



# Harnessing multi-output machine learning approach and dynamical observables from network structure to optimize COVID-19 intervention strategies

Caroline L. Alves<sup>1,\*</sup> , Katharina Kuhnert<sup>1</sup>, Francisco Aparecido Rodrigues<sup>2</sup> and Michael Moeckel<sup>1</sup> 

<sup>1</sup>Center for Scientific Services and Transfer, Aschaffenburg University of Applied Sciences, Aschaffenburg, Germany

<sup>2</sup>Institute of Mathematical and Computer Sciences (IMC), University of São Paulo (USP), São Paulo, Brazil

\*Corresponding author. Center for Scientific Services and Transfer, Aschaffenburg University of Applied Sciences, Aschaffenburg, Germany. E-mail: Caroline.Alves@th-ab.de

## Abstract

The coronavirus disease 2019 (COVID-19) pandemic has necessitated the development of accurate models to predict disease dynamics and guide public health interventions. This study leverages the COVASIM agent-based model to simulate 1331 scenarios of COVID-19 transmission across various social settings, focusing on the school, community, and work contact layers. We extracted complex network measures from these simulations and applied deep learning algorithms to predict key epidemiological outcomes, such as infected, severe, and critical cases. Our approach achieved an  $R^2$  value exceeding 95%, demonstrating the model's robust predictive capability. Additionally, we identified optimal intervention strategies using spline interpolation, revealing the critical roles of community and workplace interventions in minimizing the pandemic's impact. The findings underscore the value of integrating network analytics with deep learning to streamline epidemic modeling, reduce computational costs, and enhance public health decision-making. This research offers a novel framework for effectively managing infectious disease outbreaks through targeted, data-driven interventions.

**Keywords:** COVID-19; agent-based modeling; complex network measures; deep learning; public health interventions; COVASIM

Accurate predicting infectious disease dynamics is crucial for effective public health interventions, especially in pandemics like COVID-19. This study presents a novel approach that combines complex network analytics with deep learning to predict key epidemiological outcomes and optimize intervention strategies. Simulating various transmission scenarios using the COVASIM agent-based model across the school, community, and workplace settings achieved high predictive accuracy with a  $R^2$  value exceeding 95%. The findings underscore the significant impact of community and workplace interventions in mitigating the pandemic's spread. This framework enhances epidemic modeling by reducing computational costs and provides valuable insights for policymakers to design effective containment measures, contributing to improved management of infectious disease outbreaks.

## Introduction

The coronavirus disease 2019 (COVID-19) epidemic has disclosed difficulties and delays in the public health and societal response to emerging new diseases. While simulation tools designed to model the dynamics of infections have been quickly adapted to new virus parameters [1–3] and made available to a large community of researchers [4], ready-made expert support systems for predicting the effectiveness and impact of public health interventions have not been publicly available. Responsible public health administrators and the general public alike were lacking essential information to opt quickly for the most appropriate and least intrusive intervention.

Hence, a need for new predictive tools has become evident. This paper demonstrates a first direction to further develop agent-based disease simulators into tools for predicting optimal public health responses. As a full study of all conceivable interventions is beyond the scope of this work, a focus on contact restrictions is taken.

Mathematical models, notably Susceptible-Exposed-Infectious-Recovered (SEIR) models, have been seminal in generating predictions and guiding public health measures [5–7].

Compartmental models are non-linear models widely recognized for their effectiveness in understanding the spread of infectious diseases within a population. The SEIR model subdivides individuals into four distinct compartments based on their disease progression status: Susceptible, Exposed, Infectious, and Recovered [6, 8]. The dynamics of transitions between these compartments are governed by a set of ordinary differential equations driven by essential parameters such as transmission rate, incubation period, and recovery rate described in Equation (1) [8, 9].

$$\begin{aligned} \frac{dS}{dt} &= -\beta \frac{SI}{N} \\ \frac{dE}{dt} &= \beta \frac{SI}{N} - \alpha E \\ \frac{dI}{dt} &= \alpha E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \quad (1)$$

Where:  $S(t)$  represents the number of susceptible individuals at time  $t$ ,  $E(t)$  represents the number of exposed individuals

Received: 31 January 2025; Revised: 28 April 2025; Editorial decision: 08 May 2025; Accepted: 25 May 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [reprints@oup.com](mailto:reprints@oup.com) for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com).

(infected but not yet infectious) at time  $t$ ,  $I(t)$  represents the number of infectious individuals at time  $t$ ,  $E(t)$  represents the number of exposed individuals (infected but not yet infectious) at time  $t$ ,  $R(t)$  represents the number of recovered individuals at time  $t$ ,  $N$  is the total population size,  $\beta$  is the transmission rate (average number of contacts an infectious person makes per unit time multiplied by the probability of disease transmission in a contact),  $\alpha$  is the rate at which individuals move from the exposed to infectious compartment (1/incubation period), and  $\gamma$  is the recovery rate (1/duration of infection).

By leveraging mathematical formulations and incorporating key parameters, compartmental models have proven invaluable in guiding public health measures [5, 10, 11]. On the other hand, agent-based simulators like COVASIM [4] have emerged as powerful tools for studying the complex dynamics of infectious diseases, extending the SEIR model by allowing for the integration of fine grained disease stages, age-specific parameters, immunization models etc., enabling a more nuanced simulation of disease progression across various demographic segments. Furthermore, COVASIM implements the contact network of agents, capturing heterogeneous interaction patterns that the traditional SEIR model oversimplifies, thereby offering a more detailed and realistic representation of disease transmission [12–14].

COVASIM is a stochastic agent-based simulator, which allows for the representation of individual-level heterogeneity in behaviors and interactions, leading to more realistic epidemic predictions [15, 16]. Built on separately generated synthetic populations, this tool additionally includes country-specific demographic information on age structure and population size. Social contact networks are structured in four subgraphs (layers), referring to households, schools, workplaces, and communities [4]. Through COVASIM, it is possible to simulate a refined transmission network by a multigraph with multiple layers, capturing the complex interactions between individuals in a population [13]. The community layer accounts for random contact between individuals in the general population, mimicking casual encounters in public spaces or social gatherings [17]. Household contacts, as another layer, capture close and sustained interactions within households, which are known to be a significant source of disease transmission. The work and school layers further contribute to the complexity of the model as they reflect the specific patterns of interactions that occur in these settings [18]. Workplaces often involve dense and regular contact, while schools accommodate interactions among children and staff, influencing the spread of infections among younger populations [19]. The multigraph contact network enables a detailed and nuanced representation of disease transmission dynamics in various social contexts. The approach allows researchers and policymakers to explore the potential effectiveness of public health interventions in controlling infectious diseases [20–22] in particular those which target specifically the contact networks [13, 23–25].

One major limitation of agent-based simulators like COVASIM is their computational cost [26]. Simulating numerous scenarios with detailed contact networks can require significant computing resources and time.

This work addresses the hypothesis that between a full modeling of contact networks in agent-based simulators like COVASIM and neglecting the structure of contact networks in compartment models there is a gap to establish a computationally efficient and fast prediction method which includes some effective properties of contact networks but avoids their full inclusion. We use machine learning for a simplified substitutional model by replacing

the full contact network adjacency matrix with a description based on complex network measures, such as betweenness centrality (BC) and others based in Rodrigues et al. [27]. They serve as predictors of disease dynamics while other parameters of the COVASIM framework remain fixed.

We use COVASIM to set up the contact network adjacency matrix using its initial state initialization routine, extracted it and calculated relevant complex network measures from it. Those served as input features for a deep learning-based method to predict time series data of the disease dynamics. A limited volume of synthetic data generated from the COVASIM simulator was used for training the model. Thus, we have shown by construction that a prediction of the disease dynamics from effective network properties is possible with sufficient accuracy without incurring the computational cost of performing a full agent-based simulation on a complex contact network.

We considered contact restrictions by public health interventions which modify the contact networks in schools, community and work and the corresponding values of complex network measures. The speed-up of predictions by the substitutional model allowed to efficiently sample the space of possible contact restrictions and to quickly predict the corresponding disease dynamics. From a parameter space study an inverse model was set up to derive the manifold in intervention space which is consistent with given external constraints, for example, a maximum number of critically sick or hospitalized patients. This manifolds marks the weakest contact restrictions compatible with an externally pre-defined maximally acceptable disease dynamics. A discussion of the points of the manifold allows to open a public debate on the least intrusive and most acceptable strategy for infectious disease control.

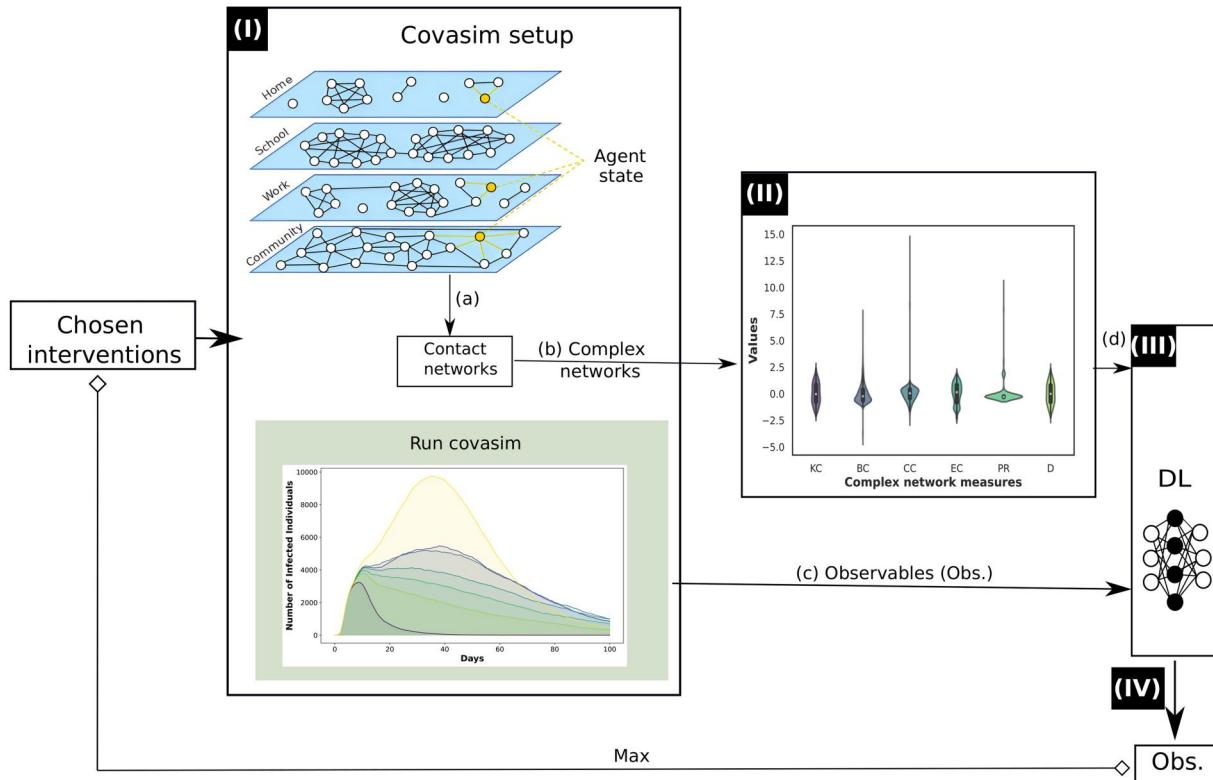
## Materials and methods

We use COVASIM to set up realistic contact networks based on a synthetic population which resembles the city of Aschaffenburg, Germany. Its roughly 70,000 inhabitants were represented by 70,000 agents and synthetic contact networks, structured as four subgraphs for households, school, work, and communities, were created to reproduce essential macroscopic statistical quantities as implemented in COVASIM. COVASIM simulations have been fitted and calibrated with respect to historic epidemiological data in previous work [28, 29]. Interaction patterns and rates within and across the subgraphs are based on the synthetic population and modeled through a stochastic process to capture the heterogeneity of disease transmission.

Analysis was conducted in Python version 3.6.15 and Docker. The Docker image includes the required Python packages, COVASIM source code, and any additional data needed for the simulations, and all code used here can be found in: <https://github.com/kathlab/covasim-covid19>.

### Creating synthetic data from COVASIM simulation

We use COVASIM to perform a parameter study for varying contact restrictions in schools (s), community (c), and work (w), which implies varying contact networks (Fig. 1-I). A total of 1331 scenarios were created by systematically varying the level of contact reduction in each of these three layers, using 11 uniformly spaced values between 0.0 (full contact removal) and 1.0 (no restriction). This 3D grid of intervention configurations is illustrated in Supplementary Appendix A (Fig. 6). For each scenario, a



**Figure 1.** Schematic representation of the methodology using COVASIM simulation. After the chosen interventions, the scenarios are simulated in COVASIM. (I) Creating synthetic data from COVASIM simulation, described in the Subsection II A: The simulation involves three contact network subgraphs—school, work, and community—jointly represented by a multigraph. Each agent is represented by one node in at least one subgraph (up to maximally one node in each subgraph) and characterized by a multi-valued agent state (in the orange box) as detailed in Kerr et al. [4]. Links on the contact network graph represent the possibility for disease transmission between agents which is randomly substantiated through a stochastic process (I.a). We consider 1331 unique variations in the contact network, each differing by the connectivity within the subgraphs. The disease dynamics is represented by time series data of observables (Obs), for example, the time dependent number of infected agents, and vectorized as  $I(t)$ . It is calculated using COVASIM for a modeling time of 100 days; exemplary simulation results are illustrated for seven scenarios (I.b). From the multigraphs, six complex network measures—KC, BC, CC, ECP, R, and D—which characterize the contact networks, were extracted for dimensionality reduction. The distribution ranges of these measures for the considered networks are depicted in (II). Substitutional ML and DL models (II C) are trained to predict the observables in (II.b) from complex network measure inputs. Finally, the manifold of minimal intervention is recovered as described in II D

new contact network was generated and the full disease dynamics were calculated over a simulation horizon of 100 days.

Contact restrictions were implemented using COVASIM's internal contact scaling mechanism. Specifically, each restriction level was applied as a scaling factor that probabilistically retains a fraction of the edges (contacts) within the corresponding network layer. For instance, a scaling factor of 0.6 means that 60% of contacts in that layer are preserved, while the remaining 40% are randomly removed. This stochastic edge reduction maintains the approximate structural characteristics of the layer while modeling decreased social interaction due to public health interventions. Each combination of restriction levels produced a unique network configuration used in the corresponding simulation.

The household subgraph was excluded from the study, as public health interventions in households are considered exceptionally intrusive and their impact comparatively less significant [30, 31]. Moreover, household networks in COVASIM are composed of small, tightly clustered groups with limited variability across scenarios. Excluding this layer also served to reduce the dimensionality of the parameter space and focus our analysis on public contact settings that are typically targeted by intervention policies. By concentrating on school, community, and work layers, we aimed to capture the most relevant aspects of transmission in the public sphere and better understand the structural effects of varying policy interventions.

While the COVASIM simulation preserves the layered structure of contact networks, for feature extraction, we aggregated the school, work, and community subgraphs into a single contact network per scenario. From this aggregated network, we computed the complex network measures used as input features for the machine learning models. This approach allowed us to capture the joint effects of modifiable social contacts while maintaining compatibility with the network analysis techniques employed in our study.

The initial infected population was set at 4.5% of the total population, reflecting early-stage epidemic conditions often observed in urban settings. This percentage was chosen based on epidemiological data suggesting an initial infection seeding rate in similar contexts. For example, the early phases of the COVID-19 pandemic in various European cities saw similar infection rates, as in [32] documented rapid increases in urban areas due to high population density and connectivity. The transmission rate, represented by the parameter  $\beta = 0.01825$ , and to differentiate between varying disease severities, we set the relative probabilities of developing severe and critical cases to 0.6558 and 0.9404, respectively. These values were derived from our previous work [28, 29] to calibrate and align with observed COVID-19 transmission dynamics, ensuring that the simulated spread closely mirrors real-world patterns. The simulations of disease dynamics were conducted for 100 days.

We calculated each scenario as an ensemble of five equivalent implementations to account for stochastic variability, enhancing our findings' statistical robustness; we calculated average outcomes across multiple runs, reducing the impact of random fluctuations and providing more reliable results [4, 33, 34].

## Complex network feature extraction for dimensionality reduction

In line with Rodrigues et al. [27], for each of these contact networks the following complex network metrics are extracted; they incorporate essential networks properties and provide dimensionality reduced input features for a machine learning model: First, BC, introduced by Freeman [35], measures the extent to which a node lies on the shortest paths between other nodes, indicating its role as a bridge or a bottleneck within the network. Closeness centrality (CC), also developed by Freeman [36], assesses how close a node is to all other nodes in the network, reflecting the efficiency with which information can spread from that node. Eigenvector centrality (EC), proposed by Bonacich [37], considers the number of connections a node has and the importance of the nodes it is connected to, providing a measure of a node's influence within the network.

In addition, PageRank (PR), originally devised by Brin and Page [38] for ranking web pages, is employed to evaluate the importance of nodes based on the quantity and quality of incoming connections. Degree (D), discussed by Doyle and Graver [39], is the simplest measure, representing the number of direct connections a node has, which can indicate its potential for interaction within the network. Finally, the k-core (KC) measure, described in [40, 41], identifies the largest subnetwork in which each node is connected to at least k other nodes, highlighting the network's cohesiveness.

The ensemble-averaged values of the complex network measures, whose distribution ranges are shown in Fig. 1-II, were used as the input features for the machine learning models to predict the corresponding  $I(t)$ . Here, ensemble average refers to averaging across five independent stochastic realizations of each scenario simulated in COVASIM. For each intervention configuration, we repeated the simulation five times and computed both the mean  $I(t)$  curve and the mean network measures across these runs, to reduce the impact of randomness in disease transmission and initialization. A summary of the covariates (input features) and response variables used in our machine learning pipeline is provided in Table 1.

## Substitutional ML and DL algorithms

We have employed a multioutput regression methodology that simultaneously predicts the full time series  $I(t) = [I_1, I_2, \dots, I_{100}]$ , where  $I_t$  denotes the number of infected individuals on day  $t$ . This vectorized formulation enables the model to learn the

temporal evolution of infections from static network features. The same approach is applied for predicting severe and critical cases over time. We implemented it for several ML algorithms and compared their performance: the support vector machine (SVM) algorithm [42]; the Random Forest (RF) algorithm [43]; and the scalable tree boosting algorithm (Xgboost) [44]. We employ grid search for hyperparameter tuning with mean  $R^2$  score as the optimizing criterion [45–50]. The set of hyperparameters and range of values considered in the grid search is shown in Table 2. The synthetic dataset was split into disjunct training (75%) and test (25%) subsets and training was performed w.r.t. the target variables  $I(t)$ .

Model performance was measured using the conventional  $R^2$  squared metrics [51–53], to assess the goodness-of-fit of our predictive models. The  $R^2$  score measures the proportion of the variance in the dependent variables that our models can elucidate.

Additionally, a previous split of 25% is made on the original dataset, reserved for final testing after the model is trained on the 10-fold cross-validation. This technique evaluates model performance while minimizing overfitting and ensuring generalization to new data.

Moreover, we implemented a fully connected neural network as deep learning model (DL). To efficiently tune the hyperparameters of this model, we used the random search optimization algorithm, which involves randomly sampling hyperparameters from a predefined range and evaluating the model's performance for each set [54]. This method effectively identifies satisfactory hyperparameter configurations without requiring exhaustive searches across the entire hyperparameter space [55]. The high dimensionality and complex interplay of hyperparameters in deep learning render traditional grid search impractical [56, 57].

Additionally, we applied dropout and L2 regularization techniques to mitigate overfitting. Dropout involves temporarily deactivating a random subset of neurons during each training iteration, preventing the model from becoming overly dependent on specific neurons and thereby enhancing generalization [58]. L2 regularization, also known as weight decay, was applied to the weights of the neural network layers. Additionally, we applied dropout and L2 regularization techniques to mitigate overfitting. Dropout involves temporarily deactivating a random subset of neurons during each training iteration, preventing the model from becoming overly dependent on specific neurons and thereby enhancing generalization [58]. L2 regularization was applied to penalize larger weight values and to reduce the risk of overfitting [59]. The architecture and random search hyperparameters for the DL algorithm are summarized in Table 3.

We consistently applied a uniform data sampling strategy across all machine learning and deep learning algorithms, using a 10-fold cross-validation approach with shuffling;  $k=10$  is a common value for this method [60–64]. This technique involves

**Table 1.** Summary of machine learning input features (covariates) and prediction targets (responses).

Type	Variable(s)	Description
Covariates	Mean PR	Average PageRank centrality of all nodes in the static contact network
	Mean CC	Average closeness centrality across all nodes
	Mean BC	Average BC across all nodes
	Mean EC	Average EC across all nodes
	Mean Degree (D)	Average degree (number of direct connections) per node
	Mean KC Index	Average k-core number, reflecting the cohesiveness of the network
Responses	$I(t)$	Time series of infected individuals from day 1 to 100, averaged across five stochastic simulation runs
	$I_{\text{severe}}(t)$	Time series of severe cases (daily counts, averaged over five runs)
	$I_{\text{critical}}(t)$	Time series of critical cases (daily counts, averaged over five runs)

**Table 2.** Hyperparameters and best values found for different machine learning models.

Model	Hyperparameter	Values tested	Description	Best value
SVM	C	0.1, 1, 10	Regularization parameter	10
	Kernel	linear, rbf	Kernel function	rbf
RF	Max depth	10, 20, 30, 40, 50	Maximum depth of each tree	5
	Max features	2, 3, 4	Maximum number of features for the best split	4
MLP	n estimators	100, 200, 300	Number of trees in the forest	50
	Hidden layer sizes	(100,), (100, 50, 100)	Neurons in the hidden layers	(100, 50, 100)
XGBoost	Alpha	0.0001, 0.001, 0.01	L2 penalty term	0.01
	Solver	adam, sgd, lbfgs	Weight optimization solver	lbfgs
	n_estimators	100, 200, 300	Number of boosting rounds	300
XGBoost	Learning rate	0.01, 0.1, 0.2	Step size shrinkage	0.1
	Max depth	3, 5, 7	Maximum depth of a tree	5

**Table 3.** DL algorithm architecture, hyperparameter search space, and best values found.

Layer	Hyperparameter	Search space	Best value
Dense (1)	Units	8 to 64 (step 4)	48
	Activation	relu, tanh, sigmoid	relu
	L2 Regularization	1e-6 to 1e-2 (log)	1.49e-6
Dense (2)	Units	16 to 128 (step 4)	104
	Activation	relu, tanh, sigmoid	tanh
	L2 Regularization	1e-6 to 1e-2 (log)	2.74e-4
Dense (3)	Use Dropout	True/False	False
	Dropout Rate	0.1 to 0.6 (step 0.01)	0.33
	Units	16 to 128 (step 4)	124
Dense (4)	Activation	relu, tanh, sigmoid	sigmoid
	L2 Regularization	1e-6 to 1e-2 (log)	1.31e-5
	Use Dropout	True/False	False
Dense (5)	Dropout Rate	0.1 to 0.6 (step 0.01)	0.30
	Units	16 to 128 (step 4)	108
	Activation	relu, tanh, sigmoid	sigmoid
Dense (5)	L2 Regularization	1e-6 to 1e-2 (log)	5.43e-4
	Use Dropout	True/False	True
	Dropout Rate	0.1 to 0.6 (step 0.01)	0.26
Dense (5)	Units	16 to 128 (step 4)	36
	Activation	relu, tanh, sigmoid	relu
	L2 Regularization	1e-6 to 1e-2 (log)	4.58e-6
Output	Use Dropout	True/False	True
	Dropout Rate	0.1 to 0.6 (step 0.01)	0.12
	Units	Equal to number of target variables	-
Optimizer	Learning rate	1e-4 to 1e-1 (log)	3.03e-4

partitioning the dataset into ten equitable folds, ensuring each fold represents a fair portion of the data. Random shuffling ensures that each batch contains a different mix of data points in every iteration. Shuffling the data before partitioning reduces the risk of any systematic patterns in the dataset influencing the performance of the model. This process is suitable to reduce a possible bias, enhance the robustness of the trained models and to improve generalization capabilities to unseen data [65–67].

Furthermore, per established preprocessing best practices, our pipeline incorporates the application of a standard scaler to normalize the features before training our multioutput model. Following the approach [68, 69], we employed this scaler to standardize the features by centering them through the removal of the mean and subsequently scaling to attain unit variance. This normalization procedure proves pivotal, especially for algorithms that hinge on the assumption of feature homogeneity. Both the input features ( $X$ ) and the corresponding multioutput target variables ( $y$ ) are subjected to this scaling protocol, which is recognized for its capacity to enhance the suitability of the data for a

diverse array of machine learning algorithms and optimization methodologies. This uniformity in scaling fosters heightened model convergence and stability and culminates in demonstrably enhanced performance.

The SHapley Additive exPlanations (SHAP) values were calculated to assess the predictive contribution of individual variables, the complex network metrics described in Subsection II A. The SHAP value methodology, rooted in cooperative game theory and popularized by Lundberg and Lee [70], allows to conclude on the relative contributions of each input feature (here: the complex network measures) to the model prediction (here: prediction variables  $I(t)$ ). For details see III A.

### Assessing the manifold of minimal intervention

For illustration, we consider a use case where disease control should be done with minimal restrictions but within a set of constraints set by the health system. Such constraints on the maximally acceptable number of infected, hospitalized or critically sick patients are given externally. For instance, the number of available beds in intensive care units can be a limiting factor which requires to keep the number of critically sick patients below that threshold. That goal, however, may be equivalently reachable by different contact network restrictions applied to one or several subgraphs of the contact network. Therefore, we consider the manifold of equivalent and minimal interventions which all imply a disease dynamics compatible with the set constraints.

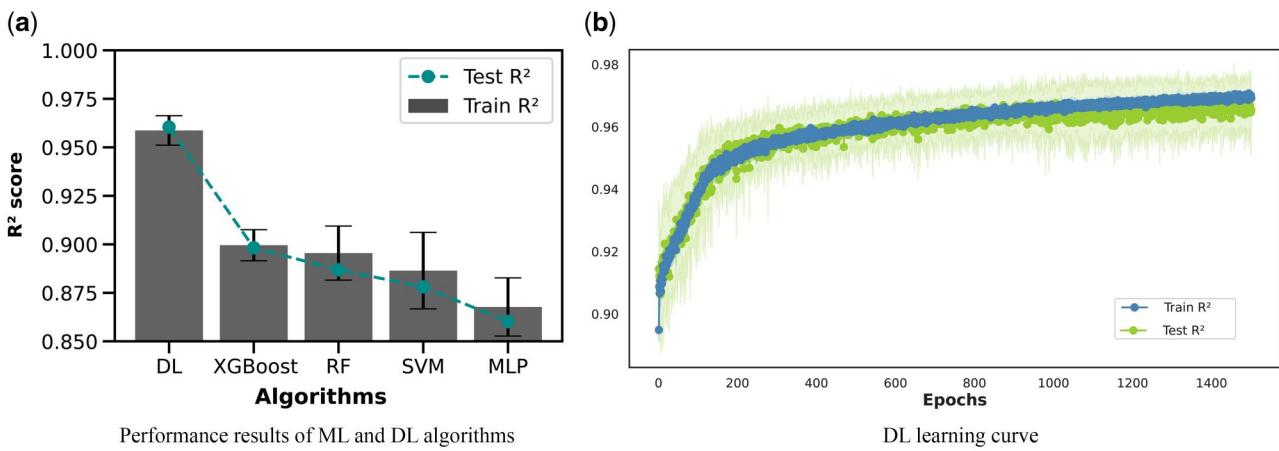
Motivated by this use case, we restrict our target variables (time series data on disease dynamics) to the maximum value of critically sick patients at any point in prediction time to facilitate the assessment of optimal interventions. We consider the three-dimensional parameter space defined by the strength of the restrictions imposed on the school, community, or work subgraphs and, as a single target value, the maximum number of the critically sick. The trained deep learning model was used to refine the parameter space discretization and to extend the synthetic dataset. Cubic splines [71] were used to finally interpolate the manifold [72] between available data points.

Figure 1 depicts our methodology summary scheme.

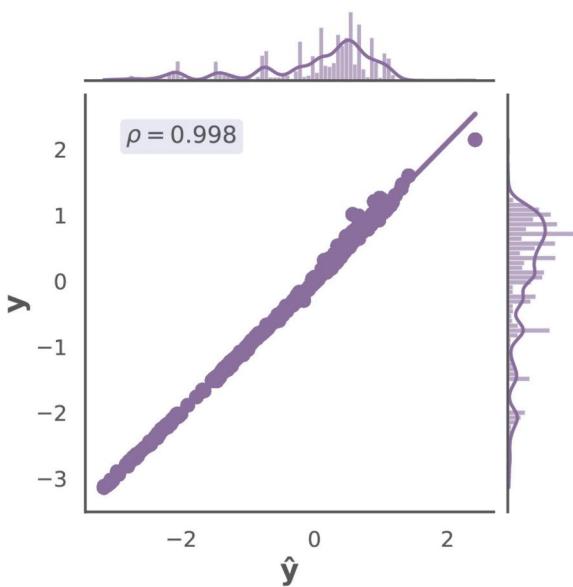
## Results

### Machine learning and deep learning results

Figure 2a depicts the results of ML and DL algorithms. By construction, we verified that the time series of observables can be predicted directly from complex network measures as input features, bypassing a full agent-based simulation on a contact



**Figure 2.** Performance Comparison of ML and DL Algorithms. (a)  $R^2$  -performance for ML and DL algorithms. The grey bar plot represents each algorithm's training performance, including error bars. The blue curve indicates the  $R^2$  performance for the test data. (b) The learning curve for the optimized DL model—the plot displays the mean  $R^2$  scores for both train (blue) and test (green) sets, shaded regions the standard deviation



**Figure 3.** Comparison of flattened predicted time series and flattened original data for DL model. The figure presents the flattened predicted time series ( $\hat{y}$ ) compared to the flattened original data ( $y$ ) across various algorithms. Both  $\hat{y}$  and  $y$  were normalized using standard scalers before the comparison. The cosine similarity ( $\rho$ ) was calculated for each comparison to quantify the similarity between the predicted and original series

network, with reasonable accuracy; all algorithms obtained  $R^2$  test higher than 85%. The DL algorithm achieved the best performance, with a mean train  $R^2$  of  $0.9587 \pm 0.0076$  and a test  $R^2$  of 0.9605.

Figure 2b illustrates the learning curves over epochs for the DL model. An epoch represents one complete pass through the entire training dataset. The curve shows the development of the model performance with ongoing training, that is, updates to the trainable weights over multiple iterations. Both the training and test performances of the DL models increase until convergence, indicating a stable and well-generalized model.

The convergence of learning curves in DL models indicates that this model achieved stable performance, which is crucial for reliable predictions and generalization to new data.

Further, we also fit the flattened predicted time series found ( $\hat{y}$ ) and compared it to the flattened original data ( $y$ ), as shown in Fig. 3. It can be seen that the cosine similarity between the predicted values and the original data is higher than 99%, demonstrating the high predictive accuracy of our DL model. We conclude that complex network measures are a suitable candidate for a reduced representation of contact networks which still maintains predictive power.

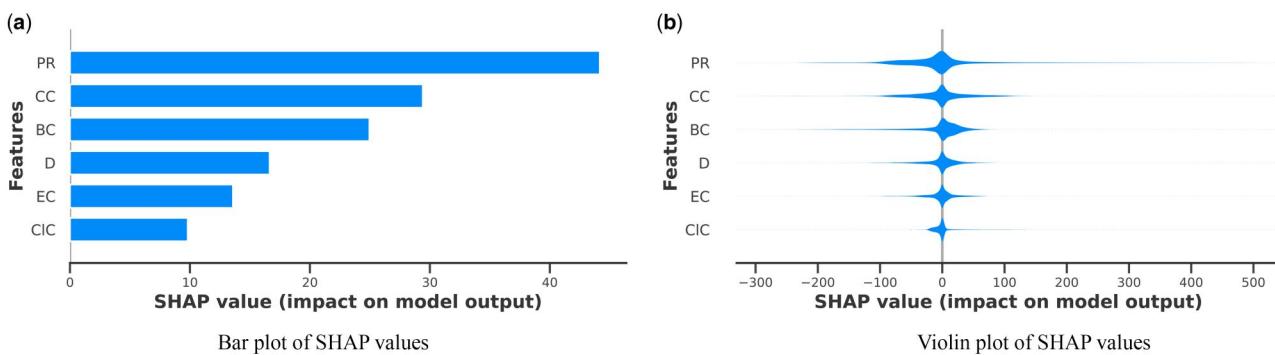
Since the DL model obtained the best performance, the SHAP value analysis is applied to this model, depicted in Fig. 4, revealing the relative importance of various complex network metrics in predicting the number of infected individuals over time,  $I(t)$ . The bar plot (Fig. 4a) indicates that PR has the highest impact on the model's output, suggesting that the influence of a node, considering the importance of its neighbors, plays a crucial role in understanding the spread of infection. Following PR, CC is also highly influential, highlighting the importance of nodes' accessibility to others within the network.

The violin plot (Fig. 4b) complements the bar plot by showing each feature's distribution of SHAP values. PR and CC have broader distributions, indicating varying impacts across different scenarios. This variability suggests that while these features are generally significant, their influence can differ significantly depending on the specific network structure and intervention scenario. The distributions for BC, D, EC, and KC are narrower, indicating more consistent impacts across different scenarios.

Further, the DL model is applied to time series for critical and severe patients and the infected patients' time series. Table 4 depicts the results of DL to the all  $I(t)$  curves in which similar results can be obtained for the infected patients. This consistency suggests that the DL models achieve stable performance across different patient categories, which is crucial for reliable predictions and generalization to new data. Further, the severe and critical time series learning curves can be seen in Supplementary Appendix B, whose curves depicted the convergence and stability of the DL model for different curves.

### Assessing the manifold of minimal intervention

As described in Subsection II D, using spline interpolation in the intervention space ( $S, c, w$ ) with the maximum number of critically sick as the target variable, we mapped out the manifold of minimal intervention. Figure 5 displays three cross sections through 3D parameter space. It illustrates the manifold for



**Figure 4.** SHAP value analysis for the deep learning model. (a) The bar plot shows the mean SHAP values for each complex network metric, indicating their average impact on model predictions. PR has the highest impact, followed by CC, BC, D, EC, and KC. (b) The violin plot shows the distribution of SHAP values for each feature, illustrating the variability in their impacts across different scenarios. Broader distributions for PR and CC suggest higher variability, while narrower distributions for BC, D, EC, and KC indicate more consistent impacts.

**Table 4.** DL performance to different  $I(t)$  curves.

$I(t)$	Train Mean $R^2$	Test $R^2$
Infected	$0.9587 \pm 0.0076$	0.9605
Severe	$0.9738 \pm 0.0037$	0.9824
Critical	$0.9841 \pm 0.0027$	0.9820

different public health interventions which restrict contacts in any combination of the dimensions community, work, and school and allows to read off the effectiveness of actions. Various intervention strategies can be discussed w.r.t. their impact on the maximum number of critical cases: The subfigures (A), (B), and (C) depict the combined impact and the interplay between restrictions in community and work (with school set to 0), community versus school (with work set to 0), and school versus work (with community set to 0), respectively. The color gradients represent the maximum critical cases observed under each intervention scenario, highlighting the regions that result in the lowest and highest number of critical patients. This visualization aids in identifying equivalent interventions which are consistent with a pre-defined maximal number of critical cases. This enables a broader discussion on the optimal public health response in a society. A full 3D visualization of the considered parameter space is depicted in [Supplementary Appendix C](#).

## Discussion

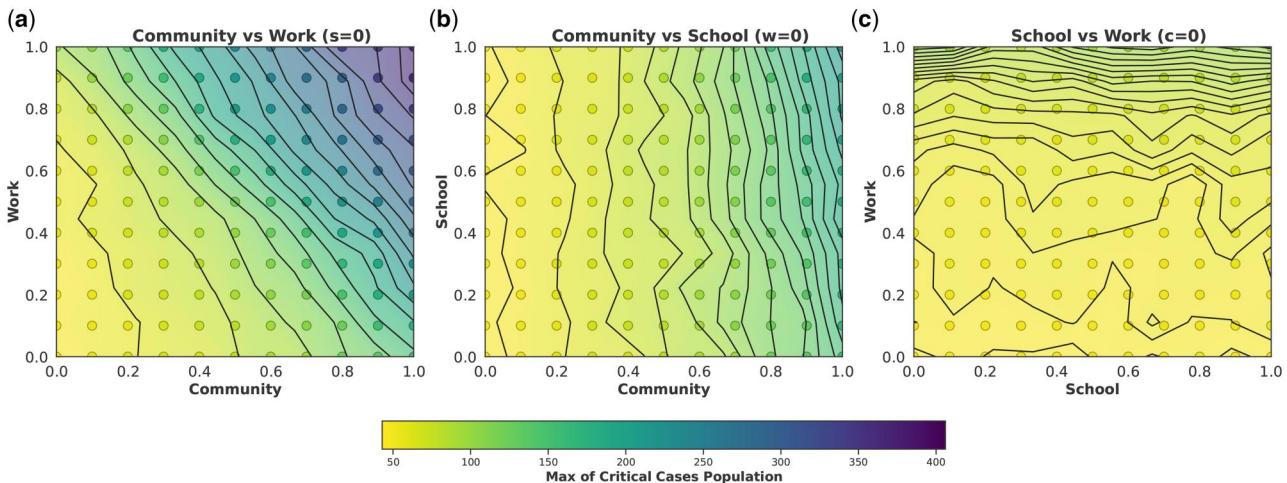
In this study, we first investigated the research question whether a full simulation of disease dynamics using an agent-based simulator is necessary to predict the time series of essential observables like the number of infected, hospitalized or critically sick patients.

By applying machine learning to a set of dimensionality-reduced features of the contact network, defined by complex network measures, we observed that models with equivalent predictive power can be constructed. A DL approach achieved the highest performance, with an  $R^2$ -adjusted value exceeding 95%. This high performance signifies the model's efficiency in predicting the time series of infected, severe, and critical cases within a sufficient approximation. This result confirmed our hypothesis that the topology of the contact network could effectively characterize the impact of the network on the disease dynamics. It is important to clarify that our study aimed to investigate how variations in the topology of contact networks—driven by

interventions targeting public contact layers—affect the progression of an epidemic. Rather than reproducing the full complexity of time-varying behavior or policy adaptation, we focused on whether the structural properties of the network alone could reliably predict disease outcomes. While this necessarily simplifies real-world dynamics, it allowed us to isolate and assess the predictive power of topological features under controlled conditions. Therefore, we successfully abstracted from the full description based on an adjacency matrix of the contact network to effective network metrics; this allows—at least partially—to circumvent the need for computationally expensive agent-based simulations.

The SHAP value results are useful to understand the relative relevance of the applied complex network measures. It demonstrates that centrality measures, mainly PR and CC, are leading in capturing the nuances of disease transmission within the network. PR is vital as it accounts for both the quantity and significance of a node's connections, aiding in identifying influential nodes [73–75]. CC measures how quickly an infection can spread from one node to all others, identifying nodes crucial for rapid dissemination [76, 77]. Our findings align with the study [78], emphasizing the predictive power of combined centrality measures. The study demonstrated that integrating normalized spectral centralities like PR with measures sensitive to graph edges, such as CC, can yield rather high predictive accuracy ( $R^2$  scores of 0.91 or higher) across various graph structures and epidemic parameters. This reinforces the notion that PR's consideration of both the quantity and quality of connections, coupled with CC's ability to measure rapid dissemination potential, makes them highly effective for identifying influential nodes and optimizing intervention strategies. Our use of these centralities indicates their applicability and robustness in network epidemiology.

Further, BC and D also show substantial contributions, emphasizing the roles of nodes in controlling information flow [35, 79] and the number of direct connections, respectively [80]. EC and KC have slightly lower SHAP values but contribute significantly to the model predictions. EC reflects the influence of nodes based on the quality of their connections [37]. At the same time, KC indicates the core-periphery structure of the network, both providing valuable insights into the network's robustness and connectivity [40]. It is important to note that network measures are not intended to be directly measured or acted upon in real-time policy implementation. Rather, they serve as interpretable indicators of how structural changes to contact networks—through public health interventions like school closures, remote work, or gathering bans—fluence epidemic dynamics. In



**Figure 5.** Cross-sections of the interpolated intervention-response surface: (a) community vs. work (school fixed at 0), (b) community vs. school (work fixed at 0), and (c) school vs. work (community fixed at 0). Each subplot shows a 2D slice of the 3D intervention space  $(s, c, w) \in [0, 1]^3$ , where two dimensions vary and the third is held constant. The axes correspond to contact retention levels in the school ( $s$ ), community ( $c$ ), and work ( $w$ ) layers. The surface illustrates the predicted maximum number of critical cases (color-coded) as estimated by the deep learning model. Smooth contours arise from spline interpolation over the model outputs and highlight how different combinations of interventions affect the disease burden

practice, these metrics help identify which types of interventions are structurally impactful, even if the network itself is only partially known or estimated. Future work could explore how real-world proxies (e.g. mobility data or contact surveys) might be used to approximate or track changes in these network properties to support planning and communication.

These findings validate our DL model's robustness and ability to effectively leverage network-based features to predict disease dynamics. This enhances our understanding of intervention strategies and their potential impacts on public health, supporting the development of more effective containment and mitigation measures. While SHAP values helped identify which network metrics most influence the model's predictions, they do not provide causal explanations for how specific interventions (e.g. reducing workplace contacts) lead to epidemiological changes. As deep learning models can function as black boxes, their use in public health decision-making must be approached with care. Future work could integrate more interpretable frameworks—such as hybrid models combining machine learning with Bayesian Networks—to improve transparency and allow clearer communication with policymakers. Our current approach should be viewed as a proof of concept for structural prediction, not a replacement for mechanistic or causally grounded epidemiological models.

A further research question directs toward a better understanding of equivalent public health interventions obtained through contact restrictions in various contexts of society. A critical aspect of our study is to identify the manifold of minimal intervention for a given set of external constraints, illustrated by a maximal number of critically sick patients. Through extensive parameter space sampling and analysis, we mapped out the relative impact of contact restrictions in schools, workplaces and community on disease dynamics (in Fig. 5). This provides a route toward improved tools for policymakers and the public alike to decide on the best strategies based on their specific public health constraints but also on an appropriate sharing of burden. The results indicate that interventions in the community and work layers are equally effective for controlling the pandemic, as demonstrated by the diagonal line representing balanced interventions between these two layers. Specifically, within the

community versus work intervention space (Fig. 5), a coordinated contact reduction in both areas yields the lowest number of critical cases. Community interventions are more effective in reducing critical cases when comparing community and school interventions (Fig. 5). When resources or compliance levels are limited, public health strategies should prioritize community-based measures over school closures. Conversely, the comparison between school and work interventions (Fig. 5) reveals that workplace interventions generate more impact in controlling the pandemic than school-based measures. This may raise questions on the relevance, necessity and appropriateness of large scale school closures to mitigate the spread of the virus. Figure 5 supports the conclusion that both community and work interventions are of comparable impact. This enables mutual trade-offs and allows for a societal debate on which combination of contact restrictions, chosen from equally effective ones, may be most appropriate for disease control in a pandemic situation.

In summary, our findings on the relative importance of contact restricting interventions indicate that among school, community, and work settings, community restrictions have the most significant impact on controlling the disease dynamics. This aligns with the general understanding that casual and frequent interactions in community settings contribute substantially to disease transmission [81, 82]. The analysis also highlights that contact restrictions at workplaces generate a higher impact than school closures. We attribute this, among other aspects, to the high density and regularity of contacts in workplace environments.

As our framework identifies a manifold of equivalent interventions it allows for an attribution of the restriction burden caused by public health interventions to different subgroups in society. However, it is beyond the scope of this simulation to include substantiated predictions on the economic and social implications of chosen public health interventions. We acknowledge that interventions which lead to comparable disease control, may have very different social and economic impacts on various social groups, for example, for low-income workers. Future extensions of our model could interface with economic and social system simulators to include these aspects and arrive at socially justified policy recommendations.

To achieve the new perspective of identifying medically equivalent public health interventions, we had to restrict the model complexity substantially. First, we included only contact restrictions as examples for public health interventions and we model changes to contact networks by just three effective parameters which represent the strength of the interventions. We maintained constant all other parameters available in COVASIM, in particular medical parameters, such as transmission and recovery rates, across all simulations. These parameters can vary significantly depending on local health conditions, population behavior, and emerging virus variants. Hence, our model does currently not account for the impact of other types of interventions like vaccinations, active contact tracing or others that have become crucial in managing pandemics. The exclusion of household contacts might oversimplify the contact dynamics, although we focused on public and communal settings to capture broader transmission patterns. While household transmission has been a significant route of spread, especially during lockdowns, the household layer in COVASIM involves a relatively small number of nodes with highly clustered and repetitive interactions. These contacts are generally stable and less influenced by public interventions than those in schools, workplaces, and communities. Nevertheless, we acknowledge this limitation and suggest that future studies include a sensitivity analysis incorporating household contacts to evaluate their added influence under more complex intervention strategies.

Furthermore, our approach relies heavily on the detection of network modifications stemming from contact restrictions by established complex network measures. While this is the case in the currently depicted contact restriction scenarios, it may happen that, in reality, different modifications of contact networks occur which may not be equally well resolved by complex network measures. A test of the approach with real network data is, unfortunately, beyond the scope of this study but could provide further evidence on the reliability of the chosen route. We mention that we have not exhausted the set of possible complex network measures and that further measures could be added to the set of input features if required.

Future work should extend this framework to include further public health interventions like vaccinations, contact tracing, and behavioral compliance to allow an expert to choose from an even more diverse set of interventions. This framework remains applicable beyond COVID-19, offering a fast and generalizable tool for modeling the spread of emerging infectious diseases and supporting policy planning in future outbreaks or resource-limited settings.

## Conclusion

The handling of the COVID-19 pandemic has highlighted the need to proceed from accurate and efficient models for predicting disease dynamics to expert support systems for guiding public health interventions. This study utilized the COVASIM agent-based model to simulate various scenarios across different social settings—specifically focusing on school, community, and workplace related contact networks. The description of contact networks has been simplified by extracting complex network measures which, despite a substantial reduction in model complexity, showed predictive power on key epidemiological outcomes when used in deep learning. This includes, for instance, the number of infected, severe, and critical cases, with a high degree of  $R^2$  values exceeding 95%.

Our approach demonstrated robust predictive capabilities and provided a framework for identifying optimal intervention strategies. We observed that community and workplace interventions are critical in minimizing the impact of the pandemic, underscoring the importance of targeted public health strategies in these areas. The integration of network analytics with deep learning offers significant advantages in epidemic modeling, including reduced computational costs and enhanced decision-making efficiency.

While our model successfully abstracted the complex dynamics of disease transmission, it is important to acknowledge certain limitations: These include the assumption of constant medical parameters, excluding vaccination effects and others. This underscores the need for ongoing research to incorporate dynamic parameters and a broader range of contact layers, which will further improve the model's applicability and robustness.

In conclusion, this study provides a novel and effective framework for planning most adequate contact restrictions to control infectious disease outbreaks. By leveraging deep learning and network measures, our approach can simplify parameter space searches and offers valuable insights for public health decision-making.

## Supplementary data

Supplementary data are available at *Biology Methods and Protocols* online.

Conflict of interest statement. None declared.

## Funding

None declared.

## Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

1. Zhang H, Song H, Wen L. et al. Forecasting tourism recovery amid covid-19. *Ann Tour Res* 2021;87:103149.
2. Booth AL, Abels E, McCaffrey P. Development of a prognostic model for mortality in covid-19 infection using machine learning. *Mod Pathol* 2021;34:522–31.
3. Barda N, Riesel D, Akivit A. et al. Developing a covid-19 mortality risk prediction model when individual-level data are not available. *Nat Commun* 2020;11:4439.
4. Kerr CC, Stuart RM, Mistry D. et al. Covasim: an agent-based model of covid-19 dynamics and interventions. *PLoS Comput Biol* 2021;17:e1009149.
5. IHME COVID-19 Forecasting Team. Modeling covid-19 scenarios for the united states. *Nat Med* 2021;27:94–105.
6. Rădulescu A, Williams C, Cavanagh K. Management strategies in a SEIR-type model of covid 19 community spread. *Sci Rep* 2020;10:21256.
7. Chang S, Pierson E, Koh PW. et al. Mobility network models of covid-19 explain inequities and inform reopening. *Nature* 2021;589:82–7.
8. Sturniolo S, Waites W, Colbourn T. et al. Testing, tracing and isolation in compartmental models. *PLoS Comput Biol* 2021;17:e1008633.

9. Scarselli D, Budanur NB, Timme M. et al. Discontinuous epidemic transition due to limited testing. *Nat Commun* 2021; **12**:2586.
10. Gao J, Heintz J, Mack C. et al. Evidence-driven spatiotemporal covid-19 hospitalization prediction with Ising dynamics. *Nat Commun* 2023; **14**:3093.
11. Jaouimaa FZ, Dempsey D, Van Osch S. et al. An age-structured SEIR model for covid-19 incidence in Dublin, Ireland with framework for evaluating health intervention cost. *PLoS One* 2021; **16**:e0260632.
12. Cattaneo A, Vitali A, Mazzoleni M. et al. An agent-based model to assess large-scale covid-19 vaccination campaigns for the Italian territory: the case study of Lombardy region. *Comput Methods Prog Biomed* 2022; **224**:107029.
13. Panovska-Griffiths J, Kerr CC, Stuart RM. et al. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second covid-19 epidemic wave in the UK: a modelling study. *Lancet Child Adolesc Health* 2020; **4**:817–27.
14. Scott N, Palmer A, Delport D. et al. Modelling the impact of relaxing covid-19 control measures during a period of low viral transmission. *Med J Aust* 2021; **214**:79–83.
15. Lorig F, Johansson E, Davidsson P. Agent-based social simulation of the covid-19 pandemic: a systematic review. *J Artif Soc Soc Simulat* 2021; **24**:5.
16. Reveil M, Chen YH. Predicting and preventing covid-19 outbreaks in indoor environments: an agent-based modeling study. *Sci Rep* 2022; **12**:16076.
17. Thompson J, McClure R, Blakely T. et al. Modelling sars-cov-2 disease progression in Australia and New Zealand: an account of an agent-based approach to support public health decision-making. *Aust N Z J Public Health* 2022; **46**:292–303.
18. Panovska-Griffiths J, Stuart R, Kerr C. et al. Modelling the impact of reopening schools in the UK in early 2021 in the presence of the alpha variant and with roll-out of vaccination against sars-cov-2. *J Math Anal Appl* 2022; **514**:126050.
19. Bosman M, Esteve A, Gabanelli L. et al. Stochastic simulation of successive waves of covid-19 in the province of Barcelona. *Infect Dis Model* 2023; **8**:145–58.
20. Shastry V, Reeves DC, Willems N. et al. Policy and behavioral response to shock events: an agent-based model of the effectiveness and equity of policy design features. *PLoS One* 2022; **17**:e0262172.
21. Wilmes P, Zimmer J, Schulz J. et al. Sars-cov-2 transmission risk from asymptomatic carriers: results from a mass screening programme in Luxembourg. *Lancet Regional Health–Europe* 2021; **4**:100056.
22. Pham QD, Stuart RM, Nguyen TV. et al. Estimating and mitigating the risk of covid-19 epidemic rebound associated with reopening of international borders in Vietnam: a modelling study. *Lancet Global Health* 2021; **9**:e916–e924.
23. Latkowski R, Dunin-Kplicz B. An agent-based covid-19 simulator: extending Covasim to the polish context. *Proc Comput Sci* 2021; **192**:3607–16.
24. Kerr CC, Mistry D, Stuart RM. et al. Controlling covid-19 via test-trace-quarantine. *Nat Commun* 2021; **12**:2993.
25. Krivorotko O, Sosnovskaia M, Vashchenko I. et al. Agent-based modeling of covid-19 outbreaks for New York state and UK: parameter identification algorithm. *Infect Dis Model* 2022; **7**:30–44.
26. Collier N, North M. Parallel agent-based simulation with repast for high performance computing. *Simulation* 2013; **89**:1215–35.
27. Rodrigues FA, Peron T, Connaughton C. et al. A machine learning approach to predicting dynamical observables from network structure. *arXiv preprint. arXiv: 1910.00544* 2019.
28. Krebs T, Moeckel MJ. Partial lockdown on unvaccinated individuals promises breaking of fourth covid-19 wave in bavaria. *medRxiv*, 2021.
29. Krebs T, von Jouanne-Diedrich H, Moeckel MJ. Covid-19 scenarios for comparing the effectiveness of age-specific vaccination regimes, exemplified for the city of aschaffenburg (Germany). *medRxiv*, 2021.
30. Haroon S, Chandan JS, Middleton J. et al. *Covid-19: breaking the chain of household transmission*. London, UK: Elsevier, 2020.
31. Wang Y, Xiong H, Liu S. et al. Simulation agent-based model to demonstrate the transmission of covid-19 and effectiveness of different public health strategies. *Front Comput Sci* 2021; **3**:642321.
32. Kucharski AJ, Russell TW, Diamond C. et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *Lancet Infect Dis* 2020; **20**:553–8.
33. Bhattacharya P, Chen J, Hoops S. et al. Data-driven scalable pipeline using national agent-based models for real-time pandemic response and decision support. *Int J High Perform Comput Appl* 2023a; **37**:4–27.
34. Ferguson NM, Laydon D, Nedjati-Gilani G. et al. Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, Vol. 16. London: Imperial College London, 2020.
35. Freeman LC. A set of measures of centrality based on betweenness. *Sociometry* 1977; **40**:35.
36. Freeman LC. Centrality in social networks conceptual clarification. *Soc Netw* 1978; **1**:215–39.
37. Bonacich P. Power and centrality: a family of measures. *Am J Sociol* 1987; **92**:1170–82.
38. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst* 1998; **30**:107–17.
39. Doyle J, Graver J. Mean distance in a graph. *Discrete Math* 1977; **17**:147–54.
40. Seidman SB. Network structure and minimum degree. *Soc Netw* 1983; **5**:269–87.
41. Newman M. *Networks: An Introduction*. Oxford: Oxford University Press, 2010.
42. Bottou L, Lin CJ. Support vector machine solvers. *Large Scale Kernel Mach* 2007; **3**:301.
43. Breiman L. Random forests. *Mach Learn* 2001; **45**:5–32.
44. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd international Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2016, 785–94.
45. Sato M, Morimoto K, Kajihara S. et al. Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma. *Sci Rep* 2019; **9**:7704.
46. Zhong Z, Yuan X, Liu S. et al. Machine learning prediction models for prognosis of critically ill patients after open-heart surgery. *Sci Rep* 2021; **11**:3384.
47. Arcadu F, Benmansour F, Maunz A. et al. Author correction: deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med* 2020; **3**:160.
48. Alves CL, Toutain TGDO, de Carvalho Aguiar P. et al. Diagnosis of autism spectrum disorder based on functional brain networks and machine learning. *Sci Rep* 2023; **13**:8072.
49. Alves CL, Cury RG, Roster K. et al. Application of machine learning and complex network measures to an eeg dataset from aya-huasca experiments. *PLoS One* 2022; **17**:e0277257.

50. Alves CL, Pineda AM, Roster K. et al. Eeg functional connectivity and deep learning for automatic diagnosis of brain disorders: Alzheimer's disease and schizophrenia. *J Phys Complex* 2022; **3**:025001.

51. Jenkins DG, Quintana-Ascencio PF. A solution to minimum sample size for regressions. *PLoS One* 2020; **15**:e0229345.

52. Car Z, Baressi Šegota S, Anelić N. et al. Modeling the spread of covid-19 infection using a multilayer perceptron. *Comput Math Methods Med* 2020; **2020**:5714714.

53. Menard S. Coefficients of determination for multiple logistic regression analysis. *Am Stat* 2000; **54**:17–24.

54. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012; **13**:281–305.

55. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 2020; **415**:295–316.

56. Roy S, Mehera R, Pal RK. et al. Hyperparameter optimization for deep neural network models: a comprehensive study on methods and techniques. *Innov Syst Softw Eng* 2023; **1**:1–12.

57. Mallik N, Bergman E, Hvarfner C. et al. Priorband: practical hyperparameter optimization in the age of deep learning. *Adv Neural Inform Proc Syst* 2023; **36**:7377–91.

58. Srivastava N, Hinton G, Krizhevsky A. et al. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; **15**:1929.

59. Ng AY. Feature selection, l1 vs. l2 regularization, and rotational invariance. In: *Proceedings of the Twenty-first International Conference on Machine learning*, 2004, p. 78.

60. Berrar D. Cross-validation. 2019, 542–5.

61. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. *J Mach Learn Res* 2004; **5**:1089.

62. Shah AA, Khan YD. Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. *Sci Rep* 2020; **10**:16913.

63. Kawamoto T, Kabashima Y. Cross-validation estimate of the number of clusters in a network. *Sci Rep* 2017; **7**:3327.

64. Chan J, Rea T, Gollakota S, Sunshine JE. Contactless cardiac arrest detection using smart devices. *NPJ Digit Med* 2019; **2**:52.

65. Farashahi S, Rowe K, Aslami Z. et al. Feature-based learning improves adaptability without compromising precision. *Nat Commun* 2017; **8**:1768.

66. Kuhn M, Johnson K. *Applied Predictive Modeling*, Vol. **26**. Berlin, Germany: Springer, 2013.

67. Brownlee J. How to choose a feature selection method for machine learning. *Mach Learn Mastery* 2019; **10**:1–7.

68. Thippa Reddy G, Swarna Priya RM, Parimala M. et al. A deep neural networks based model for uninterrupted marine environment monitoring. *Comput Commun* 2020; **157**:64–75.

69. Hartono NTP, Thapa J, Tiihonen A. et al. How machine learning can help select capping layers to suppress perovskite degradation. *Nat Commun* 2020; **11**:4172.

70. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. Cambridge, MA: The MIT Press, 2017, 30.

71. Knott GD. *Interpolating Cubic Splines*, Vol. **18**. Berlin, Germany: Springer Science & Business Media, 1999.

72. Chand AKB. Vijay Convexity-preserving rational cubic zipper fractal interpolation curves and surfaces. *Math Comput Appl* 2023; **28**:74.

73. Zhang B, Zhao X, Nie J. et al. Epidemic model-based network influential node ranking methods: a ranking rationality perspective. *ACM Comput Surveys* 2024; **56**:1.

74. Bhattacharya R, Nagwani NK, Tripathi S. Detecting influential nodes with topological structure via graph neural network approach in social networks. *Int J Inf Technol* 2023b; **15**:2233–46.

75. Ahmad A, Ahmad T, Bhatt A. Hwsmcb: a community-based hybrid approach for identifying influential nodes in the social network. *Phys A Stat Mech Appl* 2020; **545**:123590.

76. Lv Z, Zhao N, Xiong F. et al. A novel measure of identifying influential nodes in complex networks. *Phys A Stat Mech Appl* 2019; **523**:488–97.

77. Salavati C, Abdollahpouri A, Manbari Z. Ranking nodes in complex networks based on local structure and improving closeness centrality. *Neurocomputing* 2019; **336**:36–45.

78. Bucur D, Holme P. Beyond ranking nodes: predicting epidemic outbreak sizes by network centralities. *PLoS Comput Biol* 2020; **16**:e1008052.

79. Newman ME. A measure of betweenness centrality based on random walks. *Soc Netw* 2005; **27**:39–54.

80. Hébert-Dufresne L, Noël PA, Marceau V. et al. Propagation dynamics on networks featuring complex topologies. *Phys Rev E Stat Nonlin Soft Matter Phys* 2010; **82**:036115.

81. Arthur RF, Gurley ES, Salje H. et al. Contact structure, mobility, environmental impact and behaviour: the importance of social forces to infectious disease dynamics and disease ecology. *Phil Trans R Soc B* 2017; **372**:20160454.

82. Pinter-Wollman N, Jelić A, Wells NM. The impact of the built environment on health behaviours and disease transmission in social systems. *Phil Trans R Soc B* 2018; **373**:20170245.