In the previous year (2022-2023) we wanted to explore new mathematical models that can help reconstruct the dynamics of infectious diseases. We are continuing to build on this novel framework and have two research projects in progress which will take the maths forwards to novel and useful applications. We will likely report on these in the next report (2024-2025).

As we mentioned in our previous report, we have identified genetic data as a key component to understanding epidemic dynamics. In keeping without desire to develop the state-of-the-art we are in the process of creating new techniques to perform phylogenetics. Our work has led to the publication of several novel approaches that have been published in the best journals for the field of phylogenetics. First, *in Khurana et al 2023, Systematic Biology*, we explore the suitability of the most popular model in phylogenetics – the constant birth death process. When performing modelling, one needs to decide on the mechanism underpinning the process of interest, and the birth death model is the most popular mechanism in phylogenetics. In this paper we demonstrate, for the first time, the robustness of this model to a large range of scenarios and determine where it fails. In *Penn et al 2023, Genome Biology and Evolution*, we try to utilise deep learning to improve phylogenetics. The key challenge in determining a tree of relatedness is just how profoundly the number of possible trees grow. For example, say you have 40 SARS-Cov-2 genomes, a trivially small number, simply counting each possible tree relating these genomes with a hydrogen atom would ignite a star. This problem is compounded by the fact that a criterion that tells us which tree best fits the genetic data, is NP-hard (unsolvable), and the only solution is brute force or heuristics. To make progress on this challenging problem, we look to the field of deep learning, and try to use the tool that made deep learning possible, gradient descent. In our paper, we recast a discrete tree as a continuous object, and calculate a gradient (rate of change) on this object. This allows us to leap across tree space and explore the massive space well – indeed we do better than the current state of the art. To our knowledge, we are the first to have done this. Our approach is still costly and has limitations, and if we can further improve on a way to compute the gradient, it is not hyperbole to think our method could change the field. Finally in *Penn et al 2023, Systematic Biology*, we look to the practical aspects of phylogenetics. How does one even represent a tree? The universal solution is with something called a Newick string. An example for a tree with 4 leaves would be ((A, B), (C, D)) – where A and B are siblings and C and D the pairs share an ancestor. This string is very useful, but it has limitations – it is hard to compare of two trees. ((A, B), (C, D)) is identical to ((B, A), (D, C)). We wanted to take a radically different solution and decided to encode a tree not as a string, or some other linked representation, but as a vector (series) of integers. For example, the above tree is always [0,1,3] in our representation. Our representation can be used for several applications, such as inferring the correct tree, or as in the previous paper, exploring tree space. However, for this report, we showcase a plot from our paper (see below). While it may not seem exciting scientifically, our new representation requires 6 times less space to store a tree. 1 terabyte of trees stored in the current format would only need 160 gigabytes in ours. We can also perform operations such as finding unique trees several orders of magnitude faster. The world of genetics is growing constantly, more data, more trees, more everything. Our approach provides a means for numerous practical speed ups as well as avenues for new scientific contributions.

Our chair grant promised to deliver new methods for future pandemics, and to ensure that we can eventually guide policy. Prof Bhatt's previous work was the first to answer the question: how have interventions affected the spread of COVID-19. Following previous work we have written a well-received policy document, *Lison et al 2023, Lancet Public health* summarising and setting an agenda for future work. Prof Bhatt was also a part of the steering group and an author on the landmark study looking at the effect of interventions, published in the Royal Society[1].

[1] https://royalsociety.org/news-resources/projects/impact-non-pharmaceutical-interventions-on-covid-19-transmission/

(https://royalsociety.org/news-resources/projects/impact-non-pharmaceutical-interventions-on-covid-19-transmission/)

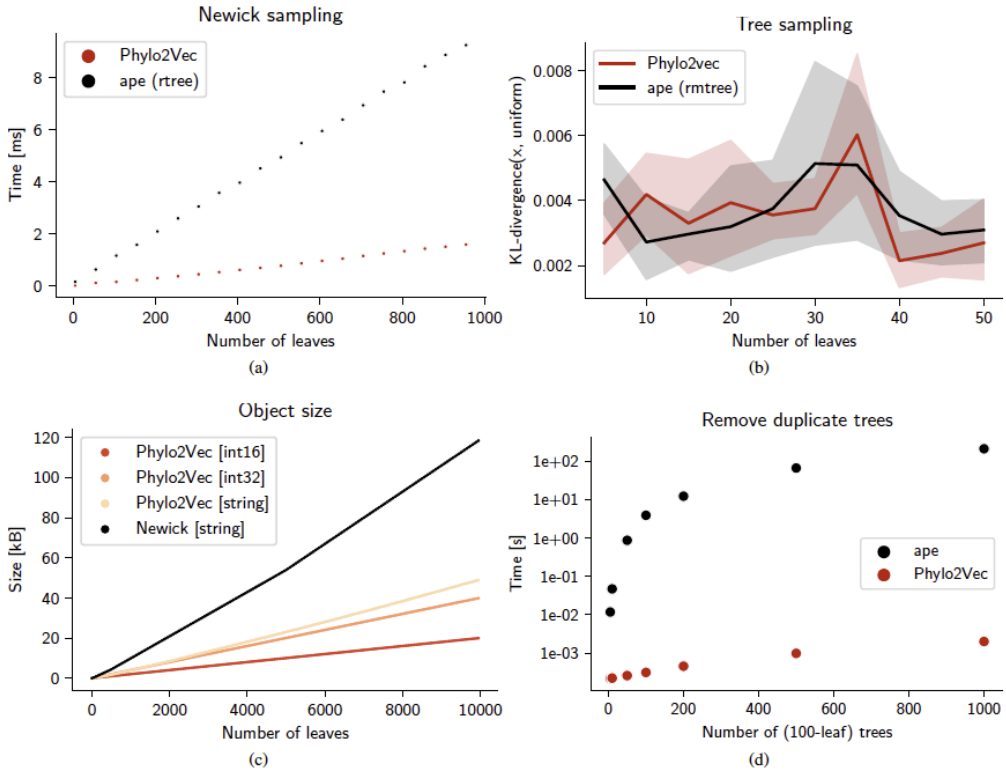

Fig. 7. Phylo2Vec allows for fast and unbiased sampling, low memory or storage, and fast comparison of trees. **(a)** Average sampling time using `phylo2vec.utils.sample` and `rtree` from `ape`. Execution time was measured over 100 executions using Python's `timeit` and R's `microbenchmark`, respectively. **(b)** Sampling bias comparison. For each size and sampler, we sample 10000 trees and converted them first to their Phylo2Vec representation, and second to an integer using a method similar to that of Rohlf (1983). We then compare the probability distributions of the integers generated by Phylo2Vec and `ape` sampling against the reference uniform distribution for each tree size using the Kullback-Leibler (KL) divergence. The lower the KL-divergence value, the more the reference distribution and the distribution of interest share similar information. **(c)** Object sizes for different tree sizes of Phylo2Vec vectors (stored as a 16- or 32-bit `numpy` integer array, or a string) compared against their Newick-format equivalents (without branch length information). **(d)** Average time for duplicate removal from a set of trees using Phylo2Vec (vectors) and the *unique.multiPhylo* function from `ape`. Execution time was measured over 30 executions using Python's `timeit` and R's `microbenchmark`, respectively.

Finally, a goal of our grant was to establish a nucleus of infectious disease expertise at the University of Copenhagen. We have, established a legal partnership with the SSI, DTU and DST to analyse COVID data and collaborate closely moving forwards. We note, we are the first group to have achieved such a broad partnership with data sharing legal agreements. As a partnership, we led a grant to Innovation Fund Denmark, and interviewed but were unsuccessful. We are currently pursuing other bits for sustained funding. We are also in the final stages of two major reports on an evaluation of COVID-19 in Denmark, which will be reported on in next years research highlights.