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 processes, both ecological and evolutionary, but the influence of these processes on biodiversity patterns often remains unclear. The aim of this work is to distinguish the ecological footprint of differing formulations of competition and environmental filtering from that of neutral processes that are invariant to species identity. We build our work on "massive eco-evolutionary synthesis simulations" (MESS), which uses information from three biodiversity axes – species richness and abundance; population genetic diversity; and trait variation in a phylogenetic context – to distinguish between processes with a mechanistic model. 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. We find that this addition is essential to properly detect and characterise competition, yielding different results than the existing simpler model that only compares species' traits to the community mean. We find from model selection that neutral forces receive much less support from real systems when trait data is available and incorporated into the inference algorithm.

Keywords

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

[*]École Normale Supérieure - PSL
[†]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 biological communities is a key goal of both ecological and evolutionary research. There are wide ranging views on how the nature and strength of species interactions influences community assembly. This has implications on, for example, the evolution of species richness (Harmon and Harrison 2015, Rabosky and Hurlbert 2015), and the ecological ramifications of evolving traits (Bassar et al. 2012). In the simplest sense, some studies characterise community assembly to be either neutral, where individuals are interchangeable (Hubbell 2001), or influenced by abiotic conditions and/or biotic interactions (Chesson 2000, HilleRisLambers et al. 2012, Adler et al. 2006, Thompson et al. 2020). We demonstrate here that competition plays a key role in community assembly.

Whether competition among species is important for structuring ecological communities has been a matter of considerable debate (HilleRisLambers et al. 2012, Macarthur and Levins 1967). The majority of studies support the idea that competition for limiting resources in communities drives niche differentiation which in turn facilitates coexistence through intraspecific negative density dependence (Chesson 2000, Adler et al. 2010, HilleRisLambers et al. 2012). Organismal traits are thought to mediate these niche-based competitive interactions (HilleRisLambers et al. 2012, Adler et al. 2013). Yet, as important as competition may seem, detecting it statistically, and therefore understanding its generality across systems, remains a challenge (HilleRisLambers et al. 2012, Barner et al. 2018, Freilich et al. 2018). Neutrality, 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, Rosindell and Phillimore 2011, Chisholm and Pacala 2010), or whether

false positives or false negatives prevail.

Part of the challenge of detecting the impact of competition and/or testing neutrality could be that limited data of only a few types have typically been used to compare model predictions to reality; multiple complementary data axes would provide more inference potential (McGill et al. 2007). To date, competition and neutrality have largely been evaluated using species abundance distributions (SAD) (Hubbell 2001, Chesson 2000, Chisholm and Pacala 2010), though increasingly other data have been separately brought to bear on the topic including: phylogenies (Webb et al. 2002, Jabot and Chave 2009, Nuismer and Harmon 2015), metabarcoding data (Baselga et al. 2015), a combination of genetic data and SADs (Overcast et al. 2019, Vellend 2005) and traits (HilleRisLambers et al. 2012, Adler et al. 2013, Ruffley et al. 2019, Aristide and Morlon 2019), even if choosing the right trait to study presents further difficulties (E-Vojtkó et al. 2020).

The Massive Eco-evolutionary Synthesis Simulations (MESS) model of Overcast et al. (2021) allows testing mechanistic hypotheses jointly 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). MESS is a simulation model that can be fitted to empirical data using machine learning procedures. Competition in the MESS model, in keeping with conventional wisdom (HilleRisLambers et al. 2012, Adler et al. 2013), is driven by evolving traits. However, an individual's fitness 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 approach delivers substantial computational gains, which is useful to run enough simu-

lations for machine learning based inference from data, but is a weak approximation for the mechanistic reality where competition is driven by interactions between individual organisms (Berger et al. 2008). Simulating mean competition may generate patterns that do not reflect real competitive processes, and so the generated data are not appropriate for detecting competition in empirical data.

Here we integrate more realistic competition models into MESS. This opens up the opportunity to interrogate different forms of competition quantitatively, discover which ones lead to distinguishable ecological community structures, and thus which forms of competition can reasonably be inferred from real world data.

2 Material & Methods

2.1 The MESS model

We expand upon the MESS model (Overcast et al. 2021), a dynamic, mechanistic individual based model of community assembly. In MESS, there are a fixed number of individuals in a local community. Each individual i is characterized by its trait value z_i, which is identical for conspecifics. At each time step, one individual dies and is instantaneously replaced by another individual, which comes from either a reproduction event within the local community or immigration from the metacommunity. The metacommunity arises from ecological and evolutionary processes, including speciation (*sensu* Hubbell (2001)). Under the assumption of neutrality, the death probability P_{UNTB} of any given individual i in the local community is computed as

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

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

Non-neutral processes are incorporated in MESS by computing, at each time step, each individual's death probability according to a chosen non-neutral community assembly model (competition or environmental filtering).

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

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

where z_E is the environmental optimum and σ_E determines how quickly fitness decays with the distance to the optimum. The probability of death in the next time step, for any given individual is given by the normalized death rate $P_{\text{filt}}(i) = \frac{q_{\text{filt}}(i)}{\sum_{j=1}^{J} q_{\text{filt}}(j)}$.

In the original MESS model (Overcast et al. 2021), competition is modeled 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

$$q_{\text{MF}}(i) = \exp\left[-\frac{1}{\sigma_E}(z_i - \bar{z})^2\right] \tag{3}$$

where $\bar{z}$ is the local community mean trait and σ_E determines how quickly competitive pressure decays with the distance between trait values.

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 each diverge away from the central mean value leading to an obvious gap around the mean. 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 favored.

In order to resolve these issues, we developed and incorporated a new competition model based on pairwise comparisons between all individuals. In this model, the death rate q_{pair} of any given individual i is based on the mean of all pairwise trait differences between individuals in the local community:

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

The added computational cost of the pairwise model was partially offset by optimizing the underlying data structures, enabling faster simulations (see Appendix A.2).

The pairwise competition model, however, does not allow refinement of the strength of intra-specific competition: individuals of the same species have the exact same trait value and thus the exponential is always equal to 1. This means that the intra-specific competition is always very strong. We therefore developed a third competition model that introduces an 'interaction matrix' parameter β_{ij} to modulate competition strength between all possible pairs of individuals. 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

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

We refer to this as the *beta*-competition model.

By allowing intra- and inter-specific competition to differ according to a parameter, we are in effect modeling differing levels of negative density dependence: $\beta_{\text{intra}} \gg \beta_{\text{inter}}$ corresponds to strong intraspecific density dependence whilst $\beta_{\text{intra}} \ll \beta_{\text{inter}}$ corresponds to no density dependence and leads invariably to mono dominance (data not shown). The three competition models are summarised in Fig. 1.

2.2 Exploration of *in silico* experiments

To explore the behavior 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), with parameter ranges following Overcast et al. (2021) (see Appendix A.3). We explored temporal trends in richness and abundance, genetic diversity, and trait variation using least square polynomial regressions using the polyfit function of Numpy v.19.0 (Oliphant 2006). 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). 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). This enabled us to distinguish between the different community assembly models.

2.3 Machine learning and inference

The MESS inference procedure uses a Random Forest (Breiman 2001) approach for model classification and parameter estimation, using python and the scikit-learn module (v0.20.3,

Pedregosa et al. (2011)). For complete details of implementation, see Overcast et al. (2021). Here we first trained a machine learning classifier in a supervised fashion on the 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 (as detailed below). A confidence percentage is associated to each model. We quantified classifier accuracy using 5-fold cross-validation, and evaluated model missclassification by compiling these results into a confusion matrix. Classifier accuracy was evaluated 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.

2.4 Study of empirical datasets

We used the datasets from Overcast et al. (2021) to compare the original MESS results to those obtained when considering the updated competition models. The data sets are: 1) a spider community from Réunion island with a 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).

Following the machine learning method from Overcast et al. (2021), 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 three clear clusters of points: environmental filtering and mean competition simulations represent one cluster, pairwise competition and β-competition simulations represent a second and purely neutral simulations form the third. Interestingly, the group formed by pairwise competition and β-competition partially overlap with the neutral simulations.

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 Fig. 6).

Results of the inferences on the empirical datasets are summarised in Fig. 5. Considering first the three datasets with SAD and genetic data. For the Reunion spider dataset, competition is detected at around 40% 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. Pairwise competition dominates over β-competition model, and over mean competition, which now receives no support. Remarkably, 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 *beta*-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

In this work, we studied the consequences for inference of simplifying competition by its mean field approximation in community assembly models. Our results show that the mean field ap-

proximation in this case is an oversimplification that can lead to underestimation of the role of competition and overestimation of the role of environmental filtering. 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. Pairwise and β-competition simulations have on average more species than the mean competition. This is expected since 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. Drawing a conclusion on the difference between pairwise and β-competition models is however not so straightforward. 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 one big indistinguishable category, the machine learning being unable to tell the two models apart. This is also consistent with the confusion matrix (Fig. 6), which hints that β_{ij} has no crucial importance. If density dependence played a major role, we would expect this to have been reflected in the ease of model selection; this was not the case, at least for the key data types we considered. Our findings here therefore question the importance of density dependence and its role in community assembly as well as the difficulty of unequivocally detecting it.

The proximity of the pairwise and β-competition simulations to the neutral simulations in our results was also striking. Both clusters clearly overlap in the PCA (Fig. 2), even though this is not consistent with the confusion matrices (Fig. 6 and 4). The random forest algorithm seems to be able to distinguish between both, yet they are indistinguishable for

the human eye in the PCA, as well as in most summary statistics (Fig. 3). Other methods might therefore misinterpret competition as neutrality and competition based simulations may resemble neutral simulations. This is consistent with previous ideas including emergent neutrality (Holt 2006) and niche-neutral models (Chisholm and Pacala 2010). Despite this potential for confusion between mechanisms, our results give hope that the combination of ecological data (abundances / traits) and evolutionary (genetic) data together with machine learning do allow us distinguish neutrality and selection much better than can be achieved with less comprehensive data sets.

The fact that, with inclusion of more nuanced competition models, the inference of environmental filtering totally disappears for these datasets compared to Overcast et al. (2021) might be explained by the improved fit of the new non-neutral models to the non-neutral signal in the data. In the previous MESS inference, all non-neutral signal was only allocated to either filtering or mean competition.

In our inferences, selection is revealed by data most often when traits are available. Indeed, neutrality is without question the best fitting model for the spider and weevil datasets that lack trait data. However, a comparison of the confusion matrices (Fig. 4 and Appendix A.4) show that the presence of trait data is not essential to the detection of filtering. This suggests that the difference we observe in our predictions might be due to a real difference in the community assembly models in action, not an artifact due to the type of data used. Nevertheless, the difference in our inferences based on the type of data used are striking, and this could have implications for the kinds of data we gather to study community assembly. Genetic data alone may not suffice to measure the selective pressure on a group, traits may also be needed(Kraft et al. 2015, McGill et al. 2006), contrary to what has been suggested

in the metabarcoding literature (Chen et al. 2019).

In this work, each simulation used to make inferences on empirical data is either neutral, competition or with environmental filtering. We do not simulate models that combine all these processes. Simulating a simultaneous combination of all the processes could be a fruitful direction 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. The testing of our method would be greatly enhanced by an empirical dataset 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 here. Such a comprehensive data set could, in the future, also be used to investigate the underlying technical reasons why no environmental filtering is inferred from data without availability of traits. For example, we could investigate whether it is the presence of trait values or the absence of species abundances distribution that influences most the percentage of inferred filtering. The β factors used in the simulations may also be refined to allow for differences among each pair of species in future work, so as to reflect species-specific interactions, which may generalise to include positive interactions as well as direct competition.

Our study underlines the importance of the range of empirical data at our disposal to detect the ecological footprint of selection and underlying processes of community assembly. We should therefore avoid jumping too quickly to conclusions about the presence or absence of selection in a community. We show that the pairwise competition model, which cannot be easily distinguished in the data from the β-competition model, represents biological reality much better than the existing mean competition model. Failure to detect pairwise competition in other data sets likely means that competition does not act this particular way, not

that competition or selection are totally absent. We hope that this work will pave the way to improved mechanistic eco-evolutionary models and associated inference procedures for community assembly as well as inspiring new data collection and placing greater emphasis on the synergistic power of genetic, abundance and trait data when analysed jointly.

5 Declarations and acknowledgment

The authors declare no conflict of interest.

This study was enabled through the financial help of the École Normale Supérieure – PSL. Much of the original research was conducted by JL during a placement to Imperial College with JR. We thank both the Imperial College research computing services (High Performance Computing) and the IBENS for the usage of their computation cluster. Through JR, this study is a contribution to Imperial College's Grand Challenges in Ecosystems and the Environment initiative. AJR was supported by US National Science Foundation grants DBI-1927319 and OIA-1849227.

The code used for the simulations is available on GitHub.

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.

Chen, W., Ren, K. et al. 2019. Stochastic processes shape microeukaryotic community assembly in a subtropical river across wet and dry seasons. 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.

E-Vojtkó, A., de Bello, F. et al. 2020. The neglected importance of floral traits in trait-based plant community assembly. Journal of Vegetation Science, 31(4): 529–539, doi:10.1111/jvs.12877.

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.

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.

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. 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. 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. bioRxiv, doi:10.1101/2020.01.30.927236.

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.

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.

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.

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

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.

6 Tables and figures

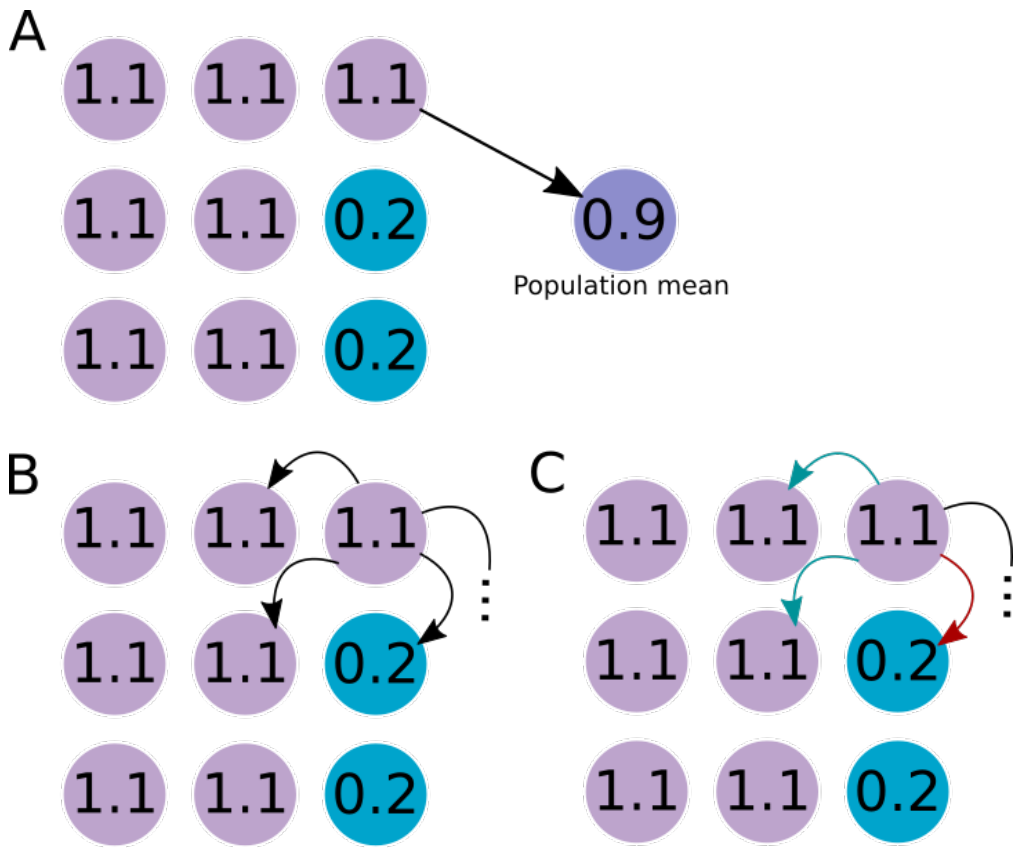

Figure 1: **Depiction of the different forms of competition.** Each circle represents an individual, with the color specifying which species it belongs to and the number its trait value. The effect of competition on fitness (symbolized by arrows) is depicted here for the individual in the top right corner. (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 or not to the same species.

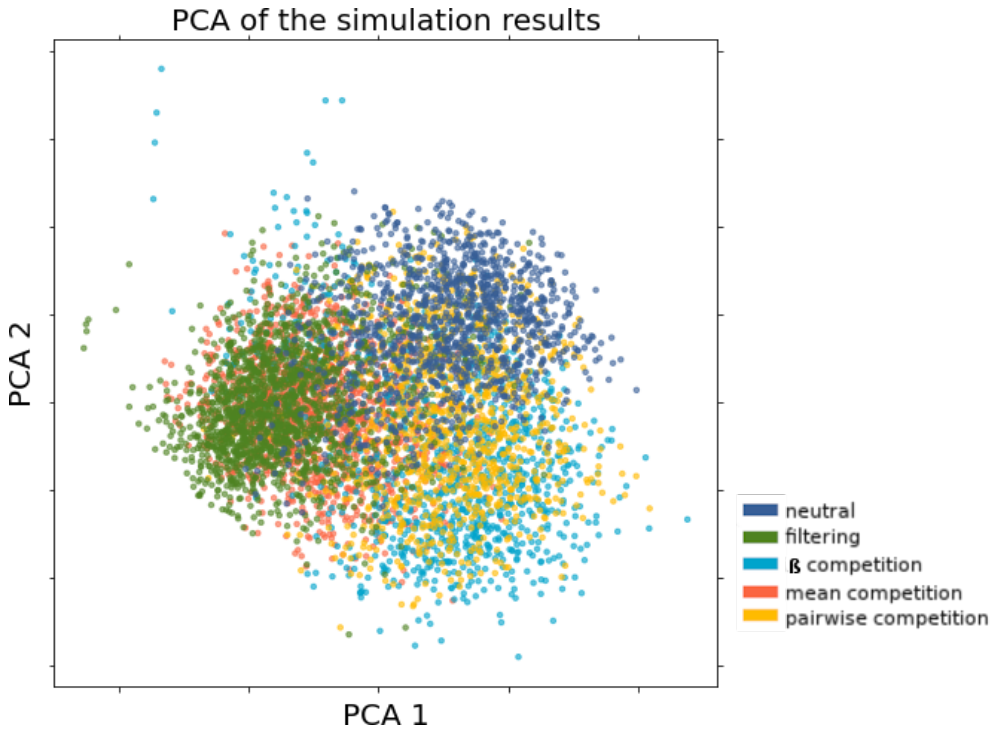

Figure 2: **Plot of the first two principal components of the simulation summary statistics.** Community assembly models depicted are neutral (dark blue), environmental filtering (green), mean competition (orange), pairwise competition (yellow) and β-competition (light blue).

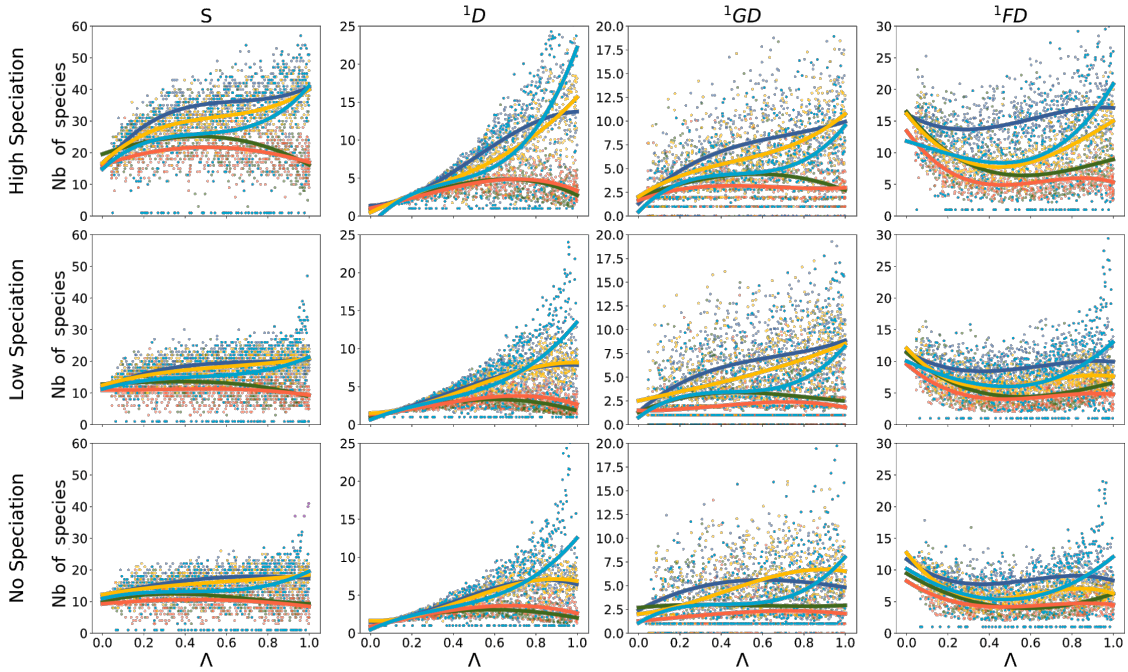

Figure 3: **Community summary statistics through time for the different community assembly models.** This plot depicts the temporal change in some summary statistics for the five focal community assembly models at three different speciation rates : No ($\nu = 0$), Low ($\nu = 0.0005$) and High ($\nu = 0.005$). 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 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.

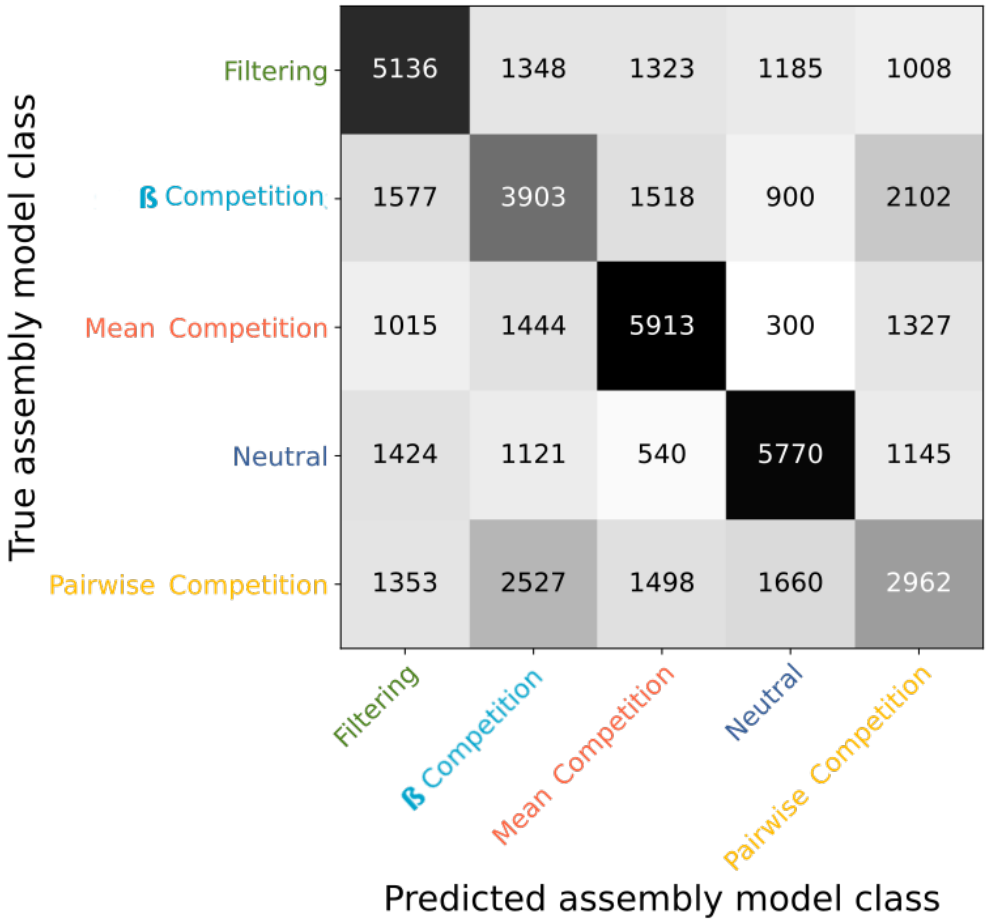

Figure 4: **Machine learning classification confusion matrix** for datasets simulated under the 5 community assembly models and classified using only trait and genetic diversity data (as is the case for the subtropical forest trees and Galapagos snails datasets). Numbers correspond to the number of datasets 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.

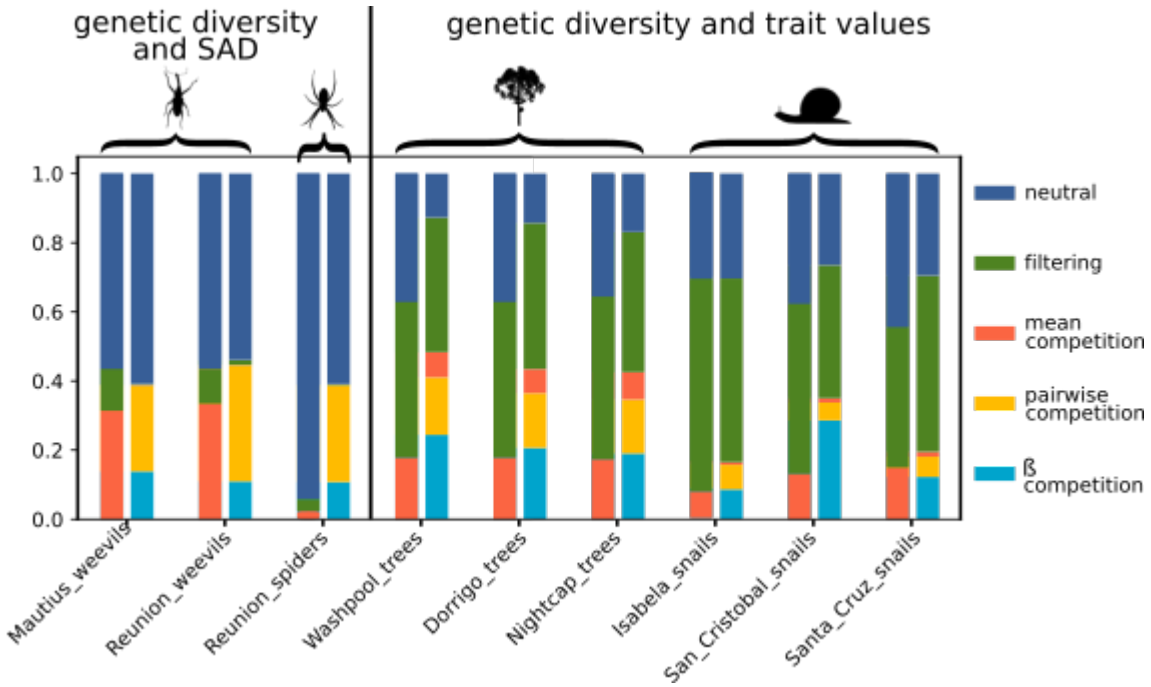

Figure 5: **Machine learning classification probabilities for each empirical community for five focal community assembly models.** For each dataset, the first bar depicts the result of the original MESS publication (Overcast et al. 2020) and the second bar the result with the new competition models. The proportion of color 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).