Title

Predicting and prioritizing species coexistence: learning outcomes via experiments

Short title

Predicting and prioritizing species coexistence

Authors

Benjamin Blonder 1,*

Oscar Godoy 2 *

Affiliations

1: Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA

(ORCID: 0000-0002-5061-2385)

2: Departamento de Biología, Instituto Universitario de Investigación Marina (INMAR), Universidad de Cádiz, Puerto Real, Spain

(ORCID: 0000-0003-4988-6626)

*: Corresponding author, benjamin.blonder@berkeley.edu, 54 Mulford Hall, Berkeley, CA, 94720 USA

Author contributions

BB conceived the idea, processed the datasets, wrote the code, and drafted the manuscript. OG contributed a dataset and contributed substantially to the manuscript.

Data accessibility statement

All data re-used in this study are publicly available. Pre-processed data and statistical analysis code are available at https://github.com/bblonder/coexistence_LOVE and will be archived upon acceptance at Zenodo.

Abstract

Predicting species coexistence can be difficult often because underlying assembly processes are unknown and data are limited. However, accurate predictions are needed for design and forecasting problems in biodiversity conservation, climate change, invasion ecology, restoration ecology, and synthetic ecology. Here we describe an approach (Learning Outcomes Via Experiments; LOVE) where a limited set of experiments are conducted and multiple community outcomes measured (richness, composition, and abundance), from which a model is trained to predict outcomes for arbitrary experiments. Across seven taxonomically datasets, LOVE predicts test outcomes with low error when trained on ~100 randomly-selected experiments. LOVE can then prioritize experiments for tasks like maximizing outcome richness or total abundance, or minimizing abundances of unwanted species. LOVE complements existing mechanism-first approaches to prediction and shows that rapid screening of communities for desirable properties may become possible.

Key words

Coexistence, coexistence outcome, prediction, machine learning, community ecology, ecological forecast

Author summary

Predicting which species will or will not coexist with each other is a central challenge for ecology. Success would allow experimental assembly of communities with desirable properties (e.g. high biodiversity). Mechanistic approaches to this problem run into data and theory limitations, as well as vast combinatorial complexity. Here we instead show to predict and prioritize coexistence without needing to identify or quantify any ecological mechanisms, based on statistical learning from randomly-selected experiments. The approach may help us to discover and then assemble ecological communities with desirable properties.

Introduction

A fundamental challenge of community ecology is to predict which species will coexist with each other, and why. Mechanism questions have been a major priority for theory, (Chesson 2000; Letten *et al.* 2017; Ellner *et al.* 2019). However, the prediction question has received less attention, though it is widely recognized to be equally important (4–6). Here we focus particularly on the challenge of predicting the *outcome* of an experiment, i.e. cases where we have assembled a community, and wish to determine what will happen next (7). Outcomes include properties like whether the resulting community is stable (resistant to local perturbations to abundance) and feasible (all species occur at non-zero abundances), the richness of the community (how many of the experimentally assembled species persist, plus any other species that may establish from a metacommunity), and the abundances of the community (which species are common, and which are rare). Outcomes can be measured at equilibrium, if one exists; or if not, at any time point of interest to us.

Several approaches to predicting outcomes have been developed. Although promising, they are limited for reasons we describe below. In the case that a dynamical model can be assumed to describe the community, some have tried to fit a model to abundance time series data (e.g. parametric models like the generalized Lotka-Volterra model (8)), then using predictions of these models to make inferences about coexistence (e.g. (8–10). This approach is limited by the need for long historical time series datasets, which are not realistic to obtain in most cases, especially for long-lived organisms, as well as the need to assume that there is a specific model that describes the population dynamics of the interacting species.

For short-lived organisms and still with the assumption of an underlying parametric model, several studies have assembled low-richness communities (e.g. singlets and pairs of species across a density gradient of the competing species), in which models are parameterized to make inferences about coexistence, e.g. (11–14). However, such approaches ignore the potential for higher order interactions (HOIs) that explain better species' fitness based on the interacting neighborhood (15,16). Alternatively, other studies have instead fit parameters using high-richness 'dropout' communities (i.e., all species except one) (17,18). Applications in large communities are also difficult because it is hard to estimate large numbers of model parameters (which scale quadratically with richness even in the simplest pairwise models without HOIs (19).

Data availability challenges also apply to approaches that do not require parameterizing a dynamical model. Empirical dynamical models / state space reconstruction (20,21) often require more time points (30-40 minimum) than are available in community studies (22). Conceptually similar are machine learning methods such as random forest and neural networks that predict species abundance across space/time based on statistical relationships for observed species abundances. These approaches still have relatively high observational data requirements (18,23–26). For instance, accurate predictions of spatial variation in abundance for 23 plant species across a 1 km transect, required 37,240 observations in (26). In contrast, a recent approach focusing on predicting abundance outcomes (7) has much lower data requirements (dozens to hundreds of observations) via a matrix pseudo-inversion approach, but cannot account for HOIs and is computationally infeasible in high richness communities.

Finally, observations of species co-occurrence can be used to make inferences about species interactions, from which mechanistic or statistical predictions of coexistence outcomes can be made via partial correlations (checkerboard patterns and related approaches involving species associations, as in joint species distribution models; e.g. (27–29). However, these methods are limited by strong linearity assumptions about the nature of species interactions, or conflation of environmental factors with species interactions (30) as well as high data requirements (29).

To bridge these gaps, we develop an approach (Learning Outcomes Via Experiments; *LOVE*) that sacrifices mechanistic understanding for predictive ability. It philosophically assumes that outcomes are more interesting than dynamics, and that it would be too difficult to identify or parameterize any dynamical model given realistic data limitations. Therefore, LOVE seeks to map directly from a set of community assembly experiment experimental conditions to coexistence outcomes (**Figure 1**). It does not require assumptions (and cannot make inferences) about community assembly mechanisms such as niche differences, indirect interactions or high-order effects (31). It can flexibly handle repeated observations, stochasticity, immigration from a metacommunity, and predict multiple types of coexistence outcomes. It has low data requirements and is computationally fast.

A critical use case for LOVE is in screening potential experiments for desirable properties. For example, in a regional pool of only n=20 species, there are $2^{20} = 1{,}048{,}576$ possible experiments (under the simplest assumption that each species only can be either present or absent at the start of the experiment). Some of these experiments will yield outcomes that are most desirable (e.g. maintain the highest biodiversity through time, maximize biomass production, or

increase biotic resistance against biological invasions). However it is not realistic to carry out every possible experiment to find the desirable outcomes, because of the vastness of the combinatorial space. This challenge is exemplified by the well-known experiments studying biodiversity-ecosystem functioning relationships such as the Cedar Creek (USA) and Jena (Europe) experiments in which far less than 1% of all possible communities are considered (32,33). Instead, LOVE could be trained on a small number of experiments (e.g. 100), and then used to predict a small set of further experiments (e.g. 20) that are statistically likely to have the desired outcome – potentially reducing by orders of magnitude the effort relative to a brute-force enumeration of the combinatorial space.

We ask (1) how well does LOVE predict outcomes at the community level depending on the number of experiments and the experimental design, compared to two alternative methods?; (2) which dataset properties explain LOVE error rates?; and (3) can LOVE prioritize experiments that yield desirable outcomes at realistic numbers of experiments? We answer these questions using experimental outcome data from seven datasets spanning a wide range of taxonomic diversity.

Methods

Concepts

We learn a function $f: X \mapsto Y$ (**Figure 2**) where X describes experimental conditions in a common environment ($X \in \{0,1\}^n$, with 0 indicating absence and 1 presence for each species as experimentally introduced), Y is a coexistence outcome that can be calculated at time *t*, and *n* is the species richness of the regional pool. Three ecological definitions of Y are considered here:

(1) the species richness of the outcome ($Y \in \{0,1,2,\ldots\}$); (2) the composition (identities) of the species that persist in the outcome ($X \in \{0,1\}^n$); and (3) the abundances at which species occur in the outcome ($X \in [0, \infty]^n$).

Modeling approaches

In the focal approach (*LOVE*), we trained random forest models to learn each of the coexistence outcome metrics Y) considered. Random forests allow for nonlinear and multiple interactions among predictors using ensembles of decision trees, and typically avoid overfitting issues (34). For models with univariate Y (richness), models were fitted using the *ranger* R package (version 0.13.1) (35). For models with multivariate Y (abundance, composition), models were trained using the *randomForestSRC* R package (36) (version 2.12.1). Models were fit using num.trees=500, mtry=ceiling(sqrt(n)), with a minimum nodesize of 1 for categorical Y and 5 for continuous Y and an unbounded maximum tree depth. For LOVE, we did not perform further hyperparameter tuning (37) or explore other machine learning algorithms, e.g. support vector machines or boosted regression trees (26), as our focus was conceptual rather than algorithmic (cf. (38) for species distribution model algorithm comparisons).

This framing assumes that species can only be present or absent at the beginning of the experiments, but the approach could be adapted to more complex situations. For example, X values describing initial abundances could be allowed to be continuous-valued ($X \in [0, \infty]^n$), or the dimensionality of X could be increased to study how order of arrival of species to a community influences assembly processes and coexistence outcomes (39), or additional

environmental covariates could be included to describe experiments in different abiotic conditions, e.g., (40).

To assess the performance of LOVE, we compare it to two alternative approaches. For the first alternative (*LOVE+reshuffle*), we repeated the above approach, but first randomly reshuffled predictor rows to break correlations between X and Y pairs in the data. For the second alternative (*naïve*), we also obtained predictions using simple heuristics. For species composition, we assumed that any species present in X (i.e. at the beginning of the experiment) would remain present in Y (i.e. when observing coexistence outcomes), and any species absent in X would remain absent in Y. For species richness, we assumed the number of species present was the same in X and in Y. For abundance, we assumed the abundance of each species in Y was equal to its mean abundance in the training data, elementwise-multiplied by X.

<u>Error definitions</u>

For multivariate categorical Y (composition), balanced accuracy for observed vs. predicted Y was calculated on a casewise basis for each assemblage and then reported as a mean across all assemblages. Error was defined as one minus the mean balanced accuracy.

For univariate continuous Y (richness), a mean absolute error (MAE) was calculated between the observed and predicted Y, treating each assemblage as a replicate. MAE was chosen instead of root mean square error as it is more robust to rare large errors. Error was defined as the MAE scaled by the regional pool species richness for each dataset.

For multivariate continuous Y (abundance), abundances were binned into ten classes, comprising 0, eight values each covering an equal quantile of the non-zero abundance values, and the maximum abundance. Quantiles and maxima were calculated across the entire dataset. Predictions were then made for these classes, after which the classes were transformed back into abundances as either 0 or the bin-mean value, respectively. This binning approach was used to minimize the impact of zero-inflation and skewed abundances. The number of bins used did not have a large impact on the error estimates (results not shown). Error was then defined as the mean MAE across all assemblages, scaled by the 95% quantile dataset abundance.

Data

Seven empirical and simulated datasets were used, which are summarized in **Table 1,** described in detail in **Text S1** and visualized in **Figure S1-S7**. Data were chosen to span a range of complexities and taxa and come from publicly available sources. For datasets generated from a dynamical model, only the predicted outcomes are used in this study; that is, *LOVE* is blind to the underlying dynamics. All datasets were pre-processed to first remove a small number of outcome abundances exceeding 10^7, which arose in a few assemblages within the 'mouse gut' dataset, and then were clipped to the (0.005,0.095) quantiles (across all assemblages within each dataset) to avoid overfitting on outlying values.

Experimental design

The ability of LOVE to predict coexistence outcomes can depend on the information available in training cases. To test this idea, we designed an in silico experiment with three levels: *low*, *high*, and *mixed* richness:

Low richness – There are $\kappa(n,k,q) = \sum_{i=1}^{k} \quad C(n,k)$ possible assemblages with richness less than or equal to *k*, where C(*n*,*k*) indicates the binomial coefficient for the number of ways to choose *k* items out of *n*. We selected a random set of cases {X,Y} cases for training, selecting only among assemblages with k=2, or k=3 pairs and triplets, named as the *low-2* and *low-3* experimental designs).

High richness – There are also $\kappa(n,k)$ possible assemblages with richness greater than or equal to *k*. We selected a random set of cases {X,Y} for training, selecting only among assemblages with *k*=*n*-1, *k*=*n*-2 (single or double dropouts, named as the *high-1* and *high-2* experimental designs).

Mixed richness – we selected a random set of cases {X,Y} from each dataset for training independent of richness (named as the *mixed* experimental design). However because $\kappa(n,k)$ is largest at $k=\lfloor n/2 \rfloor$, intermediate richness assemblages are most frequently sampled.

For each experimental design and sample size, analyses were carried out in a training-testing cycle. Error was evaluated based on comparing predicted outcomes to actual outcomes in the test-set assemblages. Each analysis was replicated 10 times to capture training case sampling variation.

Training-set sample sizes spanned from 10 to 10,000, covering 20 values evenly spaced on logarithmic scale. Analyses were skipped where sample sizes exceeded either the dataset size or the maximum number of samples available for the experimental design.

<u>Question 1 - Predictors of error</u>

We fit a linear mixed model of the form:

$$\text{scaled_error_Y} \sim \text{method} * \log_{10}(\text{num_train}) * \text{experimental_design} * \text{name} + (1|\text{name})$$

where scaled_error_Y indicates the test-set scaled error rate for the model of each coexistence outcome, method is a categorical predictor that contains the three approaches (LOVE, *LOVE+reshuffle*, *naïve*), num_train is a continuous predictor that contains the training sample size, experimental_design is a categorical predictor that contains the three levels of the experimental design (low, high and mixed richness), and name defines the dataset of interest. Dataset also was included as a random intercept to account for the nesting of replicates within other parameter combinations.

We examined the effect of method on Y via a conditional plot, conditioning on a mixed experimental design and a training sample size of 100; the effect of num_train conditioning on the LOVE method and a mixed experimental design; and the effect of experimental design conditioning on the LOVE method and a training sample size of 100. Residuals were verified as

meeting assumptions. A pseudo-R^2 was calculated as one minus the ratio of the deviance and the null deviance.

Question 2 - Dataset properties

We fit a linear mixed model of the form:

$$\text{scaled_error_Y} \sim \text{type} + \text{regional_pool_richness} + \text{num_species_delta} + \text{abundance_skewness} + (1|\text{name})$$

where type indicates whether the dataset includes natural stochasticity, regional_pool_richness indicates the number of species in the regional pool for each dataset, num_species_delta indicates the number of species gained or lost from the experiment to the outcome (based on training data), abundance_skewness indicates the skewness of the abundances of species present in the outcome (based on training data), and name is defined as in Question 1. These predictors were chosen to explore the effect of stochasticity, biodiversity, species interactions, and species dominance respectively.

Because of the results of Question 1, the model was fit using only replicates obtained for a mixed experimental design, a training sample size of 89, and the LOVE method. Conditional effect plots were used to examine the effect of each predictor, holding all nonfocal predictors to median values in the overall dataset. Mixed model goodness-of-fit was assessed as in Question 1.

Question 3 - Desirable outcomes

We explored the utility of LOVE for prioritizing experiments for desirable outcomes. For each application, we trained an abundance *LOVE* model on 89 communities from a 'mixed' experimental design replicating this across 10 samples of experimental communities. Analyses were restricted to the four datasets where the complete set of experimental outcomes were available for validation (annual plant, human gut, mouse gut, SORTIE-ND). We predicted best experiments as described below, then compared the predicted best experiments to the actual best experiments using sensitivity (true positive rate) and specificity (true negative rate) metrics.

In a 'remove unwanted' desirable outcome, we sought to predict the experimental assembly of communities that would be invasion-resistant. These were identified as cases where a focal species was present in the experiment and occurred at its 0% quantile abundance in the outcome (i.e. 0 if a species could ever be absent in at least one outcome, or its minimum abundance if never absent in any outcome). We repeated this analysis for every species in every dataset.

In a 'maximize diversity' desirable outcome, we sought to predict the experimental assembly of communities with high biodiversity, defined as Shannon's H index (41). We predicted abundance for all non-training set experiments and calculated a predicted H value. We then flagged experiments that yielded an outcome H value above the 95% quantile H value actually observed in all assemblages (note that in a real-world use, this quantile threshold's value would be unknown *a priori*, but an investigator might be able to specify a known threshold value for H that they desire to reach).

In a 'maximize abundance' desirable outcome, we sought to predict the experimental assembly of communities with high summed abundance across all species. We repeated the 'maximize diversity' analyses replacing Shannon's H with total abundance, also using a 95% quantile.

Data availability statement

Processed datasets plus code to pre-process datasets, implement LOVE, and carry out statistical analyses, are available at https://github.com/bblonder/coexistence_LOVE.[1] A technical summary of code is in **Text S2**.

Results

Question 1 - Predictors of error

Method – LOVE consistently had lower error than the two alternatives. When conditioning on a training sample size of 100 and a mixed experimental design, richness error (scaled mean absolute error) was between 0 - 0.05 for the LOVE method, compared to near 0.0 - 0.25 for the LOVE + *reshuffle* method and 0 - 0.4 for the *naïve* method (**Figure 3a**). Composition error (1 - balanced accuracy) was between 0 - 0.13 for the LOVE method, compared to 0 - 0.5 for the LOVE + *reshuffle* method and 0 - 0.2 for the *naïve* method (**Figure 3b**). Abundance error (scaled casewise mean absolute error) was between 0 - 0.15 for the LOVE method, compared to 0 - 0.25 for the LOVE + *reshuffle* method and 0 - 0.3 for the naïve method (**Figure 3c**). Some variation in error among datasets was observed. Very low error for LOVE was observed in the human gut, mouse gut, and SORTIE-ND datasets.

[1] These files will be archived upon acceptance.

Experimental design –The mixed experimental design performed best. When conditioning on a training sample size of 100 and the LOVE method, richness error for the mixed design was between 0 - 0.1, compared to 0 - 0.25 for the low-2 (pairs) and low-3 (triplets) designs, and 0 - 0.2 for the high-1 (all species except one) and 0 - 0.45 for the high-2 (all species except two) designs (**Figure 4a**). Composition error for the mixed design was between 0.05 - 0.15 compared to 0.13 - 0.4 for the low-2 and 0.1 - 0.3 for the low-3 designs, or 0.15 - 0.4 for the high-1 and high-2 designs (**Figure 4b**). Abundance error for the mixed design was between 0.05 - 0.1, compared to 0.05 - 0.1 for both the low-2 and low-3 designs, or 0.05 - 0.2 for the high-1 and 0.05 - 0.30 for the high-2 designs (**Figure 4c**). Large variation in error among datasets was observed, primarily for richness and composition.

Training sample size – Error decreased with sample size in all cases, with reasonable performance by 100 training cases. For richness, error was consistently low (**Figure 5a**); for composition, error decreased below 0.1 for all datasets by 100-200 training cases (**Figure 5b**); for abundance, error decreased at similar thresholds (**Figure 5c**), except for the annual plant, fly gut, and Cedar Creek datasets where no substantial decrease was observed.

Overall explanatory power – Overall the fixed effects in the richness error linear mixed model explained 96% of the variance in scaled error and 96% with random effects. The composition error linear mixed model explained 98% of the variance without random effects and 98% with random effects. The abundance error linear mixed model explained 84% of the variance without random effects and 98% with random effects.

There is potential concern that the performance of *LOVE* is driven by correctly predicting 0 for numerous absent species rather than correctly predicting the abundance of present species. However, scatterplots of observed vs. predicted abundance for *LOVE* using the mixed experimental design indicated that this was generally not the case (**Figure S7**). Most datasets showed relationships relatively close to the 1:1 line.

<u>Question 2 - Dataset properties</u>

The hypothesized predictors had moderate explanatory power for LOVE error rates. The richness model explained 23% of the variance in scaled error without random effects and 93% with random effects; for composition, 10% and 99%; for abundance, 16% and 99%. The role of each predictor was similar among response variables (**Figure 6**). Error was slightly higher for outcome data arising from empirical datasets rather than simulated datasets, consistent with the hypothesis of higher stochasticity and measurement error in the real world. Datasets with higher species pool richness had lower error rates, inconsistent with the hypothesis that simpler systems would be more predictable. Training datasets in which more extinctions occurred in outcomes had higher error rates, consistent with the hypothesis that extinctions can have multiple causes that are difficult to statistically disentangle. Training datasets in which outcome abundance was positively skewed also had higher error rates, consistent with the hypothesis that the absence of dominant species (i.e. a left-skewed abundance distribution) leads to more diffuse outcomes that are also difficult to statistically disentangle. However in all cases the substantial random effects variance indicates that there are additional unknown dataset properties not captured by these predictors.

373 Question 3 - desirable outcomes

374 LOVE performance for prioritizing experiments varied across desirable outcomes. For the
375 'remove unwanted' challenge, LOVE had moderate success (**Figure 7a**). In the annual plant and
376 SORTIE-ND datasets, it was possible to predict experiments that would remove an unwanted
377 species for approximately half of the species (with ~50% sensitivity and specificity) while for the
378 other half, it was not possible, with a specificity of zero. In the human gut dataset, it was never
379 possible, while in the mouse gut dataset, it was possible for approximately 20% of the species.
380 For the 'maximize diversity' and 'maximize abundance' challenges, LOVE had more consistent
381 performance (**Figure 7b**). Sensitivity ranged from 80-100% and specificity from 10-60% (10-
382 30% in the 'maximize abundance' challenge) . For both cases, the highest performance occurred
383 in the human gut dataset, and the lowest performance varied according to the challenge
384 considered (SORTIE-ND dataset for 'maximize diversity' and mouse gut dataset for 'maximize
385 abundance' (**Figure 7c**).

386

387 A visualization of the classification statistics across all experiments is shown for a best-case
388 replicate model in **Figure 8**. Summaries of classification statistics across all replicate models are
389 shown in **Figure S8**. While the performance may not appear to be satisfactory, the non-zero
390 specificities are far better than expected by random chance. For example, in the 'maximize
391 diversity' challenge with a 95% quantile threshold, random sampling of experiments would yield
392 a 5% sensitivity and 95% specificity, by definition. Thus, in many cases, the prioritization
393 process leads to a 10-fold increase in the likelihood that a selected experiment has a desired
394 outcome.

395

Discussion

Our *LOVE* implementation shows that predictions of coexistence outcomes in common environments can be obtained at low error with limited experimental data, and that predictions do not require understanding anything about the underlying dynamics or community assembly processes. With approximately 100 experiments selected randomly via mixed experimental design *LOVE* usually outperformed the approach that simplifies the relationship between the initial conditions and the coexistence outcome (*naïve*), and leveraged information in the data that was not present in the approach that broke the correlations between the initial conditions and the coexistence outcomes (*LOVE+reshuffle*).

One contribution of this work is in showing that prediction challenges sometimes can be addressed without mechanistic models or process-based understanding. Such prediction-first approaches are complementary and not exclusive to other mechanism-first or generality-oriented models (42–45). *LOVE* may succeed where some dynamical models approaches fail, because it does not need to predict dynamics, which may be highly complex, nor is it bound to particular process-based constraints on how dynamics could operate, e.g. ones that center competition (46). It is not unfair to criticize *LOVE* as a statistical fitting exercise (47,48), but it is worthwhile to also consider the value for prediction of that compromise.

A second contribution of this work is in showing that predicting outcomes is possible without long time-series datasets so long as randomized experiments can be carried out. Experimental assembly and observation of outcomes can be carried out with multiple experimental communities assembled in parallel, limiting the total time cost of the approach. This could

potentially be useful for forecasting probable outcomes in ecological restoration, invasive species management problems, or designing species mixtures for maximizing biomass production. This parallelized workflow is also possible *LOVE* when the outcomes are not measured at equilibrium but rather at a certain time point. For example, the SORTIE-ND results considered here are from an ongoing disturbance and colonization process that does not reach equilibrium. However, all the outcome data were measured after t=1000 years, so *LOVE* predictions can be interpreted as the community state at that time. In this sense, LOVE applications with non-equilibrium outcomes can be considered analogous to single-shot ecological forecasts (49).

A third contribution of this work is in showing that coexistence experiments would benefit from using mixed experimental designs, building on an emerging consensus (7,25). This approach provides more information per unit of effort, due to the complex combinations that are simultaneously explored. We found that the experimental designs with high level of species richness performed well but only at relatively high sample sizes, indicating that dropout communities alone are not useful - multiple levels of dropouts are required to resolve complex species interactions, e.g. (50). Conversely, the experimental designs with low species richness were similarly most successful in cases where it was possible to train on both pair and triplet communities, e.g. (11,51), even in the 'annual plant', 'human gut', and 'mouse gut' datasets where the underlying dynamical model involves purely pairwise interactions. This can occur for example with intransitive competition (52).

A fourth contribution of this work is showing that prioritization of experiments is possible as a shortcut to obtaining desirable outcomes. In scenarios where it is not time or cost effective to

experimentally explore all possible communities (53), *LOVE* could allow rapid screening of a subset of assemblages which will likely have desirable properties (e.g. production of a certain metabolite (18,25,54) or ecosystem function (55,56), presence of a certain subset of species (57,58), similar to guided screening approaches in drug discovery (59). In this sense, *LOVE* is guiding reachability analyses - the control theory inverse problem of identifying the set of initial experimental conditions which will eventually come to a certain outcome (60). Because of the low computational cost of LOVE, these reachability analyses can be made via enumeration.

Practical considerations

LOVE is likely to be applicable when regional species pools have richness approximately between $10 \leq n \leq 40$. If n is too low, then there may not be a sufficient number of cases (assemblages) to train machine learning models (e.g. the fly gut), or for such an exercise to be worthwhile (because experimental enumeration works better). If n is too high, then there are many possible cases (e.g. $2^{40} \approx 10^{12}$) and it may not be realistic to experimentally assemble enough training cases to adequately learn the interactions between predictors. Our analyses indicated a limited effect of regional pool richness on LOVE error rates, but we were not able to find datasets to validate results at $n>18$ species. Thus studies involving simple microcosms or predator/prey systems will likely not benefit from *LOVE*, nor will studies of highly speciose tropical forests. However, studies involving intermediate richness are candidate areas of application.

Future studies could identify additional mechanisms or properties of dynamical models that make coexistence outcomes more or less predictable, beyond those predictors we explored. For

example, a fully neutral community assembly process would yield random outcomes, and thus high *LOVE* error. Cases with more niche-based dynamics would likely yield lower error. Our identification of the number of species lost being a key predictor of *LOVE* error supports this perspective. It remains unknown how the topology of interaction networks or the nonlinearity of interactions might affect results. Without further datasets to test against, these hypotheses remain preliminary.

Uses of *LOVE* will likely require re-imagining data collection. Training data that do not explore the full combinatorial space are unlikely to yield successful applications. For example, the Cedar Creek plant dataset used here is a multi-decade high-budget effort. It includes 154 assemblages, representing 0.058% of the 2^{18} possible assemblages. Here, *LOVE* performed no better than the *naïve* method. However, the experiments were set up to investigate ecosystem functioning questions, not coexistence questions; re-investing the same budget and effort in coexistence outcomes might be able to provide more extensive coverage of the combinatorial space. Regardless, a key implication is that LOVE application would require substantially different experimental data than the observational data often available. Our study highlights the rarity (but high value) of combinatorial community assembly experiments for studying coexistence.

LOVE has limits. First, it is not guaranteed to be useful in cases where environmental conditions differ from those present during the experiment and observation of outcomes. Extrapolation to novel environmental conditions is challenging for the method. It should be possible to include environmental conditions as additional dimensions of X, but training would likely require replicating experiments across an environmental gradient. This approach may still be low-effort

relative to the alternatives. We did not have data available to assess the viability of this possibility. Second, *LOVE* may be outperformed in cases where a parameterized dynamical model is available. Rather, *LOVE* will be most useful in cases where it is not realistic to parameterize a model. Third, zero-inflation can cause challenges. In some datasets, many combinations of species will lead to the majority of species having zero abundance due to competitive exclusion, leading to strong bias in abundance predictions. Two-stage models or class weighting approaches can be used to address this in the single-species context but are not feasible in multi-species contexts due to correlations in abundance (especially zeros) across species. Here, we worked around the problem by converting the abundances to binned abundance classes, which reduces the impact of zeros. Multivariate hurdle approaches have been proposed (24) but this remains an active area of research. One could diminish the impact of true zeros by augmenting training data for a model by providing access to naïve predictions (e.g. assuming species coexist if and only if they are experimentally assembled) but this approach would fail in open communities where immigration is possible.

<u>Applications</u>

There are numerous applied ecological problems which either involve intentionally assembling a community, or involve making a perturbation to a community. In all these cases the question is: which community should be assembled, or which perturbation should be made, to optimally achieve the desired outcome? In all these cases, *LOVE* could eventually offer a toolkit for providing answers. The workflow would be to: (1) carry out a small set of experiments and train a *LOVE* model; (2) use the model to predict which experiments are most likely to yield the desired outcome; (3) invest real effort in these prioritized experiments (**Figure 1**).

A few example use cases, not explicitly explored by this theoretical work, include landscape restoration (which set of species can lead to a stable restoration outcome or high ecosystem functioning) (61,62), control or screening of invasive species (which species will be able to coexist with an extant community) (63,64), disease ecology (when can diseases persist or spread) (65), agriculture (which crops can be grown together) (66,67), and climate change (how does a shifting environment affect community composition (68,69)). Applications also exist in synthetic microbial community ecology (which set of species can be assembled to yield given useful metabolic outputs) (18,57,70) and in human gut microbiomes (which species can be introduced or removed to yield a community that supports health) (71,72).

Ethical considerations

Because many of the candidate applications of *LOVE* are in applied ecology problems relevant to land management or human health, it is necessary to also consider some of the consequent ethical issues. A key issue is that many of the potential applications involve making out-of-sample predictions that involve novel species (e.g. will a species invade). While simulations and initial experiments can build statistical confidence that out-of-sample experiments will yield the desired outcome, there is no way to guarantee it. Rare outcomes may actually occur that are undesirable and dangerous, and that may not be recoverable.Additionally, the algorithm may suggest experiments that are unsafe, i.e. transient dynamics that pass through dangerous or unwanted community states (73), such as undesired invasions, high pathogen loads, or loss of culturally important species assemblages. A healthy respect for ecological complexity (47,74), ‘unknown unknowns’ (75), and unforeseen consequences (76) is prudent and necessary.

Potential applications must also consider their political framing, i.e. whether they take a technocratic and algorithm-first approach to mediating relationships between people and nature. Such framings could be very harmful, because they would implicitly de-legitimize the value of traditional and expert knowledge, and could further support the long legacy of colonialism and white supremacy in ecology, e.g. (77,78). We suggest that instead it would be best to apply future and more powerful versions of *LOVE* to real-world problems only when the nature of the problem to be solved has been co-identified in consultations with relevant communities, and with full consideration of the unintended consequences of algorithm deployment. Once algorithms are discovered, they are hard to leave alone – but un-considered use is likely unethical.

Conclusion

LOVE provides a promising approach for predicting coexistence outcomes from experimental community assembly datasets, and complements a growing body of work on machine learning approaches in community ecology (7,17–19,24–26). Applications to experiment prioritization seem possible and could be useful if explored ethically. LOVE may be a useful step towards more mechanism-free approaches to community ecology problems.

Acknowledgments

Daniel Maynard's work was the inspiration for this study. Pierre Gaüzère, Lars Iversen, and Courtenay Ray guided initial discussions. Carl Boettiger, Jashvina Devadoss, Marcus Lapeyrolerie, Ilaíne Matos, Zachary Sunberg, Claire Tomlin, several anonymous reviewers, and others provided feedback on manuscript drafts. Andrew Letten, Daniel Maynard, William Petry, and William Sharpless provided feedback and provided guidance on datasets. Lora Murphy and Charles Canham provided input on the SORTIE-ND model. Peter Adler, Jonathan Friedman, Alison Gould, Nathan Kraft, Margaret Mayfield, Sara Mitri, Peter Reich, Alejandra Rodriguez Verdugo, Alvaro Sanchez, Jürg Spaak, and Ophelia Venturelli provided guidance on datasets. The Cedar Creek experiment was supported by NSF via grant DEB-1831944. OG acknowledges financial support provided by the Ministerio de Ciencia, Innovación y Universidades (RYC-2017-23666).

References

1. Chesson P. Mechanisms of maintenance of species diversity. Annual review of Ecology and Systematics. 2000;31(1):343–66.
2. Letten AD, Ke P, Fukami T. Linking modern coexistence theory and contemporary niche theory. Ecological Monographs. 2017;87(2):161–77.
3. Ellner SP, Snyder RE, Adler PB, Hooker G. An expanded modern coexistence theory for empirical applications. Ecology letters. 2019;22(1):3–18.
4. Keddy PA. Assembly and response rules: two goals for predictive community ecology. Journal of Vegetation Science. 1992;3(2):157–64.
5. Laughlin DC, Laughlin DE. Advances in modeling trait-based plant community assembly. Trends in plant science. 2013;18(10):584–93.
6. Houlahan JE, McKinney ST, Anderson TM, McGill BJ. The priority of prediction in ecological understanding. Oikos. 2017;126(1):1–7.
7. Maynard DS, Miller ZR, Allesina S. Predicting coexistence in experimental ecological communities. Nature Ecology & Evolution. 2020;4(1):91–100.
8. Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biology. 2016 Jun 3;17(1):121.
9. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLoS Comput Biol. 2013;9(12):e1003388.
10. Ushio M, Hsieh C hao, Masuda R, Deyle ER, Ye H, Chang CW, et al. Fluctuating interaction network and time-varying stability of a natural fish community. Nature. 2018;554(7692):360–3.
11. Vandermeer JH. The competitive structure of communities: an experimental approach with protozoa. Ecology. 1969;50(3):362–71.
12. Levine JM, HilleRisLambers J. The importance of niches for the maintenance of species diversity. Nature. 2009;461(7261):254–7.
13. Kraft NJ, Godoy O, Levine JM. Plant functional traits and the multidimensional nature of species coexistence. Proceedings of the National Academy of Sciences. 2015;112(3):797–802.
14. Balsa-Canto E, Alonso-del-Real J, Querol A. Mixed growth curve data do not suffice to fully characterize the dynamics of mixed cultures. Proceedings of the National Academy of Sciences. 2019;
15. Mayfield MM, Stouffer DB. Higher-order interactions capture unexplained complexity in diverse communities. Nature ecology & evolution. 2017;1(3):1–7.
16. Pistón N, de Bello F, Dias AT, Götzenberger L, Rosado BH, de Mattos EA, et al. Multidimensional ecological analyses demonstrate how interactions between functional traits shape fitness and life history strategies. Journal of Ecology. 2019;107(5):2317–28.
17. Ansari AF, Reddy Y, Raut J, Dixit NM. An efficient and scalable top-down method for predicting structures of microbial communities. Nature Computational Science. 2021;1(9):619–28.
18. Clark RL, Connors BM, Stevenson DM, Hromada SE, Hamilton JJ, Amador-Noguez D, et al. Design of synthetic human gut microbiome assembly and butyrate production. Nature communications. 2021;12(1):1–16.

19. Blonder B, Gauzere P, Ray C, Ke P, Petry W, Germain W, et al. Trait-based prediction and control: a roadmap for community ecology. 2022 in review;
20. Perretti CT, Munch SB, Sugihara G. Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data. Proceedings of the National Academy of Sciences. 2013;110(13):5253–7.
21. Ye H, Beamish RJ, Glaser SM, Grant SCH, Hsieh C hao, Richards LJ, et al. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proceedings of the National Academy of Sciences. 2015 Mar 31;112(13):E1569–76.
22. Chang CW, Ushio M, Hsieh C hao. Empirical dynamic modeling for beginners. Ecological research. 2017;32(6):785–96.
23. Rammer W, Seidl R. Harnessing deep learning in ecology: An example predicting bark beetle outbreaks. Frontiers in plant science. 2019;10:1327.
24. Kong S, Bai J, Lee JH, Chen D, Allyn A, Stuart M, et al. Deep hurdle networks for zero-inflated multi-target regression: Application to multiple species abundance estimation. arXiv preprint arXiv:201016040. 2020;
25. Baranwal M, Clark RL, Thompson J, Sun Z, Hero AO, Venturelli O. Deep Learning Enables Design of Multifunctional Synthetic Human Gut Microbiome Dynamics. bioRxiv. 2021 Jan 1;2021.09.27.461983.
26. Civantos-Gómez I, García-Algarra J, García-Callejas D, Galeano J, Godoy O, Bartomeus I. Fine scale prediction of ecological community composition using a two-step sequential Machine Learning ensemble. PLoS Comput Biol. 2021;
27. Connor EF, Collins MD, Simberloff D. The checkered history of checkerboard distributions. Ecology. 2013;94(11):2403–14.
28. Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM, et al. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution. 2014;5(5):397–406.
29. Morueta Holme N, Blonder B, Sandel B, McGill BJ, Peet RK, Ott JE, et al. A network approach for inferring species associations from co-occurrence data. Ecography. 2016;39(12):1139–50.
30. Blanchet FG, Cazelles K, Gravel D. Co-occurrence is not evidence of ecological interactions. Ecology Letters. 2020 Jul 1;23(7):1050–63.
31. Levine JM, Bascompte J, Adler PB, Allesina S. Beyond pairwise mechanisms of species coexistence in complex communities. Nature. 2017 print;546(7656):56–64.
32. Tilman D, Reich PB, Isbell F. Biodiversity impacts ecosystem productivity as much as resources, disturbance, or herbivory. Proceedings of the National Academy of Sciences. 2012;109(26):10394–7.
33. Weisser WW, Roscher C, Meyer ST, Ebeling A, Luo G, Allan E, et al. Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions. Basic and applied ecology. 2017;23:1–73.
34. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
35. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:150804409. 2015;
36. Ishwaran H, Kogalur UB. Fast unified random forests for survival, regression, and classification (RF-SRC). R package version. 2019;2(1).
37. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

2019;9(3):e1301.
38. Elith J, Leathwick JR. Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics. 2009;40.
39. Fukami T. Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects. Annual Review of Ecology, Evolution, and Systematics. 2015 Dec 4;46(1):1–23.
40. Pennekamp F, Pontarp M, Tabi A, Altermatt F, Alther R, Choffat Y, et al. Biodiversity increases and decreases ecosystem stability. Nature. 2018 Nov 1;563(7729):109–12.
41. Pielou EC. Shannon's formula as a measure of specific diversity: its use and misuse. The American Naturalist. 1966;100(914):463–5.
42. Levins R. The strategy of model building in population biology. American Scientist. 1966;54(4):421–31.
43. May RM. Stability and complexity in model ecosystems. Princeton University Press; 1973.
44. Evans MR, Grimm V, Johst K, Knuuttila T, De Langhe R, Lessells CM, et al. Do simple models lead to generality in ecology? Trends in ecology & evolution. 2013;28(10):578–83.
45. Tredennick AT, Hooker G, Ellner SP, Adler PB. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology. 2021 Jun 1;102(6):e03336.
46. Simha A, Hoz CPD la, Carley L. Moving beyond the "diversity paradox": the limitations of competition-based frameworks in understanding species diversity. The American Naturalist [Internet]. 2022 Mar 24 [cited 2022 May 15]; Available from: https://doi.org/10.1086/720002
47. Lawton JH. Are there general laws in ecology? Oikos. 1999;84(2):177–92.
48. Vellend M, Agrawal A. Conceptual synthesis in community ecology. The Quarterly Review of Biology. 2010;85(2):183–206.
49. Dietze MC, Fox A, Beck-Johnson LM, Betancourt JL, Hooten MB, Jarnevich CS, et al. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. Proceedings of the National Academy of Sciences. 2018;115(7):1424–32.
50. Finkel OM, Salas-González I, Castrillo G, Spaepen S, Law TF, Teixeira PJPL, et al. The effects of soil phosphorus content on plant microbiota are driven by the plant phosphate starvation response. PLoS Biology. 2019;17(11):e3000534.
51. Friedman J, Higgins LM, Gore J. Community structure follows simple assembly rules in microbial microcosms. Nature Ecology & Evolution. 2017;1:0109.
52. Gallien L, Zimmermann NE, Levine JM, Adler PB. The effects of intransitive competition on coexistence. Ecology Letters. 2017;20(7):791–800.
53. Kehe J, Kulesa A, Ortiz A, Ackerman CM, Thakku SG, Sellers D, et al. Massively parallel screening of synthetic microbial communities. Proceedings of the National Academy of Sciences. 2019;116(26):12804–9.
54. Pacheco AR, Segrè D. An evolutionary algorithm for designing microbial communities via environmental modification. Journal of the Royal Society Interface. 2021;18(179):20210348.
55. Laughlin DC. Applying trait-based models to achieve functional targets for theory-driven ecological restoration. Ecology Letters. 2014;17(7):771–84.
56. Laughlin DC, Chalmandrier L, Joshi C, Renton M, Dwyer JM, Funk JL. Generating species assemblages for restoration and experimentation: A new method that can simultaneously converge on average trait values and maximize functional diversity. Methods in Ecology and Evolution. 2018;9(7):1764–71.
57. Lindemann SR, Bernstein HC, Song HS, Fredrickson JK, Fields MW, Shou W, et al.

Engineering microbial consortia for controllable outputs. The ISME Journal. 2016 Sep 1;10(9):2077–84.
58. Angulo MT, Moog CH, Liu YY. A theoretical framework for controlling complex microbial communities. Nature Communications. 2019;10(1):1–12.
59. Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, et al. Advances in de novo drug design: from conventional to machine learning methods. International journal of molecular sciences. 2021;22(4):1676.
60. Bansal S, Tomlin C. DeepReach: A Deep Learning Approach to High-Dimensional Reachability. arXiv preprint arXiv:201102082. 2020;
61. Palmer MA, Ambrose RF, Poff NL. Ecological Theory and Community Restoration Ecology. Restoration Ecology. 1997;5(4):291–300.
62. Wainwright CE, Staples TL, Charles LS, Flanagan TC, Lai HR, Loy X, et al. Links between community ecology theory and ecological restoration are on the rise. Journal of Applied Ecology. 2018;55(2):570–81.
63. Shea K, Chesson P. Community ecology theory as a framework for biological invasions. Trends in Ecology & Evolution. 2002;17(4):170–6.
64. Gallien L, Carboni M. The community ecology of invasive species: where are we and what's next? Ecography. 2017;40(2):335–52.
65. Johnson PT, De Roode JC, Fenton A. Why infectious disease research needs community ecology. Science. 2015;349(6252):1259504.
66. Vandermeer J. The ecological basis of alternative agriculture. Annual Review of Ecology and Systematics. 1995;26(1):201–24.
67. Malézieux E. Designing cropping systems from nature. Agronomy for sustainable development. 2012;32(1):15–29.
68. HilleRisLambers J, Adler PB, Harpole WS, Levine JM, Mayfield MM. Rethinking community assembly through the lens of coexistence theory. Annual review of ecology, evolution, and systematics. 2012;43:227–48.
69. Alexander JM, Diez JM, Levine JM. Novel competitors shape species/'responses to climate change. Nature. 2015;525(7570):515–8.
70. Nalley JO, Stockenreiter M, Litchman E. Community ecology of algal biofuels: complementarity and trait-based approaches. Industrial biotechnology. 2014;10(3):191–201.
71. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial ecology: building predictive understanding of community function and dynamics. The ISME Journal. 2016 Nov 1;10(11):2557–68.
72. Sonnenburg JL, Sonnenburg ED. Vulnerability of the industrialized microbiota. Science. 2019;366(6464):eaaw9255.
73. Aswani A, Gonzalez H, Sastry SS, Tomlin C. Provably safe and robust learning-based model predictive control. Automatica. 2013;49(5):1216–26.
74. Simberloff D. Community Ecology: Is It Time to Move On? (An American Society of Naturalists Presidential Address). The American Naturalist. 2004;163(6):787–99.
75. Rumsfeld D, Myers R. DoD News Briefing - Secretary Rumsfeld and Gen. Myers [Internet]. 2002. Available from: https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636#selection-401.0-401.53
76. Crichton M. Jurassic Park. Vol. 37077. Random House; 1991.
77. Chapman MS, Oestreich WK, Frawley TH, Boettiger C, Diver S, Santos BS, et al.

Promoting equity in the use of algorithms for high-seas conservation. One Earth. 2021;4(6):790–4.
78. Wyborn C, Evans MC. Conservation needs to break free from global priority mapping. Nature Ecology & Evolution. 2021;1–3.
79. Godoy O, Kraft NJ, Levine JM. Phylogenetic relatedness and the determinants of competitive outcomes. Ecology letters. 2014;17(7):836–44.
80. Godoy O, Stouffer DB, Kraft NJB, Levine JM. Intransitivity is infrequent and fails to promote annual plant coexistence without pairwise niche differences. Ecology. 2017;98(5):1193–200.
81. Tilman D, Reich PB, Knops J, Wedin D, Mielke T, Lehman C. Diversity and productivity in a long-term grassland experiment. Science. 2001;294(5543):843–5.
82. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, et al. Microbiome interactions shape host fitness. Proc Natl Acad Sci USA. 2018 Dec 18;115(51):E11951.
83. Venturelli OS, Carr AV, Fisher G, Hsu RH, Lau R, Bowen BP, et al. Deciphering microbial interactions in synthetic human gut microbiome communities. Molecular systems biology. 2018;14(6):e8157.
84. Buffie CG, Jarchum I, Equinda M, Lipuma L, Gobourne A, Viale A, et al. Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis. Infection and immunity. 2012;80(1):62–73.
85. Pacala SW, Canham CD, Silander Jr JA. Forest models defined by field measurements: I. The design of a northeastern forest simulator. Canadian Journal of Forest Research. 1993;23(10):1980–8.
86. Pacala SW, Canham CD, Saponara J, Silander JA, Kobe RK, Ribbens E. Forest models defined by field measurements: estimation, error analysis and dynamics. Ecological Monographs. 1996;66(1):1–43.
87. Chesson PL. Geometry, heterogeneity and competition in variable environments. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1990;330(1257):165–73.
88. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006;59(10):1087–91.

Tables

Table 1. Summary of datasets used in this study. The number of states indicates the total number of experiments available in the dataset. The number of excluded cases indicates the number of assemblages excluded from training and testing due to outlier abundance values. More detail on data provenance and pre-processing is provided in **Text S2**.

Dataset name	*Taxa*	*Provenance*	*Citation*	*Number of species in regional pool (n)*	*Number of replicate experiments per assemblage*	*Number of states*	*Number of excluded cases*
Annual plant	California annual plants	Simulations from nonlinear competition / seedbank model	(79,80)	18	1	262144	23593
Cedar Creek	North American prairie plants	Experimental sowing in natural environments	(32,81)	18	1	154	14
Fly gut	Bacteria in fly gut	Experimental inoculations of germ-free flies	(82)	5	48	1536	36
Human gut	Bacteria in human gut	Simulations from Lotka-Volterra competition model	(83)	12	1	4096	246
Mouse gut	Bacteria in mouse gut	Simulations from Lotka-Volterra competition model	(9,84)	11	1	2048	256
Soil bacteria	Bacteria in soil	Experimental assembly in microcosms	(51)	8	2-14	570	23
SORTIE-ND	Eastern North American hardwood trees	Simulations of forest dynamics model	(85,86)	9	3	1536	69

Figure Legends

Figure 1. Flowchart summary of workflow for predicting coexistence outcomes and prioritizing experiments. Yellow steps indicate those that require human decision-making; blue steps indicate those that require experimental work with real organisms; purple steps those that require statistical analyses only.

Figure 2. Visualization of experimental conditions and abundance outcomes for the 'human gut' dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red. Similar plots for the other six datasets are shown in **Figure S1-S6**.

Figure 3. Test-set error rate of predictions across methods. Conditional plots are based on linear mixed models assuming a mixed experimental design and 100 training cases. Error rates are shown for **(a)** richness, **(b)** composition, and **(c)** abundance.

Figure 4. Test-set error rate of predictions across experimental designs. Conditional plots are based on linear mixed models assuming the LOVE method and 100 training cases. Error rates are shown for **(a)** richness, **(b)** composition, and **(c)** abundance.

Figure 5. Test-set error rate of predictions across training-set sample size. Conditional plots are based on linear mixed models assuming the LOVE method and mixed experimental design. Error rates are shown for **(a)** richness, **(b)** composition, and **(c)** abundance.

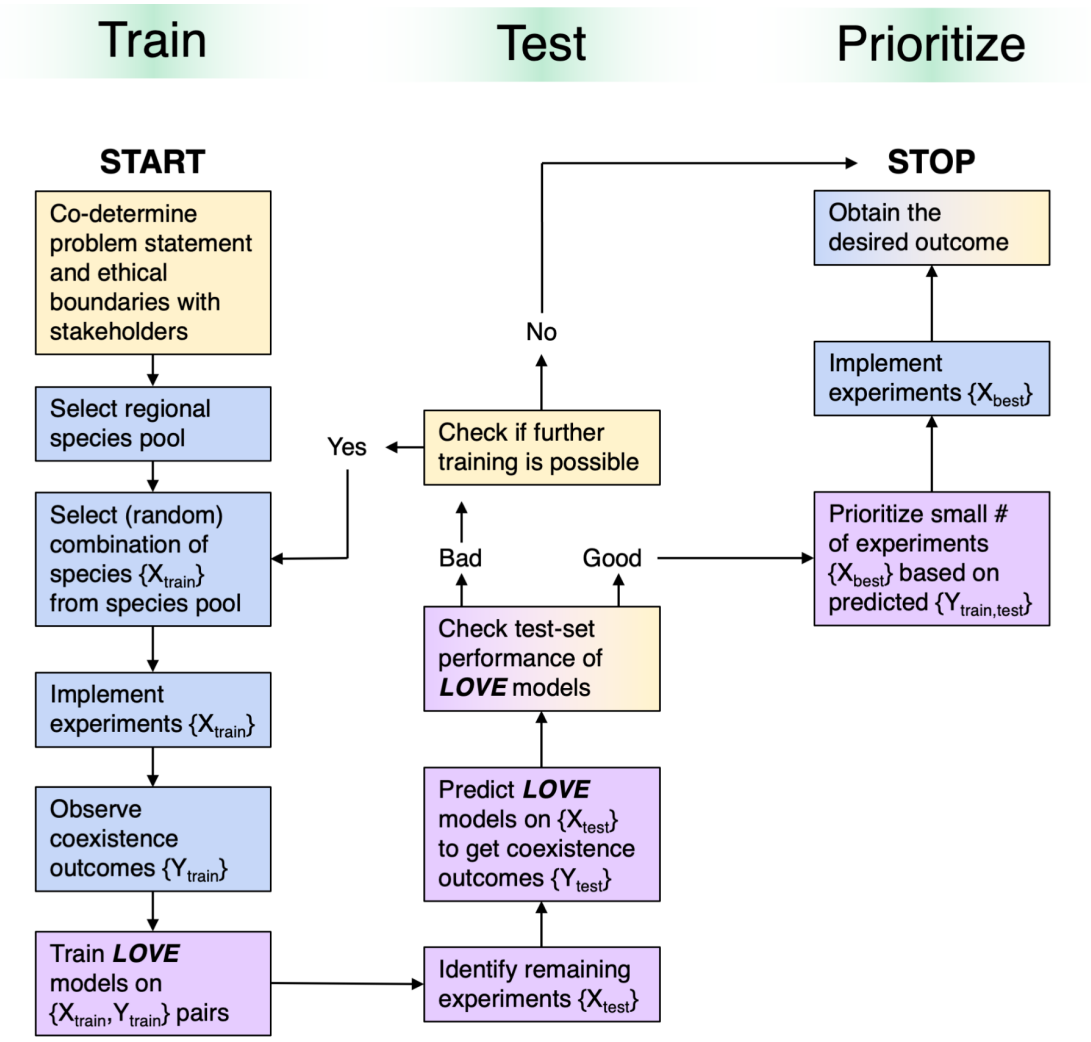
Figure 6. Test-set error rate of predictions across various dataset properties. Conditional plots are based on linear mixed models assuming the LOVE method, mixed experimental design, and 89 training experiments. Error rates are shown for **(a)** richness, **(b)** composition, and **(c)** abundance.

Figure 7. Performance of LOVE for prioritization of experiments for three tasks, **(a)** removing a species that is initially present; **(b)** obtaining a Shannon's H value greater than that for 95% of all possible experiments; and **(c)** obtaining a total abundance greater than that for 95% of all possible experiments. Boxplots indicate distributions of sensitivity and specificity of LOVE over 10 replicate runs, each trained on 89 experiments in a mixed experimental design for each dataset. Panels cover the four datasets for which complete outcome data were available. Sensitivity and specificity are calculated based on a confusion matrix for predicted vs. actual best experiments.

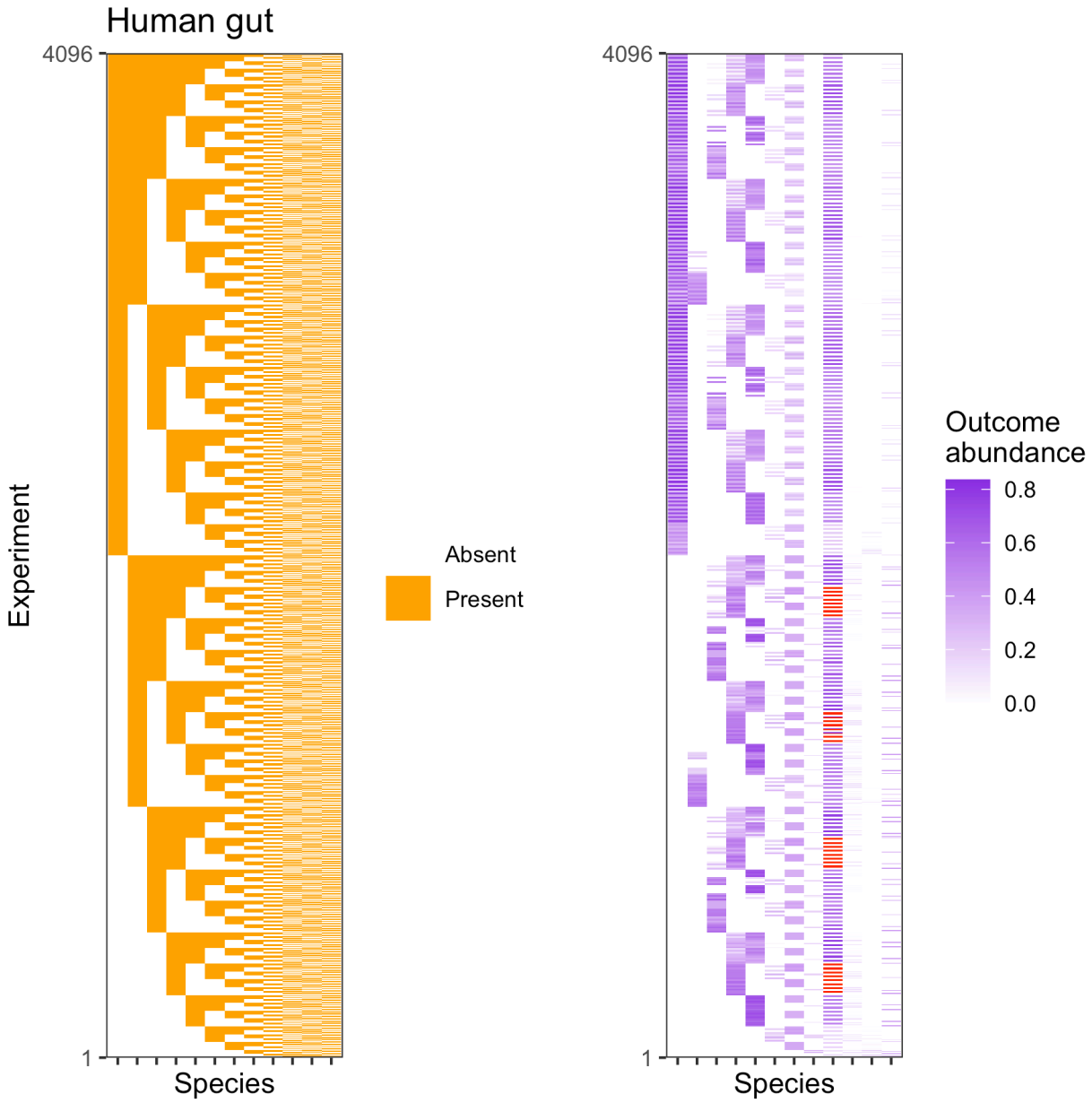
Figure 8. Best-case prioritization of experiments for three tasks, **(a)** removing a species that is initially present; **(b)** obtaining a Shannon's H value greater than that for 95% of all possible experiments; and **(c)** obtaining a total abundance greater than that for 95% of all possible experiments. Panels cover the four datasets for which complete outcome data were available; each panel indicates 1 among 10 LOVE replicates with the highest fraction of true positives. Gray indicates experimental absences, respectively; experimental presences are colored based on their classification.

Figures

Figure 1.

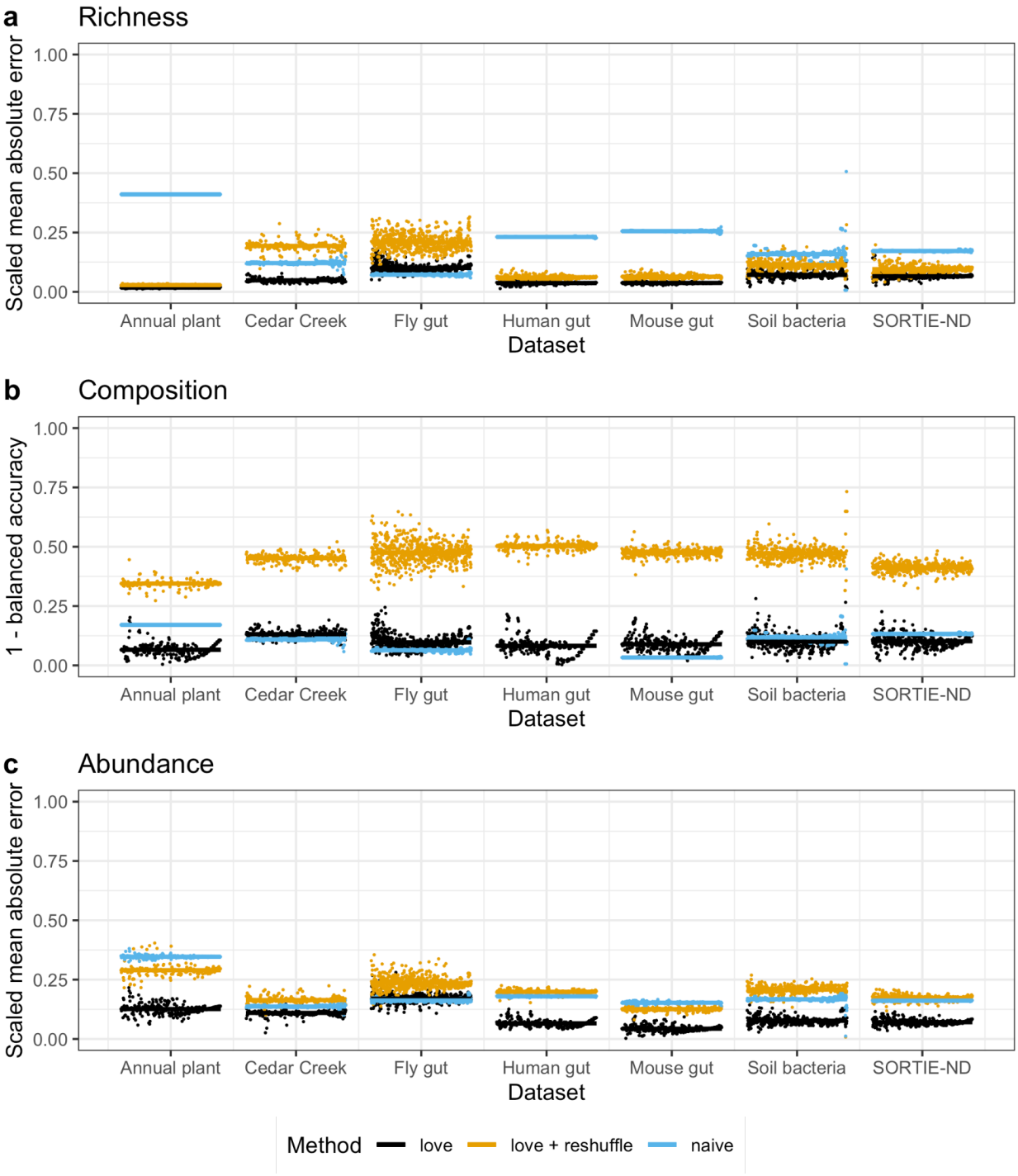


837 **Figure 2**.



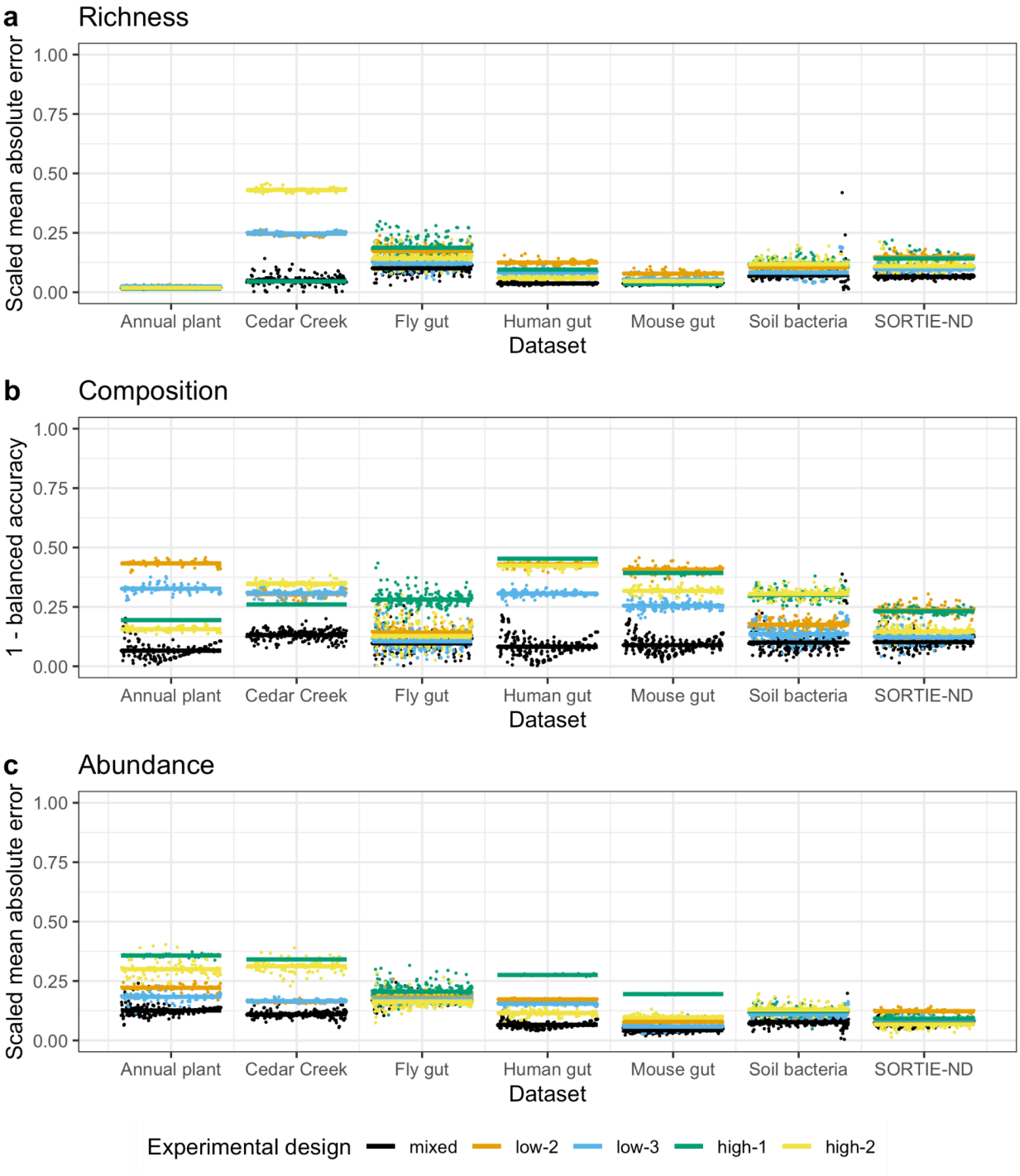
838

839 **Figure 3**.



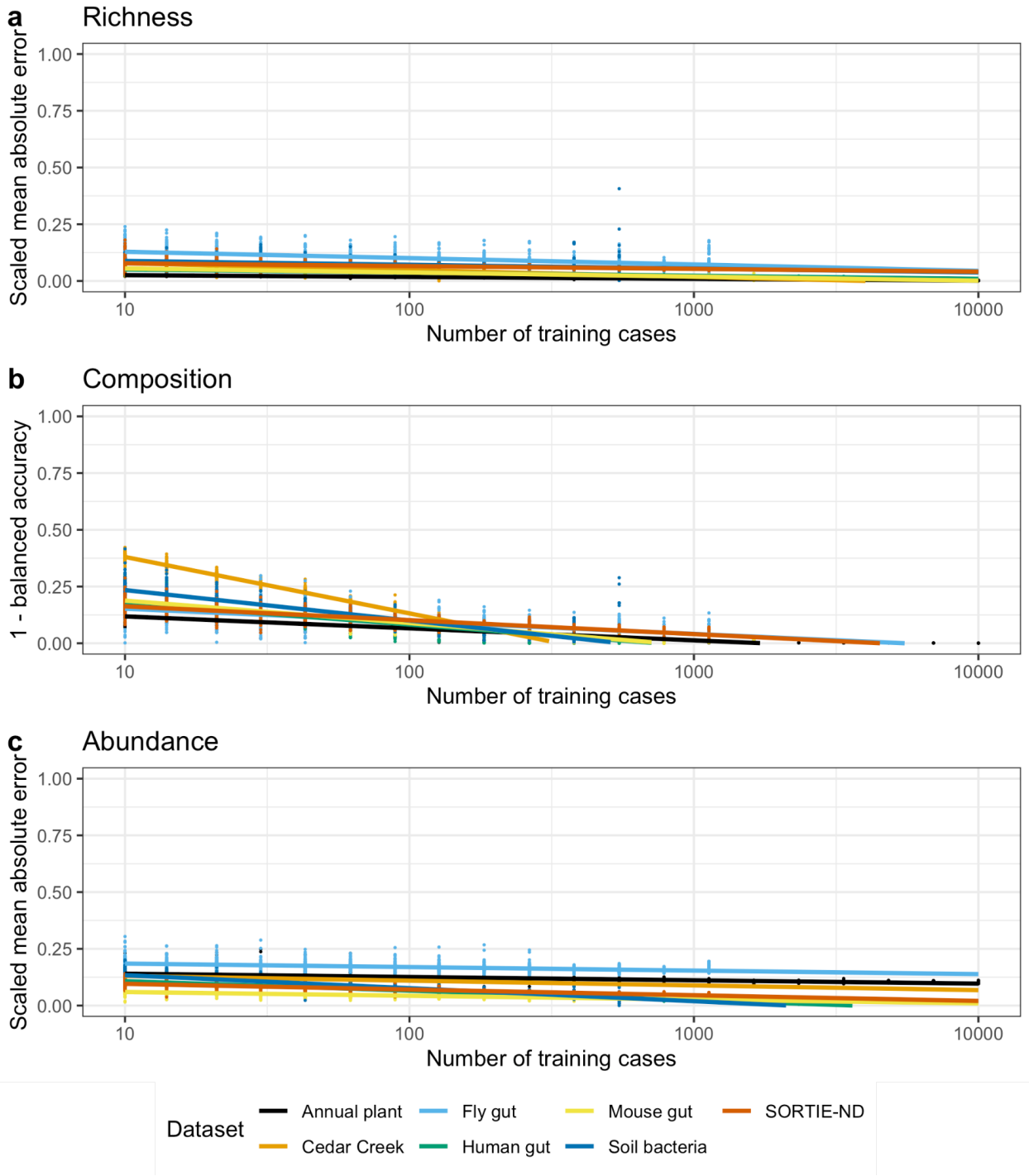
840

841 **Figure 4.**



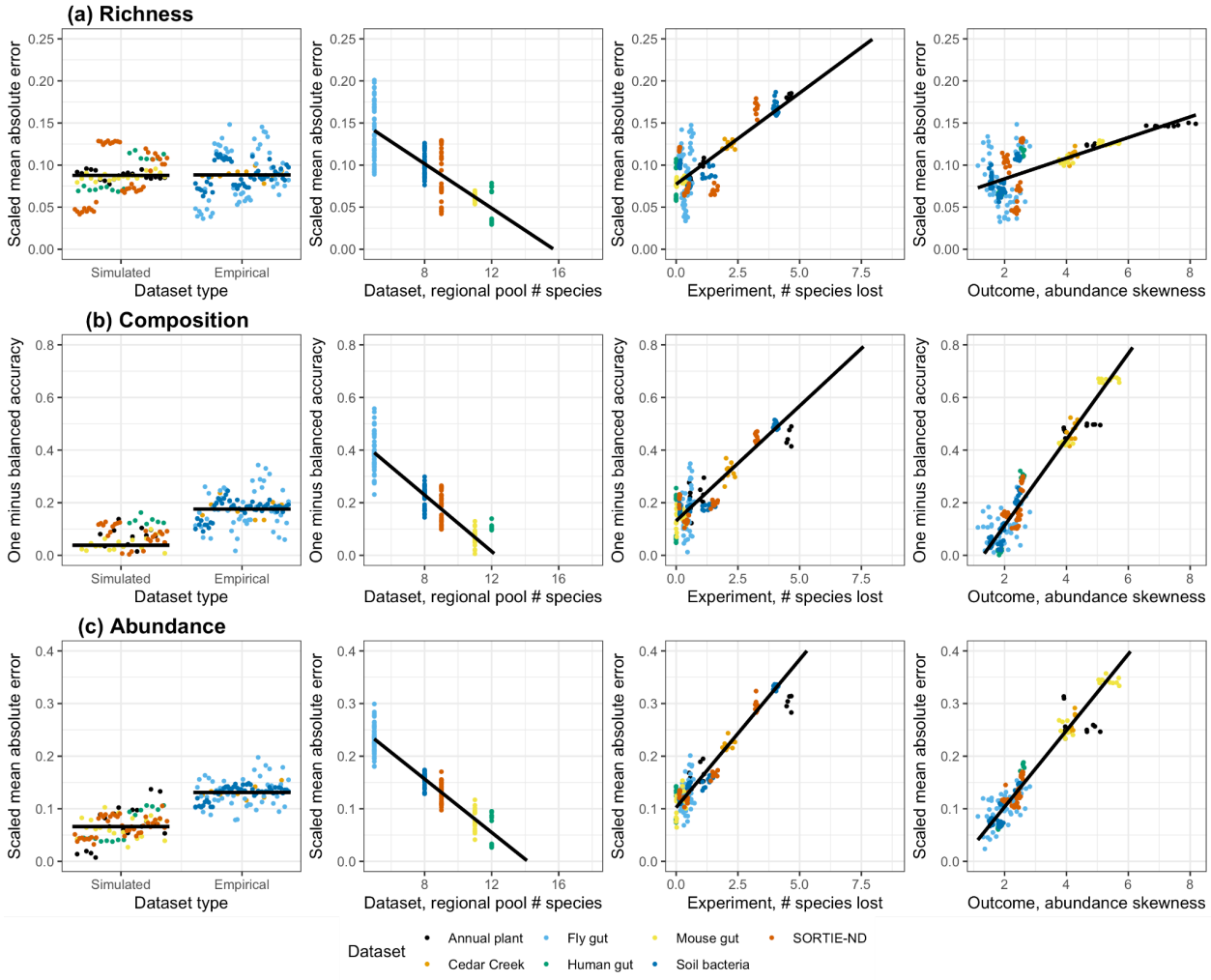
842

843 **Figure 5.**



844

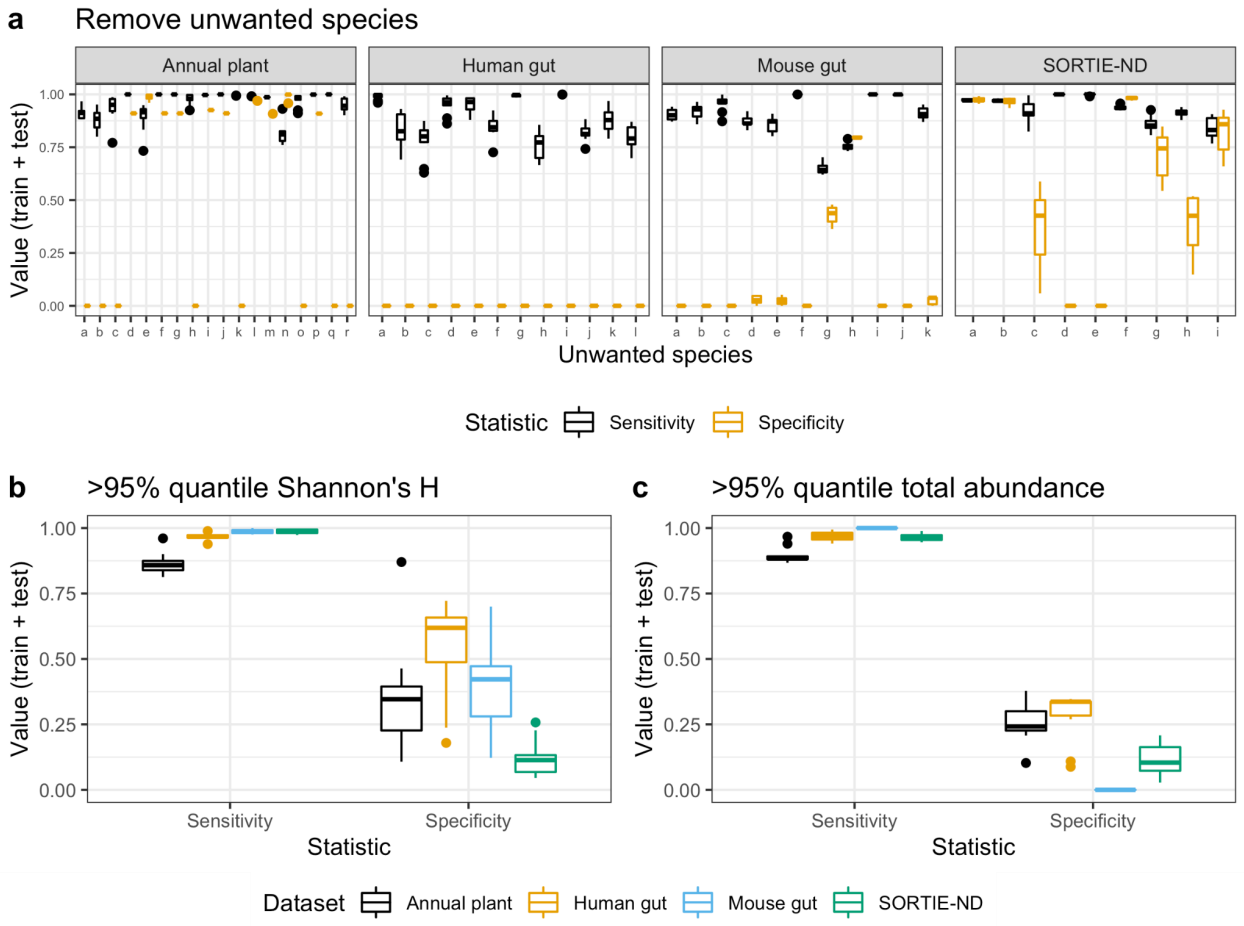
845 **Figure 6.**



846

847

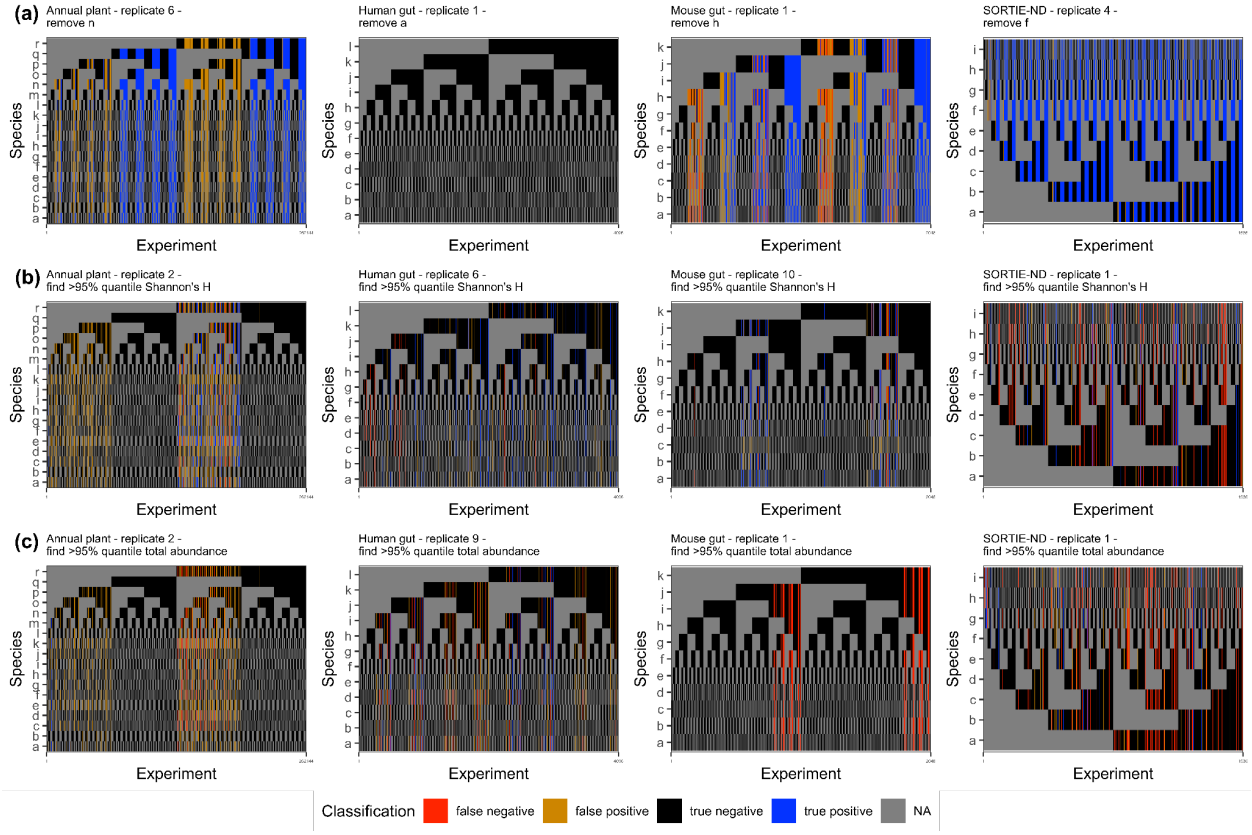
848 **Figure 7.**



849

850

Figure 8.


(a)
Annual plant - replicate 6 - remove n
Human gut - replicate 1 - remove a
Mouse gut - replicate 1 - remove h
SORTIE-ND - replicate 4 - remove f
(b)
Annual plant - replicate 2 - find >95% quantile Shannon's H
Human gut - replicate 6 - find >95% quantile Shannon's H
Mouse gut - replicate 10 - find >95% quantile Shannon's H
SORTIE-ND - replicate 1 - find >95% quantile Shannon's H
(c)
Annual plant - replicate 2 - find >95% quantile total abundance
Human gut - replicate 9 - find >95% quantile total abundance
Mouse gut - replicate 1 - find >95% quantile total abundance
SORTIE-ND - replicate 1 - find >95% quantile total abundance
Species
Experiment
Classification
false negative
false positive
true negative
true positive
NA

854 **Supporting Information**

855

856 **Table S1.** Species names for all taxa in each dataset.

Dataset	Label	Abbreviation	Name
Annual plant	a	AGHE	*Agoseris heterophylla*
Annual plant	b	AGRE	*Agoseris retrorsa*
Annual plant	c	AMME	*Amsinckia menziesii*
Annual plant	d	ANAR	*Anagallis arvensis*
Annual plant	e	CEME	*Centaurea melitensis*
Annual plant	f	CLPU	*Clarkia purpurea*
Annual plant	g	ERBO	*Erodium botrys*
Annual plant	h	ERCI	*Erodium cicutarium*
Annual plant	i	EUPE	*Euphorbia peplus*
Annual plant	j	GECA	*Geranium carolinianum*
Annual plant	k	HECO	*Hemizonia congesta* ssp. *luzulifolia*
Annual plant	l	LACA	*Lasthenia californica*
Annual plant	m	LOPU	*Lotus purshianus*
Annual plant	n	LOWR	*Lotus wrangelianus*
Annual plant	o	MEPO	*Medicago polymorpha*
Annual plant	p	NAAT	*Navarretia atractyloides*
Annual plant	q	PLER	*Plantago erecta*
Annual plant	r	SACA	*Salvia columbariae*
Cedar Creek	a	Achmi	*Achillea millefolium* (*lanulosa*)
Cedar Creek	b	Agrsm	*Agropyron smithii*
Cedar Creek	c	Amocan	*Amorpha canescens*
Cedar Creek	d	Andge	*Andropogon gerardi*
Cedar Creek	e	Asctu	*Asclepias tuberosa*
Cedar Creek	f	Elyca	*Elymus canadensis*
Cedar Creek	g	Koecr	*Koeleria cristata*
Cedar Creek	h	Lesca	*Lespedeza capitata*
Cedar Creek	i	Liaas	*Liatris aspera*
Cedar Creek	j	Luppe	*Lupinus perennis*
Cedar Creek	k	Monfi	*Monarda fistulosa*
Cedar Creek	l	Panvi	*Panicum virgatum*
Cedar Creek	m	Petpu	*Petalostemum purpureum*
Cedar Creek	n	Poapr	*Poa pratensis*
Cedar Creek	o	Queel	*Quercus ellipsoidalis*
Cedar Creek	p	Quema	*Quercus macrocarpa*
Cedar Creek	q	Schsc	*Schizachyrium scoparium*
Cedar Creek	r	Sornu	*Sorghastrum nutans*
Fly gut	a	LP	*Lactobacillus plantarum*

Fly gut	b	LB	*Lactobacillus brevis*
Fly gut	c	AP	*Acetobacter pasteurianus*
Fly gut	d	AT	*Acetobacter tropicalis*
Fly gut	e	AO	*Acetobacter orientalis*
Human gut	a	BH	*Blautia hydrogenotrophica*
Human gut	b	CA	*Collinsella aerofaciens*
Human gut	c	BU	*Bacteroides uniformis*
Human gut	d	PC	*Prevotella copri*
Human gut	e	BO	*Bacteroides ovatus*
Human gut	f	BV	*Bacteroides vulgatus*
Human gut	g	BT	*Bacteroides thetaiotaomicron*
Human gut	h	EL	*Eggerthella lenta*
Human gut	i	FP	*Faecalibacterium prausnitzii*
Human gut	j	CH	*Clostridium hiranonis*
Human gut	k	DP	*Desulfovibrio piger*
Human gut	l	ER	*Eubacterium rectale*
Mouse gut	a	Bar	*Barnesiella*
Mouse gut	b	undLac	und. Lachnospiraceae
Mouse gut	c	uncLac	uncl. Lachnospiraceae
Mouse gut	d	Oth	Other
Mouse gut	e	Bla	*Blautia*
Mouse gut	f	undMol	und. uncl. Mollicutes
Mouse gut	g	Akk	*Akkermansia*
Mouse gut	h	Cop	*Coprobacillus*
Mouse gut	i	Clodif	*Clostridium difficile*
Mouse gut	j	Ent	*Enterococcus*
Mouse gut	k	undEnt	und. Enterobacteriaceae
Soil bacteria	a	Ea	*Enterobacter aerogenes*
Soil bacteria	b	Pa	*Pseudomonas aurantiaca*
Soil bacteria	c	Pch	*Pseudomonas chlororaphis*
Soil bacteria	d	Pci	*Psuedomonas citronellolis*
Soil bacteria	e	Pf	*Pseudomonas fluorescens*
Soil bacteria	f	Pp	*Pseudomonas putida*
Soil bacteria	g	Pv	*Pseudomonas veronii*
Soil bacteria	h	Sm	*Serratia marcescens*
SORTIE-ND	a	ACRU	*Acer rubrum*
SORTIE-ND	b	ACSA	*Acer saccharum*
SORTIE-ND	c	BEAL	*Betula alleghaniensis*
SORTIE-ND	d	FAGR	*Fagus grandifolia*
SORTIE-ND	e	TSCA	*Tsuga canadensis*
SORTIE-ND	f	FRAM	*Fraxinus americana*
SORTIE-ND	g	PIST	*Pinus strobus*
SORTIE-ND	h	PRSE	*Prunus serotina*
SORTIE-ND	i	QURU	*Quercus rubra*

Text S1. Datasets and pre-processing steps.

<u>'Annual plant' dataset</u>

We obtained data from a field-parameterized plant competition model, which describes the dynamics of annual plants with seed banks (12,87). This model is more complex than the generalized Lotka-Volterra, as it includes population stage structure and nonlinear competition. Data came from 18 California annual plants (79). We modified the model reported in (79) to include multi-species competition, following (80). The modified discrete-time model describes the abundance of seeds of species *i* at time *t+1* as:

$$N_{i,t+1} = N_{i,t}[(1 - g_i)s_i + g_i F_i]$$

where:

$$F_i = \lambda_i / \left(1 + \sum_j \alpha_{ij} g_j N_{j,t}\right)$$

The modification is the inclusion in the denominator of F_i of a sum over all species, rather than the sum over only two focal species. Here, λ_i is the per germinant fecundity of species i, g_i is the germination rate of species i, s_i is the annual survival rate of ungerminated seed in the soil of species i, F_i is the number of viable seeds produced per germinated individual of species i, and α_{ij} is the per capita effect of species j on species i. 66/234 values of α_{ij} which were missing from the dataset were replaced with the mean value in the dataset per (88).

For each of the communities possible among the species pool, we initialized all species present in the experiment to $N_i(t = 0) = 1$, and ran for 1000 generations (long enough to reach

878 equilibrium). Richness and composition were calculated by flagging species with $N_i(t =$

879 $1000) \geq 0.01$.

880

881 <u>'Cedar Creek' dataset</u>

882 We obtained data from the Cedar Creek Biodiversity II 'e120' experiment. This dataset describes

883 annual aboveground biomass estimates (from 1994 to 2018) of 154 experimentally assembled

884 plant communities of varying composition (81). For each plot, we set the experimental

885 conditions (X) to whether the plot contained each of *n*=18 species (16 intentionally planted, plus

886 2 volunteer species). We then set the final abundance to each species' biomass in each plot in

887 2018. Richness and composition were calculated by flagging species with $N_i(t = 2018) > 0$.

888 This approach conflates biomass with abundance and does not account for biomass from other

889 non-focal species that colonized each plot by 2018 (e.g. numerous weeds), but is a reasonable

890 choice given the limitations of the data.

891

892 <u>'Fly gut' dataset</u>

893 We obtained data for germ-free fruit flies experimentally inoculated with each possible

894 combination of core species of fly gut bacteria at 3-day intervals, from (82). For each treatment,

895 we set the experimental conditions (X) to whether the fly had been inoculated with each of *n*=5

896 bacterial taxa. We then set the final abundances to the number of colony forming units of each

897 taxon after 10 days of experimental treatments. A total of q=48 replicate flies were used per

898 treatment. Richness and composition were calculated by flagging species with $N_i(t = 10) > 0$.

899

900 <u>'Human gut' dataset</u>

901 The generalized Lotka-Volterra model was used to simulate outcomes, based on the equation:

902 $$\frac{d\widehat{N}(t)}{dt} = diag\left(\widehat{N}(t)\right)\left(\hat{r} + A\widehat{N}(t)\right)$$

903 Here $\widehat{N}$ is a vector of abundances among species in the regional pool, $\hat{r}$ is the vector of density-
904 independent growth rates, and A is the matrix of interaction coefficients, with entry A_{ij}
905 representing the change in species i's per-capita growth rate for a unit change in the density of
906 species j.

907

908 For each of the communities possible among the species pool, and for a given set of A and r
909 parameters, we analyzed the model over the reduced dimensionality corresponding to the number
910 of species introduced in the local community. We initialized all species present in the experiment
911 to $N_i(t = 0) = 1$, then solved the differential equation up to t=1000 using the *ode* function in the
912 *deSolve* package in R. Richness and composition were calculated by flagging species with $N_i^* >$
913 0.01.

914

915 We parameterized the model for a n=11 mouse gut microbial community including the pathogen
916 *Clostridium difficile* (9) based on (84).

917

918 <u>'Mouse gut' dataset</u>

919 We followed the steps outlined for the 'human gut' dataset but using A and r parameters for a
920 n=12 synthetic human gut microbial community (83).

921

922 <u>'Soil bacteria' dataset</u>

We obtained data from experimental assembly of soil bacterial communities from (51). Communities were assembled at varying densities in microplate microcosms each comprising five cycles each comprising 48 hours of growth, followed by a 1500-fold dilution into fresh media. Data include species grown alone, in pairs, in triplets, in single-species drop-outs, and all together. Experiments were replicated from 2 to 14 times. We set the abundance of each species to its optical density after this growth period. Richness and composition were calculated by flagging species with $N_i(t = 240) > 0$.

'SORTIE-ND' dataset

We used the SORTIE-ND (version 7.0.5) model of forest dynamics, which is an individual-based forest simulator that includes demography and life history stage transitions, light competition, spatially explicit dispersal, and other processes (85,86). This model was chosen for its high complexity and stochasticity.

We obtained a parameterization of the model for *n*=9 hardwood species in eastern North America at 42°N latitude ('GMD', available by download from http://sortie-nd.org/software/7_05/sample_par_file.zip). We modified this file to change the plot size to 100 x 100 m, to run for 1000 years (200 5-year time steps) with no external disturbances, and set the parameters for Weibull seed rain and Weibull seed beta to species-specific values reported at http://sortie-nd.org/software/sample_par_file.html), as the default parameter file erroneously includes blank values (personal communication, L. Murphy and C. Canham, 2 Sept. 2021). The simulation does not come to equilibrium but rather includes fluctuations in abundance, due to the

effects of light-based competition and no self-thinning in the understorey. Some stochastic extinctions also occur.

For each local community that could be assembled from this species pool, we then ran simulations, initializing all species to an initial density of either 0 or 25 saplings ha^{-1} (default initial values) and running q=3 replicates per initial condition. We determined abundance as the absolute density of adults at t=1000. Richness and composition were calculated by flagging species with adult densities of $N_i(t = 1000) > 1$.

Text S2. Guidelines for implementing *LOVE* in R.

LOVE can be implemented using any machine learning algorithm with an input data file pairing X and Y values. This implementation assumes that input data tables have column headers

a, b, c, … , stable, feasible, richness, a.star, b.star, c.star, …

Each row in such a table corresponds to a given experiment. The variables in red correspond to the experimental conditions (X), and are 0/1 valued. The variables in orange and purple are the coexistence outcomes (Y). Orange variables are each a scalar-valued outcome, while purple variables indicate the vector-valued outcome abundance. We also pre-process the outcome abundances into a fixed number of quantile classes to avoid zero-inflation problems.

The current implementation assumes (for text parsing simplicity) that the species names are actually letters ("a", "b", "c" and so on). It is therefore also useful to generate a translation table enabling mapping between these letters and the true species names (see **Table S1** for an example).

Once a coexistence dataset has been processed and formatted this way (as all the datasets in this study have been; see the data* subfolders in the supporting code), it is straightforward to apply LOVE. Suppose the above table is called assemblages; in R, LOVE can then be minimally implemented for scalar outcomes as:

```
ranger::ranger(Y ~ a + b + c + ..., data=assemblages)
```

and for multivariate outcomes as:

```
randomForestSRC::rfsrc(Multivar(a.star, b.star, c.star, …) ~ a + b + c + …, data=assemblages)
```

Once a model object has been fit in either case, predictions on test datasets can be made using predict().

The supporting code provides a wrapper function, predict_rf(). The function takes the argument yvar, which can take value _abundance for multivariate outcomes or which can take any other value for a scalar outcome (e.g. richness). This function also provides an argument method, which implements LOVE, LOVE + reshuffle and naive predictions, as well as an argument rows_train that allows holding out certain subsets of data for training vs. testing. Class prevalence is balanced by default for univariate categorical predictions and ten quantile bins are made by default for multivariate continuous predictions.

The supporting code also provides a wrapper function, do_predictions(), which generates the cross-validation analyses across all experimental designs, methods, and response variables, for a given dataset. This function also reports model error, via balanced accuracy for univariate categorical response variables, mean absolute error for univariate continuous response variables, and casewise mean versions of the above for multivariate response variables. Outputs also include the training–set and test–set experimental conditions and outcomes.

Figure S1. Visualization of experimental conditions and abundance outcomes for the annual plant dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red.

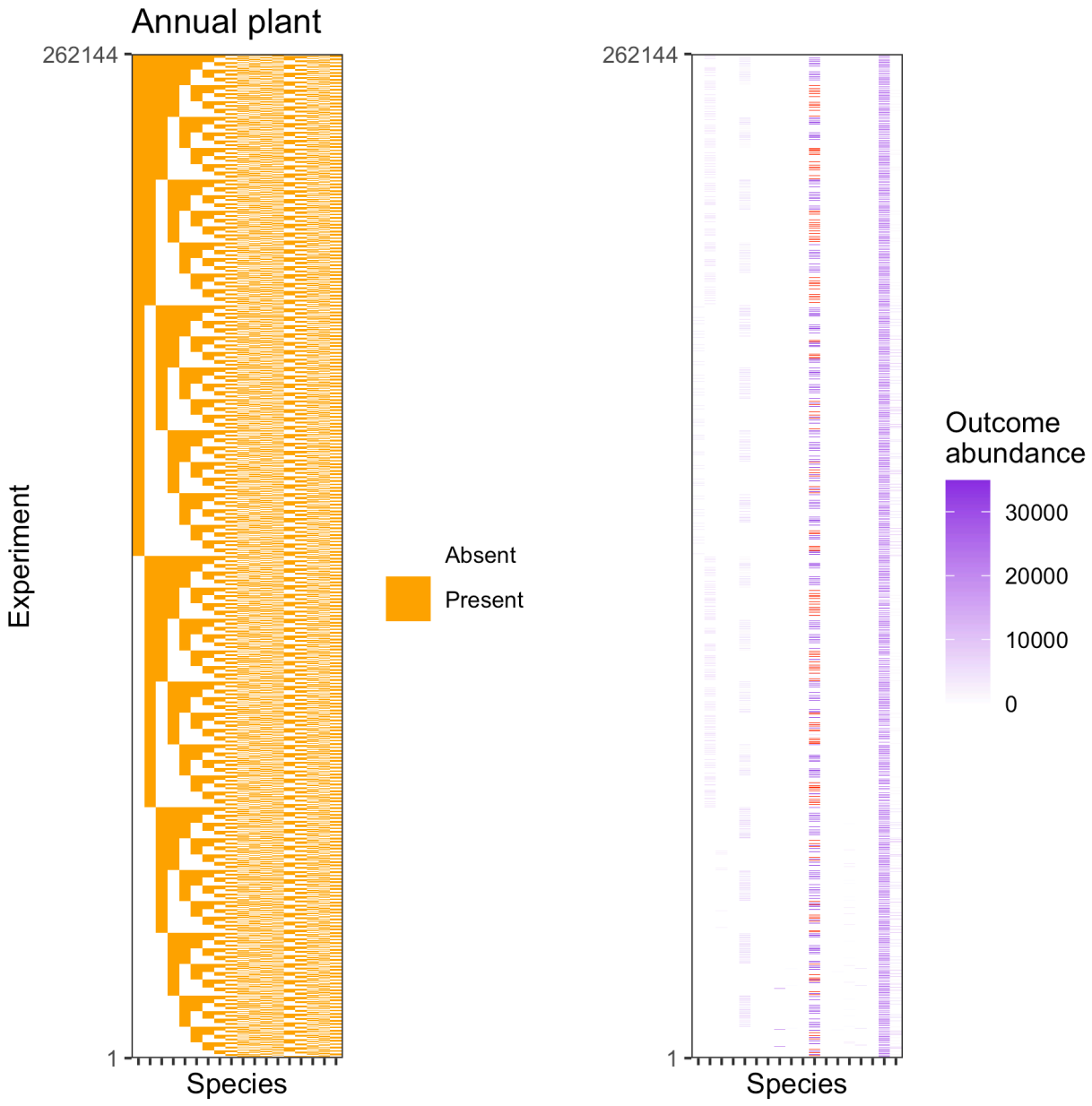
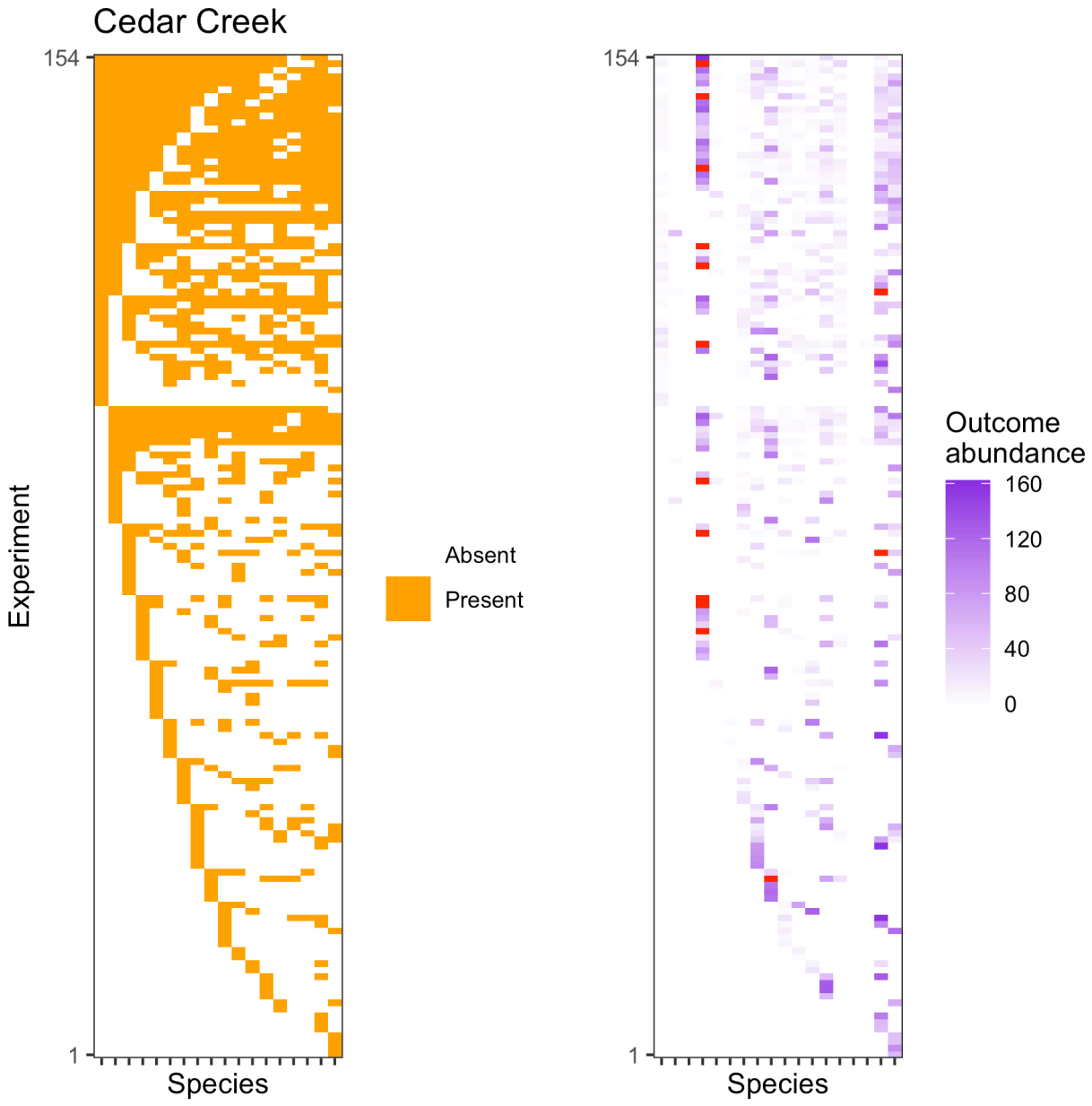


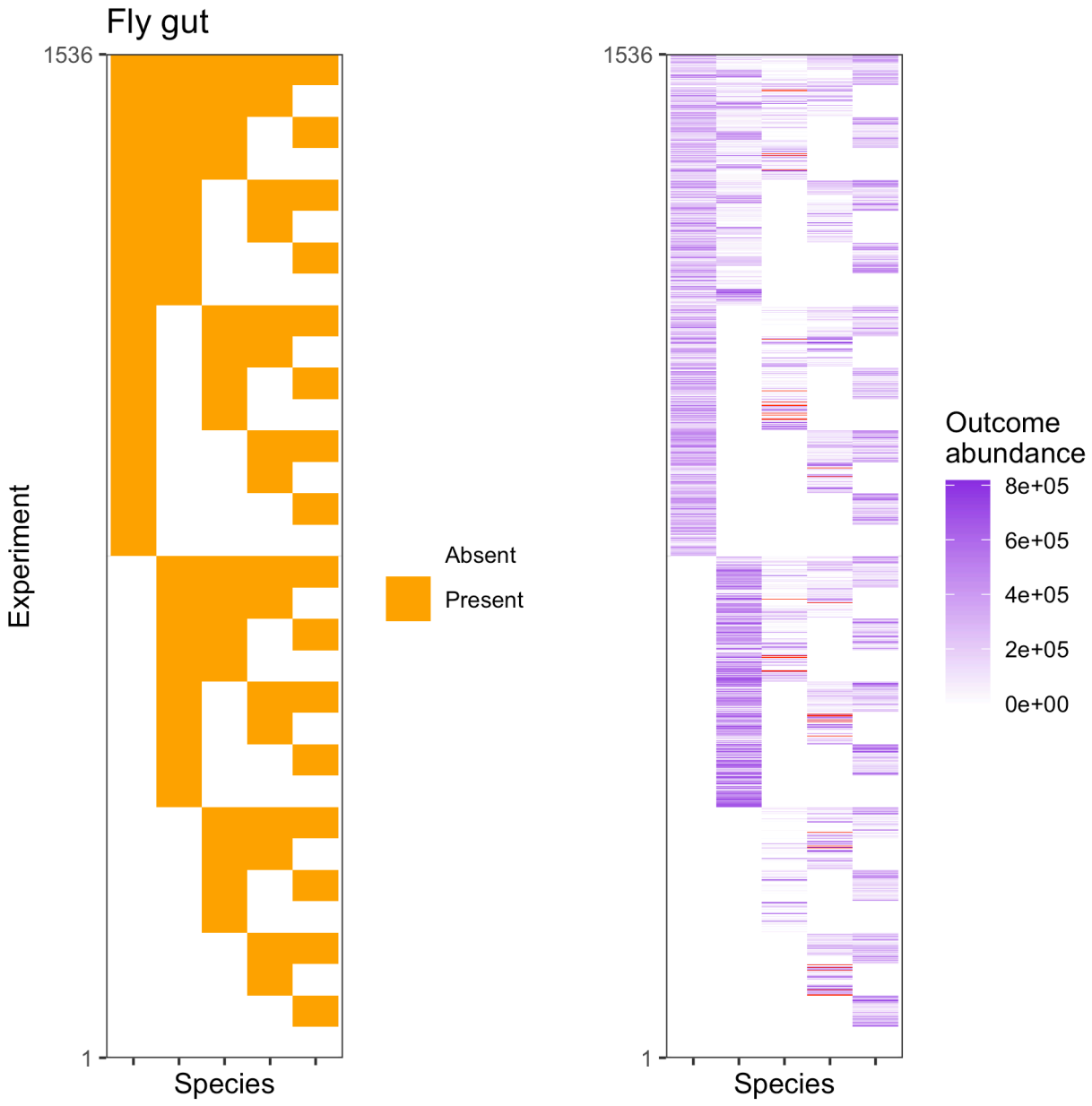
Figure S2. Visualization of experimental conditions and abundance outcomes for the Cedar Creek dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red.



1009 **Figure S3.** Visualization of experimental conditions and abundance outcomes for the fly gut

1010 dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)**

1011 outcomes. Quantile clipped values are colored red.



1012

1013

Figure S4. Visualization of experimental conditions and abundance outcomes for the mouse gut dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red.

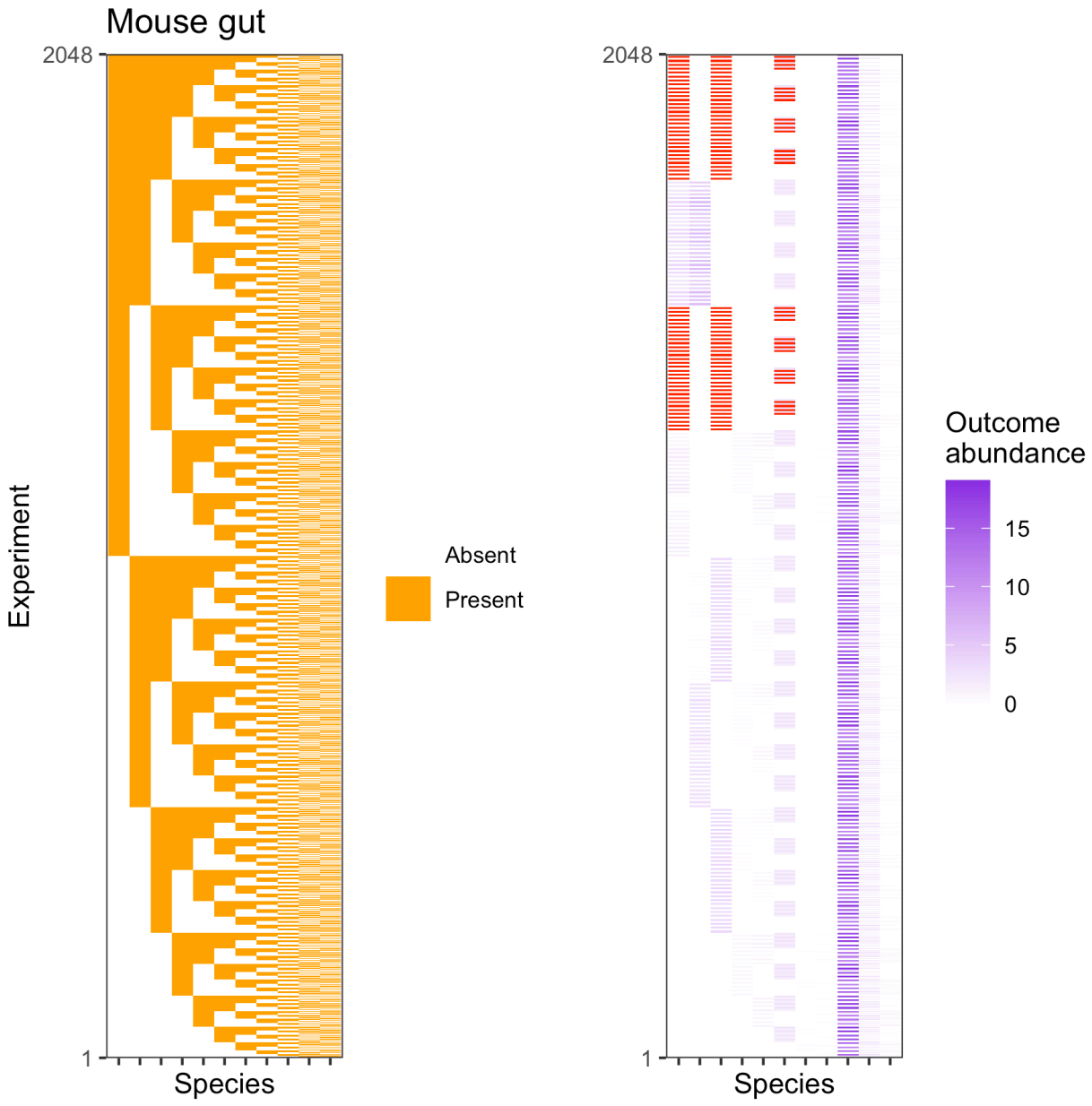


Figure S5. Visualization of experimental conditions and abundance outcomes for the soil bacteria dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red.

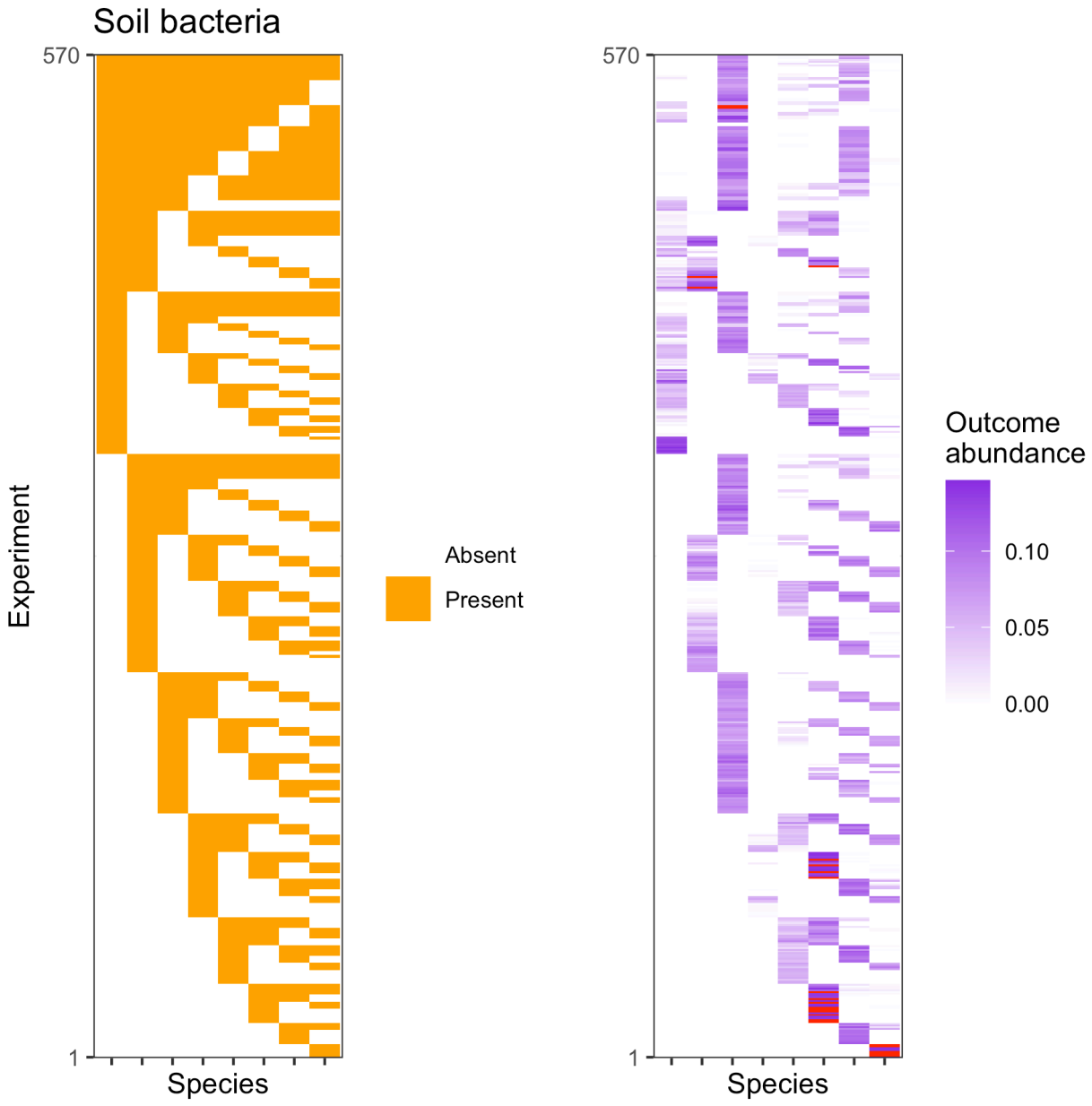


Figure S6. Visualization of experimental conditions and abundance outcomes for the SORTIE dataset. Panels show **(a)** initial species presence/absence data for each experiment and **(b)** outcomes. Quantile clipped values are colored red.

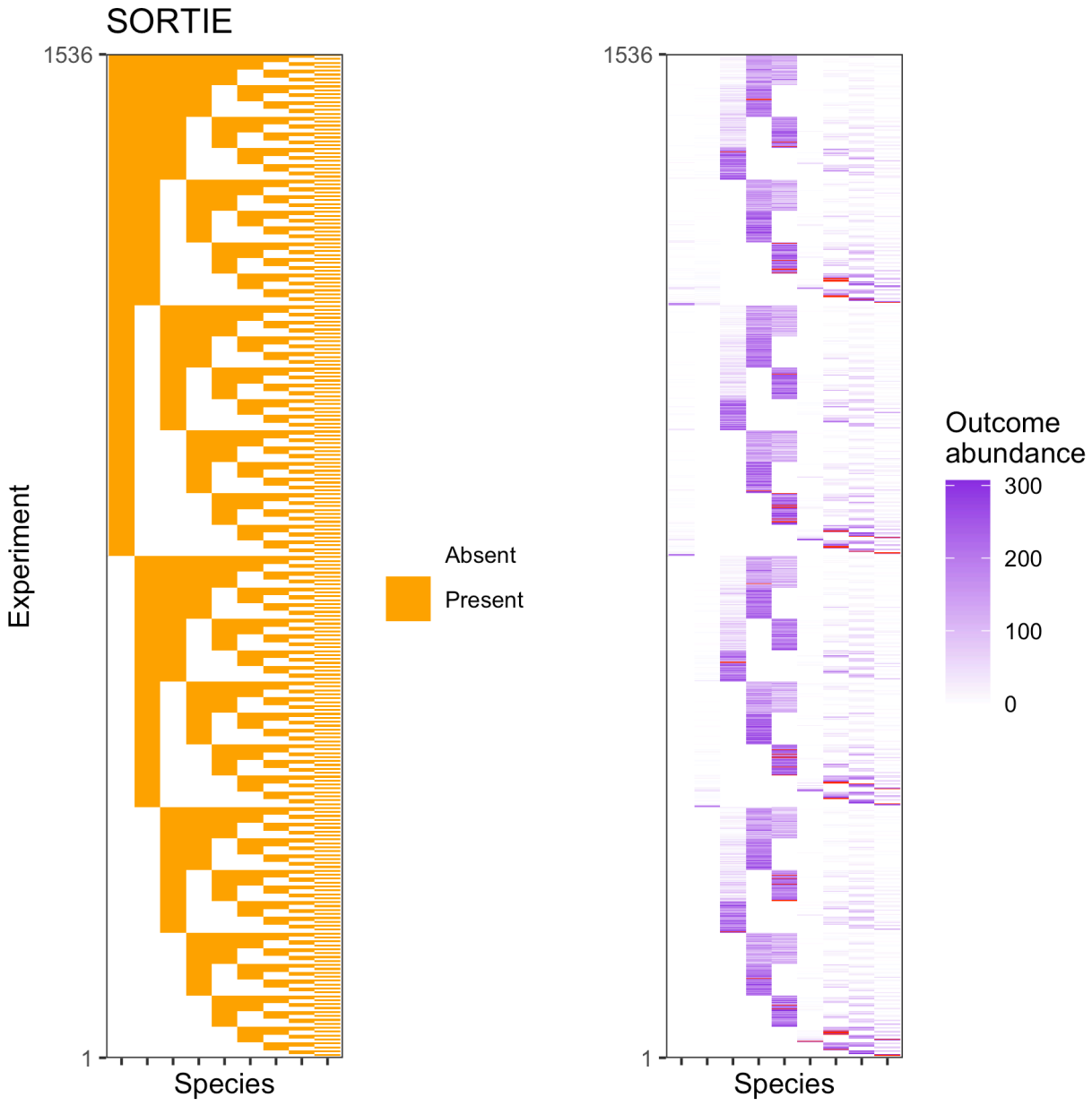


Figure S7. Test-set predictions of abundance compared to observations for *LOVE* using mixed experimental design and at a training-set sample size of *n=89* experiments for each dataset. Lines connect species-replicate combinations across all test-set experiments (only 100 randomly selected experiments per group are plotted for visual clarity). The black line indicates the 1:1 expectation.

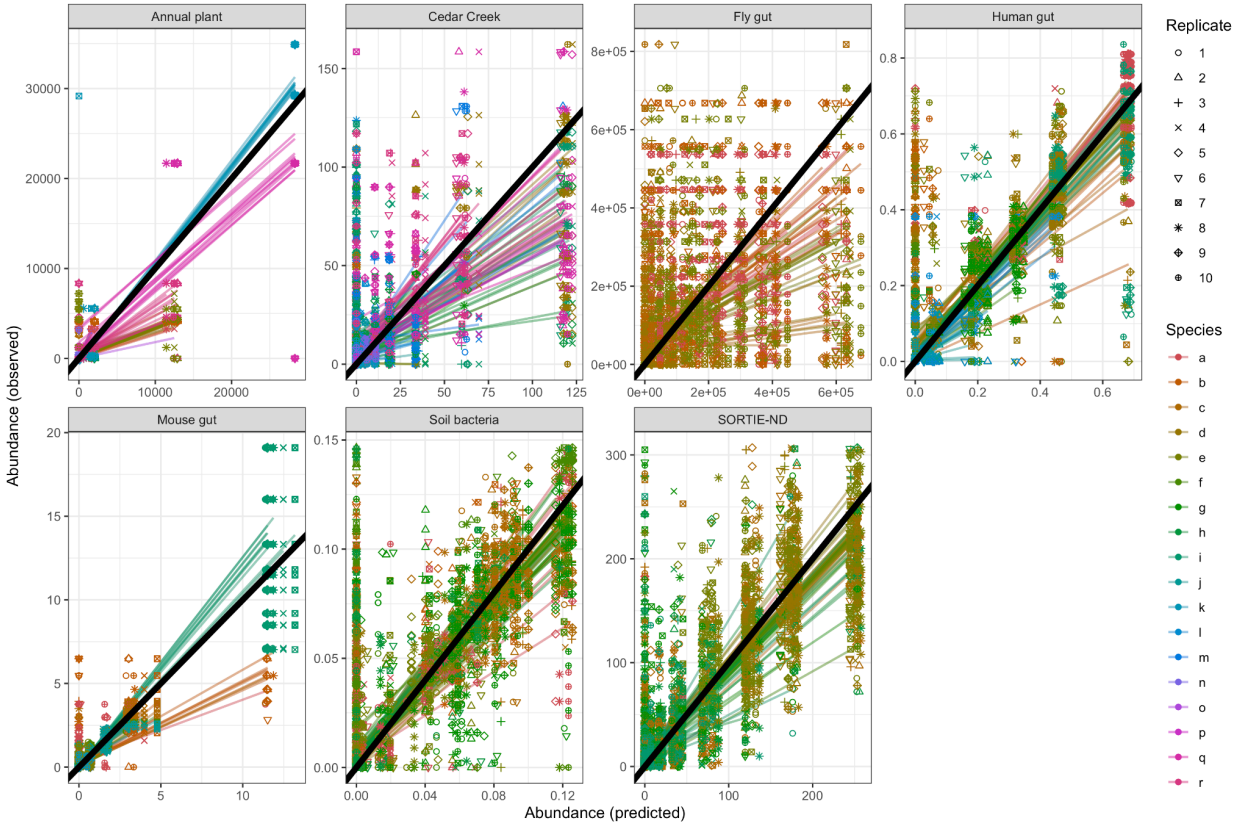
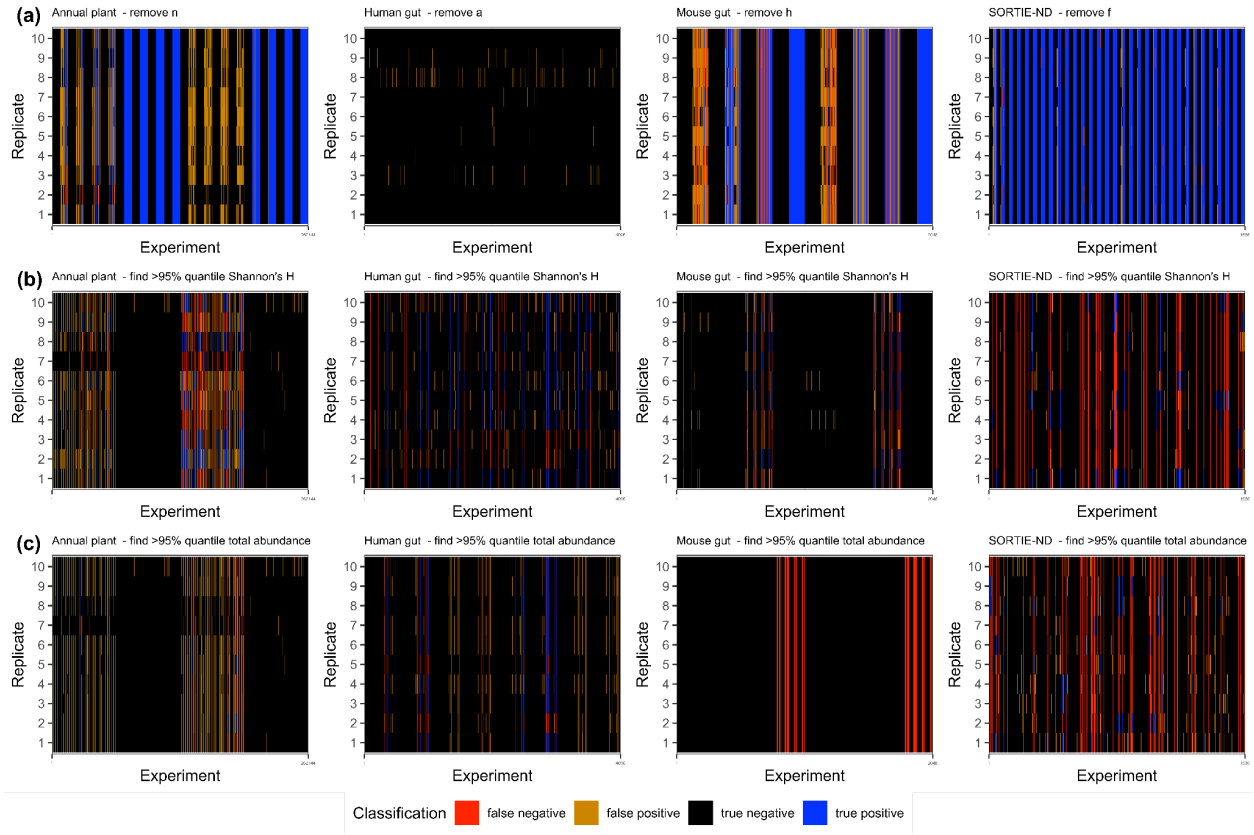



Figure S8. Performance of LOVE for prioritization of experiments for three tasks, **(a)** removing a species that is initially present; **(b)** obtaining a Shannon's H value greater than that for 95% of all possible experiments; and **(c)** obtaining a total abundance greater than that for 95% of all possible experiments. Panels indicate classification of experiments across multiple replicates.