Toward Phylogenetic Inference of Evolutionary Dynamics at Scale

Matthew Andres Moreno [2], Emily Dolson [1], and Santiago Rodriguez-Papa [1]

[1]Computer Science and Engineering
Michigan State University
East Lansing, United States
[2]Ecology and Evolutionary Biology
University of Michigan
Ann Arbor, United States
morenoma@umich.edu

Abstract

As digital evolution systems grow in scale and complexity, observing and interpreting their evolutionary dynamics will become increasingly challenging. Distributed and parallel computing, in particular, introduce obstacles to maintaining the high level of observability that makes digital evolution a powerful experimental tool. Phylogenetic analyses represent a promising tool for drawing inferences from digital evolution experiments at scale. Recent work has introduced promising techniques for decentralized phylogenetic inference in parallel and distributed digital evolution systems. However, foundational phylogenetic theory necessary to apply these techniques to characterize evolutionary dynamics is lacking. Here, we lay the groundwork for practical applications of distributed phylogenetic tracking in three ways: 1) we present an improved technique for reconstructing phylogenies from tunably-precise genome annotations, 2) we begin the process of identifying how the signatures of various evolutionary dynamics manifest in phylogenetic metrics, and 3) we quantify the impact of reconstruction-induced imprecision on phylogenetic metrics. We find that selection pressure, spatial structure, and ecology have distinct effects on phylogenetic metrics, although these effects are complex and not always intuitive. We also find that, while low-resolution phylogenetic reconstructions can bias some phylogenetic metrics, high-resolution reconstructions recapitulate them faithfully.

Introduction

Artificial life systems are a powerful technique for studying evolution and have yielded many important insights (Wilke et al., 2001; Zaman et al., 2014; Goldsby et al., 2014). Much of this power derives from the fact that these systems provide the ability to perfectly observe evolutionary dynamics. In particular, research advances have often been made possible by analyzing phylogenetic data on the evolutionary trajectories that lead to a given outcome (Lenski et al., 2003; Lalejini and Ofria, 2016; Johnson et al., 2022).

While these analyses are powerful, phylogenetic data can quickly become large and unwieldy. Thus, as artificial life systems and the questions we pose become more complex, better tools are needed. These come in two forms: 1) evolutionary summary statistics that can be used to draw abstract insights from lineages and phylogenies (Dolson et al., 2020), and 2) algorithmic techniques for efficiently collecting phylogenetic data at scale (Moreno et al., 2022a). Given that phylogenies are an abstraction that generalizes across all evolutionary contexts, they facilitate drawing conclusions that scale across different artificial life systems and biology.

Indeed, phylogeny-based metrics have already been used for varied purposes such as identifying hallmarks of open-ended evolution (Dolson et al., 2019), predicting which runs of evolutionary computation will be successful (Hernandez et al., 2022; Shahbandegan et al., 2022), quantifying patterns of tumor evolution (Scott et al., 2020; Lewinsohn et al., 2023), and identifying priorities for conservation (Forest et al., 2007). Much of the early work on quantifying phylogenetic properties has its origins in conservation biology and paleontology literature, as researchers in these fields attempted to draw inferences about the past from limited data. More recently, medical researchers have embraced measuring phylogenies in real time in an effort to understand the evolution of pathogens and cancer.

Consequently, researchers across many fields have an interest in inferring the processes that shaped a phylogeny by quantifying its topology. Currently, most research on this topic is either very abstract (e.g., the presence of negative-frequency dependent selection increases phylogenetic diversity) or very specific (e.g., phylogenetic structure can be used to infer cell division rates in solid tumors (Lewinsohn et al., 2023)). Here, we lay the groundwork for identifying the fingerprints left on phylogenetic structure by a larger range of evolutionary dynamics in a scalable and robust manner.

Historically, digital evolution systems have had the advantage of providing perfectly accurate phylogenetic data, in contrast to the reconstructed phylogenies relied upon in traditional biology. However, as digital evolution systems scale, issues of data loss and decentralization will make perfect tracking at best inefficient and at worst untenable. Thus, some systems will likely need to adopt a decentralized, reconstruction-based approach similar to biological data. Consequently, if our goal is to use phylogenetic metrics to increase the scalability of our data analysis, we must also understand how robust they are to inaccuracies introduced by reconstruction. Phylogenetic reconstructions in artificial life can be achieved through the recently-developed "hereditary stratigraphy" approach, a lightweight annotation scheme that can give tunably-precise information about the phylogenetic history between any two extant organisms.

In this paper, we build the following methodological and theoretical foundations that will be needed to use phylogenetic analyses to observe evolutionary dynamics in complex, distributed artificial life systems:

1. the ability to perform fast, accurate tree reconstructions for

very large distributed populations,
2. understanding of the relationship between evolutionary dynamics and phylogenetic metrics,
3. quantification of the effects of reconstruction-induced estimation error on phylogenetic structure, both in terms of the amount of precision required for accuracy and the amount of bias introduced by inadequate precision, and
4. quantification of the phylogenetic effects of spatial structure, as distributed computation does not support well-mixed populations.

Methods

Model System

Experiments testing the relationships between evolutionary dynamics, reconstruction error, and phylogenetic structure required a model system amenable to direct, interpretable tuning of ecology, spatial structure, and selection pressure. Additionally, in order to make findings relevant to large-scale phylogenetic analyses, computational efficiency was necessary to facilitate large population size and high generation counts. Finally, a parsimonious and generic model system was desired so that findings would better generalize across digital evolution systems.

A parsimonious model system was devised to fulfill these objectives. Genomes in this system comprised a single floating-point value, with higher magnitude corresponding to higher fitness. Population size 32,768 (2^{15}) was used for all experiments. Selection was performed using tournament selection with synchronous generations. Treatments' selection pressure was controlled via tournament size. Mutation was applied after selection, with a value drawn from a unit Gaussian distribution added to all genomes. Evolutionary runs were ended after 262,144 (2^{18}) generations. Each run required around 4 hours of compute time.

Treatments incorporating spatial structure used a simple island model. In spatially structured treatments, individuals were evenly divided among 1,024 islands and only competed in selection tournaments against sympatric population members. Islands were arranged in a one-dimensional closed ring and 1% of population members migrated to a neighboring island each generation.

Treatments incorporating ecology used a simple niche model. Population slots were split evenly between niches. Organisms were arbitrarily assigned to a niche at genesis and were only allowed to occupy population slots assigned to that niche. Therefore, individuals exclusively participated in selection tournaments with members of their own niche. In treatments also incorporating spatial structure, an even allotment of population slots was provided for every niche on every island. Every generation, individuals swapped niches with probability $3.0517578125 \times 10^{-8}$ (chosen so one niche swap would be expected every 1,000 generations).

For our main experiments, we defined the following "regimes" of evolutionary conditions:

- *plain*: tournament size 2 with no niching and no islands,
- *weak selection*: tournament size 1 with no niching and no islands,
- *strong selection*: tournament size 4 with no niching and no islands,
- *spatial structure*: tournament size 2 with no niching and 1,024 islands,
- *weak 4 niche ecology*: tournament size 2 with 4 niches and niche swap probability increased 100×,
- *4 niche ecology*: tournament size 2 with 4 niches, and
- *8 niche ecology*: tournament size 8 with 4 niches.

In follow-up experiments testing ecological dynamics with a spatial background, we defined the following additional evolutionary "regimes:"

- *plain*: tournament size 2 with no niching over 1,024 islands,
- *weak 4 niche ecology*: tournament size 2 with 4 niches and niche swap probability increased 100× over 1,024 islands,
- *4 niche ecology*: tournament size 2 with 4 niches over 1,024 islands, and
- *8 niche ecology*: tournament size 8 with 4 niches over 1,024 islands.

Finally, to foster generalizability of findings, all experiments were performed with two alternate "sensitivity" variables: evolutionary length in generations and mutation operator. Shorter runs of 32,768 and 98,304 generations were tested in addition to the full-length runs. We refer to full-length runs as completing "epoch 7" and the shorter runs as completing "epoch 0" and "epoch 2" respectively. One additional mutation operator was tested to contrast the unit Gaussian distribution: the unit exponential distribution. Under this distribution, deleterious mutations are not possible and large-effect mutations are more likely.

Across all experiments, each treatment comprised 50 replicates.

Hereditary Stratigraphic Annotations and Tree Reconstruction

Experiments testing the impact of phylogenetic inference error on phylometrics employ the recently-developed "hereditary stratigraphy" technique to facilitate phylogenetic inference (Moreno et al., 2022c). This technique works by attaching heritable annotations to individual digital genomes. Every generation, a new random "fingerprint" is generated and appended to the individuals' inherited annotations. To reconstruct phylogenetic history, fingerprints from extant organisms' annotations can be compared. Where two organisms share identical fingerprints along the record, they likely shared common ancestry. Mismatching fingerprints indicate a split in compared organisms' ancestry.

Hereditary stratigraphy enables a tunable trade-off between annotation size and estimation accuracy. Fingerprints may be discarded to decrease annotation size at the cost of reduced density of reference points to test for common (or divergent) ancestry along organisms' generational histories.

We test four levels of fingerprint retention. Each level is described as a $p\%$ "resolution" meaning that the generational distance between reference points any number of generations k back is less than $(p/100) \times k$. So, a high percentage p indicates coarse resolution and a low percentage p indicates fine resolution. In detail, at the conclusion of 262,144 generation evolutionary runs,

- at 33% resolution 68 fingerprints are retained per genome,
- at 10% resolution 170 fingerprints are retained per genome,
- at 3% resolution 435 fingerprints are retained per genome, and
- at 1% resolution 1,239 fingerprints are retained per genome.

This work uses 1 byte fingerprints, which collide with probability $1/256$. Greater space efficiency could be achieved using 1 bit fingerprints. However, this would require careful accounting for ubiquitous generation of identical fingerprints by chance and is left to future work.

Previous work with hereditary stratigraphy used UPGMA distance-based reconstruction techniques (Moreno et al., 2022b). Large-scale reconstructions required for these experiments necessitated development of a more efficient technique that did not require all pairs (i.e., $O(n^2)$ distance comparison. To accomplish this, we devised an agglomerative tree building algorithm that works by successively adding leaf organism annotations and percolating them down from the tree root along the tree path of internal nodes consistent with their fingerprint sequence, then affixing them where common ancestry ends. This new tree-building approach reduced compute time from multiple hours to around 5 minutes in most cases. Implementation materials providing the full details of this approach are included in Listing 4.

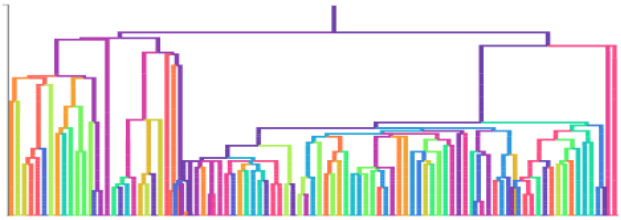
To assess the efficacy of the new agglomerative tree-building approach, we calculated all reconstructed trees' quartet distance to their respective reference. Quartet distance ranges from 0 (between identical trees) to 0.75 (between random trees), providing in this case a measure of reconstruction error. As expected, this measure of reconstruction error varied significantly with resolution for trees across all evolutionary regimes (Kruskal-Wallis tests; all $p < 10^{-20}$; Supplementary Table 2). Reconstruction error also varied significantly with evolutionary regime for each reconstruction resolution level (Kruskal-Wallis tests; all $p < 10^{-8}$; Supplementary Table 1).

For 3% and 1% resolutions, mean reconstruction error was less than 0.01 in all cases and at 10% resolution mean reconstruction error was less than 0.05 in all cases. At 33% resolution, mean reconstruction error was less than 0.12 in all cases. The largest reconstruction errors observed at 1%, 3%, 10%, and 33% resolutions were, respectively, 0.051 (weak selection regime), 0.093 (weak 4 niche ecology regime), 0.14 (plain evolutionary regime), and 0.45 (plain evolutionary regime). Supplementary Table 3 reports mean, median, standard deviation, and maxima for reconstruction error across surveyed evolutionary conditions.

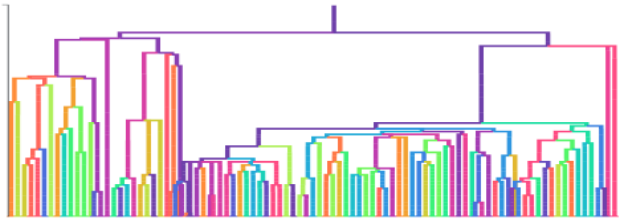
To generate reconstructed trees in experiments, we simulated the inheritance of hereditary stratigraphic annotations along a reference phylogeny to yield the set of annotations that would be attached to extant population members at the end of a run, then used our agglomerative tree building technique to infer. Thus, each reconstruction replicate has a directly-corresponding reference tree from a perfect-tree treatment replicate. Figure 1 shows a reference tree and corresponding reconstructions performed using 1%, 3%, 10%, and 33% resolution hereditary stratigraph annotations.

Phylometrics

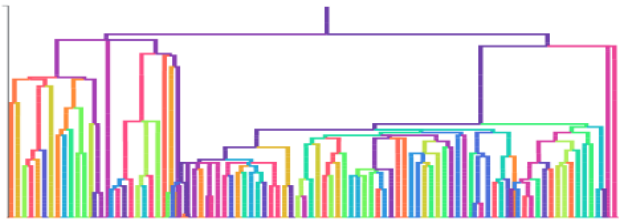
A wide range of metrics exists for quantifying the topology of a phylogeny. Tucker et. al showed that these metrics can be classified into the following three dimensions: richness, divergence, and regularity (Tucker et al., 2017). Richness metrics quantify the amount of phylogenetic diversity/evolutionary history represented by a phylogeny. Divergence metrics quantify how different the units of the phylogeny are from each other.



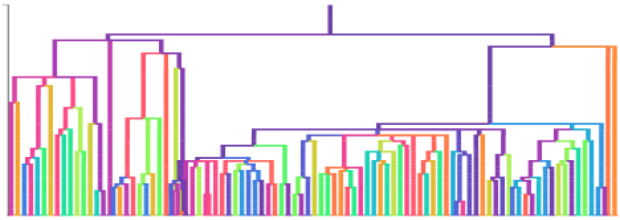
(a) reference tree



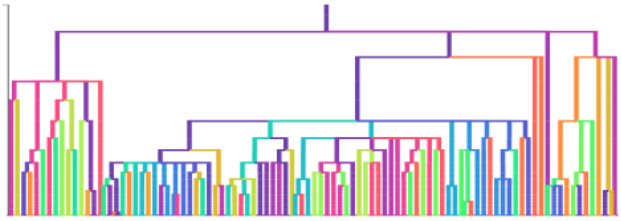
(b) 1% resolution



(c) 3% resolution



(d) 10% resolution



(e) 33% resolution

Figure 1: Comparison of phylogeny reconstructions produced by the agglomerative trie-based algorithm across different hereditary stratigraphy resolutions for the plain evolutionary regime. To maintain visual legibility, these trees contain the same sub-sample of 100 leaf nodes out of the 32,768 in the full trees. Sub-figures are arranged from top to bottom in coarsening order of reconstruction resolution. Taxon and branch color coding is consistent across subpanels. Visit `mmore500.com/hstrat-evolutionary-inference/` for mouseover-based highlighting of corresponding clades between reconstructions and reference.

Regularity metrics quantify the variance of other properties (i.e. how consistent they are across the phylogeny). Here, we focus on four metrics spread across these categories:

Number of Internal Nodes: This measurement simply counts the number of non-extant taxa in the phylogeny (i.e. the number of non-leaf nodes). Because the reconstructed trees do not contain any internal nodes not associated with a branching point (i.e. unifurcations), we strip such nodes out of the reference trees as well. It is a metric of phylogenetic richness and is closely related to Faith's classic phylogenetic diversity metric (Faith, 1992), the primary differences being that the number of internal nodes is unaffected by branch lengths and the number of leaf nodes. Consequently, we would expect it to be increased by the presence of ecology or spatial structure, as both these factors increase diversity.

Colless-like Index: The original Colless Index (Colless, 1982), also often referred to as I_c (Shao, 1990), is a measure of tree imbalance (i.e. it gets higher as the tree gets less balanced). In the context of Tucker et. al's framework, it is a regularity metric. However, the traditional Colless Index only works for strictly bifurcating trees. As our trees have multifurcations, we instead use the Colless-like Index, which is an extension of the Colless Index to multifurcating trees (Mir et al., 2018). Tree imbalance is thought to be associated with varying ecological pressures (Chamberlain et al., 2014; Burress and Tan, 2017) and has also been observed to increase in the presence of spatial structure (Scott et al., 2020).

Mean Pairwise Distance: This metric is calculated by computing the shortest distance between all pairs of leaf nodes and taking the mean of these values (Webb and Losos, 2000). Note that these distances are measured in terms of the number of nodes in between the pair, not in terms of branch lengths. Mean pairwise distance is a metric of evolutionary divergence (Tucker et al., 2017). Mean pairwise distance should be increased by scenarios that promote the long-term maintenance of distinct phylogenetic branches, such as ecology. Conversely, factors that act to reduce diversity should also reduce mean pairwise distance.

Mean Evolutionary Distinctiveness: Evolutionary distinctiveness is a metric that can be calculated for individual taxa to quantify how evolutionarily different that taxon is from all other taxa in the phylogeny (Isaac et al., 2007). To get mean evolutionary distinctiveness, we average this value across all extant taxa in the tree. Like mean pairwise distance, mean evolutionary distinctiveness is a metric of evolutionary divergence. However, it is known to capture substantially different information than mean pairwise distance (Tucker et al., 2017). Unlike our other metrics, evolutionary distinctiveness is heavily influenced by branch length. We generally expect mean evolutionary distinctiveness to be increased by similar factors to mean pairwise distance.

Software and Data Availability

Software, configuration files, and executable notebooks for this work are available at `https://github.com/mmore500/hstrat-evolutionary-inference`. Data and supplemental materials are available via the Open Science Framework `https://osf.io/vtxwd/` (Foster and Deardorff, 2017).

All hereditary stratigraph annotation, reference phylogeny generation, and phylogenetic reconstruction tools used in this work are published in the `hstrat` Python package (Moreno et al., 2022c). This project can be visited at `https://github.com/mmore500/hstrat`.

This project uses data formats and tools associated with the ALife Data Standards project (Lalejini et al., 2019) and benefited from many pieces of open-source scientific software (Ofria et al., 2020; Sand et al., 2014; Virtanen et al., 2020; Harris et al., 2020; pandas development team, 2020; Wes McKinney, 2010; Sukumaran and Holder, 2010; Cock et al., 2009).

Results and Discussion

Phylometric Signatures of Evolutionary Dynamics

The feasibility of harnessing phylogenetic analysis to detect evolutionary dynamics in digital evolution systems hinges on the premise that these dynamics induce detectable structure within the phylogenetic record. Indeed, as shown in Figure 3, dendrograms of phylogenetic histories from different evolutionary conditions exhibit striking differences.

As a first step to characterizing the phylogenetic impact of spatial structure, ecology, and selection pressure, we first tested if surveyed evolutionary conditions exhibited detectable differences in evolutionary distinctiveness, Colless-like index, pairwise distance, and ancestor count. Figure 3 summarizes the distributions of each metric across surveyed conditions. Statistical tests confirmed that each phylometric exhibited significant variation among surveyed evolutionary conditions (Kruskal-Wallis tests; all $p < 10^{-50}$; $n = 50$ per condition; Supplementary Table S1).

To better understand, we performed an all-pairs comparison of each phylometric among the seven surveyed evolutionary regimes (Wilcoxon tests with Bonferroni correction; corrected significance threshold 1.49×10^{-4}; $n = 50$ per condition; 84 comparisons per sensitivity analysis configuration; 336 comparisons total; Supplementary Table 6).

The Colless-like index is significantly depressed under all evolutionary regimes compared to the plain regime with no spatial structure, no ecology, and moderate selection pressure. Reduction in this statistic indicates that all deviations from baseline conditions increased regularity in generated phylogenies. This observation runs somewhat counter to prior results on similar tree balance metrics, in which the presence of spatial structure increased imbalance (Scott et al., 2020). Application of the metric to trees comprised individual-level taxa, instead of species-level taxa as is the case in most traditional phylogenetics work, may account for this result. Further investigation will be warranted to fully explain this outcome.

Compared to the plain regime, observed mean pairwise distance was significantly lower under strong selection and significantly higher under weak selection and with spatial structure. None of the ecological regimes induced significant changes in the mean pairwise distance phylometric compared to the plain regime.

Similarly, ancestor account had no significant relationship with ecology. However, weak selection and spatial structure significantly increased ancestor count and strong selection significantly decreased it.

Finally, mean evolutionary distinctiveness was significantly increased under all three ecological regimes compared to baseline. Additional, ecological intensity was significantly associated with

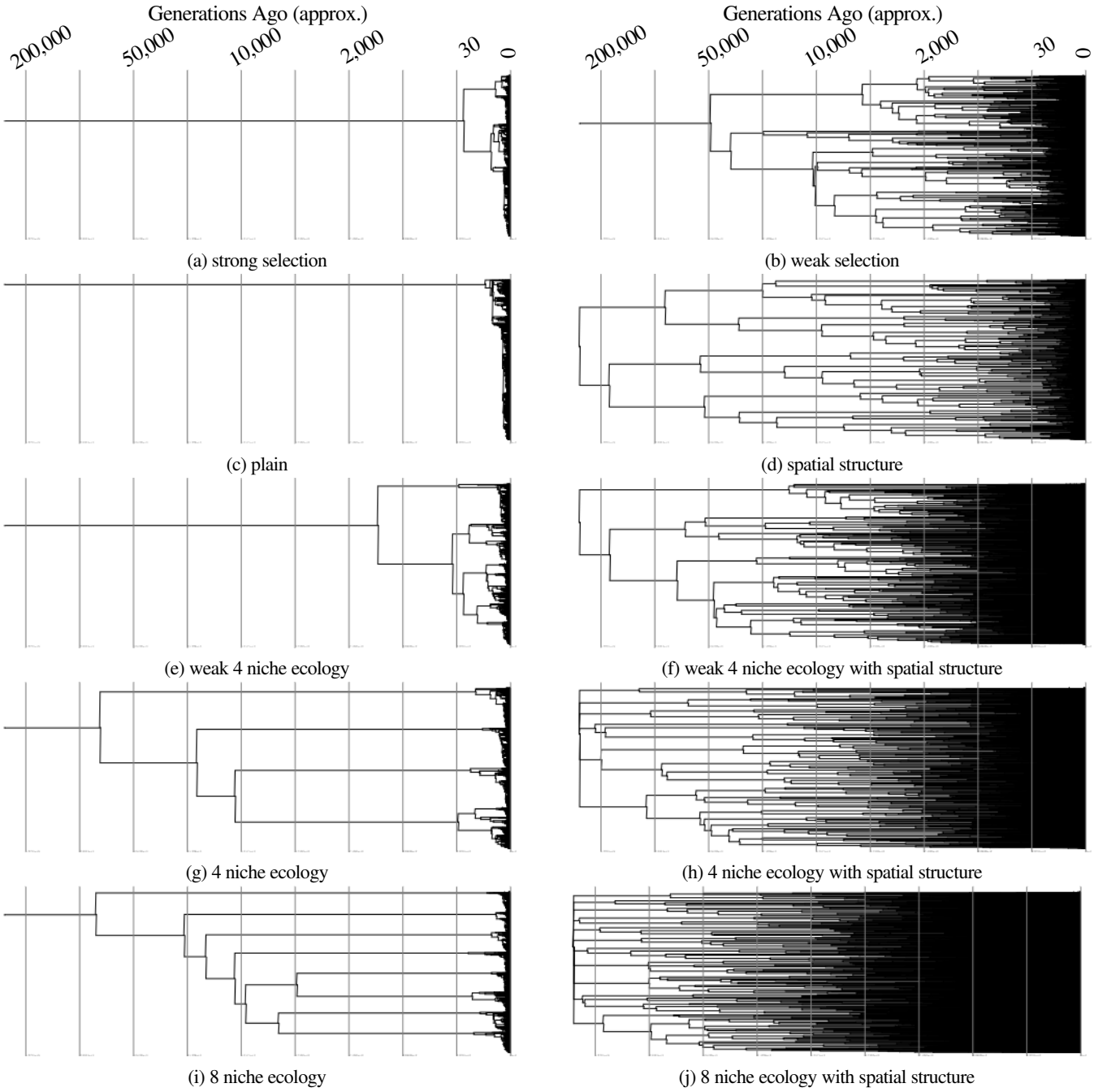


Figure 2: Sample reference phylogenies across surveyed evolutionary metrics. Each phylogeny has 32,768 leaves. Note log-scale x axis.

mean evolutionary distinctiveness, with weak 4 niche ecology having the lowest value for this phylometric and 8 niche ecology having the highest. However, evolutionary distinctiveness was more strongly driven by weak selection and even more strongly driven by spatial structure, by significant margins. Finally, strong selection significantly depressed mean evolutionary distinctiveness.

Figure 4 provides a high-level overview of the magnitude and direction of each regime's effects on evolutionary metrics compared to the plain regime. Notably, strong and weak selection both significantly decrease Colless-like index and mean pairwise distance but have opposite effects on ancestor count and mean evolutionary distinctiveness. Colless-like distance appears to be the least useful metric in distinguishing evolutionary dynamics, decreasing under all non-plain evolutionary conditions.

Ecological dynamics have significant, but relatively weak, influence on the surveyed phylometrics. So, it appears careful accounting for other evolutionary dynamics (i.e., selection pressure and spatial structure) will be essential to accurate detection of

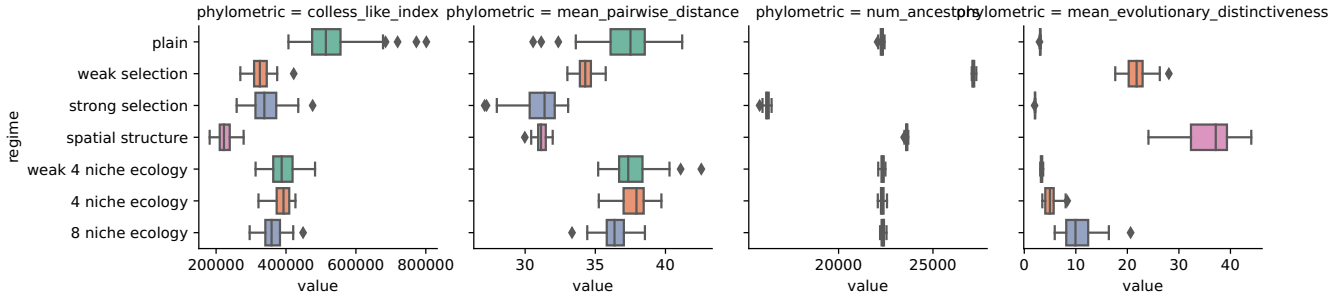


Figure 3: Distribution of tree phylometrics measured with perfect phylogenetic tracking across surveyed evolutionary regimes. Sample sizes of $n = 50$ replicates define each depicted distribution.

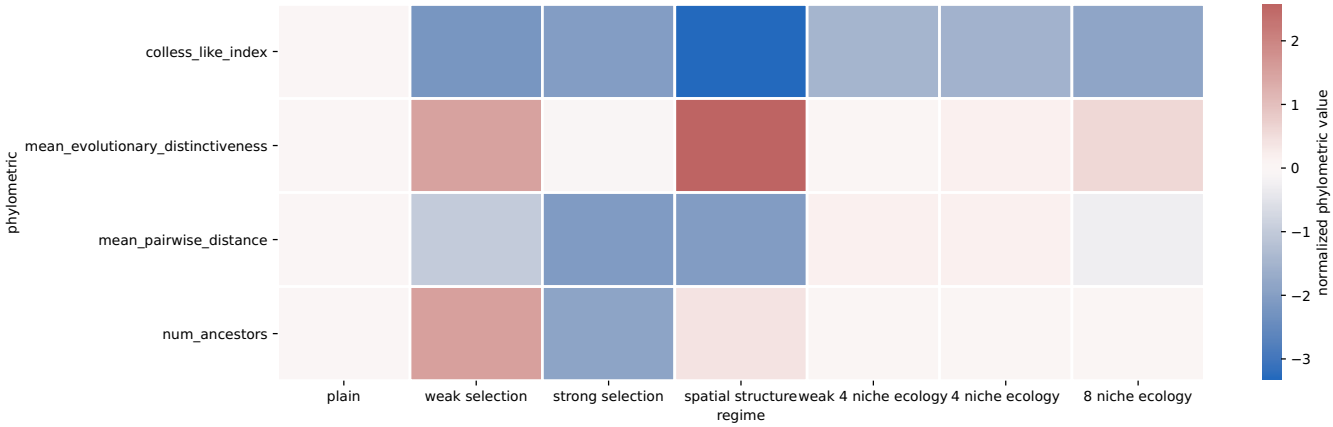


Figure 4: Heatmap of normalized tree phylometrics across surveyed evolutionary regimes, calculated on perfect-fidelity simulation phylogenetic records. See Supplementary Figure 16 for results under sensitivity analysis conditions.

ecology through phylogenetic analysis. Ancestor count and mean pairwise distance may play a role in identifying ecological dynamics, as ecological dynamics — in contrast to other factors such as spatial structure and changes in selection pressure — have little to no discernible effect on these phylometrics. Alternately, in future work it may be possible to develop phylometrics that respond more strongly — and more exclusively — to ecological dynamics.

We performed a sensitivity analysis over an alternate exponential mutation operator and earlier phylogeny sampling timepoints. We found the effects of evolutionary conditions on phylometrics to be generally consistent across surveyed conditions (Supplementary Figures 8 and 16; Supplementary Tables 6 **??**).

Phylometric Signatures of Ecological Dynamics in Spatially Structured Populations

At large scale, digital evolution populations will almost inevitably integrate spatial structure due to practical limitations of distributed computing hardware (Ackley and Small, 2014). Therefore, understanding the background effects of spatial structure on the phylogenetic signatures of other evolutionary dynamics will be essential to applications of phylogenetic inference in such applications. For this analysis, we chose to focus on ecological dynamics due to interest in how their relatively weak phylometric signatures would respond to the relatively strong influence of spatial structure.

Figure 5 summarizes the distribution of surveyed phylometrics under the three surveyed ecological regimes and the control non-ecological regime, all with spatial population structure. Statistical tests confirmed that each phylometric exhibited significant variation among these evolutionary regimes (Kruskal-Wallis tests; all $p < 1 \times 10^{-20}$; $n = 50$ per condition; Supplementary Table 10).

To explore the nature of this variation, we performed all-pairs comparisons for each phylometric among the four surveyed regimes (Wilcoxon tests with Bonferroni correction; corrected significance threshold 5.26×10^{-4}; $n = 50$ per condition; 24 comparisons per sensitivity analysis configuration; 96 comparisons total; Supplementary Table 7). Like under the spatially unstructured background, ecology drove significant increases in mean evolutionary distinctiveness. Unlike the spatially unstructured background, though, ancestor count under ecological conditions sharply exceeded ancestor count under non-ecological conditions. Ancestor count significantly increased between the 4 niche ecological regimes and the 8 niche ecological regime, as well. Additionally, the 8 niche regime significantly decreased mean pairwise distance under spatially-structured conditions. This was not the case without spatial structure.

Spatial structure appears to mediate these aspects of ecological phylogenetic structure, which do not appear in its absence.

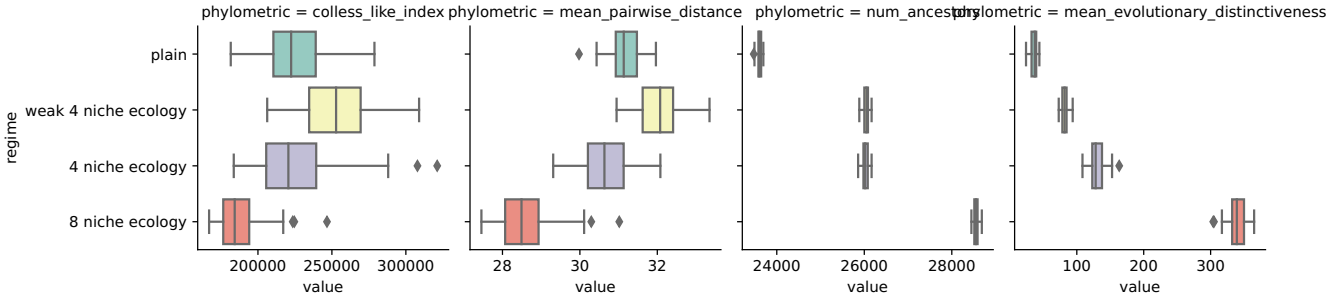


Figure 5: Distribution of phylometrics across the three surveyed ecological regimes and the control non-ecological regime—all with spatial population structure (i.e., island count 1,024). Phylometrics were calculated on perfect-fidelity simulation phylogenetic records. Results are for standard experimental conditions: gaussian mutation distribution at epoch 7 (generation 262,144). See Figure 15 for results under sensitivity analysis conditions. Sample sizes of $n = 50$ replicates define each depicted distribution.

However, spatial structure mutes the effects of ecology on the Colless-like index. Only the 8 niche regime significantly decreased this phylometric.

For these experiments, we again performed a sensitivity analysis over an alternate exponential mutation operator and earlier phylogeny sampling timepoints. We found the effects of evolutionary conditions on phylometrics to be generally consistent across surveyed conditions (Supplementary Figure 15 and Supplementary Tables 7 and 10).

Phylometric Bias of Reconstruction Error

Shifting from perfect phylogenetic tracking to approximate phylogentic reconstruction will facilitate efficiency and robust digital evolution simulations at scale, but introduces a complicating factor into phylogenetic analyses: tree reconstruction error. A clear understanding of the impact of these errors on the computed phylometrics will be necessary to ensure accurate phylogenetic analyses.

To explore this question, we compared phylometrics computed on reconstructed trees to corresponding true reference trees (Wilcoxon tests; $n = 50$ per condition; Supplementary Table 8). To err towards conservatism in detecting phylometric biases, we did not correct for multiple comparisons. Reconstructions were performed across a range of precisions, ranging from 1% relative resolution for MRCA estimates (most precise) to 33% relative resolution for MRCA estimates (least precise). Precision was manipulated by adjusting the information content of underlying hereditary stratigraphic genome annotations used to perform phylogenetic reconstruction (Moreno et al., 2022b).

For each phylometric, we sought to determine the minimum resolution required to achieve statistical non-detection (i.e., $p > 0.05$) of bias between reconstructions and their corresponding references. For most phylometrics, 3% reconstruction resolution was sufficient to achieve statistical indistinguishability between reference and reconstruction. Mean evolutionary distinctiveness was particularly robust to reconstruction error, showing no detectable bias even at only 33% reconstruction resolution. Ancestor count was highly sensitive to reconstruction error; in nearly all cases bias was still detectable at 1% reconstruction resolution. This may be due to overabundance of polytomies in reconstructed trees due to aggregation of ancestors that cannot be resolved due to estimation uncertainty. In future work, postprocessing reconstructed trees to break polytomies into arbitrary sets of bifurcations may reduce this metric's sensitivity to reconstruction.

Phylometric sensitivity to reconstruction error was broadly consistent across evolutionary regimes. Figure 6 summarizes these results.

Where detectable, estimation uncertainty bias decreased all surveyed phylometrics' numerical value. So, when testing for expected increases in phylometric values, the potential for systematic false positives due to reconstruction error can be discounted. Supplementary Figure 19 provides a full comparison the distribution of phylometric estimates on reference trees with the distributions of phylometric estimates for reconstructed trees across reconstruction resolutions.

We performed additional analyses with additional spatially structured ecological evolutionary regimes, over an alternate exponential mutation operator, and at earlier phylogeny sampling timepoints. These yielded generally similar findings for the relationship between reconstruction error and phylometric bias (Supplementary Figures 17; 21, and 20; Supplementary Table 7 and 8).

Conclusion

Traditional phylogenetic analysis revolves around accurately and precisely resolving the sequences of evolutionary events that comprise natural history. Ongoing work within the field seeks to augment this oeuvre by developing methods to extract information about evolutionary dynamics such as ecology, selection pressure, and spatial structure from these inferred records. This work contributes to the project of expanding the scope of traditional phylogenetic analysis, seeking in particular to develop the methodological and theoretical foundations that will be needed to apply phylogenetic analyses to observe evolutionary dynamics in complex, distributed artificial life systems.

First, we characterized the effects of selection pressure, ecology, and spatial population structure on phylogeny structure. Each evolutionary dynamic expressed a distinct signature across phylometrics. However, compared to spatial structure and selection pressure, ecology exerted relatively muted effects on phylogenetic metrics. Ecology had no detectable effect on ancestor count and mean pairwise distance. Ecological influence

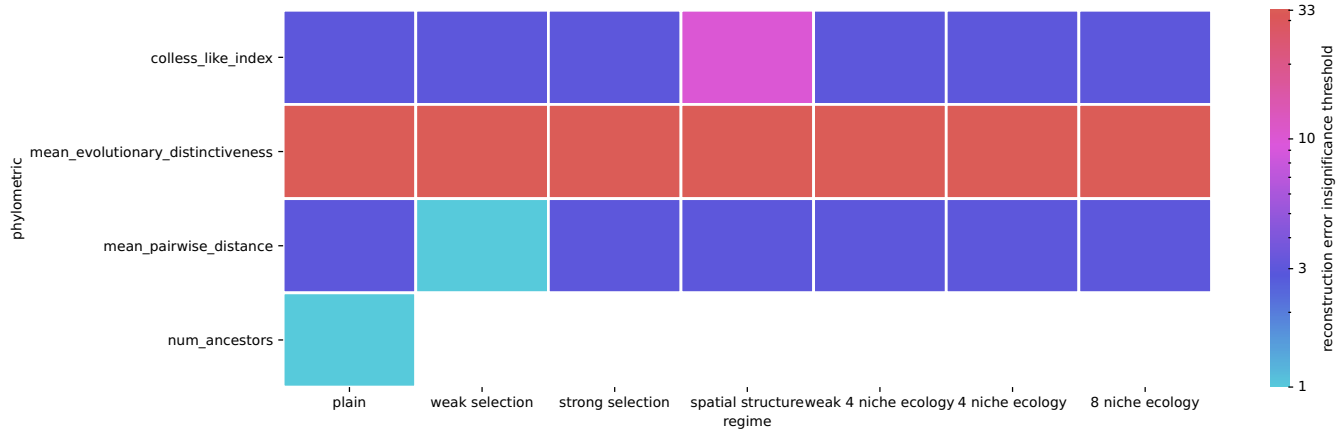


Figure 6: Reconstruction resolutions required to achieve statistical indistinguishability between reconstructions corresponding reference trees for each phylometric across surveyed evolutionary conditions. Significance level $p < 0.05$ under the Wilcoxon signed-rank test between samples of 50 replicates each is used as the threshold for statistical distinguishability. Phylometrics with looser reconstruction resolution thresholds (i.e., higher resolution percentages) are less sensitive to reconstruction error. White heat map tiles indicate that no surveyed reconstruction resolution threshold was sufficient to achieve indistinguishability from the reference tree with respect a particular phylometric. See Supplementary Figure 17 for sensitivity analysis results.

on mean evolutionary distinctiveness was significant, but small in magnitude compared to effects from the introduction of spatial structure and increased selection pressure.

Surprisingly, follow-up experiments revealed that background spatial population structure can accentuate the phylometric signature of ecology. Under these conditions, ecology induced sharp increases in ancestor count. However, the impact of ecology on Colless-like index attenuated with the addition of spatial structure. These results highlight the complexity of how interacting evolutionary dynamics impact phylogenetic structure. Even a single dynamic in isolation can influence phylometrics in opposite directions. For instance, we found that both increasing and decreasing selection pressure can significantly reduce trees' Colless-like index and mean pairwise distance.

Comparing phylometrics of reference trees against corresponding reconstructions, we found that most phylometric statistics were somewhat sensitive to reconstruction error. Ancestor count was the most sensitive, with reconstructions exhibiting significant bias compared to corresponding references at even 1% reconstruction resolution. On the other hand, mean evolutionary distinctiveness was particularly robust to reconstruction error. No bias was detected for calculations of this metric on reconstructed trees, even when reconstructed with only 33% resolution. Most phylometrics reached statistical indistinguishability between reference and reconstruction at or above 3% reconstruction resolution, suggesting a reasonable ballpark parameterization for applications of hereditary stratigraphy involving phylogenetic analysis. Additionally, phylometric bias of reconstruction error was generally consistent across different evolutionary regimes, which offers a promising simplification of future experimental considerations.

These findings contribute foundations for development of more rigorous phylogenetic assays that might eventually be capable of identifying the evolutionary conditions that produced a phylogeny. In particular, the potentially confounding impact of spatial structure and reconstruction error, which before had been poorly characterized, will be relevant to distributed digital evolution systems at scale. It is especially promising that reconstruction error can be reduced to effectively zero.

Finally, this work introduced, validated, and demonstrated a new phylogenetic reconstruction technique for large-scale populations annotated using hereditary stratigraphy. This new capability clears an important hurdle to applying hereditary stratigraphy in practice. Ultimately, the findings presented here represented a multi-pronged attack on the problem of drawing scalable evolutionary inferences from artificial life software.

Acknowledgment

This research was supported in part by NSF grants DEB-1655715 and DBI-0939454 as well as by Michigan State University through the computational resources provided by the Institute for Cyber-Enabled Research. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1424871. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ackley, D. and Small, T. (2014). Indefinitely scalable computing= artificial life engineering. In *ALIFE 14: The Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 606–613. MIT Press.

Burress, E. D. and Tan, M. (2017). Ecological opportunity alters the timing and shape of adaptive radiation. *Evolution*, pages n/a–n/a.

Chamberlain, S., Vázquez, D. P., Carvalheiro, L., Elle, E., and Vamosi, J. C. (2014). Phylogenetic tree shape and the structure of mutualistic networks. *Journal of Ecology*, 102(5):1234–1243.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Colless, D. H. (1982). Review of phylogenetics: The theory and practice of phylogenetic systematics. *Systematic Zoology*, 31(1):100–104. Publisher: [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.].

Dolson, E., Lalejini, A., Jorgensen, S., and Ofria, C. (2020). Interpreting the tape of life: Ancestry-based analyses provide insights and intuition about evolutionary dynamics. *Artificial Life*, 26(1):1–22.

Dolson, E. L., Vostinar, A. E., Wiser, M. J., and Ofria, C. (2019). The MODES toolbox: Measurements of open-ended dynamics in evolving systems. *Artificial Life*, 25(1):50–73.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10.

Forest, F., Grenyer, R., Rouget, M., Davies, T. J., Cowling, R. M., Faith, D. P., Balmford, A., Manning, J. C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T. A. J., and Savolainen, V. (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, 445(7129):757–760.

Foster, E. D. and Deardorff, A. (2017). Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2):203.

Goldsby, H. J., Knoester, D. B., Ofria, C., and Kerr, B. (2014). The evolutionary origin of somatic cells under the dirty work hypothesis. *PLoS Biol*, 12(5):e1001858.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hernandez, J. G., Lalejini, A., and Dolson, E. (2022). What can phylogenetic metrics tell us about useful diversity in evolutionary algorithms? In Banzhaf, W., Trujillo, L., Winkler, S., and Worzel, B., editors, *Genetic Programming Theory and Practice XVIII*, Genetic and Evolutionary Computation, pages 63–82. Springer Nature.

Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C., and Baillie, J. E. M. (2007). Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *Plos One*, 2(3):e296.

Johnson, K., Welch, P., Dolson, E., and Vostinar, A. E. (2022). Endosymbiosis or bust: Influence of ectosymbiosis on evolution of obligate endosymbiosis. In *ALIFE 2022: The 2022 Conference on Artificial Life*. MIT Press.

Lalejini, A., Dolson, E., Bohm, C., Ferguson, A. J., Parsons, D. P., Rainford, P. F., Richmond, P., and Ofria, C. (2019). Data standards for artificial life software. In *ALIFE 2019: The 2019 Conference on Artificial Life*, pages 507–514. MIT Press.

Lalejini, A. and Ofria, C. (2016). The evolutionary origins of phenotypic plasticity. In Gershenson, C., Froese, T., Siqueiros, J. M., Aguilar, W., Izquierdo, E. J., and Sayama, H., editors, *Proceedings of the Artificial Life Conference 2016*, pages 372–379. MIT Press.

Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

Lewinsohn, M. A., Bedford, T., Müller, N. F., and Feder, A. F. (2023). State-dependent evolutionary models reveal modes of solid tumour growth. *Nature Ecology & Evolution*, pages 1–16. Publisher: Nature Publishing Group.

Mir, A., Rotger, L., and Rosselló, F. (2018). Sound colless-like balance indices for multifurcating trees. *Plos One*, 13(9):e0203401. Publisher: Public Library of Science.

Moreno, M. A., Dolson, E., and Ofria, C. (2022a). Hereditary stratigraphy: Genome annotations to enable phylogenetic inference over distributed populations. In *ALIFE 2022: The 2022 Conference on Artificial Life*. MIT Press.

Moreno, M. A., Dolson, E., and Ofria, C. (2022b). Hereditary stratigraphy: Genome annotations to enable phylogenetic inference over distributed populations. In *ALIFE 2022: The 2022 Conference on Artificial Life*. MIT Press.

Moreno, M. A., Dolson, E., and Ofria, C. (2022c). hstrat: a python package for phylogenetic inference on distributed digital evolution populations. *Journal of Open Source Software*, 7(80):4866.

Ofria, C., Moreno, M. A., Dolson, E., Lalejini, A., Rodriguez Papa, S., Fenton, J., Perry, K., Jorgensen, S., hoffmanriley, grenewode, Baldwin Edwards, O., Stredwick, J., cgnitash, theycallmeHeem, Vostinar, A., Moreno, R., Schossau, J., Zaman, L., and djrain (2020). Empirical: C++ library for efficient, reliable, and accessible scientific software. *Zenodo*.

pandas development team, T. (2020). pandas-dev/pandas: Pandas. *Zenodo*.

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. (2014). tqdist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080.

Scott, J. G., Maini, P. K., Anderson, A. R. A., and Fletcher, A. G. (2020). Inferring tumor proliferative organization from phylogenetic tree measures in a computational model. *Systematic Biology*, 69(4):623–637.

Shahbandegan, S., Hernandez, J. G., Lalejini, A., and Dolson, E. (2022). Untangling phylogenetic diversity's role in evolutionary computation using a suite of diagnostic fitness landscapes. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Gecco '22, pages 2322–2325. Association for Computing Machinery.

Shao, K.-T. (1990). Tree balance. *Systematic Biology*, 39(3):266–276.

Sukumaran, J. and Holder, M. T. (2010). Dendropy: a python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571.

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., and Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92(2):698–715.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Webb, C. O. and Losos, A. E. J. B. (2000). Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, 156(2):145–155.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56–61.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E., and Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol*, 12(12):e1002023.