

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 parameterization 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 parameterization 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, 38, 36, 24]. 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 [40].

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 [28]. One of the benefits of using this definition is that it includes asymptomatic infections, which may also contribute substantially to spread [39, 25]. 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 [24, 34, 7, 29, 21, 22, 41, 15, 19, 42, 43, 10, 26]. 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) [16, 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 succinctly way: As we know D , it is trivial to get the reported rate α_{reported} . Then with both D and α_{reported} , it will be trivial to describe

^{*}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.

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

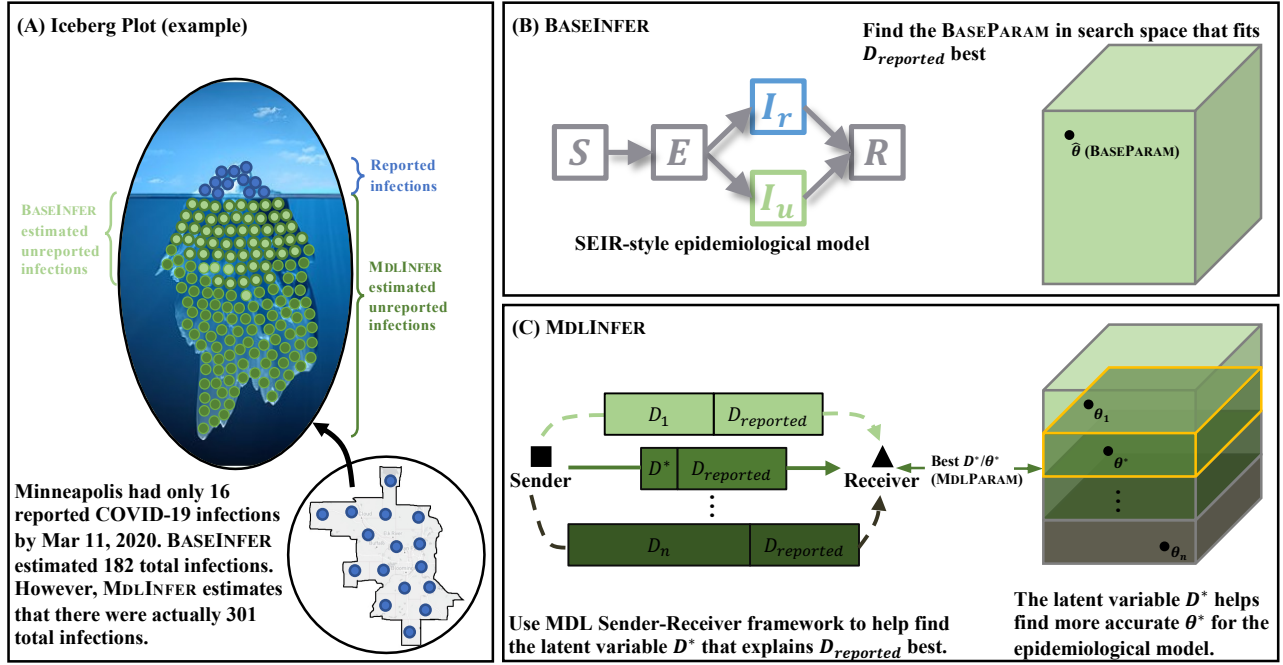


Figure 1: **Overview of our problem and methodology.** (A) We visualize the idea of reported rates using the iceberg. The visible portion above water are the reported infections, which is only a fraction of the whole iceberg representing total infections. Light green corresponds to the 182 unreported 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’ parameterization *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.

$D_{reported}$, as it is simply $D \times \alpha_{reported}$ 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 [16, 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 [37, 16, 44]. 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, 16].

2.2 Minimum Description Length framework

MDL frameworks has been widely used for numerous optimization problems ranging from network summarization [20], causality inference [9], and failure detection in critical infrastructures [5]. 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 [18]. In general, it 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 [14] or Bayesian approaches [19, 15]. BASEINFER is generally a complex, high-dimensional problem, since there are multiple parameters interacting with each other. To make matters worse, there exist many possible parameterizations that show similar performance (e.g. in RMSE, likelihood) yet correspond to vastly different reported rates, and

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})
$\hat{\Theta}$	Parameterization 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
Θ'	Parameterization 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'

BASEINFER cannot select between these competing parameterizations in a principled way.

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 we our goal 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 parameterization Θ' into MODEL. We further choose to also add $\hat{\Theta}$ estimated by BASEINFER, making our MODEL to be $(D, \Theta', \hat{\Theta})$. With $\text{MODEL} = (D, \Theta', \hat{\Theta})$, our MODEL space will be all possible daily sequences for D and all possible parameterizations for Θ' 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 Supplementary Materials.

4.1.2 Model cost With $\text{MODEL} = (D, \Theta', \hat{\Theta})$, we conceptualize the model cost by imagining that the sender S will send the $\text{MODEL} = (D, \Theta', \hat{\Theta})$ to the receiver R in three parts: (i) first send the $\hat{\Theta}$ by encoding $\hat{\Theta}$ directly (ii) next send the Θ' given $\hat{\Theta}$ by encoding $\Theta' - \hat{\Theta}$ and (iii) then send D given Θ' 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: $\text{COST}(\hat{\Theta})$, $\text{COST}(\Theta'|\hat{\Theta})$, and $\text{COST}(D|\Theta', \hat{\Theta})$. Hence

$$L(D, \Theta', \hat{\Theta}) = \text{COST}(\hat{\Theta}) + \text{COST}(\Theta' - \hat{\Theta}|\hat{\Theta}) + \text{COST}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\hat{\Theta})|\Theta', \hat{\Theta})$$

Here, the $\text{COST}(\cdot)$ 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 Supplementary Materials.

4.1.3 Data cost We need to send the $\text{DATA} = D_{\text{reported}}$ next given the MODEL. Given $\text{MODEL} = (D, \Theta', \hat{\Theta})$, 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

$$L(D_{\text{reported}}|D, \Theta', \hat{\Theta}) = \text{COST}\left(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\Theta')|D, \Theta', \hat{\Theta}\right)$$

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:

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

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_M , Calibration procedure CALIBRATE , and

- D_{reported}
- 1: Calibrate $\hat{\Theta} = \text{CALIBRATE}(O_M, D_{\text{reported}})$
- 2: The array to save the MDL cost: $\text{CostArray} = []$
- 3: **for** α_{reported} in the grid search space from 0.01 to 1 with step 0.01 **do**
- 4: $D = \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$
- 5: Calibrate $\Theta' = \text{CALIBRATE}(O_M, (D, D_{\text{reported}}))$
- 6: $\text{CostArray}[\alpha_{\text{reported}}] = L(D_{\text{reported}}, D, \Theta', \hat{\Theta})$
- 7: **end for**
- 8: $\alpha^*_{\text{reported}} = \arg \min_{\alpha_{\text{reported}}} \text{CostArray}[\alpha_{\text{reported}}]$
- 9: Find the $D^* = \arg \min_D L(D_{\text{reported}}, D, \Theta', \hat{\Theta})$. (using the Nelder-Mead algorithm).

Ensure: Total infections D^*

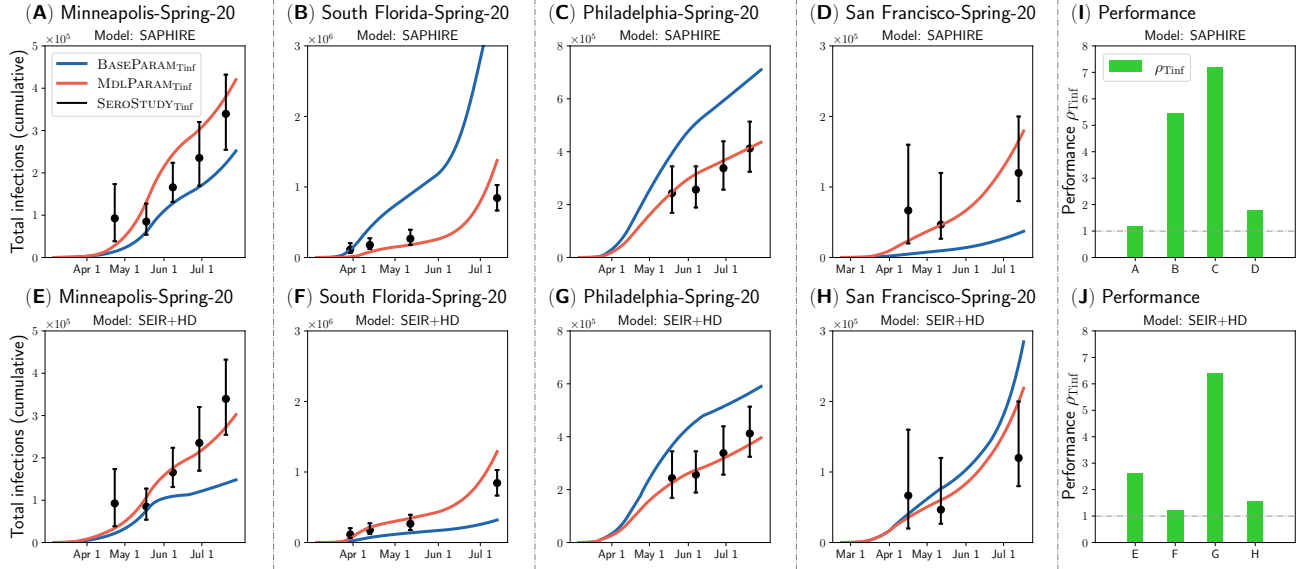


Figure 2: MDLINFER (red) gives a closer estimation of total infections to serological studies (black) than BASEINFER (blue) on various geographical regions and time periods. Note that both approaches try to fit the serological studies without being informed with them. (A)-(H) The red and blue curves represent MDLINFER’s estimation of total infections, MDLPARAM_{Tinf}, and BASEINFER’s estimation of total infections, BASEPARAM_{Tinf}, respectively. The black point estimates and confidence intervals represent the total infections estimated by serological studies [1, 16], SEROSTUDY_{Tinf}. (A)-(D) use SAPHIRE model and (E)-(H) use SEIR + HD model. (I)-(J) The performance metric, ρ_{Tinf} , comparing MDLPARAM_{Tinf} against BASEPARAM_{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_{Tinf} is closer to SEROSTUDY_{Tinf} than BASEPARAM_{Tinf}. We show more experiments in the Supplementary Materials.

5 Experiments

In this section, we will answer the following research questions

- **Question 1:** Can MDLINFER estimate the total infections accurately than BASEINFER and fit the large-scale serological studies [37, 16, 44]?
- **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 [16, 1]. The symptomatic surveillance dataset consists of point estimate $RATE_{symp}$ and standard error of the COVID-related symptomatic rate starting from April 6, 2020 [30, 35]. 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 $D_{mortality}$ (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 parameterization (BASEPARAM) $\hat{\Theta}$ and MDL parameterization (MDLPARAM) $\hat{\Theta}^*$, and (ii) forecast period, where we evaluate the forecasts generated by the parameterizations 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://anonymous.4open.science/r/MDL-ODE-Missing-05DA/> 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 [15] and SEIR + HD [19] as O_M . Following their literature [15, 19], we use Markov Chain Monte Carlo (MCMC) as the calibration procedure CALIBRATE for SAPHIRE and iterated filtering (IF) for SEIR + HD, both of which are Bayesian approaches [17]. 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 [31, 32, 12, 11] 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 $\text{SEROSTUDY}_{\text{Tinf}}$. We also plot MDLINFER's estimation of total infections, $\text{MDLPARAM}_{\text{Tinf}}$, in the same figure (red curve). To compare the performance of MDLINFER and BASEINFER with $\text{SEROSTUDY}_{\text{Tinf}}$, 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, 16]. In Figure 2, we have already accounted for this lag following CDC study guidelines [1, 16] (See Methods section for details). The vertical black lines show a 95% confidence interval for $\text{SEROSTUDY}_{\text{Tinf}}$. The blue curve represents total infections estimated by BASEINFER, $\text{BASEPARAM}_{\text{Tinf}}$. As seen in the figure, $\text{MDLPARAM}_{\text{Tinf}}$ falls within the confidence interval of the estimates given by serological studies. Significantly, in Figure 2B and Figure 2F for South Florida, BASEINFER for SAPHIRE model [15] 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, $\text{MDLPARAM}_{\text{Tinf}}$ can improve upon the $\text{BASEPARAM}_{\text{Tinf}}$

in either direction (i.e., by increasing or decreasing the total infections). Note that the $\text{MDLPARAM}_{\text{Tinf}}$ curves from both models are closer to the $\text{SEROSTUDY}_{\text{Tinf}}$ even when the $\text{BASEPARAM}_{\text{Tinf}}$ 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 2I and Figure 2J, 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} values are greater than 1 in Figure 2I and Figure 2J, which indicates that MDLINFER performs better than BASEINFER. Note that even when the value of ρ_{Tinf} is 1.20 for Figure 2A, the improvement made by $\text{MDLPARAM}_{\text{Tinf}}$ over $\text{BASEPARAM}_{\text{Tinf}}$ 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 Supplementary Materials.

5.3 Q2: Estimating reported infections Here, we first use the observed period to learn the parameterizations. 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 3A to Figure 3H, 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, $\text{MDLPARAM}_{\text{Rinf}}$. Similarly, the blue curve represents BASEINFER's estimation of reported infections, $\text{BASEPARAM}_{\text{Rinf}}$. Note that the curves to the right of the vertical grey line are future predictions. As seen in Figure 3, $\text{MDLPARAM}_{\text{Rinf}}$ aligns more closely with NYT-Rinf than $\text{BASEPARAM}_{\text{Rinf}}$, 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 $\text{MDLPARAM}_{\text{Rinf}}$ against $\text{BASEPARAM}_{\text{Rinf}}$ in a manner similar to ρ_{Tinf} . In Figure 3I and Figure 3J, 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 3A, C, E, G correspond to the early state of the COVID-19 epidemic in spring and summer 2020, and Figure 3B, D, F, H correspond to fall 2020. We can see that MDLINFER performs well in

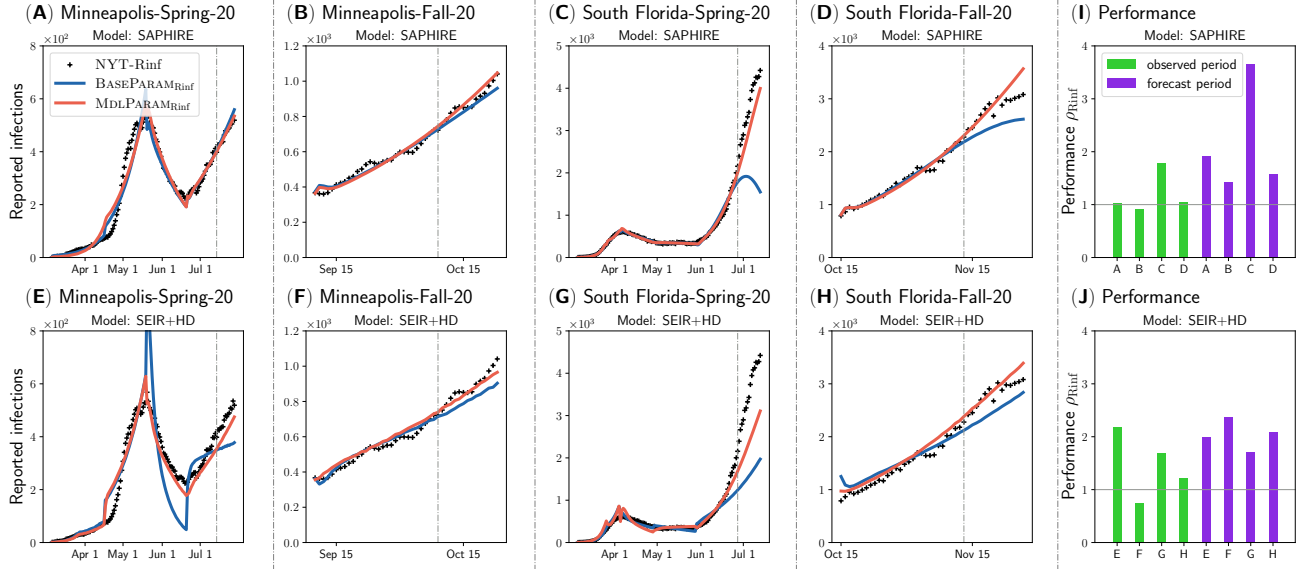


Figure 3: MDLINFER (red) gives a closer estimation of reported infections (black) than BASEINFER (blue) on various geographical regions and time periods. We use the reported infections in the observed period as inputs and try to forecast the future reported infections (forecast period). (A)-(H) The vertical 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_{Rinf}, and BASEINFER’s estimation of reported infections, BASEPARAM_{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_{Rinf} against BASEPARAM_{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_{Rinf} is closer to NYT-Rinf than BASEPARAM_{Rinf}. We show more experiments in the Supplementary Materials.

estimating temporal patterns at different stages of the COVID-19 epidemic. We show more experiments in the Supplementary Materials.

5.4 Q3: Estimating symptomatic rate trends

We validate this observation using Facebook’s symptomatic surveillance dataset [30]. 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_{Symp} and BASEPARAM_{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_{Symp} and BASEPARAM_{Symp} with the symptomatic surveillance results. We focus on trends rather than actual values because the symptomatic rate numbers could be biased [30] (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_{Symp} captures the trends of the surveyed symptomatic rate RATE_{Symp} (black plus symbols) better than BASEPARAM_{Symp}. We show more experiments in the Supplementary Materials.

To summarize, these three sets of experiments in 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 epi-

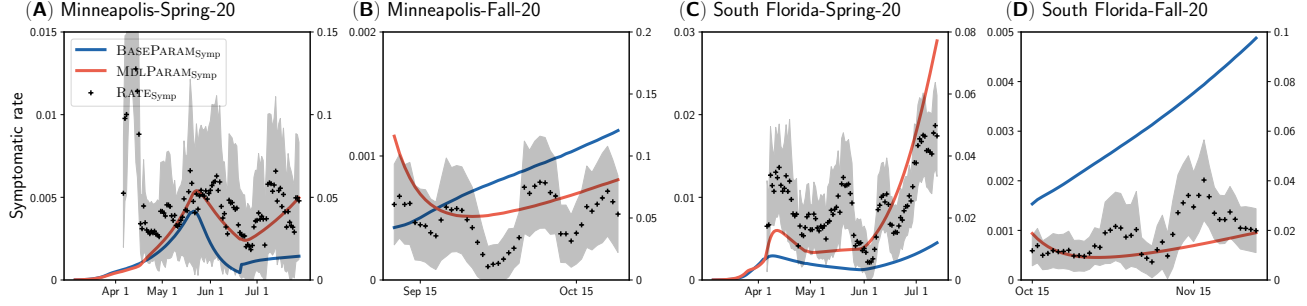


Figure 4: **MDLINFER (red) gives a closer estimation of the trends of symptomatic rate (black) than BASEINFER (blue) on various geographical regions and time periods.** (A)-(D) The red and blue curves represent MDLINFER’s estimation of symptomatic rate, MDLPARAM_{Symp}, and BASEINFER’s estimation of symptomatic rate, BASEPARAM_{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_{Symp} (the COVID-related symptomatic rates derived from the symptomatic surveillance dataset [30, 35]). They use the y-scale on the right. Note that we focus on trends instead of the exact numbers, hence MDLPARAM_{Symp}/BASEPARAM_{Symp}, and RATE_{Symp} may scale differently. We show more experiments in the Supplementary Materials.

demid 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 5A (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 5A, 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 non-pharmaceutical interventions 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 parameterization Next, we also plot the simulated reported infections generated by BASEINFER in Figure 5B (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 [25]. Therefore, we can conclude that the accuracy of non-pharmaceutical intervention simulation relies on the quality of the learned parameterization.

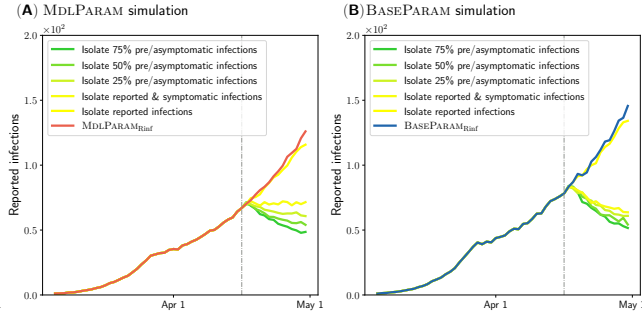


Figure 5: (A) MDLINFER reveals that non-pharmaceutical interventions (NPI) on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. Here, the red curve and other five curves represent the MDLINFER’s estimation of reported infections for no NPI scenario and 5 different 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 non-pharmaceutical intervention 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).

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 [16], and ad-hoc assumptions in epidemiological models [19, 24, 27, 15]. Overall, MDLINFER is a robust data-driven method to accurately estimate total infections, which will help data scientists, epidemiologists, and policymakers 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.

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. Vulikanti. 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] K. Budhathoki and J. Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, 2018.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] F. Gao and L. Han. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, 2012.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] M. P. Kain, M. L. Childs, A. D. Becker, and E. A. Mordecai. Chopping the tail: How preventing super-spreading can help to maintain covid-19 control. *Epidemics*, 34:100430, 2021.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] T. C. Lee. An introduction to coding theory and the two-part minimum description length principle. *International statistical review*, 69(2):169–183, 2001.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] W. H. Press and R. C. Levin. Modeling, post covid-19. *Science*, 370(6520):1015–1015, 2020.

- [30] 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.
- [31] 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.
- [32] A. Rodriguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, and B. A. Prakash. Deep-covid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting. In *AAAI*, volume 35, pages 15393–15400, 2021.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] J. Shaman. An estimation of undetected covid cases in france. *Nature*, 590:38–39, 2020.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] S. Tiwari, C. Vyasrayani, 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.
- [41] 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.
- [42] 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.
- [43] S. L. Wu, A. N. Mertens, Y. S. Crider, A. Nguyen, N. N. Pokpongkiet, 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.
- [44] 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.

Supplementary Materials

A Data

A.1 New York Times reported infections [2]

This dataset (NYT-Rinf) consists of the time sequence of reported infections D_{reported} and reported mortality $D_{\text{mortality}}$ in each county across the U.S. since the beginning of the COVID-19 pandemic (January 21, 2020) to current. For each county, the NYT-Rinf dataset provides the date, FIPS code, and the cumulative values of reported infections and mortality. Here, we use the averaged counts over 14 days to eliminate noise.

A.2 Serological studies [16, 1] This dataset consists of the point estimate and 95% confidence interval of the prevalence of antibodies to SARS-CoV-2 in 10 US locations every 3-4 weeks during March to July 2020. The serological studies use the blood specimens collected from population. For each location, CDC collects 1800 samples approximately every 3-4 weeks. Using the prevalence of the antibodies and the population, we can compute the estimated total infections and 95% confidence interval in the location. However, we cannot compare this number with the epidemiological model estimated total infection numbers directly as mentioned in the main article Methods section. We account for this problem by comparing the serological studies numbers with the estimated total infections of 7 days prior to the first day of specimen collection period (as suggested by the CDC serological studies work [16]).

A.3 Symptomatic surveillance [30] This dataset comes from Facebook's symptomatic survey [30]. The survey started on April 6, 2020 to current. As of January 28, 2021, there were a total of 16,398,000 participants, with the average daily participants number of 55,000. The survey asks a series of questions designed to help researchers understand the spread of COVID-19 and its effect on people in the United States. For the signal, they estimate the percentage of self-reported COVID-19 symptoms in population defined as fever along with either cough, shortness of breath, or difficulty breathing [30]. The dataset also includes weighted version which accounts for the differences between Facebook users and the United States population. In the experiments, we contrast the symptomatic rate trends inferred by our approach against the weighted data from the survey.

B ODE Model

B.1 SAPHIRE Model We use the SAPHIRE model [19] as one epidemiological model O_M in our experiments. The compartmental diagram of SAPHIRE

model is shown in Figure S1. The SAPHIRE model has 9 different parameters. Note that only two parameters are calibrated, while the rest are fixed.

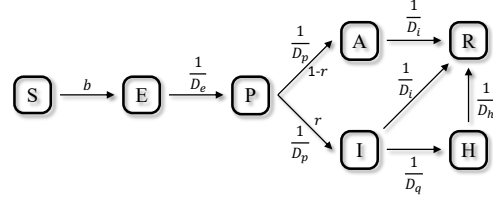


Figure S1: Compartmental diagram of SAPHIRE model [15].

In this article, we expect the epidemiological model to calibrate on both reported infections D_{reported} and candidate unreported infections $D_{\text{unreported}}$. We compute the newly reported infections and unreported infections as follows:

1. New reported infections = $\frac{\alpha P}{D_p}$: $\frac{P}{D_p}$ represents the number of new infections from presymptomatic infections every day in O_M . Here, we assume α proportion of new infections every day will be that day's new reported infections.
2. New unreported infections = $\frac{(1-\alpha)P}{D_p}$: Then, the $1 - \alpha$ proportion of new infections every day will be that day's new unreported infections.

B.2 SEIR+HD Model We also use the SEIR+HD model [19] as another epidemiological model O_M in our experiments. The compartmental diagram of SEIR+HD model is shown in Figure S2. The SEIR+HD model has of 21 different parameters. Note that only three parameters are calibrated, while the rest are fixed.

Similarly to SAPHIRE model, we still expect the epidemiological model to calibrate on reported infections D_{reported} and candidate unreported infections $D_{\text{unreported}}$. Hence we extend its calibration procedure to infer two more parameters: α and α_1 (proportion of new symptomatic infections that are reported). We compute the newly reported infections and unreported infections as follows:

1. New reported infections = $\alpha_1 \times (N_{I_P I_S} + N_{I_P I_M})$: $N_{I_P I_S} + N_{I_P I_M}$ represents the number of new symptomatic infections every day in O_M . Here, we assume α_1 proportion of new symptomatic

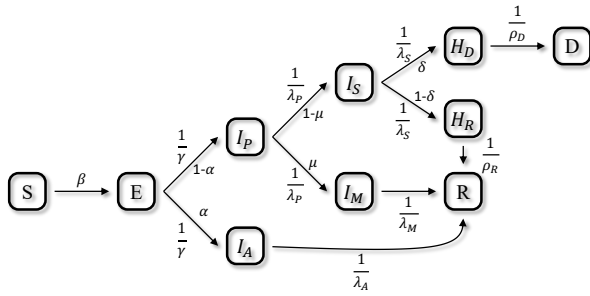


Figure S2: Compartmental diagram of SEIR+HD model [15].

infections every day will be that day's new reported infections.

2. New unreported infections = $(1 - \alpha_1) \times (N_{I_p I_s} + N_{I_p I_m}) + N_{E I_a}$:

Then, the $1 - \alpha_1$ proportion of new symptomatic infections every day and new asymptomatic infections every day will be that day's new unreported infections.

B.3 Baseline Parameterization By calibrating O_M on D_{reported} , we get the baseline parameterization $\mathbf{p}$:

$$\mathbf{p} = \text{CALIBRATE}(O_M, \{D_{\text{reported}}, \text{others}\})$$

By running the epidemiological model with $\mathbf{p}$, O_M will output the estimated reported infections $D_{\text{reported}}(\mathbf{p})$, estimated unreported infections $D_{\text{unreported}}(\mathbf{p})$, and estimated total infections $D(\mathbf{p}) = D_{\text{reported}}(\mathbf{p}) + D_{\text{unreported}}(\mathbf{p})$. We can also calculate the reported rate α_{reported} as follows:

$$\alpha_{\text{reported}} = \frac{\sum D_{\text{reported}}(\mathbf{p})}{\sum D(\mathbf{p})}$$

Here, we sum over the daily sequence $D_{\text{reported}}(\mathbf{p})$ and $D(\mathbf{p})$ to calculate a scalar as the reported rate for MDL formulation.

B.4 MDLINFER Parameterization Similarly by calibrating O_M on D_{reported} and $D_{\text{unreported}}$, we get the candidate parameterization $\mathbf{p}'$:

$$\mathbf{p}' = \text{CALIBRATE}(O_M, \{D_{\text{reported}}, D_{\text{unreported}}, \text{others}\})$$

By running the epidemiological model with $\mathbf{p}'$, O_M will output the estimated reported infections $D_{\text{reported}}(\mathbf{p}')$, estimated unreported infections $D_{\text{unreported}}(\mathbf{p}')$, and estimated total infections $D(\mathbf{p}') = D_{\text{reported}}(\mathbf{p}') +$

$D_{\text{unreported}}(\mathbf{p}')$. Similarly, we can calculate the reported rate $\alpha'_{\text{reported}}$ as follows:

$$\alpha'_{\text{reported}} = \frac{\sum D_{\text{reported}}(\mathbf{p}')}{\sum D(\mathbf{p}')}$$

With the calibration process, $\mathbf{p}$, and $\mathbf{p}'$ defined, we can next formalize the MDL cost.

C Methodology

C.1 Sender-receiver Framework Here, we use the two-part sender-receiver framework based on the Minimum Description Length (MDL) principle. The goal of the framework is to transmit the DATA in possession of the Sender S to the receiver R using a MODEL. We do this by identifying the MODEL that describes the DATA such that the total number of bits needed to encode both the MODEL and the DATA is minimized. The number of bits required to encode both the MODEL and the DATA is given by the cost function L , which has two components: (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 DATA given the MODEL.

C.2 Model Space: Other Choice In this work, the DATA is D_{reported} . One idea for defining the MODEL space is to use $\mathbf{p}$. With such a MODEL, the receiver R can easily compute first $D_{\text{reported}}(\mathbf{p})$ given $\mathbf{p}$. Then the sender S will only need to encode and send the difference between $D_{\text{reported}}(\mathbf{p})$ and D_{reported} so that the receiver can recover the DATA fully. However, this has the disadvantage that slightly different $\mathbf{p}$ could lead to vastly different $D_{\text{reported}}(\mathbf{p})$, and so the optimization problem will become hard to solve. To account for this, we propose MODEL as $\text{MODEL} = (D, \mathbf{p}', \mathbf{p})$ as described in the main article, which consists of three components.

C.3 Model Cost With the model space $\text{MODEL} = (D, \mathbf{p}', \mathbf{p})$, the sender S will send the MODEL to the receiver R in three parts: (i) first send $\mathbf{p}$, (ii) next send $\mathbf{p}'$ given $\mathbf{p}$, and then (iii) send D given $\mathbf{p}'$ and $\mathbf{p}$. Therefore, the model cost $L(D, \mathbf{p}', \mathbf{p})$ will also have three components

$$L(D, \mathbf{p}', \mathbf{p}) = \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p}'|\mathbf{p}) + \text{COST}(D|\mathbf{p}', \mathbf{p})$$

Here, we will send the first component, $\mathbf{p}$, directly, send the second component, $\mathbf{p}'$ given $\mathbf{p}$, via sending $\mathbf{p}' - \mathbf{p}$, and send the third component, D given $\mathbf{p}'$ and $\mathbf{p}$, via sending $\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\mathbf{p})$. We further write the model cost as below:

$$L(D, \mathbf{p}', \mathbf{p}) = \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p}' - \mathbf{p}|\mathbf{p}) + \text{COST}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p}', \mathbf{p})$$

C.4 Data Cost Give the $\text{MODEL} = (D, \mathbf{p}', \mathbf{p})$ and model cost above, next we will send the DATA in terms of the MODEL. Here, the DATA is D_{reported} , and the data cost will have only one component:

$$L(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p}) = \text{COST}(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p})$$

Here, we will send it via $\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p}')$, and we further write the data cost as below:

$$L(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p}) = \text{COST}\left(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p}')|D, \mathbf{p}', \mathbf{p}\right)$$

C.5 Total Cost The total cost is the sum of model cost $L(D, \mathbf{p}', \mathbf{p})$ and data cost $L(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p})$:

$$\begin{aligned} L(D_{\text{reported}}, D, \mathbf{p}', \mathbf{p}) &= L(D, \mathbf{p}', \mathbf{p}) + L(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p}) \\ &= \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p}'|\mathbf{p}) \\ &\quad + \text{COST}(D|\mathbf{p}', \mathbf{p}) + \text{COST}(D_{\text{reported}}|D, \mathbf{p}', \mathbf{p}) \\ &= \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p}' - \mathbf{p}|\mathbf{p}) \\ &\quad + \text{COST}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p}', \mathbf{p}) \\ &\quad + \text{COST}\left(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p}')|D, \mathbf{p}', \mathbf{p}\right) \end{aligned}$$

C.6 Cost Derivation Next, we derive the cost for each component and give our encoding method explicitly:

1. $\text{COST}(\mathbf{p})$: We represent $\mathbf{p}$ as a vector of real numbers (we describe our encoding later below).
2. $\text{COST}(\mathbf{p}' - \mathbf{p}|\mathbf{p})$: We will encode the difference of two vectors as a vector of real numbers.
3. $\text{COST}(\alpha'_{\text{reported}} \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p}', \mathbf{p})$: Here, we encode the difference between the two time sequences: $\alpha'_{\text{reported}} \times D$ given $D_{\text{reported}}(\mathbf{p})$.
4. $\text{COST}\left(\frac{D - D_{\text{reported}}}{1 - \alpha'_{\text{reported}}} - D(\mathbf{p}')|D, \mathbf{p}', \mathbf{p}\right)$: Again, we encode it as a difference between the two time sequences: $\frac{D_{\text{unreported}}}{1 - \alpha'_{\text{reported}}}$ given $D(\mathbf{p}')$.

Next, we describe the encoding cost of real numbers, vectors, and the difference between two time sequences.

C.6.1 Encoding Integers To encode a positive integer n , we encode both the binary representation of integer n as well as the length of the representation $\log_2 n$. Following [23], we use the cost in bits of encoding a single integer n is as follows:

$$\text{COST}(n) = \log_2 c_0 + \log^*(n).$$

where $c_0 \approx 2.865$ and $\log^*(n) = \log_2 n + \log_2 \log_2 n + \dots$ as described in [23]. There are infinite terms in $\log^*(n)$

function since after we encoded a number, we always need to encode its length as another number, which could be repeated for infinite times. Additionally, if we want to transmit an integer that can be either positive or negative, we can add another sign bit and therefore the cost in bits for integers will be

$$\text{COST}(n) = \text{COST}(|n|) + 1.$$

C.6.2 Encoding Real Numbers Note that most real numbers (e.g. π or e) need infinite number of bits to encode. Hence, we introduce a precision threshold δ . With threshold δ , we approximate a real number x with x_δ which satisfies $|x - x_\delta| < \delta$, and we encode x_δ instead. To encode x_δ , we encode both the integer part $\lfloor x \rfloor$ as well as the fractional part $x_\delta - \lfloor x \rfloor$. Hence the cost in bits of encoding a real number x is as follows:

$$\text{COST}(x) = \text{COST}(\lfloor x \rfloor) + \log_2 \frac{1}{\delta}$$

where $\lfloor x \rfloor$ is the floor of x and therefore is a integer, whose encoding cost is $\text{COST}(\lfloor x \rfloor) = \log_2 c_0 + \log^*(\lfloor x \rfloor)$. Additionally, if we want to transmit a real number that can be either positive or negative, we can add another sign bit and therefore the cost in bits for real numbers will be

$$\text{COST}(x) = \text{COST}(|x|) + 1$$

C.6.3 Encoding Vectors To encode a vector $\mathbf{p} = [\mathbf{p}[1], \mathbf{p}[2], \dots, \mathbf{p}[n]]$, we encode every components one by one as real numbers. Hence the cost in bits of encoding a vector $\mathbf{p}$ is as follows:

$$\text{COST}(\mathbf{p}) = \text{COST}(\mathbf{p}[1]) + \text{COST}(\mathbf{p}[2]) + \dots + \text{COST}(\mathbf{p}[n])$$

C.6.4 Encoding The Difference between Two Time Sequences To encode the difference $A - B = [A_{t_1} - B_{t_1}, A_{t_2} - B_{t_2}, \dots, A_{t_n} - B_{t_n}]$ between two time sequence $A = [A_{t_1}, A_{t_2}, \dots, A_{t_n}]$ and $B = [B_{t_1}, B_{t_2}, \dots, B_{t_n}]$, we encode every components one by one as real numbers. Hence the cost in bits of encoding the difference is as follows:

$$\begin{aligned} \text{COST}(A - B) &= \text{COST}(A_{t_1} - B_{t_1}) + \text{COST}(A_{t_2} - B_{t_2}) \\ &\quad + \dots + \text{COST}(A_{t_n} - B_{t_n}) \end{aligned}$$

C.7 Problem Statement Now we have derived every cost involved in our problem, and we can finally state our problem as one of estimating the total infections D as follows: Given the time sequence D_{reported} and epidemiological model O_M , find D^* that minimizes the MDL total cost:

$$D^* = \arg \min_D L(D_{\text{reported}}, D, \mathbf{p}', \mathbf{p})$$

We will give the algorithm to find such D^* as follows:

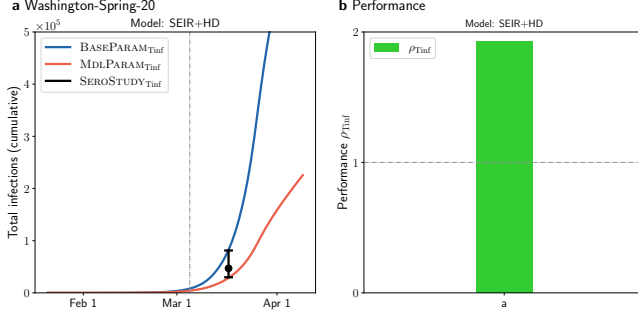


Figure S3: MDLPARAM (red) gives a closer estimation of total infections to serological studies (black) than BASEPARAM (blue). Note that the serological studies are not informed for both approaches. **a** The red and blue curves represent MDLINFER’s estimation of total infections, MDLPARAM_{Tinf}, and baseline calibration procedure’s estimation of total infections, BASEPARAM_{Tinf}, respectively. The black point estimates and confidence intervals represent the total infections estimated by serological studies [1, 16], or SEROSTUDY_{Tinf}. **b** The performance metric, ρ_{Tinf} , comparing MDLPARAM_{Tinf} against BASEPARAM_{Tinf} is shown for **a** for the SEIR+HD model. Here, the values of ρ_{Tinf} is 1.93.

C.8 Algorithms Before presenting our algorithm to find D^* , we will first address the problem of searching D^* directly. Note that D^* is a time sequence of total infections instead of a scalar, naively searching D^* directly in large search space is intractable. Hence, we propose an alternate method: First, we can find quickly a good reported rate $\alpha_{reported}^*$ since we can constrain $D = \frac{D_{reported}}{\alpha_{reported}}$ to reduce the search space. Then we can search for the optimal D^* with $\alpha_{reported}^*$ from step 1 as constraints. Here, we write down our two-step search algorithm to find the D^* as follows:

1. Step 1: We do a linear search to find a good reported rate $\alpha_{reported}^*$, which serves as an initialization in the second step.
2. Step 2: Given the $\alpha_{reported}^*$ found in step 1, we use the Nelder-Mead [13] optimization to find the D^* that minimizes $L(D_{reported}, D, \mathbf{p}', \mathbf{p})$ with $\alpha_{reported}^*$ constraints.

C.8.1 Step 1: Find the $\alpha_{reported}^*$ In step 1, we search on $\alpha_{reported}$ to find the $\alpha_{reported}^*$ as follows:

$$\alpha_{reported}^* = \arg \min_{\alpha_{reported}} L(D_{reported}, \frac{D_{reported}}{\alpha_{reported}}, \mathbf{p}', \mathbf{p})$$

To be more specific, in the first step of our algorithm, we do a linear search on different $\alpha_{reported} = [0.01, 0.02, 0.03, \dots, 0.99]$ and calibrate the O_M on $D = \frac{D_{reported}}{\alpha_{reported}}$, which means

$$\mathbf{p}' = \text{CALIBRATE}(O_M, \{D_{reported}, \frac{D_{reported}}{\alpha_{reported}} - D_{reported}, \text{others}\})$$

Then we pick the $\alpha_{reported}^*$ that corresponds to the lowest total cost $L(D_{reported}, D, \mathbf{p}', \mathbf{p})$ as the $\alpha_{reported}^*$.

C.8.2 Step 2: Find the D^* given $\alpha_{reported}^*$ With $\alpha_{reported}^*$ inferred in step 1, we will next find the D^* that minimizes the total cost.

$$D^* = \arg \min_D L(D_{reported}, D, \mathbf{p}', \mathbf{p})$$

Since we have already found $\alpha_{reported}^*$ in step 1, we will only search the D^* that satisfies

$$\sum D^* = \frac{\sum D_{reported}}{\alpha_{reported}^*}$$

To search for the optimal D^* , we leverage the popular Nelder-Mead search algorithm [13].

D Metrics

To evaluate the performance of MDLINFER, as mentioned in the main article, we use the Root Mean Squared Error (RMSE) following previous works [31, 32, 12]. Specifically, the metrics are calculated using the following equations following [33].

$$RMSE = \sqrt{\sum_t (\hat{y}_t - y_t)^2}$$

where $\hat{y}_t$ is the estimated number by either BASEINFER or MDLINFER, y_t is the ground-truth value.

E Experimental Setup

Here we describe our experimental setup in more detail and present results on additional testbeds. We also list the notations used in the experiments section in Table S1.

E.1 Total Infections The Results section in the main paper refers to BASEPARAM_{Tinf}, which represents the cumulative total infections derived from the baseline calibration procedure. It is computed as follows:

$$\text{BASEPARAM}_{Tinf} = \sum D(\mathbf{p})$$

Similarly, MDLPARAM_{Tinf}, which represents the cumulative total infections derived from MDLINFER, is computed as follows:

$$\text{MDLPARAM}_{Tinf} = \sum D(\mathbf{p}')$$

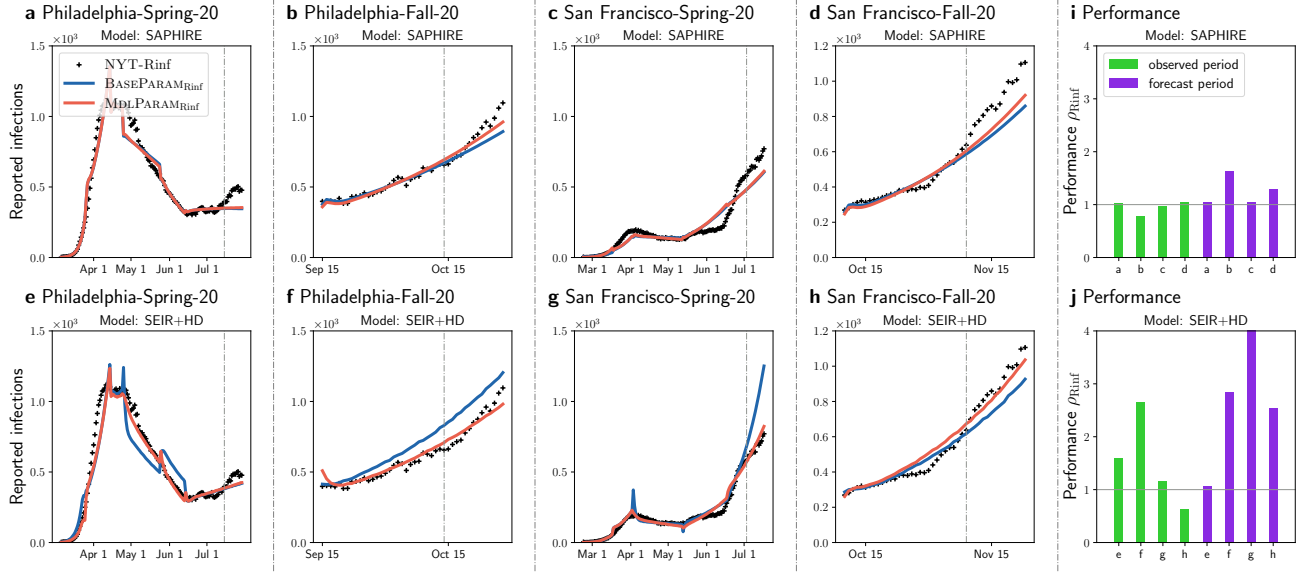


Figure S4: MDLPARAM (red) gives a closer estimation of reported infections (black) than BASEPARAM (blue) on various geographical regions and time periods. We use the reported infections in the observed period as inputs and try to forecast the future reported infections (forecast period). **a-h** The vertical grey dash line divides the observed period and forecast period. The red and blue curves represent MDLINFER's estimation of reported infections, MDLPARAM_{Rinf}, and baseline calibration procedure's estimation of reported infections, BASEPARAM_{Rinf}, respectively. The black plus symbols represent the infections reported by the New York Times (NYT-Rinf). **a-d** is for SAPHIRE model and **e-h** is for SEIR+HD model. **i-j** The performance metric, ρ_{Rinf} , comparing MDLPARAM_{Rinf} against BASEPARAM_{Rinf} is shown for **a-h** for the regions for both the SAPHIRE model in **i**, and the SEIR+HD model in **j**.

In Figure S3, we show additional results comparing the performance of MDLINFER and baseline calibration procedure in estimating total infections. Here, MDLINFER (red) gives a closer estimation of total infections to serological studies (black) than baseline calibration procedure (blue).

E.2 Reported Infections In Figure S4, we present additional results comparing the performance of MDLINFER and baseline calibration procedure in forecasting future infections (forecast period). Here, MDLINFER (red) gives a closer estimation of reported infections (black) than baseline calibration procedure (blue) on various geographical regions and time periods.

E.3 Symptomatic Rate The baseline calibration procedure and MDLINFER also estimate the symptomatic rate BASEPARAM_{Symp} and MDLPARAM_{Symp} respectively. We compare these against the Facebook symptomatic surveillance data RATE_{Symp}.

We calculate BASEPARAM_{Symp} from **p** as follows:

$$\text{BASEPARAM}_{\text{Symp}} = \frac{I_S(\mathbf{p}) + I_M(\mathbf{p})}{N}$$

where $I_S(\mathbf{p})$ is the number of infections in severe symptomatic state, $I_M(\mathbf{p})$ represents the same in mild symptomatic state, and N is the total population in this area.

Similarly MDLPARAM_{Symp} is computed as follows:

$$\text{MDLPARAM}_{\text{Symp}} = \frac{I_S(\mathbf{p}') + I_M(\mathbf{p}')}{N}$$

In Figure S5, we present additional results comparing MDLINFER and baseline calibration procedure in estimating trends of symptomatic rate. Here, MDLINFER (red) gives a closer estimation of symptomatic rate (black) than baseline calibration procedure (blue).

E.4 Cumulative Reported Rate We also calculate the a dynamic reported rate from both baseline calibration procedure and MDLINFER. Note that this cumulative reported rate is different from α_{reported} and $\alpha'_{\text{reported}}$, which are two scalars used in MDL formulation. We calculate BASEPARAM_{Rate} from BASEPARAM

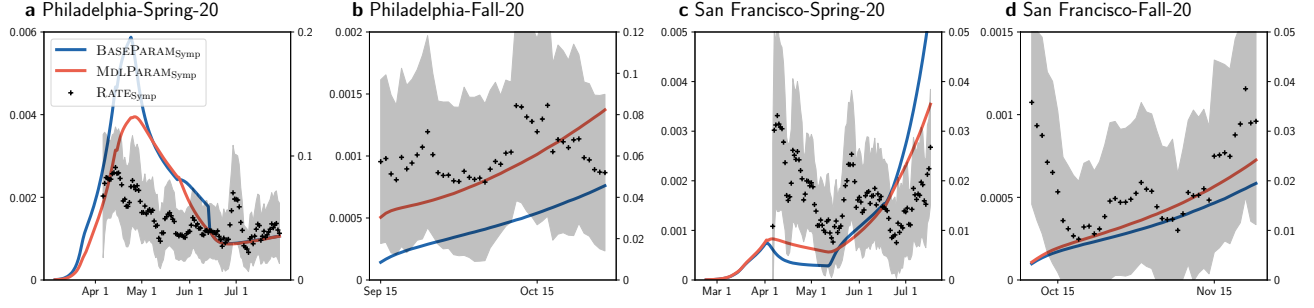


Figure S5: MDLPARAM (red) gives a closer estimation of symptomatic rate (black) than BASEPARAM (blue). The red and blue curves represent MDLINFER's estimation of symptomatic rate, $\text{MDLPARAM}_{\text{Symp}}$, and baseline calibration procedure's estimation of symptomatic rate, $\text{BASEPARAM}_{\text{Symp}}$, respectively. The black points and the shaded regions are the point estimate with standard error for $\text{RATE}_{\text{Symp}}$ (the COVID-related symptomatic rates derived from the symptomatic surveillance dataset [30, 35]).

$\mathbf{p}$ as follows:

$$\text{BASEPARAM}_{\text{Rate}} = \frac{\sum \text{NYT-Rinf}}{\sum D(\mathbf{p})}$$

Similarly we calculate $\text{MDLPARAM}_{\text{Rate}}$ from MDLPARAM $\mathbf{p}'$ as follows:

$$\text{MDLPARAM}_{\text{Rate}} = \frac{\sum \text{NYT-Rinf}}{\sum D(\mathbf{p}')}$$

E.5 Non-pharmaceutical Interventions Simulation We also use the baseline calibration procedure and MDLINFER to perform non-pharmaceutical interventions simulation on SEIR+HD model. Here, both the baseline calibration procedure and MDLINFER are estimated on the observed period, then on the future period, we will consider the following five scenarios of isolation:

1. Isolate reported infections: We isolate the α_1 fraction of severe symptomatic infections I_S and mild symptomatic infections I_M .
2. Isolate both reported infections and symptomatic infections: Note that some reported infections are included in the symptomatic infections. Here, we isolate all severe symptomatic infections I_S and mild symptomatic infections I_M .
3. Isolate 25% presymptomatic and asymptomatic infections: We isolate 25% of presymptomatic infections I_P , asymptomatic infections I_A , and all severe symptomatic infections I_S and mild symptomatic infections I_M .
4. Isolate 50% presymptomatic and asymptomatic infections: We isolate 50% of presymptomatic infections I_P , asymptomatic infections I_A , and all severe

symptomatic infections I_S and mild symptomatic infections I_M .

5. Isolate 75% presymptomatic and asymptomatic infections: We isolate 75% of presymptomatic infections I_P , asymptomatic infections I_A , and all severe symptomatic infections I_S and mild symptomatic infections I_M .

The infectiousness of the noes in isolated is reduces by 50%.

F Sensitive Analysis

We also perform sensitivity experiments to inspect the robustness of our non-pharmaceutical interventions simulations for Minneapolis-Spring-20 in Figure S6. Here, we reduce the infectiousness of the isolated infections to 3 different values: 0.4, 0.5, and 0.6, and repeat simulations in each of the scenarios.

Our results consistently show that only isolating reported or symptomatic infections is not be enough to reduce the future reported infections. However, isolating both symptomatic infections and some fraction of asymptomatic and presymptomataic infections leads to reduction in reported infections in most settings.

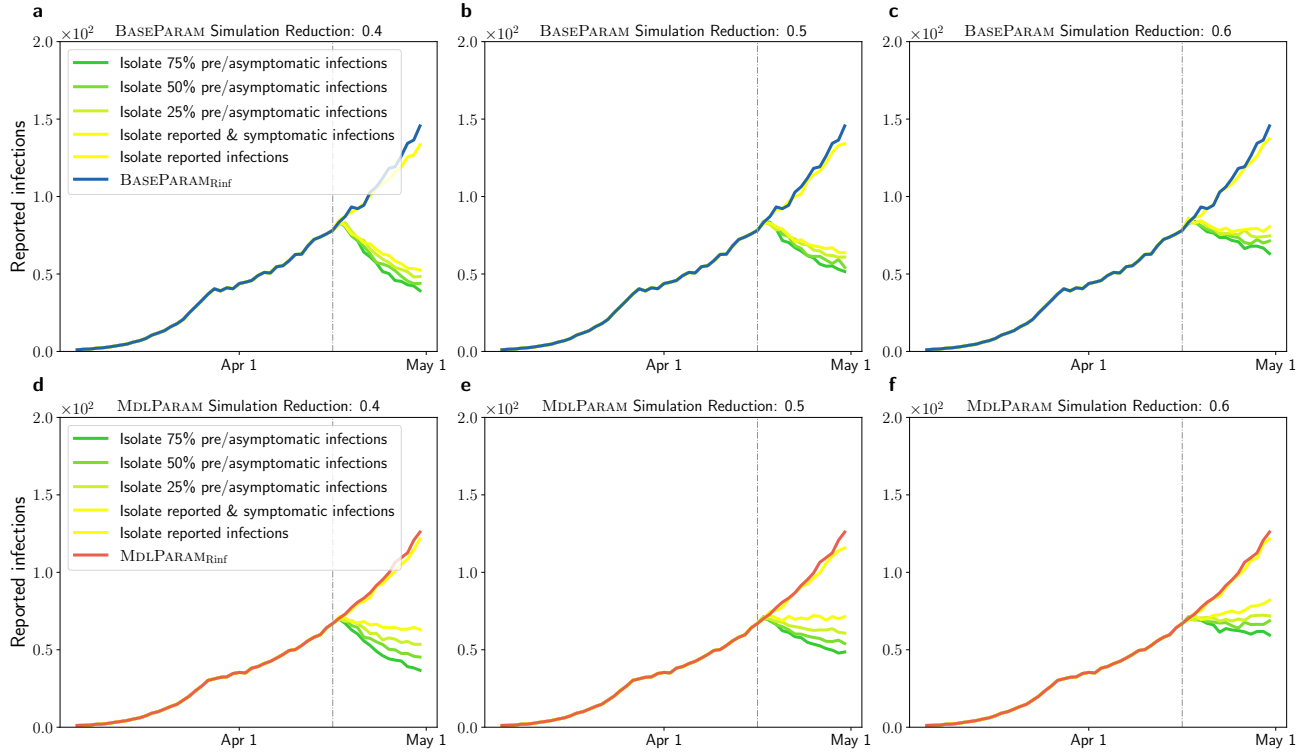


Figure S6: Our non-pharmaceutical interventions simulation results are robust. **a-c**, the vertical grey dash line divides the observed period and future period. The blue curve represents the BASEPARAM's estimation of reported infections. The other five curves represent the simulated reported infections for 5 scenarios: (i) Isolate the reported infections, (ii) symptomatic infections, symptomatic infections and (iii) 25%, (iv) 50%, (v) 75% asymptomatic and presymptomatic infections, where we reduce the infectiousness of these isolated infections to 40% in **a**, 50% in **b**, and 60% in **c** in future period. **d-f**, the vertical grey dash line divides the observed period and future period. The red curve represents the MDLPARAM's estimation of reported infections. The other five curves represent the simulated reported infections for the same 5 scenarios as in **a** to **c**. The results are for Minneapolis-Spring-20 that we shown in main article Figure 5.

Table S1: List of notations

Symbol	Description
MDLINFER	Our Minimum Description Length (MDL) framework to estimate total infections
O_M	Epidemiological models used in MDLINFER
BASEPARAM	Baseline parameterization obtained via baseline calibration procedure
MDLPARAM	Optimal parameterization identified by MDLINFER
SEROSTUDY _{Tinf}	Total infections estimated by serological studies
BASEPARAM _{Tinf}	Total infections estimated by baseline calibration procedure
MDLPARAM _{Tinf}	Total infections estimated by MDLINFER
ρ_{Tinf}	The performance metric comparing MDLINFER against baseline calibration procedure in estimating total infections
NYT-Rinf	New York Times reported infections
BASEPARAM _{Rinf}	Reported infections estimated by baseline calibration procedure
MDLPARAM _{Rinf}	Reported infections estimated by MDLINFER
ρ_{Rinf}	The performance metric comparing MDLINFER against baseline calibration procedure in estimating reported infections
RATE _{Symp}	COVID-related symptomatic rate from symptomatic surveillance data
BASEPARAM _{Symp}	Symptomatic rate estimated by baseline calibration procedure
MDLPARAM _{Symp}	Symptomatic rate estimated by MDLINFER
SEROSTUDY _{Rate}	Cumulative reported rate estimated by serological studies
BASEPARAM _{Rate}	Cumulative reported rate estimated by baseline calibration procedure
MDLPARAM _{Rate}	Cumulative reported rate estimated by MDLINFER