



OPEN

Forecasting virus outbreaks with social media data via neural ordinary differential equations

Matías Núñez^{1,2,3}, Nadia L. Barreiro⁴, Rafael A. Barrio⁵ & Christopher Rackauckas^{6,7,8}


During the Covid-19 pandemic, real-time social media data could in principle be used as an early predictor of a new epidemic wave. This possibility is examined here by employing a neural ordinary differential equation (neural ODE) trained to forecast viral outbreaks in a specific geographic region. It learns from multivariate time series of signals derived from a novel set of large online polls regarding COVID-19 symptoms. Once trained, the neural ODE can capture the dynamics of interconnected local signals and effectively estimate the number of new infections up to two months in advance. In addition, it may predict the future consequences of changes in the number of infected at a certain period, which might be related with the flow of individuals entering or exiting a region. This study provides persuasive evidence for the predictive ability of widely disseminated social media surveys for public health applications.

During a pandemic, the capacity to recognize and anticipate local viral outbreaks is critical for health experts to take proper action^{1,2}. However, the intrinsic parameters utilized by the prediction models to reflect the biological features of the virus cannot be determined until the pandemic has happened. While a pandemic is in progress, parameter estimation is fraught with uncertainty, which means that the first-principles models that rely on them inherit this uncertainty in their forecasts. According to one epidemiologist quoted in the New York Times³: “*You tell me what numbers to put in my equations, and I’ll give you the answer ...But you can’t tell me the numbers, because nobody knows them...*”, a statement that illustrates the difficulties that currently exist in predicting new infections during a pandemic.

A large amount of data is being created, either directly or indirectly, on the virus’ spread. Health Surveillance is a vital tool for forecasting, preventing, and eliminating infectious diseases and epidemics. Some of this information has been utilized for a long time in this field^{4–6}. Multiple passive^{4,7,8} and active surveillance systems^{9–11} had been established across the globe^{8,12}. In the last decade, new machine learning algorithms and vast data availability have enabled web-based surveillance as a supplement to conventional approaches^{13–15}. Internet searches^{2,16–18}, social media^{19–22}, survey data^{23,24}, contact tracking or monitoring using mobile devices^{25,26}, and contact simulations²⁷ are examples of information sources.

While digital surveillance systems offer several benefits, such as low cost, fast deployment, and extensive area coverage, they also have numerous disadvantages. For example, content and demographic bias, incomplete information, difficulty segmenting by geographic area, and complexity in digital data structure are obstacles that must be overcome^{28–30}. In addition, the use of predictive models based only on big data may be hampered by lack of reproducibility, overfitting, and the necessity to continuously update algorithms (i.e., to adjust to changes in search engines)³¹. In recent years, there has been widespread agreement on the need to mix old and new information sources to enhance health surveillance model forecasts^{31–33}.

The development of the COVID-19 pandemic has rekindled the hunt for more accurate forecasting models^{34–36}. Large technology businesses have simultaneously made resources accessible to scholars and policymakers^{37,38}. Using this information in conjunction with late indicators (such as virus positive case numbers) might be crucial to the development of more accurate outbreak prediction algorithms. The Delphi research

¹Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. ²Departamento Materiales Nucleares, Centro Atómico Bariloche, Comisión Nacional de Energía Atómica (CNEA), Bariloche, Argentina. ³Ecología cuantitativa, Instituto de Investigaciones en Biodiversidad y Medioambiente, Bariloche, Argentina. ⁴Instituto de Investigaciones Científicas y Técnicas para la Defensa (CITEDEF), Buenos Aires, Argentina. ⁵Instituto de Física, Universidad Nacional Autónoma de México, Apartado Postal 20-365, México 04510, Mexico. ⁶Computer Science & Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, MA 02142, USA. ⁷JuliaHub Inc., Cambridge, MA, USA. ⁸Pumas-AI, Baltimore, MD, USA. email: matias.nunez2@gmail.com

group at Carnegie Mellon University created an API³⁹ that gathers data from several sources and provides access to numerous COVID-specific and syndromic markers. This approach solves some of the aforementioned obstacles by providing simple access to geographic location and structured information. In addition, they presented a survey of symptoms developed in collaboration with universities and public health experts and presented via Facebook⁴⁰. This social media-advertised survey reaches a wider audience and collects more information than comparable surveys. Despite the fact that the majority of web-based surveys are inexpensive and provide quick and easy access to the population, they are limited by sample bias⁴¹. In this case, bias was partially corrected by means of weight adjustment⁴⁰.

It is evident from the preceding that digital information paired with first principles models or data-driven models might be used to enhance health surveillance systems. Several models have been reported to date that use various machine learning techniques and information sources^{42–45} as well as compartment modelling approaches^{46–49}. In this article, we examine the predictive abilities of a state-of-the-art data-driven model to anticipate COVID-19 outbreaks using data-sets from the aforementioned symptoms survey⁴⁰. In this case, time-dependent signals are retrieved from a particular geographic region. After processing several numerical indicators arising from the survey questions, each signal is generated. The acquired data pertains to the symptoms of individuals, infections within their social circle, hospital visits, number of internet searches for COVID-19, and average time away from home, among others. For example, a person's input about the number of his contacts who tested positive for COVID will be linked to the number of new cases in his location.

There is no obvious model based on fundamental principles that connects the survey components to the COVID-19 numbers. However, it is plausible to predict a correlation between the local variation in time of survey answers for an area and the emergence of new viral infections in that location. In addition, these signals have the potential to serve as early indicators^{44,50}, since they are not susceptible to delays caused by officially reported variables, local policies, or testing capacity.

In order to discover this relationship, a neural ordinary differential equation (neural ODE⁵¹) was employed to parameterize the temporal rate of change of the signal. This object employs a parameterized universal approximator to represent all conceivable phase space dynamics with a limited set of parameters that can be learnt from the training data. In this study, the neural ODE is trained on these possible early indicators and is capable of predicting viral outbreaks two months in advance. In addition, once taught, these phase space methods allow the prediction of potential future scenarios or the measurement of the uncertainty associated with changes in the number of infected in the region.

This article is divided as follows: Section “COVID-19 symptom surveys through facebook” describes the surveys and signals, as well as the arguments in favor of the notion that these signals may be utilized as early indicators. The Section “Models: first principles and data driven” provides an overview of neural ODEs and how they are used in this study. Section “Methods” describes the precise methods for incorporating the data into the neural ODEs, while “Results and discussion” section displays the predictive capability of the neural ODEs when used in this way. Finally, we examine the implications of applying these machine learning algorithms and data to health care statistics.

COVID-19 symptom surveys through facebook. Since April 2020 universities and public health officials, in collaboration with Facebook, have been conducting a massive daily survey to monitor the spread and impact of the COVID-19 pandemic in the United States. The survey⁵² is an ongoing operation that is advertised through Facebook's platform and is taken by nearly 55,000 people every day. Respondents provide information about COVID-related symptoms, contacts, prior medical conditions, risk factors, mental health, demographics and the economic effects of the pandemic. The information allows researchers to examine county-level trends across the US. Around 16 million responses have been collected so far.

The survey has four sections and it contains 35 questions. The first section gathers information about a set of symptoms used to define a condition called COVID-like illness (CLI), defined as fever of at least 100 °F, along with shortness of breath, difficulty breathing or a cough⁵³. Two key quantities are estimated with this information, for a given location and day:

1. The percentage of people with CLI,
2. The percentage of people who know someone in their local community with CLI illness (CLI-in-community).

The second section provides further information regarding testing, symptoms, and medical-seeking behavior. The third portion collects information on contacts and risk factors, while the fourth component collects demographic information. There is a sample of the exact questions asked in the supplemental materials. The numerical indicators (signals) are extracted from the collection of questions⁵² and responses, after a bias correction via weight adjustment⁴⁰, resulting in a set of time series (one for each indicator) at a specified location. The aggregated data is accessible to the public via the Delphi Group websites^{39,54}.

Surveys as early indicators. Data pertaining to the number of people who self-report CLI symptoms in a certain location may provide an early signal of COVID activity in that location. In addition, the information is not susceptible to reporting delays, unlike the formal testing metrics of confirmed daily COVID-19 cases, which are affected by testing policy and capacity.

In the Delphi Group's⁴⁰ Blog, it is shown that the “CLI-in-community” signal increases concurrently with confirmed COVID-19 cases, providing evidence that survey-based CLI signals can serve as early indications of COVID activity. Indeed, more individuals report that others in their neighborhood are ill when COVID-19 tests reveal an increase in confirmed cases. In fact, when COVID-19 testing indicate an increase in cases, more

people report that others in their community are ill. Intriguingly, the signal begins to rise dramatically days before COVID-19 cases begin a sharp increase. This study is a non-formal examination of the indicator's recall that permits the use of noisy and indirect signals as early indicators of new cases. Although the survey cannot be used to draw definitive conclusions about the true prevalence of coronavirus disease in the studied region, changes in self-reported symptoms over time could still be a meaningful reflection of the changes in coronavirus infections over time and could therefore assist in forecasting changes in the number of newly infected patients in the coming days.

Models: first principles and data driven. The use of these signals for the prediction of new cases could be done by means of a model that relates the rate of variation of the different indicators to the model's state variables. However, unlike the case of common epidemiological models, the deduction of a quantitative expression that relates the new cases as a function of the different signals extracted from the surveys is far from obvious. Even if the relations were discovered and the model was characterized, for instance, as a system of ordinary differential equations, it would undoubtedly contain unknown parameters and be subject to uncertainty. Utilizing the quantity of data and indications gathered from the surveys, a data-driven method would be a reasonable alternative for obtaining a model for forecasting new infected cases in a geographical region.

Therefore, for a particular region, we establish a vector $\vec{y}(t)$ with a sufficient collection of indicators / variables as components (including the number of new cases) and describe the model via a function that approximates the vector's temporal evolution. With such a function, the number of new cases is expressed as a function of time, empowering prediction. In the case of the classic SIR compartment model⁵⁵, for instance, the vector components are the variables number of susceptible individuals (S), number of infected individuals (I), and number of recovered individuals (R), along with their temporal variation expressed with an ordinary differential equation based on intuition and qualitative knowledge of the dynamics of contagions. However, in the case of the vector produced with the survey indicators as components, we cannot simply build such a model as it is not clear how to characterize the link between them from first principles.

Neural ordinary differential equations. Despite the absence of a known functional form that links the variables, we might examine their temporal rate of change for information. If they are represented by the vector \vec{y} and it changes by $\Delta\vec{y}$ during a time interval Δt , then the rate of variation can be expressed as $\Delta\vec{y}/\Delta t$. If Δt is sufficiently small, it may be expressed as $d\vec{y}/dt$. Now, this expression could be approximated using a parametrized function, and if a neural network NN is used, it would match perfectly to the definition of a neural ordinary differential equation (neural ODE)⁵¹.

A Neural ODE is a neural network parametrization of an ordinary differential equation which allows for learning the dynamics of any possible dynamical system due to the universal approximation theorem^{56,57} (assuming a sufficiently large neural network). In particular, we represent our dynamical system via:

$$\frac{d\vec{y}}{dt} = NN(\vec{y}, t, \theta), \quad (1)$$

where NN is a neural network given by weights θ . This neural network has an explicit t dependence since it is parameterized based on the time-dependent input signals from the data. The goal is to learn the underlying dynamics of change. The "forward pass" through a neural ODE is equivalent to solving an initial value problem where $\vec{y}(t_0)$ represents the input features and a neural network substitutes hand-crafted equations. A single forward pass gives us an entire trajectory.

Unlike other architectures used for time series like residual neural networks (RNNs)⁵⁸ or Long short-term memory (LSTMs)⁵⁹, this model is continuous in time, allowing for incorporating non-uniform data and predictions. RNNs and LSTMs are designed for uniform time data and are equivalent to neural ODEs with uniform time steps (see^{60,61}). In this sense, one could augment an RNN or LSTM with interpolations as part of the loss function; however, the accuracy of such a method for representing a continuous object is inferior to that of a differential equation solver with dense internal output (see Ref.⁶² for details). Given the constraints of the problem, it makes the most sense from a mathematical standpoint to represent the equations in this manner.

The parameters of the neural ODE are learned from the data as diagrammed in Fig. 1. The learning process is performed by minimizing the following loss function

$$L(\theta) = \sum_i |\vec{y}(t_i) - \vec{y}_{data}(t_i)|^2, \quad (2)$$

with respect to the networks parameters θ . Here, $\vec{y}_{data}(t_i)$ represents the multivariate time series as a vector whose components are the values from the chosen set of signals at time t_i while $\vec{y}(t_i)$ is given by the numerical solution to Eq. (1). Minimization is performed by gradient-based local methods, specifically ADAM⁶³. Thus in order to perform the minimization the gradient of the loss function with respect to all parameters θ must be computed. Given the large Lipschitz constants seen due to rapid changes during the onset of the growth, the adjoint technique of the original neural ODE publication is potentially unstable on the case of interest^{64–66}, and thus we opted for stabilized techniques, which avoid reverse solving^{67,68}.

The trained model could be applied to different initial conditions than those used during the learning procedure. With sufficient representative data, the neural ODE should be able to approximate the underlying dynamical system, particularly in the phase space region where the data was sampled. It is expected that this description will deteriorate as one moves further away. In particular, the trained neural network describes the vector field $\frac{d\vec{y}}{dt}$ in the sampled region of the n -dimensional space ($n = 5$ for our case, as we used 5 signals for defining the vector

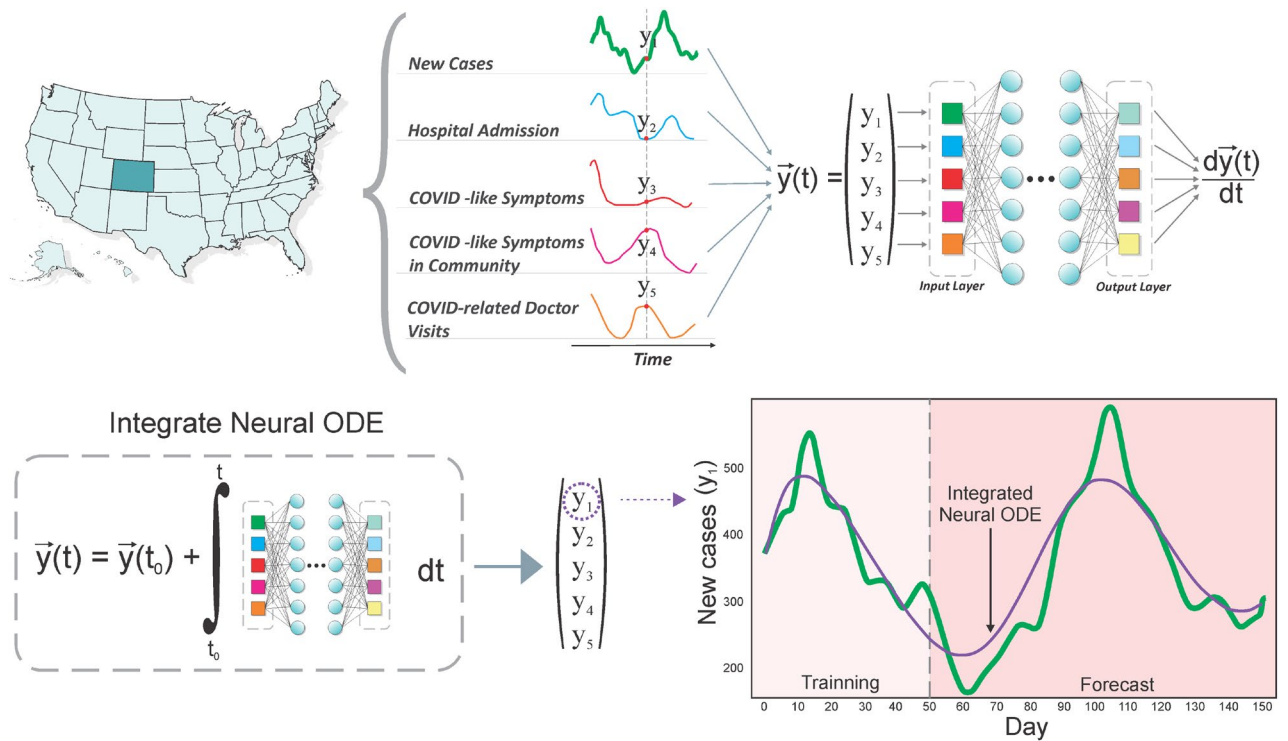


Figure 1. The Neural ODE is trained with a collection of signals/variables gathered from online surveys (shown above). By discovering the ordinary differential equation that best characterizes the data, the trained neural network can capture the dynamics of the temporal variation of the signals. The learned solution, derived by the temporal integration of the neural ODE, is displayed below against the reported data for newly infected cases in CO (signal Y_1). The solution encompasses both the training interval and the forecast.

\vec{y} , see “Methods” section). Now, if we concentrate on the initial condition \vec{y}_0 used for training, the integration of the neural ode from that point defines a flux line (the trajectory) , which is the same flux line that is followed to perform the extrapolation. By modifying the initial condition by a small amount, i.e. $\vec{y}_0 + \Delta\vec{y}$, and integrating from that point, it is anticipated that the new trajectory will be similar to the previous one (see Fig. 2). As $\Delta\vec{y}$ increases, the flow line will move away from the training region and its error with respect to the actual trajectory (as defined by the training set) will increase. Therefore, the main assumption is that the change in the initial

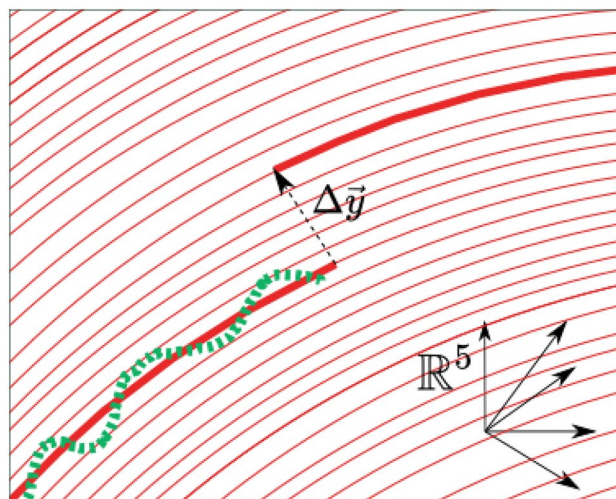


Figure 2. Flux line shift caused by a perturbation. If the state $\vec{y}(t)$ is perturbed while on a learned trajectory (thick red line), it will move to a neighboring flow line by $\Delta\vec{y}$ and continue along the solution. The green dashed line represents the five-dimensional signal extracted from the surveys and utilized to train the neural ODE. These concepts can be used to analyze disturbances in the forecast caused by abrupt changes in the number of new positive cases, which resume to changes in a single \vec{y} coordinate.

condition is small, and this justifies the use of the same neural network to describe the set of initial conditions close to the one used for training. The uncertainty could be estimated by measuring the sensitivity of the model, or the model fit error of the actual data for different $\Delta\bar{y}$. The method and rationale for defining the state \bar{y} will be described in the section that follows.

Methods

The raw signals for each USA State were downloaded using the Delphi Group API^{39,52}. A smoothing was performed via a cubic spline interpolation for all the signals/indicators⁶⁹ time series. The 7-day averaged of reported new confirmed COVID-19 cases was used as the main indicator of interest⁷⁰ for accounting for the new cases. We chose the following set of variables as components in order to build the state vector $y(t)$ for each location:

1. New daily cases (7 day averaged), (late indicator)
2. Hospital Admission, (late indicator)
3. COVID-Like Symptoms, (early indicator)
4. COVID-Like Symptoms in Community, (early indicator)
5. COVID-Related Doctor Visits (early indicator)

The collection represents an assortment of early and late correlated indicators. The complete multivariate time series is thus represented by a five-component vector $\bar{y}(t)$. Each coordinate represents a distinct preprocessed signal derived from surveys. One of the vector coordinates corresponds to the Active Cases signal, which is of interest. Nevertheless, the five components are utilized for learning and prediction.

The resulting multivariate time series were divided into a training set and a validation set. The training set was used to update the network's weights θ , whilst the validation set was used to assess over fitting and training generalization. Using a mini-batched⁷¹ variant of multiple shooting^{72–74} direct training was conducted by calculating the loss between intervals of data points. The neural ODE was solved from the point \bar{y}_i at time t_i to time t_{i+1} with the Tsit5 method⁷⁵ using the DifferentialEquations.jl implementation⁷⁶ to obtain the prediction for the following point \bar{y}_{i+1} . Then, this was compared to the actual data point $\bar{y}_{data}(t_{i+1})$. The loss was computed as the mean squared error (MSE) between the observed point $\bar{y}_{data}(t)$ and the predicted point $\bar{y}(t)$ (see Eq. 2). The adjoint implementations of the DiffEqFlux.jl package⁷⁷ were used to achieve backpropagation.

The neural networks used for parameterizing the ODE consisted of four interconnected layers with 64, 32, 16 and 8 neurons each and swish activation functions⁷⁸. Once the neural net weights θ_i are calculated, the network sets the rate of the state variables' temporal evolution (See Eq. 1). Note that this equation can be solved beyond the training interval to evaluate its forecasting accuracy. Although it is anticipated that this prediction would degrade as it goes further in phase space from the training data, we argue that a highly predictive time frame will exist.

Software. The following open source software tools were used for this work: Pandas library⁷⁹ for part of the data pre processing⁶⁹, matplotlib⁸⁰ for plotting, and InkSpace for making figures⁸¹. The Julia library DiffEqFlux^{77,82} for training the neural ODEs, and the differential equation library DifferentialEquations.jl⁷⁶ for solving the differential equations.

Results and discussion

Figure 3 displays the output and projection of the trained neural ODE for the state of Ohio. One hundred days of data were used for training. The neural ODE follows the trend of newly reported cases fifty days into the future in its forecast (see interval after the vertical dotted line). Meanwhile, Fig. 4 demonstrates the situation of the

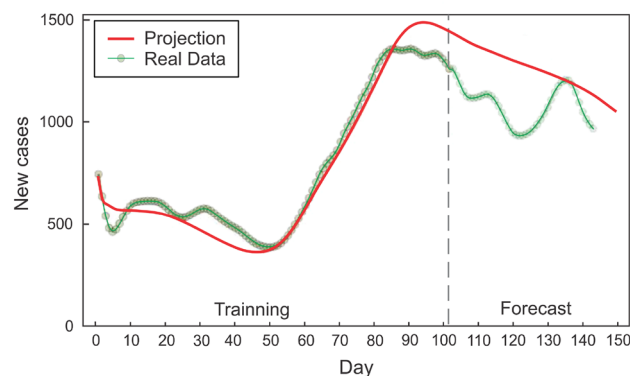


Figure 3. State of Ohio active case coordinate for the Neural ODE model. Newly reported cases are shown by dots, whereas the neural ODE solution is represented by a solid line. The vertical dashed line separates the training data set from the testing data set. New cases, Hospital Admission, COVID-Like Symptoms, COVID-Like Symptoms in the Community, and COVID-Related Doctor Visits are the variables included for this prediction⁵².

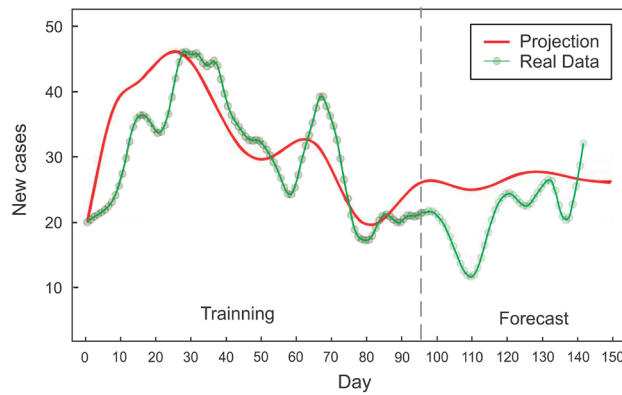


Figure 4. Prediction for the new infected cases in the state of ME (Maine), where the dynamics are not simply described by a model based on first principles, but the neural ODE is able to learn the dynamics and anticipate a rise in cases over the next fifty days. This rise is closely correlated with the recorded cases that were not utilized in the learning process (to the right of the vertical line).

state of Maine, in which the neural ODE learns to qualitatively extrapolate the new infected cases for 40 to 50 days using data from the preceding 95 days. The dynamics of the epidemic until day 95 (the last day of training), based solely on the number of new cases, indicate a decline in contagiousness; nevertheless, the neural ODE is able to forecast an increase in the number of new cases for the subsequent fifty days. Figure 5 indicates that in the current stage, the neural ode is trained with only 50 days of data, but is able to extrapolate the epidemic dynamics for the next sixty days, predicting an increase in the number of cases 15 days after the last day of training. In addition, is able to predict the day of the next peak and subsequent decline in cases.

Once the model (the neural ODE) has learned the dynamics of the local signals, it is capable of predicting new contagions and exploring potential future scenarios in the event of signal disruptions. For instance, the trained neural ODE model might estimate the influence of a sudden shift in the number of new cases at a specific time in the future. This is due to the fact that once the neural ODE learns the dynamics of the signals, it constructs parametrically a vector field parallel to the potential trajectories of the vector $\vec{y}(t)$. These trajectories define flow lines contained inside the five-dimensional space. If the state $\vec{y}(t)$ is on a flow line and is perturbed, it will move to a neighboring flow line and continue along its route (see Fig. 2). If the perturbation is large and the state moves far from the initial flow lines specified by the training points, it is anticipated that the neural ODE will not be able to represent the vector field volume where the new flow line is located; hence, the model predictions cannot be relied upon. In this instance, however, a change in the number of active cases indicates just a disruption in one of \vec{y} components. If the change is moderate, the trained model should be able to characterize the trajectory of the perturbed vector without requiring the neural network to be retrained. The disruption amounts to modifying the initial condition in the neural ODE integration and may be used to analyze prediction errors brought on by signal unpredictability related to the current number of cases. Figure 6 illustrates a forecast that accounts for such uncertainty in the present epidemic data.

People entering or leaving the modeled area will also cause a change in the trajectory of the state vector \vec{y}_0 . This is because some people will be carrying the Covid virus as they enter or leave the area. In other words, if N individuals visit the region, N_I individuals will be infected, where $N_I \leq N$. As a first approximation, we may argue that if the condition before the flow is characterized by \vec{y}_0 , then the state after the flow of people is

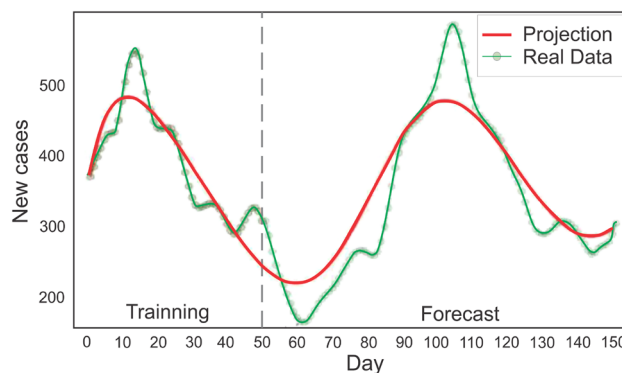


Figure 5. The training data set employed here (CO state) terminated before another outbreak of infected cases. Nevertheless, the neural ODE is able to determine the date of the outbreak's peak over 60 days after its onset.

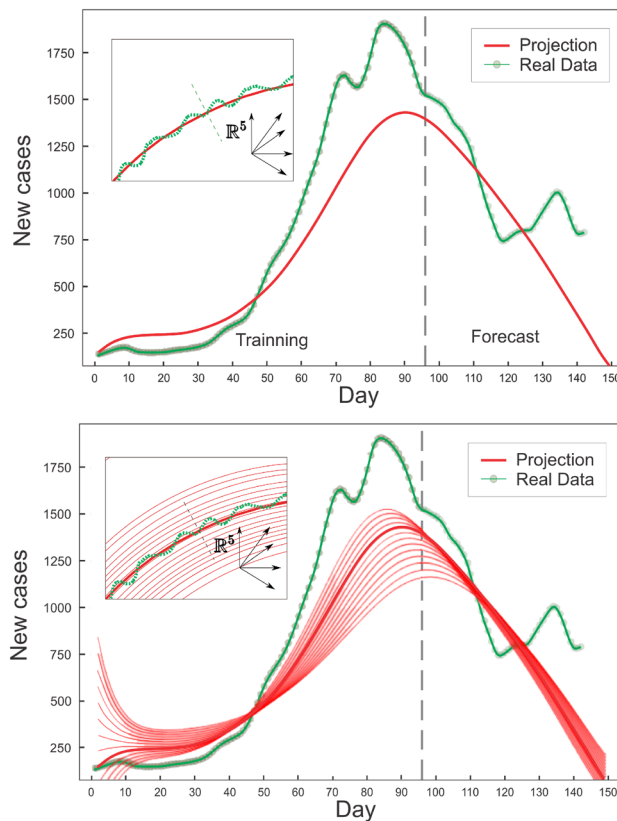


Figure 6. Once the neural ODE has learned the dynamics, its solution is capable of predicting the projection for a variety of initial conditions. Above, South Carolina's new active cases (green points) and the trained neural ODE solution (thick red line) are shown. The solution is one of the five coordinates of the full vector \vec{y} , the trajectory of which is illustrated in the inset. By integrating the same neural ODE with different initial conditions for the number of cases, the learned phase space can be explored (illustrated below) and diverse forecasts obtained. The green represents the survey signals, whereas the thick red line represents the learned solution, as shown in the inset. The thin red lines represent phase space lines that are accessible to the system as a result of a perturbation, such as a change in the initial number of Active cases. Examining signal perturbations may help measure prediction uncertainty and estimate the effect of local active case modifications.

$\vec{y}_0 + \Delta\vec{y}$, where $\Delta\vec{y} = (N_I, 0, 0, 0, 0)$ in the 5-dimensional space employed in our study. If $\Delta\vec{y}$ is not very large, we should be able to characterize the trajectory by integrating with the new beginning condition $\vec{y}_0 + \Delta\vec{y}$ (see Fig. 2). Consequently, if it is feasible to estimate the number of infected individuals among those in transit, the model might approximate the influence on the infected curve. Figure 7 showcases the varying heights of the expected peak depending on the quantity of migration chosen. By examining the perturbed solution, the change and its repercussions may be evaluated. This makes it possible for regions with strict closed borders to forecast the influence of the movement of people on the infection curve. Additionally, the impacts of immunization in the region might be measured in this manner.

The work by Mayorga et al.⁸³ provides a representative example of the calculation/estimation of the infected in the flow of persons leaving or entering cities. Models of the SEIR type depict the various geographical locations and the epidemiological dynamics of their respective populations. A flow matrix that connects the local models represents the transit of individuals. On the basis of the day flow, the number of infected I and exposed E individuals is estimated and variables are updated daily. The same flow-related reasoning can be applied to our situation by substituting the SEIR models of different locations with neural ODE models trained with data extracted from local surveys.

Conclusions and future work

Using multivariate time series connected with a geographic region, gathered by quantifying indicators from large online surveys on COVID symptoms presented via Facebook, we describe how a neural ODE can learn the dynamics that connect these variables and detect viral outbreaks in the region. We demonstrate, by analyzing data from several U.S. states, that the neural ODE is capable of forecasting up to sixty days into the future in a variety of virus-spreading scenarios.

We assert that once the neural ODE has learned the dynamics of the local signals/variables, it is capable of not only forecasting new infections in the region, but also analyzing possible future scenarios in the case of abrupt changes in the number of infected in a given day, for instance due to transit of people into or out of the analyzed

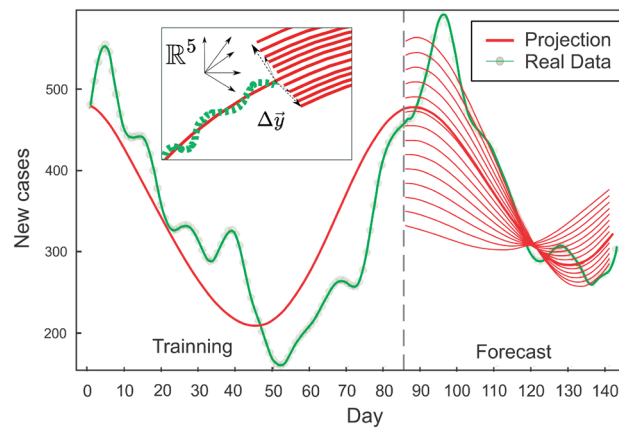


Figure 7. After being trained with Colorado data, the neural ODE can extrapolate new scenarios when the number of local new cases changes. This shift may occur as a result of the influx or outflow of travelers or a local immunization program. The neural ODE solution is represented by a thick line, whilst the reported 7-day average of new cases is represented by green dots. The test data begins after the vertical dashed line. Different shifts in the number of infected individuals are delineated, with the neural ode forecast for each instance displayed (thin line). The inset shows the full 5-dimensional space produced by the five survey variables and perturbation that defines possible flux lines in the system's coordinate space. The thick red line represents the neural ODE solution, and the green line represents survey signals.

region. This affords regions with strict closed borders the opportunity to predict the impact of the flow of people on the infection curve and, as a result, formulate policies in a controlled manner to optimize the transit of people and reduce economic stagnation during a pandemic.

In addition to the considerations mentioned earlier, future works can explore the potential of incorporating compression analysis into our research. While our current study focused on training the neural ODE using data from a single location or state, it is likely that the dynamics connecting local signals in one region to those in another region exhibit similar properties. Therefore, it would be valuable to investigate alternative training schemes where the model learns from multiple locations.

Furthermore, incorporating graphical models into the neural ODE framework, possibly through the utilization of graph neural networks, represents a promising avenue for future study. This approach would allow for a more comprehensive understanding of the interconnections and dependencies between different regions or entities within the system under consideration. By incorporating such graph-based techniques, we can potentially enhance the model's predictive capabilities and capture more nuanced dynamics.

Our neural ODE model, trained on real-time social media data, extends the principles of first principles models like the SEIR model. While the SEIR model relies on assumptions about disease transmission and population dynamics, our neural ODE model directly learns from interconnected local signals extracted from social media surveys. This data-driven approach allows us to capture the underlying dynamics of disease transmission and population behavior with greater flexibility and adaptability. Moreover, our data-driven model can be combined with first principles models, such as the SEIR model, using a scientific machine learning approach⁶⁸. By integrating the strengths of both data-driven and analytical modeling approaches, we can achieve a more comprehensive understanding of epidemic dynamics and improve the accuracy of the predictions.

This work is a preliminary phase, a *proof of concept*. It is essential to investigate various signals and combinations and evaluate their generalization capabilities. Accurate application of the uncertainty quantification requires a great deal more research before it can be utilized in public health situations. As Nobel laureate Niels Bohr remarked a century ago: “*prediction is very difficult, especially if it's about the future...*”. On the other hand, these findings provide some promising outcomes for future real-time forecasts that are based on predictive data from social media.

Data availability

All data generated or analysed during this study are included in this published article [and its supplementary information files.

Code availability

The code supporting the current manuscript can be found in the supplementary material.

Received: 3 October 2022; Accepted: 15 June 2023

Published online: 05 July 2023

References

1. Steele, L., Orefuwa, E. & Dickmann, P. Drivers of earlier infectious disease outbreak detection: A systematic literature review. *Int. J. Infect. Dis.* 53, 15. <https://doi.org/10.1016/j.ijid.2016.10.005> (2016).

2. Ning, S., Yang, S. & Kou, S. Accurate regional influenza epidemics tracking using internet search data. *Sci. Rep.* **9**, 5238 (2019).
3. McNeil Jr., D. G. *Covid-19: How much Herd Immunity is Enough*. <https://www.nytimes.com/2020/12/24/health/herd-immunity-covid-coronavirus.html> (2020)
4. Longbottom, J., Wamboga, C., Bessell, P., Torr, S. & Stanton, M. Optimising passive surveillance of a neglected tropical disease in the era of elimination: A modelling study. *PLoS Negl. Trop. Dis.* **15**, e0008599. <https://doi.org/10.1371/journal.pntd.0008599> (2021).
5. Nsubuga, P. *et al.* Disease control priorities in developing countries (2nd ed.). In *Public Health Surveillance: A Tool for Targeting and Monitoring Interventions* (eds Jamison, D. *et al.*) 997–1015 (The International Bank for Reconstruction and Development/The World Bank. Co-published by Oxford University Press, 2006).
6. Groseclose, S. L. & Buckeridge, D. L. Public health surveillance systems: Recent advances in their use and evaluation. *Annu. Rev. Public Health* **38**, 57. <https://doi.org/10.1146/annurev-publhealth-031816-044348> (2017) ((pMID: 27992726)).
7. Lombardo, J. S., Burkom, H. & Pavlin, J. ESSENCE II and the framework for evaluating syndromic surveillance systems. *Morb. Mortal. Week. Rep.* **53**, 159 (2004).
8. Project, Triple S. Assessment of syndromic surveillance in Europe. *The Lancet* **378**, 1833. [https://doi.org/10.1016/S0140-6736\(11\)60834-9](https://doi.org/10.1016/S0140-6736(11)60834-9) (2011).
9. Chen, J. *et al.* Practice and thinking of acute respiratory infection surveillance for the response of emerging respiratory diseases in Shanghai. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi* **41**, 1994–1998. <https://doi.org/10.3760/cma.j.cn112338-20200421-00616> (2020).
10. Garg, S., Bhatnagar, N. & Gangadharan, N. A case for participatory disease surveillance of the COVID-19 pandemic in India. *JMIR Public Health Surveill.* **6**, e18795. <https://doi.org/10.2196/18795> (2020).
11. Wahid, M. A., Bukhari, S. H. R., Daud, A., Awan, S. E. & Raja, M. A. Z. COVICT: An IoT based architecture for COVID-19 detection and contact tracing. *J. Ambient Intell. Human. Comput.* **14**(6), 7381–7398 (2022).
12. Henning, K. J. What is syndromic surveillance?. *Morb. Mortal. Week. Rep.* **53**, 7 (2004).
13. Shoaib, M., Haider, A., Raja, M. A. Z. & Nisar, K. S. Artificial intelligence knacks-based computing for stochastic COVID-19 SIRC epidemic model with time delay. *Int. J. Mod. Phys. B* **36**, 2250174 (2022).
14. Şerban, O., Thapen, N., Maginnis, B., Hankin, C. & Foot, V. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Inform. Process. Manag.* **56**, 1166. <https://doi.org/10.1016/j.ipm.2018.04.011> (2019).
15. Budd, J. *et al.* Digital technologies in the public-health response to COVID-19. *Nat. Med.* **26**, 1183. <https://doi.org/10.1038/s41591-020-1011-4> (2020).
16. Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. & Weinstein, R. A. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **47**, 1443. <https://doi.org/10.1086/593098> (2008).
17. Samaras, L., García-Barriocanal, E. & Sicilia, M.-A. Chapter 2 - Syndromic surveillance using web data: A systematic review. In *Innovation in Health Informatics, Series and Number Next Gen Tech Driven Personalized Med & Smart Healthcare* (eds Lytras, M. D. & Sarirete, A.) 39–77 (Academic Press, 2020).
18. Cook, S., Conrad, C., Fowlkes, A. L. & Mohebbi, M. H. Assessing google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One* **6**, 1. <https://doi.org/10.1371/journal.pone.0023610> (2011).
19. Broniatowski, D. A., Paul, M. J. & Dredze, M. National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS One* **8**, null. <https://doi.org/10.1371/journal.pone.0083672> (2013).
20. Fung, Z. T., Tse, I. C. & Fu, K. W. The use of social media in public health surveillance. *Western Pacific Surveill. Response J. WPSAR* **6**, 3. <https://doi.org/10.5365/WPSAR.2015.6.1.019> (2015).
21. Velardi, P., Stilo, G., Tozzi, A. E. & Gesualdo, F. Twitter mining for fine-grained syndromic surveillance. *Artif. Intell. Med.* **61**, 153 (2014).
22. Yousefinaghani, S., Dara, R., Poljak, Z., Bernardo, T. M. & Sharif, S. The assessment of twitter's potential for outbreak detection: Avian influenza case study. *Sci. Rep.* **9**, 18147. <https://doi.org/10.1038/s41598-019-54388-4> (2019).
23. Rossman, H. *et al.* A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nat. Med.* **26**, 634. <https://doi.org/10.1038/s41591-020-0857-9> (2020).
24. Taylor, M. & Galanis, E. Online population control surveys: A new method for investigating foodborne outbreaks. *Epidemiol. Infect.* **148**, e93. <https://doi.org/10.1017/S0950268820000837> (2020).
25. Wang, S., Ding, S. & Xiong, L. A new system for surveillance and digital contact tracing for COVID-19: Spatiotemporal reporting over network and GPS. *JMIR mHealth uHealth* **8**, e19457. <https://doi.org/10.2196/19457> (2020).
26. Wang, S., Ding, S. & Xiong, L. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *JMIR mHealth uHealth* **8**, e19457. <https://doi.org/10.2196/19457> (2020).
27. Dandekar, R. A., Henderson, S. G., Jansen, M., Moka, S., Nazarathy, Y., Rackauckas, C., Taylor, P. G. & Vuorinen, A. Safe blues: A method for estimation and control in the fight against COVID-19. *medRxiv* (2020a)
28. Abad, Z. S. H. *et al.* Digital public health surveillance: A systematic scoping review. *npj Digit. Med.* **4**, 41. <https://doi.org/10.1038/s41746-021-00407-6> (2021).
29. Brownstein, J. S., Freifeld, C. C. & Madoff, L. C. Digital disease detection - harnessing the web for public health surveillance. *N. Engl. J. Med.* **360**, 2153. <https://doi.org/10.1056/NEJMp0900702> (2009) (pMID: 19423867).
30. Choi, J., Cho, Y., Shim, E. & Woo, H. Web-based infectious disease surveillance systems and public health perspectives: A systematic review. *BMC Public Health* **16**, 1238 (2016).
31. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: Traps in big data analysis. *Science* **343**(6176), 1203–1205. <https://doi.org/10.1126/science.1248506> (2014).
32. Santillana, M. *et al.* Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput. Biol.* **11**, 1. <https://doi.org/10.1371/journal.pcbi.1004513> (2015).
33. Dolley, S. Big data's role in precision public health. *Front. Public Health* **6**, 68. <https://doi.org/10.3389/fpubh.2018.00068> (2018).
34. Reich Lab of the University of Massachusetts Amherst, *The COVID-19 forecast hub*. (2020), <https://covid19forecasthub.org/>
35. Facebook Data for Good, the Delphi Group at Carnegie Mellon University (CMU), the Joint Program on Survey Methodology at the University of Maryland (UMD), the Duke Margolis Center for Health Policy and Resolve to Save Lives, The COVID-19 symptom data challenge, (2020), <https://www.symptomchallenge.org/>
36. Bukhari, A. H., Ahmed, E., Raja, M. A. Z., Chen, Y. & Shoaib, M. A multimodal hybrid stochastic-based deterministic ARFIMA model for the sustainable analysis of COVID-19 pandemic. *Waves in Random and Complex Med.*, 1 (2023).
37. Google, *COVID-19 Community Mobility Reports*. <https://www.google.com/covid19/mobility/> (2020).
38. Facebook. *Data For Good - Covid-19 Surveys*. <https://dataforgood.fb.com/> (2020).
39. Farrow, D. C., Brooks, L. C., Rumack, A., Tibshirani, R. J. & Rosenfeld, R. (Delphi Epidata API, 2015) <https://github.com/cmu-delphi/delphi-epidata>
40. Reinhart, A. & Tibshirani, R. in *COVID-19 Symptom Surveys through Facebook* <https://delphi.cmu.edu/blog/2020/08/26/covid-19-symptom-surveys-through-facebook/> (2020).
41. Bethlehem, J. Selection bias in web surveys. *Int. Statist. Rev.* **78**, 161 (2010).

42. Yeung, A. Y., Roewer-Despres, F., Rosella, L. & Rudzicz, F. Machine learning-based prediction of growth in confirmed COVID-19 infection cases in 114 countries using metrics of nonpharmaceutical interventions and cultural dimensions: Model development and validation. *J. Med. Internet Res.* **23**, e26628 (2021).
43. ArunKumar, K., Kalaga, D. V., Kumar, C. M. S., Kawaji, M. & Brenza, T. M. Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells. *Chaos Solitons Fractals* **146**, 110861 (2021).
44. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med.* **4**, 3. <https://doi.org/10.1038/s41746-020-00372-6> (2021).
45. Dandekar, R., Rackauckas, C. & Barbastathis, G. A machine learning-aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread. *Patterns* **1**, 100145 (2020).
46. Zeb, A., Alzahrani, E., Erturk, V. S. & Zaman, G. Mathematical model for coronavirus disease 2019 (COVID-19) containing isolation class. *Biomed. Res. Int.* **2020**, 3452402. <https://doi.org/10.1155/2020/3452402> (2020).
47. Alqudah, M. A., Abdeljawad, T., Zeb, A., Khan, I. U. & Bozkurt, F. Effect of weather on the spread of COVID-19 using eigenspace decomposition. *CMC-Comput. Mater. Continua*, 3047 (2021)
48. Zhang, Z., Gul, R. & Zeb, A. Global sensitivity analysis of COVID-19 mathematical model. *Alex. Eng. J.* **60**, 565 (2021).
49. Tesfay, A. *et al.* Dynamics of a stochastic COVID-19 epidemic model with jump-diffusion. *Adv. Differ. Equ.* **2021**, 1 (2021).
50. Alvarez, E. *et al.* Estimating COVID-19 cases and outbreaks on-stream through phone calls. *R. Soc. Open Sci.* **8**(3), 202312 (2021).
51. Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* (eds Bengio, S. *et al.*) 6571–6583 (Curran Associates Inc, 2018).
52. Delphi Group, in *COVID Symptom Survey*. (2020a) <https://cmu-delphi.github.io/delphi-epidata/symptom-survey/>
53. note This definition is in line with the working definition of CLI used by the US Centers for Disease Control and Prevention (CDC) and mirrors the standard definition of influenza-like illness or ILI (defined as fever of at least 100 °F, along with sore throat or cough).
54. Delphi Group, <https://delphi.cmu.edu/covidcast> Covidcast interactive map. (2020b)
55. Kermack, W. O. & McKendrick, A. G. Contributions to the mathematical theory of epidemics-I. 1927. *Bull. Math. Biol.* **53**(1–2), 33–55. <https://doi.org/10.1007/BF02464423> (1991).
56. Winkler, D. A. & Le, T. C. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. *Mol. Inf.* **36**, 1600118 (2017).
57. Lin, H. & Jegelka, S. Resnet with one-neuron hidden layers is a universal approximator. *Adv. Neural Inform. Process. Syst.* **31**, 6169–6178 (2018).
58. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016) <https://doi.org/10.1109/CVPR.2016.902016>
59. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735 (1997).
60. Weinan, E. A proposal on machine learning via dynamical systems. *Commun. Math. Statist.* **1**, 1 (2017).
61. Habiba, M., & Pearlmutter, B. A. (2020). Neural ordinary differential equation based recurrent neural network model. In *2020 31st Irish Signals and Systems Conference (ISSC)* 1–6. (IEEE, 2020)
62. Hairer, E., Nørsett, S. P. & Wanner, G. *Solving Ordinary Differential Equations. 1, Nonstiff Problems* (Springer-Vlg, 1993).
63. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization*. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
64. Pontryagin, L. S. *Mathematical Theory of Optimal Processes* (CRC Press, 1987).
65. Gholami, A., Keutzer, K. & Biros, G. *Anode: Unconditionally Accurate Memory-efficient Gradients For Neural Odes*. arXiv preprint [arXiv:1902.10298](https://arxiv.org/abs/1902.10298) (2019)
66. Onken, D. & Ruthotto, L. *Discretize-optimize Versus Optimize-discretize For Time-series Regression and Continuous Normalizing Flows*. arXiv preprint [arXiv:2005.13420](https://arxiv.org/abs/2005.13420) (2020).
67. Serban, R. & Hindmarsh, A. C. *Cvodes: An Ode Solver with Sensitivity Analysis Capabilities*, type Tech. Rep. (2003)
68. Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A. & Edelman, A. in *Universal Differential Equations for Scientific Machine Learning*, arXiv preprint [arXiv:2001.04385](https://arxiv.org/abs/2001.04385) (2020a).
69. note Dr. Florencia Grinbladt downloaded the data and applied the smoothing.
70. Alex Reinhart, R. T. (Delphi Epidata API, 2020). <https://cmu-delphi.github.io/delphi-epidata/api/covidcast-signals/indicator-combination.html#compositional-signals-confirmed-cases-and-deaths>
71. Rackauckas, C. *et al.* *Training a Neural Ordinary Differential Equation with Mini-batching*. <https://diffeqflux.sciml.ai/stable/examples/minibatch/> (2020)
72. Morrison, D. D., Riley, J. D. & Zancanaro, J. F. Multiple shooting method for two-point boundary value problems. *Commun. ACM* **5**, 613 (1962).
73. Bock, H. G. & Plitt, K.-J. A multiple shooting algorithm for direct solution of optimal control problems. *IFAC Proc. Vol.* **17**, 1603 (1984).
74. Bock, H. G., Diehl, M. M., Leineweber, D. & Schlöder, J. P. A direct multiple shooting method for real-time optimization of nonlinear DAE processes. in: *Nonlinear Model Predictive Control* 245–267 (Springer, 2000).
75. Tsitouras, C. Runge–Kutta pairs of order 5 (4) satisfying only the first column simplifying assumption. *Comput. Math. Appl.* **62**, 770–775 (2011).
76. Rackauckas, C. & Nie, Q. Differentialequations JL-A performant and feature-rich ecosystem for solving differential equations in Julia. *J. Open Res. Softw.* **5**, 15 (2017).
77. Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D. & Ramadhan, A. *Universal Differential Equations for Scientific Machine Learning* arXiv preprint [arXiv:2001.04385](https://arxiv.org/abs/2001.04385) (2020b)
78. Ramachandran, P., Zoph, B. & Le, Q. V. *Searching for Activation Functions*. arXiv:1710.05941 [cs] (2017)
79. The pandas development team, pandas-dev/pandas: Pandas (2020) <https://doi.org/10.5281/zenodo.3509134>
80. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90. <https://doi.org/10.1109/MCSE.2007.55> (2007).
81. Harrington, B. *et al.* *Inkscape* (2004) <https://inkscape.org>
82. Rackauckas, C. *et al.* in *DiffEqFlux: Generalized Physics-Informed and Scientific Machine Learning* (sciml) (2020) <https://diffeqflux.sciml.ai/stable/>
83. Mayorga, L. *et al.* A modelling study highlights the power of detecting and isolating asymptomatic or very mildly affected individuals for COVID-19 epidemic management. *BMC Public Health* **20**, 1 (2020).

Acknowledgements

We acknowledge support from The National Autonomous University of Mexico (UNAM) and Alianza UCMX of the University of California (UC), through the project included in the Special Call for Binational Collaborative Projects addressing COVID-19. MN is partially supported by CONICET, Argentina. RAB was financially supported by Conacyt through project 283279. We thank Dr. YangQuan Chen for enlightening discussions, to Dr. Daniel Garcia for carefully reading the manuscript and Dr. Florencia Grinbladt for downloading the data and Dr. Cecilia Ventura.

Author contributions

Conceptualization—M.N. Methodology—M.N. Software—M.N., C.R. Writing—Original Draft M.N. Writing—Review & Editing M.N., N.L.B., R.A.B., C.R. Visualization—M.N., N.L.B. Supervision—M.N. Funding Acquisition—R.A.B., M.N., C.R.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-37118-9>.

Correspondence and requests for materials should be addressed to M.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023