Ensemble approaches for estimating congruence between species delimitation and morphological variation: comparing taxonomic hypotheses for the Pacific Pond Turtle (*Emys marmorata*)

Manuscript elements: Figure 1, figure 2, figure 3, figure 4, figure 5, figure 6, figure 7, figure 8, figure 9, table 1, table 2, table 3. Figures 1, 3, 4, 6, 8 are in color.

Keywords: Geometric morphometrics, machine learning, conservation, Testudinae.

Manuscript type: Article.

Prepared using the suggested LaTeX template for *Am. Nat.*

Abstract

We investigated the morphometric identification of cryptic species using machine learning approaches by examining their implications for a recently proposed cryptic turtle species (*Emys pallida*). We collected landmark-based morphometric data from 532 adult *E. marmorata/"E. pallida"* museum specimens. We assigned a classification to each specimen for six different binning schemes based on geographic occurrence data. We used an ensemble of supervised machine learning approaches to determine which classification hypothesis was best supported by the data. In addition, we applied the same approach to two clear-cut examples, one consisting of eight unambiguously distinct species closely related to *E. marmorata*, and the other consisting of two subspecies of *Trachemys scripta*. Our results indicate that there is no clear "best" grouping of *E. marmorata/"E. pallida"* based on plastron shape. Explanations for the lack of grouping in *E. marmorata* include the possibility that genetic differentiation is not associated with plastron shape variation below the species level and/or that local selective pressures (e.g., from hydrological regime) overwhelm morphological differentiation. A reconsideration of the methods used to delimit *"E. pallida,"* the lack of barriers to gene flow, the strong evidence for widespread admixture between lineages, and the fact that plastron shape can be used to delineate other emydine species and sub-species suggest that its lack of diagnosability most likely reflects the non-distinctiveness of this proposed taxon.

Introduction

Molecular systematics has repeatedly demonstrated the existence of cryptic species that can only be diagnosed using genetic data (Bickford et al., 2007; Clare, 2011; Funk et al., 2012; Pfenninger and Schwenk, 2007; Schilck-Steiner et al., 2007; Stuart et al., 2006). In attempts to streamline the documentation of biodiversity, several methods of species delimitation that rely almost entirely on genetic data have recently been proposed (Carstens and Dewey, 2010; Hausdorf and Hennig, 2010; Huelsenbeck et al., 2011; O'Meara, 2010; Pons et al., 2006; Yang and Rannala, 2010). Although strong caveats on the utility of these methods have been raised (Bauer et al., 2000; Carstens et al., 2013), they are nevertheless being used to name species (Leaché and Fujita, 2010; Spinks et al., 2014).

In contrast to those genetically-diagnosed species, the majority of extant taxa, and almost all extinct taxa, are delimited by morphology alone. This disjunction complicates interpretations of variation and diversity in deep time, as apparent morphological stasis may not reflect the true underlying diversity (Eldredge and Gould, 1972; Gould and Eldredge, 1977; Van Bocxlaer and Hunt, 2013). It also has serious implications for our records of modern biodiversity: for many museum specimens of extant taxa (e.g. those preserved in formalin), it is difficult to acquire the genetic data needed for non-morphological species delimitation methods.

These considerations have sparked interest in whether geometric morphometric analyses can capture fine-scale variation that can be used for identifying cryptic species. This would make the task of identifying and maintaining endangered or conserved groups much easier and could contribute to improved classifications of extinct taxa and populations. Most such studies focus on using morphometrics to discover differences between taxa that were identified by other means (Demandt and Bergek, 2009; Fruciano et al., 2016; Gaubert et al., 2005; Gündüz et al., 2007; Markolf et al., 2013; Polly, 2003, 2007; Zelditch et al., 2004). Additionally, there has been work on automated taxon identification and classification of taxa into groups (Baylac et al., 2003; Dobigny et al., 2003;

MacLeod, 2007; van den Brink and Bokma, 2011; Vitek et al., 2017), as well as the development of models that combine genetic, phenotypic, and geographic data to infer evolutionary units of interest (Guillot et al., 2012).

Here, we investigate the morphometric identification of cryptic species using machine learning approaches. We use an ensemble learning approach where multiple methods are used in order to look for consensus among their results. We test our approach on three datasets: plastron shape of eight species of closely related turtles, plastron shape of two subspecies of a single turtle species, and plastron shape of the *Emys marmorata* species complex. In particular, we ask whether it is possible to determine which among a set of classification hypotheses best aligns with the observed morphology, and examine the implications of our results for the *E. marmorata* complex.

Background and study system

Machine learning is an extension of known statistical methodology (Hastie et al., 2009) that emphasizes predictive accuracy and generality often at the expense of the interpretability of individual parameters. Basic statistical approaches are supplemented by randomization, sorting, and partitioning algorithms, along with the maximization or minimization of summary statistics, in order to best estimate a general model for all data, both sampled and unsampled (Hastie et al., 2009). Machine learning approaches have found use in medical research, epidemiology, economics, and automated identification of images such as handwritten zip codes (Hastie et al., 2009).

There are two major classes of machine learning method: unsupervised and supervised learning. Unsupervised learning methods are used with unlabeled data where the underlying structure is estimated; they are analogous to clustering and density estimation methods (Kaufman and Rousseeuw, 1990). Supervised learning methods are used with labeled data where the final output of data is known and the rules for going from input to output are inferred. These are analogous to classification and regression models (Breiman et al., 1984; Hastie et al., 2009). Our application of the supervised learning approaches used in this study illustrates only a sampling of the various

methods available for fitting classification models. The specific methods used in this study were chosen because they are suited for cases with more two or more response classes.

Geometric morphometric approaches to identifying differences in morphological variation between classes, including cryptic species, have mostly relied on methods like linear discriminate analysis and canonical variates analysis (Dillard, 2017; Edwards et al., 2011; Francoy et al., 2009; Gaubert et al., 2005; Gündüz et al., 2007; Mitrovski-Bogdanovic et al., 2013; Polly, 2003, 2007; Sztencel-Jabłonka et al., 2009; Zelditch et al., 2004). Because of their similarity to multivariate approaches like principal components analysis (PCA), these methods are comparatively straightforward ways of understanding the differences in morphology between classes. They also benefit from producing results that can be easily visualized, which aids in the interpretation and presentation of data and results. Most previous morphometric studies did not assess which amongst a set of alternative classification hypotheses was optimal. For example, studies such as those of Caumul and Polly (2005) and Polly (2007) focused on comparing different aspects of morphology and their fidelity to a classification scheme instead of comparing the fidelity of one aspect of morphology to multiple classification schemes. In this context, the study of Cardini et al. (2009) is noteworthy because they compared morphological variation in marmots at the population, regional, and species level and determined the fidelity of shape to divisions at each of these levels.

Here, we used an ensemble of supervised machine learning methods to compare the congruence of morphological data to different classification hypotheses. Each of these methods provide different advantages for understanding how to classify specimens, as well as the accuracy of the resulting classifications. Machine learning methods have been combined with geometric morphometric data to study shape variation in a variety of contexts, including automated taxon identification and classification of groups (Baylac et al., 2003; Dobigny et al., 2003; MacLeod, 2007; Navega et al., 2015; Van Bocxlaer and Schultheiß, 2010; van den Brink and Bokma, 2011). In the current study, we not only consider pure classification accuracy but also use a statistic of classification strength that reflects the rate at which taxa are both accurately and inaccurately classified: the area under

the Receiver Operating Characteristic curve (Hastie et al., 2009).

We analyzed the problem of whether there are distinct subspecies or cryptic species within the western pond turtle, *Emys marmorata* (Baird and Girard, 1852) (formerly *Clemmys marmorata*; see Feldman and Parham, 2002). *Emys marmorata* is distributed from northern Washington State, USA to Baja California, Mexico; populations in western Nevada may have been introduced by recent human activity or they could be a genuine part of the species' range (Bury, 2017). Traditionally, *E. marmorata* was classified into two named subspecies: the northern *E. marmorata marmorata* and the southern *Emys marmorata pallida* (Seeliger, 1945), with a central Californian intergrade zone in between. *Emys marmorata marmorata* is differentiated from *E. marmorata pallida* by the presence of a pair of triangular inguinal scales and darker neck markings. The triangular inguinal plates can sometimes be present in *E. marmorata pallida* although they are considerably smaller. Seeliger (1945) did not formally include the Baja California populations of *E. marmorata* in either taxon, implying the existence of a third distinct but unnamed subspecies.

Previous work on morphological variation in *E. marmorata* has focused primarily on differentiation between populations over a portion of the species' total range (Bury et al., 2010; Germano and Bury, 2009; Germano and Rathbun, 2008; Lubcke and Wilson, 2007); comparatively few studies have included specimens from across the entire range (Holland, 1992). Most of these studies considered how local biotic and abiotic factors may contribute to differences in carapace length, and they found that size can vary greatly between different populations (Germano and Bury, 2009; Germano and Rathbun, 2008; Lubcke and Wilson, 2007). There also has been interest in size-based sexual dimorphism in *E. marmorata* (Germano and Bury, 2009; Holland, 1992; Lubcke and Wilson, 2007), with males being on average larger than females based on total carapace length and other linear measurements. However, the quality of size as a classifier of sex can vary greatly between populations (Holland, 1992) because of the magnitude of size differences among populations (Germano and Bury, 2009; Lubcke and Wilson, 2007). The effect of sexual dimorphism on shape, *sensu* Kendall (1977), has not been assessed (Germano and Rathbun, 2008; Holland, 1992; Lubcke

and Wilson, 2007).

Of particular relevance in the context of cryptic diversity in *E. marmorata* is the morphometric analysis of carapace shape carried out by Holland (1992), who compared populations of *E. marmorata* from three areas of the species' range. Holland concluded that geographic distance was a poor indicator of morphological differentiation, and instead hypothesized that geographic features such as breaks between different drainage basis are probably more important barriers to dispersal and interbreeding. Additionally, he suggested that morphological differences were more pronounced as the magnitude of barriers and distance increased, but this variation required many variables to adequately capture, implying only very subtle morphological differentiation between putatively distinct populations. Finally, Holland concluded that *E. marmorata* is best classified as three distinct species: a northern species, a southern species, and a Columbia Basin species. This classification is similar to that of Seeliger (1945), except elevated to the species level and without recognition of a distinct Baja species.

More recently, the phylogeography of *E. marmorata* and the possibility of cryptic diversity was investigated using molecular data (Spinks and Shaffer, 2005; Spinks et al., 2014, 2010). Based on mitochondrial DNA, Spinks and Shaffer (2005) recognized four subclades within *E. marmorata*, a northern clade, a San Joaquin Valley clade, a Santa Barbara clade, and a southern clade. Analyses with nuclear DNA (Spinks et al., 2010) and single-nucleotide polymorphism (SNP) data suggest a primarily north–south division in *E. marmorata*, although these datasets differed from that of mitocondrial-based results of Spinks and Shaffer (2005) in the location of the break point (Spinks et al., 2014). All three studies discussed the potential taxonomic implications of their results, with Spinks et al. (2014) going so far as to strongly advocate for the recognition of at least two species (*E. marmorata* and *E. pallida*), and a possible third based on populations in Baja California. However, they did not discuss in detail the morphological characters that would help to diagnose these species beyond those specified by Seeliger (1945). Given that these characters are variable within the proposed species, and that Holland (1992) described shell shape variation that might be

consistent with this taxonomy, a geometric morphometric analysis of shell shape might provide a reliable way to diagnose groups (whether species or subspecies) within *E. marmorata*.

In this study, we attempt to estimate the best classification scheme of *E. marmorata* based on variation in plastron (ventral shell) shape in order to determine whether this character is consistent with any of the proposed taxonomies of the *E. marmorata* complex.

We choose to analyze plastron shape for multiple reasons. First, it is very easy to collect geometric morphometric data on plastron shape from two-dimensional pictures as the structure is virtually flat. This approach allows both museum specimens and individuals in the field to be analyzed together. Second, previous work has suggested that there are strong differences in plastron shape among traditionally-recognized emydine species (Angielczyk and Feldman, 2013; Angielczyk et al., 2011; Angielczyk and Sheets, 2007). Finally, due to these previous studies a large dataset was readily available.

In the case of the *E. marmorata* species complex, we hypothesize that if one or more of the proposed classification schemes are consistent with the morphological data then our ensemble approach fit to those hypotheses will have higher out-of-sample predictive performance than the more inconsistent hypotheses. However, if all of the classification schemes lead to equal out-of-sample predictive performance then we would conclude that the proposed hypotheses are inconsistent with whatever information is present in the morphological data. Because of unclear geographic boundaries between subgroups of *E. marmorata*, we compare multiple permutations of the (Spinks et al., 2010) and Spinks et al. (2014) hypotheses.

Methods

Specimens, sampling, morphometrics

Three different geometric morphometric datasets describing turtle plastron variation were assembled for this analysis: 1) specimens from eight distinct emydine species; 2) *T. scripta* specimens from the two main subspecies (*T. scripta elegans* and *T. scripta scripta*); and 3) *E. marmorata* specimens from across the species' geographic range. The first two datasets are intended to serve as a test of whether machine learning techniques can differentiate species-level groupings of emydine turtles using plastron shape. We expect that the first case represents a low complexity dataset because of the high level of plastron shape disparity that exists among these species (Angielczyk et al., 2011; Claude, 2006; Claude et al., 2003), whereas the second dataset should be relatively higher in complexity and more analogous to the *E. marmorata* example. We predict that the *E. marmorata* dataset should be of the highest complexity and our greatest challenge given the finding that only very subtle differences existed between geographically-distinct populations (Holland, 1992).

The first dataset we analyzed includes 578 total specimens from the following species: *Chrysemys picta*, *Clemmys guttata*, *Emys blandigii*, *Emys orbicularis*, *Glyptemys insculpta*, *Glyptemys muhlenbergii*, *Terrapene coahuila*, and *Terrapene ornata*. These specimens are a subset of those used in Angielczyk et al. (2011) and Angielczyk and Feldman (2013).

The second dataset is a compilation of 101 specimens of two subspecies of *T. scripta*: 51 specimens of *T. scripta scripta* and 50 specimens of *T. scripta elegans*. These landmark data are new to this study.

The final dataset is of 532 adult *E. marmorata* museum specimens, though not all specimens were able to be assigned a class for all schemes (Fig. 1). These specimens represent a subset of those included in Angielczyk and Sheets (2007), Angielczyk et al. (2011), and Angielczyk and Feldman (2013). Because Spinks and Shaffer (2005), Spinks et al. (2010), and Spinks et al. (2014) did not use

vouchered specimens we were not able to directly sample the individuals in their studies. Instead, our specimen classifications were based solely on the geographic information and not explicit assignment using molecular data. For each taxonomic hypothesis, specimens were assigned to one of the possible classes based on geographic occurrence data recorded in museum collections. In cases where precise latitude and longitude information were not available we estimated them from other locality information. Because the exact barriers between different biogeographic regions are unknown and unclear, we represented each hypothesis with multiple possible realizations representing the classification uncertainty for specimens present at the geographic boundaries. The taxonomic hypotheses and sub-hypotheses for *E. marmorata* used here are presented in Table 1 and Figure 1.

For Spinks et al. (2010) we used three binning schemes. All three schemes include a class for *E. marmorata* specimens from northern populations (marm) as well as a class for those assigned to *E. pallida* (pall) and an intergrade zone in the Central Coast Ranges (CCR). The schemes differ in the assignment of samples from the San Joaquin Valley (Fig. 1). Scheme SP10.1 and SP10.2 differ in the assignment of specimens from the western San Joaquin Valley to either CCR or marm reflecting uncertainty regarding their genetic affinity as explained above. In scheme SP10.3 these specimens are assigned to a San Joaquin class reflecting the mitochondrial distinctiveness shown by Spinks and Shaffer (2005). For Spinks et al. (2014) we used two binning schemes with SP14.1 being based on their phylogenetic network analysis and SP14.2 being based on their Bayesian species delimitation analysis. The latter scheme requires the addition of two new classes, "Baja" and "Foothill," to accommodate the genetic groupings recovered by the SNP Structure analysis that was used to create the guide tree for the BPP species delimitation analysis in Spinks et al. (2014). Finally, we proposed a conservative morphological hypothesis ("Morph") in order to compare the molecular hypotheses with something approximating the original taxonomic hypothesis for the group; this scheme is made up solely of the marm and pall classes from the SP10.3 scheme.

Sex was known only for a subset of the total dataset and was not included as a predictor of

classification. Instead, we estimated the degree by which specimens cluster morphologically by sex in order to determine how much of a potential biasing factor sexual dimorphism could be for our analysis of the *E. marmorata* species complex (see below).

Following previous work on plastron shape (Angielczyk and Feldman, 2013; Angielczyk et al., 2011; Angielczyk and Sheets, 2007), we used TpsDig 2.04 (Rohlf, 2005) to digitize 19 two-dimensional landmarks (Fig. 2). Seventeen of the landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the plastron. Twelve of the landmarks were symmetrical across the axis of symmetry. Because damage prevented the digitization of all the symmetric landmarks in some specimens, we reflected landmarks across the axis of symmetry (i.e. midline) prior to analysis and used the average position of each symmetrical pair. In cases where damage or incompleteness prevented symmetric landmarks from being determined, we used only the single member of the pair. We conducted all subsequent analyses on the resulting "half" plastra. We superimposed the plastral landmark configurations using generalized Procrustes analysis (Dryden and Mardia, 1998), after which we calculated the principal components (PC) of shape using the `shapes` package for R (Dryden, 2013; R Core Team, 2016). All specimens were used for superimposition, after which the subset labeled for each of the schemes were used in model training and testing (see below).

Biasing effects

We estimated the possible effect of digitization error (Arnqvist and Mårtensson, 1998; Munoz-Munoz F. and Perpinan D., 2010; von Cramon-Taubadel et al., 2007) on our results by comparing within-specimen (replicated) Procrustes distances to the distances between classification scheme centroids. Ten randomly-selected *E. marmorata* specimens were each digitized four times, with the original set of digitized coordinates serving as a fifth replicate. These 50 landmark configurations were then Procrustes superimposed. A range of four Procrustes distances was then calculated as the average of the pairwise distances between each of the replicate configurations of a given

specimen.

For each specimen, the difference in shape caused by digitization was calculated as the mean of all pairwise Procrustes distances between the five replicates of that specimen. The average distance between any two digitizations was calculated as the mean of all pairwise Procrustes distances between all replicates for all specimens. The ratio between these two values was used to assess the magnitude of variation caused by digitization. The goal of this ratio is to determine if the within group distances are on average smaller than the between individual distances; a value of 0 indicates perfect grouping, a value of 1 indicates no difference between grouping and no grouping, and a value of 1+ indicates that the grouping is counter-intuitive to the data.

Emys marmorata is known to display sexual dimorphism in plastral shape, particularly the presence of a plastra concavity in males (Seeliger, 1945). To test for biases resulting from sexual dimorphism in our *E. marmorata* dataset, we used a simple permutation test to determine if the distance between the mean female and male shapes is greater than expected when the sex labels are randomly shuffled. Because not all of our specimens have sex identifications associated with them, this analysis was done using a subset of the data (257 of 532).

Supervised learning approaches

Instead of relying on a single supervised learning method, we chose to use an ensemble approach where multiple model types are used in concert so that any congruence between them increases our support for that conclusion over another (Hastie et al., 2009). The supervised learning methods used here are named in Table 2. Each of these methods makes different assumptions, treats data differently, and can produce different classification results depending on the nature of the data (Hastie et al., 2009). For example, multinomial logistic regression is a type of generalized linear model, whereas random forest is itself an ensemble approach where multiple decision trees are fit to subsets of the full dataset and then averaged.

The maximum set of possible predictors or features used for any model of our dataset is comprised of the first 25 principal components (PCs), scaled centroid size, and the interaction between scaled centroid size and PC 1. Additional interaction terms were not considered because of model complexity/sample size concerns. Size and the interaction between size and PC 1 were included as predictors to account for known ontogenetic variation in plastron shape (Angielczyk and Feldman, 2013) as well as potential size differences between classes, even if this is unlikely (Holland, 1992; Seeliger, 1945). These data constitute a "maximum set" because the best or selected models based on five-fold cross-validation need not, and likely will not, include all predictors possible (see below). Because our supervised learning models use PCs as predictors, this approach is in many ways analogous to PCA regression. PCA regression takes advantage of reduction and orthogonality PCs to improve regression fit (Hastie et al., 2009). Because the PCs of shape are by definition orthogonal, they can easily serve as independent predictors or features of class membership without fear of collinearity.

We adopted a training and testing paradigm for selecting parsimonious models and estimating their overall error rates (Hastie et al., 2009; Kuhn and Johnson, 2013). Within-sample model performance is inherently biased upwards, so model evaluation requires overcoming this bias. With very large sample sizes, as in this study, part of the sample can be used as the "training set" and the remainder acts as the "testing set." In this approach, following all cleaning and vetting, the data are split into a training dataset and a testing dataset. The former is used for fitting the model whereas the later is used for measuring model performance, a process called model generalization. For each scheme, we limited the model training and testing to only those individuals with class labels for that scheme. In this analysis, we randomly divided 80% of samples into the training set and the remaining 20% into the testing set.

In classification studies, such as this one, a common metric of performance is the receiver operating characteristic (ROC) which is the relationship between the false and true positive rates (Hastie et al., 2009). The area under the ROC curve (AUC) is the derived estimate of the model performance;

AUC ranges from 0.5 to 1 which correspond to performance similar to random guesses and perfect classification rates, respectively (Hastie et al., 2009). Both ROC and AUC are preferable to simple classification accuracy when class membership is unbalanced, as it is in these analyses (Hastie et al., 2009). The standard ROC and AUC calculations are defined only for binary classifications, which is not the case for our eight species and *Emys* complex datasets. To generalize this approach for situations with multiple response classes, we used an all-against-one strategy where the model AUC is the average of the AUC values from the multiple binary comparisons of one class compared to all others (Hand and Till, 2001).

For a given supervised learning method, we compared the fit of 27 models as the average AUC from 10 rounds of five-fold cross-validation. Cross-validation is an approach for estimating the average out-of-sample predictive error of a model by simulating out-of-sample data from the training dataset itself (Hastie et al., 2009). In a single round of k-fold cross-validation, the training data are divided into k blocks where the model is fit to $k-1$ blocks and the values of the kth block are predicted. This is repeated for all combinations of blocks. Within each round, the predictive performance metrics are averaged across all folds. Finally, the predictive performance metric is the averaged across all rounds of k-fold cross-validation. This process was implemented using the R package `caret` (Kuhn, 2013). For a given supervised learning method, the "best" trained model is that with the highest mean AUC as estimated from five-fold cross-validation. The selected or final model, however, is the next most parsimonious model that is within one standard error of the best model; this is a variant on the "one-standard error" rule from Hastie et al. (2009). The purpose of this rule is to ameliorate the chances of selecting an overly complex model that will perform poorly when predicting the classes of out-of-sample data.

Results

Geometric morphometrics

The results of the PCA of plastron shape in both the eight species and *Trachemys* datasets demonstrate strong association between shape and the recognized classification schemes (Fig. 3).

The results of the PCA of plastron shape in the *Emys marmorata* dataset show no clear connection between plastron shape and any of the proposed classification schemes (Fig. 4). The first PC axis of shape variation appears to be primarily structured by differences in individual centroid size (Fig. 4); this was the motivation for including centroid size and its interaction with PC1 as predictors in all of the supervised learning models.

Analysis of the differences between sexes of *E. marmorata* indicates that sex does not appear to strongly structure differences in shape (Fig. 5). The difference in mean shape between the sexes is very small; the sexes overlap about has much as expected given a null distribution based on permuting the sex-labels.

Comparison of the within to between Procrustes distances of the digitization replicates gives an approximate estimate of the error between distinct groupings (Table 3). The ratio of the average within-individual distance to the average distance between individuals for the replicated datasets is 1.11; this indicates that the grouping is slightly counter-intuitive to the data and is consistent with all shapes being very similar regardless of individual identity. This value also provides a baseline by which to understand how distinct the groupings are, where other ratios are compared to the correction ratio 1.11/1.

The results from the eight species and *Trachemys* datasets indicate that both of these classification schemes are more recognizable than not given our estimate of digitization error (Table 3). In contrast, the different *E. marmorata* classification schemes appear to barely be distinct, with their within:between ratios approximating 1. This indicates that the magnitude of the differences be-

tween groupings is approximately the same as the difference between any two random individuals (Table 3).

Supervised learning

Analysis of the eight morphologically- and genetically-distinct species and the *T. scripta scripta–T. scripta elegans* datasets indicate that these taxa are sufficiently morphologically distinct to be differentiated on the basis of plastron shape. Both in-sample and out-of-sample classification have AUC values of approximately 1 for all methods, implying near-perfect classification rates (Fig. 6, 7). For both datasets, the ROC scores from testing datasets are tightly clustered near AUC = 1 (Fig. 7). These results demonstrate that when there are distinctions between the states of the classification schemes (i.e., differences in plastron shape that correlate with the different taxonomic groups), the methods used here can recover them.

AUC–based model selection revealed some important patterns of variation and congruence between the classification schemes and the actual data. Generally, the best performing models tended to include about half the total number of possible PCs (Fig. 8).

Observed AUC values for all of the optimal models are lower for the *E. marmorata* dataset than for the other two datasets (Fig. 6, 8). In most cases the different proposed classification schemes are generally poor descriptors of the observed variation. It appears that the dataset is overwhelmed by noise (likely biological and analytical), making any accurate classifications difficult at best. This observation is cemented with the generalizations of the models to the testing dataset (Fig. 9).

Mean AUC values for the model generalizations, in most cases, are approximately equal to the observed AUC values from the training dataset (Fig. 8, 9). The cases in which the AUC from the generalizations is less than the observed indicate poor model fit and a poor classification scheme. Comparison of AUC values from the model generalizations do not indicate a clear "best" classification scheme (Fig. 8, 9). Only in the case of the conservative morphological

hypothesis ("Morph") is the mean AUC value potentially distinct from that of other schemes; in this case mean AUC is lower than the average of the other five schemes which indicates that the moprhologically-based scheme performs more poorly than the molecularly-based ones. It is important to note, however, that the training and testing dataset for the "Morph" scheme is the smallest of the six schemes which may lead to poorer performance in in-sample and out-of-sample comparisons.

Discussion

As expected, our ensemble approach yields high out-of-sample classification performance for the first two datasets. These results indicate that in cases of clear class separation (Fig. 3) our approach is able to detect this and make good out-of-sample prediction.

In the case of the *E. marmorata* dataset, our results show that none of the proposed taxonomic hypotheses for the *E. marmorata* species complex are more consistent with morphological differentiation than any other proposal (Fig. 9). Both the low out-of-sample AUC values and the significant difference between the correctly and incorrectly classified observations support the conclusion that none of the hypothesized classification schemes are good descriptions of the observed plastral variation within *E. marmorata*. An analytical explanation of this result is that the level of digitization error in the *E. marmorata* dataset is so great as to swamp out any biological signal. We think this is unlikely because all of the specimens considered in our three analyses were digitized by one of us (K.D.A.), and digitization error was not a problem in the eight species or *Trachemys* examples. There are also no features of the plastron of *E. marmorata* that would make it significantly more difficult to accurately digitize than the plastra of the other speices.

Biological explanations include the possibilty that genetic differentiation is not associated with plastron shape variation and/or that local selective pressures (e.g. from hydrological regime) overwhelm morphological differentiation. Both of these options seem plausible given that shell

shape is influenced by selection for both protection and streamlining, but not necessary mate choice (Polly et al., 2016; Rivera, 2008; Rivera et al., 2014; Rivera and Stayton, 2011; Stayton, 2011), and that shell shape in *E. marmorata* is known to vary among populations inhabiting water bodies with different flow regimes (Germano and Bury, 2009; Holland, 1992; Lubcke and Wilson, 2007). Plastron shape does not seem to preserve a strong phylogenetic signal at the interspecific level in emydine turtles, at least compared to the effect of the presence or absence of a plastral hinge (Angielczyk et al., 2011), and our current results suggest that this may be the case for phylogeographic signal within emydine species as well. A final possibility (explored below) is that the proposed classification schemes themselves do not represent significant evolutionary lineages.

Despite the negative result for *E. marmorata*, it is important to note that plastron shape is an extremely effective method for differentiating classes in the additional datasets we investigated. The magnitude of shape differences between the species (measured as Procrustes distance between the eight species' mean shapes) is approximately an order of magnitude greater than the differences between the *E. marmorata* subgroups, and not surprisingly the machine learning methods had no trouble classifying the specimens correctly. However, the magnitude of the shape differences between the *T. scripta* subspecies is comparable to those separating the different *E. marmorata* subgroups, yet even in this case the machine learning methods returned an almost perfect classification. These results demonstrate that plastron shape is normally a good marker for differentiating real subgroups in close relatives of *E. marmorata*, and that our lack of results for *E. marmorata* is not simply a shortcoming of the methods we applied. Indeed, it begs the question of what factors have suppressed morphological differentiation of plastron shape in *E. marmorata* and *E. pallida* if they are distinct species. Invoking issues such as the role of the plastron in protection or the need for streamlining are insufficient because the other species are expected to be subject to similar constraints (Polly et al., 2016; Stayton, 2011). Although it may seem counterintuitive that plastron shape is both useful for species delimitation but has weak or absent phylogenetic signal, it is important to remember that these are different goals. While phylogenetically similar species may not be morphologically similar (e.g. compare the box turtles of the genus *Terrapene* to the

414 closely related spotted turtle *Clemmys guttata*), the variation within a species typically is much less than the variation between species. Therefore, the consistent plastron shapes that characterize different emydid species leads to plastron shape being a useful tool for species delimitation, even
417 when other selective factors have overprinted similarities stemming from patterns of descent from common ancestors.

Is there more than one species of Western Pond Turtle?

420 The lack of morphological support for the distinctiveness of *E. pallida* does not, on its own, preclude the recognition of this taxon. However, this apparent lack of congruence does prompt a reexamination of the methods and concepts that led to that taxonomic revision, especially
423 considering that plastron shape is demonstrably capable of differentiating species and subspecies among other emydids. In other words, before we can assess the significance of the morphological non-diagnosablity, it is essential to evaluate the methods and concepts that led to the initial
426 taxonomic revision.

Spinks et al. (2014) elevated *E. pallida* based on a species delimitation analysis of SNP data using BPP (Yang and Rannala, 2010). However, Spinks et al. (2014) did not heed the caveats about such
429 species delimitation methods raised by Carstens et al. (2013). In addition to specifically addressing the shortcomings of validation methods such as BPP that rely on guide trees and "should be interpreted with caution," Carstens et al. (2013) also strongly emphasized that "Inferences
432 regarding species boundaries based on genetic data alone are likely inadequate, and species delimitation should be conducted with consideration of the life history, geographical distribution, morphology and behaviour (where applicable) of the focal system…" These caveats evoke
435 the development of the Unified Species Concept (Dayrat, 2005; De Queiroz, 2007), Integrative Taxonomy (Padial et al., 2010), and other pluralist approaches to species delimitation. None of these considerations were brought to bear on the *E. marmorata* system until now, and in doing so
438 we find the proposal that *E. pallida* is a distinct species to be lacking.

In addition to lacking a robust morphological marker, the natural history and geographical distribution of *E. marmorata* and *E. pallida* also make the recognition of these two taxa implausible. The mitochondrial data from Spinks et al. (2014) show extensive introgression and admixture in Central California, which is expected because there are no significant barriers to gene flow in this region. They also lack sampling from the populations between the two putative species in the San Francisco Bay Area, which we predict would likely show even more genetic mixing. Combined with the well-demonstrated ability for testudinoid turtles, including emydids and even *Emys*, to hybridize (e.g. Buskirk et al. 2005; Parham et al. 2013; Spinks and Shaffer 2009) it is hard to imagine how *E. marmorata* and *E. pallida* could maintain their integrity in the face of such admixture. Any argument for the validity of *E. pallida* as a distinct species needs to address these points. Because the geography, natural history, limited sampling from key areas, demonstrated genetic admixture of *E. marmorata*, and comparisons with other morphologically diagnosable species and subspecies conflict with the recognition of *E. pallida*, we hypothesize that *E. pallida* is not a distinct species.

We fully agree with Spinks et al. (2014) that *E. marmorata* (*sensu lato*) is a species deserving of strong conservation efforts, and we do not wish to trivialize this need. Moreover, the genetic diversity uncovered by the analysis of Spinks et al. (2014) should be accounted for explicitly in any conservation plan. Given the apparent lack of morphological distinction combined with the broad range of intergradation and other problems with the species hypothesis outlined above, we recommend that the populations elevated to *E. pallida* by Spinks et al. (2014) are best considered Evolutionary Significant Units or Distinct Population Segments instead of distinct species.

Finally, it is important to note that the data and analyses we present do not let us definitively say whether the apparent lack of morphological divergence within *E. marmorata* truly reflects the presence of a single species, or if it is an artifact of plastron shape being a poor morphological marker for phylogenetic and phylogeographic divergences in the case of *E. marmorata*. This is because we could not carry out our morphometric analyses on the specimens from which the

genetic data were obtained. The comparisons with the other emydid taxa suggest that our negative result is is because *E. marmorata* is a single species. However, tests of both our preferred conclusion (*E. marmorata* as a single species) and that of Spinks et al. (2014) should include morphological and molecular analyses of the same set of voucher specimens, as well as additional tests of species delimitation using alternative methods and corroborating evidence as suggested by Carstens et al. (2013). From a morphological standpoint, support for the validity of "*E. pallida*" may come from other aspects of morphology, such as carapace shape or other features. Likewise, further investigation of the phylogeographic utility of plastron shape in other turtle species will help to clarify whether the lack of differentiation seen in *E. marmoarata*, and the strong differentiation among the other emydids, is typical or an unusual case.

References

Angielczyk, K. D., and C. R. Feldman. 2013. Are diminutive turtles miniaturized? The ontogeny of plastron shape in emydine turtles. Biological Journal of the Linnean Society 108:727–755.

Angielczyk, K. D., C. R. Feldman, and G. R. Miller. 2011. Adaptive evolution of plastron shape in emydine turtles. Evolution 65:377–394.

Angielczyk, K. D., and H. D. Sheets. 2007. Investigation of simulated tectonic deformation in fossils using geometric morphometrics. Paleobiology 33:125–148.

Arnqvist, G., and T. Mårtensson. 1998. Measurement error in geometric morphometrics: Empirical strategies to assess and reduce its impact on measures of shape.

Baird, S. F., and C. Girard. 1852. Descriptions of new species of reptiles collected by the U.S. Exploring Expedition under the command of Capt. Charles Wilkes. Proceedings of the National Academy of Sciences Philadelphia 6:174–177.

Bauer, A. M., J. F. Parham, R. M. Brown, B. L. Stuart, L. Grismer, T. J. Papenfuss, W. Bohme, J. M. Savage, S. Carranza, J. L. Grismer, P. Wagner, A. Schmitz, N. B. Ananjeva, and R. F. Inger. 2000. Availability of new Baysian-delimited gecko names and the importance of character-based species descriptions. Proceedings of the Royal Society B: Biological Sciences 278:490–492.

Baylac, M., C. Villemant, and G. Simbolotti. 2003. Combining geometric morphometrics with pattern recognition for the investigation of species complexes. Biological Journal of the Linnean Society 80:89–98.

Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, and I. Das. 2007. Cryptic species as a window on diversity and conservation. Trends in ecology & evolution 22:148–55.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. Wadsworth International Group, Belmont.

Bury, R. B. 2017. Biogeography of Western Pond Turtles in the western Great Basin: Dispersal Across a Northwest Passage ? Western Wildlife2 4:72–80.

Bury, R. B., D. J. Germano, and G. W. Bury. 2010. Population Structure and Growth of the Turtle Actinemys marmorata from the Klamath–Siskiyou Ecoregion: Age, Not Size, Matters. Copeia 2010:443–451.

Buskirk, S. W., J. F. Parham, and C. R. Feldman. 2005. On the hybridisation between two distantly related Asian turtles (Testudines: Scalia x Mauremys). Salamandra 41:21–26.

Cardini, A., D. Nagorsen, P. O'Higgins, P. D. Polly, R. W. Thorington Jr, and P. Tongiorgi. 2009. Detecting biological distinctiveness using geometric morphometrics: an example case from the Vancouver Island marmot. Ethology Ecology & Evolution 21:209–223.

Carstens, B. C., and T. A. Dewey. 2010. Species Delimitation Using a Combined Coalescent and Information-Theoretic Approach: An Example from North American Myotis Bats. Systematic Biology 59:400–414.

Carstens, B. C., T. a. Pelletier, N. M. Reid, and J. D. Satler. 2013. How to fail at species delimitation. Molecular ecology 22:4369–83.

Caumul, R., and P. D. Polly. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (Marmota, Rodentia). Evolution; international journal of organic evolution 59:2460–72.

Clare, E. L. 2011. Cryptic species? Patterns of maternal and paternal gene flow in eight neotropical bats. PloS one 6:e21460.

Claude, J. 2006. Convergence induced by plastral kinesis and geometric morphometric assessment: a geometric morphometric assessment. Fossil Turtle Research 1:34–45.

Claude, J., E. Paradis, H. Tong, and J. C. Auffray. 2003. A geometric morphometric assessment of

the effects of\nenvironment and cladogenesis on the evolution of the\nturtle shell. Biological Journal of the Linnean Society 79:485–501.

Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85:407–415.

De Queiroz, K. 2007. Species concepts and species delimitation. Systematic Biology 56:879–86.

Demandt, M. H., and S. Bergek. 2009. Identification of cyprinid hybrids by using geometric morphometrics and microsatellites. Journal of Applied Ichthyology 25:695–701.

Dillard, K. C. 2017. A comparative analysis of geometric morphometrics across two Pseudemys turtle species in east central Virginia. Masters. Virginia Commonwealth University.

Dobigny, G., L. Granjon, V. Aniskin, K. Ba, and V. Voloboulev. 2003. A new sigling species of Taterillus (Muridae, Gerbillinae) from West Agrica. Mammalian Biology 68:299–316.

Dryden, I. L. 2013. shapes: Statistical shape analysis. R package version 1.1-8.

Dryden, I. L., and K. Y. Mardia. 1998. Statistical shape analysis. Wiley, New York.

Edwards, S., J. Claude, B. J. Van Vuuren, and C. A. Matthee. 2011. Evolutionary history of the Karoo bush rat, Myotomys unisulcatus (Rodentia: Muridae): Disconcordance between morphology and genetics. Biological Journal of the Linnean Society 102:510–526.

Eldredge, N., and S. J. Gould. 1972. Punctuated equilibria: an alternative to phyletic gradualism. Pages 82–115 *in* T. J. M. Schopf, ed. Models in Paleobiology. Freeman Cooper, San Francisco.

Feldman, C. R., and J. F. Parham. 2002. Molecular phylogenetics of emydine turtles: taxonomic revision and the evolution of shell kinesis. Molecular Phylogenetics and Evolution 22:388–98.

Francoy, T. M., R. A. O. Silva, P. Nunes-Silva, C. Menezes, and V. L. Imperatriz-Fonseca. 2009. Gender identification of five genera of stingless bees (Apidae, Meliponini) based on wing morphology. Genetics and molecular research 8:207–214.

Fruciano, C., P. Franchini, F. Raffini, S. Fan, and A. Meyer. 2016. Are sympatrically speciating Midas cichlid fish special? Patterns of morphological and genetic variation in the closely related species Archocentrus centrarchus. Ecology and Evolution 6:4102–4114.

Funk, W. C., M. Caminer, and S. R. Ron. 2012. High levels of cryptic species diversity uncovered in Amazonian frogs. Proceedings of the Royal Society B: Biological Sciences 279:1806–14.

Gaubert, P., P. J. Taylor, C. a. Fernandes, M. W. Bruford, and G. Veron. 2005. Patterns of cryptic hybridization revealed using an integrative approach: a case study on genets (Carnivora, Viverridae, Genetta spp.) from the southern African subregion. Biological Journal of the Linnean Society 86:11–33.

Germano, D. J., and R. B. Bury. 2009. Variation in body size, growth, and population structure of Actinemys marmorata from lentic and lotic habitats in Southern Oregon. Journal of Herpetology 43:510–520.

Germano, D. J., and G. B. Rathbun. 2008. Growth, population structure, and reproduction of western pond turtles (Actinemys marmorata) on the Central Coast of California. Chelonian Conservation and Biology 7:188–194.

Gould, S. J., and N. Eldredge. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

Guillot, G., S. Renaud, R. Ledevin, J. Michaux, and J. Claude. 2012. A unifying model for the analysis of phenotypic, genetic, and geographic data. Systematic Biology 61:897–911.

Gündüz, I., M. Jaarola, C. Tez, C. Yeniyurt, P. D. Polly, and J. B. Searle. 2007. Multigenic and morphometric differentiation of ground squirrels (Spermophilus, Scuiridae, Rodentia) in Turkey, with a description of a new species. Molecular phylogenetics and evolution 43:916–35.

Hand, D. J., and R. J. Till. 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45:171–186.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer, New York.

Hastie, T., R. Tibshirani, F. Leisch, K. Hornik, and B. D. Ripley. 2015. mda: Mixture and Flexible Discriminant Analysis. R package version 0.4-8.

Hausdorf, B., and C. Hennig. 2010. Species delimitation using dominant and codominant multilocus markers. Systematic biology 59:491–503.

Holland, D. C. 1992. Level and pattern in morphological variation: a phylogeographic study of the western pond turtle (Clemmys marmorata). Ph.D. thesis. University of Southwestern Louisiana.

Huelsenbeck, J. P., P. Andolfatto, and E. T. Huelsenbeck. 2011. Structurama: bayesian inference of population structure. Evolutionary bioinformatics online 7:55–9.

Kaufman, L., and P. J. Rousseeuw. 1990. Finding groups in data : an introduction to cluster analysis. Wiley, New York.

Kendall, D. G. 1977. The diffusion of shape. Advances in Applied Probability 9:428–430.

Kuhn, M. 2013. caret: Classification and Regression Training. R package version 5.15-61.

Kuhn, M., and K. Johnson. 2013. Applied predictive modeling. Springer, New York, NY.

Leaché, A. D., and M. K. Fujita. 2010. Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). Proceedings. Biological sciences / The Royal Society 277:3071–7.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomforest. R News 2:18–22.

Lubcke, G. M., and D. S. Wilson. 2007. Variation in shell morphology of the Western Pond Turtle (Actinemys marmorata Baird and Giarard) from three aquativ habitats in Northern California. Journal of Herpetology 41:107–114.

MacLeod, N. 2007. Automated taxon identification in systematics: theory, approaches and applications. CRC Press, Boca Raton.

Markolf, M., H. Rakotonirina, C. Fichtel, P. von Grumbkow, M. Brameier, and P. M. Kappeler. 2013. True lemurs... true species - species delimitation using multiple data sources in the brown lemur complex. BMC Evolutionary Biology 13:233.

Mitrovski-Bogdanovic, A., A. Petrovic, M. Mitrovic, A. Ivanovic, V. Žikic, P. Starý, C. Vorburger, and Ž. Tomanovic. 2013. Identification of two cryptic species within the Praon abjectum group (Hymenoptera: Braconidae: Aphidiinae) using molecular markers and geometric morphometrics. Annals of the entomological society of America 106:170–180.

Munoz-Munoz F., and Perpinan D. 2010. Measurement error in morphometric studies: comparison between manual and computerized methods. Ann. Zool. 47:46–56.

Navega, D., R. Vicente, D. N. Vieira, A. H. Ross, and E. Cunha. 2015. Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. International Journal of Legal Medicine 129:651–659.

O'Meara, B. C. 2010. New heuristic methods for joint species delimitation and species tree inference. Systematic biology 59:59–73.

Padial, J. M., A. Miralles, I. De la Riva, and M. Vences. 2010. The integrative future of taxonomy. Frontiers in Zoology 7:1–14.

Parham, J. F., T. J. Papenfuss, P. P. V. Dijk, B. S. Wilson, C. Marte, L. R. Schettino, and W. Brian Simison. 2013. Genetic introgression and hybridization in Antillean freshwater turtles (Trachemys) revealed by coalescent analyses of mitochondrial and cloned nuclear markers. Molecular phylogenetics and evolution 67:176–87.

Pfenninger, M., and K. Schwenk. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC evolutionary biology 7:121.

Polly, P. D. 2003. Paleophylogeography of Sorex araneus: molar shape as a morphological marker for fossil shrews. Mammalia 68:233–243.

———. 2007. Phylogeographic differentiation in Sorex araneus: morphology in relation to geography and karyotype. Russian Journal of Theriology 6:73–84.

Polly, P. D., C. T. Stayton, E. R. Dumont, S. E. Pierce, E. J. Rayfield, and K. D. Angielczyk. 2016. Combining geometric morphometrics and finite element analysis with evolutionary modeling: towards a synthesis. Journal of Vertebrate Paleontology 4634.

Pons, J., T. Barraclough, J. Gomez-Zurita, A. Cardoso, D. Duran, S. Hazell, S. Kamoun, W. Sumlin, and A. Vogler. 2006. Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. Systematic Biology 55:595–609.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rivera, G. 2008. Ecomorphological variation in shell shape of the freshwater turtle Pseudemys concinna inhabiting different aquatic flow regimes. Integrative and comparative biology 48:769–87.

Rivera, G., J. N. Davis, J. C. Godwin, and D. C. Adams. 2014. Repeatability of Habitat-Associated Divergence in Shell Shape of Turtles. Evolutionary Biology pages 29–37.

Rivera, G., and C. T. Stayton. 2011. Finite element modeling of shell shape in the freshwater turtle Pseudemys concinna reveals a trade-off between mechanical strength and hydrodynamic efficiency. Journal of morphology 272:1192–203.

Rohlf, F. J. 2005. TpsDig 2.04.

Schilck-Steiner, B. C., B. Seifert, C. Stauffer, E. Christian, R. H. Crozier, and F. M. Steiner. 2007. Without morphology, cryptic species stay in taxonomic crypsis following discovery. Trends in ecology & evolution 22:391–392.

Seeliger, L. M. 1945. Variation in the Pacific Mud Turtle. Copeia 1945:150–159.

639 Spinks, P. Q., and H. B. Shaffer. 2005. Range-wide molecular analysis of the western pond turtle (Emys marmorata): cryptic variation, isolation by distance, and their conservation implications. Molecular ecology 14:2047–64.

642 ———. 2009. Conflicting mitochondrial and nuclear phylogenies for the widely disjunct Emys (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. Systematic biology 58:1–20.

645 Spinks, P. Q., R. C. Thomson, and H. Bradley Shaffer. 2014. The advantages of going large: genome wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. Molecular Ecology pages n/a–n/a.

648 Spinks, P. Q., R. C. Thomson, and H. B. Shaffer. 2010. Nuclear gene phylogeography reveals the historical legacy of an ancient inland sea on lineages of the western pond turtle, Emys marmorata in California. Molecular ecology 19:542–56.

651 Stayton, C. T. 2011. Biomechanics on the half shell: functional performance influences patterns of morphological variation in the emydid turtle carapace. Zoology (Jena, Germany) 114:213–23.

Stuart, B. L., R. F. Inger, and H. K. Voris. 2006. High level of cryptic species diversity revealed by 654 sympatric lineages of Southeast Asian forest frogs. Biology letters 2:470–4.

Sztencel-Jabłonka, A., G. Jones, and W. BogdanowicZ. 2009. Skull Morphology of Two Cryptic Bat Species: Pipistrellus pipistrellus and P. pygmaeus — A 3D Geometric Morphometrics Approach 657 with Landmark Reconstruction. Acta Chiropterologica 11:113–126.

Van Bocxlaer, B., and G. Hunt. 2013. Morphological stasis in an ongoing gastropod radiation from Lake Malawi. Proceedings of the National Academy of Sciences .

660 Van Bocxlaer, B., and R. Schultheiß. 2010. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine-learning approaches to biological discrimination. Paleobiology 36:497–515.

van den Brink, V., and F. Bokma. 2011. Morphometric shape analysis using learning vector quantization neural networks — an example distinguishing two microtine vole species. Annales Zoologici Fennici 48:359–364.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. 4th ed. Springer, New York.

Vitek, N. S., C. L. Manz, T. Gao, J. I. Bloch, S. G. Strait, and D. M. Boyer. 2017. Semi-supervised determination of pseudocryptic morphotypes using observer-free characterizations of anatomical alignment and shape. Ecology and Evolution 7:5041–5055.

von Cramon-Taubadel, N., B. C. Frazier, and M. M. Lahr. 2007. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. American journal of physical anthropology 132:535–544.

Yang, Z., and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences 107:9264–9.

Zelditch, M. L., D. L. Swiderski, and H. D. Sheets. 2004. Geometric morphometrics for biologists: a primer. Elsevier Academic Press, Amsterdam.

Tables

Table 1: Table of species delimitation hypotheses for *E. marmorata*

Abbreviation	Number of classes	citation
SP10.1	3	Spinks et al. (2010)
SP10.2	3	Spinks et al. (2010)
SP10.3	4	Spinks et al. (2010)
SP14.1	2	Spinks et al. (2014)
SP14.2	4	Spinks et al. (2014)
Morph	2	Spinks et al. (2010)

Table 2: Table of the supervised learning methods used in this analysis.

Method name	abbreviation	R package	citation
multinomial logistic regression	MLR	nnet	Venables and Ripley (2002)
linear discriminate analysis	LDA	MASS	Venables and Ripley (2002)
penalized discrminiate analysis	PDA	mda	Hastie et al. (2015)
single-hidden-layer neural network	NN	nnet	Venables and Ripley (2002)
random forests	RF	randomForest	Liaw and Wiener (2002)

Table 3: Results from the within-individual to between-individual Procrustes distances for the replicated plastron shape data. Results are presented for all three datasets analyzed here: the *Trachemys* dataset, the eight species dataset, and each of the *Emys marmorata* classification schemes.

Dataset	**Scheme**	**Ratio**	**Corrected ratio**
Replicates		1.11	
Seven species		0.33	0.37
Trachemys		0.76	0.84
E. marmorata	SP10.1	0.99	1.10
	SP10.2	1.00	1.11
	SP10.3	0.94	1.04
	SP14.1	1.01	1.12
	SP14.2	0.93	1.04
	Morph	0.99	1.09

Figure legends

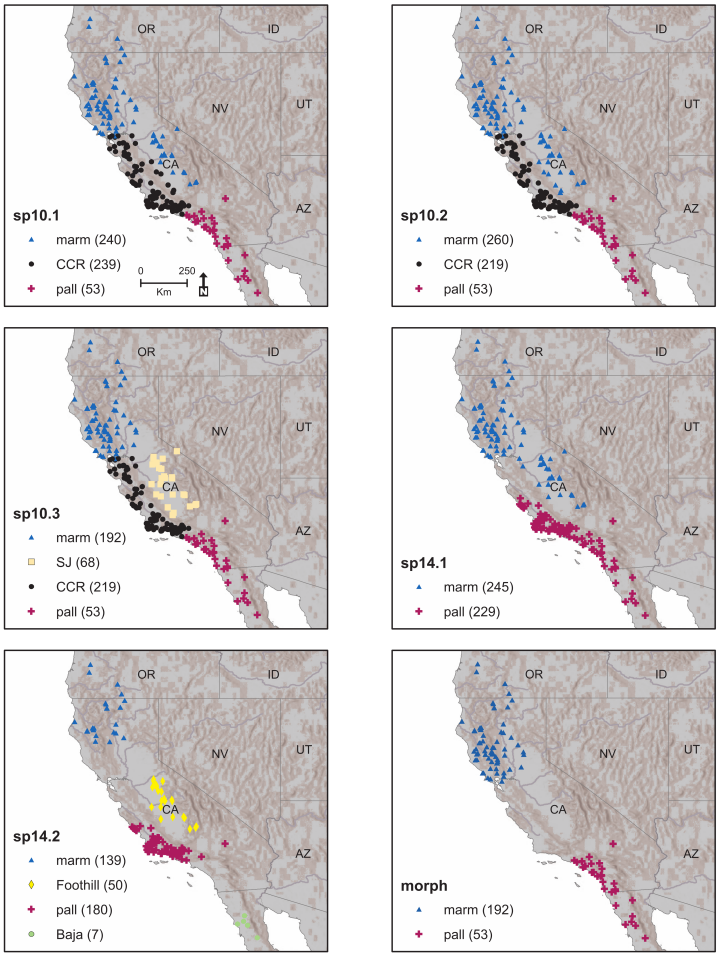

Figure 1: Geographic distribution of specimens sampled for comparing the hypothesized subdivisions of *Emys marmorata*. Each hypothesized scheme has two or more possible classes. Sample size differs between schemes because of our ability to confidently assign museum specimens to the various schemes. The number of localities shown on each map is less than the number of specimens sampled because some localities produced multiple specimens. The different classification abbreviations are as follows: *E. marmorata* = "marm", *E. pallida* = "pall", Central Coast Ranges = "CCR", San Joaquin Valley = "SJ," Baya Peninsula = "Baja,", and Sierra Foothills = "Foothill."

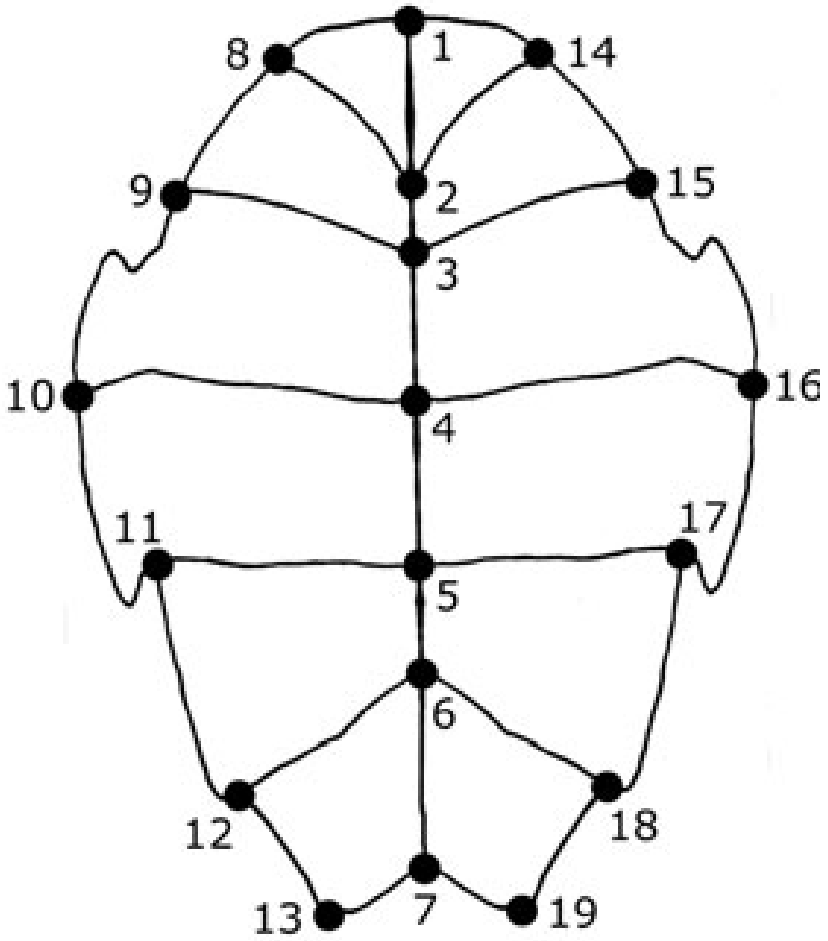

Figure 2: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmarks used in this study. Anterior is towards the top of the figure.

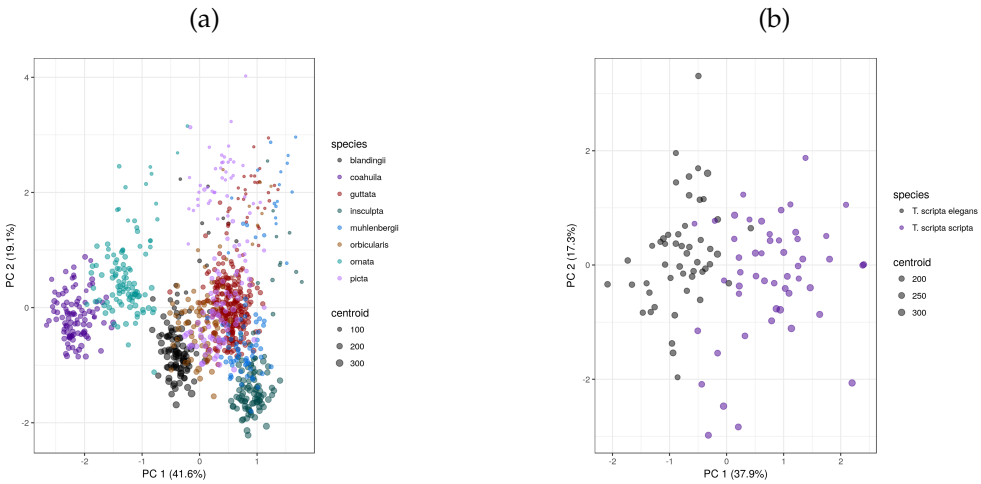

Figure 3: Two scatterplots of morphological differences from two of the three datasets analyzed in this study. (a) Scatterplot of the first two PCA axies from the landmarks from the eight different species dataset, and (b) the first two axes of variation from two subspecies of *Trachemys* dataset. Point colors correspond to the categories within each dataset while point size is proportional to individual centroid size. In parentheses next to the axis labels are the percent of total variation accounted for by that axis. For both datasets there are clear distinctions between the different categories.

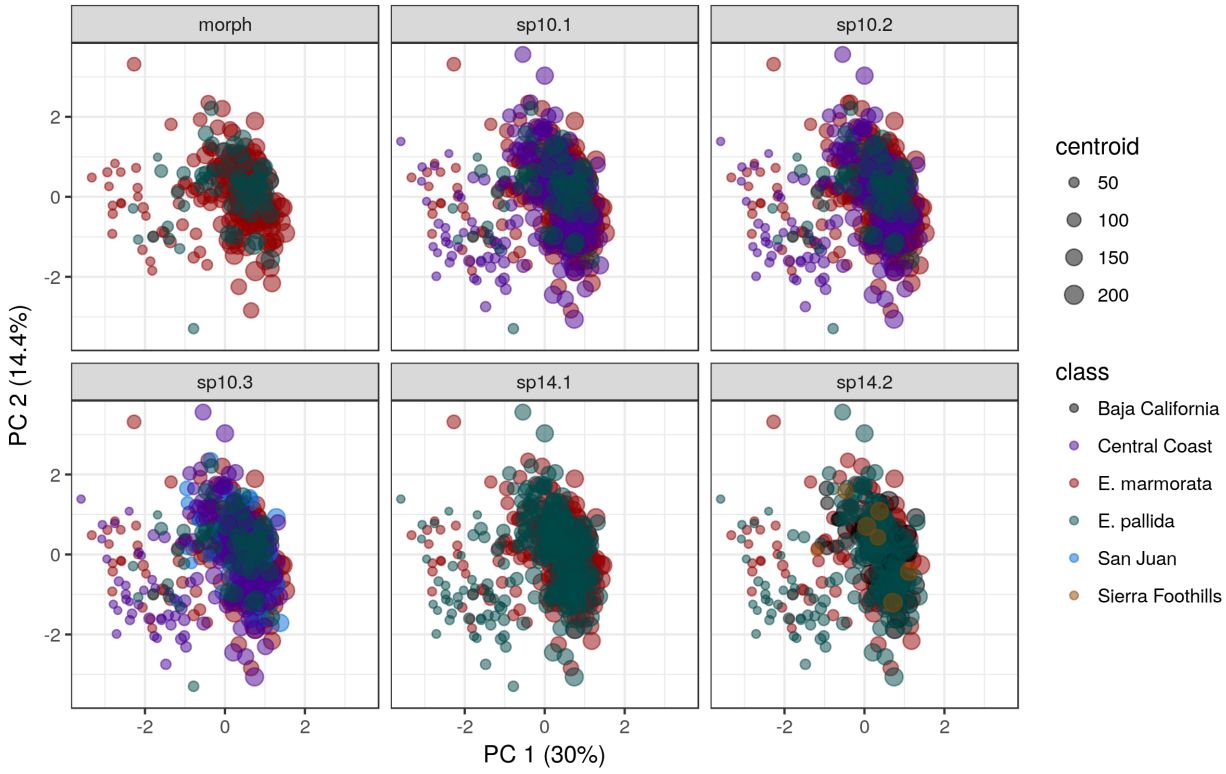

Figure 4: Scatterplot of the first two axes of morphological variation in the *Emys marmorata* dataset. Each panel corresponds to one of the six different classification schemes analyzed as part of this study (Tab. 1). Point color corresponds to the categories within each scheme, and the class names correspond to geographic regions. Point size is proportional to centroid size of that specimen and the numbers in parentheses next to the axis labels are the percent of total variation accounted for along that axis.

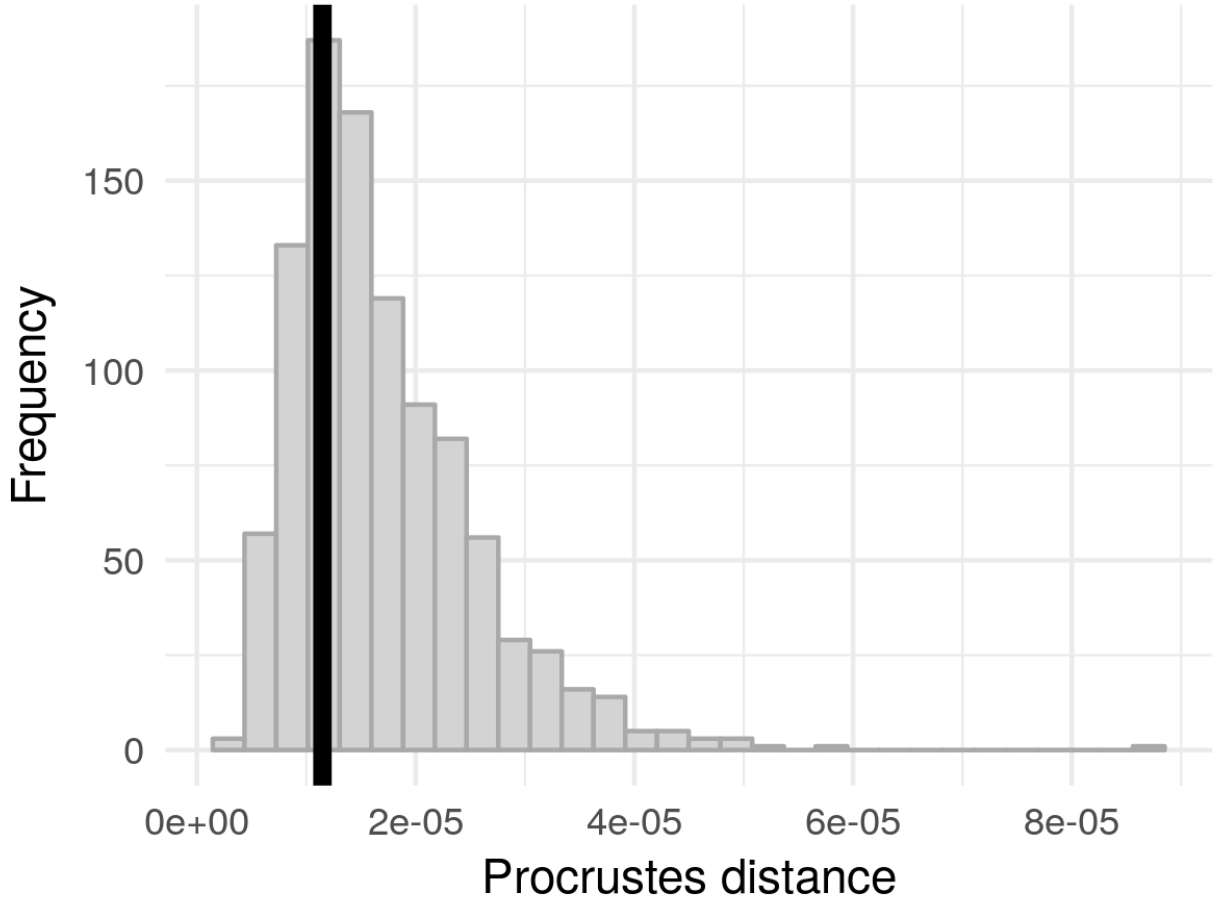

Figure 5: Comparison of observed Procrustes distance between the centroids of each sex (vertical line) to a null distribution generated from 1000 permutations of the sex-labels. This result indicates that the difference between the centroids is as small/smaller than expected by random.

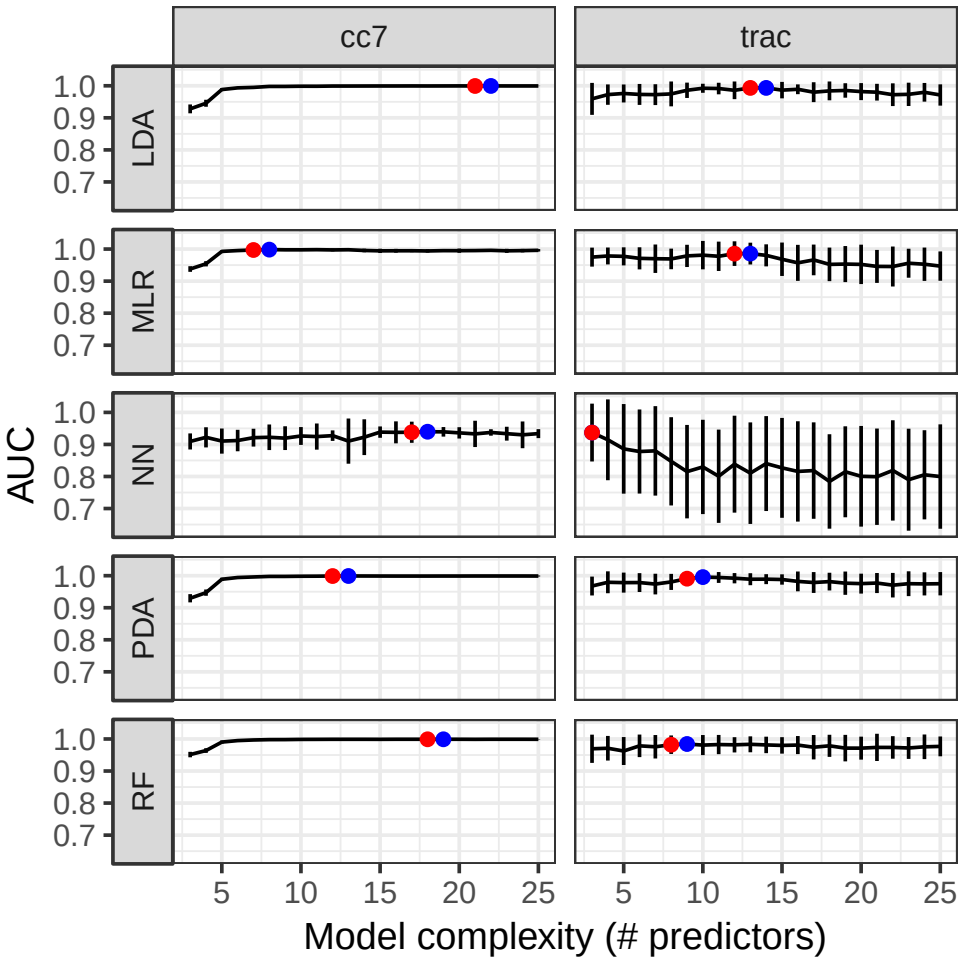

Figure 6: Comparisons of model fit to the training dataset for each of the supervised learning methods applied to the first two datasets; the results from the eight species dataset are presented in the left column, while those from the *Trachemys* dataset are presented in the right column. Models were fit to datasets of varying complexity, with the number of parameters listed along the x-axis. Model fit is measured as the area under the receiver operating characteristic (AUC), which ranges from 0.5 to 1. Error bars correspond to one standard error estimated from 10 rounds of 5-fold cross-validation. The red dot corresponds to the model fit with the highest mean AUC while the blue dot corresponds to the model selected for further analysis. In some cases, there is no difference in complexity between the best and selected models.

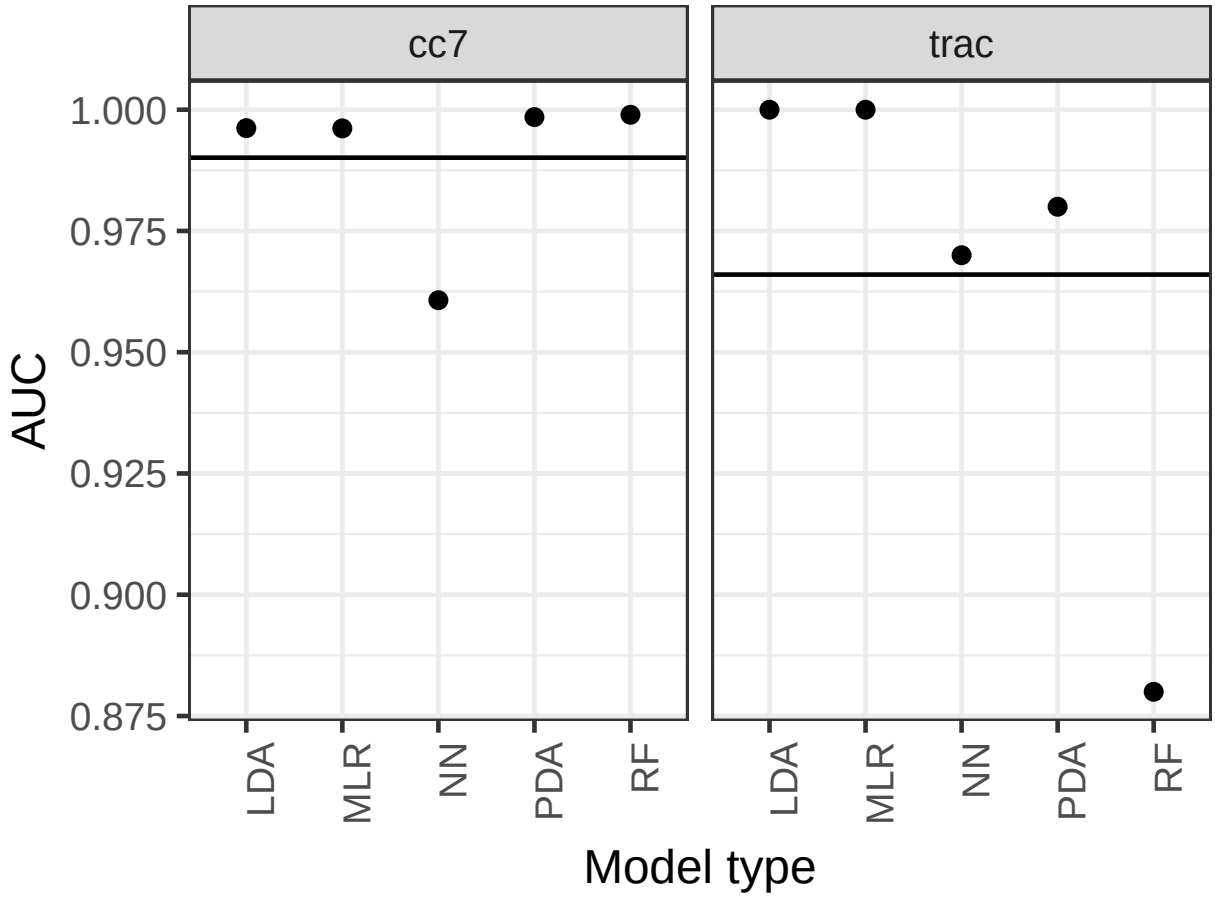

Figure 7: The results of out-of-sample predictive performance of the selected models for both the eight species (left) and *Trachemys* datasets. Predictive performance is measured as the area under the receiver operating characteristic (AUC), which ranges from 0.5 to 1. Points correspond to the individual out-of-sample predictive performance of the specific model, indicated along the x-axis. The horizontal bars correspond the average out-of-sample predictive performance of all the models.

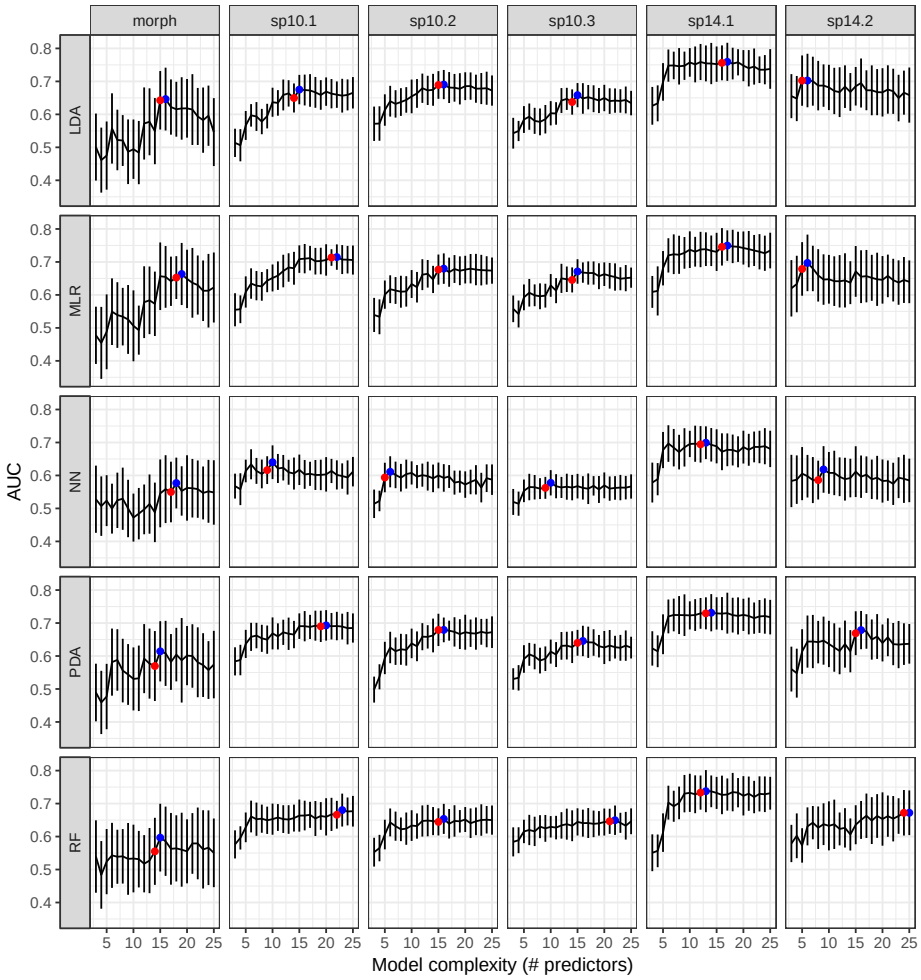

Figure 8: AUC values for models of varying complexity fit to the *Emys marmorata* training datasets for each classification scheme. The x-axis corresponds to the total number of predictors included in each model, while the y-axis corresponds to the AUC value which is a measure of goodness of fit for classification datasets. A model with a high AUC value corresponds to better classification performance than a model with a lower AUC value. Standard errors on AUC estimates are calculated from 10 rounds of 5-fold cross-validation. Indicated are the best performing and the selected models, in red and blue respectively. In some cases, there is no difference in complexity between the best and selected models.

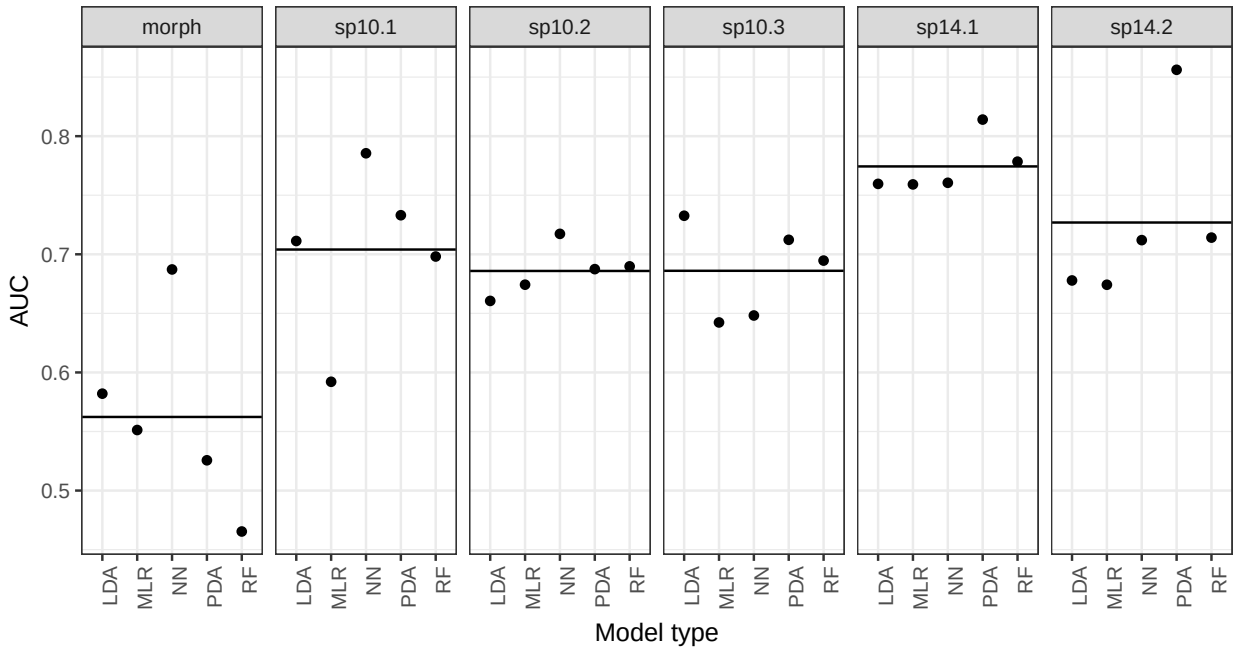

Figure 9: Comparison of out-of-sample AUC estimates from the predictions of selected models (Fig. 8), grouped by classification scheme. The horizontal line in each panel corresponds to the average AUC value across all models of that classification scheme.