A Review of Artificial Intelligence based Biological-Tree Construction: Priorities, Methods, Applications and Trends

Zelin Zang[1,2], Yongjie Xu[2], Chenrui Duan[2], Jinlin Wu[1,3], Stan Z. Li[2,†], and Zhen Lei[1,3,4†]

[1]Centre for Artificial Intelligence and Robotics (CAIR), HKISI-CAS
[2]AI Division, School of Engineering, Westlake University, Hangzhou, 310030, China
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA)
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
zangzelin@westlake.edu.cn
Correspondence: Stan.ZQ.Li@westlake.edu.cn; zhen.lei@ia.ac.cn

Abstract

Biological tree analysis serves as a pivotal tool in uncovering the evolutionary and differentiation relationships among organisms, genes, and cells. Its applications span diverse fields including phylogenetics, developmental biology, ecology, and medicine. Traditional tree inference methods, while foundational in early studies, face increasing limitations in processing the large-scale, complex datasets generated by modern high-throughput technologies. Recent advances in deep learning offer promising solutions, providing enhanced data processing and pattern recognition capabilities. However, challenges remain, particularly in accurately representing the inherently discrete and non-Euclidean nature of biological trees. In this review, we first outline the key biological priors fundamental to phylogenetic and differentiation tree analyses, facilitating a deeper interdisciplinary understanding between deep learning researchers and biologists. We then systematically examine the commonly used data formats and databases, serving as a comprehensive resource for model testing and development. We provide a critical analysis of traditional tree generation methods, exploring their underlying biological assumptions, technical characteristics, and limitations. Current developments in deep learning-based tree generation are reviewed, highlighting both recent advancements and existing challenges. Furthermore, we discuss the diverse applications of biological trees across various biological domains. Finally, we propose potential future directions and trends in leveraging deep learning for biological tree research, aiming to guide further exploration and innovation in this field.

Contents

List of Tables

List of Figures

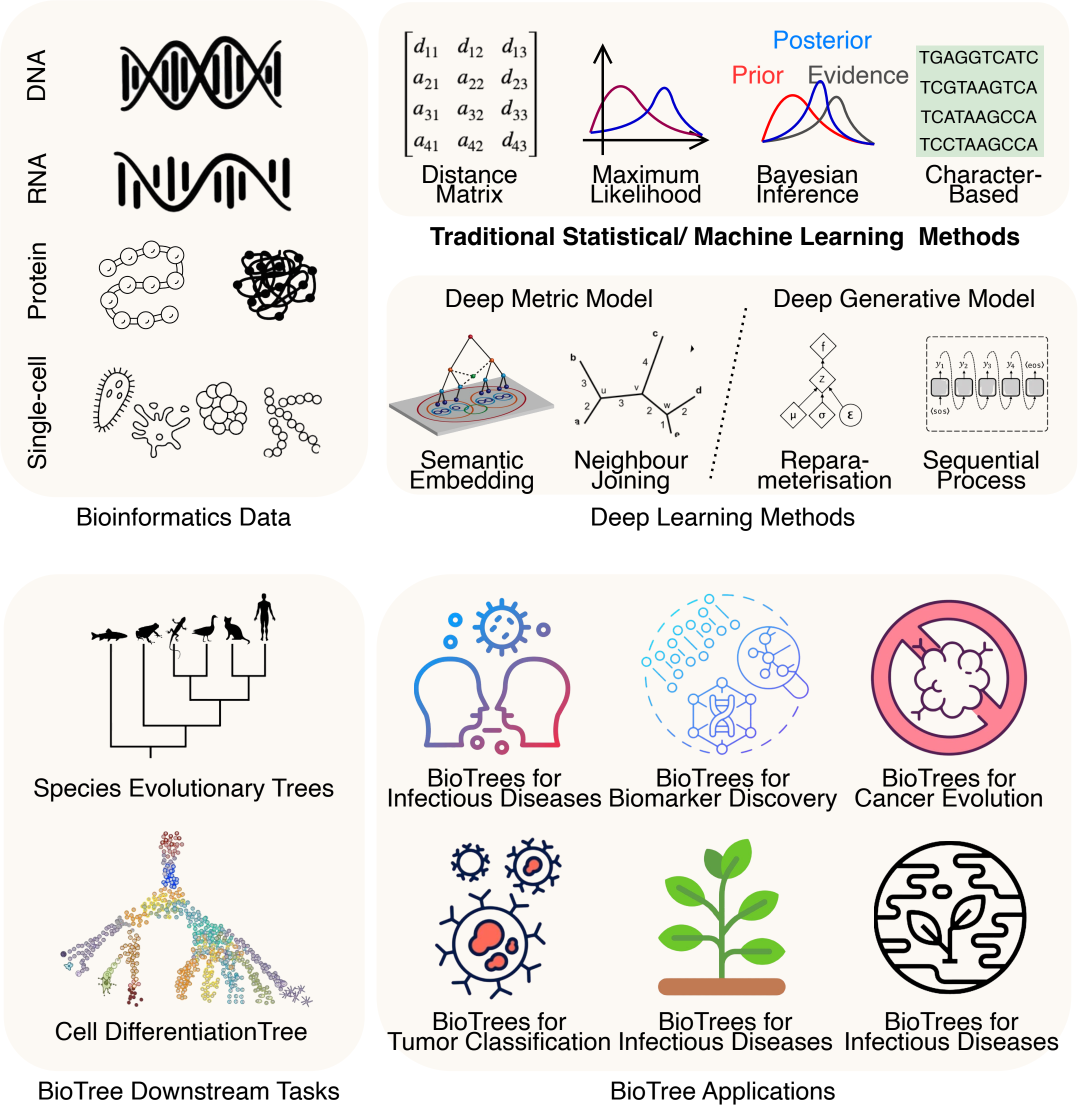

Figure 1: **The Schematic diagram of LMs for learning protein representations for PSP.** Protein structures can be leveraged as labels in supervised tasks; dashed arrows represent optional.

1 Backgrounds

Biological tree analysis methods are fundamental tools in biological research, playing a crucial role in revealing evolutionary and differentiation relationships among organisms, genes, and cells. These methods are widely used in phylogenetics, developmental biology, ecology, and medical research, helping scientists gain a deeper understanding of the origins and maintenance mechanisms of biodiversity. In **phylogenetics**, biological tree analysis involves constructing phylogenetic trees to uncover evolutionary relationships between organisms, providing a basis for taxonomists to classify and name species (Cavalli-Sforza and Edwards, 1967; Dylus et al., 2024a; Grigoriadis et al., 2024; Nei, 1987). In **developmental biology and stem cell research**, differentiation tree analysis helps researchers trace cell differentiation processes, elucidating how stem cells generate various specialized cell types (Domcke and Shendure, 2023; Trapnell et al., 2014a). In the **medical field**, biological tree analysis plays a key role in tracking the evolution of pathogens and studying the evolution and differentiation of tumor cells (Almendro et al., 2013; Graham et al., 2004; Zapatero et al., 2023).

Although traditional tree inference methods have played an important role in early research, their limitations have become increasingly apparent (Chen et al., 2023a; Delsuc et al., 2019). In **phylogenetic inference**, as the scale and complexity of modern biological data continue to grow, these traditional methods face significant challenges. They typically rely on heuristic algorithms and predefined model assumptions, performing well on small-scale datasets. However, when confronted with large-scale, multi-dimensional data generated by high-throughput sequencing technologies, these methods encounter difficulties in terms of computational complexity and scalability (McCormack et al., 2013; Szöllősi et al., 2020). Moreover, traditional methods struggle to effectively incorporate rich biological prior knowledge, particularly when dealing with complex semantics in multi-omics data, making fast and accurate tree construction a challenging task. For **differentiation tree** analysis in cell differentiation processes, current methods primarily rely on data representation and use visualization for lineage inference (Stuart et al., 2019; Wang et al., 2023). While these visualization-based methods have shown some success, they still face difficulties in generating accurate tree structures, especially when handling dynamic and complex single-cell transcriptomics and multi-omics data (Macaulay and Voet, 2017; Wagner et al., 2020; Zheng et al., 2024).

In response to these challenges, deep learning has emerged as a powerful approach to overcome the limitations of traditional methods due to its strong data processing and pattern recognition capabilities (LeCun et al., 2015a; Masoodi et al., 2023). Deep learning models, particularly Transformer models and large-scale language models, have not only achieved remarkable success in natural language processing but have also demonstrated outstanding performance in phylogenetic analysis (Jumper et al., 2021; Vaswani et al., 2017). Additionally, the powerful transfer learning capability of large-scale language models offers enhanced predictive and inferential power in genomics and protein analysis (Rives et al., 2021). Furthermore, multimodal generative models, by integrating various data types (e.g., gene sequences, protein structures, epigenomic data), provide unprecedented analytical capabilities for gene and protein phylogenetic analysis (Yang et al., 2019). It is worth emphasizing that deep learning enables the abstraction of phylogenetic and differentiation tree problems into a unified scientific framework. Although these problems focus on different levels (macro-level evolutionary relationships in phylogenetic trees versus micro-level differentiation pathways in differentiation trees), they can be addressed using similar models and methodologies.

However, using deep learning to generate tree structures remains a challenging task. While deep learning excels at representing data in Euclidean space, biological trees are inherently discrete and exhibit non-Euclidean logical structures (Chami et al., 2022). Developing deep learning models capable of capturing these complex structures is still an ongoing research problem. Current deep learning models have limitations in capturing the nonlinearity and discrete logic of biological trees, presenting a bottleneck for further advancements in this research area. To address these issues and promote the application of deep learning in biological tree research, our work makes the following contributions:

1. To facilitate the study of biological trees using deep learning and bridge the knowledge gap between deep learning researchers and biologists, we first summarize the biological priors commonly used in phylogenetic and differentiation tree analyses, helping establish a more in-depth interdisciplinary understanding (Sec. 3).
2. We systematically review the data formats and databases commonly used in biological tree analysis, providing comprehensive data resources for testing and developing new models (Sec. 2).
3. We provide a comprehensive review of traditional tree generation methods, analyzing their underlying biological priors, technical solutions, and characteristics, and summarizing their limitations in practical applications (Sec. 4).
4. We review current deep learning-based tree generation methods, summarizing recent advancements and existing challenges, offering a holistic perspective on current research directions (Sec. 5).
5. We summarize the broad applications of biological trees, highlighting their importance in fields such as phylogenetics, developmental biology, medicine, and ecology (Sec. 6).

6. Finally, we discuss the potential future directions of using deep learning for biological tree research, proposing possible research approaches and trends to guide further exploration in this field (Sec. 7).

2 Definition of the Notions and Data

2.1 Notations

Basic Terms of Data Types. Different types of biological data play a crucial role in the construction and analysis of biological trees. The following data types are frequently utilized in phylogenetic studies, each providing unique insights into the evolutionary processes and relationships being investigated.

- **Gene Sequences:** Gene sequences are the order of nucleotides in DNA or RNA that encode genetic information. They are one of the most commonly used data types in phylogenetic analysis (Li, 2017; Sanger et al., 1977).
- **Protein Sequences:** Protein sequences are chains of amino acids that build and regulate physiological processes in organisms. They are critical for studying the evolution of protein functions (Alberts et al., 2002a; Doolittle, 1981).
- **RNA Sequences:** RNA sequences are the nucleotide sequences in RNA molecules that convey and regulate genetic information, particularly significant in studying gene expression regulation and non-coding RNA (Cech and Steitz, 2014; Sharp, 1985).
- **Morphological Characteristics:** Morphological characteristics refer to the physical or structural traits of organisms, often used in phenotypic studies and classification within phylogenetic analysis (Hennig, 1966; Rieppel, 1988).
- **Single-Cell Data:** Single-cell data are sequencing or analytical data obtained from individual cells, typically used to study cell differentiation, development processes, and the cellular basis of diseases (Macosko et al., 2015; Tang et al., 2009).

Basic Terms of Algorithms and Models. The construction and analysis of biological trees require the use of various algorithms and models. Below are some of the key algorithms and models used in phylogenetic studies, each contributing to the accuracy and efficiency of tree inference.

- **Heuristic Algorithms:** Heuristic algorithms are optimization methods (Zang et al., 2019) based on empirical rules, often used to quickly generate approximate solutions but may be limited when applied to large-scale datasets.
- **Maximum Likelihood:** Maximum likelihood is a statistical method that estimates model parameters by maximizing the likelihood function given observed data, commonly used in constructing phylogenetic trees (Felsenstein, 2004).
- **Bayesian Inference:** Bayesian inference is a statistical method that updates the posterior distribution of parameters based on prior distribution and observed data, used for parameter estimation and model selection (Huelsenbeck and Ronquist, 2001).
- **Deep Learning Models:** Deep learning models are machine learning models composed of multiple layers of neural networks, excelling at handling complex pattern recognition tasks and widely applied in biological tree inference (LeCun et al., 2015b).
- **Clustering Algorithms:** Clustering algorithms partition a dataset into multiple groups or clusters,

making data points within the same cluster more similar. They have important applications in biological data classification and phylogenetic tree construction (Jain et al., 1999).

Basic Terms of Tree Concepts. Several key concepts are fundamental to understanding phylogenetic trees and the evolutionary relationships they represent. The following concepts are crucial for interpreting the structure and meaning of biological trees.

- **Common Ancestor:** A common ancestor is the earliest shared ancestor of multiple descendant species in an evolutionary tree, representing a key node in phylogenetic analysis (Maddison and Schulz, 2018).
- **Node:** A node is a point in a phylogenetic tree representing a species or evolutionary event, often used to denote the starting or ending point of divergence or evolutionary pathways (Felsenstein, 2004).
- **Branch:** A branch is a line in a phylogenetic tree that represents the relationship between an ancestor and its descendants in the evolutionary process (Felsenstein, 2004).
- **Resolution:** Resolution is the ability to distinguish between different organisms in a phylogenetic tree. High resolution means a finer distinction of evolutionary relationships (Hillis, 2019).
- **Lineage:** A lineage is a continuous pathway of evolutionary events from an ancestor to its descendants, commonly used to study the evolutionary history of species or cells (Maddison and Schulz, 2018).
- **Tree Balance:** Tree balance describes the symmetry of branch lengths or structures in a phylogenetic tree, where a balanced tree often indicates a more uniform evolutionary process (Blum and Francois, 2006).

Basic Terms of Mathematical and Statistical. Understanding the mathematical and statistical underpinnings of phylogenetic analysis is critical for interpreting results accurately. The following terms are commonly used in the quantitative aspects of biological tree inference.

- **Evolutionary Distance:** Evolutionary distance is a measure of the difference between two species or genes on an evolutionary tree, typically calculated based on gene sequence differences (Nei, 1987).
- **Support Values:** Support values are a measure of the reliability of branches in a phylogenetic tree, often obtained through bootstrap resampling (Felsenstein, 1985).
- **Topology:** Topology is the arrangement of branches and nodes in a phylogenetic tree, determining how evolutionary relationships are presented (Semple and Steel, 2003).
- **ELBO (Evidence Lower Bound):** ELBO is a key metric in variational Bayesian inference, used to approximate the lower bound of the model's log-likelihood (Blei et al., 2017).
- **KL Divergence (Kullback-Leibler Divergence):** KL divergence is an asymmetric measure of the difference between two probability distributions, often used in the design of loss functions in deep learning models (Kullback and Leibler, 1951).

2.2 Data Description

This section provides a comprehensive overview of the biological datasets used in this study, including gene data (Consortium, 2001), RNA data (Kellis et al., 2014), protein data (Consortium, 2021), and single-cell data (Zheng et al., 2017b). Each subsection details the data collection process, the technologies employed, and the format of the final datasets.

Description of Gene Data. Gene data, comprising DNA or RNA sequences, are essential for understanding the genetic basis of life and the evolutionary relationships between organisms (Consortium, 2001). These data are typically obtained through sequencing technologies (Hu et al., 2021).

The Gene data collection begins with the extraction of DNA or RNA from biological samples, which could include tissues, blood, or cell cultures (Waits and Paetkau, 2005). The extraction process typically involves lysing the cells to release nucleic acids, followed by purification steps using methods such as phenol-chloroform extraction or column-based kits like those provided by Qiagen (Barnett and Larson, 2012; Seufi and Galal, 2020). Once extracted, DNA sequencing preparation involves fragmenting the DNA into smaller pieces using sonication or enzymatic digestion (Lei et al., 2011). Adapters, which are short DNA sequences, are then ligated to both ends of each DNA fragment, serving as primers for subsequent amplification and sequencing steps (Ansorge, 2009). For RNA sequencing, the process begins with the isolation of mRNA using oligo(dT) beads that bind to the poly-A tails of eukaryotic mRNA, followed by reverse transcription into complementary DNA (cDNA) using reverse transcriptase (Ozsolak and Milos, 2011). The prepared library is then loaded onto a sequencing platform, with the choice of platform depending on the desired read length, throughput, and the complexity of the genome or transcriptome being studied (Shendure et al., 2017).

Three mainstream sequencing technologies are commonly used today. Sanger sequencing, developed in the late 1970s, uses chain-termination methods to read nucleotide sequences. Although largely replaced by high-throughput methods, it remains valuable for small-scale projects, validation, and sequencing of short DNA fragments (Sanger et al., 1977). Next-generation sequencing (NGS) technologies, such as those offered by Illumina, use massively parallel sequencing to generate large volumes of data, making them suitable for large-scale genomic studies, including whole-genome sequencing, transcriptome analysis, and targeted sequencing (Mardis, 2008). Third-generation sequencing (TGS) technologies, like Oxford Nanopore and PacBio, offer longer read lengths, which are particularly useful for resolving complex genomic regions, such as repetitive sequences and structural variants. These platforms also allow for direct RNA sequencing without the need for reverse transcription (Eid et al., 2009; Jain et al., 2016).

The final output from sequencing platforms is typically raw sequence data. A DNA sequence is mathematically represented as a string x^{gene} or x^{g} over the alphabet $\Sigma = \{A, C, G, T\}$, where each symbol corresponds to one of the four nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). For instance, a DNA sequence could be represented as x^{g}.

Description of Protein Data. Protein data, including amino acid sequences and three-dimensional structures, are critical for understanding the functional roles of proteins in biological systems (Alberts et al., 2002b).

Protein data can be represented in two primary ways: sequence data and structural data. Sequence data includes the amino acid sequences of proteins, which determine the primary structure of a protein. Proteins, as chains of amino acids, perform a wide array of functions within living organisms, including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules (Berg et al., 2002). On the other hand, structural data refers to the three-dimensional structures of proteins, which are crucial for understanding how proteins function and interact with other molecules (Alberts et al., 2002b).

The collection of protein data involves several steps. First, proteins are extracted from cells or tissues using cell lysis, followed by purification methods such as centrifugation, affinity chromatography, or ultrafiltration (Scopes, 1994). Once extracted, the proteins undergo enzymatic digestion into smaller peptides using proteolytic enzymes like trypsin, a crucial step for subsequent mass spectrometry analysis

(Olsen et al., 2006). Structural analysis is then conducted using techniques such as X-ray crystallography, cryo-electron microscopy (Cryo-EM), or nuclear magnetic resonance (NMR) spectroscopy, which provide detailed insights into the protein's function and interactions (Cheng, 2015; Rhodes, 2006; Wüthrich, 1986).

Mainstream technologies for protein analysis include mass spectrometry (MS), X-ray crystallography, and Cryo-EM. Mass spectrometry is the primary technology for identifying and quantifying proteins by analyzing peptides based on their mass-to-charge ratio, offering detailed information on protein composition and post-translational modifications (Aebersold and Mann, 2003). X-ray crystallography is the gold standard for determining the high-resolution three-dimensional structures of crystallized proteins, allowing atomic-level visualization (Rhodes, 2006). Cryo-EM, an advanced technique for studying large protein complexes and membrane proteins that are difficult to crystallize, offers near-atomic resolution without the need for crystallization (Cheng, 2015).

Protein data is typically stored in specific formats depending on whether it is sequence or structural data. Sequence data is stored in FASTA format, similar to DNA and RNA sequences. Each protein sequence can be represented as a string x^{protein} or x^{p}, where $x^{\text{protein}} = \{s_1 s_2 \ldots s_n\}$, and $s_i \in \Sigma$ for $i = \{1, 2, \ldots, n\}$, with $\Sigma = \{\text{A}, \text{C}, \text{D}, \text{E}, \ldots, \text{Y}\}$ representing the set of 20 standard amino acids. The sequence is prefixed by a header line beginning with a '>' symbol, followed by a description (Pearson and Lipman, 1990). Structural data, on the other hand, is stored in Protein Data Bank (PDB) format. A protein structure is represented by a set of atomic coordinates $C = \{(x_i, y_i, z_i)\}$, where $(x_i, y_i, z_i) \in \mathbb{R}^3$ denotes the 3D coordinates of the i-th atom in the protein. These coordinates are essential for studying protein function and interactions (Berman et al., 2000b).

Description of Single-Cell Data. Single-cell data allow researchers to explore cellular heterogeneity and study processes like differentiation and development at the single-cell level.

Single-cell data refers to the information obtained from analyzing individual cells, as opposed to bulk cell populations. This approach allows for the investigation of cellular heterogeneity, gene expression dynamics, and the identification of rare cell populations. Single-cell data can include a variety of omics layers, such as transcriptomics (RNA), genomics (DNA), epigenomics (e.g., chromatin accessibility), and proteomics (protein expression) (Kolodziejczyk et al., 2015). The collection of single-cell data involves isolating individual cells and then performing various omics analyses at the single-cell level. Cell isolation is typically achieved using techniques such as fluorescence-activated cell sorting (FACS), microfluidics (e.g., 10x Genomics Chromium), or droplet-based methods (Tung et al., 2017). After isolation, RNA is extracted from each cell and reverse-transcribed into cDNA, which is then amplified and sequenced to provide a transcriptomic profile for each individual cell (Zheng et al., 2017c). For chromatin accessibility analysis, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is adapted to single-cell analysis to identify regulatory elements within the genome (Buenrostro et al., 2015). Additionally, single-cell multi-omics techniques such as CITE-seq and ASAP-seq integrate multiple omics layers, combining scRNA-seq with surface protein expression (proteomics) or chromatin accessibility (epigenomics), enabling a more comprehensive view of cellular states and functions (Mimitou et al., 2021; Stoeckius et al., 2017).

Mainstream technologies for single-cell analysis include scRNA-seq, CITE-seq, and ASAP-seq. Technologies like 10x Genomics Chromium enable the capture and sequencing of RNA from individual cells, providing a high-resolution view of gene expression at the single-cell level, which is particularly valuable for studying cellular heterogeneity, developmental processes, and the cellular basis of diseases (Zheng et al., 2017c). CITE-seq combines scRNA-seq with surface protein profiling using oligonucleotide-tagged antibodies, allowing simultaneous measurement of transcriptomes and protein expression (Stoeckius et al., 2017). ASAP-seq integrates chromatin accessibility profiling with protein expression data, offering insights into the regulatory landscape of individual cells (Mimitou et al., 2021).

Single-cell data is typically stored in formats that accommodate the complex and high-dimensional nature of the data. As with other sequencing data, raw single-cell RNA sequencing reads are stored in FASTQ format, containing nucleotide sequences and associated quality scores. Formally, a read can be represented as a tuple (s, q), where $s \in \Sigma^n$ is the nucleotide sequence (with $\Sigma = \{A, C, G, T, N\}$) and $q \in \mathbb{Z}^n$ represents the quality scores for each base (Cock et al., 2010). After alignment to a reference genome or transcriptome, reads are stored in BAM or SAM formats. In these formats, each aligned read can be represented as a tuple (s, a), where a represents the alignment information, including the reference position and mapping quality. For ATAC-seq data, BAM files also store information on chromatin accessibility (Li et al., 2009).

In multi-omics single-cell data, each individual data point can be represented as x^{s}, where $x^{\mathrm{s}} \in \mathbb{R}^O$ captures the measured values across the different omics layers for a specific gene i in a specific cell j. Here, $x^{\mathrm{s}} = \{x_1, x_2, \ldots, x_O\}$ corresponds to the measurements for the same gene-cell pair across the O omics layers, such as transcriptomics, proteomics, and epigenomics. This compact representation allows for the integration and analysis of multiple biological dimensions within a single framework.

2.3 Commonly Used Data Set

The advancement of biological research has been greatly facilitated by the development of comprehensive datasets, which are indispensable for understanding complex biological systems. These datasets can be classified into four primary categories: gene-related, protein-related, single-cell, and image-based datasets. Each category contributes uniquely to various fields, offering resources that enable deeper insights into genetic variation, protein structure and function, cellular heterogeneity, and biodiversity.

Gene-Related Datasets. Gene-related datasets are foundational for exploring genetic variation, gene expression, and genomic annotations. The **dbSNP** database (Sherry et al., 2001) provides an extensive collection of over 150 million single nucleotide polymorphisms (SNPs) and is integral to studies of genetic variation and genome-wide association studies. Similarly, the **Gene Expression Omnibus (GEO)** (Edgar et al., 2002) offers a vast repository of gene expression datasets, allowing researchers to explore gene regulation and expression patterns across different species and conditions.

The **Human Microbiome Project (HMP)** (Consortium, 2012) is another crucial resource, advancing our understanding of the microbial communities associated with human health and disease. Meanwhile, the **Genotype-Tissue Expression (GTEx) Project** (Consortium, 2013) provides gene expression data across various human tissues, helping to uncover the relationship between genetic variation and gene expression. Furthermore, large-scale efforts like **The Cancer Genome Atlas (TCGA)** (Network, 2013) have significantly contributed to cancer research by offering comprehensive genomic profiles of multiple cancer types, aiding in the identification of molecular alterations. In population genetics, the **1000 Genomes Project** (Consortium, 2015) has been instrumental in providing whole-genome sequencing data from diverse populations, essential for understanding global genetic diversity.

Other key datasets include **Ensembl Genomes** (Kersey et al., 2018), which offers genome annotations across multiple species, and the **Genome Aggregation Database (gnomAD)** (Karczewski et al., 2020), which aggregates exome and genome data, providing crucial allele frequency information for variant interpretation in both research and clinical contexts.

Protein-Related Datasets. Understanding protein structure, function, and interactions is central to many biological processes, and protein-related datasets are critical in this context. The **Protein Data Bank (PDB)** (Berman et al., 2000a) is a fundamental resource containing a vast collection of 3D structures of proteins and nucleic acids, making it indispensable for structural biology and drug discovery efforts. Additionally, **PeptideAtlas** (Desiere et al., 2006) curates peptides identified through mass spectrometry, supporting large-scale proteomics research and protein expression studies.

For the study of protein-protein interactions, the **STRING** database (Szklarczyk et al., 2019) provides essential data on known and predicted interactions, facilitating the construction of protein interaction networks. **UniProt** (Consortium, 2019), the most comprehensive protein sequence and functional information repository, is critical for protein annotation and functional studies, offering insights into the biological roles of proteins across species.

Single-Cell Datasets. The emergence of single-cell datasets has revolutionized the understanding of cellular heterogeneity and dynamic processes at the single-cell level. Single-cell transcriptomics, particularly from **10x Genomics** (Zheng et al., 2017a), provides high-resolution gene expression data, enabling in-depth analyses of individual cell populations and their roles in tissue development and disease. The **Human Cell Atlas (HCA)** (Regev et al., 2017), aiming to create comprehensive reference maps of all human cells, serves as a vital resource for exploring cellular states and types, contributing to our understanding of human biology at an unprecedented scale.

Image-Based Datasets. Image-based datasets are pivotal for integrating computational methods Zang et al. (2023a) with biological research, particularly in biodiversity and taxonomy studies. For example, the **iNaturalist 2021 Dataset (iNat21)** (iNaturalist, 2021) leverages citizen science by compiling millions of organism images, making it an invaluable tool for biodiversity monitoring and species identification. DNA barcoding entries from **BIOSCAN-1M** (Gharaee et al., 2024) further enhance biodiversity research by enabling the mapping of global species diversity, supporting ecological studies and species discovery.

The **Encyclopedia of Life (EOL)** (of Life , EOL) aggregates taxonomic data, including images, to aid in biodiversity conservation efforts, while the **TREEOFLIFE-10M** dataset (Stevens et al., 2024) integrates image data with phylogenetic information, fostering advancements in computational biology and evolutionary studies.

The collection and integration of these diverse datasets have dramatically accelerated advancements in biological research. Gene-related datasets have facilitated the exploration of genetic variation and gene expression, while protein-related datasets provide critical insights into protein function and structure. Single-cell datasets have uncovered the complexity of cellular heterogeneity, and image-based datasets are instrumental in biodiversity monitoring and species identification. Together, these resources continue to drive discoveries in genomics, proteomics, and evolutionary biology, offering unprecedented opportunities for future research across multiple disciplines.

3 Definition and Prior Knowledge of the Tree Construction Problem

3.1 Definition of the Tree Construction Problem

Definition 1 (Tree Construction Problem)**.** Given a set $S = \{s_1, s_2, \ldots, s_n\}$ of biological entities and a corresponding set $A = \{a_1, a_2, \ldots, a_n\}$ of attribute data, the objective is to construct a tree structure $T = (V, E, L)$ that satisfies the following conditions, :

$$\begin{aligned} V &= \{v_1, v_2, \ldots, v_m\}, \\ E &= \{e_1, e_2, \ldots, e_{m-1}\}, \\ L : E &\to \mathbb{R}^+. \end{aligned}$$

Table 1: Overview of Key Datasets for Biological Research.

	Dataset Name	#Entries	Reference	URL
Gene	dbSNP	150M	Sherry et al. (2001)	https://www.ncbi.nlm.nih.gov/snp/
	GEO	100k	Edgar et al. (2002)	https://www.ncbi.nlm.nih.gov/geo/
	HMP	2.2k	Consortium (2012)	https://hmpdacc.org/
	GTEx Project	17k	Consortium (2013)	https://gtexportal.org/
	TCGA	20k	Network (2013)	https://www.cancer.gov/tcga
	Genomes Project	2,504	Consortium (2015)	https://www.internationalgenome.org/
	Ensembl Genomes	200k	Kersey et al. (2018)	https://ensemblgenomes.org/
	gnomAD	125k	Karczewski et al. (2020)	https://gnomad.broadinstitute.org/
Protein	Protein Data Bank	180k	Berman et al. (2000a)	https://www.rcsb.org/
	PeptideAtlas	2M	Desiere et al. (2006)	http://www.peptideatlas.org/
	STRING	9.6M	Szklarczyk et al. (2019)	https://string-db.org/
	UniProt	564M	Consortium (2019)	https://www.uniprot.org/
Single Cell	10x Genomics	1.3M	Zheng et al. (2017a)	https://www.10xgenomics.com/
	Human Cell Atlas	2B	Regev et al. (2017)	https://www.humancellatlas.org/
Image	iNat21	2.7M	iNaturalist (2021)	https://www.inaturalist.org/
	BIOSCAN-1M	1M	Gharaee et al. (2024)	https://www.bioscan.org/
	EOL	6.6M	of Life (EOL)	https://eol.org/
	TREEOFLIFE-10M	10.4M	Stevens et al. (2024)	https://imageomics.github.io/bioclip

where V is the set of nodes, E is the set of edges, and L is a function mapping edges to positive real numbers, where V includes both the entities in S and their inferred common ancestors or differentiated states, E connects the nodes in V, with each edge $e_k = (v_i, v_j)$ linking two nodes v_i and v_j, and L assigns a positive real number $l(e_k)$ to each edge, representing the evolutionary distance, time, or differentiation progression between the connected nodes. The tree T must be acyclic, connected, and optimize an objective function $F(T)$. In this definition, The tree structure is a connected, acyclic graph, and the objective function $F(T)$ can be likelihood, parsimony, or total branch length, depending on the specific application.

Definition 2 (Tree Construction Problem for Evolutionary Trees)**.** In the context of evolutionary trees, S represents species, genes, or proteins, and $L(e_k)$ typically represents evolutionary distance or time. The objective function $F(T)$ focuses on likelihood under a specific evolutionary model, parsimony, or total branch length.

Definition 3 (Tree Construction Problem for Differentiation Trees)**.** For differentiation trees, S represents cells or developmental states, and $L(e_k)$ represents differentiation progression. The objective function $F(T)$ aims to describe differentiation pathways or capture the most parsimonious progression.

The objective function $F(T)$ and the constraints for tree construction are determined by prior knowledge. For evolutionary trees, prior knowledge may include evolutionary models, fossil records, or molecular data. For differentiation trees, prior knowledge may include developmental biology insights, gene expression data, or known cell lineage relationships. These prior knowledge sources guide the formulation of the objective function and constraints to ensure that the resulting tree accurately reflects the underlying biological processes.

Table 2: Summary of Prior Knowledge for Phylogenetic Tree Construction: Gene Data

Prior	Descriptions	Prior Form	Knowledge Involved	References
G1	Conserved Genomic Regions	Indicator function $I(x_i^g, x_j^g)$	Regions that are relatively unchanged across species, indicating evolutionary relationships	Mount (2004), Notredame (2007), Altschul et al. (1990)
G2	Evolutionary Substitution	Transition probability matrix $P(t)$	Describes probabilistic changes in nucleotide sequences over time	Felsenstein (1981), Kimura (1980), Tavaré (1986)
G3	Genomic Linear Order of Genes	Permutation vector π	Specific order of genes along chromosomes, providing clues about evolutionary relationships	Saitou and Nei (1987), Fitch (1971)
G4	Ancestral Relationship Information	Ancestral matrix A	Known or inferred relationships between species based on shared ancestors	Maddison and Maddison (2007), Ronquist and Huelsenbeck (2003)
G5	Sequence Homology Information	Similarity matrix H	Shared ancestry between pairs of genes or sequences, critical for accurate inference	Thompson et al. (1994a), Smith and Waterman (1981a)
G6	Gene Duplication and Loss Events	Probabilistic model $P(T \mid \text{duplication, loss})$	Models gene duplication and loss events, impacting tree topology	Hahn (2009), Gu et al. (2005)
G7	Taxonomic Classification Constraints	Taxonomy tree $\mathcal{T}$	Known hierarchical relationships among species, ensuring consistency with classification	Faith (1992), Hennig (1965), Swofford et al. (1996)

3.2 Prior Knowledge

3.2.1 Prior Knowledge for Gene Phylogenetic Tree Construction

When constructing phylogenetic trees using gene sequence data, it is essential to leverage various forms of prior knowledge to enhance the accuracy and reliability of the inferred trees. This section discusses seven key types of prior knowledge, providing both a biological basis and formal mathematical descriptions, along with relevant references.

Prior G1 **Conserved Genomic Regions** Conserved regions within gene sequences are sequences that have remained relatively unchanged across different species over evolutionary time. These regions are typically under strong selective pressure, meaning that mutations in these regions are often deleterious and thus purged by natural selection. Such conserved regions can serve as reliable indicators of evolutionary relationships. Mathematically, conserved regions can be represented using an indicator function $I(x_i^g, x_j^g)$, where

$$I(x_i^g, x_j^g) = \begin{cases} 1, & \text{if sequences } x_i^g \text{ and } x_j^g \text{ share conserved regions} \\ 0, & \text{otherwise} \end{cases}$$

The similarity between these conserved regions can be quantified by the following equation:

$$d_{\text{conserved}}(x_i^g, x_j^g) = \sum_{k=1}^{L} I(x_{i,k}^g, x_{j,k}^g) \cdot d(x_{i,k}^g, x_{j,k}^g),$$

where L represents the length of the sequences, and $d(x_{i,k}^g, x_{j,k}^g)$ is a distance metric, such as Hamming distance or Jukes-Cantor distance. This approach allows for a focused analysis on regions critical to

Table 3: Summary of Prior Knowledge for Phylogenetic Tree Construction: Protein Structure and Sequence Data

Prior	Descriptions	Prior Form	Knowledge Involved	References
P1	Conserved Protein Domains	Indicator function $I(d_i^p, d_j^p)$	Conserved regions within protein sequences, indicating functional importance	Murzin et al. (1995), Marchler-Bauer et al. (2011)
P2	Evolutionary Models for Amino Acid Substitution	Substitution matrix Q	Describes the rate of amino acid substitutions over evolutionary time	Jones et al. (1992), Dayhoff et al. (1978)
P3	Protein Secondary Structure Information	Similarity matrix S	Conserved secondary structures like alpha-helices and beta-sheets	Kabsch and Sander (1983), Chothia and Finkelstein (1984)
P4	Tertiary Structure Conservation	RMSD (Root-Mean-Square Deviation)	3D structure, which is often more conserved than the primary sequence	Sali and Blundell (1994)
P5	Functional Site Conservation	Function $F(x_i^p, x_j^p)$	Conservation of critical functional sites in proteins	Bartlett et al. (2002), Thornton et al. (2000)
P6	Protein Family Classification	Classification $\mathcal{C}$	Groups proteins based on sequence and structural similarity, reflecting evolutionary origins	Bateman et al. (2002), Finn et al. (2016)
P7	Co-Evolutionary Relationships	Co-evolution matrix C	Captures the functional interdependencies of proteins through co-evolution	Göbel et al. (1994), Marks et al. (2011)

evolutionary divergence, enhancing the accuracy of tree construction (Altschul et al., 1990; Mount, 2004; Notredame, 2007).

Prior G2 **Evolutionary Substitution Models** Evolutionary models are used to describe the probabilistic changes in nucleotide sequences over time. These models take into account the rates at which one nucleotide substitutes for another and the likelihood of various evolutionary events occurring. For example, the JC69 model, a simple and commonly used model, assumes equal probabilities for all possible nucleotide substitutions and a constant rate of mutation over time. The evolutionary process can be described using a transition probability matrix $P(t)$, which specifies the probability of observing a particular sequence at time t given an ancestral sequence:

$$P(t) = \frac{1}{4} + \frac{3}{4}e^{-\mu t} \cdot I,$$

where μ is the mutation rate, and I is the identity matrix. This mathematical framework allows for the estimation of evolutionary distances between sequences, which is crucial for accurate phylogenetic inference (Felsenstein, 1981; Kimura, 1980; Tavaré, 1986).

Prior G3 **Genomic Linear Order of Genes** The linear structure of the genome, which refers to the specific order of genes along chromosomes, provides important context for phylogenetic analysis. In many cases, the relative positions of genes have been conserved throughout evolution, and this order can offer clues about evolutionary relationships. This structure can be mathematically represented using a permutation vector $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$, where each π_i indicates the position of the i-th gene in the sequence. The similarity based on genomic linear structure can be evaluated by:

$$d_{\text{linear}}(x_i^g, x_j^g) = \sum_{k=1}^{n} \delta(\pi_i(k), \pi_j(k)),$$

Table 4: Summary of Prior Knowledge for Phylogenetic Tree Construction: Single-Cell Multimodal Data

Prior	Descriptions	Prior Form	Knowledge Involved	References
S1	Gene Expression Profiles	Expression matrix E	Abundance of mRNA transcripts in single cells, indicating functional state	Trapnell et al. (2014c), Qiu et al. (2017a)
S2	RNA Velocity	Velocity vector v_i^c	Estimates the future state of individual cells based on RNA transcriptional changes	La Manno et al. (2018b), Bergen et al. (2020a)
S3	Cell Type-Specific Marker Genes	Binary matrix B	Genes uniquely expressed in specific cell types, used to identify cell identity	Tirosh et al. (2016), Plass et al. (2018)
S4	Pseudotime Ordering	Pseudotime scalar τ_i^c	Orders cells along a continuous trajectory representing differentiation progress	Trapnell et al. (2014c), Haghverdi et al. (2016)

where δ is the Kronecker delta function, which equals 1 if the gene order is the same in both sequences at position k and 0 otherwise. This method captures the evolutionary signal inherent in the gene order, which is particularly relevant for analyzing synteny (conservation of gene order) across species (Fitch, 1971; Saitou and Nei, 1987).

Prior G4 **Ancestral Relationship Information** Ancestral information refers to known or inferred relationships between species based on shared common ancestors. This information can provide strong guidance when constructing a phylogenetic tree, particularly in cases where fossil records or other historical data are available. Ancestral relationships can be encoded in an ancestral matrix A, where A_{ij} denotes the probability that species i and j share a common ancestor. The tree construction process can then be influenced by this prior knowledge, enhancing the robustness of the inferred tree topology:

$$P(\text{Tree} \mid A) = \prod_{i,j} P(\text{Tree} \mid A_{ij}) \cdot P(\text{Tree}),$$

where $P(\text{Tree} \mid A_{ij})$ is the probability of the tree given the ancestral information. Incorporating ancestral information can significantly improve the accuracy of tree reconstruction, especially in clades with well-documented evolutionary histories (Maddison and Maddison, 2007; Ronquist and Huelsenbeck, 2003).

Prior G5 **Sequence Homology Information** Homology refers to the existence of shared ancestry between a pair of genes or sequences. Homologous genes can be orthologous (derived from a common ancestor and separated by a speciation event) or paralogous (derived from a common ancestor but separated by a duplication event). Homology is a fundamental concept in evolutionary biology and is critical for accurate phylogenetic inference. The degree of homology between sequences can be represented by a similarity matrix H, where H_{ij} denotes the homology score between sequences x_i^g and x_j^g. This score can be transformed into a distance metric for phylogenetic analysis:

$$d_{\text{homology}}(x_i^g, x_j^g) = -\log(H_{ij}),$$

where the logarithm transformation helps to linearize the relationship between homology and evolutionary distance. Utilizing homology information allows for a more accurate reconstruction of evolutionary relationships, especially when dealing with complex gene families (Smith and Waterman, 1981a; Thompson et al., 1994a).

Prior G6 **Gene Duplication and Loss Events** Gene duplication and loss events are significant evolutionary processes that can greatly influence the structure of gene families and, by extension, the topology of a phylogenetic tree. Gene duplication leads to the creation of gene copies, which may diverge and acquire new functions, while gene loss can result in the elimination of certain genes in specific lineages. These events can be modeled probabilistically in the tree construction process. For a given tree topology T, the likelihood of observing a particular pattern of duplications and losses can be described by:

$$P(T \mid \text{duplication, loss}) = \prod_{d \in D} p_d \cdot \prod_{l \in L} p_l,$$

where p_d and p_l represent the probabilities of duplication and loss events, respectively, and D and L are the respective sets of these events. This probabilistic framework enables the incorporation of these complex evolutionary events into the phylogenetic analysis, which is particularly useful for studying the evolution of gene families across different species (Gu et al., 2005; Hahn, 2009).

Prior G7 **Taxonomic Classification Constraints** Taxonomic information includes known hierarchical relationships among species, such as their classification into orders, families, genera, and species. This information is often well-established based on morphological, genetic, and other types of data. Integrating taxonomic knowledge into phylogenetic tree construction provides a framework to ensure that the resulting tree is consistent with established classifications. This can be formalized by representing the taxonomy as a tree $\mathcal{T}$, and incorporating it into the tree construction process as follows:

$$P(\text{Tree} \mid \mathcal{T}) = P(\text{Tree} \mid \text{Taxonomic Constraints}) \cdot P(\text{Tree}),$$

where $P(\text{Tree} \mid \text{Taxonomic Constraints})$ is the probability of the tree given the taxonomic constraints. This approach ensures that the generated phylogenetic tree is consistent with existing taxonomies while allowing for inference in cases where taxonomic information is incomplete (Felsenstein, 2004; Hillis and Huelsenbeck, 1992).

3.2.2 Prior Knowledge for Phylogenetic Tree Construction using Protein Structure and Sequence

When constructing phylogenetic trees using protein sequences and structures, leveraging prior knowledge can significantly enhance the accuracy of the resulting trees. This section discusses key types of prior knowledge, providing both biological context and formal mathematical descriptions, along with relevant references.

Prior P1 **Conserved Protein Domains.** Conserved protein domains are specific regions within protein sequences that are preserved across different species due to their critical functional roles. These domains are often associated with essential biological functions and tend to be less variable over evolutionary time. The conservation of these domains can be mathematically represented using an indicator function $I(d_i^p, d_j^p)$, where $I(d_i^p, d_j^p) = 1$ if domains d_i^p and d_j^p are conserved across sequences, and $I(d_i^p, d_j^p) = 0$ otherwise. The similarity between conserved domains can be expressed as:

$$d_{\text{domain}}(x_i^p, x_j^p) = \sum_{k=1}^{M} I(d_{i,k}^p, d_{j,k}^p) \cdot d(d_{i,k}^p, d_{j,k}^p), \tag{1}$$

where M is the number of domains, and $d(d_{i,k}^p, d_{j,k}^p)$ is the distance metric between corresponding domains $d_{i,k}^p$ and $d_{j,k}^p$. This metric allows for the focused analysis of regions that are crucial for the protein's function and evolutionary history (Marchler-Bauer et al., 2011; Murzin et al., 1995).

Prior P2 **Evolutionary Models for Amino Acid Substitution.** Just as nucleotide substitution models are used in DNA sequence analysis, amino acid substitution models are employed to describe the changes in protein sequences over time. These models account for the biochemical properties of amino acids and the likelihood of certain substitutions occurring more frequently than others. For instance, the JTT model (Jones-Taylor-Thornton model) provides a substitution matrix Q that specifies the rate at which one amino acid is substituted for another over evolutionary time. The probability of substitution can be modeled as:

$$P(t) = e^{Qt}, \tag{2}$$

where t represents the evolutionary time, and Q is the rate matrix. This model is critical for estimating the evolutionary distances between protein sequences and constructing accurate phylogenetic trees (Dayhoff et al., 1978; Jones et al., 1992).

Prior P3 **Protein Secondary Structure Information.** Protein secondary structures, such as alpha-helices and beta-sheets, are conserved across species when they are essential to the protein's function. These structural elements can be encoded in a matrix S, where each element S_{ij}^p represents the similarity between the secondary structures of sequences x_i^p and x_j^p. The similarity can be assessed using measures such as the percentage of identical residues in aligned helices or sheets:

$$d_{\text{secondary}}(x_i^p, x_j^p) = \sum_{k=1}^{L} S(x_{i,k}^p, x_{j,k}^p), \tag{3}$$

where L is the length of the aligned sequences. This structural information enhances the accuracy of phylogenetic trees by incorporating the functional and structural constraints that shape protein evolution (Chothia and Finkelstein, 1984; Kabsch and Sander, 1983).

Prior P4 **Tertiary Structure Conservation.** The three-dimensional (tertiary) structure of a protein often provides more evolutionary information than its primary sequence alone, as structural features tend to be more conserved. The conservation of tertiary structure can be quantified by superimposing the 3D structures of two proteins and measuring the root-mean-square deviation (RMSD) between corresponding atoms:

$$d_{\text{tertiary}}(x_i^p, x_j^p) = \text{RMSD}(x_i^p, x_j^p), \tag{4}$$

where a lower RMSD indicates greater structural similarity. This metric is especially useful for inferring evolutionary relationships when sequence similarity is low but structural features are preserved (Sali and Blundell, 1994).

Prior P5 **Functional Site Conservation.** Functional sites, such as active sites in enzymes or ligand-binding sites, are critical for the protein's function and are often conserved across species. These sites can be identified and compared across sequences, with a focus on residues involved in the active or binding site. The conservation of these sites can be encoded in a function F, where $F(x_i^p, x_j^p)$ represents the similarity between the functional sites of sequences x_i^p and x_j^p:

$$d_{\text{functional}}(x_i^p, x_j^p) = \sum_{k=1}^{N} F(x_{i,k}^p, x_{j,k}^p), \tag{5}$$

where N is the number of residues in the functional site. This prior knowledge helps in accurately reconstructing phylogenetic trees that reflect the conservation of protein function across evolutionary time (Bartlett et al., 2002; Thornton et al., 2000).

Prior P6 **Protein Family Classification.** Protein family classification groups proteins based on sequence and structural similarity, often reflecting common evolutionary origins. These classifications can be used as prior knowledge to constrain the topology of phylogenetic trees. Let $\mathcal{C}$ represent the classification, and incorporate it into the tree construction as follows:

$$P(\text{Tree} \mid \mathcal{C}) = \prod_{\text{family } i \in \mathcal{C}} P(\text{Tree} \mid i), \tag{6}$$

where each family i imposes constraints on the possible tree topologies. This ensures that the resulting phylogenetic tree is consistent with known protein family groupings (Bateman et al., 2002; Finn et al., 2016).

Prior P7 **Co-Evolutionary Relationships.** Proteins often co-evolve with other proteins or molecules within the same biological pathway. These co-evolutionary relationships can be inferred from correlated mutations between interacting proteins or domains. The co-evolution can be captured in a matrix C, where C_{ij}^{p} indicates the strength of the co-evolutionary signal between sequences x_i^p and x_j^p. This can be incorporated into the tree construction as:

$$d_{\text{co-evolution}}(x_i^p, x_j^p) = -\log(C_{ij}^p), \tag{7}$$

where higher values of C_{ij}^p indicate stronger co-evolutionary signals. Incorporating co-evolutionary information helps in reconstructing trees that better reflect the functional interdependencies of proteins (Göbel et al., 1994; Marks et al., 2011).

3.2.3 Prior Knowledge for Constructing Cell Differentiation Trees using Single-Cell Multimodal Data

When constructing cell differentiation trees using single-cell multimodal data, leveraging prior knowledge is crucial for accurately modeling the complex processes of cellular differentiation. This section discusses key types of prior knowledge, providing both biological context and formal mathematical descriptions, along with relevant references.

Prior S1 *Gene Expression Profiles.* Gene expression profiles, which measure the abundance of mRNA transcripts in single cells, provide essential insights into the functional state of a cell. Highly expressed genes often indicate active biological processes and can be used to infer cellular identity and differentiation status. The expression levels of genes can be represented in a matrix E, where E_{ij}^c denotes the expression level of gene j in cell i. The similarity between cells based on their gene expression profiles can be quantified by:

$$d_{\text{expression}}(c_i, c_j) = \sum_{k=1}^{G} \left(E_{ik}^c - E_{jk}^c\right)^2, \tag{8}$$

where G is the total number of genes. This metric captures the overall difference in gene expression patterns between cells, aiding in the construction of differentiation trees (Qiu et al., 2017a; Trapnell et al., 2014c).

Prior S2 *RNA Velocity.* RNA velocity is a computational method that estimates the future state of individual cells based on the ratio of spliced and unspliced mRNA transcripts. This provides dynamic information about the direction of cell differentiation at the single-cell level. RNA velocity can be represented as a vector v_i^c for each cell i, indicating the predicted transcriptional change over time. The distance between cells based on RNA velocity can be defined as:

$$d_{\text{velocity}}(c_i, c_j) = \|v_i^c - v_j^c\|, \tag{9}$$

where $\| \cdot \|$ denotes the Euclidean norm. This approach helps in predicting the future states of cells and their differentiation trajectories (Bergen et al., 2020a; La Manno et al., 2018b).

Prior S3 *Cell Type-Specific Marker Genes.* Cell type-specific marker genes are genes that are uniquely or highly expressed in particular cell types and are used to identify cell identities during differentiation. The presence or absence of these markers can be encoded in a binary matrix B, where $B^c_{ij} = 1$ if marker gene j is expressed in cell i, and $B^c_{ij} = 0$ otherwise. The similarity between cells based on marker gene expression can be calculated as:

$$d_{\text{markers}}(c_i, c_j) = \sum_{k=1}^{M} \left| B^c_{ik} - B^c_{jk} \right| , \tag{10}$$

where M is the number of marker genes. This information is critical for accurately identifying cell types and understanding the progression of differentiation (Plass et al., 2018; Tirosh et al., 2016).

Prior S4 *Pseudotime Ordering.* Pseudotime analysis orders cells along a continuous trajectory that represents the progression of differentiation. This method infers the relative differentiation time of each cell, allowing the construction of differentiation pathways. Pseudotime can be represented as a scalar τ^c_i for each cell i, indicating its position along the differentiation trajectory. The distance between cells based on pseudotime can be calculated as:

$$d_{\text{pseudotime}}(c_i, c_j) = \left| \tau^c_i - \tau^c_j \right| , \tag{11}$$

which reflects the temporal distance between cells in their differentiation process. This prior knowledge is instrumental in visualizing and understanding cell differentiation pathways (Haghverdi et al., 2016; Trapnell et al., 2014c).

4 Classical Biological Tree Construction Methods

4.1 Classical General Tree Construction Methods

Distance-Based Methods. Distance-based methods are some of the earliest and most computationally efficient techniques for constructing phylogenetic trees, as illustrated in the top section of the diagram. These methods start by calculating genetic or evolutionary distances between sequences to create a distance matrix, which is then used to build an initial tree. Techniques such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) assume a constant rate of evolution, using the molecular clock to produce a rooted tree (Michener and Sokal, 1957). However, this assumption often does not hold in real-world scenarios, leading to potential biases. Neighbor-Joining (NJ) addresses this limitation by constructing an unrooted tree without assuming a constant rate of evolution, instead minimizing the total branch length (Saitou and Nei, 1987). More advanced methods like Minimum Evolution (ME) and Balanced Minimum Evolution (BME) further optimize tree topology to minimize overall branch lengths while balancing computational efficiency (Desper and Gascuel, 2004; Rzhetsky and Nei, 1992). While the steps in the diagram show that distance-based methods involve iterative merging of nodes to generate the final tree, they inherently reduce the complexity of the original data into pairwise distances, potentially leading to information loss. Additionally, reliance on assumptions such as the molecular clock in UPGMA can introduce biases. Therefore, these methods are best suited for preliminary analyses or scenarios with limited computational resources.

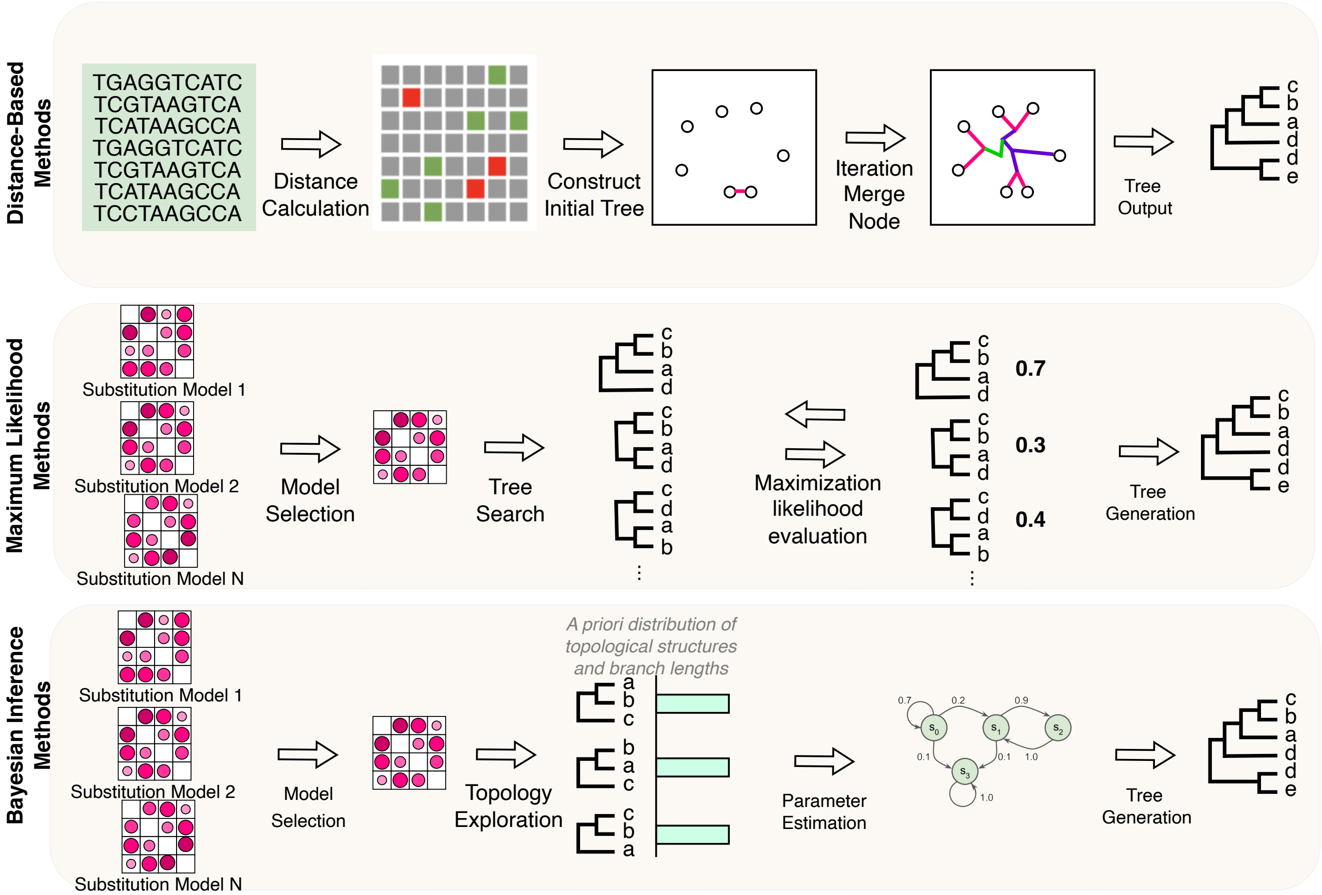

Figure 2: **Schematic diagram of Classical General Tree Construction Methods.** The methods are classified based on the type of data used, the ability to handle inconsistencies between gene and species trees, and the specific application scenarios.

Maximum Likelihood Methods. Maximum Likelihood (ML) methods, as depicted in the middle section of the diagram, offer a more statistically rigorous approach. They involve selecting an evolutionary substitution model and then searching for tree topologies and branch lengths that maximize the likelihood of the observed sequence data. The diagram illustrates a model selection phase, followed by a tree search and evaluation process, where likelihood values are computed for different tree topologies. Tools like RAxML (Randomized Axelerated Maximum Likelihood) efficiently handle large datasets, offering various evolutionary models (Stamatakis, 2014). PhyML (Phylogenetic Maximum Likelihood) balances speed and accuracy, while IQ-TREE introduces automated model selection and ultrafast bootstrap methods (Guindon and Gascuel, 2003; Nguyen et al., 2015). However, ML methods are computationally intensive, especially when processing large datasets. Moreover, as shown in the diagram, the maximization process requires careful model selection, as incorrect choices can introduce biases. Thus, while ML methods are powerful and flexible, they require adequate computational resources and evolutionary biology expertise to ensure robust results.

Bayesian Inference Methods. Bayesian Inference (BI) methods, represented in the bottom section of the diagram, offer a comprehensive probabilistic framework for tree estimation. They integrate prior information with observed data to calculate posterior probabilities for various tree topologies. The diagram outlines key

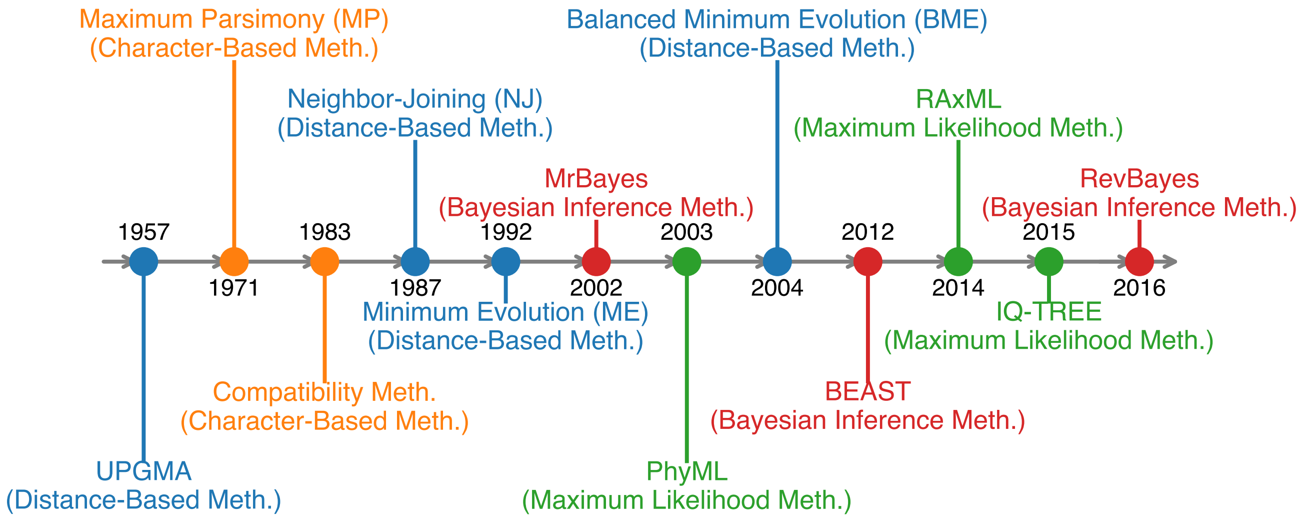

Figure 3: **The timeline of Classical General Tree Construction Methods.** The figure shows the development of tree construction methods in phylogenetics from 1957 to 2016, categorized into feature-based, distance-based, Bayesian inference, and maximum likelihood methods. Different colors indicate different categories.

steps: model selection, topology exploration, and parameter estimation. Unlike ML methods, BI methods explore the posterior distribution of trees using techniques like Markov Chain Monte Carlo (MCMC). Tools like MrBayes incorporate a wide range of evolutionary models (Ronquist et al., 2012), while BEAST (Bayesian Evolutionary Analysis by Sampling Trees) focuses on divergence time estimation (Drummond et al., 2012). RevBayes provides flexibility for modeling complex evolutionary processes (Hohna et al., 2016). The diagram indicates that BI methods use prior distributions for topological structures and branch lengths to guide the exploration process. While Bayesian methods provide probabilistic support and incorporate prior knowledge, they are computationally demanding due to the extensive MCMC sampling required. Furthermore, choosing appropriate priors and models is crucial to avoid biases. Despite these complexities, their ability to offer rigorous probabilistic estimates makes Bayesian methods invaluable for comprehensive phylogenetic studies.

In summary, the choice of phylogenetic tree construction method depends on the specific context, including data characteristics, evolutionary assumptions, and computational resources. The diagram highlights the core processes of each method, from distance matrix generation in distance-based methods to model selection and tree exploration in ML and BI methods. Distance-based methods offer computational simplicity but may lose information and introduce biases. ML and BI methods provide greater accuracy and flexibility but require careful model selection and substantial computational power. Employing a combination of these methods and critically evaluating their assumptions can ensure robust phylogenetic analysis that accurately reflects evolutionary processes.

4.2 Classical Gene-Based Phylogenetic Tree Methods

In recent years, as shown in Table 5 and Figure 4, phylogenetic inference methods have seen significant advancements, particularly in two main categories: Bayesian inference methods and alignment-free methods. Bayesian inference has evolved from the early **Markov Chain Monte Carlo (MCMC)** approach, which samples within tree space to estimate the posterior distribution of evolutionary trees. While MCMC is

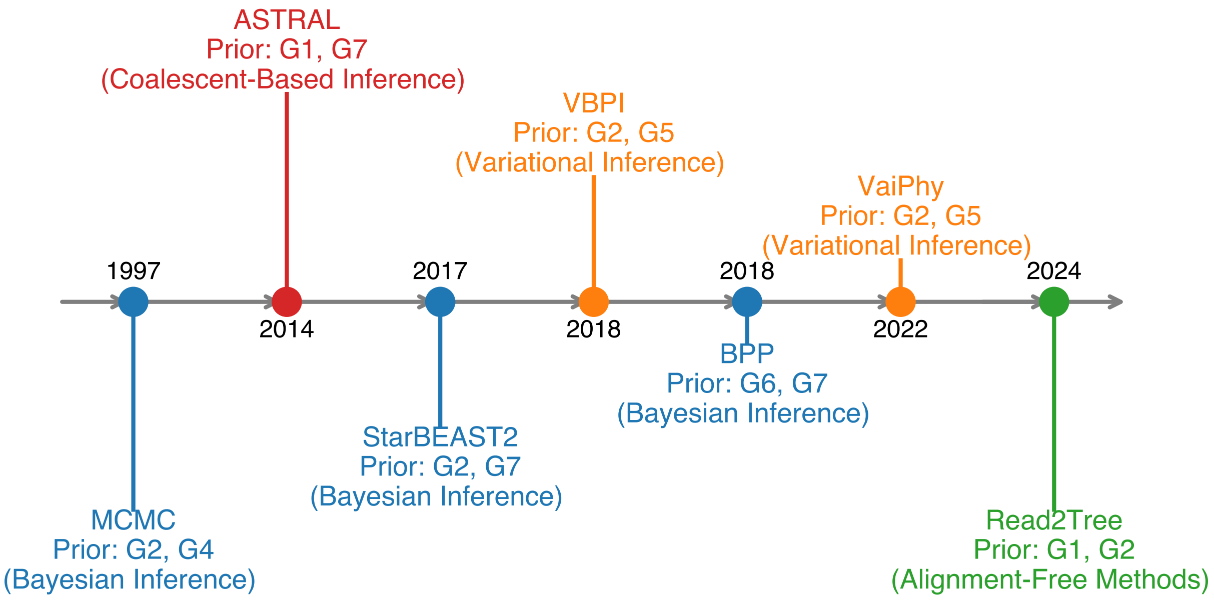

Figure 4: **The timeline of Classical Gene-Based Phylogenetic Tree Construction Methods.** The figure shows the development of gene-based tree construction methods in phylogenetics from 1994 to 2022, categorized into Bayesian inference, coalescent-based methods, and alignment-free methods. Different colors indicate different categories.

capable of handling complex evolutionary models and is well-suited for small datasets, it suffers from high computational complexity, particularly with large datasets. MCMC relies on prior knowledge such as *Evolutionary Substitution Models (G2)* (e.g., the Jukes-Cantor model) to describe evolutionary relationships between sequences (Felsenstein, 1981; Kimura, 1980). Additionally, it incorporates *Ancestral Relationship Information (G4)* to accurately infer species' evolutionary paths (Maddison and Maddison, 2007; Ronquist and Huelsenbeck, 2003).

One of the early advancements addressing the limitations of MCMC is the coalescent-based approach, represented by **ASTRAL**. This method leverages multiple gene trees to infer species trees, addressing the issue of incomplete lineage sorting (ILS) by integrating information across genes (Mirarab et al., 2014). *ASTRAL* improves computational efficiency while maintaining statistical consistency and accuracy by using prior knowledge such as *Conserved Genomic Regions (G1)* and *Taxonomic Classification Constraints (G7)* (Faith, 1992; Mount, 2004; Notredame, 2007). This method has become widely used in large-scale genomic studies.

To further address the computational challenges in large-scale data, variational inference (VI) methods have been proposed. **VBPI** (Variational Bayesian Phylogenetic Inference) is a notable method in this category. *VBPI* combines graphical models and branch length distributions, optimized via stochastic gradient ascent, which significantly reduces the inference time (Zhang and Iv, 2018). Compared to MCMC, *VBPI* uses *Evolutionary Substitution Models (G2)* by employing the transition probability matrix $P(t)$ to describe probabilistic changes in nucleotide sequences over time (Felsenstein, 1981; Kimura, 1980). This enables high accuracy on large datasets while accelerating the inference process.

Building on this, **VaiPhy** further enhances the computational efficiency of variational inference. *VaiPhy* introduces faster sampling schemes, such as the SLANTIS proposal distribution and the JC sampler, which avoid expensive auto-differentiation operations (Koptagel et al., 2022). *VaiPhy* can handle complex tree structures and leverages *Evolutionary Substitution Models (G2)* and *Sequence Homology Information (G5)* to

Table 5: Overview of the Classical Gene-based Tree Construction Mehtods.

Method Name	Description	Reference	URL
ASTRAL	A coalescent-based method for estimating species trees from multiple gene trees, known for its high accuracy	(Mirarab et al., 2014)	https://github.com/smirarab/ASTRAL/
StarBEAST2	A faster Bayesian method for species tree inference with accurate substitution rate estimates	(Ogilvie et al., 2017)	https://github.com/genomescale/starbeast2
VBPI	A variational framework for Bayesian phylogenetic analysis, using stochastic gradient ascent for posterior estimation	(Zhang and Iv, 2018)	https://github.com/tyuxie/VBPI-SIBranch
BPP	A method using genomic sequences and multispecies coalescent for species tree estimation	(Flouri et al., 2018)	https://github.com/bpp/
VaiPhy	A variational inference-based algorithm for approximate posterior inference in phylogeny	(Koptagel et al., 2022)	https://github.com/Lagergren-Lab/VaiPhy
Read2Tree	A method to infer phylogenetic trees directly from raw sequencing reads, bypassing traditional genome assembly and annotation	(Dylus et al., 2024b)	https://github.com/DessimozLab/read2tree

achieve efficient inference on large-scale datasets (Smith and Waterman, 1981a; Thompson et al., 1994a). This method significantly improves inference speed while maintaining comparable accuracy to state-of-the-art methods.

In parallel, coalescent-based Bayesian methods like **BPP** have advanced the field of species tree estimation by integrating multilocus sequence data (Flouri et al., 2018). *BPP* uses prior knowledge such as *Gene Duplication and Loss Events (G6)* and *Ancestral Relationship Information (G4)*, allowing for more precise species tree inference, particularly in the context of incomplete lineage sorting and gene flow (Gu et al., 2005; Maddison and Maddison, 2007). Similarly, methods like **StarBEAST2** combine Bayesian inference with species tree estimation, integrating *Taxonomic Classification Constraints (G7)* and evolutionary substitution rates to improve the accuracy and speed of tree inference (Ogilvie et al., 2017).

Simultaneously, alignment-free methods have rapidly gained traction, with **Read2Tree** being a prominent example. This method bypasses traditional genome assembly and sequence alignment steps, directly inferring phylogenetic trees from raw sequencing data (Dylus et al., 2024b). By utilizing prior knowledge, such as the *Genomic Linear Order of Genes (G3)* and *Conserved Genomic Regions (G1)*, Read2Tree drastically reduces computational overhead, making it particularly suitable for phylogenetic analyses of large genomic datasets (Mount, 2004; Saitou and Nei, 1987). This method not only improves inference efficiency but also maintains high accuracy when dealing with diverse and complex genomic data.

Overall, Bayesian inference methods and alignment-free approaches demonstrate a strong complementarity in phylogenetic inference. Bayesian inference excels at handling complex evolutionary models and uncertainties through the integration of prior knowledge, such as *Evolutionary Substitution Models (G2)*, *Conserved Genomic Regions (G1)*, and *Gene Duplication Events (G6)*, while alignment-free methods simplify the computational process, enabling efficient analysis of large-scale datasets. Future research should focus on integrating these methods to fully leverage multiple levels of prior knowledge, further improving the accuracy and efficiency of phylogenetic inference.

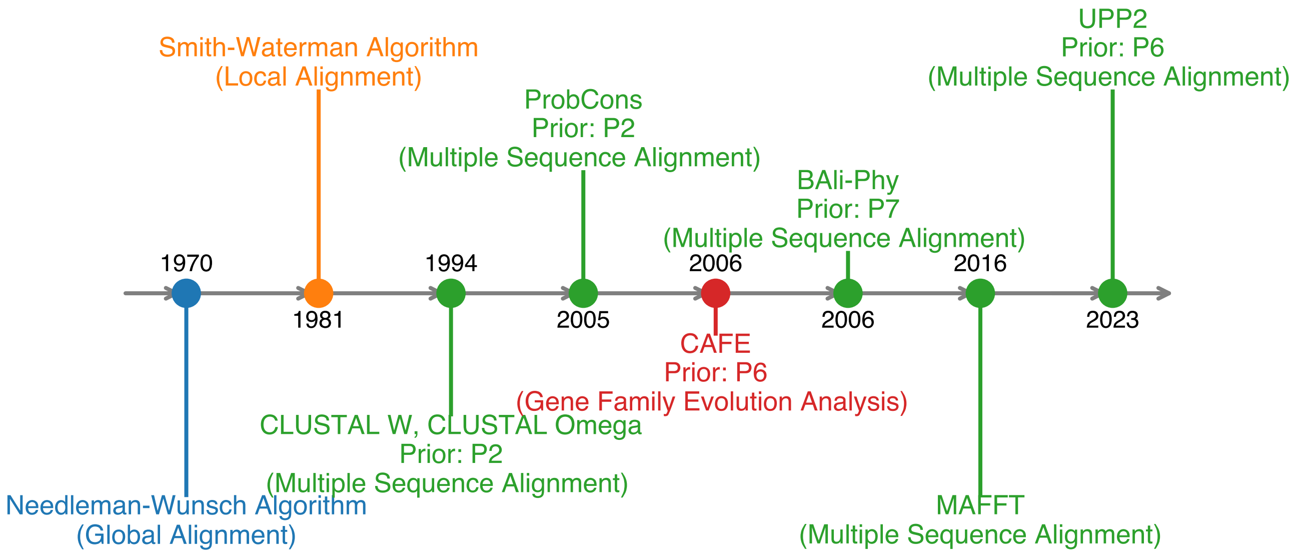
Smith-Waterman Algorithm
(Local Alignment)
ProbCons
Prior: P2
(Multiple Sequence Alignment)
BAli-Phy
Prior: P7
(Multiple Sequence Alignment)
UPP2
Prior: P6
(Multiple Sequence Alignment)
1970
1981
1994
2005
2006
2006
2016
2023
CAFE
Prior: P6
(Gene Family Evolution Analysis)
CLUSTAL W, CLUSTAL Omega
Prior: P2
(Multiple Sequence Alignment)
Needleman-Wunsch Algorithm
(Global Alignment)
MAFFT
(Multiple Sequence Alignment)

Figure 5: **The timeline of classical protein sequence-based phylogenetic tree construction methods.** The figure shows the development of protein sequence-based tree construction methods in phylogenetics from 1970 to 2023, categorized into sequence alignment, multiple sequence alignment, and gene family evolution methods. Different colors indicate different categories.

4.3 Classical Protein-Based Phylogenetic Tree Methods

The protein analysis is crucial in bioinformatics, focusing on elucidating the function, structure, and evolutionary relationships of proteins. Current protein analysis methods can be broadly divided into sequence-based and structure-based approaches. The former focuses on evolutionary inference and functional annotation through sequence similarity analysis, while the latter leverages three-dimensional (3D) structural information to provide deeper insights into functional and structural features (Holm and Sander, 1993; Zhang and Skolnick, 2005). Both types of methods have their strengths, such as computational efficiency and accuracy in identifying conserved structures, but they also face unique challenges. Each method's effectiveness often relies on specific types of prior knowledge, such as conserved domains, evolutionary models, or structural conservation. This review discusses these methods, their advantages, limitations, the prior knowledge they utilize, and future development directions.

4.3.1 Classical Protein Sequence Based Phylogenetic Tree Methods

As shown in table 6 and Figure. 5, sequence-based protein analysis methods are widely used for inferring evolutionary relationships and functional annotation. These methods utilize protein sequence information to reveal biological functions and evolutionary histories by comparing sequence similarities. **Global and local alignments** are the most fundamental sequence alignment methods. The **Needleman-Wunsch algorithm** (Needleman and Wunsch, 1970) is a classical global alignment algorithm that uses dynamic programming to find the optimal global alignment path between two protein sequences, suitable for sequences of similar length and high similarity. However, its computational cost is high, making it less practical for large datasets. In contrast, the **Smith-Waterman algorithm** (Smith and Waterman, 1981b) is designed for local alignment, capable of identifying the most similar local regions between sequences, making it suitable for sequences of different lengths or those that are only partially similar. Although it provides flexibility when dealing with

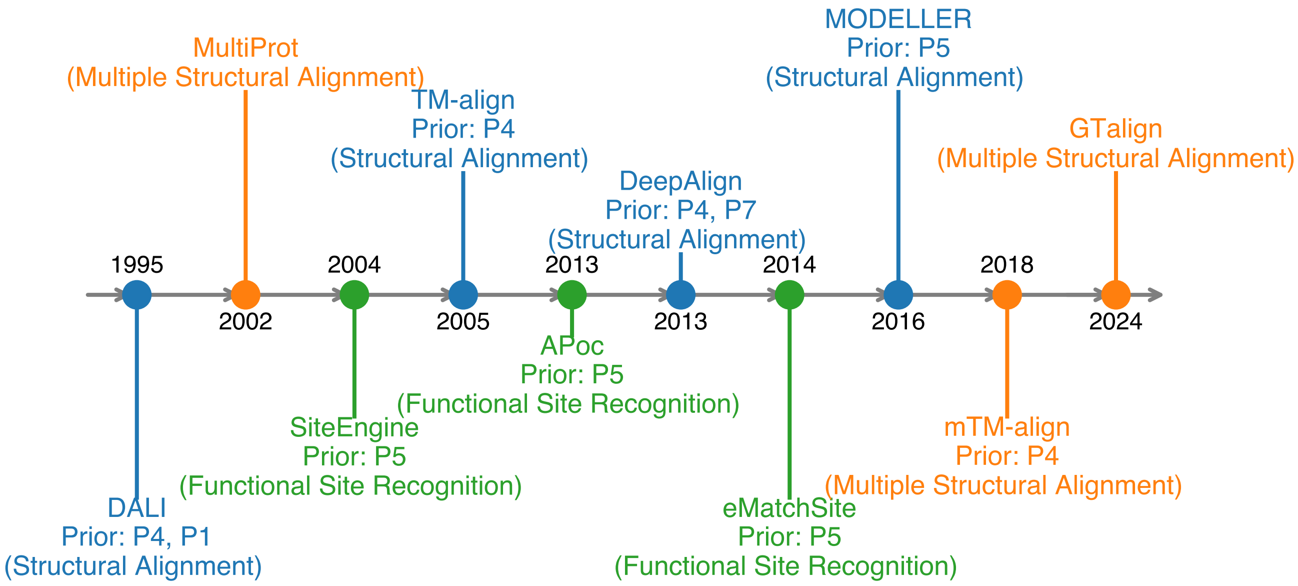

Figure 6: **The timeline of Classical Protein Structural Based Phylogenetic Tree Construction Methods.** The figure shows the development of protein structural alignment methods in phylogenetics, categorized into distance matrix-based alignment, multiple structure alignment, and functional site recognition methods. Different colors indicate different categories.

highly divergent sequences, its computational overhead is similarly high.

Multiple sequence alignment methods are crucial for studying the similarity between multiple protein sequences. **CLUSTAL W** (Thompson et al., 1994b) and **CLUSTAL Omega** (Sievers and Higgins, 2014) are representative progressive multiple sequence alignment methods that optimize alignments using techniques such as progressive weighting and position-specific gap penalties, making them suitable for large-scale sequence datasets. These methods use prior knowledge of *Evolutionary Models for Amino Acid Substitution (P2)*, such as substitution matrices (e.g., the JTT matrix or Dayhoff matrix) (Dayhoff et al., 1978; Jones et al., 1992), to model evolutionary relationships and guide the alignment process. However, they may lead to suboptimal alignments when dealing with sequences containing many insertions or deletions (indels). In contrast, **ProbCons** (Do et al., 2005) uses a probabilistic consistency-based model that also relies on *evolutionary models (P2)*, but with a more sophisticated approach to account for sequence divergence, effectively capturing complex interactions between sequences during alignment and demonstrating higher accuracy. Nevertheless, the computational complexity of these statistical and probabilistic models remains a significant challenge when handling very large datasets.

The **MAFFT** program (Katoh and Standley, 2016) introduces a new feature that addresses the issue of over-alignment, where unrelated segments are erroneously aligned. Traditional **MAFFT** is known for its sensitivity in aligning conserved regions in remote homologs, but this sensitivity can lead to over-alignment, especially with low-quality or noisy sequences. The improved **MAFFT** uses a variable scoring matrix for different pairs of sequences (or groups) within a single multiple sequence alignment, based on the global similarity of each pair. This approach reduces over-alignment and improves the overall reliability of the alignment, especially in databases increasingly populated by noisy sequences.

Similarly, **UPP2** (Park et al., 2023) is an advancement of the Ultra-large multiple sequence alignment method that deals with fragmentary sequences using an ensemble of Hidden Markov Models (eHMMs) to represent an

Table 6: Overview of classical protein-based sequence alignment methods.

Method Name	Description	Reference	URL
CLUSTAL Omega	Fast multiple sequence alignment for large datasets using an advanced algorithm.	(Sievers and Higgins, 2014)	N/A
CLUSTAL W	Progressive multiple sequence alignment with position-specific gap penalties.	(Thompson et al., 1994b)	N/A
ProbCons	Probabilistic multiple sequence alignment using hidden Markov models for higher accuracy.	(Do et al., 2005)	N/A
CAFE	Gene family evolution modeling with random birth and death process.	(De Bie et al., 2006)	N/A
BAli-Phy	Bayesian sequence alignment and phylogenetic inference in one framework.	(Suchard and Redelings, 2006)	`https://www.bali-phy.org/`
MAFFT	Fast multiple sequence alignment with sensitivity for remote homologs.	(Katoh and Standley, 2016)	`http://www.blast2go.de`
UPP2	Ultra-large sequence alignment with phylogeny-aware profiles and HMMs for fragmentary sequences.	(Park et al., 2023)	`https://github.com/gillichu/sepp`

estimated alignment on the full-length sequences in the input, and then adds the remaining sequences using selected HMMs from the ensemble. It significantly improves accuracy, especially in datasets with substantial sequence length heterogeneity. The use of *Phylogeny-aware Profiles (P6)* as prior knowledge allows **UPP2** to adaptively handle large datasets with varying sequence lengths, which makes it particularly effective in handling incomplete or highly divergent sequences, compared to other leading MSA methods.

Beyond sequence alignment, tools for gene family evolution and evolutionary analysis, such as **CAFE** (De Bie et al., 2006) and **BAli-Phy** (Suchard and Redelings, 2006), play important roles in studying gene family expansion, gene loss, and protein functional evolution. **CAFE** models gene family evolution by simulating a random birth and death process for gene family size, aiding in the study of gene family dynamics. This method incorporates *Protein Family Classification (P6)* as prior knowledge to define gene family groups and model their evolutionary trajectories based on sequence and structural similarities (Bateman et al., 2002; Finn et al., 2016). However, its effectiveness heavily depends on the accuracy of the input phylogenetic tree. **BAli-Phy**, on the other hand, integrates sequence alignment and phylogenetic inference within a **Bayesian framework**, using priors like *Co-Evolutionary Relationships (P7)* that capture the interdependencies between proteins through co-evolution (Göbel et al., 1994; Marks et al., 2011). This integration reduces the biases that may arise from separate analyses but has high computational complexity, limiting its application to large-scale datasets.

4.3.2 Classical Protein Structure Based Phylogenetic Tree Methods

As shown in table 7 and Figure. 6, protein structure analysis is a critical component of bioinformatics, as it provides deeper insights into protein function, interactions, and evolutionary relationships that sequence-based methods alone cannot offer. Unlike sequence-based methods that rely solely on primary amino acid sequences, structure-based methods utilize three-dimensional (3D) structural information of proteins to capture more complex evolutionary and functional relationships. These methods typically require prior knowledge, such as conserved tertiary structures and functional site conservation. The following content discusses the development of structural alignment methods, functional site recognition techniques, and structural comparison algorithms in chronological and logical order, along with their applications.

Development of Protein Structural Alignment Methods. Early structural alignment methods, such as **DALI** (Holm and Sander, 1995), used distance matrix-based alignment to compare protein structures, aiming

Table 7: Overview of classical protein structural based tree construction methods.

Method Name	Description	Reference	URL
DALI	Distance matrix-based structural alignment for detecting global and local similarities.	(Holm and Sander, 1995)	N/A
MultiProt	Multiple structure alignment using geometric cores, suitable for partial alignments.	(Shatsky et al., 2002)	N/A
SiteEngine	Functional site recognition by comparing protein surface binding sites.	(Shulman-Peleg et al., 2004)	`https://bio.tools/siteengine`
TM-align	TM-score-based pairwise structural alignment with high speed and accuracy.	(Zhang and Skolnick, 2005)	`https://zhanggroup.org/TM-align/`
APoc	Large-scale structural comparison for identifying pockets on protein surfaces.	(Gao and Skolnick, 2013)	`http://cssb.biology.gatech.edu/APoc`
DeepAlign	Protein structure alignment combining spatial proximity with evolutionary information.	(Wang et al., 2013)	`https://github.com/realbigws/DeepAlign`
eMatchSite	Binding site alignment tolerant to structural distortions in protein models.	(Brylinski, 2014)	`http://www.brylinski.org/ematchsite`
MODELLER	Comparative protein structure modeling based on sequence alignment with templates.	(Webb and Sali, 2016)	`https://salilab.org/modeller/`
mTM-align	Extension of TM-align for multiple structure alignment with improved accuracy and speed.	(Dong et al., 2018)	`https://github.com/CSB5/CaDRReS`
GTalign	Spatial index-driven multiple structure alignment with high efficiency for large datasets.	(Margelevičius, 2024)	`https://github.com/openCONTRABASS/CONTRABASS`

to detect both global and local structural similarities. DALI implemented a network-based tool for protein structure comparison, leveraging prior knowledge of *Tertiary Structure Conservation (P4)* and *Conserved Protein Domains (P1)* to effectively identify remote homologs and functionally similar proteins. DALI laid the foundation for the field of protein structural alignment, especially in uncovering distant evolutionary relationships that are not easily detectable by sequence analysis alone. However, its computational complexity limits its application to large-scale datasets.

As the demand for computational efficiency grew, **TM-align** (Zhang and Skolnick, 2005) was introduced. TM-align uses the TM-score rotation matrix combined with dynamic programming to achieve optimal pairwise structural alignment, offering higher speed and better alignment accuracy than DALI and CE methods. TM-align focuses on *Tertiary Structure Conservation (P4)* (e.g., RMSD) to ensure that alignments reflect conserved 3D structures. Its significant computational efficiency and accuracy have led to its widespread use in practical applications, particularly for rapid and precise comparison of large protein structure databases.

With the need for multiple protein structure alignments, the **MultiProt** algorithm (Shatsky et al., 2002) provided a solution for multiple structural alignments. Unlike the previous methods, MultiProt identifies common geometric cores among proteins without requiring all molecules to participate in the alignment. Its advantage lies in handling highly variable datasets, especially in scenarios involving diverse structures and partial alignments. However, its computational cost increases significantly with larger data size and complexity.

In the 2010s, to address the growing number of protein structures and improve the accuracy of multiple alignments, **mTM-align** (Dong et al., 2018) was developed. mTM-align is an extension of the TM-align method, designed to tackle the challenge of aligning more than two protein structures simultaneously. This method retains the advantages of *Tertiary Structure Conservation (P4)* and has been benchmarked on widely used datasets, demonstrating consistent superiority in alignment accuracy and computational efficiency. It is particularly useful for large-scale proteomic datasets where accurate and rapid multiple structural alignments are critical.

The most recent multiple structure alignment method, **GTalign** (Margelevičius, 2024), employs a spatial index-driven strategy to achieve optimal superposition at high speeds. GTalign focuses on providing rapid and accurate structural comparisons using its spatial indexing approach. Its high efficiency in parallel processing and rapid computation makes it highly applicable in modern biological research, especially when dealing with large-scale datasets. However, the requirement for pre-indexing structures can pose a challenge when new data is frequently added to the analysis pipeline.

Development of Functional Site Recognition Techniques. Functional site recognition is another critical aspect of structure-based protein analysis. The early method, **SiteEngine** (Shulman-Peleg et al., 2004), identifies regions on one protein surface that are similar to a binding site on another protein. SiteEngine does not require sequence or fold similarities; instead, it uses prior knowledge in the form of *Functional Site Conservation (P5)* to recognize similar binding sites. This method is particularly advantageous for predicting molecular interactions and aiding in drug discovery. However, its dependency on high-quality protein structures can limit its application in cases where experimental data is sparse or noisy.

The **APoc** method (Gao and Skolnick, 2013) is another tool designed for large-scale structural comparison, particularly for identifying pockets on protein surfaces. APoc uses a scoring function called the Pocket Similarity Score (PS-score) to measure the similarity between different protein pockets and employs statistical models to assess the significance of these similarities. It leverages *Functional Site Conservation (P5)* to enhance its predictive power in classifying ligand-binding sites and predicting protein molecular function. While robust, its performance is influenced by the quality of input data, especially when the structures are predicted models rather than experimentally determined ones.

eMatchSite (Brylinski, 2014) introduced a new sequence order-independent method for binding site alignment in protein models, capable of constructing accurate local alignments. eMatchSite shows high tolerance to structural distortions in weakly homologous protein models and uses *Functional Site Conservation (P5)* as prior knowledge, providing new perspectives for studying drug-protein interaction networks, especially in system-level applications such as polypharmacology and rational drug repositioning.

Comparative Modeling and Other Methods. **MODELLER** (Webb and Sali, 2016) is a traditional tool for comparative protein structure modeling. It predicts 3D structures based on sequence alignment with known templates and uses *Tertiary Structure Conservation (P4)* as key prior knowledge. While effective for modeling proteins with known homologs, MODELLER's performance diminishes for novel proteins without suitable templates.

The **DeepAlign** method (Wang et al., 2013) takes a different approach by combining spatial proximity with evolutionary information and hydrogen-bonding similarity, providing a more comprehensive alignment perspective that accounts for both geometric and evolutionary constraints.

4.4 Classical Single-Cell-Based Lineage Tree Methods

In single-cell RNA sequencing (scRNA-seq) analysis, inferring developmental and differentiation trajectories is essential for unraveling complex biological processes. This involves three core tasks: trajectory, pseudo-time, and lineage inference. Various computational methods have been developed for these purposes, primarily falling into two categories: trajectory & pseudo-time inference methods and lineage inference methods.

4.4.1 Classical Single-cell Trajectory & Pseudotime Inference Methods

As shown in Table. 8, Figure. 7 and Figure. 8, the trajectory inference methods aim to reconstruct the differentiation pathways of cells by organizing them along potential developmental trajectories. These methods

Table 8: Overview of Dimensionality Reduction, Probabilistic, and RNA Velocity-based Methods for Trajectory and Pseudotime Inference.

Method Name	Description	Reference	URL
TSCAN	Clusters cells based on gene expression and constructs an MST for trajectory identification.	(Ji and Ji, 2016)	https://github.com/zji90/TSCAN
Monocle 2	Enhances Monocle with a reversed graph embedding for linear and trajectories.	(Qiu et al., 2017b)	https://cole-trapnell-lab.github.io/monocle-release/
FORKS	Infers bifurcating and linear trajectories using Steiner trees.	(Sharma et al., 2017)	https://github.com/macsharma/FORKS
Scanpy	Offers a framework for single-cell analysis, including trajectory methods.	(Wolf et al., 2018)	https://scanpy.readthedocs.io/
Seurat	Comprehensive tool for single-cell RNA-seq trajectory inference.	(Stuart and Satija, 2019)	https://satijalab.org/seurat/
PAGA	Creates an abstracted graph of cellular relationships to refine trajectories.	(Wolf et al., 2019)	https://github.com/theislab/paga
Monocle 3	Combines Monocle 2, UMAP, and PAGA for managing complex branching trajectories.	(Cao et al., 2019)	https://cole-trapnell-lab.github.io/monocle3
SoptSC	Constructs a cell similarity graph for pseudotemporal ordering.	(Wang et al., 2019)	https://github.com/WangShuxiong/SoptSC
Waddington-OT	Applies optimal transport to infer trajectories from scRNA-seq data.	(Schiebinger et al., 2019)	https://github.com/zsteve/gWOT
PoincaréMaps	Estimates pseudotime using hyperbolic distances in hyperbolic space.	(Klimovskaia et al., 2020)	https://github.com/facebookresearch/PoincareMaps
VIA	Employs random walks and MCMC simulations for trajectory reconstruction.	(Stassen et al., 2021)	https://github.com/ShobiStassen/VIA
LineageOT	Models lineage progression using optimal transport theory.	(Forrow and Schiebinger, 2021)	https://github.com/aforr/LineageOT
GeneTrajectory	Uses optimal transport metrics to infer gene trajectories.	(Qu et al., 2024)	https://github.com/KlugerLab/GeneTrajectory
SCUBA	Bifurcation analysis for trajectory inference in gene space.	(Marco et al., 2014)	https://github.com/gcyuan/SCUBA
BGP	Estimates branching times for individual genes.	(Boukouvalas et al., 2018)	https://github.com/ManchesterBioinference/BranchedGP
CSHMMs	Extends probabilistic methods to continuous trajectories.	(Lin and Bar-Joseph, 2019)	http://www.andrew.cmu.edu/user/chiehl1/CSHMM/
Ouija	Models gene expression along pseudotemporal trajectories.	(Campbell and Yau, 2019)	https://github.com/kieranrcampbell/ouija
RNA velocity	Analyzes spliced and unspliced transcripts to capture transcriptional dynamics.	(La Manno et al., 2018a)	http://velocyto.org/
scVelo	Generalizes RNA velocity analysis to diverse kinetics.	(Bergen et al., 2020b)	https://scvelo.readthedocs.io/
CellRank	Integrates RNA velocity with pseudotime inference to identify lineage drivers.	(Lange et al., 2022)	https://cellrank.readthedocs.io/
TFvelo	Integrates gene regulatory data to extend RNA velocity analysis.	(Li et al., 2024)	https://github.com/xiaoyeye/TFvelo

use prior information *Cell Type-Specific Marker Genes (S3)* to identify continuous progression and branching points that represent different lineage decisions. In contrast, pseudo-time inference, based on the prior assumption *Pseudotime Ordering (S4)*, focuses on ordering cells along a temporal axis, estimating the relative progression of individual cells through a dynamic process. While pseudo-time methods do not necessarily infer explicit branching lineages, they capture the gradual changes in cell states over time. Both approaches are primarily grounded in prior knowledge high-dimension *Cell Type-Specific Marker Genes (S3)*. The existing computational methods can be broadly categorized into three groups. The first two (dimensionality reduction and gene space-based probabilistic methods) link cells over time using gene expression, while the third (RNA velocity) relies on data from spliced and unspliced transcripts.

Dimensionality Reduction-based Methods for Trajectory & Pseudotime Inference. Dimensionality reduction-based methods leverage lower-dimensional representations of cells to infer spanning trees or other graphical structures, which are then used to map cells and reconstruct trajectories. These methods allow for the simultaneous reconstruction of cellular trajectories and the visualization of cell distributions in an interpretable and accessible manner. The existing methods can generally be classified into three main

Table 9: Overview of Classical Single-cell Lineage Inference & Tree Construction Methods.

Method Name	Description	Reference	URL
cellTree	Uses a probabilistic framework to model gene expression data and construct a tree-like structure outlining hierarchical differentiation.	duVerle et al. (2016)	https://github.com/tidwall/celltree
Slingshot	Constructs lineage trees by embedding cells into a reduced dimensional space and connecting clusters through minimum spanning trees.	Street et al. (2018)	https://github.com/kstreet13/slingshot
Monocle DDRTree	Builds a tree structure representing cell lineages using dimensionality reduction combined with reversed graph embedding.	Qiu et al. (2017b)	https://cole-trapnell-lab.github.io/monocle-release
PAGA trees	Constructs a graph representing clusters of cells and abstracts it into a tree structure to capture hierarchical branching.	Wolf et al. (2019)	https://dynverse.org/reference/dynmethods/other/ti_paga_tree/
PROSSTT	Simulates single-cell RNA-seq datasets for differentiation processes to generate lineage trees for benchmarking lineage inference methods.	Papadopoulos et al. (2019)	https://github.com/soedinglab/prosstt
SoptSC	Builds a lineage tree by clustering and lineage inference using cell-to-cell similarity matrices.	Wang et al. (2019)	https://github.com/WangShuxiong/SoptSC
CALISTA	Integrates clustering, lineage progression, transition gene identification, and pseudotime ordering into a unified framework to construct lineage trees.	Papili Gao et al. (2020)	https://github.com/CABSEL/CALISTA

categories: dimensionality reduction methods, dimensionality reduction combined with graph-based methods, and dimensionality reduction integrated with pseudo-time analysis.

For **dimensionality reduction methods**, high-dimensional *Cell Type-Specific Marker Genes (S3)* are reduced to a lower-dimensional space for trajectory inference directly. For instance, **ForceAtlas2** (Jacomy et al., 2014) positions nodes in a graph by simulating a physical system where nodes repel each other like charged particles, while edges act like springs pulling connected nodes together, leading to a balanced and visually meaningful network structure for trajectory inference. The **Monocle** (Trapnell et al., 2014b) orders cells in pseudotime using independent component analysis (ICA) and constructs a spanning tree to infer linear trajectories. **Monocle 2** (Qiu et al., 2017b) enhances Monocle with a reversed graph embedding technique to create a principal graph, enabling robust handling of both linear and branching trajectories. **FORKS** (Sharma et al., 2017) infers bifurcating and linear trajectories using Steiner trees, enhancing robustness against noise and complexity. **TSCAN** (Ji and Ji, 2016) clusters cells based on gene expressions and constructs a minimum spanning tree (MST) for trajectory identification. **Slingshot** (Street et al., 2018) fits smooth curves in the reduced-dimensional space for simultaneous pseudotime and lineage inference. **PAGA** (Wolf et al., 2019) creates an abstracted graph of cellular relationships to capture both continuous and discrete transitions before refining the trajectories. **Monocle 3** (Cao et al., 2019) combines the strengths of Monocle 2, UMAP, and PAGA to manage complex branching trajectories with improved accuracy and scalability. **SoptSC** (Wang et al., 2019) constructs a cell similarity graph for pseudotime ordering and uses the shortest path for trajectory inference. **PoincaréMaps** (Klimovskaia et al., 2020) estimates pseudotime ordering using hyperbolic distances within hyperbolic space. **Waddington-OT** (Schiebinger et al., 2019) applies optimal transport to infer trajectories from scRNA-seq data. **LineageOT** (Forrow and Schiebinger, 2021) models lineage progression using optimal

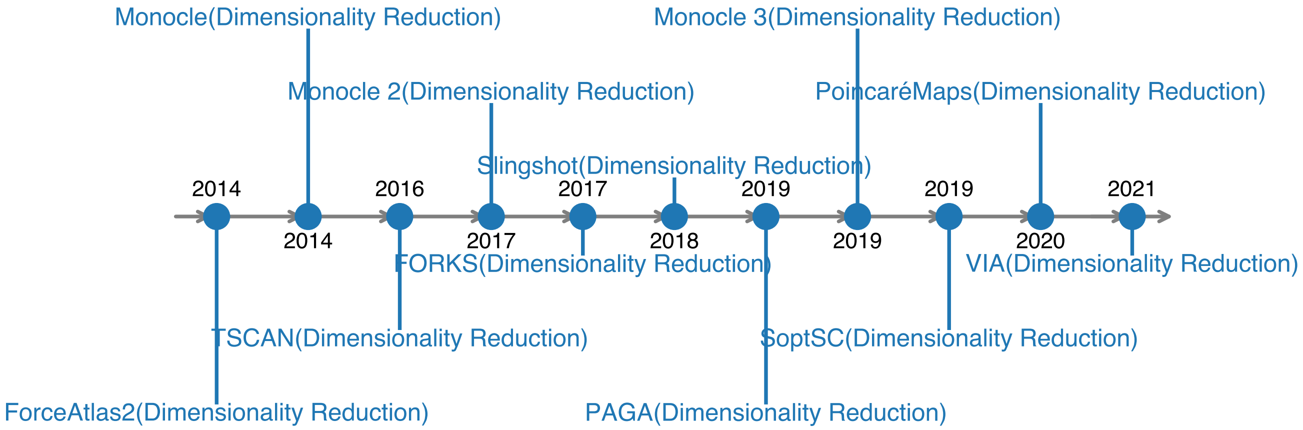

Figure 7: **The timeline of Dimensionality Reduction based Classical Single Cell Trajectory Inference Methods.** The figure shows the chronological development of trajectory inference methods based on single-cell RNA sequencing data. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in cell development analysis.

transport theory. **GeneTrajectory** (Qu et al., 2024) employs optimal transport metrics to infer gene trajectories. **Seurat** (Stuart and Satija, 2019) and **Scanpy** (Wolf et al., 2018) are comprehensive tools for single-cell RNA-seq trajectory inference. In addition, **VIA** successfully identifies elusive lineages and rare cell fates across various prior knowledge, including *Protein Expression Levels* and *Epigenetic Modification*. It (Stassen et al., 2021) employs random walks and MCMC simulations for trajectory reconstruction.

Probabilistic Models in Gene Space. Dimensionality reduction has the potential downside of inferring trajectories from only the most abundantly *Cell Type-Specific Marker Genes (S3)*, which could hinder the ability to distinguish and accurately reconstruct cell state clusters that have fewer cells. Several methods have been proposed to overcome this limitation by inferring pseudotime and trajectories directly from the *Gene Expression Profiles (S1)*. **SCUBA** (Marco et al., 2014) uses bifurcation analysis to model trajectories in gene space. **CSHMMs** (Lin and Bar-Joseph, 2019) extend probabilistic methods to continuous trajectories, allowing cells to be assigned to any position along the trajectory graph. **BGP** (Boukouvalas et al., 2018) estimates branching times for individual genes, while **Ouija** (Campbell and Yau, 2019) models gene expression along pseudotemporal trajectories.

RNA Velocity-based Methods. RNA velocity-based methods further utilize prior information *RNA Velocity (S2)* to analyze spliced and unspliced transcripts, capturing transcriptional dynamics within cells. **RNA velocity** (La Manno et al., 2018a) provides insights into a cell's future trajectory by calculating the ratio of spliced and unspliced mRNAs. **scVelo** (Bergen et al., 2020b) generalizes RNA velocity analysis to diverse transcriptional kinetics. **CellRank** (Lange et al., 2022) integrates RNA velocity with pseudotime inference to identify lineage drivers. **TFvelo** (Li et al., 2024) extends RNA velocity analysis by integrating gene regulatory data, enhancing the accuracy of cell dynamics and trajectory inference.

4.4.2 Classical Single-cell Lineage Inference & Tree Construction Methods

As shown in Table.9 and Figure.9, single-cell lineage inference aims to reconstruct the hierarchical relationships between individual cells by analyzing their *Gene Expression Profiles (S1)*. Its primary goal is to generate a lineage tree that represents the developmental paths cells take as they divide and differentiate. Each branch of

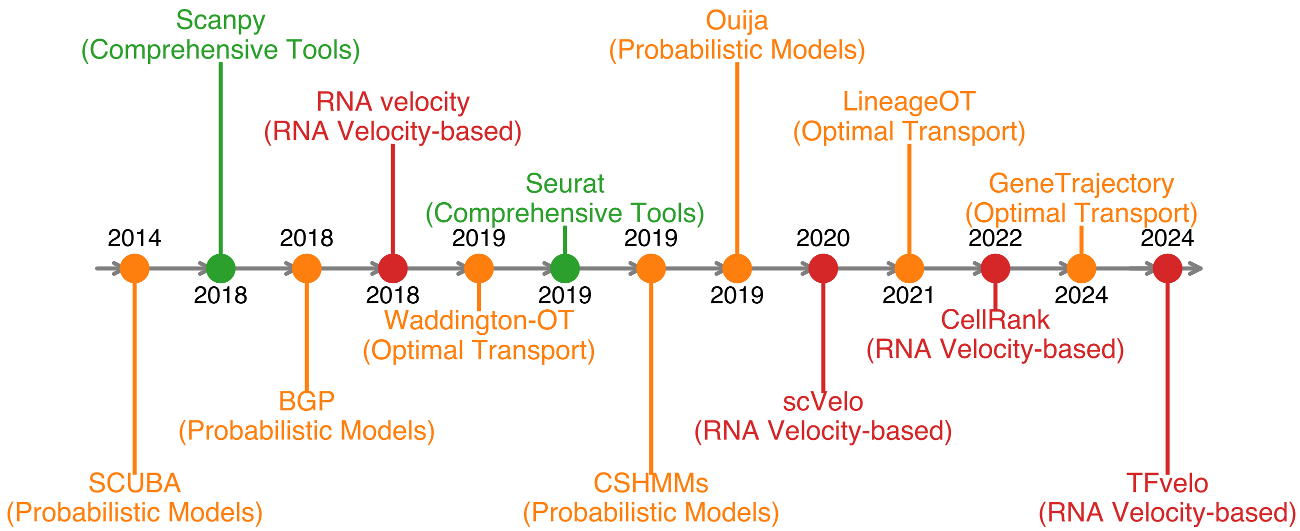

Figure 8: **The timeline of Classical Single Cell Trajectory Inference Methods.** The figure shows the chronological development of trajectory inference methods based on single-cell RNA sequencing data. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in cell development analysis.

the tree reflects how cells progress from a common progenitor to various specialized cell types.

Dimensionality reduction-based methods map cell data into a low-dimensional space to reconstruct complex lineage trees with multiple branches, allowing for pseudotime inference and better noise handling during cell differentiation analysis. **Slingshot** (Street et al., 2018) constructs lineage trees by embedding cells into a reduced dimensional space and connecting clusters through minimum spanning trees, thereby capturing the branching structure of cell lineages in the form of a tree. **Monocle DDRTree** (Qiu et al., 2017b) explicitly builds a tree structure to represent cell developmental lineages by combining discriminative dimensionality reduction with reversed graph embedding, enabling the inference of cell trajectories from gene expression data within a tree framework.

Graph-based methods utilize graph abstraction techniques to model relationships between cells and reconstruct lineage trees. **PAGA trees** (Wolf et al., 2019) constructs a graph where nodes represent clusters of cells and edges represent the connectivity probabilities between them. By abstracting this graph into a simplified tree structure, PAGA enables the reconstruction of complex lineage topologies, capturing the hierarchical branching patterns inherent in cell differentiation processes.

Simulation-based methods provide synthetic datasets with known lineage topologies to test and develop lineage reconstruction tools. **PROSSTT** (Papadopoulos et al., 2019) simulates single-cell RNA-seq datasets for differentiation processes, generating lineage trees of any desired complexity, noise level, noise model, and size. By producing datasets with predefined tree structures, PROSSTT allows for benchmarking and evaluating the accuracy of lineage inference methods in reconstructing the true underlying tree topology.

Similarity matrix-based methods utilize a cell-to-cell similarity matrix to analyze relationships between cells and construct lineage trees based on these similarities. **SoptSC** (Wang et al., 2019) builds a lineage tree by performing clustering and lineage inference using cell-cell relationships derived from a similarity matrix, effectively capturing the hierarchical differentiation paths in a tree structure.

Statistical/Probabilistic model-based methods rely on statistical or probabilistic models to account for

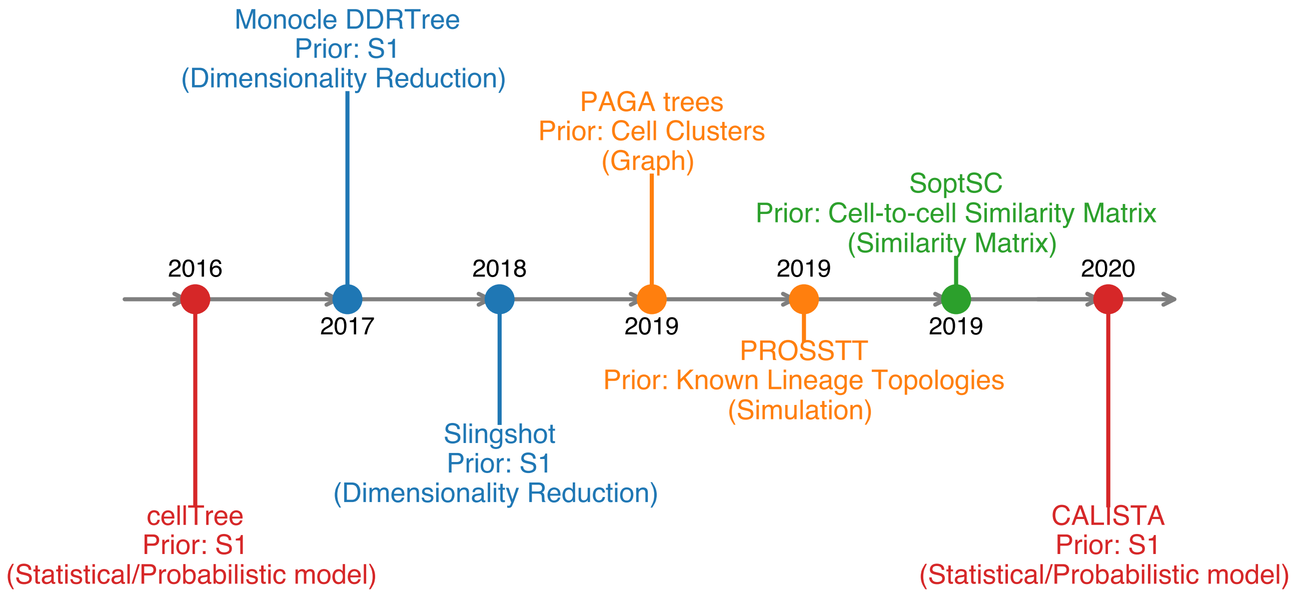
Monocle DDRTree
Prior: S1
(Dimensionality Reduction)
PAGA trees
Prior: Cell Clusters
(Graph)
SoptSC
Prior: Cell-to-cell Similarity Matrix
(Similarity Matrix)
2016
2018
2019
2020
2017
2019
2019
PROSSTT
Prior: Known Lineage Topologies
(Simulation)
Slingshot
Prior: S1
(Dimensionality Reduction)
cellTree
Prior: S1
(Statistical/Probabilistic model)
CALISTA
Prior: S1
(Statistical/Probabilistic model)

Figure 9: **The Timeline of Classical Single Cell Tree Construction Methods.** The figure shows the chronological development of tree-based methods for cell differentiation analysis based on single-cell RNA sequencing data. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in cell development analysis.

noise and stochasticity in gene expression profiles while constructing lineage trees. **cellTree** (duVerle et al., 2016) models the gene expression data using a probabilistic framework to construct a tree-like structure that outlines hierarchical differentiation, explicitly representing cell lineages as branches of a tree. **CALISTA** (Papili Gao et al., 2020) integrates clustering, lineage progression, transition gene identification, and pseudotime ordering into a unified framework, constructing lineage trees that represent the developmental trajectories of cells based on statistical modeling of gene expression patterns.

4.5 Limitations of Traditional Tree Methods

Computational Complexity: Traditional tree construction methods, such as Maximum Likelihood (ML) and Bayesian Inference, have been foundational in phylogenetics due to their robust statistical frameworks (Felsenstein, 1981). However, these methods can be computationally intensive, particularly when handling large datasets. As the number of sequences increases, the number of possible tree topologies grows exponentially, making exhaustive searches impractical. While heuristic approaches like RAxML (Stamatakis, 2014) and MrBayes (Ronquist et al., 2012) have been developed to mitigate these challenges, they still require significant computational resources. This can result in long processing times and high memory usage when applied to datasets containing thousands or millions of sequences, potentially limiting their scalability in the context of high-throughput sequencing.

Scalability Challenges: Traditional phylogenetic methods are primarily designed for analyzing single types of sequence data, such as DNA or protein sequences. With the rise of multi-omics approaches that integrate genomic, transcriptomic, proteomic, and epigenomic data, these methods may encounter scalability and adaptability issues (McCormack et al., 2013). The diverse data types and complex interrelationships present challenges in constructing accurate trees that fully capture the biological context. Nonetheless, ongoing methodological advancements are gradually enhancing the ability of traditional approaches to accommodate more complex and high-dimensional datasets.

Model Dependency: The reliance on predefined evolutionary models is a hallmark of traditional phylogenetic methods. These models typically assume uniform evolutionary rates and patterns across all sequences (Jukes and Cantor, 1969). While this simplifies the modeling process, it may not always reflect the true evolutionary dynamics, where rates can vary across lineages and different genes may experience distinct selective pressures (Yang, 1994). This model dependency can introduce biases or inaccuracies in tree inferences. However, advancements in model selection and the development of more flexible models are progressively addressing these limitations, allowing for more accurate representations of evolutionary processes.

Handling Uncertain and Noisy Data: In real-world datasets, issues such as sequencing errors, gene loss, and missing data are common and can impact the accuracy of tree construction (Lemey et al., 2009). Traditional methods like Maximum Likelihood are sensitive to these uncertainties, which may result in less robust tree topologies. Despite these challenges, various strategies, including data preprocessing, error correction, and the incorporation of uncertainty into model frameworks, have been employed to improve the resilience of traditional methods against noisy and incomplete data.

5 Deep Learning Based Biological Tree Construction Methods

5.1 Deep General Tree Construction Methods

Tree generation is a critical research problem with wide applications in biological evolution analysis, lineage inference, and hierarchical classification system construction. Unlike general graph generation tasks, tree generation assumes a clear evolutionary direction or hierarchical relationship, necessitating the capturing of structural features and strict constraints such as acyclicity and single-root properties. This complexity makes tree generation more challenging. In this section, we review recent advances in deep learning-based tree generation methods, focusing on models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and autoregressive models. Each method's characteristics and applications in tree generation are discussed, along with their strengths and limitations. Figure 10 provides an overview of the three common tree generation methods discussed in this section.

GAN-Based Tree Generation Methods. Generative Adversarial Networks (GANs) employ adversarial training between a generator that produces graph structures and a discriminator that evaluates their realism, playing a significant role in graph generation tasks. Classical models like **NetGAN** generate graphs by learning random walk sequences on existing graphs, showcasing effectiveness in network reconstruction tasks (Bojchevski and Günnemann, 2018). Building on this, MolGAN extends the GAN framework to molecular graphs, focusing on chemical properties, which has significant applications in drug design (De Cao and Kipf, 2018).

More sophisticated GANs, such as **Hierarchical GANs**, introduce complex generative structures, including **GAN-Tree** and **Hierarchical GAN-Tree**, to handle multimodal data distributions and multi-label classification tasks (Kundu et al., 2019; Wang et al., 2022). The GAN-Tree model incrementally learns a hierarchical generative structure for multimodal data, offering a versatile framework for multimodal data generation. This incremental learning of tree-like structures enables it to effectively handle image generation and multi-label classification tasks, outperforming traditional GAN models in these scenarios.

Further advancing the GAN-based approach, **HC-MGAN** introduces a hierarchical generation strategy using multi-generator GANs (MGANs) for deep clustering (Mello et al., 2022). It achieves hierarchical data

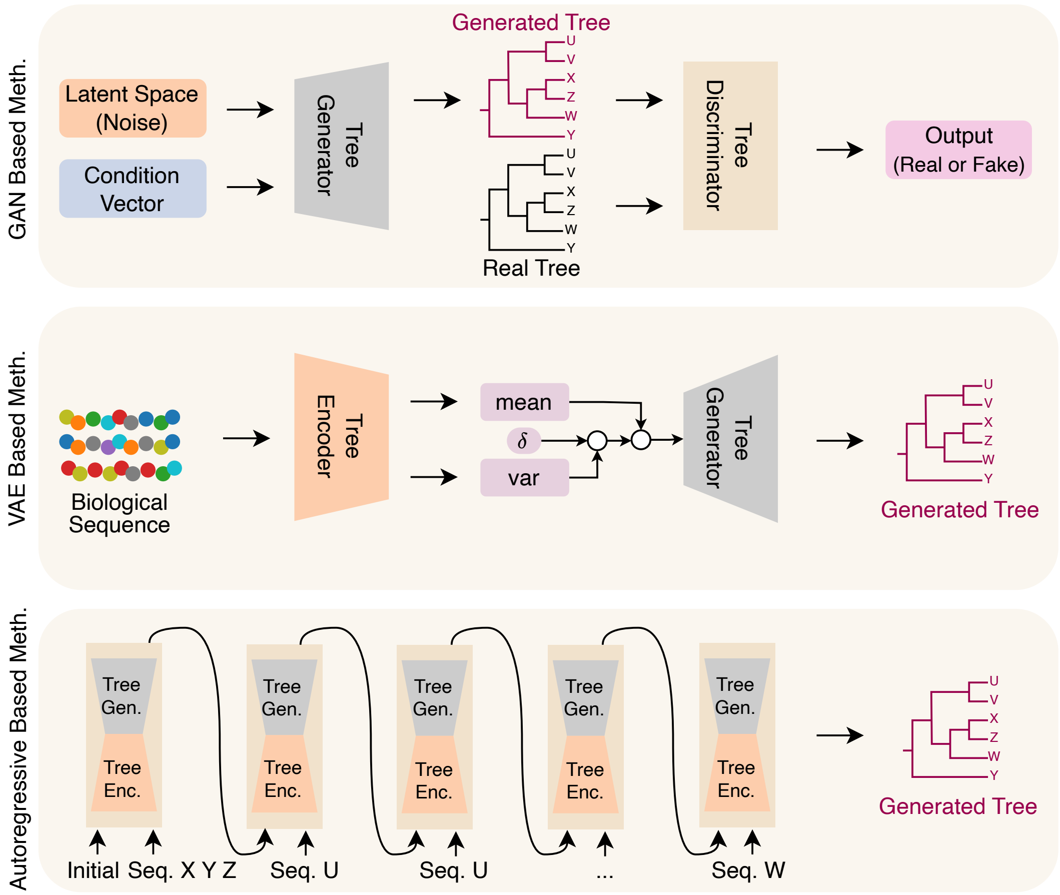

Figure 10: **The Deep Learning-Based Tree Generation Methods.** This figure summarizes three common tree generation methods for biological sequence analysis: GAN-Based Method, which uses a latent space and a condition vector to generate trees, with a discriminator distinguishing real from generated trees; VAE-Based Method, which encodes sequences into a latent space and generates trees by sampling from it; and Autoregressive-Based Method, which iteratively generates trees from an initial sequence and subsequent sequences using an autoregressive model.

organization through top-down clustering trees, offering meaningful clustering of real data distributions and a novel method for tree structure generation tasks. Additionally, the **Hierarchical GAN-Tree (HGT)** model combines bidirectional capsule networks to enhance feature generation through unsupervised divisive clustering, addressing mode collapse issues commonly found in traditional GANs (Wang et al., 2022).

These GAN-based tree generation methods excel in managing complex data distributions and hierarchical structures. However, they still face challenges under strict tree structure constraints, such as acyclicity. Their performance can potentially be enhanced by integrating other generative strategies like VAEs or autoregressive models, especially for generating larger and more intricate tree structures.

VAE-Based Tree Generation Methods. Variational Autoencoders (VAEs) offer a probabilistic approach to learning latent representations of graph structures, providing potential solutions for generating specific tree structures. Although traditional VAEs, like **VGAE**, have shown great performance in graph representation and link prediction tasks (Kipf and Welling, 2016b), their unconstrained generation process can result in structures that do not adhere to the hierarchy and acyclicity requirements of trees.

To address these constraints, the **Tree Variational Autoencoder (TreeVAE)** introduces a generative hierarchical clustering model that learns a flexible tree-based posterior distribution over latent variables (Manduchi et al., 2023). This model enables the generation of samples while preserving the hierarchical structure, proving effective in data clustering and generation tasks. Similarly, the **Junction Tree Variational Autoencoder (JTVAE)** tackles the challenge of chemical graph generation by converting the problem into tree generation (Jin et al., 2018a). It first generates a tree-structured scaffold, followed by a message-passing network that reconstructs the molecular graph. This two-step method ensures chemical validity and has demonstrated superiority over previous state-of-the-art methods in various molecular design tasks.

Diffuse-TreeVAE further enhances VAE-based tree generation by integrating it into the framework of Denoising Diffusion Probabilistic Models (DDPMs) for image generation (Gonçalves et al., 2024). This approach generates root embeddings for a learned latent tree structure, propagating through hierarchical paths, and uses a second-stage DDPM to refine and produce high-quality images. It overcomes the limitations of traditional VAE models, contributing to advancements in clustering-based generative modeling. Additionally, researchers have emphasized uncertainty quantification (UQ) in generative models. For instance, **Leveraging Active Subspaces for Epistemic Model Uncertainty** captures model uncertainty in the JT-VAE model by leveraging low-dimensional active subspaces without altering the model architecture (Abeer et al., 2024). This method has shown effectiveness in molecular optimization tasks.

Overall, VAE-based methods, particularly those employing hierarchical structures like TreeVAE and JTVAE, address the constraints required for tree generation. However, they still need refinement in scaling to larger and more complex tree structures.

Autoregressive Tree Generation Methods. Autoregressive models, such as **GraphRNN**, treat graph generation as a sequential process, where nodes and edges are generated step-by-step (You et al., 2018b). This sequential nature allows for fine-grained control over hierarchical relationships and dependencies inherent in tree structures. By explicitly modeling the generation order, GraphRNN ensures the preservation of acyclicity and hierarchical properties, making it particularly suited for generating trees.

Applications of GraphRNN to tree generation include the construction of biological family trees and evolutionary trees, where maintaining hierarchical information is crucial. The stepwise approach of autoregressive models offers advantages in controlling the generated structure's complexity and depth, providing flexibility in the creation of diverse tree structures. However, the inherent sequential process can be computationally intensive, particularly as the tree size increases.

In summary, deep learning-based tree generation methods offer diverse approaches, each with its own set of strengths and limitations. GAN-based models are powerful in handling complex data distributions but face challenges in strictly adhering to tree constraints. VAE-based methods provide a probabilistic framework suitable for hierarchical clustering and molecular design but require further enhancement to scale to larger tree structures. Autoregressive models, while maintaining strict control over hierarchical generation, may encounter computational limitations as tree complexity grows. Future research may benefit from combining these methods to leverage their individual strengths, creating more robust and scalable solutions for tree generation tasks.

5.2 Deep Gene-Based Phylogenetic Tree Construction Methods

As shown in Table. 10 and Figure. 11, recent advances in deep learning have significantly advanced the field of phylogenetics, leading to the development of novel algorithms and techniques that improve the accuracy, efficiency, and scalability of phylogenetic inference. These methods rely on deep learning architectures to address challenges faced by traditional approaches. Based on the prior knowledge they utilize and the problems

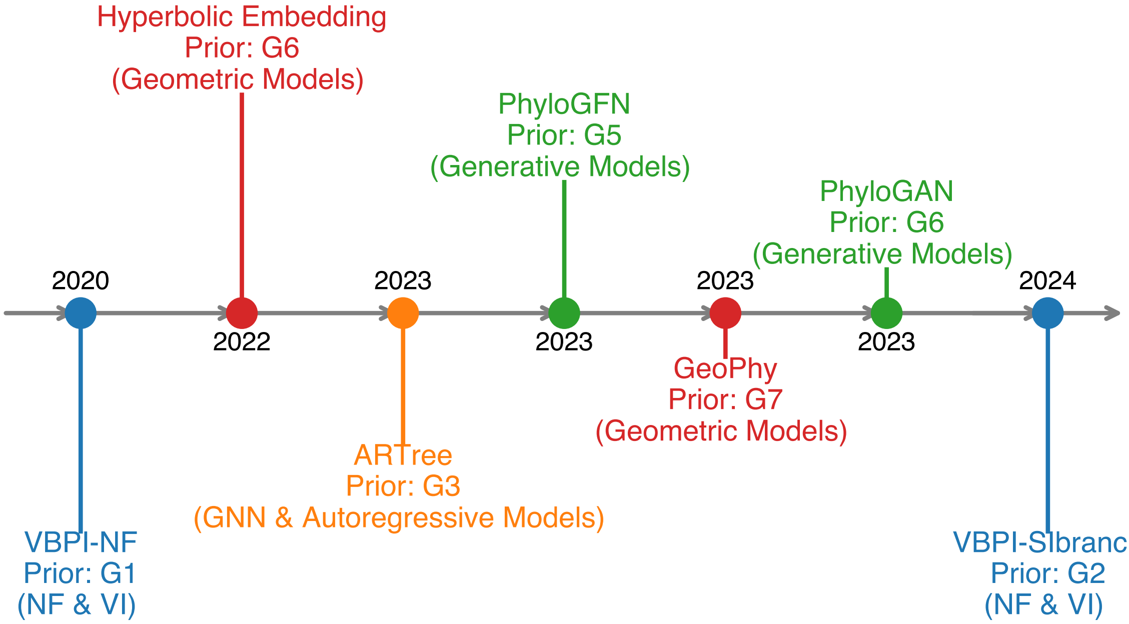

Figure 11: **The Timeline of Deep Gene Tree Construction Methods.** The figure shows the development of deep learning-based gene tree construction methods in phylogenetics from 2020 to 2024, categorized into normalizing flows and variational inference methods, graph neural network (GNN) and autoregressive models, and geometric and generative models. Different colors indicate different categories.

they tackle, existing deep learning methods can be categorized into three main groups: normalizing flows and variational inference methods, graph neural network (GNN) and autoregressive models, and geometric and generative models.

First, normalizing flows and variational inference methods have shown great promise in handling the complex, non-Euclidean tree space in phylogenetic inference. Traditional methods often struggle to handle uncertainty in tree variables, particularly when dealing with large-scale gene data. **VBPI-NF** (Zhang, 2020) employs normalizing flows to model branch length distributions across different tree topologies, using *conserved genomic regions (G1)* as model input to ensure consistency in evolutionary relationships (Altschul et al., 1990; Mount, 2004; Notredame, 2007). **VBPI-SIbranch** (Xie et al., 2024) further improves the efficiency of handling non-Euclidean branch length spaces using GNNs and incorporates *evolutionary substitution models (G2)* to model changes in nucleotide sequences over time (Felsenstein, 1981; Kimura, 1980; Tavaré, 1986).

Second, graph neural network (GNN) and autoregressive models have become crucial tools for developing flexible probabilistic models. **ARTree** (Xie and Zhang, 2023) decomposes tree topologies into node addition operations and uses GNNs to model conditional distributions, avoiding the need for manually engineered heuristic features. This method combines *evolutionary substitution model (G2)* to train a deep network.

Geometric and generative models form the third major category, focusing on representing tree topologies in continuous geometric spaces and leveraging deep learning frameworks. **PhyloGFN** (Zhou et al., 2023) is based on generative flow networks (GFlowNets) and is designed to sample tree topologies and evolutionary distances from a multimodal posterior distribution. This method leverages *sequence homology information (G5)* to improve the diversity and quality of evolutionary hypotheses and address parsimony-based and Bayesian inference challenges (Smith and Waterman, 1981a; Thompson et al., 1994a). The **hyperbolic embedding method** (Jiang et al., 2022) embeds gene sequences into hyperbolic space to reduce distortion compared to

Table 10: Overview of the Classical Gene-based Tree Construction Methods.

Method Name	Description	Reference	URL
VBPI-NF	Uses normalizing flows to model branch length distributions across tree topologies, improving flexibility in non-Euclidean tree space.	Zhang (2020)	`https://github.com/zcrabbit/vbpi-nf`
Hyperbolic Embedding	Embeds gene sequences into hyperbolic spaces to reduce distance distortion, improving species tree distance modeling.	Jiang et al. (2022)	`https://github.com/yueyujiang/hdepp`
ARTree	Autoregressive model that decomposes tree topology into sequences of leaf node additions, using GNNs for tree topology estimation.	Xie and Zhang (2023)	`https://github.com/tyuxie/ARTree`
PhyloGFN	Utilizes generative flow networks (GFlowNets) to sample from the multimodal posterior distribution over tree topologies and evolutionary distances.	Zhou et al. (2023)	`https://github.com/zmy1116/phylogfn`
Geophy	Fully differentiable method for phylogenetic inference in continuous geometric spaces, incorporating chromatin accessibility data.	Mimori and Hamada (2023)	`https://github.com/m1m0r1/geophy`
PhyloGAN	Generative adversarial network (GAN) model for inferring phylogenetic relationships by generating data similar to real evolutionary data.	Smith and Hahn (2023)	`https://github.com/meganlsmith/phyloGAN/`
VBPI-SIbranch	Applies graph neural networks (GNNs) to handle non-Euclidean branch length space with improved computational efficiency.	Xie et al. (2024)	`https://github.com/tyuxie/vbpi-sibranch`

Euclidean space, improving the accuracy of phylogenetic placement. It incorporates *the genomic linear order of genes (G3)* and *gene duplication and loss events (G6)* to model species tree distances effectively (Gu et al., 2005; Hahn, 2009). Another method, GeoPhy Mimori and Hamada (2023) introduces **GeoPhy**, a fully differentiable model for phylogenetic inference that embeds tree topologies into continuous geometric spaces, including Euclidean and hyperbolic spaces. It eliminates the need for preselected topologies, enabling end-to-end optimization by leveraging geometric transformations and variance reduction techniques.

Finally, deep learning models based on generative adversarial networks (GANs), such as **PhyloGAN** (Smith and Hahn, 2023), represent the frontier of phylogenetic inference. PhyloGAN generates evolutionary data and uses *gene duplication and loss events (G6)* to optimize data generation, providing a heuristic search mechanism for exploring complex model spaces that traditional methods struggle to navigate (Gu et al., 2005; Hahn, 2009).

In summary, deep learning has greatly advanced phylogenetic inference by handling large, complex datasets, optimizing computational efficiency, and improving the accuracy of evolutionary hypothesis generation. These methods incorporate a variety of prior knowledge, such as the genomic linear order of genes, evolutionary substitution models, and homology information, to enhance the inference of phylogenetic trees and networks. Future research should focus on integrating these deep learning methods with traditional phylogenetic models to address more complex evolutionary processes, thereby offering a more comprehensive understanding of biological evolution.

5.3 Deep Protein-Based Phylogenetic Tree Construction Methods

As shown in Table 11 and Figure 12, phylogenetic inference methods based on protein sequence and structure have made significant advances, particularly in improving efficiency and accuracy when handling large-scale datasets. These methods can be broadly categorized into two main types: sequence-based and structure-based inference methods. As data volume continues to grow, traditional methods have encountered challenges related to computational complexity, which have prompted the introduction of novel algorithms, prior

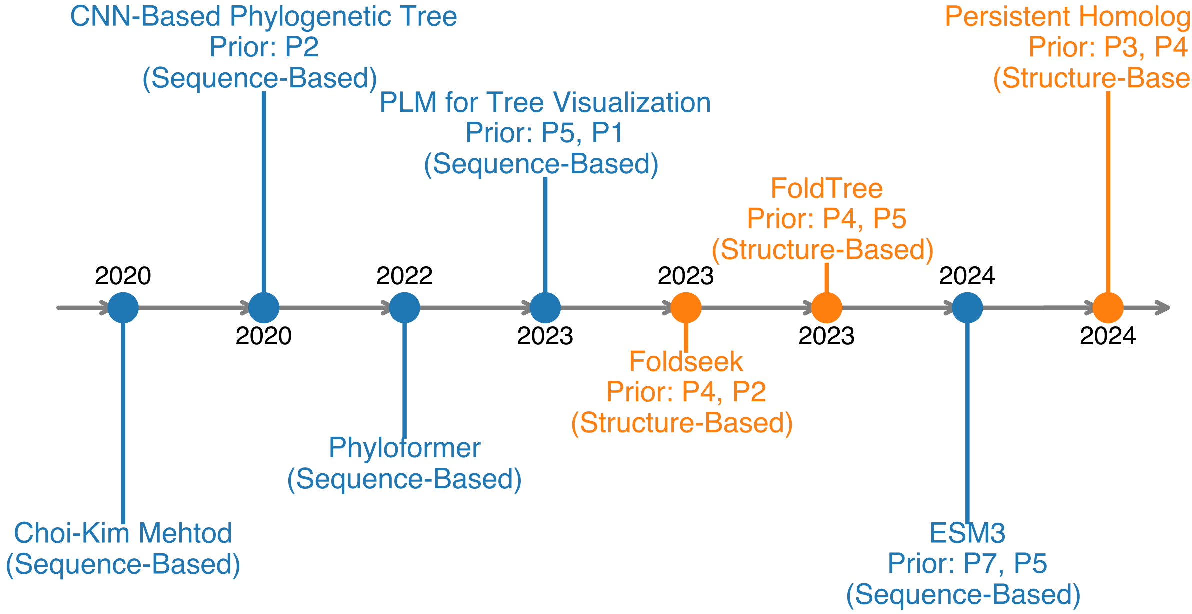

Figure 12: **The Timeline of Deep Protein Tree Construction Methods.** The figure shows the chronological development of phylogenetic tree inference methods based on protein sequence and structural information. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in evolutionary analysis.

knowledge, and deep learning techniques to drive further innovation in the field of phylogenetic inference.

5.3.1 Deep Protein Sequence-Based Phylogenetic Tree Methods.

First, the **Choi-Kim Method** Choi and Kim (2020) used whole-proteome data to construct a tree of life, revealing the evolutionary relationships among extant organisms. They applied information-theoretic methods to construct a topologically stable tree and proposed the concept of a deep burst of organismal diversity near the root of the evolutionary tree. This approach provided a new perspective on large-scale evolutionary studies, particularly in reconstructing the evolutionary history of diverse species. In this method, *Conserved Protein Domains (P1)* were used as prior knowledge, employing the indicator function $I(d_i^p, d_j^p)$ to identify conserved regions within protein sequences, reflecting their functional importance Marchler-Bauer et al. (2011); Murzin et al. (1995). To address challenges related to large datasets, Suvorov et al. (2020) proposed a convolutional neural network (CNN)-based approach to infer phylogenetic tree topologies from multiple sequence alignments (**CNN-Based Phylogenetic Tree**). This method leveraged features extracted from multiple sequence alignments to optimize the inference process. The CNN-based approach utilized *Evolutionary Models for Amino Acid Substitution (P2)*, described by the substitution matrix Q, to account for the rate of amino acid substitutions over evolutionary time Dayhoff et al. (1978); Jones et al. (1992), thus enhancing the accuracy of phylogenetic tree inference.

Deep learning frameworks have also shown great potential in combining different methods for phylogenetic inference. For example, **Phyloformer** Nesterenko et al. (2022) uses a transformer-based architecture to predict evolutionary distances between sequences and reconstruct tree topologies using standard distance-based algorithms. Yeung et al. (2023) developed a sequence embedding tree visualization method based on protein language models (**PLM for Tree Visualization**), aimed at improving the functional clustering of diverse

Table 11: Overview of the Classical Protein-based Tree Construction Methods.

Method Name	Description	Reference	URL
Choi-Kim Mehtod	Sequence-based method using whole-proteome data and evolutionary substitution models to infer phylogenetic relationships.	Choi and Kim (2020)	https://github.com/jaejinchoi/FFP
CNN-Based Phylogenetic Tree	CNN-based method for inferring tree topologies from multiple sequence alignments, improving accuracy and speed.	Suvorov et al. (2020)	https://github.com/SchriderLab/Tree_learning
Phyloformer	Transformer-based network architecture that predicts evolutionary distances between sequences, allowing for rapid tree topology reconstruction.	Nesterenko et al. (2022)	https://github.com/lucanest/Phyloformer
PLM for Tree Visualization	Embedding-based tree visualization to enhance functional clustering of protein sequences.	Yeung et al. (2023)	github.com/esbgkannan/chumby
Foldseek	Converts protein structures into structural alphabets for fast search and alignment.	van Kempen et al. (2023)	https://github.com/steineggerlab/foldseek
FoldTree	Infers relationships using tertiary structure and functional site conservation.	Moi et al. (2023)	https://github.com/DessimozLab/fold_tree
ESM3	Language model for simulating protein evolution using co-evolutionary relationships.	Hayes et al. (2024)	https://www.evolutionaryscale.ai/blog/esm3-release
Persistent Homology (PH)	Applies topological data analysis to capture structural phylogenetic signals.	Bou Dagher et al. (2024)	N/A

protein superfamilies. By generating tree-like structures, the method effectively captures global topological relationships and local functional clustering, making it particularly powerful for visualizing and classifying high-dimensional protein sequence data. In these methods, *Protein Family Classification (P6)* serves as key prior knowledge, grouping proteins based on sequence and structural similarity to reflect their evolutionary origins Bateman et al. (2002); Finn et al. (2016).

Additionally, Hayes et al. (2024) introduced **ESM3**, a multimodal generative language model that further advanced sequence-based phylogenetic inference. ESM3 could simulate evolutionary processes over hundreds of millions of years, generating functional proteins that were significantly divergent from known proteins. This method showcased its ability to tackle complex evolutionary tasks and large datasets, providing unprecedented computational efficiency and innovation in generating new functional proteins. Here, *Functional Site Conservation (P5)* was utilized as prior knowledge, using the function $F(x_i^p, x_j^p)$ to identify and measure critical functional sites within proteins Bartlett et al. (2002); Thornton et al. (2000), thereby enhancing the accuracy of evolutionary tree construction.

5.3.2 Deep Structure-Based Phylogenetic Tree Methods.

In structure-based methods, protein structure information has provided deeper insights into evolutionary relationships. van Kempen et al. (2023) proposed **Foldseek**, a method that converts protein tertiary structure into structural alphabets to significantly improve structure search speed. Foldseek relied on structural alignment to enable fast inference across large protein structure datasets. In these methods, *Tertiary Structure Conservation (P4)* serves as crucial prior knowledge, with the root-mean-square deviation (RMSD) used to measure the conservation of protein 3D structure, which is often more conserved than the primary sequence Sali and Blundell (1994).

Building on structural analysis, Bou Dagher et al. (2024) introduced **Persistent Homology (PH)** for phylogenetic inference, marking the first application of topological data analysis in this field. PH calculated the topological features of protein tertiary structures to measure evolutionary distances. This method captured

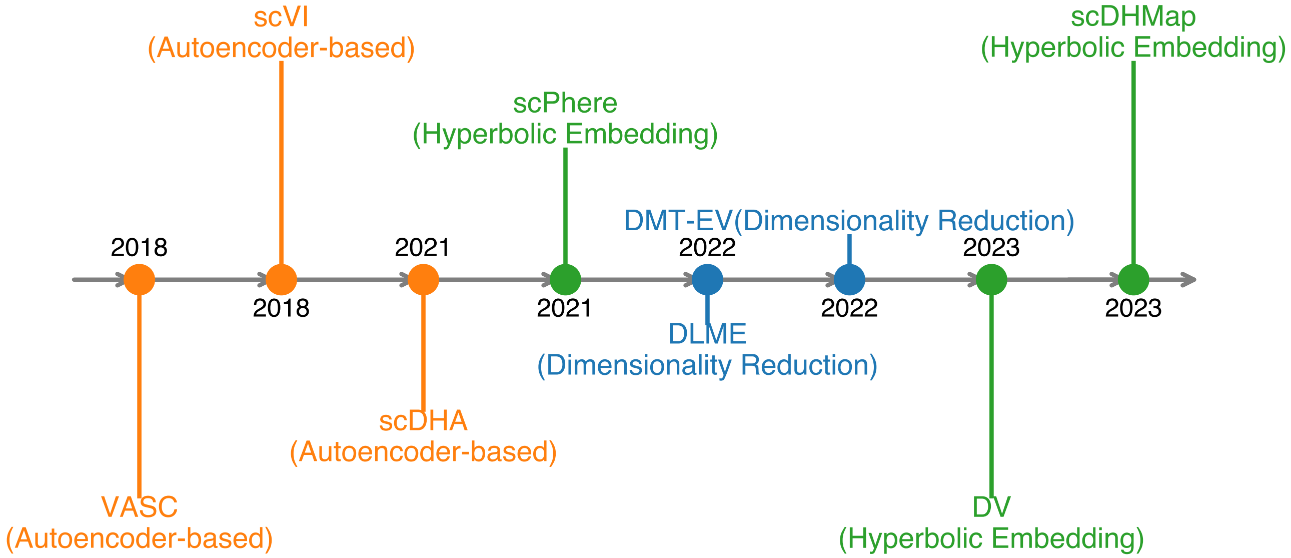

Figure 13: **The Timeline of Deep Single Cell Tree Construction Methods.** The figure shows the chronological development of trajectory inference methods based on single-cell RNA sequencing data. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in cell development analysis.

strong phylogenetic signals within protein structures, offering a novel approach for analyzing evolutionary relationships at both small and large evolutionary scales. Here, *Protein Secondary Structure Information (P3)* was utilized as prior knowledge, employing the similarity matrix S to identify conserved secondary structures such as alpha-helices and beta-sheets, reflecting important evolutionary features Chothia and Finkelstein (1984); Kabsch and Sander (1983).

Moi et al. (2023) extended structure-based methods with **FoldTree**, a method designed to infer evolutionary relationships between proteins with large evolutionary distances. The application of FoldTree in studying the evolutionary diversification of protein families demonstrated its strength in handling complex evolutionary histories by combining structural conservation and functional site information. In this context, *Functional Site Conservation (P3)* was again used as prior knowledge, leveraging the function $F(x_i^p, x_j^p)$ to identify critical functional sites within proteins Bartlett et al. (2002); Thornton et al. (2000), thus improving the accuracy of phylogenetic tree construction.

5.4 Deep Single-Cell-Based Lineage Tree Construction Methods

As shown in Table.12 and Figure.13, in the field of single-cell RNA sequencing (scRNA-seq), deep learning techniques have emerged as powerful tools for handling high-dimensional and sparse data, particularly in inferring cellular differentiation pathways and generating differentiation trees. By integrating dimensionality reduction methods with pseudo-time analysis, researchers have gained deeper insights into the transitions of cellular states and the dynamics of gene expression. This section focuses on various approaches, including dimensionality reduction-based methods, deep generative models, and RNA velocity-based methods, illustrating how these techniques leverage the strengths of deep learning to significantly enhance the accuracy and interpretability of differentiation tree construction, providing novel perspectives and tools for understanding biological processes.

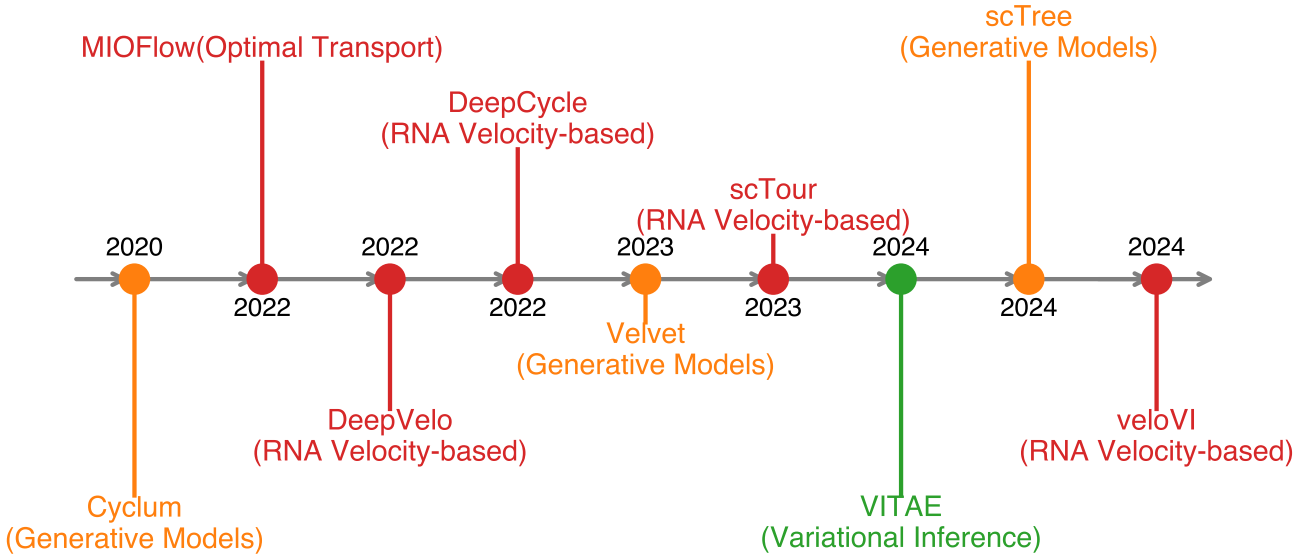

Figure 14: **The Timeline of Deep Single Cell Tree Construction Methods.** The figure shows the chronological development of trajectory inference methods based on single-cell RNA sequencing data. These methods have evolved by incorporating different types of prior knowledge to improve accuracy and computational efficiency in cell development analysis.

Dimensionality Reduction-based Methods. The existing methods can generally be classified into two main categories: dimensionality reduction methods and dimensionality reduction integrated with pseudo-time analysis, both contributing to the generation of differentiation trees by capturing the hierarchical structure of cell states.

For **dimensionality reduction methods**, high-dimensional *Cell Type-Specific Marker Genes (S3)* are projected into a lower-dimensional space, which serves as a foundation for constructing the differentiation tree by identifying distinct cellular states. Deep manifold learning methods have been increasingly utilized for dimensionality reduction in single-cell data analysis, thereby aiding in the generation of differentiation trees. **DMAGE (deep manifold attributed graph embedding)** (Zang et al., 2021) effectively captures both structural and feature information in latent spaces by leveraging node-to-node geodesic similarities. This allows for a more accurate reconstruction of cellular relationships, which is crucial for inferring cell differentiation pathways. Their subsequent works, **DLME (deep local-flatness manifold embedding)** (Zang et al., 2022b), address the challenges posed by under-sampled data through data augmentation (Zang et al., 2024b) and local flatness constraints, further enhancing the accuracy of cell state embeddings and thus improving differentiation tree construction. Similarly, **UDRN (unified dimensional reduction neural-network)** (Zang et al., 2023b) integrates feature selection and feature projection, ensuring that the essential cellular features are preserved in the reduced space, facilitating the differentiation tree generation process. **DMT-EV** (Zang et al., 2022a) enhances both performance and explainability by using manifold-based loss functions to maintain cellular hierarchical structures in the latent space, which directly benefits the generation of differentiation trees.

Autoencoder-based methods, such as **VASC** (Wang and Gu, 2018) and **scVI** (Lopez et al., 2018), encode high-dimensional *Gene Expression Profiles (S1)* into lower-dimensional latent spaces, capturing key information about cellular states. These methods not only improve the visualization and clustering of cells but also support the construction of differentiation trees by revealing the underlying branching patterns of cell lineages. **scDHA (single-cell decomposition using hierarchical autoencoder)** (Tran et al., 2021; Zhang et al., 2023) filters

Table 12: Overview of Deep Learning Methods in Single-Cell Trajectory Inference.

Method Name	Description	Reference	URL
	Dimensionality Reduction-based Methods		
VASC	Models scRNA-seq data distribution and clusters latent space for improved dimensionality reduction.	Wang and Gu (2018)	https://github.com/wang-research/VASC
scVI	Applies VAE to single-cell transcriptomic data, addressing noise and dropout events.	Lopez et al. (2018)	https://github.com/YosefLab/scVI
scDHA	Uses a non-negative kernel autoencoder for filtering insignificant genes in scRNA-seq data.	Tran et al. (2021)	https://github.com/duct317/scDHA
scPhere	Uses deep hyperbolic embedding to compute pseudotime in hyperbolic space.	Ding and Regev (2021)	https://github.com/klarman-cell-observatory/scPhere
DLME	Addresses under-sampled data through data augmentation and local flatness constraints.	Zang et al. (2022b)	https://github.com/zangzelin/code_ECCV2022_DLME
DMT-EV	Enhances dimensionality reduction performance and explainability using manifold-based loss functions.	Zang et al. (2022a)	https://github.com/zangzelin/code_EVNet_DMTEV
MIOFlow	Aligns geodesic distances on the data manifold to accurately reconstruct trajectories.	Huguet et al. (2022)	https://github.com/KrishnaswamyLab/MIOFlow
VITAE	Combines hierarchical models with VAEs to map the latent space of single-cell data.	Du et al. (2024)	https://github.com/jaydu1/VITAE
	Deep Generative Models		
Cyclum	Uses autoencoders to identify cyclic trajectories in gene expression data.	Liang et al. (2020)	https://github.com/KChen-lab/cyclum
scTree	VAE-based method integrating hierarchical clustering with batch correction.	Vandenhirtz et al. (2023)	https://github.com/mvandenhi/sctree-public
Velvet	Models gene expression dynamics using a VAE and neural stochastic differential equation system.	Maizels et al. (2023)	https://github.com/rorymaizels/velvet
	RNA Velocity-based Methods		
DeepVelo	Uses neural network-based ODE framework to model transcriptional dynamics and RNA velocity.	Chen et al. (2022)	https://github.com/bowang-lab/DeepVelo
DeepCycle	Analyzes cell cycle gene regulation dynamics in scRNA-seq data using deep learning.	Riba et al. (2022)	https://github.com/andreariba/DeepCycle
scTour	Infers cellular dynamics using a VAE and neural ODE framework, minimizing batch effects.	Li (2023)	https://github.com/LiQian-XC/sctour
veloVI	Shares information across all cells to learn kinetic parameters and latent time for RNA velocity inference.	Gayoso et al. (2024)	https://github.com/YosefLab/velovi

insignificant genes and projects data into a lower-dimensional space, providing a more focused view of the essential differentiation trajectories.

Dimensionality reduction integrated with pseudo-time analysis incorporates prior information on *Pseudotime Ordering (S4)*, facilitating differentiation tree generation by tracking the transitions between cell states over time. Deep hyperbolic embedding methods, such as **scPhere** (Ding and Regev, 2021) and **scDHMap** (Tian et al., 2023), compute hyperbolic distances in latent space to infer pseudotime, effectively reconstructing differentiation pathways. By integrating pseudo-time and cell embeddings, these methods generate more accurate differentiation trees that represent the temporal progression and branching of cellular differentiation processes. Additionally, **VITAE (variational inference for trajectory by autoEncoder)** (Du et al., 2024) provides a hierarchical model that assigns edge scores to cell transitions, directly informing the construction of the differentiation tree's backbone.

Deep Generative Models. Deep generative models, such as autoencoders and VAEs, focus on capturing the latent distribution of *Gene Expression Profiles (S1)* to simulate cell state transitions, thereby serving as critical tools in differentiation tree generation. For instance, **Cyclum** (Liang et al., 2020) uses autoencoders to identify cyclic trajectories in gene expression, helping to elucidate differentiation cycles within the differentiation tree. **scTree** (Vandenhirtz et al., 2023) integrates hierarchical clustering with batch correction to enhance the identification of cellular hierarchies, using a tree-structured approach to represent differentiation paths.

Similarly, **Velvet** (Maizels et al., 2023) models global gene expression dynamics in latent space, providing a comprehensive view of the differentiation landscape.

RNA Velocity-based Methods. Several methods estimate *RNA Velocity (S2)* to model cellular trajectories and generate differentiation trees. **veloVI** (Gayoso et al., 2024) shares information across cells and genes to learn latent time and kinetic parameters, improving the accuracy of inferred differentiation paths. **scTour** (Li, 2023) uses a deep learning architecture built on VAE and neural ODEs to estimate pseudotime and map cells into a latent space, facilitating differentiation tree generation. By modeling continuous transcriptional dynamics, **DeepVelo** (Chen et al., 2022) provides a refined view of gene expression changes, directly contributing to the construction of high-resolution differentiation trees. **DeepCycle** (Riba et al., 2022) fits cycling patterns observed in the unspliced-spliced RNA space, offering a detailed map of differentiation processes during the cell cycle.

6 Applications of Biological Trees

Phylogenetic trees, also known as evolutionary trees or phylogeny, have widespread applications in biology, spanning from species evolution analysis to molecular phylogenetics. This section provides a detailed overview of these applications along with specific examples.

6.1 Biological Trees in the Study of Infectious Diseases

Phylogenetic trees play a pivotal role in infectious disease research, serving as essential tools for tracing the origins, transmission, and evolutionary dynamics of pathogens across various biological scales. By integrating molecular data with evolutionary models, these analyses offer insights into the complex processes underlying the emergence and spread of infectious agents, with significant implications for public health interventions.

At the molecular and evolutionary level, phylogenetic analyses are indispensable for identifying the origins and reconstructing the evolutionary trajectories of viral pathogens. One prominent example is the classification of SARS-CoV-2 as a novel coronavirus, achieved through comprehensive phylogenetic analyses that revealed its close genetic relationship to bat coronaviruses. This classification provided the foundation for understanding SARS-CoV-2 as the causative agent of the COVID-19 pandemic (Gorbalenya et al., 2020). Furthermore, phylogenetic methods have been critical in tracking the evolutionary divergence of SARS-CoV-2 variants, including the Omicron subvariants BA.4, BA.5, and XBB. These analyses not only traced the lineage-specific mutations that differentiated these variants but also shed light on their global spread and potential public health impacts, aiding in the timely identification of new threats (Tamura et al., 2023; Tegally et al., 2022).

Beyond the molecular level, phylogenetic tools have been extensively applied to monitor virus transmission dynamics within and between populations. These analyses provide critical insights into how pathogens adapt and evolve over time, often revealing the complex interplay between viral evolution and transmission patterns. For example, research on the spread of highly pathogenic avian influenza A (H5N1) among marine mammals and seabirds in Peru utilized phylogenetic trees to trace genetic reassortments that facilitated cross-species transmission, highlighting the zoonotic potential of these viruses and underscoring the importance of phylogenetic analysis in predicting future spillover events (Leguia et al., 2023). Similarly, studies on SARS-CoV-2 transmission within immunocompromised individuals have demonstrated how intrahost viral evolution can contribute to the emergence of new variants, further complicating efforts to control the pandemic and emphasizing the role of phylogenetics in understanding viral persistence and adaptation in specific host populations (Gonzalez-Reiche et al., 2023).

In addition to its application in pandemic contexts, phylogenetic analysis has been employed to explore co-infections involving non-pandemic viruses, broadening its utility in virology. A notable case is the investigation of Adeno-associated virus type 2 (AAV2) in U.S. children with acute severe hepatitis, where phylogenetic methods were used to assess viral relationships and explore the role of co-infections in disease severity. This example demonstrates the versatility of phylogenetic tools beyond pandemic viruses, showcasing their broader applicability in elucidating complex viral interactions (Servellita et al., 2023).

Overall, phylogenetic trees are invaluable in infectious disease research, providing detailed insights into pathogen evolution, transmission dynamics, and cross-species interactions. By tracing the evolutionary pathways of pathogens and predicting future outbreaks, phylogenetic analyses are instrumental in informing public health strategies and shaping global responses to emerging infectious diseases.

6.2 Biological Trees in the Study of Biomarker Discovery

The integration of phylogenetic trees in biomarker discovery has emerged as a powerful analytical approach across various biological levels, offering insights into evolutionary relationships that guide the identification and validation of biomarkers. Spanning scales from microbial communities to gene family diversification, population genetics, and species-level comparative genomics, phylogenetic analysis enriches our biological understanding while presenting new opportunities for applications in precision medicine, agriculture, and environmental conservation.

At the microbial and environmental level, phylogenetic trees have become indispensable tools in metagenomics and environmental microbiology. By reconstructing evolutionary relationships within microbial communities, these trees help elucidate the functional roles of microbes in ecosystems and their potential as disease biomarkers. For instance, phylogenetic analysis has been applied to study sulfur metabolic genes in the human gut microbiome, where specific microbial genes were identified as potential biomarkers for colorectal cancer (Wolf et al., 2022; Zang et al., 2023c). This approach demonstrates how the evolutionary study of microbial genes can provide actionable insights for disease diagnosis and treatment. Similarly, the discovery of novel circular DNA viruses through phylogenetic analyses highlights the method's capacity to uncover viral diversity in previously uncharacterized environments, broadening our understanding of virology (Tisza et al., 2020). Such findings underscore the crucial role of phylogenetic trees in expanding our knowledge of microbial evolution and their application in biomarker discovery within environmental and health-related contexts.

As research transitions from microbial ecosystems to gene-level analyses, phylogenetic trees continue to play a crucial role in exploring the evolutionary history and diversification of gene families. This line of research has significant implications for identifying biomarkers related to disease resistance and functional gene evolution. For example, the structural evolution of the LRR-RLK gene family, which drives diversification in plant defense mechanisms, was explored through phylogenetic methods, offering insights into the genetic underpinnings of disease resistance (Man et al., 2020). Similarly, the evolutionary expansion of the CHS-L gene family in *Senna tora* was linked to the biosynthesis of anthraquinones, a class of compounds with pharmaceutical relevance (Kang et al., 2020). These studies demonstrate how phylogenetic analysis of gene family diversity and structural evolution can inform functional genomics and facilitate the discovery of potential biomarkers.

At the population genetics level, phylogenetic trees provide a framework for uncovering genetic diversity and structural variations associated with disease susceptibility. By integrating phylogenetic analyses with genomic data, researchers can identify population-specific biomarkers and uncover the genetic bases for gene-environment interactions. For instance, the combination of phylogenetic and structural variation analysis in diverse human populations has led to the identification of population-specific biomarkers, revealing how genetic diversity impacts disease susceptibility (Ebert et al., 2021). Furthermore, stress-responsive genes in

Nitraria tangutorum were identified through genome-wide analysis, shedding light on the genetic mechanisms underlying adaptation to environmental stressors (Zhu et al., 2023b). These studies highlight how phylogenetic trees can reveal complex genetic structures and their implications for population health and adaptation.

On a broader, species-level scale, phylogenetic trees play a fundamental role in comparative genomics, enabling the identification of species-specific biomarkers related to adaptive traits. Through cross-species comparisons, researchers can trace the evolutionary conservation and divergence of genes across species, which is crucial for understanding trait evolution and adaptation. For example, phylogenetic mapping of resistance genes in winter wheat provided valuable insights into gene conservation at the species level, with direct implications for crop improvement and disease resistance (Kale et al., 2022). In a similar vein, studies exploring gene transfer mechanisms across domains revealed evolutionary connections between archaea and eukaryotes, emphasizing the utility of phylogenetic trees in tracing gene function evolution and speciation events (Ghaly et al., 2022; Moi et al., 2022). These investigations demonstrate the power of phylogenetic analysis in revealing the evolutionary forces shaping species and their potential for informing biomarker discovery related to environmental adaptation.

In summary, phylogenetic trees serve as critical tools across multiple biological scales, offering a comprehensive approach to biomarker discovery that integrates evolutionary insights from microbial ecosystems to species-wide genomic comparisons. Whether analyzing microbial community dynamics, gene family diversification, population genetics, or species-level evolution, phylogenetic analysis provides a robust framework for understanding the complex biological processes underlying biomarker discovery. These applications not only expand our understanding of biodiversity and evolutionary mechanisms but also offer practical strategies for advancing fields such as precision medicine, agricultural enhancement, and environmental conservation.

6.3 Biological Trees in the Study of Cancer Evolution and Tumor Classification

The application of evolutionary approaches in cancer research has significantly enhanced our understanding of the onset, progression, and therapeutic resistance of tumors. Phylogenetic trees, in particular, have proven to be indispensable tools, providing deeper insights into cancer resistance mechanisms, tumor evolution under selective pressures, and the functional genomics of cancer driver genes. This section categorizes the applications of evolutionary trees in cancer research into three major areas: understanding cancer resistance mechanisms, analyzing tumor evolution and therapeutic resistance, and exploring cancer driver mechanisms through functional genomics.

Understanding Cancer Resistance Mechanisms through Evolutionary Trees. Phylogenetic trees have been instrumental in investigating natural cancer resistance mechanisms in various species. These studies aim to uncover how evolutionary adaptations, such as duplications in tumor suppressor genes, contribute to reduced cancer risk in certain species. By tracing the evolutionary pathways of these adaptations, researchers can better understand the genetic foundations of cancer resistance and potentially apply these findings to human cancer therapies.

One such study by Vazquez et al. (2022) explored the parallel evolution of reduced cancer risk in Xenarthran lineages, such as sloths and armadillos, through phylogenetic analyses. The research found that bursts of tumor suppressor gene duplications coincided with reduced cancer risk, suggesting that these genetic duplications play a pivotal role in enhancing natural cancer resistance. Similarly, Kolora et al. (2021) examined Pacific Ocean rockfish species, identifying genetic determinants associated with longevity and cancer resistance. Their findings highlighted the role of positive selection in DNA repair pathways, illustrating how evolutionary innovations contribute to cancer resistance. In another study, Wang et al. (2024) introduced PhyloVelo, a computational tool that integrates phylogenetic analysis to infer cell differentiation trajectories. This tool tracks lineage-specific adaptations and evolutionary dynamics, advancing our understanding of the molecular mechanisms underlying cancer resistance.

Collectively, these studies demonstrate how evolutionary trees can elucidate the genetic basis of natural cancer resistance, offering a foundation for developing new cancer therapies based on these insights.

Uncovering Tumor Evolution and Therapeutic Resistance through Phylogenetic Analysis. Phylogenetic trees are also employed to study tumor evolution, particularly in the context of therapeutic resistance. By reconstructing the evolutionary trajectories of tumors, researchers gain a deeper understanding of how tumors adapt to therapeutic interventions and develop resistance over time. This knowledge is crucial for designing more effective treatment strategies that target the evolutionary dynamics of cancer cells.

For example, Fisk et al. (2022) used phylogenetic analysis to study mutational processes in EGFR-driven lung adenocarcinoma. The research revealed that both endogenous factors, such as mutator gene mutations, and exogenous factors, such as mutagenic therapies, contribute to the emergence of therapeutic resistance. The study underscored the importance of considering the evolutionary pressures exerted on cancer cells when designing treatment strategies. Similarly, Kwon et al. (2020) traced the lineage dynamics of transmissible cancer in Tasmanian devils, uncovering how cancer cells adapt to different environmental and parasitic niches. This research highlighted the significance of understanding tumor evolution to combat the persistence and spread of cancer. In another example, Schmidt et al. (2023) developed the zero-agnostic copy number transformation (ZCNT) model, which optimizes tumor phylogeny inference and reveals gene changes associated with therapeutic resistance. The model represents a computational advancement in accurately modeling the evolutionary processes that lead to resistance.

These studies highlight the critical role of phylogenetic analysis in understanding the complex evolutionary processes that tumors undergo, particularly in the face of therapeutic pressures. By uncovering these dynamics, researchers can better predict resistance patterns and develop targeted treatment strategies.

Exploring Cancer Driver Mechanisms through Functional Genomics Based on Evolutionary Trees. In addition to studying cancer resistance and tumor evolution, phylogenetic trees are used to explore the functional genomics of cancer driver genes. By analyzing the evolutionary conservation and divergence of key genes, researchers can identify potential therapeutic targets and gain insight into the molecular mechanisms driving tumor progression.

For instance, Jonsson et al. (2024) investigated the role of the gene CLEC18A in clear cell renal cell carcinoma (ccRCC), utilizing phylogenetic analysis to trace its evolutionary conservation and functional divergence in cancer. This study provided insights into how CLEC18A is regulated within the tumor microenvironment and its role in tumor progression. Similarly, Zhang et al. (2024) explored the evolutionary dynamics of DNA transposable elements (TEs) in cancer cells, offering insights into genome engineering for cancer therapy. These studies underscore the value of evolutionary trees in understanding gene function evolution in the context of cancer. Furthermore, Edogbanya et al. (2021) examined the evolutionary history of the gene C1ORF112, revealing its role in DNA replication and DNA damage response, key processes implicated in cancer development. The study by Julca et al. (2023) provided a comprehensive genomic and metabolomic analysis of the medicinal plant *Oldenlandia corymbosa*, revealing biosynthetic pathways with anticancer properties, which offers a unique perspective on the evolutionary basis of therapeutic compounds. Lastly, Schmidt et al. (2023) applied the ZCNT model in functional genomics to better understand cancer driver mechanisms within complex genomic datasets.

These studies demonstrate how phylogenetic trees can be applied to uncover the evolutionary dynamics of cancer driver genes, shedding light on their roles in tumor progression and offering new avenues for therapeutic development.

In summary, phylogenetic trees have become essential tools in cancer research, enabling scientists to investigate the evolution of cancer resistance, the mechanisms underlying tumor progression and therapeutic resistance, and the functional genomics of cancer driver genes. By integrating evolutionary insights with modern computational tools, researchers can develop more effective strategies for cancer diagnosis, treatment, and prevention, paving the way for improved outcomes in cancer therapy.

6.4 Biological Trees in the Study of Agriculture and Crop Improvement

Evolutionary trees are integral to plant science research, serving as a foundational tool for evolutionary analysis across a broad spectrum of applications. They are widely used to study genomic diversity, pathogen evolution, ecosystem management, and the functional evolution of plant genes. By constructing and analyzing phylogenetic trees, researchers can uncover the evolutionary relationships among species, the patterns of genome evolution, and the adaptive strategies plants employ in diverse ecological environments. This section reviews the methodologies and applications of evolutionary trees in plant science, underscoring their essential role in advancing the field.

Application of Evolutionary Trees in Plant Genomic Diversity and Domestication Traits. In studying plant genomic diversity and domestication traits, evolutionary trees are extensively employed to analyze structural variations in genomes and to trace the evolutionary relationships of specific genes. Pangenome analysis, for example, constructs a composite genome from multiple species or varieties and integrates evolutionary trees to reveal how selective pressures and adaptive changes have shaped different genes during evolution. Chen et al. (2023b) utilized this approach to identify genetic variations associated with domestication traits in broomcorn millet, providing key insights into the genomic changes that occurred during the domestication process. Similarly, phylogenomic methods apply large-scale genomic data to build evolutionary trees that unravel the complexity of species diversity and phylogenetic relationships, offering a deeper understanding of plant evolutionary history. Guo et al. (2023) demonstrated how these phylogenetic analyses could support plant taxonomy and agricultural enhancement by identifying genetic diversity critical to adaptation and crop improvement. Additionally, co-expression network analysis, in conjunction with evolutionary trees, has been used to investigate the co-evolution and functional clustering of genes, offering molecular insights into plants' environmental adaptability and multicellular development (Feng et al., 2024). These examples underscore the utility of evolutionary trees in providing a comprehensive picture of plant genome evolution and their role in improving domestication practices.

Application of Evolutionary Trees in Plant Pathogen Evolution and Ecosystem Management. In the realm of plant pathogen evolution and ecosystem management, evolutionary trees serve as crucial tools for understanding pathogen diversity and tracing ecological dissemination pathways. Phylogenetic meta-analysis, which integrates molecular sequence data from plant pathogens, uses evolutionary trees to reveal the distribution patterns and evolutionary relationships of different pathogens. For example, Bourret et al. (2023) employed evolutionary tree analysis to study the distribution and ecological risks of plant pathogens in California, offering vital data to inform plant protection strategies. The use of evolutionary models in combination with ecological management approaches provides insights into pest evolution and resistance patterns, helping optimize management strategies in agricultural ecosystems. Thrall et al. (2011) used evolutionary tree-based models to study the mechanisms of pathogen evolution, which enabled the development of proactive management tools aimed at mitigating pest threats in agro-ecosystems. Further research, such as the work by Kan et al. (2024), explored the co-evolution of plant genomes and their interactions with pathogens, emphasizing how evolutionary trees can elucidate the molecular mechanisms behind ecological adaptation and pathogen resistance in plants.

Application of Evolutionary Trees in Plant Genomic Evolution and Functional Studies. Evolutionary trees are also pivotal in investigating plant genomic evolution and functional studies, particularly in revealing the adaptive mechanisms that underpin plant survival across diverse environments. Cytonuclear interaction analyses, which focus on the co-evolution of nuclear and organellar genomes, rely on evolutionary trees to trace how these genetic systems evolve in coordination. By analyzing whole-genome data, Kan et al. (2024) demonstrated that the co-evolution of nuclear and organellar genes plays a critical role in maintaining genomic stability during polyploidization, a process that has significantly influenced the diversification of Brassica species. Multi-omics approaches, which integrate genomic, transcriptomic, and proteomic data, further utilize evolutionary trees to explore the functional evolution of genes, shedding light on how plants adapt to environmental stresses (Jia et al., 2023). For instance, evolutionary analysis combined with chromosome-level genome assembly has been employed to study gene family expansion and evolutionary patterns, revealing the molecular underpinnings of plant ecological adaptations and behaviors, such as predation, as shown by Yuan et al. (2023). These applications demonstrate the versatility of evolutionary trees in studying plant genomic evolution and function, providing critical insights into both basic plant biology and applied agricultural science.

In summary, evolutionary trees are indispensable tools in plant research, offering profound insights into the mechanisms underlying genomic diversity, pathogen evolution, and functional gene adaptation. Their application spans multiple biological scales, from studying individual gene evolution to managing large-scale ecological systems. Through the construction and interpretation of evolutionary trees, researchers can uncover the intricate relationships that drive plant evolution, enabling advancements in agricultural improvement, ecosystem management, and the broader understanding of plant sciences. As plant science continues to evolve, the role of phylogenetic trees in uncovering the molecular mechanisms of plant adaptation and survival will remain essential, contributing to both theoretical research and practical applications in the field.

6.5 Ecology and Environmental Studies

Evolutionary biology seeks to uncover the origins of species, their relationships, and the adaptive changes they undergo. Recent advancements in molecular phylogenetics, genomics, and ecology have enabled researchers to probe the complexity of species evolution and their responses to ecological and environmental contexts more deeply. This review focuses on three central themes in current research: phylogenetic reconstruction and evolutionary relationships, genomic evolution and adaptive studies, and species diversity and biogeography. These themes help elucidate the mechanisms behind biodiversity, ecological adaptation strategies, and the role of environmental factors in shaping species evolution.

Phylogenetics and Evolutionary Relationship Reconstruction. Phylogenetic reconstruction is essential for understanding the evolutionary history of species and their adaptations to ecological pressures. By analyzing molecular data and constructing evolutionary trees, researchers can infer species relationships and divergence patterns, providing insights into how species respond to environmental challenges.

Recent studies highlight the importance of taxon sampling in evolutionary inference, as small changes in sampling can significantly alter phylogenetic outcomes. For instance, Bernot et al. (2023) revised the phylogeny of crustaceans and hexapods, showing that variations in sampling influence tree topologies and, consequently, our understanding of species' ecological adaptations. This study challenges existing phylogenetic hypotheses and underscores the significance of environmental diversity in evolutionary relationship studies. Similarly, Eme et al. (2023) reconstructed the evolutionary relationships between Asgard archaea and eukaryotes, shedding light on gene duplication and loss during early life evolution, providing insights into species' adaptations to different ecological niches.

Phylogenetic analyses have also been applied to clarify the evolutionary positions of rare species. For example, Lax and Keeling (2023) employed single-cell transcriptomics and phylogenetic tools to study *Dolium sedentarium*, confirming its unique evolutionary position in specific ecological contexts. These studies demonstrate how molecular phylogenetic methods can resolve uncertainties in evolutionary histories, offering a pathway for more precise species classification. Furthermore, studies like Maurya et al. (2023), which examined the phytogeographic history of *Capparis*, reveal how species differentiation and migration are influenced by environmental factors, further contributing to our understanding of species evolution and reclassification.

Genomic Evolution and Adaptive Studies. Research on genomic evolution investigates how structural and functional changes in genomes drive species' adaptations to diverse environments. Trait innovations, gene expansions, and genome rearrangements are key processes in ecological adaptation and diversification.

For example, the comparative genomics of multicellular algae and land plants studied by Feng et al. (2024) revealed that specific gene expansions and signaling network modifications were crucial for plant adaptation to terrestrial environments. These findings provide a theoretical foundation for understanding how genomic changes facilitate ecological adaptation. Similarly, research by Blaimer et al. (2023) on the phylogeny of Hymenoptera insects demonstrated how trait innovations like parasitism and phytophagy drive species diversification in response to environmental conditions.

In addition, studies of genome rearrangements have revealed how structural changes enable the evolution of new phenotypic traits. For instance, Marletaz et al. (2023) analyzed the genome of the little skate, uncovering how regulatory networks and genome rearrangements facilitated the evolution of its wing-like fins. These studies suggest that environmental changes are key drivers of genomic evolution and highlight the importance of understanding these dynamics for evolutionary biology.

Species Diversity and Evolutionary Biogeography. Research in species diversity and evolutionary biogeography integrates ecological and environmental data to understand how historical processes and environmental changes shape species adaptation and diversification. This approach reveals how geographical environments influence evolutionary pathways and species distributions.

The impact of human activities on species diversity and evolution has been a major focus of recent studies. Chen et al. (2023c) explored the domestication history of yaks, taurine cattle, and their hybrids on the Tibetan Plateau, showing how human activities and natural selection have jointly shaped these species' ecological adaptations. Similarly, Guo et al. (2023) analyzed the phylogeny of flowering plants, revealing the influence of whole-genome duplication and hybridization on species biogeography, further illustrating how evolutionary processes differ across ecological environments.

Genomic studies on plant domestication have also contributed to our understanding of species adaptation to environmental changes. For instance, Chen et al. (2023b) conducted a pangenome analysis of broomcorn millet, linking genomic variations to domestication traits and offering critical data for crop improvement and ecological adaptation research. These studies emphasize how environmental conditions and genomic changes interact to influence species' evolutionary trajectories, demonstrating the importance of evolutionary biogeography in understanding species diversity.

Research in phylogenetics, genomic evolution, and species diversity plays a pivotal role in modern evolutionary biology, offering a comprehensive view of biodiversity formation and species adaptation. By integrating phylogenetic reconstruction, genomic analysis, and biogeographical methods, researchers can reveal the mechanisms underlying evolutionary processes, particularly in response to changing ecological environments. These studies not only advance evolutionary biology theories but also provide essential insights for ecological conservation, biodiversity management, environmental monitoring, and agricultural development. Future research will benefit from further integration of ecological and molecular data, offering an increasingly dynamic understanding of biological evolution.

7 Current Limitations of Classical and Deep Learning-based Tree Construction Methods

7.1 Limitations of Classical Biological Tree Construction Methods

The limitations of classical biological tree construction methods in phylogenetic analysis arise from the inherent characteristics of their algorithms and theoretical assumptions, as well as a mismatch between the complexity of biological data and the requirements of modern scientific research. Understanding these limitations helps in refining current methods and developing new tools that better address the challenges of contemporary bioinformatics.

A major issue is scalability and computational complexity, which significantly limits the application of classical methods to large-scale datasets. Methods such as Maximum Likelihood (ML) and Bayesian Inference, while effective on smaller datasets, rely on algorithms that require exhaustive searches through possible tree structures. As the dataset size increases and the number of taxa grows, the combinatorial explosion effect results in computation time and resource consumption that grows exponentially. This makes large-scale phylogenetic analysis difficult to perform within traditional computational frameworks, slowing the pace of biological discovery and limiting the downstream applications of evolutionary trees, such as ecosystem conservation strategies and drug target identification (Stamatakis, 2014). In fields like metagenomics or environmental genomics, where large volumes of genetic sequence data need to be analyzed efficiently, the computational limitations of conventional methods hinder the process of discovery and the practical use of evolutionary trees.

Another significant limitation is the inadequate handling of uncertainty and missing data, which reflects the high requirements for data integrity in classical methods. Biological data, especially genomic data from field samples or historical specimens, often contain missing information. Methods like Maximum Likelihood and Bayesian Inference may yield unreliable inferences when faced with incomplete data, as these methods typically assume high-quality, complete input and cannot effectively account for uncertainty when dealing with missing data or noise (Guindon and Gascuel, 2003). This can lead to phylogenetic trees that deviate significantly from actual evolutionary histories, introducing biases in studies on species relationships, evolutionary pathways, and timescales. For example, in viral evolution research, where high mutation rates and incomplete genomic sequences are common, classical methods may severely underestimate or misinterpret the timing and pathways of key evolutionary events.

The dependence on specific model assumptions further affects the applicability and accuracy of classical methods. Traditional phylogenetic inference methods rely on specific evolutionary models, such as the molecular clock hypothesis or constant substitution rate models. These assumptions often do not align with real biological evolution processes. Heterogeneity in evolutionary rates, lineage-specific substitution patterns, and complex evolutionary events (such as horizontal gene transfer and genome duplication) are frequent in biological evolution, and classical methods struggle to capture these complexities (Ronquist et al., 2012). Bayesian methods, for instance, allow for uncertainty through prior settings, but their effectiveness depends heavily on the appropriateness of model selection and priors. If the model choice is incorrect, it can lead to significant biases and erroneous inferences. This is especially problematic for biological groups with complex evolutionary histories, such as polyploidy events in plants or recombination events in pathogens, where inferences based on inaccurate models can distort our understanding of biological history.

Classical methods also have limited capability to handle data complexity and diversity, further underscoring their shortcomings in dealing with the diverse challenges posed by modern biological data. A current trend in biological research is the integration and analysis of multi-omics data, combining genomic, transcriptomic, epigenomic, and metabolomic data to achieve a comprehensive understanding of organismal function and evolutionary history. Most traditional biological tree construction methods are designed for single-type data analysis and lack mechanisms for integrating multiple data sources (Nguyen et al., 2015). When evolutionary signals between different omics data are inconsistent or conflicting, traditional methods often fail to provide a reliable integrated result. This deficiency in integrated analysis may lead to misinterpretations of biological evolutionary processes and hinder a holistic understanding of multi-level biological systems.

These limitations stem from a mismatch between the design assumptions and analytical frameworks of classical biological tree construction methods and the complexity of modern biological data. The consequences are not just inaccuracies in inferring evolutionary relationships but also broader misguidance in biological research, impacting both foundational studies and practical applications. Addressing these challenges requires innovation and optimization across multiple dimensions, including computational methods, model construction, and data integration. Future directions may involve incorporating more flexible and robust statistical models, developing efficient algorithms to reduce computational costs, and exploring the integration of deep learning and other advanced computational methods with traditional phylogenetics to meet the challenges and demands of modern bioinformatics research.

7.2 Challenges of Deep Learning-Based Methods for Tree Construction Methods

Deep learning-based methods have emerged as powerful alternatives for constructing phylogenetic trees due to their ability to learn complex patterns from high-dimensional data. However, several challenges remain in effectively applying these methods.

One significant challenge is the interpretability of deep learning models. Deep learning models, such as deep neural networks, generative adversarial networks, and variational autoencoders, are often considered "black boxes." These models learn highly non-linear and complex feature representations during training, making it difficult to provide intuitive explanations for the results or understand the underlying biological processes. This challenge arises mainly from the high complexity and multi-layered architecture of these models, which capture hidden patterns in data but struggle to clearly link these patterns to biological phenomena (Bojchevski and Günnemann, 2018; Jin et al., 2018b; Kipf and Welling, 2016b; You et al., 2018a). The lack of interpretability can lead to misunderstandings of evolutionary relationships, especially when precise evolutionary pathways between species need to be identified or specific biological mechanisms of evolutionary processes need to be interpreted. Furthermore, this lack of interpretability may undermine the credibility of research findings, particularly when validating results against traditional biological theories.

Data requirements and generalization capabilities pose another significant challenge for deep learning models. These models typically require large amounts of labeled data to achieve high performance. However, biological datasets are often limited, incomplete, or biased, which can lead to overfitting. Overfitting occurs when the model performs well on training data but poorly on unseen or differently characterized datasets (Jin et al., 2018b; Li et al., 2018; You et al., 2018a,b). The consequence of overfitting is limited generalization capability, which affects the reliability of the model in practical applications. In biological applications, this limitation can lead to erroneous phylogenetic tree inferences, potentially impacting the understanding of evolutionary processes and guiding future research.

Integration with biological prior knowledge is another challenge. While deep learning models can capture complex patterns, they often lack effective integration with domain-specific knowledge, such as evolutionary constraints or phylogenetic priors. This absence of prior knowledge integration can result in tree structures generated by the models that lack biological plausibility, affecting their interpretability in biological research (De Cao and Kipf, 2018; Jin et al., 2018b; Manduchi et al., 2023; You et al., 2018a). Deep learning models typically rely on a data-driven learning process and do not inherently incorporate existing theories and knowledge

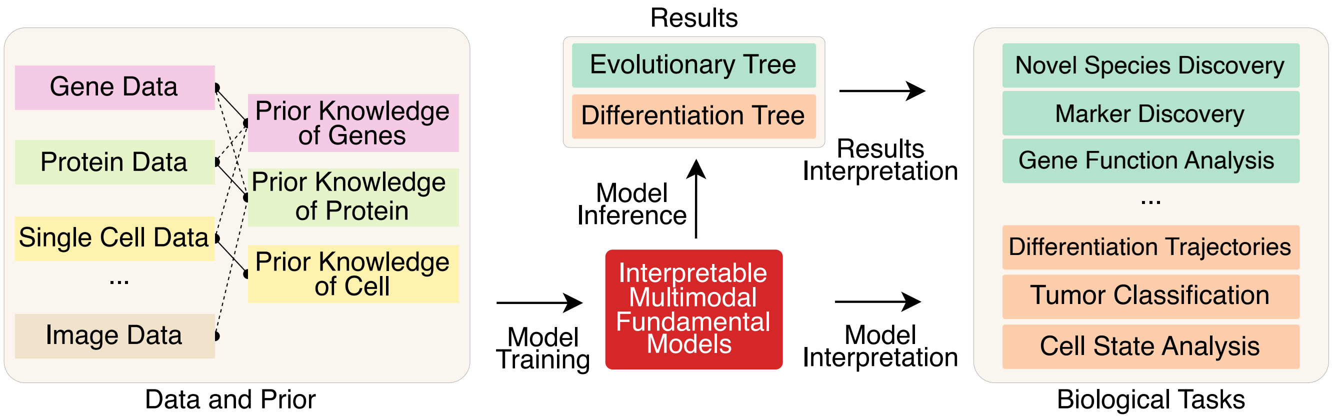

Figure 15: **Integrative Framework for Interpretable Multimodal Deep Learning in Biological Research.** The figure illustrates the integration of multimodal biological data and prior knowledge in deep learning models to enhance model interpretability and transparency. By combining multimodal data and prior knowledge, deep learning models can provide accurate predictions while uncovering biological knowledge through interpretable results.

from evolutionary biology, which can lead to results deviating from biological reality. Moreover, when dealing with complex evolutionary questions, failure to consider biological prior knowledge may result in inferences inconsistent with biological facts, affecting the accuracy and applicability of scientific conclusions.

The computational costs and resource limitations associated with deep learning models present another challenge. Training deep learning models often requires substantial computational resources, including high-performance GPUs and large memory capacity. This computational demand is particularly high when dealing with large-scale biological datasets (Kundu et al., 2019; Li et al., 2018; Wang et al., 2022; You et al., 2018b). Such high computational costs can pose a barrier for many research teams, especially those lacking these resources. Additionally, limitations in computational resources can lead to extended training times, impacting the efficiency and progress of research. In practical applications, these computational constraints can hinder the development and validation of new algorithms and models, thereby slowing down scientific progress.

While deep learning-based methods offer promising new avenues for phylogenetic analysis, these challenges indicate a need for further improvement in interpretability, data integration, model development, and computational efficiency to fully harness their potential in biological research. The development of hybrid approaches that combine the strengths of both classical and deep learning methods could provide a path forward, addressing these challenges while leveraging the advantages of each approach.

8 Opportunities in Tree Construction Methods

8.1 Enhancing Interpretability of Deep Learning Models

While deep learning models excel at learning complex patterns from high-dimensional data, their "black-box" nature limits their acceptance in evolutionary biology research. Therefore, improving the interpretability and transparency of these models is a crucial direction for future research (see Figure 15). By training deep learning models with multimodal biological data (e.g., gene, protein, single-cell, and image data) and their prior knowledge, we can not only improve the prediction accuracy of these models but also perform various downstream tasks (e.g., evolutionary tree and differentiation tree construction, species discovery, gene function analysis) based on the model outputs and their interpretable results. This approach enables us to achieve high-accuracy predictions while uncovering biological knowledge through deep models.

New neural network architectures, such as attention-based models and self-explainable neural networks, provide methods for automatically explaining or visualizing important features, thereby enhancing model interpretability. Techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) can quantify the contribution of each input feature to the final prediction outcome. These methods help uncover the biological patterns learned by deep learning models and verify whether these results are consistent with existing biological knowledge, thus avoiding potential misunderstandings. For example, in applications such as the Junction Tree Variational Autoencoder (JT-VAE) for molecular graph generation, interpretability can provide insights into how the model captures chemical substructures and their contributions to molecular biological functions (Jin et al., 2018b; Lundberg and Lee, 2017; Zang et al., 2024a).

Interpretability techniques can generally be divided into two categories: model-intrinsic interpretability and post-hoc interpretability. Post-hoc interpretability methods mainly target trained deep learning models and provide explanations by analyzing input-output relationships. SHAP and LIME are two common post-hoc interpretation techniques that reveal the decision-making mechanism of the model by assessing the marginal contribution of each input feature to the prediction outcome (Lundberg and Lee, 2017). These methods are particularly suitable for high-dimensional data analysis, such as gene expression data and protein interaction network data. In the task of phylogenetic tree generation, SHAP and LIME can help identify which genes or molecular features have significant contributions to the tree's branching decisions, which is crucial for understanding the prediction principles of deep learning models and for conducting further biological analysis tasks.

However, despite the usefulness of post-hoc interpretability methods like SHAP and LIME in providing model explanations to some extent, they often fall short in terms of stability and effectiveness for practical biological discovery. These methods rely on the relationships between perturbations in input data and model outputs, making their results highly sensitive to data distribution and model changes. Consequently, post-hoc interpretability methods may exhibit inconsistencies across different datasets or model architectures, limiting their application in complex biological problems. Therefore, to better meet the needs of biological discovery, it is essential to design more interpretable and robust deep learning models that can provide stable and reliable interpretative results while handling high-dimensional and diverse biological data.

To further enhance the interpretability of deep learning models, integrating biological prior knowledge into the model architecture design and training processes could be considered. For example, introducing domain-specific evolutionary constraints or priors in biological tree construction, combined with a hierarchical interpretation framework, can provide a clearer explanation path for complex biological evolutionary processes. This combination can significantly improve the credibility and application value of deep learning models. Thus, by adopting diverse interpretability techniques and leveraging the outputs of deep models for various downstream biological tasks (see Figure 15), future deep learning models can provide strong biological explanations while improving prediction accuracy, thereby promoting their widespread application in bioinformatics, phylogenetics, and other related fields.

8.2 Integration of Multi-Omics and Multi-Modal Data

Currently, a significant number of studies focus primarily on single data types. In the construction of evolutionary trees, research is often centered on gene sequences, protein sequences, and some morphological image data; in differentiation tree construction, the focus is mainly on single-cell transcriptome RNA sequencing data. However, relying solely on single-modal data may not fully capture the complexity and diversity of

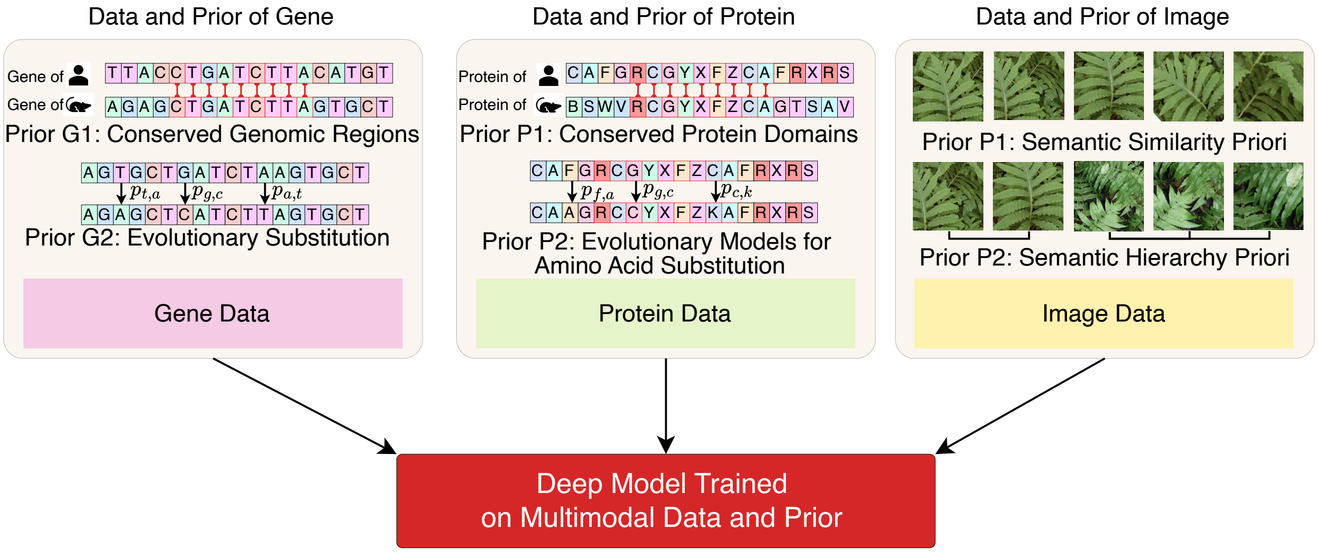

Figure 16: **The futurework for multimodal deep learning in biological research.** The figure illustrates the integration of multi-modal data in deep learning models for biological research, combining genomic, proteomic, transcriptomic, metabolomic, and epigenetic data to enhance model performance and uncover comprehensive biological information.

biological systems. Biological processes are inherently multi-layered and multi-scaled, involving interactions across genes, proteins, metabolites, cells, and tissues. Utilizing multiple modalities of data simultaneously to train larger and more complex deep learning models may lead to significant performance improvements (Rao et al., 2021).

Different modalities of data can introduce prior knowledge from various dimensions. For example, *genomic data* provide genetic information, including mutations, copy number variations, and structural variations; *proteomic data* reflect protein expression levels, post-translational modifications, and interaction networks; *transcriptomic data* reveal gene expression levels and regulatory relationships; *metabolomic data* indicate the state of cellular metabolic activities; and *epigenetic data* (such as DNA methylation and histone modifications) offer additional layers of gene regulation. By integrating multiple modalities of data, we can learn more robust and comprehensive biological information, uncover associations between different biological layers, and enhance the generalization ability and predictive accuracy of models (Hasin et al., 2017). This is crucial for constructing more accurate and reliable biological evolutionary and differentiation trees, providing a more holistic reflection of the evolution and development processes of organisms.

In the current mainstream deep learning field, the development of multi-modal models is advancing rapidly, offering abundant techniques and methods that can be leveraged and transferred. For instance, the recently proposed *BLIP* (Bootstrapping Language-Image Pre-training) model (Li et al., 2023, 2022), minigpt-4 (Zhu et al., 2023a), and Hugginggpt (Shen et al., 2024) has successfully fused vision and language modalities through a unified pre-training framework, achieving remarkable results in tasks such as image captioning and visual question answering. BLIP employs contrastive learning and generative modeling to effectively learn cross-modal feature representations, capturing deep associations between different modalities. Additionally, the *CLIP* (Contrastive Language-Image Pre-training) model (Radford et al., 2021) has also demonstrated outstanding performance in multi-modal tasks by pre-training on large-scale image-text pair data. The success of these models provides new ideas for integrating multi-modal data in bioinformatics.

Simultaneously, *Graph Neural Networks* (GNNs) have been utilized to integrate multi-omics data, capturing complex relationships within biological networks (Kipf and Welling, 2016a; Veličković et al., 2017). For example, DeepOmics employs graph convolutional networks to combine multi-omics data for cancer type prediction, achieving better performance than single-modal methods (Huang et al., 2020). Moreover, generative models like *Variational Autoencoders* (VAEs) and *Generative Adversarial Networks* (GANs) have been applied to multi-modal data fusion and representation learning (Goodfellow et al., 2014; Kingma and Welling, 2013). Applying these advanced multi-modal deep learning techniques to the field of bioinformatics can effectively integrate multi-omics data, uncover associations and mechanisms across different biological layers.

However, integrating multi-modal data also presents numerous challenges. Different modalities may have varying scales, noise levels, and degrees of data missingness, requiring robust algorithms to handle these issues. Additionally, heterogeneity between multi-modal data makes effective alignment and integration a key problem (Wang et al., 2020). Researchers have proposed various strategies to address these challenges. For instance, using *aligned embeddings* to map data from different modalities into a common feature space (Zhang et al., 2018); adopting *cross-modal attention mechanisms* to dynamically weigh and fuse information from each modality (Tsai et al., 2019); and incorporating biological prior knowledge, such as gene pathways and protein-protein interaction networks, to guide model construction and training (Meng et al., 2019).

Through these efforts, deep learning models can play a more significant role in bioinformatics and systems biology, promoting advancements in related research. For example, in disease studies, multi-modal data integration can help identify key biomarkers, facilitating the development of personalized medicine (Picard et al., 2019); in evolutionary biology, integrating genomic, phenotypic, and ecological data can lead to a more comprehensive understanding of species' evolutionary mechanisms (Zhang and Yang, 2020).

8.3 Leveraging Cellular and Species-Level Information for Downstream Tasks

In biological research, evolutionary trees (phylogenetic trees) and differentiation trees (developmental pathways) are fundamental tools for understanding the evolutionary relationships among species and the developmental differentiation pathways of cells. However, traditional studies often consider these two aspects separately, focusing either on macro-level species evolution or micro-level cell differentiation. Combining evolutionary trees with differentiation trees and leveraging information from both cellular and species levels provide new opportunities for accomplishing downstream tasks (see Figure **??**).

Integrating information from evolutionary and differentiation trees allows for a more comprehensive understanding of the complexity of biological processes. For example, in *species discovery*, combining genetic information with cellular differentiation patterns helps identify new species or subspecies and gain deeper insights into their evolutionary pathways (Bock, 1959; Spain et al., 2023). In *gene function analysis*, correlating the evolutionary conservation of genes with their roles in cell differentiation can reveal the functions of key genes and their regulatory networks.

The mutual validation between evolutionary and differentiation trees helps improve the reliability and accuracy of models. For instance, discrepancies between the two may indicate that existing models need refinement or suggest the existence of new biological phenomena. In *disease progression modeling*, understanding the abnormal evolution and differentiation processes of cancer cells can aid in discovering new diagnostic biomarkers and therapeutic targets, optimizing disease diagnosis and treatment strategies (Zhao et al., 2023).

Furthermore, integrating these two types of information can facilitate advanced tasks such as *differentiation trajectory mapping* and *tumor classification*. By analyzing how evolutionary mutations affect cellular differentiation pathways, we can improve tumor classification and predict progression trends (Cowan et al., 2023). In *personalized medicine*, treatment plans can be more precisely formulated based on patients' genetic backgrounds and cellular differentiation characteristics, enhancing therapeutic efficacy.

To achieve this integration, deep learning models need to handle multi-modal data and capture complex hierarchical relationships. Techniques such as *hierarchical attention networks* and *multi-task learning* can be employed to model interactions between evolutionary and differentiation data (Yariv et al., 2023). Additionally, incorporating *biological prior knowledge* into model architectures and training processes can enhance interpretability and ensure alignment with biological principles.

However, integrating multi-level data also presents several challenges. Different data modalities may have scale differences, varying noise levels, and missing values, requiring robust preprocessing and normalization methods (Chai, 2023). Advanced data integration and alignment algorithms, such as *canonical correlation analysis* and *manifold alignment*, can be used to reconcile differences between datasets and achieve effective integration.

Through the mutual validation of evolutionary and differentiation trees, future deep learning models are expected to achieve higher predictive accuracy and provide deeper biological insights. This comprehensive approach will promote advancements in fields such as phylogenetics, developmental biology, and medical diagnostics, offering a new perspective for deeply understanding the complexity of life.

9 Data Availability

No data was generated or used in this review.

10 Code Availability

No code was generated or used in this review.

11 Acknowledgements

This work was supported by the National Key R&D Program of China (No.2022ZD0115100), the National Natural Science Foundation of China Project (No. U21A20427), and Project (No. WU2022A009) from the Center of Synthetic Biology and Integrated Bioengineering of Westlake University. We thank the Westlake University HPC Center for providing computational resources.

12 Author Contributions

Stan Z. Li and Zelin Zang proposed this research. Zelin Zang, Yongjie Xu, and Chenrui Duan collected the information. Zelin Zang, Yongjie Xu, and Chenrui Duan wrote the manuscript. Jinlin Wu, Stan Z. Li, and Zhen Lei provided valuable suggestions on the manuscript. All authors discussed the results, revised the draft manuscript, and read and approved the final manuscript.

13 Competing Interests

The authors declare no competing interests.

References

Abeer, A. N. M. N., Jantre, S., Urban, N. M., and Yoon, B.-J. (2024). Leveraging Active Subspaces to Capture Epistemic Model Uncertainty in Deep Generative Models for Molecular Design. arXiv:2405.00202 [cs, q-bio, stat].

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002a). *Molecular Biology of the Cell*. Garland Science, 4th edition.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002b). *Molecular Biology of the Cell*. Garland Science.

Almendro, V., Marusyk, A., and Polyak, K. (2013). Cellular heterogeneity and molecular evolution in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 8:277–302.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Ansorge, W. J. (2009). Next-generation dna sequencing techniques. *New biotechnology*, 25(4):195–203.

Barnett, R. and Larson, G. (2012). A phenol–chloroform protocol for extracting dna from ancient samples. *Ancient DNA: Methods and Protocols*, pages 13–19.

Bartlett, G. J., Porter, C. T., Borkakoti, N., and Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, 324(1):105–121.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., et al. (2002). The pfam protein families database. *Nucleic Acids Research*, 30(1):276–280.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*. W H Freeman.

Bergen, V., Lange, M., Peidli, S., et al. (2020a). Generalizing rna velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020b). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*, 38(12):1408–1414. Publisher: Nature Publishing Group.

Berman, H. M. et al. (2000a). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000b). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Bernot, J. P., Owen, C. L., Wolfe, J. M., Meland, K., Olesen, J., and Crandall, K. A. (2023). Major Revisions in Pancrustacean Phylogeny and Evidence of Sensitivity to Taxon Sampling. *Molecular Biology and Evolution*, 40(8):msad175.

Blaimer, B. B., Santos, B. F., Cruaud, A., Gates, M. W., Kula, R. R., Mikó, I., Rasplus, J.-Y., Smith, D. R., Talamas, E. J., Brady, S. G., and Buffington, M. L. (2023). Key innovations and the diversification of Hymenoptera. *Nature Communications*, 14(1):1212. Publisher: Nature Publishing Group.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blum, M. G. and Francois, O. (2006). Random processes of tree growth and statistical tests of tree imbalance. *Evolution*, 60(6):1138–1150.

Bock, W. J. (1959). Preadaptation and multiple evolutionary pathways. *Evolution*, pages 194–211.

Bojchevski, A. and Günnemann, S. (2018). Netgan: Generating graphs via random walks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Bou Dagher, L., Madern, D., Malbos, P., and Brochier-Armanet, C. (2024). Persistent homology reveals strong phylogenetic signal in 3D protein structures. *PNAS nexus*, 3(4):pgae158. Publisher: Oxford University Press US.

Boukouvalas, A., Hensman, J., and Rattray, M. (2018). BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biology*, 19(1):65.

Bourret, T. B., Fajardo, S. N., Frankel, S. J., and Rizzo, D. M. (2023). Cataloging Phytophthora Species of Agriculture, Forests, Horticulture, and Restoration Outplantings in California, U.S.A.: A Sequence-Based Meta-Analysis. *Plant Disease*. Publisher: The American Phytopathological Society TLDR: A meta-analysis of Phytophthora detections within the state was conducted using publicly available sequences as a primary source of data rather than published records to better understand threats to California plant health.

Brylinski, M. (2014). eMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models. *PLoS Computational Biology*, 10(9):e1003829. TLDR: eMatchSite is a new method for constructing sequence order-independent alignments of ligand binding sites in protein models that opens up the possibility to investigate drug-protein interaction networks for complete proteomes with prospective systems-level applications in polypharmacology and rational drug repositioning.

Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., Aryee, M. J., Majeti, R., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Campbell, K. R. and Yau, C. (2019). A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics*, 35(1):28–35.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., and Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.

Cavalli-Sforza, L. and Edwards, A. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570.

Cech, T. R. and Steitz, J. A. (2014). The noncoding rna revolution—trashing old rules to forge new ones. *Cell*, 157(1):77–94.

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553.

Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C., and Murphy, K. (2022). Machine learning on graphs: A model and comprehensive taxonomy. *Journal of Machine Learning Research*, 23(89):1–64.

Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al. (2023a). Tbtools-ii: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Molecular plant*, 16(11):1733–1742.

Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., Chen, W., Liao, Y., Zhang, W., Gao, Y., Dong, K., Ren, R., Yang, T., Zhang, L., Qi, M., Li, Z., Zhao, M., Wang, H., Wang, J., Qiao, Z., Li, H., Jiang, Y., Liu, G., Song, X., Deng, Y., Li, H., Yan, F., Dong, Y., Li, Q., Li, T., Yang, W., Cui, J., Wang, H., Zhou, Y., Zhang, X., Jia, G., Lu, P., Zhi, H., Tang, S., and Diao, X. (2023b). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nature Genetics*, 55(12):2243–2254. Publisher: Nature Publishing Group.

Chen, N., Zhang, Z., Hou, J., Chen, J., Gao, X., Tang, L., Wangdue, S., Zhang, X., Sinding, M.-H. S., Liu, X., Han, J., Lü, H., Lei, C., Marshall, F., and Liu, X. (2023c). Evidence for early domestic yak, taurine cattle, and their hybrids on the Tibetan Plateau. *Science Advances*, 9(50):eadi6857. Publisher: American Association for the Advancement of Science.

Chen, Z., King, W. C., Hwang, A., Gerstein, M., and Zhang, J. (2022). DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Science Advances*, 8(48):eabq3745. Publisher: American Association for the Advancement of Science.

Cheng, Y. (2015). Single-particle cryo-em at crystallographic resolution. *Cell*, 161(3):450–457.

Choi, J. and Kim, S.-H. (2020). Whole-proteome tree of life suggests a deep burst of organism diversity. *Proceedings of the National Academy of Sciences*, 117(7):3678–3686. Publisher: Proceedings of the National Academy of Sciences TLDR: The main features of a whole-proteome ToL for 4,023 species with known complete or almost complete genome sequences on grouping and kinship among the groups at deep evolutionary levels are described.

Chothia, C. and Finkelstein, A. V. (1984). Principles that determine the structure of proteins. *Annual Review of Biochemistry*, 53(1):537–572.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771.

Consortium, . G. P. (2015). A global reference for human genetic variation. *Nature*, 526:68–74.

Consortium, G. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585.

Consortium, H. M. P. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207–214.

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Consortium, T. U. (2021). Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489.

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515.

Cowan, A., Ferrari, F., Freeman, S. S., Redd, R., El-Khoury, H., Perry, J., Patel, V., Kaur, P., Barr, H., Lee, D. J., et al. (2023). Personalised progression prediction in patients with monoclonal gammopathy of undetermined significance or smouldering multiple myeloma (pangea): a retrospective, multicohort study. *The Lancet Haematology*, 10(3):e203–e212.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). *Atlas of protein sequence and structure*. National Biomedical Research Foundation.

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, 22(10):1269–1271. Publisher: Oxford University Press.

De Cao, N. and Kipf, T. (2018). Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*.

Delsuc, F., Philippe, H., and Douzery, E. J. (2019). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 20(1):1–12.

Desiere, F. et al. (2006). The peptideatlas project. *Nucleic Acids Research*, 34(suppl_1):D655–D658.

Desper, R. and Gascuel, O. (2004). The balanced minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 21(3):587–598.

Ding, J. and Regev, A. (2021). Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nature communications*, 12(1):2554.

Do, C. B., Mahabhashyam, M. S., Brudno, M., and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2):330–340. Publisher: Cold Spring Harbor Lab.

Domcke, S. and Shendure, J. (2023). A reference cell tree will serve science better than a reference cell atlas. *Cell*, 186(6):1103–1114.

Dong, R., Peng, Z., Zhang, Y., and Yang, J. (2018). mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 34(10):1719–1725. TLDR: The proposed multiple structure alignment algorithm (mTM-align) was proposed, which is an extension of the highly efficient pairwise structure alignment program TM-align, and benchmarked on four widely used datasets, showing that mTM- align consistently outperforms other algorithms.

Doolittle, R. F. (1981). Protein evolution. *Science*, 214(4517):149–159.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with beauti and the beast 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

Du, J.-H., Chen, T., Gao, M., and Wang, J. (2024). Joint trajectory inference for single-cell genomics using deep learning with a mixture prior. *Proceedings of the National Academy of Sciences*, 121(37):e2316256121. Publisher: Proceedings of the National Academy of Sciences.

duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H., and Tsuda, K. (2016). CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics*, 17(1):363.

Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J., and Dessimoz, C. (2024a). Inference of phylogenetic trees directly from raw sequencing reads using read2tree. *Nature Biotechnology*, 42(1):139–147.

Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. J., and Dessimoz, C. (2024b). Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nature Biotechnology*, 42(1):139–147. Publisher: Nature Publishing Group.

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., Santamarina, M., Höps, W., Ashraf, H., Chuang, N. T., Yang, X., Munson, K. M., Lewis, A. P., Fairley, S., Tallon, L. J., Clarke, W. E., Basile, A. O., Byrska-Bishop, M., Corvelo, A., Evani, U. S., Lu, T.-Y., Chaisson, M. J. P., Chen, J., Li, C., Brand, H., Wenger, A. M., Ghareghani, M., Harvey, W. T., Raeder, B., Hasenfeld, P., Regier, A. A., Abel, H. J., Hall, I. M., Flicek, P., Stegle, O., Gerstein, M. B., Tubio, J. M. C., Mu, Z., Li, Y. I., Shi, X., Hastie, A. R., Ye, K., Chong, Z., Sanders, A. D., Zody, M. C., Talkowski, M. E., Mills, R. E., Devine, S. E., Lee, C., Korbel, J. O., Marschall, T., and Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117. Publisher: American Association for the Advancement of Science.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Edogbanya, J., Tejada-Martinez, D., Jones, N. J., Jaiswal, A., Bell, S., Cordeiro, R., van Dam, S., Rigden, D. J., and de Magalhães, J. P. (2021). Evolution, structure and emerging roles of C1ORF112 in DNA replication, DNA damage responses, and cancer. *Cellular and Molecular Life Sciences*, 78(9):4365–4376. TLDR: Gene expression data show that, among human tissues, C1ORF112 is highly expressed in the testes and overexpressed in various cancers when compared to healthy tissues, and protein models suggest that C1ORN112 is an alpha-helical protein.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., Stott, M. B., Schramm, A., Kjeldsen, K. U., Teske, A. P., Baker, B. J., and Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, 618(7967):992–999. Publisher: Nature Publishing Group.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). *Phylogenies and the Comparative Method*, volume 125. University of Chicago Press.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates.

Feng, X., Zheng, J., Irisarri, I., Yu, H., Zheng, B., Ali, Z., de Vries, S., Keller, J., Fürst-Jansen, J. M. R., Dadras, A., Zegers, J. M. S., Rieseberg, T. P., Dhabalia Ashok, A., Darienko, T., Bierenbroodspot, M. J., Gramzow, L., Petroll, R., Haas, F. B., Fernandez-Pozo, N., Nousias, O., Li, T., Fitzek, E., Grayburn, W. S., Rittmeier, N., Permann, C., Rümpler, F., Archibald, J. M., Theißen, G., Mower, J. P., Lorenz, M., Buschmann, H., von Schwartzenberg, K., Boston, L., Hayes, R. D., Daum, C., Barry, K., Grigoriev, I. V., Wang, X., Li, F.-W., Rensing, S. A., Ben Ari, J., Keren, N., Mosquna, A., Holzinger, A., Delaux, P.-M., Zhang, C., Huang, J., Mutwil, M., de Vries, J., and Yin, Y. (2024). Genomes of multicellular algal sisters to land plants illuminate signaling network evolution. *Nature Genetics*, 56(5):1018–1031. Publisher: Nature Publishing Group.

Finn, R. D., Coggill, P., Eberhardt, R. Y., et al. (2016). The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

Fisk, J. N., Mahal, A. R., Dornburg, A., Gaffney, S. G., Aneja, S., Contessa, J. N., Rimm, D., Yu, J. B., and Townsend, J. P. (2022). Premetastatic shifts of endogenous and exogenous mutational processes support consolidative therapy in EGFR-driven lung adenocarcinoma. *Cancer Letters*, 526:346–351. TLDR: Mutational signature analyses within clinically annotated cancer chronograms are applied to detect and describe the shifting mutational processes caused by both endogenous and exogenous factors between tumor sampling timepoints to inform therapeutic decision making and retrospective assessment of disease etiology.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution*, 35(10):2585–2593.

Forrow, A. and Schiebinger, G. (2021). Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature communications*, 12(1):4940.

Gao, M. and Skolnick, J. (2013). APoc: large-scale identification of similar protein pockets. *Bioinformatics*, 29(5):597–604. TLDR: This work introduces a computational method, APoc (Alignment of Pockets), for the large-scale, sequence order-independent, structural comparison of protein pockets, and demonstrates that APoc has better performance than the geometric hashing-based method SiteEngine.

Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., Theis, F. J., and Yosef, N. (2024). Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nat Methods*, 21(1):50–59. Publisher: Nature Publishing Group.

Ghaly, T. M., Tetu, S. G., Penesyan, A., Qi, Q., Rajabal, V., and Gillings, M. R. (2022). Discovery of integrons in Archaea: Platforms for cross-domain gene transfer. *Science Advances*, 8(46):eabq6376. Publisher: American Association for the Advancement of Science.

Gharaee, Z., Gong, Z., Pellegrino, N., Zarubiieva, I., Haurum, J. B., Lowe, S., McKeown, J., Ho, C., McLeod, J., and et al., Y.-Y. W. (2024). A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *Advances in Neural Information Processing Systems*, 36. Accessed: 2024-09-17.

Gonzalez-Reiche, A. S., Alshammary, H., Schaefer, S., Patel, G., Polanco, J., Carreño, J. M., Amoako, A. A., Rooker, A., Cognigni, C., Floda, D., van de Guchte, A., Khalil, Z., Farrugia, K., Assad, N., Zhang, J., Alburquerque, B., Sominsky, L. A., Gleason, C., Srivastava, K., Sebra, R., Ramirez, J. D., Banu, R., Shrestha, P., Krammer, F., Paniz-Mondolfi, A., Sordillo, E. M., Simon, V., and van Bakel, H. (2023). Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nature Communications*, 14(1):3235. Publisher: Nature Publishing Group.

Gonçalves, J. d. S., Manduchi, L., Vandenhirtz, M., and Vogt, J. E. (2024). Structured Generations: Using Hierarchical Clusters to guide Diffusion Models. In *ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*.

Goodfellow, I. et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680.

Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., Ziebuhr, J., and Coronaviridae Study Group of the International Committee on

Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, 5(4):536–544. Publisher: Nature Publishing Group.

Graham, C. H., Ron, S. R., Santos, J. A., Schneider, C. J., and Moritz, C. (2004). Habitat history improves prediction of biodiversity in rain forest fauna. *Proceedings of the National Academy of Sciences*, 101(3):543–548.

Grigoriadis, K., Huebner, A., Bunkum, A., Colliver, E., Frankell, A. M., Hill, M. S., Thol, K., Birkbak, N. J., Swanton, C., Zaccaria, S., et al. (2024). Conipher: a computational framework for scalable phylogenetic reconstruction with error correction. *Nature Protocols*, 19(1):159–183.

Gu, X., Zhang, Z., and Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences*, 102(3):707–712.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

Guo, C., Luo, Y., Gao, L.-M., Yi, T.-S., Li, H.-T., Yang, J.-B., and Li, D.-Z. (2023). Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology*, 65(2):299–323. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jipb.13415.

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317.

Haghverdi, L., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848.

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100(5):605–617.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, 18(1):83.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y., Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. (2024). Simulating 500 million years of evolution with a language model. Pages: 2024.07.01.600583 Section: New Results TLDR: This work presents ESM3, a frontier multimodal generative language model that reasons over the sequence, structure, and function of proteins, and prompts ESM3 to generate fluorescent proteins with a chain of thought.

Hennig, W. (1965). Phylogenetic systematics. *Annual Review of Entomology*, 10(1):97–116.

Hennig, W. (1966). *Phylogenetic Systematics*. University of Illinois Press.

Hillis, D. M. (2019). The tree of life: Resolving the relationships of the majority of living species. *Systematic Biology*, 68(5):896–900.

Hillis, D. M. and Huelsenbeck, J. P. (1992). Phylogeny and the evolution of hiv. *Science*, 257(5079):1159–1163.

Hohna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138. Publisher: Elsevier.

Holm, L. and Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20(11):478–480.

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811.

Huang, S., Chaudhary, K., and Garmire, L. X. (2020). Fusion of multi-omics data and deep learning for cancer patient survivability prediction. *Methods*, 166:28–37.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Huguet, G., Magruder, D. S., Tong, A., Fasina, O., Kuchroo, M., Wolf, G., and Krishnaswamy, S. (2022). Manifold Interpolating Optimal-Transport Flows for Trajectory Inference. *Advances in Neural Information Processing Systems*, 35:29705–29718.

iNaturalist (2021). inaturalist 2021 dataset (inat21). Accessed: 2024-09-17.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, 9(6):e98679. Publisher: Public Library of Science.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323. Overview of clustering algorithms and their applications in various fields, including biological data classification.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2016). The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):1–11.

Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117.

Jia, X., Wang, L., Zhao, H., Zhang, Y., Chen, Z., Xu, L., and Yi, K. (2023). The origin and evolution of salicylic acid signaling and biosynthesis in plants. *Molecular Plant*, 16(1):245–259. Publisher: Elsevier TLDR: 10 core protein families in SA signaling and biosynthesis across green plant lineages are identified and it is revealed that the ancient abnormal inflorescence meristem 1 (AIM1)-based beta-oxidation pathway is crucial for the biosynthesis of SA in chlorophyte algae, and this biosynthesis pathway may have facilitated the adaptation of early-diverging green algae to the high-light-intensity environment on land.

Jiang, Y., Tabaghi, P., and Mirarab, S. (2022). Learning Hyperbolic Embedding for Phylogenetic Tree Placement and Updates. *Biology*, 11(9):1256. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute TLDR: It is shown how the conventional (Euclidean) deep learning methods developed for phylogenetics can benefit from using hyperbolic geometry, and the appropriate geometry for faithfully representing tree distances while embedding gene sequences is examined.

Jin, W., Barzilay, R., and Jaakkola, T. (2018a). Junction Tree Variational Autoencoder for Molecular Graph Generation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2323–2332. PMLR. ISSN: 2640-3498.

Jin, W., Barzilay, R., and Jaakkola, T. (2018b). Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning (ICML)*.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8(3):275–282.

Jonsson, G., Hofmann, M., Mereiter, S., Hartley-Tassell, L., Sakic, I., Oliveira, T., Hoffmann, D., Novatchkova, M., Schleiffer, A., and Penninger, J. M. (2024). CLEC18A interacts with sulfated GAGs and controls clear cell renal cell carcinoma progression. Pages: 2024.07.08.602586 Section: New Results TLDR: A key role is reported of the CLEC18 family of C-type lectins in the progression of clear cell renal cell carcinoma (ccRCC) and the potential benefit of modulating CLEC18 expression in the renal tumor microenvironment is highlighted.

Jukes, T. H. and Cantor, C. R. (1969). Evolutionary clocks and their settings. *Mammalian Protein Metabolism*, 3:21–132.

Julca, I., Mutwil-Anderwald, D., Manoj, V., Khan, Z., Lai, S. K., Yang, L. K., Beh, I. T., Dziekan, J., Lim, Y. P., Lim, S. K., Low, Y. W., Lam, Y. I., Tjia, S., Mu, Y., Tan, Q. W., Nuc, P., Choo, L. M., Khew, G., Shining, L., Kam, A., Tam, J. P., Bozdech, Z., Schmidt, M., Usadel, B., Kanagasundaram, Y., Alseekh, S., Fernie, A., Li, H. Y., and Mutwil, M. (2023). Genomic, transcriptomic, and metabolomic analysis of Oldenlandia corymbosa reveals the biosynthesis and mode of action of anti-cancer metabolites. *Journal of Integrative Plant Biology*, 65(6):1442–1466. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jipb.13469 TLDR: It is revealed that ursolic acid causes mitotic catastrophe in cancer cells and three high-confidence protein binding targets by Cellular Thermal Shift Assay (CETSA) and reverse docking will allow us to further develop this valuable compound.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kale, S. M., Schulthess, A. W., Padmarasu, S., Boeven, P. H. G., Schacht, J., Himmelbach, A., Steuernagel, B., Wulff, B. B. H., Reif, J. C., Stein, N., and Mascher, M. (2022). A catalogue of resistance gene homologs and a chromosome-scale reference sequence support resistance gene mapping in winter wheat. *Plant Biotechnology Journal*, 20(9):1730–1742. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.13843.

Kan, S., Liao, X., Lan, L., Kong, J., Wang, J., Nie, L., Zou, J., An, H., and Wu, Z. (2024). Cytonuclear Interactions and Subgenome Dominance Shape the Evolution of Organelle-Targeted Genes in the Brassica Triangle of U. *Molecular Biology and Evolution*, 41(3):msae043. TLDR: This study investigates the evolutionary pattern of organelle-targeted genes in Brassica carinata and 2 varieties of Brassica juncea at the whole-genome level, with particular focus on cytonuclear enzyme complexes and highlights an important role for subgenome dominance in allopolyploid genome evolution, even in genes whose function depends on separately inherited molecules.

Kang, S.-H., Pandey, R. P., Lee, C.-M., Sim, J.-S., Jeong, J.-T., Choi, B.-S., Jung, M., Ginzburg, D., Zhao, K., Won, S. Y., Oh, T.-J., Yu, Y., Kim, N.-H., Lee, O. R., Lee, T.-H., Bashyal, P., Kim, T.-S., Lee, W.-H., Hawkins, C., Kim, C.-K., Kim, J. S., Ahn, B. O., Rhee, S. Y., and Sohng, J. K. (2020). Genome-enabled discovery of anthraquinone biosynthesis in Senna tora. *Nature Communications*, 11(1):5875. Publisher: Nature Publishing Group.

Karczewski, K. J. et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443.

Katoh, K. and Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, 32(13):1933–1942. TLDR: A new feature of the MAFFT multiple alignment program for suppressing over-alignment (aligning unrelated segments) by utilizing a variable scoring matrix for different pairs of sequences (or groups) in a single multiple sequence alignment, based on the global similarity of each pair.

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., et al. (2014). Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138.

Kersey, P. J. et al. (2018). Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, 46(D1):D802–D808.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):2966.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular Cell*, 58(4):610–620.

Kolora, S. R. R., Owens, G. L., Vazquez, J. M., Stubbs, A., Chatla, K., Jainese, C., Seeto, K., McCrea, M., Sandel, M. W., Vianna, J. A., Maslenikov, K., Bachtrog, D., Orr, J. W., Love, M., and Sudmant, P. H. (2021). Origins and evolution of extreme life span in Pacific Ocean rockfishes. *Science*, 374(6569):842–847. Publisher: American Association for the Advancement of Science TLDR: Genomes generated from rockfish species of different life spans elucidates the genetic determinants of aging and highlights the genetic innovations that underlie life history trait adaptations and, in turn, how they shape genomic diversity.

Koptagel, H., Kviman, O., Melin, H., Safinianaini, N., and Lagergren, J. (2022). VaiPhy: a Variational Inference Based Algorithm for Phylogeny. In *Advances in Neural Information Processing Systems*.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kundu, J. N., Gor, M., Agrawal, D., and Babu, R. V. (2019). Gan-tree: An incrementally learned hierarchical generative framework for multi-modal data distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8191–8200.

Kwon, Y. M., Gori, K., Park, N., Potts, N., Swift, K., Wang, J., Stammnitz, M. R., Cannell, N., Baez-Ortega, A., Comte, S., Fox, S., Harmsen, C., Huxtable, S., Jones, M., Kreiss, A., Lawrence, C., Lazenby, B., Peck, S., Pye, R., Woods, G., Zimmermann, M., Wedge, D. C., Pemberton, D., Stratton, M. R., Hamede, R., and Murchison, E. P. (2020). Evolution and lineage dynamics of a transmissible cancer in Tasmanian devils. *PLOS Biology*, 18(11):e3000926. Publisher: Public Library of Science TLDR: Overall, DFT1 is a remarkably stable lineage whose genome illustrates how cancer cells adapt to diverse environments and persist in a parasitic niche.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., et al. (2018a). Rna velocity of single cells. *Nature*, 560(7719):494–498.

La Manno, G., Soldatov, R., Zeisel, A., et al. (2018b). Rna velocity of single cells. *Nature*, 560(7719):494–498.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. (2022). CellRank for directed single-cell fate mapping. *Nat Methods*, 19(2):159–170. Publisher: Nature Publishing Group.

Lax, G. and Keeling, P. J. (2023). Molecular phylogenetics of sessile Dolium sedentarium, a petalomonad euglenid. *Journal of Eukaryotic Microbiology*, 70(5):e12991. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jeu.12991.

LeCun, Y., Bengio, Y., and Hinton, G. (2015a). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Bengio, Y., and Hinton, G. (2015b). Deep learning. *Nature*, 521(7553):436–444.

Leguia, M., Garcia-Glaessner, A., Muñoz-Saavedra, B., Juarez, D., Barrera, P., Calvo-Mac, C., Jara, J., Silva, W., Ploog, K., Amaro, L., Colchao-Claux, P., Johnson, C. K., Uhart, M. M., Nelson, M. I., and Lescano, J. (2023). Highly pathogenic avian influenza A (H5N1) in marine mammals and seabirds in Peru. *Nature Communications*, 14(1):5489. Publisher: Nature Publishing Group.

Lei, H., Mi, L., Zhou, X., Chen, J., Hu, J., Guo, S., and Zhang, Y. (2011). Adsorption of double-stranded dna to graphene oxide preventing enzymatic digestion. *Nanoscale*, 3(9):3888–3892.

Lemey, P., Salemi, M., and Vandamme, A.-M. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Li, H. (2017). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Li, J., Pan, X., Yuan, Y., and Shen, H.-B. (2024). TFvelo: gene regulation inspired RNA velocity estimation. *Nat Commun*, 15(1):1387. Publisher: Nature Publishing Group.

Li, Q. (2023). scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics. *Genome Biology*, 24(1):149.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. (2018). Learning deep generative models of graphs. In *International Conference on Machine Learning (ICML)*.

Liang, S., Wang, F., Han, J., and Chen, K. (2020). Latent periodic process inference from single-cell RNA-seq data. *Nat Commun*, 11(1):1441. Publisher: Nature Publishing Group.

Lin, C. and Bar-Joseph, Z. (2019). Continuous-state HMMs for modeling time-series single-cell RNA-Seq data. *Bioinformatics*, 35(22):4707–4715.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Macaulay, I. C. and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends in Genetics*, 33(2):155–168.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Maddison, D. R. and Schulz, K.-S. (2018). The tree of life. *Systematic Biology*, 67(5):719–729.

Maddison, W. P. and Maddison, D. R. (2007). Mesquite: a modular system for evolutionary analysis. *Evolutionary Bioinformatics*, 3:47–50.

Maizels, R. J., Snell, D. M., and Briscoe, J. (2023). Deep dynamical modelling of developmental trajectories with temporal transcriptomics. Pages: 2023.07.06.547989 Section: New Results.

Man, J., Gallagher, J. P., and Bartlett, M. (2020). Structural evolution drives diversification of the large LRR-RLK gene family. *New Phytologist*, 226(5):1492–1505. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.16455.

Manduchi, L., Vandenhirtz, M., Ryser, A., and Vogt, J. E. (2023). Tree Variational Autoencoders. In *Advances in Neural Information Processing Systems*.

Marchler-Bauer, A., Lu, S., Anderson, J. B., et al. (2011). Cdd: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39(suppl_1):D225–D229.

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. (2014). Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A*, 111(52):E5643–5650.

Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387–402.

Margelevičius, M. (2024). GTalign: spatial index-driven protein structure alignment, superposition, and search. *Nature Communications*, 15(1):7305. Publisher: Nature Publishing Group.

Marks, D. S., Hopf, T. A., and Sander, C. (2011). Protein 3d structure computed from evolutionary sequence variation. *PloS One*, 6(12):e28766.

Marletaz, F., Calle, E., Acemel, R., Paliou, C., Naranjo, S., Martinez, P., Cases, I., Sleight, V., Hirschberger, C., Marcet, M., Navon, D., Andrescavage, A., Skvortsova, K., Duckett, P., Gonzalez, A., Bogdanovic, O., Gibcus, J., Yang, L., Gallardo, L., and Gomez, J. (2023). The little skate genome and the evolutionary emergence of wing-like fins. *Nature*, 616:1–9.

Masoodi, F., Quasim, M., Bukhari, S., Dixit, S., and Alam, S. (2023). *Applications of machine learning and deep learning on biological data*. CRC Press.

Maurya, S., Cornejo, X., Lee, C., Kim, S.-Y., Hai, D. V., and Choudhary, R. K. (2023). Molecular phylogenetic tools reveal the phytogeographic history of the genus *Capparis* L. and suggest its reclassification. *Perspectives in Plant Ecology, Evolution and Systematics*, 58:125720.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular phylogenetics and evolution*, 66(2):526–538.

Mello, D. P. M. d., Assunção, R. M., and Murai, F. (2022). Top-Down Deep Clustering with Multi-Generator GANs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7770–7778. Number: 7.

Meng, C., Jin, S., Wang, L., and Guo, F. (2019). Gene ontology-based transfer learning for gene function prediction. *IEEE Access*, 7:54995–55007.

Michener, C. D. and Sokal, R. R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11(2):130–162.

Mimitou, E. P., Cheng, A., Montalbano, A., Hao, Y., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2021). Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature Methods*, 18(5):527–537.

Mimori, T. and Hamada, M. (2023). GeoPhy: Differentiable Phylogenetic Inference via Geometric Gradients of Tree Topologies. In *Advances in Neural Information Processing Systems*.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz, C. (2023). Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. Pages: 2023.09.19.558401 Section: New Results TLDR: It is demonstrated that structure-informed phylogenies can outperform sequence-only ones not only for distantly related proteins but also, remarkably, for more closely related ones.

Moi, D., Nishio, S., Li, X., Valansi, C., Langleib, M., Brukman, N. G., Flyak, K., Dessimoz, C., de Sanctis, D., Tunyasuvunakool, K., Jumper, J., Graña, M., Romero, H., Aguilar, P. S., Jovine, L., and Podbilewicz, B. (2022). Discovery of archaeal fusexins homologous to eukaryotic HAP2/GCS1 gamete fusion proteins. *Nature Communications*, 13(1):3880. Publisher: Nature Publishing Group.

Mount, D. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.

Murzin, A. G., Brenner, S. E., Hubbard, T. J., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press.

Nesterenko, L., Boussau, B., and Jacob, L. (2022). Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. Pages: 2022.06.24.496975 Section: New Results TLDR: This work presents a radically different approach with a transformer-based network architecture that, given a multiple sequence alignment, predicts all the pairwise evolutionary distances between the sequences, which in turn allow us to accurately reconstruct the tree topology with standard distance-based algorithms.

Network, C. G. A. R. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, 3(8):e123.

of Life (EOL), E. (2024). Encyclopedia of life (eol) dataset. Accessed: 2024-09-17.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution*, 34(8):2101–2114.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–648.

Ozsolak, F. and Milos, P. M. (2011). Rna sequencing: advances, challenges and opportunities. *Nature reviews genetics*, 12(2):87–98.

Papadopoulos, N., Gonzalo, P. R., and Söding, J. (2019). PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics*, 35(18):3517–3519.

Papili Gao, N., Hartmann, T., Fang, T., and Gunawan, R. (2020). CALISTA: Clustering and LINEAGE Inference in Single-Cell Transcriptional Analysis. *Front. Bioeng. Biotechnol.*, 8. Publisher: Frontiers.

Park, M., Ivanovic, S., Chu, G., Shen, C., and Warnow, T. (2023). UPP2: fast and accurate alignment of datasets with fragmentary sequences. *Bioinformatics*, 39(1):btad007. TLDR: UPP2 is presented, a direct improvement on UPP that produces more accurate alignments compared to leading MSA methods on datasets exhibiting substantial sequence length heterogeneity and is among the most accurate otherwise.

Pearson, W. R. and Lipman, D. J. (1990). Rapid and sensitive sequence comparison with fastp and fasta. *Methods in Enzymology*, 183:63–98.

Picard, M., Scott-Boyer, M.-P., Pérusse, L., Tremblay, A., Bouchard, C., Brisson, D., Despres, J.-P., and Vohl, M.-C. (2019). Integrative multi-omics data analysis reveals novel biomarkers and mechanisms for lung cancer. *Scientific Reports*, 9(1):16975.

Plass, M., Solana, J., Wolf, F. A., et al. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391):eaaq1723.

Qiu, X., Hill, A., Packer, J., et al. (2017a). Single-cell mrna quantification and differential analysis with census. *Nature Methods*, 14(3):309–315.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*, 14(10):979–982.

Qu, R., Cheng, X., Sefik, E., Stanley III, J. S., Landa, B., Strino, F., Platt, S., Garritano, J., Odell, I. D., Coifman, R., Flavell, R. A., Myung, P., and Kluger, Y. (2024). Gene trajectory inference for single-cell data by optimal transport metrics. *Nat Biotechnol*, pages 1–11. Publisher: Nature Publishing Group.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rao, A., Barkley, D., Franca, G. S., and Yanai, I. (2021). Deep learning for spatially resolved data in single-cell omics. *Annual Review of Biomedical Data Science*, 4:123–142.

Regev, A., Teichmann, S. A., et al. (2017). The human cell atlas. *eLife*, 6:e27041.

Rhodes, G. (2006). *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. Academic Press.

Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes, W. M., and Molina, N. (2022). Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nat Commun*, 13(1):2865. Publisher: Nature Publishing Group.

Rieppel, O. (1988). Fundamentals of comparative biology. *The Quarterly Review of Biology*, 63(3):319–320.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118.

Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Rzhetsky, A. and Nei, M. (1992). The minimum evolution approach to distance-based phylogenetic analysis: Theory and practice. *Molecular Biology and Evolution*, 9(5):945–967.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Sali, A. and Blundell, T. L. (1994). Comparative protein modeling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.

Schmidt, H., Sashittal, P., and Raphael, B. J. (2023). A zero-agnostic model for copy number evolution in cancer. *PLOS Computational Biology*, 19(11):e1011590. Publisher: Public Library of Science TLDR: The zero-agnostic copy number transformation (ZCNT) model is introduced, a simplification of the CNT model that allows the amplification or deletion of regions with zero copies and an algorithm, Lazac, is developed for solving the large parsimony problem on copy number profiles.

Scopes, R. K. (1994). *Protein Purification: Principles and Practice*. Springer Science & Business Media.

Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press.

Servellita, V., Sotomayor Gonzalez, A., Lamson, D. M., Foresythe, A., Huh, H. J., Bazinet, A. L., Bergman, N. H., Bull, R. L., Garcia, K. Y., Goodrich, J. S., Lovett, S. P., Parker, K., Radune, D., Hatada, A., Pan, C.-Y., Rizzo, K., Bertumen, J. B., Morales, C., Oluniyi, P. E., Nguyen, J., Tan, J., Stryke, D., Jaber, R., Leslie, M. T., Lyons, Z., Hedman, H. D., Parashar, U., Sullivan, M., Wroblewski, K., Oberste, M. S., Tate, J. E., Baker, J. M., Sugerman, D., Potts, C., Lu, X., Chhabra, P., Ingram, L. A., Shiau, H., Britt, W., Gutierrez Sanchez, L. H., Ciric, C., Rostad, C. A., Vinjé, J., Kirking, H. L., Wadford, D. A., Raborn, R. T., St. George, K., and Chiu, C. Y. (2023). Adeno-associated virus type 2 in US children with acute severe hepatitis. *Nature*, 617(7961):574–580. Publisher: Nature Publishing Group.

Seufi, A. M. and Galal, F. H. (2020). Fast dna purification methods: Comparative study: Dna purification. *WAS Science Nature (WASSN) ISSN: 2766-7715*, 3(1).

Sharma, M., Li, H., Sengupta, D., Prabhakar, S., and Jayadeva (2017). FORKS: Finding Orderings Robustly using k-means and Steiner trees.

Sharp, P. A. (1985). On the origin of rna splicing and introns. *Cell*, 42(2):397–400.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2002). MultiProt — A Multiple Protein Structural Alignment Algorithm. In Guigó, R. and Gusfield, D., editors, *Algorithms in Bioinformatics*, pages 235–250, Berlin, Heidelberg. Springer. TLDR: A fully automated highly efficient technique which detects the multiple structural alignments of protein structures and presents new multiple structural alignment results of protein families from the All beta proteins class in the SCOP classification.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. (2024). Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353.

Sherry, S. T. et al. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2004). Recognition of Functional Sites in Protein Structures. *Journal of Molecular Biology*, 339(3):607–633. TLDR: A novel method is described, SiteEngine, that assumes no sequence or fold similarities and is able to recognize proteins that have similar binding sites and may perform similar functions, and which may aid in assigning a function and in classification of binding patterns.

Sievers, F. and Higgins, D. G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology*, 1079:105–116. Publisher: Springer.

Smith, M. L. and Hahn, M. W. (2023). Phylogenetic inference using generative adversarial networks. *Bioinformatics*, 39(9):btad543. TLDR: PhyloGAN is developed, a GAN that infers phylogenetic relationships among species and uses an evolutionary model as the generator, and infers a phylogenetic tree either considering or ignoring gene tree heterogeneity.

Smith, T. F. and Waterman, M. S. (1981a). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Smith, T. F. and Waterman, M. S. (1981b). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Spain, L., Coulton, A., Lobon, I., Rowan, A., Schnidrig, D., Shepherd, S. T., Shum, B., Byrne, F., Goicoechea, M., Piperni, E., et al. (2023). Late-stage metastatic melanoma emerges through a diversity of evolutionary pathways. *Cancer discovery*, 13(6):1364–1385.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stassen, S. V., Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K., and Tsia, K. K. (2021). Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nat Commun*, 12(1):5528. Publisher: Nature Publishing Group.

Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., Dong, L., Dahdul, W. M., Stewart, C., Berger-Wolf, T., Chao, W.-L., and Su, Y. (2024). Bioclip: A vision foundation model for the tree of life. *arXiv preprint arXiv:2311.18803*. Accessed: 2024-09-17.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. I., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21.

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nature reviews genetics*, 20(5):257–272.

Suchard, M. A. and Redelings, B. D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16):2047–2048. Publisher: Oxford University Press.

Suvorov, A., Hochuli, J., and Schrider, D. R. (2020). Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. *Systematic Biology*, 69(2):221–233.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). *Phylogenetic inference*, pages 407–514. Sinauer Associates.

Szklarczyk, D. et al. (2019). String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. (2020). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Nature Ecology & Evolution*, 4(8):1160–1165.

Tamura, T., Ito, J., Uriu, K., Zahradnik, J., Kida, I., Anraku, Y., Nasser, H., Shofa, M., Oda, Y., Lytras, S., Nao, N., Itakura, Y., Deguchi, S., Suzuki, R., Wang, L., Begum, M. M., Kita, S., Yajima, H., Sasaki, J., Sasaki-Tabata, K., Shimizu, R., Tsuda, M., Kosugi, Y., Fujita, S., Pan, L., Sauter, D., Yoshimatsu, K., Suzuki, S., Asakura, H., Nagashima, M., Sadamasu, K., Yoshimura, K., Yamamoto, Y., Nagamoto, T., Schreiber, G., Maenaka, K., Hashiguchi, T., Ikeda, T., Fukuhara, T., Saito, A., Tanaka, S., Matsuno, K., Takayama, K., and Sato, K. (2023). Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nature Communications*, 14(1):2800. Publisher: Nature Publishing Group.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6:377–382.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of dna sequence. *Lecture of Mathematics for Life Science*, 17:57.

Tegally, H., Moir, M., Everatt, J., Giovanetti, M., Scheepers, C., Wilkinson, E., Subramoney, K., Makatini, Z., Moyo, S., Amoako, D. G., Baxter, C., Althaus, C. L., Anyaneji, U. J., Kekana, D., Viana, R., Giandhari, J., Lessells, R. J., Maponga, T., Maruapula, D., Choga, W., Matshaba, M., Mbulawa, M. B., Msomi, N., Naidoo, Y., Pillay, S., Sanko, T. J., San, J. E., Scott, L., Singh, L., Magini, N. A., Smith-Lawrence, P., Stevens, W., Dor, G., Tshiabuila, D., Wolter, N., Preiser, W., Treurnicht, F. K., Venter, M., Chiloane, G., McIntyre, C., O'Toole, A., Ruis, C., Peacock, T. P., Roemer, C., Kosakovsky Pond, S. L., Williamson, C., Pybus, O. G., Bhiman, J. N., Glass, A., Martin, D. P., Jackson, B., Rambaut, A., Laguda-Akingba, O., Gaseitsiwe, S., von Gottberg, A., and de Oliveira, T. (2022). Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nature Medicine*, 28(9):1785–1790. Publisher: Nature Publishing Group.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994a). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994b). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680. Publisher: Oxford University Press.

Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. (2000). From sequence to function: methods and applications. *Current Opinion in Structural Biology*, 10(3):374–380.

Thrall, P. H., Oakeshott, J. G., Fitt, G., Southerton, S., Burdon, J. J., Sheppard, A., Russell, R. J., Zalucki, M., Heino, M., and Ford Denison, R. (2011). Evolution in agriculture: the application of evolutionary approaches to the management of biotic interactions in agro-ecosystems. *Evolutionary Applications*, 4(2):200–215. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1752-4571.2010.00179.x TLDR: Biotic interactions involving pests and pathogens are focused on as exemplars of situations where integration of agronomic, ecological and evolutionary perspectives has practical value and the use of predictive frameworks based on evolutionary models as pre emptive management tools are advocated.

Tian, T., Zhong, C., Lin, X., Wei, Z., and Hakonarson, H. (2023). Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning. *Genome Research*, 33(2):232–246.

Tirosh, I., Venteicher, A. S., Hebert, C., et al. (2016). Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313.

Tisza, M. J., Pastrana, D. V., Welch, N. L., Stewart, B., Peretti, A., Starrett, G. J., Pang, Y.-Y. S., Krishnamurthy, S. R., Pesavento, P. A., McDermott, D. H., Murphy, P. M., Whited, J. L., Miller, B., Brenchley, J., Rosshart, S. P., Rehermann, B., Doorbar, J., Ta'ala, B. A., Pletnikova, O., Troncoso, J. C., Resnick, S. M., Bolduc, B., Sullivan, M. B., Varsani, A., Segall, A. M., and Buck, C. B. (2020). Discovery of several thousand highly diverse circular DNA viruses. *eLife*, 9:e51971. Publisher: eLife Sciences Publications, Ltd.

Tran, D., Nguyen, H., Tran, B., La Vecchia, C., Luu, H. N., and Nguyen, T. (2021). Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nature communications*, 12(1):1029.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014a). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014b). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*, 32(4):381–386.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S.-R., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014c). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Tsai, Y.-H. H. et al. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Tung, P.-Y., Blischak, J. D., Hsiao, C.-J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., and Gilad, Y. (2017). Single-cell transcriptomics reveals gene expression heterogeneity in biological processes. *Genome Biology*, 18(1):1–14.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Söding, J., and Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(February). Publisher: Springer US.

Vandenhirtz, M., Barkmann, F., Manduchi, L., Vogt, J. E., and Boeva, V. (2023). sctree: Discovering cellular hierarchies in the presence of batch effects in scrna-seq data. *arXiv preprint arXiv:2304.12345*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vazquez, J. M., Pena, M. T., Muhammad, B., Kraft, M., Adams, L. B., and Lynch, V. J. (2022). Parallel evolution of reduced cancer risk and tumor suppressor duplications in Xenarthra. *eLife*, 11:e82558. Publisher: eLife Sciences Publications, Ltd.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wagner, A., Regev, A., and Yosef, N. (2020). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 38(12):1401–1414.

Waits, L. P. and Paetkau, D. (2005). Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *The Journal of Wildlife Management*, 69(4):1419–1433.

Wang, B., Hu, X., Zhang, C., Li, P., and Yu, P. S. (2022). Hierarchical GAN-Tree and Bi-Directional Capsules for multi-label image classification. *Knowledge-Based Systems*, 238:107882.

Wang, D. and Gu, J. (2018). Vasc: dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, Proteomics and Bioinformatics*, 16(5):320–331.

Wang, K., Hou, L., Wang, X., Zhai, X., Lu, Z., Zi, Z., Zhai, W., He, X., Curtis, C., Zhou, D., and Hu, Z. (2024). PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes. *Nature Biotechnology*, 42(5):778–789. Publisher: Nature Publishing Group TLDR: Applying PhyloVelo to seven lineage-traced scRNA-seq datasets, generated using CRISPR-Cas9 editing, lentiviral barcoding or immune repertoire profiling, demonstrates its high accuracy and robustness in inferring complex lineage trajectories while outperforming RNA velocity.

Wang, R., Li, X., Wang, W., Kang, M., Wu, Y., Chen, J., and Lu, Y. (2020). Deep learning in multi-omics integration: A review of applications in cancer. *Frontiers in Genetics*, 11:412.

Wang, R., Zhang, R., Khodaverdian, A., and Yosef, N. (2023). Theoretical guarantees for phylogeny inference from single-cell lineage tracing. *Proceedings of the National Academy of Sciences*, 120(12):e2203352120.

Wang, S., Karikomi, M., MacLean, A. L., and Nie, Q. (2019). Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Research*, 47(11):e66.

Wang, S., Ma, J., Peng, J., and Xu, J. (2013). Protein structure alignment beyond spatial proximity. *Scientific Reports*, 3(1):1448. Publisher: Nature Publishing Group TLDR: Experimental results show that DeepAlign can generate structure alignments much more consistent with manually-curated alignments than other automatic tools especially when proteins under consideration are remote homologs, implying that in addition to geometric similarity, evolutionary information and hydrogen-bonding similarity are essential to aligning two protein structures.

Webb, B. and Sali, A. (2016). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59.

Wolf, P. G., Cowley, E. S., Breister, A., Matatov, S., Lucio, L., Polak, P., Ridlon, J. M., Gaskins, H. R., and Anantharaman, K. (2022). Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer. *Microbiome*, 10(1):64.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. John Wiley & Sons.

Xie, T., Matsen IV, F. A., Suchard, M. A., and Zhang, C. (2024). Variational Bayesian Phylogenetic Inference with Semi-implicit Branch Length Distributions. arXiv:2408.05058 [cs, stat].

Xie, T. and Zhang, C. (2023). ARTree: A Deep Autoregressive Model for Phylogenetic Inference. In *Advances in Neural Information Processing Systems*.

Yang, K. K., Wu, Z., and Arnold, F. H. (2019). Machine learning-guided directed evolution for protein engineering. *Nature Methods*, 16:687–694.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39:105–111.

Yariv, B., Yariv, E., Kessel, A., Masrati, G., Chorin, A. B., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2023). Using evolutionary data to make sense of macromolecules with a "face-lifted" consurf. *Protein Science*, 32(3):e4582.

Yeung, W., Zhou, Z., Mathew, L., Gravel, N., Taujale, R., O'Boyle, B., Salcedo, M., Venkat, A., Lanzilotta, W., Li, S., and Kannan, N. (2023). Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Briefings in Bioinformatics*, 24(1):bbac619. TLDR: This work develops workflows and visualization methods for the classification of protein families using sequence embedding derived from protein language models and proposes a new hierarchical classification for the S-Adenosyl-L-Methionine enzyme superfamily which has been difficult to classify using traditional alignment-based approaches.

You, J., Liu, B., Ying, R., Pande, V., and Leskovec, J. (2018a). Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

You, J., Ying, R., Ren, X., Hamilton, W., and Leskovec, J. (2018b). Graphrnn: Generating realistic graphs with deep auto-regressive models. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Yuan, R., Zheng, B., Li, Z., Ma, X., Shu, X., Qu, Q., Ye, X., Li, S., Tang, P., and Chen, X. (2023). The chromosome-level genome of Chinese praying mantis Tenodera sinensis (Mantodea: Mantidae) reveals its biology as a predator. *GigaScience*, 12:giad090. TLDR: The high-quality genome assembly of the praying mantis provides a valuable repository for studying the evolutionary patterns of the mantis genomes and the gene expression profiles of insect predators.

Zang, Z., Cheng, S., Xia, H., Li, L., Sun, Y., Xu, Y., Shang, L., Sun, B., and Li, S. Z. (2022a). Dmt-ev: An explainable deep network for dimension reduction. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1710–1727.

Zang, Z., Cheng, S., Xia, H., Li, L., Sun, Y., Xu, Y., Shang, L., Sun, B., and Li, S. Z. (2024a). DMT-EV: An Explainable Deep Network for Dimension Reduction. *IEEE transactions on visualization and computer graphics*, 30(3):1710–1727. TLDR: A deep neural network method called DMT-EV is developed, which provides not only excellent performance in structural maintainability but also explainability to the DR therein, and consistently outperforms the state-of-the-art methods in both performance measures and explainability.

Zang, Z., Li, S., Wu, D., Guo, J., Xu, Y., and Li, S. Z. (2021). Unsupervised Deep Manifold Attributed Graph Embedding. *arXiv:2104.13048 [cs]*. arXiv: 2104.13048 version: 1.

Zang, Z., Li, S., Wu, D., Wang, G., Wang, K., Shang, L., Sun, B., Li, H., and Li, S. Z. (2022b). Dlme: Deep local-flatness manifold embedding. pages 576–592. Springer, Cham.

Zang, Z., Luo, H., Wang, K., Zhang, P., Wang, F., Li, S. Z., and You, Y. (2024b). DiffAug: Enhance Unsupervised Contrastive Learning with Domain-Knowledge-Free Diffusion-based Data Augmentation. In *International Conference on Machine Learning*.

Zang, Z., Shang, L., Yang, S., Wang, F., Sun, B., Xie, X., and Li, S. Z. (2023a). Boosting Novel Category Discovery Over Domains with Soft Contrastive Learning and All in One Classifier. pages 11824–11833. IEEE Computer Society. TLDR: A framework named Soft-contrastive All-in-one Network (SAN) is proposed for ODA and UNDA tasks, which includes a novel data-augmentation-based soft contrastive learning (SCL) loss to fine-tune the backbone for feature transfer and a more human-intuitive classifier to improve new class discovery capability.

Zang, Z., Wang, W., Song, Y., Lu, L., Li, W., Wang, Y., and Zhao, Y. (2019). Hybrid Deep Neural Network Scheduler for Job-Shop Problem Based on Convolution Two-Dimensional Transformation. *Computational Intelligence and Neuroscience*, 2019(Research Article):1–19. TLDR: A hybrid deep neural network scheduler (HDNNS) is proposed to solve job-shop scheduling problems (JSSPs) and the results show that the MAKESPAN index of HDNNS is 9% better than that of HNN and the index is also 4% betterthan that of ANN in ZLP dataset.

Zang, Z., Xu, Y., Lu, L., Geng, Y., Yang, S., and Li, S. Z. (2023b). Udrn: unified dimensional reduction neural network for feature selection and feature projection. *Neural Networks*, 161:626–637.

Zang, Z., Xu, Y., Lu, L., Geng, Y., Yang, S., and Li, S. Z. (2023c). Udrn: unified dimensional reduction neural network for feature selection and feature projection. *Neural Networks*, 161:626–637. Publisher: Pergamon.

Zapatero, M. R., Tong, A., Opzoomer, J. W., O'Sullivan, R., Rodriguez, F. C., Sufi, J., Vlckova, P., Nattress, C., Qin, X., Claus, J., et al. (2023). Trellis tree-based analysis reveals stromal regulation of patient-derived organoid drug responses. *Cell*, 186(25):5606–5619.

Zhang, C. (2020). Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows. In *Advances in Neural Information Processing Systems*, volume 33, pages 18760–18771. Curran Associates, Inc.

Zhang, C., Fu, H., Hu, Q., Cao, X., Liu, Q., and Tian, Q. (2018). Multi-view multiple clusterings via deep matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):90–103.

Zhang, C. and Iv, F. A. M. (2018). Variational Bayesian Phylogenetic Inference. In *Advances in Neural Information Processing Systems*.

Zhang, T., Tan, S., Tang, N., Li, Y., Zhang, C., Sun, J., Guo, Y., Gao, H., Cai, Y., Sun, W., Wang, C., Fu, L., Ma, H., Wu, Y., Hu, X., Zhang, X., Gee, P., Yan, W., Zhao, Y., Chen, Q., Guo, B., Wang, H., and Zhang, Y. E. (2024). Heterologous survey of 130 DNA transposons in human cells highlights their functional divergence and expands the genome engineering toolbox. *Cell*, 187(14):3741–3760.e30. Publisher: Elsevier TLDR: It is found that the Tc1/mariner superfamily exhibits elevated activity, potentially explaining their pervasive horizontal transfers and highlights the varied transposition features and evolutionary dynamics of DNA TEs and increases the TE toolbox diversity.

Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309. Publisher: Oxford University Press.

Zhang, Y., Tran, D., Nguyen, T., Dascalu, S. M., and Harris, F. C. (2023). A robust and accurate single-cell data trajectory inference method using ensemble pseudotime. *BMC Bioinformatics*, 24(1):55.

Zhang, Y. and Yang, Z. (2020). Machine learning in phylogenetics and its applications. *Briefings in Bioinformatics*, 21(2):636–647.

Zhao, A., Sun, J., and Liu, Y. (2023). Understanding bacterial biofilms: From definition to treatment strategies. *Frontiers in cellular and infection microbiology*, 13:1137947.

Zheng, G.-C. et al. (2017a). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049.

Zheng, G.-C., Terry, J., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., et al. (2017b). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. (2017c). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049.

Zheng, Y., Liu, Y., Yang, J., Dong, L., Zhang, R., Tian, S., Yu, Y., Ren, L., Hou, W., Zhu, F., et al. (2024). Multi-omics data integration using ratio-based quantitative profiling with quartet reference materials. *Nature biotechnology*, 42(7):1133–1149.

Zhou, M. Y., Yan, Z., Layne, E., Malkin, N., Zhang, D., Jain, M., Blanchette, M., and Bengio, Y. (2023). PhyloGFN: Phylogenetic inference with generative flow networks. In *The Twelfth International Conference on Learning Representations*.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023a). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, L., Wu, J., Li, M., Fang, H., Zhang, J., Chen, Y., Chen, J., Cheng, T., Zhu, L., Wu, J., Li, M., Fang, H., Zhang, J., Chen, Y., Chen, J., and Cheng, T. (2023b). Genome-wide discovery of CBL genes in *Nitraria tangutorum* Bobr. and functional analysis of *NtCBL1-1* under drought and salt stress. *Forestry Research*, 3(1). Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Maximum Academic Press Number: FR-2023-0028 Primary_atype: Forestry Research Publisher: Maximum Academic Press Subject_term: ARTICLE Subject_term_id: ARTICLE.