Detecting the ecological footprint of selection

Juliette Luiselli[*], Isaac Overcast[†], Andrew Rominger[‡], Megan Ruffley[§],

Hélène Morlon[†] and James Rosindell[¶]

ABSTRACT

The structure of communities is influenced by many ecological and evolutionary processes, but the way this manifests in classic biodiversity patterns often remains unclear. Here we aim to distinguish the ecological footprint of selection through competition or environmental filtering, from that of neutral processes that are invariant to species identity. We build on existing Massive Eco-evolutionary Synthesis Simulations (MESS), which uses information from three biodiversity axes – species abundances; genetic diversity; and trait variation – to distinguish between mechanistic processes. In order to correctly detect and characterise competition, we add a new form of competition to MESS that explicitly compares the traits of each pair of individuals, allowing us to distinguish between inter- and intra-specific competition. Our results are qualitatively different to those of previous work that only compares each individual's trait to the community mean. We find that neutral forces receive much less support from real systems when trait data is available and incorporated into the inference algorithm. We conclude that gathering more different types of data could be the key to unravelling the mechanisms of community assembly.

Keywords

Eco-evolutionary Model, Community Ecology, Neutral Theory, Competition, Selection

* École Normale Supérieure – PSL & INSA Lyon
† Institut de Biologie de l'École Normale Supérieure
‡ University of Maine
§ Department of Plant Biology, Carnegie Institution for Science
¶ Imperial College London

1 Introduction

Understanding the assembly of ecological communities is a key goal of research in both ecology and evolution. Some studies characterise community assembly as either neutral, where individual species identities are interchangeable (Hubbell 2001), or under selection (*sensu* Vellend 2010), where species identities have influence, for example through abiotic conditions or biotic interactions (Chesson 2000, HilleRisLambers et al. 2012, Adler et al. 2006, Thompson et al. 2020). Such selective interactions may have varying strengths, building a continuum from neutrality (no selection) to strong selection (Gravel et al. 2006). The type and strength of species' interactions has been shown to influence the evolution of species richness (Harmon and Harrison 2015, Rabosky and Hurlbert 2015), and the ecological ramifications of evolving traits (Bassar et al. 2012). Despite recent advances, it remains challenging to characterise selection from empirical data, leading to varied opinions and conclusions.

Whether competition among species is important for structuring ecological communities has been a matter of particular debate (HilleRisLambers et al. 2012, Macarthur and Levins 1967). Many studies support the idea that the competition for limiting resources is the driving factor of niche differentiation, which facilitates coexistence of different species due to a high intra-specific competition, also known as density-dependence (Chesson 2000, Adler et al. 2010, HilleRisLambers et al. 2012). These niche-based competitive interactions are thought to be mediated by organismal traits (HilleRisLambers et al. 2012, Adler et al. 2013). Yet, detecting such competition statistically, and therefore understanding its generality across systems, remains a challenge (HilleRisLambers et al. 2012, Barner et al. 2018, Freilich et al. 2018). In contrast, neutral theory, as the prevailing alternative model to niche-based competition, is much easier to test statistically because it is a low-complexity model (Rosindell and Phillimore 2011), but it is unclear whether tests that reject or fail to reject neutrality do so for valid reasons (McGill et al. 2007, Chisholm and Pacala 2010, Rosindell et al. 2012), or whether false positives or false negatives prevail.

Being able to retrieve the competitive factors from empirical data would help us to improve our understanding of competitive interactions, in ecology (shorter timescales and individual interactions) as well as in evolution (longer timescale and species interactions). One of the reasons why this has proved elusive may be that only limited data of a few types have been used to compare model predictions to reality. Multiple complementary data axes should provide more inference potential (McGill et al. 2007). To date, competition and neutrality have largely been evaluated using species abundance distributions (SAD), as this data is historically the easiest to collect (Hubbell 2001, Chesson 2000, Chisholm and Pacala 2010). Other data have been used including phylogenies, which account for the evolutionary history of the local species and their past interactions (Webb et al. 2002, Jabot and Chave 2009, Nuismer and Harmon 2015), metabarcoding data, which gather abundances and genomic proximity information (Baselga et al. 2015), a combination of genetic data and SADs (Overcast et al. 2019, Vellend 2005), and traits, which can inform on the interactions between the species and with their environment (HilleRisLambers et al. 2012, Adler et al. 2013, Ruffley et al. 2019, Aristide and Morlon 2019). Yet, these data are generally used separately.

The Massive Eco-evolutionary Synthesis Simulations (MESS) model of Overcast et al. (2021) allows testing mechanistic hypotheses across a combination of three data axes: species abundances, population genetic variation and trait values. These three axes reflect a variety of processes operating over a variety of time scales, from a few generations (abundances) to several tens of thousands of generations (genetic variation). Moreover, traits and genetic variation can account for the information present in phylogeny data, and SAD and some genetic variability can be recovered from metabarcoding data: the three chosen axes cover the available and collectable data and diverse biological phenomena. MESS is a simulation model that can be fitted to empirical data using machine learning procedures, and thus is an ideal tool to study the eventual traces of selection in community assembly data.

Selection in the MESS model, in keeping with conventional wisdom (HilleRisLambers et al. 2012, Adler et al. 2013), is driven by evolving traits and interactions with either the environment or other individuals. However, an individual's fitness in the competition model of MESS is determined by the distance of its trait

to the mean trait value of individuals in the local community, rather than by pairwise distances between the individual's trait and all other traits in the community. This "mean competition" is attractive because it delivers substantial computational gains, which are important to run enough simulations for machine learning based inference from data. Mean competition is often used to model the probability of persistence of a species (Ruffley et al. 2019) and has the advantage of modelling a diffuse biotic background, with which individuals interact (McGill et al. 2006). It is, however, a weak approximation for the mechanistic reality where competition is driven by interactions between individual organisms (Berger et al. 2008, Vázquez et al. 2007). Simulating mean competition may thus generate patterns that do not reflect real competitive processes, and so the generated data would not be appropriate for detecting competition in empirical data.

Here we investigate the importance of competition in community assembly and our ability to recover it from empirical data through simulation models. To do this, we apply MESS with a new and more realistic competition model. We find that previous conclusions about the presence and strength of selection may be artefacts of the mean competition simulation method. We find, consistent with current thinking, that more data types enhance inference power. We also find that trait data are most helpful in detection of selective forces as an alternative to neutral ones.

2 Material & Methods

2.1 The MESS model

Our simulations are individual based with a distinct metacommunity and local or island community (Hubbell 2001, Rosindell and Harmon 2013, Overcast et al. 2019). Simulations are run as a time series, enabling the study of both dynamic equilibrium and non-equilibrium behaviour. A single trait value is associated with each

species identity, which can be used in different ways to model selection and non-neutral dynamics. After the community simulation is completed and population size fluctuations for each species are known, this information is used to constrain a coalescence-based simulation of genetic variation within each species (Kelleher et al. 2016).

Following the MESS model of Overcast et al. (2021), we simulate a fixed number of individuals in the local community. Each individual i is characterized by its trait value z_i . At each time step, one individual dies and is replaced by another individual, which comes either from immigration from the metacommunity, at rate m, or from a reproduction event within the local community. Speciation occurs by point mutation with probability v at each reproduction event. The metacommunity is modelled as a very large regional pool, which is fixed with respect to the timescale of the assembly process in the local community. It arises from ecological and evolutionary processes, including speciation *sensu* Hubbell (2001). Under the assumption of neutrality, the death probability $P_{neutral}$ of any given individual i in the local community at each time step is computed as

$$P_{neutral}(i)=\frac{1}{J} \tag{1}$$

where J is the size of the local community measured as the number of individuals.

Selection is incorporated in MESS by computing, at each time step, each individual's probability of death according to a chosen community assembly model (competition or environmental filtering).

In the environmental filtering model, the trait value of each individual is compared to an optimal trait value that depends solely on the environment. The death rate q_{filt} of any given individual i is computed as

$$q_{filt}(i)=1-\exp\left[-s_E\left(z_i-z_E\right)^2\right] \tag{2}$$

where z_E is the environmental optimum and s_E determines the strength of the filtering. Intraspecific variation

is assumed to be negligible in face of interspecific variation, and all individuals of the species a have the same trait value z_a which represents the mean phenotype of the species. The probability of death in the next time step, for any given individual is given by the normalized death rate $P_{filt}(i)=\frac{q_{filt}(i)}{\sum_{i=1}^{J} q_{filt}(j)}$.

In Overcast et al. (2021), competition is modelled by a mean-field approximation: the trait value of an individual is compared to the mean trait value of the local community. The death rate q_{MF} of any given individual i is then computed as

$$q_{MF}(i)=\exp\left[-s_E\left(z_i-\bar{z}\right)^2\right] \tag{3}$$

where $\bar{z}$ is the local community mean trait and s_E determines how quickly competitive pressure decays with the distance between trait values. As previously, and for each further death rate q, we derive the death probability P through normalization : $P_{MF}(i)=\frac{q_{MF}(i)}{\sum_{j=1}^{J} q_{MF}(j)}$

The mean-field approach collapses all trait differences into one value and can therefore generate counter-intuitive results. For example, the distribution of species across the trait axis might be bimodal as two groups of species diverge away from the central mean value, leading to an obvious gap around the mean (see Appendix A.3). The area around the mean in trait space is thus free from species and competition but is still the most penalised trait, while denser areas, further away from the mean but with more species, are favoured.

Here, we correct this artefact by using a new competition model based on pairwise comparisons between all individuals. In our model, the death rate q_{pair} of any given individual i is based on the mean of all pairwise trait differences with the other individuals in the local community:

$$q_{pair}(i) = \sum_{j=1, j \neq i}^{J} \exp\left[-s_E (z_i - z_j)^2\right] \tag{4}$$

The added computational cost of the pairwise model was partially offset by optimizing the underlying data structures of the original MESS model (Overcast et al. 2021), enabling faster simulations (see Appendix A.2). In contrast to the mean competition model, the pairwise competition model is expected to produce uniformly and regularly distributed species along the trait axis, which is confirmed in our simulations (see Appendix A.3). The pairwise competition model does not, however, allow us to refine the strength of intra-specific competition: individuals of the same species have the exact same trait value and thus the exponential in equation (4) is always equal to 1.

We also implement a third "β-competition" model that introduces an interaction matrix parameter β_{ij} to modulate competition strength between all possible pairs of individuals. Larger values of β_{ij} increase the strength of competition between individuals i and j. We set $\beta_{ij} = \beta_{\text{intra}}$ when individuals with indexes i and j are conspecific, and $\beta_{ij} = \beta_{\text{inter}}$ when they are heterospecific. The resulting death rate is given by

$$q_{\beta}(i) = \sum_{j=1, j \neq i}^{J} \beta_{ij} \exp\left[-s_E (z_i - z_j)^2\right] \tag{5}$$

By allowing intra- and inter-specific competition to differ according to a parameter, we are in effect modelling differing levels of negative density dependence: $\beta_{\text{intra}} >> \beta_{\text{inter}}$ corresponds to strong intraspecific density dependence whilst $\beta_{\text{intra}} << \beta_{\text{inter}}$ corresponds to no density dependence. We leave the $\beta_{\text{intra}} << \beta_{\text{inter}}$ case for future work, noting that preliminary tests suggest the model will lead to mono dominance. The three competition models that we study here are summarised in Fig. 1. Notably, the death probability, computed from the given death rates, for each individual in the cases of filtering or competition converges toward a neutral probability $\frac{1}{J}$ for each individual as the strength of selection s_E converges toward 0.

2.2 Exploration of *in silico* experiments

To explore the behaviour of the proposed competition models and understand how the different models affect the outcome of community assembly, we ran 10 000 simulations for each of the five community assembly models (neutral, filtering, mean competition, pairwise competition and β-competition), covering wide ranges of possibilities for the main parameters of the simulations : the age of the community (through Λ, a parameter used to quantify the progress of the simulation toward equilibrium), the number of individuals J, the strength of the ecological filtering or competition s_E, the strength of inter-individuals interactions β, the migration rate m, the speciation rate ν, and the abundance/effective population scaling factor α (see Appendix A.1).

Each simulation is characterised by a number of summary statistics along each data axis (species abundances, population genetic variation and trait values): the first moments of each community-wide distribution, Spearman rank correlation among all data axes, differences between metacommunity and local community values of trait mean and standard deviation, and Hill numbers to quantify the shape of each distribution (Chao et al. 2014). The temporal trends are studied in terms of Λ, a parameter used to quantify the progress of the simulation toward equilibrium (Overcast et al. 2019), consistent with the original MESS model Overcast et al. (2021). A community is considered at equilibrium, and Λ=1, when the initial conditions are no longer detectable in the system, and this advancement toward equilibrium is measured as the proportion of individuals in the community descending from a lineage that colonized during the simulation. We visually inspected the resulting simulations by collapsing simulated summary statistics into principal components using the built-in PCA function of MESS (Overcast et al. 2021) (Fig. 2). This enabled us to distinguish between the different community assembly models.

2.3 Machine learning and inference

We follow the same procedure as Overcast et al. (2021) for model classification and parameter estimation: Random Forest (Breiman 2001) with python and the scikit-learn module (v0.20.3, Pedregosa et al. (2011)). We first train a machine learning classifier in a supervised fashion on 50,000 simulated datasets (10,000 for each assembly model). We then use the trained classifier to predict model class probabilities for each of the empirical datasets. A confidence percentage is associated to each model. We quantified classifier accuracy using 5-fold cross-validation on simulated data, and evaluated model misclassification by compiling these results into a confusion matrix. We evaluated classifier accuracy using two different suites of simulated data axes, one composed of only SAD and genetic data, and the other composed of trait values and genetic data, to reflect the data configurations of our empirical datasets. We describe an ideal case scenario, with all three data axis, in the appendix, which shows that an extensive gathering of empirical data would substantially improve the accuracy of our predictions (Appendix, Fig. 7).

2.4 Study of empirical datasets

We used the empirical datasets following Overcast et al. (2021): 1) a spider community from Réunion island with standardized sampling for abundance and genetic diversity of ten 50 m x 50 m plots and 1282 individuals sequenced for one 500bp mtDNA region (COI) (Emerson et al. 2017); 2) two weevil communities from two Mascarene islands (one from Réunion and one from Mauritius) which have been densely sampled for abundance and sequenced for one mtDNA region (600bp COI) at the community-scale (Kitson et al. 2018); 3) three subtropical rain forest tree communities scored for multiple continuous traits and shotgun sequenced for whole cpDNA (Rossetto et al. 2015); 4) Galapagos snail communities collected from all major islands (three in total), sampled for one mtDNA region (500bp COI; (Kraemer et al. 2019)) and scored for two continuous traits (Triantis et al. 2016).

We compared summary statistics linked to the SAD, genetic diversity and traits computed on the empirical data to those computed on 50,000 simulations (10,000 for each community assembly model).

3 Results

The PCA (see Fig. 2) shows that across time, simulations progressively differentiate themselves into clusters according to their community assembly model: the community assembly model is easier to discriminate in older communities. β-competition points are broadly spread in the figure, and especially hard to distinguish from pairwise competition. However, the first two PCA components only account for around 30% of the variance. The groups formed by pairwise competition and β-competition partially overlap with the neutral simulation group. The filtering and mean competition groups resemble one another during the approach to equilibrium ($\Lambda < 1$).

The same clustering is also found in the temporal dynamics of the summary statistics for the three data axes along time (Fig 3): the summary statistics from the mean competition and environmental filtering simulations most often follow similar trajectories. Trajectories from the β-competition and pairwise competition simulations were also similar to each other (but distinct from the former). The neutral simulations most closely resembled the β-competition and pairwise competition simulations.

Results of the misclassification rates when using trait values and genetic diversity confirm that the community assembly model can be correctly recovered from the simulation results in around 50% of the cases, while a random classifier would only be correct in 20% of the cases (see Fig. 4). The greatest confusion is between pairwise competition and β-competition, and the neutral model was the best recovered. Filtering and mean competition models were also easily distinguished by the inference procedure. A confusion matrix with SAD and genetic diversity data shows similar results (See Appendix A.4, Fig. 6). The best classification is nevertheless achieved when all three data types are used (See Appendix A.4, Fig. 7), but the combination of all three are not yet available for empirical communities.

We consider first the three datasets with SAD and genetic data: for the Reunion spider dataset, competition is detected at around 40% (Fig. 5) while it was not inferred in Overcast et al. (2021). For the two Mascarene weevil datasets, the confidence percentage predicted for the neutral model remains the same as in the previous MESS analysis by Overcast et al. (2021), but the circa 40% predicted for both mean competition and filtering in

the original analysis is overtaken by the combination of pairwise competition and β-competition (Fig. 5). Pairwise competition dominates over β-competition model, and over mean competition, which now receives no support. With inclusion of more nuanced competition models, the inference of environmental filtering totally disappears for these datasets compared to Overcast et al. (2021).

Environmental filtering is only significantly detected in the subtropical rain forest tree and Galapagos snail communities, which are also the datasets that contain trait measurements. For empirical data that include trait information, the β -competition model was overall a better fit than the other competition models.

For all datasets, the added percentages of all three competition models is greater than the percentage predicted for competition in the original MESS publication. Among competition models, the mean competition is under-represented, and in many cases totally absent when other competition models are available as an alternative.

4 Discussion

We studied the power of inference in community assembly models given different combinations of empirical data. A key advance compared to previous work is the investigation of a more sophisticated competition model that considers the interaction between pairs of individuals instead of making a mean field approximation. Our results show that the mean field approximation in this case can lead to underestimation of the role of competition and overestimation of the role of environmental filtering. This is consistent with our observation that mean competition and environmental filtering produce very similar results in our PCA on approach to equilibrium ($\Lambda < 1$) (Fig. 2). The reason may be because mean competition produces a bimodal trait distribution that is effectively filtering against midpoints in the trait space.

In our empirical data analysis, the mean competition model receives almost no support when pairwise and β-competition models are added to the analysis as alternatives. This is consistent with the intuition that those two models better reflect biological reality of competition. Indeed, mechanistic simulations with the pairwise competition model were mostly classified by the original MESS inference method (Overcast et al. 2021) as

mean competition (Appendix A.4, Table 3) though sometimes classified as neutral or environmental filtering. This demonstrates that competition can be mistaken for neutrality or environmental filtering if the model of competition is of insufficient complexity. The disappearance of support for the mean competition in our new classifier further supports the hypothesis that pairwise competition is a better description of to the empirical data.

Pairwise and β-competition simulations have on average more species than the mean competition simulations (see Fig. 3). This is expected because selecting for evenly distributed species across the trait space, as in these competition cases, allows for more diversity than selecting for two diverging groups of species, as in the mean competition case. As β-competition depicts density-dependence more accurately, we could expect it to have a significant advantage over pairwise competition. However, the PCA results (Fig. 2) show that simulation outcomes mostly overlap between pairwise and β-competition: they could be interpreted as a single indistinguishable category. This is further supported by the confusion matrices (Fig. 6 and Appendix A.4), which suggest that β_{ij} has no crucial importance in the simulation outcome. Density dependence may therefore not play a major role because it was not easily detected by the model selection process. Future work could add further parameters and retrieved summary statistics to better model and better detect density-dependence, but will likely come at a high computational cost.

The striking proximity of the pairwise and β-competition simulations to the neutral simulations in our PCA results (Fig. 2) was not apparently consistent with our confusion matrices (Fig. 4 and Appendix A.4). The random forest algorithm seems to be able to distinguish between neutral and non-neutral models, which are indistinguishable for the human eye in the PCA, as well as in most summary statistics (Fig. 3). It is easy to see, therefore, how weaker inference procedures, backed with less detailed empirical data, could misinterpret competition as neutrality and consequently that competition based simulations may often resemble neutral simulations in terms of the community properties studied. This may be an example of emergent neutrality (Holt 2006), and consistent with niche-neutral models (Chisholm and Pacala 2010) where communities

consist of multiple niches but with individuals of multiple species interacting neutrally within each niche. Despite the potential for confusion between mechanisms, our results show that the combination of ecological data (abundances / traits) and evolutionary (genetic) data, together with machine learning, remains a most promising approach to distinguish neutrality and selection much better than could be achieved with data sets consisting of only one type of data.

The striking difference in our inferences based on the type of data used have implications for the kinds of data we gather to study community assembly. Selection was revealed best by our inference procedure when trait data are available. Indeed, the neutral model was the best fitting for the spider and weevil datasets that lack trait data. Contrary to what has been suggested in the metabarcoding literature (Chen et al. 2019), this result shows that genetic data alone may not suffice to measure the selective pressure on a group and traits may be needed (Kraft et al. 2015, McGill et al. 2006). A comparison of the confusion matrices shows, however, that the presence of trait data is not essential for detecting filtering or competition (Fig. 4 and Appendix A.4). The differences in our inference results between systems with different data types might therefore be due to real community differences or due to an artefact of the types of data available.

Our results could also be extended to the macroevolutionary models: if competition is already taken into account in phenotypic evolution models (Drury et al. 2016), it often relies only on the species trait value. However, if competitive interactions are mostly pairwise interactions, then the relative abundance of each species would also be relevant when modelling evolution using their trait value, as it would largely influence the strength of intra- and inter-specific competition undergone.

During our inference process on empirical data, the selected model is either neutral, competitive (in one of a number of ways) or with environmental filtering. There was not a single model simulated that combines all these processes in varying amounts. A fruitful direction for future work would be to simulate a simultaneous combination of all the processes in a single model. This would enable us to verify that our inferences (choosing between starkly contrasting models) correspond to what would be predicted by a more nuanced and

continuous view of the community assembly models. Another possible direction would be to add intraspecific trait variation. Such variation could enable a different handling of the distinction between inter- and intra-specific competition and could allow several species to occupy the same niche (Scheffer and van Nes 2006). In future work, the β factors used in the simulations could also be refined to allow for differences among each pair of species, so as to reflect species-specific interactions, which may generalise to include positive interactions as well as direct competition. This would however necessitate a wide parameter exploration in the simulation, at an exponentially increasing computational cost.

We hope that future empirical studies will provide datasets with all three types of data (genetic diversity, SAD and trait values) rather than having to rely on only two of these three as we did in our present work. Such a comprehensive dataset could also be used to shed light on the underlying technical reasons why no environmental filtering is inferred from data without availability of traits: is it the presence of trait values, or the absence of species abundances distributions, that most influences the lack of support for environmental filtering? It could also be used to test how much of the inferred results depend on the type of data available and how much are resulting from ecological differences in the community under study.

Our study underlines the importance of the range of empirical data available to detect the ecological footprint of selection, in contrast to neutrality. Our results reiterate a warning that we should not jump too quickly to conclusions about the presence or absence of selection, especially when only one type of data is available. We show that the pairwise competition model, which cannot be easily distinguished in the data from the β-competition model, is a clear improvement of the previously used mean competition model. Failure to detect pairwise competition in other data sets likely means that competition does not act this way, not that competition, or selection in a broader sense, are absent. We hope that this work will pave the way to improved mechanistic eco-evolutionary models and associated inference procedures for community assembly. We also hope to inspire new data collection and place greater emphasis on the synergistic power of genetic, abundance and trait data when analysed jointly.

References

Adler, P. B., Ellner, S. P. et al. 2010. Coexistence of perennial plants: an embarrassment of niches. *Ecology Letters*, 13(8): 1019–1029, doi:10.1111/j.1461-0248.2010.01496.x.

Adler, P. B., Fajardo, A. et al. 2013. Trait-based tests of coexistence mechanisms. *Ecology Letters*, 16(10): 1294–1306, doi:10.1111/ele.12157.

Adler, P. B., HilleRisLambers, J. et al. 2006. Climate variability has a stabilizing effect on the coexistence of prairie grasses. *Proceedings of the National Academy of Sciences*, 103(34): 12793–12798, doi:10.1073/pnas.0600599103.

Aristide, L. and Morlon, H. 2019. Understanding the effect of competition during evolutionary radiations: an integrated model of phenotypic and species diversification. *Ecology Letters*, 22(12): 2006–2017, doi:10.1111/ele.13385.

Barner, A. K., Coblentz, K. E. et al. 2018. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology*, 99(3): 557–566, doi:10.1002/ecy.2133.

Baselga, A., Gómez-Rodríguez, C. et al. 2015. Multi-hierarchical macroecology at species and genetic levels to discern neutral and non-neutral processes. *Global Ecology and Biogeography*, 24(8): 873–882, doi:10.1111/geb.12322.

Bassar, R. D., Ferriere, R. et al. 2012. Direct and indirect ecosystem effects of evolutionary adaptation in the trinidadian guppy (poecilia reticulata). *The American Naturalist*, 180(2): 167–185, doi:10.1086/666611.

Berger, U., Piou, C. et al. 2008. Competition among plants: Concepts, individual-based modelling approaches, and a proposal for a future research strategy. *Perspectives in Plant Ecology*, Evolution and Systematics, 9(3): 121 – 135, doi:10.1016/j.ppees.2007.11. 002.

Breiman, L. 2001. Random forests. Machine Learning, 45(1): 5–32, doi:10.1023/A: 1010933404324.

Chao, A., Chiu, C-H. et al. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution and Systematics*, 45:1, 297-324, doi:10.1146/annurev-ecolsys-120213-091540

Chen, W., Ren, K. et al. 2019. Stochastic processes shape microeukaryotic community assembly in a subtropical river across wet and dry seasons. *Microbiome* 7(1): 138, doi:10.1186/ s40168-019-0749-8.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, 31(1): 343–366, doi:10.1146/annurev.ecolsys.31.1.343.

Chisholm, R. A. and Pacala, S. W. 2010. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proceedings of the National Academy of Sciences*, 107(36): 15821–15825, doi:10.1073/ pnas.1009387107.

Drury, J., Clavel, J., et al. 2016. Estimating the effect of competition on trait evolution using Maximum Likelihood Inference. *Systematic Biology*, 65(4):700–710, doi:10.1093/sysbio/syw020

Emerson, B. C., Casquet, J. et al. 2017. A combined field survey and molecular identification protocol

for comparing forest arthropod biodiversity across spatial scales. *Molecular Ecology Resources*, 17(4): 694–707, doi:10.1111/1755-0998.12617.

Freilich, M. A., Wieters, E. et al. 2018. Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? *Ecology*, 99(3): 690–699, doi:10.1002/ecy.2142.

Harmon, L. J. and Harrison, S. 2015. Species diversity is dynamic and unbounded at local and continental scales. *The American Naturalist*, 185(5): 584–593, doi:10.1086/680859.

HilleRisLambers, J., Adler, P. et al. 2012. Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics*, 43(1): 227–248, doi:10.1146/annurev-ecolsys-110411-160411.

Holt, R. 2006. Emergent neutrality. *Trends in Ecology & Evolution*, 21(10): 531–533, doi:10.1016/j.tree.2006.08.003.

Hubbell, S. P. 2001. The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32). *Princeton University Press*, ISBN 978-0-691-02128-7.

Gravel, D., Canham, C.D. et al. 2006. Reconciling niche and neutrality: the continuum hypothesis. *Ecology Letters*, 9: 399-409, doi:10.1111/j.1461-0248.2006.00884.x

Jabot, F. and Chave, J. 2009. Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters*, 12(3): 239–248, doi:10.1111/j.1461-0248.2008.01280.x.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5), e1004842, doi:10.1371/journal.pcbi.1004842.

Kitson, J. J. N., Warren, B. H. et al. 2018. Community assembly and diversification in a species-rich radiation of island weevils (coleoptera: Cratopini). *Journal of Biogeography*, 45(9): 2016–2026, doi:10.1111/jbi.13393.

Kraemer, A. C., Philip, C. W. et al. 2019. Trade-offs direct the evolution of coloration in galapagos land snails. *Proceedings of the Royal Society B: Biological Sciences*, 286(1894): 20182278, doi:10.1098/rspb.2018.2278.

Kraft, N. J. B., Godoy, O. et al. 2015. Plant functional traits and the multidimensional nature of species coexistence. *Proceedings of the National Academy of Sciences*, 112(3): 797–802, doi:10.1073/pnas.1413650112.

Macarthur, R. and Levins, R. 1967. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, 101(921): 377–385, doi:10.1086/282505.

McGill, B. J., Enquist, B. J. et al. 2006. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, 21(4): 178–185, doi:10.1016/j.tree.2006.02.002.

McGill, B. J., Etienne, R. S. et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10): 995–1015, doi:10.1111/j.1461-0248.2007.01094.x.

Nuismer, S. L. and Harmon, L. J. 2015. Predicting rates of interspecific interaction from phylogenetic trees. *Ecology letters*, 18(1): 17–27, doi:10.1111/ele.12384.

Oliphant, T. E. 2006. A guide to NumPy, vol. 1. *Trelgol Publishing USA*.

Overcast, I., Emerson, B. C. et al. 2019. An integrated model of population genetics and community ecology. *Journal of Biogeography*, 46(4): 816–829, doi:10.1111/jbi.13541.

Overcast, I., Ruffley, M. et al. 2021. A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. *Molecular Ecology Resources*, 21(8): 2782-2800, doi:10.1111/1755-0998.13514.

Pedregosa, F., Varoquaux, G. et al. 2011. Scikit-learn: Machine learning in python. Machine Learning in Python, 6.

Rabosky, D. L. and Hurlbert, A. H. 2015. Species richness at continental scales is dominated by ecological limits. *The American Naturalist*, 185(5): 572–583, doi:10.1086/ 680850.

Rosenberg, N. A. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3: 380–390, doi:10.1038/nrg795

Rosindell, J. and Harmon, L. J. 2013. A unified model of species immigration, extinction and abundance on islands. *Journal of Biogeography*, 40: 1107–1118, doi:10.1111/jbi.12064

Rosindell, J., Hubbell, S. P. et al. 2012. The case for ecological neutral theory. *Trends in Ecology & Evolution,* 27(4): 203-208, doi:10.1016/j.tree.2012.01.004

Rosindell, J. and Phillimore, A. B. 2011. A unified model of island biogeography sheds light on the zone of radiation: A unified model of island biogeography. *Ecology Letters*, 14(6): 552–560, doi:10.1111/j.1461-0248.2011.01617.x.

Rossetto, M., McPherson, H. et al. 2015. Where did all the trees come from? a novel multispecies approach reveals the impacts of biogeographical history and functional diversity on rain forest assembly. *Journal of Biogeography*, 42(11): 2172–2186, doi:10. 1111/jbi.12571.

Ruffley, M., Peterson, K. et al. 2019. Identifying models of trait-mediated community assembly using random forests and approximate bayesian computation. *Ecology and Evolution*, 9(23): 13218–13230, doi:10.1002/ece3.5773.

Scheffer, M. and van Nes, E.H. 2006. Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences*, 103 (16) 6230-6235, doi: 10.1073/pnas.0508024103

Thompson, P. L., Guzman, L. M. et al. 2020. A process-based metacommunity framework linking local and regional scale community ecology. *Ecology Letters*, 23(9): 1314–1329, doi:10.1111/ele.13568.

Triantis, K. A., Rigal, F. et al. 2016. Discordance between morphological and taxonomic diversity: land snails of oceanic archipelagos. *Journal of Biogeography*, 43(10): 2050–2061, doi:10.1111/jbi.12757.

Vázquez, D.P., Melián, C.J. et al. 2007, Species abundance and asymmetric interaction strength in ecological networks. *Oikos*, 116: 1120-1127. doi:10.1111/j.0030-1299.2007.15828.x

Vellend, M. 2005. Species diversity and genetic diversity: Parallel processes and correlated patterns. *The American Naturalist*, 166(2): 199–215, doi:10.1086/431318.

Vellend, M. 2010. Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology*. 85(2): 183-206, doi:10.1086/652373.

Webb, C. O., Ackerly, D. D. et al. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33(1): 475–505, doi:10.1146/annurev.ecolsys.33. 010802.150448.

5 Tables and figures

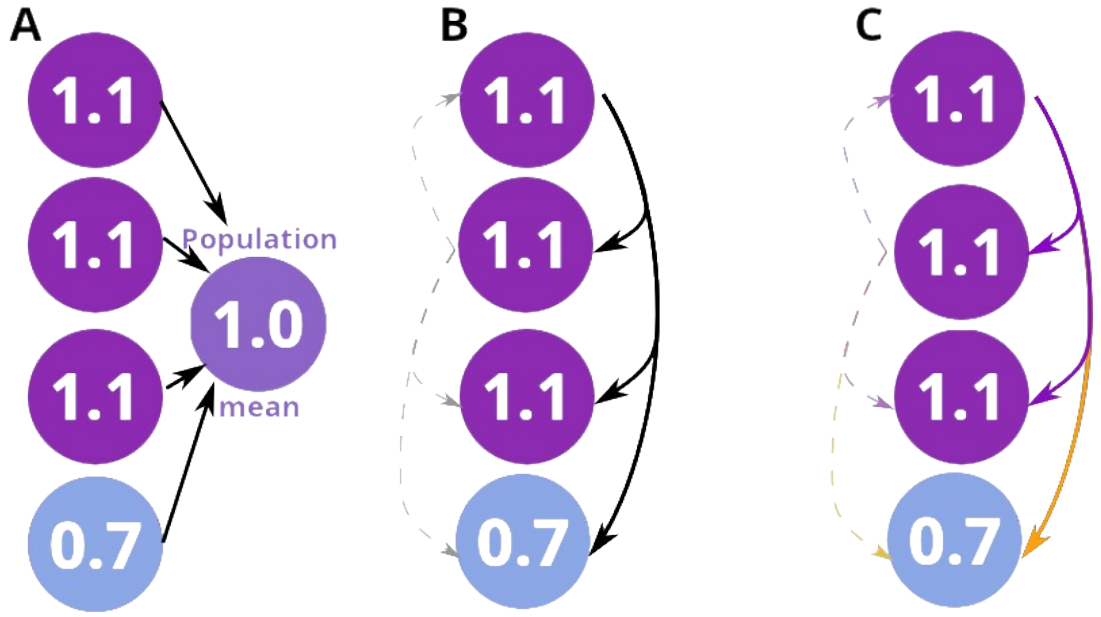

Figure 1: **Depiction of the different forms of competition.**

Each circle represents an individual, with the colour specifying which species it belongs to and the number its trait value. The effect of competition on fitness (symbolized by arrows) is depicted here for the individual at the top (solid arrows) and second from top (dotted arrows). (A) Mean competition model: the trait value of the individual is compared to the mean trait value for the community. (B) Pairwise competition: the trait value of the individual is compared individually to each other individual's trait value. (C) β-competition: the trait value of the individual is compared individually to each other individual's trait value, weighted by a factor depending on whether the pair of individuals belong to the same species.

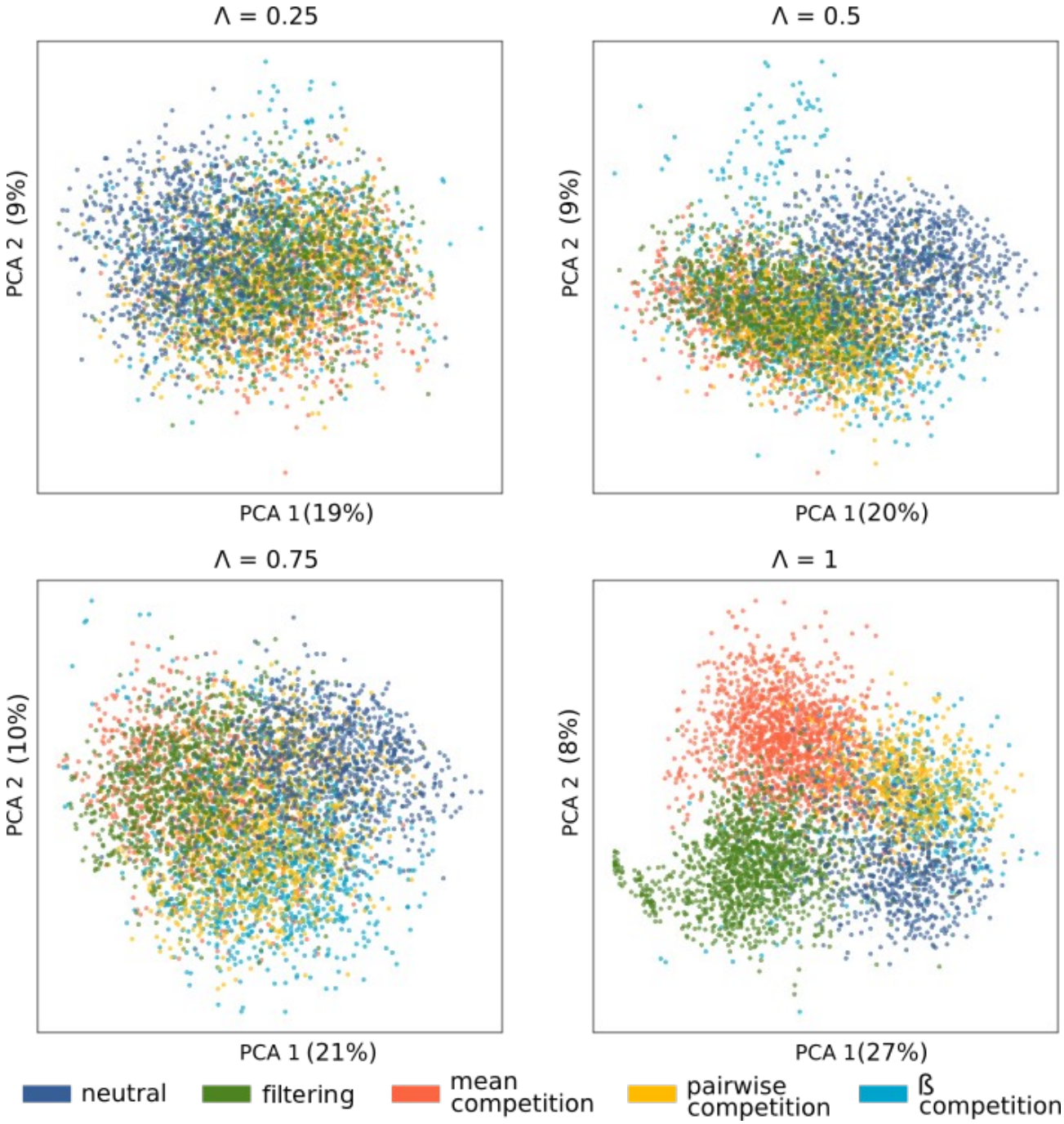

Figure 2: **Plot of the first two principal components of the simulation summary statistics at different equilibrium stages (Λ).**

Community assembly models depicted are neutral (dark blue), environmental filtering (green), mean competition (orange), pairwise competition (yellow) and β-competition (light blue).The percentage of variance explained is indicated for each component.

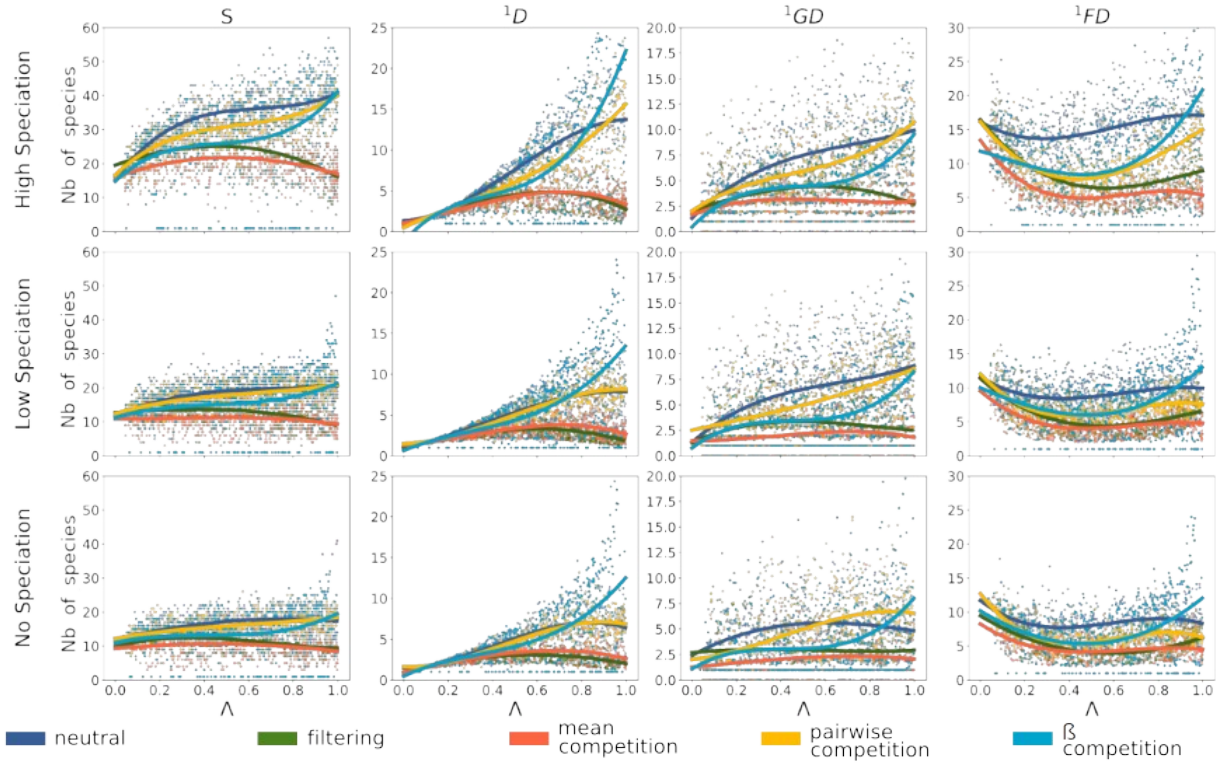

Figure 3: **Community summary statistics through time for the different community assembly models.**

This plot shows the temporal change in some summary statistics for the five focal community assembly models at three different speciation rates: No (ν = 0), Low (ν = 0.0005) and High (ν = 0.005). Community assembly models depicted are neutral (dark blue), environmental filtering (green), mean competition (orange), pairwise competition (yellow) and β-competition (light blue). Each subpanel shows the resultant summary statistic computed at equally spaced time points for over 1500 simulations for each model, with a community size J = 1000, an ecological strength s_E = 0.1 and a migration rate $m = 5e - 3$. Simulated values are depicted as points and a least square polynomial is fit with the poly fit function of NumPy v.19.0 (Oliphant 2006) to better illustrate the trajectory. The far left column of panels illustrate species richness on the y-axis (S). The y-axes of the other columns illustrate the Hill number of order 1 for abundance, genetic diversity, and trait values, respectively.

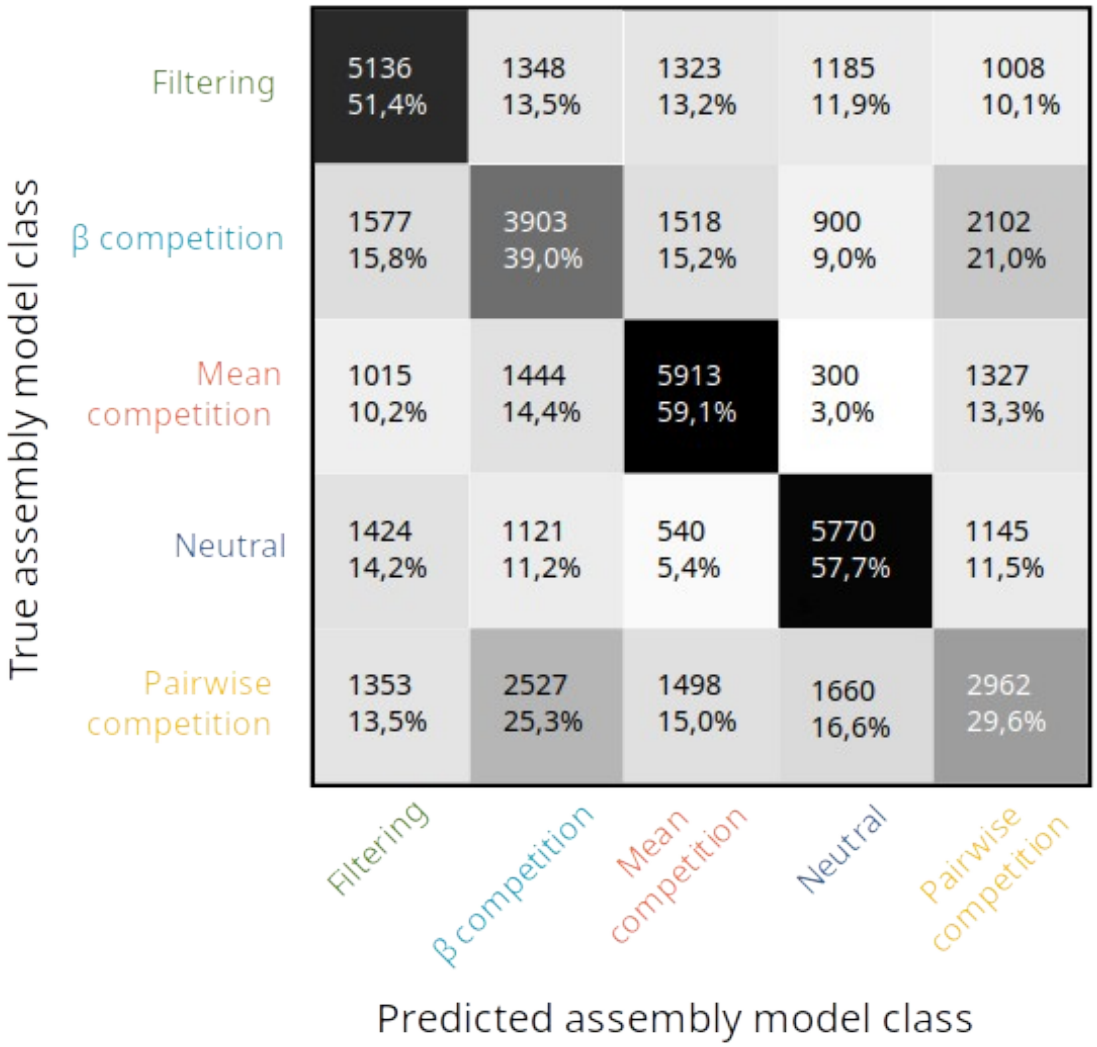

Figure 4: **Machine learning classification confusion matrix** for datasets simulated under the 5 community assembly models and classified using only trait and genetic diversity data (as is the case for the subtropical forest trees and Galapagos snails datasets). Numbers correspond to the number of datasets simulated under a given community assembly model (rows) that are classified in each model (column). In the case of perfect classification, all values would fall along the diagonal. Percentages indicate the proportion of simulations run with one given class (row) assigned to the column class.

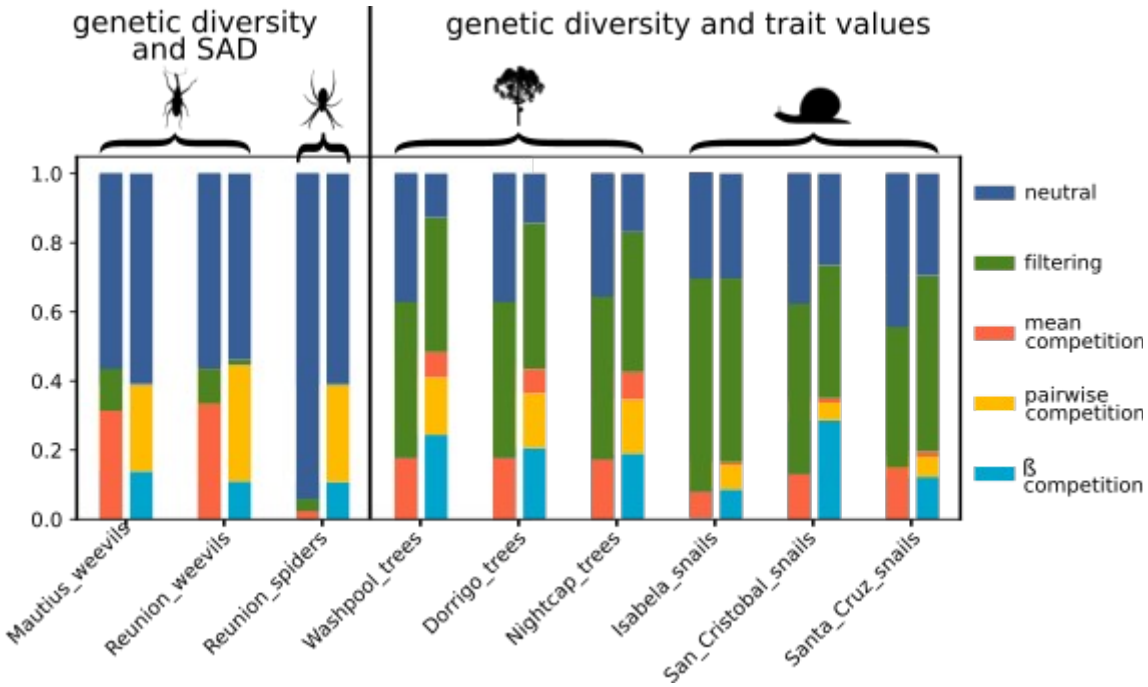

Figure 5: **Machine learning classification probabilities for each empirical community for five focal community assembly models.** For each dataset, the first bar depicts the result of the original MESS model (Overcast et al. 2021) and the second bar the result with our new competition models. The proportion of colour within each bar represents the proportional predicted model class for neutrality (dark blue), environmental filtering (green), mean competition (orange), pairwise competition (yellow) and β-competition (light blue).