection. J Bioinform Comput Biol [Internet]. 2020 Jun 1 [cited 2021 Dec 9];18(3). Available from: https://pubmed.ncbi.nlm.nih.gov/32698725/

88. Johansson-Åkhe I, Mirabello C, Wallner B. InterPepRank: Assessment of Docked Peptide Conformations by a Deep Graph Network. Frontiers in Bioinformatics [Internet]. 2021 Oct 25 [cited 2021 Dec 9];0:60. Available from: https://www.frontiersin.org/articles/10.3389/fbinf.2021.763102/full

89. Cao Y, Shen Y. Energy-based Graph Convolutional Networks for Scoring Protein Docking Models. Proteins: Structure, Function and Bioinformatics [Internet]. 2019 Dec 28 [cited 2021 Dec 9];88(8):1091–9. Available from: http://arxiv.org/abs/1912.12476

90. Yang F, Fan K, Song D, Lin H. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. BMC Bioinformatics [Internet]. 2020 Jul 21 [cited 2021 Dec 9];21(1):1–16. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03646-8

91. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics [Internet]. 2018 Jul 1 [cited 2021 Dec 9];34(13):i457–66. Available from: https://academic.oup.com/bioinformatics/article/34/13/i457/5045770

92. Huang YA, Hu P, Chan KCC, You ZH. Graph convolution for predicting associations between miRNA and drug resistance. Bioinformatics [Internet]. 2020 Feb 1 [cited 2021 Dec 9];36(3):851–8. Available from: https://pubmed.ncbi.nlm.nih.gov/31397851/

93. Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics [Internet]. 2020 Dec 30 [cited 2021 Dec 9];36(Supplement_2):i911–8. Available from: https://academic.oup.com/bioinformatics/article/36/Supplement_2/i911/6055929

94. Hwang D, Jeon M, Kang J. A drug-induced liver injury prediction model using transcriptional response data with graph neural network. Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020. 2020 Feb 1;323–9.

95. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43(D1):D447–52.

96. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Systems Biology. 2012;

97. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, Fitzgerald GA, Dolinski K, Grosser T, Troyanskaya OG. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet [Internet]. 2015 May 27 [cited 2021 Dec 9];47(6):569–76. Available from: https://pubmed.ncbi.nlm.nih.gov/25915600/

98. Zitnik M, Leskovec J. Predicting multicellular function through multi-layer tissue networks. Bioinformatics [Internet]. 2017 Jul 15 [cited 2021 Dec 9];33(14):i190–8. Available from: https://academic.oup.com/bioinformatics/article/33/14/i190/3953967

99. Rousseaux MWC, de Haro M, Lasagna-Reeves CA, de Maio A, Park J, Jafar-Nejad P, Al-Ramahi I, Sharma A, See L, Lu N, Vilanova-Velez L, Klisch TJ, Westbrook TF, Troncoso JC, Botas J, Zoghbi HY. TRIM28 regulates the nuclear accumulation and toxicity of both alpha-synuclein and tau. Elife [Internet]. 2016 Oct 25 [cited 2022 Jan 18];5(OCTOBER2016). Available from: https://pubmed.ncbi.nlm.nih.gov/27779468/

100. Lasagna-Reeves CA, de Haro M, Hao S, Park J, Rousseaux MWC, Al-Ramahi I, Jafar-Nejad P, Vilanova-Velez L, See L, de Maio A, Nitschke L, Wu Z, Troncoso JC, Westbrook TF, Tang J, Botas J, Zoghbi HY. Reduction of Nuak1 Decreases Tau and Reverses Phenotypes in a

Tauopathy Mouse Model. Neuron [Internet]. 2016 Oct 19 [cited 2022 Jan 18];92(2):407–18. Available from: https://pubmed.ncbi.nlm.nih.gov/27720485/

101. Chouhan AK, Guo C, Hsieh YC, Ye H, Senturk M, Zuo Z, Li Y, Chatterjee S, Botas J, Jackson GR, Bellen HJ, Shulman JM. Uncoupling neuronal death and dysfunction in Drosophila models of neurodegenerative disease. Acta Neuropathol Commun [Internet]. 2016 Jun 23 [cited 2022 Jan 18];4(1):62. Available from: https://pubmed.ncbi.nlm.nih.gov/27338814/

102. Genin E, Hannequin D, Wallon D, Sleegers K, Hiltunen M, Combarros O, Bullido MJ, Engelborghs S, de Deyn P, Berr C, Pasquier F, Dubois B, Tognoni G, Fiévet N, Brouwers N, Bettens K, Arosio B, Coto E, del Zompo M, Mateo I, Epelbaum J, Frank-Garcia A, Helisalmi S, Porcellini E, Pilotto A, Forti P, Ferri R, Scarpini E, Siciliano G, Solfrizzi V, Sorbi S, Spalletta G, Valdivieso F, Vepsäläinen S, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Boss P, Hanon O, Piccardi P, Annoni G, Seripa D, Galimberti D, Licastro F, Soininen H, Dartigues JF, Kamboh MI, van Broeckhoven C, Lambert JC, Amouyel P, Campion D. APOE and Alzheimer disease: a major gene with semi-dominant inheritance. Mol Psychiatry [Internet]. 2011 Sep [cited 2021 Dec 9];16(9):903–7. Available from: https://pubmed.ncbi.nlm.nih.gov/21556001/

103. Spangenberg E, Severson PL, Hohsfield LA, Crapser J, Zhang J, Burton EA, Zhang Y, Spevak W, Lin J, Phan NY, Habets G, Rymar A, Tsang G, Walters J, Nespi M, Singh P, Broome S, Ibrahim P, Zhang C, Bollag G, West BL, Green KN. Sustained microglial depletion with CSF1R inhibitor impairs parenchymal plaque development in an Alzheimer's disease model. Nature Communications 2019 10:1 [Internet]. 2019 Aug 21 [cited 2021 Dec 9];10(1):1–21. Available from: https://www.nature.com/articles/s41467-019-11674-z

104. Mashkaryan V, Siddiqui T, Popova S, Cosacak MI, Bhattarai P, Brandt K, Govindarajan N, Petzold A, Reinhardt S, Dahl A, Lefort R, Kizil C. Type 1 Interleukin-4 Signaling Obliterates Mouse Astroglia in vivo but Not in vitro. Front Cell Dev Biol [Internet]. 2020 Feb 26 [cited 2021 Dec 9];8. Available from: https://pubmed.ncbi.nlm.nih.gov/32181251/

105. Yang F, Diao X, Wang F, Wang Q, Sun J, Zhou Y, Xie J. Identification of Key Regulatory Genes and Pathways in Prefrontal Cortex of Alzheimer's Disease. Interdiscip Sci [Internet]. 2020 Mar 1 [cited 2021 Dec 9];12(1):90–8. Available from: https://pubmed.ncbi.nlm.nih.gov/32006383/

106. Myers AJ, Kaleem M, Marlowe L, Pittman AM, Lees AJ, Fung HC, Duckworth J, Leung D, Gibson A, Morris CM, de Silva R, Hardy J. The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. Hum Mol Genet [Internet]. 2005 Aug 15 [cited 2021 Dec 9];14(16):2399–404. Available from: https://pubmed.ncbi.nlm.nih.gov/16000317/

107. Lu T, Aron L, Zullo J, Pan Y, Kim H, Chen Y, Yang TH, Kim HM, Drake D, Liu XS, Bennett DA, Colaiácovo MP, Yankner BA. REST and stress resistance in ageing and Alzheimer's disease. Nature [Internet]. 2014 [cited 2021 Dec 9];507(7493):448–54. Available from: https://pubmed.ncbi.nlm.nih.gov/24670762/

108. Ofengeim D, Mazzitelli S, Ito Y, DeWitt JP, Mifflin L, Zou C, Das S, Adiconis X, Chen H, Zhu H, Kelliher MA, Levin JZ, Yuan J. RIPK1 mediates a disease-associated microglial response in

Alzheimer's disease. Proc Natl Acad Sci U S A [Internet]. 2017 Oct 10 [cited 2021 Dec 9];114(41):E8788–97. Available from: https://pubmed.ncbi.nlm.nih.gov/28904096/

109. Yamakawa H, Cheng J, Penney J, Gao F, Rueda R, Wang J, Yamakawa S, Kritskiy O, Gjoneska E, Tsai LH. The Transcription Factor Sp3 Cooperates with HDAC2 to Regulate Synaptic Function and Plasticity in Neurons. Cell Rep [Internet]. 2017 Aug 8 [cited 2021 Dec 9];20(6):1319–34. Available from: https://pubmed.ncbi.nlm.nih.gov/28793257/

110. Saleem S, Biswas SC. Tribbles Pseudokinase 3 Induces Both Apoptosis and Autophagy in Amyloid-β-induced Neuronal Death. J Biol Chem [Internet]. 2017 Feb 17 [cited 2021 Dec 9];292(7):2571–85. Available from: https://pubmed.ncbi.nlm.nih.gov/28011637/

111. Long JM, Holtzman DM. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. Cell. 2019 Oct 3;179(2):312–39.

112. Pavel M, Imarisio S, Menzies FM, Jimenez-Sanchez M, Siddiqi FH, Wu X, Renna M, O'Kane CJ, Crowther DC, Rubinsztein DC. CCT complex restricts neuropathogenic protein aggregation via autophagy. Nature Communications 2016 7:1 [Internet]. 2016 Dec 8 [cited 2021 Dec 9];7(1):1–18. Available from: https://www.nature.com/articles/ncomms13821

113. Wang BJ, Her GM, Hu MK, Chen YW, Tung YT, Wu PY, Hsu WM, Lee H, Jin LW, Hwang SPL, Chen RPY, Huang CJ, Liao YF. ErbB2 regulates autophagic flux to modulate the proteostasis of APP-CTFs in Alzheimer's disease. Proc Natl Acad Sci U S A [Internet]. 2017 Apr 11 [cited 2021 Dec 9];114(15):E3129–38. Available from: https://pubmed.ncbi.nlm.nih.gov/28351972/

114. Miranda AM, Herman M, Cheng R, Nahmani E, Barrett G, Micevska E, Fontaine G, Potier MC, Head E, Schmitt FA, Lott IT, Jiménez-Velázquez IZ, Antonarakis SE, di Paolo G, Lee JH, Hussaini SA, Marquer C. Excess Synaptojanin 1 Contributes to Place Cell Dysfunction and Memory Deficits in the Aging Hippocampus in Three Types of Alzheimer's Disease. Cell Rep [Internet]. 2018 Jun 5 [cited 2021 Dec 9];23(10):2967–75. Available from: https://pubmed.ncbi.nlm.nih.gov/29874583/

115. Uberti D, Lanni C, Racchi M, Govoni S, Memo M. Conformationally altered p53: a putative peripheral marker for Alzheimer's disease. Neurodegener Dis [Internet]. 2008 Mar [cited 2021 Dec 9];5(3–4):209–11. Available from: https://pubmed.ncbi.nlm.nih.gov/18322392/

116. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JSK, Younkin S, Hazrati L, Collinge J, Pocock J, Lashley T, Williams J, Lambert JC, Amouyel P, Goate A, Rademakers R, Morgan K, Powell J, st. George-Hyslop P, Singleton A, Hardy J. TREM2 Variants in Alzheimer's Disease . New England Journal of Medicine [Internet]. 2013 Jan 10 [cited 2021 Dec 14];368(2):117–27. Available from: https://www.nejm.org/doi/full/10.1056/nejmoa1211851

117. Campion D, Charbonnier C, Nicolas G. SORL1 genetic variants and Alzheimer disease risk: a literature review and meta-analysis of sequencing data. Acta Neuropathol [Internet]. 2019 Aug 1 [cited 2021 Dec 14];138(2):173–86. Available from: https://pubmed.ncbi.nlm.nih.gov/30911827/

118. Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, Sundseth SS, Huentelman MJ, Welsh-Bohmer KA, Reiman EM. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J [Internet]. 2010 Oct [cited 2021 Dec 9];10(5):375–84. Available from: https://pubmed.ncbi.nlm.nih.gov/20029386/

119. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res [Internet]. 2020 Jan 1 [cited 2021 Dec 9];48(D1):D845–55. Available from: https://pubmed.ncbi.nlm.nih.gov/31680165/

120. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: Update 2019. Nucleic Acids Research. 2019;47(D1):D948–54.

121. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, Burgess JD, Chai HS, Crook J, Eddy JA, Li H, Logsdon B, Peters MA, Dang KK, Wang X, Serie D, Wang C, Nguyen T, Lincoln S, Malphrus K, Bisceglio G, Li M, Golde TE, Mangravite LM, Asmann Y, Price ND, Petersen RC, Graff-Radford NR, Dickson DW, Younkin SG, Ertekin-Taner N. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016 Oct 11;3:160089.

122. de Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, Klein HU, White CC, Peters MA, Lodgson B, Nejad P, Tang A, Mangravite LM, Yu L, Gaiteri C, Mostafavi S, Schneider JA, Bennett DA. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142.

123. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, Taga M, Klein HU, Patrick E, Komashko V, McCabe C, Smith R, Bradshaw EM, Root DE, Regev A, Yu L, Chibnik LB, Schneider JA, Young-Pearse TL, Bennett DA, de Jager PL. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci. 2018;21(6):811–9.

124. Wan YW, Al-Ouran R, Mangleburg CG, Perumal TM, Lee T v, Allison K, Swarup V, Funk CC, Gaiteri C, Allen M, Wang M, Neuner SM, Kaczorowski CC, Philip VM, Howell GR, Martini-Stoica H, Zheng H, Mei H, Zhong X, Kim JW, Dawson VL, Dawson TM, Pao PC, Tsai LH, Haure-Mirande JV, Ehrlich ME, Chakrabarty P, Levites Y, Wang X, Dammer EB, Srivastava G, Mukherjee S, Sieberts SK, Omberg L, Dang KD, Eddy JA, Snyder P, Chae Y, Amberkar S, Wei W, Hide W, Preuss C, Ergun A, Ebert PJ, Airey DC, Mostafavi S, Yu L, Klein HU, Accelerating Medicines Partnership-Alzheimer's Disease Consortium, Carter GW, Collier DA, Golde TE, Levey AI, Bennett DA, Estrada K, Townsend TM, Zhang B, Schadt E, de Jager PL, Price ND, Ertekin-Taner N, Liu Z, Shulman JM, Mangravite LM, Logsdon BA. Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. Cell Rep. 2020;32(2):107908.

125. Logsdon BA, Perumal TM, Swarup V, Wang M, Funk C, Gaiteri C, Allen M, Wang X, Dammer E, Srivastava G, Mukherjee S, Sieberts SK, Omberg L, Dang KD, Eddy JA, Snyder P, Chae Y,

Amberkar S, Wei W, Hide W, Preuss C, Ergun A, Ebert PJ, Airey DC, Carter GW, Mostafavi S, Yu L, Klein HU, Consortium the AA, Collier DA, Golde T, Levey A, Bennett DA, Estrada K, Decker M, Liu Z, Shulman JM, Zhang B, Schadt E, Jager PL de, Price ND, Ertekin-Taner N, Mangravite LM. Meta-analysis of the human brain transcriptome identifies heterogeneity across human AD coexpression modules robust to sample collection and methodological approach. bioRxiv [Internet]. 2019 Jan 3 [cited 2021 Dec 9];510420. Available from: https://www.biorxiv.org/content/10.1101/510420v1

126. AMP-AD. AD Knowledge Portal. https://www.synapse.org/#!Synapse:syn2580853/. 2016;

127. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2014;

128. Lisewski AM, Lichtarge O. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. Physica A [Internet]. 2010 Aug 15 [cited 2021 Dec 9];389(16):3250–3. Available from: https://pubmed.ncbi.nlm.nih.gov/20625477/

129. Lisewski AM, Quiros JP, Ng CL, Adikesavan AK, Miura K, Putluri N, Eastman RT, Scanfeld D, Regenbogen SJ, Altenhofen L, Llinás M, Sreekumar A, Long C, Fidock DA, Lichtarge O. Supergenomic network compression and the discovery of exp1 as a glutathione transferase inhibited by artesunate. Cell. 2014;

130. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics 2017 18:9 [Internet]. 2017 Jun 12 [cited 2021 Dec 9];18(9):551–62. Available from: https://www.nature.com/articles/nrg.2017.38

131. Pham M, Lichtarge O. Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks. Pac Symp Biocomput. 2020;

132. Revilla S, Ursulet S, Álvarez-López MJ, Castro-Freire M, Perpiñá U, García-Mesa Y, Bortolozzi A, Giménez-Llort L, Kaliman P, Cristòfol R, Sarkis C, Sanfeliu C. Lenti-GDNF gene therapy protects against Alzheimer's disease-like neuropathology in 3xTg-AD mice and MC65 cells. CNS Neurosci Ther [Internet]. 2014 Nov 1 [cited 2021 Dec 9];20(11):961–72. Available from: https://pubmed.ncbi.nlm.nih.gov/25119316/

133. Konishi Y, Yang LB, He P, Lindholm K, Lu B, Li R, Shen Y. Deficiency of GDNF Receptor GFRα1 in Alzheimer's Neurons Results in Neuronal Death. J Neurosci [Internet]. 2014 Sep 24 [cited 2021 Dec 9];34(39):13127–38. Available from: https://pubmed.ncbi.nlm.nih.gov/25253858/

134. de Ferrari and G v., Inestrosa NC. Wnt signaling function in Alzheimer's disease. Brain Research Reviews. 2000 Aug 1;33(1):1–12.

135. Wang BJ, Her GM, Hu MK, Chen YW, Tung YT, Wu PY, Hsu WM, Lee H, Jin LW, Hwang SPL, Chen RPY, Huang CJ, Liao YF. ErbB2 regulates autophagic flux to modulate the proteostasis of APP-CTFs in Alzheimer's disease. Proc Natl Acad Sci U S A [Internet]. 2017 Apr 11 [cited 2021 Dec 13];114(15):E3129–38. Available from: https://www.pnas.org/content/114/15/E3129

136. Chaudhury AR, Gerecke KM, Wyss JM, Morgan DG, Gordon MN, Carroll SL. Neuregulin-1 and erbB4 immunoreactivity is associated with neuritic plaques in Alzheimer disease brain and in a transgenic model of Alzheimer disease. J Neuropathol Exp Neurol [Internet]. 2003 Jan 1 [cited 2021 Dec 13];62(1):42–54. Available from: https://pubmed.ncbi.nlm.nih.gov/12528817/

137. Hudák A, Kusz E, Domonkos I, Jósvay K, Kodamullil AT, Szilák L, Hofmann-Apitius M, Letoha T. Contribution of syndecans to cellular uptake and fibrillation of α-synuclein and tau. Scientific Reports 2019 9:1 [Internet]. 2019 Nov 12 [cited 2021 Dec 13];9(1):1–19. Available from: https://www.nature.com/articles/s41598-019-53038-z

138. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Anagnostopoulos A, Asabor R, Baldarelli RM, Beal JS, Bello SM, Blodgett O, Butler NE, Christie KR, Corbani LE, Creelman J, Dolan ME, Drabkin HJ, Giannatto SL, Hale P, Hill DP, Law M, Mendoza A, McAndrews M, Miers D, Motenko H, Ni L, Onda H, Perry M, Recla JM, Richards-Smith B, Sitnikov D, Tomczuk M, Tonorio G, Wilming L, Zhu Y. Mouse Genome Database (MGD) 2019. Nucleic Acids Research [Internet]. 2019 Jan 8 [cited 2021 Dec 9];47(Database issue):D801. Available from: /pmc/articles/PMC6323923/

139. Huichalaf CH, Al-Ramahi I, Park KW, Grunke SD, Lu N, de Haro M, El-Zein K, Gallego-Flores T, Perez AM, Jung SY, Botas J, Zoghbi HY, Jankowsky JL. Cross-species genetic screens to identify kinase targets for APP reduction in Alzheimer's disease. Hum Mol Genet [Internet]. 2019 Jun 15 [cited 2022 Jan 18];28(12):2014–29. Available from: https://pubmed.ncbi.nlm.nih.gov/30753434/

140. Lee W, Lavery L, Rousseaux MWC, Rutledge EB, Jang Y, Wan Y, Wu S, Kim W, Al-Ramahi I, Rath S, Adamski CJ, Bondar V v, Tewari A, Soleimani S, Mota S, Yalamanchili HK, Orr HT, Liu Z, Botas J, Zoghbi HY. Dual targeting of brain region-specific kinases potentiates neurological rescue in Spinocerebellar ataxia type 1. EMBO J [Internet]. 2021 Apr [cited 2022 Jan 18];40(7). Available from: https://pubmed.ncbi.nlm.nih.gov/33709453/

141. Huichalaf CH, Al-Ramahi I, Park KW, Grunke SD, Lu N, de Haro M, El-Zein K, Gallego-Flores T, Perez AM, Jung SY, Botas J, Zoghbi HY, Jankowsky JL. Cross-species genetic screens to identify kinase targets for APP reduction in Alzheimer's disease. Hum Mol Genet [Internet]. 2019 Jun 15 [cited 2022 Jan 18];28(12):2014–29. Available from: https://pubmed.ncbi.nlm.nih.gov/30753434/

142. Rousseaux MWC, Vázquez-Vélez GE, Al-Ramahi I, Jeong HH, Bajić A, Revelli JP, Ye H, Phan ET, Deger JM, Perez AM, Kim JY, Lavery LA, Xu Q, Li MZ, Kang H, Kim JJ, Shulman JM, Westbrook TF, Elledge SJ, Liu Z, Botas J, Zoghbi HY. A Druggable Genome Screen Identifies Modifiers of α-Synuclein Levels via a Tiered Cross-Species Validation Approach. J Neurosci [Internet]. 2018 Oct 24 [cited 2022 Jan 18];38(43):9286–301. Available from: https://pubmed.ncbi.nlm.nih.gov/30249792/

143. Al-Ramahi I, Lu B, di Paola S, Pang K, de Haro M, Peluso I, Gallego-Flores T, Malik NT, Erikson K, Bleiberg BA, Avalos M, Fan G, Rivers LE, Laitman AM, Diaz-García JR, Hild M, Palacino J, Liu Z, Medina DL, Botas J. High-Throughput Functional Analysis Distinguishes Pathogenic,

Nonpathogenic, and Compensatory Transcriptional Changes in Neurodegeneration. Cell Syst [Internet]. 2018 Jul 25 [cited 2022 Jan 18];7(1):28-40.e4. Available from: https://pubmed.ncbi.nlm.nih.gov/29936182/

144. Lu XH, Mattis VB, Wang N, Al-Ramahi I, van den Berg N, Fratantoni SA, Waldvogel H, Greiner E, Osmand A, Elzein K, Xiao J, Dijkstra S, de Pril R, Vinters H v., Faull R, Signer E, Kwak S, Marugan JJ, Botas J, Fischer DF, Svendsen CN, Munoz-Sanjuan I, Yang XW. Targeting ATM ameliorates mutant Huntingtin toxicity in cell and animal models of Huntington's disease. Sci Transl Med [Internet]. 2014 Dec 24 [cited 2022 Jan 18];6(268):268ra178. Available from: https://pubmed.ncbi.nlm.nih.gov/25540325/

145. Park J, Al-Ramahi I, Tan Q, Mollema N, Diaz-Garcia JR, Gallego-Flores T, Lu HC, Lagalwar S, Duvick L, Kang H, Lee Y, Jafar-Nejad P, Sayegh LS, Richman R, Liu X, Gao Y, Shaw CA, Arthur JSC, Orr HT, Westbrook TF, Botas J, Zoghbi HY. RAS-MAPK-MSK1 pathway modulates ataxin 1 protein levels and toxicity in SCA1. Nature [Internet]. 2013 [cited 2022 Jan 18];498(7454):325–31. Available from: https://pubmed.ncbi.nlm.nih.gov/23719381/

146. Han KM, Kang RJ, Jeon H, Lee HJ, Lee JS, Park HH, Gak Jeon S, Suk K, Seo J, Hoe HS. Regorafenib Regulates AD Pathology, Neuroinflammation, and Dendritic Spinogenesis in Cells and a Mouse Model of AD. Cells [Internet]. 2020 Jul 9 [cited 2021 Dec 9];9(7). Available from: /pmc/articles/PMC7408082/

147. Zhang P, Kishimoto Y, Grammatikakis I, Gottimukkala K, Cutler RG, Zhang S, Abdelmohsen K, Bohr VA, Sen JM, Gorospe M, Mattson MP. Senolytic therapy alleviates Aβ-associated oligodendrocyte progenitor cell senescence and cognitive deficits in an Alzheimer's disease model. Nature Neuroscience [Internet]. [cited 2021 Dec 9]; Available from: https://doi.org/10.1038/s41593-019-0372-9

148. Angelopoulou E, Piperi C. Beneficial Effects of Fingolimod in Alzheimer's Disease: Molecular Mechanisms and Therapeutic Potential. NeuroMolecular Medicine 2019 21:3 [Internet]. 2019 Jul 16 [cited 2021 Dec 9];21(3):227–38. Available from: https://link.springer.com/article/10.1007/s12017-019-08558-2

149. Grammas P, Martinez J, Sanchez A, Yin X, Riley J, Gay D, Desobry K, Tripathy D, Luo J, Evola M, Young A. A new paradigm for the treatment of Alzheimer's disease: targeting vascular activation. J Alzheimers Dis [Internet]. 2014 [cited 2021 Dec 9];40(3):619–30. Available from: https://pubmed.ncbi.nlm.nih.gov/24503617/

150. la Barbera L, Vedele F, Nobili A, Krashia P, Spoleti E, Latagliata EC, Cutuli D, Cauzzi E, Marino R, Viscomi MT, Petrosini L, Puglisi-Allegra S, Melone M, Keller F, Mercuri NB, Conti F, D'Amelio M. Nilotinib restores memory function by preventing dopaminergic neuron degeneration in a mouse model of Alzheimer's Disease. Prog Neurobiol [Internet]. 2021 Jul 1 [cited 2021 Dec 9];202. Available from: https://pubmed.ncbi.nlm.nih.gov/33684513/

151. Turner RS, Hebron ML, Lawler A, Mundel EE, Yusuf N, Starr JN, Anjum M, Pagan F, Torres-Yaghi Y, Shi W, Mulki S, Ferrante D, Matar S, Liu X, Esposito G, Berkowitz F, Jiang X, Ahn J, Moussa C. Nilotinib Effects on Safety, Tolerability, and Biomarkers in Alzheimer's Disease.

Ann Neurol [Internet]. 2020 Jul 1 [cited 2021 Dec 9];88(1):183–94. Available from: https://pubmed.ncbi.nlm.nih.gov/32468646/

152. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, Choi D, Coté AG, Daley M, Deimling S, Desbuleux A, Dricot A, Gebbia M, Hardy MF, Kishore N, Knapp JJ, Kovács IA, Lemmens I, Mee MW, Mellor JC, Pollis C, Pons C, Richardson AD, Schlabach S, Teeking B, Yadav A, Babor M, Balcha D, Basha O, Bowman-Colin C, Chin SF, Choi SG, Colabella C, Coppin G, D'Amata C, de Ridder D, de Rouck S, Duran-Frigola M, Ennajdaoui H, Goebels F, Goehring L, Gopal A, Haddad G, Hatchi E, Helmy M, Jacob Y, Kassa Y, Landini S, Li R, van Lieshout N, MacWilliams A, Markey D, Paulson JN, Rangarajan S, Rasla J, Rayhan A, Rolland T, San-Miguel A, Shen Y, Sheykhkarimli D, Sheynkman GM, Simonovsky E, Taşan M, Tejeda A, Tropepe V, Twizere JC, Wang Y, Weatheritt RJ, Weile J, Xia Y, Yang X, Yeger-Lotem E, Zhong Q, Aloy P, Bader GD, de Las Rivas J, Gaudet S, Hao T, Rak J, Tavernier J, Hill DE, Vidal M, Roth FP, Calderwood MA. A reference map of the human binary protein interactome. Nature. 2020 Apr 16;580(7803):402–8.

153. Zuena AR, Casolini P, Lattanzi R, Maftei D. Chemokines in Alzheimer's disease: New insights into prokineticins, chemokine-like proteins. Frontiers in Pharmacology. 2019;10(MAY):622.

154. Kern DM, Cepeda MS, Lovestone S, Seabrook GR. Aiding the discovery of new treatments for dementia by uncovering unknown benefits of existing medications. Alzheimer's & Dementia : Translational Research & Clinical Interventions [Internet]. 2019 Jan 1 [cited 2021 Dec 9];5:862. Available from: /pmc/articles/PMC6909196/

155. Chumakov I, Nabirotchkin S, Cholet N, Milet A, Boucard A, Toulorge D, Pereira Y, Graudens E, Traore S, Foucquier J, Guedj M, Vial E, Callizot N, Steinschneider R, Maurice T, Bertrand V, Scart-Gres C, Hajj R, Cohen D. Combining two repurposed drugs as a promising approach for Alzheimer's disease therapy. Scientific Reports 2015 5:1 [Internet]. 2015 Jan 8 [cited 2021 Dec 9];5(1):1–12. Available from: https://www.nature.com/articles/srep07608

156. Valencia RG, Mihailovska E, Winter L, Bauer K, Fischer I, Walko G, Jorgacevski J, Potokar M, Zorec R, Wiche G. Plectin dysfunction in neurons leads to tau accumulation on microtubules affecting neuritogenesis, organelle trafficking, pain sensitivity and memory. Neuropathol Appl Neurobiol [Internet]. 2021 Feb 1 [cited 2021 Dec 9];47(1):73–95. Available from: https://pubmed.ncbi.nlm.nih.gov/32484610/

157. Knuesel I, Riban V, Zuellig RA, Schaub MC, Grady RM, Sanes JR, Fritschy JM. Increased vulnerability to kainate-induced seizures in utrophin-knockout mice. Eur J Neurosci. 2002 May;15(9):1474–84.

158. B Johnson EC, Dammer EB, Duong DM, Ping L, Zhou M, Yin L, Higginbotham LA, Guajardo A, White B, Troncoso JC, Thambisetty M, Montine TJ, Lee EB, Trojanowski JQ, Beach TG, Reiman EM, Haroutunian V, Wang M, Schadt E, Zhang B, Dickson DW, Ertekin-Taner N, Golde TE, Petyuk VA, Jager PL, Bennett DA, Wingo TS, Rangaraju S, Hajjar I, Shulman JM, Lah JJ, Levey AI, Seyfried NT. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. [cited 2021 Dec 14]; Available from: https://doi.org/10.1038/s41591-020-0815-6

159. Drummond E, Nayak S, Faustin A, Pires G, Hickman RA, Askenazi M, Cohen M, Haldiman T, Kim C, Han X, Shao Y, Safar JG, Ueberheide B, Wisniewski T. Proteomic differences in amyloid plaques in rapidly progressive and sporadic Alzheimer's disease. Acta Neuropathol [Internet]. 2017 Jun 1 [cited 2021 Dec 14];133(6):933–54. Available from: https://pubmed.ncbi.nlm.nih.gov/28258398/

160. Ginsberg SD, Hemby SE. Expression profile of transcripts in Alzheimer's disease tangle-bearing CA1 neurons Neurobiology of schizophrenia and antipsychotic effects View project Neurobiology of drug abuse View project. Article in Annals of Neurology [Internet]. 2000 [cited 2021 Dec 9]; Available from: https://www.researchgate.net/publication/319529436

161. Merlo P, Frost B, Peng S, Yang YJ, Park PJ, Feany M. P53 prevents neurodegeneration by regulating synaptic genes. Proc Natl Acad Sci U S A [Internet]. 2014 Dec 16 [cited 2021 Dec 14];111(50):18055–60. Available from: https://pubmed.ncbi.nlm.nih.gov/25453105/

162. Gao S, Zhang X, Song Q, Liu J, Ji X, Wang P. POLD1 deficiency is involved in cognitive function impairment in AD patients and SAMP8 mice. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie [Internet]. 2019 Jun 1 [cited 2021 Dec 14];114. Available from: https://pubmed.ncbi.nlm.nih.gov/30978525/

163. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet [Internet]. 2012 Aug 10 [cited 2021 Dec 9];91(2):224–37. Available from: https://pubmed.ncbi.nlm.nih.gov/22863193/

164. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nature Reviews Genetics 2019 20:12 [Internet]. 2019 Oct 11 [cited 2021 Dec 13];20(12):747–59. Available from: https://www.nature.com/articles/s41576-019-0177-4

165. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. American Journal of Human Genetics [Internet]. 2014 Jul 3 [cited 2021 Dec 13];95(1):5. Available from: /pmc/articles/PMC4085641/

166. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, Bi W, Gabrielsen ME, Daly MJ, Neale BM, Hveem K, Abecasis GR, Willer CJ, Lee S. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nature Genetics 2020 52:6 [Internet]. 2020 May 18 [cited 2021 Dec 13];52(6):634–9. Available from: https://www.nature.com/articles/s41588-020-0621-6

167. Chhetri J, King AE, Gueven N. Alzheimer's Disease and NQO1: Is there a Link? Curr Alzheimer Res [Internet]. 2018 Feb 23 [cited 2021 Dec 14];15(1):56–66. Available from: https://pubmed.ncbi.nlm.nih.gov/28164770/

168. SantaCruz KS, Yazlovitskaya E, Collins J, Johnson J, DeCarli C. Regional NAD(P)H:quinone oxidoreductase activity in Alzheimer's disease. Neurobiol Aging [Internet]. 2004 [cited 2021 Dec 14];25(1):63–9. Available from: https://pubmed.ncbi.nlm.nih.gov/14675732/

169. Wang Y, Santa-Cruz K, Decarli C, Johnson JA. NAD(P)H:quinone oxidoreductase activity is increased in hippocampal pyramidal neurons of patients with Aalzheimer's disease. Neurobiol Aging [Internet]. 2000 Jul [cited 2021 Dec 14];21(4):525–31. Available from: https://pubmed.ncbi.nlm.nih.gov/10924765/

170. Raina AK, Templeton DJ, Deak JC, Perry G, Smith MA. Quinone reductase (NQO1), a sensitive redox indicator, is increased in Alzheimer's disease. Redox Rep [Internet]. 1999 [cited 2021 Dec 14];4(1–2):23–7. Available from: https://pubmed.ncbi.nlm.nih.gov/10714272/

171. Katsonis P, Lichtarge O. CAGI5: Objective performance assessments of predictions based on the Evolutionary Action equation. Hum Mutat [Internet]. 2019 Sep 1 [cited 2022 May 3];40(9):1436–54. Available from: https://pubmed.ncbi.nlm.nih.gov/31317604/

172. Kim YW, Al-Ramahi I, Koire A, Wilson SJ, Konecki DM, Mota S, Soleimani S, Botas J, Lichtarge O. Harnessing the paradoxical phenotypes of APOE ε2 and APOE ε4 to identify genetic modifiers in Alzheimer's disease. Alzheimer's and Dementia. 2021;

173. Koire A, Katsonis P, Kim YW, Buchovecky C, Wilson SJ, Lichtarge O. A method to delineate de novo missense variants across pathways prioritizes genes linked to autism. Science Translational Medicine [Internet]. 2021;13(594):eabc1739. Available from: https://stm.sciencemag.org/lookup/doi/10.1126/scitranslmed.abc1739

174. Sambamoorthy G, Raman K. Understanding the evolution of functional redundancy in metabolic networks. Bioinformatics [Internet]. 2018 Sep 1 [cited 2021 Dec 9];34(17):i981–7. Available from: https://academic.oup.com/bioinformatics/article/34/17/i981/5093206

175. John L. Hartman I, Garvik B, Hartwell L. Principles for the Buffering of Genetic Variation. Science (1979) [Internet]. 2001 Feb 9 [cited 2021 Dec 9];291(5506):1001–4. Available from: https://www.science.org/doi/abs/10.1126/science.1056072

176. Güell O, Sagués F, Serrano MÁ. Essential Plasticity and Redundancy of Metabolism Unveiled by Synthetic Lethality Analysis. PLOS Computational Biology [Internet]. 2014 [cited 2021 Dec 9];10(5):e1003637. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003637

177. Butkiewicz M, Blue EE, Leung YY, Jian X, Marcora E, Renton AE, Kuzma A, Wang LS, Koboldt DC, Haines JL, Bush WS. Functional annotation of genomic variants in studies of late-onset Alzheimer's disease. Bioinformatics. 2018 Aug 15;34(16):2724–31.

178. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics. 2005;76(5):887–93.

179. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2).

180. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics [Internet]. 2010 Nov 15 [cited 2022 May 4];26(22):2867–73. Available from:

https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq559

181. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genetics. 2006 Dec;2(12):2074–93.

182. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research [Internet]. 2019 Jan 8 [cited 2021 Dec 9];47(D1):D607–13. Available from: https://academic.oup.com/nar/article/47/D1/D607/5198476

183. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res [Internet]. 2003 Nov [cited 2021 Dec 9];13(11):2498–504. Available from: https://pubmed.ncbi.nlm.nih.gov/14597658/

184. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, Griffith M, Griffith OL, Wagner AH. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res [Internet]. 2021 Jan 8 [cited 2021 Dec 13];49(D1):D1144–51. Available from: https://pubmed.ncbi.nlm.nih.gov/33237278/

185. Lasagna-Reeves CA, de Haro M, Hao S, Park J, Rousseaux MWC, Al-Ramahi I, Jafar-Nejad P, Vilanova-Velez L, See L, de Maio A, Nitschke L, Wu Z, Troncoso JC, Westbrook TF, Tang J, Botas J, Zoghbi HY. Reduction of Nuak1 Decreases Tau and Reverses Phenotypes in a Tauopathy Mouse Model. Neuron. 2016 Oct 19;92(2):407–18.

186. Chouhan AK, Guo C, Hsieh YC, Ye H, Senturk M, Zuo Z, Li Y, Chatterjee S, Botas J, Jackson GR, Bellen HJ, Shulman JM. Uncoupling neuronal death and dysfunction in Drosophila models of neurodegenerative disease. Acta Neuropathol Commun. 2016;4(1):62.

187. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011 Aug 31;12:357.

188. Al-Ramahi I, Giridharan SSP, Chen YC, Patnaik S, Safren N, Hasegawa J, de Haro M, Gee AKW, Titus SA, Jeong H, Clarke J, Krainc D, Zheng W, Irvine RF, Barmada S, Ferrer M, Southall N, Weisman LS, Botas J, Marugan JJ. Inhibition of PIP4Kγ ameliorates the pathological effects of mutant huntingtin protein. Elife [Internet]. 2017 Dec 26 [cited 2022 May 3];6. Available from: https://pubmed.ncbi.nlm.nih.gov/29256861/

189. Onur TS, Laitman A, Zhao H, Keyho R, Kim H, Wang J, Mair M, Wang H, Li L, Perez A, de Haro M, Wan YW, Allen G, Lu B, Al-Ramahi I, Liu Z, Botas J. Downregulation of glial genes involved in synaptic function mitigates Huntington's disease pathogenesis. Elife. 2021;10.

190. Manrubia S, Cuesta JA, Aguirre J, Ahnert SE, Altenberg L, Cano A v., Catalán P, Diaz-Uriarte R, Elena SF, García-Martín JA, Hogeweg P, Khatri BS, Krug J, Louis AA, Martin NS, Payne JL, Tarnowski MJ, Weiß M. From genotypes to organisms: State-of-the-art and perspectives of a

cornerstone in evolutionary dynamics. Phys Life Rev [Internet]. 2021 Sep 1 [cited 2022 Jun 19];38:55–106. Available from: https://pubmed.ncbi.nlm.nih.gov/34088608/

191. Johnston IG, Dingle K, Greenbury SF, Camargo CQ, Doye JPK, Ahnert SE, Louis AA. Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. Proc Natl Acad Sci U S A [Internet]. 2022 Mar 15 [cited 2022 Jun 19];119(11). Available from: https://pubmed.ncbi.nlm.nih.gov/35275794/

192. Iwasa Y. Free fitness that always increases in evolution. J Theor Biol [Internet]. 1988 Dec 7 [cited 2022 Jun 19];135(3):265–81. Available from: https://pubmed.ncbi.nlm.nih.gov/3256719/

193. Sella G, Hirsh AE. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A [Internet]. 2005 Jul 5 [cited 2022 Jun 18];102(27):9541–6. Available from: https://pubmed.ncbi.nlm.nih.gov/15980155/

194. de Vladar HP, Barton NH. The contribution of statistical physics to evolutionary biology. Trends Ecol Evol [Internet]. 2011 Aug [cited 2022 Jun 18];26(8):424–32. Available from: https://pubmed.ncbi.nlm.nih.gov/21571390/

195. Barton NH, de Vladar HP. Statistical mechanics and the evolution of polygenic quantitative traits. Genetics [Internet]. 2009 Mar [cited 2022 Jun 18];181(3):997–1011. Available from: https://pubmed.ncbi.nlm.nih.gov/19087953/

196. Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, Yadav A, Banerjee N, Gillies CE, Damask A, Liu S, Bai X, Hawes A, Maxwell E, Gurski L, Watanabe K, Kosmicki JA, Rajagopal V, Mighty J, Jones M, Mitnaul L, Stahl E, Coppola G, Jorgenson E, Habegger L, Salerno WJ, Shuldiner AR, Lotta LA, Overton JD, Cantor MN, Reid JG, Yancopoulos G, Kang HM, Marchini J, Baras A, Abecasis GR, Ferreira MAR. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature [Internet]. 2021 Nov 25 [cited 2022 Jun 18];599(7886):628–34. Available from: https://pubmed.ncbi.nlm.nih.gov/34662886/

197. Curtis D. Analysis of 200 000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. J Med Genet [Internet]. 2022 Jun 1 [cited 2022 Jun 18];59(6):597–604. Available from: https://pubmed.ncbi.nlm.nih.gov/33910933/

198. Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, Wong E, Liu D, Wade Davis J, Haefliger C, Katrina Loomis A, Mikkilineni R, Noh HJ, Wadhawan S, Bai X, Hawes A, Krasheninina O, Ulloa R, Lopez AE, Smith EN, Waring JF, Whelan CD, Tsai EA, Overton JD, Salerno WJ, Jacob H, Szalma S, Runz H, Hinkle G, Nioi P, Petrovski S, Miller MR, Baras A, Mitnaul LJ, Reid JG, Moiseyenko O, Rios C, Saha S, Abecasis G, Banerjee N, Beechert C, Boutkov B, Cantor M, Coppola G, Economides A, Eom G, Forsythe C, Fuller ED, Gu Z, Habegger L, Jones MB, Lanche R, Lattari M, LeBlanc M, Li D, Lotta LA, Manoochehri K, Mansfield AJ, Maxwell EK, Mighty J, Nafde M, O'Keeffe S, Orelus M, Padilla MS, Panea R, Polanco T, Pradhan M, Rasool A, Schleicher TD, Sharma D, Shuldiner A, Staples JC, van Hout C v., Widom L, Wolf SE, John S, Chen CY, Sexton D, Kupelian V, Marshall E, Swan T, Eaton S, Liu

JZ, Loomis S, Jensen M, Duraisamy S, Tetrault J, Merberg D, Badola S, Reppell M, Grundstad J, Zheng X, Deaton AM, Parker MM, Ward LD, Flynn-Carroll AO, Austin C, March R, Pangalos MN, Platt A, Snowden M, Matakidou A, Wasilewski S, Wang Q, Deevi S, Carss K, Smith K, Sogaard M, Hu X, Chen X, Ye Z. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet [Internet]. 2021 Jul 1 [cited 2022 Jun 18];53(7):942–8. Available from: https://pubmed.ncbi.nlm.nih.gov/34183854/

199. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LDW, Porubsky D, Mao Y, Gordon D, Vollger MR, Lewis AP, Munson KM, Devogelaere E, Armstrong J, Diekhans M, Walker JA, Tomlinson C, Graves-Lindsay TA, Kremitzki M, Salama SR, Audano PA, Escalona9 M, Maurer NW, Antonacci F, Mercuri L, Maggiolini FAM, Catacchio CR, Underwood JG, O'Connor DH, Sanders AD, Korbel JO, Ferguson B, Kubisch HM, Picker L, Kalin NH, Rosene D, Levine J, Abbott DH, Gray25 SB, Sanchez MM, Kovacs-Balint ZA, Kemnitz JW, Thomasy SM, Roberts JA, Kinnally EL, Capitanio JP, Skene JHP, Platt M, Cole SA, Green RE, Ventura M, Wiseman RW, Paten B, Batzer MA, Rogers J, Eichler EE. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science [Internet]. 2020 Dec 18 [cited 2022 Jun 18];370(6523). Available from: https://pubmed.ncbi.nlm.nih.gov/33335035/

200. Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput Biol [Internet]. 2018 Dec 1 [cited 2022 Jun 18];14(12). Available from: https://pubmed.ncbi.nlm.nih.gov/30550564/

201. Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. Proc Natl Acad Sci U S A [Internet]. 2003 Feb 2 [cited 2022 Jun 18];100(3):776. Available from: /pmc/articles/PMC298677/

202. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes & Development [Internet]. 2018 Oct 10 [cited 2022 Jun 18];32(19–20):1267. Available from: /pmc/articles/PMC6169832/

203. Hsu TK, Asmussen J, Koire A, Choi BK, Gadhikar MA, Huh E, Lin CH, Konecki DM, Kim YW, Pickering CR, Kimmel M, Donehower LA, Frederick MJ, Myers JN, Katsonis P, Lichtarge O. A general calculus of fitness landscapes finds genes under selection in cancers. Genome Res [Internet]. 2022 May 1 [cited 2022 Jun 18];32(5):916–29. Available from: https://pubmed.ncbi.nlm.nih.gov/35301263/

204. McLendon R, Friedman A, Bigner D, van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K, Yung WKA, Bogler O, Weinstein JN, VandenBerg S, Berger M, Prados M, Muzny D, Morgan M, Scherer S, Sabo A, Nazareth L, Lewis L, Hall O, Zhu Y, Ren Y, Alvi O, Yao J, Hawes A, Jhangiani S, Fowler G, San Lucas A, Kovar C, Cree A, Dinh H, Santibanez J, Joshi V, Gonzalez-Garay ML, Miller CA, Milosavljevic A, Donehower L, Wheeler DA, Gibbs RA, Cibulskis K, Sougnez C, Fennell T, Mahan S, Wilkinson J, Ziaugra L, Onofrio R, Bloom T, Nicol R, Ardlie K, Baldwin J, Gabriel S, Lander ES, Ding L, Fulton RS, McLellan MD, Wallis J, Larson DE, Shi X, Abbott R, Fulton L, Chen K, Koboldt DC, Wendl MC, Meyer R, Tang Y, Lin L, Osborne JR, Dunford-Shore BH, Miner TL, Delehaunty K, Markovic C, Swift G,

Courtney W, Pohl C, Abbott S, Hawkins A, Leong S, Haipek C, Schmidt H, Wiechert M, Vickery T, Scott S, Dooling DJ, Chinwalla A, Weinstock GM, Mardis ER, Wilson RK, Getz G, Winckler W, Verhaak RGW, Lawrence MS, O'Kelly M, Robinson J, Alexe G, Beroukhim R, Carter S, Chiang D, Gould J, Gupta S, Korn J, Mermel C, Mesirov J, Monti S, Nguyen H, Parkin M, Reich M, Stransky N, Weir BA, Garraway L, Golub T, Meyerson M, Chin L, Protopopov A, Zhang J, Perna I, Aronson S, Sathiamoorthy N, Ren G, Yao J, Wiedemeyer WR, Kim H, Sek WK, Xiao Y, Kohane IS, Seidman J, Park PJ, Kucherlapati R, Laird PW, Cope L, Herman JG, Weisenberger DJ, Pan F, van den Berg D, van Neste L, Joo MY, Schuebel KE, Baylin SB, Absher DM, Li JZ, Southwick A, Brady S, Aggarwal A, Chung T, Sherlock G, Brooks JD, Myers RM, Spellman PT, Purdom E, Jakkula LR, Lapuk A v., Marr H, Dorton S, Yoon GC, Han J, Ray A, Wang V, Durinck S, Robinson M, Wang NJ, Vranizan K, Peng V, van Name E, Fontenay G v., Ngai J, Conboy JG, Parvin B, Feiler HS, Speed TP, Gray JW, Brennan C, Socci ND, Olshen A, Taylor BS, Lash A, Schultz N, Reva B, Antipin Y, Stukalov A, Gross B, Cerami E, Wei QW, Qin LX, Seshan VE, Villafania L, Cavatore M, Borsu L, Viale A, Gerald W, Sander C, Ladanyi M, Perou CM, Hayes DN, Topal MD, Hoadley KA, Qi Y, Balu S, Shi Y, Wu J, Penny R, Bittner M, Shelton T, Lenkiewicz E, Morris S, Beasley D, Sanders S, Kahn A, Sfeir R, Chen J, Nassau D, Feng L, Hickey E, Barker A, Gerhard DS, Vockley J, Compton C, Vaught J, Fielding P, Ferguson ML, Schaefer C, Zhang J, Madhavan S, Buetow KH, Collins F, Good P, Guyer M, Ozenberger B, Peterson J, Thomson E. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature [Internet]. 2008 Oct 23 [cited 2022 Jun 18];455(7216):1061–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18772890/

205. Lawrence MS, Sougnez C, Lichtenstein L, Cibulskis K, Lander E, Gabriel SB, Getz G, Ally A, Balasundaram M, Birol I, Bowlby R, Brooks D, Butterfield YSN, Carlsen R, Cheng D, Chu A, Dhalla N, Guin R, Holt RA, Jones SJM, Lee D, Li HI, Marra MA, Mayo M, Moore RA, Mungall AJ, Robertson AG, Schein JE, Sipahimalani P, Tam A, Thiessen N, Wong T, Protopopov A, Santoso N, Lee S, Parfenov M, Zhang J, Mahadeshwar HS, Tang J, Ren X, Seth S, Haseley P, Zeng D, Yang L, Xu AW, Song X, Pantazi A, Bristow CA, Hadjipanayis A, Seidman J, Chin L, Park PJ, Kucherlapati R, Akbani R, Casasent T, Liu W, Lu Y, Mills G, Motter T, Weinstein J, Diao L, Wang J, Hong Fan Y, Liu J, Wang K, Auman JT, Balu S, Bodenheimer T, Buda E, Hayes DN, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Kimes PK, Liu Y, Marron JS, Meng S, Mieczkowski PA, Mose LE, Parker JS, Perou CM, Prins JF, Roach J, Shi Y, Simons J v., Singh D, Soloway MG, Tan D, Veluvolu U, Walter V, Waring S, Wilkerson MD, Wu J, Zhao N, Cherniack AD, Hammerman PS, Tward AD, Pedamallu CS, Saksena G, Jung J, Ojesina AI, Carter SL, Zack TI, Schumacher SE, Beroukhim R, Freeman SS, Meyerson M, Cho J, Noble MS, DiCara D, Zhang H, Heiman DI, Gehlenborg N, Voet D, Lin P, Frazer S, Stojanov P, Liu Y, Zou L, Kim J, Muzny D, Doddapaneni HV, Kovar C, Reid J, Morton D, Han Y, Hale W, Chao H, Chang K, Drummond JA, Gibbs RA, Kakkar N, Wheeler D, Xi L, Ciriello G, Ladanyi M, Lee W, Ramirez R, Sander C, Shen R, Sinha R, Weinhold N, Taylor BS, Aksoy BA, Dresdner G, Gao J, Gross B, Jacobsen A, Reva B, Schultz N, Sumer SO, Sun Y, Chan TA, Morris LG, Stuart J, Benz S, Ng S, Benz C, Yau C, Baylin SB, Cope L, Danilova L, Herman JG, Bootwalla M, Maglinte DT, Laird PW, Triche T, Weisenberger DJ, van den Berg DJ, Agrawal N, Bishop J, Boutros PC, Bruce JP, Byers LA, Califano J, Carey TE, Chen Z, Cheng H, Chiosea SI, Cohen E, Diergaarde B, Egloff AM, El-Naggar AK, Ferris RL, Frederick MJ, Grandis JR, Guo Y, Haddad RI, Harris T, Hui ABY, Lee JJ, Lippman SM, Liu FF, McHugh JB, Myers J, Ng PKS, Perez-Ordonez B, Pickering CR, Prystowsky M, Romkes M, Saleh AD, Sartor

MA, Seethala R, Seiwert TY, Si H, van Waes C, Waggott DM, Wiznerowicz M, Yarbrough WG, Zhang J, Zuo Z, Burnett K, Crain D, Gardner J, Lau K, Mallery D, Morris S, Paulauskis J, Penny R, Shelton C, Shelton T, Sherman M, Yena P, Black AD, Bowen J, Frick J, Gastier-Foster JM, Harper HA, Leraas K, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Baboud J, Jensen MA, Kahn AB, Pihl TD, Pot DA, Srinivasan D, Walton JS, Wan Y, Burton RA, Davidsen T, Demchok JA, Eley G, Ferguson ML, Mills Shaw KR, Ozenberger BA, Sheth M, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Saller C, Tarvin K, Chen C, Bollag R, Weinberger P, Golusiński W, Golusiński P, Ibbs M, Korski K, Mackiewicz A, Suchorska W, Szybiak B, Curley E, Beard C, Mitchell C, Sandusky G, Ahn J, Khan Z, Irish J, Waldron J, William WN, Egea S, Gomez-Fernandez C, Herbert L, Bradford CR, Chepeha DB, Haddad AS, Jones TR, Komarck CM, Malakh M, Moyer JS, Nguyen A, Peterson LA, Prince ME, Rozek LS, Taylor EG, Walline HM, Wolf GT, Boice L, Chera BS, Funkhouser WK, Gulley ML, Hackman TG, Hayward MC, Huang M, Rathmell WK, Salazar AH, Shockley WW, Shores CG, Thorne L, Weissler MC, Wrenn S, Zanation AM, Brown BT, Pham M. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature [Internet]. 2015 Jan 29 [cited 2022 Jun 18];517(7536):576–82. Available from: https://pubmed.ncbi.nlm.nih.gov/25631445/

206. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Sander C, Stuart JM, Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN, Chu A, Chuah E, Chun HJE, Dhalla N, Guin R, Hirst M, Hirst C, Holt RA, Jones SJM, Lee D, Li HI, Marra MA, Mayo M, Moore RA, Mungall AJ, Robertson AG, Schein JE, Sipahimalani P, Tam A, Thiessen N, Varhol RJ, Beroukhim R, Bhatt AS, Brooks AN, Cherniack AD, Freeman SS, Gabriel SB, Helman E, Jung J, Meyerson M, Ojesina AI, Pedamallu CS, Saksena G, Schumacher SE, Tabak B, Zack T, Lander ES, Bristow CA, Hadjipanayis A, Haseley P, Kucherlapati R, Lee S, Lee E, Luquette LJ, Mahadeshwar HS, Pantazi A, Parfenov M, Park PJ, Protopopov A, Ren X, Santoso N, Seidman J, Seth S, Song X, Tang J, Xi R, Xu AW, Yang L, Zeng D, Auman JT, Balu S, Buda E, Fan C, Hoadley KA, Jones CD, Meng S, Mieczkowski PA, Parker JS, Perou CM, Roach J, Shi Y, Silva GO, Tan D, Veluvolu U, Waring S, Wilkerson MD, Wu J, Zhao W, Bodenheimer T, Hayes DN, Hoyle AP, Jeffreys SR, Mose LE, Simons J v., Soloway MG, Baylin SB, Berman BP, Bootwalla MS, Danilova L, Herman JG, Hinoue T, Laird PW, Rhie SK, Shen H, Triche T, Weisenberger DJ, Carter SL, Cibulskis K, Chin L, Zhang J, Sougnez C, Wang M, Getz G, Dinh H, Doddapaneni HV, Gibbs R, Gunaratne P, Han Y, Kalra D, Kovar C, Lewis L, Morgan M, Morton D, Muzny D, Reid J, Xi L, Cho J, Dicara D, Frazer S, Gehlenborg N, Heiman DI, Kim J, Lawrence MS, Lin P, Liu Y, Noble MS, Stojanov P, Voet D, Zhang H, Zou L, Stewart C, Bernard B, Bressler R, Eakin A, Iype L, Knijnenburg T, Kramer R, Kreisberg R, Leinonen K, Lin J, Liu Y, Miller M, Reynolds SM, Rovira H, Shmulevich I, Thorsson V, Yang D, Zhang W, Amin S, Wu CJ, Wu CC, Akbani R, Aldape K, Baggerly KA, Broom B, Casasent TD, Cleland J, Dodda D, Edgerton M, Han L, Herbrich SM, Ju Z, Kim H, Lerner S, Li J, Liang H, Liu W, Lorenzi PL, Lu Y, Melott J, Nguyen L, Su X, Verhaak R, Wang W, Wong A, Yang Y, Yao J, Yao R, Yoshihara K, Yuan Y, Yung AK, Zhang N, Zheng S, Ryan M, Kane DW, Aksoy BA, Ciriello G, Dresdner G, Gao J, Gross B, Jacobsen A, Kahles A, Ladanyi M, Lee W, Lehmann K van, Miller ML, Ramirez R, Rätsch G, Reva B, Schultz N, Senbabaoglu Y, Shen R, Sinha R, Sumer SO, Sun Y, Taylor BS, Weinhold N, Fei S, Spellman P, Benz C, Carlin D, Cline M, Craft B, Goldman M, Haussler D, Ma S, Ng S, Paull E, Radenbaugh A, Salama S, Sokolov A, Swatloski T, Uzunangelov V, Waltman P, Yau C, Zhu J, Hamilton SR, Abbott S, Abbott R, Dees

ND, Delehaunty K, Ding L, Dooling DJ, Eldred JM, Fronick CC, Fulton R, Fulton LL, Kalicki-Veizer J, Kanchi KL, Kandoth C, Koboldt DC, Larson DE, Ley TJ, Lin L, Lu C, Magrini VJ, Mardis ER, McLellan MD, McMichael JF, Miller CA, O'Laughlin M, Pohl C, Schmidt H, Smith SM, Walker J, Wallis JW, Wendl MC, Wilson RK, Wylie T, Zhang Q, Burton R, Jensen MA, Kahn A, Pihl T, Pot D, Wan Y, Levine DA, Black AD, Bowen J, Frick J, Gastier-Foster JM, Harper HA, Helsel C, Leraas KM, Lichtenberg TM, McAllister C, Ramirez NC, Sharpe S, Wise L, Zmuda E, Chanock SJ, Davidsen T, Demchok JA, Eley G, Felau I, Sheth M, Sofia H, Staudt L, Tarnuzzer R, Wang Z, Yang L, Zhang J, Omberg L, Margolin A, Raphael BJ, Vandin F, Wu HT, Leiserson MDM, Benz SC, Vaske CJ, Noushmehr H, Wolf D, Veer LVT, Anastassiou D, Yang THO, Lopez-Bigas N, Gonzalez-Perez A, Tamborero D, Xia Z, Li W, Cho DY, Przytycka T, Hamilton M, McGuire S, Nelander S, Johansson P, Jörnsten R, Kling T. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet [Internet]. 2013 Oct 1 [cited 2022 Jun 18];45(10):1113–20. Available from: https://pubmed.ncbi.nlm.nih.gov/24071849/

207. Marciano DC, Wang C, Hsu TK, Bourquard T, Atri B, Nehring RB, Abel NS, Bowling EA, Chen TJ, Lurie PD, Katsonis P, Rosenberg SM, Herman C, Lichtarge O. Evolutionary action of mutations reveals antimicrobial resistance genes in Escherichia coli. Nat Commun [Internet]. 2022 Dec 9 [cited 2022 Jun 18];13(1):3189. Available from: https://pubmed.ncbi.nlm.nih.gov/35680894/

208. Poirel L, Jayol A, Nordmanna P. Polymyxins: Antibacterial Activity, Susceptibility Testing, and Resistance Mechanisms Encoded by Plasmids or Chromosomes. Clin Microbiol Rev [Internet]. 2017 Apr 1 [cited 2022 Jun 18];30(2):557–96. Available from: https://pubmed.ncbi.nlm.nih.gov/28275006/

209. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim BJ, Kim YJ, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, MacArthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Neale BM, Daly MJ. The mutational

constraint spectrum quantified from variation in 141,456 humans. Nature [Internet]. 2020 May 28 [cited 2022 Jun 18];581(7809):434–43. Available from: https://pubmed.ncbi.nlm.nih.gov/32461654/

210. Lek M, Karczewski KJ, Minikel E v., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Abboud HE, Abecasis G, Aguilar-Salinas CA, Arellano-Campos O, Atzmon G, Aukrust I, Barr CL, Bell GI, Bergen S, Bjørkhaug L, Blangero J, Bowden DW, Budman CL, Burtt NP, Centeno-Cruz F, Chambers JC, Chambert K, Clarke R, Collins R, Coppola G, Córdova EJ, Cortes ML, Cox NJ, Duggirala R, Farrall M, Fernandez-Lopez JC, Fontanillas P, Frayling TM, Freimer NB, Fuchsberger C, García-Ortiz H, Goel A, Gómez-Vázquez MJ, González-Villalpando ME, González-Villalpando C, Grados MA, Groop L, Haiman CA, Hanis CL, Hattersley AT, Henderson BE, Hopewell JC, Huerta-Chagoya A, Islas-Andrade S, Jacobs SB, Jalilzadeh S, Jenkinson CP, Moran J, Jiménez-Morale S, Kähler A, King RA, Kirov G, Kooner JS, Kyriakou T, Lee JY, Lehman DM, Lyon G, MacMahon W, Magnusson PK, Mahajan A, Marrugat J, Martínez-Hernández A, Mathews CA, McVean G, Meigs JB, Meitinger T, Mendoza-Caamal E, Mercader JM, Mohlke KL, Moreno-Macías H, Morris AP, Najmi LA, Njølstad PR, O'Donovan MC, Ordóñez-Sánchez ML, Owen MJ, Park T, Pauls DL, Posthuma D, Revilla-Monsalve C, Riba L, Ripke S, Rodríguez-Guillén R, Rodríguez-Torres M, Sandor P, Seielstad M, Sladek R, Soberón X, Spector TD, Tai SE, Teslovich TM, Walford G, Wilkens LR, Williams AL. Analysis of protein-coding genetic variation in 60,706 humans. Nature [Internet]. 2016 Aug 17 [cited 2022 Jun 18];536(7616):285–91. Available from: https://pubmed.ncbi.nlm.nih.gov/27535533/

211. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature [Internet]. 2001 May 3 [cited 2022 Jun 18];411(6833):41–2. Available from: https://pubmed.ncbi.nlm.nih.gov/11333967/

212. Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J. Heritability of alternative splicing in the human genome. Genome Res [Internet]. 2007 [cited 2022 Jun 18];17(8):1210–8. Available from: https://pubmed.ncbi.nlm.nih.gov/17671095/

213. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, Scott E, Ciancanelli MJ, Lafaille FG, Markle JG, Martinez-Barricarte R, de Jong SJ, Kong XF, Nitschke P, Belkadi A, Bustamante J, Puel A, Boisson-Dupuis S, Stenson PD, Gleeson JG, Cooper DN, Quintana-Murci L, Claverie JM, Zhang SY, Abel L, Casanova JL. The human gene damage index as a gene-level approach to prioritizing exome variants. Proc Natl Acad Sci U S A [Internet]. 2015 Nov 3 [cited 2022 Jun 18];112(44):13615–20. Available from: https://pubmed.ncbi.nlm.nih.gov/26483451/

214. Ast G. How did alternative splicing evolve? Nat Rev Genet [Internet]. 2004 Oct [cited 2022 Jun 18];5(10):773–82. Available from: https://pubmed.ncbi.nlm.nih.gov/15510168/

215. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet [Internet]. 2006 Mar [cited 2022 Jun 18];7(3):211–21. Available from: https://pubmed.ncbi.nlm.nih.gov/16485020/

216. Fort P, Kajava A v., Delsuc F, Coux O. Evolution of proteasome regulators in eukaryotes. Genome Biol Evol [Internet]. 2015 May 1 [cited 2022 Jun 18];7(5):1363–79. Available from: https://pubmed.ncbi.nlm.nih.gov/25943340/

217. Zuin A, Isasa M, Crosas B. Ubiquitin signaling: extreme conservation as a source of diversity. Cells [Internet]. 2014 Jul 10 [cited 2022 Jun 18];3(3):690–701. Available from: https://pubmed.ncbi.nlm.nih.gov/25014160/

218. Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud N v., Williams LD. History of the ribosome and the origin of translation. Proc Natl Acad Sci U S A [Internet]. 2015 Dec 15 [cited 2022 Jun 18];112(50):15396–401. Available from: https://pubmed.ncbi.nlm.nih.gov/26621738/

219. Timsit Y, Sergeant-Perthuis G, Bennequin D. Evolution of ribosomal protein network architectures. Sci Rep [Internet]. 2021 Dec 1 [cited 2022 Jun 18];11(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33436806/

220. Romano AH, Conway T. Evolution of carbohydrate metabolic pathways. Res Microbiol [Internet]. 1996 [cited 2022 Jun 18];147(6–7):448–55. Available from: https://pubmed.ncbi.nlm.nih.gov/9084754/

221. Unden G, Bongaerts J. Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors. Biochim Biophys Acta [Internet]. 1997 Jul 4 [cited 2022 Jun 18];1320(3):217–34. Available from: https://pubmed.ncbi.nlm.nih.gov/9230919/

222. OMIM - Online Mendelian Inheritance in Man [Internet]. [cited 2022 Jun 18]. Available from: https://www.omim.org/

223. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet [Internet]. 2013 Aug [cited 2022 Jun 18];9(8). Available from: https://pubmed.ncbi.nlm.nih.gov/23990802/

224. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev M v., Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN, Osterman AL. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol [Internet]. 2003 Oct [cited 2022 Jun 18];185(19):5673–84. Available from: https://pubmed.ncbi.nlm.nih.gov/13129938/

225. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. The Essential Genome of Escherichia coli K-12. mBio [Internet]. 2018 Jan 1 [cited 2022 Jun 18];9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29463657/

226. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol [Internet]. 2006 May 16 [cited 2022 Jun 18];2. Available from: https://pubmed.ncbi.nlm.nih.gov/16738554/

227. Keseler IM, Gama-Castro S, Mackie A, Billington R, Bonavides-Martínez C, Caspi R, Kothari A, Krummenacker M, Midford PE, Muñiz-Rascado L, Ong WK, Paley S, Santos-Zavaleta A, Subhraveti P, Tierrafría VH, Wolfe AJ, Collado-Vides J, Paulsen IT, Karp PD. The EcoCyc Database in 2021. Front Microbiol [Internet]. 2021 Jul 28 [cited 2022 Jun 18];12. Available from: https://pubmed.ncbi.nlm.nih.gov/34394059/

228. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA. Phenotypic landscape of a bacterial cell. Cell [Internet]. 2011 Jan 7 [cited 2022 Jun 18];144(1):143–56. Available from: https://pubmed.ncbi.nlm.nih.gov/21185072/

229. Rousset F, Cabezas-Caballero J, Piastra-Facon F, Fernández-Rodríguez J, Clermont O, Denamur E, Rocha EPC, Bikard D. The impact of genetic diversity on gene essentiality within the Escherichia coli species. Nat Microbiol [Internet]. 2021 Mar 1 [cited 2022 Jun 18];6(3):301–12. Available from: https://pubmed.ncbi.nlm.nih.gov/33462433/

230. Preissner SC, Hoffmann MF, Preissner R, Dunkel M, Gewiess A, Preissner S. Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. PLoS One [Internet]. 2013 Dec 10 [cited 2022 Jun 18];8(12). Available from: https://pubmed.ncbi.nlm.nih.gov/24340040/

231. Mehboob H, Tahir IM, Iqbal T, Saleem S, Perveen S, Farooqi A. Effect of UDP-Glucuronosyltransferase (UGT) 1A Polymorphism (rs8330 and rs10929303) on Glucuronidation Status of Acetaminophen. Dose Response [Internet]. 2017 Jul 13 [cited 2022 Jun 18];15(3). Available from: https://pubmed.ncbi.nlm.nih.gov/28932176/

232. Yan C, Wang R, Li J, Deng Y, Wu D, Zhang H, Zhang H, Wang L, Zhang C, Sun H, Zhang X, Wang J, Yang H, Li S. HLA-A gene polymorphism defined by high-resolution sequence-based typing in 161 Northern Chinese Han people. Genomics Proteomics Bioinformatics [Internet]. 2003 [cited 2022 Jun 18];1(4):304–9. Available from: https://pubmed.ncbi.nlm.nih.gov/15629059/

233. Jin P, Wang E. Polymorphism in clinical immunology - From HLA typing to immunogenetic profiling. J Transl Med [Internet]. 2003 Nov 18 [cited 2022 Jun 23];1(1). Available from: https://pubmed.ncbi.nlm.nih.gov/14624696/

234. Hughes GM, Teeling EC, Higgins DG. Loss of olfactory receptor function in hominin evolution. PLoS One [Internet]. 2014 Jan 2 [cited 2022 Jun 18];9(1). Available from: https://pubmed.ncbi.nlm.nih.gov/24392153/

235. Trimmer C, Keller A, Murphy NR, Snyder LL, Willer JR, Nagai MH, Katsanis N, Vosshall LB, Matsunami H, Mainland JD. Genetic variation across the human olfactory receptor repertoire alters odor perception. Proc Natl Acad Sci U S A [Internet]. 2019 May 7 [cited 2022 Jun 18];116(19):9475–80. Available from: https://pubmed.ncbi.nlm.nih.gov/31040214/

236. Ramesh A, Darko S, Hua A, Overman G, Ransier A, Francica JR, Trama A, Tomaras GD, Haynes BF, Douek DC, Kepler TB. Structure and Diversity of the Rhesus Macaque Immunoglobulin Loci through Multiple De Novo Genome Assemblies. Front Immunol [Internet]. 2017 Oct 27 [cited 2022 Jun 18];8(OCT). Available from: https://pubmed.ncbi.nlm.nih.gov/29163486/

237. Doxiadis GGM, Otting N, de Groot NG, Noort R, Bontrop RE. Unprecedented polymorphism of Mhc-DRB region configurations in rhesus macaques. J Immunol [Internet]. 2000 Mar 15 [cited 2022 Jun 18];164(6):3193–9. Available from: https://pubmed.ncbi.nlm.nih.gov/10706710/

238. Goeders N, van Melderen L. Toxin-antitoxin systems as multilevel interaction systems. Toxins (Basel) [Internet]. 2014 [cited 2022 Jun 18];6(1):304–24. Available from: https://pubmed.ncbi.nlm.nih.gov/24434905/

239. Vanchurin V, Wolf YI, Koonin E v., Katsnelson MI. Thermodynamics of evolution and the origin of life. Proc Natl Acad Sci U S A [Internet]. 2022 Feb 8 [cited 2022 Jun 18];119(6). Available from: https://pubmed.ncbi.nlm.nih.gov/35131858/

240. Barton NH, Coe JB. On the application of statistical physics to evolutionary biology. J Theor Biol [Internet]. 2009 Jul 21 [cited 2022 Jun 18];259(2):317–24. Available from: https://pubmed.ncbi.nlm.nih.gov/19348811/

241. Novikov IB, Wilkins AD, Lichtarge O. An Evolutionary Trace method defines functionally important bases and sites common to RNA families. PLoS Comput Biol [Internet]. 2020 [cited 2022 Jun 18];16(3). Available from: https://pubmed.ncbi.nlm.nih.gov/32208421/

242. Pei G, Hu R, Jia P, Zhao Z. DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue- and cell type-specific manner. Nucleic Acids Res [Internet]. 2021 Jul 2 [cited 2022 Jun 18];49(W1):W131–9. Available from: https://pubmed.ncbi.nlm.nih.gov/34048560/

243. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods [Internet]. 2015 Sep 29 [cited 2022 Jun 18];12(10):931–4. Available from: https://pubmed.ncbi.nlm.nih.gov/26301843/

244. Trieu T, Martinez-Fundichely A, Khurana E. DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. Genome Biol [Internet]. 2020 Mar 26 [cited 2022 Jun 18];21(1). Available from: https://pubmed.ncbi.nlm.nih.gov/32216817/

245. van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K v., Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics [Internet]. 2013 [cited 2022 Jun 18];43(1110). Available from: https://pubmed.ncbi.nlm.nih.gov/25431634/

246. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA van der, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ, Neale B, MacArthur DG, Banks E. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv [Internet]. 2018 Jul 24 [cited 2022 Jun 18];201178. Available from: https://www.biorxiv.org/content/10.1101/201178v3

247. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res [Internet]. 2010 Sep [cited 2022 Jun 18];20(9):1297–303. Available from: https://pubmed.ncbi.nlm.nih.gov/20644199/

248. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021 Feb 16 [cited 2022 May 16];10(2). Available from: https://pubmed.ncbi.nlm.nih.gov/33590861/

249. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81(3):559–75.

250. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genetics [Internet]. 2006 Dec [cited 2022 Jun 18];2(12):e190. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020190

251. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 2006 38:8 [Internet]. 2006 Jul 23 [cited 2022 Jun 18];38(8):904–9. Available from: https://www.nature.com/articles/ng1847

252. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 Jun 6 [cited 2022 Jun 18];17(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27268795/

253. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin A v., Sirotkin A v., Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol [Internet]. 2012 May 1 [cited 2022 Jun 18];19(5):455–77. Available from: https://pubmed.ncbi.nlm.nih.gov/22506599/

254. Lomsadze A, Gemayel K, Tang S, Borodovsky M. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. Genome Res [Internet]. 2018 Jul 1 [cited 2022 Jun 18];28(7):1079–89. Available from: https://pubmed.ncbi.nlm.nih.gov/29773659/

255. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, da Rin Fioretto L, Davidson C, Dodiya K, Donaldson S, el Houdaigui B, el Naboulsi T, Fatima R, Giron CG, Genez T, Martinez JG, Guijarro-Clarke C, Gymer A, Hardy

M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh DN, Parker A, Parton A, Perry M, Pilizota I, Prosovetskaia I, Sakthivel MP, Salam AIA, Schmitt BM, Schuilenburg H, Sheppard D, Perez-Silva JG, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner MM, Szpak M, Thormann A, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt SE, Iisley GR, Loveland JE, Martin FJ, Moore B, Mudge JM, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion SJ, Dyer S, Harrison PW, Howe KL, Yates AD, Zerbino DR, Flicek P. Ensembl 2022. Nucleic Acids Res [Internet]. 2022 Jan 7 [cited 2022 Jun 18];50(D1):D988–95. Available from: https://pubmed.ncbi.nlm.nih.gov/34791404/

256. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A pietro, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik D v., Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods [Internet]. 2020 Mar 1 [cited 2022 Jun 18];17(3):261–72. Available from: https://pubmed.ncbi.nlm.nih.gov/32015543/

257. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res [Internet]. 2003 Nov [cited 2022 Jun 18];13(11):2498–504. Available from: https://pubmed.ncbi.nlm.nih.gov/14597658/

258. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res [Internet]. 2019 Feb 1 [cited 2022 Jun 18];18(2):623–32. Available from: https://pubmed.ncbi.nlm.nih.gov/30450911/

259. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics [Internet]. 2011 Nov 9 [cited 2022 Jun 18];12. Available from: https://pubmed.ncbi.nlm.nih.gov/22070249/

260. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun [Internet]. 2017 Dec 1 [cited 2022 Jun 18];8(1). Available from: https://pubmed.ncbi.nlm.nih.gov/29184056/

261. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS. ClinGen--the Clinical Genome Resource. N Engl J Med [Internet]. 2015 Jun 4 [cited 2022 Jun 18];372(23):2235–42. Available from: https://pubmed.ncbi.nlm.nih.gov/26014595/

262. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;

263. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421--7.

264. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, Als TD, Bigdeli TB, Børte S, Bryois J, Charney AW, Drange OK, Gandal MJ, Hagenaars SP, Ikeda M, Kamitaki N, Kim M, Krebs K, Panagiotaropoulou G, Schilder BM, Sloofman LG, Steinberg S, Trubetskoy V, Winsvold BS, Won HH, Abramova L, Adorjan K, Agerbo E, al Eissa M, Albani D, Alliey-Rodriguez N, Anjorin A, Antilla V, Antoniou A, Awasthi S, Baek JH, Bækvad-Hansen M, Bass N, Bauer M, Beins EC, Bergen SE, Birner A, Bøcker Pedersen C, Bøen E, Boks MP, Bosch R, Brum M, Brumpton BM, Brunkhorst-Kanaan N, Budde M, Bybjerg-Grauholm J, Byerley W, Cairns M, Casas M, Cervantes P, Clarke TK, Cruceanu C, Cuellar-Barboza A, Cunningham J, Curtis D, Czerski PM, Dale AM, Dalkner N, David FS, Degenhardt F, Djurovic S, Dobbyn AL, Douzenis A, Elvsåshagen T, Escott-Price V, Ferrier IN, Fiorentino A, Foroud TM, Forty L, Frank J, Frei O, Freimer NB, Frisén L, Gade K, Garnham J, Gelernter J, Giørtz Pedersen M, Gizer IR, Gordon SD, Gordon-Smith K, Greenwood TA, Grove J, Guzman-Parra J, Ha K, Haraldsson M, Hautzinger M, Heilbronner U, Hellgren D, Herms S, Hoffmann P, Holmans PA, Huckins L, Jamain S, Johnson JS, Kalman JL, Kamatani Y, Kennedy JL, Kittel-Schneider S, Knowles JA, Kogevinas M, Koromina M, Kranz TM, Kranzler HR, Kubo M, Kupka R, Kushner SA, Lavebratt C, Lawrence J, Leber M, Lee HJ, Lee PH, Levy SE, Lewis C, Liao C, Lucae S, Lundberg M, MacIntyre DJ, Magnusson SH, Maier W, Maihofer A, Malaspina D, Maratou E, Martinsson L, Mattheisen M, McCarroll SA, McGregor NW, McGuffin P, McKay JD, Medeiros H, Medland SE, Millischer V, Montgomery GW, Moran JL, Morris DW, Mühleisen TW, O'Brien N, O'Donovan C, Olde Loohuis LM, Oruc L, Papiol S, Pardiñas AF, Perry A, Pfennig A, Porichi E, Potash JB, Quested D, Raj T, Rapaport MH, DePaulo JR, Regeer EJ, Rice JP, Rivas F, Rivera M, Roth J, Roussos P, Ruderfer DM, Sánchez-Mora C, Schulte EC, Senner F, Sharp S, Shilling PD, Sigurdsson E, Sirignano L, Slaney C, Smeland OB, Smith DJ, Sobell JL, Søholm Hansen C, Soler Artigas M, Spijker AT, Stein DJ, Strauss JS, Świątkowska B, Terao C, Thorgeirsson TE, Toma C, Tooney P, Tsermpini EE, Vawter MP, Vedder H, Walters JTR, Witt SH, Xi S, Xu W, Yang JMK, Young AH, Young H, Zandi PP, Zhou H, Zillich L, Adolfsson R, Agartz I, Alda M, Alfredsson L, Babadjanova G, Backlund L, Baune BT, Bellivier F, Bengesser S, Berrettini WH, Blackwood DHR, Boehnke M, Børglum AD, Breen G, Carr VJ, Catts S, Corvin A, Craddock N, Dannlowski U, Dikeos D, Esko T, Etain B, Ferentinos P, Frye M, Fullerton JM, Gawlik M, Gershon ES, Goes FS, Green MJ, Grigoroiu-Serbanescu M, Hauser J, Henskens F, Hillert J, Hong KS, Hougaard DM, Hultman CM, Hveem K, Iwata N, Jablensky A v., Jones I, Jones LA, Kahn RS, Kelsoe JR, Kirov G, Landén M, Leboyer M, Lewis CM, Li QS, Lissowska J, Lochner C, Loughland C, Martin NG, Mathews CA, Mayoral F, McElroy SL, McIntosh AM, McMahon FJ, Melle I, Michie P, Milani L, Mitchell PB, Morken G, Mors O, Mortensen PB, Mowry B, Müller-Myhsok B, Myers RM, Neale

BM, Nievergelt CM, Nordentoft M, Nöthen MM, O'Donovan MC, Oedegaard KJ, Olsson T, Owen MJ, Paciga SA, Pantelis C, Pato C, Pato MT, Patrinos GP, Perlis RH, Posthuma D, Ramos-Quiroga JA, Reif A, Reininghaus EZ, Ribasés M, Rietschel M, Ripke S, Rouleau GA, Saito T, Schall U, Schalling M, Schofield PR, Schulze TG, Scott LJ, Scott RJ, Serretti A, Shannon Weickert C, Smoller JW, Stefansson H, Stefansson K, Stordal E, Streit F, Sullivan PF, Turecki G, Vaaler AE, Vieta E, Vincent JB, Waldman ID, Weickert TW, Werge T, Wray NR, Zwart JA, Biernacka JM, Nurnberger JI, Cichon S, Edenberg HJ, Stahl EA, McQuillin A, di Florio A, Ophoff RA, Andreassen OA. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet [Internet]. 2021 Jun 1 [cited 2022 Jun 18];53(6):817–29. Available from: https://pubmed.ncbi.nlm.nih.gov/34002096/

265. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. Circ Res [Internet]. 2018 Feb 2 [cited 2022 Jun 18];122(3):433–43. Available from: https://pubmed.ncbi.nlm.nih.gov/29212778/

266. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet [Internet]. 2010 Jun [cited 2022 Jun 18];11(6):446–50. Available from: https://pubmed.ncbi.nlm.nih.gov/20479774/

267. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet [Internet]. 2009 Jun [cited 2022 Jun 18];10(6):392–404. Available from: https://pubmed.ncbi.nlm.nih.gov/19434077/

268. Yi HC, You ZH, Huang DS, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. Brief Bioinform [Internet]. 2022 Jan 1 [cited 2022 Jun 18];23(1). Available from: https://pubmed.ncbi.nlm.nih.gov/34471921/

269. da Pozzo E, la Pietra V, Cosimelli B, da Settimo F, Giacomelli C, Marinelli L, Martini C, Novellino E, Taliani S, Greco G. P53 functional inhibitors behaving like pifithrin-β counteract the alzheimer peptide non-β-amyloid component effects in human sh-sy5y cells. ACS Chemical Neuroscience [Internet]. 2014 May 21 [cited 2022 Jun 18];5(5):390–9. Available from: https://pubs.acs.org/doi/abs/10.1021/cn4002208

270. Farmer KM, Ghag G, Puangmalai N, Montalbano M, Bhatt N, Kayed R. P53 aggregation, interactions with tau, and impaired DNA damage response in Alzheimer's disease. Acta Neuropathologica Communications 2020 8:1 [Internet]. 2020 Aug 10 [cited 2022 Jun 18];8(1):1–21. Available from: https://actaneurocomms.biomedcentral.com/articles/10.1186/s40478-020-01012-6

271. Abate G, Frisoni GB, Bourdon JC, Piccirella S, Memo M, Uberti D. The pleiotropic role of p53 in functional/dysfunctional neurons: focus on pathogenesis and diagnosis of Alzheimer's disease. Alzheimer's Research and Therapy [Internet]. 2020 Dec 1 [cited 2022 Jun 18];12(1):1–10. Available from: https://alzres.biomedcentral.com/articles/10.1186/s13195-020-00732-0

272. Fuchs P, Zörer M, Reipert S, Rezniczek GA, Propst F, Walko G, Fischer I, Bauer J, Leschnick MW, Lüscher B, Thalhammer JG, Lassmann H, Wiche G. Targeted inactivation of a developmentally regulated neural plectin isoform (plectin 1c) in mice leads to reduced motor nerve conduction velocity. Journal of Biological Chemistry. 2009;

273. Jaynes ET. Information Theory and Statistical Mechanics. Physical Review [Internet]. 1957 May 15 [cited 2022 Jun 18];106(4):620. Available from: https://journals.aps.org/pr/abstract/10.1103/PhysRev.106.620

274. Jaynes ET. Information Theory and Statistical Mechanics. II. Physical Review [Internet]. 1957 Oct 15 [cited 2022 Jun 18];108(2):171. Available from: https://journals.aps.org/pr/abstract/10.1103/PhysRev.108.171

275. Franchini LF, Pollard KS. Human evolution: The non-coding revolution. BMC Biology [Internet]. 2017 Oct 2 [cited 2022 Jun 18];15(1):1–12. Available from: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0428-9

276. Parvandeh S, Donehower LA, Panagiotis K, Hsu TK, Asmussen JK, Lee K, Lichtarge O. EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants. Nucleic Acids Res [Internet]. 2022 Apr 12 [cited 2022 Jun 18]; Available from: https://pubmed.ncbi.nlm.nih.gov/35412634/

VITA

Yashwanth Lagisetty was born to Sai Ramesh Lagisetty and Vani Shree Lagisetty. He received a Bachelor of Sciences in Physics at the University of Michigan in 2015. During his undergraduate studies, Yash conducted research in the lab of Dr. Thomas Schwarz where he worked primarily on direct detection of Dark Matter experiments and beyond standard model Higgs physics. Following this, Yash joined the Medical Scientist Training Program at the McGovern Medical school and The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.