I det forgangne år (2022-2023) ønskede vi at udforske nye matematiske modeller, der kan hjælpe med at rekonstruere udviklingen af smitsomme sygdomme. Vi fortsætter med at bygge videre på denne nye model og har to igangværende forskningsprojekter, som vil føre matematikken videre til nye og nyttige anvendelser. Vi vil sandsynligvis rapportere om disse i den næste rapport (2024-2025).

Som vi nævnte i vores tidligere rapport, så har vi udpeget genetiske data som en nøglekomponent til at forstå epidemiers dynamik. I overensstemmelse med vores ønske om at udvikle state-of-the-art metoder er vi i gang med at skabe nye phylogenetiske modeller. Vores arbejde har ført til udgivelsen af flere nye metoder, der er blevet offentliggjort i nogle af de bedste tidsskrifter inden for fylogenetik. *I Khurana et al 2023, Systematic Biology*, undersøger vi for første gang egnetheden af den mest populære model inden for fylogenetik - den konstante fødsel-død-proces. Når man modellerer indenfor fylogenetik, er man nødt til at beslutte sig for, hvilken mekanisme der ligger til grund for den proces, man er interesseret i, og fødselsdødsmodellen er den mest populære mekanisme. I denne artikel demonstrerer vi for første gang denne models robusthed over for en lang række scenarier og fastslår, hvor den fejler.

I *Penn et al 2023, Genome Biology and Evolution*, forsøger vi at bruge *deep learning* til at forbedre fylogenetikken. Den største udfordring ved at bestemme et slægtskabstræ (fylogeni) er, hvor meget antallet af potentielle træer vokser. Dette problem forværres af, at et kriterium, der fortæller os, hvilket træ der passer bedst til de genetiske data, er NP-hårdt (uløseligt), og den eneste løsning er brute force eller heuristik. For at gøre fremskridt med dette udfordrende problem forsøger vi at bruge det værktøj, der gjorde *deep learning* mulig, nemlig gradientnedstigning (*gradient descent)*. I vores artikel omdanner vi et diskret træ til et kontinuerligt objekt og beregner en gradient (ændringshastighed) på dette objekt. Det giver os mulighed for at springe over trærummet og udforske det massive rum på en god måde - endda bedre end det nuværende tekniske niveau. Så vidt vi ved, er vi de første, der har gjort dette. Vores tilgang er fortsat dyr og har begrænsninger, og hvis vi yderligere kan forbedre vores måde at beregne gradienten på, er det ikke en overdrivelse at tro, at vores metode kan ændre feltet.

Endelig ser vi i *Penn et al 2023, Systematic Biology*, på de praktiske aspekter af fylogenetikken. Hvordan repræsenterer man overhovedet et træ? Den gængse løsning er noget, der kaldes en Newick-streng. Et eksempel på et træ med 4 blade ville være ((A, B), (C, D)) - hvor A og B er søskende, og C og D er par, der deler en forfader. Denne streng er meget nyttig, men den har sine begrænsninger - det kan være svært at sammenligne to træer. ((A, B), (C, D)) er identisk med ((B, A), (D, C)). Vi ønskede at anvende en radikalt anderledes løsning og besluttede at kode et træ ikke som en streng eller en anden kædet repræsentation, men som en vektor (serie) af heltal. For eksempel er ovenstående træ altid [0,1,3] i vores repræsentation. Vores repræsentation kan bruges til flere formål, såsom at udlede det korrekte træ, eller som i den tidligere artikel, at udforske træets rum. Men i denne rapport viser vi et plot fra vores artikel (se nedenfor). Selvom det måske ikke virker spændende rent videnskabeligt, kræver vores nye repræsentation 6 gange mindre plads til at gemme et træ. 1 terabyte træer gemt i det nuværende format ville kun kræve 160 gigabyte i vores. Vi kan også udføre operationer som at finde unikke træer flere størrelsesordener hurtigere. Genetikkens verden vokser konstant, flere data, flere træer, mere af det hele. Vores tilgang giver mulighed for adskillige praktiske hastighedsforøgelser såvel for nye videnskabelige bidrag.

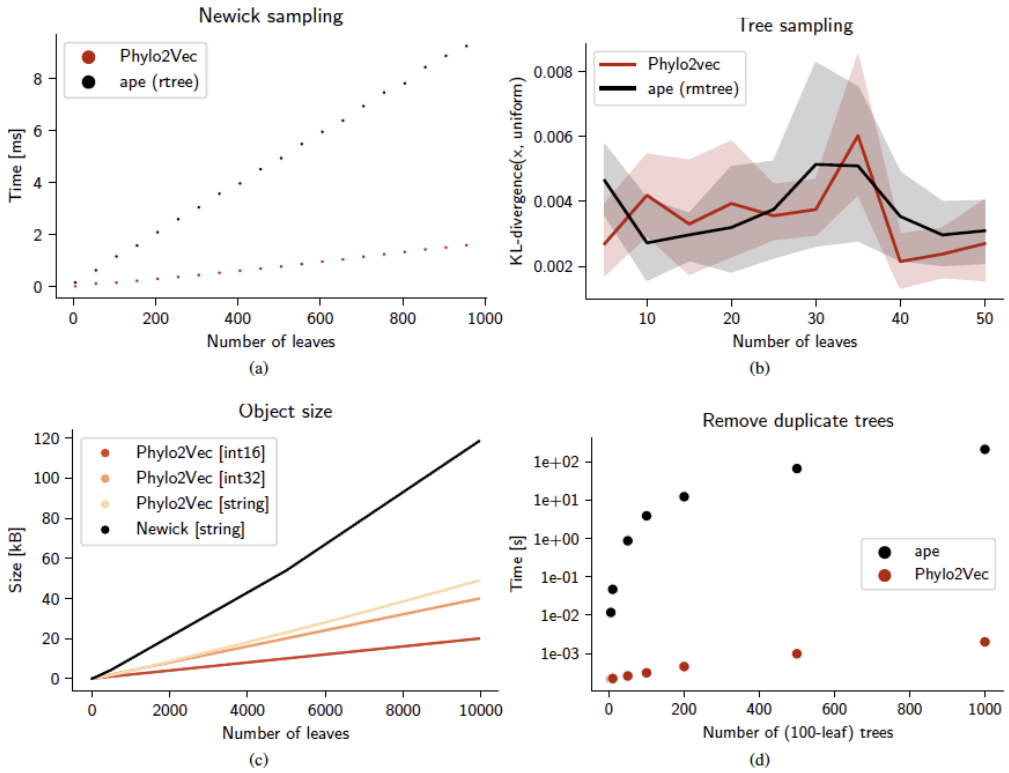

Fig. 7. Phylo2Vec allows for fast and unbiased sampling, low memory or storage, and fast comparison of trees. **(a)** Average sampling time using `phylo2vec.utils.sample` and `rtree` from `ape`. Execution time was measured over 100 executions using Python's `timeit` and R's `microbenchmark`, respectively. **(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.

Vores bevilling lovede at levere nye metoder til fremtidige pandemier og at sikre, at vi i sidste ende kan guide fremtidige retningslinjer. Prof. Bhatts tidligere arbejde var den første til at besvare spørgsmålet: hvordan har interventionsforanstaltninger (e.g., nedlukninger) påvirket spredningen af COVID-19? I forlængelse af vores tidligere arbejde har vi skrevet et velmodtaget politisk dokument, *Lison et al 2023, Lancet Public Health* , som opsummerer og sætter en dagsorden for det fremtidige forskningarbejde. Prof. Bhatt var også en del af styregruppen og forfatter til den skelsættende undersøgelse af effekten af forskellige COVID-19 interventioner, som blev offentliggjort i Royal Society (https://royalsociety.org/news-resources/projects/impact-non-pharmaceutical-interventions-on-covid-19-transmission/).

Endelig var et af målene med vores bevilling at etablere en kerne af ekspertise inden for infektionssygdomme på Københavns Universitet. Vi har etableret et juridisk partnerskab med SSI, DTU og DST for at analysere COVID-data og arbejde tæt sammen fremadrettet. Vi bemærker, at vi er den første gruppe, der har opnået et så bredt partnerskab med juridiske aftaler om datadeling. Som et partnerskab søgte vi en bevilling til Innovationsfonden og blev interviewet, men uden succes. Vi forfølger i øjeblikket andre muligheder for vedvarende finansiering. Vi er også ved at lægge sidste hånd på to store rapporter om en evaluering af COVID-19 i Danmark, som vil blive omtalt i næste års forskningshøjdepunkter.