HHS Public Access

Author manuscript

Annu Rev Ecol Evol Syst. Author manuscript; available in PMC 2023 December 15.

Published in final edited form as:

Annu Rev Ecol Evol Syst. 2022 November ; 53(1): 113–136. doi:10.1146/annurev-ecolsys-102320-093722.

Simulation Tests of Methods in Evolution, Ecology, and Systematics: Pitfalls, Progress, and Principles

Katie E. Lotterhos[1], **Matthew C. Fitzpatrick**[2], **Heath Blackmon**[3]

[1]Department of Marine and Environmental Sciences, Northeastern University, Nahant, Massachusetts, USA

[2]Appalachian Lab, University of Maryland Center for Environmental Science, Frostburg, Maryland, USA

[3]Department of Biology, Texas A&M University, College Station, Texas, USA

Abstract

Complex statistical methods are continuously developed across the fields of ecology, evolution, and systematics (EES). These fields, however, lack standardized principles for evaluating methods, which has led to high variability in the rigor with which methods are tested, a lack of clarity regarding their limitations, and the potential for misapplication. In this review, we illustrate the common pitfalls of method evaluations in EES, the advantages of testing methods with simulated data, and best practices for method evaluations. We highlight the difference between method evaluation and validation and review how simulations, when appropriately designed, can refine the domain in which a method can be reliably applied. We also discuss the strengths and limitations of different evaluation metrics. The potential for misapplication of methods would be greatly reduced if funding agencies, reviewers, and journals required principled method evaluation.

Keywords

evaluation; validation; domain of applicability; area under the curve; benchmark data sets; equifinality

1. INTRODUCTION

The fields of ecology, evolution, and systematics (EES) have been revolutionized by advances in computation, data storage, sampling technology, and genomics over the last thirty years. With these advancements, biologists can address increasingly complex questions spanning topics from genes to ecosystems. The big data that result from these

k.lotterhos@northeastern.edu .

AUTHOR CONTRIBUTIONS

This project was conceptualized and administered by K.E.L. Funding was obtained by K.E.L. and M.C.F. The visualizations were developed by K.E.L. and H.B. Data curation (Supplemental Material) was performed by K.E.L. All authors contributed to the literature review, writing, and revision of the manuscript.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

inquiries increasingly are being analyzed using novel and often complex statistical methods. However, methods often see usage beyond their intended application and before their behavior is fully understood.

Method:

a statistical model, machine learning algorithm, and/or pipeline that is used in hypothesis testing, model selection, prediction, and/or forecasting

New methods are constantly being developed because, as scientists, we seek tools that accommodate the complexities of the real world. The vast variety of life forms and ways of studying them (and funding constraints) often result in nonrandom experimental designs (e.g., block designs or stratified sampling) and/or nonindependence in the data that must be accounted for in statistical analyses. Even random sampling designs may have nonindependence, due to spatial and/or temporal autocorrelation, as well as shared eco-evolutionary history among species, populations, individuals, and genes. New methods are also being developed to better account for nonlinearities and interactions, which are common in biological data sets, and machine learning increasingly is being applied to deal with these unique challenges (Lucas 2020, Schrider & Kern 2018). For these reasons, EES are emerging as fields in which knowledge of statistics is as important as knowledge of natural history so that methods can be reliably applied to complex data sets (Austin et al. 2006). However, proper application of methods and interpretation of the inferences they provide requires rigorous evaluation to fully understand the methods' behavior, strengths, and limitations.

Nonrandom:

samples were not collected with equal probability

Nonindependence:

samples are related in some way (e.g., via evolutionary history, ecological interactions, spatial or temporal proximity)

Inference:

a conclusion reached on the basis of evidence and reasoning

Evaluation:

a quantitative comparison of how well one or more methods perform on different data sets and/or relative to each other

1.1. Evaluation Versus Validation

The goal of method evaluations is to quantify the reliability of methods in different scenarios. In comparing and testing methods, the terms evaluation and validation are often used interchangeably, but they are not the same. Evaluation is the process of assessing method(s) on ground-truth data set(s). Ideally, method evaluation is based on simulated data that capture the complexities of empirical data, as well as simulated data that are generated from alternative processes that have the potential to confound the method (Figure 1). The result of a successful evaluation is an understanding of the strengths and limitations of the method(s).

Validation:

providing a body of evidence that a method meets a set of a priori criteria for a specific real-world application

Ground-truth data set:

information that is known to be real or true, provided by direct simulation (in silico) or measurement (in situ)

Validation, however, is the process of determining the degree to which a method provides accurate inference about the real world based on the intended uses of the method (AIAA 1998). The best model in an evaluation study may still have high error rates or low predictability for a particular situation. In other words, the best model may still be worse than no model at all. The result of a successful validation is an understanding of how well the method is expected to perform for a specific application or data set (Figure 1). Thus, important components of validation are the *a priori* designation of the minimum criteria a method should meet for a specific application and a quantitative evaluation of whether the method achieves these criteria (Natl. Res. Counc. 2012).

1.2. Why Test Methods with Simulations?

In this review, we illustrate common pitfalls of method evaluations in EES, the advantages of testing methods with simulated data, and the role of carefully planned evaluations in preventing the misuse of faulty methods. Testing methods with simulations offers a powerful approach because the true underlying process that generated the data is known (Lotterhos et al. 2018, Zurell et al. 2010). Much can be learned about black-box methods just by looking at what data go in and what results come out (Bergstrom & West 2021, Lucas 2020). Simulations also have the benefit of conducting many replicate experiments *in silico* when traditional experiments may be more timely and costly or impossible. For example, Felsenstein (1985) used simulated data to illustrate the failure of traditional statistical approaches and to evaluate a new method, phylogenetic independent contrasts, that could account for phylogenetic relationships among data points. This application of simulations not only revealed the reason that an existing method failed but also catalyzed the development of phylogenetic comparative methods for decades.

2. METHODS IN EVOLUTION, ECOLOGY, AND SYSTEMATICS: A CONSTANTLY SHIFTING LANDSCAPE

2.1. Methods Developed First, Broken Later

> Now, here, you see, it takes all the running you can do, to keep in the same place.
>
> —The Red Queen, *Through the Looking Glass* by Lewis Carroll

New methods that seem to do the same thing as existing methods—but have nuanced differences that may be critically important—are being developed constantly. For example, as of 2019, over 400 methods were available to analyze single-cell RNA sequencing data (Weber et al. 2019). Here, we illustrate the constantly shifting landscape of methods development, application (approximated as citation rate), and evaluation on three categories of widely used methods in EES: genome scans for local adaptation via outlier tests for genetic differentiation (Hoban et al. 2016), tests of differential diversification in systematics (Ng & Smith 2014), and species distribution models (SDMs) in ecology (Elith & Leathwick 2009) (for detailed methods, see Supplemental Methods). For each category, we plotted citation rates for different methods relative to when one or more highly cited evaluation study was published (Figure 2).

Species distribution models (SDMs):

empirical models that relate species occurrences to environmental predictor variables to understand and predict species distributions across space and time

These three scenarios illustrate different dynamics between method development, application, and evaluation. For example, in the category of outlier tests for genetic differentiation, a method called FDIST and its extensions (LOSITAN) was widely used for ~15 years, until four evaluations between 2010 and 2014 (Lotterhos & Whitlock 2014, Narum & Hess 2011, Pérez-Figueroa et al. 2010, Whitlock & Lotterhos 2015) used simulations to show that the algorithm had high false positive rates under more realistic species demographies. The citation rates for FDIST and LOSITAN started to decrease between 2015 and 2020 (Figure 2a). Likewise in systematics, a method called BiSSE was widely used for ~5 years before a 2015 evaluation of state dependent diversification methods (Rabosky & Goldberg 2015) demonstrated that BiSSE had an inflated false positive rate because it frequently chose the most complex model among all models considered (Figure 2b). This led to an increase in the use of alternatives like BAMM and HiSSE, which do not suffer from this shortcoming (Figure 2b) (Beaulieu & O'Meara 2016). For SDMs, use of a method known as MaxEnt (Phillips & Dudík 2008) exploded in part due to its favorable performance in a large model evaluation study (Elith et al. 2006) and its implementation in a user-friendly software interface (Figure 2c). These examples demonstrate that although methods are constantly being developed, evaluation studies can lead to decreased citation rates for poorly performing methods and increased citation rates for highly performing methods, which reflects the frequency with which the method is used. There is, however,

typically a several-year lag between the publication of an evaluation and any change in citation rate.

2.2. Pitfalls in Developing Methods

Different (not mutually exclusive) hypotheses may explain the shifting landscape of methods development and use. On one hand, the scientific process itself is cyclical: As hypotheses are tested, the outcomes lead to new hypotheses to test. On the other hand, if new methods are not rigorously evaluated when they are first developed, they can become widely used for some time until their limitations are known, after which they fall out of use. Published studies comparing a new method to existing methods may be (intentionally or unintentionally) biased in favor of the new method (Boulesteix et al. 2017, 2018; Weber et al. 2019).

A number of pitfalls can occur during method development that result in a biased evaluation. For example, rather than using ground-truth data sets, it is not uncommon for methods to be evaluated using only empirical data (as in Elith et al. 2006) (Figure 2c) or simulations that align with the assumptions of the method (which leads to verification rather than evaluation or validation) (Natl. Res. Counc. 2012). An extension of the latter case is subjectivity in the choice of simulated (or empirical) data sets that are used to test the proposed method, which results in bias in favor of the proposed method (Weber et al. 2019). Other common pitfalls result from familiarity with the proposed method (when the developer tunes the parameters for the proposed method but uses default parameters for the other methods) and subjectivity in the choice of evaluation metrics used to compare methods, both of which can bias results in favor of the proposed method (Weber et al. 2019).

Verification:

process of determining how accurately a computer program correctly implements the equations of the mathematical or statistical model

Once a method is published, it becomes available for application, and if the assumptions and limitations of the method—and how to determine them—are not well described, then the method could be applied incorrectly. There are unquantified costs associated with the improper use of methods, such as conducting analyses to understand why changing a method led to a different result on the same data set, conducting assays to confirm results, and in worst case scenarios, using erroneous results to guide decision-making in health or conservation contexts.

3. GOALS IN TESTING METHODS

> Every algorithm, no matter how well it performs generally, may have an Achilles' heel—some weakness that will cause it to stumble when faced with a certain problem.
>
> —Natl. Res. Counc. (2012, p. 24)

The lack of standards in EES for method evaluations creates challenges for researchers who seek to apply methods to their data. These challenges share the common question of whether the user is making the correct inference based on the method output. Here, we review four common challenges that arise when applying methods to data, and the associated goal of method evaluation that is important for addressing each challenge (Table 1), with examples. To illustrate how method evaluations have helped to achieve these goals, we reviewed 37 evaluation studies in EES (Supplemental Table 1).

3.1. Challenge #1: How Does One Know if Applying a Given Method to the Data Gave a True Result in a Hypothesis Test?

In the null hypothesis statistical framework, one tests the probability of obtaining the data, given a null hypothesis. Accurate results of a null hypothesis test depend on whether the data meet the assumptions of the method. A major goal when testing methods is to determine the frequency with which a method returns a true positive or false positive result (Table 1).

This is a common goal in the evaluation of genome scan methods, which are statistical tests that are used to determine the genetic basis of a trait or detect selection in sequence data. In this case, simulations have been used to evaluate how the degree of adaptation, population demography, recombination, admixture, and sampling design affect error rates in genome scans (Capblancq et al. 2018; Forester et al. 2018; Lotterhos 2019; Lotterhos & Whitlock 2014, 2015; Luu et al. 2017) (Supplemental Table 1). For example, Forester et al. simulated genomes under selection by different types of spatial heterogeneity and used these data sets to compare univariate and multivariate methods for detecting selection (Forester et al. 2016). They showed that multivariate methods had lower false positive rates, but also lower power, than univariate methods (Forester et al. 2016).

SDMs also are frequently evaluated for their ability to discriminate presence from nonpresence (i.e., true absence, pseudoabsence, or background data). In one such study, Qiao et al. (2015) tested 8 SDM algorithms using 14 virtual species simulated with different niches in environmental space, with and without barriers to dispersal in geographic space. They found that no single algorithm performed best in all instances, with performance depending on the particular traits of the virtual species being modeled.

3.2. Challenge #2: Is the Prediction from This Method Accurate and Precise?

Methods are used to make predictions in a variety of ways, ranging from parameter estimation on a biological process [e.g., branch length estimation in phylogenetic trees (Brown et al. 2010)] to the development of a model that predicts response variable(s) from explanatory variable(s) [as in statistics and machine learning (Lucas 2020)]. For methods like SDMs that are routinely used to inform management decisions, it is especially important to determine whether models assign the correct probabilities to a given outcome (Chivers et al. 2014). Thus, another goal of testing methods is to quantify the bias and precision of a method prediction (Table 1).

This is a common goal of methods that estimate population parameters (such as population size, niche breadth, or recombination rate) or individual values (such as fitness or a trait value). For instance, the threshold model for discrete character evolution estimates an

unmeasured continuous trait value that leads to the observed discrete trait. In implementing this method, simulations provided ground-truth data sets for which this unobserved value was known, and this allowed for an evaluation of the method that would otherwise be impossible (Revell 2014).

This challenge also extends to the Bayesian model selection framework. For instance, phylogenetics methods seek to accurately infer the posterior probabilities for the correct tree topology among a number of hypothesized topologies. The Bayesian star tree paradox occurs when (informative) data with many nucleotides lead to an inaccurate phylogenetic tree with a high posterior probability, while (uninformative) data with a single nucleotide lead to accurate posterior probabilities for a set of hypothesized trees (Lewis et al. 2005). This result was paradoxical because the performance of methods is expected to improve with the use of a larger number of nucleotides. Simulations have been used to show that the paradox occurred because the true tree from which data were generated was not among the hypotheses being considered in the Bayesian analysis and to demonstrate how better algorithms can lead to more accurate results (Lewis et al. 2005).

3.3. Challenge #3: Equifinality: Did the Pattern in the Data Result from a Process Other Than the Process Originally Assumed?

Research in EES often aims to infer the process(es) responsible for a given pattern based on statistical analysis of that pattern alone. This can be challenging because a number of different processes can result in the same pattern in the data (Dormann et al. 2012, Oreskes et al. 1994). This is known as the problem of equifinality, which can confound method evaluation, for example, when several parameterizations of the model are statistically unidentifiable (many parameter estimates produce an equally good fit to the data) (Dormann et al. 2012). Thus, another goal of testing methods is to develop an understanding of confounding factors, hidden variables, and other processes that could change the inference of the underlying phenomenon (Table 1).

Equifinality:

when several parameterizations exist that equally fit a given data set (also known as nonidentifiability)

The use of simulations can help to identify these confounding processes and demonstrate (or rule out) equifinality, which can lead to more realistic interpretations of empirical data. In phylogenetics, Louca & Pennell (2020) used simulations to show that when diversification is correlated with fluctuating environments, a single phylogenetic tree can be consistent with a myriad of diversification histories (Supplemental Table 1). In ecological genomics, empirical studies have found a negative relationship between population size and a machine-learning prediction of genomic vulnerability, which was inferred to be driven by signals of selection in the genome (Bay et al. 2018, Ruegg et al. 2018). However, a simulation study showed that neutral demographic processes resulting from genetic drift could produce the same pattern (Láruson et al. 2022), illustrating that the empirical observation may not have been driven by the process of selection, as was first inferred (Supplemental Table 1).

Another example comes from the use of so-called joint SDMs that infer signals of biotic interactions from residual covariation between species (Ovaskainen et al. 2017). Patterns of residual covariation in joint SDMs could also arise from statistical artifacts related to missing environmental covariates or poor model fit. In a simulation study, whether residual covariation from joint SDMs reflected biotic interactions depended on the spatial scale, prevalence of the modeled species, the nature of the interaction between species, and other factors (Zurell et al. 2018) (Supplemental Table 1).

3.4. Challenge #4: Were Correct Decisions Made When Designing the Experiment, Filtering Data, or Choosing a Method?

Inferences gained from research in EES are also sensitive to the numerous nuanced decisions made when designing experiments and analyzing data. Failing to detect an effect in a study (e.g., a nonsignificant result) raises questions about whether there truly was an effect and whether the experimental design had enough power to detect an effect. Thus, the fourth major goal of testing methods is to illustrate best practices for the scientific method, including experimental design and data filtering steps, among other nuanced decisions (Table 1).

Simulation studies have been useful for evaluating nuanced decisions. For example, simulations have been used to show that experimental designs in ecology with many levels and fewer replicates per level (but that were not necessarily fully factorial) can yield accurate parameter estimates while also requiring less experimental effort (Blanquart et al. 2013, Cottingham et al. 2005, Inouye 2001) (Supplemental Table 1). Simulations are also becoming commonplace for evaluating SDMs and the numerous methodological decisions that go into fitting such models (Meynard et al. 2019). One consequential decision often encountered when fitting SDMs is determining how, where, and how many pseudoabsences (also known as background data in certain contexts) should be generated to build reliable models. Several studies have used simulations to inform selection strategies and have shown that the best approach depends on the nature of the occurrence data, study region, and statistical algorithm being used to fit SDMs, among other factors (Barbet-Massin et al. 2012, Liu et al. 2019, Lobo & Tognelli 2011).

Method evaluations using simulations have also played an important role in genomic research that requires a large number of choices at each step in the pipeline. For example, simulations have been used to illustrate how (*a*) bias in allele frequencies arises from the way tissue was prepared for sequencing (Arnold et al. 2013, Flanagan & Jones 2018, Gautier et al. 2013), (*b*) sequencing depth can influence error rates from different sequencing platforms (Harismendy et al. 2009), (*c*) bias in genetic diversity statistics is generated from de novo transcriptome assembly (Freedman et al. 2021, Hölzer & Marz 2019), and (*d*) failure to filter data properly for linkage disequilibrium can lead to false positive results in genome scans for selection (Lotterhos 2019) (for details on these simulation studies, see Supplemental Table 1). Simulations are critical for determining the context and boundaries within which data interpretations are valid, as well as providing guidance on decision-making for research.

4. PRINCIPLED METHOD EVALUATION AND VALIDATION FOR EVOLUTION, ECOLOGY, AND SYSTEMATICS

While a mathematical understanding of the algorithms involved in a method can provide a conceptual understanding of its pitfalls and how it could lead to false inference, it is not essential. However, it is essential that simulations be carefully designed to ensure the simulated data provide a proper test of the variety of applications that a method may encounter. This includes considering which features of the data may confound a method and what kind of processes could lead to similar patterns in data.

4.1. Phase I: Plan Development

The first phase in a simulation study is to develop a conceptual plan that describes the problem to be addressed. Based on the assumptions of the methods or the way they are parameterized, it should be possible to hypothesize a domain of applicability that delineates the specific situations and data sets to which that method can be applied (sensu Natl. Res. Counc. 2012). For example, Zurell et al. (2009) simulated a spatially explicit multi-species dynamic population model with range shifts and tested how well SDMs predicted the simulated changes in species distributions. They found that some SDMs could accurately forecast range shifts when species were able to rapidly track climate change (via long-distance dispersal, a slow rate of climate change, and/or a range contraction rather than shift), which helped to define the domain of applicability for SDMs in forecasting range dynamics (Zurell et al. 2009) (Supplemental Table 1).

Domain of applicability:

a region of a domain space in which a method returns acceptably accurate results

We also postulate that a domain of inference (what could be inferred from the method output) should be hypothesized. For example, a horseshoe pattern in a principal components analysis of microbial data was commonly inferred to be a statistical artifact for which a correction should be applied (Morton et al. 2017). However, Morton et al. (2017) showed with simulated data that the horseshoe pattern could be caused by niche differentiation along an ecological gradient, which altered the domain of inference (Supplemental Table 1).

Domain of inference:

the processes that can be deduced or concluded from a method based on evidence and reasoning

These domains should be informed by the goal of the methods to be evaluated (Table 1). Once these domains are hypothesized, the next step is to brainstorm ways to challenge the methods with adversarial data sets (sensu Molnar 2021) that will help to define these domains and also capture biological realism inherent in the empirical data that the methods may encounter in the future. In machine learning, the term adversarial is commonly used to describe challenging data sets that substantially degrade performance and thereby help

determine the domain of applicability (Hendrycks et al. 2021, Molnar 2021). Here, we use adversarial to refer to data sets that capture the extreme characteristics of data that users may choose to analyze with a given method, although they might not cause performance to degrade.

Adversarial data sets:

data sets that challenge a method because they capture the extreme characteristics of data that might be used in a method

4.2. Phase II: Simulation

The second phase in a simulation study is to create a set of ground-truth data sets that will be used for evaluating methods. How to construct simulations from a biologically reasonable description of the system has been reviewed elsewhere (Otto & Day 2011, chapter 2). Here, we review helpful techniques for focusing the simulations in the face of real-world complexity by using contrasting cases or response surface designs and by simulating signals on real data.

4.2.1. Contrasting cases.—Contrasting cases are a set of simulations that differ in important ways that may challenge methods, but these cases avoid confounding by controlling for important features of the data. As an example of contrasting cases, Lotterhos & Whitlock (2014) contrasted the performance of genome scans to identify outliers for the metric F_{ST}, which describes the degree of genetic differentiation between two or more populations at a locus. Spatially divergent selection can cause affected loci to have larger values of F_{ST} than the genome-wide background, and outlier identification can be used to discover the genetic basis of local adaptation. They used simulations that had the same mean F_{ST} but different variance in F_{ST}, which was created by different demographies (Figure 3a). Importantly, to create data sets with the same mean F_{ST}, they did not use the same dispersal or deme size among demographies (Lotterhos & Whitlock 2014). A challenge with contrasting cases is to determine which features of the data to control for. While many of the method evaluations in our review fit into the category of using contrasting cases (Supplemental Table 1), few described whether features of the data were controlled for in different scenarios.

4.2.2. Response surface designs.—In the absence of prior knowledge on the simulation behavior across a range of input parameters, a simple rule of thumb is to explore as much of the input region as possible. Different approaches to response surface designs include sequential designs (Ranjan et al. 2008), Latin hypercube designs (McKay et al. 1979), and variants thereof (Lin et al. 2010, Tang 1993). In the context of statistical sampling, a Latin square contains one simulated data set in each row and each column, while the rows contain different levels of one parameter and the columns contain different levels of a second parameter (Figure 3b). A Latin hypercube maintains these criteria in many dimensions and allows the parameter space to be covered at a reduced effort compared to fully factorial designs.

In our literature review, sequential designs, in which several values of each parameter are simulated, were the most used type of response surface design (Supplemental Table 1). For example, Liu et al. (2019) generated virtual species at three levels of prevalence (the proportion of sites occupied) and then fit ten different SDM algorithms to these virtual species using a varying number of training presences and pseudoabsences. They showed that the number of pseudoabsences that maximized model performance depended on the modeling technique, species prevalence, and the number of training presences (Supplemental Table 1). In general, Latin hypercube designs are rarely used in EES research (but for recent examples, see Mellin et al. 2016, Santini et al. 2021), indicating that existing packages for experimental design in both R (Carnell 2021) and Python (Lee 2015) could be leveraged for this purpose. A caveat with response surface designs is that changing a single parameter may affect multiple features of the data in a way that is confounding to the analysis (e.g., if migration rate is increased, both the mean and the variance of the F_{ST} distribution change).

4.2.3. Simulating on real data.—Simulating signals on real data has the advantage of capturing features of the data, such as nonindependence among samples, that are difficult to model. For example, signals of selection can be simulated on landscape genomic data by generating allele frequency shifts on real data as a function of fake environments (e.g., Berg & Coop 2014), by simulating data using an observed covariance structure among populations (e.g., Gautier 2015), or by drawing from posterior estimates of demographic parameters from empirical data (e.g., Harris et al. 2018) (Supplemental Table 1). Similarly, trait values can be simulated on empirically inferred phylogenies and used to investigate methods in phylogenic inference (Figure 3c) (Rabosky & Goldberg 2015) (Supplemental Table 1). While we recommend that this technique is employed whenever possible, also note that there can be subjectivity and bias in the way signals are generated.

4.3. Phase III: Evaluation

In the evaluation phase, each method is evaluated across different simulations and possibly compared to other methods. The metrics that are used in evaluation depend on the type of data used for ground truth, the type of metric output by the method, and the goal of the method evaluation.

4.3.1. Evaluation metrics: discrimination for binary classification.—Many methods output a statistic that represents the strength of evidence for a specific hypothesis (the signal, e.g., P value, Bayes factor, or similar probability). Hypothesis testing is often based on a binary classification (e.g., null hypothesis rejected or not; comparison of one hypothesis to another, as in Bayes factors; species present or absent, as in an SDM). In this case, we wish to evaluate the performance of the method in its ability to discriminate between true positives (which should have the most extreme signals) and true negatives (which should have the least extreme signals). Although true positive rates, false positive rates, and false discovery rates are commonly used to evaluate binary classifiers, these metrics depend on arbitrary thresholds. For this reason, the performance of methods is best compared using threshold-independent metrics, most notably the area under the precision-recall curve (AUC-PR) and the area under the receiver operating characteristic curve (AUC-

ROC) (Davis & Goadrich 2006). AUC-PR is based on a plot of precision (the proportion of positive hits that are true positives) on the y-axis versus true positive rate (sensitivity or recall) on the x-axis. AUC-ROC is based on a plot of true positive rate (recall) on the y-axis versus false positive rate on the x-axis.

We illustrate the differences between AUC-PR and AUC-ROC in Figure 4, which compares four hypothetical methods (the Perfect Method, Good Method A, Good Method B, and the Poor Method) applied to an imbalanced data set with 10 positive cases and 90 negative cases (e.g., 10 loci affected by selection and 90 neutral loci in a genomics data set, or 10 species presences and 90 species absences in an SDM). Note that here we use positive cases and negative cases to refer to the true state—not the test result. In both analogies, outcomes with the highest signals output by a method (e.g., top hits) are inferred to be positive cases. The Perfect Method correctly ranks all 10 positive cases as having the highest signals and has an AUC-PR and AUC-ROC of 1 (Figure 4). In contrast, the Poor Method randomly ranks positive cases among negative cases and has an AUC-PR of 0.1 and an AUC-ROC of 0.55 (an AUC-ROC of 0.5 is expected by chance) (Figure 4).

Which metric to use depends on the characteristics of the data being tested and the relative importance of different types of errors on model evaluation. For imbalanced data, namely few positive cases and many negative cases (as in Figure 4), AUC-PR can better capture method performance than AUC-ROC. This is because AUC-ROC incorporates false positive rates (i.e., negative cases that are falsely inferred to be positive cases), which can remain comparatively low when negative cases far outnumber positive cases (Fawcett 2004, Saito & Rehmsmeier 2015, Sofaer et al. 2019). Although AUC-ROC remains widely used in SDM, simulation studies have demonstrated how AUC-ROC can be inflated when analyses incorporate a large number of pseudoabsences or background points distributed beyond the species range; this led to some methods outperforming others solely due to different numbers of background points (Jiménez-Valverde 2012, Sofaer et al. 2019). Sofaer et al. (2019) used simulations to show that AUC-PR can be a useful complement to AUC-ROC when evaluating model fit using unbalanced data, but suggested that no single evaluation metric can fully characterize method performance (Supplemental Table 1).

The advantage of the AUC-ROC score is that it more accurately reflects the costs of validation compared to AUC-PR. In the analogy of a genome scan in which each hit would be validated with gene editing, the locus with the highest signal (top hit) would be tested first, followed by the next highest signal, and so on. Ideally, one would want to discover as many true positives (e.g., loci that affect the trait) as possible with the minimum number of validation tests. In Figure 4, Good Method A returned five of the ten loci under selection in the top six hits, but the remaining five were interspersed with neutral loci. Good Method B returned eight of the ten loci under selection in the top 20 hits, but only one of them was in the top five hits. Good Method A has a higher AUC-PR than Good Method B, which reflects that more of the top hits are true positives for Good Method A compared to Good Method B (Figure 4a). However, Good Method B has a higher AUC-ROC than Good Method A, which reflects that a greater proportion of true positives would be discovered with a lower number of total validation tests for Good Method B compared to Good Method A (Figure 4b; for calculations, see Supplemental Appendix 1).

Our literature review of methods evaluations revealed that the application of AUC metrics in EES was highly variable. In the evaluation of genome scans, method performance was based more often on arbitrary thresholds of true positive rates, false positive rates, and false discovery rates (Supplemental Table 1). In SDM evaluations, AUC-ROC curves have been widely applied (despite underlying issues for imbalanced data, as described above), but AUC-PR was rarely used. When relevant, requiring both AUC-PR and AUC-ROC for publication would increase the rigor of method evaluations.

4.3.2. Evaluation metrics: accuracy.—Accuracy is how close a prediction from a method is to the true value. If the goal of evaluation is to determine the accuracy of a prediction, the approach depends on the type of data that is used as ground truth and the type of prediction made by the method. If the ground-truth data are categorical and the method prediction is also categorical, then classification accuracy is typically calculated as the proportion of ground-truth classes that fall into each predicted class (Figure 5a). For example, machine learning has been used in genomic research to predict the type of positive selection that affects different regions of the genome. Schrider & Kern (2016) trained a machine learning algorithm to classify the genome into hard sweeps, soft sweeps, and neutral regions. However, they evaluated the algorithm with data simulated under an equilibrium demography, and when the algorithm was trained with data simulated under more realistic demographies (including bottlenecks and migration) the performance of the algorithm substantially degraded (Harris et al. 2018) (Supplemental Table 1).

If the ground-truth data are categorical and the method prediction is a probability, the accuracy of the method can be assessed in terms of whether the method correctly predicts the class probability (Pearce & Ferrier 2000). Methods that predict a continuous probability of discrete outcomes are commonly used in SDMs, which calculate the probability that a species is present in a particular habitat. In this case, metrics like the Brier score (Brier 1950) and the recently proposed validation metric applied to probabilistic predictions (VMAPP) (Chivers et al. 2014) can be used for accuracy evaluation. The Brier score is equivalent to the mean squared error and ranges from 0 to 1, where 0 indicates complete agreement between observation and prediction, and 1 indicates complete disagreement (Figure 5b), whereas VMAPP is a goodness-of-fit metric that measures the magnitude and direction of bias and how bias changes across the prediction range. In contrast, metrics of discrimination like AUC-ROC and AUC-PR quantify the ability of the model to correctly distinguish between classes (i.e., for SDMs, the extent to which predicted probabilities for occupied sites are higher than those for unoccupied sites), regardless of the accuracy of the predicted class probability. Given the different insights that assessments of discrimination capacity and accuracy provide, model evaluation ideally should consider both of these components of predictive performance (Norberg et al. 2019, Pearce & Ferrier 2000). For example, Maguire et al. (2016) used both the AUC-ROC and Brier score to evaluate different strategies (single-species versus community-level modeling) for predicting shifts in the distributions of species in response to climate change.

If the ground-truth data are numerical and the method prediction is also numerical, the method can be assessed in terms of the accuracy and precision of the estimation. Evaluation metrics such as mean absolute error and root mean square error (RMSE) (Figure 5c)

increase as the prediction becomes more biased and imprecise but do not reflect the direction of bias (for examples and R code, see Supplemental Appendix 2). For example, Flagel et al. (2019) simulated DNA sequences to evaluate whether convolutional neural networks could be trained to estimate population genetic parameters from alignment images. They used RMSE to compare the algorithm's estimate of historical population recombination rate to the true value (Flagel et al. 2019). Evaluation metrics such as mean error or percent error (mean error scaled by the true value) increase as the prediction becomes more biased and also reflect the direction of bias but are unaffected by dispersion (Supplemental Appendix 2). For instance, in an evaluation of methods for building extremely large phylogenies, Beaulieu & O'Meara (2018) used percent error in branch-length estimates to show that branch-length estimation was robust in the face of missing data, except for very short branches.

Sometimes when both the ground-truth data and prediction are numerical, the method prediction is not in the same units as the ground-truth data. For example, methods that are used to predict the degree of maladaptation when genotypes are moved to a new environment could be based on genetic distance, environmental distance, or both (Láruson et al. 2022, Rellstab et al. 2021). In this case, the units of the ground-truth data is the change in fitness of the genotype when moved to a new environment, but the units of the method prediction depend on what model was used. Thus, RMSE would be inappropriate as an evaluation metric because it assumes the method prediction and ground-truth data are both in the same units. Instead, the correlation between each predicted degree of maladaptation and fitness (the ground truth from simulated data) can be used to compare the performance of methods (for an example, see Láruson et al. 2022). It is important to keep in mind that the use of correlation as an evaluation metric is limited because it measures the extent to which two variables are linearly related (e.g., a measure of precision) and does not capture other aspects of the evaluation such as bias or different rates of change.

A number of other summary statistics, such as sum of absolute log-ratios, log-modulus, and cross-entropy, can also be used to summarize model fit to the evaluation data (these metrics are reviewed in further detail in Weber et al. 2019).

4.3.3. Evaluation metrics: model comparison.—

Many methods require users to choose among models as part of the analysis [e.g., in phylogenetic inference, a DNA model is often chosen (Lanfear et al. 2017)]. This is a crucial step, since inference using incorrect models can lead to systematic errors (Cunningham et al. 1998). Akaike information criterion (AIC), Bayesian information criterion (BIC), Bayes factors, and likelihood ratio tests are the most commonly applied metrics to choose among competing models. AIC and BIC have become increasingly common in EES partly because these metrics do not require models to be nested. When comparing two models with these metrics, general rules are typically applied where differences greater than 2.5 (AIC) or 2.0 (BIC) indicate at least some support for the model with the lower score (Anderson & Burnham 2004). Alternatively, Akaike model weights have been used to provide relative support for competing models. For instance, in an ancestral state estimate of ant colony traits, Akaike model weights were used to pick a best model for inference (Borowiec et al. 2021). Though less common, Akaike model weights can also be used with model averaging to produce parameter estimates that

do not rely on any single model (Anderson & Burnham 2004). This approach was recently illustrated for analyses of genetic architecture (Blackmon & Demuth 2016).

Model averaging:

combining multiple models to make forecasts or perform inference; often models are weighted based on explanatory power

4.3.4. Completing the evaluation process.—

The evaluation process is complete when the domain of applicability and domain of inference for a method have been well defined. Critically, these domains can be well defined only if there are evaluation results for data sets both within and outside of the domains.

4.4. Phase IV: Validation

Validation is the process of assessing whether or not quantities of interest (validation criteria) are within an a priori specified validation tolerance for a specific application. The goal of validation is to determine how well the method performs for a specific application and involves two steps. In the first step, the method is parameterized by one set of data, and then in the second step, the method is tested against an independent set of empirical holdout data from the specific application (Figure 1). While method validation is not a focus of this review, we cover it briefly because the term is often used interchangeably with evaluation in the EES literature. Principles in method validation—and its differences from evaluation—have been widely developed in the engineering and physical sciences (AIAA 1998, Natl. Res. Counc. 2012, Thacker et al. 2004) but rarely applied to EES. Method validation is especially crucial when models are to be used to inform decision making, as is often the case with SDMs. The validation process determines whether there is sufficient evidence that the model is accurate for its intended use—it cannot prove that a model is correct and accurate for all possible scenarios (Thacker et al. 2004).

Validation criteria:

quantities of interest that are used to assess the accuracy of a method based on its intended use

Validation tolerance:

the a priori required level of accuracy for validation criteria

Holdout data:

data not used in training an algorithm that provides a final estimate of method performance; also called testing data

The first step of the validation process is to define the intended use of the method, the holdout data that will be used in validation, the validation criteria that will be used to

determine performance, and the level of performance needed (or validation tolerance) for that intended use (Figure 1). This last step should include consultation with stakeholders and others with subject-matter expertise. By necessity, the validation process should be closely integrated with empirical research within the proposed intended use (Guillera-Arroita et al. 2015). Careful a priori consideration of validation tolerances and preregistration of validation plans and experiments (Nosek et al. 2018) provide extra rigor for a validation study.

4.4.1. Completing the validation process.—The method is considered suitable for its intended use when acceptable agreement between the holdout data and method results is achieved (Thacker et al. 2004). Care should be taken to conduct the validation on the holdout data only once to avoid circularity by fine-tuning the model to fit the holdout data. Ideally, multiple data sets and metrics are used in determining acceptable agreement, and multiobjective optimization (Miettinen 2012) may prove useful in this determination. Even complex models are often only a rough approximation to reality: The validation process determines if they are close enough to be useful (Box 1979). Although it may be possible to prove a method is invalid for a specific situation, the generic term valid model does not make sense (Natl. Res. Counc. 2012, Oreskes et al. 1994). There is, at most, a body of evidence that can be presented to suggest that the model produces results that are consistent with reality (Natl. Res. Counc. 2012).

Acceptable agreement:

a prespecified tolerance within which holdout data are determined to validate a method

4.4.2. The role of simulation tests in validation.—Simulations are probably most commonly used to invalidate specific methods for specific applications. However, phenomena in EES that operate across large spatial and/or temporal extents, such as species diversification, demographic inference, or forecasts of range shifts, cannot be directly validated (Oreskes et al. 1994) and therefore necessitate the use of simulations for quasi-validation. For example, posterior predictive simulations for a DNA evolution model can play a key role in determining whether the model selected is adequate for the analyzed data (Gelman et al. 1996, Höhna et al. 2018). With this approach, model parameters are estimated from the empirical data and used to generate thousands of simulated data sets. A range of summary statistics are then chosen and calculated for both the empirical and simulated data sets. If a summary statistic for the empirical data is an outlier relative to the distribution of the statistic calculated for the simulated data, the model fails to capture some important aspect(s) of the actual biological process generating the data (e.g., acceptable agreement is not reached). By examining a range of summary statistics, this approach elucidates when models work, how models fail, and how to design more appropriate models for future analysis (Pennell et al. 2015, Rice & Mayrose 2021). Note, however, that using simulations in this way for method validation does not necessarily protect against cases of equifinality, and potential for equifinality should still be addressed as part of the larger validation process.

5. BEST PRACTICES

In contrast to research on clinical trial design, there has been little research on how to minimize cognitive bias and conflicts of interest in the development of statistical methods (Boulesteix et al. 2018). Here, we review best practices and cautionary principles for method evaluation and validation.

5.1. Use Ground-Truth Benchmark Data Sets, Including Adversarial Examples That Can Refine the Domains of Applicability and Inference

In EES it is common for developers to create custom simulations to evaluate a new method. As a result, methods are not compared on a common ground. In machine learning and image recognition, advancements have been achieved using standard benchmark data sets. For example, the MNIST database of handwritten digits (LeCun et al. 1998) allowed for advancements in computational image recognition. Properly designed benchmarks that include adversarial data sets within and outside of a method's domain of applicability and inference could drive similar advances in EES. While the development of benchmarks in EES is growing and now includes a standard library of population genetic models (Adrion et al. 2020) and variation benchmarks (Sarkar et al. 2020, Thompson et al. 1999), the standardized use of benchmarks for methods development and evaluation is still rare in EES.

Benchmark data sets:

simulated or empirical data sets for which the underlying truth is known and that are used as a standard for evaluating methods

5.2. Conduct the Comparison Neutrally with Double Blinding

Anecdotal evidence supports the idea that differences in user expertise can lead to a variety of outcomes despite the same method being applied to the same data set (Gilbert et al. 2012, Lotterhos et al. 2016). A neutral comparison (sensu Boulesteix et al. 2017) is a comparison that is designed such that user expertise does not favor certain methods. Thus, evaluators should report the steps they took to apply an equivalent level of parameter tuning for each method, such that each method is compared under its best performance. A best practice, therefore, is for the user to simulate data sets blind to the methods and for the evaluator to tune the method blind to the ground truth in the simulated data (analogous to double-blind clinical trials). Among the 37 method evaluations in our literature review, none of them reported double blinding, illustrating that this should be encouraged by journals and reviewers.

5.3. Separate Holdout Data and Training Data, if Applicable

Some methods require data to train an algorithm. In this case, it is necessary to partition each simulated data set into a subset that serves as training data for parameterization (which may include cross-validation) and a subset that serves as holdout data for evaluation. Evaluators should avoid the circularity of using the same data for training and evaluation, which can inflate the performance statistics of the method (Grimm et al. 2015, Molnar 2021). Take, for example, simulated data sets of virtual species that are used to test SDMs:

a sample of observations from each simulation should be used to train the SDM (training data), and a different subset of observations from the same simulation should be used to evaluate whether the algorithm correctly predicts the species distribution (holdout data) (e.g., Meynard & Quinn 2007, Norberg et al. 2019). How best to partition the data into training and holdout subsets [e.g., using spatial blocking versus random selection (Norberg et al. 2019) or leave-one-population-out selection (Rellstab et al. 2021, Roberts et al. 2017)] is an active area of research for different methods. As a cautionary principle, cross-validation is not always the same as method validation because it is only a summary of model fit, which can still suffer from equifinality; furthermore, model accuracy is only a small component of the validation process (Lotterhos et al. 2018).

5.4. Consider Biological Relevance When Assessing Method Performance

Biological relevance implies a biological effect of interest that is considered important based on expert judgment (EFSA Scientific Committee 2011, Martínez-Abraín 2008, Whitlock & Schluter 2020). Evaluations can assess the biological relevance of method outputs in two ways, and well-designed evaluation studies can help elucidate cases when the method output can be used to make robust inference.

Biological relevance:

an effect considered by expert judgment to be important and meaningful

The first way evaluations can assess the biological relevance of method outputs is to test whether a causal pattern is inferred when none exists. Simulation studies have shown this is a significant problem for methods that are sensitive to unrelated information, sometimes due to correlation in the data. In a particularly funny example, Fourcade et al. (2018) showed that classical paintings (as in works of art) georeferenced to the extent of European geographic space could successfully be used as predictors in SDMs, sometimes outperforming real environmental variables, despite having no biological relevance. Similarly, Robosky & Goldberg (2015) showed that the numbers of letters in a species name can predict diversification rates. Where applicable, including these types of nonsensical data sets in an evaluation study can highlight the limitations of methods (or lack thereof) and/or the risk of assuming correlation is the same as causation.

A second way evaluations can assess the biological relevance of method predictions is to evaluate the relationship between method performance and effect size. The effect size refers to the magnitude of a response in the data. Ideally, a method has good performance across a range of effect sizes, including effect sizes of zero (e.g., a null hypothesis). For example, methods that detect correlated trait evolution for discrete characters can infer causality between traits when it does not exist (Maddison & FitzJohn 2015). In general, it is as important for methods to perform well when the effect size is so small that it has little biological interest as it is for methods to perform well when effect sizes are large. Evaluation studies that simulate data across a range of effect sizes, with different amounts of intrinsic noise, help users understand method performance across this continuum.

In general, biological relevance should be considered along with performance metrics when interpreting the results of an evaluation. It is possible to calculate the theoretical maximum performance of an ideal method from simulated data, although to our knowledge this is not practiced. For example, if the best method in an evaluation study explains 10% percent of variation in the data when an ideal method could explain 50%, evaluators should assess whether this is enough variation explained for the method to be useful in its intended application. For some applications, explaining 10% of variance may be an improvement, while for others it may not be explaining enough of the biology to be useful.

5.5. Summarizing Variation in the Simulations

Evaluators should think carefully about how to present variation in method performance among data sets simulated with the same parameters that vary due to the amount of intrinsic error simulated. On one hand, for a simulation study of a method that shows a mismatch between the estimate and reality, it is relevant to show that enough simulations were done to ensure robust inference. The standard error of simulation results is an important measure of whether enough replicate simulations were run. On the other hand, standard error can be artificially decreased by running more replicates, which could make it seem that the variability in performance is much lower than it is. For this reason, method performance should be presented in a way that captures the variability in the simulations for each set of parameters (e.g., standard deviation), in addition to justification that enough replicates were done (e.g., standard error).

5.6. Additional Best Practices

Other best practices include comparing multiple methods on the same data sets, making the data set and code publicly available (Meynard et al. 2019), and assessing the assumptions of each method prior to evaluating them. Of the 37 evaluation studies we reviewed, 34 of them compared multiple methods, and 24 of them made the data and/or code publicly available (Supplemental Table 1). These studies, however, rarely stated whether the assumptions of each method were assessed prior to evaluation (Supplemental Table 1).

6. FUTURE DIRECTIONS AND CHALLENGES

6.1. Establishing Domains of Applicability and Inference for Decision-Making

The domains of applicability and inference are helpful for communicating the conditions for which methods can be trusted. Establishing these domains will continue to challenge EES researchers, because even accurately modeled physical systems lack a formal mathematical framework for defining such a domain (Natl. Res. Counc. 2012). Thus, the development of logical principles to establish these domains is an important area for future research.

Method evaluation and validation can assist decision makers in making informed choices about an intended application. While others have reviewed how models should be used as scientific support for decision-making (reviewed in Natl. Res. Counc. 2007, Starfield 1997), less attention has been paid to how error rates and/or uncertainty from method evaluation should be incorporated into the decision-making process. Decision makers must weigh the potential benefits of applying a method to the potential risks of the method

returning an inaccurate result. An important area of future research is therefore how to design the evaluation and validation process such that the results can be incorporated into decision theory and multiobjective optimization. Decision theory offers a formal framework for determining the optimal choice in the face of these costs and benefits, given their relative probabilities, and includes frameworks for making decisions under ignorance (Peterson 2009, Pratt et al. 2008). Multiobjective optimization involves the development of mathematical optimization functions for determining the optimal decision in the presence of trade-offs between two or more conflicting objectives (Miettinen 2012).

6.2. Requiring Principled Method Evaluation for Publication

It is unrealistic to believe that, during method development, one could foresee every way a method could be (mis)applied. However, some authors have called for systematic evaluation of new methods using simulations before those methods are applied to real data (Austin et al. 2006). Developing robust methods involves practicing evaluation and validation as a cyclical process (Figure 1) until the domains of applicability and inference are fully understood. With increased integration between method development, evaluation, empirical research, and validation, we can reduce the hidden costs associated with methods being developed first and broken later.

Achieving this increased integration will continue to challenge EES until standardized criteria for method evaluation are widely used. The current lack of standardized criteria leads to high variability in the rigor with which methods are evaluated. What constitutes a comprehensive set of benchmarks for groups of methods and how to conduct principled evaluation on those benchmarks required for publication are important areas for future research. Comparisons of the performance of different methods on benchmark data sets could be made more efficient through a shared database system, but funding for developing such a system is scarce in EES. Developing a culture of principled method evaluation will continue to be a major challenge in EES until it is supported and required by funding agencies, reviewers, and journals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We would like to thank Thais Bittar, Remy Gatins, and Editorial Committee Member Michael Whitlock for helpful comments that improved the quality of this manuscript. The authors were supported by funding from the National Science Foundation [grants 1655701 (to K.E.L. and M.C.F.), 2043905 (to K.E.L.), 1656099 (to M.C.F.), and 1856450 (to M.C.F.)] and National Institutes of Health [grant R35GM138098 (to H.B.)].

DATA AVAILABILITY

The 37 evaluation studies reviewed for this article are listed in Supplemental Table 1. The R code used to produce the AUC plots is included as Supplemental Appendix 1. The R code used to explore the effects of bias and error on metrics is in Supplemental Appendix 2.

LITERATURE CITED

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, et al. 2020. A community-maintained standard library of population genetic models. eLife 9:54967

AIAA (Am. Inst. Aeronaut. Astronaut.). 1998. Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (AIAA G-077–1998(2002)). Reston, VA: Am. Inst. Aeronaut. Astronaut.

Anderson D, Burnham K. 2004. Model Selection and Multi-Model Inference. New York: Springer. 2nd Ed.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol 22(11):3179–90 [PubMed: 23551379]

Austin MP, Belbin L, Meyers JA, Doherty LM. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. Ecol. Model 199(2):197–216Describes the use of virtual species for testing ecological theory related to modeling plant distributions.

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol 3(2):327–38

Bay RA, Harrigan RJ, Underwood VL, Gibbs HL, Smith TB, Ruegg K. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science 359(6371):83–86 [PubMed: 29302012]

Beaulieu JM, O'Meara BC. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol 65(4):583–601 [PubMed: 27016728]

Beaulieu JM, O'Meara BC. 2018. Can we build it? Yes we can, but should we use it? Assessing the quality and value of a very large phylogeny of campanulid angiosperms. Am. J. Bot 105(3):417–32 [PubMed: 29746717]

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. PLOS Genet 10(8):e1004412 [PubMed: 25102153]

Bergstrom CT, West JD. 2021. Calling Bullshit: The Art of Skepticism in a Data-Driven World New York:Random House

Blackmon H, Demuth JP. 2016. An information-theoretic approach to estimating the composite genetic effects contributing to variation among generation means: moving beyond the joint-scaling test for line cross analysis. Evolution 70(2):420–32 [PubMed: 26704183]

Blanquart F, Kaltz O, Nuismer SL, Gandon S. 2013. A practical guide to measuring local adaptation. Ecol. Lett16(9):1195–205 [PubMed: 23848550]

Borowiec ML, Cover SP, Rabeling C. 2021. The evolution of social parasitism in *Formica* ants revealed by a global phylogeny. PNAS 118(38):e2026029118 [PubMed: 34535549]

Boulesteix A-L, Binder H, Abrahamowicz M, Sauerbrei W, Simul. Panel STRATOS Initiat. 2018. On the necessity and design of studies comparing statistical methods. Biom. J 60(1):216–18 [PubMed: 29193206]

Boulesteix A-L, Wilson R, Hapfelmeier A. 2017. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. BMC Med. Res. Methodol 17(1):138 [PubMed: 28888225] Highlights how principles of clinical trial design can be applied to method evaluations.

Box GEP. 1979. Robustness in the strategy of scientific model building. In Robustness in Statistics, ed. Launer RL, Wilkinson GN, pp. 201–36. New York: Academic

Brier GW. 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev 78(1):1–3

Brown JM, Hedtke SM, Lemmon AR, Lemmon EM. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol 59(2):145–61 [PubMed: 20525627]

Capblancq T, Luu K, Blum MGB, Bazin E. 2018. Evaluation of redundancy analysis to identify signatures of local adaptation. Mol. Ecol. Resour 18(6):1223–33 [PubMed: 29802785]

Carnell R 2021. lhs: Latin Hypercube Samples Stat. Softw. Package, CRAN-R Proj. https://CRAN.R-project.org/package=lhs

Chivers C, Leung B, Yan ND. 2014. Validation and calibration of probabilistic predictions in ecology. Methods Ecol. Evol 5(10):1023–32

Cottingham KL, Lennon JT, Brown BL. 2005. Knowing when to draw the line: designing more informative ecological experiments. Front. Ecol. Env 3(3):145–52

Cunningham CW, Zhu H, Hillis DM. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52:978–87 [PubMed: 28565216]

Davis J, Goadrich M. 2006. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, ed. Cohen WW, Moore A, pp. 233–40. New York: Assoc. Comput. Mach.

Dormann CF, Schymanski SJ, Cabral J, Chuine I, Graham C, et al. 2012. Correlation and process in species distribution models: bridging a dichotomy. J. Biogeogr 39(12):2119–31

EFSA Sci. Comm. 2011. Statistical significance and biological relevance. Europ. Food Safety Auth. Journal 9(9):2372

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129–51

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst 40:677–97

Fawcett T 2004. ROC graphs: notes and practical considerations for data mining researchers Tech. Rep. HPL-2003–4, HP Lab., Palo Alto, CA. https://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf

Felsenstein J 1985. Phylogenies and the comparative method. Am. Nat 125(1):1–15

Flagel L, Brandvain Y, Schrider DR. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol. Biol. Evol 36(2):220–38 [PubMed: 30517664]

Flanagan SP, Jones AG. 2018. Substantial differences in bias between single-digest and double-digest RAD-seq libraries: a case study. Mol. Ecol. Resour 18(2):264–80 [PubMed: 29120082]

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR. 2016. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Mol. Ecol 25(1):104–20 [PubMed: 26576498]

Forester BR, Lasky JR, Wagner HH, Urban DL. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Mol. Ecol 27(9):2215–33 [PubMed: 29633402]

Fourcade Y, Besnard AG, Secondi J. 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr 27(2):245–56

Freedman AH, Clamp M, Sackton TB. 2021. Error, noise and bias in de novo transcriptome assemblies. Mol.Ecol. Resour 21(1):18–29 [PubMed: 32180366]

Gautier M 2015. Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics 201(4):1555–79 [PubMed: 26482796]

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, et al. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol 22(11):3165–78 [PubMed: 23110526]

Gelman A, Meng X-L, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Stat. Sin 6(4):733–60

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, et al. 2012. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. Mol. Ecol 21(20):4925–30 [PubMed: 22998190]

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, et al. 2015. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat 36(5):513–23 [PubMed: 25684150]

Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr 24(3):276–92

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):R32 [PubMed: 19327155]

Harris RB, Sackman A, Jensen JD. 2018. On the unfounded enthusiasm for soft selective sweeps II: examining recent evidence from humans, flies, and viruses. PLOS Genet 14(12):e1007859 [PubMed: 30592709]

Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. 2021. Natural adversarial examples. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15257–66. Piscataway, NJ: IEEE

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, et al. 2016. Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. Am. Nat 188(4):379–97 [PubMed: 27622873]

Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. 2018. P3: phylogenetic posterior prediction in RevBayes. Mol. Biol. Evol 35(4):1028–34 [PubMed: 29136211] Demonstrates the use of posterior predictive simulations to evaluate models and validate their use with specific data sets.

Hölzer M, Marz M. 2019. *De novo* transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8(5):giz039

Inouye BD. 2001. Response surface experimental designs for investigating interspecific competition. Ecology 82(10):2696–706

Jiménez-Valverde A 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecol. Biogeogr 21:498–507

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol 34(3):772–73 [PubMed: 28013191]

Láruson ÁJ, Fitzpatrick MC, Keller SR, Haller BC, Lotterhos KE. 2022. Seeing the forest for the trees: assessing genetic offset predictions from gradient forest. Evolutionary Appl 15(3)403–16

LeCun Y, Cortes C, Burges CJC. 1998. The MNIST database of handwritten digits http://yann.lecun.com/exdb/mnist/

Lee A. 2015. pyDOE: the experimental design package for Python Softw. Package. https://pythonhosted.org/pyDOE/

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol 54(2):241–53 [PubMed: 16012095]

Lin CD, Bingham D, Sitter RR, Tang B. 2010. A new and flexible method for constructing designs for computer experiments. Ann. Stat 38(3):1460–77

Liu C, Newell G, White M. 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. Ecography 42(3):535–48

Lobo JM, Tognelli MF. 2011. Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. J. Nat. Conserv 19(1):1–7

Lotterhos KE. 2019. The effect of neutral recombination variation on genome scans for selection. G3 Genes Genomes Genet 9(6):1851–67

Lotterhos KE, François O, Blum MGB. 2016. Not just methods: User expertise explains the variability of outcomes of genome-wide studies. bioRxiv 055046. 10.1101/055046

Lotterhos KE, Moore JH, Stapleton AE. 2018. Analysis validation has been neglected in the Age of Reproducibility. PLOS Biol 16(12):e3000070 [PubMed: 30532167]

Lotterhos KE, Whitlock MC. 2014. Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. Mol. Ecol 23(9):2178–92 [PubMed: 24655127]

Lotterhos KE, Whitlock MC. 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol. Ecol 24(5):1031–46 [PubMed: 25648189]

Louca S, Pennell MW. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580(7804):502–5 [PubMed: 32322065]

Lucas TCD. 2020. A translucent box: interpretable machine learning in ecology. Ecol. Monogr 90(4):e01422

Luu K, Bazin E, Blum MGB. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour 17(1):67–77 [PubMed: 27601374]

Maddison WP, FitzJohn RG. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst. Biol 64(1):127–36 [PubMed: 25209222]

Maguire KC, Nieto-Lugilde D, Blois JL, Fitzpatrick MC, Williams JW, et al. 2016. Controlled comparison of species- and community-level models across novel climates and communities. Proc. R Soc. B 283(1826):20152817

Martínez-Abraín A 2008. Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. Acta Oecol 34(1):9–11

McKay MD, Beckman RJ, Conover WJ. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(2):239–45

Mellin C, Lurgi M, Matthews S, MacNeil MA, Caley MJ, et al. 2016. Forecasting marine invasions under climate change: Biotic interactions and demographic processes matter. Biol. Conserv 204:459–67

Meynard CN, Leroy B, Kaplan DM. 2019. Testing methods in species distribution modelling using virtual species: What have we learnt and what are we missing? Ecography 42(12):2021–36Review of testing SDM methods and methodological decisions using virtual species.

Meynard CN, Quinn JF. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. J. Biogeogr 34(8):1455–69

Miettinen K 2012. Nonlinear Multiobjective Optimization New York: Springer Sci. Bus. Media

Molnar C 2021. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable https://christophm.github.io/interpretable-ml-book/index.html

Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017. Uncovering the horseshoe effect in microbial analyses. mSystems 2(1):e00166–16 [PubMed: 28251186]

Narum SR, Hess JE. 2011. Comparison of F_{ST} outlier tests for SNP loci under selection. Mol. Ecol. Resour11(1):184–94 [PubMed: 21429174]

Natl. Res. Counc. 2007. Models in Environmental Regulatory Decision Making Washington, DC: Natl. Acad. Press

Natl. Res. Counc. 2012. Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification Washington, DC: Natl. Acad. PressReview of principles in model evaluation for physics and engineering.

Ng J, Smith SD. 2014. How traits shape trees: new approaches for detecting character state-dependent lineage diversification. J. Evol. Biol 27(10):2035–45 [PubMed: 25066512]

Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecol. Monogr 89(3):e01370

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. PNAS 115(11):2600–6 [PubMed: 29531091]

Oreskes N, Shrader-Frechette K, Belitz K. 1994. Verification, validation, and confirmation of numerical models in the Earth sciences. Science 263(5147):641–46 [PubMed: 17747657]

Otto SP, Day T. 2011. A Biologist's Guide to Mathematical Modeling in Ecology and Evolution Princeton, NJ: Princeton Univ. Press

Ovaskainen O, Tikhonov G, Norberg A, Blanchet FG, Duan L, et al. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett 20(5):561–76 [PubMed: 28317296]

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model 133:225–45

Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. 2015. Model adequacy and the macroevolution of angiosperm functional traits. Am. Nat 186(2):E33–50 [PubMed: 26655160] Shows how different summary statistics can highlight different aspects of model performance.

Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A. 2010. Comparing three different methods to detect selective loci using dominant markers. J. Evol. Biol 23(10):2267–76 [PubMed: 20796133]

Peterson M 2009. An Introduction to Decision Theory Cambridge, UK: Cambridge Univ. Press. 1st ed.

Phillips SJ, Dudík M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2):161–75

Pratt J, Raiffa H, Schlaifer R. 2008. Introduction to Statistical Decision Theory Cambridge, MA: MIT Press

Qiao H, Soberón J, Peterson AT. 2015. No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. Methods Ecol. Evol 6(10):1126–36

Rabosky DL, Goldberg EE. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Syst. Biol 64(2):340–55 [PubMed: 25601943] Illustrates the process and advantages of simulating data in conjunction with real data.

Ranjan P, Bingham D, Michailidis G. 2008. Sequential experiment design for contour estimation from complex computer codes. Technometrics 50(4):527–41

Rellstab C, Dauphin B, Exposito-Alonso M. 2021. Prospects and limitations of genomic offset in conservation management. Evol. Appl 14(5):1202–12 [PubMed: 34025760]

Revell LJ. 2014. Ancestral character estimation under the threshold model from quantitative genetics. Evolution 68(3):743–59 [PubMed: 24152239]

Rice A, Mayrose I. 2021. Model adequacy tests for probabilistic models of chromosome-number evolution. New Phytol 229(6):3602–13 [PubMed: 33226654]

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–29

Ruegg K, Bay RA, Anderson EC, Saracco JF, Harrigan RJ, et al. 2018. Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. Ecol. Lett 21(7):1085–96 [PubMed: 29745027]

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3):e0118432 [PubMed: 25738806]

Santini L, Benítez-López A, Maiorano L, Čengić M, Huijbregts MAJ. 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib 27(6):1035–50

Sarkar A, Yang Y, Vihinen M. 2020. Variation benchmark datasets: update, criteria, quality and applications. Database 2020:baz117

Schrider DR, Kern AD. 2016. S/HIC: robust identification of soft and hard sweeps using machine learning. PLOS Genet 12(3):e1005928 [PubMed: 26977894]

Schrider DR, Kern AD. 2018. Supervised machine learning for population genetics: a new paradigm. Trends Genet 34(4):301–12 [PubMed: 29331490]

Sofaer HR, Hoeting JA, Jarnevich CS. 2019. The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol. Evol 10(4):565–77

Starfield AM. 1997. A pragmatic approach to modeling for wildlife management. J. Wildl. Manag 61(2):261–70

Tang B 1993. Orthogonal array-based Latin hypercubes. J. Am. Stat. Assoc 88(424):1392–97

Thacker BH, Doebling SW, Hemez FM, Anderson MC, Pepin JE, Rodriguez EA. 2004. Concepts of model verification and validation LA-14167-MS. Los Alamos Natl. Lab., Los Alamos, NM. https://www.osti.gov/servlets/purl/835920/Review of concepts in model evaluation for physics and engineering.

Thompson JD, Plewniak F, Poch O. 1999. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15(1):87–88 [PubMed: 10068696]

Weber LM, Saelens W, Cannoodt R, Soneson C, Hapfelmeier A, et al. 2019. Essential guidelines for computational method benchmarking. Genome Biol 20(1):125 [PubMed: 31221194] Review of approaches in model evaluation for genomics.

Whitlock MC, Lotterhos KE. 2015. Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. Am. Nat 186(S1):S24–36 [PubMed: 26656214]

Whitlock MC, Schluter D. 2020. The Analysis of Biological Data New York: W. H. Freeman. 3rd ed.

Zurell D, Berger U, Cabral JS, Jeltsch F, Meynard CN, et al. 2010. The virtual ecologist approach: simulating data and observers. Oikos 119(4):622–35Review of the use of simulated data to evaluate methods in ecology.

Zurell D, Jeltsch F, Dormann CF, Schröder B. 2009. Static species distribution models in dynamically changing systems: How good can predictions really be? Ecography 32(5):733–44

Zurell D, Pollock LJ, Thuiller W. 2018. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogeneous environments? Ecography 41(11):1812–19

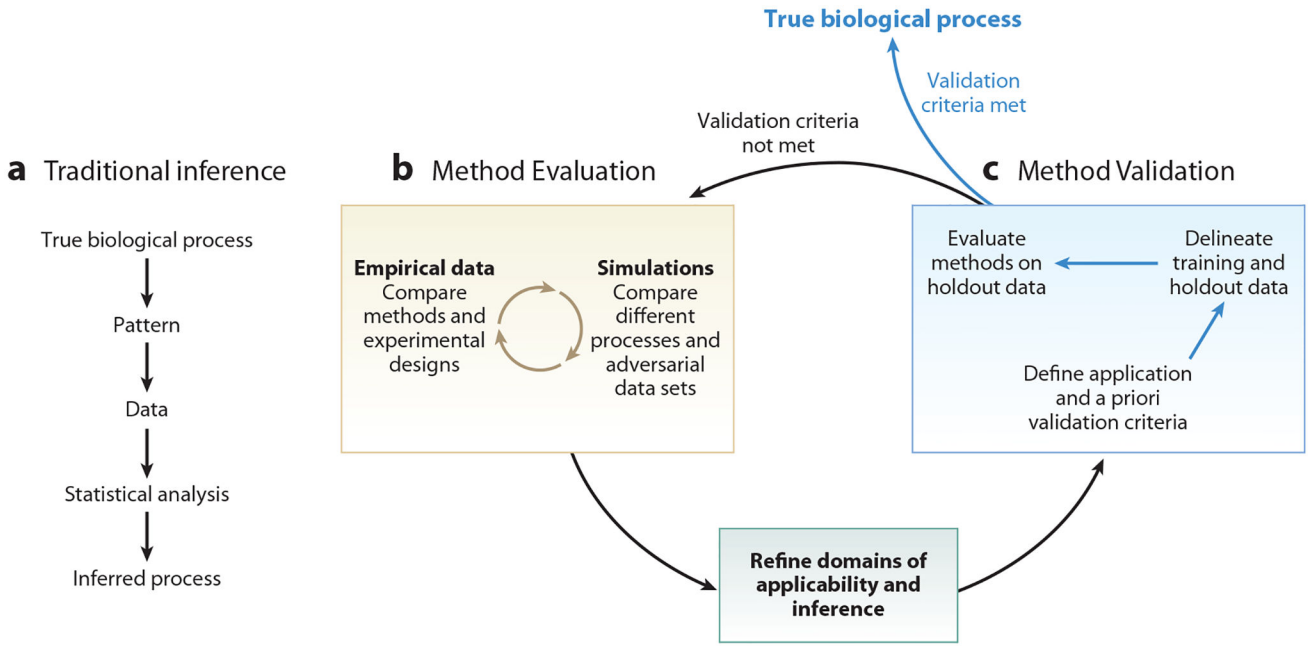

Figure 1.

An overview of method evaluation and validation in the sciences. (*a*) In traditional inference, it is not always known whether the outcome of an analysis is representative of the true underlying biological process. (*b*) Method evaluation is the process of comparing methods on data sets for which the truth is known and that include patterns of realism observed in empirical data, as well as adversarial data sets (e.g., null or challenging data sets) that help to define the domains of applicability and inference for the methods. (*c*) Method validation is the process of determining whether a method is good enough to apply to a specific situation and involves defining the application and a set of a priori validation criteria (e.g., the required level of accuracy) for that application. When the evaluation and validation criteria are met, the results add to the body of evidence that the method gives robust inference or predictions about the true underlying biological process.

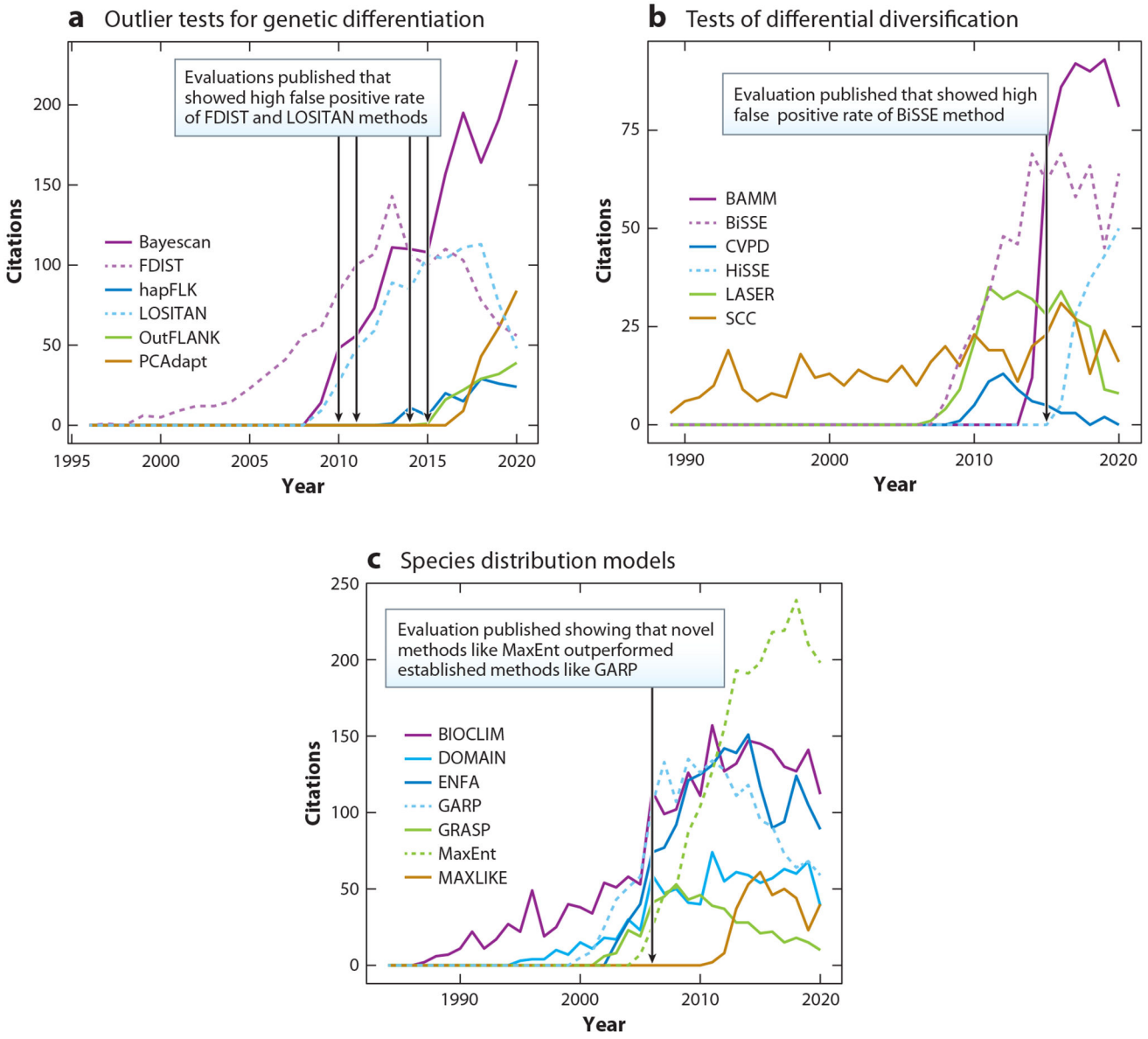

Figure 2.

The citation rates for methods in different ecology, evolution, and systematics applications, in relation to publication of the evaluations of those methods (indicated by *arrows*). The methods indicated with a dotted line showed good or poor performance in the evaluation. (*a*) F_{ST} outlier methods (Bayescan, FDIST, hapFLK, LOSITAN, OutFLANK, PCAdapt) relative to when negative evaluations of FDIST/LOSITAN were published. F_{ST} is a measure of genetic differentiation. (*b*) Diversification tests (BAMM, BiSSE, CVPD, HiSSE, LASER, SCC) relative to a negative evaluation of BiSSE. (*c*) Species distribution models (BIOCLIM, DOMAIN, ENFA, GARP, GRASP, MaxEnt, MAXLIKE) relative to a positive evaluation of MaxEnt.

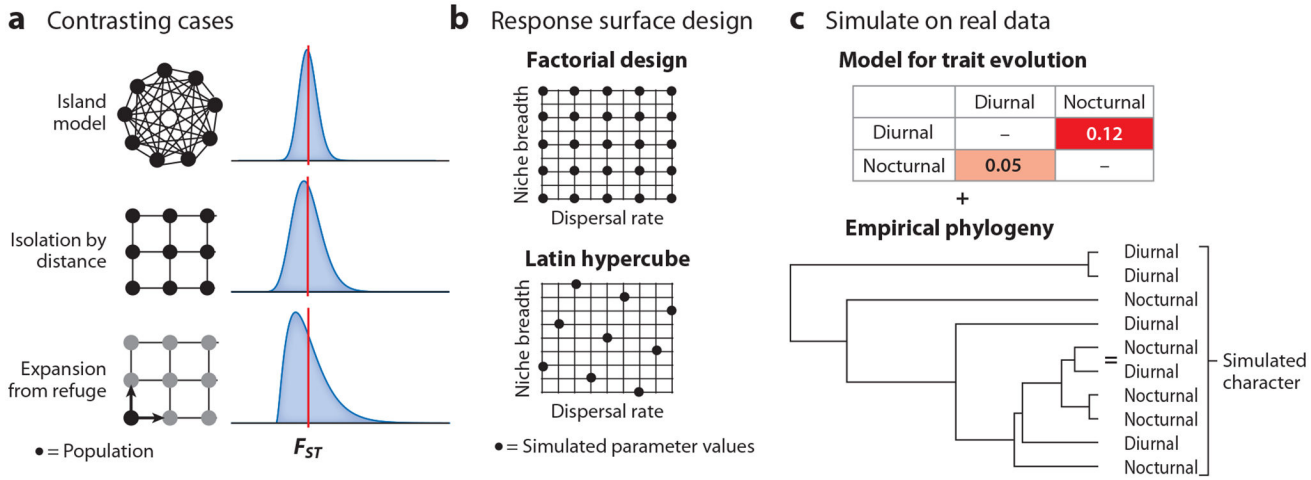

Figure 3.

Techniques that can be used to guide simulations. (*a*) Contrasting cases present different scenarios while controlling for important features of the data. In this example, the distribution of F_{ST} (genetic differentiation) is contrasted for different demographies, while the overall mean F_{ST} (*vertical red line*) is controlled for. The demography that produces each distribution of F_{ST} is visualized on the left, with populations represented by dots and paths of migration represented by lines. In the expansion from refuge (*bottom*), the population initially occupies the black location and spreads to the grey locations over time. (*b*) Response surface designs study the response variable across many levels of the explanatory variables. In this example, a method evaluation seeks to explore the performance of a species distribution model across the dispersal rate and niche breadth of a virtual species. The matrix represents the total parameter space, while the dots represent the combination of parameter values that were simulated. Simulation levels could follow a factorial design (*top*) or a Latin hypercube that requires fewer total levels to cover a similar response space (*bottom*). (*c*) Simulations can also be performed on real data. In this example, a Markov model of trait evolution is based on a matrix that describes the probability that one trait (in *rows*) mutates to another trait (in *columns*) per unit time. Here, the traits are diurnal or nocturnal behavior. This model is then simulated on an empirical phylogeny to produce a data set of simulated traits that can be used to test methods.

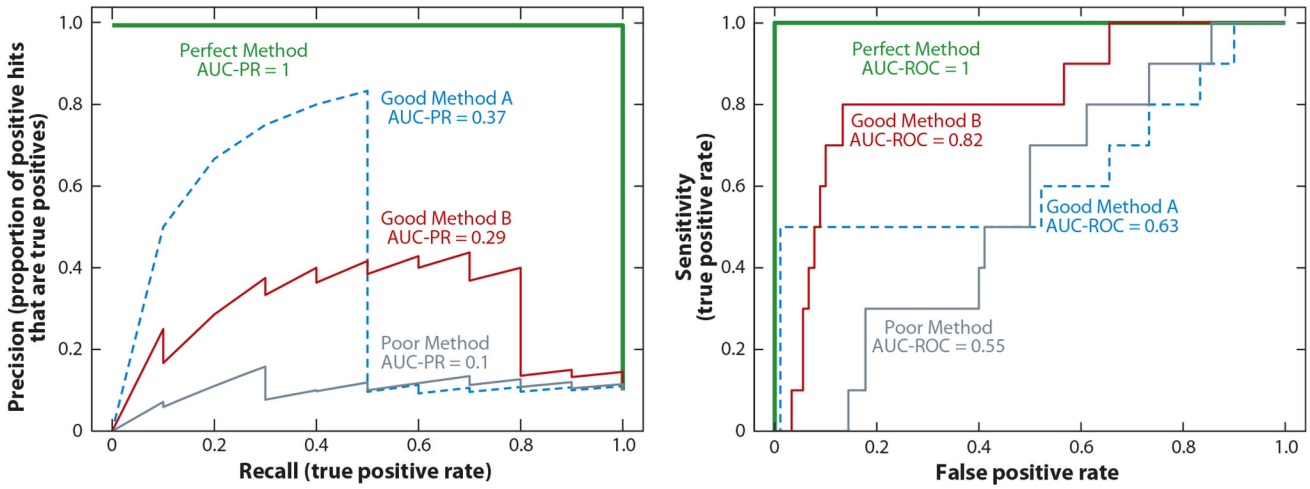

Figure 4.

Approaches for quantifying the performance of a method for binary outcomes (hypothesis test). Comparison of (*a*) AUC-PR and (*b*) AUC-ROC for different methods applied to the same data set. In this data set, there are 10 positive cases (e.g., loci under selection, species presences) and 90 negative cases (e.g., neutral loci, species absences), to which four methods are applied. Cases refer to the true state—not the test result. Note that recall (the x-axis in panel *a*) and sensitivity (the y-axis in panel *b*) are different names for the true positive rate. The true positive rate is the proportion of positive cases that the method infers are positive cases. The false positive rate is the proportion of negative cases that the method infers are positive cases. Abbreviations: AUC-PR, area under the precision-recall curve; AUC-ROC, area under the receiver operating characteristic curve.

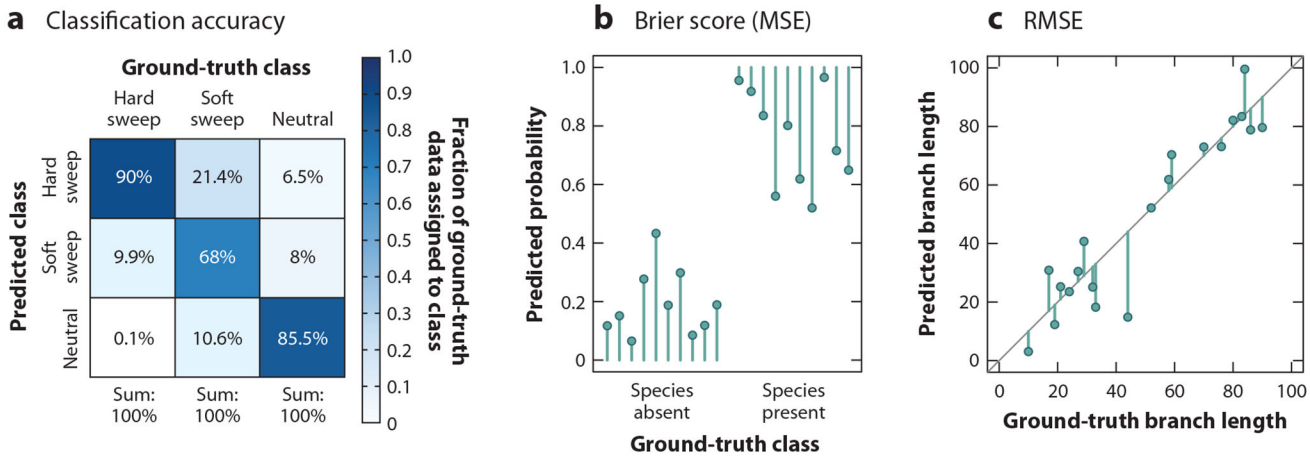

Figure 5.

Approaches for calculating prediction accuracy. (*a*) If the ground-truth data are categorical and the method prediction is also categorical, classification accuracy is calculated for each true class as the fraction of ground-truth classes that fall into each predicted class. On the diagonal are true positive rates. In this example, a method seeks to classify data corresponding to a region of the genome into hard sweep, soft sweep, and neutral classes. (*b*) If the ground-truth data are categorical and the method prediction is a probability (*dots*), the Brier score is calculated as the MSE between the probability of the ground-truth class (in this example, 0 means species absent, and 1 means species present) and the probability predicted by the method. The MSE is visually represented as squaring the lines corresponding to each dot and then taking the mean. (*c*) If the ground-truth parameter is numerical and the method predicts an estimate of the parameter (in this example, branch length on a phylogenetic tree), then the RMSE is calculated between the ground truth and the prediction. The RMSE is visually represented as squaring the lines corresponding to each dot, taking the mean, and then taking the square root. Abbreviations: MSE, mean square error; RMSE, root mean square error.

Table 1

Challenges, goals, and metrics in method evaluation

Challenge in applying method to data	Goal of method evaluation	Evaluation metrics
Does applying this method to the data give a true result in a hypothesis test?	Evaluate errors in hypothesis test Determine which method(s) have the lowest error rates Determine whether the error rates are acceptable for application	True positive rate False positive rate False discovery rate Area Under the Precision-Recall Curve Area Under the Receiver-Operator Curve
Is the prediction from this method accurate and precise?	Quantify bias and error of method prediction Evaluate whether the prediction is accurate and precise	If ground truth is categorical and method prediction is categorical: Classification accuracy If ground truth is categorical and method prediction is a probability: Brier score Validation metric applied to probabilistic predictions (VMAPP) If ground truth is continuous and method prediction is continuous: Root mean square error Accuracy/bias Precision
Did the pattern in the data result from a different process from that assumed?	Evaluate equifinality Develop an understanding of confounding factors, hidden variables, and other possible processes	Model comparison metrics: Akaike information criteria Bayes information criteria Likelihood ratio test Bayes factor Posterior probability
Were correct decisions made when designing the experiment, filtering data, and choosing a method?	Evaluate nuanced decisions Illustrate best practices in the scientific method	Any of the above metrics, depending on the context