Supporting Information for

Rapid Phylogeographic Inference of Regional Epidemic Dynamics for Routine Genomic Surveillance

Leke Lyu[a], Nicola F. Müller[b], Zewen Liu[c], Lambodhar Damodaran[d], Wei Jin[c], Max S. Y. Lau[a]*
a. Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322;
b. Division of HIV, Infectious Diseases & Global Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA 94143;
c. Department of Computer Science, Emory University, Atlanta, GA 30322;
d. Department of Biology, Emory University, Atlanta, GA 30322;

*Max S. Y. Lau
Email: msy.lau@emory.edu

This PDF file includes:

Simulation model and conventions
Tables S1 to S6

Drivers of the Source-Sink Score
Figures S1 to S3

STEPHY Enables Accurate and Rapid Inference
Figures S4

STEPHY Ranks Transmission Hotspots and Generalizes Across Population Scales
Figures S5 to S7

Real-World Application on Denmark
Figures S8 to S9

SI References

Simulation model and conventions

Datasets described in **Supplementary Tables 1–6** were generated using the stochastic SIR metapopulation model described in the Materials and Methods section of the main text. Simulations were performed with the Gillespie algorithm implemented in ReMASTER (BEAST2) (1, 2). Unless explicitly stated otherwise, all datasets follow the shared conventions described below. Each table reports only the parameter settings specific to that dataset.

For each region i, N_i denotes the population size, R_{0i} denotes the basic reproduction number, γ_i denotes the recovery rate, and δ denotes the sampling rate. The migration rate from region i to region j is denoted by m_{ij}. The transmission rate β_i is parameterized as $\beta_i = R_{0i}(\gamma_i + \delta)$. All rates are expressed in day^{-1}.

Each simulation is initialized with a single infectious individual in one region selected uniformly at random. Thus, the seeded region has $I_i(0) = 1$, whereas all other regions have $I_i(0) = 0$. All regions start with $R_i(0)$ and $S_i(0) = N_i - I_i(0)$.

Region-specific values of R_{0i} and γ_i are sampled under a bounded-difference constraint, so that no two regions differ by more than the specified maximum. The sampling rate δ is shared across regions. Population sizes N_i and the asymmetric migration matrix m_{ij}, with zero diagonal entries, are sampled independently.

Simulation duration is specified in units of the mean infectious period. A value τ is drawn from the range specified in each table and converted to calendar time as $T = \tau/\bar{\gamma}$, where $\bar{\gamma}$ is the mean recovery rate across regions.

Each simulation terminates when either the maximum simulation time T is reached or the sampled tree size reaches the specified cap $nTips$, whichever occurs first. A simulated genealogy is retained only if every region contributes more than 30 sampled tips. Otherwise, the parameter configuration is discarded, a new configuration is sampled, and the simulation is repeated for up to 50 attempts.

Supplementary Table 1 defines the primary benchmark dataset used to train and evaluate STEPHY, whereas **Supplementary Table 2** defines the Denmark-calibrated dataset used for empirical inference. The remaining tables provide additional simulation datasets for targeted evaluations, including analyses of source–sink score (SSS) behavior under different epidemiological settings (**Supplementary Table 3**) and model generalization across population scales (**Supplementary Tables 4–6**).

Supplementary Table 1.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	12	-
Population size	N_i	$U(5000,50000)$	individuals
Basic reproduction number	R_{0i}	$U(2,8),\ \max diff \leq 2$	-
Recovery rate	γ_i	$U(0.05,0.25),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	$U(0.0002,0.0028)$	day^{-1}
Migration rate	m_{ij}	$U(0.0005,0.0045)$	day^{-1}
Horizon	τ	$U(6,18) \rightarrow T = \tau/\bar{\gamma}$	Infectious periods
Tree size cap	$nTips$	5,000 tips	-

Supplementary Table 2.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	5 (Hovedstaden, Midtjylland, Syddanmark, Sjælland, Nordjylland)	-
Population size	N_i	1,860,000 / 1,330,000 / 1,220,000 / 840,000 / 590,000	individuals
Basic reproduction number	R_{0i}	$Beta(2,3.5)\ rescaled\ to\ [0.5,4], max\ diff \leq 1$	-
Recovery rate	γ_i	$U(0.07,0.23),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	$U(0.001,0.01)$	day^{-1}
Migration rate	m_{ij}	$U(0.001,0.012)$	day^{-1}
Horizon	T	$U(20,240)$	days
Tree size cap	$nTips$	8,000 tips	-

Supplementary Table 3.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	12	-
Population size	N_i	$U(20000,35000)$	individuals
Basic reproduction number	R_{0i}	$U(2,8),\ \max diff \leq 4$	-
Recovery rate	γ_i	$U(0.05,0.25),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	$U(0.0002,0.0028)$	day^{-1}
Migration rate	m_{ij}	$U(0.0005,0.0045)$	day^{-1}
Horizon	τ	$U(6,18) \rightarrow T = \tau/\bar{\gamma}$	Infectious periods
Tree size cap	$nTips$	5,000 tips	-

Supplementary Table 4.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	12	-
Population size	N_i	$U(5000,50000)$	individuals
Basic reproduction number	R_{0i}	$U(2,8),\ \max diff \leq 2$	-
Recovery rate	γ_i	$U(0.05,0.25),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	0.0015	day^{-1}
Migration rate	m_{ij}	$U(0.0005,0.0045)$	day^{-1}
Horizon	τ	$U(6,18) \rightarrow T = \tau/\bar{\gamma}$	Infectious periods
Tree size cap	$nTips$	5,000 tips	-

Supplementary Table 5.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	12	-
Population size	N_i	$U(10000,100000)$	individuals
Basic reproduction number	R_{0i}	$U(2,8),\ \max diff \leq 2$	-
Recovery rate	γ_i	$U(0.05,0.25),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	0.00075	day^{-1}
Migration rate	m_{ij}	$U(0.0005,0.0045)$	day^{-1}
Horizon	τ	$U(6,18) \rightarrow T = \tau/\bar{\gamma}$	Infectious periods
Tree size cap	$nTips$	5,000 tips	-

Supplementary Table 6.

Parameter	Symbol	Distribution / Value	Units
Num of regions	n	12	-
Population size	N_i	$U(15000,150000)$	individuals
Basic reproduction number	R_{0i}	$U(2,8),\ \max diff \leq 2$	-
Recovery rate	γ_i	$U(0.05,0.25),\ \max diff \leq 0.01$	day^{-1}
Sampling rate	δ	0.0005	day^{-1}
Migration rate	m_{ij}	$U(0.0005,0.0045)$	day^{-1}
Horizon	τ	$U(6,18) \rightarrow T = \tau/\bar{\gamma}$	Infectious periods
Tree size cap	$nTips$	5,000 tips	-

Drivers of the Source-Sink Score

For each simulated outbreak, we summarize the migratory role of every region using the reconstructed transmission history. Let a_{ij} denote the number of realized transmission events from region i to region j. The viral outflow and inflow of region i are $out_i = \sum_j a_{ij}$ and $in_i = \sum_j a_{ji}$, respectively, where the sums are taken over all $j \neq i$). The Source-Sink Score is defined as $SSS_i = (out_i - in_i)/(out_i + in_i)$. SSS_i ranges from −1 to +1. Values near +1indicate net sources of infection, values near −1 indicate net sinks, and values near 0 indicate approximately balances inflow and outflow (3, 4).

We next asked which epidemiological characteristics determine a region's realized SSS_i. We considered three candidate predictors: population size (N_i), local transmissibility (R_{0i}), and migration asymmetry. Larger populations can sustain more transmission and may therefore export more infections. Higher R_0 accelerates epidemic growth, increasing the pool of infections available for export. Migration asymmetry is summarized by a migration index, $MigIdx_i = (\sum_j m_{ij} - \sum_j m_{ji})/(\sum_j m_{ij} + \sum_j m_{ji})$, which measures the source-sink tendency implied by the migration rates alone.

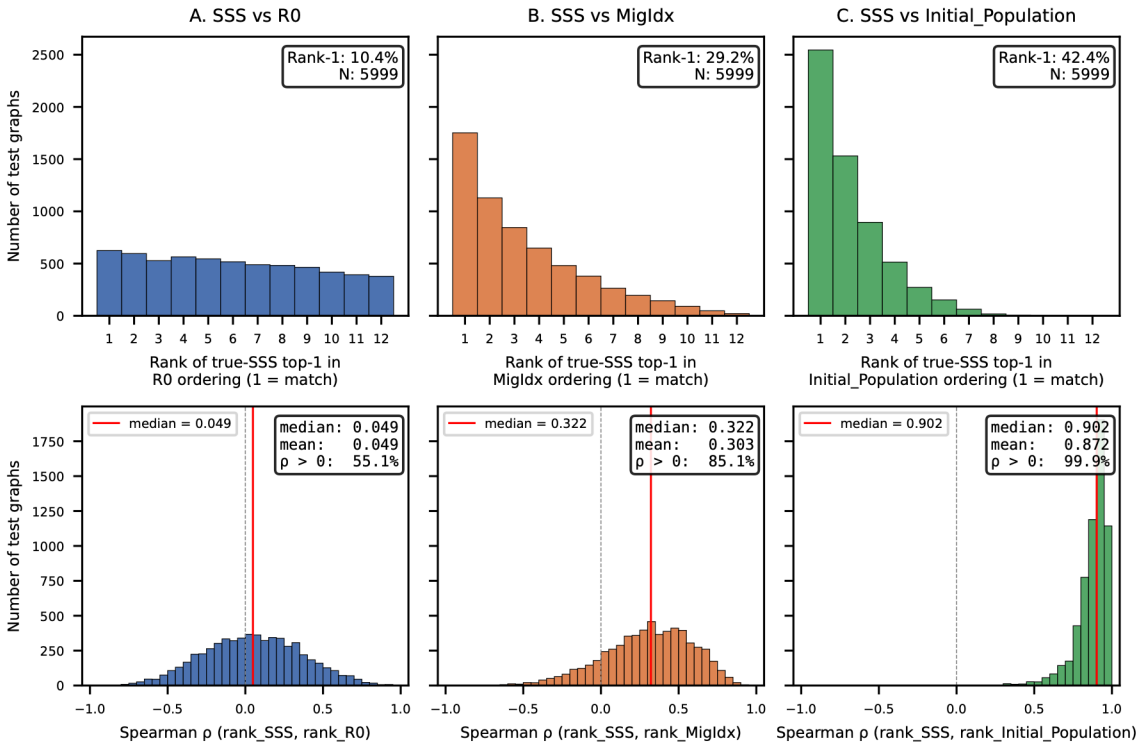
To assess how well each candidate predicts SSS_i, we evaluated N_i, R_{0i}, and $MigIdx_i$ independently for each simulated genealogy. For every genealogy, we ranked regions by each predictor and recorded two summaries: whether the top-ranked region matched the region with the highest realized SSS_i, and the Spearman rank correlation (ρ) between each predictor and SSS_i across regions. **Supplementary Figures 1-3** report the distributions of these summaries across genealogies in the corresponding datasets.

In the primary benchmark dataset (**Supplementary Table 1**), regional population sizes span an order of magnitude, $N_i \sim U(5000,50000)$, whereas R_{0i} is constrained to differ by at most 2 across regions. Under this regime, population size is the strongest predictor of SSS_i, most frequently identifying the top source region and yielding the highest median Spearman correlation (**Supplementary Figure 1**).

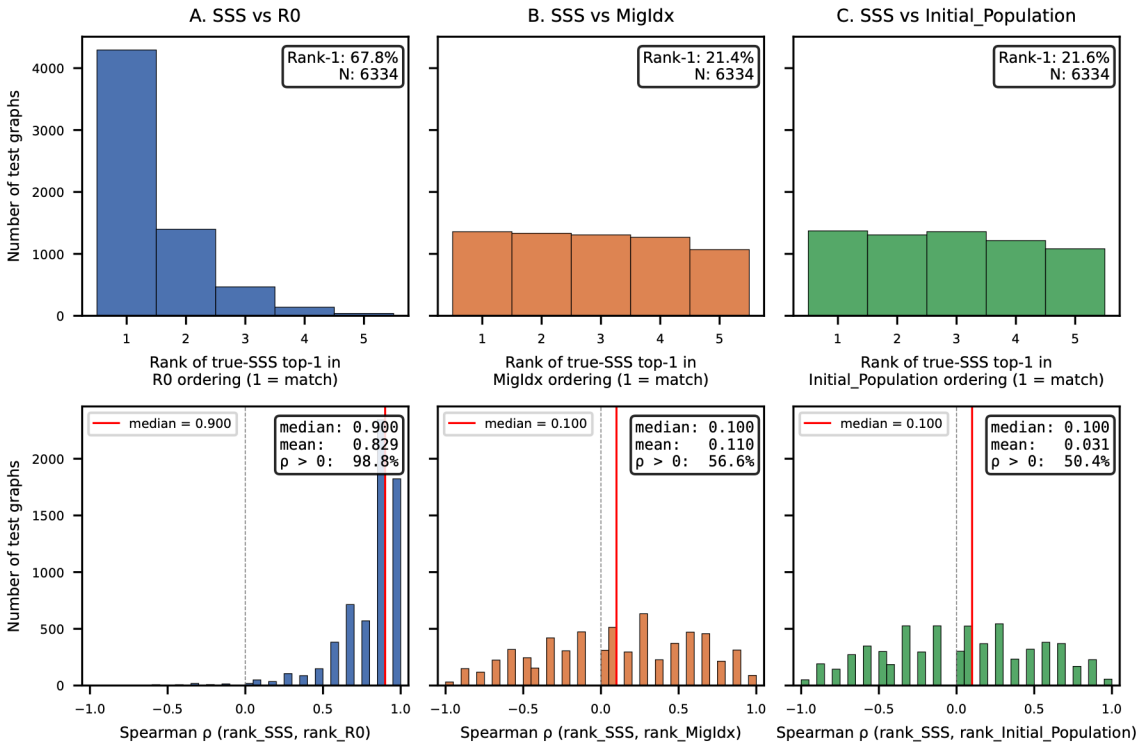
In the Denmark-calibrated dataset (**Supplementary Table 2**), population sizes are fixed to the five Danish regions and are large enough that susceptible depletion is negligible over the simulated horizon. Early growth in each region is therefore governed primarily by transmissibility. Accordingly, R_{0i}, drawn per region from a Beta distribution rescaled to [0.5, 4], is the strongest predictor of SSS_i (**Supplementary Figure 2**).

In the compressed-population dataset (**Supplementary Table 3**), population sizes are drawn from a narrower range, $N_i \sim U(20000,35000)$, making regions approximately equal in size. In this regime, migration asymmetry becomes the dominant structural determinant, and $MigIdx_i$ is the strongest predictor of SSS_i (**Supplementary Figure 3**).

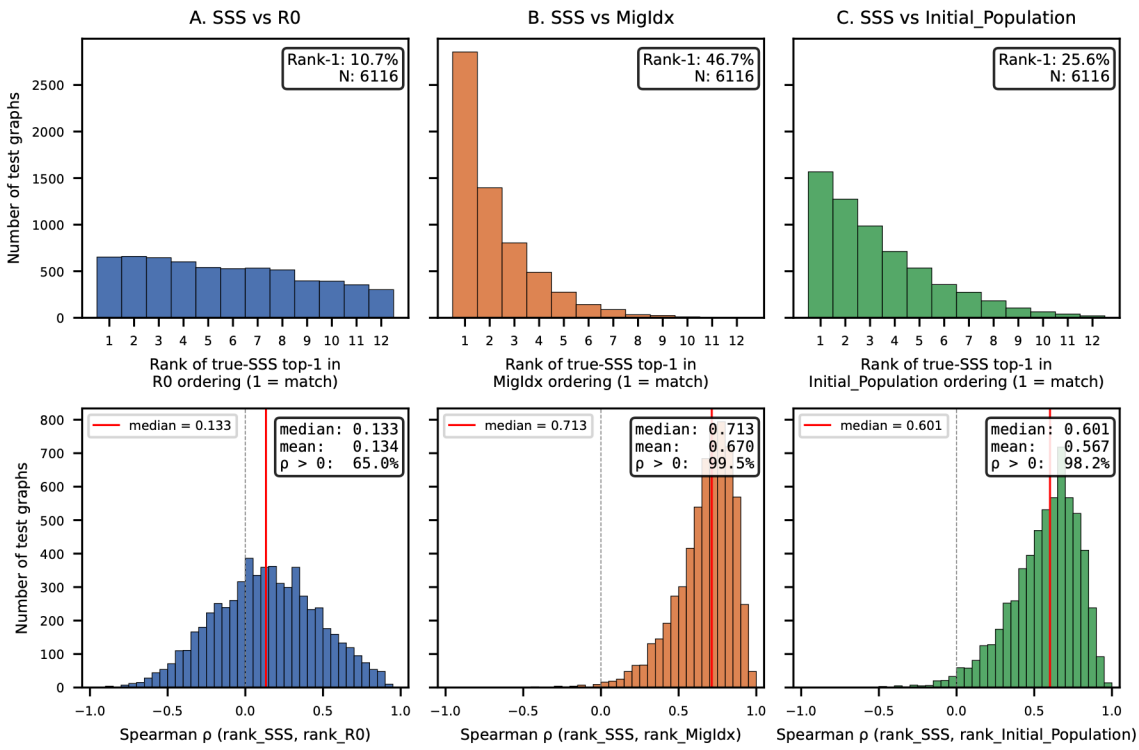
Together, these analyses show that the dominant determinant of SSS_i depends on the epidemiological regime. Population size, transmissibility, and migration asymmetry each become the strongest predictor when variation in the other factors is constrained.



Supplementary Fig. S1. Exploring predictors of the Source-Sink Score in the primary benchmark dataset.



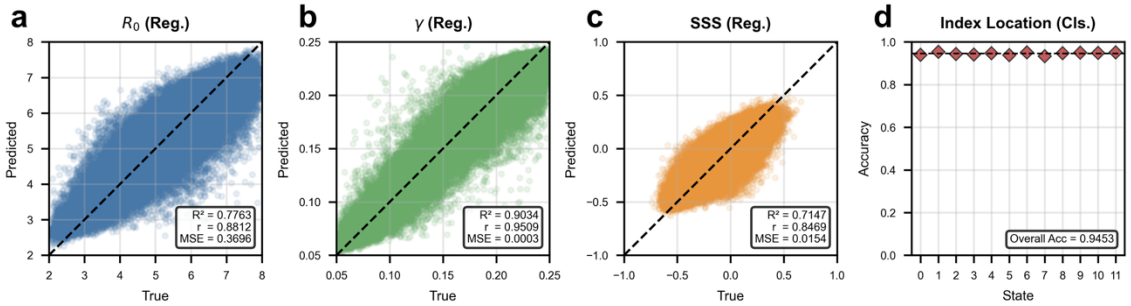
Supplementary Fig. S2. Exploring predictors of the Source-Sink Score in the Denmark-calibrated dataset.



Supplementary Fig. S3. Exploring predictors of the Source-Sink Score in the compressed-population dataset.

STEPHY Enables Accurate and Rapid Inference

Here in **Supplementary Fig. S4**, we present the performance of CBLV-CNN on the primary benchmark dataset described in **Supplementary Table 1**, using the same regression and classification tasks evaluated in the main text. The figure follows the style of **Fig. 2a–d**. Panels **a–c** show predicted versus true values for the three regression targets: R_0, γ, and SSS. The dashed line marks the identity line, and the insets report the coefficient of determination (R^2), Pearson correlation coefficient (r), and mean squared error (MSE), all computed using the median quantile prediction from the CQR head. Panel **d** shows the per-region accuracy for index-location inference, with the dashed horizontal line marking the overall accuracy reported in the inset. Overall, **Supplementary Fig. S4** shows that CBLV-CNN performs worse than STEPHY on the regression tasks

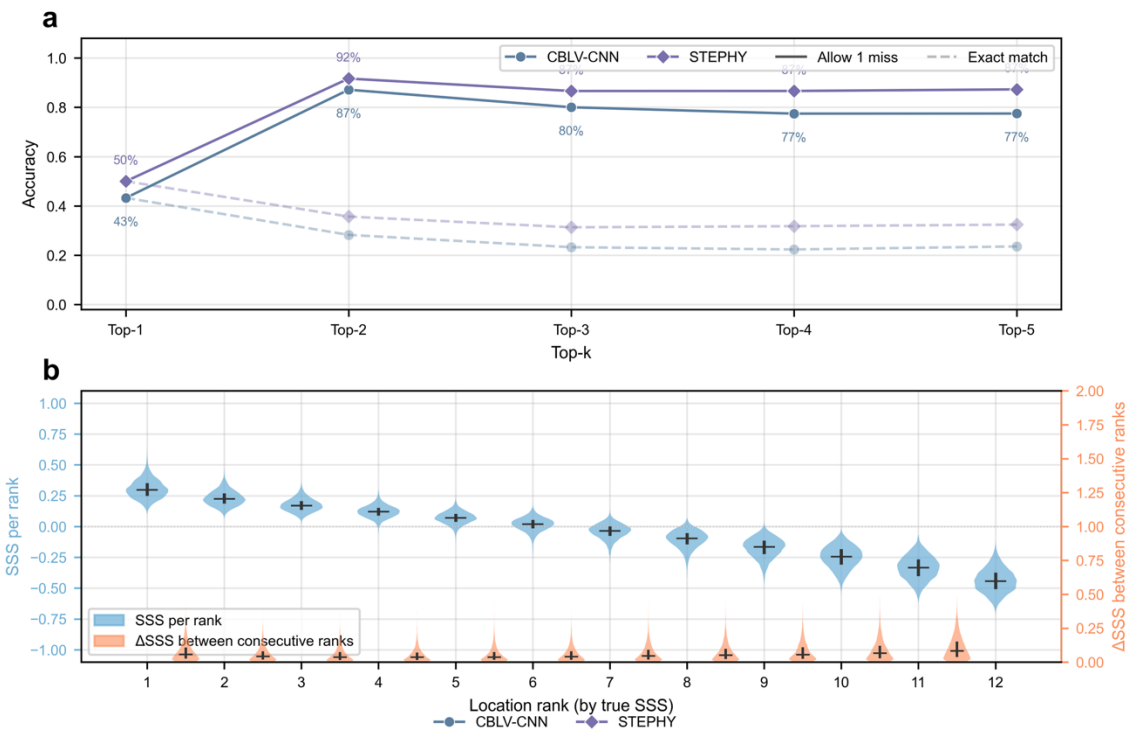


Supplementary Fig. S4. CBLV-CNN performance across regression and classification tasks.

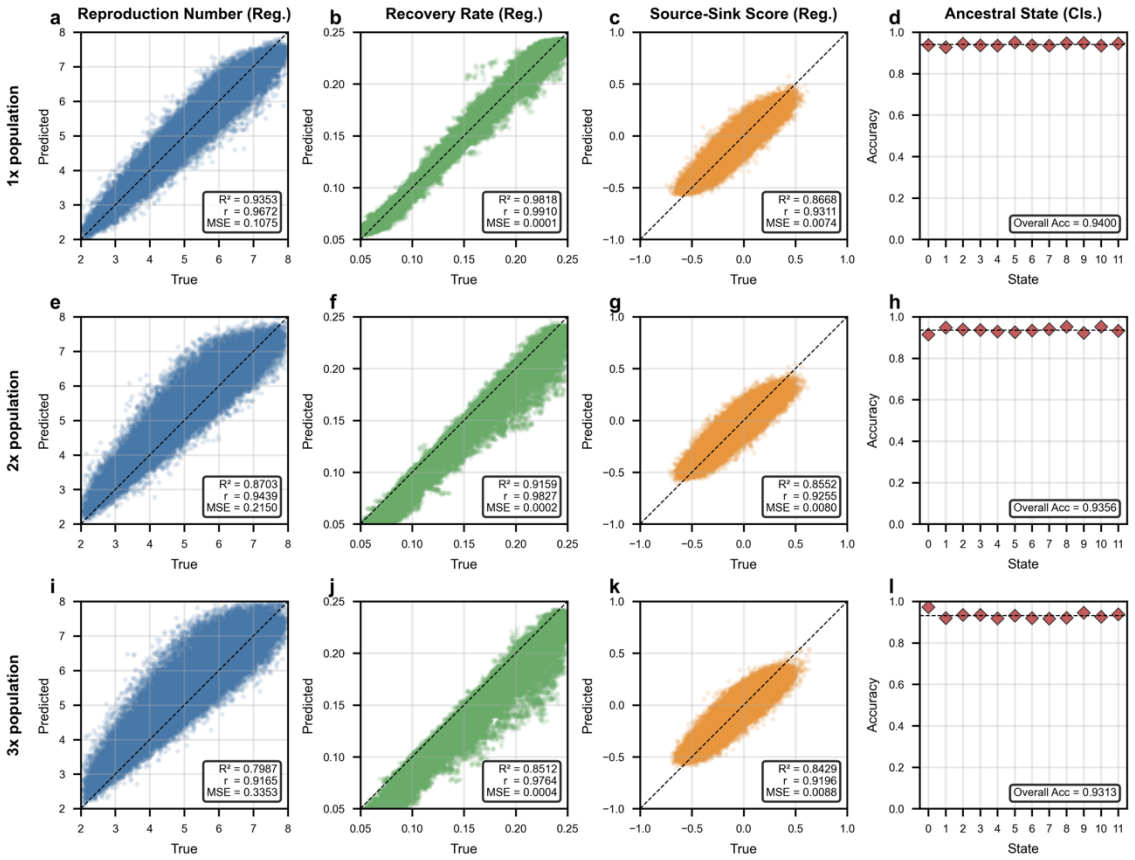
STEPHY Ranks Transmission Hotspots and Generalizes Across Population Scales

In **Supplementary Fig. S5**, we present STEPHY's ability to identify regions with high SSS_i values on the primary benchmark dataset described in **Supplementary Table 1**. Similar to **Fig. 3a–b**, this analysis evaluates whether the model can recover the top-ranked regions rather than only estimate their absolute values. Panel **a** shows top-k accuracy for identifying regions with the highest SSS_i values across CBLV-CNN and STEPHY. Solid lines allow one mismatch, whereas dashed lines indicate exact-match accuracy. Panel **b** shows the distribution of true SSS_i values by descending rank across all test outbreaks. Blue violins show SSS_i values at each rank, and orange violins show the gaps between adjacent ranks. Thick vertical lines indicate the interquartile range, and horizontal lines indicate the median. Overall, STEPHY shows similar ranking performance for SSS_i as observed for R_0, indicating that it can reliably identify the regions that act as the strongest transmission sources.

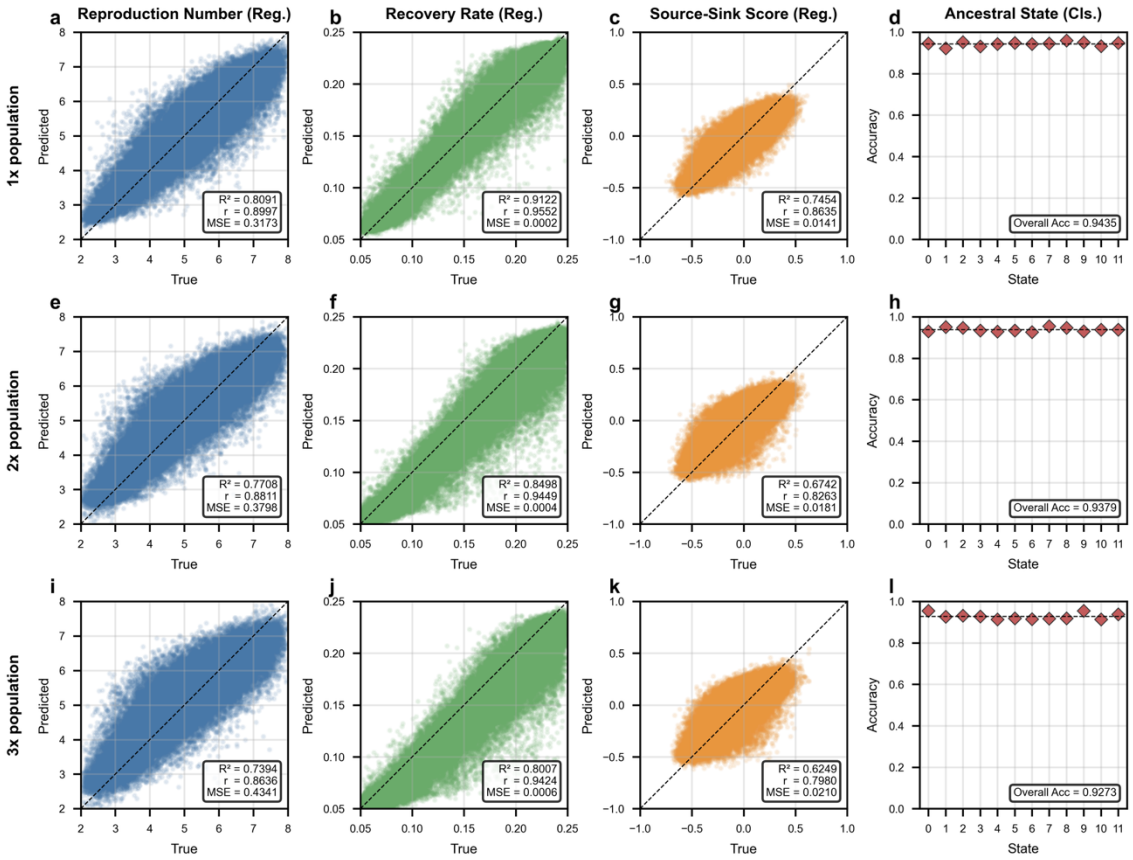
We further provide a detailed evaluation of population-scale generalization for STEPHY and CBLV-CNN. **Supplementary Fig. S6** shows the results for STEPHY, and **Supplementary Fig. S7** shows the corresponding results for CBLV-CNN. In both figures, the first row shows the baseline setting, where population sizes and sampling rates match the training scale (population size x1; sampling rate x1; **Supplementary Table 4**). The second row shows a twofold population-scale shift, with population sizes doubled and sampling rates reduced by half to maintain comparable tree sizes (population size x2; sampling rate x0.5; **Supplementary Table 5**). The third row shows a threefold population-scale shift, with population sizes tripled and sampling rates reduced to one-third (population size x3; sampling rate x0.33; **Supplementary Table 6**). Overall, both models show declining performance as population size shifts beyond the training range, but STEPHY consistently performs better than CBLV-CNN across regression targets. In particular, SSS_i estimation remains much more stable for STEPHY across population scales. We also observe a systematic prediction bias as population size increases: R_{0i} tends to be overestimated, whereas γ_i tends to be underestimated.



Supplementary Fig. S5. STEPHY ranking performance for identifying regions with the highest Source-Sink Scores.



Supplementary Fig. S6. STEPHY generalization across increasing population scales.

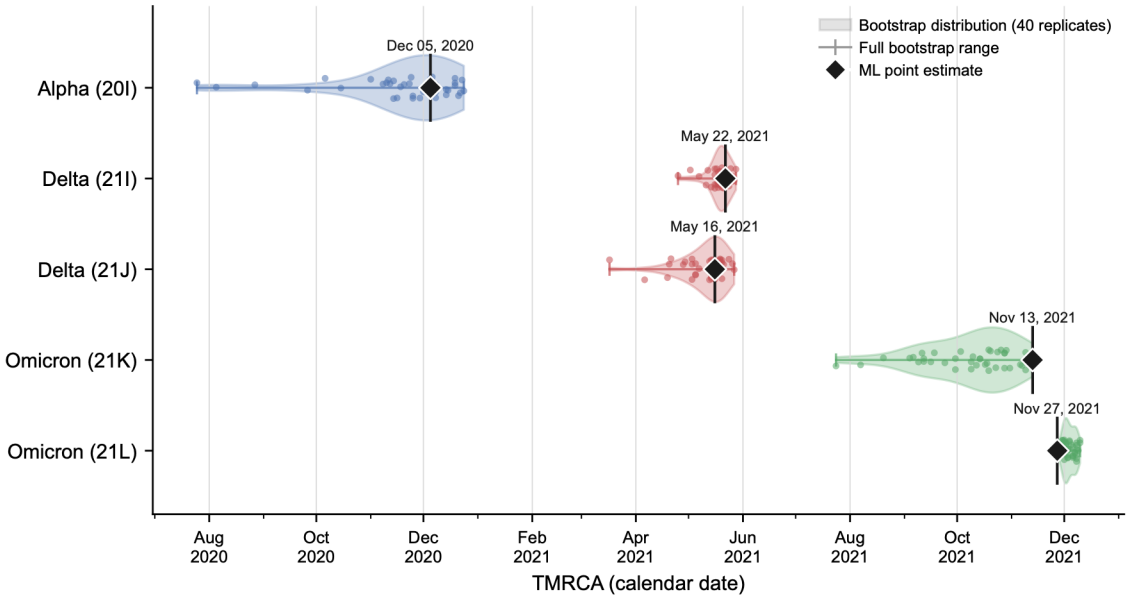


Supplementary Fig. S7. CBLV-CNN generalization across increasing population scales.

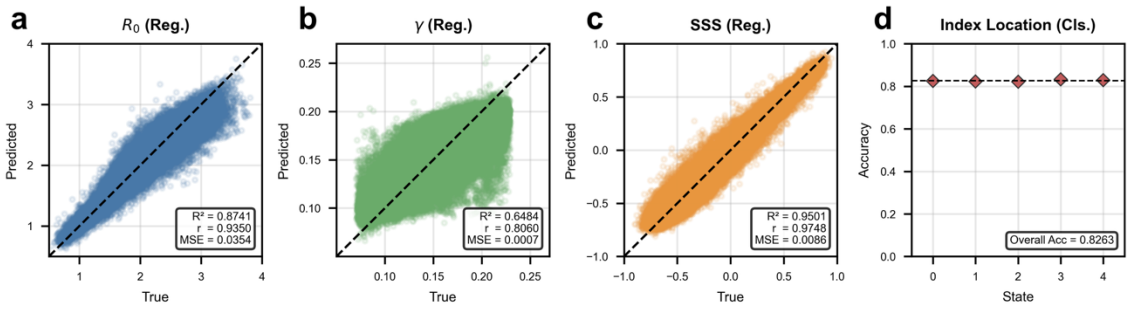
Real-World Application on Denmark

In **Supplementary Fig. S8**, we summarize phylogenetic uncertainty in the inferred TMRCA of each target lineage. For each lineage, the figure shows the distribution of TMRCA estimates across 40 bootstrap trees, together with the TMRCA estimated from the maximum-likelihood tree. The maximum-likelihood estimate generally falls within the bootstrap distribution. Among the five lineages, Alpha 20I and Omicron 21K show the widest TMRCA ranges, suggesting greater uncertainty near the base of these lineages.

For inference in Denmark, we trained a Denmark-specific STEPHY model using simulations tailored to the five Danish regions and their population sizes (**Supplementary Table 3**). The simulated outbreaks included five interconnected populations corresponding to Hovedstaden, Midtjylland, Syddanmark, Sjælland, and Nordjylland. In **Supplementary Fig. S9**, we show the performance of this Denmark-specific model using the same format as **Fig. 2**. Panels **a-d** summarize held-out test performance for R_0, γ, SSS, and index-location inference.



Supplementary Fig. S8. Phylogenetic uncertainty in TMRCA estimates across target lineages.



Supplementary Fig. S9. Performance of the Denmark-specific STEPHY model.

SI References

1. R. Bouckaert, *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).

2. T. G. Vaughan, ReMASTER: improved phylodynamic simulation for BEAST 2.7. *Bioinformatics* **40**, btae015 (2024).

3. L. Lyu, *et al.*, Tracing SARS-CoV-2 clusters across local scales using genomic data. *Proc. Natl. Acad. Sci.* **122**, e2501435122 (2025).

4. L. Lyu, *et al.*, Characterizing spatial epidemiology in a heterogeneous transmission landscape using the spatial transmission count statistic. *Commun. Med.* **5**, 165 (2025).