



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

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

This paper introduces PhyloDeep, a deep learning tool for predicting phylodynamic parameters under three birth-death models. The paper introduces machine learning models based on two different representations of timed phylogenies: (1) a summary statistic based representation, which represents timed phylogenies based on extended feature set first described by Saulnier et al., and (2) a vector based representation that uniquely describes timed phylogenies, which are rooted, binary, edge-labeled trees. On simulated data of modest size (of 50-200 and 200-500 leaves), the paper demonstrates that (1) the two representations have equivalent performance for all birth-death models, (2) have similar performance as BEAST2 for the simplest model, (3) but better performance than BEAST2 for the more complicated birth-death models. Finally, the paper illustrates a use case of PhyloDeep on a HIV datasets with 200 leaves, showing that the predicted birth-death model parameters are more in line with another previous study than BEAST's. Overall, the paper is well written and thorough, and will be of interest to the readership of this journal. I have a couple of comments that could strengthen the manuscript.

1. How do learned models generalize?

I would like to know how well learned models generalize to other smaller datasets. For instance, does the learned model for the larger instances (200-500 tips) perform well on the smaller instances (50-200 tips)?

2. Guidance on which representation to use?

Related to above, I'd like to see a more thorough investigation of the differences between the two representations? Can you find scenarios where one representation would be preferable over the other? Does one take longer to train? Is there a difference in generalizability?

3. What is the impact of the additional features added to the SS representation.

The authors introduced new summary statistics. I'd like to know more about the performance with and without these additional features. Also, how well does a joint representation perform, where you combine the two current representations (based on summary statistics, and based on tree topology/branch lengths).

4. Compute likelihood of solutions identified by PhyloDeep

To better understand where the improvement of performance relative to BEAST2 comes from, it would be good to evaluate the likelihood of the solutions identified by PhyloDeep. Do they have larger likelihood than BEAST2 solutions?

5. Additional real dataset. Preferably a large-scale one.

Finally, I think the impact of the method/paper can be largely increased if you would consider an additional large-scale dataset.

Minor comments:

\* More emphasis on PhyloDeep

The first mention of PhyloDeep occurs very late in the paper, almost like an afterthought. I think it should be featured more prominently, e.g. in the Abstract and Introduction, especially if you want this to be a tool to be used by the community.

\* Line 452: maximal => maximum or largest

Reviewer #2:  
Remarks to the Author:  
Review

Deep learning from phylogenies to uncover the transmission dynamics of epidemics

By Voznica et al.

The authors propose an original likelihood-free, simulation-based approach grounded on deep learning to infer the dynamics of epidemics from genetic data. They compare different versions of their approach to existing approaches currently used in the domain. They apply the method to an existing real data set dealing with the HIV epidemic in Zurich, which was already analyzed and leads the authors to refine the knowledge about the determinants of HIV transmission in Zurich. The short discussion gives several perspectives for extending the application domain of their approach. Codes implementing the approach are provided within the Python package PhyloDeep (note that I have not tested the codes).

This study is particularly well designed and many aspects are explored (most of the interrogations that I had during the reading of the main text are actually treated in the supplementary material). The main text is clear and, as I write above, generally adequately complemented by supporting information. The methods are well described and relevant.

The following concerns more specific points.

The title, DEEP LEARNING FROM PHYLOGENIES TO UNCOVER THE TRANSMISSION DYNAMICS OF EPIDEMICS, induces a confusion since the inference of the transmission dynamics may refer to the estimation of 'who infected whom', whereas the authors' objective is upstream: selecting a transmission model and estimating its parameters. Hence, I wonder if a title like DEEP LEARNING FROM PHYLOGENIES TO INFER THE DETERMINANTS OF DISEASE TRANSMISSION DYNAMICS (or something approaching) would be more adequate.

To facilitate the reading across this rich piece of work, supplementary figures and tables should be ordered as they appear in the main text.

After reading the introduction, I was unsure whether you were embedding your method in the framework of ABC or not. If I understood correctly, you do not, and you simply mention ABC, and more specifically the paper by Saulnier et al., because you recycle summary statistics that were proposed by Saulnier and her colleagues. At l.68, the term 'rejection-free' puts the reader on the track that you do not develop an ABC approach, but you should make it more explicit at the transition from the paragraph about the Saulnier's paper (l.60-67) to the next paragraph (l.68-80).

l.122: The authors should more exactly specify here what they call the 'sampling probability' and what means 'known'. This is clear later in the paper, but is ambiguous at this stage of the paper.

l.196-197: The convergence issue only concerns BEAST2, doesn't it?

In the discussion, the authors should evoke the question about how their approaches scale up with larger data sets than those considered in the application and simulations, with larger trees, lower sampling probability, and models with more parameters. In particular, is there a need for a much larger number of simulations (than 4M) for training the deep learning tools that the authors proposed to use in these cases? Supp. Fig. 3 and 4 partly tackle this question and, if my interpretation is correct, SF3 states that the performance is relatively stable in terms of model selection when the tree size increases, and SF4 states that the performance in terms of parameter estimation accuracy increases with the number of simulations. The authors could make a synthesis of this type of results in the discussion and extrapolate (to some extents) for answering the other dimensions of the above-mentioned question.

Fig. 4 and Supp. Fig. 8: the authors interestingly show that observed summary statistics for HIV are within the 'simulated envelope' of summary statistics throughout an analysis of the first two axes of a PCA. It would be interesting as well to perform the a priori check for the row summary statistics (without PCA). Since there are many summary statistics, the authors could provide a relatively concise table indicating, for each SS (i.e., marginally), to which quantile the observed value corresponds. This table could also be summarized into an histogram providing the distribution of the afore-mentioned quantiles. The table and the histogram would more precisely indicate how the class of used models represent real data.

Supp. Table 1 should include the information given at l.823-825 that a different prior is used for the infectious period in the numerical experiment and in the application. Looking at Fig. 4, it seems to me that a different prior is also used for  $X_{\{SS\}}$ , but I am maybe wrong and I maybe missed this information in the text.

Samuel Soubeyrand, INRAE, BioSP.

April 22, 2022

Dear R ,

We would like to thank you for your comments on our manuscript "Deep learning from phylogenies to uncover the transmission dynamics of epidemics", submitted to Nature Communications. We have uploaded a revised version to the journal's website. We apologize for the delay in reviving this manuscript. It was a complicated time for all of us with Covid-19 and Jakub Voznica (first author) moving to another institution after his thesis defence.

Your comments helped us to improve the methods, the PhyloDeep software and the original manuscript. Following the comments of referee 1, we have considerably extended the range of application of our neural networks, making them capable of analysing very large phylogenies in a few minutes, thanks to a novel decomposition of large pathogen phylogenies into sub-epidemics (sub-trees). We also assessed the generalization capabilities and the likelihood performance of our approach. Following the comments of referee 2, the discussion has been completed and we changed the title to "Deep learning from phylogenies to infer the epidemiological dynamics of outbreaks", which is clearer.

We are confident that the new version is much improved thanks to your comments, all of which have been taken into account. In what follows, you will find our point-by-point responses/changes, as well as a highlighted version of the main manuscript where all changes are marked in blue. We have also uploaded highlighted versions of the Methods and Supplementary Information onto the journal website.

We look forward to any further comments you may have.

Sincerely, the authors

## REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This paper introduces PhyloDeep, a deep learning tool for predicting phylodynamic parameters under three birth-death models. The paper introduces machine learning models based on two different representations of timed phylogenies: (1) a summary statistic based representation, which represents timed phylogenies based on extended feature set first described by Saulnier et al., and (2) a vector based representation that uniquely describes timed phylogenies, which are rooted, binary, edge-labeled trees. On simulated data of modest size (of 50-200 and 200-500 leaves), the paper demonstrates that (1) the two representations have equivalent performance for all birth-death models, (2) have similar performance as BEAST2 for the simplest model, (3) but better performance than BEAST2 for the more complicated birth-death models. Finally, the paper illustrates a use case of PhyloDeep on a HIV datasets with 200 leaves, showing that the predicted birth-death model parameters are more in line with another previous study than BEAST's. Overall, the paper is well written and thorough, and will be of interest to the readership of this journal. I have a couple of comments that could strengthen the manuscript.

1. How do learned models generalize?

I would like to know how well learned models generalize to other smaller datasets. For instance, does the learned model for the larger instances (200-500 tips) perform well on the smaller instances (50-200 tips)?

In statistical learning theory [31], generalization relates to the ability to predict new samples drawn from the same distribution as the training instances. Generalization is opposed to rote learning and overfitting, where the learned classifier or regressor predicts the training instances accurately, but new instances extracted from the same distribution or population poorly. The generalization capability of our NNs was extensively assessed in the submitted version of the manuscript, using large, independent testing sets (Fig. 3).

However, we agree with the referee that extending the study to samples that differ from the training distribution is clearly of interest in phylodynamics, in particular when the input tree is smaller than the training trees (as he/she suggested), but also, most importantly, when the input tree is larger than the training trees. We added results along this line. To summarize:

- We estimated the parameters of small trees (50-199 tips) using NNs trained with large trees (200-500 tips), and vice versa the parameters of large trees with NNs trained with small trees. The results (Supplementary Fig. 4) were surprisingly good from a machine learning standpoint, as the testing trees clearly departed from the training distribution. In particular, the accuracy obtained with FFNN-SS (summary statistics) was not affected very much by this strong violation of standard machine learning assumptions, while the accuracy of CNN-CBLV (combinatorial tree representation) was impacted but remained relatively high.

- However, in these experiments all trees are still of moderate size ( $\leq 500$  tips), while very large trees will become increasingly common in the near future with viral pathogens (see the current epidemics...). We thus explored another use of our NNs (pages 7, 9-10, Fig. 4), where a 'huge' input tree (5,000 to 10,000 tips in our experiments) is first decomposed into a set of disjoint subtrees (50 to 500 tips), which cover most of the huge-tree branches. Then, we apply the NNs for predictions on each subtree and combine the results using weighted averages. The results are impressive, for both FFNN-SS and CNN-CBLV. The prediction requires  $\sim 1$  CPU minute and the accuracy obtained with these huge trees is clearly higher than the one obtained with large trees (200-500 tips), with an error drop of a factor of 2 to 3 (Fig. 4). When applying this decomposition method to the prediction of large trees using NNs trained with small trees, the error became nearly identical to the error obtained with the right NNs (Supplementary Fig. 4).
- We believe that this capacity of NNs, made possible by their predictive speed, opens the way to many applications, which cannot be addressed today by any existing method. In particular, it is now possible to analyse extremely large phylogenies, and the approach could be used to track the evolution of parameters (e.g.  $R_0$ ) in different regions (sub-trees) of a global tree, as a function of dates (as in Bayesian skyline models), geographical areas, viral variants etc. This new decomposition approach and the corresponding algorithm, named 'subtree picker', have been added to PhyloDeep and described in Methods (page 7).

## 2. Guidance on which representation to use?

Related to above, I'd like to see a more thorough investigation of the differences between the two representations? Can you find scenarios where one representation would be preferable over the other? Does one take longer to train? Is there a difference in generalizability?

Thank you for this point, we added the following subsection, addressing all these issues (see also above comments and changes regarding generalizability):

### ***SS is simpler, but CBLV has high potential for application to new models***

*FFNN-SS and CNN-CBLV show similar accuracy across all settings (Fig. 3, Supplementary Tab. 1-2), including when predicting huge trees from their subtrees (Fig. 4). The only exception is the prediction of large trees using NNs trained with small trees (Supplementary Fig. 4), where FFNN-SS is superior to CNN-CBLV, but this goes beyond the recommended use of the approach, as only a part of the (large) query tree is given to the (small) CNN-CBLV.*

*However, the use of the two representations is clearly different, and it is likely that with new models and scenarios their accuracy will differ. SS requires a simpler architecture (FFNN) and is trained faster (e.g., 5 hours with large BDSS trees), with less training instances (Supplementary Fig. 6). However, this simplicity is obtained at cost of a long preliminary work to design appropriate summary statistics for each new model, as was confirmed in our analyses of BDSS simulations. To estimate the parameters of this model, we added summary statistics on*

*transmission chains on top of the SS taken from Saulnier et al. [19]. This improved the accuracy of superspreading fraction estimates of the FFNN-SS, so that it was comparable to the CNN-CBLV, while the accuracy for the other parameters remained similar (Supplementary Fig. 7). The advantage of the CBLV is its generality, meaning there is no loss of information between the tree and its representation in CBLV regardless of which model the tree was generated under. However, CBLV requires more complex architectures (CNN), more computing time in the learning phase (150 hours with large BDSS trees) and more training instances (Supplementary Fig. 6). Such an outcome is expected. With raw CBLV representation, the convolutional architecture is used to “discover” relevant summary statistics (or features, in machine learning terminology), which has a computational cost.*

*In fact, the two representations should not be opposed. An interesting direction for further research would be to combine them (e.g., during the FFNN phase), to possibly obtain even better results. Moreover, SS are still informative and useful (and quickly computed), in particular to perform sanity checks, both a priori and a posteriori (Fig. 5, Supplementary Fig. 8), or to quickly evaluate the predictability of new models and scenarios.*

3. What is the impact of the additional features added to the SS representation.

The authors introduced new summary statistics. I'd like to know more about the performance with and without these additional features.

In the revised version, the accuracy of all parameter estimates for BDSS is provided in Supplementary Fig. 7, with and without these additional features, and compared to CBLV. To summarize: the accuracy for the superspreading fraction (the most difficult parameter) is substantially improved with the new features and becomes similar to CNN-CBLV's, while the accuracy for the other parameters remains similar. Note, moreover, that the results for BD and BDEI were obtained with SS including these new features. All this is provided and explained in the revised version (see new subsection above and Supplementary Fig. 7).

Also, how well does a joint representation perform, where you combine the two current representations (based on summary statistics, and based on tree topology/branch lengths).

Combining both representations is certainly an interesting direction for further research. However, this imposes more complex NN architectures; for example, to incorporate the SS in the FFNN phase, after CNN and feature extraction from CBLV. Note, moreover, that the predictions of both approaches are highly correlated (close to 1 for most parameters of the three models), meaning that there is likely little room for improvement. Thus, we decided to leave this research direction for future works, and to give some indications in the Discussion (page 15).



#### 4. Compute likelihood of solutions identified by PhyloDeep

To better understand where the improvement of performance relative to BEAST2 comes from, it would be good to evaluate the likelihood of the solutions identified by PhyloDeep. Do they have larger likelihood than BEAST2 solutions?

We fully understand this demand, but a problem is that computing the likelihood is generally difficult (if not impossible) in phylodynamics, hence the numerous ABC and likelihood-free methods.

However, with the simplest birth-death (BD) model we have a closed form solution to compute the likelihood function, and we applied Referee's suggestion to our 'large' dataset, where BEAST2 and our NNs have similar accuracy (Fig. 3). We also computed the likelihood for the 'true' parameter values used to simulate the trees, in order to have an independent and solid assessment of the performance of the various methods. If a given method tends to produce higher likelihood than the one obtained with the true parameters values, then it performs "well enough" in terms of likelihood optimization, as optimizing further should not result in higher accuracy. The results (Supplementary Tab. 3) were as follows: (i) all methods (BEAST2, FFNN-SS and CNN-CBLV) obtained higher likelihood values than those obtained with true parameter values for ~70% of the trees, with a significant average difference; (ii) the difference of likelihood values between BEAST2, FFNN-SS and CNN-CBLV was non-significant, which explains their similar accuracy. These results are remarkable, as the NNs do not explicitly optimize the likelihood function associated with the model but use a radically different simulation-based learning approach.

Applying the same to BDEI and BDSS turned out to be impossible, as we do not have closed form solutions, and BEAST2 does not converge for several datasets due to numerical issues in likelihood computation and possible local optima (Fig. 3). Using BEAST2, we were unable to compute the likelihood value of our estimates and the one of the true parameter values, for a large fraction of trees. However, for the partial results we obtained (not shown), the figure seems to be similar to that with BD: the NNs obtain highly likely solutions, with similar likelihood as BEAST2's (when it converges and produces reasonable estimates), and significantly higher likelihood than that of the true parameter values.

All this is explained and detailed in the manuscript (pages 8-9), Methods (page 18) and Supplementary Tab. 3.

#### 5. Additional real dataset. Preferably a large-scale one.

Finally, I think the impact of the method/paper can be largely increased if you would consider an additional large-scale dataset.

We agree with the referee on the importance of analysing large data sets and trees, as they are becoming increasingly common today. However, hardly any existing method can accurately

estimate phylodynamics models with trees having (say) >1,000 tips (see our difficulties with BEAST2 and trees with <500 tips). When we submitted the first version of the paper, we were not sure how our NNs could be applied to very large trees, especially with CBLV (with SS it is still possible to summarize big trees using a few dozens of well-chosen features, but with the possible risk of losing essential information). In the revised version, we proposed, implemented and evaluated a solution based on disjoint subtrees extraction, estimation and averaging (see above). To assess this novel approach, we decided to use 'huge' simulated trees (5,000 to 10,000 tips) rather than a real tree, where the actual value of the parameters is often questionable and subject to debate. Results (pages 7, 9-10, Fig. 4) are quite convincing, with remarkably fast and accurate inference (see above), meaning that this approach open the way for new applications of phylodynamics, which were just impossible before. We thank the referee for his/her suggestion, which prompted us to further developments and clearly improved the paper in our opinion.

Minor comments:

\* More emphasis on PhyloDeep

The first mention of PhyloDeep occurs very late in the paper, almost like an afterthought. I think it should be featured more prominently, e.g. in the Abstract and Introduction, especially if you want this to be a tool to be used by the community.

Done, in both Abstract and Introduction, we thank the referee for his/her suggestion.

\* Line 452: maximal => maximum or largest

Done.

---

## Reviewer #2 (Remarks to the Author):

Review

Deep learning from phylogenies to uncover the transmission dynamics of epidemics

By Voznica et al.

The authors propose an original likelihood-free, simulation-based approach grounded on deep learning to infer the dynamics of epidemics from genetic data. They compare different versions of their approach to existing approaches currently used in the domain. They apply the method to an existing real data set dealing with the HIV epidemic in Zurich, which was already analysed and leads the authors to refine the knowledge about the determinants of HIV transmission in Zurich. The short discussion gives several perspectives for extending the application domain of

their approach. Codes implementing the approach are provided within the Python package PhyloDeep (note that I have not tested the codes).

This study is particularly well designed and many aspects are explored (most of the interrogations that I had during the reading of the main text are actually treated in the supplementary material). The main text is clear and, as I write above, generally adequately complemented by supporting information. The methods are well described and relevant.

The following concerns more specific points.

The title, DEEP LEARNING FROM PHYLOGENIES TO UNCOVER THE TRANSMISSION DYNAMICS OF EPIDEMICS, induces a confusion since the inference of the transmission dynamics may refer to the estimation of 'who infected whom', whereas the authors' objective is upstream: selecting a transmission model and estimating its parameters. Hence, I wonder if a title like DEEP LEARNING FROM PHYLOGENIES TO INFER THE DETERMINANTS OF DISEASE TRANSMISSION DYNAMICS (or something approaching) would be more adequate.

Thank you for this point, we agree that "transmission" can be confusing and changed the title to:

DEEP LEARNING FROM PHYLOGENIES TO INFER THE EPIDEMIOLOGICAL DYNAMICS OF OUTBREAKS

To facilitate the reading across this rich piece of work, supplementary figures and tables should be ordered as they appear in the main text.

Done.

After reading the introduction, I was unsure whether you were embedding your method in the framework of ABC or not. If I understood correctly, you do not, and you simply mention ABC, and more specifically the paper by Saulnier et al., because you recycle summary statistics that were proposed by Saulnier and her colleagues. At l.68, the term 'rejection-free' puts the reader on the track that you do not develop an ABC approach, but you should make it more explicit at the transition from the paragraph about the Saulnier's paper (l.60-67) to the next paragraph (l.68-80).

This has been clarified. In fact, our approach is a continuation of regression-based ABC. We wrote (blue part is new; page 4):

*...To address this issue Saulnier et al. [19] developed a large set of summary statistics. In addition, they used a regression step to select the most relevant statistics and to correct for the discrepancy between the simulations retained in the rejection step and the analyzed phylogeny. They observed that the sensitivity to the rejection parameters were greatly attenuated thanks to regression (see also Blum et al. [20]).*

*Our work is a continuation of regression-based ABC, and aims at overcoming its main limitations. Using the approximation power of currently available neural network architectures, we propose a likelihood-free method relying on deep learning from millions of trees of varying size simulated within a broad range of parameter values. By doing so, we bypass the rejection step, which is both time consuming with large simulation sets, and sensitive to the choice of the distance function and summary statistics...*

I.122: The authors should more exactly specify here what they call the 'sampling probability' and what means 'known'. This is clear later in the paper, but is ambiguous at this stage of the paper.

This has been clarified (page 6).

I.196-197: The convergence issue only concerns BEAST2, doesn't it?

Yes, this has been clarified (page 8).

In the discussion, the authors should evoke the question about how their approaches scale up with larger data sets than those considered in the application and simulations, with larger trees, lower sampling probability, and models with more parameters. In particular, is there a need for a much larger number of simulations (than 4M) for training the deep learning tools that the authors proposed to use in these cases? Supp. Fig. 3 and 4 partly tackle this question and, if my interpretation is correct, SF3 states that the performance is relatively stable in terms of model selection when the tree size increases, and SF4 states that the performance in terms of parameter estimation accuracy increases with the number of simulations. The authors could make a synthesis of this type of results in the discussion and extrapolate (to some extents) for answering the other dimensions of the above-mentioned question.

Thank you for raising these points, which are addressed here and there, and summarized in the Discussion, where we added (pages 15-16):

*A key issue in both phylodynamics and machine learning applications is scalability. Our results show that very large phylogenies can be analysed very efficiently (~1 minute for 10,000 tips), with resulting estimates more accurate than with smaller trees (Fig. 4), as predicted by learning theory. Again, as expected, more complex models require more training instances, especially BDSS using CBLV (Supplementary Fig. 3), but the ratio remains reasonable, and it is likely that complex (but identifiable) models will be handled efficiently with manageable training sets. Surprisingly, we did not observe a substantial drop of the accuracy with lower sampling probabilities (results not shown). To analyse very large trees, we used a decomposition into smaller, disjoint subtrees. In fact, all our NNs were trained with trees of moderate size (<500 tips). Another approach would be to learn directly from large trees. This is an interesting direction for further research, but this poses several difficulties. The first is that we need to simulate these very large trees, and a large number of them (millions or more). Then, SS is the easiest representation to learn, but at the risk of losing essential information, which means that new summary statistics will likely be needed for sufficiently complete representation of very*

*large phylogenies. Similarly, with CBLV more complex NN architectures (e.g., with additional and larger kernels in the convolutional layers) will likely be needed, imposing larger training sets. Combining both representations (e.g., during the FFNN phase) is certainly an interesting direction for further research. Note, however, that the predictions of both approaches for the three models we studied are highly correlated (Pearson coefficient nearly equal to 1 for most parameters), which means that there is likely little room for improvement (at least with these models).*

Fig. 4 and Supp. Fig. 8: the authors interestingly show that observed summary statistics for HIV are within the `simulated envelope' of summary statistics throughout an analysis of the first two axes of a PCA. It would be interesting as well to perform the a priori check for the row summary statistics (without PCA). Since there are many summary statistics, the authors could provide a relatively concise table indicating, for each SS (i.e., marginally), to which quantile the observed value corresponds. This table could also be summarized into an histogram providing the distribution of the afore-mentioned quantiles. The table and the histogram would more precisely indicate how the class of used models represent real data.

Thanks for the suggestion; we have added this functionality to PhyloDeep (page 19 in Methods). The SSs of the input tree are provided to the user, along with the corresponding [min, max] values in our simulations. With the HIV dataset, some SSs rejected the BD and BDEI models, which consistently have probability 0 in model selection (Supplementary Tab. 5 and Fig. 8).

Supp. Table 1 should include the information given at l.823-825 that a different prior is used for the infectious period in the numerical experiment and in the application. Looking at Fig. 4, it seems to me that a different prior is also used for  $X_{\{SS\}}$ , but I am maybe wrong and I maybe missed this information in the text.

In Fig. 5 (previously Fig. 4) we display the posterior distributions of the parameters, not the priors which were the same as in the simulations (in fact, our "priors" correspond to the simulation parameters, displayed in Supplementary Table 1). This has been clarified in the figure legend.

1 **DEEP LEARNING FROM PHYLOGENIES TO UNCOVER**  
2 **THE EPIDEMIOLOGICAL DYNAMICS OF OUTBREAKS**

3

4 **AUTHORS**

5 Voznica J<sup>1,2,3\*</sup>, Zhukova A<sup>1,4,5,6\*</sup>, Boskova V<sup>7</sup>, Saulnier E<sup>1</sup>, Lemoine F<sup>1,4</sup>, Moslonka-Lefebvre M<sup>1</sup>, Gascuel O<sup>1,8\*</sup>

6

7 **AFFILIATIONS**

8 <sup>1</sup> Institut Pasteur, Université Paris Cité, Unité Bioinformatique Evolutive, Paris, FRANCE

9 <sup>2</sup> Université de Paris, Paris, FRANCE

10 <sup>3</sup> Institut de Biologie de l'École Normale Supérieure, Ecole Normale Supérieure, CNRS, INSERM, Université Paris

11 Sciences et Lettres, Paris, FRANCE

12 <sup>4</sup> Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, FRANCE

13 <sup>5</sup> Institut Pasteur, Université Paris Cité, Epidemiology and Modelling of Antibiotic Evasion, Paris, FRANCE

14 <sup>6</sup> Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif, FRANCE

15 <sup>7</sup> Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of

16 Vienna, Vienna, AUSTRIA

17 <sup>8</sup> Institut de Systématique, Evolution, Biodiversité (UMR 7205 - CNRS, Muséum National d'Histoire Naturelle, SU,

18 EPHE, UA), Paris, FRANCE

19 **\* CO-CORRESPONDING AUTHORS**

20 jakub.voznica@pasteur.fr (JV), anna.zhukova@pasteur.fr (AZ), olivier.gascuel@mnhn.fr (OG)

21 **ABSTRACT**

22 Widely applicable, accurate and fast inference methods in phylodynamics are needed to fully profit from the richness  
23 of genetic data in uncovering the dynamics of epidemics. Standard methods, including maximum-likelihood and  
24 Bayesian approaches, generally rely on complex mathematical formulae and approximations, and do not scale with  
25 dataset size. We develop a likelihood-free, simulation-based approach, which combines deep learning with (1) a large  
26 set of summary statistics measured on phylogenies or (2) a complete and compact representation of trees, which avoids  
27 potential limitations of summary statistics and applies to any phylodynamics model. Our method enables both model  
28 selection and estimation of epidemiological parameters from very large phylogenies. We demonstrate its speed and  
29 accuracy on simulated data, where it performs better than the state-of-the-art methods. To illustrate its applicability,  
30 we assess the dynamics induced by superspreading individuals in an HIV dataset of men-having-sex-with-men in  
31 Zurich. Our tool *PhyloDeep* is available on [github.com/evolbioinfo/phylodeep](https://github.com/evolbioinfo/phylodeep).

32 **KEYWORDS**

33 Phylodynamics; molecular epidemiology; tree representation; neural networks; HIV.

## 34 INTRODUCTION

35 Pathogen phylodynamics is a field combining phylogenetics and epidemiology<sup>[1]</sup>. Viral or bacterial samples from  
36 patients are sequenced and used to infer a phylogeny, which describes the pathogen's spread among patients. The tips  
37 of such phylogenies represent sampled pathogens, and the internal nodes transmission events. Moreover, transmission  
38 events can be dated and thereby provide hints on transmission patterns. Such information is extracted by phylodynamic  
39 methods to estimate epidemiological and population dynamic parameters<sup>[2-4]</sup>, assess the impact of population  
40 structure<sup>[2,5]</sup>, and reveal the origins of epidemics<sup>[6]</sup>.

41 Birth-death models<sup>[7]</sup> incorporate easily interpretable parameters common to standard infectious-disease  
42 epidemiology, such as basic reproduction number  $R_0$ , infectious period, *etc.* In contrast to the standard epidemiological  
43 models, the birth-death models can be applied to estimate parameters from phylogenetic trees<sup>[8]</sup>. In these models,  
44 births represent transmission events, while deaths represent removal events for example due to treatment or recovery.  
45 Upon a patient's removal, their pathogens can be sampled, producing tips in the tree.

46 Here we focus on three specific, well-established birth-death models (**Fig. 1**): birth-death model (BD)<sup>[8,9]</sup>, birth-death  
47 model with exposed and infectious classes (BDEI)<sup>[5,10,11]</sup>, and birth-death model with superspreading (BDSS)<sup>[5,12]</sup>.  
48 These models were deployed using BEAST2<sup>[12,13]</sup> to study the phylodynamics of such diverse pathogens as Ebola  
49 virus<sup>[10]</sup>, Influenza virus<sup>[12]</sup>, Human Immunodeficiency Virus (HIV)<sup>[5]</sup>, Zika<sup>[14]</sup> or SARS-CoV-2<sup>[15]</sup>. Using these  
50 models, we will demonstrate the reliability of our deep learning-based approach.

51 While a great effort has been invested in the development of new epidemiological models in phylodynamics, the field  
52 has been slowed down by the mathematical complexity inherent to these models. BD, the simplest model, has a closed  
53 form solution for the likelihood formula of a tree for a given set of parameters<sup>[8,10]</sup>, but more complex models (*e.g.*,  
54 BDEI and BDSS) rely on a set of ordinary differential equations (ODEs) that cannot be solved analytically. To estimate  
55 parameter values through maximum-likelihood and Bayesian approaches, these ODEs must be approximated  
56 numerically for each tree node<sup>[5,10-12]</sup>. These calculations become difficult as the tree size increases, resulting in  
57 numerical instability and inaccuracy<sup>[12]</sup>, as we will see below.

58 Inference issues with complex models are typically overcome by approximate Bayesian computation (ABC)<sup>[16,17]</sup>.  
59 ABC is a simulation-based technique relying on a rejection algorithm<sup>[18]</sup>, where from a set of simulated phylogenies



60 within a given prior (values assumed for parameter values), those closest to the analysed phylogeny are retained and  
61 give the posterior distribution of the parameters. This scheme relies on the definition of a set of summary statistics  
62 aimed at representing a phylogeny and on a distance measure between trees. This approach is thus sensitive to the  
63 choice of the summary statistics and distance function (*e.g.*, Euclidean distance). To address this issue Saulnier *et*  
64 *al.*<sup>[19]</sup> developed a large set of summary statistics. In addition, they used a regression step to select the most relevant  
65 statistics and to correct for the discrepancy between the simulations retained in the rejection step and the analysed  
66 phylogeny. They observed that the sensitivity to the rejection parameters were greatly attenuated thanks to regression  
67 (see also Blum *et al.*<sup>[20]</sup>).

68 Our work is a continuation of regression-based ABC, and aims at overcoming its main limitations. Using the  
69 approximation power of currently available neural network architectures, we propose a likelihood-free method relying  
70 on deep learning from millions of trees of varying size simulated within a broad range of parameter values. By doing  
71 so, we bypass the rejection step, which is both time consuming with large simulation sets, and sensitive to the choice  
72 of the distance function and summary statistics. To describe simulated trees and use them as input for the deep learner,  
73 we develop two tree representations: (1) a large set of summary statistics mostly based on Saulnier *et al.*<sup>[19]</sup>, and (2) a  
74 complete and compact vectorial representation of phylogenies, including both the tree topology and branch lengths.  
75 The summary statistics are derived from our understanding and knowledge of the epidemiological processes.  
76 However, they can be incomplete and thus miss some important aspects of the studied phylogenies, which can  
77 potentially result in low accuracy during inference. Moreover, it is expected that new phylodynamic models will  
78 require design of new summary statistics, as confirmed by our results with BDSS. In contrast, our vectorial  
79 representation is a raw data representation that preserves all information contained in the phylogeny and thus should  
80 be accurate and deployable on any new model, provided the model parameters are identifiable. Our vectorial  
81 representation naturally fits with deep learning methods, especially the convolutional architectures, which have  
82 already proven their ability to extract relevant features from raw representations, for example in image analysis<sup>[21,22]</sup>  
83 or weather prediction<sup>[23]</sup>.

84 In the following, we introduce our vectorial tree representation and the new summary statistics designed for BDSS.  
85 We then present the deep learning architectures trained on these representations and evaluate their accuracy on  
86 simulated datasets in terms of both parameter estimation and model selection. We show that our approach applies not

87 only to trees of the same size as the training instances, but also to very large trees with thousands of tips through the  
88 analysis of their subtrees. The results are compared to those of the gold standard method, BEAST2<sup>[12,13]</sup>. Lastly, we  
89 showcase our methods on an HIV dataset<sup>[24,25]</sup> from the men-having-sex-with-men (MSM) community from Zurich.  
90 All technical details are provided in **Methods**. Our methods and tools are implemented in the PhyloDeep software,  
91 which is available on GitHub ([github.com/evolbioinfo/phylodeep](https://github.com/evolbioinfo/phylodeep)), PyPi ([pypi.org/project/phylodeep](https://pypi.org/project/phylodeep)) and Docker  
92 Hub ([hub.docker.com/r/evolbioinfo/phylodeep](https://hub.docker.com/r/evolbioinfo/phylodeep)).

## 93 RESULTS

94 Neural networks are trained on numerical vectors from which they can learn regression and classification tasks. We  
95 trained such networks on phylogenetic trees to estimate epidemiological parameters (regression) and select  
96 phylodynamic models (classification). We undertook two strategies for representing phylogenetic trees as numerical  
97 vectors, which we describe first, before showing the results with simulated and real data.

98 **Summary statistics (SS) representation.** We used a set of 83 SS developed by Saulnier *et al.*<sup>[19]</sup>: 26 measures of  
99 branch lengths, such as median of both internal and tip branch lengths; 8 measures of tree topology, such as tree  
100 imbalance; 9 measures on the number of lineages through time, such as time and height of its maximum; and 40  
101 coordinates representing the lineage-through-time (LTT) plot. To capture more information on the phylogenies  
102 generated by the BDSS model, we further enriched these SS with 14 new statistics on transmission chains describing  
103 the distribution of the duration between consecutive transmissions (internal tree nodes). Our SS are diverse,  
104 complementary and somewhat redundant. We used feed-forward neural networks (FFNN) with several hidden layers  
105 (**Fig. 2 b (i)**) that select and combine relevant information from the input features. In addition to SS, we provide both  
106 the tree size (*i.e.*, number of tips) and the sampling probability used to generate the tree, as input to our FFNN (**Fig. 2**  
107 **a (vi)**). We will refer to this method as FFNN-SS.

108 **Compact vectorial tree representation.** While converting raw information in the form of a phylogenetic tree into a  
109 set of SS, information loss is unavoidable. This means not only that the tree cannot be fully reconstructed from its SS,  
110 but also that depending on how much useful and relevant information is contained in the SS, the neural network may  
111 fail to solve the problem at hand. As an alternative strategy to SS, and to prevent information loss in the tree  
112 representation, we developed a representation called ‘Compact Bijective Ladderized Vector’ (CBLV).

113 Several vectorial representations of trees based either on polynomial<sup>[26,27]</sup>, Laplacian spectrum<sup>[28]</sup> or F matrices<sup>[29]</sup> have  
114 been developed previously. However, they represent the tree shape but not the branch lengths<sup>[26]</sup> or may lose  
115 information on trees<sup>[28]</sup>. In addition, some of these representations require vectors or matrices of quadratic size with  
116 respect to the number of tips<sup>[29]</sup>, or are based on complex coordinate systems of exponential size<sup>[27]</sup>.

117 Inspired by these approaches, we designed our concise, easily computable, compact, and bijective (i.e. 1-to-1) tree  
118 representation that applies to trees of variable size and is appropriate as machine learning input. To obtain this  
119 representation, we first ladderize the tree, that is, for each internal node, the descending subtree containing the most  
120 recently sampled tip is rotated to the left, **Fig. 2 a (ii)**. This ladderization step does not change the tree but facilitates  
121 learning by standardizing the input data. Moreover, it is consistent with trees observed in real epidemiological datasets,  
122 for example Influenza, where ladder-like trees reflect selection and are observed for several pathogens<sup>[1]</sup>. Then, we  
123 perform an inorder traversal<sup>[30]</sup> of the ladderized tree, during which we collect in a vector for each visited internal  
124 node its distance to the root and for each tip its distance to the previously visited internal node. In particular, the first  
125 vector entry corresponds to the tree height. This transformation of a tree into a vector is bijective, in the sense that we  
126 can unambiguously reconstruct any given tree from its vector representation (**Supplementary Fig. 1**). The vector is  
127 as compact as possible, and its size grows linearly with the number of tips. We complete this vector with zeros to  
128 reach the representation length of the largest tree contained in our simulation set, and [we add the sampling probability](#)  
129 [used to generate the tree \(or an estimate of it when analysing real data; Fig. 2 a \(v\), b \(i\)\)](#).

130 Bijectivity combined with ladderization facilitates the training of neural networks, which do not need to learn that  
131 different representations correspond to the same tree. However, unlike our SS, this full representation does not have  
132 any high-level features. In CBLV identical subtrees will have the same representation in the vector whenever the roots  
133 of these subtrees have the same height, while the vector representation of the tips in such subtrees will be the same no  
134 matter the height of the subtree's root. Similar subtrees will thus result in repeated patterns along the representation  
135 vector. We opted for Convolutional Neural Networks (CNN), which are designed to extract information on patterns  
136 in raw data. Our CNN architecture (**Fig. 2 b (ii)**) includes several convolutional layers that perform feature extraction,  
137 as well as maximum and average pooling layers that select relevant features and keep feature maps of reasonable  
138 dimensions. The output of the CNN is then fed into a FFNN that combines the patterns found in the input to perform  
139 predictions. In the rest of the manuscript, we refer to this method as CNN-CBLV.

## 140 **Simulated datasets**

141 For each phylodynamic model (BD, BDEI, BDSS), we simulated 4 million trees, covering a large range of values for  
142 each parameter of epidemiological interest ( $R_0$ , infectious period:  $1/\gamma$ , incubation period:  $1/\epsilon$ , the fraction at  
143 equilibrium of superspreading individuals:  $f_{SS}$ , and the superspreading transmission ratio:  $X_{SS}$ ). Of the 4 million trees,  
144 3.99 million were used as a training set, and 10,000 as a validation set for early stopping in the training phase<sup>[31]</sup>.  
145 Additionally, we simulated another 10,000 trees, which we used as a testing set, out of which 100 were also evaluated  
146 with the gold standard methods, BEAST2 and TreePar, which are more time consuming. Another 1 million trees were  
147 used to define confidence intervals for estimated parameters. For BD and BDEI we considered two settings: one with  
148 small trees (50 to 199 tips, in **Supplementary Fig. 2**) and a second with large trees (200 to 500 tips, **Fig. 3**). For  
149 BDSS, we considered only the setting with large trees, as the superspreading individuals are at a low fraction and  
150 cannot be detected in small trees (results not shown). *Lastly, we investigated the applicability of our approach to very*  
151 *large data sets, which are increasingly common with viral pathogens. To this goal, we generated for each model 10,000*  
152 *‘huge’ trees, with 5,000 to 10,000 tips each and with the same parameter ranges as used with the small and large trees.*  
153 *To estimate the parameter values of a huge tree, we extracted a nearly complete coverage of this tree by disjoint*  
154 *subtrees with 50 to 500 leaves. Then, we predicted the parameter values for every subtree using our NNs, and averaged*  
155 *subtree predictions to obtain parameter estimates for the huge tree.*

156 To increase the generality of our approach and avoid the arbitrary choice of the time scale (one unit can be a day, a  
157 week, or a year), we rescaled all trees and corresponding epidemiological parameters, such that the average branch  
158 length in a tree was equal to 1. After inference, we rescaled the estimated parameter values back to the original time  
159 scale.

## 160 **Neural networks yield more accurate parameter estimates than gold standard methods**

161 We compared accuracy of parameter estimates yielded by our deep learning methods and those yielded by two state-  
162 of-the-art phylodynamics inference tools, BEAST2<sup>[12,13]</sup> and TreePar<sup>[5]</sup>. The comparison shows that our deep learning  
163 methods trained with SS and CBLV are either comparable (BD) or more accurate (BDEI and BDSS) than the state-  
164 of-the-art inference methods (**Fig. 3, Supplementary Tab. 1**). The simple BD model has a closed form solution for  
165 the likelihood function, and thus BEAST2 results are optimal in theory<sup>[8,9]</sup>. Our results with BD are similar to those  
166 obtained with BEAST2, and thus nearly optimal as well. For BDEI and BDSS our results are more accurate than

167 BEAST2, which is likely explained by numerical approximations of likelihood calculations in BEAST2<sup>[5,10,11]</sup> for  
168 these models. These approximations can lead BEAST2 to a lack of convergence (2% cases for BDEI and 15% cases  
169 for BDSS) or a convergence to local optima. We suspect BEAST2 of converging to local optima when it converged  
170 to values with high relative error ( $>1.0$ ; 8% cases for BDEI and 11% cases for BDSS, **Fig. 3 b-c**). Furthermore, our  
171 deep learning approaches showed a lower bias in parameter estimation than BEAST2 (**Supplementary Tab. 2**). As  
172 expected, both approaches, FFNN-SS and CNN-CBLV, get more accurate with larger trees (**Supplementary Fig. 3**).

173 We tried to perform maximum likelihood estimation (MLE) implemented in the TreePar package<sup>[5]</sup> on the same trees  
174 as well. While MLE under BD model on simulations yielded as accurate results as BEAST2, for more complex models  
175 it showed overflow and underflow issues (*i.e.*, reaching infinite values of likelihood) and yielded inaccurate results,  
176 such as more complex models (BDEI, BDSS) having lower likelihood than a simpler, nested one (BD) for a part of  
177 simulations (results not shown). These issues were more prominent for larger trees. TreePar developers confirmed  
178 these limitations and suggested using the latest version of BEAST2 instead.

179 To further explain the performance of our NNs, we computed the likelihood value of their parameter estimates. This  
180 was easy with the BD model since we have a closed form solution for the likelihood function. The results with this  
181 model (**Supplementary Tab. 3**, using TreePar) showed that the likelihoods of both FFNN-SS and CNN-CBLV  
182 estimates are similar to BEAST2's, which explains the similar accuracy of the three methods (**Fig. 3**). We also  
183 computed the likelihood of the 'true' parameter values used to simulate the trees, in order to have an independent and  
184 solid assessment. If a given method tends to produce higher likelihood than that of the true parameter values, then it  
185 performs well in terms of likelihood optimization, as optimizing further should not result in higher accuracy. The  
186 results (**Supplementary Tab. 3**) were again quite positive, as BEAST2 and our NNs achieved a higher likelihood  
187 than the true parameter values for ~70% of the trees, with a significant mean difference. With BDEI and BDSS,  
188 applying the same approach proved difficult due to convergence and numerical issues, with both BEAST2 and TreePar  
189 (see above). For the partial results we obtained (not shown), the overall pattern seems to be similar to that with BD:  
190 the NNs obtain highly likely solutions, with similar likelihood as BEAST2's (when it converges and produces  
191 reasonable estimates), and significantly higher likelihood than that of the true parameter values. All these results are  
192 remarkable, as the NNs do not explicitly optimize the likelihood function associated to the models, but use a radically  
193 different learning approach, based on simulation.

## 194 **Neural networks are fast inference methods**

195 We compared the computing time required by each of our inference methods. All computing times were estimated for  
196 a single thread of our cluster, except for the training of neural architectures where we used our GPU farm. Neural  
197 networks require heavy computing time in the learning phase; for example, with BDSS (the most complex model),  
198 simulating 4M large trees requires ~800 CPU hours, while training FFNN-SS and CNN-CBLV requires ~5 and ~150  
199 hours, respectively. However, with NNs, inference is almost instantaneous and takes ~0.2 CPU seconds per tree on  
200 average, including encoding the tree in SS or CBLV, which is the longest part. For comparison, BEAST2 inference  
201 under the BD model with 5 million MCMC steps takes on average ~0.2 CPU hours per tree, while inference under  
202 BDEI and BDSS with 10 million MCMC steps takes ~55 CPU hours and ~80 CPU hours per tree, respectively. In  
203 fact, the convergence time of BEAST2 is usually faster (~6 CPU hours with BDEI and BDSS), but can be very long  
204 in some cases, to the point that convergence is not observed after 10 million steps (see above).

## 205 **Neural networks have high generalization capabilities and apply to very large data sets**

206 In statistical learning theory<sup>[31]</sup>, generalization relates to the ability to predict new samples drawn from the same  
207 distribution as the training instances. Generalization is opposed to rote learning and overfitting, where the learned  
208 classifier or regressor predicts the training instances accurately, but new instances extracted from the same distribution  
209 or population poorly. The generalization capabilities of our NNs were demonstrated, as we used independent testing  
210 sets in all our experiments (**Fig. 3**). However, we expect poor results with trees that depart from the training  
211 distribution, for example showing very high  $R_0$ , while our NNs have been trained with  $R_0$  in the range [1, 5]. If, for a  
212 new study, larger or different parameter ranges are required, we must retrain the NNs with *ad hoc* simulated trees.  
213 However, a strength of NNs is that thanks to their flexibility and approximation power, very large parameter ranges  
214 can be envisaged, to avoid repeating training sessions too often.

215 Another sensible issue is that of the size of the trees. Our NNs have been trained with trees of 50-to-199 tips (small)  
216 and 200-to-500 tips (large), that is, trees of moderate size (but already highly time consuming in a Bayesian setting,  
217 for the largest ones). Thus, we tested the ability to predict the parameters of small trees using NNs trained on large  
218 trees, and vice versa, the ability to predict large trees with NNs trained on small trees. The results (**Supplementary**  
219 **Fig. 4**) are surprisingly good, especially with summary statistics (FFNN-SS) which are little impacted by these changes  
220 of scale as they largely rely on means (*e.g.*, of branch lengths<sup>[19]</sup>). This shows unexpected generalization capabilities

221 of the approach regarding tree size. Most importantly, the approach can accurately predict huge trees (**Fig. 4**) using  
222 their subtrees and the means of the corresponding parameter estimates, in ~1 CPU minute. This extends the  
223 applicability of the approach to data sets that cannot be analysed today, unless using similar tree decomposition and  
224 very long calculations to analyse all subtrees.

#### 225 **Neural networks are accurate methods for model selection**

226 We trained CNN-CBLV and FFNN-SS on simulated trees to predict the birth-death model under which they were  
227 simulated (BD or BDEI for small trees; BD, BDEI or BDSS for large trees). Note that for parameters shared between  
228 multiple models, we used identical parameter value ranges across all these models (**Supplementary Tab. 4**). Then,  
229 we assessed the accuracy of both of our approaches on 100 simulations obtained with each model and compared it  
230 with the model selection under BEAST2 based on Akaike information criterion through Markov Chain Monte Carlo  
231 (AICM)<sup>[32,33]</sup>. The AICM, similar to deviance information criterion (DIC) by Gelman *et al.*<sup>[32]</sup>, does not add  
232 computational load and is based on the average and variance of posterior log-likelihoods along the Markov Chain  
233 Monte Carlo (MCMC).

234 FFNN-SS and CNN-CBLV have similar accuracy (**Supplementary Tab. 5**), namely 92% for large trees (BD vs BDEI  
235 vs BDSS), and accuracy of 91% and 90%, respectively, for small trees (BD vs BDEI). BEAST2 yielded an accuracy  
236 of 91% for large trees and 87% for small trees. The non-converging simulations were not considered for any of these  
237 methods (*i.e.*, 5% simulations for small trees and 24% for large trees).

238 The process of model selection with a neural network is as fast as the parameter inference (~0.2 CPU seconds per  
239 tree). This represents a practical, fast and accurate way to perform model selection in phylodynamics.

#### 240 **Neural networks are well suited to learn complex models**

241 To assess the complexity of learned models, we explored other inference methods, namely: (1) linear regression as a  
242 baseline model trained on summary statistics (LR-SS); (2) FFNN trained directly on CBLV (FFNN-CBLV); (3) CNN  
243 trained on Compact Random Vector (CNN-CRV), for which the trees were randomly rotated, instead of being  
244 ladderized as in **Fig. 2 (ii)**; and (4) two “null models”.

245 LR-SS yielded inaccurate results even for the BD model (**Supplementary Tab. 1**), which seems to contrast with  
246 previous findings<sup>[19]</sup>, where LR approach combined with ABC performed only slightly worse than BEAST2. This can

247 be explained by the lack of rejection step in LR-SS, which enables to locally reduce the complexity of the relation  
248 between the representation and the inferred values to a linear one<sup>[18]</sup>. However, the rejection step requires a metric  
249 (*e.g.*, the Euclidean distance), which may or may not be appropriate depending on the model and the summary  
250 statistics. Moreover, rejection has a high computational cost with large simulation sets.

251 Neural networks circumvent these problems with rejection and allow for more complex, non-linear relationships  
252 between the tree representation and the inferred values to be captured. This is also reflected in our results with FFNN-  
253 CBLV and CNN-CRV, which both proved to be generally more accurate than LR-SS. However, FFNN-CBLV was  
254 substantially less accurate than CNN-CBLV (**Supplementary Tab. 1, Supplementary Fig. 5**). This indicates the  
255 presence of repeated patterns that may appear all along the vectorial representation of trees, such as subtrees of any  
256 size, which are better extracted by CNN than by FFNN. In its turn, CNN-CRV required larger training sets to reach  
257 an accuracy comparable to CNN-CBLV (**Supplementary Tab. 1, Supplementary Fig. 5**), showing that the  
258 ladderization and bijectivity of the CBLV helped the training.

259 To assess how much information is actually learned, we also measured the accuracy of two “null models”: FFNN  
260 trained to predict randomly permuted target values; and a random predictor, where parameter values were sampled  
261 from prior distributions. Results show that the neural networks extract a considerable amount of information for most  
262 of the estimated parameters (**Supplementary Tab. 1**). The most difficult parameter to estimate was the fraction of  
263 superspreading individuals in BDSS model, with accuracy close to random predictions with small trees, but better  
264 performance as the tree size increases (**Fig. 4, Supplementary Fig. 3**).

### 265 **SS is simpler, but CBLV has high potential for application to new models**

266 FFNN-SS and CNN-CBLV show similar accuracy across all settings (**Fig. 3, Supplementary Tab. 1-2**), including  
267 when predicting huge trees from their subtrees (**Fig. 4**). The only exception is the prediction of large trees using NNs  
268 trained with small trees (**Supplementary Fig. 4**), where FFNN-SS is superior to CNN-CBLV, but this goes beyond  
269 the recommended use of the approach, as only a part of the (large) query tree is given to the (small) CNN-CBLV.

270 However, the use of the two representations is clearly different, and it is likely that with new models and scenarios  
271 their accuracy will differ. SS requires a simpler architecture (FFNN) and is trained faster (*e.g.*, 5 hours with large  
272 BDSS trees), with less training instances (**Supplementary Fig. 6**). However, this simplicity is obtained at cost of a



273 long preliminary work to design appropriate summary statistics for each new model, as was confirmed in our analyses  
274 of BDSS simulations. To estimate the parameters of this model, we added summary statistics on transmission chains  
275 on top of the SS taken from Saulnier *et al.*<sup>[19]</sup>. This improved the accuracy of superspreading fraction estimates of the  
276 FFNN-SS, so that it was comparable to the CNN-CBLV, while the accuracy for the other parameters remained similar  
277 (**Supplementary Fig. 7**). The advantage of the CBLV is its generality, meaning there is no loss of information between  
278 the tree and its representation in CBLV regardless of which model the tree was generated under. However, CBLV  
279 requires more complex architectures (CNN), more computing time in the learning phase (150 hours with large BDSS  
280 trees) and more training instances (**Supplementary Fig. 6**). Such an outcome is expected. With raw CBLV  
281 representation, the convolutional architecture is used to “discover” relevant summary statistics (or features, in machine  
282 learning terminology), which has a computational cost.

283 In fact, the two representations should not be opposed. An interesting direction for further research would be to  
284 combine them (*e.g.* during the FFNN phase), to possibly obtain even better results. Moreover, SS are still informative  
285 and useful (and quickly computed), in particular to perform sanity checks, both a priori and a posteriori (**Fig. 5**,  
286 **Supplementary Fig. 8**), or to quickly evaluate the predictability of new models and scenarios.

### 287 **Showcase study of HIV in MSM subpopulation in Zurich**

288 The Swiss HIV Cohort is densely sampled, including more than 16,000 infected individuals<sup>[24]</sup>. Datasets extracted  
289 from this cohort have often been studied in phylodynamics<sup>[8,25]</sup>. We analysed a dataset of an MSM subpopulation from  
290 Zurich, which corresponds to a cluster of 200 sequences studied previously by Rasmussen *et al.*<sup>[25]</sup>, who focused on  
291 the degree of connectivity and its impact on transmission between infected individuals. Using coalescent approaches,  
292 they detected the presence of highly connected individuals at the beginning of the epidemic and estimated  $R_0$  to be  
293 between 1.0 and 2.5. We used their tree as input for neural networks and BEAST2.

294 To perform analyses, one needs an estimate of the sampling probability. We considered that: (1) the cohort is expected  
295 to include around 45% of Swiss individuals infected with HIV<sup>[24]</sup>; and (2) the sequences were collected from around  
296 56% of individuals enrolled in this cohort<sup>[34]</sup>. We used these percentages to obtain an approximation of sampling  
297 probability of  $0.45 \times 0.56 \sim 0.25$  and used this value to analyse the MSM cluster. To check the robustness of our  
298 estimates, we also used sampling probabilities of 0.2 and 0.3 in our estimation procedures.

299 First, we performed a quick sanity check considering the resemblance of HIV phylogeny with simulations obtained  
300 with each model. Two approaches were used, both based on SS (**Supplementary Fig. 8**). Using principal component  
301 analysis (PCA), all three considered birth-death models passed the check. However, when looking at the 97 SS values  
302 in detail, namely checking whether the observed tree SS were within the [min, max] range of the corresponding  
303 simulated values, the BD and BDEI models were rejected for some of the SS (5 for both models, all related to branch  
304 lengths). Then, we performed model selection (BD vs BDEI vs BDSS) and parameter estimation using our two  
305 methods and BEAST2 (**Fig. 5 a-b**). Finally, we checked the model adequacy with a second sanity check, derived from  
306 the inferred values and SS (**Fig. 5 c, Supplementary Fig. 8**).

307 Model selection with CNN-CBLV and FFNN-SS resulted in the acceptance of BDSS (probability of 1.00 versus 0.00  
308 for BD and BDEI), and the same result was obtained with BEAST2 and AICM. These results are consistent with our  
309 detailed sanity check, and with what is known about HIV epidemiology, namely, the presence of superspreading  
310 individuals in the infected subpopulation<sup>[35]</sup> and the absence of incubation period without infectiousness such as is  
311 emulated in BDEI<sup>[36]</sup>.

312 We then inferred parameter values under the selected BDSS model (**Fig. 5 a-b**). The values obtained with FFNN-SS  
313 and CNN-CBLV are close to each other, and the 95% CI are nearly identical. We inferred an  $R_0$  of 1.6 and 1.7, and  
314 an infectious period of 10.2 and 9.8 years, with FFNN-SS and CNN-CBLV, respectively. Transmission by  
315 superspreading individuals was estimated to be around 9 times higher than by normal spreaders and superspreading  
316 individuals were estimated to account for around 7-8% of the population. Our  $R_0$  estimates are consistent with the  
317 results of a previous study<sup>[8]</sup> performed on data from the Swiss cohort, and the results of Rasmussen *et al.*<sup>[25]</sup> with this  
318 dataset. The infectious period we inferred is a bit longer than that reported by Stadler *et al*, who estimated it to be 7.74  
319 [95% CI 4.39-10.99] years<sup>[8]</sup>. The infectious period is a multifactorial parameter depending on treatment efficacy and  
320 adherence, the times from infection to detection and to the start of treatment, *etc*. In contrast to the study by Stadler *et*  
321 *al*, whose data were sampled in the period between 1998 and 2008, our dataset covers also the period between 2008  
322 and 2014, during which life expectancy of patients with HIV was further extended<sup>[37]</sup>. This may explain why we found  
323 a longer infectious period (with compatible CIs). Lastly, our findings regarding superspreading are in accordance with  
324 those of Rasmussen *et al.*<sup>[25]</sup>, and with a similar study in Latvia<sup>[5]</sup> based on 40 MSM sequences analysed using a  
325 likelihood approach. Although the results of the latter study may not be very accurate due to the small dataset size,

326 they still agree with ours, giving an estimate of a superspreading transmission ratio of 9, and 5.6% of superspreading  
327 individuals. Our estimates were quite robust to the choice of sampling probability (*e.g.*,  $R_0 = 1.54, 1.60$  and  $1.66$ , with  
328 FFNN-SS and a sampling probability of  $0.20, 0.25$  and  $0.30$ , respectively, **Fig. 5 b**).

329 Compared to BEAST2, the estimates of the infectious period and  $R_0$  were similar for both approaches, but BEAST2  
330 estimates were higher for the transmission ratio ( $14.5$ ) and the superspreading fraction ( $10.6\%$ ). These values are in  
331 accordance with the positive bias of BEAST2 estimates that we observed in our simulation study for these two  
332 parameters, while our estimates were nearly unbiased (**Supplementary Tab. 2**).

333 Finally, we checked the adequacy of BDSS model by resemblance of HIV phylogeny to simulations. Using inferred  
334 95% CI, we simulated 10,000 trees and performed PCA on SS, to which we projected the SS of our HIV phylogeny.  
335 This was close to simulations, specifically close to the densest swarm of simulations, supporting the adequacy of both  
336 the inferred values and the selected model (**Fig. 5 c**). *When looking at the 97 SS in detail, some of the observed values  
337 where not in the [min, max] range of the 10,000 simulated values. However, these discordant SS were all related to  
338 the lineage-through-time plot (LTT; *e.g.*,  $x$  and  $y$  coordinates of this plot; **Supplementary Fig. 8**), consistent with the  
339 fact that the probabilistic, sampling component of the BDSS model is an oversimplification of actual sampling  
340 schemes, which depend on contact tracing, sampling campaigns and policies, etc.*

## 341 **DISCUSSION AND PERSPECTIVES**

342 In this manuscript, we presented new methods for parameter inference and model selection in phylodynamics based  
343 on deep learning from phylogenies. Through extensive simulations, we established that these methods are at least as  
344 accurate as state-of-the-art methods *and capable of predicting very large trees in minutes, which cannot be achieved  
345 today by any other existing method*. We also applied our deep learning methods to the Swiss HIV dataset from MSM  
346 and obtained results consistent with current knowledge of HIV epidemiology.

347 Using BEAST2, we obtained inaccurate results for some of the BDEI and BDSS simulations. While BEAST2 has  
348 been successfully deployed on many models and tasks, it clearly suffers from approximations in likelihood  
349 computation with these two models. However, these will likely improve in near future. In fact, we already witnessed  
350 substantial improvements done by BEAST2 developers to the BDSS model, while carrying out this research.

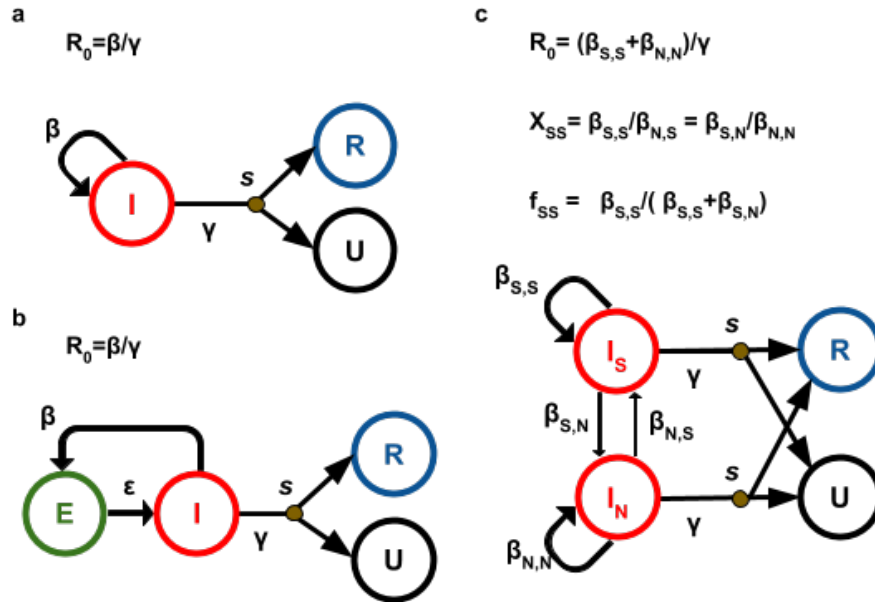
351 Both of our neural network approaches circumvent likelihood computation and thereby represent a new way of using  
352 molecular data in epidemiology, without the need to solve large systems of differential equations. This opens the door  
353 to novel phylodynamics models, which would make it possible to answer questions previously too complex to ask.  
354 This is especially true for CBLV representation, which does not require the design of new summary statistics, when  
355 applied to trees generated by new mathematical models. A direction for further research would be to explore such  
356 models, for example based on structured coalescent<sup>[38,39]</sup>, or to extend the approach to macroevolution and species  
357 diversification models<sup>[40]</sup>, which are closely related to epidemiological models. Other fields related to phylodynamics,  
358 such as population genetics, have been developing likelihood-free methods<sup>[41]</sup>, for which our approach might serve as  
359 a source of inspiration.

360 A key issue in both phylodynamics and machine learning applications is scalability. Our results show that very large  
361 phylogenies can be analysed very efficiently (~1 minute for 10,000 tips), with resulting estimates more accurate than  
362 with smaller trees (**Fig. 4**), as predicted by learning theory. Again, as expected, more complex models require more  
363 training instances, especially BDSS using CBLV (**Supplementary Fig. 3**), but the ratio remains reasonable, and it is  
364 likely that complex (but identifiable) models will be handled efficiently with manageable training sets. Surprisingly,  
365 we did not observe a substantial drop of accuracy with lower sampling probabilities (results not shown). To analyse  
366 very large trees, we used a decomposition into smaller, disjoint subtrees. In fact, all our NNs were trained with trees  
367 of moderate size (<500 tips). Another approach would be to learn directly from large trees. This is an interesting  
368 direction for further research, but this poses several difficulties. The first is that we need to simulate these very large  
369 trees, and a large number of them (millions or more). Then, SS is the easiest representation to learn, but at the risk of  
370 losing essential information, which means that new summary statistics will likely be needed for sufficiently complete  
371 representation of very large phylogenies. Similarly, with CBLV more complex NN architectures (*e.g.*, with additional  
372 and larger kernels in the convolutional layers) will likely be needed, imposing larger training sets. Combining both  
373 representations (*e.g.*, during the FFNN phase) is certainly an interesting direction for further research. Note, however,  
374 that the predictions of both approaches for the three models we studied are highly correlated (Pearson coefficient  
375 nearly equal to 1 for most parameters), which means that there is likely little room for improvement (at least with  
376 these models).

377 A key advantage of the deep learning approaches is that they yield close to immediate estimates and apply to trees of  
378 varying size. Collection of pathogen genetic data became standard in many countries, resulting in densely sampled  
379 infected populations. Examples of such datasets include HIV in Switzerland and UK<sup>[24,42]</sup>, 2013 Ebola epidemics<sup>[6]</sup>,  
380 several Influenza epidemics and the 2019 SARS-Cov-2 pandemic ([www.gisaid.org](http://www.gisaid.org))<sup>[43]</sup>. For many such pathogens,  
381 trees can be efficiently and accurately inferred<sup>[44-46]</sup> and dated<sup>[47-49]</sup> using standard approaches. When applied to such  
382 dated trees, our methods can perform model selection and provide accurate phylodynamic parameter estimates within  
383 a fraction of a second. Such properties are desirable for phylogeny-based real-time outbreak surveillance methods,  
384 which must be able to cope with the daily influx of new samples, and thus increasing size of phylogenies, as the  
385 epidemic unfolds, in order to study local outbreaks and clusters, and assess and compare the efficiency of healthcare  
386 policies deployed in parallel. *Moreover, thanks to the subtree picking and averaging strategy, it is now possible to*  
387 *analyse extremely large phylogenies, and the approach could be used to track the evolution of parameters (e.g.,  $R_0$ ) in*  
388 *different regions (sub-trees) of a global tree, as a function of dates (as in Bayesian skyline models<sup>[4]</sup>), geographical*  
389 *areas, viral variants etc.*

390 **ACKNOWLEDGEMENT:** We would like to thank Dr Kary Ocaña and Tristan Dot for initiating experiments on  
391 machine learning and phylogenetic trees in our laboratory. We would like to thank Quang Tru Huynh for  
392 administrating the GPU farm at Institut Pasteur and the INCEPTION program (Investissement d’Avenir grant ANR-  
393 16-CONV-0005) that financed the GPU farm. We would like to thank Dr Christophe Zimmer from Institut Pasteur,  
394 Sophia Lambert and Dr H el ene Morlon from Institut de Biologie de l’Ecole Normale Sup erieure IBENS and Dr Guy  
395 Baele from Katholieke Universiteit KU Leuven for useful discussions and Dr Isaac Overcast from IBENS for critical  
396 reading of the manuscript. We would like to thank Dr Tanja Stadler and J er emie Scir e for their help with BEAST2  
397 and MLE approaches. JV is supported by Ecole Normale Sup erieure Paris-Saclay and by ED Fronti eres de l’Innovation  
398 en Recherche et Education, Programme Bettencourt. VB would like to thank Swiss National Science Foundation for  
399 funding (Early PostDoc mobility grant P2EZP3\_184543). OG is supported by PRAIRIE (ANR-19-P3IA-0001).

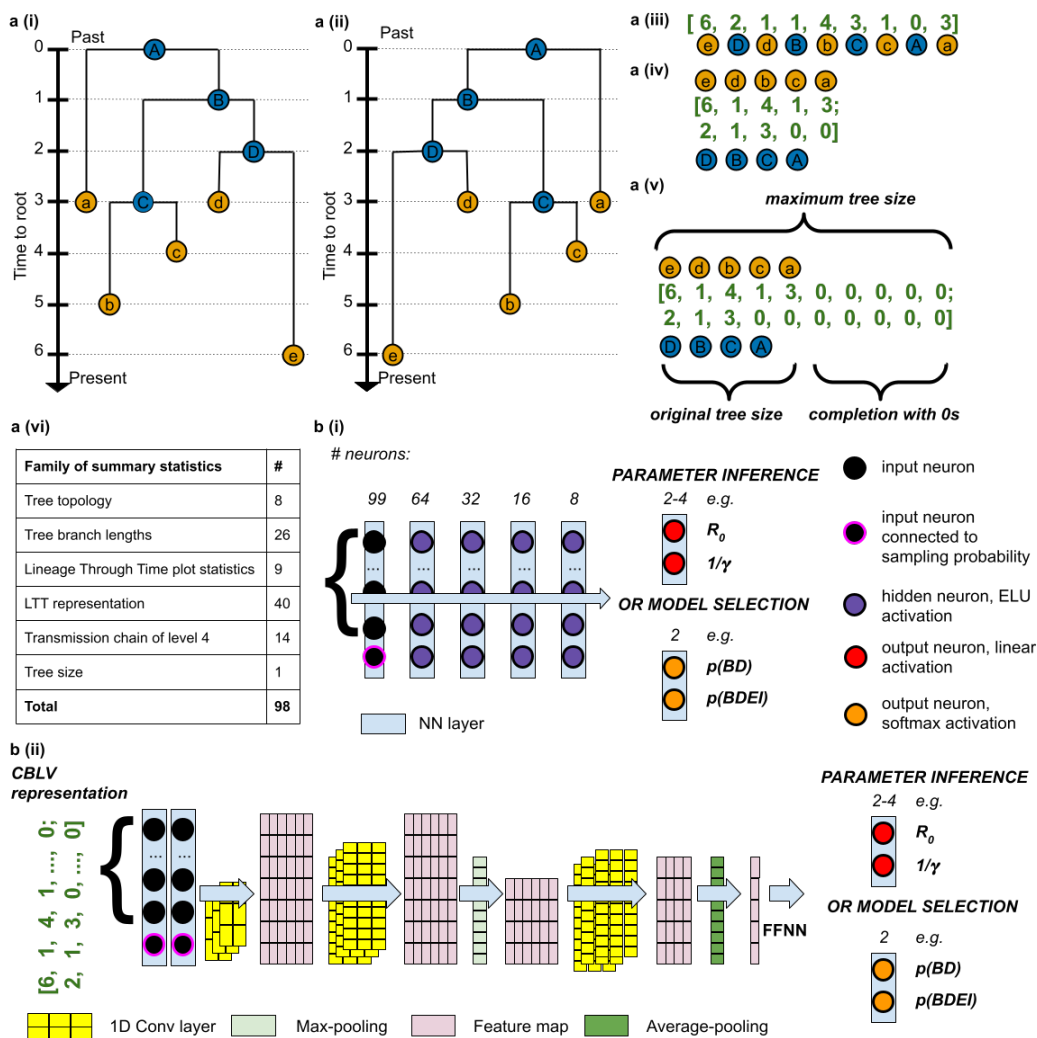
401 Fig. 1: Birth-death models



402

403 **Note to Fig. 1.** **a** Birth-death model (BD)<sup>[8,9]</sup>, **b**, birth-death model with Exposed-Infected individuals (BDEI)<sup>[5,10,11]</sup>  
 404 and **c**, birth-death model with SuperSpreading (BDSS)<sup>[5,12]</sup>. BD is the simplest generative model, used to estimate  $R_0$   
 405 and the infectious period  $(1/\gamma)$ <sup>[8,9]</sup>. BDEI and BDSS are extended version of BD. BDEI enables to estimate latency  
 406 period  $(1/\epsilon)$  during which individuals of exposed class E are infected, but not infectious<sup>[5,10,11]</sup>. BDSS includes two  
 407 populations with heterogeneous infectiousness: the so-called superspreading individuals (S) and normal spreaders (N).  
 408 Superspreading individuals are present only at a low fraction in the population ( $f_{ss}$ ) and may transmit the disease at a  
 409 rate that is multiple times higher than that of normal spreaders (rate ratio =  $X_{ss}$ )<sup>[5,12]</sup>. Superspreading can have various  
 410 complex causes, such as the heterogeneity of immune response, disease progression, co-infection with other diseases,  
 411 social contact patterns or risk behaviour, *etc.* Infectious individuals I (superspreading infectious individuals  $I_S$  and  
 412 normal spreaders  $I_N$  for BDSS), transmit the disease at rate  $\beta$  ( $\beta_{X,Y}$  for an individual of type X transmitting to an  
 413 individual of type Y for BDSS), giving rise to a newly infected individual. The newly infected individual is either  
 414 infectious right away in BD and BDSS or goes through an exposed state before becoming infectious at rate  $\epsilon$  in BDEI.  
 415 Infectious individuals are removed at rate  $\gamma$ . Upon removal, they can be sampled with probability  $s$ , becoming of  
 416 removed sampled class R. If not sampled upon removal, they move to non-infectious unsampled class U.

417 **Fig. 2: Pipeline for training neural networks on phylogenies**

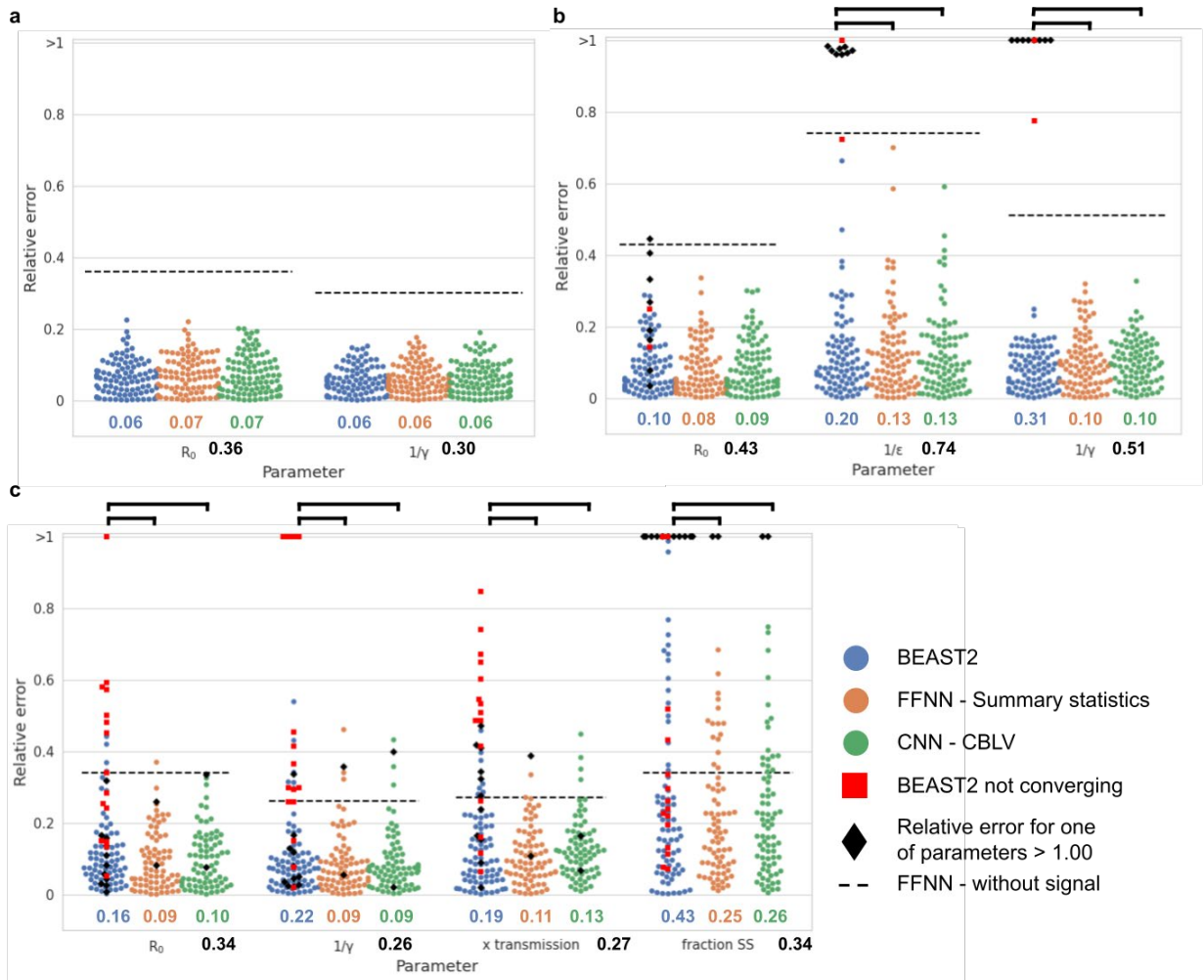


418

419 **Note to Fig.2. Tree representations:** **a (i)**, simulated binary trees. Under each model from Fig. 1, we simulate many  
 420 trees of variable size (50 to 200 tips for ‘small trees’ and 200 to 500 tips for ‘large trees’). For illustration, we have  
 421 here a tree with 5 tips. We encode the simulations into two representations, either **a (ii-v)**, in a complete and compact  
 422 tree representation called ‘Compact Bijective Ladderized Vector’ abbreviated as CBLV or **a (vi)** with summary  
 423 statistics (SS). CBLV is obtained through **a (ii)** ladderization or sorting of internal nodes so that the branch supporting  
 424 the most recent leaf is always on the left and **a (iii)** an inorder tree traversal, during which we append to a real-valued  
 425 vector for each visited internal node its distance to the root and for each visited tip its distance to the previously visited  
 426 internal node. We reshape this representation into **a (iv)**, an input matrix in which the information on internal nodes  
 427 and leaves is separated into two rows. Finally, **a (v)**, we complete this matrix with zeros so that the matrices for all  
 428 simulations have the size of largest simulation matrices. For illustration purpose, we here consider that the maximum  
 429 tree size covered by simulations is 10, and the representation is thus completed with 0s accordingly. SS consists of **a**  
 430 **(vi)**, a set of 98 statistics: 83 published in *Saulnier et al*<sup>[19]</sup>, 14 on transmission chains and 1 on tree size. The  
 431 information on sampling probability is added to both representations. **b: Neural networks** are trained on these  
 432 representations to estimate parameter values or to select the underlying model. For SS, we use, **b (i)**, a deep feed-  
 433 forward neural network (FFNN) of funnel shape (we show the number of neurons above each layer). For the CBLV  
 434 representation we train, **b (ii)**, Convolutional Neural Networks (CNN). The CNN is added on top of the FFNN. The  
 435 CNN combines convolutional, maximum pooling and global average pooling layers, as described in detail in **Methods**.



436 **Fig. 3: Assessment of deep learning accuracy**



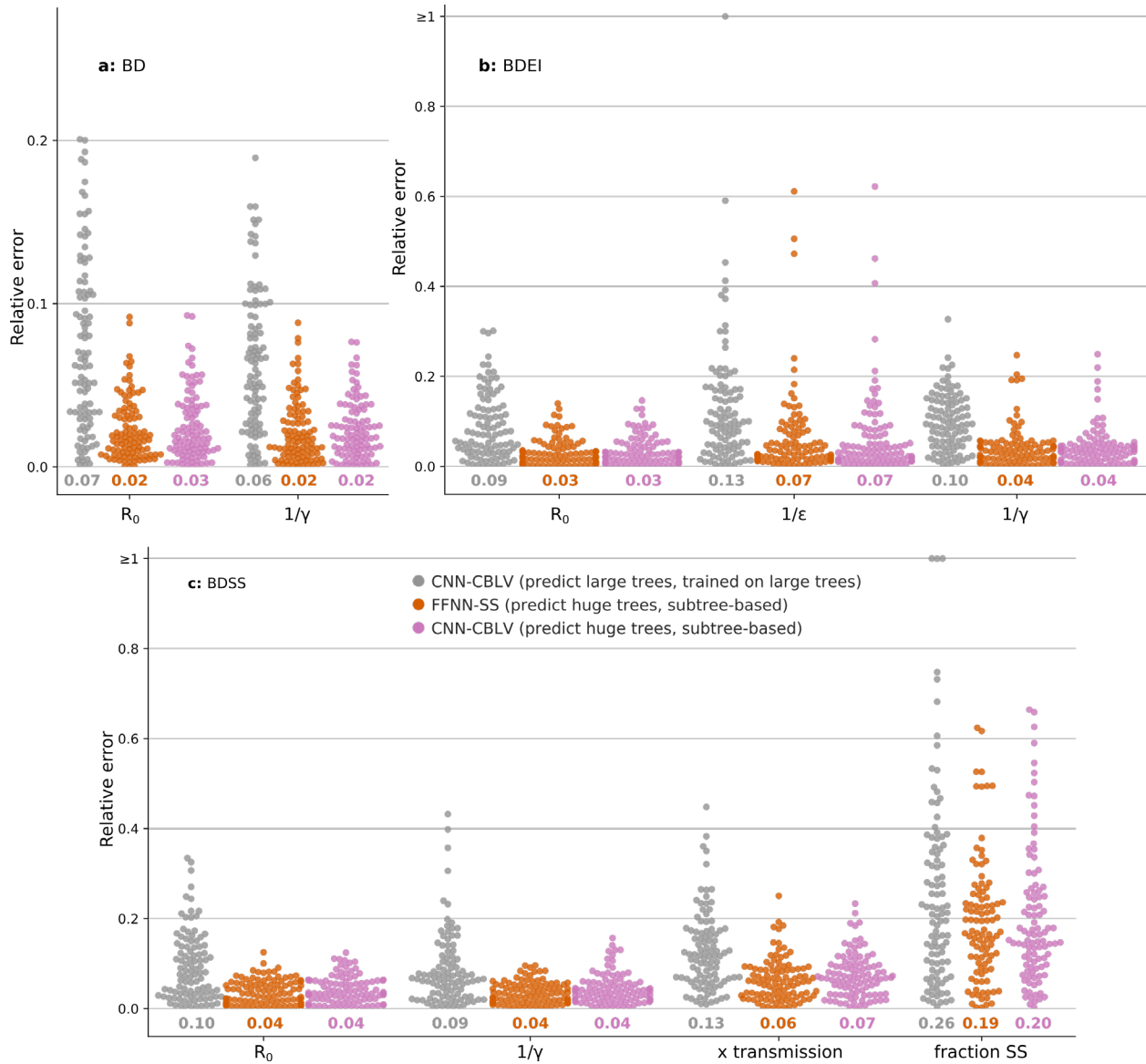
437

438 **Note to Fig. 3.** Comparison of inference accuracy by BEAST2 (in blue), deep neural network trained on SS (in orange)  
 439 and convolutional neural network trained on the CBLV representation (in green) on 100 test trees. The size of training  
 440 and testing trees was uniformly sampled between 200 and 500 tips. We show the relative error for each test tree. The  
 441 error is measured as the normalized distance between the median *a posteriori* estimate by BEAST2 or point estimates  
 442 by neural networks and the target value for each parameter. We highlight simulations for which BEAST2 did not  
 443 converge and whose values were thus set to median of the parameter subspace used for simulations by depicting them  
 444 as red squares. We further highlight the analyses with a high relative error (>1.00) for one of the estimates as black  
 445 diamonds. We compare the relative errors for **a**, BD-simulated, **b**, BDEI-simulated and **c**, BDSS-simulated trees.  
 446 Average relative error is displayed for each parameter and method in corresponding colour below each figure. The  
 447 average error of a FFNN trained on summary statistics but with randomly permuted target is displayed as black dashed  
 448 line and its value is shown in bold black below the x-axis. The accuracy of each method is compared by paired z-test;  
 449  $P < 0.05$  is shown as thick full line; non-significant is not shown.

450



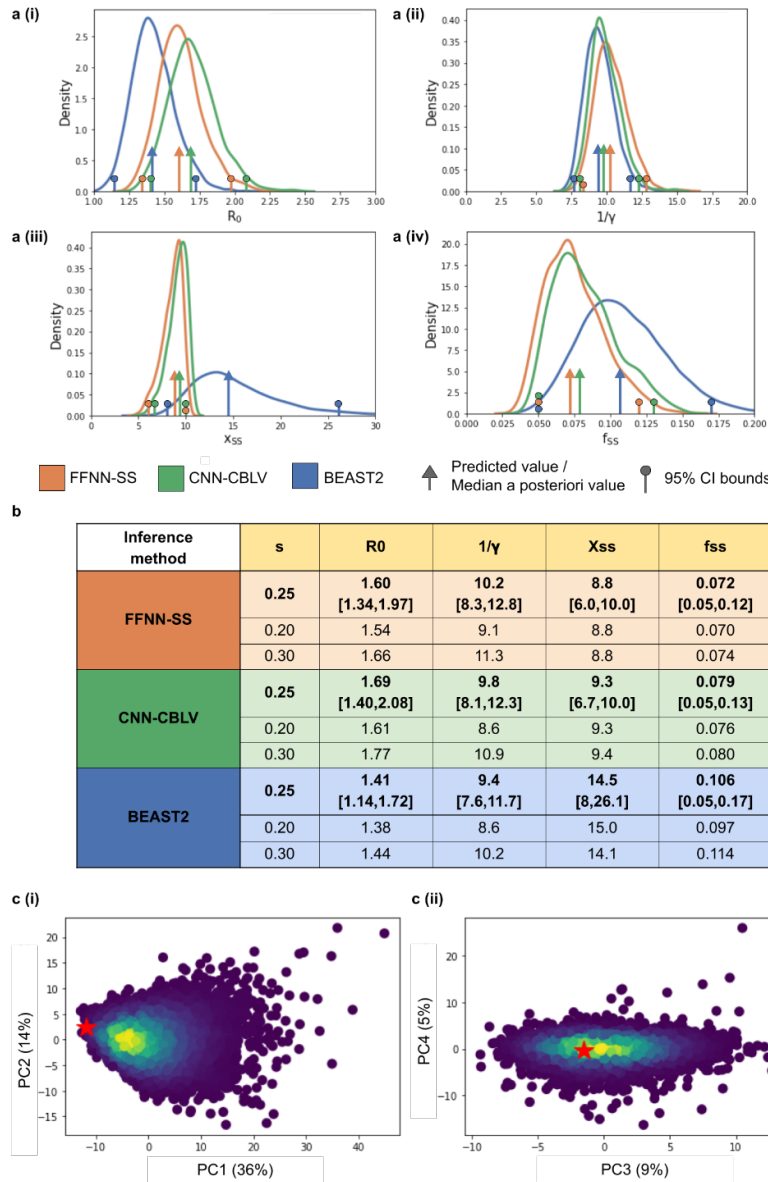
451 **Fig. 4: Deep learning accuracy with ‘huge’ trees**



452

453 **Note to Fig. 4.** Comparison of inference accuracy by neural networks trained on large trees in predicting large trees  
 454 (CNN-CBLV, in grey, same as in Fig. 3) and huge trees (FFNN-SS, in orange, and CBLV-NN, in pink) on 100 large  
 455 and 100 huge test trees. The training and testing large trees are the same as in Fig. 3 (between 200 and 500 tips each).  
 456 The huge testing trees were generated for the same parameters as the large training and testing trees, but their size  
 457 varied between 5,000 and 10,000 tips. We show the relative error for each test tree. The error is measured as the  
 458 normalized distance between the point estimates by neural networks and the target values for each parameter. We  
 459 compare the relative errors for **a**, BD-simulated, **b**, BDEI-simulated and **c**, BDSS-simulated trees. Average relative  
 460 error is displayed for each parameter and method in corresponding colour below each plot.

461 **Fig. 5: Parameter inference on HIV data sampled from MSM Zurich**



462

463 **Note to Fig. 5.** Using BDSS model with BEAST2 (in blue), FFNN-SS (in orange), and CNN-CBLV (in green) we  
464 infer, **a (i)**, basic reproduction number, **a (ii)**, infectious period (in years), **a (iii)**, superspreading transmission ratio  
465 and, **a (iv)**, superspreading fraction. For FFNN-SS and CNN-CBLV, we show the **posterior** distributions and the 95%  
466 CIs obtained with a fast approximation of the parametric bootstrap (**Methods**). For BEAST2, the **posterior**  
467 distributions and 95% CI were obtained considering all reported steps (9,000 in total) excluding the 10% burn-in.  
468 Arrows show the position of the original point estimates obtained with FFNN-SS and CNN-CBLV and the median *a*  
469 *posteriori* estimate obtained with BEAST2. Circles show lower and upper boundaries of 95% CI. **b**, these values are  
470 reported in a table, together with point estimates obtained while considering lower and higher sampling probabilities  
471 (0.20 and 0.30). **c**, 95% CI boundaries obtained with FFNN-SS are used to perform an *a posteriori* model adequacy  
472 check. We simulated 10,000 trees with BDSS while resampling each parameter from a uniform distribution, whose  
473 upper and lower bounds were defined by the 95% CI. We then encoded these trees into SS, performed PCA and  
474 projected SS obtained from the HIV MSM phylogeny (red stars) on these PCA plots. We show here the projection  
475 into **c (i)**, first two components of PCA, **c (ii)**, the 3<sup>rd</sup> and 4<sup>th</sup> components, together with the associated percentage of  
476 variance displayed in parentheses. Warm colours correspond to high density of simulations.

477 **CITATIONS**

- 478 1. Grenfell, B.T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**,  
479 327-332 (2004).
- 480 2. Volz, E.M., Kosakovsky Pond, S.L., Ward, M.J., Leigh Brown, A.J., Frost, S.D. Phylodynamics of infec-  
481 tious disease epidemics. *Genetics* **183**, 1421-30 (2009).
- 482 3. Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G. Bayesian Coalescent Inference of Past Population  
483 Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, **22**, 1185–1192 (2005).
- 484 4. Stadler, T. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C  
485 virus (HCV). *Proceedings of the National Academy of Sciences* **110**, 228-233 (2013)
- 486 5. Stadler, T., Bonhoeffer, S. Uncovering epidemiological dynamics in heterogeneous host populations using  
487 phylogenetic methods. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368(1614)**,  
488 20120198 (2013).
- 489 6. Gire, S.K. et al Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 out-  
490 break. *Science* **345**, 1369-72 (2014).
- 491 7. Boskova, V., Bonhoeffer, S., Stadler, T. Inference of Epidemiological Dynamics Based on Simulated Phy-  
492 logenies Using Birth-Death and Coalescent Models. *PLOS Computational Biology* **10(11)**, e1003913  
493 (2014).
- 494 8. Stadler, T. *et al.* Estimating the Basic Reproductive Number from Viral Sequence Data. *Mol. Biol. Evol.*  
495 **29**, 347–57 (2012).
- 496 9. Leventhal, G.E., Günthard, H.F., Bonhoeffer, S., Stadler, T. Using an Epidemiological Model for Phyloge-  
497 netic Inference Reveals Density Dependence in HIV Transmission. *Mol. Biol. Evol.* **31**, 6–17 (2014).
- 498 10. Stadler, T., Kuhnert, D., Rasmussen, D.A., du Plessis, L. Insights into the early epidemic spread of Ebola in  
499 sierra leone provided by viral sequence data. *PLoS Curr.* **6**, (2014).
- 500 11. Kühnert, D., Stadler, T., Vaughan, T.G., Drummond, A.J. Phylodynamics with Migration: A Computa-  
501 tional Framework to Quantify Population Structure from Genomic Data. *Mol. Biol. Evol.* **33**, 2102-16  
502 (2016).

- 503 12. Sciré, J., Barido-Sottani, J., Kühnert, D., Vaughan, T.G., Stadler, T. Improved multi-type birth-death phylo-  
504 dynamic inference in BEAST 2 (2020). Preprint at [https://www.biorxiv.org/con-](https://www.biorxiv.org/content/10.1101/2020.01.06.895532v1.full.pdf)  
505 [tent/10.1101/2020.01.06.895532v1.full.pdf](https://www.biorxiv.org/content/10.1101/2020.01.06.895532v1.full.pdf)
- 506 13. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computa-*  
507 *tional Biology* **10(4)**, e1003537 (2014).
- 508 14. Boskova, V., Stadler, T., Magnus, C. The influence of phylodynamic model specifications on parameter  
509 estimates of the Zika virus epidemic. *Virus Evolution* **4(1)**, vex044 (2018).
- 510 15. Vaughan, T.G., Sciré, J., Nadeau, S.A., Stadler, T. Estimates of outbreak-specific SARS-CoV-2 epidemio-  
511 logical parameters from genomic data (2020). Preprint at [https://www.medrxiv.org/con-](https://www.medrxiv.org/content/10.1101/2020.09.12.20193284v1.full.pdf)  
512 [tent/10.1101/2020.09.12.20193284v1.full.pdf](https://www.medrxiv.org/content/10.1101/2020.09.12.20193284v1.full.pdf)
- 513 16. Rubin, D.B. Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *The*  
514 *Annals of Statistics* **12**, 1151-72 (1984).
- 515 17. Beaumont, M.A., Zhang, W., Balding, D.J. Approximate Bayesian Computation in Population Genetics.  
516 *Genetics* **164**, 2025-35 (2002).
- 517 18. Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O. Approximate Bayesian Computation (ABC) in  
518 practice. *Trends in Ecology & Evolution* **25**, 410-8 (2010).
- 519 19. Saulnier, E., Gascuel, O., Alizon, S. Inferring epidemiological parameters from phylogenies using regres-  
520 sion-ABC: A comparative study. *PLoS Comp. Biol.* **13(3)**, e1005416 (2017).
- 521 20. Blum, M.G.B. *Handbook Of Approximate Bayesian Computation Ch. Regression approaches for ABC.* 71–  
522 85. (Chapman and Hall/CRC Press, Boca Raton, 2018).
- 523 21. LeCun, Y., Kavukcuoglu, K., Farabet, F. Convolutional networks and applications in vision. *Proc. IEEE*  
524 *Int. Symp. Circuits Syst.* 253-6 (2010).
- 525 22. Krizhevsky, K., Sutskever, I., Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Net-  
526 works. *Advances in neural information processing systems* 1097-105 (2012).
- 527 23. Chattopadhyay, A., Hassanzadeh, P., Pasha, S. Predicting clustered weather patterns: A test case for appli-  
528 cations of convolutional neural networks to spatio-temporal climate data. *Sci. Rep.* **10**, 1317 (2020)
- 529 24. The Swiss HIV Cohort Study *et al.* Cohort Profile: The Swiss HIV Cohort Study. *International Journal of*  
530 *Epidemiology* **39**, 1179–89 (2010).

- 531 25. Rasmussen, D.A., Kouyos, R., Günthard, H.F., Stadler, T. Phylodynamics on local sexual contact networks.  
532 *PLOS Comp. Biol.* **13(3)**, e1005448 (2017).
- 533 26. Colijn, C. & Plazzotta, G. A metric on phylogenetic tree shapes. *Systematic Biology* **67**, 113–26 (2018).
- 534 27. Liu, P., Gould, M., Colijn, C. Analyzing Phylogenetic Trees with a Tree Lattice Coordinate System and a  
535 Graph Polynomial, *Systematic Biology*, in press (2022). Preprint at <https://doi.org/10.1093/sysbio/syac008>
- 536 28. Lewitus, E. & Morlon, H. Characterizing and Comparing Phylogenies from their Laplacian Spectrum. *Sys-*  
537 *tematic Biology* **65**, 495-507 (2016).
- 538 29. Kim, J., Rosenberg, N.A., Palacios, J.A. Distance metrics for ranked evolutionary trees. *Proceedings of the*  
539 *National Academy of Sciences* **117**, 28876-86 (2020).
- 540 30. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C. *Introduction To Algorithms*. 286-307 (The MIT  
541 Press, Cambridge, 2009).
- 542 31. Bengio, Y. *Neural Networks: Tricks Of The Trade*, Ch. *Practical Recommendations for Gradient-Based*  
543 *Training of Deep Architectures*. (Springer, Berlin, Heidelberg 2002).
- 544 32. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. *Bayesian Data Analysis: Second Edition*. (Chapman and  
545 Hall/CRC Press, Boca Raton, 2004).
- 546 33. Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while ac-  
547 commodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157-67 (2012).
- 548 34. Kouyos, R.D. *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmis-  
549 sion in Switzerland. *J. Infect. Dis.* **201**, 1488-97 (2010).
- 550 35. May, R.M. & Anderson, R.M. Transmission dynamics of HIV infection. *Nature* **326**, 137–142 (1987).
- 551 36. Brenner, B.G. *et al.* Quebec Primary HIV Infection Study Group. High rates of forward transmission events  
552 after acute/early HIV-1 infection. *J. Infect. Dis.* **195**, 951-9 (2007).
- 553 37. Gueler, A. *et al.* Swiss National Cohort Life expectancy in HIV-positive persons in Switzerland. *AIDS* **31**,  
554 427-436 (2017).
- 555 38. Rasmussen, D.A., Volz, E.M., Koelle, K. Phylodynamic Inference for Structured Epidemiological Models.  
556 *PLoS Comput. Biol.* **10(4)**, e1003570 (2014).
- 557 39. Volz, E.M. & Siveroni, I. Bayesian phylodynamic inference with complex models. *PLoS Comput. Biol.*  
558 **14(11)**, e1006546 (2018).

- 559 40. MacPherson, A., Louca, S., McLaughlin, A., Joy, J.B., Pennell, M.W. Unifying Phylogenetic Birth–Death  
560 Models in Epidemiology and Macroevolution, *Systematic Biology*, **71(1)**, 172–189 (2022).
- 561 41. Sanchez, T., Cury, J., Charpiat, G., Jay, F. Deep learning for population size history inference: Design,  
562 comparison and combination with approximate Bayesian computation. *Mol. Ecol. Resour.* **00**, 1-16. (2020).
- 563 42. Dunn, D. & Pillay, D. UK HIV drug resistance database: background and recent outputs. *J. HIV Ther.* **12**,  
564 97–8 (2007).
- 565 43. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality.  
566 *Euro Surveill.* **22**, 30494 (2017).
- 567 44. Minh, B.Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic  
568 era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- 569 45. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A. RAxML-NG: a fast, scalable and user-  
570 friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
- 571 46. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing  
572 the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-21 (2010).
- 573 47. Sagulenko, P., Puller, V., Neher, R.A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus*  
574 *Evol.* **4(1)**, vex042 (2018).
- 575 48. To, T.H., Jung, M., Lycett, S., Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst*  
576 *Biol.* **65**, 82-97 (2016).
- 577 49. Volz, E.M. & Frost, S.D.W. Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3(3)**, vex025 (2017).

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

I thank the authors for their detailed response to my previous comments, which have been satisfactorily addressed in the revised manuscript. This work will be of interest to the readership of this journal.