

# Similarities, differences and biases in cophylogenetic models for host-symbiont coevolution

Gabriele Di Palma<sup>1,2</sup>[0009–0006–0928–5892], Catherine  
Matias<sup>3,4,5</sup>[0000–0001–6665–2421], and Blerina Sinimeri<sup>2</sup>[0000–0002–9797–7592]

<sup>1</sup> Università Campus Bio-Medico di Roma, Rome, Italy, [dipalmag@luiss.it](mailto:dipalmag@luiss.it)

<sup>2</sup> LUISS University, Rome, Italy

<sup>3</sup> Sorbonne Université

<sup>4</sup> Université Paris Cité

<sup>5</sup> CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005  
Paris, France

**Abstract.** Synthetic data are becoming increasingly important for computational studies of cophylogeny, including machine learning, benchmarking, and method testing. Several generators have been proposed to produce such data, but each relies on different assumptions about host-symbiont coevolution. These assumptions are often implicit and rarely examined, even though results can depend strongly on the synthetic model being used. In this article, we present a systematic structural analysis of representative cophylogeny generators under controlled scenarios. The goal is to make their assumptions explicit and to understand how these choices shape the synthetic data they produce as well as the conclusions that may be drawn from them.

**Keywords:** cophylogeny · synthetic data · host-symbiont evolution.

## 1 Introduction

Reconstructing the shared evolutionary history of symbionts and their hosts is central in several application domains, including the identification and tracking of emerging infectious diseases [12,14,22]. The growing availability of public sequence data has made these analyses increasingly feasible at scale. A standard way to formalize host-symbiont coevolution is through *cophylogeny* models, which aim to explain the histories of hosts and symbionts using their evolutionary trees (typically inferred from DNA sequences). In this framework, coevolution is often expressed as the problem of mapping the symbiont phylogeny onto the host phylogeny (see, e.g., [21,6,9,16,28,30]). This mapping, called a *reconciliation*, describes the relationship between the two trees via biologically meaningful events such as cospeciation, host-switching, duplication, and loss. Because real host-symbiont datasets are limited in number, size, and diversity, synthetic data are an important resource for developing and validating computational methods

in cophylogeny. Several synthetic generators for host–symbiont systems have been proposed, based on different modeling assumptions and simulation strategies. These generators are used to evaluate reconciliation methods and statistical tests of congruence, and they are expected to become even more important for data-driven pipelines, especially machine learning, where large, controlled datasets are needed for training and benchmarking. Despite this growing role, synthetic generators are often treated as interchangeable, and their outputs are rarely analyzed in detail beyond the specific task they are used for. In practice, different generators may encode different assumptions about evolutionary processes and may constrain the space of host–symbiont structures they can produce. As a result, datasets generated under comparable evolutionary scenarios may still differ substantially in their trees structures, detectable coevolutionary signal, and association network topology. Such differences can directly affect the conclusions drawn from synthetic-based computational studies.

In this paper we study the structural properties of synthetic host–symbiont datasets produced by commonly used cophylogeny generators. Thus, we adopt a regime-based evaluation framework, considering cospeciation-dominated, host-switch-dominated, and mixed coevolutionary regimes, and analyze each simulated dataset at three complementary levels. **First**, we examine coevolution-based summaries of the generated data, focusing on the relative frequencies of key events (in particular cospeciation and host-switching) to verify how closely each generator follows the intended regime. **Second**, we characterize the host and symbiont phylogenies through size and shape descriptors, including the number of leaves and standard tree-balance indices such as Cherry, Sackin, and Colless. **Third**, we study the association structure by analysing the bipartite network defined by the host–symbiont nowadays associations. We measure classical network parameters such as the connectivity and heterogeneity (e.g., density, assortativity, and hosts hotspot concentration). Finally, we compare these structural characteristics of the synthetic data to those observed in real host–symbiont datasets, which we use more as a reference region rather than as ground truth. This study is guided by two questions:

- (Q1) When we fix the same high-level coevolutionary regime, do different generators produce datasets with comparable structure, or do they systematically induce different tree shapes, congruence signals, and association-network patterns?
- (Q2) How much do the structural patterns produced by each generator overlap with those observed in real host–symbiont datasets, and do the generators fully cover this structural space or leave significant regions occupied by real data unexplored?

Our goal is not to identify the “best” generator but to explore how different modeling choices determine the structure of synthetic cophylogenetic data. By making these differences clear, we aim to help researchers choose synthetic data in a careful way for cophylogeny studies, and to point out the biases that can come from using one generator instead of another.

## 2 Methods

Here we focus on the host-symbiont setting, where the generators simulate a host phylogeny, a symbiont phylogeny, and the associations between them. Existing generators in the literature follow different modeling choices (for example in the set of macro-evolutionary events they include, or whether they use time and branch lengths). We therefore compare the following host-symbiont generators, namely **Coala** [4], **treeduck** [10] and the model described in [1] that has no name and that we call **cophylo** (from the name of the main function provided by the authors). We also include **AsymmeTree** [26], originally developed in the gene-species setting. Although its gene-family simulation builds on ideas from tools such as SaGePhy [13], **AsymmeTree** provides a simpler “tree evolving on a tree” framework that produces reconciliation-style histories without introducing additional layers such as domain evolution or population processes. This makes it structurally closer to host-symbiont generators and allows a more direct comparison in our setting. We do not consider gene-species simulators more broadly, since many of them include processes specific to gene evolution, such as horizontal gene transfer with replacement or incomplete lineage sorting, which do not have clear counterparts in host-symbiont macro-evolutionary models. Conversely, host-symbiont systems may involve *multiple associations*, where a symbiont is linked to more than one host at a time, a situation that does not arise in the gene-species setting.

### 2.1 Description of the evaluated generators

We now describe the 4 generators considered in this study. Notice that most of these generators are by-products of more global coevolution or cophylogenetics methods that we do not describe here. We rather only focus on the parts of these tools that enable generating a pair of host and symbiont trees together with coevolution history and extant leaves associations. Some of these tools only generate a symbiont tree *conditional* on a given input host tree (and additional parameters), while other *jointly* generate the host and symbiont tree (in general, from a birth and death process). In order to get a full generating process of host and symbiont pair, we combine the former tools with a simple birth-death process with parameters  $(\lambda_H, \mu_H)$  (speciation and extinction, respectively) on the host tree when it should be given as input.

**Coala.** The Java software **Coala** [4,5] contains a generator (**TGLGenerator.jar**) for simulating symbiont coevolution along a given host tree and allowing 4 different types of events: *cospeciation*, *duplication*, *host-switch* and *loss*. The input host tree does not need to be dated (if branch lengths exist, they will be discarded) so we rely on the function **TreeSim** from the **DendroPy** python library [18,17], using parameters  $(\lambda_H, \mu_H)$  and conditioned on the number of extant leaves to produce the input host tree. The output symbiont tree will similarly be undated. The generator uses 3 parameters resulting in a vector  $\langle p_c, p_d, p_s, p_l \rangle$  of 4 probabilities constrained to sum to 1. The model is based on

the Duplication–Transfer–Loss (DTL) model [29,3]. Conditional on a given host tree  $H$ , the generation of the symbiont tree proceeds by recursively considering its unmapped nodes that are (temporarily) “positioned” on branches and for which one of the 4 different events occurs. The process starts by positioning the symbiont root on the branch before the root of  $H$  (by default, although the user may choose any host branch), where this initial branch is fictitious and serves only to initialise the simulation. For every unmapped symbiont node that is temporarily positioned on a host branch, the 3 events cospeciation, duplication and switch (so except in case of a loss) will induce its mapping on the host node below the branch it was positioned on, as well as its speciation. Now, in case of a cospeciation (occurring with probability  $p_c$ ), each descendant symbiont lineage is positioned on one of the 2 host branches below the associated host. Conversely, when duplication occurs (with probability  $p_d$ ), each descendant symbiont lineage stays positioned on the same host branch its parent was on. A host-switch (occurring with probability  $p_s$ ) induces one descendant lineage to be positioned on the same host branch its parent was on, while the other descendant is positioned on a host branch chosen randomly, under the constraint that this does not violate the time feasibility of the reconstruction so far [28]. If such a branch does not exist, the switch does not happen and a new event  $i$  is drawn among the 3 others (namely  $i \in \{c, d, l\}$ ) with rescaled probability  $p_i/(p_c + p_d + p_l)$ . Finally when a loss occurs (with probability  $p_l = 1 - p_c - p_d - p_s$ ), the symbiont node is not mapped and rather re-positioned on one of the 2 host branches (chosen randomly) below the branch it was. Note that a loss event can represent three distinct yet indistinguishable scenarios: (a) speciation of the host species independently of the symbiont, which then remains associated with only one of the new host species; (b) cospeciation of the host and symbiont, “immediately” followed by the extinction of one of the newly formed symbiont species; or (c) the same as (b), but with a failure to detect the symbiont in one of the two new host species. The process continues until no unmapped symbiont nodes remain; the evolution of a lineage stops once it is mapped to a host leaf. For a fixed host tree  $H$  and a given set of event probabilities, the simulator generates multiple symbiont trees (100 by default). A representative tree is then selected as the median tree, i.e., the tree that minimises the distance to the others under a chosen tree-distance measure. Note that the set of events included in **Coala** result in host-symbiont coevolved trees with no multiple associations between their leaves (so that a symbiont has at most one host). The method **AmoCoala** [27] later developed to overcome this limitation relies on an initial pair of host-symbiont trees and their leaves associations to set additional local probabilities of new events (called *spreads*) that produce these multiple associations and are defined at every internal node of the host tree. This very-fine tuning of the model (allowing heterogeneous probabilities of spreads along the tree) becomes a drawback in our study, as the sheer number of additional choices would complicate our analysis. For this reason, we choose not to include this generator in our comparison.

**treeducken.** The `sim_cophyBD` function from the R package **treeducken** [11] implements a cophylogenetic birth-death process that simulates, in forward time,

Table 1: Correspondence between biological processes, event names and parameters across models. AS stands for associated symbionts; MA for multiple associations.

Coevolution event	TreeSim +Coala	treeducken	species_tree_n_age +cophylo	AsymmeTree
(Independent) Host speciation	host speciation $\lambda_H$ ; induces a loss for AS	host speciation $\lambda_H$ ; induces MA for AS	host speciation $\lambda_H$ or no cospeciation $1 - c$ ; induces either a loss (if specialist) or MA (if generalist) for AS	species tree speciation $\lambda_H$
(Independent) Host extinction	host extinction $\mu_H$	host extinction $\mu_H$	host extinction $\mu_H$	species tree extinction $\mu_H$
Independent symbiont speciation	duplication $p_d$	symbiont speciation $\lambda_S$	symbiont speciation $\lambda_S$	duplication $\lambda_D$
Independent symbiont extinction	–	symbiont extinction $\mu_S$	symbiont extinction $\mu_S$	loss $\lambda_L$
Induced symbiont extinction	–	host extinction ( $\mu_H$ ) of the unique host (depends on past history) of that symbiont	–	–
Cospeciation (joint host-symbiont speciation)	cospeciation $p_c$	cospeciation $\lambda_c$	cospeciation $c$	(gene) speciation; prob. 1
Symbiont speciates and one descendant moves to another host lineage	switch $p_s$	switch $\chi$	–	horizontal transfer $\lambda_{HGT}$
Association gain at host split (one symbiont associated to both descendant hosts)	–	spread $\chi$	switch $s$ or no cospeciation ( $1 - c$ ) of a generalist symbiont ( $1 - f_d(1) \simeq 0.54$ by default)	–
Association loss at host split	loss $p_l$	–	no cospeciation ( $1 - c$ ) of a specialist symbiont ( $f_d(1) \simeq 0.46$ by default)	–

a pair of phylogenies (host tree  $H$  and symbiont tree  $S$ ) together with their extant ecological interactions [10]. Note that it's the only tool out of the four tested here that simulates the two trees jointly. The model allows hosts and symbionts to undergo speciation and extinction independently, and it also includes coupled events such as *cospeciation* (with explicit parameter) and *coextinction* (implicit).

The generator is based on a continuous-time Markov process and is conditioned on a simulation time horizon. Both trees will start at the same date and evolve during that time horizon, creating trees with the same stem age. The generator contains a total of six competing events, parameterized by their rates  $(\lambda_H, \mu_H, \lambda_S, \mu_S, \chi, \lambda_c)$ . Symbiont evolution is governed by the speciation rate  $\lambda_S$  and the extinction rate  $\mu_S$ . A symbiont speciation event is similar to a duplication-like event from *Coala*, in which both descendant lineages inherit the ancestral host associations, while a symbiont extinction corresponds to a (particular case of a) loss-like event in that setting. The model includes a *host-expansion* event that takes 2 different forms (*switch* and *spread*), both being (additional) symbiont speciation events occurring with rate  $\chi$ . The switch event creates one symbiont speciation with one descendant that has a new host association (among hosts not associated to its parent) while the other descendant retains the ancestral host repertoire. In spread event, one randomly chosen descendant gains a novel host association in addition to the ancestral host repertoire (thus creating a multiple association). The tool has three options (accessible through `hs_mode`  $\in \{\text{"switch"}, \text{"spread"}, \text{"both"}\}$ ), where only switches, spreads or both events (with half probability for each) may occur, respectively. The user may also enforce a *host limit*, i.e., a maximum number of host associations per symbiont lineage at any given time. Cospeciation occurs at rate  $\lambda_c$  when a host lineage bifurcates and induces a simultaneous symbiont speciation event, triggered on one randomly selected symbiont lineage currently associated with that host. After cospeciation, each descendant host lineage is associated with one of the two new symbiont descendants, and remaining ancestral associations are redistributed at random among descendants. Hosts can also speciate independently of symbionts at rate  $\lambda_H$  and any associated symbiont will then remain associated to each descendant (thus creating a multiple association). This is analogous to a failure-to-diverge event as described in [7] and implemented in [9,15]. Finally, host extinction occurs at rate  $\mu_H$  and may implicitly induce coextinction of symbiont lineages that are left without any associated hosts.

*cophylo*. The article [1] comes with a code as Supplementary Information [2]. It includes a C program (`main_cophylo.c`) generator for simulating a symbiont tree  $S$ , conditional on the input host  $H$ , using both a continuous-time Markov process with three competing events: *speciation*, *extinction* and *host shift*; as well as an additional process of *cospeciation* that may occur at each host speciation. The input host should be a dated tree and we simulate it using the auxiliary function `species_tree_n_age` from the tool *AsymmeTree* (described below), that generates a birth-death process, conditional on the number of extant leaves but also on a time horizon  $T$ . Then the conditional simulation of the symbiont tree is obtained by successively running the former Markov process on

Table 2: Comparison of simulation design across cophylogeny generators.

Model	Host phylo.	Symb.phylo.	Associations	Time	Symb. root
Coala	fixed input (indep.)	cond.	cond.	no	mapped to a host node
treeducken	joint	joint	joint	yes	same origin as host
cophylo	fixed input (indep.)	cond.	cond. with constraint on nb of hosts per symbiont at any time (input possible)	yes	independent origin time
AsymmeTree	birth-death process	cond.	cond.	yes	same origin as host

indep. = simulated independently; cond. = conditionally simulated; joint = jointly simulated.

each time interval defined by two successive host speciations. The generator has 5 evolutionary parameters:  $\lambda_S, \mu_S$  and  $s$  are the symbiont speciation, extinction and switch rates, respectively. Additionally,  $c$  is a cospeciation probability and  $T_{MRCA}$  a time to the most recent common ancestor (TMRCA). It enables to start simulating the symbiont tree at a different age than the host tree.

The first symbiont is linked to a random number  $d$  (drawn from parameter distribution  $f_d$  on the number of hosts per symbiont at any time) of contemporary hosts, or all available hosts if  $d$  exceeds their number. From the initial time to the next host speciation event, or between any 2 successive host speciation events, a continuous-time Markov process with the 3 above mentioned competing events is run. With rate  $\lambda$  (resp.  $\mu$ ) times the number of existing symbionts, a speciation (resp. extinction) event occurs to one symbiont randomly chosen. In case of a speciation, a first descendant inherits all associations of its parent while the other inherits a random number  $d$  (drawn with  $f_d$ ) among a subsample of these. In case of an extinction, that symbiont is simply removed. With rate  $s$  times the total number of current associations, a switch event occurs, randomly chosen among symbionts weighted by their number of current hosts. This symbiont gains a new host among those not yet associated to it. Note that in that switch event, contrarily to what happens in **Coala** or **treeducken**, the symbiont does not undergo a speciation. So that this event creates a multiple association. It rather corresponds to the spread event from **treeducken**. Additionally, at each host speciation time, their associated symbionts may cospeciate with probability  $c$ , in which case two symbiont descendants are created and each one follows one of the two descendant host. If there is no cospeciation, then the symbiont may either be a *specialist* (with probability  $f_d(1)$  which is user-chosen or  $\approx 0.46$  by default) or a *generalist*. A specialist will follow only one of the two descendant hosts, while a generalist will follow both and thus create a new multiple association. Note that in the latter case, the symbiont did not speciate and remains the same in the two descendant hosts, differentiating this case from the cospeciation one.

**AsymmeTree.** The article [26] contains a general model for gene family history simulation along a species tree with *duplication, loss, horizontal transfer and con-*

*version.* Putting aside the conversion event that is specific to the gene-species context, the model is similar to DTL mentioned above and may be used in the host-symbiont context. In the corresponding `python` package [25], we focus here on two functions, namely `species_tree_n_age` that uses a birth-death model conditioned on time and number of extant leaves for simulating a species tree (host tree in our context), and `dated_gene_tree` that simulates a gene tree (symbiont tree in our context), conditional on a dated species tree. Note that in the former, the branches leading to extinct species are pruned from the output. We use the latter with only three events: duplication with rate  $\lambda_D$ , loss with rate  $\lambda_L$  and horizontal (gene) transfer with rate  $\lambda_{HGT}$ . The root of the gene (symbiont) tree is placed at the root of the species (host) tree. The host-speciation events are ordered and processed sequentially. The simulation alternates between two cases: a host-speciation event and random events occurring in the symbiont tree during two successive host speciation times. When host-speciation occurs, the gene necessarily “cospeciates” at the same time; in our host-symbiont case, that means the the symbiont cospeciates with its host with probability 1. Then, and similarly to `cophylo`, between two successive host-speciations, a continuous-time Markov process is run with the three competing events: duplication, loss and horizontal transfer. Duplication is the same event as in `Coala`, corresponding to symbiont speciation in `treeducken` or `cophylo`. A loss event, is the same as a loss in `Coala` or as a symbiont extinction in `treeducken` or `cophylo`. Finally, horizontal transfer corresponds to a switch in `Coala` or `treeducken` (and does not have an equivalent event in `cophylo`, since it implies a symbiont speciation). Note that this generator never produces multiple associations (that do not exist in the gene-species context).

*Summary of the generators characteristics.* Table 1 summarizes the characteristics of the tools with a focus on the different coevolutionary events, while Table 2 contrasts these characteristics by emphasizing joint or conditional simulation of the trees as well as the role of time. Note that in the former, the event “Induced symbiont extinction” only exists for `treeducken` because the host and symbiont trees are jointly simulated, while for the other tools, the host tree is generated and pruned for extinct branches before being input in the conditional generator for the symbiont tree.

## 2.2 Measures used to compare the generators

In this section, we describe 2 different type of characteristics that we have measured on each pair of host and symbiont trees, across the different generators. While the coevolution-based measures where used in a calibration process, to tune the parameters so as to obtain comparable situations across generators, the tree-based and association-based measures are used to understand the characteristics of the data they produce.

*Coevolution-based measures.* For each simulated host–symbiont pair, we record the number of cospeciation and host-switch events appearing in the coevolution



process between the two trees. We focus on these two events because they are present, with comparable interpretation, across all generators. This persistence and stability also enables us to rigorously analyze simulation regimes defined by their relative frequencies (see Section 3.1).

*Tree-based measures.* For each simulated pair of host and symbiont trees, we first record basic size and scale characteristics. In particular, for each tree (host or symbiont) we measure the number of leaves and the tree height (i.e. the distance from the root to its deepest leaf). For each simulated pair, we also compute the difference between the number of leaves in the host and symbiont tree. We then compute standard shape and balance indices, through normalized versions of the *Cherry*, *Colless* [8] and *Sackin* [24] indices (see Appendix A for definitions).

All distances and heights reported in those measures are defined in the topological sense (number of edges), rather than using branch lengths. This is because not all generators produce comparable temporal information, and some models enforce a fixed time horizon while others allow simulated times to extend beyond it. Thus, using topological distances ensures that the measures are comparable across generators.

*Association-based measures.* For each simulated dataset, we consider the bipartite host-symbiont interaction network defined on the leaves of the two trees. On this interaction network, we measure the *density*, the *degree-assortativity* and the frequency of *host hotspots* (see Appendix A for definitions). The density is the average number of connections in the ecological interaction network of hosts and symbionts. The degree-assortativity [19] measures the tendency of generalist (resp. specialist) species to be associated to other generalist (resp. specialists). It corresponds to a correlation so that it can exhibit positive or negative values. The frequency of hosts hotspots captures the presence of hosts that are highly connected [20,23].

### 2.3 Experimental setup

We explored a high-cospeciation (regime R1), a high host-switch (regime R2), and a mixed situation (regime R3). The exact parameter choices are given in Table 4 in Appendix B. These parameters were tuned so that each generator reproduces as closely as possible the intended coevolutionary regime, defined in terms of the relative prevalence of cospeciation and host-switch events. Comparability across generators was therefore monitored using these coevolution-based measures (see Section 3.1). Moreover, to make the tools as comparable as possible, we choose to set up the time  $T_{MRC\bar{A}}$  in `cophylo` equal to the horizon time used to generate the symbiont tree. As a consequence, the host and the symbiont trees have the same age (as in the two other tools where time is considered). We also need to set up the distribution  $f_d$  on the number of hosts per symbiont at any time for that tool. Following the suggestion of the authors on their own dataset example, we rely on a power law with exponent parameter  $\alpha = 1.6$  and support in  $\{1, \dots, n_H\}$  where  $n_H$  is the number of extant hosts (i.e. the probability to have  $k$  hosts is proportional to  $k^{-\alpha}$ ).

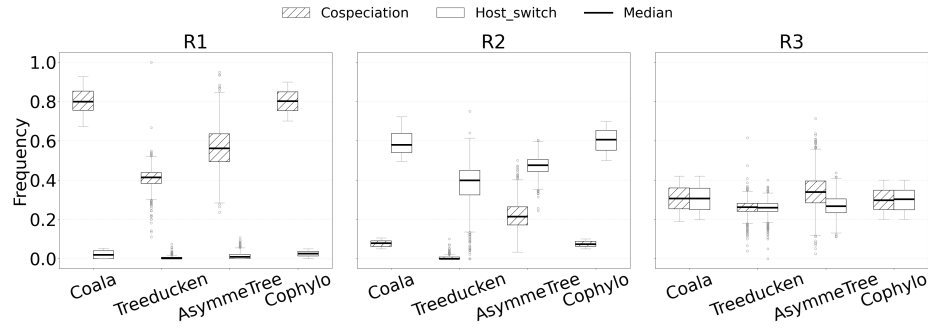


Fig. 1: Boxplots of cospeciation and host-switch frequencies across generators in regimes R1–R3, shown after parameter tuning. The figure indicates how closely each generator attains the intended regime.

### 3 Results and discussions

#### 3.1 Coevolution-based analysis

Fig. 1 shows how the observed frequencies of cospeciation and host-switch events vary across generators and regimes. We stress that these measures were first used to tune the parameters of the generators (described in Table 4) in order to best reproduce each regime. Hence, the results shown correspond to the best match obtained for each generator. Because cospeciation and host-switch frequencies were used during parameter tuning, Fig. 1 should be interpreted mainly as showing how closely each generator matches the intended regime after calibration. By contrast, the analyses in the following subsections rely on measures that were not used for tuning and therefore more directly reflect structural differences between generators.

In R1, where cospeciation is expected to dominate, **Coala** and **cophylo** closely follow the intended regime, producing high cospeciation frequencies with very low host-switch values and relatively concentrated distributions. In contrast, **treeducken** generates substantially lower cospeciation frequencies and almost no host-switches, indicating that a significant portion of events is allocated to other processes not shown in the figure. **AsymmTree** produces intermediate cospeciation values with a broader spread and more outliers, suggesting less tightly controlled outcomes. In R2, where host-switches are expected to dominate, **Coala** and **cophylo** again reproduce the intended behaviour with high host-switch frequencies and low cospeciation. **treeducken** still shows a noticeable deviation, with host-switch frequencies that are lower and more dispersed than in the other models and cospeciation values close to zero, while **AsymmTree** displays intermediate values with moderate dispersion. In R3, corresponding to a mixed regime, all generators move toward more balanced frequencies between cospeciation and host-switches. Here the differences between generators are smaller: **treeducken** produces relatively concentrated distributions, while **AsymmTree** exhibits the

widest spreads and more outliers, and **Coala** and **cophylo** remain comparatively stable.

### 3.2 Tree-based analysis

*Basic tree parameters.* Figs. 2, and 3 show how strongly the regime parameters overall constrain tree growth across the different regimes and generators. In R1, high cospeciation and low host-switch should keep symbiont and host tree sizes similar, and this is what we observe for **Coala**, **cophylo** and **treeducken**. However **treeducken** still produces much larger trees, indicating that additional events drive tree growth even in scenarios where cospeciation dominates. In R2, frequent switches allow symbionts to diversify independently, and this may lead to a dramatic increase in symbiont size for **AsymmeTree**, while **Coala** and **cophylo** remain constrained. In R3, differences are smaller but **treeducken** continues to generate the largest trees, showing that its growth is less tightly controlled by the regime. Finally, especially in the R2 and R3 regimes, the simulated symbiont trees are often larger than the corresponding host trees, indicating that under these parameter settings, symbiont lineages diversify more rapidly than host lineages. In R1, this effect is less pronounced overall, but it is still very visible for **AsymmeTree**, where symbiont trees can grow substantially even under high cospeciation. This behaviour is expected in that model, which was originally designed for gene-species evolution, where gene lineages can diversify more independently from the species tree.

*Shape and (Im)Balance-measures.* Fig. 4 shows that, across the three regimes, the balance indices reflect how strongly each generator constrains the symbiont tree to follow the host tree. Generators that simulate host and symbiont jointly (such as **treeducken**) tend to produce more similar host and symbiont shapes, since both trees are built under the same process and remain tightly coupled. In contrast, generators where the host is simulated independently and the symbiont is added conditionally (such as **Coala** and **cophylo**), or where additional independent events are allowed (such as **AsymmeTree**), show larger differences between the two trees. In the high-cospeciation regime (R1), host and symbiont shapes are generally close for **Coala** and **treeducken**, indicating that strong cospeciation induces strong topology similarities, while **cophylo** produces more imbalanced symbiont trees. In the switch-dominated regime (R2), the coupling weakens: frequent switches allow the symbiont tree to diverge more from the host, and several generators produce noticeably more imbalanced symbiont topologies. The mixed regime (R3) lies between these behaviours, with moderate similarity but persistent model-dependent differences. We can conclude that matching cospeciation and switch settings does not guarantee similar tree shapes: the resulting topology depends both on how the trees are generated (jointly or independently) and on how strongly symbiont evolution is constrained by the host.

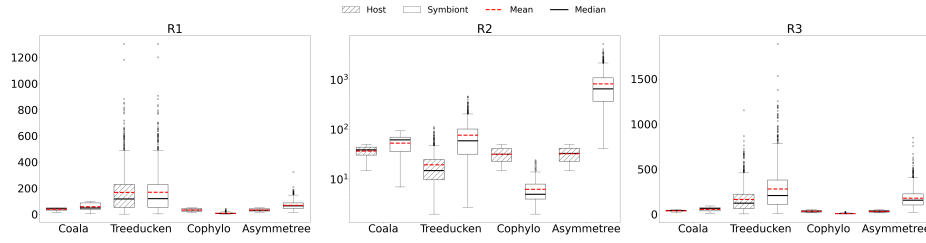


Fig. 2: Boxplots of the number of leaves for each regime R1-R3

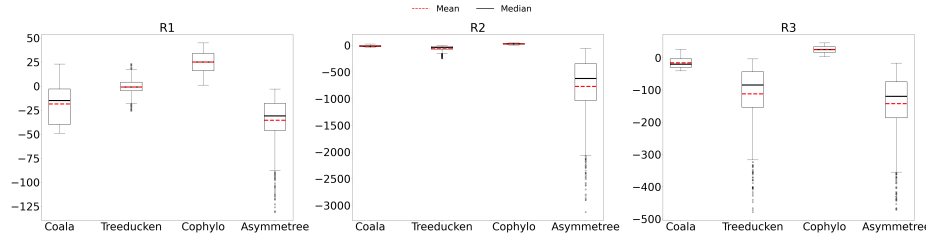


Fig. 3: Boxplots of the difference in the number of leaves between host and symbiont trees for each simulated dataset, across regimes R1–R3 (without 1% extremes; see Fig 8 for complete image).

### 3.3 Association-based measures

In Table 3 and Figs. 5–6, the network-level measures show that each generator produces markedly different host–symbiont association structures across regimes, reflecting how associations are constructed in the simulation. Notice that `cophylo` is excluded from this analysis. This is because we observed a memory-related bug in `cophylo` when reporting host–symbiont associations: in some cases, the output includes hosts or symbionts that do not appear in the corresponding phylogenetic trees. We reported this issue to the authors, but a fix is not yet available. Because this problem affects only the association output, it may bias network-based metrics. For this reason, we use `cophylo` results only for the previously defined tree-based measures, and exclude it from the association-based comparisons. In `Coala`, each symbiont is associated with exactly one host, which keeps the association network sparse across regimes. Thus, the degree variation comes mainly from hosts accumulating symbionts, so the proportion of host hotspots remains similar in R1–R3. Assortativity is consistently negative, indicating that edges tend to link nodes with different degrees. Because each symbiont can attach to only one host, this dis-assortative structure is expected and remains stable across regimes. In `treeducken`, density stays low in all regimes, but it is much more stable in the R2 than in the other two. Indeed, when switching dominates, events mainly change which host a lineage is linked to, so repeated runs tend to give similar density values, whereas the high-cospeciation and mixed regimes lead to larger run-to-run differences in how associations are

distributed. Assortativity is close to zero in the high-cospeciation and mixed regimes, but becomes clearly negative in the high-switch regime, because frequent switching produces a stronger generalist–specialist structure where edges tend to connect nodes with different degrees. Host hotspots are similar in the high-cospeciation and high-switch regimes but drop in the mixed regime, because R1 and R2 produce more uneven host degrees, while R3 makes degrees more even, leaving fewer hosts above  $\text{mean}+1\text{sd}$ . In **AsymmeTree**, density stays low in all regimes, but it is lowest in the R2 regime. A simple way to read this is that, under high switching, the generator tends to increase the number of possible host–symbiont pairs faster than it increases the number of realised links, so the network becomes sparser. Assortativity is negative in all regimes, and it is most negative in R2. This matches the regime meaning: when switching dominates, links are more likely to connect nodes with very different degrees, which pushes assortativity down; when cospeciation dominates, this effect is weaker and assortativity moves closer to zero. Host hotspots follow the same general idea: they are lowest in R1, highest in R2, and remain relatively high in R3, which is consistent with switching creating a more uneven host-degree distribution (more “outlier” hosts above the  $\text{mean}+1\text{sd}$  threshold), while cospeciation spreads associations more evenly across hosts.

### 3.4 Comparison to real data

We considered 37 real host–symbiont datasets from the literature. These datasets span a broad range of biological systems, including plant–fungus interactions, insect–parasite and vertebrate–parasite associations, as well as endosymbiotic relationships (e.g., *Wolbachia*), and display substantial variation in tree sizes and association patterns. To compare real and synthetic datasets in one quantitative view, we embed each dataset in a common feature space and project it to two dimensions using PCA. Each dataset is represented by a seven-dimensional vector of tree size/shape summaries (host–symbiont leaf difference and the Cherry, Sackin, and Colless indices). Network features such as assortativity and host hotspots could also be included, but here we use only tree-based features so that **cophylo** can be compared fairly. Features are standardized (zero mean, unit variance), PCA is fit on the union of real and synthetic datasets, and we report the first two principal components. It is important to note that the real datasets do not come with regime labels (e.g., “cospeciation-dominated” vs “switch-dominated”), so we cannot assign each real dataset to a specific simulated setting. A second issue is that we cannot fairly evaluate a generator by sampling its parameters at random: most random combinations produce biologically implausible trees or association patterns, so the task would be computationally hard. That’s why we restrict attention to three interpretable regimes that represent common coevolutionary scenarios.

In Fig. 7, the real datasets (black) sit in a fairly compact area (delineated by the dashed ellipse). The synthetic datasets do not all fall there: where they land depends much more on which generator produced them than on whether we used one of the three regimes. **cophylo** (green) overlaps the real region the

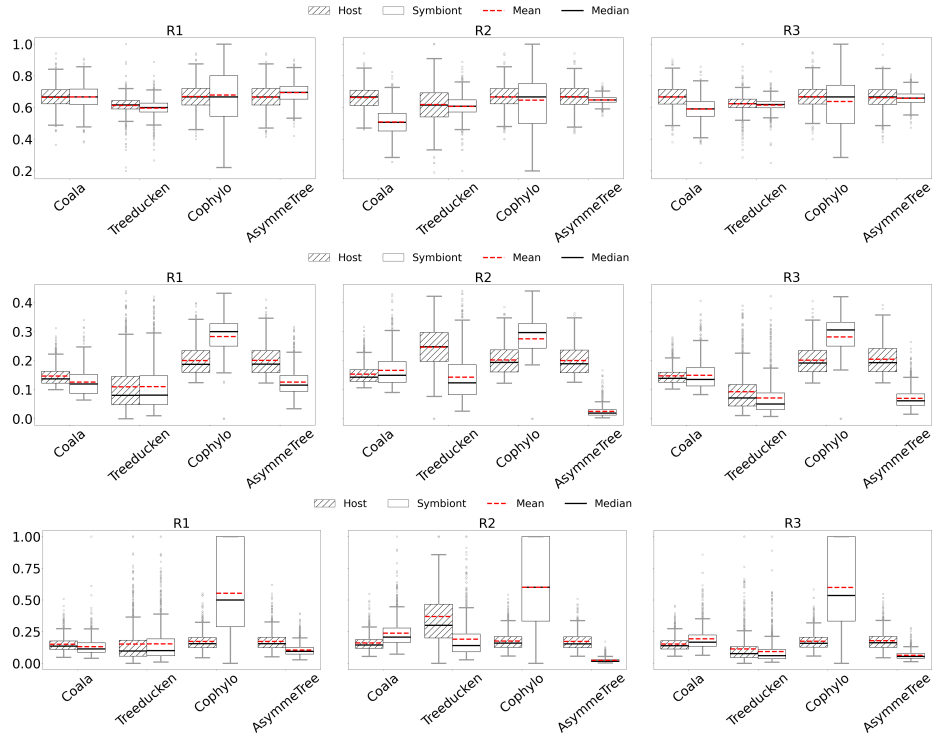


Fig. 4: Normalized Cherry (first row), Sackin (second row) and Colless (third row) indices across regimes R1–R3.

most, but also shows a clear downward extension (lower PC2) outside the ellipse. **treeduck** (red) is the most dispersed, reaching the real region but also extending far upward/right with several outliers. **Coala** (orange) is the most concentrated, clustering in the lower-left/central area and covering only part of the real region. **AsymmeTree** (blue) is mostly shifted left (negative PC1) and slightly higher on PC2, with only limited overlap near the left edge of the real region.

Table 3: Mean, Median, and Standard Deviation of association network density across regimes R1-R3.

	Mean/Median (sd)	Mean/Median (sd)	Mean/Median (sd)
Method	R1	R2	R3
Coala	0.032 / 0.028 (0.014)	0.041 / 0.032 (0.018)	0.041 / 0.033 (0.019)
treeduck	0.039 / 0.016 (0.067)	0.030 / 0.025 (0.019)	0.033 / 0.018 (0.054)
AsymmeTree	0.031 / 0.027 (0.011)	0.022 / 0.019 (0.008)	0.026 / 0.023 (0.01)

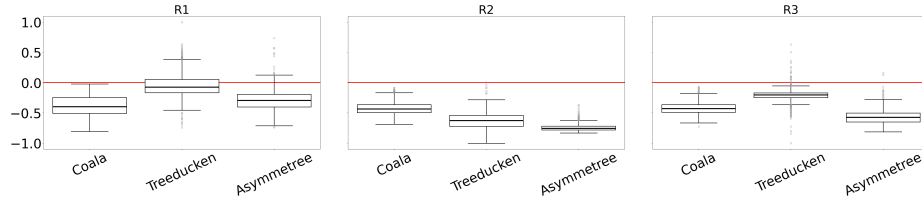


Fig. 5: Degree assortativity across regimes R1-R3.

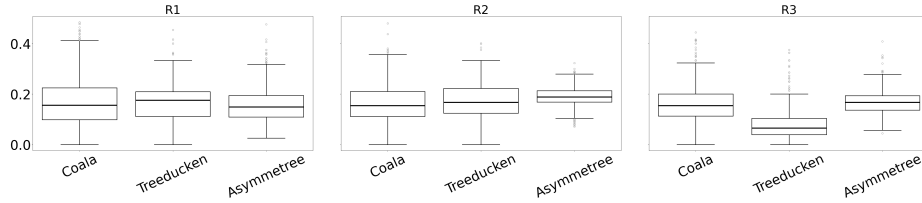


Fig. 6: Host hotspots frequency across regimes R1-R3.

## 4 Conclusion

This study shows that modeling choices affect the synthetic datasets used in cophylogeny. By comparing four common generators (**Coala**, **treeduck**, **cophylo**, and **AsymmeTree**) under the same three regimes, we observe systematic differences in the resulting feature profiles, including tree-shape summaries and, when available, association structure. This should be taken into account when using synthetic data to evaluate methods, because benchmarks may mix the effect of the intended regime with generator-specific patterns. A clear next step is to make “realism” and “coverage” measurable: define simple feature-based distances to quantify how close simulated datasets are to real ones and how well each generator covers the region occupied by real data, and use these scores to guide parameter calibration and to build comparable benchmark sets. This would make simulation-based comparisons more reliable and easier to interpret across studies. The code for reproducing those experiments as well as the real-datasets used are available at [https://github.com/Gabbo240900/synthetic\\_cophylo](https://github.com/Gabbo240900/synthetic_cophylo)

**Disclosure of Interests.** The authors have no competing interests.

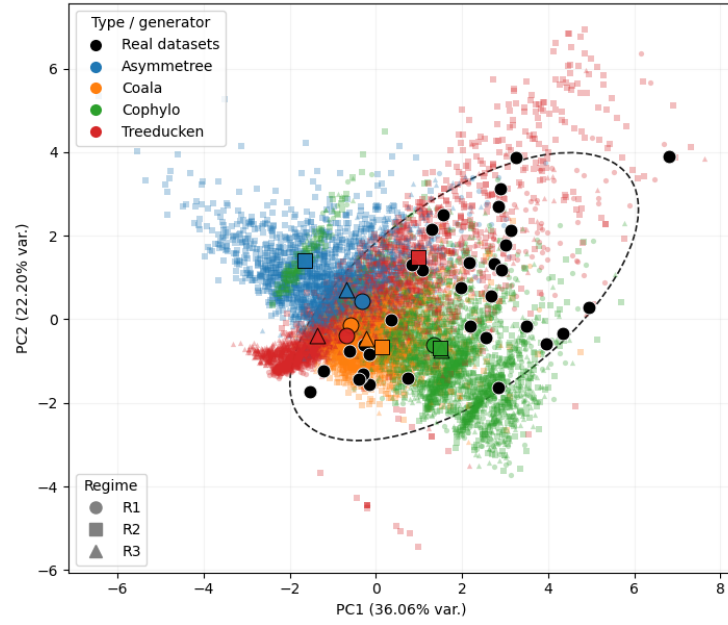


Fig. 7: Projection of real (black) and synthetic datasets (colored by generator, shaped by regime) onto PC1-PC2 from standardized feature vectors. Large markers show generator-regime centroids; the dashed ellipse summarizes real-data dispersion.



## References

- Alcala, N., Jenkins, T., Christe, P., Vuilleumier, S.: Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters* **20**, 1014–1024 (2017)
- Alcala, N., Jenkins, T., Christe, P., Vuilleumier, S.: Simulation software for host-parasite cophylogenies (2017), <https://onlinelibrary.wiley.com/doi/full/10.1111/ele.12799>, Supplementary Information
- Bansal, M.S., Alm, E., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* **28**(12), i283–i291 (2012)
- Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., Sagot, M.F.: Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology* **64**(3), 416–431 (2014)
- Baudet, C., Donati, B., Urbini, L., Crescenzi, P., Gautier, C., Matias, C., Sagot, M.F., Sinaimeri, B.: AmoCoala (2023), <https://github.com/sinaimeri/AmoCoala>, Java software
- Charleston, M.A.: Recent results in cophylogeny mapping. *Advances in Parasitology* **54**, 303–330 (2003)
- Charleston, M.A.: A new likelihood method for cophylogenetic analysis (2009), <http://www.it.usyd.edu.au/research/tr/tr636.pdf>, unpublished.
- Colless, D.H.: Review of: Phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology* **31**(1), 100–104 (1982)
- Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R.: Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology* **5**(1) (2010)
- Dismukes, W., Heath, T.A.: treeduck: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution* **12**(8), 1358–1364 (2021)
- Dismukes, W., Justison, J.: treeduck (2021), <https://github.com/wadedismukes/treeduck/tree/main>, R package version 1.1.0
- Etherington, G.J., Ring, S.M., Charleston, M.A., Dicks, J., Rayward-Smith, V.J., Roberts, I.N.: Tracing the origin and co-phylogeny of the caliciviruses. *Journal of General Virology* **87**(5), 1229–1235 (2006)
- Kundu, S., Bansal, M.S.: SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics* **35**(18), 3496–3498 (2019)
- Lei, B.R., Olival, K.J.: Contrasting Patterns in Mammal–Bacteria Coevolution: Bartonella and Leptospira in Bats and Rodents. *PLOS Neglected Tropical Diseases* **8**(3), 1–11 (2014)
- Libeskind-Hadas, *et al.*: Jane 4 (2019), <https://www.cs.hmc.edu/~hadas/jane/>, Java software
- Merkle, D., Middendorf, M., Wieseke, N.: A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics* **11**(Suppl 1), S60 (2010)
- Moreno, M.A., Holder, M.T., Sukumaran, J.: DendroPy 5 (2024), <https://jeetsukumaran.github.io/DendroPy/index.html>, Python library - version 5.0.8
- Moreno, M.A., Holder, M.T., Sukumaran, J.: Dendropy 5: a mature python library for phylogenetic computing. *Journal of Open Source Software* **9**(101), 6943 (2024)
- Newman, M.E.J.: Mixing patterns in networks. *Physical Review E* **67**(2) (2003)
- Newman, M.: *Networks*. Oxford University Press (2018)
- Page, R.D.M.: Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics* **10**(2), 155–173 (1994)

22. Pennington, P.M., Messenger, L.A., Reina, J., Juárez, J.G., Lawrence, G.G., Dotson, E.M., Llewellyn, M.S., Córdón-Rosales, C.: The Chagas disease domestic transmission cycle in Guatemala: Parasite-vector switches and lack of mitochondrial co-diversification between *Triatoma dimidiata* and *Trypanosoma cruzi* subpopulations suggest non-vectorial parasite dispersal across the Motagua valley. *Acta Tropica* **151**, 80–87 (2015), ecology and diversity of *Trypanosoma cruzi*
23. Poulin, R.: Network analysis shining light on parasite ecology and diversity. *Trends in Parasitology* **26**(10), 492–498 (2010)
24. Sackin, M.J.: “Good” and “bad” phenograms. *Systematic Biology* **21**(2), 225–226 (1972)
25. Schaller, D., Hellmuth, M., Stadler, P.F.: *AsymmeTree* (2022), <https://github.com/david-schaller/AsymmeTree>, Python package version 2.2
26. Schaller, D., Hellmuth, M., Stadler, P.F.: *AsymmeTree*: A Flexible Python Package for the Simulation of Complex Gene Family Histories. *Software* **1**(3), 276–298 (2022)
27. Sinaimeri, B., Urbini, L., Sagot, M.F., Matias, C.: Cophylogeny reconstruction allowing for multiple associations through approximate Bayesian computation. *Systematic Biology* **72**(6), 1370–1386 (2023)
28. Stolzer, M.L., Lai, H., Xu, M., Sathaye, D., Vernot, B., Durand, D.: Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**(18), i409–i415 (2012)
29. Tofigh, A., Hallett, M., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. on Comput. Biol. Bioinf.* **8**(2), 517–535 (2011)
30. Wang, Y., Mary, A., Sagot, M.F., Sinaimeri, B.: Cappybara: equivalence class enumeration of cophylogeny event-based reconciliations. *Bioinformatics* **36**(14), 4197–4199 (2020)

## A Formal definitions of measures

For any rooted binary tree  $T$ , let  $L(T)$  and  $V(T)$  denote its sets of leaves and vertices, respectively.

A *cherry* is a pair of leaves that are adjacent to a common ancestor node. The *normalized Cherry index* of a tree  $T$  is defined as the number of cherries divided by  $|L(T)|/2$  (corresponding to maximal value). The *normalized Colless index* [8] is defined as

$$C(T) = \frac{2}{|L(T)|(|L(T)| - 1)} \sum_{v \in \mathring{V}(T)} |n_L(v) - n_R(v)|,$$

where  $\mathring{V}(T)$  is the set of internal nodes of  $T$ , and  $n_L(v)$  and  $n_R(v)$  denote the number of descendant leaves in the left and right subtrees of node  $v$ . The *normalized Sackin index* [24] is defined as

$$S(T) = \frac{1}{|L(T)|^2} \sum_{\ell \in L(T)} d(r, \ell),$$

where  $d(r, \ell)$  is the distance from the root  $r$  to leaf  $\ell$ . Smaller values of  $S(T)$  correspond to more balanced trees.

For a pair of host-symbiont trees  $(H, S)$ , let  $A \subseteq L(H) \times L(S)$  denote the set of associations, where  $(h, s) \in A$  if symbiont  $s$  is associated with host  $h$ . We write  $G = (L(H) \cup L(S), A)$  for the resulting bipartite graph, which corresponds to the ecological interaction network of interest.

The *density* of  $G$  is defined as

$$\delta(G) = \frac{|A|}{|L(H)| \cdot |L(S)|}.$$

Let  $d(v)$  denote the degree of a node  $v$  in  $G$ . The *degree assortativity* (see Eq. (21) in [19]) of  $G$  measures the tendency of nodes to connect to other nodes with similar degree and is defined as the Pearson correlation coefficient between the degrees at the endpoints of edges:

$$r(G) = \frac{\sum_{(h,s) \in A} (d(h) - \mu_H)(d(s) - \mu_S)}{\sqrt{\sum_{(h,s) \in A} (d(h) - \mu_H)^2} \sqrt{\sum_{(h,s) \in A} (d(s) - \mu_S)^2}},$$

where

$$\mu_H = \frac{1}{|A|} \sum_{(h,s) \in A} d(h), \quad \mu_S = \frac{1}{|A|} \sum_{(h,s) \in A} d(s).$$

A *host hotspot* is defined as a host leaf  $h \in L(H)$  whose degree exceeds the average host degree plus one standard deviation, a standard threshold used to

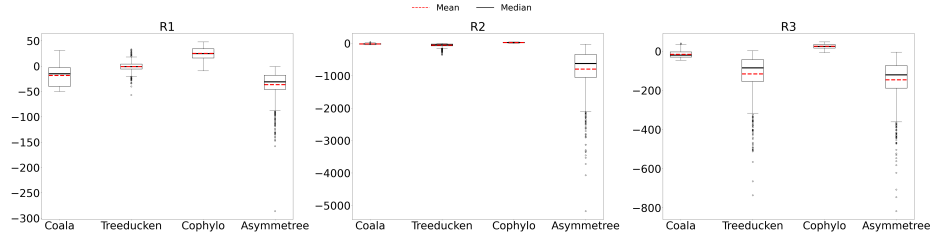


Fig. 8: Boxplots of the difference in the number of leaves between host and symbiont trees for each simulated dataset, across regimes R1–R3

identify highly connected nodes in ecological and interaction networks. Formally, letting

$$\mu = \frac{1}{|L(H)|} \sum_{h \in L(H)} d(h), \quad \sigma = \sqrt{\frac{1}{|L(H)|} \sum_{h \in L(H)} (d(h) - \mu)^2},$$

we define the set of host hotspots as

$$\text{Hot}(G) = \{h \in L(H) : d(h) > \mu + \sigma\},$$

and record the frequency  $|\text{Hot}(G)|/|V(H)|$ .

## B Parameter settings

Table 4 shows the parameter choices for the generators across the three regimes.

## C Removal of extreme values

In order to increase the readability of the plots, Fig 8 was produced on filtered data, where 1% extreme values were removed from all generators. The most important example is **AsymmeTree** in regime R2 where the range of data removed goes from -5000 to -3000.

## D Real data charts

In this section, we provide the characteristics measured on the set of 37 real host-symbiont datasets.

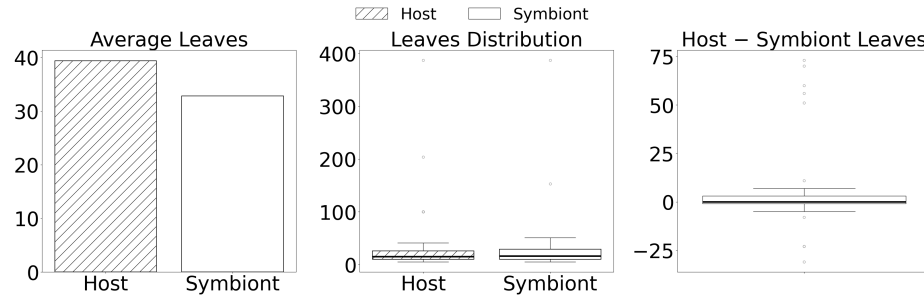


Fig. 9: Distribution of the number of leaves on the 37 real datasets (average, boxplot and difference between Host and Symbiont).

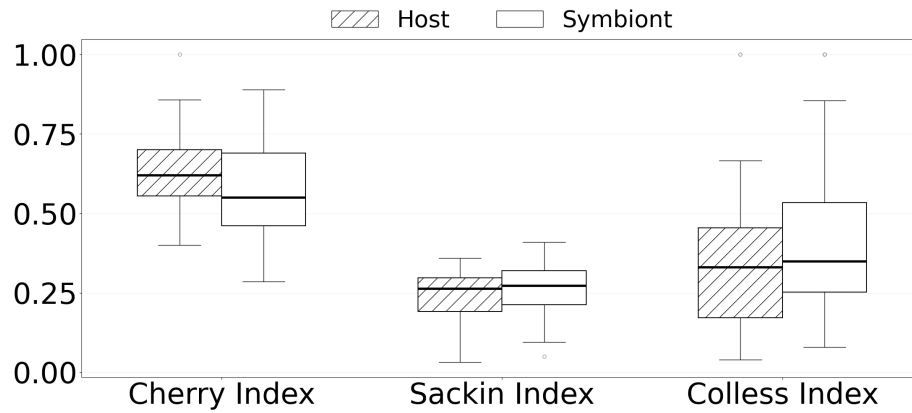


Fig. 10: Boxplots of the (Im)Balance indices measured on the 37 real datasets.

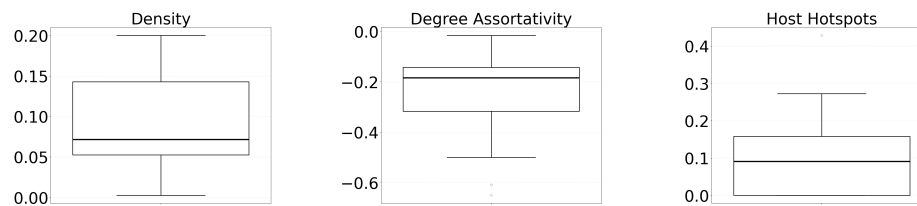


Fig. 11: Boxplots of the association-based measures for the 37 real datasets.

Table 4: Parameter settings for regimes R1–R3 across generators.

<b>R1: high cospeciation and low switching</b>	
<b>Generator</b>	<b>Parameter settings</b>
Coala	$\lambda_H = 0.7$ ; $\mu_H = 0.63$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $p_c \sim U(0.7, 0.9)$ ; $p_s \sim U(0, 0.05)$ ; $(p_d, p_l) \sim (1 - p_c - p_s) * \text{Dirichlet}(1, 1)$ .
treeducken	$\lambda_H = 0.7$ ; $\mu_H = 0.63$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.645$ ; $\lambda_c \sim U(2, 2.5)$ ; $\chi \sim U(0, 0.05)$ ; $T = 2$ ; hs-mode="both"
cophylo	$\lambda_H = 0.7$ ; $\mu_H = 0.63$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.645$ ; $c \sim U(0.7, 0.9)$ ; $s \sim U(0, 0.05)$ ; $f_d(k) \propto k^{-1.6}$ ; $T = T_{MRCA}2$ (host and symbiont).
AsymmeTree	$\lambda_H = 0.7$ ; $\mu_H = 0.63$ ; $\lambda_D = 0.7$ ; $\lambda_L = 0.645$ ; $\lambda_{HGT} \sim U(0, 0.05)$ ; $T = 2$ (host and symbiont).
<b>R2: high host-switches</b>	
Coala	$\lambda_H = 0.7$ ; $\mu_H = 0.24$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $p_c \sim U(0.05, 0.1)$ ; $p_s \sim U(0.5, 0.7)$ ; $(p_d, p_l) \sim (1 - p_c - p_s) * \text{Dirichlet}(1, 1)$ .
treeducken	$\lambda_H = 0.7$ ; $\mu_H = 0.24$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.66$ ; $\lambda_c \sim U(0, 0.05)$ ; $\chi \sim U(2, 2.5)$ ; $T = 2$ ; hs-mode="both"
cophylo	$\lambda_H = 0.7$ ; $\mu_H = 0.24$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.66$ ; $c \sim U(0.05, 0.1)$ ; $s \sim U(0.5, 0.7)$ ; $f_d(k) \propto k^{-1.6}$ ; $T = T_{MRCA}2$ (host and symbiont).
AsymmeTree	$\lambda_H = 0.7$ ; $\mu_H = 0.24$ ; $\lambda_D = 0.7$ ; $\lambda_L = 0.66$ ; $\lambda_{HGT} \sim U(2, 2.5)$ ; $T = 2$ (host and symbiont).
<b>R3: mixed coevolution</b>	
Coala	$\lambda_H = 0.7$ ; $\mu_H = 0.45$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $p_c \sim U(0.2, 0.4)$ ; $p_s \sim U(0.2, 0.4)$ ; $(p_d, p_l) \sim (1 - p_c - p_s) * \text{Dirichlet}(1, 1)$ .
treeducken	$\lambda_H = 0.7$ ; $\mu_H = 0.45$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.6$ ; $\lambda_c \sim U(1.5, 1.8)$ ; $\chi \sim U(1.5, 1.8)$ ; $T = 2$ ; hs-mode="both"
cophylo	$\lambda_H = 0.7$ ; $\mu_H = 0.45$ ; Nb extant host leaves $\sim X \in [15, 50]$ ; $\lambda_S = 0.7$ ; $\mu_S = 0.6$ ; $c \sim U(0.2, 0.4)$ ; $s \sim U(0.2, 0.4)$ ; $f_d(k) \propto k^{-1.6}$ ; $T = T_{MRCA}2$ (host and symbiont).
AsymmeTree	$\lambda_H = 0.7$ ; $\mu_H = 0.45$ ; $\lambda_D = 0.7$ ; $\lambda_L = 0.6$ ; $\lambda_{HGT} \sim U(1.5, 1.8)$ ; $T = 2$ (host and symbiont).