Information Theoretic Model Selection for Accurately Estimating Unreported COVID-19 Infections

Jiaming Cui[a], Arash Haddadan[b], A S M Ahsan-Ul Haque[c], Bijaya Adhikari[d], Anil Vullikanti[b,c], and B. Aditya Prakash[a,1]

[a]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
[b]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22904
[c]Department of Computer Science, University of Virginia, Charlottesville, VA 22904
[d]Department of Computer Science, The University of Iowa, Iowa City, IA 52242

Supplementary Information

[1]To whom correspondence should be addressed. E-mail: badityap@cc.gatech.edu

Supplementary Methods

In this section, we first describe the data used in the paper. These include the ones used to infer model parameters, BaseParam and MdlParam, and the ones used for comparison and validation. We then describe the epidemiological model used in the paper in detail. Then, we move on to our Minimum Description Length optimization formulation and the two-step algorithm, both of which we briefly described in the main paper. Finally, we present the results which were omitted from the main paper.

Data

New York Times reported infections dataset

The New York Times reported infections, NYT-Rinf, dataset [1] 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). 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.

Serological studies

The serological studies [3, 6] consists of the point estimate and 95% confidence interval of the prevalence of antibodies to SARS-CoV-2. Using the prevalence of the antibodies and the population, we can compute the estimated total infections and 95% confidence interval in the location. For Santa Clara, CA, the estimated prevalence of antibodies is 2.8% (95% confidence interval: 1.3%-4.7%). Therefore, the estimated total infections is 54000 (95% confidence interval: 25000-91000) [3]. For Bucks, PA, and Western Washington the estimated prevalence of antibodies are 3.2% and 1.1% respectively (95% confidence interval: 1.7%-5.2% and 0.7%-1.9% respectively). The population of Bucks county is 628270 and the population of Western Washington is 4273500. These imply that the total infections for Bucks, PA and Western Washington are 20100 (95% confidence interval: 10680-32670) and 47000 (95% confidence interval: 29900-81200) [6] respectively.

Symptomatic surveillance data

The symptomatic surveillance data comes from Facebook's symptomatic survey [2]. The survey started on April 6, 2020. 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 defined as fever along with either cough, shortness of breath, or difficulty breathing [2]. The data also includes weighted version which accounts for the differences between Facebook users and the United States population. In some experiments, we contrast the symptomatic rate inferred by our approach against the weighted data from the survey.

Epidemiological Model

Base Epidemiological Model

The epidemiological model described in [4] serves as the base epidemiological model O_{B} in our experiments. The compartmental diagram of O_{B} is shown in Fig. S5. As seen in the figure, it

consists of the following 10 states.

1. S: Susceptible
2. E: Exposed
3. I_P: Pre-symptomatic
4. I_S: Symptomatic, severe
5. I_M: Symptomatic, mild
6. I_A: Asymptomatic
7. H_D: Hospitalized, eventual death
8. H_R: Hospitalized, eventual recover
9. R: Recovered
10. D: Dead

O_{B}, as described in [4], has of 21 different parameters, which are listed below. Note that only three parameters are calibrated, while the rest are fixed.

1. C_A: Relative infectiousness of asymptomatic (fixed)
2. C_P: Relative infectiousness of presymptomatic (fixed)
3. C_M: Relative infectiousness of mild symptomatic (fixed)
4. C_S: Relative infectiousness of severe symptomatic (fixed)
5. γ: Preinfectious period (fixed)
6. λ_P: Presymptomatic duration (fixed)
7. λ_A: Infectious period for asymptomatic infections (fixed)
8. λ_S: Time from symptom onset to hospitalizations (severe) (fixed)
9. λ_M: Time from symptom onset to recovery (mild) (fixed)
10. ρ_R: Time from hospitalization to recovery (fixed)
11. ρ_D: Time from hospitalization to death (fixed)
12. N: Population (fixed)
13. Start date: Start date of the epidemic (fixed)
14. Work From Home start date: Work from home start date (fixed)
15. σ_{WFH}: Work from home proportion of contacts remaining (fixed)
16. E_0: Number of initial infections that began the epidemic (calibrated)

17. C_A: Relative infectiousness of asymptomatic infections (fixed)
18. α: Proportion of infections that are asymptomatic (fixed)
19. $1-\mu$: Proportion of symptomatic infections that require hospitalization (fixed)
20. β_0: Transmission rate in the absence of interventions (calibrated)
21. σ: The proportional reduction on β_0 under shelter-in-place (calibrated)

Next we describe the calibration process described in [4] in detail. Before doing so, we define $I_{\text{new sympt}}$ as the time sequence of daily new symptomatic infections, D_{new} as the day level time sequence of mortality, and H_{new} as the time sequence of newly hospitalized infections everyday. These can be calculated from the states of O_{B} as follows:

1. $I_{\text{new sympt}} = dI_P I_S + dI_P I_M$.
2. $D_{\text{new}} = dH_D D$.
3. $H_{\text{new}} = dI_S H_R + dI_S H_D$.

The calibration process in [4] minimizes the LogLikelihood between the observed $D_{\text{mortality}}$ and D_{new} as predicted by O_{B}.

$$\text{LogLikelihood} = dpois(D_{\text{mortality}}, D_{\text{new}}) \tag{1}$$

where

$$dpois(a, b) = \log(P(X = a | X \sim Poisson(\lambda = b))) \tag{2}$$

The process infers a set of parameters including the transmission rate β_0 (the transmission rate in the absence of interventions), σ (the proportional reduction on β_0 under shelter-in-place), and E_0 (number of initial infections that began the epidemic). The optimization problem that the calibration vies to solve can be written as follows:

$$[\beta_0, \sigma, E_0] = \arg\min_{\beta_0, \sigma, E_0} \{dpois(D_{\text{mortality}}, D_{\text{new}})\} \tag{3}$$

Extended Epidemiological Model

Note that the calibration process defined above depends only on $D_{\text{mortality}}$. In this work, we expect the epidemiological model to calibrate on reported infections D_{reported} and candidate unreported infections $D_{\text{unreported}}$. Hence we extend O_{B} to a slightly different epidemiological model O_{M}. However, to ensure that the epidemiological model structure of O_{M} is similar to that of O_{B}, we just add two additional term on top of parameters and states defined by O_{B}, without removing anything. These new terms include: new reported infections everyday $D_{\text{new reported}}$, and new unreported infections everyday $D_{\text{new unreported}}$. These are defined as follows, which is similar to [7]:

1. $D_{\text{new reported}} = \alpha_1 \times (dI_P I_S + dI_P I_M)$:

   $I_{\text{new sympt}} = dI_P I_S + dI_P I_M$ represents the number of new symptomatic infections everyday in O_{M}. Here, we assume α_1 proportion of new symptomatic infections everyday will be that day's new reported infections.

2. New unreported infections $D_{\text{new unreported}} = (1-\alpha_1) \times (dI_P I_S + dI_P I_M) + dEI_A$:

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

Note that the parameter α_1 above is different than the parameter α, which represents the proportion of infections that are asymptomatic. In our calibration process, we make both α and α_1 (proportion of new symptomatic infections that are reported) learnable. Therefore, the extended epidemiological model O_{M} now calibrates by minimizing the LogLikelihood between D_{new} to $D_{\text{mortality}}$, $D_{\text{new reported}}$ to D_{reported}, and $D_{\text{new unreported}}$ to $D_{\text{unreported}}$

$$\begin{aligned} \textsc{LogLikelihood} = & \, w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) \\ & + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \\ & + w_{\text{unreported}} \times dpois(D_{\text{unreported}}, D_{\text{new unreported}}) \end{aligned} \tag{4}$$

The calibration process infers values of a set of parameters including β_0, σ, E_0, α, and α_1. The calibration process can be written as follows:

$$\begin{aligned} [\beta_0, \sigma, E_0, \alpha, \alpha_1] = \arg \min_{\beta_0, \sigma, E_0, \alpha, \alpha_1} \{ & w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) \\ & + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \\ & + w_{\text{unreported}} \times dpois(D_{\text{unreported}}, D_{\text{new unreported}}) \} \end{aligned} \tag{5}$$

With the calibration of the epidemiological model O_{M} introduced, next we will define baseline parameterization BaseParam and MdlInfer parameterization MdlParam.

Baseline Parameterization (BaseParam)

By calibrating O_{M} on D_{reported} and $D_{\text{mortality}}$, we get the baseline paramterization $\mathbf{p}$:

$$\mathbf{p} = \textsc{Calibrate}(O_{\text{M}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{6}$$

by minimizing

$$\textsc{LogLikelihood} = w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \tag{7}$$

We write this procedure as follows:

$$\begin{aligned} \mathbf{p} = [\mathbf{p}[\beta_0], \mathbf{p}[\sigma], \mathbf{p}[E_0], \mathbf{p}[\alpha], \mathbf{p}[\alpha_1]] = \arg \min_{\beta_0, \sigma, E_0, \alpha, \alpha_1} \{ & w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) \\ & + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \} \end{aligned} \tag{8}$$

From $\mathbf{p}$, we can generate the O_{M} reported infections $D_{\text{reported}}(\mathbf{p})$, unreported infections $D_{\text{unreported}}(\mathbf{p})$, and total infections $D(\mathbf{p}) = D_{\text{reported}}(\mathbf{p}) + D_{\text{unreported}}(\mathbf{p})$. We can also calculate the reported rate $\mathbf{p}[\alpha_{\text{reported}}]$ as follows:

$$\mathbf{p}[\alpha_{\text{reported}}] = \frac{\sum_t D_{\text{reported}}(\mathbf{p})}{\sum_t D(\mathbf{p})} = \frac{\mathbf{p}[\alpha_1] \times \sum_t (dI_P I_S + dI_P I_M)}{\sum_t (dI_P I_S + dI_P I_M + dEI_A)} \tag{9}$$

MdlInfer Parameterization (MdlParam)

By calibrating O_{M} on $D_{\text{unreported}}$, D_{reported}, and $D_{\text{mortality}}$, we get the MdlInfer paramterization **p'**:

$$\mathbf{p'} = \textsc{Calibrate}(O_{\mathrm{M}}, D_{\text{unreported}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{10}$$

by minimizing

$$\begin{aligned} \textsc{LogLikelihood} = & \, w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) \\ & + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \\ & + w_{\text{unreported}} \times dpois(D_{\text{unreported}}, D_{\text{new unreported}}) \end{aligned} \tag{11}$$

We write this procedure as follows:

$$\begin{aligned} \mathbf{p'} = [\mathbf{p'}[\beta_0], \mathbf{p'}[\sigma], \mathbf{p'}[E_0], \mathbf{p'}[\alpha], \mathbf{p'}[\alpha_1]] = \arg \min_{\beta_0, \sigma, E_0, \alpha, \alpha_1} \{ & w_{\text{mortality}} \times dpois(D_{\text{mortality}}, D_{\text{new}}) \\ & + w_{\text{reported}} \times dpois(D_{\text{reported}}, D_{\text{new reported}}) \\ & + w_{\text{unreported}} \times dpois(D_{\text{unreported}}, D_{\text{new unreported}}) \} \end{aligned} \tag{12}$$

Similarly, from **p'**, we can similarly generate the O_{M} reported infections $D_{\text{reported}}(\mathbf{p'})$, unreported infections $D_{\text{unreported}}(\mathbf{p'})$, and total infections $D(\mathbf{p'}) = D_{\text{reported}}(\mathbf{p'}) + D_{\text{unreported}}(\mathbf{p'})$ by just simulating the O_{M}. We can also calculate the reported rate $\mathbf{p'}[\alpha_{\text{reported}}]$ as follows:

$$\mathbf{p'}[\alpha_{\text{reported}}] = \frac{\sum_t D_{\text{reported}}(\mathbf{p'})}{\sum_t D(\mathbf{p'})} = \frac{\mathbf{p'}[\alpha_1] \times \sum_t (dI_P I_S + dI_P I_M)}{\sum_t (dI_P I_S + dI_P I_M + dE I_A)} \tag{13}$$

With the calibration process, **p**, and **p'** defined, we can next formalize the MDL cost.

Methodology

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:

1. Model Cost L(Model): The cost of encoding the Model.
2. Data Cost L(Data|Model): The cost of encoding Data given the Model.

Model Space

In this work, the Data is D_{reported}. Therefore, one of the most natural Model would have been Model = (**p**), as one can directly compute $D_{\text{reported}}(\mathbf{p})$ and use it to encode D_{reported}. This can be written as follows:

$$L(\textsc{Model}) = \textsc{Cost}(\mathbf{p}) \tag{14}$$

and

$$L(\textsc{Model}) = \textsc{Cost}(D_{\text{reported}} | \mathbf{p}) \tag{15}$$

However, this model space suffers from the frangibility that slightly different values of $\mathbf{p}$ could lead to vastly different costs. To account for this, we define the MODEL as MODEL $= (D, \mathbf{p'}, \mathbf{p})$, which consists of three components. The intuition here is that we use D to reparameterize the $\mathbf{p'}$, use $\mathbf{p}$ to send the $\mathbf{p'}$, and finally use $\mathbf{p'}[\alpha_{\text{reported}}] \times D$ to encode D_{reported}.

Model Cost

With the model space MODEL $= (D, \mathbf{p'}, \mathbf{p})$ above, the sender S will send the MODEL to the receiver R in three parts:

1. First send $\mathbf{p}$.
2. Next send $\mathbf{p'}$ given $\mathbf{p}$.
3. Then send D given $\mathbf{p'}$ and $\mathbf{p}$.

Therefore, the MDL Model Cost $L(D, \mathbf{p'}, \mathbf{p})$ will also have three components

$$L(D, \mathbf{p'}, \mathbf{p}) = \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p'}|\mathbf{p}) + \text{COST}(D|\mathbf{p'}, \mathbf{p}) \tag{16}$$

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 $\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})$. We further write the MDL Model Cost in Eq. 16 as below:

$$L(D, \mathbf{p'}, \mathbf{p}) = \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p'} - \mathbf{p}|\mathbf{p}) + \text{COST}(\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p'}, \mathbf{p}) \tag{17}$$

Data Cost

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

$$L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) = \text{COST}(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) \tag{18}$$

Here, we will send it via $\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})$, and we further write the MDL Data Cost in Eq. 18 as below:

$$L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) = \text{COST}(\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})|D, \mathbf{p'}, \mathbf{p}) \tag{19}$$

Total MDL Cost

The Total MDL Cost is the sum of MDL Model Cost $L(D, \mathbf{p'}, \mathbf{p})$ and MDL Data Cost $L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p})$:

$$\begin{aligned} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) &= L(D, \mathbf{p'}, \mathbf{p}) + L(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) \\ &= \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p'}|\mathbf{p}) + \text{COST}(D|\mathbf{p'}, \mathbf{p}) + \text{COST}(D_{\text{reported}}|D, \mathbf{p'}, \mathbf{p}) \\ &= \text{COST}(\mathbf{p}) + \text{COST}(\mathbf{p'} - \mathbf{p}|\mathbf{p}) + \text{COST}(\mathbf{p'}[\alpha_{\text{reported}}] \times D - D_{\text{reported}}(\mathbf{p})|\mathbf{p'}, \mathbf{p}) \\ &+ \text{COST}(\frac{D - D_{\text{reported}}}{1 - \mathbf{p'}[\alpha_{\text{reported}}]} - D(\mathbf{p'})|D, \mathbf{p'}, \mathbf{p}) \end{aligned} \tag{20}$$

Cost Derivation

Next, we derive the cost for each component and give our encoding method explicitly:

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

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

Encoding Integers

To encode a positive integer n, we need to encode both the binary representation of integer n as well as the length of the representation $\log_2 n$ [8]. Hence the cost of encoding a single integer n is as follows:

$$\textsc{Cost}(n) = \log_2 c_0 + \log^*(n). \tag{21}$$

where $c_0 \approx 2.865$ and $\log^*(n)$ is

$$\log^*(n) = \log_2 n + \log_2 \log_2 n + ... \tag{22}$$

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 (encoding length in bits) for integers will be

$$\textsc{Cost}(n) = \textsc{Cost}(|n|) + 1. \tag{23}$$

Encoding Real Numbers

Note that most real numbers (e.g., π or e) need infinite number of bits to encode. Hence, we need to 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 need to encode both the integer part $\lfloor x \rfloor$ as well as the fractional part $x_\delta - \lfloor x \rfloor$. Hence the cost of encoding a real number x is as follows:

$$\textsc{Cost}(x) = \textsc{Cost}(\lfloor x \rfloor) + \log_2 \frac{1}{\delta} \tag{24}$$

where $\lfloor x \rfloor$ is the floor of x and therefore is a integer, whose encoding cost is

$$\textsc{Cost}(\lfloor x \rfloor) = \log_2 c_0 + \log^*(\lfloor x \rfloor) \tag{25}$$

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 (encoding length in bits) for real numbers will be

$$\textsc{Cost}(x) = \textsc{Cost}(|x|) + 1 \tag{26}$$

Encoding Vectors

To encode a vector $\mathbf{p}$, we need to encode every components one by one as real numbers. Hence the cost of encoding a vector $\mathbf{p}$ is as follows:

$$\textsc{Cost}(\mathbf{p}) = \textsc{Cost}(\mathbf{p}[\alpha]) + \textsc{Cost}(\mathbf{p}[\beta]) + ... \tag{27}$$

Encoding The Difference between Two Time Sequences

To encode the difference $A - B = [A_{t_1} - B_{t_1}, A_{t_2} - B_{t_2}, ..., A_{t_n} - B_{t_n}]$ between two time sequence $A = [A_{t_1}, A_{t_2}, ..., A_{t_n}]$ and $B = [B_{t_1}, B_{t_2}, ..., B_{t_n}]$, we need to encode every components one by one as real numbers. Hence the cost of encoding the difference is as follows:

$$\textsc{Cost}(A - B) = \textsc{Cost}(A_{t_1} - B_{t_1}) + \textsc{Cost}(A_{t_2} - B_{t_2}) + ... \tag{28}$$

Problem Statement

Now we have derived every cost involved in our problem, and we can finally state our problem as one of searching for the best total infections D^* as follows:
Given the time sequence D_{reported}, epidemiological model O_{M}, find the best D^* that minimizes the MDL total cost:

$$D^* = \arg\min_{D} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{29}$$

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

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, naively searching D^* directly in large search space is intractable. Hence, we turn to use an alternate method: First, we can find quickly a good reported rate $\alpha^*_{\text{reported}}$ since we can constrain $D = \frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ to reduce the search space. Then we can search for the optimal D^* with $\alpha^*_{\text{reported}}$ 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 $\alpha^*_{\text{reported}}$, which serves as an initialization in the second step.
2. Step 2: Given the $\alpha^*_{\text{reported}}$ found in step 1, we use the Nelder-Mead [5] optimization to find the D^* that minimizes $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$ with $\alpha^*_{\text{reported}}$ constraints.

Step 1: Find the $\alpha^*_{\text{reported}}$

In step 1, we search on α_{reported} to find the $\alpha^*_{\text{reported}}$ as follows:

$$\alpha^*_{\text{reported}} = \arg\min_{\alpha_{\text{reported}}} L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{30}$$

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

$$\mathbf{p'} = \textsc{Calibrate}(O_{\text{M}}, \frac{D_{\text{reported}}}{\alpha_{\text{reported}}} - D_{\text{reported}}, D_{\text{reported}}, D_{\text{mortality}}) \tag{31}$$

Then we pick the $\alpha^*_{\text{reported}}$ that corresponds to the lowest MDL Total Cost $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$ as the $\alpha^*_{\text{reported}}$. Specifically, we will use $D(\mathbf{p'})$ instead of $\frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ as D when measuring the MDL Total Cost $L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p})$. The reason behind it is that $\frac{D_{\text{reported}}}{\alpha_{\text{reported}}}$ magnifies the noise of reported infections by $\frac{1}{\alpha_{\text{reported}}}$ times. Hence smaller α_{reported} tends to magnify the noise larger and leads to a bias towards higher MDL Total Cost. To cancel this bias introduced by the noise and get stable and robust α_{reported}, we use the more smooth $D(\mathbf{p'})$ as D in step 1.

Step 2: Find the D^* given $\alpha^*_{\text{reported}}$

With $\alpha^*_{\text{reported}}$ inferred in step 1, we will next find the D^* that minimizes the MDL Total Cost.

$$D^* = \arg\min_D L(D_{\text{reported}}, D, \mathbf{p'}, \mathbf{p}) \tag{32}$$

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

$$\sum_t D^* = \frac{\sum_t D_{\text{reported}}}{\alpha^*_{\text{reported}}} \tag{33}$$

To search for the optimal D^*, we leverage the popular Nelder-Mead search algorithm [5].

Experimental Setup

Here we describe our experimental setup in more detail and present results on additional testbeds.

Total Infections

The Results section in the main paper refers to $\text{BaseParam}_{\text{Tinf}}$, which represents the reported infections derived from the baseline parameterization BaseParam. It is computed as follows:

$$\text{BaseParam}_{\text{Tinf}} = \sum_t D(\mathbf{p}) \tag{34}$$

Similarly, $\text{MdlParam}_{\text{Tinf}}$, which represents the reported infections derived from MdlParam, is computed as follows:

$$\text{MdlParam}_{\text{Tinf}} = \sum_t D(\mathbf{p'}) \tag{35}$$

In Fig. S6, we show additional results comparing the performance MdlParam and BaseParam in estimating total infections. In Philadelphia, PA, New York City, and South Florida, the total infections $\text{MdlParam}_{\text{Tinf}}$ estimated by MdlParam is within the confidence interval given by serological studies, while $\text{BaseParam}_{\text{Tinf}}$ is not. As for Hennepin, MN, although both MdlParam and BaseParam are within the confidence interval, BaseParam is still susceptible to underestimate the total infections.

Symptomatic Rate

The baseline parameterization and MdlInfer also estimate the number of symptomatic rate $\text{BaseParam}_{\text{Symp}}$ and $\text{MdlParam}_{\text{Symp}}$ respectively. We compare these against the Facebook symptomatic surveillance data $\text{Rate}_{\text{Symp}}$.

We calculate $\text{BASEPARAM}_{\text{Symp}}$ from BASEPARAM $\mathbf{p}$ as follows:

$$\text{BASEPARAM}_{\text{Symp}} = \frac{I_S(\mathbf{p}) + I_M(\mathbf{p})}{N} \tag{36}$$

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

Similarly $\text{MDLPARAM}_{\text{Symp}}$ is computed as follows:

$$\text{MDLPARAM}_{\text{Symp}} = \frac{I_S(\mathbf{p'}) + I_M(\mathbf{p'})}{N} \tag{37}$$

Next we define ρ_{Symp} as a metric to compare $\text{BASEPARAM}_{\text{Symp}}$ and $\text{MDLPARAM}_{\text{Symp}}$ based on how well they approximate $\text{RATE}_{\text{Symp}}$. It is defined as follows.

$$\rho_{\text{Symp}} = \frac{\text{RMSE}(\text{BASEPARAM}_{\text{Symp}}, \text{RATE}_{\text{Symp}})}{\text{RMSE}(\text{MDLPARAM}_{\text{Symp}}, \text{RATE}_{\text{Symp}})} \tag{38}$$

Here, the values of ρ_{Symp} larger than 1 indicates that $\text{MDLPARAM}_{\text{Symp}}$ is closer to the $\text{RATE}_{\text{Symp}}$ than $\text{BASEPARAM}_{\text{Symp}}$, and the values of ρ_{Symp} smaller than 1 indicates that $\text{MDLPARAM}_{\text{Symp}}$ is further to the $\text{RATE}_{\text{Symp}}$ than $\text{BASEPARAM}_{\text{Symp}}$.

In Fig. S7, we present additional results comparing the accuracy MDLPARAM against BASEPARAM in estimating symptomatic rate. In Hennepin, MN, Philadelphia, PA, and South Florida, $\text{MDLPARAM}_{\text{Symp}}$ fits the symptomatic surveillance data better than $\text{BASEPARAM}_{\text{Symp}}$. However for New York City, both $\text{MDLPARAM}_{\text{Symp}}$ and $\text{BASEPARAM}_{\text{Symp}}$ diverge from $\text{RATE}_{\text{Symp}}$. For New York City, we doubt whether the New York City symptomatic surveillance data itself is of high quality (Note that New York City was one of the major COVID "hotspot" in April 2020, the 2% COVID-19 related symptomatic rate may be suspicious [9]).

Reported Infections

The baseline parameterization and MDLINFER also estimate the number of reported infections $\text{BASEPARAM}_{\text{Rinf}}$ and $\text{MDLPARAM}_{\text{Rinf}}$ respectively. We compare these against the New York Times reported infections NYT-Rinf.

Next we define ρ_{Rinf} as a metric to compare $\text{BASEPARAM}_{\text{Rinf}}$ and $\text{MDLPARAM}_{\text{Rinf}}$ based on how well they approximate NYT-Rinf. It is defined as follows:

$$\rho_{\text{Rinf}} = \frac{\text{RMSE}(\text{BASEPARAM}_{\text{Rinf}}, \text{NYT-Rinf})}{\text{RMSE}(\text{MDLPARAM}_{\text{Rinf}}, \text{NYT-Rinf})} \tag{39}$$

Here, the values of ρ_{Rinf} larger than 1 indicates that $\text{MDLPARAM}_{\text{Rinf}}$ is closer to the NYT-Rinf than $\text{BASEPARAM}_{\text{Rinf}}$, and the values of ρ_{Rinf} smaller than 1 indicates that $\text{MDLPARAM}_{\text{Rinf}}$ is further to the NYT-Rinf than $\text{BASEPARAM}_{\text{Rinf}}$.

In Fig. S8, we present additional results comparing the accuracy MDLPARAM against BASEPARAM in estimating reported infections. In Philadelphia, PA, New York City, South Florida, and Western Washington (Fall), $\text{MDLPARAM}_{\text{Rinf}}$ fits the NYT-Rinf reported infections better than $\text{BASEPARAM}_{\text{Rinf}}$. However for Western Washington (Spring), $\text{MDLPARAM}_{\text{Rinf}}$ slightly fail to fit the NYT-Rinf reported infections better than $\text{BASEPARAM}_{\text{Rinf}}$ for the observed period. This could be explained by the fact that the reported infections during the observed period for Western Washington (Spring) is too small and therefore sensitive to small fluctuations.

Reported Rate

We also calculate the dynamic reported rate from both baseline parameterization and MdlInfer. We calculate $\text{BaseParam}_{\text{Rate}}$ from BaseParam $\mathbf{p}$ as follows:

$$\text{BaseParam}_{\text{Rate}} = \frac{\sum_t \text{NYT-Rinf}}{\sum_t D(\mathbf{p})} \tag{40}$$

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

$$\text{BaseParam}_{\text{Rate}} = \frac{\sum_t \text{NYT-Rinf}}{\sum_t D(\mathbf{p'})} \tag{41}$$

In Fig. S9, we present additional results comparing the MdlParam against BaseParam in estimating temporal reported rate. In Bucks, PA, Western Washington, Philadelphia, PA, New York City, and South Florida, the reported rate estimated by MdlParam is within the confidence interval provided by serological studies while the reported rate estimated by BaseParam is outside or barely fits the interval.

Non-pharmaceutical Interventions Simulation

We also use the both baseline parameterization and MdlInfer to perform non-pharmaceutical interventions simulation. Here, both the baseline parameteriztion $\mathbf{p}$ and MdlInfer inferred $\mathbf{p'}$ 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%. Fig. S10 shows additional results on non-pharmaceutical interventions described above. We can still see the same results that MdlParam leads to more realistic non-pharmaceutical intervention simulations than BaseParam and non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic.

Sensitive Analysis

We also perform sensitivity experiments to inspect the robustness of our non-pharmaceutical interventions simulations in Fig. S11. Here, we reduce the infectiousness of the isolated infections to 3 different values, and repeat simulations in each of the scenarios. Our results 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.

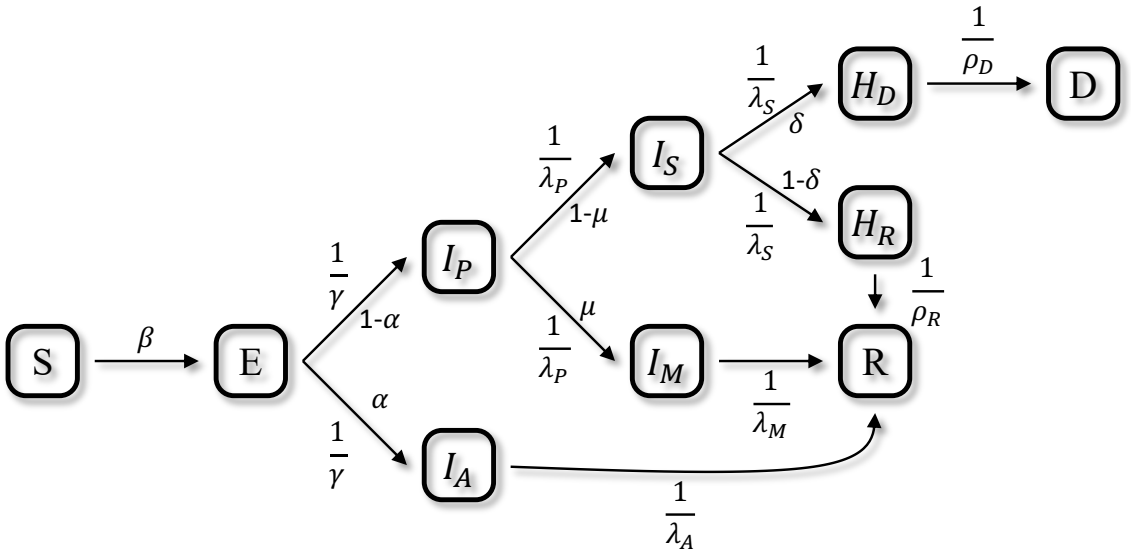

Figure S5: Compartmental diagram of epidemiological model O_{B} [4].

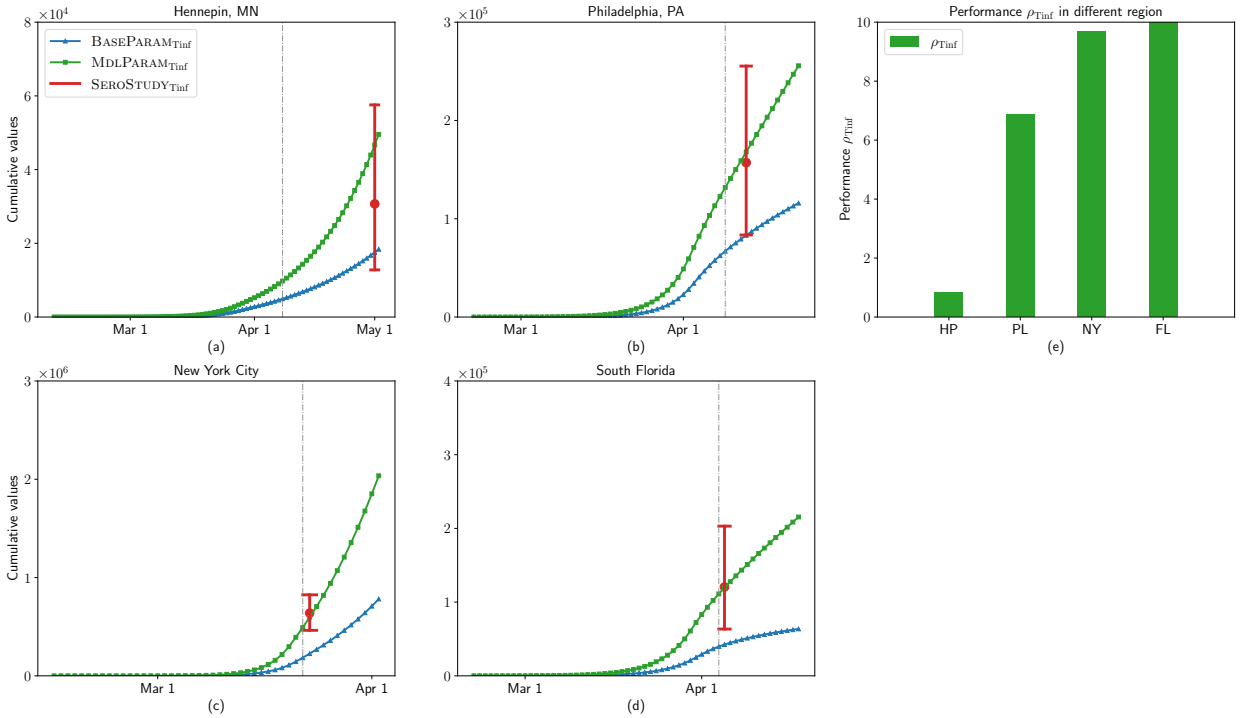

Figure S6: MdlParam estimates total infections more accurately than BaseParam. (a)-(d) The grey dash line divides the observed period (used to train BaseParam as well as MdlParam) and the future period (which was not accessible to the model while training). Blue curve and green curve represent the total infections estimated by baseline parameterization, BaseParam$_{\text{Tinf}}$, and the total infections estimated by MdlInfer parameterization, MdlParam$_{\text{Tinf}}$ respectively. The red point estimate SeroStudy$_{\text{Tinf}}$ and confidence interval represent the total infections estimated by serological studies [3, 6]. Note that each plot corresponds to a different geographic region, and the scales are different. (e) The performance metric, ρ_{Tinf}, comparing MdlParam against BaseParam in estimating total infections are shown for the regions in (a)-(d). Here the values of ρ_{Tinf} are 0.82, 6.87, 9.69, and 78.04, implying that MdlParam generally performs better in estimating total infections than BaseParam.

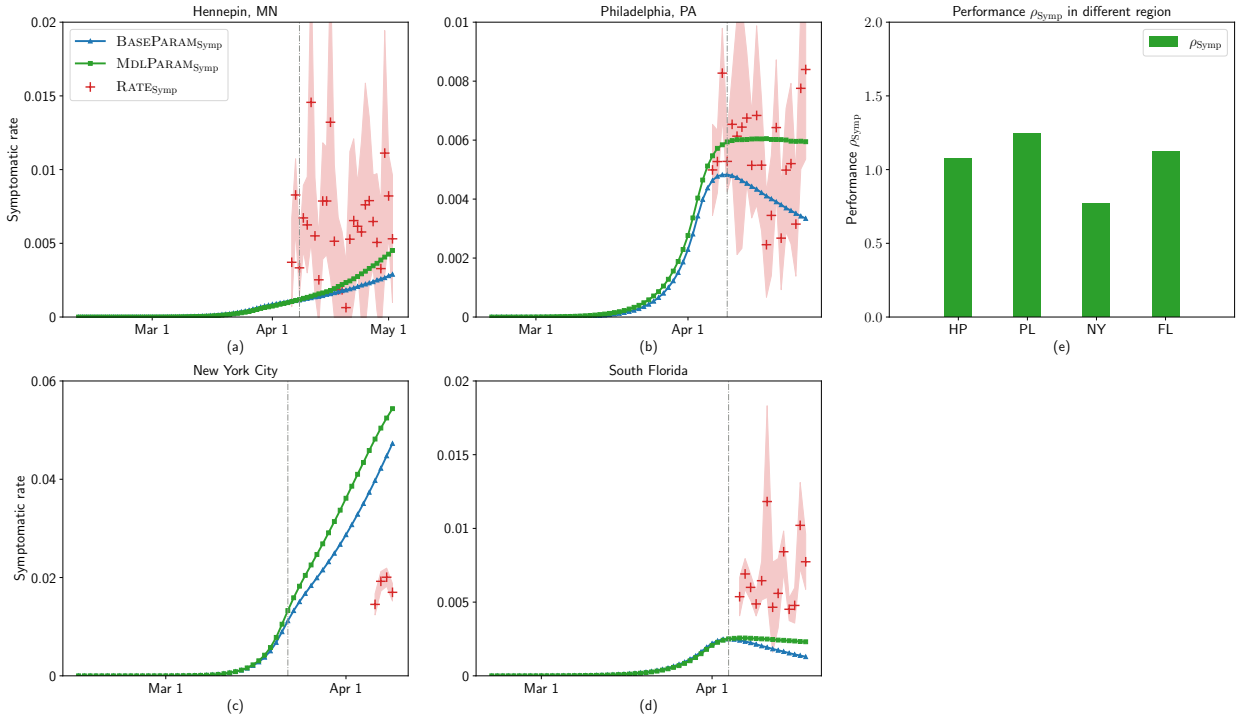

Figure S7: MdlParam estimates more accurate symptomatic rate than BaseParam. (a)-(d) Blue curve and green curve represent the symptomatic rate estimated by baseline parameterization, BaseParam$_{\text{Symp}}$, and symptomatic rate estimated by MdlInfer parameterization, MdlParam$_{\text{Symp}}$. The red point estimate Rate$_{\text{Symp}}$ and confidence interval represent the COVID-related symptomatic rate from Facebook's symptomatic surveillance data [2]. Each plot corresponds to a different geographic region, and the scales are different. (e) The performance metric, ρ_{Symp}, comparing MdlParam against BaseParam in estimating symptomatic rate are shown for the regions in (a)-(d). Here the values of ρ_{Symp} are 1.07, 1.25, 0.77, and 1.12, implying that MdlParam generally performs better in estimating symptomatic rate than BaseParam.

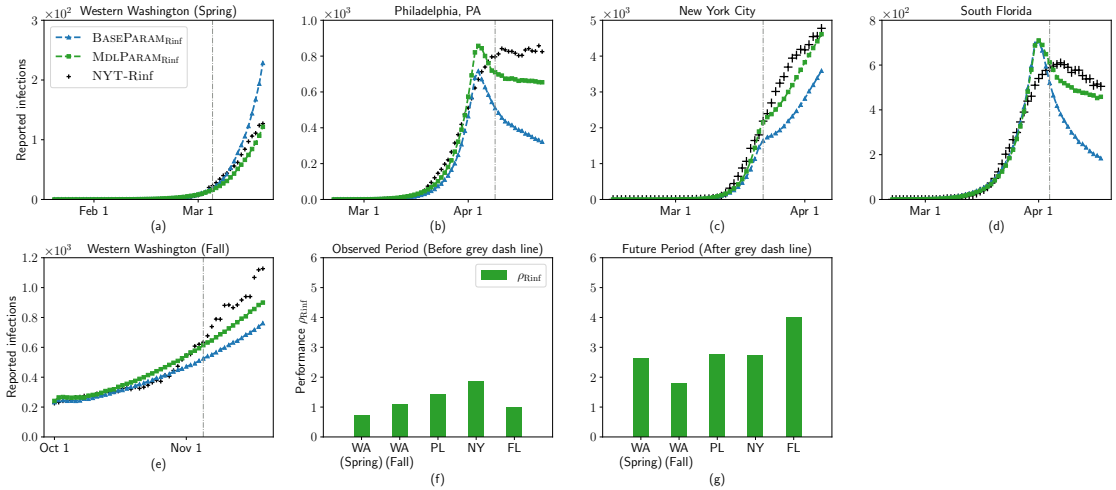

Figure S8: MdlParam leads to better fit and projection than BaseParam. (a)-(e) The grey dash line divides the observed period and future period. Black plus symbols, blue curve, and green curve represent the reported infections NYT-Rinf, reported infections estimated by baseline parameterization, $\text{BaseParam}_{\text{Rinf}}$, and reported infections estimated by MdlInfer parameterization, $\text{MdlParam}_{\text{Rinf}}$ respectively. Note that each plot corresponds to a different geographic region. As seen in the figure, $\text{MdlParam}_{\text{Rinf}}$ aligns much closer with NYT-Rinf than $\text{BaseParam}_{\text{Rinf}}$. Besides, $\text{MdlParam}_{\text{Rinf}}$ projects the future trends better than $\text{BaseParam}_{\text{Rinf}}$. (f)-(g) The performance metric, ρ_{Rinf}, comparing MdlParam against BaseParam in estimating reported infections are shown for the regions for both observed period (f), and future period (g). In both observed and future period, MdlInfer estimates reported infections better than BaseParam, and ρ_{Rinf} in future period is even larger than ρ_{Rinf} in observed period.

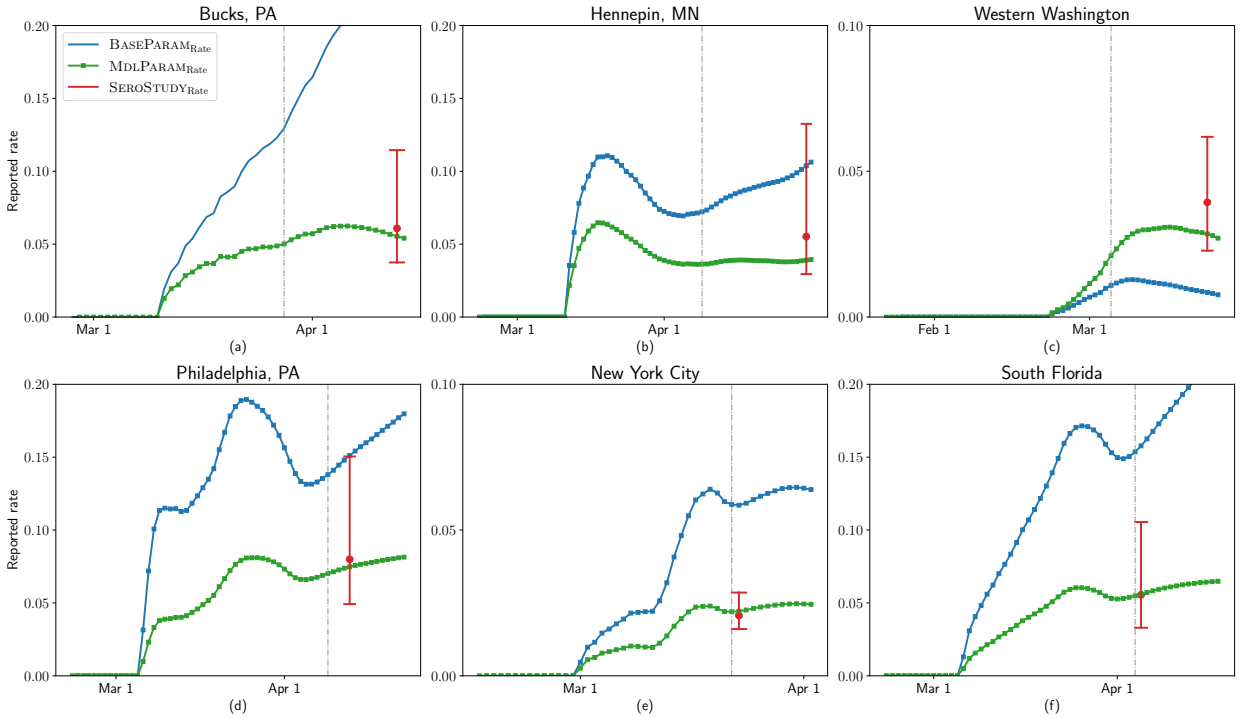

Figure S9: MdlParam estimates reported rate more accurately than BaseParam. The grey dash line divides the observed period and future period. Blue curve and green curve represent the reported rate estimated by baseline parameterization, $\text{BaseParam}_{\text{Rate}}$, and reported rate estimated by MdlInfer parameterization, $\text{MdlParam}_{\text{Rate}}$ respectively. The red point estimate $\text{SeroStudy}_{\text{Rate}}$ and confidence interval represent the reported rate estimated by serological studies [3, 6]. Note that each plot corresponds to a different geographic region. As seen in the figure, the reported rate estimated by MdlParam is within the confidence interval provided by serological studies while the reported rate estimated by BaseParam is outside or barely in the interval.

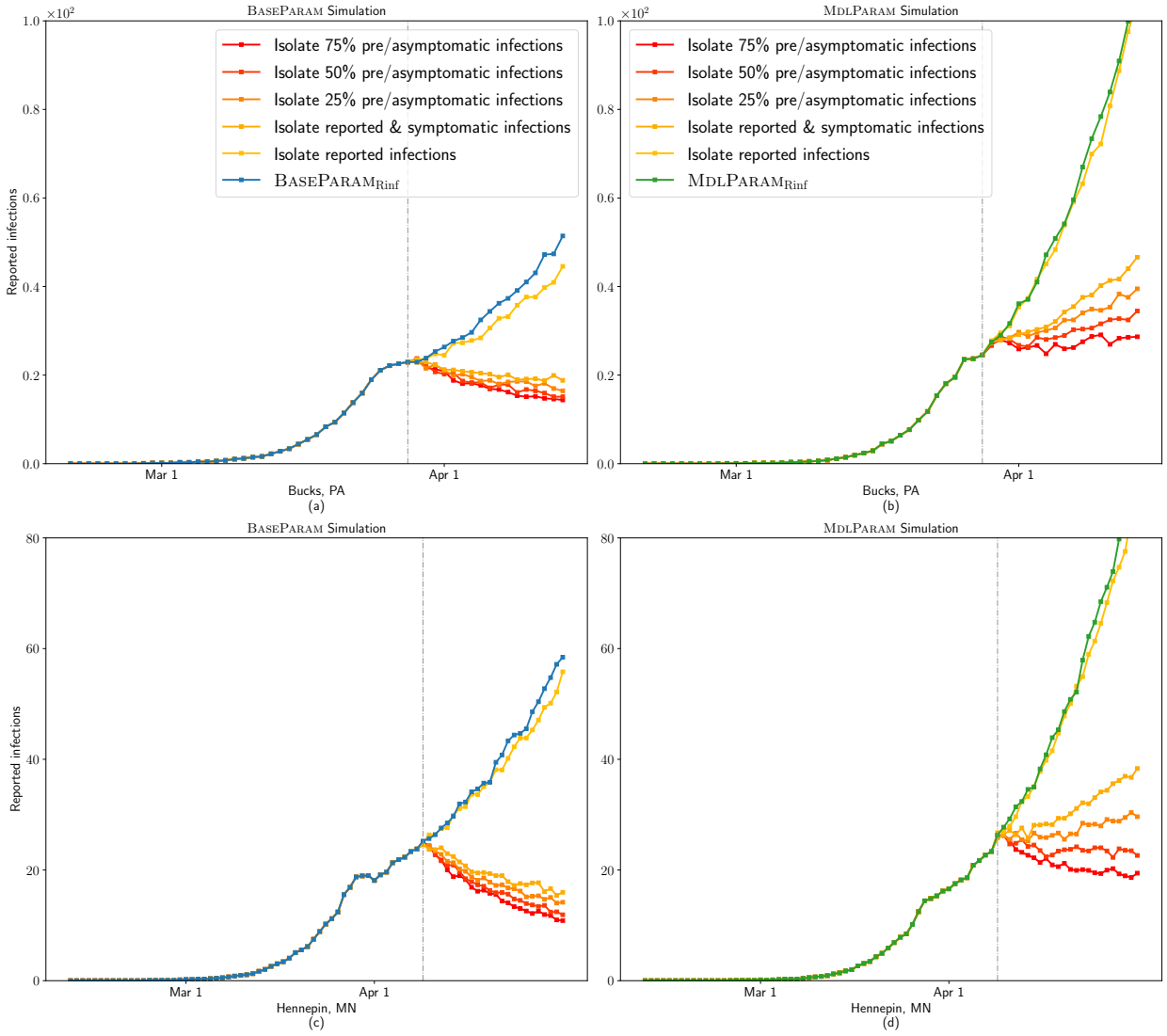

Figure S10: MdlParam leads to more realistic non-pharmaceutical intervention simulations than BaseParam. (a)&(c) Accuracy of non-pharmaceutical intervention simulations relies on the good inference of unreported infections. The grey dash line divides the observed period and future period. The blue curve represents the reported infections estimated by BaseParam. The other five curves represent the simulated reported infections for 5 scenarios: Isolate the reported infections, symptomatic infections, symptomatic infections and 25%, 50%, 75% asymptomatic and presymptomatic infections, where we reduce the infectiousness of these isolated infections to half in future period. (b)&(d) Non-pharmaceutical interventions on asymptomatic and presymptomatic infections are essential to control the COVID-19 epidemic. The green curve represents the reported infections estimated by MdlParam.

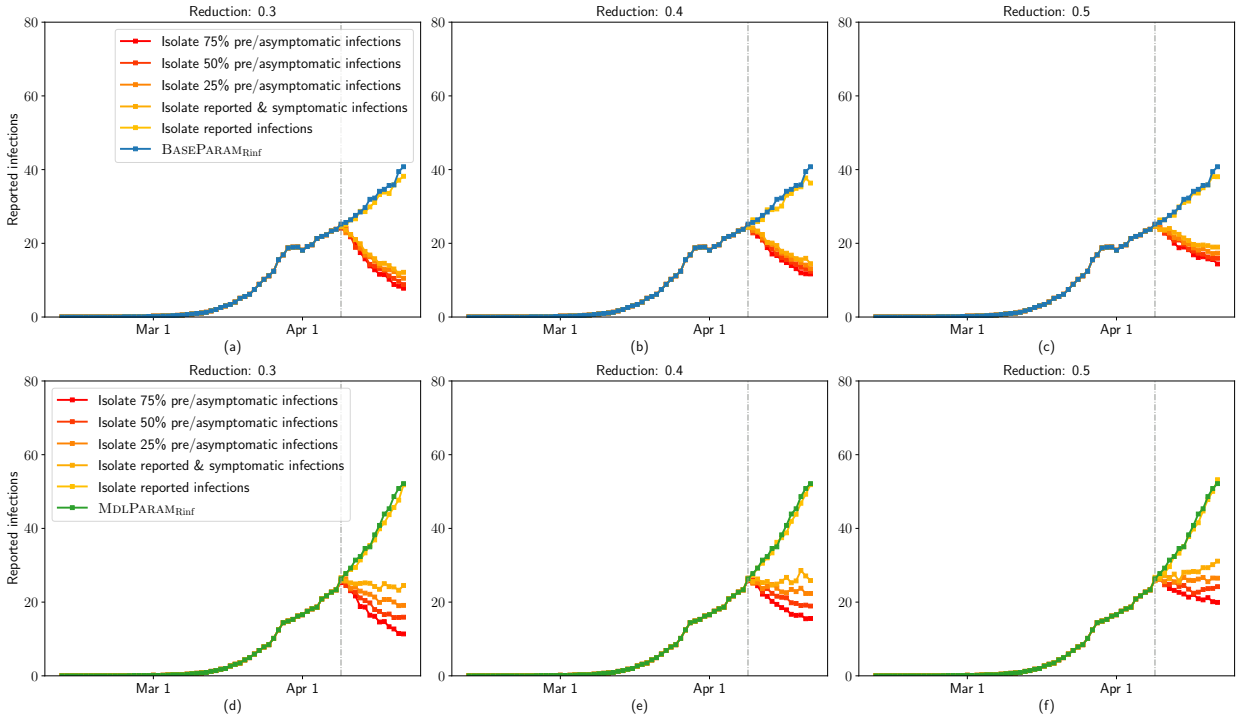

Figure S11: Our non-pharmaceutical interventions simulation results are robust. In (a) to (c), the grey dash line divides the observed period and future period. The blue curve represents the reported infections estimated by BaseParam. The other five curves represent the simulated reported infections for 5 scenarios: Isolate the reported infections, symptomatic infections, symptomatic infections and 25%, 50%, 75% asymptomatic and presymptomatic infections, where we reduce the infectiousness of these isolated infections to 30% in (a), 40% in (b), and 50% in (c) in future period. In (d) to (f), the grey dash line divides the observed period and future period. The green curve represents the reported infections estimated by MdlParam. The other five curves represent the simulated reported infections for 5 scenarios: Isolate the reported infections, symptomatic infections, symptomatic infections and 25%, 50%, 75% asymptomatic and presymptomatic infections, where we reduce the infectiousness of these isolated infections to 30% in (d), 40% in (e), and 50% in (f) in future period.

Table S1: List of notations

[h!] Symbol	Description
MdlInfer	Our Minimum Description Length (MDL) framework to estimate total infections
O_{M}	Epidemiological model [4] used in MdlInfer
BaseParam	Parameterizaton of base epidemiological model
MdlParam	MdlInfer parameterizaton inferred by our MdlInfer framework
$\mathrm{SeroStudy}_{\mathrm{Tinf}}$	Total infections estimated by serological studies
$\mathrm{BaseParam}_{\mathrm{Tinf}}$	Total infections estimated by baseline parameterization
$\mathrm{MdlParam}_{\mathrm{Tinf}}$	Total infections estimated by MdlInfer parameterizaton
ρ_{Tinf}	The performance metric comparing MdlInfer against baseline parameterizaton in estimating total infections
$\mathrm{Rate}_{\mathrm{Symp}}$	COVID-related symptomatic rate from Facebook's symptomatic surveillance data
$\mathrm{BaseParam}_{\mathrm{Symp}}$	Symptomatic rate estimated by baseline parameterizaton
$\mathrm{MdlParam}_{\mathrm{Symp}}$	Symptomatic rate estimated by MdlInfer parameterizaton
ρ_{Symp}	The performance metric comparing MdlInfer against baseline parameterizaton in estimating symptomatic rate
NYT-Rinf	New York Times reported infections
$\mathrm{BaseParam}_{\mathrm{Rinf}}$	Reported infections estimated by baseline parameterizaton
$\mathrm{MdlParam}_{\mathrm{Rinf}}$	Reported infections estimated by MdlInfer parameterizaton
ρ_{Rinf}	The performance metric comparing MdlInfer against baseline parameterizaton in estimating reported infections
$\mathrm{SeroStudy}_{\mathrm{Rate}}$	Reported rate estimated by serological studies
$\mathrm{BaseParam}_{\mathrm{Rate}}$	Reported rate estimated by baseline parameterizaton
$\mathrm{MdlParam}_{\mathrm{Rate}}$	Reported rate estimated by MdlInfer parameterizaton

References

[1] Coronavirus in the U.S.:Latest Map and Case Count, 2020.

[2] Delphi's COVID-19 Surveys, 2020.

[3] Bendavid, E., Mulaney, B., Sood, N., Shah, S., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Bogan, A., et al. Covid-19 antibody seroprevalence in santa clara county, california. *International journal of epidemiology 50*, 2 (2021), 410–419.

[4] Childs, M. L., Kain, M. P., Kirk, D., Harris, M., Couper, L., Nova, N., Delwel, I., Ritchie, J., and Mordecai, E. A. The impact of long-term non-pharmaceutical interventions on covid-19 epidemic dynamics and control. *medRxiv* (2020).

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

[6] Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., 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 (2020), 1576–1586.

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

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

[9] Thompson, C. N., Baumgartner, J., Pichardo, C., Toro, B., Li, L., Arciuolo, R., Chan, P. Y., Chen, J., Culp, G., Davidson, A., et al. Covid-19 outbreak—new york city, february 29–june 1, 2020. *Morbidity and Mortality Weekly Report 69*, 46 (2020), 1725.