













# 1 phylo2vec: a library for vector-based phylogenetic tree 2 manipulation

3 Neil Scheidwasser <sup>1\*</sup>, Ayush Nag<sup>2\*</sup>, Matthew J Penn <sup>1</sup>, Anthony MV  
4 Jakob <sup>4</sup>, Frederik Mølkjær Andersen <sup>1</sup>, Mark P Khurana <sup>1</sup>, Landung  
5 Setiawan <sup>2</sup>, Madeline Gordon <sup>2</sup>, David A Duchêne <sup>1</sup>, and Samir  
6 Bhatt <sup>1,3¶</sup>

7 1 Section of Health Data Science and AI, University of Copenhagen, Copenhagen, Denmark 2 eScience  
8 Institute, University of Washington, Seattle, United States 3 MRC Centre for Global Infectious Disease  
9 Analysis, Imperial College London, London, United Kingdom 4 Independent researcher ¶ Corresponding  
10 author \* These authors contributed equally.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

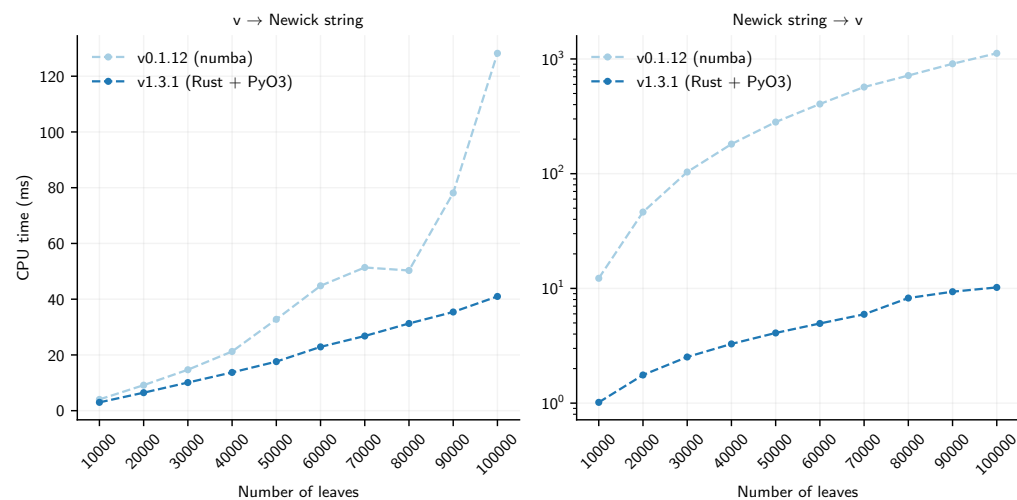
## License

Authors of papers retain copyright  
and release the work under a  
Creative Commons Attribution 4.0  
International License ([CC BY 4.0](#)).

## 11 Summary

12 Phylogenetics is a fundamental component of evolutionary analysis frameworks in biology (Yang,  
13 2014) and linguistics (Atkinson & Gray, 2005). Recently, the advent of large-scale genomics and  
14 the SARS-CoV-2 pandemic has highlighted the necessity for phylogenetic software to handle  
15 large datasets (Attwood et al., 2022; Kapli et al., 2020; Khurana et al., 2024; Kraemer et al.,  
16 2025). While significant efforts have focused on scaling optimisation algorithms (De Maio et  
17 al., 2023; Sanderson, 2021; Turakhia et al., 2021), visualization (Sanderson, 2022), and lineage  
18 identification (McBroome et al., 2024), an emerging body of research has been dedicated to  
19 efficient representations of data for genomes (Deorowicz et al., 2023) and phylogenetic trees  
20 such as phylo2vec (Chauve et al., 2025; Penn et al., 2024; Richman et al., 2025). Compared  
21 to the traditional Newick format which represents trees using strings of nested parentheses  
22 (Felsenstein, 2004), modern tree representations utilize integer vectors to define the tree  
23 topology traversal. This approach offers several advantages, including easier manipulation,  
24 increased memory efficiency, and applicability to machine learning.

25 Here, we present the latest release of phylo2vec (or Phylo2Vec), a high-performance software  
26 package for encoding, manipulating, and analysing binary phylogenetic trees. At its core,  
27 the package is based on the phylo2vec (Penn et al., 2024) representation of binary trees,  
28 and is designed to enable fast sampling and tree comparison. This release features a core  
29 implementation in Rust for improved performance and memory efficiency (Figure 1), with  
30 wrappers in R and Python (superseding the original release (Penn et al., 2024)), making it  
31 accessible to a broad audience in the bioinformatics community.



**Figure 1:** Benchmark times for converting a phylo2vec vector to a Newick string (left) and vice versa (right). Execution time was measured over at least 20 runs per size, comparing Python functions in the latest release (via Rust bindings with [PyO3](#)) against the previous release ([Penn et al., 2024](#)) based on Numba ([Lam et al., 2015](#)). Benchmarks ran on an on an AMD Ryzen Threadripper PRO 5995WX (64 cores, 7 GHz, 256 GB RAM).

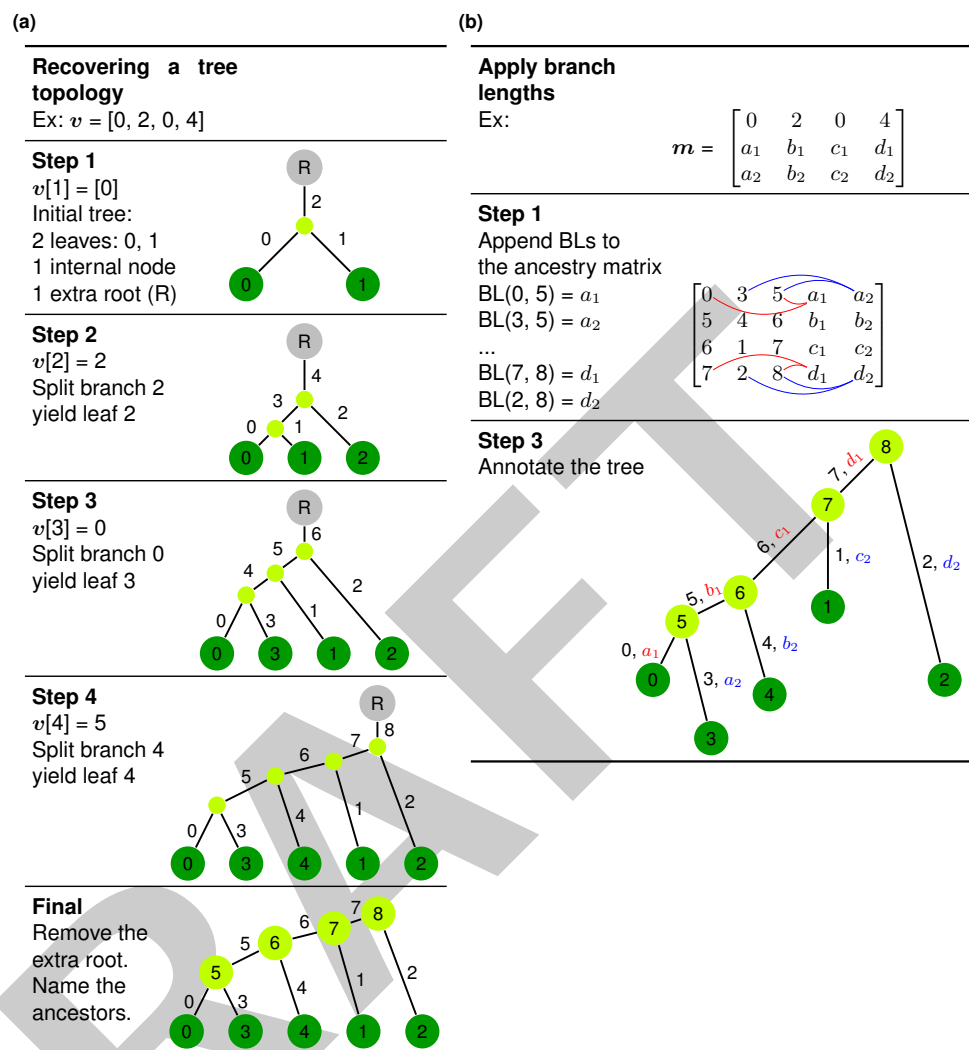
## Statement of need

The purpose of the phylo2vec library is threefold. First, it provides robust phylogenetic tree manipulation in Rust, complementing other efforts such as `light_phylogeny` ([Duchemin et al., 2018](#)) for reconciled phylogenies ([Nakhleh, 2013](#)), and `rust-bio` ([Köster, 2016](#)), which does not yet cover phylogenetics. Second, it complements existing libraries such as `ape` ([Paradis & Schliep, 2019](#)) in R, and `ete3` ([Huerta-Cepas et al., 2016](#)) and `DendroPy` ([Moreno et al., 2024](#)) in Python, by providing fast tree sampling, fast tree comparison and efficient tree data compression ([Penn et al., 2024](#)). Third, the phylo2vec representation offers a pathway to using new optimisation frameworks for phylogenetic inference. A notable example is `GradME` ([Penn et al., 2023](#)), a gradient descent-based algorithm which uses a continuous relaxation of the phylo2vec representation.

## Features

The presented release of phylo2vec addresses optimisations limitations of ([Penn et al., 2024](#)) with  $\mathcal{O}(n \log n)$  implementations for vector-to-Newick and Newick-to-vector conversions, leveraging Adelson-Velsky and Landis (AVL) trees ([Adelson-Velsky & Landis, In Russian. English translation in Soviet Mathematics Doklady, 3:1259–1263, 1962](#)) and Fenwick trees ([Fenwick, 1994](#)), respectively.

New features include supporting branch length annotations by extending the vector representation of size  $n - 1$  leaves to a matrix of size  $(n - 1) \times 3$  ([Figure 2](#)), leaf-level operations (pruning, placement, MRCA identification), fast cophenetic distance matrix calculation, and various optimisation schemes based on phylo2vec tree representations, notably hill-climbing ([Penn et al., 2024](#)) and `GradME` ([Penn et al., 2023](#)). Bayesian MCMC inference is enabled through a likelihood function that leverages tree representation similarities with `BEAGLE` ([Ayres et al., 2012](#); [Suchard & Rambaut, 2009](#)). Finally, user-friendliness is enhanced with step-by-step demos of the inner workings of phylo2vec's representations and their conversion to common phylogenetic tree formats.



**Figure 2:** Recovering a tree from a phylo2vec vector: example for  $v = [0, 2, 0, 4]$ . (a) Leaf placement algorithm from (Penn et al., 2024). (b) Augmenting  $v$  into a matrix  $m$  with branch lengths. We use an intermediary ancestry matrix whereby each row encodes a cherry (two children + parent) with two extra columns for branch lengths. The node with the smallest descendant has the branch length in the second column, and the other the branch length in the third column.

## Maintenance

A strong focus of this release is to support long-term maintenance through implementing recommended software practices into its project structure and development workflow. The project is structured with a Rust API containing core algorithms with language bindings to avoid tight coupling and enable easy language additions. Additionally, we have established a robust continuous integration (CI) pipeline using GitHub Actions, which features:

- Unit test frameworks for Rust (cargo), Python (pytest), and R (testthat (Wickham, 2011))
- Benchmarking on the Rust code (criterion) and its Python bindings (pytest-benchmark)

Lastly, to complement Jupyter Notebook demos, comprehensive documentation is provided using Jupyter Book and Read The Docs.

## Acknowledgements

S.B. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. S.B. acknowledges support from the Danish National Research Foundation via a chair grant (DNRF160, also supporting N.S. and M.P.K.), The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345, also supporting M.J.P. and F.M.A.), and the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). D.A.D. is supported by a Data Science - Emerging researcher award from Novo Nordisk Fonden (NNF23OC0084647).

## References

- Adelson-Velsky, Georgii, & Landis, E. (In Russian. English translation in Soviet Mathematics Doklady, 3:1259–1263, 1962). An algorithm for the organization of information. *Proc. USSR Acad. Sci.*, 146, 263–266.
- Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4), 513–526. <https://doi.org/10.1080/10635150590950317>
- Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., & Pybus, O. G. (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics*, 23(9), 547–562. <https://doi.org/10.1038/s41576-022-00483-8>
- Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., & others. (2012). BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, 61(1), 170–173. <https://doi.org/10.1093/sysbio/syr100>
- Chauve, C., Colijn, C., & Zhang, L. (2025). A vector representation for phylogenetic trees. *Philosophical Transactions B*, 380(1919), 20240226. <https://doi.org/10.1098/rstb.2024.0226>
- De Maio, N., Kalaghatgi, P., Turakhia, Y., Corbett-Detig, R., Minh, B. Q., & Goldman, N. (2023). Maximum likelihood pandemic-scale phylogenetics. *Nature Genetics*, 55(5), 746–752. <https://doi.org/10.1038/s41588-023-01368-0>
- Deorowicz, S., Danek, A., & Li, H. (2023). AGC: Compact representation of assembled genomes with fast queries and updates. *Bioinformatics*, 39(3), btad097. <https://doi.org/10.1093/bioinformatics/btad097>
- Duchemin, W., Gence, G., Arigon Chifolleau, A.-M., Arvestad, L., Bansal, M. S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A. A., Dessimoz, C., Dylus, D., Hasic, D., Mallo, D., Planel, R., Posada, D., Scornavacca, C., Szöllösi, G., Zhang, L., ... Daubin, V. (2018). RecPhyloXML: A format for reconciled gene trees. *Bioinformatics*, 34(21), 3646–3652. <https://doi.org/10.1093/bioinformatics/bty389>
- Felsenstein, J. (2004). *Inferring phylogenies* (Vol. 2). Sinauer Associates.
- Fenwick, P. M. (1994). A new data structure for cumulative frequency tables. *Software: Practice and Experience*, 24(3), 327–336. <https://doi.org/10.1002/spe.4380240306>
- Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6), 1635–1638. <https://doi.org/10.1093/molbev/msw046>

- 114 Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age.  
115 *Nature Reviews Genetics*, 21(7), 428–444. <https://doi.org/10.1038/s41576-020-0233-0>
- 116 Khurana, M. P., Curran-Sebastian, J., Scheidwasser, N., Morgenstern, C., Rasmussen, M.,  
117 Fonager, J., Stegger, M., Tang, M.-H. E., Juul, J. L., Escobar-Herrera, L. A., & oth-  
118 ers. (2024). High-resolution epidemiological landscape from ~ 290,000 SARS-CoV-2  
119 genomes from Denmark. *Nature Communications*, 15(1), 7123. <https://doi.org/10.1038/s41467-024-51371-0>
- 120
- 121 Köster, J. (2016). Rust-bio: A fast and safe bioinformatics library. *Bioinformatics*, 32(3),  
122 444–446. <https://doi.org/10.1093/bioinformatics/btv573>
- 123 Kraemer, M. U., Tsui, J. L.-H., Chang, S. Y., Lytras, S., Khurana, M. P., Vanderslott, S., Bajaj,  
124 S., Scheidwasser, N., Curran-Sebastian, J. L., Semenova, E., & others. (2025). Artificial  
125 intelligence for modelling infectious disease epidemics. *Nature*, 638(8051), 623–635.  
126 <https://doi.org/10.1038/s41586-024-08564-w>
- 127 Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based Python JIT compiler.  
128 *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 1–6.  
129 <https://doi.org/10.1145/2833157.2833162>
- 130 McBroome, J., Bernardi Schneider, A. de, Roemer, C., Wolfinger, M. T., Hinrichs, A. S.,  
131 O'Toole, A. N., Ruis, C., Turakhia, Y., Rambaut, A., & Corbett-Detig, R. (2024). A  
132 framework for automated scalable designation of viral pathogen lineages from genomic  
133 data. *Nature Microbiology*, 9(2), 550–560. <https://doi.org/10.1038/s41564-023-01587-5>
- 134 Moreno, M. A., Holder, M. T., & Sukumaran, J. (2024). DendroPy 5: a mature Python  
135 library for phylogenetic computing. *Journal of Open Source Software*, 9(101), 6943.  
136 <https://doi.org/10.21105/joss.06943>
- 137 Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree  
138 reconciliation. *Trends in Ecology & Evolution*, 28(12), 719–728. <https://doi.org/10.1016/j.tree.2013.09.004>
- 139
- 140 Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics  
141 and evolutionary analyses in R. *Bioinformatics*, 35, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- 142
- 143 Penn, M. J., Scheidwasser, N., Khurana, M. P., Duchêne, D. A., Donnelly, C. A., & Bhatt, S.  
144 (2024). Phylo2Vec: A vector representation for binary trees. *Systematic Biology*, syae030.  
145 <https://doi.org/10.1093/sysbio/syae030>
- 146 Penn, M. J., Scheidwasser, N., Penn, J., Donnelly, C. A., Duchêne, D. A., & Bhatt, S. (2023).  
147 Leaping through tree space: Continuous phylogenetic inference for rooted and unrooted trees.  
148 *Genome Biology and Evolution*, 15(12), evad213. <https://doi.org/10.1093/gbe/evad213>
- 149 Richman, H., Zhang, C., & IV, F. A. M. (2025). *Vector encoding of phylogenetic trees by*  
150 *ordered leaf attachment*. <https://doi.org/10.48550/arXiv.2503.10169>
- 151 Sanderson, T. (2021). Chronumetal: Time tree estimation from very large phylogenies. In  
152 *bioRxiv* (pp. 2021–2010). Cold Spring Harbor Laboratory. <https://doi.org/10.1101/2021.10.27.465994>
- 153
- 154 Sanderson, T. (2022). Taxonium, a web-based tool for exploring large phylogenetic trees.  
155 *eLife*, 11. <https://doi.org/10.7554/eLife.82392>
- 156 Suchard, M. A., & Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics.  
157 *Bioinformatics*, 25(11), 1370–1376. <https://doi.org/10.1093/bioinformatics/btp244>
- 158 Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler,  
159 D., & Corbett-Detig, R. (2021). Ultrafast sample placement on existing tRees (USHER)  
160 enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, 53(6),

- 161 809–816. <https://doi.org/10.1038/s41588-021-00862-7>
- 162 Wickham, H. (2011). Testthat: Get started with testing. *The R Journal*, 3(1), 5–10.  
163 <https://doi.org/10.32614/rj-2011-002>
- 164 Yang, Z. (2014). *Molecular evolution: A statistical approach*. Oxford University Press.  
165 <https://doi.org/10.1093/acprof:oso/9780199602605.001.0001>

DRAFT