

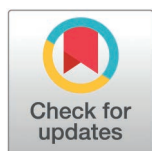
RESEARCH ARTICLE

Predicting missing links in COVID-19 infection networks

Pavlos Alexandros Dimitriou^{1,2*}, Valentinos Silvestros³, Elisavet Constantinou³, Costas Pitris^{1,2}, Panayiotis Kolios^{1,4}

1 University of Cyprus, Department of Electrical and Computer Engineering, Nicosia, Cyprus **2** KIOS Research and Innovation Center of Excellence, Nicosia, Cyprus, **3** Cyprus Ministry of Health, Nicosia, Cyprus, **4** University of Cyprus, Department of Computer Science, Nicosia, Cyprus

* pdimit05@ucy.ac.cy



Abstract

Real-world infection networks are often incomplete since contact tracing cannot capture all infection events. As a result, many infected individuals appear as isolated cases with no recorded epidemiological links. Link prediction methods can be used to reconstruct these missing links, allowing for a more complete representation of the infection network. In this study, two approaches were evaluated for predicting missing links during the first four waves of COVID-19 in Cyprus. The first approach relied on classical machine learning classifiers trained on engineered edge-level features. For each pair of cases, node-level epidemiological attributes, such as age, gender, infection date, NACE, and residency, were combined using various operators to form an edge feature vector. Feature selection was performed using Minimum Redundancy Maximum Relevance (mRMR) followed by greedy forward feature selection, while permutation importance was used to assess the contribution of the selected features. Random Forest and Gradient Boosting consistently achieved the strongest performance, with F1-scores ranging from approximately ~0.82 to ~0.90 and Mean Reciprocal Rank (MRR) values between 0.55 and 0.64 for the first two waves. For the third and fourth waves, F1-scores ranged from 0.68 to 0.75 and MRR values from 0.23 to 0.33. The second approach employed graph representation learning, using a GraphSAGE model. Node embeddings were generated from both the node attributes and the observed network topology, and several embedding-combination strategies were evaluated including concatenation, absolute difference, squared difference, Hadamard product, and dot product. Link prediction with graph representation learning achieved F1-scores ranging from ~0.70 to 0.79 and MRR values from 0.23 to 0.42 across all pandemic waves. After the best classifier was validated on known cases, it was applied to unlinked nodes to infer their most likely infectors, resulting in reconstructed networks with fewer isolated components and larger connected structures. The results confirmed plausible extension of connections while maintaining epidemiologically relevant characteristics, such as the outdegree. These results demonstrate that

OPEN ACCESS

Citation: Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P (2026) Predicting missing links in COVID-19 infection networks. PLOS Complex Syst 3(8): e0000124. <https://doi.org/10.1371/journal.pcsy.0000124>

Editor: Hocine Cherifi, Université de Bourgogne: Université de Bourgogne, FRANCE

Received: February 26, 2026

Accepted: August 3, 2026

Published: August 28, 2026

Copyright: © 2026 Dimitriou et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The source code used in this study, together with an example dataset is publicly available on Zenodo DOI: <https://doi.org/10.5281/zenodo.21619477>.

Funding: P.A.D., C.P. and P.K. have been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under Grant

Agreement No. 739551 (<https://cordis.europa.eu/project/id/739551>). P.A.D., C.P., P.K., V.S., E.C. have been funded by the CIPHIS project (Cyprus Innovative Public Health ICT System), C1.112, of the NextGenerationEU programme under the Republic of Cyprus Recovery and Resilience Plan (https://commission.europa.eu/projects/cyprus-innovative-public-health-in-formation-and-communications-technology-ict-system-ciphis_en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

machine learning and graph representation learning can identify missing infection links, thus identifying likely sources of infection. Such tools can assist epidemiologists by providing a more complete picture of the spread of the disease, particularly during small epidemic waves, to allow for more informed and targeted interventions.

Author summary

Understanding how infectious diseases spread requires accurate contact information. However, real-world contact tracing is often incomplete, leaving many cases without any recorded links. These missing connections limit the ability of public health teams to understand transmission patterns and to design effective interventions. This study explores how network analysis, machine learning, and graph-based methods can help fill this gap by predicting likely infection links between infected individuals. Two approaches were evaluated. The first employed traditional machine learning models trained on the epidemiological characteristics of each pair of cases, such as age, infection date, workplace category, and geographic proximity. The second approach used graph representation learning, where a GraphSAGE neural network was trained on patterns directly from the structure of the observed infection network. Both methods were evaluated across multiple pandemic waves. The results suggest that the proposed methods are capable of identifying plausible missing links. Reconstructed networks contained fewer isolated cases and revealed possible infection pathways while maintaining epidemiologically important network characteristics, such as outdegree. These findings demonstrate that data-driven link prediction can support public health decision-making by providing a more complete picture of how infections spread, even when contact tracing is incomplete.

1. Introduction

During the surge of the COVID-19 pandemic, one of the major challenges faced by public health authorities was on one hand to effectively collect contact tracing data but also analyze the large volume of data collected during such investigations. In Cyprus, the third largest island in the Mediterranean, with a population of approximately one million, COVID-19 surveillance during the first two pandemic waves relied heavily on manual contact tracing procedures. Epidemiologists conducted telephone interviews with infected individuals and manually recorded exposure histories and symptoms in simple datasheets. Subsequently, contact tracing officers searched for contact information and compared multiple lists across different computer systems in order to match infected cases with their contacts and determine the direction of transmission. As case numbers rapidly increased, these manual comparisons made it increasingly difficult to identify specific individuals, trace their contacts, and cross-reference relevant information in a timely manner. Similar limitations of contact

tracing systems were also reported in Germany [1]. In general, manual contact tracing is time-consuming, resource-intensive, and less effective, especially in situations where the infection rates were rapidly escalating [2,3].

As the pandemic progressed, electronic contact tracing platforms were introduced to facilitate case investigation and data management. These systems significantly improved data availability and standardization. However, they also revealed a critical limitation: a substantial proportion of confirmed cases were unable to identify their source of infection, either because (i) they could not recall the individuals they had met, (ii) or chose not to disclose all their contacts, (iii) or had been infected by individuals unknown to them. Such cases, are referred to in the literature as unlinked, orphan, isolated, or sporadic cases. They hamper the reconstruction of infection networks, obscure the identification of super-spreading events, and limit the effectiveness of targeted interventions, while also delaying contact tracing investigations. The effectiveness of contact tracing in controlling COVID-19 transmission has been shown to critically depend on the speed with which cases were identified and their contacts isolated. Even short delays substantially reduced the likelihood of interrupting the transmission [2,4]. These findings underscore that rapid case detection, minimal delays, and sustained tracing capacity are essential prerequisites for maintaining epidemic control during the early stages of COVID-19 outbreaks. To assist healthcare administrators in improving contact tracing activities network visualization tools were developed [5,6], helping to prioritize contacts for evaluation [7]. Beyond visualization, network-based approaches enable the identification of transmission patterns across age groups, occupations, and geographic districts [8,9], facilitate the detection of superspreading individuals, and support the estimation of key epidemiological parameters such as the basic reproduction number.

Most existing studies of infection networks exclude the unlinked cases from their analysis, as these cases appear as isolated nodes. However, unlinked cases have emerged as a critical indicator of declining contact tracing performance [10]. Evidence from Korea and Japan showed that increases in the proportion of unlinked cases were associated with increases in confirmed cases, and that sustained upward trends in unlinked cases could serve as an early warning signal indicating the need for expanded testing and contact tracing efforts [10]. In another study, researchers categorized traced and untraced infections, providing a real-time indicator of transmission potential that accounts for significant missing data within the infection tree [11]. Complementary analyses from Hong Kong [12], further revealed that unlinked cases were typically detected later, were associated with more severe clinical outcomes, and correlated strongly with accelerated epidemic growth, reflecting their role in sustaining hidden transmission chains. These unlinked infections frequently seed clusters in households, eateries, and workplaces, highlighting the need to prevent silent spread in high-risk social environments. In addition, it was shown that linked cases had shorter interval from onset to hospital than unlinked cases [13]. A recent study has investigated the impact of missing information in contact tracing using stochastic agent-based models [14]. The study showed that missing information could significantly alter transmission structure, increase network diameter, and reduce the effectiveness of contact tracing [14].

Link prediction, the task of estimating the probability that an edge exists between two nodes of the network, is currently being investigated as a possible solution to the extensive lack of accurate contact tracing. Link prediction is extensively used in various domains [15,16], including social and information networks [17,18], to predict potential friends [15], and in recommender systems, to suggest new products or services to the customers [19]. In biological networks [20], link prediction can be utilized to identify potential protein-protein interactions or gene regulatory relationships. There are two main approaches to performing link prediction: (a) similarity-based methods and (b) learning-based methods. A similarity-based approach computes a similarity score for every potential pair of nodes based on their structural proximity within the network. These scores are then ranked in descending order, with pairs receiving higher scores being considered more likely to form links. In contrast, a learning-based approach treats link prediction as a classification problem, assigning a score or probability to each pair of unconnected nodes that reflects the likelihood of a link existing between them. In general, link prediction can be viewed as a binary classification task, where each node pair is categorized as either a positive (existing) or negative (non-existing) link [21]. Therefore, typical machine learning models can be employed to address this problem.

Despite its success in other domains, the application of link prediction methods to COVID-19 infection networks remains limited. In one study, researchers applied link prediction techniques to infer missing connections between disconnected clusters. They trained machine learning classifiers on simple features, including the number of days between reported infections, the age difference between individuals, and a binary indicator of whether they resided in the same housing block [1]. Graph neural networks were also used to predict whether a COVID-19–infected individual would cause additional future infections. The model incorporated interaction-related information, such as contact order, contact duration, movement routes, and individual-level characteristics including symptom status [22]. Similarity-based link predictions methods were applied to COVID-19 contact network to reconstruct the contact network [23]. Transmission networks were constructed so that nodes represented either locations or combinations of locations and age groups. Link prediction was then performed to estimate which pairs of nodes (locations or age-group–location combinations) were likely to become connected through new infections in the near future [24]. Another approach used maximum likelihood to reconstruct transmission chains between index cases and their close contacts who were confirmed positive [25].

The limited adoption of link prediction methods in infectious disease networks is partly attributable to the structural properties of these networks. Empirically constructed infection networks typically consist of numerous small, fragmented components rather than a single large, well-connected graph [8,26]. Such structural characteristics hinder the effectiveness of traditional similarity-based link prediction approaches, which rely primarily on network topology and often fail when applied to sparse or disconnected graphs without incorporating epidemiological attributes of individuals [27]. Another major barrier is related to data availability and quality. Individual-level infection data are difficult to obtain, require substantial effort to curate, and frequently contain missing or incomplete information. As a result, many existing studies exclude unlinked cases from their analyses, further fragmenting the observed transmission network and potentially biasing epidemiological inferences.

The aim of this study was to develop a link prediction framework to deduce the most likely sources of infection for unlinked cases, using available demographic and temporal information. By reconstructing plausible infection pathways for unlinked cases, this work contributes to a more comprehensive understanding of COVID-19 transmission dynamics and highlights the potential of network-based machine learning methods to support epidemic response and preparedness in future outbreaks. Improving the completeness of infection networks and developing of automated, data driven, tools, can enable more effective and efficient epidemiological investigations. The overall efficiency and timeliness of contact tracing activities, particularly during periods of high case incidence, when manual procedures become unsustainable, can be improved, leading both to reduced workload and, more importantly, to more timely and precise interventions.

2. Methods

2.1. Ethics statement

All data analysis was performed in accordance to relevant national and EU guidelines and regulations. This work was approved by the Cyprus National Bioethics Committee (Approval number: CNBC 2023.01.146).

2.2. Data

Infection data from 32,244 confirmed COVID-19 cases recorded in Cyprus, between March 2020 and May 2021, were used to construct infection networks. This period covers the first four pandemic waves. Among these cases, 11,191 individuals had no documented source of infection and were therefore classified as unlinked cases. The data were collected by the Ministry of Health of the Republic of Cyprus as part of the mandated contact tracing investigations during the COVID-19 pandemic. During the first two waves, contact tracing was conducted primarily through telephone interviews. From the third wave onward, investigations were conducted using an electronic reporting system. Identified contacts were subsequently tested, and those who tested positive were epidemiologically linked to the confirmed cases. For this study,

Table 1. Dates and case counts for each epidemic wave.

	Period	Linked cases	Unlinked cases
Wave 1	18/03/2020-17/04/2020	451	222
Wave 2	12/07/2020-14/08/2020	202	68
Wave 3	13/12/2020-12/01/2021	7308	4702
Wave 4	04/04/2021-04/05/2021	13092	6199

<https://doi.org/10.1371/journal.pcsy.0000124.t001>

the Ministry of Health provided anonymized infection records with demographic information including age, gender, occupation, test result day, and district of residence for each infected individual, for 30 days around the peak of each wave. The dates and the number of linked and unlinked cases for each wave are summarized in [Table 1](#). Occupational information was also provided using the Statistical Classification of Economic Activities in the European Community (NACE) coding system.

2.3. Graph construction and preprocessing

For each pandemic wave, the infection network was constructed as a directed graph $G = (V, E)$ where nodes V corresponded to confirmed COVID-19 cases. Directed edges $(u, v) \in E$ indicated that case u infected case v . Each node was associated with a set of attributes, including the date of positive testing, age group, gender, economic activity (identified according to the NACE code), and residential postcode. To prepare the data for the supervised link prediction task, the observed infection network was randomly split into training and test sets. Specifically, 15% of the observed edges were randomly removed and treated as positive test examples for testing, while the remaining edges constituted the training graph. Negative samples were generated by randomly sampling node pairs with no observed infection link based on temporal and spatial constraints. These constraints included a time difference of no more than 15 days between cases, and a maximum geographical distance of 50 km. They were applied to assure that the models were trained on epidemiologically plausible data. The number of negative samples was set 50 times the number of positive samples to better represent the class imbalance in the classification task for Method 1.

2.4. Prediction method 1: Feature-based supervised learning for link prediction

Extending prior work, two distinct link prediction frameworks were implemented and evaluated in this study [28]. The first method was based on feature engineering and machine learning. Initially, node-level attributes were extracted from the infection graphs. These attributes included demographic, temporal, and network characteristics of infected individuals. For each possible pair of nodes, edge-level feature vectors were then constructed using pairwise transformations of node features (e.g., differences, absolute differences, product, and other relational descriptors) to generate a rich set of pairwise features. To reduce redundancy and improve predictive performance, features were ranked using the minimum Redundancy Maximum Relevance (mRMR) algorithm [29]. The highest-ranked features were used to train various supervised machine learning classifiers. Following model training, permutation feature importance analysis [30] was conducted to assess the contribution of each feature to the predictive performance. [Fig 1](#) is a block diagram illustrating the link prediction framework.

2.4.1. Edge feature engineering. To capture the complex relationships between pairs of cases, a set of engineered edge features was constructed for each candidate node pair. Two categories of features were considered: i) generic pairwise transformations (describing mathematical relationships between node attributes), including the difference, absolute difference, squared difference, product, and raw values of the corresponding variables for both nodes; ii) Domain-specific epidemiological features (elucidating epidemiological and localization relationships), including geographical distance between residential postcodes (computed using the Haversine formula), municipality indicator

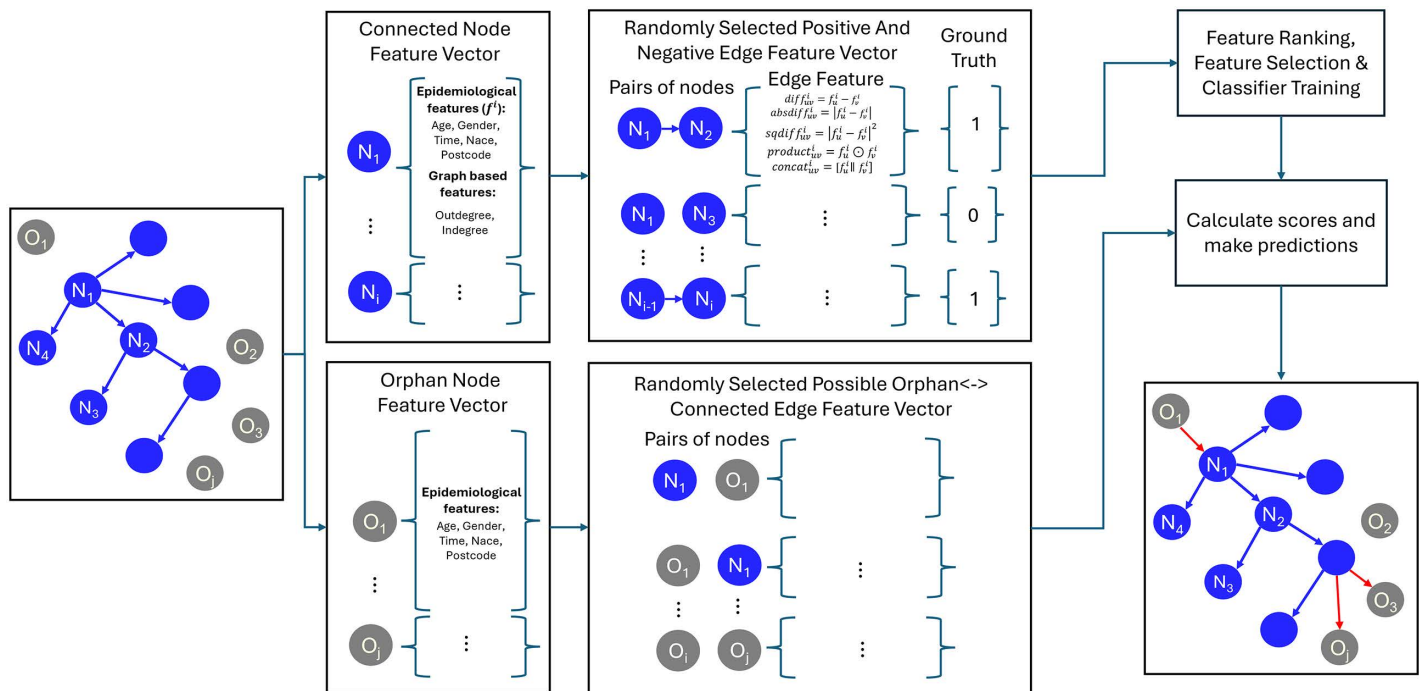


Fig 1. Workflow for predicting missing links in COVID-19 infection networks using feature engineering and machine learning.

<https://doi.org/10.1371/journal.pcsy.0000124.g001>

denoting whether both nodes resided in the same municipality, age-group difference, and NACE-section difference [8]. The complete feature construction scheme is provided in [S1 Table](#) of the Supplementary Materials. In total, 52 edge features were generated.

2.4.2. Feature ranking and selection. To reduce redundancy and improve model generalization, a two-stage feature selection strategy was applied. An initial ranking of candidate edge features was performed using the minimum Redundancy Maximum Relevance (mRMR) algorithm [29], which selects features that are highly informative with respect to the target variable, while minimizing mutual redundancy. Features were ordered according to their mRMR importance score. Features with very low scores (< 0.01) were considered to have negligible contribution and were therefore excluded, resulting in a more compact and stable feature set. Subsequently, a greedy forward selection procedure was applied. To reduce the substantial computational cost of performing greedy feature selection for every classifier, Gradient Boost was used as the classifier for the greedy selection. Features were added sequentially to the model only if they improved the cross-validated Mean Reciprocal Rank (MRR)-score during training. Because model training was repeated with multiple random seeds, each producing different train/test splits, the mRMR ranking and the final feature sets, after the greedy selection procedure, differed across seeds.

During experimentation, fluctuations in performance metrics (MRR, F1-score) was observed, suggesting that some mRMR-selected features may not capture stable transmission patterns. By removing indegree and outdegree features the performance was improved, thereby the degree features were excluded. To identify the optimal feature set for each epidemic wave, all models were subsequently retrained using each selected feature set, and the feature set yielding the highest MRR-score was selected for the final analysis. Finally, permutation feature importance was employed to quantify the contribution of each selected feature to the final predictions. This method measures the decrease in predictive performance when individual features are randomly permuted, thereby quantifying their true contribution to model predictions.

2.4.3. Classifier training and cross-validation. The predictive performance of multiple supervised classifiers was evaluated. The evaluated models included Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Logistic Regression (LR), LightGBMRanker (LGBMRanker) and Gradient Boosting (GB). For each classifier–feature set combination, 5-fold cross-validation was performed on the training edges. Model performance was assessed using Mean Reciprocal Rank (MRR), F1-score (Top-15), top k-accuracy and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). For each epidemic wave, the final classifier–feature set combination was selected based on the highest mean MRR across the cross-validation folds. An F1-score adapted for the source-ranking task was computed from the ranked candidate lists. A true transmission link was counted as correctly recovered if it appeared within the top $k = 15$ ranked candidates and exceeded the probability threshold. Links receiving scores above the threshold but ranked below the top K were treated as false positives, whereas links below the probability threshold were considered false negatives. The F1-score balances precision and recall, which is particularly important for imbalanced datasets, while the ROC AUC summarizes the model's ability to discriminate true links from non-links. Mean Reciprocal Rank (MRR) evaluates how highly the true infection source is ranked among all candidate sources, whereas top-k accuracy measures the proportion of cases for which the true source appears within the top-k ranked predictions.

To evaluate robustness, the entire procedure was repeated across multiple random seeds, with each seed resulting in a different train–test split and, consequently, a slightly different feature subset after the greedy mRMR-based selection. For each seed, cross-validation MRR-scores were averaged across folds, allowing the identification of the best-performing classifier independently of the specific feature subset selected for that seed. Once the classifier with the highest mean MRR-score across seeds was determined, the feature subset generated by the greedy selection procedure for that specific classifier was used. This process ensured that the final model has the most consistently predictive. Performance metrics are reported as mean $\pm$ standard deviation across seeds.

To ensure a fair comparison among classifiers, hyperparameter tuning was performed for each model via grid search within the cross-validation loop. Optimized parameters included the number of trees for Random Forest, maximum depth for Decision Tree, regularization strength for Logistic Regression and SVM, number of neighbors for K-Nearest Neighbors, and learning rate for Gradient Boosting. The complete list of hyperparameter values explored for each classifier is provided in [S2 Table](#) in the Supplementary Materials.

2.5. Prediction Method 2: Graph neural network approach for link prediction

The second approach employed a graph neural network–based link prediction model. A two-layer GraphSAGE encoder was used to learn node embeddings by aggregating information from local network neighborhoods, allowing each representation to capture both individual attributes and surrounding transmission structure. Node features included one-hot encoded categorical variables (sex, economic activity sector, and municipality) and standardized numerical variables (age and infection time index). The learned embeddings were then combined pairwise and passed to a multilayer perceptron (MLP) decoder to estimate the probability of a transmission link between two cases. Multiple embedding-combination operators were evaluated (e.g., concatenation, difference-based, and element-wise operations). In addition, pairwise epidemiological edge features, including temporal difference, geographical distance, shared postcode, shared municipality, and shared economic activity sector, were incorporated into the decoder. This workflow is shown in the [Fig 2](#). The model was trained in a supervised setting using documented transmission links as positive examples and epidemiologically plausible hard negative examples satisfying temporal and geographic constraints. Training and evaluation followed a strict inductive node split, in which the evaluated test transmission edges were excluded from the message-passing graph to prevent information leakage. Early stopping and validation-based threshold optimization were applied. During evaluation, each target case was ranked against all epidemiologically plausible candidate infectors, and performance was assessed.

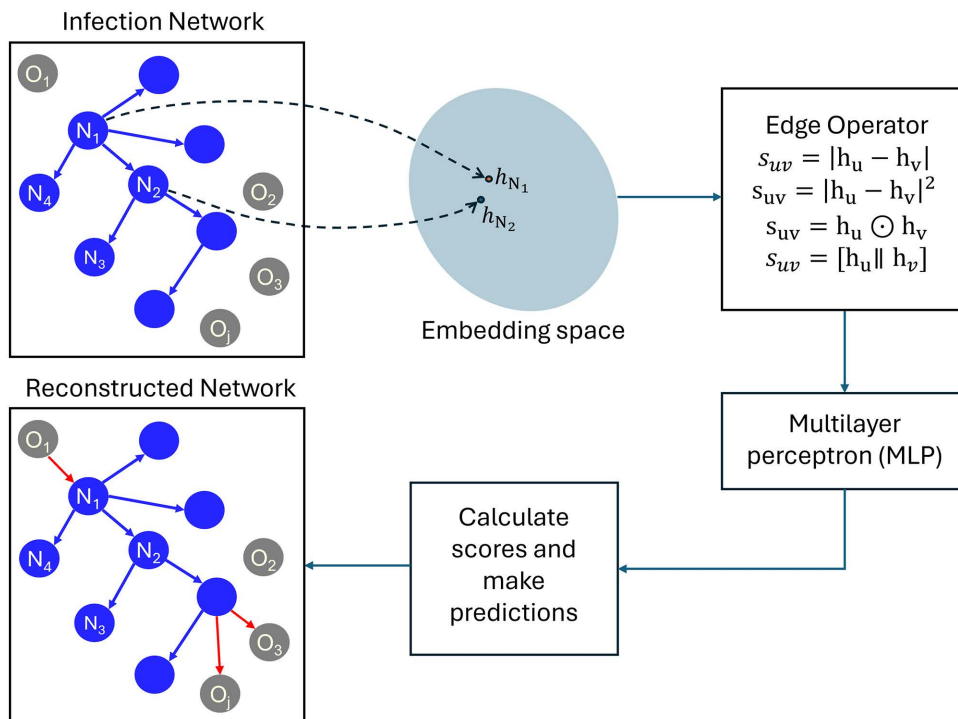


Fig 2. Workflow for predicting missing links in COVID-19 infection networks using GraphSAGE.

<https://doi.org/10.1371/journal.pcsy.0000124.g002>

The final model was used to infer plausible infection sources for unlinked cases, restricting candidate links to pairs consistent with a biologically plausible infectious time window.

2.5.1. Link prediction via inductive representation learning (GraphSAGE). A GraphSAGE graph neural network architecture was employed to learn latent transmission patterns from the observed infection network [31]. The encoder consisted of two GraphSAGE convolution layers, where node embeddings were updated iteratively according to (1)

$$\mathbf{h}_i^{(k)} = \sigma \left(\mathbf{W}^{(k)} \cdot \text{AGG}^{(k)} \left(\{\mathbf{h}_i^{(k-1)}\} \cup \{\mathbf{h}_j^{(k-1)} : j \in \mathcal{N}(i)\} \right) \right) \quad (1)$$

where, $\mathcal{N}(i)$ denotes the neighbours of node i , $\mathbf{W}^{(k)}$ represents the trainable weight matrix at layer k , and σ is a nonlinear activation function. For $k=0$, $\mathbf{h}_i^{(0)}$ represents the initial encoded node features, for $k=1$, $\mathbf{h}_i^{(1)}$ represents the embedding after passing from the first convolution layer, and $k=2$, $\mathbf{h}_i^{(2)}$ are the final embeddings ($\mathbf{z}_i$) used for link prediction.

For a candidate infection pair (u, v) , node embeddings were combined using several operators $\phi(\mathbf{z}_u, \mathbf{z}_v)$, including concatenation, absolute difference, squared difference, average, Hadamard product and dot product. The resulting edge representation was processed through a Multilayer Perceptron (MLP) to estimate the transmission probability (2).

$$\hat{p}_{uv} = \sigma \left(\text{MLP} \left(\phi(\mathbf{z}_u, \mathbf{z}_v) \right) \right) \quad (2)$$

For each documented transmission event, the true infection source was retained as the positive example, while up to 500 epidemiologically plausible alternative source cases were included as negative candidates. Candidate sources were required to satisfy a biologically plausible transmission interval (0–15 days) and a maximum geographical distance of 50 km from the target case. The model was optimized using a group-wise cross-entropy ranking loss, encouraging the true infection source

to receive the highest score among all candidate sources for the same target. Early stopping based on validation MRR was used to reduce overfitting. Model robustness was assessed across multiple random seeds and embedding combination operators using ranking-based metrics, including MRR and top-k accuracy, together with ROC AUC and F1-score.

2.5.2. Implementation and computational details. The GraphSAGE encoder was configured with a hidden layer dimension of 256. Training was conducted for a maximum of 300 epochs using the Adam optimizer with a learning rate of 0.01 and a weight decay of 0.0001 to mitigate overfitting. A patience-based early stopping mechanism was employed, where training was terminated if the validation loss failed to improve for 75 consecutive epochs. The model state associated with the lowest validation loss was preserved for final evaluation on the test set. All experiments were parallelized on a High-Performance Computing (HPC) cluster. The models were implemented using the PyTorch Geometric library and accelerated via GPU resources. Computational resources were provided by the High Performance Computing facility of the University of Cyprus (UCY HPC).

2.6. Estimating infection links for unlinked cases

After both link prediction approaches were evaluated, the best-performing model for each epidemic wave, based on the highest MRR-score, was selected to infer potential infection links for unlinked cases. This exercise was performed to examine the effect of including the unlinked cases on the network characteristics of the infection network. For each wave, all epidemiologically plausible source–target pairs were generated by restricting candidate sources to those occurring within a biologically plausible transmission interval (0–15 days) before the target case. Each candidate pair was then scored using the selected model to estimate its transmission probability. Finally, for each unlinked target case, only the candidate source with the highest predicted probability was retained, resulting in a single inferred infection source for every target.

3. Results

3.1. Method 1: Feature-based supervised learning for link prediction

Feature rankings from mRMR, for three representative seeds from Wave 1 are provided in the Supplementary Materials ([S1–S3 Figs](#)), illustrating that although the ranking of features varied slightly across seeds, the most informative features were consistently selected. [S4 Fig](#) in the supplementary materials shows the frequency with which each feature exceeded the importance threshold across all seeds and waves. [S5–S8 Figs](#) in the Supplementary Material present the permutation importance scores for the best-performing classifier in each epidemic wave following greedy feature selection. Across all waves, physical distance consistently exhibited high importance across random seeds, highlighting spatial proximity as a robust predictor of transmission likelihood in the reconstructed infection networks. Since in-degree and out-degree information is unavailable for unlinked cases during inference, all degree-based features were excluded from the final model configuration to maintain consistency between training and inference. As shown in [S3 Table](#), removing all degree-related features, also led to improved performance across all epidemic waves ([S4 Table](#)). To assess the effect of class imbalance, negative-to-positive sampling ratios of 50:1, 300:1, and 500:1 were evaluated for Wave 3. The 50:1 ratio achieved the best overall predictive performance and was therefore adopted for all subsequent experiments ([S3 Table](#), Wave 3). The complete results for the 300:1 and 500:1 sampling ratio are provided in [S5 Table](#).

Each seed produced a slightly different combination of selected features during the greedy selection process ([S5–S8 Figs](#)). Consequently, a unique feature set was obtained for each seed, resulting in slight variations in the performance of the evaluated classifiers. To identify the best-performing model, the performance of each classifier was averaged across all selected feature sets, and the classifier demonstrating the most consistent overall performance was selected. As shown in [S3 Table](#), the best-performing classifiers were Gradient Boosting for Waves 1 and 2 and Random Forest for Waves 3 and 4, although the differences in performance between the top models were relatively small.

To identify the most robust feature set, the ten feature subsets obtained for the various seeds for the best performing classifier in each wave were extracted. The full training pipeline was re-run using each of these candidate sets without

repeating the mRMR ranking and greedy feature selection procedure. This process ensures the identification of both the most stable classifier and the most stable feature set for each respective wave. The resulting performance metrics are summarized in Table 2, while Table 3 presents the final feature sets selected for each epidemic wave. For Wave 1, Random Forest obtained the highest F1-score (0.86 ± 0.02), while Gradient Boosting achieved the highest MRR (0.56 ± 0.04) and AUC (0.93 ± 0.02). Similar trends were observed for Wave 2, where Random Forest achieved the highest F1-score (0.90 ± 0.04), whereas Gradient Boosting provided the highest-ranking performance (MRR = 0.64 ± 0.06) and AUC (0.94 ± 0.02). For the larger epidemic waves (Waves 3 and 4), the overall performance of all classifiers decreased, reflecting the substantially greater size and complexity of the transmission networks. As the number of cases increased, contact tracing became more difficult, leaving the observed transmission networks more fragmented and reducing the information

Table 2. Test-set performance of the best-performing classifiers for the feature-based supervised learning (Method 1) across all epidemic waves, based on AUC and F1-score.

	Wave 1				Wave 2			
	F1	AUC	MRR	Top15 ACC	F1	AUC	MRR	Top15 ACC
DT	0.59 ± 0.07	0.70 ± 0.04	0.30 ± 0.07	0.52 ± 0.07	0.76 ± 0.09	0.79 ± 0.06	0.46 ± 0.08	0.73 ± 0.10
GB	0.82 ± 0.04	0.93 ± 0.02	0.56 ± 0.04	0.83 ± 0.04	0.87 ± 0.05	0.94 ± 0.02	0.64 ± 0.06	0.92 ± 0.03
KNN	0.65 ± 0.04	0.72 ± 0.02	0.27 ± 0.04	0.53 ± 0.04	0.64 ± 0.06	0.71 ± 0.03	0.34 ± 0.06	0.59 ± 0.07
LDA	0.75 ± 0.05	0.86 ± 0.03	0.37 ± 0.02	0.70 ± 0.05	0.49 ± 0.12	0.77 ± 0.04	0.31 ± 0.07	0.70 ± 0.10
LightGBMRanker	0.77 ± 0.05	0.92 ± 0.01	0.54 ± 0.03	0.84 ± 0.03	0.72 ± 0.08	0.93 ± 0.02	0.60 ± 0.06	0.92 ± 0.02
LR	0.77 ± 0.04	0.87 ± 0.02	0.39 ± 0.01	0.74 ± 0.03	0.71 ± 0.06	0.78 ± 0.04	0.36 ± 0.06	0.76 ± 0.09
NB	0.79 ± 0.05	0.86 ± 0.03	0.39 ± 0.03	0.72 ± 0.05	0.35 ± 0.12	0.74 ± 0.04	0.26 ± 0.07	0.73 ± 0.09
RF	0.86 ± 0.02	0.92 ± 0.02	0.55 ± 0.03	0.83 ± 0.04	0.90 ± 0.04	0.93 ± 0.02	0.60 ± 0.06	0.91 ± 0.03
SVM	0.49 ± 0.14	0.50 ± 0.13	0.16 ± 0.09	0.33 ± 0.14	0.57 ± 0.14	0.50 ± 0.11	0.18 ± 0.07	0.41 ± 0.14
	Wave 3				Wave 4			
	F1	AUC	MRR	Top15 ACC	F1	AUC	MRR	Top15 ACC
DT	0.68 ± 0.01	0.90 ± 0.00	0.22 ± 0.01	0.52 ± 0.01	0.71 ± 0.01	0.85 ± 0.00	0.27 ± 0.01	0.56 ± 0.01
GB	0.68 ± 0.01	0.92 ± 0.00	0.23 ± 0.01	0.53 ± 0.01	0.75 ± 0.01	0.96 ± 0.00	0.32 ± 0.01	0.60 ± 0.01
KNN	0.61 ± 0.02	0.81 ± 0.01	0.14 ± 0.01	0.44 ± 0.02	0.54 ± 0.04	0.89 ± 0.04	0.17 ± 0.01	0.37 ± 0.03
LDA	0.66 ± 0.01	0.89 ± 0.00	0.20 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.89 ± 0.00	0.15 ± 0.01	0.35 ± 0.01
LightGBMRanker	0.67 ± 0.01	0.93 ± 0.00	0.21 ± 0.01	0.52 ± 0.01	0.75 ± 0.01	0.96 ± 0.00	0.32 ± 0.01	0.60 ± 0.01
LR	0.67 ± 0.01	0.90 ± 0.00	0.21 ± 0.01	0.51 ± 0.01	0.52 ± 0.01	0.90 ± 0.00	0.15 ± 0.01	0.35 ± 0.01
NB	0.67 ± 0.01	0.90 ± 0.01	0.21 ± 0.01	0.51 ± 0.01	0.40 ± 0.01	0.87 ± 0.00	0.10 ± 0.01	0.25 ± 0.01
RF	0.69 ± 0.01	0.93 ± 0.00	0.23 ± 0.01	0.54 ± 0.01	0.75 ± 0.01	0.96 ± 0.00	0.33 ± 0.01	0.61 ± 0.01
SVM	0.29 ± 0.29	0.73 ± 0.15	0.08 ± 0.08	0.20 ± 0.22	0.11 ± 0.09	0.60 ± 0.07	0.03 ± 0.02	0.06 ± 0.06

<https://doi.org/10.1371/journal.pcsy.0000124.t002>

Table 3. Final feature sets selected for each epidemic wave. The reported features correspond to the best-performing feature set identified after evaluating the candidate feature sets generated across ten random seeds.

Wave	Final selected features
Wave 1	Age group difference, Day difference, Day–distance product, Day squared difference, Gender difference, Source gender, Municipality, NACE difference, NACE distance, NACE distance–time interaction, Physical distance, Postcode squared difference
Wave 2	Age difference, Absolute age difference, Age group difference, Age product, Day difference, Day product, Source gender, NACE distance, NACE distance–time interaction, Physical distance, Absolute postcode difference, Postcode product
Wave 3	Day–distance product, Target gender, Municipality, Target NACE sector, Physical distance, Squared physical distance
Wave 4	Age absolute difference, Age difference, Age group difference, Age product, Age squared difference, Day absolute difference, Day difference, Day–distance product, Gender absolute difference, Gender squared difference, Source gender, Target gender, Same municipality, NACE absolute difference, NACE distance, Squared physical distance, Postcode product.

<https://doi.org/10.1371/journal.pcsy.0000124.t003>

available for accurate link prediction. In Wave 3, Random Forest achieved the highest F1-score (0.69 ± 0.01), AUC (0.93 ± 0.00), MRR (0.23 ± 0.01), and top-15 accuracy (0.54 ± 0.01), closely followed by Gradient Boosting. ROC curves for the final selected models are provided in the Supplementary Material (S9–S12 Figs), demonstrating improved discrimination compared to the temporal-only and geography-only baseline models.

3.2. Method 2: Graph Neural Network Approach for Link prediction

Table 4 summarizes the different strategies used to combine node embeddings generated by the GraphSAGE model before they were passed to the MLP for link prediction. For Wave 1, the Diff2 operator achieved the highest F1-score (0.79 ± 0.07) and AUC (0.89 ± 0.02), while the Dot operator obtained the highest MRR (0.43 ± 0.06). Similarly, for Wave 2, Diff2 achieved the highest F1-score (0.74 ± 0.12), AUC (0.84 ± 0.04), and shared the highest MRR (0.41 ± 0.08) with the Hadamard operator. In contrast, for the larger Waves 3 and 4, the Average operator achieved the highest MRR (0.30 ± 0.01 and 0.23 ± 0.00 , respectively), together with the highest F1-scores (0.74 ± 0.01 and 0.70 ± 0.01). In Wave 1 and Wave 2, the performance of all strategies was lower than the feature-engineering/machine learning method. For the third wave, however, the graph neural network-based approach achieved higher MRR performance than the machine learning approach, and comparable MRR for Wave 4. ROC curves for the final selected GraphSAGE models, together with the temporal-only and geography-only baseline models, are provided in the Supplementary Material (S13–S16 Figs).

3.3. Application to unlinked cases

After the best model for each pandemic wave was identified, the top-performing classifier was applied to infer transmission links for unlinked cases. For each unlinked case, all epidemiologically plausible candidate infection sources were scored by the selected model. The candidate with the highest predicted probability was then assigned as the inferred infection source, ensuring that each target case was linked to a single most likely infector. The reconstructed infection networks for the first two epidemic waves are shown in Fig 3, while the reconstructed networks for Waves 3 and 4 are

Table 4. Test-set performance of the best-performing classifiers for Graph Neural Network Approach (Method 2) across all epidemic waves, based on AUC and F1-score.

	Wave 1				Wave 2			
	F1	AUC	MRR	Top15 ACC	F1	AUC	MRR	Top15 ACC
avg	0.64 ± 0.19	0.81 ± 0.06	0.40 ± 0.06	0.71 ± 0.07	0.52 ± 0.25	0.72 ± 0.09	0.36 ± 0.08	0.70 ± 0.13
concat	0.64 ± 0.13	0.79 ± 0.07	0.36 ± 0.05	0.63 ± 0.09	0.63 ± 0.12	0.75 ± 0.06	0.35 ± 0.06	0.75 ± 0.09
diff	0.60 ± 0.13	0.75 ± 0.05	0.34 ± 0.05	0.57 ± 0.08	0.42 ± 0.11	0.68 ± 0.04	0.27 ± 0.05	0.58 ± 0.09
diff2	0.79 ± 0.07	0.89 ± 0.02	0.42 ± 0.04	0.76 ± 0.04	0.74 ± 0.12	0.84 ± 0.04	0.41 ± 0.08	0.81 ± 0.08
dot	0.77 ± 0.14	0.88 ± 0.03	0.43 ± 0.06	0.75 ± 0.05	0.61 ± 0.31	0.76 ± 0.09	0.37 ± 0.12	0.74 ± 0.21
hadamard	0.76 ± 0.08	0.87 ± 0.04	0.41 ± 0.04	0.76 ± 0.04	0.61 ± 0.13	0.77 ± 0.06	0.41 ± 0.04	0.78 ± 0.07
	Wave 3				Wave 4			
	F1	AUC	MRR	Top15 ACC	F1	AUC	MRR	Top15 ACC
avg	0.74 ± 0.01	0.87 ± 0.01	0.30 ± 0.01	0.60 ± 0.01	0.70 ± 0.01	0.89 ± 0.01	0.23 ± 0.00	0.54 ± 0.01
concat	0.67 ± 0.03	0.91 ± 0.00	0.28 ± 0.01	0.57 ± 0.02	0.67 ± 0.01	0.91 ± 0.01	0.22 ± 0.01	0.52 ± 0.01
diff	0.63 ± 0.06	0.81 ± 0.02	0.23 ± 0.02	0.48 ± 0.05	0.56 ± 0.02	0.77 ± 0.01	0.17 ± 0.01	0.39 ± 0.02
diff2	0.71 ± 0.02	0.93 ± 0.01	0.23 ± 0.02	0.55 ± 0.03	0.66 ± 0.01	0.91 ± 0.01	0.19 ± 0.01	0.49 ± 0.01
dot	0.73 ± 0.02	0.93 ± 0.01	0.27 ± 0.02	0.59 ± 0.03	0.69 ± 0.03	0.91 ± 0.01	0.21 ± 0.02	0.53 ± 0.04
hadamard	0.73 ± 0.02	0.92 ± 0.01	0.28 ± 0.02	0.58 ± 0.02	0.63 ± 0.03	0.91 ± 0.01	0.19 ± 0.01	0.48 ± 0.02

Note: Avg: element-wise averaging of source and target embeddings; Concat: vector concatenation; Diff: the element-wise difference; Diff2: the squared element-wise difference; Dot: the dot product between embeddings; and Hadamard: the element-wise (Hadamard) product. These combined representations are used as input to the MLP predictor.

<https://doi.org/10.1371/journal.pcsy.0000124.t004>

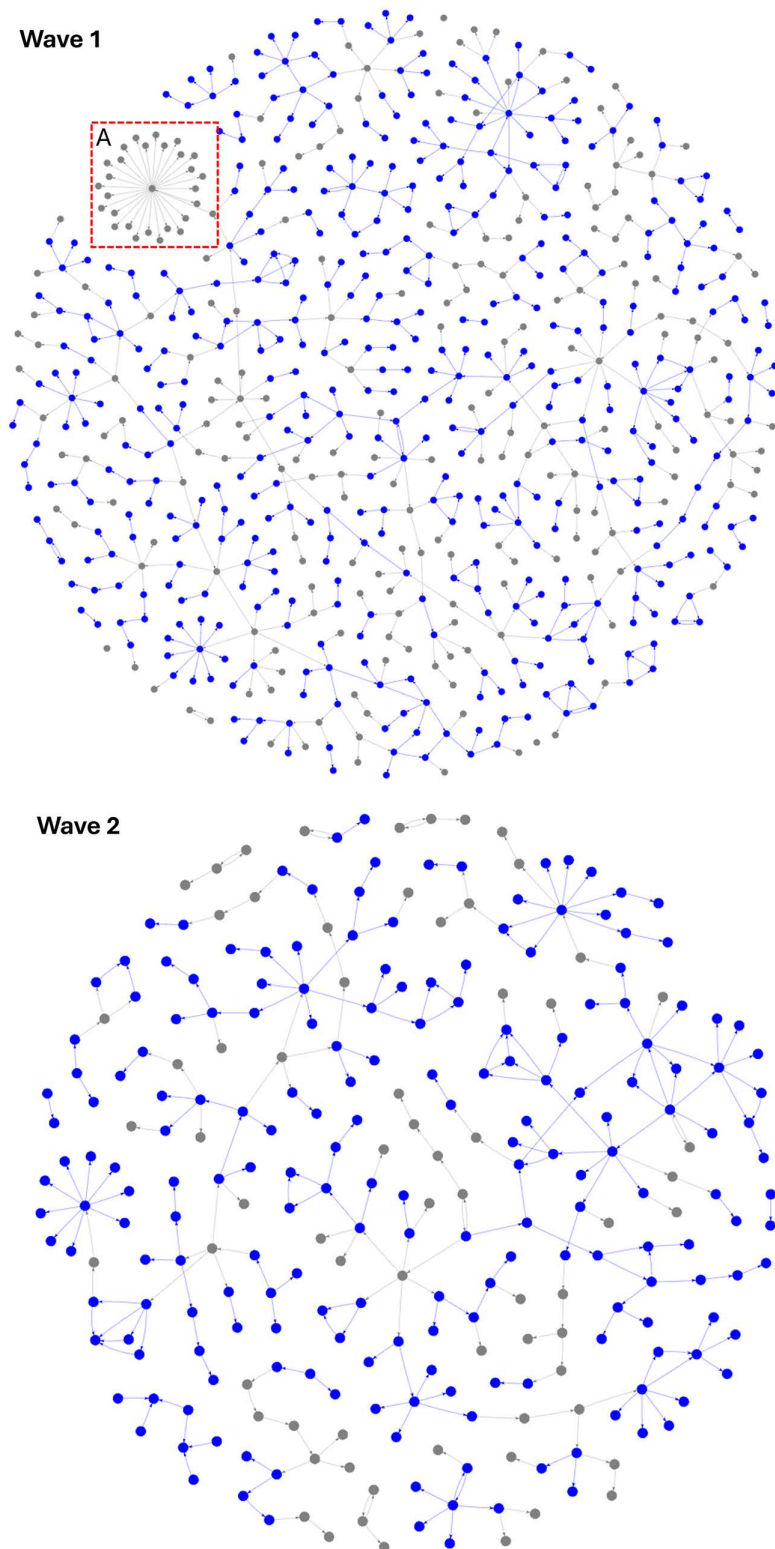


Fig 3. Visualization of the infection network for Wave 1 and Wave 2 including predicted links for unlinked cases. Blue nodes represent individuals with known epidemiological information, while grey nodes correspond to unlinked cases. Grey edges indicate predicted transmission links and blue edges the actual transmission links.

<https://doi.org/10.1371/journal.pcsy.0000124.g003>

provided in the Supplementary Material (S17 Fig). Cases with confirmed epidemiological information are displayed as blue nodes, whereas unlinked cases are displayed as grey nodes. Predicted transmission links are represented by grey edges, and confirmed transmissions are represented by blue edges.

Since Waves 3 and 4 included a large number of cases, the reconstructed transmission chains cannot be visually appreciated from the full network plots (S17 Fig). To provide a clearer illustration of how the inferred links integrate unlinked cases into the infection structure, Fig 4 shows an example of a reconstructed infection chain from Wave 4. This example highlights several types of inferred connections, including unlinked–unlinked links and connections between known cases and unlinked cases. Together, these patterns demonstrate how the link prediction models extend the observed transmission structure and incorporate previously isolated cases or trees into plausible chains.

For each reconstructed network, key structural properties were computed, including the out-degree distribution and the distribution of shortest-path lengths. These metrics were then compared with those of the original observed network as shown in Fig 5. The reconstructed networks exhibited out-degree distributions that closely resembled those of the observed infection networks, suggesting that the reconstruction procedure preserved the overall transmission structure. As expected, the reconstructed networks showed longer shortest path lengths than the observed networks. This increase is expected since the inferred transmission links connect previously disconnected nodes and transmission chains, resulting in longer and more complete transmission pathways. Consequently, the shortest path length provides a useful structural check, indicating that the inferred links extend the observed transmission network while maintaining a plausible network topology.

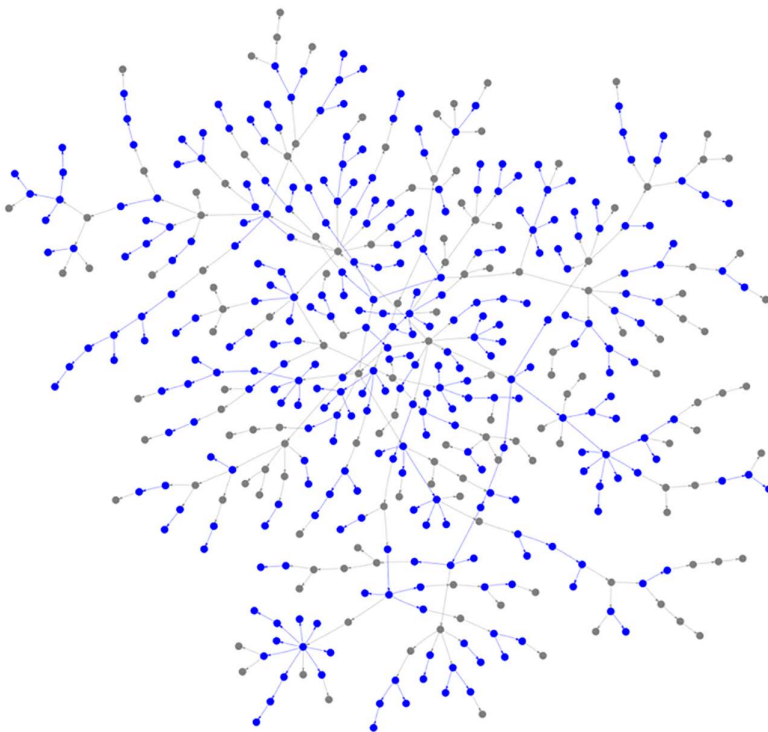


Fig 4. Visualization of an infection chain in Wave 4. Blue nodes represent individuals with known epidemiological information, while grey nodes correspond to unlinked cases. Grey edges indicate predicted transmission links and blue edges the actual transmission links.

<https://doi.org/10.1371/journal.pcsy.0000124.g004>

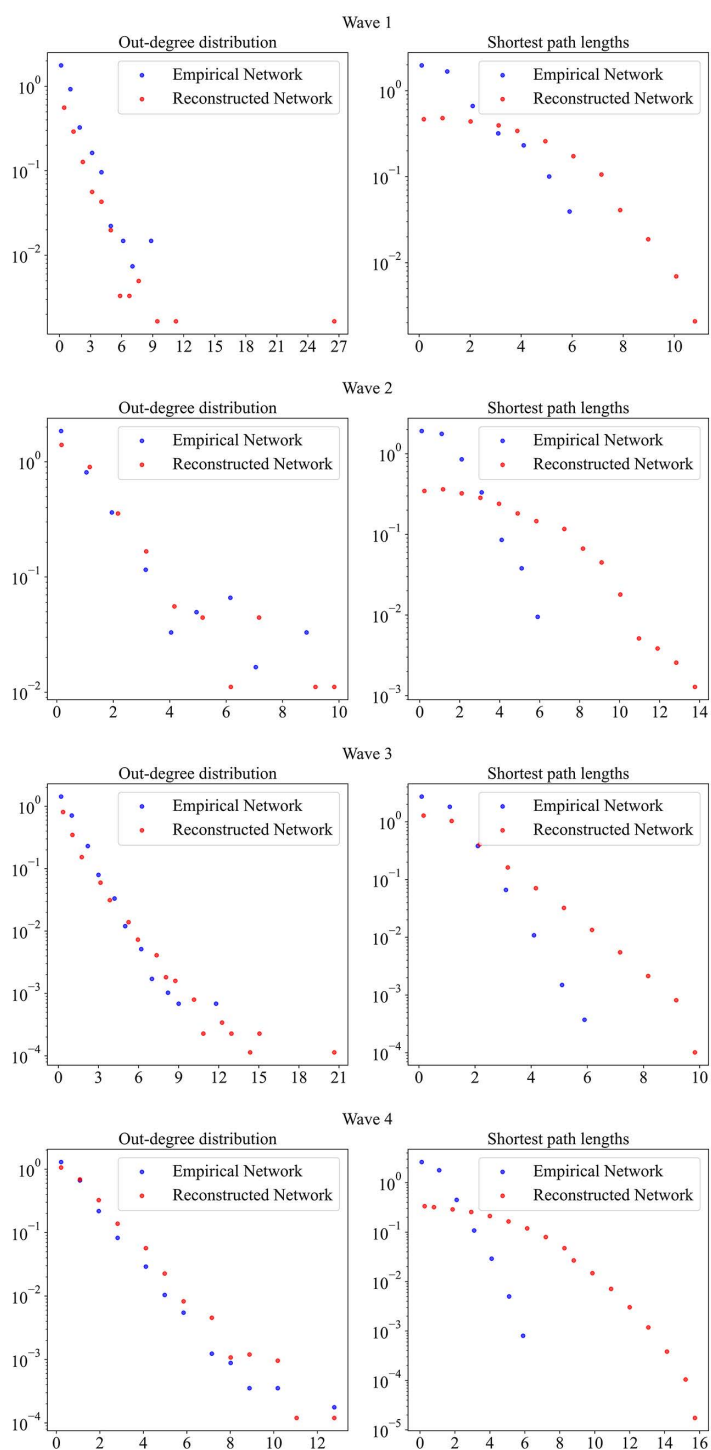


Fig 5. Comparison of out-degree and shortest-path distributions between the original and the reconstructed network.

<https://doi.org/10.1371/journal.pcsy.0000124.g005>

4. Discussion

Overall, the classical machine learning approach achieved the highest predictive performance, with Gradient Boosting producing the best results in Waves 1 and 2 (MRR=0.56 and 0.64, F1-score=0.82 and 0.87, ROC AUC=0.93 and 0.94, and top-15 accuracy=0.83 and 0.92, respectively). In Wave 3, the GraphSAGE-based approach achieved the best performance (MRR=0.30, F1-score=0.74, ROC AUC=0.87, and top-15 accuracy=0.60), whereas in Wave 4, Random Forest achieved the highest predictive performance (MRR=0.33, F1-score=0.75, ROC AUC=0.96, and Top-15 accuracy=0.61). Although predictive performance was highest during the smaller epidemic waves (Waves 1 and 2) and declined as the transmission networks became larger (Waves 3 and 4), both approaches continued to provide useful predictions, demonstrating their potential to support outbreak investigations during large-scale epidemics. From a public health perspective, these results indicate that the performance of proposed methods is adequate to provide meaningful epidemiological guidance, at least for prioritising the investigation of cases with unknown infection sources. For example, during the first two pandemic waves, the top-15 accuracy reached 92%, meaning that in up to 92% of cases, the true infection source appeared among the 15 highest-ranked candidates. In practice, this could help epidemiologists narrow their investigation from a large number of possible sources to a small, prioritised, group of individuals.

Previous approaches to transmission network reconstruction have relied primarily on pathogen genomic data, using phylogenetic and Bayesian methods such as TransPhylo [32] and Outbreaker2 [33] to infer transmission links. While these methods have proven valuable, they require whole-genome sequencing data which is not always available. In this work, the focus is on reconstructing plausible transmission chains from incomplete epidemiological data, highlighting the important implications for public-health surveillance. In many real-world outbreaks, contact tracing records are fragmented, and a substantial proportion of cases remain unlinked. Incomplete network data arise across diverse domains, with link prediction algorithms having been applied to partially observed social networks and brain connectivity data [34]. By integrating networks, machine learning predictions with epidemiological constraints, the proposed framework provides a way to identify plausible links and generate multiple plausible realizations of the underlying transmission network. These reconstructed networks can reveal how isolated cases may connect to known chains, identify potential bridging events between clusters, and highlight individuals who may play superspreader roles. This is consistent with a work demonstrating that graph neural networks can be applied to contact tracing as an online network exploration problem, for detecting superspreaders [35]. The reconstructed networks also allow direct comparison with the original observed networks, making it possible to quantify how missing information alters inferred transmission dynamics and to assess the epidemiological role of unlinked cases during an outbreak. Beyond retrospective analysis, these reconstructed networks could potentially support epidemic modelling by providing empirically-derived contact structures, offering an alternative to synthetic topologies such as random, scale-free, or small-world networks. This is particularly relevant given that contact network topology has been shown to significantly affect epidemic simulation outcomes [36]. Such reconstructed networks could potentially improve the predictive power of simulation models and enhance situational awareness during emerging epidemics empowering public health authorities with a clearer and more complete picture of transmission chains.

As illustrated in Fig 3 and Fig 4, the reconstructed networks connected previously disconnected transmission chains, enabling short chains in the observed data to merge into longer transmission pathways. This resulted in a more complete representation of disease transmission by revealing epidemiologically plausible links that were not identified through routine contact tracing investigations. A representative example is shown in Fig 3A for Wave 1, where the inferred transmission cluster corresponds to a superspreading event that was confirmed by epidemiologists and further validated through public press announcements [37,38]. All individuals within this cluster worked in the same NACE sector (C10, Manufacture of food products) and resided in the same municipality, providing additional epidemiological support for the inferred transmission links. The reconstructed transmission chain associated with this Wave 1 cluster is presented in the Supplementary Material (S18 Fig), where nodes are coloured according to their NACE category. The visualisation clearly shows

that the reconstructed cluster consists of individuals from the same occupational sector. The successful identification of this previously undocumented transmission cluster, which was independently confirmed through epidemiological investigation and public reports, provides strong evidence that the proposed reconstruction method can accurately recover missing transmission links.

Despite the strong reconstruction performance observed in the Wave 1, the inferred links should be interpreted as the most likely transmission hypotheses generated by the model rather than confirmed transmission events. Their accuracy depends on the quality and completeness of the underlying surveillance data used for training. In the case of Cyprus, the dataset used in this study is of relatively high quality for several reasons: the high volume of daily cases was systematically recorded, and as an island nation with controlled points of entry, the vast majority of infections were documented through the national surveillance system. However, even well-curated contact tracing data are not entirely free from bias. Recall bias during patient interviews, undetected asymptomatic transmission, and incomplete contact tracing may all contribute to missing or incorrect transmission links. These limitations should be considered when interpreting the reconstructed networks and when applying this framework to surveillance systems with lower tracing capacity.

Although GraphSAGE achieved reliable performance during the larger epidemic waves, its effectiveness was constrained by the sparse and fragmented structure of the observed infection networks. Graph neural networks learn node representations by aggregating information from neighbouring nodes. However, many transmission chains consisted of small, disconnected components, limiting the amount of structural information available for message passing. This limitation became even more pronounced when transmission links were withheld during evaluation, further reducing graph connectivity [28]. Under these conditions, epidemiological attributes such as temporal proximity, physical distance, age, and occupational similarity provide stronger predictive signals than the network structure itself. A promising direction for future research would be the use of heterogeneous epidemiological networks, in which individuals remain connected through additional relationship types, such as shared age groups, geographic regions, occupational sectors, or documented contact networks. Under this framework, even when transmission edges are removed for validation or inference, message passing can still exploit these auxiliary relationships to learn richer and more informative node embeddings, potentially improving link prediction performance in sparse epidemiological networks.

Beyond transmission network reconstruction, the inferred networks provide opportunities for additional epidemiological analyses. The inferred transmission networks can be compared with the observed networks not only through global structural metrics, such as out-degree distributions and shortest-path lengths, but also through epidemiologically meaningful attributes, including age groups, districts, and occupational categories. Such comparisons would help quantify how the inference process alters transmission pathways across age group, occupations and spatial patterns, and reveal whether certain groups are disproportionately affected by missing data. Additional embedding strategies, such as attention-based GNNs, contrastive learning, or uncertainty-aware node representations, can be explored to improve generalization to unlinked cases and enhance robustness under sparse network conditions. Finally, incorporating individual-level contact data, when available, would provide a richer substrate for link prediction and could substantially improve the fidelity of reconstructed transmission chains. Integrating these improvements would further strengthen the utility of data-driven network reconstruction for epidemiological analysis and outbreak response.

An important consideration is the applicability of the proposed framework to unlinked cases. Since the true transmission source of unlinked cases is unknown, direct validation of the reconstructed links is not possible. Although the inductive evaluation strategy used in this study approximates this real-world task, held-out linked cases may not fully represent unlinked cases, which may differ in terms of testing delays, interview completeness, undocumented exposures, or social behaviour. A promising approach for further validation would be to use epidemic simulation frameworks to generate transmission networks and complete infection trees for which the true source of each infection is known [39–41]. Selected transmission links could then be removed to create simulated unlinked cases, and the proposed link-prediction methods could be evaluated according to their ability to recover the known infection sources. Such an approach would enable

direct and controlled assessment of reconstruction accuracy under different epidemic conditions, network structures, levels of missingness, and intervention scenarios.

5. Conclusion

This study demonstrates that partially observed infection networks can be successfully reconstructed, supporting outbreak investigations and strengthening preparedness for future epidemics. Building upon previous work [28], the proposed framework expands the feature engineering strategy and introduces a GraphSAGE-based approach, showing that both classical machine learning models and graph representation learning can be used to infer missing transmission links among unlinked cases. The resulting reconstructed networks capture a wide range of plausible transmission pathways, recover long and complex chains, identify superspreading events, and provide a more complete view of plausible outbreak dynamics than is possible with observed data alone.

Supporting information

S1 Table. Overview of engineered edge-features used for link prediction.

(PDF)

S1 Fig. mRMR feature-ranking results for Wave 1 (seed=2025).

(TIF)

S2 Fig. mRMR feature-ranking results for Wave 1 (seed=2).

(TIF)

S3 Fig. mRMR feature-ranking results for Wave 1 (seed=99).

(TIF)

S4 Fig. Frequency of features selected across seeds for each wave during mRMR ranking.

(TIF)

S5 Fig. Permutation-importance scores for Wave 1 using the Gradient Boosting classifier.

(TIF)

S6 Fig. Permutation-importance scores for Wave 2 using the Gradient Boosting classifier.

(TIF)

S7 Fig. Permutation-importance scores for Wave 3 using the Random Forest. classifier.

(TIF)

S8 Fig. Permutation-importance scores for Wave 4 using the Random Forest classifier.

(TIF)

S2 Table. Hyperparameters for Method 1.

(PDF)

S3 Table. Test-set performance of the best-performing classifiers for Method 1 across all epidemic waves, based on AUC, F1-score, MRR and top-15 accuracy. All outdegree and indegree features were excluded.

(PDF)

S4 Table. Test-set performance of the best-performing classifiers for Method 1 across all epidemic waves, based on AUC, F1-score, MRR and top-15 accuracy. All outdegree and indegree features were included.

(PDF)

S5 Table. Performance of the evaluated classifiers on Wave 3 using a 300:1 and 500:1 negative-to-positive sampling ratio.

(PDF)

S9 Fig. Receiver operating characteristic (ROC) curves for Method 1 using the final selected feature set for Wave 1.

(TIFF)

S10 Fig. Receiver operating characteristic (ROC) curves for Method 1 using the final selected feature set for Wave 2.

(TIFF)

S11 Fig. Receiver operating characteristic (ROC) curves for Method 1 using the final selected feature set for Wave 3.

(TIFF)

S12 Fig. Receiver operating characteristic (ROC) curves for Method 1 using the final selected feature set for Wave 4.

(TIFF)

S13 Fig. Receiver operating characteristic (ROC) curves for Method 2 (GraphSAGE) for Wave 1.

(TIFF)

S14 Fig. Receiver operating characteristic (ROC) curves for Method 2 (GraphSAGE) for Wave 2.

(TIFF)

S15 Fig. Receiver operating characteristic (ROC) curves for Method 2 (GraphSAGE) for Wave 3.

(TIFF)

S16 Fig. Receiver operating characteristic (ROC) curves for Method 2 (GraphSAGE) for Wave 4.

(TIFF)

S17 Fig. Visualization of the predicted infection network for Wave 3 and 4.

(TIF)

S18 Fig. Node colors represent NACE categories: green (healthcare), red (food industry), beige (retired), and orange (all other sectors).

(TIF)

Acknowledgments

This work is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 739551 (KIOS CoE) and the Government of the Republic of Cyprus through the Cyprus Deputy Ministry of Research, Innovation and Digital Policy. It was also supported by the CIPHIS project (Cyprus Innovative Public Health ICT System), C1.1I2, of the NextGenerationEU programme under the Republic of Cyprus Recovery and Resilience Plan, based on the research collaboration agreement with the Cyprus Ministry of Health.

Author contributions

Conceptualization: Pavlos Alexandros Dimitriou, Costas Pitris.

Data curation: Pavlos Alexandros Dimitriou, Valentinos Silvestros, Elisavet Constantinou, Costas Pitris, Panayiotis Kolios.

Formal analysis: Pavlos Alexandros Dimitriou, Costas Pitris.

Funding acquisition: Panayiotis Kolios.

Methodology: Pavlos Alexandros Dimitriou.

Supervision: Elisavet Constantinou, Costas Pitris, Panayiotis Kolios.

Validation: Pavlos Alexandros Dimitriou, Valentinos Silvestros, Elisavet Constantinou, Costas Pitris.

Visualization: Pavlos Alexandros Dimitriou.

Writing – original draft: Pavlos Alexandros Dimitriou.

Writing – review & editing: Pavlos Alexandros Dimitriou, Costas Pitris, Panayiotis Kolios.

References

- Antweiler D, Sessler D, Rosknecht M, Abb B, Ginzel S, Kohlhammer J. Uncovering chains of infections through spatio-temporal and visual analysis of COVID-19 contact traces. *Comput Graph*. 2022;106:1–8. <https://doi.org/10.1016/j.cag.2022.05.013> PMID: [35637696](#)
- Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *Lancet Public Health*. 2020;5(8):e452–9. [https://doi.org/10.1016/S2468-2667\(20\)30157-2](https://doi.org/10.1016/S2468-2667(20)30157-2)
- Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *J Epidemiol Community Health*. 2020;74(10):861–6. <https://doi.org/10.1136/jech-2020-214051> PMID: [32576605](#)
- Hellewell J, Hellewell J, Hellewell J. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob Health*. 2020;8(4):e488–96. [https://doi.org/10.1016/S2214-109X\(20\)30074-7](https://doi.org/10.1016/S2214-109X(20)30074-7)
- Saraswathi S, Mukhopadhyay A, Shah H, Ranganath TS. Social network analysis of COVID-19 transmission in Karnataka, India. *Epidemiol Infect*. 2020;148:1–10. <https://doi.org/10.1017/S095026882000223X>
- Nagarajan K, Muniyandi M, Palani B, Sellappan S. Social network analysis methods for exploring SARS-CoV-2 contact tracing data. *BMC Med Res Methodol*. 2020;20(1):233. <https://doi.org/10.1186/s12874-020-01119-3> PMID: [32942988](#)
- Andre M, Ijaz K, Tillinghast JD, Krebs VE, Diem LA, Metchock B, et al. Transmission network analysis to complement routine tuberculosis contact investigations. *Am J Public Health*. 2007;97(3):470–7. <https://doi.org/10.2105/AJPH.2005.071936> PMID: [17018825](#)
- Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P. Network epidemiological analysis of COVID-19 transmission patterns by age, occupation and residence across four waves in Cyprus. *Sci Rep*. 2025;15(1):28416. <https://doi.org/10.1038/s41598-025-14267-7> PMID: [40760155](#)
- Liu Y, Gu Z, Liu J. Uncovering transmission patterns of COVID-19 outbreaks: A region-wide comprehensive retrospective study in Hong Kong-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/> 2021
- Jeong Y, Kang S, Kim B, Gil YJ, Hwang S-S, Cho S-I. Utilization of the Unlinked Case Proportion to Control COVID-19: A Focus on the Non-pharmaceutical Interventional Policies of the Korea and Japan. *J Prev Med Public Health*. 2023;56(4):377–83. <https://doi.org/10.3961/jpmph.23.056> PMID: [37551076](#)
- Lee H, Choi H, Lee H, Lee S, Kim C. Uncovering COVID-19 transmission tree: identifying traced and untraced infections in an infection network. *Front Public Health*. 2024;12:1362823. <https://doi.org/10.3389/fpubh.2024.1362823> PMID: [38887240](#)
- Chong KC, et al. Characterization of unlinked cases of COVID-19 and implications for contact tracing measures: Retrospective analysis of surveillance data. *JMIR Public Health Surveill*. 2021;7(11). <https://doi.org/10.2196/30968>
- Wong NS, Lee SS, Kwan TH, Yeoh E-K. Settings of virus exposure and their implications in the propagation of transmission networks in a COVID-19 outbreak. *Lancet Reg Health West Pac*. 2020;4:100052. <https://doi.org/10.1016/j.lanwpc.2020.100052> PMID: [34013218](#)
- Chae M.-K, Son W-S, Lee SH. The missing links: Evaluating contact tracing with incomplete data in large metropolitan areas during an epidemic. *Jan*. 2026, Accessed: Jan. 29, 2026. <https://arxiv.org/pdf/2601.14632>
- Daud NN, Ab Hamid SH, Saadoon M, Sahran F, Anuar NB. Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications*. 2020;166:102716. <https://doi.org/10.1016/j.jnca.2020.102716>
- Linyuan LL, Zhou T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*. 2011;390(6):1150–70. <https://doi.org/10.1016/J.PHYSA.2010.11.027>
- Link prediction in relational data. In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. <https://dl.acm.org/doi/10.5555/2981345.2981428>
- Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *J Am Soc Inf Sci*. 2007;58(7):1019–31. <https://doi.org/10.1002/asi.20591>
- Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans Knowl Data Eng*. 2005;17(6):734–49. <https://doi.org/10.1109/tkde.2005.99>

20. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A*. 2003;100(8):4372–6. <https://doi.org/10.1073/pnas.0735871100> PMID: 12676999
21. Al Hasan M, Chaoji V, Salem S, Zaki M. Link Prediction using Supervised Learning. <http://www.elsevier.com>
22. Song K, Park H, Lee J, Kim A, Jung J. COVID-19 infection inference with graph neural networks. *Sci Rep*. 2023;13(1):11469. <https://doi.org/10.1038/s41598-023-38314-3> PMID: 37454206
23. Li M, Cui J, Zhang J, Pei X, Sun G. Transmission characteristic and dynamic analysis of COVID-19 on contact network with Tianjin city in China. *Physica A: Statistical Mechanics and its Applications*. 2022;608:128246. <https://doi.org/10.1016/J.PHYSA.2022.128246>
24. Kwon O, Jo HH. Clustering and link prediction for mesoscopic COVID-19 transmission networks in Republic of Korea. *Chaos*. 2023;33(1). <https://doi.org/10.1063/5.0130386>
25. Pei S, et al. Contact tracing reveals community transmission of COVID-19 in New York City. *Nat Commun*. 2022;13(1):1–8. <https://doi.org/10.1038/s41467-022-34130-x>
26. Yang Z, Zhang J, Gao S, Wang H. Complex Contact Network of Patients at the Beginning of an Epidemic Outbreak: An Analysis Based on 1218 COVID-19 Cases in China. *Int J Environ Res Public Health*. 2022;19(2):689. <https://doi.org/10.3390/ijerph19020689> PMID: 35055511
27. Hibshman JI, Weninger T. Inherent Limits on Topology-Based Link Prediction. 2023.
28. Dimitriou PA, Silvestros V, Constantinou E, Pitris C, Kolios P. Reconstructing COVID-19 infection networks: A link prediction approach to connecting orphan cases. In: *Complex Networks & Their Applications XIV: Proceedings of The Fourteenth International Conference on Complex Networks and their Applications: COMPLEX NETWORKS 2025*. <https://link.springer.com/book/9783032166449>
29. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(8):1226–38. <https://doi.org/10.1109/TPAMI.2005.159> PMID: 16119262
30. Altmann A, Tološi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–7. <https://doi.org/10.1093/bioinformatics/btq134> PMID: 20385727
31. Inductive representation learning on large graphs. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. <https://dl.acm.org/doi/10.5555/3294771.3294869>
32. Didelot X, Fraser C, Gardy J, Colijn C, Malik H. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol Biol Evol*. 2017;34(4):997–1007. <https://doi.org/10.1093/MOLBEV/MSW275>
33. Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: a modular platform for outbreak reconstruction. *BMC Bioinformatics*. 2018;19(Suppl 11):363. <https://doi.org/10.1186/s12859-018-2330-z> PMID: 30343663
34. Wu Q, Zhang Z, Ma T, Waltz J, Milton D, Chen S. Link predictions for incomplete network data with outcome misclassification. *Stat Med*. 2021;40(6):1519–34. <https://doi.org/10.1002/sim.8856> PMID: 33482688
35. Tan CW, Yu P-D, Chen S, Poor HV. DeepTrace: Learning to Optimize Contact Tracing in Epidemic Networks With Graph Neural Networks. *IEEE Trans on Signal and Inf Process over Networks*. 2025;11:97–113. <https://doi.org/10.1109/tsipn.2025.3530346>
36. Ahmad R, Xu KS. Effects of contact network models on stochastic epidemic simulations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2017. p. 101–10. https://doi.org/10.1007/978-3-319-67256-4_10
37. COVID19: Cyprus supermarket chain and bakery facility hit by virus. *Financial Mirror*. <https://www.financialmirror.com/2020/04/14/covid19-cyprus-supermarket-chain-and-bakery-facility-hit-by-virus/> Accessed 2026 July 29.
38. Zorbas A. Announcement by A. Zorbas & Sons Ltd on 12 Covid cases. <https://en.philenews.com/local/announcement-by-a-zorbas-sons-ltd-on-12-covid-cases/> Accessed 2026 July 29.
39. COVID-19 Modelling Aotearoa public projects. <https://gitlab.com/cma-public-projects/pain> Accessed 2026 July 29.
40. Turnbull SM, Hobbs M, Gray L, Harvey EP, Scarrold WML, O’Neale DRJ. Investigating the transmission risk of infectious disease outbreaks through the Aotearoa Co-incidence Network (ACN): a population-based study. *Lancet Reg Health West Pac*. 2022;20:100351. <https://doi.org/10.1016/j.lanwpc.2021.100351> PMID: 35024675
41. Vattiato G. Modelling Aotearoa New Zealand’s COVID-19 protection framework and the transition away from the elimination strategy. *R. Soc. Open Sci*. 2023;10(2). <https://doi.org/10.1098/RSOS.220766> PMID: 35024675