Accurately Estimating Unreported Infections using Information Theory

Jiaming Cui[*†‡] Bijaya Adhikari[§] Arash Haddadan[†¶] A S M Ahsan-Ul Haque[†]
Jilles Vreeken[‖] Anil Vullikanti[†] B. Aditya Prakash[*]

Abstract

One of the most significant challenges in combating against the spread of infectious diseases was the difficulty in estimating the true magnitude of infections. Unreported infections could drive up disease spread, making it very hard to accurately estimate the infectivity of the pathogen, therewith hampering our ability to react effectively. Despite the use of surveillance-based methods such as serological studies, identifying the true magnitude is still challenging. This paper proposes an information theoretic approach for accurately estimating the number of total infections. Our approach is built on top of Ordinary Differential Equations (ODE) based models, which are commonly used in epidemiology and for estimating such infections. We show how we can help such models to better compute the number of total infections and identify the parametrization by which we need the fewest bits to describe the observed dynamics of reported infections. Our experiments on COVID-19 spread show that our approach leads to not only substantially better estimates of the number of total infections but also better forecasts of infections than standard model calibration based methods. We additionally show how our learned parametrization helps in modeling more accurate what-if scenarios with non-pharmaceutical interventions. Our approach provides a general method for improving epidemic modeling which is applicable broadly.

Keywords: Information theory, Ordinary differential-based Equations, Modeling, Forecasting

1 Introduction

One of the most significant challenges in combating against the spread of infectious diseases in population is estimating the number of total infections. Our inability in estimating unreported infections allows them to drive up disease transmission. For example, in the COVID-19 pandemic, a significant number of COVID-19 infections were unreported, due to various factors such as the lack of testing and asymptomatic infections [8, 6, 39, 37, 25]. There were only 23 reported infections in five major U.S. cities by March 1, 2020, but it has been estimated that there were in fact more than 28,000 total infections by then [4], and spread the COVID-19 to the whole US. Similar trends were observed in other countries, such as in Italy, Germany, and the UK [41].

In fact, an accurate estimation of the number of total infections is a fundamental epidemiological question and critical for pandemic planning and response. Therefore, epidemiologists use *reported rate* (α_{reported}) to capture total infections, which is defined as the ratio of reported infections to total infections [29]. One of the benefits of using this definition is that it includes asymptomatic infections, which may also contribute substantially to spread [40, 26]. To estimate the reported rate, data scientists and epidemiologists have devoted much time and effort to using epidemiological models. There are many carefully constructed Ordinary Differential Equation (ODE) based models that capture the transmission dynamics of different infectious diseases [25, 35, 7, 30, 22, 23, 42, 16, 20, 43, 44, 11, 27]. However, these models still suffer from estimating accurate reported rates, leading to suboptimal total infections estimation. For example, as shown in Figure 1, the Minneapolis Metro Area had only 16 COVID-19 reported infections by March 11, 2020. Although epidemiologists estimate that there were 182 total infections (light green part in the iceberg) using epi models, later studies revealed that there were actually around 300 total infections (iceberg below the sea level) [17, 1], which is much larger than the epi model estimated values.

To tackle this, we propose a new information theory-based approach named MdlInfer to estimate the reported rate. It is based on the following central intuition: Suppose an "oracle" gives us the time series of the number of *total infections* D, we should be able to describe D_{reported} in a succinct way: As we know

[*] College of Computing, Georgia Institute of Technology, Email: {jiamingcui1997, badityap}@gatech.edu.

[†] Department of Computer Science, University of Virginia, Email: {zej9va,ah8zf,ah3wj,vsakumar}@virginia.edu.

[‡] Department of Computer Science, Virginia Tech, Email: jiamingcui@vt.edu.

[§] Department of Computer Science, The University of Iowa, Email: bijaya-adhikari@uiowa.edu.

[¶] Modeling and Optimization, Amazon, Email: ahaddada@amazon.com

[‖] CISPA Helmholtz Center for Information Security, Email: vreeken@cispa.de

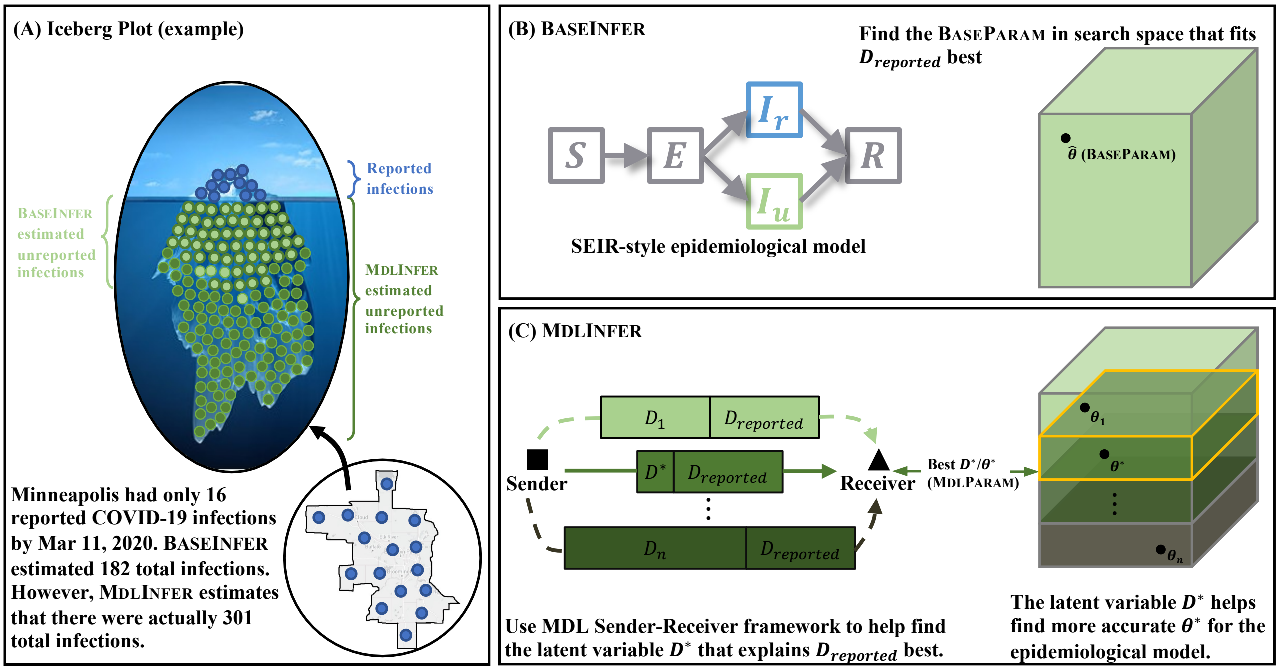
(A) Iceberg Plot (example)
Reported infections
BASEINFER estimated unreported infections
MDLINFER estimated unreported infections
Minneapolis had only 16 reported COVID-19 infections by Mar 11, 2020. BASEINFER estimated 182 total infections. However, MDLINFER estimates that there were actually 301 total infections.
(B) BASEINFER
Find the BASEPARAM in search space that fits $D_{reported}$ best
S
E
I_r
I_u
R
SEIR-style epidemiological model
$\hat{\theta}$ (BASEPARAM)
(C) MDLINFER
Sender
Receiver
D_1
$D_{reported}$
D^*
$D_{reported}$
D_n
$D_{reported}$
Best D^*/θ^* (MDLPARAM)
θ_1
θ^*
θ_n
Use MDL Sender-Receiver framework to help find the latent variable D^* that explains $D_{reported}$ best.
The latent variable D^* helps find more accurate θ^* for the epidemiological model.

Figure 1: **Overview of our problem and methodology.** (**A**) We visualize the idea of reported rates using the iceberg. The visible portion above water are the reported infections, which is only a fraction of the whole iceberg representing total infections. Light green corresponds to the 182 unreported infections estimated by typical current practice used by researchers. We call it as the basic approach, or BASEINFER. In contrast, dark green corresponds to the more accurate and much larger 301 unreported infections found by our approach MDLINFER. (**B**) The usual practice is to calibrate an epidemiological model to reported data and compute the reported rate from the resultant parameterization of the model. Here, an SEIR-style model with explicit compartments for reported-vs-unreported infection is shown in the figure as an example. (**C**) Our new approach MDLINFER instead aims to compute a more accurate reported rate by finding a 'best' parametrization *for the same epidemiological model* (i.e., SEIR-style model in this example) using a principled information theoretic formulation - two-part 'sender-receiver' framework. Assume that a hypothetical Sender S wants to transmit the reported infections as the DATA to a Receiver R in the cheapest way possible. Hence S will find/solve for the best D^*, intuitively, the MODEL that takes the fewest number of bits to encode the DATA. Using D^*, we can find the best Θ^* by exploring a smaller search space.

D, it is trivial to get the reported rate α_{reported}. Then with both D and α_{reported}, it will be trivial to describe D_{reported}, as it is simply $D \times \alpha_{\text{reported}}$ plus a little bit of noise.

In practice, we are of course not given D, but we could estimate D as a latent variable. Specifically, as shown in Figure 1(C), we use Minimum Description Length (MDL) principle to estimate D, which allows us to most succinctly describe (i.e., most accurately encode/reconstruct) the dynamics of D_{reported}. Here, MDLINFER gives an estimate of 301 total infections in Minneapolis as shown below the sea level, which is much closer to the ground truth total infections numbers of around 300 [17, 1].

Our main contributions are summarized below.

- We propose an MDL-based approach on top of ODE-based epidemiological models, which are harder to formulate and optimize. To the best of our knowledge, we are the first to propose an MDL-based approach on top of ODE-based epidemiological models.
- Our proposed MDL-based approach MDLINFER performs superior to the state of the art epidemiological model in estimating total infections and predicting the future reported infections.
- We also show that MDLINFER can aid policy making by analyzing counter-factual non-pharmaceutical interventions, while inaccurate epidemiological model estimates may lead to wrong non-pharmaceutical intervention conclusions.

The rest of the paper is organized in the following way: Section 2 discusses the related works. In section 3, we introduce the current ODE model calibration method to estimate the reported rate, and the background of MDL framework. We then introduce our MDLINFER framework in section 4 and explain how we use it to estimate the total infections. In section 5, we evaluate the performance of MdlInfer. We then discuss future work and conclude in section 6.

2 Related work

2.1 Reported rate estimation One of the most effective current methods to identify the reported rate in a region is through large-scale serological studies [38, 17, 45]. These surveys use blood tests to identify the prevalence of antibodies against target pandemic in a large population. While serological studies can give an accurate estimation, they are expensive and are not sustainable in the long run [3]. Furthermore, it is also challenging to obtain real-time data using such studies since there are unavoidable delays between sample collection and laboratory tests [1, 17].

2.2 Minimum Description Length framework MDL frameworks has been widely used for numerous optimization problems ranging from network summarization [21], causality inference [10], and failure detection in critical infrastructures [5]. They are also used in machine learning as regularization to help with model selection and avoid overfitting. [9] However, these works are built on networks and agent-based models. To the best of our knowledge, we are the first to propose an MDL-based approach on top of ODE-based epidemiological models.

3 Preliminaries

3.1 ODE-based Models An ODE-based epidemiological model uses ordinary differential equations to describe the spread of diseases by modeling changes in populations (e.g., susceptible, infected, recovered) over time [19]. In general, the O_{M} has a set of parameters Θ that need to estimate from *observed data* using a so-called calibration procedure, Calibrate. In practice, the data we use for calibration can be the time series of the number of reported infections, or D_{reported}. To estimate the number of total infections, these models often explicitly include reported rate as one of their parameters, or include multiple parameters that jointly account for it. We call it as BaseInfer in later sections for brief. There are many calibration procedures commonly used in literature, such as RMSE-based [15] or Bayesian approaches [20, 16]. BaseInfer is generally a complex, high-demensional problem, since there are multiple parameters interacting with each other. To make matters worse, there exist many possible parametrizations that show similar performance (e.g. in RMSE, likelihood) yet correspond to vastly different reported rates, and BaseInfer cannot select between these competing parametrizations in a principled way.

Table 1: List of notations

Notation	Description
O_{M}	ODE model
Θ	ODE model parameters to infer
D_{reported}	Reported infections
α_{reported}	Reported rate
BaseInfer	Baseline ODE calibration procedure (calibrated on only D_{reported})
$\hat{\Theta}$	Parametrization estimated by BaseInfer
$\hat{\alpha}_{\text{reported}}$	Reported rate in $\hat{\Theta}$
$D_{\text{reported}}(\hat{\Theta})$	ODE simulated reported infections using $\hat{\Theta}$
$D(\hat{\Theta})$	ODE simulated total infections using $\hat{\Theta}$
MdlInfer	Our framework
D	Candidate total infections in MdlInfer
$\Theta^{'}$	Parametrization estimated by MdlInfer when calibrating on both D and D_{reported}
$\alpha'_{\text{reported}}$	Reported rate in Θ'
$D_{\text{reported}}(\Theta')$	ODE simulated reported infections using Θ'
$D(\Theta')$	ODE simulated total infections using Θ'

3.2 Two-part sender-receiver MDL framework In this work, we use this framework to identify the total infections. The conceptual goal of the framework is to transmit the Data from the possession of the hypothetical sender S to the hypothetical receiver R. We assume the sender does this by first sending a Model and then sending the Data under this Model. In this MDL framework, we want to minimize the number of bits for this process. We do this by identifying the Model that encodes the Data such that the total number of bits needed to encode both the Model and the Data is minimized. Hence, our cost function in the total number of bits needed is composed of two parts: (i) model cost $L(\text{Model})$: The cost in bits of encoding the Model and (ii) data cost $L(\text{Data}|\text{Model})$: The cost in bits of encoding the Data given the Model. Intuitively, the idea is that a good Model will lead to a fewer number of bits needed to encode both Model and Data. The general MDL optimization problem can be formulated as follows: Given the Data, $L(\text{Model})$, and $L(\text{Data}|\text{Model})$, find Model^* such that

$$\text{Model}^* = \arg\min_{\text{Model}} L(\text{Model}) + L(\text{Data}|\text{Model})$$

4 MdlInfer

4.1 MDL formulation In our situation, the Data is the reported infections D_{reported}, which is the only real-world data given to us. As for the Model, intuitively it should be $(D, \alpha'_{\text{reported}})$ since our goal is to find the total infections D with the corresponding reported rate $\alpha'_{\text{reported}}$. Note that as two-part MDL (and MDL in general) does not assume the nature of the Data or the Model, our MdlInfer can be applied to any ODE model. Next, we give more details how to formulate our problem of estimating total infections D. We also list the notations in Table 1

4.1.1 Model space As described above, our Model is intuitively $(D, \alpha'_{\text{reported}})$. Note that reported rate is actually one of the parameters for the ODE model O_{M}, we choose to include its corresponding parametrization $\Theta^{'}$ into Model. We further choose to add $\hat{\Theta}$ estimated by BaseInfer, making our Model to be $(D, \Theta^{'}, \hat{\Theta})$. With Model $= (D, \Theta^{'}, \hat{\Theta})$, our Model space will be all possible daily sequences for D and all possible parametrizations for $\Theta^{'}$ and $\hat{\Theta}$. The MDL framework will search in this space to find the Model*. We also discuss other alternative Models and why $(D, \Theta^{'}, \hat{\Theta})$ is better in Appendix[1].

4.1.2 Model cost With Model $= (D, \Theta^{'}, \hat{\Theta})$, we conceptualize the model cost by imagining that the sender S will send the Model $= (D, \Theta^{'}, \hat{\Theta})$ to the receiver R in three parts: (i) first send the $\hat{\Theta}$ by encoding $\hat{\Theta}$ directly (ii) next send the $\Theta^{'}$ given $\hat{\Theta}$ by encoding $\Theta^{'} - \hat{\Theta}$ and (iii) then send D given $\Theta^{'}$ and $\hat{\Theta}$ by encoding $\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})$. Intuitively, both $\alpha'_{\text{reported}} \times D$ and $D_{\text{reported}}(\hat{\Theta})$ should be close to D_{reported}, and the receiver could recover the D using $\hat{\Theta}$, $\alpha'_{\text{reported}}$, and $D_{\text{reported}}(\hat{\Theta})$ as they have already been sent. We term the model cost as $L(D, \Theta^{'}, \hat{\Theta})$ with three components: $\textsc{Cost}(\hat{\Theta})$, $\textsc{Cost}(\Theta^{'}|\hat{\Theta})$, and $\textsc{Cost}(D|\Theta^{'}, \hat{\Theta})$. Hence

$$\begin{aligned} L(D, \Theta^{'}, \hat{\Theta}) &= \textsc{Cost}(\hat{\Theta}) + \textsc{Cost}(\Theta^{'} - \hat{\Theta}|\hat{\Theta}) \\ &\quad + \textsc{Cost}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})|\Theta^{'}, \hat{\Theta}) \end{aligned}$$

Here, the $\textsc{Cost}(\cdot)$ function gives the total number of bits we need to spend in encoding each term. The details of the encoding method can be found in the Appendix.

4.1.3 Data cost We need to send the Data $= D_{\text{reported}}$ next given the Model. Given Model $= (D, \Theta^{'}, \hat{\Theta})$, we send Data by encoding $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta^{'})$. Intuitively, $D - D_{\text{reported}}$ corresponds to the unreported infections, and $1 - \alpha'_{\text{reported}}$ is the unreported rate. Therefore, $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}}$ should be close to the total infections D and $D(\Theta^{'})$. The receiver could also recover the D_{reported} using D, $\alpha'_{\text{reported}}$, and $D(\Theta^{'})$ as they have already been sent. We term data cost as $L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta})$ and formulate it as follows

$$L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta}) = \textsc{Cost}(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta^{'})|D, \Theta^{'}, \hat{\Theta})$$

4.1.4 Total cost With $L(D, \Theta^{'}, \hat{\Theta})$ and $L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta})$ above, the total cost $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ will be:

$$\begin{aligned} L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta}) &= L(D, \Theta^{'}, \hat{\Theta}) + L(D_{\text{reported}}|D, \Theta^{'}, \hat{\Theta}) \\ &= \textsc{Cost}(\hat{\Theta}) + \textsc{Cost}(\Theta^{'} - \hat{\Theta}|\hat{\Theta}) \\ &\quad + \textsc{Cost}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})|\Theta^{'}, \hat{\Theta}) \\ &\quad + \textsc{Cost}(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta^{'})|D, \Theta^{'}, \hat{\Theta}) \end{aligned}$$

4.2 Problem statement

Note that our main objective is to estimate the total infections D. With $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$, we can state the problem as: Given the time sequence D_{reported}, epidemiological model O_{M}, and a calibration procedure Calibrate, find D^{*} that minimizes the MDL total cost i.e.

$$D^{*} = \arg\min_{D} \; L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$$

4.3 Algorithm

Next, we will present our algorithm to solve the problem in section 4.2. Note that directly searching D^{*} naively is intractable since D^{*} is a daily sequence not a scalar. Instead, we propose first finding a "good enough" reported rate $\alpha^{*}_{\text{reported}}$ quickly with the constraint $D = \frac{D_{\text{reported}}}{\alpha^{*}_{\text{reported}}}$ to reduce the search space. Then with this $\alpha^{*}_{\text{reported}}$, we can search for the optimal D^{*}. Hence we propose a two-step algorithm: (i) do a linear search to find a good reported rate $\alpha^{*}_{\text{reported}}$ (ii) given the $\alpha^{*}_{\text{reported}}$ found above, use an optimization method to find the D^{*} that minimizes $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$ with $\alpha^{*}_{\text{reported}}$ constraints. The pseudo-code is given in Algorithm 1.

Algorithm 1 MdlInfer

```
Require: $O_{\text{M}}$, Calibration procedure Calibrate, and $D_{\text{reported}}$
Calibrate $\hat{\Theta} = \textsc{Calibrate}(O_{\text{M}}, D_{\text{reported}})$
The array to save the MDL cost: CostArray = [ ]
for $\alpha_{\text{reported}}$ in the grid search space from 0.01 to 1 with step 0.01 do
    $D = \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$
    Calibrate $\Theta^{'} = \textsc{Calibrate}(O_{\text{M}}, (D, D_{\text{reported}}))$
    CostArray[$\alpha_{\text{reported}}$] = $L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$
end for
$\alpha^{*}_{\text{reported}} = \arg\min_{\alpha_{\text{reported}}}$ CostArray[$\alpha_{\text{reported}}$]
Find the $D^{*} = \arg\min_{D} L(D_{\text{reported}}, D, \Theta^{'}, \hat{\Theta})$. (using the Nelder-Mead algorithm).
Ensure: Total infections $D^{*}$
```

5 Experiments

In this section, we will answer the following research questions

- **Question 1:** Can MdlInfer estimate the total

[1] https://github.com/AdityaLab/MDL-ODE-Missing/blob/main/Appendix.pdf

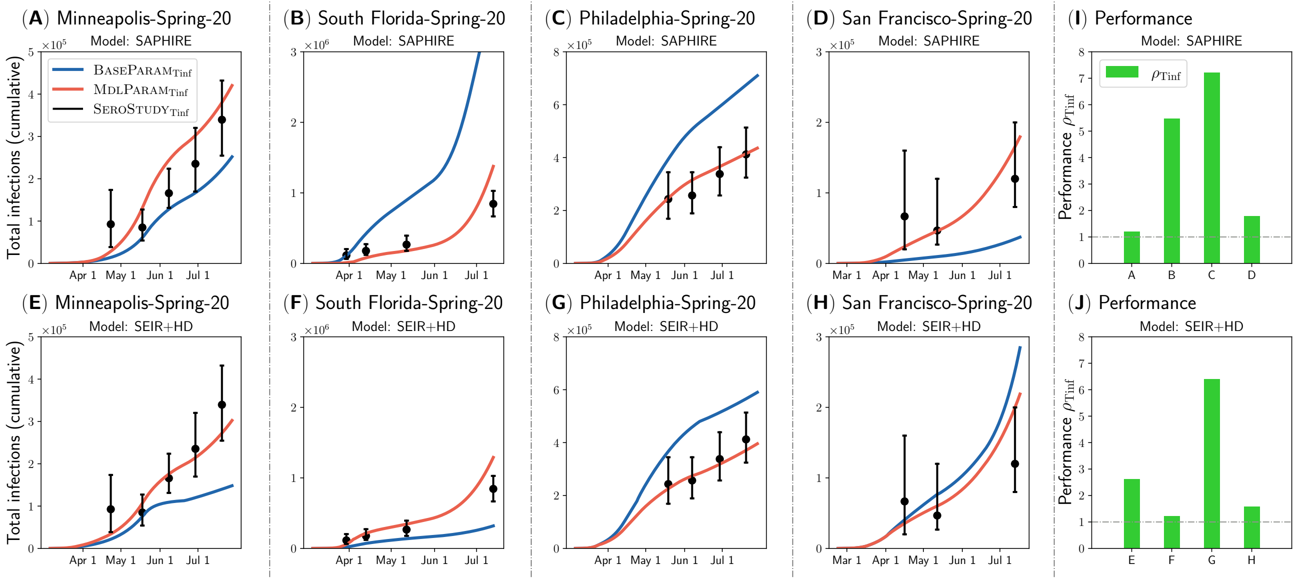

Figure 2: **MdlInfer (red) gives a closer estimation of total infections to serological studies (black) than BaseInfer (blue) on various geographical regions and time periods.** Note that both approaches try to fit the serological studies without being informed with them. (**A**)-(**H**) The red and blue curves represent MdlInfer's estimation of total infections, MdlParam$_{\text{Tinf}}$, and BaseInfer's estimation of total infections, BaseParam$_{\text{Tinf}}$, respectively. The black point estimates and confidence intervals represent the total infections estimated by serological studies [1, 17], SeroStudy$_{\text{Tinf}}$. (**A**)-(**D**) use SAPHIRE model and (**E**)-(**H**) use SEIR + HD model. (**I**)-(**J**) The performance metric, ρ_{Tinf}, comparing MdlParam$_{\text{Tinf}}$ against BaseParam$_{\text{Tinf}}$ in fitting serological studies is shown for each region. (**I**) is for SAPHIRE model in (**A**)-(**D**), and (**J**) is for SEIR + HD model in (**E**)-(**H**). Here, the values of ρ_{Tinf} are 1.20, 5.47, 7.21, and 1.79 in (**I**), and 2.62 ,1.22, 6.39, and 1.58 in (**J**). Note that ρ_{Tinf} larger than 1 means that MdlParam$_{\text{Tinf}}$ is closer to SeroStudy$_{\text{Tinf}}$ than BaseParam$_{\text{Tinf}}$. We show more experiments in the Appendix.

infections accurately than BaseInfer and fit the large-scale serological studies [38, 17, 45]?

- **Question 2:** Can MdlInfer fit the reported infection D_{reported} and forecast future infections accurately?
- **Question 3:** How does MdlInfer captures the trends of symptomatic rate?
- **Question 4:** How can MdlInfer help to evaluate the non-pharmaceutical interventions?

5.1 Setup

5.1.1 Dataset We choose 8 regions and periods based on the severity of the outbreak and the availability of serological studies and symptomatic surveillance data. The serological studies dataset consists of the point and 95% confidence interval estimates of the prevalence of antibodies to SARS-CoV-2 in these locations every 3–4 weeks from March to July 2020 [17, 1]. The symptomatic surveillance dataset consists of point estimate Rate$_{\text{Symp}}$ and standard error of the COVID-related symptomatic rate starting from April 6, 2020 [31, 36]. The reported infections are from New York Times [2], which consists of the daily time sequence of reported COVID-19 infections D_{reported} and the mortality $D_{\text{mortality}}$ (cumulative values) for each county in the US starting from January 21, 2020. In each region, we divide the timeline into two time periods: (i) observed period, when only the number of reported infections are available, and both BaseInfer and MdlInfer are used to learn the baseline parametrization (BaseParam) $\hat{\Theta}$ and MDL parametrization (MdlParam) Θ^*, and (ii) forecast period, where we evaluate the forecasts generated by the parametrizations learned in the observed period. To handle the time-varying reported rates, we divide the observed period into multiple sub-periods and learn different reported rates for each sub-period separately. Our data and code have been deposited in https://github.com/AdityaLab/MDL-ODE-Missing, which can be run on other datasets. A demo is also deposited there.

5.1.2 ODE model We compare MdlInfer and BaseInfer using two different ODE-based epidemiological models: SAPHIRE [16] and SEIR + HD [20] as O_{M}. Following their literature [16, 20], we use Markov Chain Monte Carlo (MCMC) as the calibration procedure Calibrate for SAPHIRE and iterated filtering (IF) for SEIR + HD, both of with are Bayesian

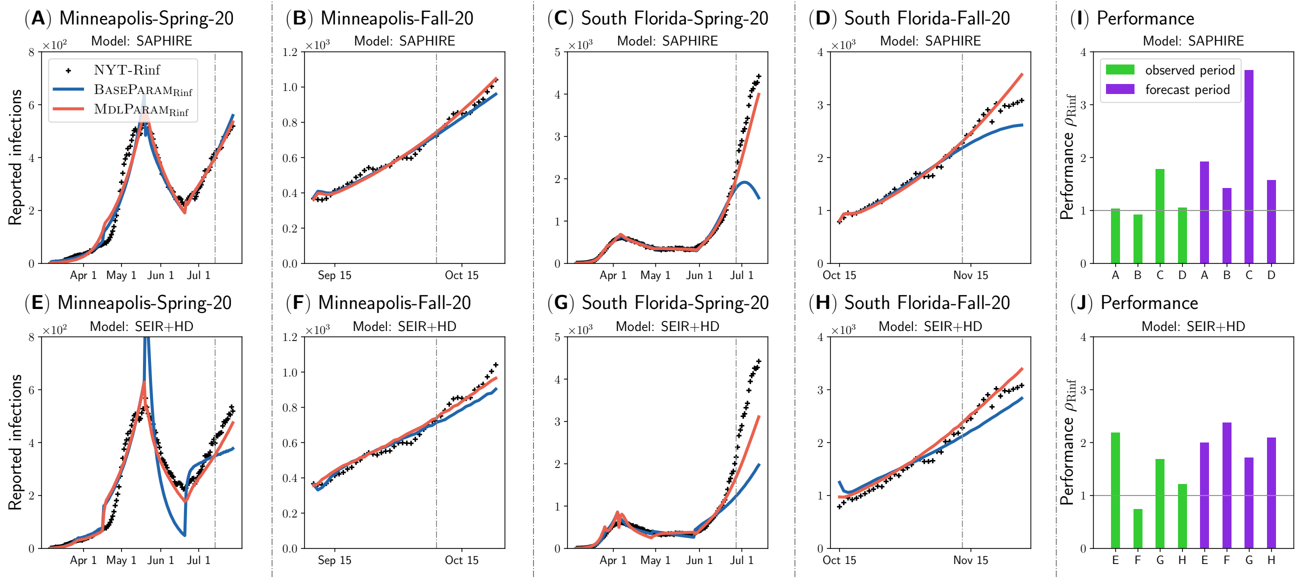

Figure 3: **MdlInfer (red) gives a closer estimation of reported infections (black) than BaseInfer (blue) on various geographical regions and time periods.** We use the reported infections in the observed period as inputs and try to forecast the future reported infections (forecast period). (**A**)-(**H**) The vertical grey dash line divides the observed period (left) and forecast period (right). The red and blue curves represent MdlInfer's estimation of reported infections, MdlParam$_{\text{Rinf}}$, and BaseInfer's estimation of reported infections, BaseParam$_{\text{Rinf}}$, respectively. The black plus symbols represent the reported infections collected by the New York Times (NYT-Rinf). (**A**)-(**D**) use SAPHIRE model and (**E**)-(**H**) use SEIR + HD model. (**I**)-(**J**) The performance metric, ρ_{Rinf}, comparing MdlParam$_{\text{Rinf}}$ against BaseParam$_{\text{Rinf}}$ in fitting reported infections is shown for each region. (**I**) is for SAPHIRE model in (**A**)-(**D**), and (**J**) is for SEIR + HD model in (**E**)-(**H**). Note that ρ_{Rinf} larger than 1 means that MdlParam$_{\text{Rinf}}$ is closer to NYT-Rinf than BaseParam$_{\text{Rinf}}$. We show more experiments in the Appendix.

approaches[18]. Both these epidemiological models have previously been shown to perform well in fitting reported infections and provided insight that was beneficial for the COVID-19 response.

5.1.3 Metrics To quantify the performance gap between the two approaches, we use the root mean squared error (RMSE) following the previous work [32, 33, 13, 12] for evaluation. To further demonstrate the performance, we further compute the ratio ρ as the fraction of the RMSE errors of BaseInfer over MdlInfer. Specifically, when the ratio is greater than 1, it implies that the MdlInfer is performing ρ times better than BaseInfer.

5.2 Q1: Estimating total infections

Here, we use the point estimates of the total infections calculated from serological studies as the ground truth (black dots shown in Figure 2). We call it SeroStudy$_{\text{Tinf}}$. We also plot MdlInfer's estimation of total infections, MdlParam$_{\text{Tinf}}$, in the same figure (red curve). To compare the performance of MdlInfer and BaseInfer with SeroStudy$_{\text{Tinf}}$, we use the cumulative value of estimated total infections. Note that values from the serological studies are not directly comparable with the total infections because of the lag between antibodies becoming detectable and infections being reported [1, 17]. In Figure 2, we have already accounted for this lag following CDC study guidelines [1, 17] (See Methods section for details). The vertical black lines shows a 95% confidence interval for SeroStudy$_{\text{Tinf}}$. The blue curve represents total infections estimated by BaseInfer, BaseParam$_{\text{Tinf}}$. As seen in the figure, MdlParam$_{\text{Tinf}}$ falls within the confidence interval of the estimates given by serological studies. Significantly, in Figure 2**B** and Figure 2**F** for South Florida, BaseInfer for SAPHIRE model [16] overestimates the total infections, while for SEIR + HD model underestimates the total infections. However, MdlInfer consistently estimates the total infections correctly. This observation shows that as needed, MdlParam$_{\text{Tinf}}$ can improve upon the BaseParam$_{\text{Tinf}}$ in either direction (i.e., by increasing or decreasing the total infections). Note that the MdlParam$_{\text{Tinf}}$ curves from both models are closer to the SeroStudy$_{\text{Tinf}}$ even when the BaseParam$_{\text{Tinf}}$ curves are different. The results of better accuracy in spite of various geographical regions and time periods show that MdlInfer is consistently able to estimate total infections more accurately.

In Figure 2**I** and Figure 2**J**, we plot $\rho_{\text{Tinf}} = \frac{\text{Rmse}(\text{BaseParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}{\text{Rmse}(\text{MdlParam}_{\text{Tinf}},\text{SeroStudy}_{\text{Tinf}})}$. Overall, the ρ_{Tinf} val-

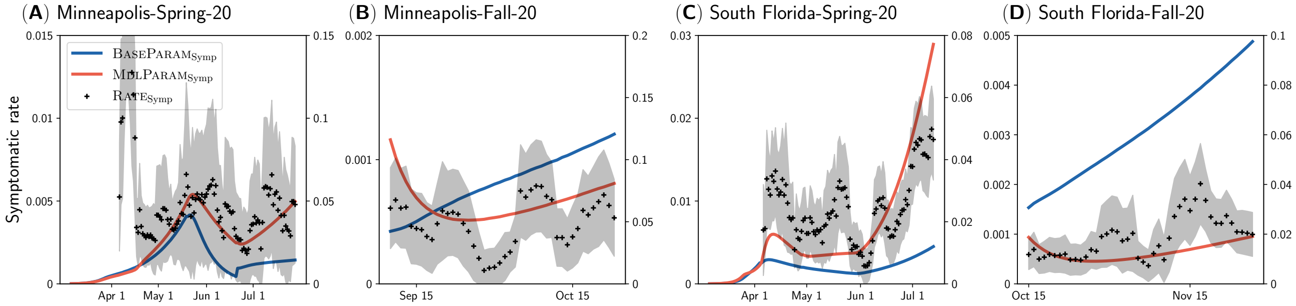

Figure 4: **MdlInfer (red) gives a closer estimation of the trends of symptomatic rate (black) than BaseInfer (blue) on various geographical regions and time periods.** (**A**)-(**D**) The red and blue curves represent MdlInfer's estimation of symptomatic rate, MdlParam$_{\text{Symp}}$, and BaseInfer's estimation of symptomatic rate, BaseParam$_{\text{Symp}}$, respectively. They use the y-scale on the left. The black points and the shaded regions are the point estimate with standard error for Rate$_{\text{Symp}}$ (the COVID-related symptomatic rates derived from the symptomatic surveillance dataset [31, 36]). They use the y-scale on the right. Note that we focus on trends instead of the exact numbers, hence MdlParam$_{\text{Symp}}$/BaseParam$_{\text{Symp}}$, and Rate$_{\text{Symp}}$ may scale differently. We show more experiments in the Appendix.

ues are greater than 1 in Figure 2**I** and Figure 2**J**, which indicates that MdlInfer performs better than BaseInfer. Note that even when the value of ρ_{Tinf} is 1.20 for Figure 2**A**, the improvement made by MdlParam$_{\text{Tinf}}$ over BaseParam$_{\text{Tinf}}$ in terms of RMSE is about 12091. Hence, one can conclude that MdlInfer is indeed superior to BaseInfer, when it comes to estimating total infections. We show more experiments in the Appendix.

5.3 Q2: Estimating reported infections

Here, we first use the observed period to learn the parametrizations. We then *forecast* the future reported infections (i.e., forecast periods), which were *not* accessible to the model while training. The results are summarized in Figure 3. In Figure 3**A** to Figure 3**H**, the vertical grey dash line divides the observed and forecast period. The black plus symbols represent reported infections collected by the New York Times, NYT-Rinf. The red curve represents MdlInfer's estimation of reported infections, MdlParam$_{\text{Rinf}}$. Similarly, the blue curve represents BaseInfer's estimation of reported infections, BaseParam$_{\text{Rinf}}$. Note that the curves to the right of the vertical grey line are future predictions. As seen in Figure 3, MdlParam$_{\text{Rinf}}$ aligns more closely with NYT-Rinf than BaseParam$_{\text{Rinf}}$, indicating the superiority of MdlInfer in fitting and forecasting reported infections.

We use a similar performance metric $\rho_{\text{Rinf}} = \frac{\text{Rmse}(\text{BaseParam}_{\text{Rinf}},\text{NYT-Rinf})}{\text{Rmse}(\text{MdlParam}_{\text{Rinf}},\text{NYT-Rinf})}$ to compare MdlParam$_{\text{Rinf}}$ against BaseParam$_{\text{Rinf}}$ in a manner similar to ρ_{Tinf}. In Figure 3**I** and Figure 3**J**, we plot the ρ_{Rinf} for the observed and forecast period. In both periods, we notice that the ρ_{Rinf} is close to or greater than 1. This further shows that MdlInfer has a better or at least closer fit for reported infections than BaseInfer. Additionally, the ρ_{Rinf} for the forecast period is even greater than ρ_{Rinf} for the observed period, which shows that MdlInfer performs even better than BaseInfer while forecasting.

Note that Figure 3**A**, **C**, **E**, **G** correspond to the early state of the COVID-19 epidemic in spring and summer 2020, and Figure 3**B**, **D**, **F**, **H** correspond to fall 2020. We can see that MdlInfer performs well in estimating temporal patterns at different stages of the COVID-19 epidemic. We show more experiments in the Appendix.

5.4 Q3: Estimating symptomatic rate trends

We validate this observation using Facebook's symptomatic surveillance dataset [31]. We plot MdlInfer's and BaseInfer's estimated symptomatic rate over time and overlay the estimates and standard error from the symptomatic surveillance data in Figure 4. The red and blue curves are MdlInfer's and BaseInfer's estimation of symptomatic rates, MdlParam$_{\text{Symp}}$ and BaseParam$_{\text{Symp}}$ respectively. Note that SAPHIRE model does not contain states corresponding to the symptomatic infections. Therefore, we only focus on SEIR + HD model. We compare the trends of the MdlParam$_{\text{Symp}}$ and BaseParam$_{\text{Symp}}$ with the symptomatic surveillance results. We focus on trends rather than actual values because the symptomatic rate numbers could be biased [31] (see Methods section for a detailed discussion) and therefore cannot be compared directly with model outputs like what we have done for serological studies. As seen in Figure 4, MdlParam$_{\text{Symp}}$ captures the trends of the surveyed symptomatic rate Rate$_{\text{Symp}}$ (black plus symbols) better than BaseParam$_{\text{Symp}}$. We show more experiments in the Appendix.

To summarize, these three sets of experiments in

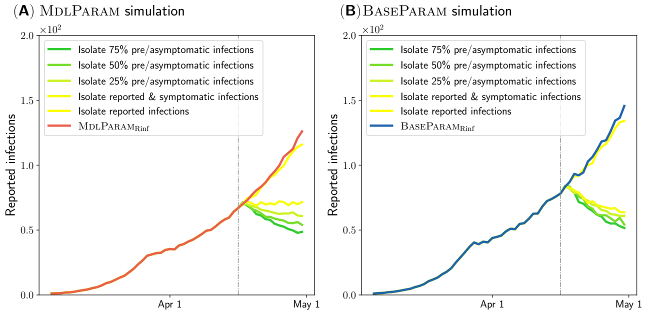

Figure 5: (**A**) MdlInfer reveals that non-pharmaceutical interventions (NPI) on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. Here, the red curve and other five curves represent the MdlInfer's estimation of reported infections for no NPI scenario and 5 different NPI scenarios described in the Results section. The vertical grey dash line divides the observed period (left) and forecast period (right). (**B**) Inaccurate estimation by BaseInfer may lead to wrong NPI conclusions. The blue curve and other five curves represent the BaseInfer's estimation of reported infections for no NPI scenario and the same 5 scenarios in (**B**).

section 5.2 to section 5.4 together demonstrate that BaseInfer fail to accurately estimate the total infections including unreported ones. On the other hand, MdlInfer estimates total infections closer to those estimated by serological studies and better fits reported infections and symptomatic rate trends.

5.5 Q4: Evaluate the effect of non-pharmaceutical Interventions

We have already shown that MdlInfer is able to estimate the number of total infections accurately. In the following three observations, we show that such accurate estimations are important for evaluating the effect of non-pharmaceutical interventions.

5.5.1 Non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic

Our simulations show that non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control COVID-19. Here, we plot the simulated reported infections of MdlParam in Figure 5**A** (red curve). We then repeat the simulation of reported infections for 5 different scenarios: (i) isolate just the reported infections, (ii) isolate just the symptomatic infections, and isolate symptomatic infections in addition to (iii) 25%, (iv) 50%, and (v) 75% of both asymptomatic and presymptomatic infections. In our setup, we assume that the infectivity reduces by half when a person is isolated. As seen in Figure 5**A**, when only the reported infections are isolated, there is almost no change in the "future" reported infections. However, when we isolate both the reported and symptomatic infections, the reported infections decreases significantly. Even here, the reported infections are still not in decreasing trend. On the other hand, non-pharmaceutical interventions for some fraction of asymptomatic and presymptomatic infections make reported infections decrease. Thus, we can conclude that NPIs on asymptomatic infections are essential in controlling the COVID-19 epidemic.

5.5.2 Accuracy of non-pharmaceutical intervention simulations relies on the good estimation of parametrization

Next, we also plot the simulated reported infections generated by BaseInfer in Figure 5**B** (blue curve). As seen in the figure, based on BaseInfer, we can infer that only non-pharmaceutical interventions on symptomatic infections are enough to control the COVID-19 epidemic. However, this has been proven to be incorrect by prior studies and real-world observations [26]. Therefore, we can conclude that the accuracy of non-pharmaceutical intervention simulation relies on the quality of the learned parametrization.

6 Conclusion

This study proposes MdlInfer, a data-driven model selection approach that automatically estimates the number of total infections using epidemiological models. Our approach leverages the information theoretic Minimum Description Length (MDL) principle and addresses several gaps in current practice including the long-term infeasibility of serological studies [17], and ad-hoc assumptions in epidemiological models [20, 25, 28, 16]. Overall, MdlInfer is a robust data-driven method to accurately estimate total infections, which will help data scientists, epidemiologists, and policy-makers to further improve existing ODE-based epidemiological models, make accurate forecasts, and combat future pandemics. More generally, MdlInfer opens up a new line of research in epidemic modeling using information theory.

Acknowledgements: This paper was partially supported by the NSF (Expeditions CCF-1918770 and CCF-1918656, CAREER IIS-2028586, RAPID IIS-2027862, Medium IIS-1955883, Medium IIS-2106961, Medium IIS-2403240, IIS-1931628, IIS-1955797, IIS-2027848, IIS-2331315), NIH 2R01GM109718, CDC MInD program U01CK000589, ORNL, Dolby faculty research award, and funds/computing resources from Georgia Tech and GTRI. B. A. was in part supported by the CDC MInD-Healthcare U01CK000531-Supplement. A.V.'s work is also supported in part by grants from the UVA Global Infectious Diseases Institute (GIDI).

References

[1] Commercial laboratory seroprevalence survey data. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/commercial-lab-surveys.html.

[2] Coronavirus in the u.s.:latest map and case count. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html.

[3] Covid-19 test price information, https://www.questdiagnostics.com/business-solutions/health-plans/covid-19/pricing.

[4] Hidden outbreaks spread through u.s. cities far earlier than americans knew, estimates say. https://www.nytimes.com/2020/04/23/us/coronavirus-early-outbreaks-cities.html.

[5] B. Adhikari, P. Rangudu, B. A. Prakash, and A. Vullikanti. Near-optimal mapping of network states using probes. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 108–116. SIAM, 2018.

[6] J. B. Aguilar, J. S. Faust, L. M. Westafer, and J. B. Gutierrez. Investigating the impact of asymptomatic carriers on covid-19 transmission. *MedRxiv*, 2020.

[7] A. N. Angelopoulos, R. Pathak, R. Varma, and M. I. Jordan. On identifying and mitigating bias in the estimation of the covid-19 case fatality rate. *arXiv preprint arXiv:2003.08592*, 2020.

[8] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang. Presumed asymptomatic carrier transmission of covid-19. *JAMA*, 323(14):1406–1407, 2020.

[9] L. Blier and Y. Ollivier. The description length of deep learning models. *Advances in Neural Information Processing Systems*, 31, 2018.

[10] K. Budhathoki and J. Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, 2018.

[11] Q. Cao and B. Heydari. Micro-level social structures and the success of covid-19 national policies. *Nature Computational Science*, 2(9):595–604, 2022.

[12] E. Y. Cramer, Y. Huang, Y. Wang, E. L. Ray, M. Cornell, J. Bracher, A. Brennen, A. J. C. Rivadeneira, A. Gerding, K. House, et al. The united states covid-19 forecast hub dataset. *Scientific data*, 9(1):462, 2022.

[13] E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, et al. Evaluation of individual and ensemble probabilistic forecasts of covid-19 mortality in the united states. *PNAS*, 119(15):e2113561119, 2022.

[14] F. Gao and L. Han. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, 2012.

[15] V. Gopalakrishnan, S. Pethe, S. Kefayati, R. Srinivasan, P. Hake, A. Deshpande, X. Liu, E. Hoang, M. Davila, S. Bianco, et al. Globally local: Hyper-local modeling for accurate forecast of covid-19. *Epidemics*, 37:100510, 2021.

[16] X. Hao, S. Cheng, D. Wu, T. Wu, X. Lin, and C. Wang. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584(7821):420–424, 2020.

[17] F. P. Havers, C. Reed, T. Lim, J. M. Montgomery, J. D. Klena, A. J. Hall, A. M. Fry, D. L. Cannon, C.-F. Chiang, A. Gibbons, et al. Seroprevalence of antibodies to sars-cov-2 in 10 sites in the united states, march 23-may 12, 2020. *JAMA internal medicine*, 180(12):1576–1586, 2020.

[18] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King. Inference for dynamic and latent variable models via iterated, perturbed bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719–724, 2015.

[19] H. Jang, S. Justice, P. M. Polgreen, A. M. Segre, D. K. Sewell, and S. V. Pemmaraju. Evaluating architectural changes to alter pathogen dynamics in a dialysis unit. In *ASONAM*, pages 961–968. IEEE, 2019.

[20] M. P. Kain, M. L. Childs, A. D. Becker, and E. A. Mordecai. Chopping the tail: How preventing superspreading can help to maintain covid-19 control. *Epidemics*, 34:100430, 2021.

[21] D. Koutra, U. Kang, J. Vreeken, and C. Faloutsos. Vog: Summarizing and understanding large graphs. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 91–99. SIAM, 2014.

[22] M. U. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[23] S. Lai, N. W. Ruktanonchai, L. Zhou, O. Prosper, W. Luo, J. R. Floyd, A. Wesolowski, M. Santillana, C. Zhang, X. Du, et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. *Nature*, 585(7825):410–413, 2020.

[24] T. C. Lee. An introduction to coding theory and the two-part minimum description length principle. *International statistical review*, 69(2):169–183, 2001.

[25] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493, 2020.

[26] S. M. Moghadas, M. C. Fitzpatrick, P. Sah, A. Pandey, A. Shoukat, B. H. Singer, and A. P. Galvani. The implications of silent transmission for the control of covid-19 outbreaks. *Proceedings of the National Academy of Sciences*, 117(30):17513–17515, 2020.

[27] P. Padmanabhan, R. Desikan, and N. M. Dixit. Modeling how antibody responses may determine the efficacy of covid-19 vaccines. *Nature Computational Science*, 2(2):123–131, 2022.

[28] S. Pei, S. Kandula, and J. Shaman. Differential effects of intervention timing on covid-19 spread in the united states. *Science advances*, 6(49):eabd6370, 2020.

[29] S. Pei, T. K. Yamana, S. Kandula, M. Galanti, and J. Shaman. Burden and characteristics of covid-19 in the united states during 2020. *Nature*, 598(7880):338–

341, 2021.

[30] W. H. Press and R. C. Levin. Modeling, post covid-19. *Science*, 370(6520):1015–1015, 2020.

[31] A. Reinhart, L. Brooks, M. Jahja, A. Rumack, J. Tang, S. Agrawal, W. Al Saeed, T. Arnold, A. Basu, J. Bien, et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[32] A. Rodríguez, J. Cui, N. Ramakrishnan, B. Adhikari, and B. A. Prakash. Einns: epidemiologically-informed neural networks. In *AAAI*, volume 37, pages 14453–14460, 2023.

[33] A. Rodriguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, and B. A. Prakash. Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *AAAI*, volume 35, pages 15393–15400, 2021.

[34] P. Roy, S. Sarkar, S. Biswas, F. Chen, Z. Chen, N. Ramakrishnan, and C.-T. Lu. Deep diffusion-based forecasting of covid-19 by incorporating network-level mobility information. In *ASONAM*, pages 168–175, 2021.

[35] T. W. Russell, N. Golding, J. Hellewell, S. Abbott, L. Wright, C. A. Pearson, K. van Zandvoort, C. I. Jarvis, H. Gibbs, Y. Liu, et al. Reconstructing the early global dynamics of under-ascertained covid-19 cases and infections. *BMC medicine*, 18(1):1–9, 2020.

[36] J. A. Salomon, A. Reinhart, A. Bilinski, E. J. Chua, W. La Motte-Kerr, M. M. Rönn, M. B. Reitsma, K. A. Morris, S. LaRocca, T. H. Farag, et al. The us covid-19 trends and impact survey: Continuous real-time measurement of covid-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118(51), 2021.

[37] J. Shaman. An estimation of undetected covid cases in france. *Nature*, 590:38–39, 2020.

[38] N. Sood, P. Simon, P. Ebner, D. Eichner, J. Reynolds, E. Bendavid, and J. Bhattacharya. Seroprevalence of sars-cov-2–specific antibodies among adults in los angeles county, california, on april 10-11, 2020. *JAMA*, 323(23):2425–2427, 2020.

[39] S. Stockmaier, N. Stroeymeyt, E. C. Shattuck, D. M. Hawley, L. A. Meyers, and D. I. Bolnick. Infectious diseases and social distancing in nature. *Science*, 371(6533), 2021.

[40] R. Subramanian, Q. He, and M. Pascual. Quantifying asymptomatic infection and transmission of covid-19 in new york city using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[41] S. Tiwari, C. Vyasarayani, and A. Chatterjee. Data suggest covid-19 affected numbers greatly exceeded detected numbers, in four european countries, as per a delayed seiqr model. *Scientific reports*, 11(1):1–12, 2021.

[42] C. R. Wells, P. Sah, S. M. Moghadas, A. Pandey, A. Shoukat, Y. Wang, Z. Wang, L. A. Meyers, B. H. Singer, and A. P. Galvani. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences*, 117(13):7504–7509, 2020.

[43] B. Wilder, M. Charpignon, J. A. Killian, H.-C. Ou, A. Mate, S. Jabbari, A. Perrault, A. N. Desai, M. Tambe, and M. S. Majumder. Modeling between-population variation in covid-19 dynamics in hubei, lombardy, and new york city. *Proceedings of the National Academy of Sciences*, 117(41):25904–25910, 2020.

[44] S. L. Wu, A. N. Mertens, Y. S. Crider, A. Nguyen, N. N. Pokpongkiat, S. Djajadi, A. Seth, M. S. Hsiang, J. M. Colford, A. Reingold, et al. Substantial underestimation of sars-cov-2 infection in the united states. *Nature communications*, 11(1):1–10, 2020.

[45] W. Zhang, J. P. Govindavari, B. D. Davis, S. S. Chen, J. T. Kim, J. Song, J. Lopategui, J. T. Plummer, and E. Vail. Analysis of genomic characteristics and transmission routes of patients with confirmed sars-cov-2 in southern california during the early stage of the us covid-19 pandemic. *JAMA network open*, 3(10):e2024191, 2020.