Received: 13 October 2021 | Revised: 15 October 2021 | Accepted: 18 October 2021

DOI: 10.1111/1755-0998.13532

MOLECULAR ECOLOGY RESOURCES WILEY

EDITORIAL

Machine learning in molecular ecology

1 | INTRODUCTION

Advances in next-generation sequencing (NGS) platforms are allowing researchers to routinely collate large genome-wide data sets to address a variety of ecological questions. However, with this big data comes big analytical challenges that are increasingly addressed using machine learning (for a review, see Schrider & Kern, 2018). Machine learning is a subfield of artificial intelligence and represents a conglomeration of methods where predictive accuracy is the primary goal (e.g., Belcaid & Toonen, 2015; Breiman, 2001; Elith et al., 2008; Lucas, 2020). Machine learning assumes that the data-generating process is unknown and complex and finds the dominant patterns by learning the relationships between inputs and responses (Elith et al., 2008). Broadly, machine learning differs from other statistical approaches in two important ways. The first is that predictive performance drives model formulation rather than model selection or expert opinion, and the second is there is less emphasis on model selection (Breiman, 2001; Lucas, 2020). For these reasons, machine learning has the reputation for being less interpretable and difficult to apply rigorously (Elith et al., 2008; Lucas, 2020; Molnar, 2018). However, in parallel with the revolution of sequencing techniques, there has also been a revolution in data science in terms of predictive performance and techniques to interpret machine learning models (Fountain-Jones et al., 2019; Lucas, 2020; Molnar, 2018). There are now streamlined R and Python packages that make the robust use of algorithms from support vector machines (SVMs) to neural networks readily achievable (e.g., Abadi et al., 2015; Kuhn & Wickham, 2020, see Text Box 1 for some important machine learning terminology). Moreover, other statistical paradigms such as approximate Bayesian computation (ABC) are being applied side-by-side or within machine learning frameworks to enhance the utility of these approaches (e.g., Carlson, 2020; Raynal et al., 2019).

The ability of machine learning algorithms to build powerful predictive models that capture complex nonlinear responses with minimal statistical assumptions has been harnessed by most molecular ecology subdisciplines for decades. For example, machine learning models were developed before the turn of the millennium to classify normal or cancerous tissue based on transcription profiles (Furey et al., 2000). Not long after gradient boosting models (GBMs) were developed (e.g., Hastie et al., 2009), researchers were applying the approach to classify population genetics models based on a suite of summary statistics such as Tajima's θ_π (Lin et al., 2011). In addition, extensions of the popular random forest algorithm have been utilized in ecological genetics to untangle the drivers of climate adaptation (Fitzpatrick & Keller, 2015). Generally, however, advances in computer science and machine learning are slow to filter down to ecologists (Belcaid & Toonen, 2015; Elith & Hastie, 2008; Fountain-Jones et al., 2019), partly through unfamiliarity with these types of approaches but also because of the rapid rate of advance in the data science field.

This Special Issue aims to help expand the use of machine learning approaches and to help bring advances in data science to the toolkits of molecular ecologists. This issue comprises 17 papers grouped into four sections covering a diverse variety of molecular ecology subdisciplines. The first section covers how machine learning can be applied to make inferences about population demography. We further group these papers algorithmically with four papers utilizing random forest architecture and the remaining four using neural networks. The second section highlights how machine learning can detect signatures of selection across loci. The third section highlights how these methods can be applied to untangle the complex ecological drivers of genomic change (‘ecological genomics’) and species community dynamics. The last section explores how advances in machine learning can provide insights into species limits and contribute to biodiversity monitoring.

2 | SUMMARY OF THE SPECIAL FEATURE

2.1 | Demographic inference

2.1.1 | Random forests

Random forests (RF) are supervised machine-learning methods that perform classification and regression analyses. They were initially proposed by Ho (1995), then formally developed by Breiman (2001). In the context of demographic inference in population genomics, they are used to determine which of several demographic models best explains a given observed data set. Then, for this chosen model, RF can be used to estimate the posterior distribution for each parameter, yielding point estimates and credibility intervals for these parameters. In population genomics, RF methods are used in the context of ABC (Beaumont et al., 2002). The training data sets consist of summary statistics computed on data sets simulated under the different assumed demographic models. For each simulation, parameter values are drawn from prior distributions. The RF algorithm then assigns observed data sets to one of the assumed demographic models and estimates the parameter values under this

BOX 1 Key machine learning terminology used throughout this Special Issue

Term	Definition
Artificial neural networks	Networks consisting of an interconnected network of nodes, designed to mimic the interconnected neurons that compose animal neural systems
CART	Classification and Regression Trees (CART) are used to perform classification and regression by constructing a decision tree that maximizes predictive accuracy on a labelled training data set
Cross-validation	An approach to quantifying power and generalizability of an algorithm by applying the algorithm to data sets with known labels or generating parameters (often simulated data sets), and quantifying the accuracy of inference
Deep learning	Machine learning algorithms that use multiple layers to learn features of the training data. Generally, early layers extract simple features, while layers applied later extract higher order features from the input data
Features	Features are the input to many machine learning algorithms (e.g., Random Forests [RF]), and are generally a set of summary statistics meant to capture the useful information in the data. Some approaches (e.g., deep learning) perform automated feature extraction and can be trained on raw data rather than features determined a priori to be informative
Interpretable machine learning	The field of interpretable machine learning aims to address the 'black-box' criticism of machine learning by extracting information about how machine learning algorithms map input to output
Random Forests (RF)	A machine learning approach (derived from CART) that uses a collection of decision trees to perform classification or regression
Supervised learning	Machine learning algorithms that require labelled training data. Often in molecular ecology, these training data are generated using simulations under known models and parameters.
Unsupervised learning	Machine learning algorithms that do not require labelled training data. Such approaches are often used for dimensionality reduction and clustering in molecular ecology

model. Compared to other ABC algorithms, ABC with random forest (ABCRF, Pudlo et al., 2016; Raynal et al., 2019) is efficient with smaller training sets, reducing the computational expense. The algorithm also appears to be much less sensitive to correlations among variables.

In this context, Collin et al. (2021) introduce a new computer package named DIYABC Random Forest version 1.0, based on the existing program DIYABC version 2.1.0 (Cornuet et al., 2014). This program offers a user-friendly interface that allows simulation of various kinds of genetic data (microsatellites, DNA sequences or single nucleotide polymorphisms [SNPs]) under demographic scenarios defined by the user. These data are then used to train an RF algorithm, which can be applied to observed data points. The program also allows the power and accuracy of inferences to be evaluated. They demonstrate that their program can be readily applied to real data, using the example of a set of human populations, for which they infer well-known events such as the out-of-Africa dispersion.

Other authors developed their own simulation programs for specific demographic scenarios. Fortes-Lima et al. (2021) introduce a software package METHIS to simulate various scenarios of admixture between two populations. Due to a publication error, Fortes-Lima et al. (2021) was published in Volume 21, Issue 4, but the paper remains part of this special issue. They also developed several summary statistics, specifically designed for distinguishing different admixture scenarios and inferring the parameters of the chosen scenario. They use the R package *ABCRF* to perform model choice and the R package *ABC* (Csilléry et al., 2012) for parameter inference, using in this case a neural network algorithm instead of RF. Through a cross-validation approach, they demonstrate the validity of their method, which fails mainly when highly nested scenarios are considered. They apply their approach to two human populations (African Americans and Barbadians) and infer the most likely admixture scenario for these two populations, a scenario of continuous decrease over time of the contributions of the source populations.

The availability of whole genome data is likely to increase our ability to infer precisely the demographic history of human populations. Under the *ABCRF* framework, several statistics need to be developed to synthesize the information provided by such data. In this context, Ghirotto et al. (2021) develop a new set of summary statistics, the frequency distribution of segregating sites (FDSS). They demonstrate on simulated data the efficiency of these statistics for distinguishing complex historical scenarios. They also apply these statistics to two real data sets. They consider first a set of human genomes from modern and ancient individuals (Neanderthal and Denisovan) to investigate whether current non-African populations resulted from one or several dispersal events and conclude that the multiple dispersal model is more likely. They also apply their method to a set of orangutan (*Pongo pygmaeus*) genomes, and find support for a scenario of separation and subsequent migration between populations from Sumatra and Borneo.

Finally, Fraïsse et al. (2021) provide a framework named *DILS* (Demographic Inference with Linked Selection). They propose a hierarchical approach that first infers the best historical scenario for one or two populations, with scenarios such as divergence between populations and reconnection through migration. Their method then estimates whether there is heterogeneity among loci in effective population size or migration rates, corresponding respectively to background selection in low-recombination and gene-dense regions and to selection against migrants at markers associated with barriers to gene flow. Finally, their method allows the genomic regions most associated with these barriers to be identified. They show the effectiveness of their method on a real example from *Mytilus* mussels. This method is of particular interest because it allows the use to infer both demographic and selective processes concurrently.

Altogether, these papers offer innovative methods that take advantage of the efficiency of the RF algorithm to infer demographic (and selective) processes. They also show that these methods can be applied readily to genetic data, including microsatellites, SNPs and NGS data. All these methods use the ABC framework, the main advantage of which is its flexibility, making it possible to consider any type of population history model for which data can be simulated. There is little doubt, therefore, that these methods will be used extensively in the future.

2.1.2 | Deep learning

While RF and related algorithms have led to substantial advances in population genetics and phylogeography, these approaches still require that large-dimensional genomic data be summarized using a set of summary statistics or features. While summary statistics may be powerful for answering some questions, for others it remains unclear which statistics may be the most useful, and it is unclear whether any set of statistics can fully capture the information content of genomic data sets (Flagel et al., 2019). Supervised deep learning may allow researchers to circumvent the need to calculate summary statistics when making inferences from genomic data. In comparison to RF, deep learning inference begins similarly. Typically, researchers define a set of models, and simulate data under these models. These data are then either summarized or used directly to train a neural network, and this network is applied to empirical data to predict which model (and/or parameters) generated the data. The major difference compared to RF is that the data need not be summarized using summary statistics for input into deep learning algorithms. For example, convolutional neural networks (CNNs; LeCun & Bengio, 1995) can be trained directly from SNP data simulated under various evolutionary scenarios (e.g., Chan et al., 2018; Flagel et al., 2019). In this Special Issue, four papers explore the use of deep learning to perform model selection or infer parameters from genomic data. These papers span the continuum from population genetics to phylogeography to phylogenetics and highlight the flexibility of deep learning approaches.

In an in-depth comparison of ABC and several deep learning algorithms, Sanchez et al. (2020) highlight the power of deep learning to infer complex population size histories. They consider several deep learning algorithms, including Multi-layer Perceptron Networks (MLPs; Rumelhart et al., 1986), several CNNs and a custom architecture that they term SPIDNA (sequence position informed deep neural architecture). Notably, SPIDNA is optimized for genomic data in that it is invariant to permutations of haplotypes, adaptive to varying numbers of SNPs and accounts for long-range dependencies between SNPs. Sanchez et al. (2020) demonstrate that their approach performs as well as or better than other network architectures considered, and show that, by combining their approach with ABC, they can achieve even higher accuracies. Sanchez et al. (2020) combine careful design of network architectures and Bayesian hyperparameter optimization to take advantage of the potential power of deep learning architectures, and their paper is a valuable guide to all wishing to apply deep learning in population genomics.

While Sanchez et al. (2020) focus on inferring population size histories, Fonseca et al. (2021) confront a problem common to phylogeographic studies, where model spaces may include population size histories, divergences and gene flow. They confront a complex model space using CNNs and a hierarchical approach to infer the phylogeographic history of South American lizards (genus *Norops*). As a first step, they compare models of population size history for each of the four populations in their study. Using these inferred population histories, they design a model set to explore evolutionary relationships among the four populations. They find that CNNs and their hierarchical approach have more power to distinguish amongst these complex models compared to traditional ABC, and their results suggest a complex history of population size changes, divergence and gene flow in *Norops* lizards.

Continuing in the direction of inferences over deeper timescales, Blischak et al. (2021) use deep learning to compare models of interspecific admixture and hybridization. Specifically, they demonstrate the power of CNNs to determine whether hybrid speciation or admixture has occurred. Rather than using SNPs as input directly, as in Sanchez et al. (2020) and Fonseca et al. (2021), Blischak et al. (2021) summarize genomic data by calculating d_{xy} across all species pairs within windows of the genome. This summarization incorporates phylogenetic information content and linkage information, both of which may be helpful in distinguishing hybridization scenarios. They compare their approach to ABCRF using a set of predefined summary statistics and find that CNNs offer increased power to distinguish amongst modes of introgression. Finally, they apply their approach to detect an ancient introgression event in *Heliconius* butterflies.

While the previous three papers use deep learning algorithms trained on data sets generated under a wide range of parameters in a manner analogous to many ABC algorithms, Wang et al. (2021) explore an approach that permits a heuristic exploration of parameter space. In their contribution, they introduce a method (*pg-gan*) for inferring demographic parameters using generative adversarial networks (GANs; Goodfellow et al., 2014). Briefly, GANs consist of a generator and a discriminator. The discriminator learns to distinguish real data from simulated data while the generator learns to produce simulated data that are difficult to distinguish from real data. The ultimate goal of a GAN is to generate fake data that are indistinguishable from real data. By using a parameterized evolutionary model as a generator, Wang et al. (2021) are able to adapt GANs to infer evolutionary parameters. The GAN is trained until the generator produces realistic data, and the evolutionary parameters that produce these realistic data are then inferred to be the true evolutionary parameters. Wang et al. (2021) demonstrate that this approach can accurately estimate parameters under a two-population isolation-with-migration model using simulation studies. They then apply *pg-gan* to estimate parameters across several human populations and demonstrate the power and flexibility of this approach.

These four papers highlight the flexibility of deep learning as a tool in molecular ecology. Inferences can be made across various time frames, from inferring relatively recent population size histories, to inferring divergences among populations, and to inferring admixture events between species. They can be applied directly to SNP alignments, or alignments can be summarized using high-dimensional statistics that would overwhelm ABC and even RF algorithms. Finally, deep learning algorithms can be used in a manner analogous to ABC, where training data sets are simulated from prior distributions meant to span the range of reasonable evolutionary parameters, or, using GANs, deep learning can be used to heuristically explore parameter space. Certainly, these papers offer a promising glimpse of the diversity of potential applications of deep learning in population genomics, phylogeography and phylogenetics.

2.2 | Selection

While the number of studies aimed at inferring demographic processes from genomic data using machine learning methods has increased steadily, as shown in the previous section, studies aimed at detecting loci under natural selection using such methods are so far much less common. We mentioned above the *DILS* method (Fraïsse et al., 2021) that aims to infer jointly demographic and selective processes. The two methods that we describe in this section focus more on selection, assuming a given demographic model.

Isildak et al. (2021) propose a model for detecting selection on genes at intermediate frequencies and distinguishing between recent balancing selection and incomplete sweeps. This is a challenging task, as these two types of selection leave quite similar signatures on the genome. As pointed out by the authors, summary statistics are not sufficient to distinguish between them. Conversely, they demonstrate the efficiency of neural networks, either artificial neural networks (ANNs) or CNNs. They also demonstrate that CNNs show a lower false positive rate compared to ANNs. Finally, they apply their method to European populations sequenced for the MEFV gene region, known to be linked with familial Mediterranean fever, for which they demonstrate that several variants underwent incomplete sweeps.

Luqman et al. (2021) provide an alternative method based on ABC that aims to estimate the demographic parameters for each locus. They consider that loci that show significant deviations in terms of their demographic parameters are affected by selective processes such as local adaptation. A simulation study allows them to show the increased efficiency of their method compared to recently developed methods such as *PCAdapt* (Luu et al., 2017) or *outFLANK* (Whitlock & Lotterhos, 2015). They also show the efficiency of their method on a real case (*Antirrhinum majus*).

These approaches are promising and there is no doubt that they will become as common as the methods aiming to infer demographic processes. Unlike neutral processes, which can be modelled very efficiently backward in time through the coalescent process, selective processes can, in most cases, only be simulated through individual-based forward-in-time simulation approaches. Obtaining large training sets can therefore be quite computationally intensive. The increasing availability of computing power along with the development of more efficient algorithms (e.g. Haller et al., 2019) should overcome these current limitations.

2.3 | Ecological genomics

Machine learning approaches, such as RF, are also increasingly recognized as important tools to untangle how environmental and landscape factors shape genomic change (Fitzpatrick & Keller, 2015; Fountain-Jones et al., 2017). Landscape and ecological genomics attempt to understand individual, population and community responses and adaptation to abiotic and biotic factors (e.g., Aguirre-Liguori et al., 2021; Allen & Banfield, 2005; Fitzpatrick & Keller, 2015; Steane et al., 2014; Ungerer et al., 2008). These responses are, by definition, complex, nonlinear and impacted by a variety of confounding factors (Ungerer et al., 2008). To handle this complexity and scale from 'molecule to landscape', spatial biogeographical methods, such as species distribution models (SDMs), and multivariate statistical techniques commonly employed to quantify community ecological problems are often used (e.g., Fitzpatrick & Keller, 2015; Forester et al., 2018; Jay et al., 2012). For example, constrained ordination approaches such as redundancy analysis (RDA, Legendre & Legendre, 2012) are a powerful method to identify gene–environment associations (GEAs) (Capblancq & Forester, 2021; Forester et al., 2018). However, analyses such as RDA are less able to capture nonlinear gene–environment relationships, and constructing individual SDMs for large numbers of loci is impractical (Fitzpatrick & Keller, 2015). The first three papers of this section address methodological challenges inherent to landscape and ecological genomics using innovative and contrasting approaches. The last paper in this section, in contrast, demonstrates how population genetic data can be used to infer the ecological and evolutionary drivers of community assembly using machine learning.

Gain and François (2021) introduce *LEA 3*, which expands the functionality of the LEA R package (Frichot & François, 2015) (Landscape and Ecological Association studies). The LEA R package harnesses latent factor mixed models (LFMMs) to quantify GEAs (Frichot et al., 2013). The authors harness LFMMs as unsupervised machine learning methods to quantify population structure and detect adaption without assumptions of the underlying biological processes. In the *LEA* framework, the 'latent' variables (or unobserved variables) represent data generated by population structure or statistical artefacts (Frichot & François, 2015; Frichot et al., 2013). Latent factor methods can detect structure in multivariate data sets by reducing high-dimensional matrices into a form suitable for hypothesis testing and inference (Warton et al. 2015). However, computationally efficient inference and parameter estimation have been challenging for latent factor methods (Niku et al., 2017). *LEA 3* overcomes this challenge through the implementation of LFMM ridge estimation algorithms (Caye et al., 2019), and the authors show that this technique has improved statistical performance compared to the previous Markov chain Monte Carlo (MCMC) methods. Moreover, *LEA 3* allows users to impute missing genotype data, detect outlier loci and calculate genetic offset statistics, a substantial increase in functionality compared to previous *LEA* iterations.

Fitzpatrick et al. (2021) take a different machine learning approach to calculate genetic offsets, utilizing the increasingly used Gradient Forest (GF) algorithm. GF is a multivariate extension of RF (Ellis et al., 2012) that captures nonlinear GEAs by quantifying compositional turnover in allele frequencies along environmental gradients (Fitzpatrick & Keller, 2015). GF turnover functions can be used to transform environmental features into 'genomic space' that captures expected variability in the genetic makeup of populations in different environments and project these relationships into the future (Fitzpatrick & Keller, 2015). The Euclidean distance between each population in current and future genomic spaces, or the 'genetic offset', is a metric of how vulnerable a population is to rapid environmental change (such as under future climate change scenarios) (Fitzpatrick & Keller, 2015). In this paper, Fitzpatrick et al. (2021) experimentally test the power and utility of GF-based genetic offsets using a common garden approach and compare the performance of GF to detect outlier SNPs to other methods such as LFMMs. They exposed balsam poplar (*Populus balsamifera* L.) from different populations throughout the species range to novel climates and assessed performance (e.g., growth and mortality) compared to the estimated genetic offset. The authors found that that populations with larger GF-based offsets had reduced performance compared to populations with smaller offsets, providing evidence for the utility of the approach to quantify climate maladaptation. Further, Fitzpatrick et al. (2021) highlight the GF routine as a promising approach to detect outlier SNPs without the restrictive assumptions of other models (e.g., GF is nonparametric and accounts for interactions between features).

Fountain-Jones et al. (2021) expand upon the GF algorithm by utilizing recent advances in multimodel inference and interpretable machine learning. The authors introduce *MrIML* ('Mister iml'; Multi-response Interpretable Machine Learning) to provide a powerful and flexible machine learning architecture to understand how landscape and environmental factors shape genomic change. The *MrIML* approach allows the user to compare the predictive performance of various models beyond RF from generalized linear models to neural networks in a streamlined and mathematically coherent way. Importantly, *MrIML* also provides a rich suite of interpretable machine learning tools to allow users to explore and better interpret complex machine learning models. The authors test the *MrIML* approach on simulated landscape genetic data sets (Landguth et al., 2020) and two contrasting empirical systems (balsam poplar alleles from the GIGANTEA-5 [*GI5*] gene, Fitzpatrick & Keller, 2015; and bobcat [*Lynx rufus*] feline immunodeficiency virus SNPs; Fountain-Jones et al., 2017; Lee et al., 2012). Similarly to Fitzpatrick et al. (2021), Fountain-Jones et al. (2021) assess the performance of the *MrIML* algorithm to detect outlier loci. For simulated data, *MrIML* correctly identified the known loci under selection, even when the relationship was complex and nonlinear, further highlighting the utility of the approach. The method also garnered new insights into the empirical data and identified plausible feature interactions that are difficult to capture with other methods. Moreover, *MrIML* can be used in other ecological subdisciplines to, for example, predict microbiome composition, explore molecular epidemiological patterns (Fountain-Jones et al., 2017, 2018) and untangle ecological community structure.

Lastly in this section, Overcast et al., (2021) demonstrate how population genetic data can be utilized to understand how ecological communities are organized. The authors introduce the *Massive Eco-evolutionary Synthesis Simulations* (MESS), which uses concepts in island biogeography to offer a unified model for community assembly. The MESS model generates predictions on species richness and abundance, population genetic diversity and phylogenetic trait variation. Importantly the MESS model parameters are fitted to empirical data using supervised machine learning, and similarly to *MrIML* there is flexibility with respect to which particular algorithm is employed. Population genetic data are rarely utilized in community assembly models, and the authors show that this unified approach could lead to mechanistic insights into how immigration, speciation and environmental filtering shape communities.

The papers in this section highlight the different machine learning approaches used to understand ecological genomic and community ecology problems. With the ever-increasing availability of large ecological and community genomic data sets, methods such as those highlighted in this section will be crucial not only in identifying complex patterns but also for predicting population and community response to change.

2.4 | Biodiversity, and species limits

Machine learning can also be used to better quantify and understand the biodiversity present on Earth, an important goal given the current biodiversity crisis. Two papers in the special issue use machine learning to improve our understanding of biodiversity. Martin et al. (2021) evaluate the use of unsupervised machine learning algorithms in species delimitation of North American box turtles, by employing a suite of unsupervised machine learning methods first applied to species delimitation by Derkarabetian et al. (2019) in combination with a supervised machine learning approach (Smith & Carstens, 2020) and a more traditional coalescent-based approach to species delimitation (Leaché et al., 2014). With respect to unsupervised machine learning algorithms, they compare the use of RF, T-distributed Stochastic Neighbor Embedding (Maaten & Hinton, 2008) and Variational Auto-Encoders (Kingma & Welling, 2013) to cluster individuals into putative species while using a variety of missing data thresholds, minor allele frequency filters and methods for selecting the best number of populations or species. They find that these approaches are very sensitive to relaxed missing data filters, and that filtering based on minor allele frequencies improves performance. Overall, their results suggest that, while machine learning is a powerful approach to delimiting species, researchers should take care to filter for missing data and minor allele frequencies and to evaluate various algorithms. This paper highlights both the power of machine learning approaches and the need for cautious application and interpretation of results.

Barrow et al. (2020) take a broader look at biodiversity by focusing on intraspecific diversity across Nearctic amphibians. Recent studies have begun to use RF to better understand the predictors of genetic diversity over large spatial scales (e.g., Pelletier & Carstens, 2018), and Barrow et al. (2020) use this approach to better understand the determinants of intraspecific genetic diversity, an important metric for understanding species resilience and assessing conservation concern. They use RF to analyse repurposed genetic data, life history data, and geographical data from 137 species of amphibians. They found that across the 137 Nearctic amphibians included in the study, taxonomic family, sample size and some bioclimatic variables were the best predictors of intraspecific diversity. They then focused on salamanders, and found that sample size and the minimum latitude of the species range were important predictors. This work highlights the role of machine learning in utilizing public databases to ask questions spanning large taxonomic groups and geographical areas in ways that traditional approaches could not.

Of course, with any large public database, errors may prove challenging for downstream inference. Barcoding data have become an important tool for biologists (Hebert et al., 2003), as evidenced by, for example, the use of barcode sequences in Barrow et al. (2020). However, errors can result in inflated estimates of diversity when not corrected. To address this issue, Nugent et al. (2021) introduce *debar*, an approach for denoising COI-5P DNA barcode data using machine learning. *Debar* uses a Profile Hidden Markov model (PHMM) to detect indel errors in *COI* barcoding data. The PHMM is trained on a set of high-quality filtered sequences, and then used to detect and correct indel errors in other sequences. Nugent et al. (2021) demonstrate the accuracy of this approach on both simulated and real data sets. Such approaches are promising in an era where more data are being collected than ever before, and error assessment can become incredibly time-consuming and prohibitive in the absence of accurate and easy-to-apply tools for detecting and correcting errors.

The papers in this section highlight that machine learning can contribute to our assessments of biodiversity in myriad ways. Machine learning approaches can help to accurately delimit species, providing better estimates of species-level diversity (Martin et al., 2021). Machine learning approaches also allow prediction of patterns of biodiversity at large geographical scales by facilitating the combination of genomic, ecological and geographical data in novel ways (Barrow et al., 2020). Finally, machine learning can improve our estimates of biodiversity by allowing efficient error correction of barcoding data sets (Nugent et al., 2021). Again, the diversity of approaches presented here highlights the potential of machine learning to enhance our understanding of biodiversity across temporal and spatial scales.

3 | CONCLUDING REMARKS

The studies compiled as part of this special issue highlight how machine learning approaches can be used to tackle a wide variety of challenges in molecular ecology. The efficiency, predictive power and ability to model complex nonlinear patterns with minimal assumptions make machine learning algorithms a natural analytical choice for many data sets, in particular very large data sets such as whole-genome sequences. Merging machine learning approaches with other statistical paradigms such as ABC are a particularly attractive option. Enhanced resources in, for example, R and python with user-friendly syntax, as provided by the studies compiled in this issue, allow molecular ecologists to use and compare a wide variety of cutting-edge algorithms.

Furthermore, advances in data science over the last 10 years (e.g., Molnar, 2018) mean that there is no need to consider these algorithms as 'black boxes'; interpretable machine learning can garner novel insights into model structure that can help gain new insights into data. The real challenge for ecologists is simply keeping up with the rapid pace of change in data science. Belcaid and Toonen (2015) wrote a review on computer and data science in molecular ecology, stressing the need for more integration between the fields, and this Special Issue highlights how fruitful this integration can be. The increase in computing power and the possibility of massive parallel computing thanks to GPUs should make it possible to tackle increasingly complex problems in the future, by making optimal use of the very large data sets that will become increasingly available and by allowing the integration of different types of data, such as genomic, epigenomic, phenotypic and ecological data sets.

Nicholas M. Fountain-Jones[1]
Megan L. Smith[2]
Frédéric Austerlitz[3]

[1]*School of Natural Sciences, University of Tasmania, Hobart, TAS, Australia*
Email: nfountainjones@gmail.com
[2]*Department of Biology, Indiana University, Bloomington, Indiana, USA*
[3]*UMR 7206 Eco-anthropologie, MNHN, CNRS, Université de Paris, Paris, France*

ORCID

Nicholas M. Fountain-Jones https://orcid.org/0000-0001-9248-8493
Megan L. Smith https://orcid.org/0000-0002-6362-9354
Frédéric Austerlitz https://orcid.org/0000-0001-8031-455X

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software Available from tensorflow.org

Aguirre-Liguori, J. A., Ramírez-Barahona, S., & Gaut, B. S. (2021). The evolutionary genomics of species' responses to climate change. *Nature Ecology & Evolution*, *5*, 1350–1360. https://doi.org/10.1038/s41559-021-01526-9

Allen, E., & Banfield, J.(2005). Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, *3*, 489–498.

Barrow, L. N., Masiero da Fonseca, E., Thompson, C. E. P., & Carstens, B. C. (2020). Predicting amphibian intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data. *Molecular Ecology Resources*, *21*(8), 2818–2831. https://doi.org/10.1111/1755-0998.13303

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Belcaid, M., & Toonen, R. J. (2015). Demystifying computer science for molecular ecologists. *Molecular Ecology*, *24*, 2619–2640.

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2021). Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *Molecular Ecology Resources*, *21*(8), 2676–2688. https://doi.org/10.1111/1755-0998.13355

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Capblancq, T., & Forester, B. R. (2021). Redundancy Analysis (RDA): A Swiss Army knife for landscape genomics. *Methods in Ecology and Evolution*, https://doi.org/10.1111/2041-210X.13722

Carlson, C. J. (2020). embarcadero: Species distribution modelling with Bayesian additive regression trees in r. *Methods in Ecology and Evolution*, *11*(7), 850–858. https://doi.org/10.1111/2041-210X.13389

Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution*, *36*, 852–860.

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S., & Song, Y. S. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in Neural Information Processing Systems*, *31*, 8594.

Collin, F.-D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J.-M., & Estoup, A. (2021). Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Molecular Ecology Resources*, *21*(8), 2598–2613. https://doi.org/10.1111/1755-0998.13413

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., Marin, J.-M., & Estoup, A. (2014). DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, *30*(8), 1187–1189.

Csillery, K., Francois, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., & Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*, *139*, 106562.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: Calculating importance gradients on physical predictors. *Ecology*, *93*(1), 156–168. https://doi.org/10.1890/11-0252.1

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., & Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Molecular Ecology Resources*, *21*(8), 2749–2765. https://doi.org/10.1111/1755-0998.13374

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16.

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238.

Fonseca, E. M., Colli, G. R., Werneck, F. P., & Carstens, B. C. (2021). Phylogeographic model selection using convolutional neural networks. *Molecular Ecology Resources*, *21*(8), 2661–2675. https://doi.org/10.1111/1755-0998.13427

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, *27*, 2215–2233.

Fortes-Lima, C. A., Laurent, R., Thouzeau, V., Toupance, B., & Verdu, P. (2021). Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. *Molecular Ecology Resources*, *21*(4), 1098–1117. https://doi.org/10.1111/1755-0998.13325

Fountain-Jones, N. M., Craft, M. E., Funk, W. C., Kozakiewicz, C., Trumbo, D. R., Boydston, E. E., Lyren, L. M., Crooks, K., Lee, J. S., VandeWoude, S., & Carver, S. (2017). Urban landscapes can change virus gene flow and evolution in a fragmentation-sensitive carnivore. *Molecular Ecology*, *26*(22), 6487–6498.

Fountain-Jones, N. M., Kozakiewicz, C. P., Forester, B. R., Landguth, E. L., Carver, S., Charleston, M., Gagne, R. B., Greenwell, B., Kraberger, S., Trumbo, D. R., Mayer, M., Clark, N. J., & Machado, G. (2021). MrIML: Multi-response interpretable machine learning to model genomic landscapes. *Molecular Ecology Resources*, *21*(8), 2766–2781. https://doi.org/10.1111/1755-0998.13495

Fountain-Jones, N. M., Machado, G., Carver, S., Packer, C., Recamonde-Mendoza, M., & Craft, M. E. (2019). How to make more from exposure data? An integrated machine learning pipeline to predict pathogen exposure. *Journal of Animal Ecology*, *88*, 1447–1461. https://doi.org/10.1111/1365-2656.13076

Fountain-Jones, N. M., Pearse, W. D., Escobar, L. E., Alba-Casals, A., Carver, S., Davies, T. J., Kraberger, S., Papeş, M., Vandegrift, K., Worsley-Tonks, K., & Craft, M. E. (2018). Towards an ecophylogenetic framework for infectious disease ecology. *Biological Reviews*, *93*, 950–970.

Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., Loire, É., Simon, A., Galtier, N., Duret, L., Bierne, N., Vekemans, X., & Roux, C. (2021). DILS: Demographic inferences with linked selection by using ABC. *Molecular Ecology Resources*, *21*(8), 2629–2644. https://doi.org/10.1111/1755-0998.13323

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*, 925–929. https://doi.org/10.1111/2041-210X.12382

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30*(7), 1687–1699.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*, 906–914.

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*, *21*(8), 2738–2748. https://doi.org/10.1111/1755-0998.13366

Ghirotto, S., Vizzari, M. T., Tassi, F., Barbujani, G., & Benazzo, A. (2021). Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. *Molecular Ecology Resources*, *21*(8), 2614–2628. https://doi.org/10.1111/1755-0998.13263

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 52–566. https://doi.org/10.1111/1755-0998.12968

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. *In: The Elements of Statistical Learning*. Springer, NY, pp. 337–387.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & Dewaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1512), 313–321.

Ho, T. K. (1995). *Random decision forests*. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. https://ieeexplore.ieee.org/document/598994

Isildak, U., Stella, A., & Fumagalli, M. (2021). Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Molecular Ecology Resources*, *21*(8), 2706–2718. https://doi.org/10.1111/1755-0998.13379

Jay, F., Manel, S., Alvarez, N., Durand, E. Y., Thuiller, W., Holderegger, R., Taberlet, P., & François, O. (2012). Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, *21*, 2354–2368. https://doi.org/10.1111/j.1365-294X.2012.05541.x

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes*. In: Proceedings of the International Conference on Learning Representations (ICLR). arXiv:1312.6114 [stat.ML].

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org

Landguth, E. L., Forester, B. R., Eckert, A. J., Shirk, A. J., Menon, M., Whipple, A., Day, C. C., & Cushman, S. A. (2020). Modelling multilocus selection in an individual-based, spatially-explicit landscape genetics framework. *Molecular Ecology Resources*, *20*, 605–615. https://doi.org/10.1111/1755-0998.13121

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic Biology*, *63*(4), 534–542.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (Vol. *3361*(10), pp. 255–258). The MIT Press.

Lee, J. S., Ruell, E. W., Boydston, E. E., Lyren, L. M., Alonso, R. S., Troyer, J. L., & Vandewoude, S. (2012). Gene flow and pathogen transmission among bobcats (*Lynx rufus*) in a fragmented urban landscape. *Molecular Ecology*, *21*(7), 1617–1631. https://doi.org/10.1111/j.1365-294X.2012.05493.x

Legendre, P., & Legendre, L. (2012). *Numerical ecology*. Elsevier.

Li, J., Li, H., Jakobsson, M., Li, S., Sjödin, P., & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? *Molecular Ecology*, *21*, 28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x

Lin, K., Li, H., Schlötterer, C., & Futschik, A. (2011). Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics*, *187*(1), 229–244.

Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, *90*(4), e01422.

Luqman, H., Widmer, A., Fior, S., & Wegmann, D. (2021). Identifying loci under selection via explicit demographic models. *Molecular Ecology Resources*, *21*(8), 2719–2737. https://doi.org/10.1111/1755-0998.13415

Luu, K., Bazin, E., & Blum, M. G. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Martin, B. T., Chafin, T. K., Douglas, M. R., Placyk, J. S. Jr, Birkhead, R. D., Phillips, C. A., & Douglas, M. E. (2021). The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (*Terrapene* spp.). *Molecular Ecology Resources*, *21*(8), 2801–2817. https://doi.org/10.1111/1755-0998.13350

Molnar, C. (2018). *Interpretable machine learning*. Retrieved from https://christophm.github.io/interpretable-ml-book/

Niku, J., Warton, D. I., Hui, F. K. C., & Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, *22*, 498–522. https://doi.org/10.1007/s13253-017-0304-7

Nugent, C. M., Elliott, T. A., Ratnasingham, S., Hebert, P. D., & Adamowicz, S. J. (2021). Debar: A sequence-by-sequence denoiser for COI-5P DNA barcode data. *Molecular Ecology Resources*, *21*(8), 2832–2846. https://doi.org/10.1111/1755-0998.13384

Overcast, I., Ruffley, M., Rosindell, J., Harmon, L., Borges, P. A. V., Emerson, B. C., Etienne, R. S., Gillespie, R., Krehenwinkel, H., Mahler, D. L., Massol, F., Parent, C. E., Patiño, J., Peter, B., Week, B., Wagner, C., Hickerson, M. J., & Rominger, A. (2021). A unified model of species abundance, genetic diversity, and functional diversity reveals the mechanisms structuring ecological communities. *Molecular Ecology Resources*, *21*(8), 2782–2800. https://doi.org/10.1111/1755-0998.13514

Pelletier, T. A., & Carstens, B. C. (2018). Geographical range size and latitude predict population genetic structure in a global survey. *Biology Letters*, *14*(1), 20170566.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, *32*(6), 859–866.

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. Mcclelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition volume 1: Foundations* (pp. 318–362). MIT Press.

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2020). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*, *21*(8), 2645–2660. https://doi.org/10.1111/1755-0998.13224

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312.

Smith, M. L., & Carstens, B. C. (2020). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, *74*(2), 216–229.

Steane, D. A., Potts, B. M., McLean, E., Prober, S. M., Stock, W. D., Vaillancourt, R. E., & Byrne, M. (2014). Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology*, *23*, 2500–2513.

Ungerer, M., Johnson, L., & Herman, M. (2008). Ecological genomics: understanding gene and genome function in the natural environment. *Heredity*, *100*, 178–183.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H. H., Mathieson, I., & Mathieson, S. (2021). Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, *21*(8), 2689–2705. https://doi.org/10.1111/1755-0998.13386

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of FST. *American Naturalist*, *186*(Suppl 1), S24–S36.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007