Graph Neural Networks to Identify Genetic Modifiers of Rare Complex Inheritable Diseases

Eliatan Niktab

2023

A thesis submitted to
the Victoria University of Wellington
in fulfillment of the requirements for
the degree of Doctor of Philosophy

Preface

For thou who fell, so I did not.

For thou who did not have, so I did!

Acknowledgments

I want to extend my heartfelt gratitude to Paul, Andrew, Mark, Ingrid, Shona, Gillian, and Patricia for their unwavering support throughout this journey.

I am immensely grateful to Callaghan Innovation, the Parseghian Medical Research Fund, Bethlehem Griffiths Research Foundation, Maurice Wilkins Centre, and Google Cloud for generously providing the necessary funding for this project.

My sincere appreciation goes to Calvary Public Hospital and Royal Melbourne Hospital for their assistance in patient recruitment.

Lastly, and most importantly, I am profoundly thankful to the Huntington's disease and Niemann-Pick type C patients who participated in this study for placing their trust in me and allowing me to work with their genomic data. Your belief in me has been invaluable in making this project a reality.

Thesis abstract

Genome-wide association analyses (GWAS) studies based on frequentist statistics have often proven ineffective in deriving biological insights from sequencing data. These GWAS lack the machinery to safeguard against technical noise inherent to high throughput sequencing platforms and are not conceptually designed for processing large sets of high-dimensional genomic data. However, such shortcomings are not peculiar to GWAS and have been studied in other fields of science, such as signal processing and computer science, for a long time. In particular, machine learning techniques, especially deep learning models, have proven highly successful in dealing with noisy high-dimensional data. Recently it has been shown that these techniques can be effective for handling genomic data even when directly transferred from modern computer vision and natural language processing applications.

This thesis builds off the existing suites of such methodologies and presents a robust computational pipeline to functionally annotate whole-genome sequencing data. Moreover, it discusses and presents a data solution to efficiently process the large, heterogeneous datasets required for such analyses. The main objective of this thesis is to put forward a solution to identify variants that modify disease-causing mutations of complex heritable diseases. This is not a trivial problem given that the current gold standard approach, GWAS methodology, suffers not only from the drawbacks just described but is also underpowered by multiple testing (not useful for rare diseases) and fails to account for the epistatic nature of genetic interactions responsible for the onset and manifestation of complex diseases.

Here, a set of cell-specific Gene Regulatory Networks (GRNs) inferred from dynamic genomic data was constructed. Most attempts to construct GRNs delineating such complex interactions relied on combining non-standardized high-throughput static datasets that contained false positive interactions and missing data points without insights into cell developmental states. To illuminate these intricate dynamic regulatory interconnections of the genome, specific to a tissue or a cell type, the Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks (NS-DIMCORN) that allows unrestricted neural network architectures (to accommodate arbitrary depth increase for larger sets of genes) and training without partitioning the data dimensions was developed. NS-DIMCORN was trained on not-homogenized bulk tissue-specific RNA-seq and single-cell RNA-seq as a surrogate for cells' continuous developmental states and modeled these highly dynamic systems with a set of ordinary differential equations. NS-DIMCORN yielded a continuous-time invertible generative model with unbiased density estimation only from RNA-seq read-count data and allowed time-flexible sampling of each gene's expression level for *ab initio* assembly of genes regulatory network of specific cells.

Secondly, Precise Graph-based Genome-Wide Annotation Sofware (PG-GWAS) was developed. For this purpose, embedding was used to map genomic variables to a vector of continuous numbers. Thus, each genomic variant was assigned a unique contextualized score that encoded the likelihood of effects on its respective gene products. These scores were pan-genomic by constructing a k-mer representation of all the haplotypes, independent of any "reference genome," and were based only on each variant's evolutionary constraints. Next, a graph representation of individuals' genomes was constructed that integrated genomic variation scores, tissue-specific gene-gene interaction, and regulatory networks (assembled from GRNs) to allow the study of the genomic variants in aggregate and accounting for epistasis. Utilizing the Graph Attention mechanism identified these networks' most critical interactions and

allowed annotating the entire whole-genome graphs to determine the most prominent genomic features (i.e., groups of interacting genes) within each genome that could be responsible for different symptoms and onset in patients with the same disease-causing mutations. Eventually, to demonstrate the efficacy of this approach, PG-GWAS was tested on new sets of sequencing data, where the result improved in standard GWAS and provided insight into disease epistasis.

Contents

List of used abbreviations

A	Adenine
ATP	Adenosine triphosphate
AII	All Interactions Identified
AUPRC	Area Under the Precision-Recall Curve
AUROC	Area Under the Receiver Operating Characteristic
F1-S	Balanced F-score
CC	Compound-Complex
CDiR	Correct Directions
Cy	Cyclic
C	Cytosine
DC	Degree Centrality
Den	Density
DCL	Diagnostic Confidence Level
DE	Differentially Expressed
EPR	Early Precision Ratio
GRNs	Gene Regulatory Networks
GWAS	Genome-wide association analyses
GII	Genuine Identified Interactions
GAT	Graph Attention Network
GCE	Graph Clustering Coefficient
GD	Graph Degree

GDm	Graph Diameter
GDi	Graph Distance
GFid	Graph Fidelity
GM	Graph Modularity
GPL	Graph Path Length
GTNs	Ground Truth Networks
G	Guanine
HMC	Hamiltonian Monte Carlo
HD	Huntington's disease
INDELs	Insertion-Deletion
KTSNs	Known to be Truth Sub-Networks of experimental data
LD	Linkage Disequilibrium
Li	Long Linear
lncRNA	long non-coding RNA
LoF	Loss-of-Function
MAD	Mean Absolute Deviation
mRNAs	messenger RNAs
MAF	Minor Allele Frequency
MEI	Missing Expected Interactions
MI	Mutual Information
MPNNs	Message Passing Neural Networks
MT	Mitochondrial DNA
NLP	Natural Language Processing
NGS	Next Generation Sequencing
NS-DIMCORN	Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks
NES	Normalized Embedding Score
NPC-C	Niemann-Pick type C

NSS	Normalized Severity Score
NVs	Nucleotide Variant
OMIM	Online Mendelian Inheritance in Man
ODEs	Ordinary Differential Equations
PG-GWAS	Precise Graph-based Genome-Wide Annotation Sofware
Pr	Precision
ROC	Receiver Operating Characteristic
RNNs	Recurrent Neural Networks
RNA	Ribonucleic Acid
tRNAs	Transfer RNAs
SNPs	Single Nucleotide Polymorphisms
scDNA-seq	single-cell DNA sequencing
D	The Averaged Degree
T	Thymine
T5KI	Top 5000 Identified Interactions
TFs	Transcription Factors
Tr	Trifurcating
UHDRS	Unified Huntington Disease Rating Scale

List of Figures

List of Tables

Chapter 1

Technical background and thesis outline

1.1 A primer on human genetics

The entire genetic code of humans (the genome) of 32 00 000 000 nucleotides is found in 23 separate linear molecules called DNA that are tightly packed in chromosomes.[1] Chromosomes comprise DNA wrapped in histones (chromatin) that are methylated in specific places[2]. A nucleotide is the basic building block of DNA that is attached to a phosphate group and a nitrogen-containing base, Adenine (A), Cytosine (C), Guanine (G), or Thymine (T), where the order of these four nucleotides determines the encoded message for making specific proteins[3]. The DNA code is transcribed, where the genetic information stored in DNA is converted into an RNA molecule, specifically messenger RNA (mRNA), encoding hundreds of thousands of protein-specifying codes[4]. This system permutes three nucleotides at a time in the mRNA to make a linear sequence of amino acids specifying codons. A decoding system uses the 20 amino-acid-specific transfer RNAs (tRNAs)[3]. The latter have base-pairing anticodons, one for each of the 20 amino acids, causing translation into specific proteins of linear chains of amino acid residues[4]. Proteins perform essential functions within organisms, including a plethora of enzyme activities, such as those for DNA replication and RNA transcription, the respiratory cycle, and the generation of Adenosine triphosphate (ATP)[3]. They also provide structural elements (e.g., collagen, ECM, microtubules, microfibrils) of various natures. But apart from encoding proteins, certain RNA species may also be used as essential regulatory elements such as miRNAs and long non-coding RNA (lncRNA)[3]. Intriguingly the trillions of cells making up the human body each perform a different function at any given time, entirely dependent on which genes are expressed and in what quantity[5–7]. This intricate dynamic system is self-controlled through an exquisite regulatory system of genes, cis-regulatory-element, and sets of trans-regulatory-elements such as Transcription Factors (TFs), histone methylations, and de-methylations[3]. In this

perfectly balanced system, any DNA sequence variation (genetic variant) can drastically affect cell morphologies, development, or normal function and can contribute directly or indirectly to diseases[6]. These variations can be at single loci genomic positions, comprise longer nucleotide sequences, or include copy number variations, translocations, and inversions.

1.1.1 DNA mutation and genetic variants

A Nucleotide Variant (NV), depending on its type (single nucleotide polymorphisms/ insertion-deletion) or its location (coding/non-coding region), can increase, reduce or completely stop the expression of a gene or affect its downstream expression (e.g. glycosylation, multimerization through sulfhydryl bonds, ionic and van der Waals bridges, hydrophobic interactions), to affect protein folding to functional entities[3]. In Single Nucleotide Polymorphisms (SNPs), only one nucleotide in DNA base pairs is changed and there are several types of SNPs. A missense SNP may occur in the protein-coding region of DNA and such a variation at a given locus may result in coding different amino acids, eventually altering the properties of that gene's downstream mature protein products[3]. A nonsense SNP if produces a stop codon, terminates protein synthesis prematurely and causes a loss of function in that protein[3]. Insertion-Deletion (INDELs) involves adding or removing one or few nucleotides into the DNA sequence, changing the gene's reading frame during RNA transcription[3]. Nucleotide variants in non-protein-coding regions can contribute to irregular cell functions by impacting mRNA processing, chromatin interactions, and DNA expression by altering transcription factor binding sites and lncRNA abnormalities[8]. Effects of genetic variants are among the main considerations of inheritable disease susceptibility, also called broad sense heritability (H^2)[9].

1.1.2 Gene Regulatory Networks (GRNs)

Large numbers of genes work together as a highly interconnected dynamic system, known as gene regulatory networks, to regulate and execute every cellular function in cells, such as differentiation, metabolism, the cell cycle, signal transduction, and so on[5]. Although an oversimplification of the problem, GRNs are often modeled by perturbations where a combination of two or a few genes are systematically knocked out (removed) from the genome to investigate their contributory relationship[10]. More recently, transcriptomic profiling of individual cells (single-cell DNA sequencing) has allowed more realistic modeling of GRNs, by comparing gene expression in different cells at different developmental stages or under the effect of different stressors[11]. Chapter 2 reviews various computational models developed for GRN inference and analysis and presents novel tools to infer GRNs from single-cell DNA sequencing (scRNA-seq) data.

1.1.3 Genetic architecture of diseases

Historically, heritable human diseases have been broadly classified as Mendelian or complex disorders using oversimplified groupings/clustering of genes to explain the underlying genetic architecture of diseases[12]. Genetic architecture describes all attributes of genetic contributions to a given phenotype (e.g., disease symptoms), such as genetic variants influencing the phenotype, their effect size, frequency, and interactions with each other and the environment[12]. A disease is called Mendelian if disease-causing genomic variants segregate according to Mendel's inheritance laws. These disorders are usually caused by rare genetic variations with high penetrance (effect size) as they are negatively selected for in the population. A Mendelian disease is termed monogenic if heritable Loss-of-Function (LoF) single variants (mutations) are responsible for the observed phenotypic trait. A complex disease, by contrast, does not follow Mendelian inheritance patterns and can result from any

combination of multiple genetic factors or interactions of these factors. These diseases are considered polygenic, a term conveying that many genetic variants contribute to phenotypic variability observed in patients with the disease. However, given the implications of highly interconnected GRNs, more recent studies have introduced the concept of universal pleiotropy for complex symptoms of disease[13–15]. Pleiotropy is when a location on the genome affects two or more unrelated phenotypic traits. In the contemporary era of disease etiology research, the complex and highly comorbid symptoms of diseases are viewed as results of an omnigenic architecture, that is, every gene expressed in a cell relevant to the disease contributes to the disease's manifestation.[16,17]. In 2016, a novel genome sequencing analysis identified apparently healthy individuals (genetic superheroes) resilient to the effects of LoF mutations that cause eight Mendelian diseases[18].

1.1.4 Genome Wide Association Analyses

Genome Wide Association Analyses (GWAS) was developed as a new approach based on the success of vast amounts of next-generation sequencing of individuals[19]. GWAS is a technique that allows individuals' entire genomes to be fine-mapped readily on an unprecedented scale to a phenotype. Hence it allowed association by statistical analysis to fine-grained phenotypes. A typical GWAS investigates the association of genetic variation to the phenotype of interest with one or up to million genetic markers. Frequency differences of these markers in healthy and affected groups of individuals are compared to implicate their relationship to phenotypes. Thus, for each genetic marker, one at a time, a linear regression model (logistic regression model for binary phenotypes and linear regression for quantitive phenotypes) is fitted to predict the target phenotype using influential genetic variants and their covariates as input. The coefficient for the genetic variant term is then tested for significance using statistical hypothesis tests such as the likelihood ratio test or Wald's test. However, unlike the extensive success of GWAS in implicating

high penetrance disease-causing variants for diseases with monogenic and oligogenic (phenotype as a result of few variants only) causes, there has been little success in identifying the role of the numerous disease modifying variants or variants with small effect sizes in heritable diseases[19]. Identifying modifiers is particularly difficult since modifiers, unlike causal variants, are not necessarily rare, given that their action is epistatic with the disease-causing gene. Epistasis is a phenomenon in which the effect of a genetic variant is dependent on the presence or absence of other genetic variations due to gene buffering and regulatory effects of part of the genome on other parts. Shortcomings of GWAS are discussed in more detail in Chapter 3, but briefly, these include the over-conservativeness of multiple testing correction, ignorance of population stratification, the confounding effect of Linkage Disequilibrium (LD) between neighboring markers, arbitrary LoF scores, the inclusion of only common variants, ignoring epistasis, and lastly the power reduction of some models owing to over-fitting of linear regression for genetic markers separately.

1.1.5 Non-parametric tests and deep learning in genomics

Deep learning is now used extensively in genomics research as it can perform the necessary dimensionality reduction required for dealing with high dimensional-omics data, hence capturing the complex nature of biological functions for analysis[20,21]. The input into a neural network, the main form of deep learning, is typically matrices of vectorized values, and outcomes are predictions related to the input. In genomics, this input can be a DNA/RNA sequence or -omics data after being processed into a format that can be fed into neural networks (e.g., One-hot encoding[22]) for particular conditions such as diseases or a phenotype. On the other hand, outputs can be predictions about gene expression ratios, effects of genetic variations on a cellular process, or disease manifestations. However, the pitfall of deep learning is the need for a large amount of training data for models to learn how to predict the outcome from the inputs. Such data are often scarce and sometimes not even available for

many biological questions because there is no known answer or human recognizable pattern that can use for labeling of the training data[20].

1.2 Thesis outline

1.2.1 Overview

Indeed, the main contribution of this thesis is not developing new deep learning algorithms (although improved architectures were devised and distributed data pipelines were developed to accommodate working with terabytes of genomic data) but is the innovative agglomeration and transformation of biological data in a way that allows utilizing state-of-the-art deep learning algorithms efficiently.

In Chapter 2, I used a flavor of the generative models[23] to infer GRNs from single-cell sequencing data. Single-cell sequencing data capture the ratio of each gene's expression in thousands of cells at different developmental stages. The rationale for this choice was that generating single-cell sequencing data is unrestricted (one can sequence many cells at different developmental stages), but labeling this data (i.e., what are the most active genes in each cell at an exact time during a developmental stage) is yet technically impossible[24]. Unsupervised learning methods such as generative models have the potential to leverage these large pools of unlabeled data (thousands of cells at many different developmental stages). Particularly for this approach, a large amount of data is collected, and then the model is trained to generate data like the input by trying to learn the input data distribution function. More accurately, by training a neural network with significantly fewer parameters than the data dimension, the generative model is forced to learn the essence of the data to generate it efficiently.

In Chapter 3, Natural Language Processing (NLP) is utilized to capture the deleterious effect of genetic variants. Although NLP has been used for predicting consequences of genetic variants[25], here, by representing whole genome DNA sequences as a graph comprising the complete sets of possible combinations of nucleotides, my contribution is looking at genetic variants, not in comparison to one

reference genome or trying to predict their deleterious effects in isolation but in the genome as a whole. Briefly, I assigned a predictive pan-genomic[26], contextualized (aware of variants before or after) meaning to each genetic variant that is quantifiable in the sense of deleterious effect on downstream products and is understandable to a computer.

Thirdly, Graph Attention Network (GAT), a neural network architecture that operates on graph-structured data, was used to combine information about a pan-genomic score of genetic variants (nodes of the graph) and genetic interactions (edges of the graph) to model omnigenic architecture of the Mendelian diseases with varying complex symptoms as a result of genetic modifiers. My other notable contribution here is an assembly of the whole genome as a graph. As a result, GAT can leverage masked self-attentional layers, which allow nodes of the graph to be weighted depending on their importance (here, contribution to trait) and pool information across the graph, allowing disease to be studied truly under the omnigenic paradigm.

1.2.2 Aims

There are more than 6,000 Mendelian diseases for which there is no cure or effective therapy[27]. These diseases were believed to be monogenic, i.e., caused by LoF mutations in one gene. More than 4.2 million SNPs and INDELs have been identified in the human genome, of which around 176 000 are predicted to result in LoF[28–30]. However, recent large-scale sequencing and analyses of individuals have identified genetic variants that modify the onset and manifestation of some Mendelian diseases and, in some cases, buffer the deleterious effects of the responsible high penetrance LoFs and confer resilience in some individuals. This outcome of accumulated research further emphasizes the recent theoretical development in the modeling of disease heritability, suggesting that all complex traits of heritable diseases result from a sufficiently interconnected network of genes and hence share a single 'omnigenic' architecture. Consequently, the success of conventional GWAS (still the most common method of studying disease heritability) though limited, is yet to be matched in identifying variants that modify complex traits of Mendelian diseases[19].

My aim in this thesis is to annotate whole-genome sequencing data functionally, look at genomic variations in the aggregate, and investigate Mendelian diseases with complex onset and manifestation under the monogenetic paradigm. Therefore

1. I aim to construct an accurate GRN that encapsulates all relevant pairwise cis- and trans-regulatory interactions[31] in a genome, hence allowing epistasis modeling.

2. I aim to develop a tool for precise annotation of the whole genome that is the limiting factor in identifying druggable cellular functions that underpin the onset and manifestation of complex inheritable diseases.

3. I aim to recruit and sequence the whole genome of rare and ultra-rare Mendelian disease[32] patients with varying ages of onset/symptoms to identify known and

new potential disease-modifying genetic variants in the genome of the patients.

Chapter 2

Ordinary differential equations to construct invertible generative models of cell type and tissue-specific regulatory networks

2.1 Abstract

The most recent model describing complex symptoms of "Mendelian" diseases posits an omnigenic architecture for the heritability of these traits. Thus analyses of the underpinning biological functions resulting from the self-regulated system of all gene-gene interactions among the active genes of the affected tissues, rather than loss of function mutation on a few genes, gained momentum. As a practical oversimplification, to date, most attempts to construct a model delineating such interactions relied on integrated, high-throughput data that contained ubiquitous false positives, resulting in the reportage of burgeoning, inconsistent and spurious functional groupings. Additionally, intricate regulatory interconnections of the genome, specific to a tissue or a specific cell developmental stage, remain unsolved owing to the fact that the functionality of particular groups of genes is dynamically dependent on the cell life cycle/environment and is often impossible to infer from convoluted aggregated data. Nonetheless, advancements in RNA sequencing technologies allowed inferring tissue/cell developmental stage-specific Gene Regulatory Networks (GRNs), but an accurate assembly of this network remained a strenuous task. Here we developed the Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks (NS-DIMCORN) to address such drawbacks. The main advantage of NS-DIMCORN is constructing continuous distributions of gene expression from RNA-seq data that allow sampling of missing readouts to estimate the gene expression trajectories dependent on other genes. We systematically demonstrated that the proposed approach is scalable and compares favorably to the state-of-the-art algorithms in recovering genome-wide *ab initio* genetic interactions, whether from synthetic or empirical data. Overall, we put forward a path to contrive tissue-specific, directed, time-aware GRNs purely from data and without relying on expert knowledge or prior assumptions.

2.2 Introduction

Historically, GRNs are represented where genes are nodes, edges are pairs of genetic interactions, and networks are overlapping pairs. GRNs inference has used data from gene deletion (perturbation) screens and experiments that compared non-perturbed (healthy) versus perturbed (diseased tissues)[33]. The recent emergence of high-depth multi-sample (bulk RNA-seq) and single-cell RNA-sequencing (scRNA-seq) data allowed more accurate GRN inference by modeling dynamic RNA expressions for different tissue types or cells at different developmental stages (time points)[34]. Distilling informative *ab initio* genome-wide GRNs from whole-genome RNA expression is especially useful given that Next Generation Sequencing (NGS) can conveniently achieve statistical power using pseudo-bulk techniques and high-depth single-cell RNA sequencing[35]. NGS has already produced substantial tissue and cell-specific RNA-seq data that would allow comparing genetic interactions in different experimental settings (e.g., healthy/diseases), different cellular processes, and different developmental stages, all at tissue/single-cell level resolution[36].

In RNA-seq experiments, the expression level of transcripts is calculated from the number of sequenced reads that map to the codon responsible for those transcripts at the same snapshot[37]. Thus, Differentially Expressed (DE) transcripts can also indicate the direction and strength of their correlation with other genes and samples[38]. However, it should be noted that RNA-seq data from NGS still contains high technical noise, which can be exacerbated by sequencing-specific data features such as sample heterogeneity, variation in sequencing depth, and sparsity mapped reads[34]. Additionally, regulatory interactions are deemed far more interconnected than simple protein-protein or gene-gene interactions due to the discovery of various small Ribonucleic Acid (RNA) sequences that play active roles in the machinery that regulates cellular processes[39]. Nonetheless, as reviewed by Pratapa *et al.,* construct-

ing informative models that recapitulate the complete and accurate set of genetic interactions from RNA-seq data has been a very active area of research in the past decade[40].

Initial *ab initio* GRNs based on Boolean logic[41] have been successfully used to model high-level monotonic interactions of cellular mechanisms[42]. Even so, these models failed to capture cascades of complex events, such as promoter recognition and the dynamic self-regulatory protein translations across the entire genome, over the differentiating lifetime of a cell[43]. Therefore most recently, enhanced Boolean logic[44], regularized linear regressions[45], Bayesian networks[34,46], partial correlation, semi-partial correlation[47,48], Pearson correlation[49], tree[43], entropy[50] based approaches[51] and Gaussian graphical models[52] have all been utilized to model intricate interactions between genes and gene products (e.g., proteins), but accurate GRNs inference remains a challenging problem[40]. Briefly, enhanced Boolean logic, regression, and correlation-based approaches fail to capture higher-order and more complex gene-gene relationships (e.g., non-additive interactions)[41]. Though such interactions can be elucidated using mutual information entropy, they require homogeneous data or hyper-parameter re-tuning to avoid overestimating interactions' significance[53]. Spanning tree-based approaches can be used but are computationally expensive, extremely sensitive to changes in data, and inadequate for predicting continuous values as observed in empirical data[54]. Traversing Bayesian networks may also be used, but calculating the conditional probability of edges is not a trivial problem when a large number of interactions are involved, and these networks are only suitable for steady-state data in their vanilla form[34,46,52]. Finally, Gaussian graphical methods depend on the Gaussianity assumption, which also implies linear dependencies between genes; additionally, most implementations of this approach to date are incapable of handling directional interactions as in cell differentiation[48].

Another promising approach for inferring GRNs from time-stamped data (like

RNA-seq) is borrowed predominantly from other fields of natural sciences (e.g., Physics) that have a long history of studying dynamic systems and lending to representing the dynamic process of a cell using systems of Ordinary Differential Equations (ODEs)[55]. This method has been demonstrated to generate more realistic behavior and better describe genuine regulatory relationships of genome-wide interactions in the cell[36,52]. In ODE-based methods, derivatives of differential equations describing the system can be estimated by difference approximation[56,57] or regularised differentiation[58], and then be solved by linear methods[59] by fitting a mechanistic of nonlinear functions[55,56] or nonparametric techniques[60]. In both of these approaches, the main constraint of solving ODEs for a biological system with tens of thousands of genes is their high-dimensional parameter spaces that require larger sample sizes that can rapidly become computationally intractable as the addition of new samples adds up[34]. Here we demonstrate that this problem can be dealt with satisfactorily by

1. utilizing a highly flexible and scalable method that can model functional relationships of dynamic data
2. bypassing the error-prone derivative estimation, and
3. using a nonrestrictive and scalable model architecture that allows cheap computation.

Therefore, we developed the Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks (NS-DIMCORN) that allows unrestricted neural network architectures (i.e., arbitrary depth increase) and training the model without partitioning or ordering the data dimensions. NS-DIMCORN yields a continuous-time invertible generative model with unbiased density estimation by one-pass sampling, allowing scalability and end-to-end training of larger ODEs-based models (Figure 2.1). Furthermore, NS-DIMCORN only requires scRNA-seq read count data as an input, and time points are automatically inferred by estimating probability distributions

for the continuous gene expression trajectories instead of probability distributions for the derivatives. This allows easy sampling of the continuous trajectories using Hamiltonian Monte Carlo and calculates nonlinear gene dependency based on conditional Mutual Information (MI)[61]. To this end, we demonstrated that NS-DIMCORN, on average, outperforms other state-of-art algorithms[47,49–51,57,62–65] in inferring GRN from synthetic, bulk, and single-cell RNA-seq data.

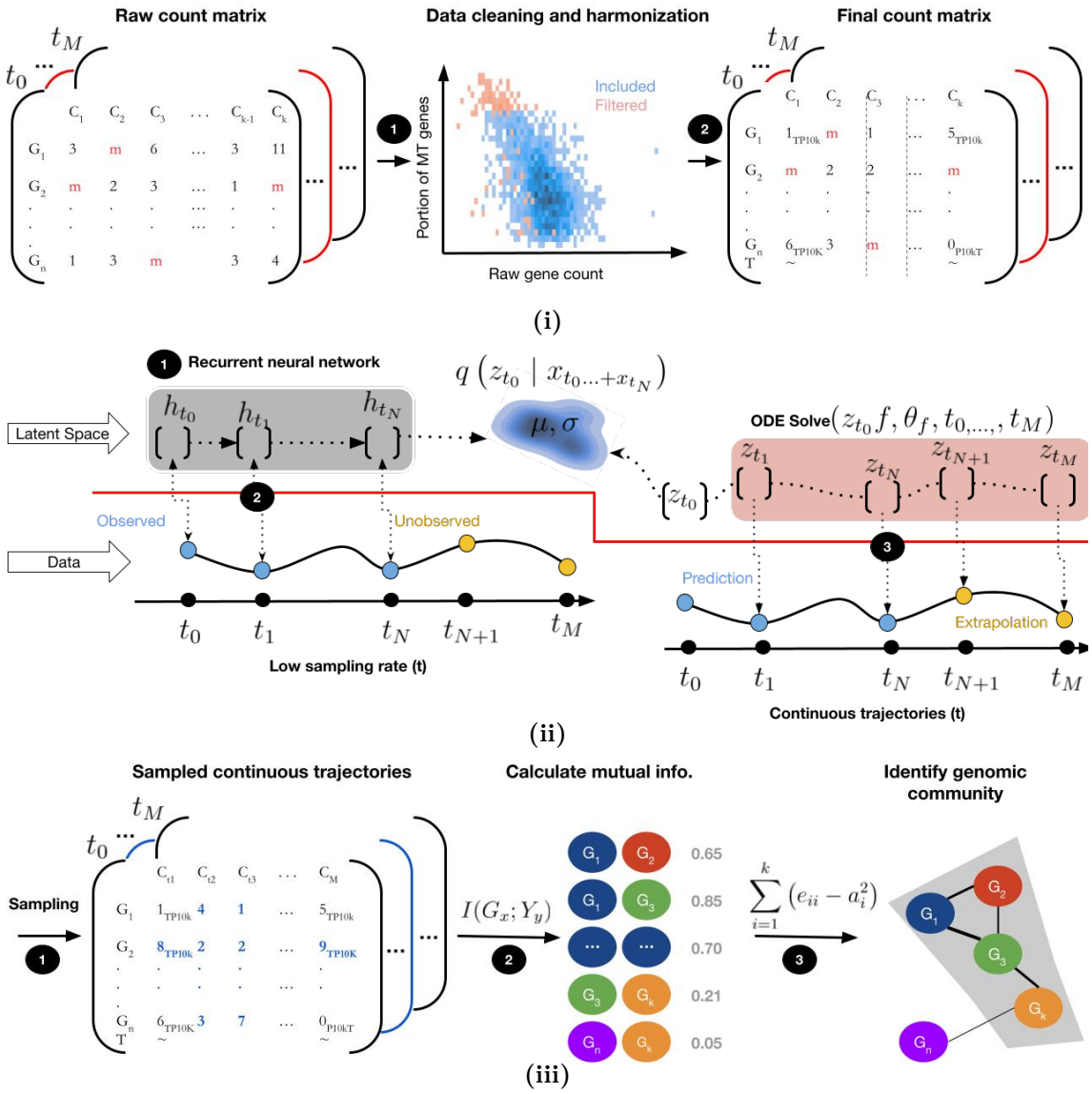


Figure 2.1 | Overview of NS-DIMCORN. i: Data cleaning and harmonization where mitochondrial genes (MT), cells with low gene count, batch effects, and genes with low depth were removed. ii: training a neural network for bijective mapping from a latent trajectory of learned continuous dynamics to data. iii: sampling and constructing sample-specific GRN's for community detection, from TP10K as defined at Section 2.5.2.

2.3 Results

2.3.1 Overview of the algorithm

Given a set of irregularly sampled (missing reads indicated in red in Figure 2.1i) scRNA-seq data for a specific tissue or cells, the goal of NS-DIMCORN is to model gene expression across cellular process trajectories (i.e., cell lineage differentiation trajectories). To this end, the read counts for all the scRNA-seq samples are normalized (Figure 2.1i-1); samples with spurious or low-quality reads are removed; counts are standardized, and the data set is harmonized to remove any confounding variables such as batch effects (Figure 2.1i-2). NS-DIMCORN represents different cell states by continuous latent trajectory (Figure 2.1ii-1) and defines a bijective map from the latent learned latent space to data by integrating latent variables (Figure 2.1ii-2). Latent trajectories are computed by solving an initial value problem by an ODEs solver that is parameterized by a neural network and a given initial state, z_{t0}. The output of the last layer of the neural network is the solution to the initial value problem, the hidden units are parameterized as a continuous function of time, and the parameters of nearby "layers" are automatically tied together. This learned model then, in turn, allows continuous sampling of reading counts even for missing states/genes (blue states and reads) arbitrarily far forwards or backward in time (Figure 2.1iii-1). Continuous sampling from cell states allows accurate estimation of conditional mutual information between each set of gene pairs as *ab initio* (translated to weights of edges between two genes) and signs of co-variances as an indicator of interaction directions in the inferred networks of tissue-specific gene-gene interactions (Figure 2.1iii-2).

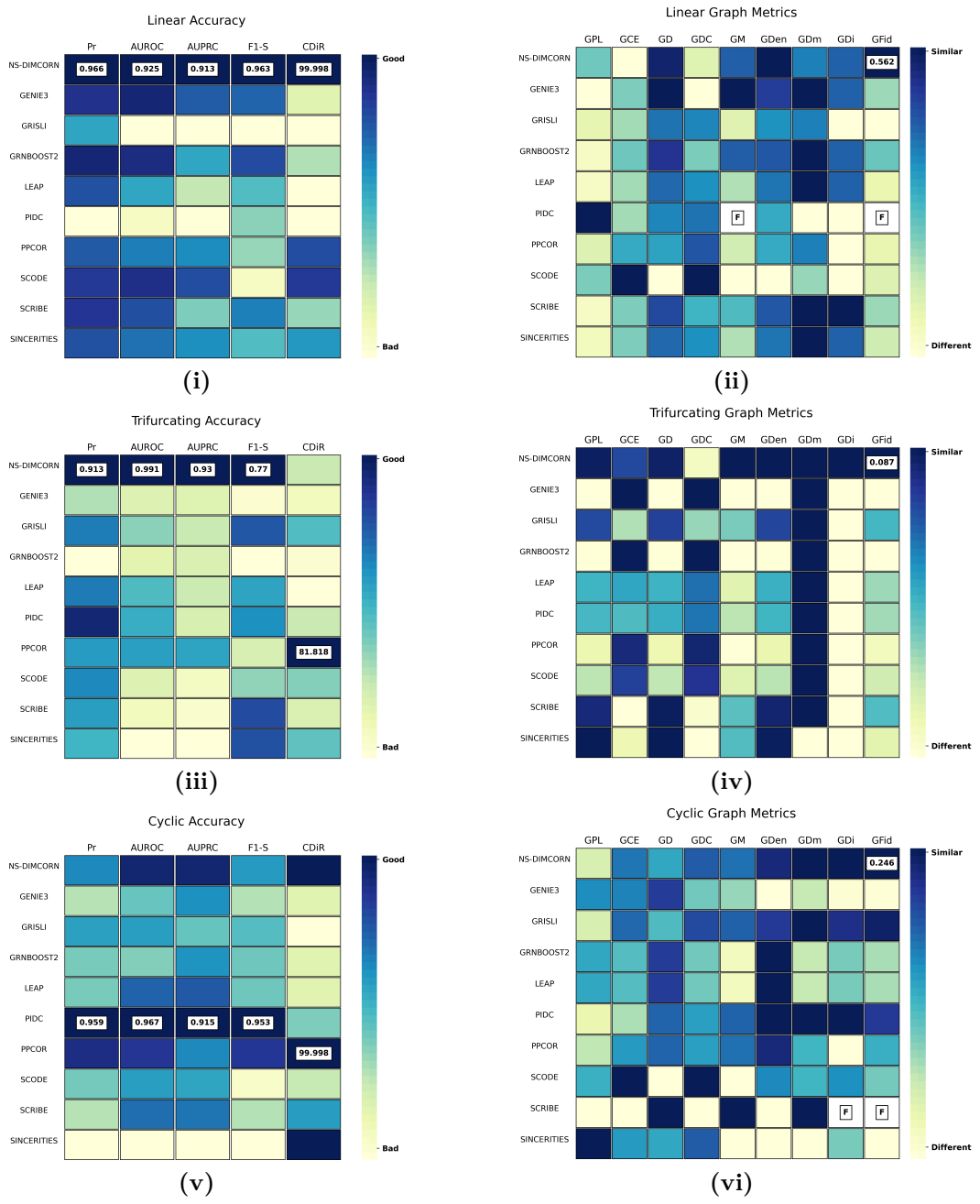


Figure 2.2 | Comparison of different algorithms when inferring GRNs from synthetic data. In i, iii, and v, the six measurements (columns) represent a measure of accuracy (as described in the methods section) for each algorithm (rows). In ii, iv, and vi Absolute Minkowski distance of eight fundamental graph metrics between the ground truth network and inferred network and the Jensen-Shannon divergence of Laplacian spectra of the two graphs (GDi) are plotted. Column values for are panels are normalized, where blue represents better accuracy/faithfulness than yellow. Blocks with the highest values and more than three significant figures of difference are annotated, and F indicates an undefined number. Overall, the heatmaps presented show NS-DIMCORN stays faithful to the ground truth network topology and demonstrates the best accuracy for simulated long linear and trifurcating development processes. PIDC slightly outperforms NS-DIMCORN in the sense of accuracy by less than 1 percent but NS-DIMCORN constructed the most accurate GRN topologically.

2.3.2 Benchmarking against simulated data

Simulated data for Long Linear (Li), Trifurcating (Tr), and Cyclic (Cy) cellular processes trajectories

In a closed dynamic system, functional changes result from the interchange of information and interactions between constructional elements of the system (i.e., structural elements can switch each other on/off or self-regulate to steer the fate of that system)[66]. In the context of GRNs, nuanced and continuously evolving changes of these dynamic systems can be surveyed employing network topological features, which can elucidate the type and strength of pairs and groups of interactions between genes in a network[67]. Many metrics have been suggested to study a network's topology and estimate the structural distance (similarity) between two networks[68].

Here we focused only on the most robust and commonly used measures of network topology[69], namely Graph Path Length (GPL), Graph Degree (GD), Graph Modularity (GM), Graph Diameter (GDm), and Graph Clustering Coefficient (GCE) as defined in the Methods (Section 2.5.8). For ease of reporting, we equipped the vector space of these metrics with a norm–as opposed to using the topology metrics directly; and reported the Minkowski distance of each one of these metrics for the inferred network and their respective Ground Truth Networks (GTNs) in synthetic examples and Known to be Truth Sub-Networks of experimental data (KTSNs) only including interactions experimentally validated interactions for empirical data (Section 2.5.7). In addition to topology metrics, it has been demonstrated that the Laplacian eigenvalues of graphs also capture the local and global properties of networks[70]; therefore, we further included Graph Distance (GDi) metrics which capture the structural distance of the compared networks in terms of the Jensen-Shannon distance of graph Laplacian matrices (Section 2.5.8). Finally, combining all the proposed metrics, we calculated Graph Fidelity Distance (GFid) that nor-

malizes and averages measurements in more than one graph (Section 2.5.8)[69]. Using these considerations, we compared ten tools for inferring *ab initio* GRNs, including NS-DIMCORN. We sought to determine if GRNs inferred by an algorithm from a database would exhibit dynamics identical to the known underpinning network. To this end, we executed each algorithm on pre-processed expression data as described in the Methods using the recommended hyper-parameters from the original publications[47,49–51,57,62–65]. This step resulted in a ranked edge list of the inferred network connections (edges). After removing all the self-loops, the inferred networks were analyzed for faithfulness to the true network and accuracy as described in the Methods (Sections 2.5.7 to 2.5.8), we then reported two sets of statistics indicating the performance of each given algorithm.

To avoid the pitfall of technical and instrumental noise in scRNA-seq data[40,71,72], and for ease of interpretation[73], we first focused on synthetic data with a known GRNs that could serve as the ground truth (Figures 2.2 to 2.3). This initial focus also allowed us to avoid any limitations of pseudo-time dependent inference that could potentially affect our benchmarking results, especially for GRISLI[62], LEAP[49], PPCOR[47], SCODE[64], SCRIBE[65], and SINCERITIES[51] which required pseudo-times provided separately as input. Three different temporal trajectories here, namely Linear, Cyclic, and Trifurcating, are constructed as described in BoolODE package[74] represent the different possible dynamic cellular processes[40], whether the trajectories relate to metabolism, cellular reprogramming, reproduction, differentiation, or apoptosis through cell developmental stages. As for the linear trajectory here, we included a long cascade of intermediate genes to attain enough complexity but ensured the linear trajectory still resulted in one distinct final steady state for each initial state.

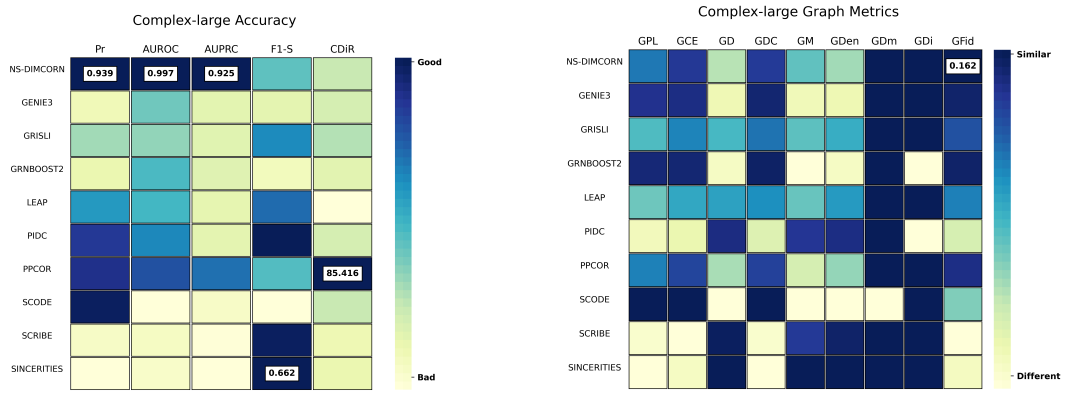
We observed that GRNBOOST2[63] and SCODE identified the highest number of genuine regulatory interactions for long linear trajectories. At the same time,

PIDC failed to identify any interaction owing to its approach to calculating mutual information between genes. However, GRNBOOST2 and SCODE, notably SCODE, showed low discrimination power, resulting in many false positives, spurious interactions, and consequently lower AUROC and AUPRC –indicators of discrimination power (Figure 2.2i). For the same linear trajectory, NS-DIMCORN only misclassified one of the authentic regulatory interactions but with far fewer false positives than other methodologies and inferred the most accurate regulatory network, with the highest AUROC and AUPRC. The non-perfect Direction scores (CDiR) for NS-DIMCORN also resulted from the misclassified genuine interactions. Otherwise, the correct direction was inferred for every identified genuine interaction (Figure 2.2i). Expectedly, topology analysis of the inferred graph confirmed the above model accuracy metric, and NS-DIMCORN showed the highest fidelity toward the ground truth. We hypothesized that the unanticipated high difference in GCE (see Methods for metric definitions) for NS-DIMCORN can be attributed to the fact that the clustering step of NS-DIMCORN creates artificial groupings within the harmonized structure of the linear trajectory and hence reduces the statistical power of the NS-DIMCORN inference. Indeed omitting the clustering step improved the performance of the NS-DIMCORN for the linear trajectories (Supplementary Table S1), but we believe a larger sample size would be a less error-prone approach for resolving this issue, primarily when cell trajectories are not known or suspected to be complex (Figure 2.2ii).

We also looked at trifurcating trajectories of cellular processes where mutual regulation motifs involving more than one gene result in a few distinct steady states from common initial states. As illustrated in Figure 2.2iii, benchmarking NS-DIMCORN against the nine other algorithms for trifurcating trajectories demonstrated the highest precision and superior discrimination power of NS-DIMCORN, based largely on the highest AUROC (Section 2.5.6) and AUPRC (Section 2.5.6).

Nonetheless, NS-DIMCORN was mostly unsuccessful in identifying the direction of trajectory interaction. This was because of the arbitrary Cartesian coordinate of each final steady state. Indeed NS-DIMCORN intrinsically would not have any notion of the origin and would also not capture this information in the latent space. PPCOR, on the other hand, was able to identify directions better, although for the fewer correctly identified genuine interactions, using partial direction correlation between two pairs and relying on the direction of pseudo-times (Figure 2.2iv).

NS-DIMCORN has struggled mostly with oscillatory circuits that yield linear trajectories where the final state coincides with the initial state. This behavior was mapped out here in the synthetic cyclic data with zero steady-state. NS-DIMCORN was only the second-best inference algorithm based on AUROC as well as AUPRC and demonstrated lower precision than PIDC and PPCOR. Circularity introduced by the absence of initial/final temporal distinction among cells most likely underlies this lack of performance. Regardless, NS-DIMCORN still successfully captures the actual topology of the original graph better than all the other methods studied here (Figure 2.2v).



(i) **(ii)**

Figure 2.3 | Comparison of different algorithms inferring GRNs from cells with complex dynamics (linear, trifurcating, and cyclic) with large sample sizes and a high number of active genes dynamic.

Simulated data for Compound-Complex (CC) cellular processes trajectories:

Networks comprising only single-mode trajectories conveniently allow identifying strengths and shortcomings of an inference method in isolation but are hardly representative of actual biological RNA-seq data[75]. Thus, we combined three datasets with linear, cycling, and trifurcating trajectories to simulate more real-world-like, complex, and large datasets. In addition, this larger dataset tests the scalability of each inference method as it requires each algorithm to consider three times more cells while inferring the underpinning GRN of the dataset. As expected, NS-DIMCORN successfully captured the actual topology of the original graph and inferred the most accurate regulatory network, with the highest AUROC and AUPRC (Figure 2.3i). The higher F1-S for SCRIBE and SINCERITIES indicated a high data imbalance when considered with extremely low Pr (Section 2.5.6), AUROC, and AUPRC scores. Higher modularity (Section 2.5.8) in the inferred GRN from these methods suggests that SCRIBE and SINCERITIES focus on only part of the data, which reduces their overall score (Figure 2.3). For the reasons mentioned above, the PPCOR algorithm was better at identifying directions of correctly identified genuine interactions, as it was observed for cyclic trifurcating trajectories.

2.3.3 Benchmarking against empirical data

We did not expect that the relatively simple rules used by BoolODE generating synthetic data, even in their complex mode, would sufficiently mimic the properties of the real biological data. So we sought to determine if a GRN inferred by an algorithm from an empirical dataset would exhibit dynamics and steady states identical to the original underpinning network. Empirical RNA-seq data is often categorized into three distinct groups (steady-state, bulk, and single-cell sequencing data)[76]. The steady-state data refer to the expression level of genes after introduc-

ing gene knockouts or essential gene perturbation under the assumption that only meaningful pairwise interactions exist[77]. Even after excluding non-characterized transcriptomes, this approach requires more than 200 million observations to assay all pairs of genetic interactions for the remaining $\approx 20,000$ human genes[78]. At the time of writing that paper, the largest dataset of this type only included around 0.1% gene pairs[79] of possible interactions.

The other two types of data, namely Bulk RNA-seq data and scRNA-seq data, provided sample data sets comprising snapshots of different cellular process states within different cells or tissues[80]. Consequently, these data can represent the dynamic interactions of genes and cellular process trajectories beyond genetically modified or chemically perturbed, which is the focus of this study[76]. While scRNA-seq data does not suffer the loss of information inherent to averaging processes in bulk RNA sequencing methodology[80], it is more noise-prone[81].

To establish that NS-DIMCORN is scalable but at the same time is also sensitive enough to detect nuanced dynamics specific to different tissues in the human body among the noise, we included a larger dataset of brain cells with more specialized cell types and a smaller dataset of heart cells. Empirical ground truths mainly rely on prior expert knowledge and, due to their binary and agglomerative nature, do not encapsulate signal type, direction, or time dependency of the functional genomics nexus they describe[82,83]. Moreover, pleiotropic effects of genes and noisy high throughput data used for the curation of these networks, combined with the subjective allocation of genes to a network, cause truth networks to differ in number and type of genes[84] in different studies and may include many spurious interactions. To alleviate the effects of false positive or false negative interactions from empirical ground truth data sets on our experiment, we constructed a ground truth network comprising only experimentally validated Transcription Factors (TFs) and genetic interactions that are involved in essential and well-characterized metabolic pathways

(Section 2.5.1).

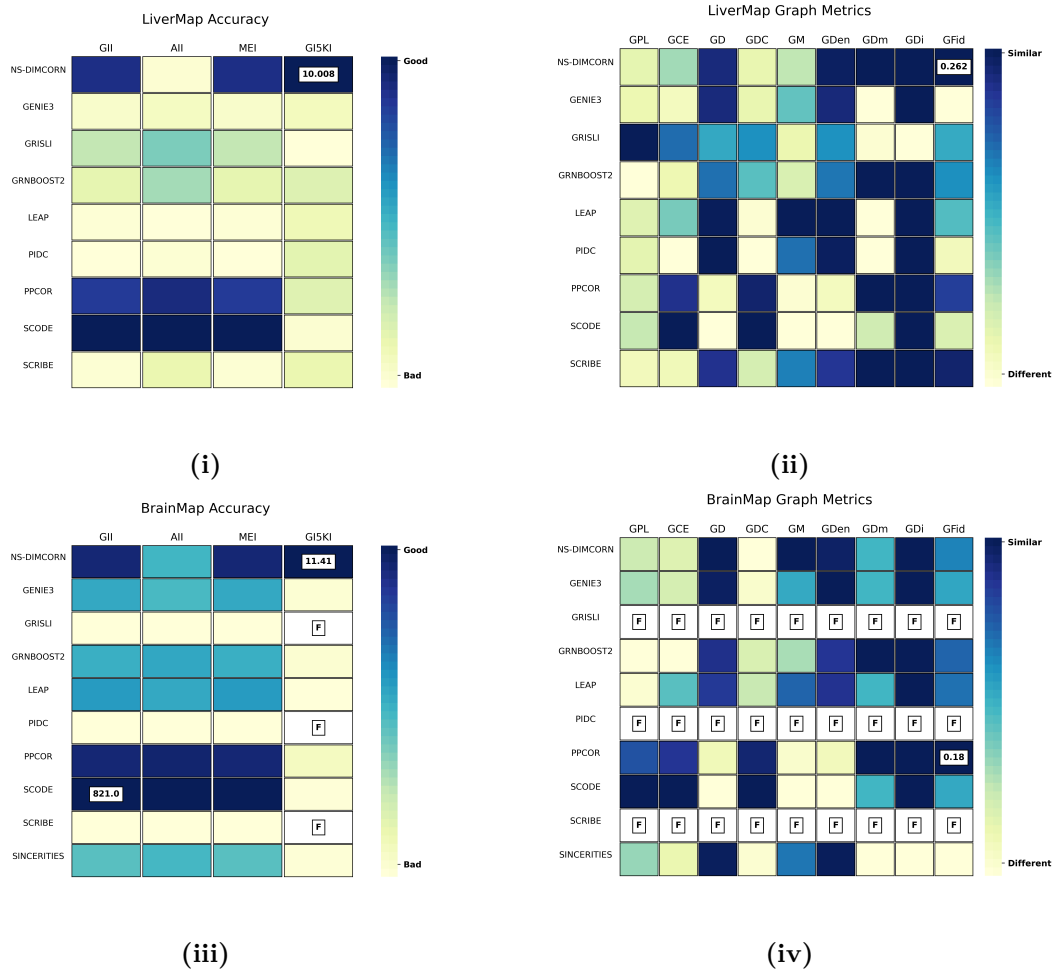


Figure 2.4 | NS-DIMCORN is scalable and identifies more known genuine metabolic interactions than any other algorithm in brain and liver single-cell sequencing data. a;c, The four measurements (columns) represent a measure of accuracy for each algorithm (rows) where only true positives are partially known. b;d, Absolute Minkowski distance of eight fundamental graph metrics between sub-networks of partially known genuine metabolic interactions and inferred networks in addition to Jensen-Shannon divergence of Laplacian spectra of the two graphs (GDi) are measured. Column values are normalized, darker blue represents better accuracy, and lighter yellow represents lower accuracy. Blocks with the highest values and more than three significant figures of difference are annotated, and F indicates an undefined number.

NS-DIMCORN inferred network from sc-RNA-seq data is specific:

In the liver, SCODE identified the highest GII (Section 2.5.7) compared to its respective KTSN (Section 2.5.1) and many other interactions. The high number of interactions identified might indicate many false positives; therefore, a worse T5KI (Section 2.5.7) score was given to SCODE (Figure 2.4i). PPCOR more or less showed the same behavior as SCODE, which is in contrast to NS-DIMCORN, which attained a better GII score and showed good sensitivity indicated by low AII (Section 2.5.7) and high T5KI (Section 2.5.7). Comparing the topological fidelity distance (Section 2.5.8) of the KTSN and the inferred GRN again demonstrated the superior performance of the NS-DIMCORN. GM (Section 2.5.8) and GD (Section 2.5.8) scores are also consistent with our rationale that SCODE connected most of the genes in the network and lost the modular texture of real GRN (Figure 2.4ii).

In the brain, NS-DIMCORN again achieved the highest GI5KI but showed a more significant topological distance to the KTSN than PPCOR. Graph path length (Section 2.5.8) and gene centralities appeared to contribute most to this observation (Figure 2.4iii, Figure 2.4iv). The extra interactions identified by PPCOR appear to be non-overlapping with the KTSN, but the same genes, in general, having more interactions as indicated by graph clustering coefficient and greater reach regardless of their authenticity would explain this observation. It is noted that GRISLI, PIDC, and SCRIBE failed to generate an output, given the size of the brain dataset and the number of genes involved in the network. GRISLI ran out of memory on a high-performance 72-core computer with more than 2TB of RAM, SCRIBE could not generate output in more than seven days on the same computer, and PIDC produced only undefined values for the output (Figure 2.4iii, Figure 2.4iv).

NS-DIMCORN is scalable and allows specificity for bulk RNA sequences and aggregated data:

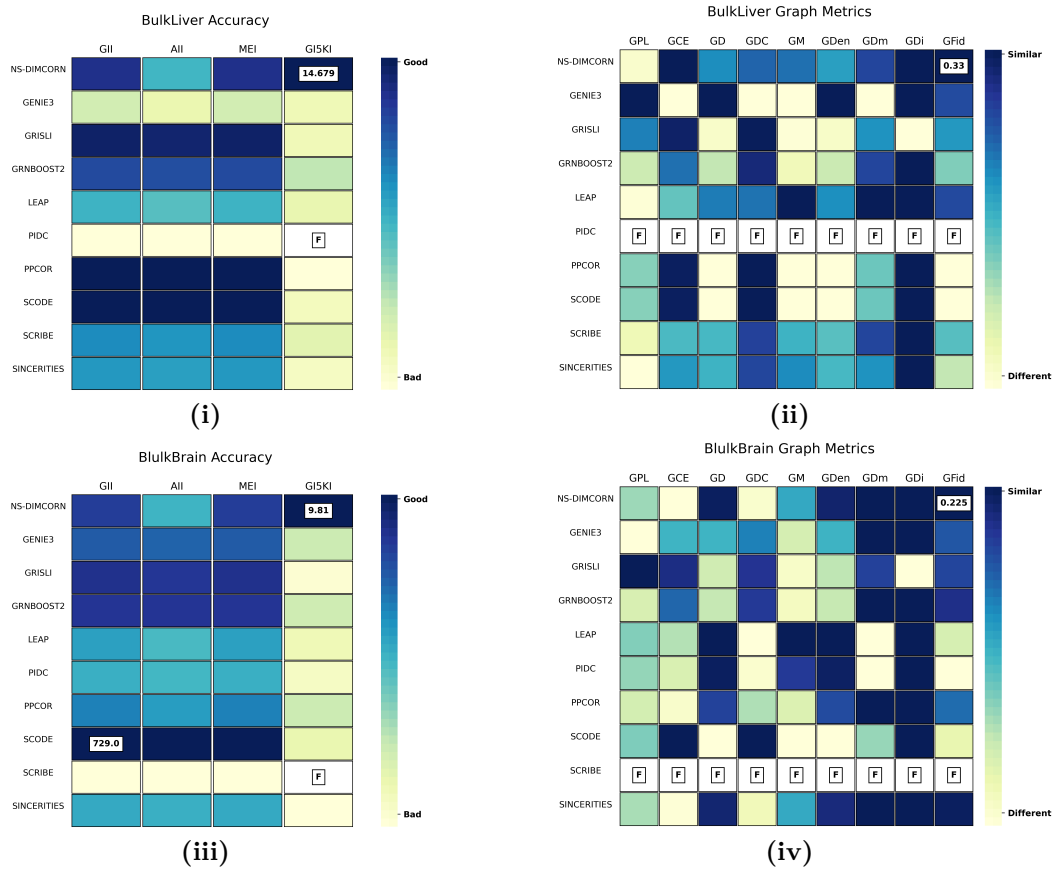


Figure 2.5 | NS-DIMCORN is less sensitive to diffident cell types in the same tissue in comparison to different but similar tissues.

For the bulk RNA-seq data, despite the fact that the brain has more specialized cells, only a few people with different sub-brain tissue information were obtainable. Liver samples were of a larger dataset, although the number of individuals sequenced for the liver data was still relatively small compared to the number of cells in the scRNA-seq dataset. Our rationale for choosing bulk RNA-seq data was to investigate if NS-DIMCORN can still successfully infer the best network when the data is highly aggregated–bulk, many genes are involved–the brain and only partial data exist–

few sub-tissues. Our observations were largely in accordance with the result from scRNA-seq data. The only difference is that we recovered a more accurate topology for the bulk brain, which is consistent with the literature stating bulk RNA might still be a better choice for identifying interactions with smaller effect sizes[80]. We also observed SCRIBE was the only algorithm that failed on the brain data, which suggests, as opposed to GRISLI and PIDC, the algorithm performance is dependent on the network size (Figure 2.5iv).

2.4 Discussion

We excluded four algorithms that we included initially in this study, namely SCRIBE[65], SINGE[85] GRNVBEM[46] and SCNS[44] given that they failed to produce an output for most or all of the datasets studied here. Speed, memory, or inflexible implementation were the main drawbacks of these methods on our high-performance computer with 2TB of RAM, 72 core CPUs, and 4 Nvidia A100-80GB GPUs. Although these specs are vastly better than an ordinary desktop computer, not every method could achieve concurrency or efficiently utilize all the provided computation resources, so the same result would likely have been obtained on low-spec computers. We did not attempt to optimize the run time of any of these methods and terminated any process after a week if no output was produced. Of these methods, NS-DIMCORN was the only method cable of utilizing GPUs for array operations and model training, being entirely based on TensorFlow[86] and cuPy[87] (a GPU based implementation of NumPy). We observed that including 2500 genes for around 8000 cells during training, the NS-DIMCORN model requires roughly thirty hours to model the data. However, our method's run-time and memory usage heavily depend on the number of genes and samples used as input.

NS-DIMCORN primarily relies on DESC for clustering the related cells before training the generative model, but this might introduce some limitations when study-

ing linear cellular trajectories. DESC originally initialized clustering regions using the Leiden clustering algorithm[88] and then optimized the clusters by stochastic gradient descent[89]. Here we swapped Leiden clustering for DensMAP, believing UMAP better preserves the fine details of the data manifold[88]; therefore, more lenient clustering configurations for UMAP would improve NS-DIMCORN performance for linear data[90].

Regulatory interaction between genes in different cells can be statistically defined by information theory[53]. Although most gene pairs satisfy linear or monotonic relationships, Mutual information is often used as a generalized correlation measure[91,92]. It has been suggested that bi-weight mid-correlation transformed via the topological overlap transformation is a more robust correlation measure and attains better accuracy in identifying interacting gene sets in terms of Gene Ontology enrichment[91]. We believe this assumption naively overlooked the implicit reliance of the used MI methodology on the local uniformity of the underlying joint distribution, which is not the case for strongly dependent variables such as gene expression level[50]. NS-DIMCORN uses covariance to identify genes interaction direction and Non-parametric entropy estimation to compute the degree of this correlation. The basic idea behind non-parametric entropy estimation is that locally estimating the log probability density at each data point, then averaging these estimates, allows accurate estimation of MI between two strongly dependent variables[93].

In summary, we presented the Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks (NS-DIMCORN) and systematically evaluated NS-DIMCORN with synthetic data representing different cellular trajectories, bulk, and scRNA-seq data from different tissues and sample sizes. We demonstrated NS-DIMCORN scalability due to its unrestricted neural network architectures and showed its superior performance compared to the state-of-the-art algorithms for *ab initio* GRN inference based on cellular trajectories. We showed that not only does NS-DIMCORN

estimate the chronological ordering of the cellular trajectories unsupervised from the data, but it also offers high sensitivity and specificity due to its invertible generative model that allows unbiased density estimation using continuous sampling.

2.5 Methods

2.5.1 Datasets

Empirical data

For single-cell sequencing data, the Allen brain map dataset[94] comprising 76,533 total nuclei from the primary motor cortex of two coronal post-mortem human brain specimens was chosen. Those authors have described details of performed DNA-seq sequencing and preliminary data processing steps such as case inclusion criteria, nucleus dissociation/sorting, and RNA-sequencing methodology (barcode extraction, mapping, alignment, filtering and annotating BAM file with gene tags)[94]. We obtained the raw gene expression count matrix as a CSV file and applied filtering and clustering as described later in (Section 2.5.2). To study the tissue specificity of our method, we also included the Human Protein Atlas dataset[95] that comprised 8439 total nuclei derived from parenchymal and non-parenchymal of fresh hepatic tissue of five human livers[96]. We obtained the raw gene expression count matrix that was prepared as described[96,97] and applied filtering and clustering specific to this study (Section 2.5.2). Bulk RNA-seq data in this study was obtained from the GTEx Consortium atlas[98] portal (dbGaP Accession phs000424.v8.p2). For all the individuals for which data from the liver and brain was available, we downloaded read counts for the RNA-Seq data and processed the data as described in more detail here (Section 2.5.2).

Known to be true sub-network of empirical data (KTSN)

We only included experimentally validated transcription factors (TFs)[99] and genetic interactions involved in essential and well-characterized metabolic pathways[100,101] in our ground truth as described earlier. We further filtered interactions of the network to "super pathways" only if they have been captured analogously in KEGG[100], Reactome[101] and WikiPathway[102], the three most cited datasets in published -omics studies[84] to indicate biological evidence.

Synthetic data

Expression data were simulated for 500 cells following linear, cyclic, or trifurcating two-dimensional projections by converting their Boolean GTN interaction matrix into noisy nonlinear ordinary differential equations described by Pratapa et al.,[40,103] before. Random Gaussian noise[104] was added to ensure the intrinsic stochasticity of the data was conserved in the simulated data[103].

2.5.2 RNA-seq data prepossessing

We obtained the count matrix (Allen brain, Human protein atlas dataset) and OMNI SNP Array Intensity files (GTEx brain and liver) and then read those files into an AnnData object with hierarchical data format[105] for the downstream processing. Genes detected in less than a threshold number of cells/tissue samples were not included due to the low sampling rate. This threshold was determined based on the average sequencing depth for each dataset so that the majority of high-confidence reads were retrieved (Allen Brain dataset = 100 cells, Human Protein Atlas dataset = 50 cells, GTEx brain = 25 samples, and GTEx liver = 10 samples). To further remove poor-quality cells, we calculated the total RNA read counts and the percentage of those counts relating to mitochondrial genes and then removed cells and samples without enough RNA reads ($3 \times$ Mean Absolute Deviation (MAD)), gene coverage ($4 \times MAD$), or a high portion of mitochondrial RNA ($3 \times MAD$)

(Figures 2.6 and 2.7).

To make sure all cells/samples have the same ratio of genes expressed for the deep learning model and other downstream analyses, sample counts for each cell were normalized by the total counts over all genes so that it sums up to 10×10^4 and then pseudo-log-transformed these values as follows:

$$log(\text{TP10k} + 1) = log((\text{transcripts}/1 \times 10^4) + 1) \tag{2.1}$$

While datasets that combine microarray data[106], the expression state of a large number of genes, are advantageous in achieving statistical power, samples from different batches that were obtained and prepared at varying locations or times (e.g., GTEx) or comprise a large number of cells (e.g., Allen Brain Map) suffer from non-biological experimental variation or "batch effects"[107]. Batch effects often impose serious computational challenges and result in spurious outcomes and conclusions; hence, to address this problem, we utilized DESC, an unsupervised deep embedding algorithm capable of removing batch effects that are smaller than the actual biological variation[89]. DESC also assigns cells with a more similar experimental setting into a soft cluster in an unsupervised manner[89]. For the DESC step, we adopted the default setting but only considered the maximum of 10 neighborhoods instead of 25 and incorporated densMAP community detection methodology for establishing seed clustering boundaries to better preserve the topological features of transcriptomic variability data[88,108].

GENIE3, GRNBoost2, PIDC and NS-DIMCORN account for cell sub-types and underlying cell state changes directly. For the rest of the algorithms that, in addition to RNA-seq data, require trajectory inference data as input, we inferred the progression of cells through geodesic distance along the coarse-grained map of the sequencing data manifold, based on the connectivity of manifold partitions [109] and provided these values to the algorithms as pseudo-times for mimicking real

sequential cell states (c,d - Figures 2.6 and 2.7).

In order to avoid mischaracterizing sub-tissues as the states of cells by NS-DIMCORN, we further refined and merged the soft clusters previously assigned by DESC into larger sets dependent on consensus cell-type-specific marker genes of the brain and liver. Namely, brain cells were combined into five distinct groups of astrocytes, endothelial, microglia, neuron, and oligodendrocyte[110]; and liver cells were placed into either of the cholangiocytes, blood, mesenchymal, epithelial, immune and bud-hepatic sub-cluster[96]. Fine-grained final cell clusters (e,f - Figures 2.6 and 2.7) used to profile the robustness of NS-DIMCORN also accounted for developmental cell stages G_1/S, S, G_2 and G_2/M, using well-characterized marker genes of essential cell cycle processes (DNA replication, chromosome segregation, and cell adhesion)[111].

We included only the most relevant genetic drivers of each GRN, thus increasing the power of downstream network inference algorithms, by identifying the set of most characteristic genes[78] of each tissue/sample type and ranked them based on their variances across each single-cell dataset. To further explain, after data standardization with a regularized standard deviation (i.e., z-score normalization per feature) and controlling for the relationship between mean expression and variability, the pseudo-log-transformed normalized variance of each gene was calculated as the variance of each gene; genes were then ranked by this variance[75], and only the most highly variable 2500 genes were included in downstream analysis.

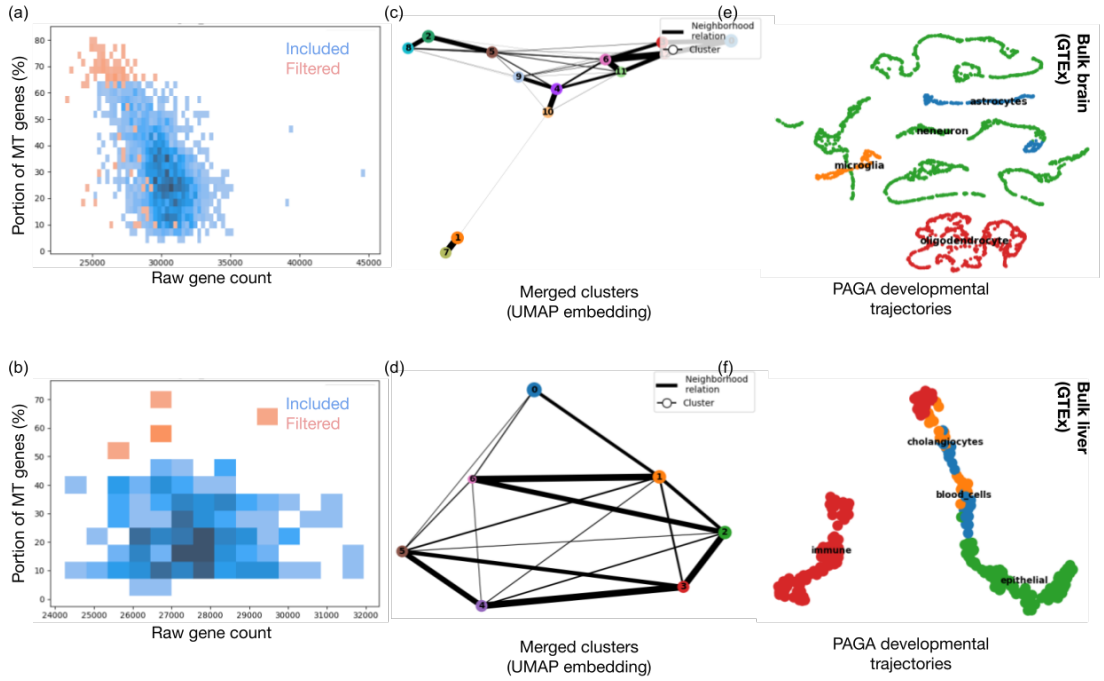


Figure 2.6 | Bulk RNA-seq data prepossessing: cells with less than a threshold RNA-seq reads count or cells with a high portion of mitochondrial genes were removed (a,b). Cells with high-quality reads were then clustered on the expression profile, and PAGA developmental trajectories were calculated (c, d). Obtained clusters were refined into merged sets of the same developmental stage and cell subtype(c, d).

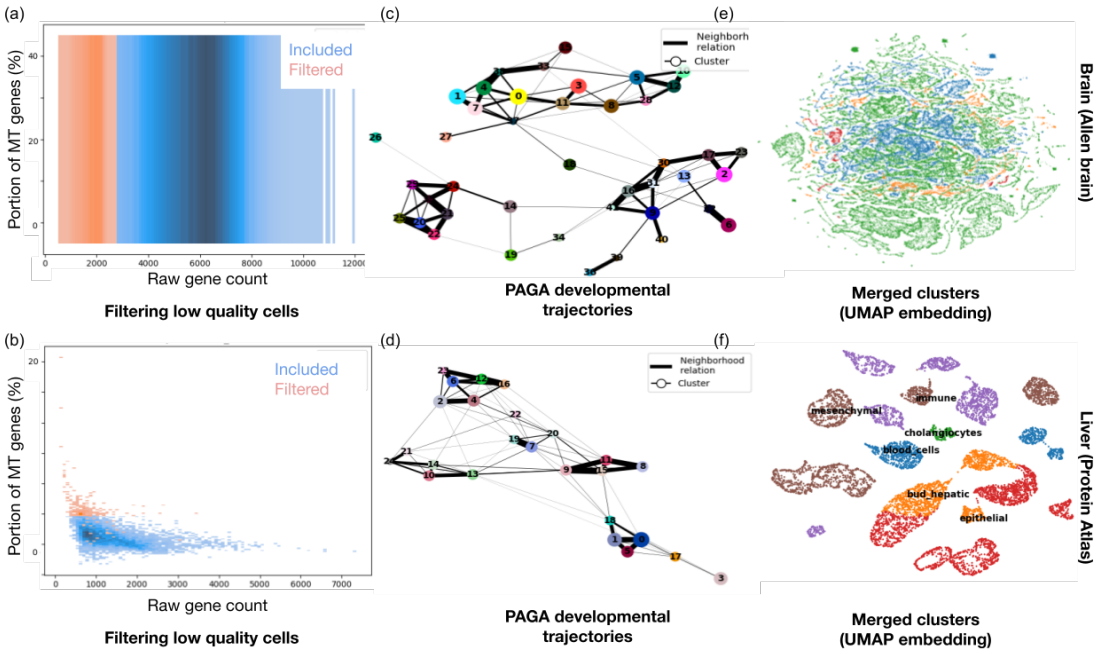


Figure 2.7 | Single cells RNA-seq data prepossessing: cells with less than a threshold RNA-seq reads count or cells with a high portion of mitochondrial genes were removed (a,b). Cells with high-quality readds were then clustered on expression profile, and PAGA developmental trajectories were calculated (c, d). Obtained clusters were refined into merged sets of the same developmental stage and cell subtype(c, d).

2.5.3 Estimating RNA-seq data real distribution

The main advantage of NS-DIMCORN is estimating continuous distributions of gene expression trajectories from RNA-seq data that allows sampling of the trajectory states between the available readouts as well as the observed results. We achieved the above by designing and optimizing an ODE network[112,113] that defines a continuous bijective map between vector field (latent variables, $\mathbf{y}$) to RNA-seq data $\mathbf{x}$, such that formally:

$$\mathcal{T}_\theta : \mathbf{x} \rightarrow \mathbf{y}, \mathcal{T}_\theta^1 : \mathbf{y} \rightarrow \mathbf{x} \tag{2.2}$$

Sampling from high dimensional space of the RNA-seq data would be computationally expensive or infeasible, hence given the invertible function above, instead of directly parameterizing the distribution of RNA-seq data, we specified the data distribution implicitly by warping a base distribution $\mathbf{Z} \sim p_{\mathbf{z}}(\mathbf{z})$, with an invertible (bijective) function. For RNA-seq data $\mathbf{x}_{n,m}$with m genes, n observations and $\mathbf{x} \in \mathbf{R}^D$ we chose $\mathbf{z}_0$ as the base distribution where $\mathbf{z}_0$ is a multivariate normal with $\mu = \{\mu_0, \cdots, \mu_m\}$ and $\sigma = \{\sigma_0, \cdots, \sigma_m\}$. Here μ_0 is equal to the normalized mean of the first gene from n observations, and σ_0 is its normalized variance among all the readouts for that gene. It should be noted that when the number of observations was less than 1280 (chosen based on the number of batches that could be fitted optimally on the available A100 Nvidia GPUs with 80 GB of memory) for a tissue/type of cell, in order to allow the model to converge with minimum fluctuation during the training, we augmented the dataset by generating new readouts within range of 0.1 (just an arbitrary choice) of variance of each gene for randomly selected RNA-seq observations.

Thus, if the parameterized continuous dynamics of genes, trajectories using invertible ODE parametric function were specified by a neural network:

$$\frac{d\mathbf{z}(t)}{dt} = f(\mathbf{z}(t), t; \theta) \tag{2.3}$$

For training the network, NS-DIMCORN firstly takes samples from the base distribution $\mathbf{Z}_0 \sim p_{z_0}(\mathbf{Z}_0)$; then solves the initial value problem below:

$$\mathbf{z}(t_0) = \mathbf{z}_0, \partial\mathbf{z}(t)/\partial t = f(\mathbf{z}(t), t; \theta) \tag{2.4}$$

for $\mathbf{Z}_{t1}$ using the Dormand-Prince explicit solver of non-stiff ODEs[114] given the observations in the RNA-seq dataset. To calculate the value of $\mathbf{Z}(t_1) \in \mathbf{R}^D$, the main challenge is computing the determinant of the Jacobian of the $\frac{\partial f}{\partial \mathbf{z}}$, which can restrict the architecture of the neural network used[115]. Here we use an instantaneous change of variables Equation (2.5)[112] as described by Chen et al. for this calculation that allows the gradients to be computed efficiently using the adjoint sensitivity method[112]:

$$\frac{\partial \log p(\mathbf{z}(t))}{\partial t} = -\operatorname{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) \tag{2.5}$$

$$\log p(\mathbf{z}(t_1)) = \log p(\mathbf{z}(t_0)) - \int_{t_0}^{t_1} \operatorname{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) dt \tag{2.6}$$

Also, to avoid the major pitfall of the RNA inference algorithms, instead of incorporating error-prone pseudo-time, we allowed the solver to choose t_{k-1} and t_k, the time between two different observation, which was then integrated over time in Equation (2.6) as stated in Equation (2.5).

Eventually for every gene expression readout, NS-DIMCORN computed $\mathbf{Z}_0$ that generates that readout as well its likelihood using:

$$\underbrace{\begin{bmatrix} \mathbf{z}_0 \\ \log p(\mathbf{x}) - \log p_{z_0}(\mathbf{z}_0) \end{bmatrix}}_{\text{solutions}} = \underbrace{\int_{t_1}^{t_0} \begin{bmatrix} f(\mathbf{z}(t), t; \theta) \\ -\operatorname{Tr}\left(\frac{\partial f}{\partial \mathbf{z}(t)}\right) \end{bmatrix} dt}_{\text{dynamics}}, \underbrace{\begin{bmatrix} \mathbf{z}(t_1) \\ \log p(\mathbf{x}) - \log p(\mathbf{z}(t_1)) \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix}}_{\text{initial values}} \tag{2.7}$$

Given that now we can efficiently calculate the Jacobian of $\mathcal{T} = \frac{\partial \mathcal{T}_\theta(\mathbf{y})}{\partial \mathbf{y}}$ as follows we can keep track of the deformations using the change of variable formula (Equation (2.8)), and transfer the notion of probability onto $\mathbf{x}$ and invert it again if needed[115] as follows:

$$\log p_{\mathbf{x}}(\mathbf{x}) = \log p_{\mathbf{y}}(\mathbf{y}) - \log \det \left| \frac{\partial \mathcal{T}_\theta(\mathbf{y})}{\partial \mathbf{y}} \right| \tag{2.8}$$

2.5.4 Co-variance Estimation

The inverse of the covariance matrix (precision matrix) is proportional to the partial independence relationship between matrix columns (genes). Under the assumption that only linear relationships exist between genes, if two genes are independent conditionally, all the corresponding coefficients in the precision matrix for those genes will be zero[116]. The covariance matrix of each sampled dataset can be calculated empirically. But inversion of the covariance matrix is computationally expensive and sometimes numerically impossible. Moreover, for high dimensional data or uncentered data samples, the precision matrix obtained from the inversion of the covariance matrix is not accurate (The Maximum Likelihood Estimator is not a good estimator of the eigenvalues of the covariance matrix). Consequently, estimating the precision matrix directly from data is the next best logical step[117]. To this end, we utilized a Hamiltonian Monte Carlo (HMC) sampler with adaptive step size[118] where the target log probability was a multivariate normal, parameterized by Chelosky factors of the precision matrix and a Wishart distribution as prior distribution (conjugate prior of multivariate normal). The full implementation of the

HMC sampler is described by MCMC using Hamiltonian dynamics paper[119] and its implementation is described elsewhere[120].

2.5.5 Mutual information (MI)

MI is a quantitative measurement of how much concurrent information exists about two variables. MI is a better measure of nonlinear interaction[92] and consequently a good candidate for measuring non-linear interactions in GRNs. For two genes (G_n, G_m distributed according to some joint probability density $\mu(g_n, g_m)$, where marginal densities of g_m is equal to $\mu_{g_m}(g_m) = \int dg_n \mu(g_m, g_n)$ and the marginal densities of g_n is equal to $\mu_{g_n}(g_n) = \int dg_m \mu(g_n, g_m)$ the MI is defined as

$$I(G_n, G_m) = \iint dg_n dg_m \mu(g_n, g_m) \log \frac{\mu(g_n, g_m)}{\mu_{g_n}(g_n)\mu_{g_m}(g_m)} \tag{2.9}$$

For a more efficient estimation of MI for strongly dependent variables (such are precursors of protein in a pathway), we tweaked the Kozachenko-Leonenko estimator[92] for local nonuniformity correction such that if $\mathcal{V}(i) \subset \mathbf{R}^d$ is the volume of k nearest neighbors of a sample point g^i in some space; we assumed that there is some subset, $\overline{\mathcal{V}}(i) \subseteq \mathcal{V}(i)$ with volume $V(i) \leq V(i)$ which density is constant as described by Geo et al., genes[93] .

$$\hat{I}_{LNC}(\mathbf{G}) = \hat{I}(\mathbf{g}) - \frac{1}{N}\sum_{i=1}^{N} \log \frac{\bar{V}(i)}{V(i)} \tag{2.10}$$

Thus, this correction term will improve the estimate of $V(i)$ for strongly correlated interactions.

2.5.6 Model accuracy metrics for synthetic data

For synthetic data, where we were sure about the majority of the actual interactions (besides noise) in a GRE, we evaluated the result of each algorithm using the following criteria. We assigned the edges in the relevant network the true positive label and ranked edges from each method as the predictions. Beforehand, we omitted all self-

loops given that some methods always assigned the highest rank to self-regulating genes. Some other methods, such as SINGE, and critically NS-DIMCORN, ignored them.

Precision (Pr)

The precision is the ratio

$$tp/(tp + fp) \tag{2.11}$$

where tp is the number of true positives and fp is the number of false positives.

The Area Under the Precision-Recall Curve (AUPRC)

AUPRC was calculated using the average_precision_ score function from sklearn[121], which summarises the precision-recall curve as the weighted mean of precisions achieved at different thresholds.

The Area Under the Receiver Operating Characteristic (AUROC)

AUROC was calculated using the roc_auc_score function from sklearn[121]. The Receiver Operating Characteristic (ROC) curve is created by plotting the true positive rate vs. the fraction of false positives out of the negative false positive rate at various threshold settings. A receiver operating characteristic curve, or ROC curve, illustrates the diagnostic ability of the classifier as its discrimination threshold changes. AUROC varies between 0 and 1, with 0.5 being an uninformative model.

Balanced F-score (F1-S)

F1 scores compute the harmonic mean of precision and recall so that

$$\frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})} \tag{2.12}$$

where tp is the number of true positives fn is false negatives and fp is the number of false positives. An F1 score of 1 is the best score and 0 is the worst possible F1

score.

Correct Directions (CDiR)

CDiR is the portion of inferred interactions that also showed the direction of interaction between genes correctly in the biological sense this can be interpreted as inhibitory vs. excitatory regulation.

2.5.7 Model accuracy metrics for empirical data

Given that we could not fully establish which gene-gene interactions are false positives, true positives, and false negatives for empirical datasets, instead of the conventional accuracy metrics introduced above, we adopted a new set of modified accuracy measures for the bulk and the single-cell sequencing data. AUC and AUPRC estimates are sensitive to noisy data[122] yet are linearly related to observed accuracy[123] and it is also proven that both are closely related to the Wilcoxon test of ranks[124,125]. Using the above line of reasoning, we assumed that we could extrapolate the accuracy of a model from a portion of identified true positives and false negatives. Hence for predictor model f, we defined True in the Top 5000 Identified Interactions (T5KI) as an unbiased surrogate metric of the model's discrimination and calibration. T5KI can be interpreted like the Early Precision Ratio (EPR) metric that was previously used for benchmarking GRN inference algorithms[40].

All Interactions Identified (AII)

AII counts the number of all the identified interactions.

Genuine Identified Interactions (GII)

In the context of empirical data, AII measures the number of identified interactions identical to those in our curated super pathway dataset.

Missing Expected Interactions (MEI)

MEI is the offset of unidentified interactions from the supers pathway dataset.

Genuine Interaction in the Top 5000 Identified Interactions (T5KI)

T5KI measures the number of AII only in the top-ranked 5000 inferred interactions while incurring the penalty for missing known-to-be true interactions (MEI). Formally

$$T5KI(f) = \frac{\sum_{t_0 \in \mathcal{M}^0} \sum_{t_1 \in \mathcal{M}^1} \mathbf{1}\left[f\left(t_0\right) < f\left(t_1\right)\right]}{|\mathcal{M}^0| \cdot |\mathcal{M}^1|} \tag{2.13}$$

here $\mathbf{1}\left[f\left(t_0\right) < f\left(t_1\right)\right]$ denotes an indicator function which returns 1 iff $f\left(t_0\right) < f\left(t_1\right)$ otherwise return $0; \mathcal{M}^0$ is the set of negative examples such as MEI, and $\mathcal{M}^1$ is the set of positive examples such as AII.

2.5.8 Graph topology metrics

For the strictly defined weighted Graph G with sets of vertices $G \subseteq \{v_1, ..., v_n\}$, n nodes and m edges

Graph Path Length (GPL)

is the normalized sum of path lengths $d(s,t)$ between all pairs of nodes, and it measures the efficiency of information flow for a network. Here we reported the averaged PL of each sub-graph when disconnected graphs were observed.

$$a = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \tag{2.14}$$

The Averaged Degree (D)

is the averaged number of adjacent edges to nodes, while Averaged Degree Centrality (DC) is the portion of nodes connected to each node. The Density (Den) of network G is defined as

$$d = \frac{m}{n(n-1)} \tag{2.15}$$

Together, D, DC, and Den were used as a surrogate for the first moment and second moment of the Degree distribution, which indicates the number and size of the network hubs.

Graph Modularity (GM)

uses Clauset−Newman−Moore greedy modularity maximization and calculates the strength of divisions of the network into clusters.

Graph clustering Coefficient (GCE)

has been calculated as the geometric average of the sub-graphs normalized edge weights and measures the degree to which nodes in a graph tend to cluster together. Formally

$$c_u = \frac{1}{\deg(u)(\deg(u)-1))} \sum_{vw} \left(\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw}\right)^{1/3} \tag{2.16}$$

Where $\hat{w}_{uv} = w_{uv}/\max(w)$

Graph Diameter (GDm)

is the measure of the graph's eccentricity. In other words, the maximum distance between a vertex to all other vertices is called the diameter, Dm. Dm is in contrast to the more behavioral metrics discussed beforehand (how each node behaves), M, Ce, and Dm focus on the topological level of a network.

Graph Distances (GDi)

We calculated structural distance $D\left(\Gamma_1, \Gamma_2\right)$ between two different graphs in terms of the Jensen-Shannon distance J − S of Laplacian spectra of the two graphs[126]. Briefly if Gaussian kernel $g(x,\lambda)\colon 1/\sqrt{2\pi\sigma^2}\exp\left(-\left(x-m_x\right)^2/2\sigma^2\right)$ exists the function of convoluted spectrum of a network with $\sigma = .01$ is defined as

$$f(x) = \int g(x,\lambda)\sum_k \delta\left(\lambda,\lambda_k\right)d\lambda = \sum_k g\left(x,\lambda_k\right) \quad \text{and} \quad 0 < \int f(x)dx < \infty \tag{2.17}$$

spectral density f^* was then calculated by normalizing f as:

$$f^*(x) = \frac{f(x)}{\int f(y)dy} \tag{2.18}$$

and distance is equal to

$$D\left(\Gamma_1, \Gamma_2\right) = \sqrt{JS\left(f_1^*, f_2^*\right)} \tag{2.19}$$

Graph Fidelity (GFid)

Fi combines the properties of a complex inferred graph and ground truth networks to compare their similarity overall with a single numerical calculate fidelity metric δ as described by Alexandru Topirceanu[69]. Fidelity measures the averages over an arbitrary number of measurements for a graph.

2.5.9 Overview of the benchmarked algorithms

GENIE3

GENIE3[57] was the top performer algorithm for inferring regulatory networks for bulk transcriptional data in the DREAM4 challenge. GENIE3 regresses the expression profile of genes one at a time and then ranks each gene's importance in predicting other genes' profiles using random forests. It then constructs regulatory networks by aggregating these weights such that the level of importance becomes the edge weights in the network.

GRISLI

GRISLI[62] uses linear ODEs to calculate how gene expression values change during the cell sampled cell states for the provided experimental times (here, input pseudo-times).

GRNBoost2

GRNBoost2[63] uses regression and tree-based models like GENIE3 but incorporates stochastic gradient boosting and early stopping to achieve better speed for bigger networks with more genes under study.

LEAP

LEAP[49] calculates asymmetric Pearson correlation of RNA-seq read counts between permuted different experimental times. The Maximum Pearson's correlation between two pairs indicates the directed edge weights in the network.

PIDC

PIDC[50] calculates unique mutual information between two genes such that the relationship between two genes is proportional to the relationship of those genes to all the other genes in the network.

PPCOR

PPCOR[47] estimates the pairwise partial correlation coefficients given all the other genes' expressions and computes a P-value for each correlation. Negative correlations here are deemed inhibitory and positive correlations are considered activating.

SCODE

SCODE[64] is essentially the data dimensionality and regress for linear ODEs to describe how gene-gene interactions result in observed gene expression dynamics.

SCRIBE

SCRIBE[65] computes mutual information between the past state of a regulator gene and the current state of a regulated gene, given the state of the regulated gene at the last experimental time, then excludes interactions relating to other indirect effects.

SINCERITIES

SINCERITIES[51] infers gene-gene weights in GRNs by computing changes in the distributions of gene expressions between two consecutive experimental times using the Kolmogorov–Smirnov statistic and ridge regression. Partial correlation analyses between pairs of genes then indicate the direction interactions in these networks.

2.6 Supplementary information

Tissue	Method	Pr	F1-S	AUROC	AUPRC	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
Linear	NS-DIMCORN	0.966	0.963	0.925	0.913	99.998	-0.271	1.000	2.828	211.132	0.088	0.048	3.000	0.330
Linear	GENIE3	0.858	0.830	0.898	0.750	16.666	0.381	0.714	2.449	243.949	0.016	0.143	2.000	0.330
Linear	GRISLI	0.538	0.467	0.469	0.247	0.000	0.143	0.762	4.899	121.474	0.249	0.286	3.000	0.430
Linear	GRNBOOST2	0.909	0.865	0.884	0.614	28.571	0.286	0.690	3.464	172.205	0.085	0.190	2.000	0.330
Linear	LEAP	0.760	0.699	0.721	0.417	0.000	0.286	0.762	4.690	126.920	0.216	0.238	2.000	0.330
Linear	PIDC	0.000	0.643	0.500	0.250	0.000	-1.238	0.762	5.292	112.389	nan	0.333	5.000	0.430
Linear	PPCOR	0.732	0.633	0.776	0.670	79.998	0.095	0.590	6.000	99.000	0.235	0.333	3.000	0.430
Linear	SCODE	0.833	0.500	0.878	0.777	85.713	-0.238	0.238	10.583	55.695	0.296	0.667	4.000	0.430
Linear	SCRIBE	0.842	0.798	0.830	0.496	33.333	0.286	0.714	4.000	149.000	0.164	0.190	2.000	0.295
Linear	SINCERITIES	0.760	0.699	0.789	0.665	59.999	0.238	0.714	4.690	126.920	0.216	0.238	2.000	0.330

Table S1 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms for synthetic data that represents linear trajectory for cellular process as depicted at fig. 2.2

.

Tissue	Method	Pr	F1-S	AUROC	AUPRC	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
Trifurcating	NS-DIMCORN	0.913	0.770	0.991	0.930	47.368	0.336	0.541	39.497	57.233	0.323	0.301	0.400	0.470
Trifurcating	GENIE3	0.725	0.181	0.605	0.269	40.625	0.868	0.448	93.936	23.485	0.666	0.833	0.400	0.629
Trifurcating	GRISLI	0.827	0.629	0.688	0.324	57.692	0.430	0.801	47.166	47.764	0.535	0.395	0.400	0.629
Trifurcating	GRNBOOST2	0.649	0.166	0.596	0.280	37.931	0.865	0.452	93.545	23.587	0.666	0.830	0.400	0.629
Trifurcating	LEAP	0.828	0.505	0.743	0.317	36.667	0.593	0.668	65.197	34.278	0.609	0.558	0.400	0.629
Trifurcating	PIDC	0.898	0.541	0.771	0.284	47.368	0.604	0.710	64.395	34.717	0.575	0.569	0.400	0.629
Trifurcating	PPCOR	0.804	0.286	0.793	0.580	81.818	0.799	0.486	87.216	25.371	0.664	0.764	0.400	0.629
Trifurcating	SCODE	0.818	0.376	0.605	0.190	53.125	0.723	0.530	79.402	27.967	0.606	0.688	0.400	0.629
Trifurcating	SCRIBE	0.802	0.650	0.562	0.162	45.000	0.361	0.942	38.670	58.478	0.513	0.326	0.400	0.629
Trifurcating	SINCERITIES	0.781	0.642	0.521	0.131	56.250	0.325	0.885	37.745	59.936	0.508	0.308	1.400	0.629

Table S2 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms for synthetic data that represents trifurcating trajectory for cellular process as depicted at fig. 2.2

.

Tissue	Method	Pr	F1-S	AUROC	AUPRC	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
Cyclic	NS-DIMCORN	0.838	0.816	0.944	0.883	99.998	0.800	0.417	3.162	157.114	0.049	0.133	0.000	0.284
Cyclic	GENIE3	0.714	0.714	0.722	0.687	33.332	0.350	0.444	2.000	249.000	0.111	1.000	1.500	0.572
Cyclic	GRISLI	0.813	0.772	0.789	0.565	20.000	0.800	0.389	3.464	143.338	0.043	0.200	0.000	0.314
Cyclic	GRNBOOST2	0.752	0.755	0.700	0.677	33.332	0.433	0.611	2.000	249.000	0.153	0.067	1.500	0.464
Cyclic	LEAP	0.752	0.755	0.856	0.771	33.332	0.433	0.611	2.000	249.000	0.153	0.067	1.500	0.464
Cyclic	PIDC	0.959	0.953	0.967	0.915	49.999	0.867	0.750	2.449	203.124	0.044	0.067	0.000	0.284
Cyclic	PPCOR	0.929	0.908	0.911	0.697	99.998	0.733	0.500	2.449	203.124	0.049	0.133	1.000	0.579
Cyclic	SCODE	0.754	0.635	0.789	0.645	39.999	0.467	0.156	5.292	93.491	0.167	0.400	1.000	0.395
Cyclic	SCRIBE	0.714	0.714	0.844	0.729	66.664	1.000	1.000	1.414	352.553	0.000	1.000	0.000	nan
Cyclic	SINCERITIES	0.619	0.619	0.544	0.306	99.995	-0.033	0.500	3.162	157.114	0.167	1.000	2.000	0.464

Table S3 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms for synthetic data that represents cyclic trajectory for cellular process as depicted at fig. 2.2

.

Tissue	Method	Pr	F1-S	AUROC	AUPRC	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
Compound	NS-DIMCORN	0.939	0.350	0.997	0.925	61.224	0.352	0.280	180.159	19.537	0.719	0.765	1.200	0.395
Compound	GENIE3	0.892	0.192	0.726	0.217	60.000	0.247	0.257	199.369	17.559	0.765	0.871	1.200	0.395
Compound	GRISLI	0.903	0.462	0.697	0.226	62.500	0.464	0.382	155.603	22.778	0.718	0.653	1.200	0.395
Compound	GRNBOOST2	0.893	0.155	0.759	0.228	58.696	0.221	0.239	204.828	17.064	0.779	0.896	1.200	0.395
Compound	LEAP	0.918	0.505	0.766	0.210	53.488	0.499	0.435	146.219	24.305	0.721	0.619	1.200	0.395
Compound	PIDC	0.932	0.661	0.836	0.214	60.000	0.664	0.643	106.558	33.723	0.660	0.454	1.200	0.395
Compound	PPCOR	0.934	0.362	0.899	0.676	85.416	0.363	0.299	176.969	19.908	0.751	0.754	1.200	0.395
Compound	SCODE	0.938	0.106	0.542	0.132	61.224	0.187	0.211	212.186	16.438	0.779	0.930	2.200	0.395
Compound	SCRIBE	0.889	0.651	0.573	0.089	57.143	0.693	0.708	98.271	36.651	0.661	0.424	1.200	0.395
Compound	SINCERITIES	0.886	0.662	0.565	0.085	57.692	0.712	0.675	96.287	37.427	0.642	0.405	1.200	0.395

Table S4 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms for synthetic data that represents compound trajectories for cellular processes as depicted at fig. 2.2

.

Tissue	Method	GII	AII	MEI	GI5KI	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
BrainMap	NS-DIMCORN	772	34035	-49	11.410	47.798	0.835	0.521	3575.325	9.321	0.325	0.487	1.000	0.233
BrainMap	GENIE3	447	32965	-374	1.032	74.049	0.818	0.510	3649.886	9.110	0.413	0.471	1.000	0.233
BrainMap	GRISLI	0	0	-821	nan	0.000	-0.337	0.783	85.364	431.269	nan	0.012	2.000	0.233
BrainMap	GRNBOOST2	429	37507	-392	1.119	57.343	0.887	0.585	3940.637	8.364	0.459	0.537	0.000	0.233
BrainMap	LEAP	483	37183	-338	0.862	95.652	0.881	0.418	4064.006	8.080	0.377	0.533	1.000	0.233
BrainMap	PIDC	0	0	-821	nan	0.000	-0.337	0.783	85.364	431.269	nan	0.012	2.000	0.233
BrainMap	PPCOR	771	64977	-50	1.686	17.769	0.711	0.264	6673.072	4.530	0.510	0.940	0.000	0.233
BrainMap	SCODE	821	68265	0	0.931	100.000	0.663	0.217	7012.493	4.262	0.518	0.988	1.000	0.233
BrainMap	SCRIBE	0	0	-821	nan	0.000	-0.337	0.783	85.364	431.269	nan	0.012	2.000	0.233
BrainMap	SINCERITIES	369	33375	-452	0.977	69.106	0.812	0.537	3617.587	9.200	0.385	0.477	2.000	0.245

Table S5 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms using brain scRNA-seq data obtained from BrainMap fig. 2.5

.

Tissue	Method	GII	AII	MEI	GI5KI	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
LiverMap	NS-DIMCORN	1384	46983	-53	10.008	55.058	0.797	0.528	4833.467	7.958	0.457	0.485	0.111	0.225
LiverMap	GENIE3	918	49412	-519	2.433	77.233	0.822	0.611	4831.999	7.961	0.419	0.511	1.111	0.225
LiverMap	GRISLI	1032	64163	-405	1.741	100.000	-0.031	0.387	6498.063	5.664	0.486	0.668	1.089	0.469
LiverMap	GRNBOOST2	974	60753	-463	3.219	61.499	0.943	0.597	5784.176	6.486	0.472	0.631	0.111	0.225
LiverMap	LEAP	901	45906	-536	2.629	96.892	0.777	0.502	4549.114	8.518	0.293	0.473	1.111	0.225
LiverMap	PIDC	894	46708	-543	3.051	84.116	0.794	0.639	4499.431	8.623	0.359	0.482	1.111	0.225
LiverMap	PPCOR	1358	90160	-79	3.160	14.948	0.744	0.325	8520.335	4.082	0.509	0.944	0.111	0.225
LiverMap	SCODE	1437	93961	0	1.935	100.000	0.703	0.285	8882.588	3.875	0.514	0.985	0.889	0.225
LiverMap	SINCERITIES	904	52193	-533	2.817	68.363	0.852	0.604	5007.700	7.647	0.367	0.540	0.111	0.225

Table S6 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms using brain scRNA-seq data obtained from Protein Atlas fig. 2.5

.

Tissue	Method	GII	AII	MEI	GI5KI	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
GTExBrain	NS-DIMCORN	617	33700	-112	9.810	59.643	0.771	0.658	3529.475	9.341	0.524	0.494	0.167	0.237
GTExBrain	GENIE3	552	49199	-177	3.280	44.203	0.997	0.451	5167.502	6.063	0.589	0.726	0.167	0.237
GTExBrain	GRISLI	647	57676	-82	1.498	100.000	0.300	0.286	6146.067	4.939	0.618	0.853	0.333	0.560
GTExBrain	GRNBOOST2	636	58207	-93	3.222	35.220	0.862	0.361	6039.692	5.043	0.613	0.861	0.167	0.237
GTExBrain	LEAP	418	32146	-311	2.215	94.498	0.739	0.542	3465.000	9.534	0.437	0.470	1.167	0.237
GTExBrain	PIDC	380	32989	-349	1.840	72.105	0.760	0.581	3526.582	9.350	0.465	0.483	1.167	0.237
GTExBrain	PPCOR	484	39033	-245	3.290	36.570	0.851	0.640	4050.023	8.012	0.593	0.573	0.167	0.237
GTExBrain	SCODE	729	66795	0	2.456	100.000	0.733	0.249	6906.650	4.285	0.628	0.989	0.833	0.237
GTExBrain	SCRIBE	0	0	-729	nan	0.000	-0.267	0.751	76.211	477.934	nan	0.011	1.833	0.237
GTExBrain	SINCERITIES	395	34760	-334	1.302	83.038	0.787	0.655	3640.856	9.025	0.525	0.509	0.167	0.237

Table S7 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms using brain RNA-seq data obtained from GTEx fig. 2.5

.

Tissue	Method	GII	AII	MEI	GI5KI	CDiR	GPL	GCE	GD	GDC	GM	GDen	GDm	GDi
GTExLiver	NS-DIMCORN	1109	60832	-133	14.679	52.119	0.939	0.210	5449.034	8.047	0.388	0.489	0.714	0.210
GTExLiver	GENIE3	263	15753	-979	1.581	100.000	-0.558	0.861	2296.129	20.471	0.512	0.119	2.711	0.210
GTExLiver	GRISLI	1199	116005	-43	1.509	100.000	-0.040	0.231	10436.040	3.724	0.511	0.942	1.214	0.258
GTExLiver	GRNBOOST2	1000	96794	-242	3.794	36.000	0.644	0.403	8663.514	4.691	0.495	0.785	0.714	0.210
GTExLiver	LEAP	635	55230	-607	2.018	88.189	0.986	0.582	5086.391	8.693	0.337	0.443	0.286	0.210
GTExLiver	PIDC	0	0	-1242	nan	0.000	-0.561	0.779	111.760	440.122	nan	0.010	2.714	0.210
GTExLiver	PPCOR	1242	121771	0	0.000	0.000	0.439	0.221	10845.712	3.546	0.512	0.990	1.714	0.210
GTExLiver	SCODE	1242	121771	0	1.216	100.000	0.439	0.221	10845.712	3.546	0.512	0.990	1.714	0.210
GTExLiver	SCRIBE	790	73887	-452	2.394	86.709	0.832	0.551	6678.529	6.382	0.423	0.597	0.714	0.210
GTExLiver	SINCERITIES	747	70082	-495	1.013	99.331	1.000	0.472	6486.310	6.601	0.401	0.565	1.214	0.210

Table S8 | Benchmarking NS-DIMCORN against the other state-of-art GRN inference algorithms using liver RNA-seq data obtained from GTEx fig. 2.5

.

Chapter 3

Precise graph-based annotation of the whole genome of patients with complex heritable diseases

3.1 Abstract

Precise annotation of the genome is the limiting factor in identifying druggable cellular functions that underpin the complex onset and manifestation of inheritable diseases from DNA sequencing data. Although many *ad hoc* methods have been developed for agglomerating DNA sequencing data with other *-omic* information to shed light on the etiology and source of these conditions, there is yet to be a uniform approach for combining tissue-specific epistasis and the deleterious impact of disease-related genetic variants. Recent advancements in deep learning and the availability of datasets comprising hundreds of thousands of whole genome sequences of patients accompanied by longitudinal clinical records provide new prospects for devising models with enough depth to capture the intricacy of genotype-phenotype relationships in complex heritable diseases. Here we developed the Precise Graph-based Genome-Wide Annotation Software (PG-GWAS), a new method based on graph attention networks (a combination of a graph neural network and attention layers) for annotating sequencing data and identifying deleterious genetic variations in their biological context. The main component of PG-GWAS's is a pan-genomic graph of the whole genome, augmented with cell type/developmental state-specific regulatory networks. In these graphs, nodes capture gene-specific deleterious effects of genetic variation, and edges are weighted by gene-gene interaction strength, enabling effective pooling of information from genes' local deleterious burden and effects from distal sets of interacting genes. Annotation of the genome of patients with rare complex inheritable "Mendelian" diseases shows that PG-GWAS successfully identifies damaging variants. Furthermore, it prioritizes genes involved in pathways paramount for disease progression and symptoms. Finally, PG-GWAS supports processing genome-scale datasets using distributed, GPU-accelerated data architecture and implementation.

3.2 Introduction

The onset and most phenotypic manifestations of heritable diseases can not be explained by a small number of genomic loci or only by the additive combination of loci associated with the diseases[127–129]. Consequently, studying such progressions/manifestations is more exacting and does not conform to the classic Mendelian paradigm that a few rare genomic variants are responsible for a disease,[130–132]. However, the advancement of next-generation sequencing and population-scale biobanks, such as the UK Biobank[133], "All of Us"[134] and Biobank Japan[135] presented an unprecedented opportunity to investigate the common heritable contributors of complex diseases[136]. These data sets contain health records for hundreds of thousands of individuals and genomic data to allow scaled mapping between genome and clinical phenotypes under the assumption of poly- or omnigenic models of complex diseases[133,134].

To date, the most common approach for studying complex diseases, Genome-wide association studies (GWAS), analyzes the differences between three sets of genetic markers, namely Single Nucleotide Polymorphisms (SNPs), Sequence Variations (SVs), or copy-number variants in individuals with the trait and a matching control samples cohort[137]. This association between a genetic variant and an outcome is quantified as an odds ratio (OR) by comparing the occurrence of genetic variants in the case and control group with linear or logistic regression[137] Critically, after 15 years of GWAS and despite sample sizes of some studies exceeding a million participants[138,139], most of the genomic risk loci associated with phenotypic traits using GWAS explained only a relatively small proportion of complex diseases heritability leading to the term "missing heritability"[140–145].

The GWAS methodology's shortcoming is two-fold[146–148]. Firstly, on the intrinsic level, GWAS requires accurate genotyping, yet the quality of called genetic

markers depends on the sequencing instrument used and the multi-stage, error-prone data processing involved for determining variant markers[149,150]. Moreover, even with accurate genotyping, GWAS depends on the 'reference genome,' meaning variant calling is bedeviled by the presence of population stratification, such as for people with different racial ancestries (genetic backgrounds)[151]. Secondly, due to the cosegregation of chunks of DNA during meiotic recombination, some neighboring genetic variants tend to be inherited together and appear to be correlated with a phenotype misleadingly[152]. This phenomenon is referred to as linkage disequilibrium (LD) and results in biased test statistics and inflated statistical association[153]. This situation is only exacerbated with GWAS design that tests for millions of associations one at a time (multiple testing), resulting spurious associations[154–158], and the proposed remedy for this situation, stringent multiple-testing thresholds, removes genuine associations with small effect sizes[159,160]. Most importantly, classic GWAS does not consider the type of effects caused by genomic variants (e.g., loss of function, buffering of genes[161], regulatory activity of non-coding DNA) or the interactome networks that mediate genotype-phenotype relationships in specific cell types that drive a disease[162–166] Therefore it is not surprising that the current approach to GWAS has been criticized for lacking a theoretical basis and being only a statistical convenience[167–170]. Specifically, complex heritable traits that are influenced by high-order synergistic (epistatic) interactions between genes in all expressed genomic elements in the cell must also be accounted for[144,167,171,172].

As a result, recent GWAS attempted to tackle these challenges by 1) functionally annotating downstream effects of genomic variants and 2) integrating GWAS results with other -omics information such as gene expression, chromatin activity, and regulatory networks between and among risk factors[98,127,128,150,166,167,173,174]. The first approach can involve variant-level identification of genomic coordinates on the reference genome. These genomic coordinates are localized to protein-coding or non-

coding regions of the DNA by converging the coordinates on proximal known genes' location on the assembled reference genome[175]. The impacts of these genomic variations can then be examined for expression pattern and protein integrity at a gene level[25]. In this regard, tools like Variant Effect Predictor (VEP)[176] infer whether the variant is exonic, intronic, splicing, 3 -untranslated region (UTR), 5 -UTR, intergenic, synonymous, non-synonymous or is a frameshift insertion/deletion. These tools provide a gene loss of function or protein perturbation score based on some particular nucleotide-based criteria[177] available from coordinate base datasets such as ANNOVAR[178] and CAAD[179]. Indeed, from an information-theory point of view, protein functional information is encoded within its primary DNA sequence[180]. But protein's primary sequence also determines functional three-dimensional shape[180], and methods based on Deep Language models have been shown to learn this structural information accurately from sequencing data[25,181]. In the second approach, context-specific functional gene expression is based on the observation that specific diseases primarily affect specific types of cells[182]. Indeed, it is equally noteworthy that while core genes directly affect disease pathogenesis, peripheral genes modify or buffer these conditions differently in individuals (i.e., each person has a unique genetic background[183]). This type of variation may be discovered by grouping (guilt-by-association) methods in a function-specific manner[184,185]. The genetic regulatory network comprising identified gene-gene interactions can then be consolidated, organizing these individual interactions into biologically meaningful groups that allow testing various genetic scenarios by specific gene variant queries[186–188]. However, delineating such complex interactions requires integrating data from multiple modalities with distinct feature spaces in a way that can be flexibly augmented and thoroughly analyzed[189]. To our knowledge, there has not been a unified framework that addresses the shortcomings mentioned above in capturing the deleterious effect of genetic variants, thence proceeding to a holistic examination of its context-

specific consequences of such annotated variation is required[190–193]. Here we present our Precise Graph-based Genome-Wide Annotation Software (PG-GWAS), a graph attention neural network model[194] designed to take into account all regulatory information available for a diseased cell. PG-GWAS consequently improves predictions about the deleterious effect of genomic variants. In PG-GWAS graphs, nodes represent genetic variants with contextualized embedding (converting high-dimensional data to low-dimensional data) scores that encode the likelihood of deleterious effects on their respective gene products in their biological context.

The human reference genome is a unified linear representation of a human genome, created by melding haplotypes of about 25 people, with a single individual of Caucasian ancestry dominating this composite[195]. This "reference genome" cannot represent the broad spectrum of human genome population stratification[196]. However, the well-practiced method of minimizing errors and limiting the number of genetic variants is to filter variable sites with a minor allele frequency (MAF) computed against a reference genome. This practice has been shown to confound the results by ignoring more common variants with smaller effect sizes or common variants with high epistasis (*i.e.,* rare combination of two common or rare variants)[197]. By contrast, deleterious scores from PG-GWAS are pan-genomic, meaning they are not dependent on the reference genome. Moreover, PG-GWAS operates on graph-structured data allowing any epistasis to be included in the analysis of genomic variants. It utilizes an attention-based architecture that effectively pools networks' heterogeneous local features, elucidating how genomic variants are organized into functional pathways and contribute to cellular dysfunctions. Finally, PG-GWAS is built on an optimized, distributed, and GPU-accelerated back-end computing architecture to enable the processing of extremely large data sets required for this type of analysis. Here, we showed PG-GWAS was accurate and robust in identifying deleterious variants using a curated set of known pathogenic variants in the human

population. Furthermore, we provide a proof-of-concept for rare heritable complex diseases, in which we sequenced and annotated the whole genomes of cohorts of Huntington's disease (HD) and of Niemann-Pick type C1 (NP-C1) patients with the distinct onset and manifestation of the disease and successfully identified genetic modifiers consistent variants with the current literature for both diseases.

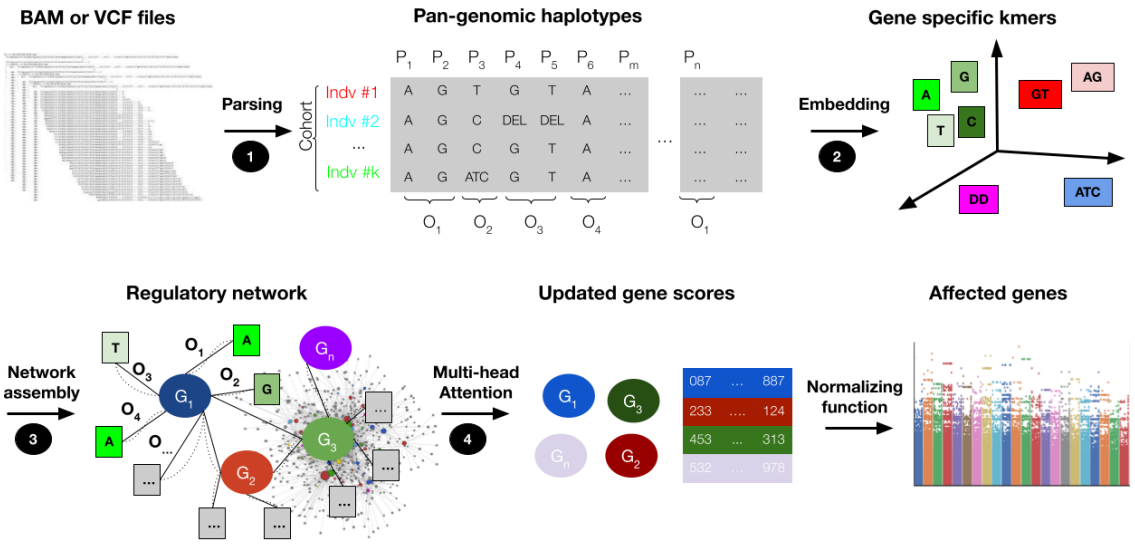


Figure 3.1 | Overview of PG-GWAS. 1: Binary Sequence Alignment Map (BAM) or Variant Call Format (VCF) files are first parsed into pan-genomic haplotypes where each nucleotide of a fixed length P_m string is grouped into a gene-specific k-mer with coordinate O_n. 2: K-mers then receive unique contextualized embedding scores that encode the likelihood of deleterious effects on respective gene products. 3: Graph representation of individuals' genomes, comprising genes as nodes, contextualized k-mers embedding as node features, and tissue-specific gene-gene interactions as edges, are used to assemble regulatory networks. 4: Multi-head graph attention network mechanism sums up and normalizes the collective effects of all interactions in the network and assigns each node a new score.

3.3 Results

3.3.1 Overview of the algorithm

PG-GWAS is a supervised machine-learning method for annotating and prioritizing local and global tissue-specific genomic features (a combination of variants and genomic interactions) that contribute to the manifestation and onset of complex heritable diseases. PG-GWAS operates on a graph where nodes represent genetic variations contextualized embedding, and edges represent functional relationships between nodes and output that are pooled and updated numerical values of genetic variation and genomic interactions (Section 3.3.3). These features in latent space are of much lower dimension than the original gene node features comprising all possible nucleotides for a sequencing region, allowing them to be meaningfully analyzed and compared.

The functional consequence of genetic variants depends on both effects of amino acid substitutions, regulatory effects, metabolic pathways, and epistasis between genes[198]. PG-GWAS uses multi-head graph attention neural networks, a combination of densely connected convolutional neural networks[199] layers and attention[200] layers to combine the effects of genetic variation and interactions of any provided context as the backbone network. As illustrated in (Figure 3.1), PG-GWAS requires aligned sequencing data as an input indicating the deletion or insertion of a nucleotide or sets of nucleotides at a given position $P \in N$ of the whole genome sequence. These haplotypes are then converted into gene-specific k-mers (defined at Section 3.5.1) and constructed as constrained (no overlapping k–mers) De Bruijn graphs[201] for efficient storage and retrieval, where each node on the graph is assigned a gene-specific index $O \in N$ and embedding score. Embedding scores represent deleterious effects of the k-mers in their biological context (Section 3.5.2) and are computed by training a model to learn gene-specific embedding of genetic

variants using the dataset of known relationships among human genome variations and phenotypes[202]. Any prior knowledge of the node's interactions (here a *de-novo* cell type-specific regulatory network) can then be incorporated in the final network, whether each node is further branched with various gene specif k-mers or any other information. Final scores are then computed by pooling and normalizing values of all nodes by the multi-head graph attention network mechanism[194] as described in the Methods, (Section 3.3.3) to output the final predictive Normalized Embedding Score (NES) of each gene for the phenotypes under investigation.

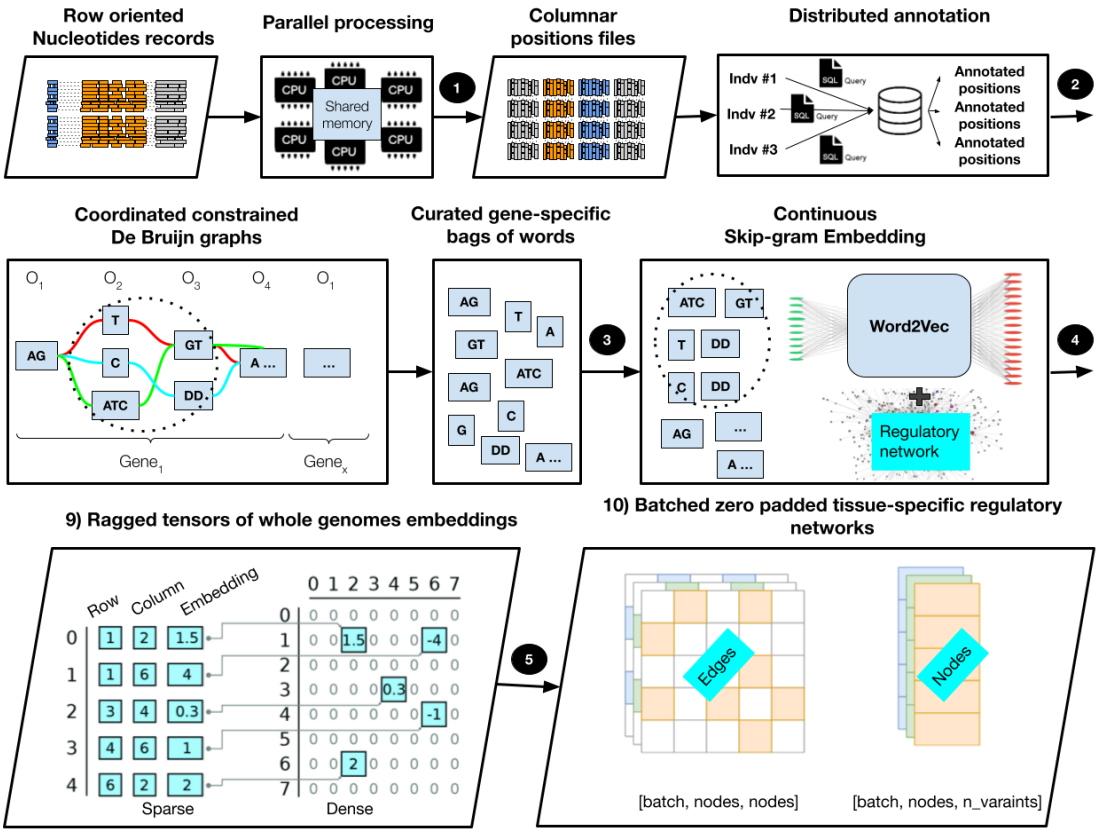


Figure 3.2 | Overview of PG-GWAS data architecture. 1: Row-oriented genomic records for each nucleotide and nucleotide position are processed using a shared memory multi-threaded implementation and stored as column-oriented records. 2: These records are then annotated with gene tags using a distributed SQL query engine. 3: Unique combinations of nucleotides for distinct tags are assembled into constrained De Bruijn graphs where each node in the graph (O_x) represents a gene-specific k-mer. 4: K-mers were assigned a unique contextualized embedding score that encoded the likelihood of deleterious effects on its respective gene and the results were stored as sparse matrices. 5: By incorporating the tissue-specific gene regulatory networks, personalized graph representations of individuals' genomes that integrate contextualized embedding scores are constructed and batched by zero padding.

3.3.2 Benchmarking

Multi-node parallel architecture allows fast data retrieval, annotation, and batch processing:

The human nuclear genome is approximately 3 200 000 000 nucleotides long, divided into 23 homologous pairs of linear chromosomes, hence requiring the processing of terabytes of data, depending on the sequencing depth and genome coverage[203] (approximately 150 terabytes of data to analyze 1000 individuals' genomes genetic variants only[204,205]). Indeed many software tools have been developed for querying, variant calling, and file format conversion and manipulation of these data files. However, tools are not particularly designed with batch processing in mind, as would be required for efficient hardware utilization in deep learning-based analysis[206]. Therefore, this study set out to analyze approximately 67 000 whole genome sequences devising a data architecture and developing a suite of helper software to allow almost linear computing scalability depending on the number of CPU cores available.

Firstly we optimized and reimplemented the currently fastest option available BAM/VCF read-write[207,208] C native library by adding shared-memory multiprocessing[209] capability and converted row-oriented sequencing alignment and variant formats to columnar-oriented files. This further enabled us to utilize distributed SQL query engines[210] for annotating gene-specific k-mers and assembling constrained De Bruijn graphs of these k-mers (Figure 3.2-1,2). Moreover, the original implementation of the Word2Vec algorithm used here for extracting low-dimensional vector representations of the k-mers is parallelized for multi-core CPU architectures but is based on vector-vector operations that do not efficiently use computational resources[211]. To remedy this issue, we instead adopted the Hogbatch approach[212] that uses mini batching, negative sample sharing, and expressing the problem using matrix multiplication operations (Figure 3.2-3). Altogether these approaches allowed us

to scale up the computation time nearly linearly across cores and nodes as opposed to the exponential time required for single-node architecture (Figure 3.3).

A complete representation of a cell's regulatory network can amount to 23 000 gene nodes connected to more than 30 million nucleotides (a large adjacency matrix or incidence matrix representing all the edges plus a matrix of node features). Therefore, another critical issue was storing these heterogeneous graphs consisting of several disjoint node sets $V_1 \ldots, V_n$ and edge sets$E_1 \ldots, E_m$ in a way that allows batch processing for optimized GPU usage. Purposefully, we represented and stored regulatory networks as ragged tensors of integers (tensors with non-uniform shapes), in the form of Protocol Buffers, language-neutral, platform-neutral, serialized structured data[213], which could readily be converted to zero padded batches of data (Figure 3.2-5). Courtesy of this approach, on average, we stored final regulatory graphs with only 1.6 MB per whole human genome (as appose to $\approx$ 30 millions $\times$ 8-16 bit per nucleotide excluding indices) . This allowed training our graph attention network on batches of 448 whole human genomes on 80 GB A100 GPUs.

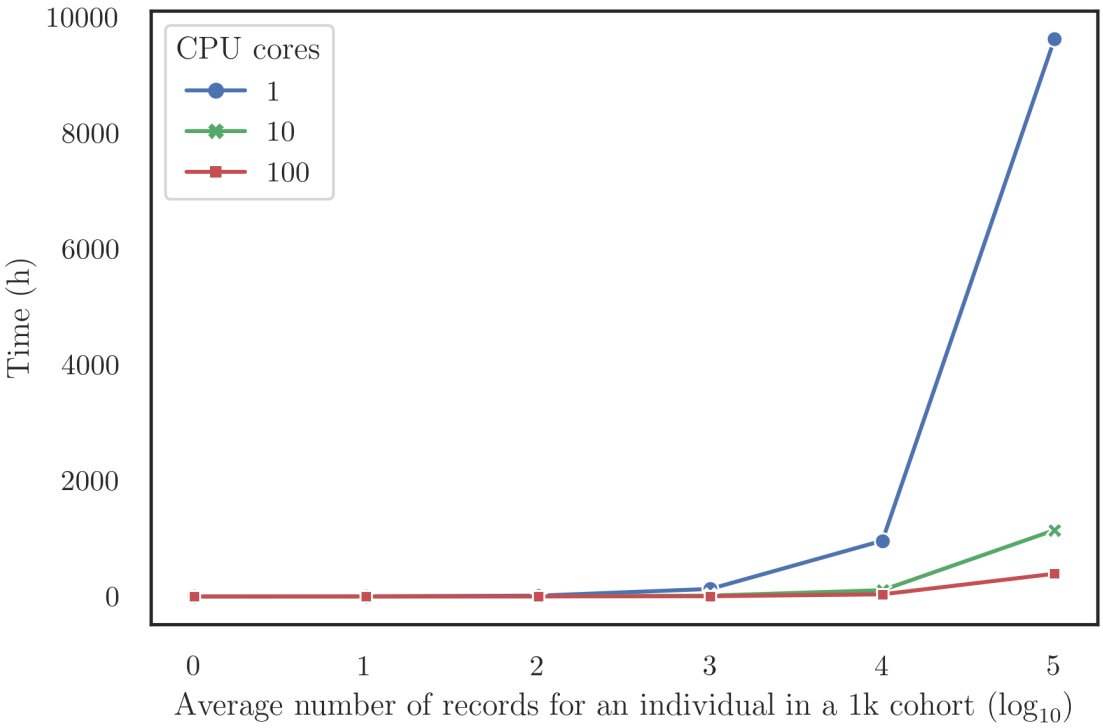
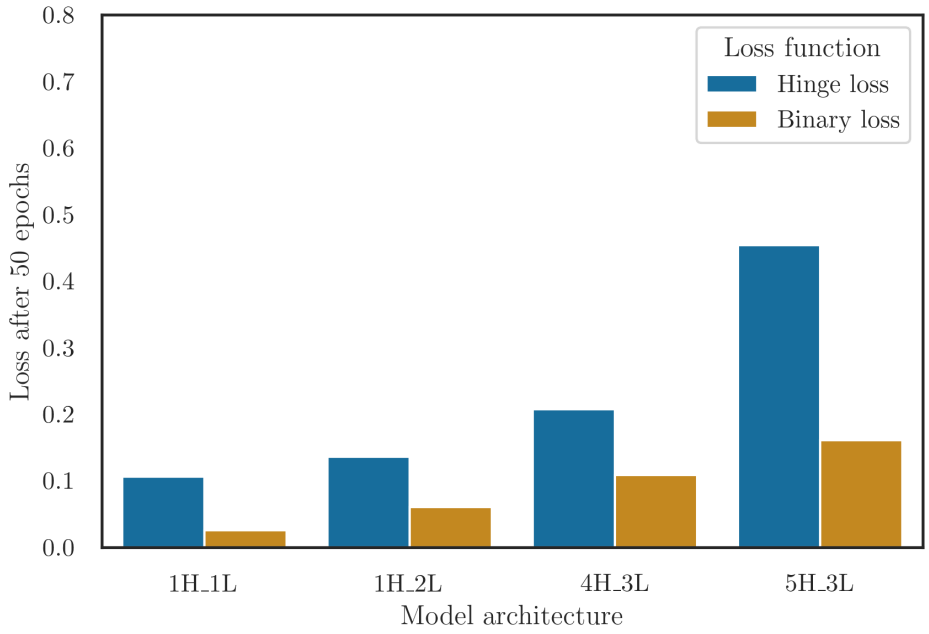
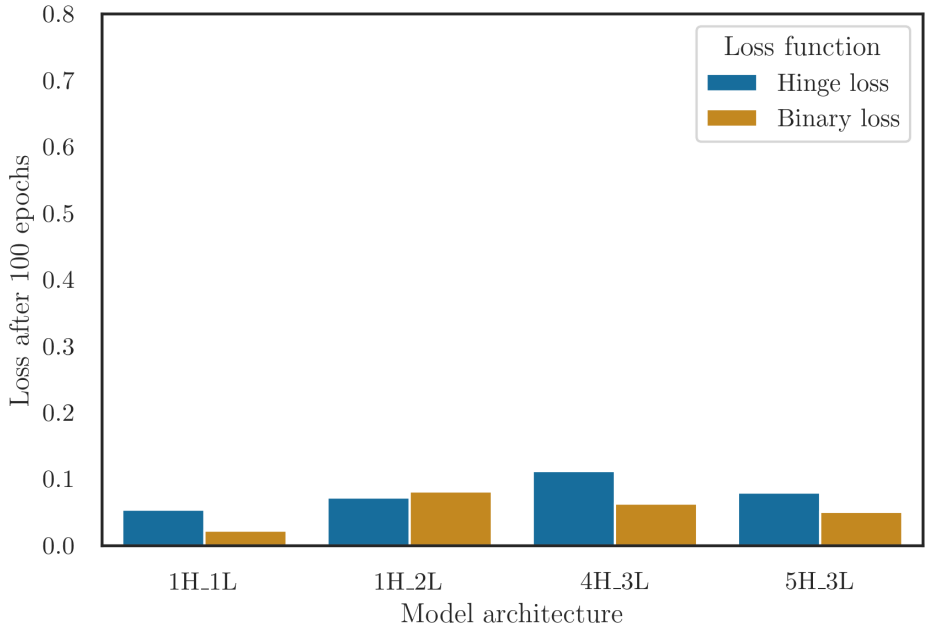


Figure 3.3 | Data processing speed for different numbers of CPU cores. Line plot shows exponential growth of the time required in hours for processing nucleotide records of 1000 individuals stored at separate files into De Bruijn graphs using a single-core implementation. Green and orange lines display the same process with a much gentler slob by using the PG-GWAS shared memory multi-threaded data processing implementation.



(i)



(ii)

Figure 3.4 | Loss function value for different graph attention network architectures used in PG-GWAS. i: Bar plot illustrate Hinge loss and Binary cross entropy loss after 50 epochs for one head one layer (1H_1L), one head two layers (1H_2L), four head three layers (4H_3L) and five head three layers (5H_3L) architectures of the multi-head graph attention network. ii: illustrates Hinge loss and Binary cross entropy loss for the same set of architectures as (i) after 100 epochs. All the different architectures of the models converged before 100 epochs.

PG-GWAS is accurate and stable during training:

We curated a training dataset comprising pathogenic mutations, genetic variants with clinical significance plus benign missense variants from Online Mendelian Inheritance in Man (OMIM)[28], and ClinVar datasets[202] but including only the variants with clinical evidence. To balance the positive and negative sets, we randomly added variants from the ClinVar dataset to the genome of 3110 healthy individuals from different genetic backgrounds (Section 3.5.1). In total, two cohorts of 50, 000 individuals with 24, 600 complete penetrance nonsense and frameshift mutations, 154 824 missense, nonsense, and frameshift genetic variants with known clinical significance, and 159 304 benign missense mutations covering 18458 genes were simulated in this manner for training (80, 20, training validation splits, where test data set were patients data).

We implemented the model and training algorithms using TensorFlow to implement best practices for data automation, model tracking, performance monitoring, and model retraining[86]. We used a stochastic gradient descent with momentum algorithm[214] to update the model's parameters at an initial learning rate of 7e-5 (momentum=0.9)[215]. We applied early stopping with hinge loss[216] for training algorithms on datasets with loss function optimization as a metric to avoid overfitting and experimented with different architectures to study the model's accuracy. We observed that the model loss function would only stay stable with up to 5 attention heads and three dense layers with 14,320 trainable parameters as defined in the Methods (Section 3.5.1) or returned undefined loss. On the other hand, the limiting factor for increasing the number of hidden units in the dense layers of the model was the limited memory of the used GPU. Nonetheless, as illustrated in (Figure 3.4), all models with different architectures are converged after roughly 100 epochs indicating attention architecture utility in learning outcomes of genomewide graphs. Overall our model with five attention heads and three dense layers seems to outper-

form other architectures by a considerable margin, with the AUC-ROC equal to 0.71 compared to the second-best architecture with four attention heads and three dense layers only achieving AUC-ROC equal to 0.67, a somewhat moderate improvement given the small architectural difference. A model with one attention head and two dense layers with AUC-ROC equal to 0.54 showed the worst performance in comparison to the model with only one attention head plus one dense neural network layer with AUC-ROC equal to 0.60, most likely due to the shallowness of the dense layer (low number of trainable variables) in comparison to the multi-head attention layers of the model and overfitting[217].

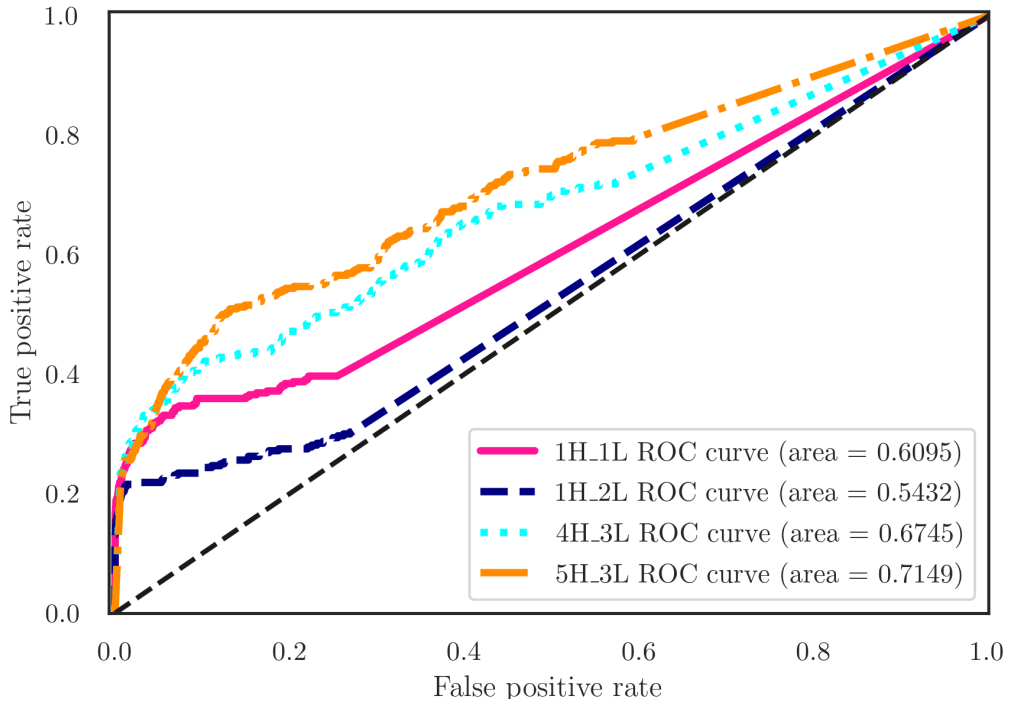
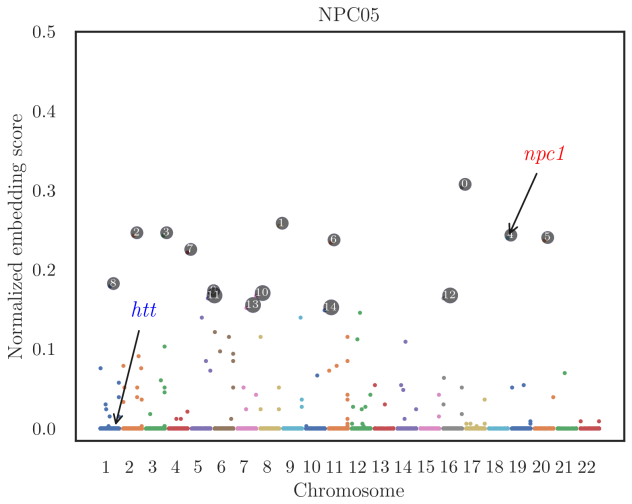


Figure 3.5 | Receiver Characteristic Operator (ROC) and Area Under the Curve (AUC-ROC) for PG-GWAS classifications. ROC plots the true positive rates against false positive rates of of affected genes at various threshold values. Coordinate (0,1) of the ROC space presents 100% sensitivity (no false negatives), 100% specificity (no false positives) and a model with higher AUC-ROC has a better classifications performance. In contrast black line mimics the performance of a random classifier. ROC plotted for one head one layer (1H_1L), one head two layers (1H_2L), four head three layers (4H_3L) and five head three layers (5H_3L) architectures of the multi-head graph attention network used in PG-GWAS and AUC-ROC indicated as ”area” in the graph’s legend shows 5H_3L yields highest accuracy.

3.3.3 Precise graph-based annotation

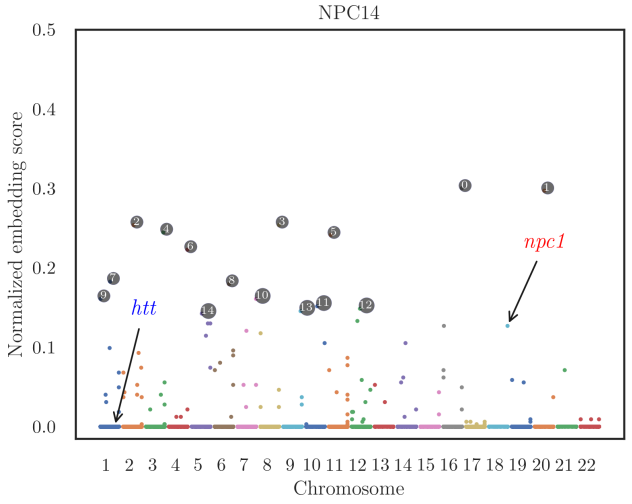
PG-GWAS accurately capture INDEL and missense deleterious mutation in causal genes:

By allowing PG-GWAS to assign predictive Normalized Embedding Score (NES) to deleterious genetic variants, we expected the full penetrance causal mutation to be annotated with the highest score in a genome graph. To test this hypothesis, we sequenced the whole genome of patients in two distinct cohorts of rare and ultra-rare complex heritable diseases. We annotated the genome of these patients with PG-GWAS because PG-GWAS had never seen these data before and was not specially trained on disease-causing mutations of genes for these diseases. We first sequenced and annotated the genome of 29 newly recruited patients of Huntington's disease (HD) as described in the Methods. HD is a rare inheritable disease characterized by multiple insertions of adjacent CAG repeats in the *HTT* gene[218]. Indeed for HD, all patients with longer than normal CAG repeats length in their genome received a high NES for the mutated *htt*. Provided that the NES is not only affected by the deleterious effect of a genetic variant, we expected to observe that some patients with longer CAG repeats receive a higher NES. This difference is partly explainable by the stochastic nature of trained networks to predict the deleterious effect of genetic variants[214,219]. More importantly, the NES is updated by pooling information from the entire network (i.e., the deleterious scores of neighboring nodes in the graph), so it is not independent of other variants in the genome.



Rank	Chr	Gene	Score
0	16	PRKCB	0.304
1	8	PENK	0.255
2	2	ZEB2	0.243
3	3	CCK	0.243
4	18	NPC1	0.24
5	20	SNAP25	0.237
6	11	CRYAB	0.234
7	4	SPP1	0.222
8	1	RGS4	0.179
9	5	FGF1	0.17
10	7	STX1A	0.167
11	5	GABRA1	0.164
12	16	NUPR1	0.164
13	7	ADCY1	0.152
14	10	VIM	0.149

(i)



Rank	Chr	Gene	Score
0	16	PRKCB	0.3
1	20	SNAP25	0.297
2	2	ZEB2	0.254
3	8	PENK	0.254
4	3	CCK	0.245
5	11	CRYAB	0.241
6	4	SPP1	0.223
7	1	RGS4	0.183
8	6	SOD2	0.18
9	1	ACTN2	0.161
10	7	STX1A	0.161
11	10	CTNNA3	0.152
12	12	GRIN2B	0.149
13	9	LPAR1	0.146
14	5	MEF2C	0.142

(ii)

Figure 3.6 | Precise annotation of the genome-wide graph of NPC-C patients with varying *npc1* mutation. Manhattan plot shows each gene's predictive normalized embedding score after applying the multi-head graph attention mechanism. Each gene node in the input file was connected to K-mers with contextualized embedding scores, encoding the likelihood of deleterious effects on that gene, while edges incorporated the tissue-specific gene regulatory interactions of brain cells. Pooled scores are aggregated, normalized, and plotted on the y-axis. i: shows the highest predictive score for *npc1* gene in this cohort was assigned to patient NPC05 with the T/TTTTT frameshift insertion mutation at Chr 18:23563989. ii: the lowest predictive score for *npc1* gene in this cohort was assigned to patient NPC14 with disease-causing single nucleotide variant G/A at Chr18:23520540.

Bearing in mind that HD often involves long insertion-deletion mutations (INDELs) in the *htt* gene, we suspected that these could result in inflated deleterious scores for the k-mer representing INDELS variations (due to high entropy change of DNA sequence after INDELs)[220], biasing our estimation of PG-GWAS accuracy by ballooned NESs. Thus, we turned our attention toward another heritable complex disease and sequenced the whole genome of 13 newly recruited Niemann-Pick type C (NPC-C) patients with one homologous or two heterozygous loss of function mutations on *npc1*. Unlike the relationship of CAG repeat lengths and disease severity in HD, there is generally no linear relationship between the type of loss of function *NPC1* gene mutation and juvenile, late-onset, visceral and neurological symptoms of NPC1 disease[221]. More than 400 disease-causing mutations covering the protein sequence of *NPC1* have been described, but these deleterious genetic variations are mostly missense mutations[202], allowing us to narrow the investigation of the robustness of PG-GWAS for single nucleotide genetic variations. In so doing, we successfully identified damaging *npc1* genes in all the NPC-C patients using the same trained model that was used for annotating HD patient genomes and observed normalized embedding score between 0.132 (Figure 3.6ii) to 0.238 (Figure 3.6i) for *npc1* and a normalized embedding score of 0 for *HTT*. This result further indicated that PG-GWAS was robust in identifying and prioritizing full penetrance deleterious mutation among whole genomes. Notably, conventional GWAS, having no notion of the deleterious effects of each variation, fails to identify both the high penetrance causal genes and disease-modifying variants in a small sample size[222] such as our NPC-C and HD cohorts due to multiple testing and conservative correction. This has been illustrated in Figure S2, Figure S1 and Table S1 where we used GWAS with Linear, Linear Mixed a Bayesian model and different null hypothesis testing techniques as described in the Methods, all failing to identify any variants above the recommended threshold[223] of $p_{val} < 5 \times 10^{7.8}$ after multiple testing correction[224].

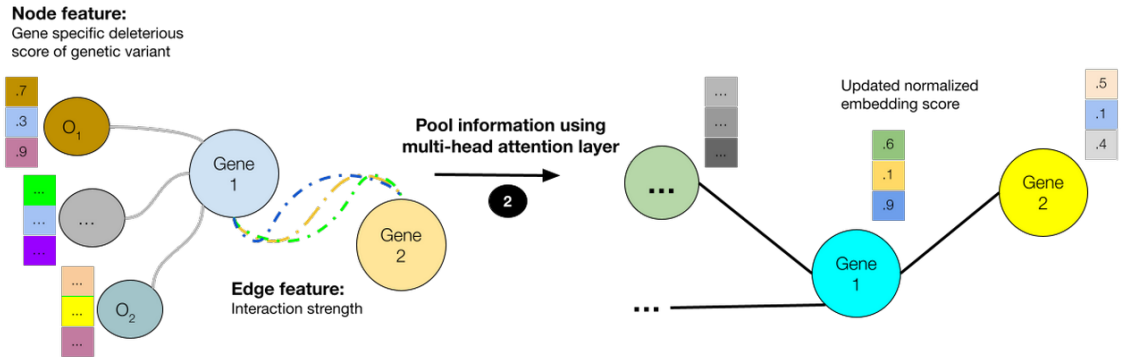


Figure 3.7 | An overview of normalized embedding score. PG-GWAS operates on a graph representing deleterious genetic variants in a biological context, defined as the regulatory network of gene-gene interactions. On the left genes of interest are the slightly bigger nodes connected to gene-specific k-mers, each encoding the deleterious effect of the genetic variants in a latent space as a vector. One the right we used three dense layers and five attention heads to compute normalized embedding scores for each gene by pooling the effects of all genetic variants and interacting neighbors represented with an embedding score next to each gene node.

Patient id	*htt*	CAG repeat length	UHDRS	AGE AT DCL4	Normalized severity
HD001	0.454	43	NA	51	1
HD002	0.256	42	34	55	0.6
HD003	0.245	41	7	72	0.3
HD004	0.189	42	NA	56	0.3
HD005	0.253	43	7	52	0.3
HD006	0.253	43	NA	50	0.6
HD007	0.332	52	45	28	0.9
HD008	0.255	44	NA	48	0.6
HD009	0.263	41	31	65	0.2
HD010	0.505	46	55	40	0.6
HD011	0.243	42	24	50	0.2
HD012	0.256	41	10	60	0.3
HD013	0.229	42	10	55	0.5
HD014	0.242	40	13	52	0.2
HD015	0.227	41	2	NA	0.5
HD016	0.194	41	19	61	0.2
HD017	0.136	40	12	45	0.2
HD018	0.179	41	29	55	0
HD019	0.251	43	63	60	0
HD021	0.03	39	11	67	0.2
HD022	0.273	42	6	35	0.5
HD023	0.246	41	NA	67	0.5
HD024	0.263	43	32	59	0.2
HD025	0.262	41	22	52	0.3
HD026	0.169	49	59	46	0.5
HD027	0.258	43	5	51	0.2
HD028	0.292	47	60	43	0.5
HD029	0.241	40	NA	41	0.6
HD051	0.245	41	NA	NA	NA

Table 3.1 | Clinical information of Huntington's disease cohort. CAG repeat length, normalized embedding scores of deleterious *HTT* gene from PG-GWAS, UHDRS, DCL-4 onset age and normalized severity score (NSS) were tabulated. The effect of CAG repeat length had been accounted for using Langbehn et al. model where patients with a more severe HD or an earlier motor onset received a higher NSE dependent on CAG repeat length.

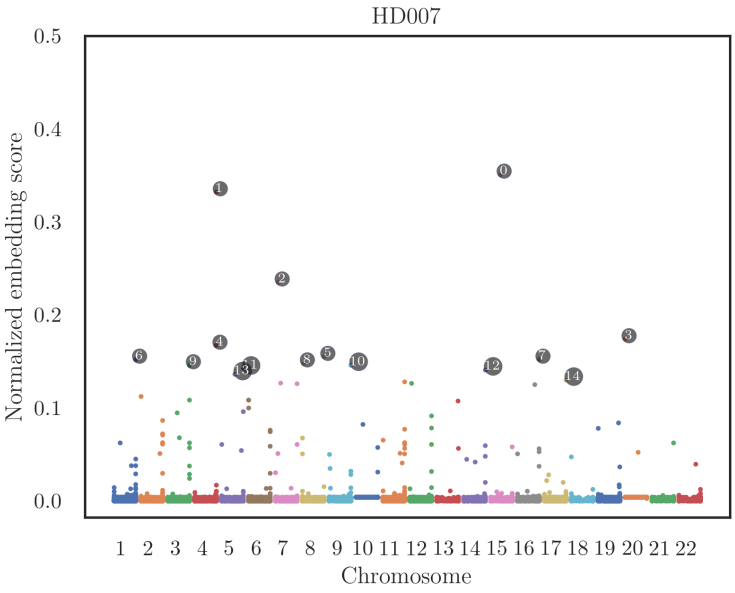
PG-GWAS identifies known modifiers of complex diseases:

Huntington's disease is responsible for the progressive breakdown (degeneration) of nerve cells in the brain and broadly impacts a person's functional abilities[225]. HD symptoms can vary in onset and manifestation and commonly appear when individuals are in their 30s or 40s. In rare cases, such as "juvenile HD", symptoms develop before age 20, and in 4.4-11.5% of individuals, they appear only at over 60 years of age (late-onset patients). HD "age at onset" and disease severity in individuals predominantly result from the uninterrupted longer-than-normal length of the CAG repeats; however, depending on what phenotypic measure is used, this high penetrance genetic abnormality only explains approximately 56% of the heritability for HD varying phenotypes[226].

To identify potential loci responsible for this missing heritability, we first quantitatively measured the age of onset and disease severity in our HD patients cohort and defined a Normalized Severity Score (NSS). This was a CAG length-independent score normalized by the age of the patients. Here we incorporated the Unified Huntington Disease Rating Scale (UHDRS), a standardized assessment consisting of 31 items rated on a scale from 0 to 4 to capture the age of onset plus a wide range of phenotypic measures across motor, cognitive, behavioral, and functional symptoms at different time points along the course of the disease[227] (Table 3.1). Furthermore, for calculating this NSS, we included age at a Diagnostic Confidence Level equal to 4 (DCL4), a clinical score that indicates unmistakable signs of HD-related motor impairments with 99% confidence. At the outcome, we accounted for the length of CAG repeats contribution to disease severity based on Langbehn et al. model[228]. We imputed DCL4 or UHDRS if either was missing and computed the final NSS score values as described in the Methods section. Two patients (HD0051 and HD0029) were eventually excluded for lack of sufficient clinical information (Table 3.1).

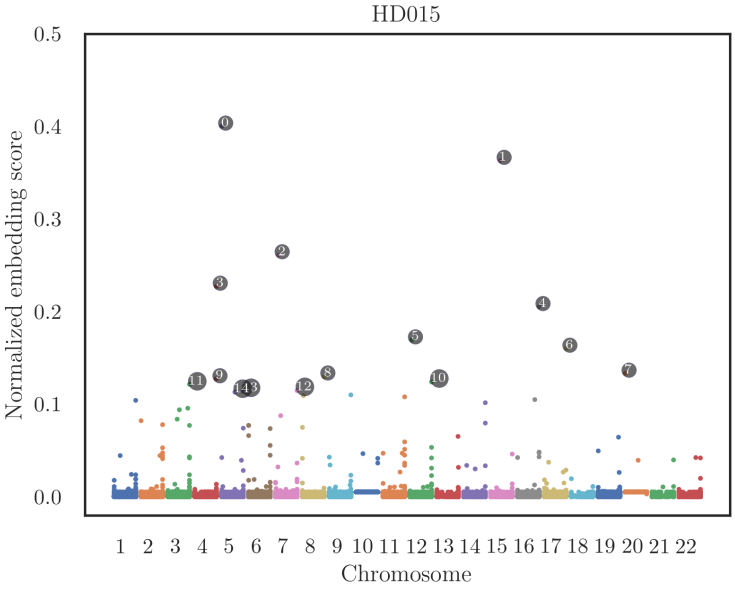
We then identified known modifiers of HD diseases in the annotated networks of

each patient and tried to explain the low or high NSS assigned to each patient using known functions associated with each gene. Significantly *fan1* has been established as a modifier of HD[218,229]. The *FAN1* role is thought to be in DNA interstrand cross-link repair and encodes a protein with endonuclease (cleaving the phosphodiester bond within a polynucleotide chain) and exonuclease (cleaving nucleotides one at a time from the end of a polynucleotide chain) activity[230]. Taking into account that HD is not caused by a simple loss of function of the *HTT* alleles, we surmised it is possible that *FAN1* can play a part in decreasing the effect of toxic polyglutamine in HD patients[231,232]. Therefore, the earliest age at DLC-4 in our cohort of patients for patient HD007, and higher than the median NES score for patient HD015, might be explained by the high predictive normalized embedding score assigned to the *fan1* gene in the genomes (Figures 3.8i to 3.8ii). Similarly, *rhoa* we identified in patient HD006 a high predictive normalized embedding score that has been shown to affect many highly expressed genes in the nervous system contributing to neurodegeneration in Parkinson's, Alzheimer's and Huntington's disease[233,234]. Notably, *tars2* was also unique in this cohort, with a high predictive embedding score in the genome of patient HD006, suggesting the possibility of mitochondrial-related functions in determining the onset and symptoms of the diseases (Figure 3.8iii).



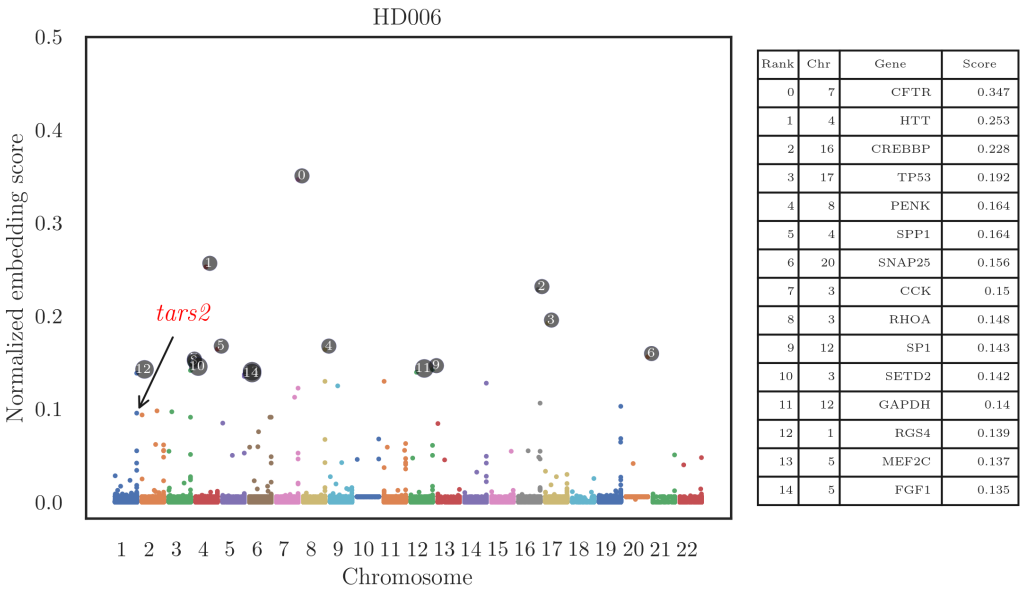
Rank	Chr	Gene	Score
0	15	FAN1	0.351
1	4	HTT	0.332
2	7	CFTR	0.235
3	20	SNAP25	0.174
4	4	SPP1	0.167
5	8	PENK	0.155
6	1	RGS4	0.152
7	16	CREBBP	0.152
8	8	NDRG1	0.148
9	3	CCK	0.146
10	9	LPAR1	0.146
11	5	MEF2C	0.142
12	14	CHGA	0.141
13	5	FGF1	0.136
14	17	TP53	0.13

(i)



Rank	Chr	Gene	Score
0	5	MSH3	0.4
1	15	FAN1	0.363
2	7	CFTR	0.261
3	4	HTT	0.227
4	16	CREBBP	0.205
5	12	GAPDH	0.169
6	17	TP53	0.16
7	20	SNAP25	0.133
8	8	PENK	0.13
9	4	SPP1	0.127
10	12	SP1	0.124
11	3	CCK	0.121
12	7	STX1A	0.115
13	5	MEF2C	0.114
14	5	FGF1	0.113

(ii)



Rank	Chr	Gene	Score
0	7	CFTR	0.347
1	4	HTT	0.253
2	16	CREBBP	0.228
3	17	TP53	0.192
4	8	PENK	0.164
5	4	SPP1	0.164
6	20	SNAP25	0.156
7	3	CCK	0.15
8	3	RHOA	0.148
9	12	SP1	0.143
10	3	SETD2	0.142
11	12	GAPDH	0.14
12	1	RGS4	0.139
13	5	MEF2C	0.137
14	5	FGF1	0.135

(iii)

Figure 3.8 | Precise annotation of the genome-wide graph of Huntington's disease patients with varying CAG repeat lengths. Manhattan plot shows each gene's predictive normalized embedding score after applying the multi-head. Each gene node in the input file was connected to a K-mers as contextualized embedding score, encoding the likelihood of deleterious effects on that gene for a given k-mer. Edges incorporated the tissue-specific gene regulatory interactions of brain cells in these genome-wide graphs. Pooled scores are aggregated, normalized, and plotted on the y-axis. i: deleterious *htt* gene was assigned the second highest score of 0.332 for H007 with 52 CAG repeat length, just below *fan1*(Table 3.1). ii: shows the two genes with the highest scores, including *fan1* HD015. iii: *rhoa* and *tars2* were identified as susceptibility genes.

3.4 Discussion

Here we introduced PG-GWAS, a DNA sequencing data annotation method based on a graph attention network. PG-GWAS allowed the pooling of information from genetic variant deleterious effects on the downstream gene, regulatory interaction, and epistasis within a cell and showed robustness in identifying full penetrance causal and disease-modifying genes. PG-GWAS success relied on two main underlying components: a tissue cell type-specific gene-gene regulatory network capturing interaction among genes and the skip-gram model[235] encoding deleterious effects of genetic variants in a way suitable for assembling each person's genome by augmenting the backbone regulatory network.

As explained in chapter one, we obtained the count matrix of single-cell RNA-seq data from publicly available data to construct the backbone regulatory network used for the whole genome graph of patients. Regardless, data sets, such as above, comprising data from the often healthy cells at different developmental stages from broad tissue sections, can misrepresent the dynamic of diseased cells[236]. We tried to address this shortcoming by clustering cells into relevant groups using DESC[89]. However, it is plausible that obtaining spatially resolved RNA-seq[237,238] data from each patient where possible or models of HD in tissue culture should improve the accuracy of the constructed regulatory networks.

We trained the skip-gram model of PG-GWAS on a balanced data set of known benign genetic variants, variants with function loss, and variants with clinical consequences, made into whole genomes by incorporating variants randomly into the genome of healthy individuals. This was under the assumption that the frequency of each variant was determined by evolutionary constraints (more damaging mutations occur less often). While that assumption is valid, this approach did not expose the model to a training data set representing the actual frequency of a combination of

variants and did not capture the co-causation of comorbid clinical effects for genetic variants. We were aware of this shortcoming and had trained PG-GWAS on DNA sequences of more the 70,000 patients of rare diseases with comprehensive clinical records. This dataset contained real haplotypes with real medical consequences, including comorbid phenotypes. Due to privacy issues, we could not present those data here and are in the process of identifying solutions for exporting the trained model without jeopardizing patients' privacy per the ethics approval of the third party owning that dataset.

Another point to address is using the Word2Vec algorithm for embedding genetic variants. Indeed like natural language, naturally evolved DNA sequences are molecular elements exhibiting "word" frequency, preserving the strings entropy, and preserving semantic information about the input sequence. Nonetheless, a protein made from the recipe encoded in the genome is much more than a code in the primary sequence of letters but also is a three-dimensional structure affected by order of particular amino acids affecting proper functional folding[25]. Therefore, we used neural networks that are capable of learning long-range intra-molecular dependencies. In prediction applications from primary sequence, RNNs have been shown adept in the prediction of the folding of proteins in native three-dimensional structures, successfully matching crystallographic and NMR methods[239]. Notably, the prediction of protein structures as a graph inference problem in 3D space in which residues distance define the edges of the graph[181] achieved extremely high accuracy. We expect incorporating such models into the architecture of PG-GWAS would be desirable.

Lastly, here for practical reasons, such as training time and memory limitation when feeding the sparse tensors during training (input tensors had to be zero-padded before batching due to how Cuda cores are arranged in GPUs[240]), we constrained the scope of this analysis to only the coding region of the genome. Indeed, tackling

the more challenging problem, prediction of functional consequences of non-coding variants would only be possible with enough end-to-end training data to confidently annotate the effects of these mutations on downstream function. This is conceptually possible within the PG-GWAS framework of the approach presented here. Although it is a daunting task to obtain a fully observable network of the whole genome with its enormous information load and complex semantics of irregular interaction between roughly 3 200 000 000 nucleotides (nodes) per person, effective heterogeneous network sampling is possible. This may be achieved by considering the conditional dependency of node types and link types that have proven efficient for dealing with such networks[241]. Maybe, for now, scaling the graph attention network (GAT) models to large graphs though difficult and an active area of research, surely a suite of techniques will soon be available for representing arbitrary heterogeneous graphs that can scale to whole-genome with billions of nodes of edges[242].

3.5 Methods

3.5.1 Datasets

Cohorts of patients with rare complex inheritable diseases

To investigate the robustness and predictive efficiency of PG-GWAS, we obtained two separate ethics approval from the Human Research Ethics Committee at the Royal Melbourne Hospital (HREC Project 2018.266) and recruited 29 symptomatic participants with confirmed mutant *htt* genes (CAG repeat > 36) plus 13 genetically defined NPC1 patients. Written consent was obtained for all participants in this research, and genomic DNA extracted from venous blood was used for whole-genome sequencing using the Illumina TruSeq DNA PCR-Free library preparation and Illumina HiSeq X Ten sequencer, generating 150 bp paired-end reads with a mean coverage of 30X for 95% of the genome at Australian Genome Research Facility. Sequencing reads were mapped to the RefSeq release 215[243] reference human genome hg38 using BWA-MEM[244] filtered to only include nucleotide coordinates with > 8x coverage and Phred consensus quality > 25. As described by Yun et al.,[245] variants were called using DeepVariant[149] version 1.4.

Whole genome training data set

Whole 30X genome sequences of 1572 females and 1538 males with different genetic backgrounds, including 218 Gambian Mandinka, 114 Southern Han Chinese, 113 Luhya, 112 Telugu, 112 Toscani, 111 Tamil, 111 Yoruba, 109 Dai Chinese, 109 Gujarati, 109 Esan, 107 Puerto Rican, 106 Han Chinese, 105 Colombian, 105 Iberian, 104 Japanese, 104 Punjabi, 104 British, 102 Bengali, 102 Finnish, 101 Kinh Vietnamese, 101 CEPH, 100 Gambian Jola, 100 Gambian Fula, 100 Gambian Wolof, 98 African Caribbean, 96 Mende, 91 Peruvian, 73 Mexican ancestries and 68 African ancestries were obtained from[246]. UCSC liftOver was used to convert genomic co-

ordinates to RefSeq release 215.

Gene specific k-mers corpus of pan-genomic haplotypes

1,053,623,523 genetic variants validated in the RefSNP dataset were used to form every possible unique uninterrupted set of nucleotides (haplotype) along the coding region of RefSeq release 215[243]. These haplotypes were clustered into gene-specific bags of k-mers and added as a distinct vocabulary of the corpus of pan-genomic haplotypes.

3.5.2 Genomic variants contextualized skip-gram embedding

We implemented parallelized[247] Word2Vec's[248] skip-gram model to compute haplotypes' (as defined in Section 3.5.1) contextualized embeddings within a gene-specific bag of k-mers. Under the assumption that the less frequent haplotypes are under evolutionary pressure[249], the training objective of the skip-gram model was to find haplotypes' representations in vector space such that common variants of a genomic locus are placed closer to each other to indicate their less likely deleterious effect on the gene (*i.e.,* predict the more common haplotypes for the more common context(neighboring) haplotypes). The model is trained on skip-grams of haplotypes, which are n-grams (n indicating a window range) that allow neighboring haplotypes to be skipped. Here n is set to 10 as recommended by Mikolov *et. al.,* for s small corpus[250]. To this end, we simulated a corpus comprising gene strings with different arrangements of haplotypes, where the global Minor Allele Frequency (MAF) of a variant determined the rate of occurrence of haplotypes containing genomic variants. Sequences of haploid indices (list of integers) were then transformed into tuples of words of the (haplotype, in the same window), labeled 1–positive samples and (haplotype, random haplotype from the gene corpus), labeled 0–negative samples and used for training the skip-gram model. Formally, for a gene with the arrangement

of haplotypes such as $h_1, h_2, h_3, \ldots, h_t$, the goal was to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p\left(h_{t+j} \mid h_t\right) \tag{3.1}$$

where c is the size of the training context and $p\left(h_{t+j} \mid h_t\right)$ is defined as the soft-max function:

$$p\left(h_O \mid h_I\right) = \frac{\exp\left({v'_{h_O}}^{\top} v_{h_I}\right)}{\sum_{h=1}^{H}\exp\left({v'_h}^{\top} v_{h_I}\right)} \tag{3.2}$$

Here, v_h and v'_h are the input and output vector representations of h, and H is the number of unique haplotypes in the gene corpus.

Although for computational convenience herein, the above function was approximated by negative sampling[250], a simplified version of noise contrastive estimation[251], defined as

$$\log\sigma\left({v'_{w_O}}^{\top} v_{w_I}\right) + \sum_{i=1}^{k}\mathbb{E}_{w_i\sim P_n(w)}\left[\log\sigma\left(-{v'_{w_i}}^{\top} v_{w_I}\right)\right] \tag{3.3}$$

for $f(x) = 1/(1+\exp(-x))$ and with the noise distribution $P_n(w)$ as a free parameter.

3.5.3 Graph attention network (GAT)

The inputs for the attention layer are gene nodes with a set of features, namely contextualized gene-specific haplotypes from the word2vec model

$$\mathbf{vh} = \left\{\vec{vh}'_1, \vec{vh}'_2, \ldots, \vec{vh}'_N\right\}, \vec{vh}_i \in \mathbb{R}^F \tag{3.4}$$

where N is the number of genes, and F is the number of genetic variants. This layer outputs

$$\mathbf{vh}'' = \left\{\vec{vh}''_1, \vec{vh}''_2, \ldots, \vec{vh}''_N\right\}, \vec{vh}''_i \in \mathbb{R}^{F\prime} \tag{3.5}$$

where $F\prime$ cardinality was set to 20. Initially, multiple linear transformations (dense

layer) parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$ was used for preprocessing every node, and masked attention (only attending neighboring nodes one stride away, including the node) was implemented such that the normalized importance of node i's features to node j were computed as

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k\right]\right)\right)} \tag{3.6}$$

We Incorporated multi-attention heads as above, and once normalized attention coefficients were computed for each head then, these numbers were used to output a linear combination of the features corresponding to them, named here as predictive normalized embedding score features for every node by averaging

$$\vec{h}_i'' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right) \tag{3.7}$$

3.5.4 Normalized Severity Score (NSS)

To calculate the normalized severity score as an indication of HD disease onset and severity, we used DLC4 and UHDRS, two standardized clinical measures of HD. We combined both of these scores (imputed one from the other if one was missing as described elsewhere[252]). We then accounted for the contribution of CAG length using[228], as formally defined below.

$$\mathbf{NSS} = \frac{\frac{\mathbf{dlc4} - \frac{1}{N}\sum_{i=1}^{N}(dlc4_i)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(dlc4_i - \frac{1}{N}\sum_{i=1}^{N}(dlc4_i)\right)^2}} + \frac{\mathbf{uhdrs} - \frac{1}{N}\sum_{i=1}^{N}(uhdrs_i)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(uhdrs_i - \frac{1}{N}\sum_{i=1}^{N}(uhdrs_i)\right)^2}}}{e^{9.556 - 0.1460 \times \mathbf{CAGl}}} \tag{3.8}$$

3.5.5 Linear models for conventional GWAS

For the linear mixed model, we tested for alternative hypothesis $H_1 : \beta \neq 0$ against the null hypothesis $H_0 : \beta = 0$ for each genetic variant one at a time, using one of

the three commonly used test statistics (Wald, likelihood ratio or score) by fitting univariate linear mixed mode of the following form:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta + \mathbf{u} + \boldsymbol{\epsilon}; \quad \mathbf{u} \sim \mathrm{MVN}_n\left(0, \lambda\tau^{-1}\mathbf{K}\right), \quad \boldsymbol{\epsilon} \sim \mathrm{MVN}_n\left(0, \tau^{-1}\mathbf{I}_n\right) \quad (3.9)$$

where y is a vector of n dimensions for n individuals; $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_c)$ is an $n \times c$ matrix of fixed effects; α is $c - vector$ of coefficients and intercepts; x is a vector of genetic variants; β is the effect size of the genetic variants; β is the effect size of the variant, u is vector of random effects; $\boldsymbol{\epsilon}$ is an $n - vector$ of errors; $\mapsto MVN_n$ represent the n-dimensional multivariate normal distribution; τ^- is the variance of the residual errors; λ is the ratio between the two variance components and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

3.5.6 Bayesian Sparse Linear Mixed Model

For Bayesian Sparse Linear Mixed Model, a linear model of the following form was fitted as described elsewhere[253]. Briefly,

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon} \quad (3.10)$$

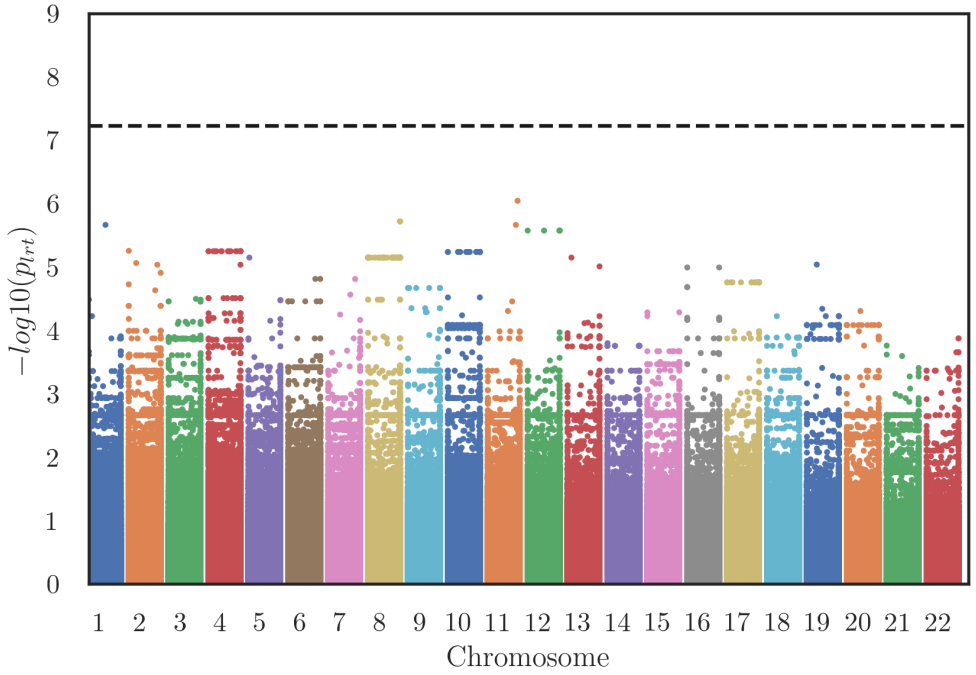
$$\beta_i \sim \pi\mathrm{N}\left(0, \sigma_a^2\tau^{-1}\right) + (1-\pi)\delta_0 \quad (3.11)$$

$$\mathbf{u} \sim \mathrm{MVN}_n\left(0, \sigma_b^2\tau^{-1}\mathbf{K}\right) \quad (3.12)$$

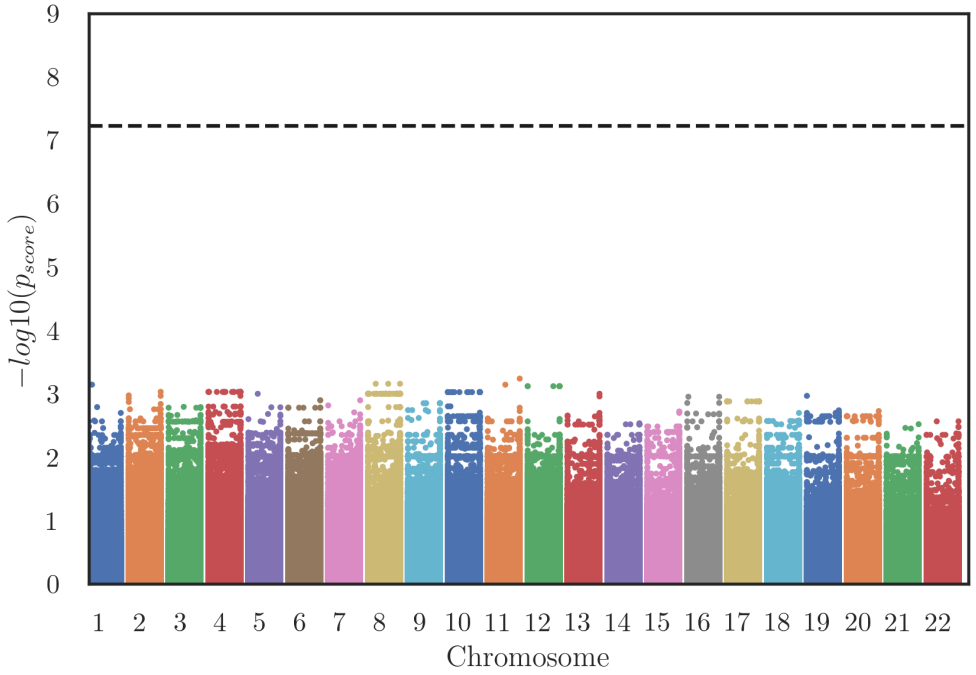
$$\boldsymbol{\epsilon} \sim \mathrm{MVN}_n\left(0, \tau^{-1}\mathbf{I}_n\right) \quad (3.13)$$

here μ is the NSS, $\mathbf{X}$ is an $n \times p$ matrix of genotypes measured on n individuals at p genetic markers.

3.6 Supplementary information



(i)



(ii)

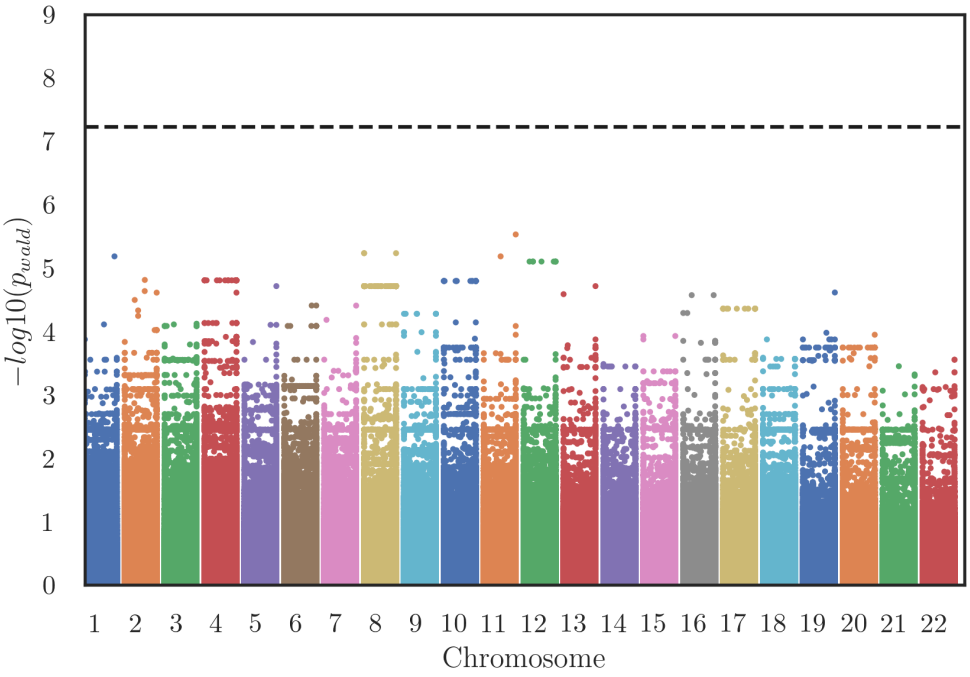
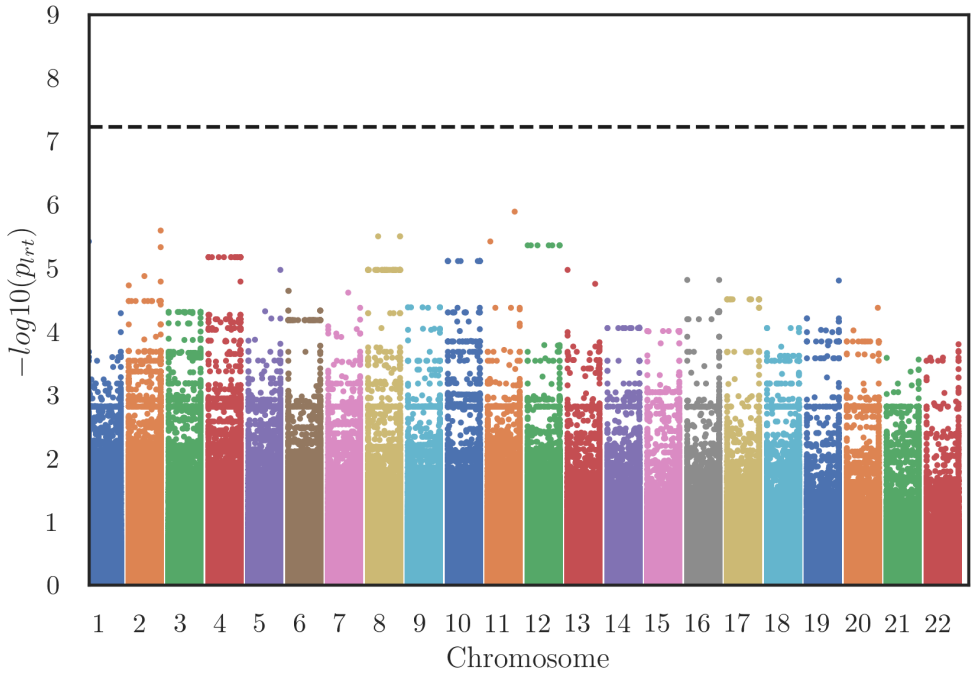
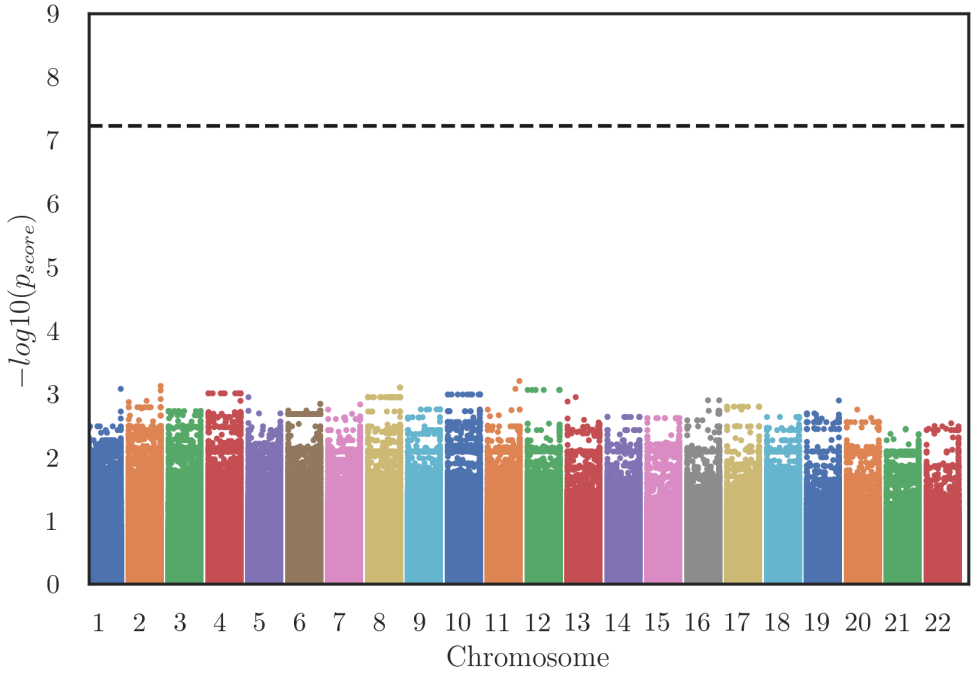


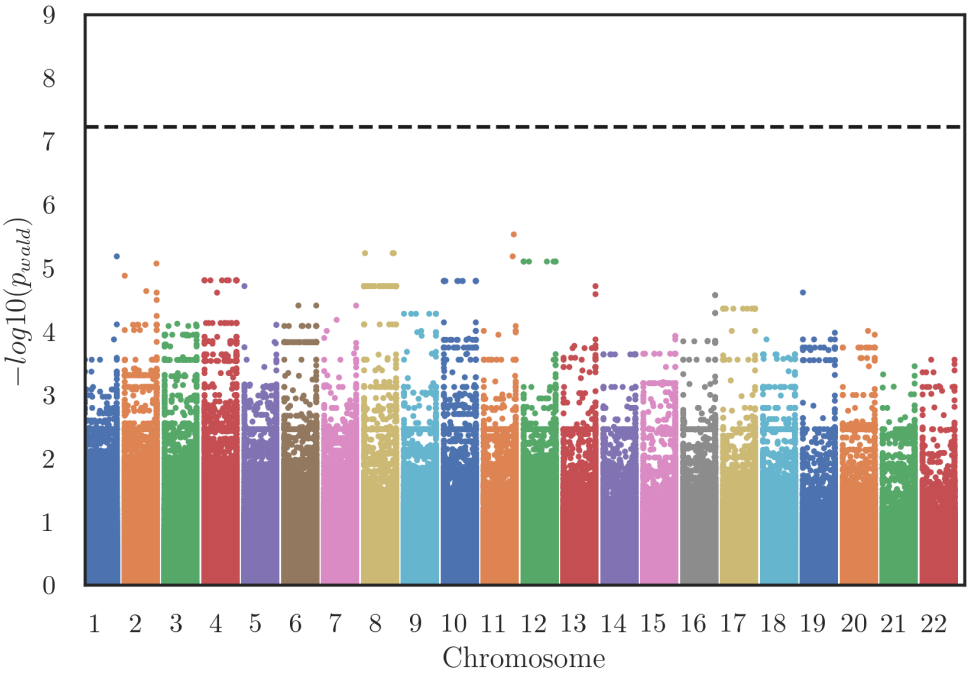
Figure S1 | linear effect of SNPs. Manhattan plot shows associations between single nucleotide polymorphisms and NSE by fitting a univariate linear model to test alternative hypothesis $H_1 : \beta \neq 0$ against the null hypothesis $H_1 : \beta = 0$ for each SNP one at a time. P_{val} using i: likelihood ratio ii:p_{score} and iii: Wald test were calculated but did not identify any SNPs at the set significance threshold (dotted black line).


9
8
7
6
5
4
3
2
1
0
$-log10(p_{lrt})$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

(i)


9
8
7
6
5
4
3
2
1
0
$-log10(p_{score})$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome

(ii)



(iii)

Figure S2 | Mixed linear effect of SNPs. Manhattan plot shows associations between single nucleotide polymorphisms and Huntington's NSE by fitting a univariate linear mixed model to test alternative hypothesis $H_1 : \beta \neq 0$ against the null hypothesis $H_1 : \beta = 0$ for each SNP in turn. i: likelihood ratio ii:p_{score} and iii: Wald were used but did not identify any SNPs at the set significance threshold (dotted black line).

Chr	Alpha	Beta	Gamma	Eff	Overlapped Gene	Type
5	2.779883E-06	0.1255081	0.0135	0.0016	RASGEF1C	Protein coding
19	2.509945E-06	0.1380574	0.00925	0.0012	ZNF724	Protein coding
12	3.565862E-06	0.3278935	0.0023	0.0007	KSR2	Protein coding
4	1.852247E-06	0.06034145	0.00972	0.0005	CAMK2D	Protein coding
19	2.14203E-06	0.3531355	0.00162	0.0005	FAM129C	Protein coding
4	1.513999E-06	0.05736716	0.00968	0.0005	SPOCK3	Protein coding
6	3.181277E-06	0.230313	0.00231	0.0005	LINC01611	None
11	2.927772E-06	0.3885618	0.00113	0.0004	DEAF1	Protein coding
9	2.031202E-06	0.139296	0.00264	0.0003	ASTN2	Protein coding
4	1.733706E-06	0.1134032	0.00288	0.0003	LEF1-AS1	None
3	2.087646E-06	0.1943299	0.00161	0.0003	RETNLB	None
6	2.448271E-06	0.2390765	0.00127	0.0003	NRN1	None
2	2.368347E-06	0.1558697	0.00182	0.0002	ANKRD44	Protein coding
1	1.530364E-06	0.1917339	0.00146	0.0002	COL24A1	None
2	2.030374E-06	0.1651637	0.00169	0.0002	BAZ2B	Protein coding
8	2.912861E-06	0.3739305	0.00073	0.0002	ZFAT	Protein coding
18	2.304442E-06	0.1322678	0.00193	0.0002	RN7SL50P	None
1	2.517616E-06	0.1864179	0.00131	0.0002	GLIS1	Protein coding
1	1.625957E-06	0.2117134	0.00109	0.0002	SMYD3	Protein coding
6	2.135392E-06	0.180409	0.00122	0.0002	LINC01626	None

Table S1 | Top 1% identified variants falling within known gene coordinates with the Bayesian Sparse Linear Mixed Model as described in the Methods.

Chr	Overlapped Gene	Beta	Se	p_wald	p_lrt	p_score	Beta	p_wald
11	TMEM80	0.4701248109	0.0775787532	2.9406899E-06	8.9809362E-07	0.0005751219	0.4701248109	2.94E-06
11	EPS8L2	0.4701248109	0.0775787532	2.9406899E-06	8.9809362E-07	0.0005751219	0.4701248109	2.94E-06
8	ZFAT	0.4596179128	0.0794653818	5.8104301E-06	1.897773E-06	0.00069388287	0.4596179128	5.81E-06
11	LRP5	0.4193406999	0.0730913877	6.5239228E-06	2.1552839E-06	0.00071734522	0.4193406999	6.52E-06
1	ST6GALNAC3	0.4193406999	0.0730913877	6.5239228E-06	2.1552839E-06	0.00071734522	0.4193406999	6.52E-06
12	KSR2	0.370497793	0.0654411763	7.8750518E-06	2.6503819E-06	0.00075782748	0.370497793	7.87E-06
2	LRP1B	0.3485625088	0.0646196827	1.537583E-05	5.5276009E-06	0.00092940527	0.3485625088	1.53E-05

Table S2 | Top 1% identified variants falling within known gene coordinates with the Linear Model as described in the Methods.

Chr	Overlapped Gene	Beta	Se	logl_H1	l_remle	l_mle	p_wald	p_lrt	p_score
11	TMEM80	0.470	0.077	18.981	9.999E-06	100000	2.9406931E-06	1.283847E-06	0.0006280
11	EPS8L2	0.470	0.077	18.981	9.999E-06	100000	2.9406931E-06	1.283847E-06	0.0006280
8	ZFAT	0.459	0.079	18.118	9.999E-06	100000	5.8104361E-06	3.1507971E-06	0.0007940
11	LRP5	0.419	0.073	17.940	9.999E-06	100000	6.5239319E-06	3.7961511E-06	0.0008358
1	ST6GALNAC3	0.419	0.073	17.940	9.999E-06	100000	6.5239319E-06	3.7961511E-06	0.0008358
12	KSR2	0.370	0.065	17.804	9.999E-06	100000	7.8750618E-06	4.37169E-06	0.0008694
2	LRP1B	0.351	0.062	18.321	21.003	100000	8.4769617E-06	2.55142E-06	0.0007501

Table S3 | Top 1% identified variants falling within known gene coordinates with the Linear Mixed Model as described in the Methods.

Chapter 4

Summary

4.0.1 Synthesis

This study aimed to develop a pipeline that allows identifying rare and common disease-modifying genetic variants from DNA sequencing data. While genome-wide association studies are the most common approach employed for such goal, GWAS methodology in the core is not equipped to handle the technical noise inherent in high-throughput sequencing platforms and is not conceptually designed to process large quantities of high-dimensional genomic data representing a complex nexus of gene regulatory networks[140,186]. Moreover, ad-hoc measures to enhance the power of GWAS have been adequate to a certain extent but introduced unique challenges[187]. For example, filtering GWAS results with pathway analysis heavily relied on off-the-shelf gene-gene interaction networks while constructing GRNs to illustrate complex interactions typically involved merging non-standardized high-throughput static datasets, resulting in a high number of false positive interactions and lacking data points or insights into cellular developmental stages[186].

Logically I set out 1) to construct an accurate Gene Regulatory Network (GRN) that encompasses all significant cis- and trans-regulatory interactions in a genome, thereby enabling epistasis modeling, 2) to develop a precise whole-genome annotation tool, which is crucial for identifying cellular functions associated with the onset and manifestation of complex heritable diseases and 3) to discover known and novel potential disease-modifying genetic variants in patients' genomes as proof of concept for my strategy.

For first aim to better understand these complex dynamic regulatory relationships within the genome, specific to tissues or cell types, the Non-Stiff Dynamic Invertible Model of CO-Regulatory Networks was created here. This model permitted unrestricted neural network structures and training, allowing for more extensive gene sets, and was successfully trained on non-homogenized bulk tissue-specific RNA-seq or single-cell RNA-seq, representing the continuous developmental states

of cells. Given that the NS-DIMCORN model utilized ordinary differential equations to simulate these highly dynamic systems by warping a multivariate distribution, a continuous-time invertible model with unbiased density estimation was generated solely based on RNA-seq read-count data, enabling time-flexible sampling of each gene's expression level for de-novo construction of cell-specific gene regulatory networks.

For the second and third aims, positional embedding techniques were employed to convert each person's genomic variations into continuous numerical vectors, assigning each genomic variant a unique context-specific score representing the likelihood of its impact on related gene products. Unique to this thesis, these scores were pan-genomic and constructed using a k-mer representation of all haplotypes, independent of any "reference genome" based solely on the evolutionary constraints of each variant. Next, using the algorithmic provisions developed and described before a graph representation of individual genomes was created, integrating genomic variation scores, tissue-specific gene-gene interactions, and regulatory networks (derived from GRNs) to collectively facilitate the study of genomic variants while accounting for epistasis. In the last chapter, Precise Graph-based Genome-Wide Annotation Software (PG-GWAS) was developed and exhibited promising results in annotating every person's genome unaffected by sequencing cohorts' statistical moments which was in line with my third aim. It is to note that the graph attention mechanism was the primary mechanism utilized in PG-GWAS to identify the most critical interactions within these networks, enabling the annotation of whole-genome graphs and the determination of the most significant genomic features (i.e., interacting gene groups) within each genome that could be responsible for varying symptoms and disease onset in patients with the same causative mutations.

4.0.2 Future directions

Although I have successfully benchmarked and demonstrated the efficacy of the proposed methodology in this thesis, I believe the following would be the next logical steps in improving the overall usability of the my method for clinical diagnosis and prognosis:

1. PG-GWAS was initially developed and tested using whole genome sequences from over 90,000 healthy and diseased individuals with comprehensive longitudinal clinical records. I hypothesized that even if a genomic variation is structurally deleterious, its pathological effects might remain benign in different individuals or exhibit varied influences on disease manifestation and onset due to the buffering or alteration by other genetic complications. My preliminary results, obtained from the aforementioned datasets, supported this hypothesis; however, I could not further evaluate these findings due to privacy concerns stipulated in the contract between Genomic England (sequencing providers and clinical data providers) and the study participants. Acquiring such a dataset is crucial for further exploration of my methodology. Alternative datasets, such as those from All of US[134] are currently available and, depending on their access policies may be utilized in future research.

2. Preliminary natural language processing embedding techniques, such as the Skip-Gram mode of Word2Vec, have been effectively employed to embed genomic variations while preserving their semantic relationships positionally[248]. In this approach, the Skip-Gram model predicts the context nucleotide within a specified window size (gene) for each target k-mer comprising the variants. This method performs well for smaller datasets and generates satisfactory embeddings for infrequent words[211,248]. Nevertheless, employing a Transformer-based deep learning architecture, which enables the processing of long-range

dependencies and context through multiple layers and many parameters, could better capture the intricacies, patterns, and structures in DNA and protein language[25]. Utilizing such a model would also enable the inclusion of non-coding regions of the genome (excluded herein for technical reasons) and the pre-training of the model with various information modalities, such as crystallography, as exemplified in the critically acclaimed AlphaFold paper[181].

3. GATs used here use attention mechanisms to determine the importance of each genetic variation when aggregating information from the local interacting genes with structural variations. In GATs, an attention mechanism is computed based on a shared linear transformation followed by a non-linear activation function that is then transformed into attention coefficients using a softmax function, ensuring standardized output[194]. The primary advantage of GATs is their ability to adaptively focus on the most relevant neighbors, allowing for the efficient handling of complex graph structures and heterogeneous node features[194]. However, the attention mechanism is computationally expensive compared to other graph convolutional networks and hence is limited by a set number of hops considered[194]. Message Passing Neural Networks (MPNNs) is another class of graph-based deep learning models that focus on the aggregation and update of node features through message passing such that at message passing phase, each node sends a message to its neighbors based on its current state and the state of the edge connecting them[254]. These messages are then aggregated for each node, and the aggregated message is combined with the node's current state during and update phase. MPNNs are thought to be more effective in capturing the complex relationships between nodes in a graph, with the long-range driver effect of one node, as they consider every edge but are not as successful as GATs in prioritizing network edges and crucial dynamic weighting of neighbors[255]. Hence, we suggest us-

ing a message-passing neural network combined with an attention mechanism that will allow more efficient and accurate prediction of topological properties of the genome-wide graph of individuals. Indeed we are aware that such architecture will impose challenging technical difficulties and computational optimizations that need to be considered in future works.

Bibliography

1. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. en. *Science* **376,** eabl3533 (Apr. 2022).
2. *DNA Packaging: Nucleosomes and Chromatin* en. https://www.nature.com/scitable/topicpage/dna-packaging-nucleosomes-and-chromatin-310/. Accessed: 2023-4-26.
3. Alberts, B. *et al. Protein Function* (Garland Science, 2002).
4. *Translation: DNA to mRNA to Protein* en. https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/. Accessed: 2023-4-26.
5. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. en. *Nat. Rev. Mol. Cell Biol.* **9,** 770–780 (Oct. 2008).
6. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. en. *Nat. Rev. Genet.* **21,** 71–87 (Feb. 2020).
7. Levental, I. & Lyman, E. Regulation of membrane protein structure and function by their lipid nano-environment. en. *Nat. Rev. Mol. Cell Biol.* **24,** 107–122 (Feb. 2023).
8. Aliperti, V., Skonieczna, J. & Cerase, A. Long Non-Coding RNA (lncRNA) Roles in Cell Biology, Neurodevelopment and Neurological Disorders. en. *Noncoding RNA* **7** (June 2021).
9. Hill, W. G. in *Brenner's Encyclopedia of Genetics (Second Edition)* (eds Maloy, S. & Hughes, K.) 432–434 (Academic Press, San Diego, Jan. 2013).

10. Molinelli, E. J. *et al.* Perturbation biology: inferring signaling networks in cellular systems. en. *PLoS Comput. Biol.* **9,** e1003290 (Dec. 2013).

11. Luo, Q., Yu, Y. & Lan, X. SIGNET: single-cell RNA-seq-based gene regulatory network prediction using multiple-layer perceptron bagging. en. *Brief. Bioinform.* **23** (Jan. 2022).

12. Yong, S. Y., Raben, T. G., Lello, L. & Hsu, S. D. H. Genetic architecture of complex traits and disease risk predictors. en. *Sci. Rep.* **10,** 12055 (July 2020).

13. Xu, Q., Yang, C. & Pei, Y.-F. Editorial: Genetic Pleiotropy in Complex Traits and Diseases. en. *Front. Genet.* **13,** 897383 (Apr. 2022).

14. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. en. *Trends Genet.* **29,** 66–73 (Feb. 2013).

15. Ratnakumar, A., Weinhold, N., Mar, J. C. & Riaz, N. Protein-Protein interactions uncover candidate 'core genes' within omnigenic disease networks. en. *PLoS Genet.* **16,** e1008903 (July 2020).

16. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. en. *Cell* **169,** 1177–1186 (June 2017).

17. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. en. *Cell* **173,** 1573–1580 (June 2018).

18. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. en. *Nat. Biotechnol.* **34,** 531–538 (May 2016).

19. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. en. *Nat. Rev. Genet.* **20,** 467–484 (Aug. 2019).

20. Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. en. *Nat Biomed Eng* (Oct. 2022).

21. Alharbi, W. S. & Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. en. *Hum. Genomics* **16,** 26 (July 2022).

22. Hancock, J. T. & Khoshgoftaar, T. M. Survey on categorical data for neural networks. en. *Journal of Big Data* **7,** 1–41 (Apr. 2020).

23. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density estimation using Real NVP. arXiv: 1605.08803 [cs.LG] (May 2016).

24. Ahmed, R. *et al.* Single-Cell RNA Sequencing with Spatial Transcriptomics of Cancer Tissues. en. *Int. J. Mol. Sci.* **23** (Mar. 2022).

25. Ofer, D., Brandes, N. & Linial, M. The language of proteins: NLP, machine learning & protein sequences. en. *Comput. Struct. Biotechnol. J.* **19,** 1750–1758 (Mar. 2021).

26. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. en. *Nat. Rev. Genet.* **21,** 243–254 (Apr. 2020).

27. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. en. *Am. J. Hum. Genet.* **97,** 199–215 (Aug. 2015).

28. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. en. *Nucleic Acids Res.* **33,** D514–7 (Jan. 2005).

29. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. en. *Nucleic Acids Res.* **38,** e164 (Sept. 2010).

30. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. en. *Nat. Commun.* **13,** 3895 (July 2022).

31. Wang, Q. *et al.* Evolution of cis- and trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. en. *BMC Genomics* **20,** 933 (Dec. 2019).

32. Sardella, M. & Belcher, G. Pharmacovigilance of medicines for rare and ultrarare diseases. en. *Ther Adv Drug Saf* **9,** 631–638 (Nov. 2018).

33. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. en. *Nat. Rev. Genet.* **20,** 631–656 (Nov. 2019).

34. Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L. & Gonçalves, J. Gene regulatory network inference from sparsely sampled noisy data. en. *Nat. Commun.* **11,** 3493 (July 2020).

35. Yu, L., Fernandez, S. & Brock, G. Power analysis for RNA-Seq differential expression studies. en. *BMC Bioinformatics* **18,** 234 (May 2017).

36. Zhao, M., He, W., Tang, J., Zou, Q. & Guo, F. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. en. *Brief. Bioinform.* **22** (Sept. 2021).

37. Singh, D., Singh, P. K., Chaudhary, S., Mehla, K. & Kumar, S. in *Advances in Genetics* (eds Friedmann, T., Dunlap, J. C. & Goodwin, S. F.) 87–121 (Academic Press, Jan. 2012).

38. Cardoso, T. F. *et al.* RNA-seq based detection of differentially expressed genes in the skeletal muscle of Duroc pigs with distinct lipid profiles. en. *Sci. Rep.* **7,** 40005 (Feb. 2017).

39. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. en. *Nat. Rev. Mol. Cell Biol.* **22,** 96–118 (Feb. 2021).

40. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. en. *Nat. Methods* **17,** 147–154 (Feb. 2020).

41. Alberch, P. Kauffman, S. A. The origins of order. Self-organization and selection in evolution. Oxford University Press (1993). Price: f17.95 (pb), f51.00 (hb). ISBN:

0-19-505811-9 (hb) and 0-19-507951-5 (pb). en. *J. Evol. Biol.* **7,** 518–519 (July 1994).

42. Borriello, E. & Daniels, B. C. The basis of easy controllability in Boolean networks. en. *Nat. Commun.* **12,** 5227 (Sept. 2021).

43. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. en. *PLoS One* **5** (Sept. 2010).

44. Woodhouse, S., Piterman, N., Wintersteiger, C. M., Göttgens, B. & Fisher, J. SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. en. *BMC Syst. Biol.* **12,** 59 (May 2018).

45. Omranian, N., Eloundou-Mbebi, J. M. O., Mueller-Roeber, B. & Nikoloski, Z. Gene regulatory network inference using fused LASSO on multiple data sets. en. *Sci. Rep.* **6,** 20533 (Feb. 2016).

46. Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C. & Huang, Y. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. en. *Bioinformatics* **34,** 964–970 (Mar. 2018).

47. Kim, S. Ppcor: An R package for a fast calculation to semi-partial correlation coefficients. en. *Commun. Stat. Appl. Methods* **22,** 665–674 (Nov. 2015).

48. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. en. *BMC Syst. Biol.* **1,** 37 (Aug. 2007).

49. Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. en. *Bioinformatics* **33,** 764–766 (Mar. 2017).

50. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. en. *Cell Syst* **5,** 251–267.e3 (Sept. 2017).

51. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. en. *Bioinformatics* **34,** 258–266 (Jan. 2018).

52. Zhao, H. & Duan, Z.-H. Cancer Genetic Network Inference Using Gaussian Graphical Models. en. *Bioinform. Biol. Insights* **13,** 1177932219839402 (Apr. 2019).

53. Bennasar, M., Hicks, Y. & Setchi, R. Feature selection using Joint Mutual Information Maximisation. *Expert Syst. Appl.* **42,** 8520–8532 (Dec. 2015).

54. Aljabbouli, H., Albizri, A. & Harfouche, A. Tree-Based Algorithm for Stable and Efficient Data Clustering. en. *Informatics* **7,** 38 (Sept. 2020).

55. Aderhold, A., Husmeier, D. & Grzegorczyk, M. Approximate Bayesian inference in semi-mechanistic models. en. *Stat. Comput.* **27,** 1003–1040 (2017).

56. Casadiego, J., Nitzan, M., Hallerberg, S. & Timme, M. Model-free inference of direct network interactions from nonlinear collective dynamics. en. *Nat. Commun.* **8,** 2192 (Dec. 2017).

57. Huynh-Thu, V. A. & Geurts, P. dynGENIE3: dynamical GENIE3 for the inference of gene networks from time series expression data. en. *Sci. Rep.* **8,** 3384 (Feb. 2018).

58. Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2,** 52–63 (June 2016).

59. Bansal, M., Della Gatta, G. & di Bernardo, D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. en. *Bioinformatics* **22,** 815–822 (Apr. 2006).

60. Penfold, C. A., Shifaz, A., Brown, P. E., Nicholson, A. & Wild, D. L. CSI: a nonparametric Bayesian approach to network inference from multiple perturbed time series gene expression data. en. *Stat. Appl. Genet. Mol. Biol.* **14,** 307–310 (June 2015).

61. Paninski, L. Estimation of entropy and mutual information. en. *Neural Comput.* **15,** 1191–1253 (June 2003).

62. Aubin-Frankowski, P.-C. & Vert, J.-P. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. en. *Bioinformatics* **36,** 4774–4780 (Sept. 2020).

63. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. en. *Bioinformatics* **35,** 2159–2161 (June 2019).

64. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. en. *Bioinformatics* **33,** 2314–2321 (Aug. 2017).

65. Qiu, X. *et al.* Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. en. *Cell Syst* **10,** 265–274.e11 (Mar. 2020).

66. Irwin, M. & Wang, Z. *Dynamic Systems Modeling* Aug. 2017.

67. Bahadorian, M. *et al.* A topology-dynamics-based control strategy for multi-dimensional complex networked dynamical systems. en. *Sci. Rep.* **9,** 19831 (Dec. 2019).

68. Sutherland, W. A. *Introduction to Metric and Topological Spaces (Oxford Mathematics)* 2nd ed. en (Oxford University Press, Oct. 2009).

69. Topirceanu, A., Udrescu, M. & Vladutiu, M. *Network Fidelity: A Metric to Quantify the Similarity and Realism of Complex Networks* in *2013 International Conference on Cloud and Green Computing* (Sept. 2013), 289–296.

70. Preciado, V. M., Jadbabaie, A. & Verghese, G. C. Structural Analysis of Laplacian Spectral Properties of Large-Scale Networks. *IEEE Trans. Automat. Contr.* **58,** 2338–2343 (Sept. 2013).

71. Lozoya, O. A., Santos, J. H. & Woychik, R. P. A Leveraged Signal-to-Noise Ratio (LSTNR) Method to Extract Differentially Expressed Genes and Multivariate Patterns of Expression From Noisy and Low-Replication RNAseq Data. en. *Front. Genet.* **9,** 176 (May 2018).

72. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. en. *Nat. Biotechnol.* **37,** 547–554 (May 2019).

73. Gilpin, L. H. *et al. Explaining Explanations: An Overview of Interpretability of Machine Learning* in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (Oct. 2018), 80–89.

74. *BoolODE — BEELINE documentation* en. https://murali-group.github.io/Beeline/BoolODE.html. Accessed: 2023-4-14.

75. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. en. *Cell* **177,** 1888–1902.e21 (June 2019).

76. Hong, M. *et al.* RNA sequencing: new technologies and applications in cancer research. en. *J. Hematol. Oncol.* **13,** 166 (Dec. 2020).

77. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. en. *Nat. Methods* **18,** 635–642 (June 2021).

78. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. en. *Genome Res.* **22,** 1760–1774 (Sept. 2012).

79. Horlbeck, M. A. *et al.* Mapping the Genetic Landscape of Human Cells. en. *Cell* **174,** 953–967.e22 (Aug. 2018).

80. Li, X. & Wang, C.-Y. From bulk, single-cell to spatial RNA sequencing. en. *Int. J. Oral Sci.* **13,** 36 (Nov. 2021).

81. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. en. *Nat. Methods* **18,** 723–732 (July 2021).

82. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. en. *Nucleic Acids Res.* **49,** D605–D612 (Jan. 2021).

83. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (Oct. 2005).

84. Mubeen, S. *et al.* The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. en. *Front. Genet.* **10,** 1203 (Nov. 2019).

85. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. en. *Cell Rep.* **38,** 110333 (Feb. 2022).

86. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

87. *cupy: NumPy & SciPy for GPU* en.

88. Narayan, A., Berger, B. & Cho, H. Assessing single-cell transcriptomic variability through density-preserving data visualization. en. *Nat. Biotechnol.* (Jan. 2021).

89. Li, X. *et al.* Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. en. *Nat. Commun.* **11,** 2338 (May 2020).

90. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426 [stat.ML] (Feb. 2018).

91. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. en. *BMC Bioinformatics* **13,** 328 (Dec. 2012).

92. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. en. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69,** 066138 (June 2004).

93. Gao, S., Ver Steeg, G. & Galstyan, A. Efficient Estimation of Mutual Information for Strongly Dependent Variables. arXiv: 1411.2003 [cs.IT] (Nov. 2014).

94. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. en. *Nature* **573,** 61–68 (Sept. 2019).

95. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. en. *Protein Sci.* **27,** 233–244 (Jan. 2018).

96. MacParland, S. A. *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. en. *Nat. Commun.* **9,** 4383 (Oct. 2018).

97. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. en. *Science* **347,** 1260419 (Jan. 2015).

98. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. en. *Science* **369,** 1318–1330 (Sept. 2020).

99. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. en. *Science* **376,** eabj5089 (Apr. 2022).

100. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. en. *Nucleic Acids Res.* **44,** D457–62 (Jan. 2016).

101. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. en. *Nucleic Acids Res.* **50,** D687–D692 (Jan. 2022).

102. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. en. *Nucleic Acids Res.* **46,** D661–D667 (Jan. 2018).

103. Ocone, A., Haghverdi, L., Mueller, N. S. & Theis, F. J. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. en. *Bioinformatics* **31,** i89–96 (June 2015).

104. Qi, J., Sun, L., Li, K. & Wang, L. Gaussian noise parameter estimation based on multiple singular value decomposition and non-linear fitting. en. *IET Image Proc.* **16,** 3025–3038 (Sept. 2022).

105. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. en. *Genome Biol.* **23,** 42 (Feb. 2022).

106. Selvaraj, S. & Natarajan, J. Microarray data analysis and mining tools. en. *Bioinformation* **6,** 95–99 (Apr. 2011).

107. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. en. *Biostatistics* **8,** 118–127 (Jan. 2007).

108. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. en. *Sci. Rep.* **9,** 5233 (Mar. 2019).

109. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. en. *Genome Biol.* **20,** 59 (Mar. 2019).

110. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. en. *Sci. Rep.* **8,** 8868 (June 2018).

111. Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell cycle and their expression in tumors. en. *Mol. Biol. Cell* **13,** 1977–2000 (June 2002).

112. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. Neural ordinary differential equations. arXiv: 1806.07366 [cs.LG] (June 2018).

113. Grathwohl, W., Chen, R. T. Q., Bettencourt, J., Sutskever, I. & Duvenaud, D. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. arXiv: 1810.01367 [cs.LG] (Oct. 2018).

114. Shampine Lawrence, f. Some practical Runge-Kutta formulas. *Math. Comput.* **46,** 135–150 (Jan. 1986).

115. Oliva, J. B. *et al.* Transformation Autoregressive Networks. arXiv: 1801.09819 [stat.ML] (Jan. 2018).

116. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. en. *Biostatistics* **9,** 432–441 (July 2008).

117. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88,** 365–411 (Feb. 2004).

118. Beskos, A., Pillai, N. S., Roberts, G. O., Sanz-Serna, J. M. & Stuart, A. M. Optimal tuning of the Hybrid Monte-Carlo Algorithm. arXiv: 1001.4460 [math.PR] (Jan. 2010).

119. Neal, R. M. MCMC using Hamiltonian dynamics. arXiv: 1206.1901 [stat.CO] (June 2012).

120. *tfp.mcmc.HamiltonianMonteCarlo* en. https://www.tensorflow.org/probability/api_docs/python/tfp/mcmc/HamiltonianMonteCarlo. Accessed: 2023-4-18.

121. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

122. Lobo, J. M., Jiménez-Valverde, A. & Real, R. AUC: a misleading measure of the performance of predictive distribution models. en. *Glob. Ecol. Biogeogr.* **17,** 145–151 (Mar. 2008).

123. Flach, P. A., Hernández-Orallo, J. & Ramirez, C. F. *A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance* Jan. 2011.

124. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27,** 861–874 (June 2006).

125. Calders, T. & Jaroszewicz, S. *Efficient AUC Optimization for Classification* in *Knowledge Discovery in Databases: PKDD 2007* (Springer Berlin Heidelberg, 2007), 42–53.

126. Jurman, G., Visintainer, R., Filosi, M., Riccadonna, S. & Furlanello, C. *The HIM glocal metric and kernel for network comparison and classification* in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (Oct. 2015), 1–10.

127. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. en. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (June 2009).

128. Fóthi, Á., Pintér, C., Pollner, P. & Lőrincz, A. Peripheral gene interactions define interpretable clusters of core ASD genes in a network-based investigation of the omnigenic theory. en. *NPJ Syst Biol Appl* **8,** 28 (Aug. 2022).

129. González-Seoane, B., Ponte-Fernández, C., González-Domínguez, J. & Martín, M. J. PyToxo: a Python tool for calculating penetrance tables of high-order epistasis models. en. *BMC Bioinformatics* **23,** 117 (Apr. 2022).

130. Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. en. *Genetics* **211,** 1125–1130 (Apr. 2019).

131. Jha, K., Saha, S. & Singh, H. Prediction of protein-protein interaction using graph neural networks. en. *Sci. Rep.* **12,** 8360 (May 2022).

132. Kreitmaier, P., Katsoula, G. & Zeggini, E. Insights from multi-omics integration in complex disease primary tissues. en. *Trends Genet.* (Sept. 2022).

133. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. en. *Nature* **562,** 203–209 (Oct. 2018).

134. All of Us Research Program Investigators *et al.* The "All of Us" Research Program. en. *N. Engl. J. Med.* **381,** 668–676 (Aug. 2019).

135. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. en. *J. Epidemiol.* **27,** S2–S8 (Mar. 2017).

136. Alipanahi, B. *et al.* Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. en. *Am. J. Hum. Genet.* **108,** 1217–1230 (July 2021).

137. Cano-Gamez, E. & Trynka, G. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. en. *Front. Genet.* **11,** 424 (May 2020).

138. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. en. *Nat. Genet.* **50,** 1112–1121 (July 2018).

139. Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. en. *Nat. Genet.* **51,** 394–403 (Mar. 2019).

140. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. en. *Am. J. Hum. Genet.* **101,** 5–22 (July 2017).

141. Siminovitch, K. A. PTPN22 and autoimmune disease. en. *Nat. Genet.* **36,** 1248–1249 (Dec. 2004).

142. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. en. *Nat. Genet.* **50,** 1219–1224 (Sept. 2018).

143. Moschen, A. R., Tilg, H. & Raine, T. IL-12, IL-23 and IL-17 in IBD: immunobiology and therapeutic targeting. en. *Nat. Rev. Gastroenterol. Hepatol.* **16,** 185–196 (Mar. 2019).

144. Wang, K. *et al.* Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. en. *Am. J. Hum. Genet.* **84,** 399–405 (Mar. 2009).

145. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. en. *Nature* **461,** 747–753 (Oct. 2009).

146. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. *Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery* en. May 2022.

147. Sun, B. B. *et al.* Genetic associations of protein-coding variants in human disease. en. *Nature,* 1–8 (Feb. 2022).

148. Zhao, B. *et al.* Common genetic variation influencing human white matter microstructure. en. *Science* **372** (June 2021).

149. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. en. *Nat. Biotechnol.* **36,** 983–987 (Nov. 2018).

150. Zeng, S. *et al.* G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers. en. *Nucleic Acids Res.* **49,** W228–W236 (July 2021).

151. Zverinova, S. & Guryev, V. Variant calling: Considerations, practices, and developments. en. *Hum. Mutat.* **43,** 976–985 (Aug. 2022).

152. Slatkin, M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. en. *Nat. Rev. Genet.* **9,** 477–485 (June 2008).

153. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nat. Genet.* **47,** 291–295 (Mar. 2015).

154. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. en. *Nat. Genet.* **36,** 512–517 (May 2004).

155. Novembre, J. *et al.* Genes mirror geography within Europe. en. *Nature* **456,** 98–101 (Nov. 2008).

156. Lawson, D. J. *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? en. *Hum. Genet.* **139,** 23–41 (Jan. 2020).

157. Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. en. *Sci Adv* **6,** eaay0328 (Apr. 2020).

158. Kerminen, S. *et al.* Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. en. *Am. J. Hum. Genet.* **104,** 1169–1181 (June 2019).

159. Wei, Z., Sun, W., Wang, K. & Hakonarson, H. Multiple testing in genome-wide association studies via hidden Markov models. en. *Bioinformatics* **25,** 2802–2808 (Nov. 2009).

160. Joo, J. W. J., Hormozdiari, F., Han, B. & Eskin, E. Multiple testing correction in linear mixed models. en. *Genome Biol.* **17,** 62 (Apr. 2016).

161. Steenwyk, J. L. *et al.* An orthologous gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. en. *Sci Adv* **8,** eabn0105 (May 2022).

162. Luck, K. *et al.* A reference map of the human binary protein interactome. en. *Nature* **580,** 402–408 (Apr. 2020).

163. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. en. *Science* **337,** 1190–1195 (Sept. 2012).

164. Rolland, T. *et al.* A proteome-scale map of the human interactome network. en. *Cell* **159,** 1212–1226 (Nov. 2014).

165. Diaz-Gallo, L.-M. *et al.* Systematic approach demonstrates enrichment of multiple interactions between non-HLA risk variants and HLA-DRB1 risk alleles in rheumatoid arthritis. en. *Ann. Rheum. Dis.* **77,** 1454–1462 (Oct. 2018).

166. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6,** 26 (2014).

167. Diaz-Gallo, L.-M., Brynedal, B., Westerlind, H., Sandberg, R. & Ramsköld, D. Understanding interactions between risk factors, and assessing the utility of the additive and multiplicative models through simulations. en. *PLoS One* **16,** e0250282 (Apr. 2021).

168. Kendler, K. S. & Gardner, C. O. Interpretation of interactions: guide for the perplexed. en. *Br. J. Psychiatry* **197,** 170–171 (Sept. 2010).

169. Clayton, D. Commentary: reporting and assessing evidence for interaction: why, when and how? en. *Int. J. Epidemiol.* **41,** 707–710 (June 2012).

170. Weinberg, C. R. Less is more, except when less is less: Studying joint effects. en. *Genomics* **93,** 10–12 (Jan. 2009).

171. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. en. *Nature* **494,** 234–237 (Feb. 2013).

172. Gilbert-Diamond, D. & Moore, J. H. Analysis of gene-gene interactions. en. *Curr. Protoc. Hum. Genet.* **Chapter 1,** Unit1.14 (July 2011).

173. Uffelmann, E. & Posthuma, D. Emerging Methods and Resources for Biological Interrogation of Neuropsychiatric Polygenic Signal. en. *Biol. Psychiatry* **89,** 41–53 (Jan. 2021).

174. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. en. *Am. J. Hum. Genet.* **94,** 559–573 (Apr. 2014).

175. Hebbar, P. & Sowmya, S. K. *Genomic Variant Annotation: A Comprehensive Review of Tools and Techniques* in *Intelligent Systems Design and Applications* (Springer International Publishing, 2022), 1057–1067.

176. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. en. *Genome Biol.* **17,** 122 (June 2016).

177. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. en. *Nucleic Acids Res.* **47,** D886–D894 (Jan. 2019).

178. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. en. *Nat. Protoc.* **10,** 1556–1566 (Oct. 2015).

179. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. en. *Nat. Genet.* **46,** 310–315 (Mar. 2014).

180. Ptitsyn, O. B. How does protein synthesis give rise to the 3D-structure? en. *FEBS Lett.* **285,** 176–181 (July 1991).

181. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. en. *Nature* (July 2021).

182. Barrio-Hernandez, I. & Beltrao, P. Network analysis of genome-wide association studies for drug target prioritisation. en. *Curr. Opin. Chem. Biol.* **71,** 102206 (Sept. 2022).

183. Gerlai, R. in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* 660–663 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).

184. Oti, M. & Brunner, H. G. The modular nature of genetic diseases. en. *Clin. Genet.* **71,** 1–11 (Jan. 2007).

185. Carter, H., Hofree, M. & Ideker, T. Genotype to phenotype via network analysis. en. *Curr. Opin. Genet. Dev.* **23,** 611–621 (Dec. 2013).

186. White, M. J. *et al.* Strategies for Pathway Analysis Using GWAS and WGS Data. en. *Curr. Protoc. Hum. Genet.* **100,** e79 (Jan. 2019).

187. Kao, P. Y. P., Leung, K. H., Chan, L. W. C., Yip, S. P. & Yap, M. K. H. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. en. *Biochim. Biophys. Acta Gen. Subj.* **1861,** 335–353 (Feb. 2017).

188. Attrill, H. *et al.* Annotation of gene product function from high-throughput studies using the Gene Ontology. en. *Database* **2019** (Jan. 2019).

189. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. en. *Nat. Biotechnol.* **40,** 1458–1466 (Oct. 2022).

190. Slim, L., Chatelain, C., Azencott, C.-A. & Vert, J.-P. Novel methods for epistasis detection in genome-wide association studies. en. *PLoS One* **15,** e0242927 (Nov. 2020).

191. Misra, G. *et al.* Genome-wide association coupled gene to gene interaction studies unveil novel epistatic targets among major effect loci impacting rice grain chalkiness. en. *Plant Biotechnol. J.* **19,** 910–925 (May 2021).

192. Chang, Y.-C. *et al.* GenEpi: gene-based epistasis discovery using machine learning. en. *BMC Bioinformatics* **21,** 68 (Feb. 2020).

193. Caylak, G., Tastan, O. & Cicek, A. E. A Tool for Detecting Complementary Single Nucleotide Polymorphism Pairs in Genome-Wide Association Studies for Epistasis Testing. en. *J. Comput. Biol.* **28,** 378–380 (Apr. 2021).

194. Veličković, P. *et al.* Graph Attention Networks. arXiv: 1710.10903 [stat.ML] (Oct. 2017).

195. Green, R. E. *et al.* A draft sequence of the Neandertal genome. en. *Science* **328,** 710–722 (May 2010).

196. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. en. *Nature* **604,** 437–446 (Apr. 2022).

197. Linck, E. & Battey, C. J. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. en. *Mol. Ecol. Resour.* **19,** 639–647 (May 2019).

198. Surendran, P. *et al.* Rare and common genetic determinants of metabolic individuality and their effects on human health. en. *Nat. Med.* (Nov. 2022).

199. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. arXiv: 1608.06993 [cs.CV] (Aug. 2016).

200. Vaswani, A. *et al.* Attention is all you need. arXiv: 1706.03762 [cs.CL] (June 2017).

201. Muggli, M. D., Alipanahi, B. & Boucher, C. Building large updatable colored de Bruijn graphs via merging. en. *Bioinformatics* **35,** i51–i60 (July 2019).

202. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. en. *Nucleic Acids Res.* **46,** D1062–D1067 (Jan. 2018).

203. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. en. *Nat. Rev. Genet.* **15,** 121–132 (Feb. 2014).

204. Krumm, N. & Hoffman, N. Practical estimation of cloud storage costs for clinical genomic data. en. *Pract Lab Med* **21,** e00168 (Aug. 2020).

205. Cook, D. E. & Andersen, E. C. VCF-kit: assorted utilities for the variant call format. en. *Bioinformatics* **33,** 1581–1582 (May 2017).

206. Pereira, R., Oliveira, J. & Sousa, M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. en. *J. Clin. Med. Res.* **9** (Jan. 2020).

207. Bonfield, J. K. *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. en. *Gigascience* **10** (Feb. 2021).

208. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. en. *Gigascience* **10** (Feb. 2021).

209. Dagum, L. & Menon, R. *OpenMP: An Industry- Standard API for Shared- Memory Programming* https://ucbrise.github.io/cs262a-spring2018/notes/OpenMP.pdf. Accessed: 2022-11-23.

210. Van Wouw, S., Viña, J., Iosup, A. & Epema, D. An Empirical Performance Evaluation of Distributed SQL Query Engines, 123–131 (Jan. 2015).

211. Ji, S., Satish, N., Li, S. & Dubey, P. Parallelizing Word2Vec in Shared and Distributed Memory. arXiv: 1604.04661 [cs.DC] (Apr. 2016).

212. Sallinen, S., Satish, N., Smelyanskiy, M., Sury, S. S. & Ré, C. *High Performance Parallel Stochastic Gradient Descent in Shared Memory* in *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (May 2016), 873–882.

213. *Protocol Buffers: Google's Data Interchange Format* en. https://opensource.googleblog.com/2008/07/protocol-buffers-googles-data.html. Accessed: 2022-11-29.

214. Nakerst, G., Brennan, J. & Haque, M. Gradient descent with momentum — to accelerate or to super-accelerate? arXiv: 2001.06472 [cs.LG] (Jan. 2020).

215. Sutskever, I., Martens, J., Dahl, G. & Hinton, G. *On the importance of initialization and momentum in deep learning* in *Proceedings of the 30th International Conference on Machine Learning* (eds Dasgupta, S. & McAllester, D.) **28** (PMLR, Atlanta, Georgia, USA, 2013), 1139–1147.

216. Gentile, C. & Warmuth, M. K. *Linear hinge loss and average margin* https://proceedings.neurips.cc/paper/1998/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf. Accessed: 2022-12-2.

217. Xiao, M. *et al.* Addressing Overfitting Problem in Deep Learning-Based Solutions for Next Generation Data-Driven Networks. en. *Proc. Int. Wirel. Commun. Mob. Comput. Conf.* **2021** (Aug. 2021).

218. Lee, J.-M. *et al.* Genetic modifiers of Huntington disease differentially influence motor and cognitive domains. en. *Am. J. Hum. Genet.* **109,** 885–899 (May 2022).

219. Ruder, S. An overview of gradient descent optimization algorithms. arXiv: 1609.04747 [cs.LG] (Sept. 2016).

220. Bentz, C., Alikaniotis, D., Cysouw, M. & Ferrer-i-Cancho, R. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. en. *Entropy* **19,** 275 (June 2017).

221. Shammas, H., Kuech, E.-M., Rizk, S., Das, A. M. & Naim, H. Y. Different Niemann-Pick C1 Genotypes Generate Protein Phenotypes that Vary in their Intracellular Processing, Trafficking and Localization. en. *Sci. Rep.* **9,** 5292 (Mar. 2019).

222. Sherva, R. & Farrer, L. A. Power and pitfalls of the genome-wide association study approach to identify genes for Alzheimer's disease. en. *Curr. Psychiatry Rep.* **13,** 138–146 (Apr. 2011).

223. Chen, Z., Boehnke, M., Wen, X. & Mukherjee, B. Revisiting the genome-wide significance threshold for common variant GWAS. en. *G3* **11** (Feb. 2021).

224. Noble, W. S. How does multiple testing correction work? en. *Nat. Biotechnol.* **27,** 1135–1137 (Dec. 2009).

225. Roos, R. A. C. Huntington's disease: a clinical review. en. *Orphanet J. Rare Dis.* **5,** 40 (Dec. 2010).

226. Gusella, J. F. & MacDonald, M. E. Huntington's disease: the case for genetic modifiers. en. *Genome Med.* **1,** 80 (Aug. 2009).

227. Unified Huntington's Disease Rating Scale: reliability and consistency. Huntington Study Group. en. *Mov. Disord.* **11,** 136–142 (Mar. 1996).

228. Langbehn, D. R., Hayden, M. R., Paulsen, J. S. & and the PREDICT-HD Investigators of the Huntington Study Group. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. en. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B,** 397–408 (Mar. 2010).

229. Ambrose, C. M. *et al.* Structure and expression of the Huntington's disease gene: evidence against simple inactivation due to an expanded CAG repeat. en. *Somat. Cell Mol. Genet.* **20,** 27–38 (Jan. 1994).

230. Airik, M. *et al.* Persistent DNA damage underlies tubular cell polyploidization and progression to chronic kidney disease in kidneys deficient in the DNA repair protein FAN1. en. *Kidney Int.* **102,** 1042–1056 (Nov. 2022).

231. Schulte, J. & Littleton, J. T. The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. en. *Curr. Trends Neurol.* **5,** 65–78 (Jan. 2011).

232. Goold, R. *et al.* FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. en. *Hum. Mol. Genet.* **28,** 650–661 (Feb. 2019).

233. Narayanan, K. L., Chopra, V., Rosas, H. D., Malarick, K. & Hersch, S. Rho Kinase Pathway Alterations in the Brain and Leukocytes in Huntington's Disease. en. *Mol. Neurobiol.* **53,** 2132–2140 (May 2016).

234. Schmidt, S. I., Blaabjerg, M., Freude, K. & Meyer, M. RhoA Signaling in Neurodegenerative Diseases. en. *Cells* **11** (May 2022).

235. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. arXiv: 1310.4546 [cs.CL] (Oct. 2013).

236. Al-Dalahmah, O. *et al.* Single-nucleus RNA-seq identifies Huntington disease astrocyte states. en. *Acta Neuropathol Commun* **8,** 19 (Feb. 2020).

237. Marx, V. Method of the Year: spatially resolved transcriptomics. en. *Nat. Methods* **18,** 9–14 (Jan. 2021).

238. Avior, Y., Sagi, I. & Benvenisty, N. Pluripotent stem cells in disease modelling and drug discovery. en. *Nat. Rev. Mol. Cell Biol.* **17,** 170–182 (Mar. 2016).

239. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. arXiv: 2006.15222 [cs.CL] (June 2020).

240. Walden, A., Zubair, M., Stone, C. P. & Nielsen, E. J. *Memory optimizations for sparse linear algebra on GPU hardware* in *2021 IEEE/ACM Workshop on Memory Centric High Performance Computing (MCHPC)* (IEEE, St. Louis, MO, USA, Nov. 2021).

241. Yang, C.-L., Kung, P.-H., Li, C.-T., Chen, C.-A. & Lin, S.-D. *Sampling heterogeneous networks* in *2013 IEEE 13th International Conference on Data Mining* (IEEE, Dallas, TX, USA, Dec. 2013).

242. Mayer, B. & Perozzi, B. *Scalable Heterogeneous Graph Sampling with GCP and Dataflow For Graph Neural Networks* en. https://cloud.google.com/blog/products/ai-machine-learning/scaling-heterogeneous-graph-sampling-gnns-google-cloud-dataflow. Accessed: 2022-12-3. July 2022.

243. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. en. *Nucleic Acids Res.* **44,** D733–45 (Jan. 2016).

244. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. *Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems* in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (May 2019), 314–324.

245. Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. en. *Bioinformatics* **36,** 5582–5589 (Jan. 2021).

246. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. en. *Nucleic Acids Res.* **48,** D941–D947 (Jan. 2020).

247. Ji, S., Satish, N., Li, S. & Dubey, P. Parallelizing Word2Vec in Shared and Distributed Memory. arXiv: 1604.04661 [cs.DC] (Apr. 2016).

248. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781 [cs.CL] (Jan. 2013).

249. Helsen, J. *et al.* Gene Loss Predictably Drives Evolutionary Adaptation. en. *Mol. Biol. Evol.* **37,** 2989–3002 (Oct. 2020).

250. Mikolov, T., Sutskever, I., View, M. & View, M. Distributed Representations of Words and Phrases and their Compositionality.

251. Gutmann, M. & Hyvärinen, A. *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models* https://proceedings.mlr.press/v9/gutmann10a/gutmann10a.pdf. Accessed: 2022-11-29.

252. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. en. *J. Stat. Softw.* **45,** 1–67 (Dec. 2011).

253. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. en. *PLoS Genet.* **9,** e1003264 (Feb. 2013).

254. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv: 1704.01212 [cs.LG] (Apr. 2017).

255. Karagiannakos, S. *Best Graph Neural Network architectures: GCN, GAT, MPNN and more* en. https://theaisummer.com/gnn-architectures/. Accessed: 2023-4-26. Sept. 2021.