Graph-Based Reconstruction and Analysis of Disease Transmission Networks using Viral Genomic Data

Ziqi Ke[1]* and Haris Vikalo[1]

[1]Department of Electrical and Computer Engineering

The University of Texas at Austin

*To whom correspondence should be addressed.

E-mail: ziqike@utexas.edu, hvikalo@ece.utexas.edu

May 11, 2023

Keywords: disease transmission networks, deep learning, auto-encoders, convolutional neural networks

Abstract: Understanding the patterns of viral disease transmissions helps establish public health policies and aids in controlling and ending a disease outbreak. Classical methods for studying disease transmission dynamics that rely on epidemiological data, such as times of sample collection and duration of exposure intervals, struggle to provide desired insight due to limited informativeness of such data. A more precise characterization of disease transmissions may be acquired from sequencing data that reveals genetic distance between viral genomes in patient samples. Indeed, genetic distance between viral strains present in hosts contains valuable information about transmission history, thus motivating the design of methods that rely on genomic data to reconstruct a directed disease transmission network, detect transmission clusters, and identify significant network nodes (e.g., super-spreaders). In this paper, we present a novel end-to-end framework for the analysis of viral transmissions utilizing viral genomic (sequencing) data. The proposed framework groups infected hosts into transmission clusters based on the reconstructed viral strains infecting them; the genetic distance between a pair of hosts is calculated using Earth Mover's Distance, and further used to infer transmission direction between the hosts. To quantify the significance of a host in the transmission network, the importance score is calculated by a graph convolutional autoencoder. The viral transmission network is represented by a directed minimum spanning tree utilizing the Edmond's algorithm modified to incorporate constraints on the importance scores of the hosts. The proposed framework outperforms state-of-the-art techniques for the analysis of viral transmission dynamics in several experiments on semi-experimental as well as experimental data. Source codes are available at https://github.com/WuLoli/AutoNet

1 Introduction

Understanding the spread of a pathogen across a network of hosts assists in the development of effective public health interventions and disease prevention, containment and eradication strategies. Examples include studies of infectious disease transmissions in support of reconstructing a transmission network, detecting transmission clusters, and identifying super-spreaders in the network.

Traditional methods for infectious disease outbreak analysis that rely on epidemiological data such as the time of testing and duration of exposure suffer from labour-intensive contact tracing (Hall et al., 2016), and generally struggle to provide desired insight due to limited informativeness of such data. For example, the time of testing is an unreliable indicator of the time of infection, especially for a disease that may be asymptomatic long after the infection (e.g., in the case of the original COVID-19 strain, the symptoms occur 2-14 days following the infection). With the advance of next-generation sequencing (NGS) technologies, rapid and accurate reconstruction of viral populations is feasible and affordable. Since genomic evolutionary distance between viral strains present in different hosts contains valuable information about transmission history, analysis of genomic data collected by NGS technologies may provide significant insight into disease transmission patterns.

Existing methods for studying disease transmission patterns can be classified as (i) epidemiological data driven, (ii) genomic data driven, and (iii) methods that rely on the combination of both epidemiological and genomic data. Prior work on utilizing epidemiological data includes a social network analysis of Mycobacterium Tuberculosis transmissions using patient medical records and contact interview forms (Cook et al., 2007); stochastic mathematical models describing disease transmission process using both behavioral and environmental data (Grassly and Fraser, 2008); analysis of social networks built utilizing existing clinical informatics resources aiming to explore the implications of patient-healthcare worker interactions on disease transmission (Gundlapalli et al., 2009); study of a human contact network formed using close proximity interaction data, meant to provide insight into transmissions of an influenza-like disease during a typical day at an American high school (Salathe et al., 2010); a stochastic eco-epidemiological model for the analysis of dengue transmission using seasonal and spatial dynamics (Otero and Solari, 2010); a 2009 H1N1 pandemic influenza transmission model estimated via Markov chain Monte Carlo sampling of demographic and clinical data (Cauchemez et al., 2011); likelihood-based methods for the analysis of influenza A(H1N1) transmission in a school population using clinical symptom and contact data (Hens et al., 2012); mathematical modeling of the transmission of lumpy skin disease virus using direct and indirect contact information of cattle exhibiting typical clinical signs of the disease (Magori-Cohen et al., 2012); network models of transmission dynamics in wild animal and livestock populations using contact data (Craft, 2015); a statistical inference method for the construction of the influenza A virus transmission tree in a college-based population utilizing epidemiological, clinical and contact tracing data (Zhang and De Angelis, 2016); a model enabling reconstruction of the full-spectrum dynamics of COVID-19 using epidemiological data (Hao et al., 2020); statistical modeling for the reconstruction of transmission pairs for COVID-19 utilizing detailed demographic characteristics, travel history, social relationships and epidemiological timelines (Xu et al., 2020); and a visualization technique based on individual reports of epidemiological data to construct disease transmission graphs for the COVID-19 epidemic (Luo et al., 2021).

As an alternative, methods that aim to go beyond traditional techniques for outbreak analysis by relying on genomic data have recently started gaining attention, including a graph based technique for reconstructing transmission trees utilizing genomic data from the early stages of the A/H1N1 influenza pandemic (Jombart et al., 2011); minimum spanning tree based methods for estimating relationships among individual strains or isolates in molecular epidemiology (Salipante and Hall, 2011); a Bayesian approach for reconstructing densely sampled outbreaks from whole-genome sequence data and inferring a transmission network via a Monte Carlo Markov chain (Didelot et al., 2014); an approach for analyzing genetic distance between pathogen strains to estimate routes of transmission in bacterial disease outbreaks (Worby et al., 2014); a method for molecular detection of hepatitis C virus transmissions in outbreak settings (Campo et al., 2016); stochastic epidemic models for investigating person-to-person communicable disease transmission with densely sampled genomic data (Worby et al., 2016); a statistical framework to infer host-to-host transmissions built around a computationally efficient model of pathogen evolution (De Maio et al., 2016); algorithms to infer genetic relatedness, detect possible transmissions, and analyze clusters' structure validated using experimental sequencing data from HCV outbreaks (Glebova et al., 2017); an approach incorporating shared genetic variants and phylogenetic distance data to identify transmission routes from pathogen deep-sequence data (Worby et al., 2017); a graph-based method for modeling viral evolution and epidemic spread via evolutionary analysis of intra-host viral populations (Skums et al., 2018); a Bayesian approach for transmission inference that explicitly models evolution of pathogen populations in an outbreak (De Maio et al., 2018); a statistical learning approach with a pseudo-evolutionary model to infer epidemiological links from deep sequencing data for human, animal and plant diseases (Alamil et al., 2019); an algorithm based on Earth mover's distance for viral outbreak investigations utilizing raw NGS reads (Melnyk et al., 2020), and a maximum-likelihood framework that integrates phylogenetic and random graph models of genomic data (Skums et al., 2022).

Methods relying on the combination of epidemiological and genomic data include a minimum spanning tree model to identify the history of transmission of hepatitis C virus in an outbreak (Spada et al., 2004); a maximum-likelihood approach for the analysis of HIV-1 transmissions utilizing clinical, epidemiological and phylogenomic data (Fisher et al., 2010); a stochastic infectious disease model for the analysis of the spread of infectious salmon anaemia among salmon farms utilizing genetic and space-time data (Aldrin et al., 2011); a likelihood-based framework drawing upon temporal, geographical and genomic data in an epidemic of avian influenza A (H7N7) in The Netherlands in 2003 (Ypma et al., 2012); Bayesian inference frameworks to reconstruct most likely transmission patterns and infection dates for the analysis of UK epidemics of foot-and-mouth disease virus (Cottam et al., 2008; Morelli et al., 2012; Lau et al., 2015); a statistical framework to infer key epidemiological and mutational parameters by simultaneously estimating the phylogenetic and transmission tree in an outbreak of foot-and-mouth disease (Ypma et al., 2013); a statistical method exploiting both pathogen sequences and collection dates to reveal dynamics of densely sampled outbreaks for the analysis of the 2003 Singaporean outbreak of Severe Acute Respiratory Syndrome (SARS) (Jombart et al., 2014); a Bayesian inference method for the reconstruction of HIV transmission trees from viral sequences and uncertain infection time data (Montazeri et al., 2018); a systematic Bayesian transmission network model to reconstruct the transmission network of the foot-and-mouth disease epidemic in Japan in 2010 (Hayama et al., 2019); a Bayesian methodology that uses contact data for the inference of transmission trees in a statistically rigorous manner, alongside genomic data and temporal data (Campbell et al., 2019), and an analysis of mumps virus transmission in the US utilizing epidemiological data from public health investigations and mumps virus whole genome sequences (Wohl et al., 2020).

In this paper, we present an end-to-end framework for the reconstruction and interpretation of disease transmissions from genomic data; the framework is motivated by an observation that the reconstruction task would benefit from using not only the genetic distances between hosts but also from quantifying and leveraging the importance of a host in a viral transmission network. The proposed framework aims to first divide patients into different clusters, where patients inside the same cluster are infected by the same type of viral variant of the infectious disease; to this end, we utilize a modified version of the viral population reconstruction method TenSQR (Ahn et al., 2018). Next, the genetic distance between a pair of patients inside the same cluster is calculated based on the Earth Mover's Distance between corresponding k-mer distributions (Rubner et al., 1998; Melnyk et al., 2020), which reflects the minimal amount of work that must be performed to transform one distribution into another by moving 'distribution mass' around. Third, possible transmission directions between pairs of patients are determined using the calculated genetic distances. Fourth, the importance score of every patient is calculated via a custom-designed graph convolutional auto-encoder inspired by (Salha et al., 2019). Auto-encoders are neural networks that can be trained to automatically extract salient low-dimensional representations of high-dimensional data in an unsupervised manner (Goodfellow et al., 2016); graph convolutional auto-encoders are a variant of auto-encoders specifically designed to analyze graph-structured data. Finally, a directed minimum spanning tree, incorporating the local and global information provided by the genetic distances and the importance scores, is constructed by imposing importance scores constraints in the classical Edmonds' algorithm (Edmond, 1967).

Our main contributions are summarized as follows:

- We developed an end-to-end computational framework that utilizes sequencing data to detect disease transmission clusters, reconstruct a directed disease transmission network, and quantify the significance of a viral host within the transmission network.

- Our proposed method reconstructs the directed disease transmission network by leveraging not only the local information captured by the genetic distances but also the global information provided by the importance scores of each host extracted by a designed graph convolutional auto-encoder.

- We conducted several experiments on experimental COVID-19 data and semi-experimental hepatitis C virus data, and compared the results with existing state-of-the-art methods, demonstrating the ability of the proposed framework to efficiently and accurately reconstruct a disease transmission network, outperforming state-of-the-art methods.

2 Methods

2.1 Problem Formulation

The aims of the framework for disease transmission network analysis developed in this paper include:

1. Discovery of transmission clusters, where the clusters collect hosts infected by the same pathogen variant.

2. Inference of a directed transmission network of hosts based on genomic information about viral pathogens infecting them.

3. Identification of super-spreaders/critical hosts in the network.

To detect clusters of hosts/patients, we adapt to the current problem the viral population reconstruction method TenSQR we previously co-developed in (Ahn et al., 2018). Specifically, we first successively cluster hosts into communities represented by the consensus sequences of cluster-specific viral genomes, and then construct a weighted directed acyclic graph $G = (V, E, \mathbf{w})$ for each cluster, where V is the set of nodes representing hosts, E is the set of edges indicating disease transmission directions, and $\mathbf{w}$ is the set of edge weights. For instance, node v_i denotes the i^{th} host, edge $e_{ij} = (v_i, v_j)$ indicates that the j^{th} host may have been infected by the i^{th} host, and weight w_{ij} is the genetic distance between the pathogens infecting the two hosts. Following (Melnyk et al., 2020), we represent the genetic distance separating two pathogens via the Earth Mover's Distance between corresponding k-mer distributions. While the genetic distances provide useful local information about transmissions, we further design a graph convolutional auto-encoder to obtain global information in form of importance scores quantifying how influential hosts are in the network. The importance scores are in turn used to guide the search for a directed minimum spanning tree revealing transmissions along the network. Fig. 1 illustrates the architecture of the proposed end-to-end framework.

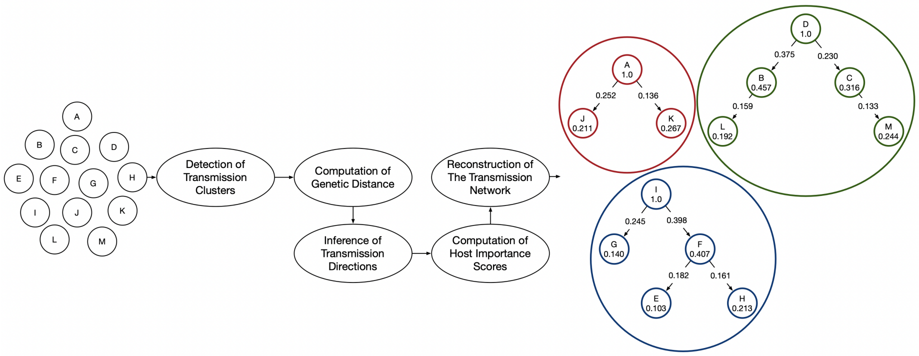
A
B
C
D
E
F
G
H
I
J
K
L
M
Detection of Transmission Clusters
Computation of Genetic Distance
Inference of Transmission Directions
Computation of Host Importance Scores
Reconstruction of The Transmission Network
A 1.0
0.252
0.136
J 0.211
K 0.267
D 1.0
0.375
0.230
B 0.457
C 0.316
0.159
0.133
L 0.192
M 0.244
I 1.0
0.245
0.398
G 0.140
F 0.407
0.182
0.161
E 0.103
H 0.213

Figure 1: Architecture of the proposed end-to-end framework for disease transmission analysis. Differently colored nodes illustrate three disease transmission clusters, with arrows showing the transmission directions. The numbers shown in nodes and on edges are the importance scores and the genetic distances, respectively.

2.2 Detection of Transmission Clusters

As already mentioned, the method used to detect disease transmission clusters builds upon TenSQR proposed in our prior work (Ahn et al., 2018). Let P denote an $n \times l$ pathogen matrix where n is the number of hosts and l is the (maximum) length of the viral genomes; note that the entries in P may be erroneous due to sequencing errors. After organizing hosts into transmission clusters, where hosts in the same cluster are infected by the same variant type of the pathogen, the consensus sequence is formed to represent each cluster. The reconstructed consensus sequences form a $k \times l$ consensus matrix C, where k denotes the number of transmission clusters, which our method can determine automatically (to this end we rely on the framework introduces in our prior work (Ahn et al., 2018), see Subsection 2.4). To ensure the distances between nucleotides are consistent, we denote nucleotides by 4-dimensional standard unit vectors $\mathbf{e}_i^{(4)}$, $1 \leq i \leq 4$, with 0s in all positions except the i^{th} one that has value 1 (e.g., $\mathbf{e}_1^{(4)} = [1\ 0\ 0\ 0]$, $\mathbf{e}_2^{(4)} = [0\ 1\ 0\ 0]$, and so on). The pathogen matrix P can be re-written as a binary tensor $\mathcal{P}$ whose fibers represent nucleotides and horizontal slices correspond to pathogen strains. The transmission cluster detection can then be formulated as a tensor factorization problem since $\mathcal{P}$ can be thought of as being obtained by multiplying an $n \times k$ pathogen strain membership indicator matrix M and a binary tensor $\mathcal{C}$ that encodes the consensus sequence of each cluster. Fibers of $\mathcal{C}$ are standard unit vectors $\mathbf{e}_i^{(4)}$ representing alleles, while each lateral slice of $\mathcal{C}$ is one of the k consensus sequences representing cluster centroids. Note that the pathogen strain membership indicator matrix M has standard unit vectors $\mathbf{e}_i^{(k)}$, $1 \leq i \leq k$, for rows; if the j^{th} row of $\mathbf{M}$ is $\mathbf{e}_i^{(k)}$, that indicates the j^{th} host is assigned to the i^{th} transmission cluster. To proceed, we formulate the transmission network clustering problem as a collection of $k-1$ tensor factorization problems; after each factorization, pathogens associated with the most dominant transmission cluster are removed from $\mathcal{P}$ and the factorization (of smaller dimension) is performed anew until only one cluster remains.

Fig. 2 illustrates the detection of transmission clusters. To formalize the

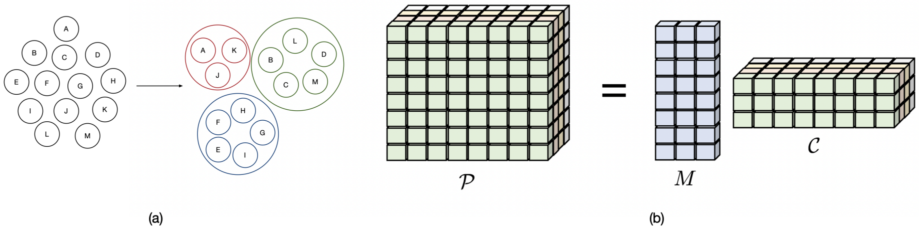

Figure 2: (a) Detecting transmission clusters. (b) Identifying viral species via tensor factorization.

tensor factorization representation of the problem, let $\mathbf{P} \in \{0,1\}^{n \times 4l}$ and $\mathbf{C} \in \{0,1\}^{4l \times k}$ denote the mode-1 unfoldings of tensors $\mathcal{P}$ and $\mathcal{C}$, respectively. The transmission network clustering problem can be written as

$$\min_{\mathbf{M},\mathbf{C}} \frac{1}{2} \|\mathbf{P} - \mathbf{M}\mathbf{C}^\top\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This is a non-convex optimization that can be approximately solved via alternating minimization. For further details regarding finding a solution to the above optimization problem and thus successively identifying pathogens associated with the transmission clusters, we refer a reader to (Ahn et al., 2018).

2.3 Computation of Genetic Distances

Given a cluster of hosts, we build a graph in which the edge weights reflect genetic similarities between pathogens infecting the hosts. The Earth Mover's Distance (EMD) used to measures the difference between two probability distributions is calculated as a solution to the transportation problem (i.e., the Monge-Kantorovich problem), which is readily formalized as a linear program (Rubner et al., 1998, 2000). Let

$$P = \{(p_1, w_{p1}), (p_2, w_{p2}), ..., (p_m, w_{pm})\}$$

denote the normalized k-mer frequencies of a viral strain, where p_i represents the i^{th} k-mer and w_{pi} is its frequency; following parameter tuning, we set $k = 4$ in our experiments.

Similarly, let us denote the normalized k-mer frequencies of another viral strain by

$$Q = \{(q_1, w_{q1}), (q_2, w_{q2}), ..., (q_n, w_{qn})\}.$$

Let $D = [d_{ij}]$ be the shortest path distance between k-mers p_i and q_j in the undirected De Bruijn graph formed by all the pathogens in a transmission cluster. [Fig. 3 illustrates a sample de Bruijn graph constructed from two sequences with $k = 4$.] The Earth Mover's Distance between P and Q is defined as

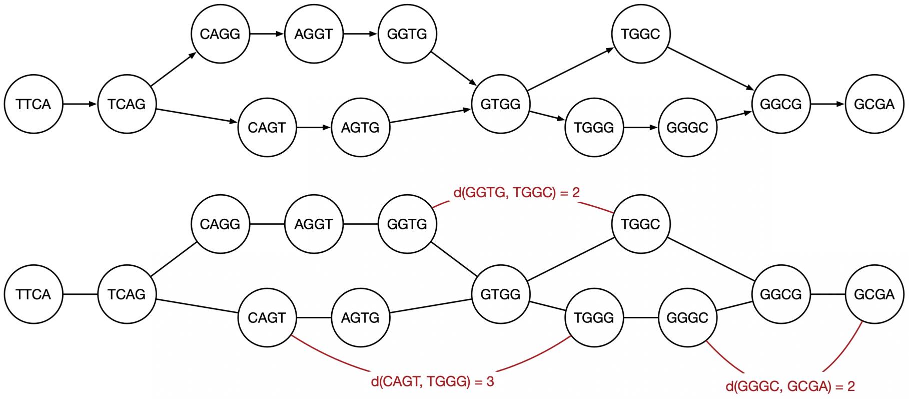

Figure 3: An example of a de Bruijin graph constructed from sequences TTCAGTGGGCGA and TTCAGGTGGCGA. A few selected pairwise distances between k-mers are also indicated.

$$\text{EMD}(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}},$$

where the total flow $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = 1$. The aforementioned linear program is focused on finding $F = [f_{ij}]$ minimizing

$$\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}$$

subject to

$$\begin{aligned}
f_{ij} \geq 0, \quad 1 \leq i \leq m \quad \text{and} \quad 1 \leq j \leq n \\
\sum_{j=1}^{n} f_{ij} \leq w_{pi},\ 1 \leq i \leq m, \quad \sum_{i=1}^{m} f_{ij} \leq w_{qi},\ 1 \leq j \leq n \\
\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\{\sum_{i=1}^{m} w_{pi}, \sum_{j=1}^{n} w_{qj}\}.
\end{aligned}$$

where f_{ij} denotes the flow between p_i and q_j.

2.4 Inference of Transmission Directions

For two hosts A and B with normalized k-mer frequencies $(f_1^A, f_2^A, ..., f_n^A)$ and $(f_1^B, f_2^B, ..., f_n^B)$, where n denotes the total number of unique k-mers in A and B, the maximum mean k-mer distribution is defined as (Melnyk et al., 2020)

$$\text{Mean}(A, B) = \left(\frac{f_1^{max}}{\sum_{i=1}^{n} f_i^{max}}, \ldots, \frac{f_n^{max}}{\sum_{i=1}^{n} f_i^{max}} \right),$$

where $f_i^{max} = \max\{f_i^A, f_i^B\}$ for all $i = 1, 2, \ldots, n$. Then the transmission direction between hosts A and B is assumed to be from A to B if

$$\text{EMD}(\text{Mean}(A, B), A) < \text{EMD}(\text{Mean}(A, B), B),$$

and from B to A otherwise; as before, $\text{EMD}(\cdot, \cdot)$ denotes the Earth Mover's Distance between its arguments. Fig. 4 illustrates inference of the transmission direction between hosts A and B.

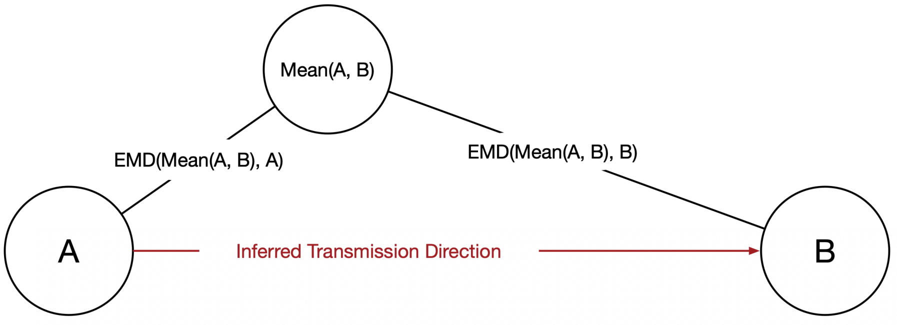

Figure 4: An illustration of inferring transmission direction between host A and host B.

2.5 Computing Host Importance Scores via Graph Convolutional Auto-Encoder

Consider a host network represented by the weighted directed graph $G = (V, E, \mathbf{w})$, where $|V| = n$ is the number of hosts and where weights $\mathbf{w}$ reflect genetic distances between hosts. Let X denote the $n \times n$ matrix of distances between pathogens associated with the hosts, and let A denote the adjacency matrix of G. Motivated by (Salha et al., 2019), we design the graph convolutional encoder (GCE) that learns to construct an $n \times (d+1)$ node embedding matrix $|Z, M|$, where $|\cdot|$ denotes matrix concatenation and the dimensions of Z and M are $n \times d$ and $n \times 1$, respectively. The first $d \ll f$ dimensions of the embedding correspond to the latent feature representation of a host, whereas the last dimension corresponds to a mass parameter $m_i \in R^+$ of the host; note that m_i is reflective of the impact of host i on the graph flow and thus signifies its importance in the spread of the disease. By drawing parallels to the Newton's theory of universal gravitation (Salha et al., 2019), acceleration $a_{i \to j} = \frac{Gm_j}{r^2}$ (where $r^2 = \|z_i - z_j\|_2^2$) can be interpreted as an indicator of the likelihood that the pathogen associated with host i is a genetic ancestor of the pathogen associated with host j. A graph convolutional encoder with L layers, where

$L \geq 2$, and $|Z, M| = H^{(L)}$ can be summarized as

$$\begin{aligned} H^{(0)} &= X \\ H^{(l)} &= \mathrm{ReLU}(D_{out}^{-1}(A+I)H^{(l-1)}W^{(l-1)} + B^{(l-1)}) \\ H^{L} &= D_{out}^{-1}(A+I)H^{(L-1)}W^{(L-1)} + B^{(L-1)}, \end{aligned}$$

where $l \in \{1, 2, ..., L-1\}$, D_{out} denotes the diagonal out-degree matrix of $A+I$, I represents the identity matrix, and W and B are learnable weight and bias matrix, respectively. In summary, a graph convolutional encoder learns node embeddings from A and X as $|Z, M| = \mathrm{GCE}(A, X)$. A graph decoder is then leveraged to reconstruct A from Z and M while allowing for asymmetric connectivity between hosts. The reconstructed adjacency matrix can be represented as $\hat{A}_{ij} = \sigma(\log a_{i \to j}) = \sigma(\log Gm_j - \log \|z_i - z_j\|_2^2)$, where $\sigma(x) = 1/(1+e^{-x})$ denotes the sigmoid function and m_j is the importance score of the j^{th} host; note that, in general, $\hat{A}_{ij} \neq \hat{A}_{ji}$. The loss function in the form of weighted cross entropy is given by $L = -\sum_{i,j} A_{ij} \log \hat{A}_{ij}$. We build a two-layer graph convolutional encoder with layer dimensions set to $\lceil \frac{n}{2} \rceil$ and $\lceil \frac{n}{4} \rceil$. The learning rate is 0.001, and the number of training epoch is 200.

2.6 Reconstructing the Transmission Network

In the final stage of our framework, the weighted directed graph is pruned to obtain the disease transmission network. To this end, we modify the classical Edmond's algorithm (Edmond, 1967) to enable leveraging both the local and global information about the transmission dynamics; the former is provided by the genetic distances while the latter is given by the node importance scores. In particular, given a weighted directed graph $G = (V, E, \mathbf{w}, \mathbf{s})$, where V is the set of vertices representing the hosts, E is the set of directed edges, $\mathbf{w}$ is the set of edge weights, and $\mathbf{s}$ is the set of importance scores of the hosts, we would like to find a directed minimum spanning tree T having the smallest weight and satisfying the constraint that transmission directions are from the hosts with higher importance scores to the hosts with lower importance scores. Let $w(e)$ be the weight on edge e, $w(u, v)$ denote the weight on the edge from u to v, and $s(u)$ be the importance score of host u. A pseudo-code of our modification of the Edmond's algorithm, which for convenience we refer to as EDMOND+SCORE, is given below.

Algorithm [Edmond+Score]:

(1) Select a host with the highest importance score as the root r, and remove all edges whose destination is r.

(2) For each node v except r, keep the edge with the lowest weight incoming to v from $\pi(v)$, where $s(\pi(v)) \geq s(v)$, and remove other edges whose destination is v. If the set of edges $P = \{(\pi(v), v) | v \in V \setminus r\}$ does not contain cycles, the desired directed minimum spanning tree is found. Otherwise, go to step (3).

(3) If there is at least one cycle in P, select one such cycle and denote it by C. Define a new weighted directed graph $G' = (V', E', \mathbf{w}', \mathbf{s})$ and treat C as a single (virtual) node, v_C. For any edge from $u \notin C$ to $v \in C$, add a new edge $e = (u, v_C)$ to E' and add $w'(e) = w(u, v) - w(\pi(v), v)$ to $\mathbf{w}'$, where $\pi(v)$ is the source of v in C. For any edge from $u \in C$ to $v \notin C$, add a new edge $e = (v_C, v)$ to E' and add $w'(e) = w(u, v)$ to $\mathbf{w}'$. For an edge from $u \notin C$ to $v \notin C$, add a new edge $e = (u, v)$ to E' and add $w'(e) = w(u, v)$ to $\mathbf{w}'$.

(4) For all edges incoming to v_C, identify (u, v) with the lowest edge weight and remove $(\pi(v), v)$ from C to break the cycle.

(5) Repeat steps (3) and (4) until all the cycles in G are broken.

Fig. 5 shows an example of pruning a weighted directed graph with the Edmond+Score algorithm.

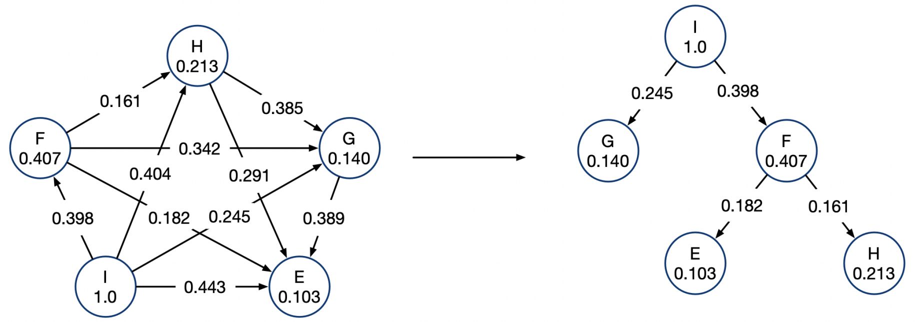

Figure 5: An example of a directed minimum spanning tree obtained by leveraging both local and global information about the transmission dynamics.

3 Results

In this section, we report results of several benchmarking tests that compare the performance of the introduced methodology with state-of-the-art techniques for transmission network inference; for convenience, we refer to the proposed end-to-end framework as AutoNet, as it relies on scores provided by an auto-encoder to infer the network.

3.1 Performance on Semi-Experimental Hepatitis C Virus Data

Performance of AutoNet is first tested on semi-experimental hepatitis C data and compared with state-of-the-art methods for the reconstruction of disease transmission networks from viral genomic data, including k-MED (Melnyk et al., 2020), QUENTIN (Skums et al., 2018), MinDist (Campo et al., 2016) and SOPHIE (Skums et al., 2022). k-MED (Melnyk et al., 2020) is a heuristic for viral outbreak investigations that relies on the earth mover's distance metric; QUENTIN (Skums et al., 2018) is a graph-based method that models viral evolution and epidemic spread; MinDist (Campo et al., 2016) is a heuristic that detects viral transmission based on a pre-defined threshold using minimal Hamming distances; finally, SOPHIE (Skums et al., 2022) is an algorithmic framework that relies on integrated phylogenetic and random graph models to infer viral transmission networks from genomic data. The semi-experimental data is generated using FAVITES (Moshiri et al., 2019), a software for simulating viral transmission and evolution which we applied to experimentally obtained sequencing data of pathogens collected from a cardiac surgeon and five patients infected by the surgeon (Esteban et al., 1996). In particular, pathogens associated with the surgeon and five patients are represented by a sequence of 188 nucleotides encompassing the first hypervariable region (HVR-1) at the junction between the coding regions for envelope glycoproteins E1 and E2. The mean and standard deviation of the Hamming distance between the 188-nucleotide sequences are 11.0 and 3.84, respectively. We follow the steps below to repeatedly synthesize disease transmission networks using the default settings of FAVITES:

1. Generate a complete contact network which consists of 6 nodes representing the cardiac surgeon and five patients.

2. Set the pathogen representing the cardiac surgeon as the seed sequence.

3. Simulate a series of transmission events under the transmission model.

4. Determine the patient sequencing events, update the viral phylogenetic tree, and specify mutation rates.

5. Generate erroneous reads from the sequences on the viral phylogenetic trees,

and use the obtained reads to simulate assembly of those sequences. This creates a tree-like disease transmission network that we aim to reconstruct from the 188 nucleotides-long pathogen sequence information.

The considered methods are compared in terms of the transmission direction accuracy, defined as the ratio of the number of pairs of hosts with correctly predicted transmission direction to the total number of host pairs, and the source identification accuracy, defined as the fraction of the transmission networks with correctly predicted sources. Table 1 shows the results on a semi-experimental dataset containing 5000 disease transmission networks, where the transmission direction accuracy is calculated by aggregating all the possible host pairs in 5000 disease transmission networks. As can be seen, AutoNet achieves both the highest transmission direction accuracy as well as the highest source identification accuracy.

Table 1: The performance comparison of AutoNet and the competing methods on the semi-experimental dataset with 5000 disease transmission networks. Note that MinDist cannot infer transmission directions (hence, N/A).

	Transmission Direction Accuracy	Source Identification Accuracy
AutoNet (this paper)	**93.4%**	**85.2%**
k-MED	89.5%	79.4%
QUENTIN	86.2%	72.5%
MinDist	N/A	45.2%
SOPHIE	90.1%	82.8%

3.2 Performance on Experimental City-Level COVID-19 Data

Next, we compare performance of AutoNet with the selected state-of-the-art methods on experimental city-level COVID-19 data from the Global Initiative

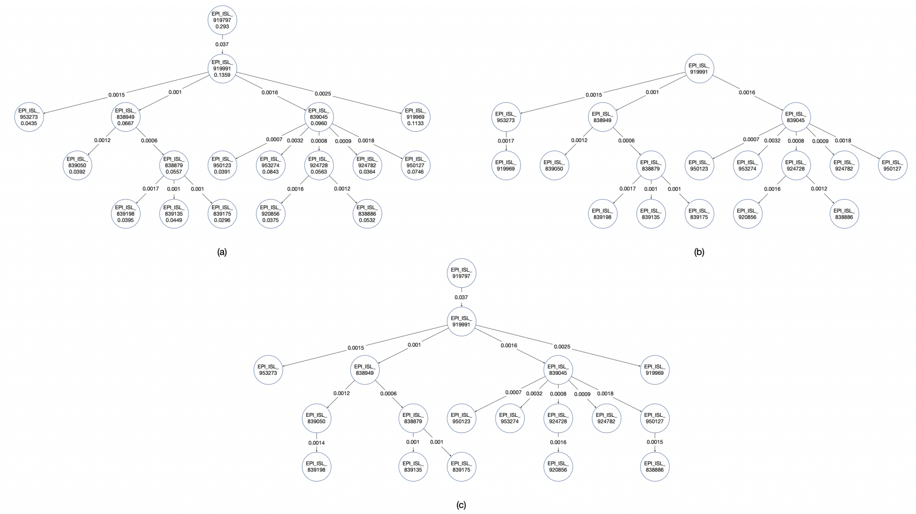

Figure 6: Experiments on city-level COVID-19 data. (a) The disease transmission network reconstructed by AutoNet; note the GISAID accession IDs and the importance scores for each host node. (b) The result of k-MED. (c) The result of SOPHIE.

on Sharing All Influenza Data (GISAID) (Shu and McCauley, 2017). The GISAID promotes rapid sharing of data from all influenza viruses and the coronavirus causing COVID-19. This includes genetic sequences and related clinical and epidemiological data associated with human viruses, as well as geographical and species-specific data associated with the avian and other animal viruses, to help researchers understand how viruses evolve and spread during epidemics and pandemics (https://www.gisaid.org). For each host/patient, the experimental COVID-19 data includes the collection date, submission date, location, gender, patient age, virus variant type, specimen source, sequencing technology, assembly method, sequencing coverage, originating lab, submitting lab and so on. We focus on data collected in London between January 1st and January 31st of 2021, which includes 4252 hosts and 2 virus variant types, $B.1.1.7$ and $B.1.177$. We first test the performance of TenSQR (Ahn et al., 2018) adapted to the task of of detecting transmission clusters using the reported virus variant

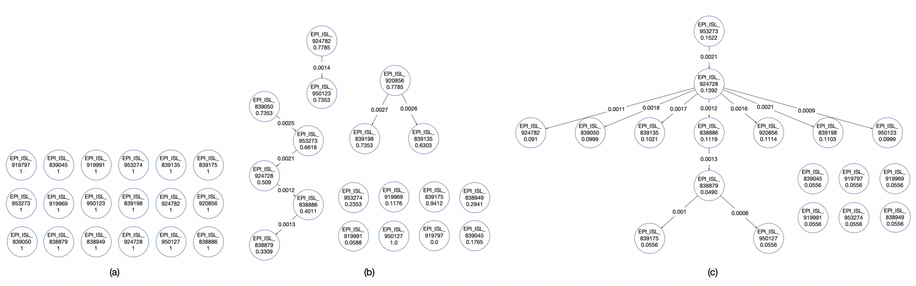

Figure 7: The disease transmission networks reconstructed from city-level COVID-19 data by AutoNet with the importance scores replaced by the standard centrality measures. Note the GISAID accession IDs and centrality measures for each host node. (a) Degree centrality. (b) Closeness centrality. (c) Betweenness centrality.

types as the ground truth, and then report the spanning trees reconstructed by AutoNet and k-MED (Shu and McCauley, 2017). Since the sequences of reported pathogens associated with the 4252 hosts are shorter than the sequence lengths of the COVID-19 reference genome, the host sequences are first aligned to the COVID-19 reference genome (NCBI Reference Sequence: NC_045512.2) using BLAST (McGinnis and Madden, 2004); TenSQR is then deployed to detect transmission clusters. The accuracy, precision and recall of the clustering results achieved by TenSQR are 0.998, 0.995 and 0.984, respectively. The F-1 score and the area under the ROC curve (AUC) are 0.989 and 0.992, respectively. Figures 6 (a), (b) and (c) illustrate the sub-graphs of the disease transmission networks reconstructed by AutoNet, k-MED and SOPHIE, respectively. Note that we only show the results achieved by AutoNet, k-MED and SOPHIE because QUENTIN could not complete the task in 48 hours, while MinDist cannot infer transmission directions. It is worth pointing out that our AutoNet and k-MED reconstruct similar networks. k-MED identifies the host marked with EPI_ISL_919991 as the source, while our AutoNet and SOPHIE identify the host marked with EPI_ISL_919797 as the source and infer that it infected the host marked with EPI_ISL_919991.

Moreover, we conducted an ablation study to evaluate the significance of the novel importance scores proposed within our AutoNet framework. In particular, we systematically replaced these importance scores by the widely-used centrality measures including degree centrality, closeness centrality, and betweenness centrality, while keeping the remainder of our framework unaltered. The findings are visually presented in Fig.7 (a), (b), and (c), which showcase the disease transmission networks reconstructed from the city-level COVID-19 data using AutoNet with varied centrality measures. As shown in Fig.7 (a), when the importance scores are replaced by degree centrality, the transmission network reconstruction fails; this is caused by the failure to satisfy the condition that transmissions happen from nodes with higher importance scores to those with lower importance scores. Similarly, as shown in Fig.7 (b), employing closeness centrality resulted in three fragmented transmission networks and the exclusion of eight hosts. Furthermore, Fig.7 (c) demonstrates that replacing the importance scores with betweenness centrality leads to identifying EPI_ISL_953273 as the source but excludes six hosts in the network. These empirical results highlight the pivotal role played by our proposed importance scores in capturing the intricate dynamics of disease transmission.

3.3 Performance on Experimental Country-Level COVID-19 Data

Next, we compare performance of AutoNet and the considered state-of-the-art methods on experimental country-level COVID-19 data aiming to discover how the pathogens are transmitted between different cities. In addition to London's, COVID-19 data collected in Alderley Edge, Milton Keynes, Cambridge, Glasgow, Oxford and Edinburgh between January 1st and January 31st of 2021

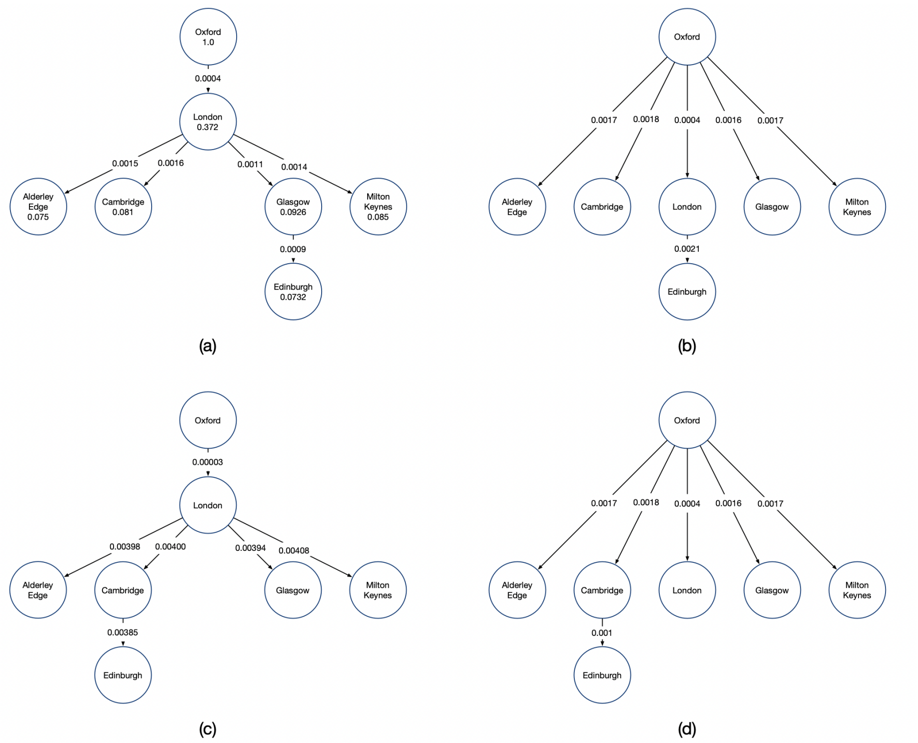

Figure 8: Experiments on country-level COVID-19 data. (a) The disease transmission network reconstructed by AutoNet. (b) The result of k-MED. (c) The result of QUENTIN. (d) The result of SOPHIE.

are also analyzed. The number of hosts present in the data for these additional cities is 16433, 9977, 7494, 4644, 1085 and 154, respectively. We rely on the consensus sequence of the dominant viral variant type to represent each city, and then perform the reconstruction task. The host sequences are aligned to the COVID-19 reference genome (NCBI Reference Sequence: NC_045512.2) before obtaining the consensus sequence for each city. Fig. 8 (a), (b), (c) and (d) show the reconstruction results obtained by AutoNet, k-MED, QUENTIN and SOPHIE, respectively. It is worth pointing out that our AutoNet and QUENTIN reconstruct similar networks, and that all the methods identify Oxford as the source of the spread, while our AutoNet and QUENTIN both identify London

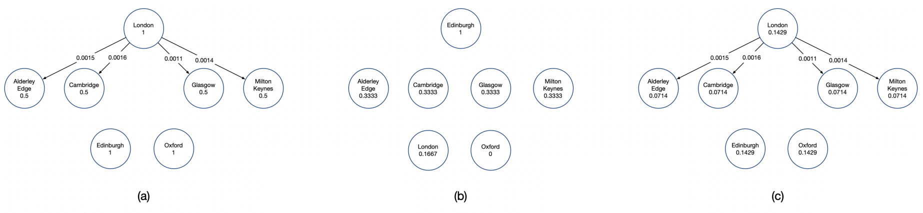

Figure 9: The disease transmission networks reconstructed from country-level COVID-19 data by AutoNet with the importance scores replaced by the standard centrality measures. Note the cities and centrality measures for each node. (a) Degree centrality. (b) Closeness centrality. (c) Betweenness centrality.

as the second source.

Finally, we conducted another ablation study to assess the significance of our proposed importance scores. We replaced these scores with standard centralities, including degree centrality, closeness centrality, and betweenness centrality, while keeping the rest of our proposed framework unaltered. Our study findings are reported in Fig. 9 (a), (b), and (c), depicting the disease transmission networks reconstructed from country-level COVID-19 data using AutoNet with different centrality measures. As seen in Fig. 9 (a) and (b), the disease transmission networks reconstructed based on degree centrality and betweeness centrality are identical, with London identified as the source of the spread to Alderley Edge, Cambridge, Glasgow, and Milton Keynes. However, Edinburgh and Oxford are not included in the network as AutoNet only allows transmissions from the nodes with higher scores to nodes with lower scores. Moreover, Fig. 9 (c) demonstrates that replacing the importance scores with closeness centrality leads to a network reconstruction failure. This further emphasizes the critical role of our proposed importance scores in enabling AutoNet to accurately capture disease transmission dynamics.

4 Conclusions

In this paper we presented AutoNet, an end-to-end framework for the detection of disease transmission clusters, reconstruction of directed transmission networks, and the discovery of super-spreaders from genomic data. The framework first clusters infected viral hosts into groups where the hosts in a group are infected by the same pathogen variant. After quantifying similarity between pairs of viral hosts in a group, directions of transmission between hosts are estimated and the importance score of each host is calculated. Finally, a directed minimum spanning tree is reconstructed by leveraging both the local and global information about transmissions, provided by the genetic similarity between hosts and the hosts' importance scores calculated via a graph autoencoder. Benchmarking on semi-experimental and experimental data shows that the proposed framework is capable of reconstructing disease transmission networks efficiently and accurately, outperforming state-of-the-art competing techniques. Future work includes extending the proposed framework to enable joint processing of both genomic and epidemiological data.

Authorship Contribution Statement

Algorithms and experiments were designed by ZK and HV. Algorithm code was implemented and tested by ZK. The manuscript was written by ZK and HV. Both authors read and approved the final manuscript.

Author Disclosure Statement

No competing financial interests exist.

Funding Statement

This work was funded in part by the NSF grant CCF-2109983.

References

Ahn, S., Ke, Z., Vikalo, H., et al. Viral quasispecies reconstruction via tensor factorization with successive read removal. *Bioinformatics*, 34(13):i23–i31, 2018.

Alamil, M., Hughes, J., Berthier, K., et al. Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1776):20180258, 2019.

Aldrin, M., Lyngstad, T., Kristoffersen, A., et al. Modelling the spread of infectious salmon anaemia among salmon farms based on seaway distances between farms and genetic relationships between infectious salmon anaemia virus isolates. *Journal of the Royal Society Interface*, 8(62):1346–1356, 09 2011. doi: 10.1098/rsif.2010.0737. URL https://pubmed.ncbi.nlm.nih.gov/21325314.

Campbell, F., Cori, A., Ferguson, N., et al. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol*, 15(3):e1006930, 2019.

Campo, D., Xia, G., Dimitrova, Z., et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *The Journal of Infectious Diseases*, 213(6):957–965, 5/15/2021 2016. doi: 10.1093/infdis/jiv542. URL https://doi.org/10.1093/infdis/jiv542.

Cauchemez, S., Bhattarai, A., Marchbanks, T., et al. Role of social networks in shaping disease transmission during a community outbreak of 2009 h1n1 pandemic influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):2825–2830, 2011.

Cook, V., Sun, S., Tapia, J., et al. Transmission network analysis in tuberculosis contact investigations. *The Journal of Infectious Diseases*, 196(10):1517–1527, 2007.

Cottam, E., Thébaud, G., Wadsworth, J., et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings. Biological Sciences*, 275(1637):887–895, 2008.

Craft, M. E. Infectious disease transmission and contact networks in wildlife and livestock. *Philos Trans R Soc Lond B Biol Sci*, 370(1669):20140107, 2015.

De Maio, N., Wu, C., Wilson, D., et al. Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology*, 12(9):e1005130, 2016.

De Maio, N., Worby, C., Wilson, D., et al. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput Biol*, 14(4): e1006117, 2018.

Didelot, X., Gardy, J., and Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, 31(7):1869–1879, 5/15/2021 2014. doi: 10.1093/molbev/msu121. URL https://doi.org/10.1093/molbev/msu121.

Edmond, J. Optimum branchings. *Journal of Research of the National Bureau of Standards Section B*, 71B(4):233–240, 1967.

Esteban, J., Gómez, J., Martell, M., et al. Transmission of hepatitis c virus by a cardiac surgeon. *The New England Journal of Medicine*, 334(9):555–560, 1996.

Fisher, M., Pao, D., Brown, A., et al. Determinants of hiv-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *AIDS*, 24(11):1739–1747, 2010. URL `https://journals.lww.com/aidsonline/Fulltext/2010/07170/Determinants_of_HIV_1_transmission_in_men_who_have.15.aspx`.

Glebova, O., Knyazev, S., Melnyk, A., et al. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC genomics*, 18(Suppl 10):918, 12 2017. doi: 10.1186/s12864-017-4274-5. URL `https://pubmed.ncbi.nlm.nih.gov/29244009`.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Grassly, N. and Fraser, C. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487, 2008.

Gundlapalli, A., Ma, X., Benuzillo, J., et al. Social network analyses of patient-healthcare worker interactions: implications for disease transmission. *Annual Symposium proceedings / AMIA Symposium*, pages 213–217, 2009.

Hall, M., Woolhouse, M., Rambaut, A., et al. Using genomics data to reconstruct transmission trees during disease outbreaks. *Revue scientifique et technique (International Office of Epizootics)*, 35(1):287–296, 2016. doi: 10.20506/rst.35.1.2433. URL `http://europepmc.org/abstract/MED/27217184`.

Hao, X., Cheng, S., Wu, D., et al. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584(7821):420–424, 2020. doi: 10.1038/s41586-020-2554-8. URL https://doi.org/10.1038/s41586-020-2554-8.

Hayama, Y., Firestone, S., Stevenson, M., et al. Reconstructing a transmission network and identifying risk factors of secondary transmissions in the 2010 foot-and-mouth disease outbreak in japan. *Transboundary and emerging diseases*, 66(5):2074–2086, 2019.

Hens, N., Calatayud, L., Kurkela, S., et al. Robust reconstruction and analysis of outbreak data: Influenza a(h1n1)v transmission in a school-based population. *American journal of epidemiology*, 176(3):196–203, 2012.

Jombart, T., Eggo, R., Dodd, P., et al. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, 2011. doi: 10.1038/hdy.2010.78. URL https://doi.org/10.1038/hdy.2010.78.

Jombart, T., Cori, A., Didelot, X., et al. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol*, 10(1):e1003457, 2014.

Lau, M., Marion, G., Streftaris, G., et al. A systematic bayesian integration of epidemiological and genetic data. *PLoS Comput Biol*, 11(11):e1004633, 2015.

Luo, C., Ma, Y., Jiang, P., et al. The construction and visualization of the transmission networks for covid-19: A potential solution for contact tracing and assessments of epidemics. *Scientific Reports*, 11:8605, 2021. doi: 10.1038/s41598-021-87802-x. URL https://doi.org/10.1038/s41598-021-87802-x.

Magori-Cohen, R., Louzoun, Y., Herziger, Y., et al. Mathematical modelling and evaluation of the different routes of transmission of lumpy skin disease virus. *Veterinary Research*, 43(1):1, 2012. doi: 10.1186/1297-9716-43-1. URL `https://doi.org/10.1186/1297-9716-43-1`.

McGinnis, S. and Madden, T. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, 32(Web Server issue): W20–W25, 07 2004. doi: 10.1093/nar/gkh435. URL `https://pubmed.ncbi.nlm.nih.gov/15215342`.

Melnyk, A., Knyazev, S., Vannberg, F., et al. Using earth mover's distance for viral outbreak investigations. *BMC Genomics*, 21(5):582, 2020. doi: 10.1186/s12864-020-06982-4. URL `https://doi.org/10.1186/s12864-020-06982-4`.

Montazeri, H., Little, S., Mozaffarilegha, M., et al. Bayesian reconstruction of transmission trees from genetic sequences and uncertain infection times. *Stat Appl Genet Mol Biol.*, 2018.

Morelli, M., Thébaud, G., Chadœuf, J., et al. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol*, 8(11):e1002768, 2012.

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J., et al. Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35(11):1852–1861, 2019.

Otero, M. and Solari, H. Stochastic eco-epidemiological model of dengue disease transmission by aedes aegypti mosquito. *Mathematical Biosciences*, 223(1): 32–46, 2010.

Rubner, Y., Tomasi, C., and Guibas, L. A metric for distributions with applications to image databases. In *Sixth International Conference on*

Computer Vision (IEEE Cat. No.98CH36271), pages 59–66, 1998. doi: 10.1109/ICCV.1998.710701.

Rubner, Y., Tomasi, C., and Guibas, L. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000. doi: 10.1023/A:1026543900054. URL https://doi.org/10.1023/A:1026543900054.

Salathe, M., Kazandjieva, M., Lee, J., et al. A high-resolution human contact network for infectious disease transmission. *Proc Natl Acad Sci*, 107(51): 22020–22025, 2010.

Salha, G., Limnios, S., Hennequin, R., et al. Gravity-inspired graph autoencoders for directed link prediction. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 589–598, 2019.

Salipante, S. J. and Hall, B. G. Inadequacies of minimum spanning trees in molecular epidemiology. *Journal of Clinical Microbiology*, 49(10):3568–3575, 2011.

Shu, Y. and McCauley, J. Gisaid: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*, 22(13):30494, 03 2017. doi: 10.2807/1560-7917.ES.2017.22.13.30494. URL https://pubmed.ncbi.nlm.nih.gov/28382917.

Skums, P., Zelikovsky, A., Singh, R., et al. Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1): 163–170, 5/15/2021 2018. doi: 10.1093/bioinformatics/btx402. URL https://doi.org/10.1093/bioinformatics/btx402.

Skums, P., Mohebbi, F., Tsyvina, V., et al. Sophie: Viral outbreak investigation and transmission history reconstruction in a joint phylogenetic and network theory framework. *Cell Systems*, 13(10):844–856, 2022.

Spada, E., Sagliocca, L., Sourdis, J., et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis c virus infection. *Journal of clinical microbiology*, 42(9):4230–4236, 09 2004. doi: 10.1128/JCM.42.9.4230-4236.2004. URL https://pubmed.ncbi.nlm.nih.gov/15365016.

Wohl, S., Metsky, H., Schaffner, S., et al. Combining genomics and epidemiology to track mumps virus transmission in the united states. *PLoS Biol*, 18(2): e3000611, 2020.

Worby, C., Lipsitch, M., Hanage, W., et al. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS computational biology*, 10(3):e1003549, 2014.

Worby, C., O'Neill, P., Kypraios, T., et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*, 10(1):395–417, 03 2016. doi: 10.1214/15-aoas898. URL https://pubmed.ncbi.nlm.nih.gov/27042253.

Worby, C., Lipsitch, M., Hanage, W., et al. Shared genomic variants: Identification of transmission routes using pathogen deep-sequence data. *American Journal of Epidemiology*, 186(10):1209–1216, 5/15/2021 2017. doi: 10.1093/aje/kwx182. URL https://doi.org/10.1093/aje/kwx182.

Xu, X., Liu, X., Wu, Y., et al. Reconstruction of transmission pairs for novel coronavirus disease 2019 (covid-19) in mainland china: Estimation of superspreading events, serial interval, and hazard of infection. *Clinical infectious diseases*, 71(12):3163–3167, 5/15/2021 2020. doi: 10.1093/cid/ciaa790. URL https://doi.org/10.1093/cid/ciaa790.

Ypma, R., Bataille, A., Stegeman, A., et al. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, 279(1728):444–450, 2012.

Ypma, R., van Ballegooijen, W., Wallinga, J., et al. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3): 1055–1062, 2013.

Zhang, X. and De Angelis, D. Construction of the influenza a virus transmission tree in a college-based population: co-transmission and interactions between influenza a viruses. *BMC Infectious Diseases*, 16(1):38, 2016. doi: 10.1186/s12879-016-1373-x. URL `https://doi.org/10.1186/s12879-016-1373-x`.