

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

Peter D Smits¹, Kenneth D Angielczyk², James F Parham³, and Bryan L Stuart⁴

¹Department of Integrative Biology, University of California – Berkeley

²Integrative Research Center, Field Museum of Natural History

³John D. Cooper Archaeological and Paleontological Center, Department of Geological Sciences, California State University, Fullerton

⁴Section of Research and Collections, North Carolina Museum of Natural Sciences

December 4, 2017

Corresponding author: Peter D Smits, Department of Integrative Biology, University of California – Berkeley, 3040 Valley Life Sciences Building #5151, Berkeley, CA, 94720, USA; E-mail: psmits@berkeley.edu

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 recorded in museum collection archives. 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. In contrast, the analyses of the clear-cut examples produced near perfect classifications, demonstrating that the methods can recover correct results when an appropriate signal

exists. 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 (Stuart et al. 2006; Bickford et al. 2007; Schilck-Steiner et al. 2007; Pfenninger and Schwenk 2007; Clare 2011; Funk et al. 2012). In attempts to streamline the documentation of biodiversity, several methods of species delimitation that rely almost entirely on genetic data have recently been proposed (Pons et al. 2006; Carstens and Dewey 2010; Hausdorf and Hennig 2010; O’Meara 2010; Yang and Rannala 2010; Huelsenbeck et al. 2011). 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 (Polly 2003; Zelditch et al. 2004; Gaubert et al. 2005; Gündüz et al. 2007; Polly 2007; Demandt and Bergek 2009; Markolf et al. 2013; Fruciano et al. 2016). 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

54 shape of eight species of closely related turtles, plastron shape of two subspecies of a single
55 turtle species, and plastron shape of the *Emys marmorata* species complex. In particular, we
56 ask whether it is possible to determine which among a set of classification hypotheses best
57 aligns with the observed morphology, and examine the implications of our results for the *E.
58 marmorata* complex.

Background and study system

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

68 There are two major classes of machine learning method: unsupervised and supervised learning.
69 Unsupervised learning methods are used with unlabeled data where the underlying structure
70 is estimated; they are analogous to clustering and density estimation methods (Kaufman and
71 Rousseeuw 1990). Supervised learning methods are used with labeled data where the final
72 output of data is known and the rules for going from input to output are inferred. These are
73 analogous to classification and regression models (Breiman et al. 1984; Hastie et al. 2009).

74 Our application of the supervised learning approaches used in this study illustrates only
75 a sampling of the various methods available for fitting classification models. The specific
76 methods used in this study were chosen because they are suited for cases with more two or
77 more response classes.

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

94 Here, we used an ensemble of supervised machine learning methods to compare the congruence
96 of morphological data to different classification hypotheses. Each of these methods provide
98 different advantages for understanding how to classify specimens, as well as the accuracy of
100 the resulting classifications. Machine learning methods have been combined with geometric
102 morphometric data to study shape variation in a variety of contexts, including automated
104 taxon identification and classification of groups (Baylac et al. 2003; Dobigny et al. 2003;
MacLeod 2007; Van Bocxlaer and Schultheiß 2010; van den Brink and Bokma 2011; Navega
et al. 2015). 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.

118 Previous work on morphological variation in *E. marmorata* has focused primarily on differentiation
120 between populations over a portion of the species' total range (Lubcke and Wilson
2007; Germano and Rathbun 2008; Germano and Bury 2009; Bury et al. 2010); comparatively
few studies have included specimens from across the entire range (Holland 1992). Most of
122 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
124 (Lubcke and Wilson 2007; Germano and Rathbun 2008; Germano and Bury 2009). There also
has been interest in size-based sexual dimorphism in *E. marmorata* (Holland 1992; Lubcke
126 and Wilson 2007; Germano and Bury 2009), with males being on average larger than females
based on total carapace length and other linear measurements. However, the quality of size
128 as a classifier of sex can vary greatly between populations (Holland 1992) because of the
magnitude of size differences among populations (Lubcke and Wilson 2007; Germano and
130 Bury 2009). The effect of sexual dimorphism on shape, *sensu* Kendall (1977), has not been
assessed (Holland 1992; Lubcke and Wilson 2007; Germano and Rathbun 2008).

132 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
136 features such as breaks between different drainage basis are probably more important barriers
to dispersal and interbreeding. Additionally, he suggested that morphological differences were
138 more pronounced as the magnitude of barriers and distance increased, but this variation
required many variables to adequately capture, implying only very subtle morphological
140 differentiation between putatively distinct populations. Finally, Holland concluded that *E.
marmorata* is best classified as three distinct species: a northern species, a southern species,
142 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.

144 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. 2010, 2014).
146 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
148 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
150 these datasets differed from that of mitochondrial-based results of Spinks and Shaffer (2005) in
the location of the break point (Spinks et al. 2014). All three studies discussed the potential
152 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
154 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
156 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
158 this taxonomy, a geometric morphometric analysis of shell shape might provide a reliable
way to diagnose groups (whether species or subspecies) within *E. marmorata*.

160 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
162 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
164 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
166 be analyzed together. Second, previous work has suggested that there are strong differences
in plastron shape among traditionally-recognized emydine species (Angielczyk and Sheets
168 2007; Angielczyk et al. 2011; Angielczyk and Feldman 2013). Finally, due to these previous
studies a large dataset was readily available.

170 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
172 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
174 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
176 unclear geographic boundaries between subgroups of *E. marmorata*, we compare multiple permutations of the (Spinks et al. 2010) and Spinks et al. (2014) hypotheses.

178 **MATERIALS AND METHODS**

Specimens, sampling, morphometrics

180 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
182 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
184 represents a low complexity dataset because of the high level of plastron shape disparity that exists among these species (Claude et al. 2003; Claude 2006; Angielczyk et al. 2011), whereas
186 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
188 complexity and our greatest challenge given the finding that only very subtle differences existed between geographically-distinct populations (Holland 1992).

190
192 The first dataset we analyzed includes 578 total specimens from the following species: *Chrysemys picta*, *Clemmys guttata*, *Emys blandigii*, *Emys orbicularis*, *Glyptemys insculpta*,
194 *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).

196 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
198 are new to this study.

200 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
202 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 voucherized specimens we were not able to directly sample the
204 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,
206 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
208 represented each hypothesis with multiple possible realizations representing the classification uncertainty for specimens present at the geographic boundaries. The taxonomic hypotheses
210 and sub-hypotheses for *E. marmorata* used here are presented in Table 1 and Figure 1.

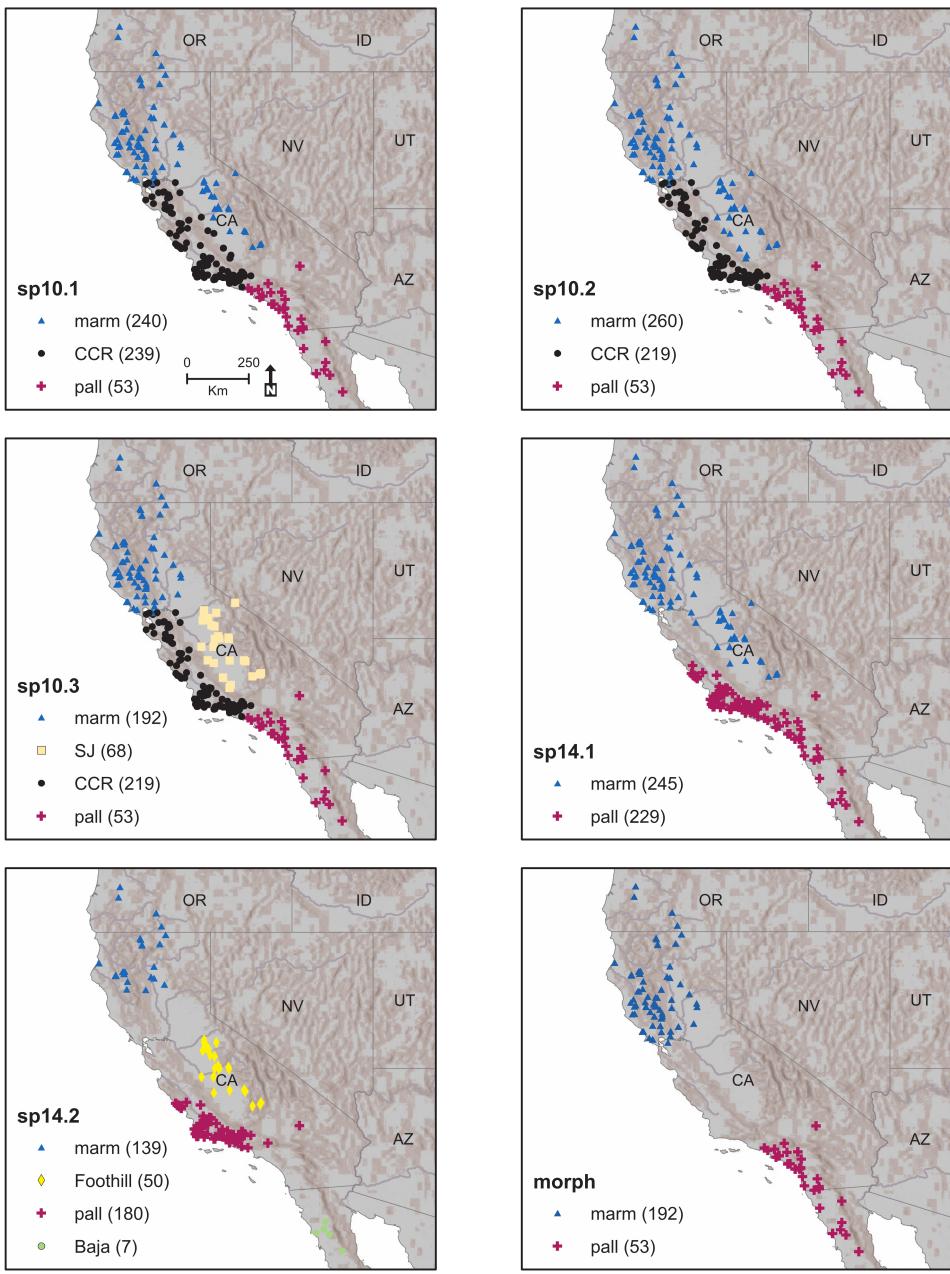


Figure 1: Geographic distribution of specimens sampled for comparing the hypothesized subdivisions of *Emys marmorata*. Each hypothesized scheme has two or more possible classes. Sample size differs between schemes because of our ability to confidently assign museum specimens to the various schemes. 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.”

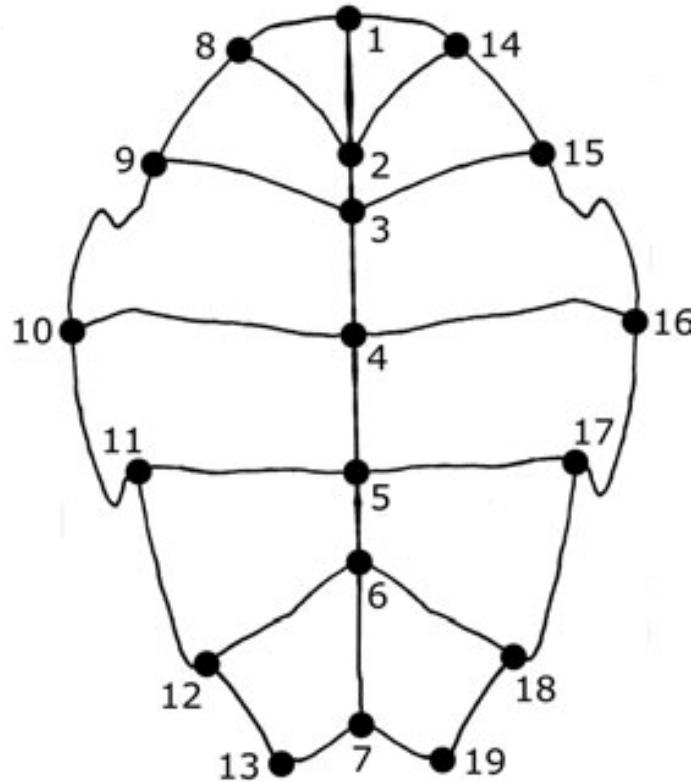


Figure 2: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmarks used in this study. Anterior is towards the top of the figure.

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)

For Spinks et al. (2010) we used three binning schemes. All three schemes include a class
 214 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
 216 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
 218 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
 220 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
 222 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
 224 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
 226 morphological hypothesis (“Morph”) in order to compare the molecular hypotheses with
 something approximating the original taxonomic hypothesis for the group; this scheme is
 228 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
 230 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
 232 be for our analysis of the *E. marmorata* species complex (see below).

Following previous work on plastron shape (Angielczyk and Sheets 2007; Angielczyk et al.
 234 2011; Angielczyk and Feldman 2013), we used TpsDig 2.04 (Rohlf 2005) to digitize 19
 two-dimensional landmarks (Fig. 2). Seventeen of the landmarks are at the endpoints or
 236 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
 238 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
 240 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
 242 on the resulting “half” plastra. We superimposed the plastral landmark configurations using
 generalized Procrustes analysis (Dryden and Mardia 1998), after which we calculated the

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 discriminant analysis	LDA	MASS	Venables and Ripley (2002)
penalized discriminant 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)

²⁴⁴ principal components (PC) of shape using the **shapes** package for R (R Core Team 2016; Dryden 2013). All specimens were used for superimposition, after which the subset labeled
²⁴⁶ for each of the schemes were used in model training and testing (see below).

Biasing effects

²⁴⁸ We estimated the possible effect of digitization error (Arnqvist and Mårtensson 1998; von Cramon-Taubadel et al. 2007; Munoz-Munoz F. and Perpinan D. 2010) on our results
²⁵⁰ by comparing within-specimen (replicated) Procrustes distances to the distances between classification scheme centroids. Ten randomly-selected *E. marmorata* specimens were each
²⁵² 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
²⁵⁴ Procrustes distances was then calculated as the average of the pairwise distances between each of the replicate configurations of a given specimen.

²⁵⁶ 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
²⁵⁸ 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
²⁶⁰ 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
²⁶² 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
²⁶⁴ to the data.

Emys marmorata is known to display sexual dimorphism in plastral shape, particularly the
²⁶⁶ presence of a plastral concavity in males (Seeliger 1945). To test for biases resulting from sexual
²⁶⁸ 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
²⁷⁰ associated with them, this analysis was done using a subset of the data (257 of 532).

Supervised learning approaches

²⁷² 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 (Seeliger 1945; Holland 1992). 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 k th 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.

330

RESULTS

Geometric morphometrics

332 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).

336 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 338 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.

340 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 342 sexes is very small; the sexes overlap about has much as expected given a null distribution based on permuting the sex-labels.

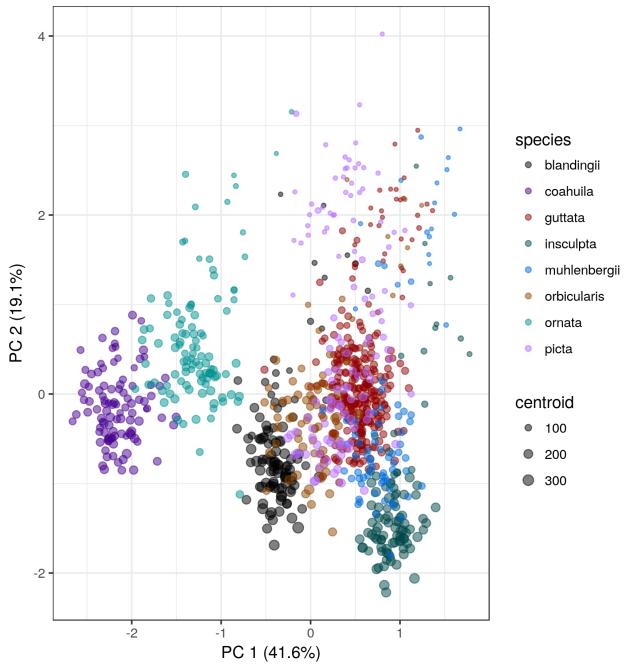
344 Comparison of the within to between Procrustes distances of the digitization replicates gives 346 an approximate estimate of the error between distinct groupings (Table 3). The ratio of the average within-individual distance to the average distance between individuals for the replicated datasets is 1.11; this indicates that the grouping is slightly counter-intuitive to 348 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, 350 where other ratios are compared to the correction ratio 1.11/1.

352 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: 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
<i>Trachemys</i>		0.76	0.84
<i>E. marmorata</i>	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

(a)



(b)

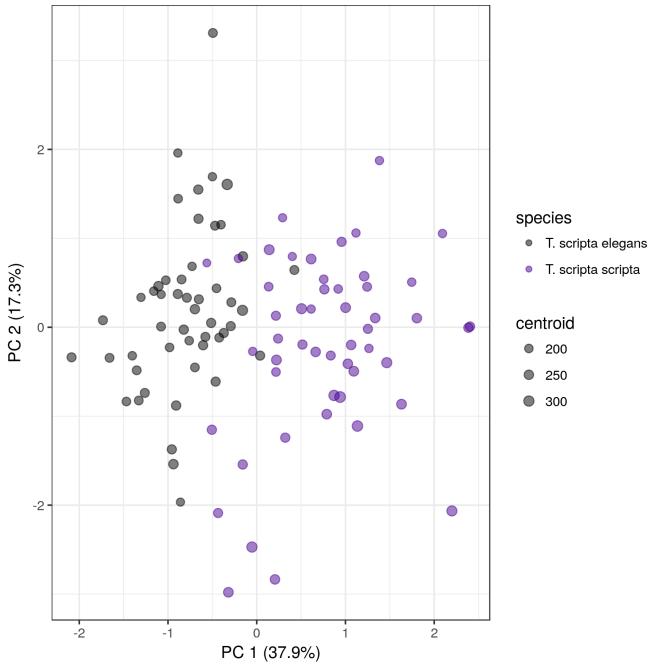


Figure 3: Two scatterplots of morphological differences from two of the three datasets analyzed in this study. (a) Scatterplot of the first two PCA axes from the landmarks from the eight different species dataset, and (b) the first two axes of variation from two subspecies of *Trachemys* dataset. Point 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.

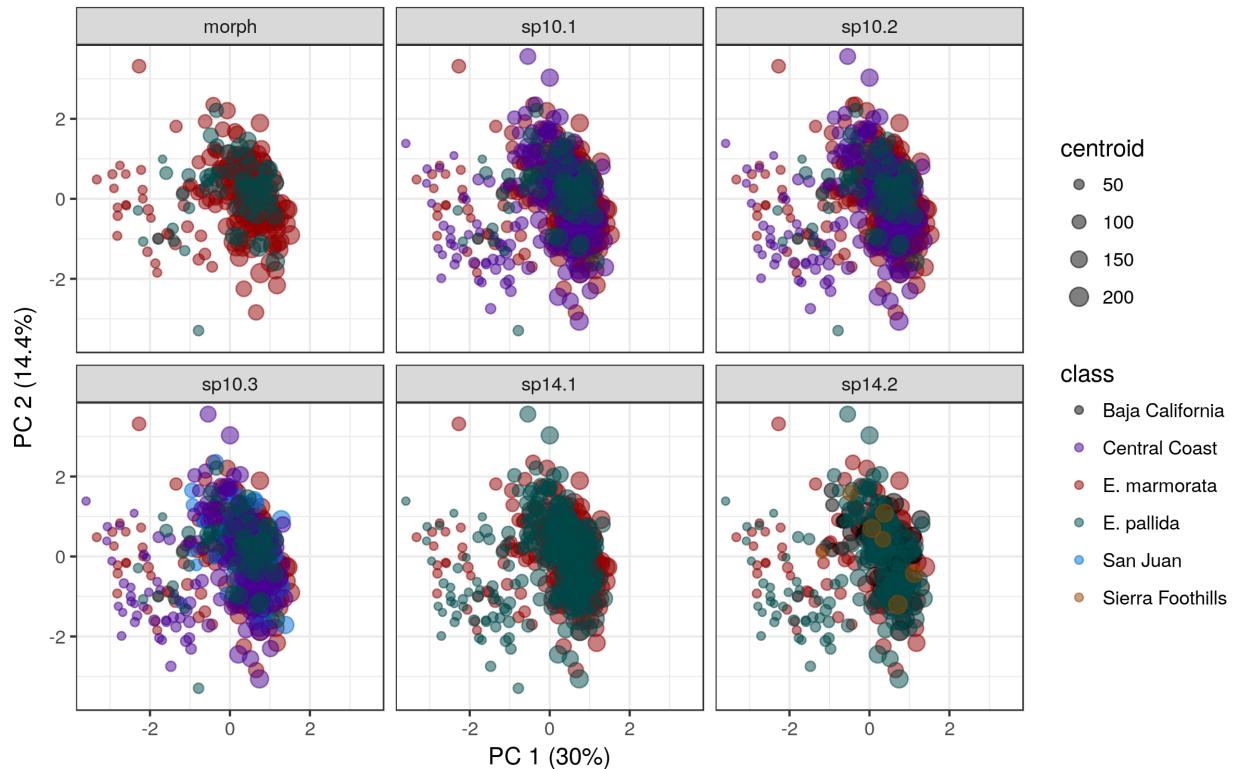


Figure 4: Scatterplot of the first two axes of morphological variation in the *Emys marmorata* dataset. Each panel corresponds to one of the six different classification schemes analyzed as part of this study (Tab. 1). Point color corresponds to the categories within each scheme, and the class names 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.

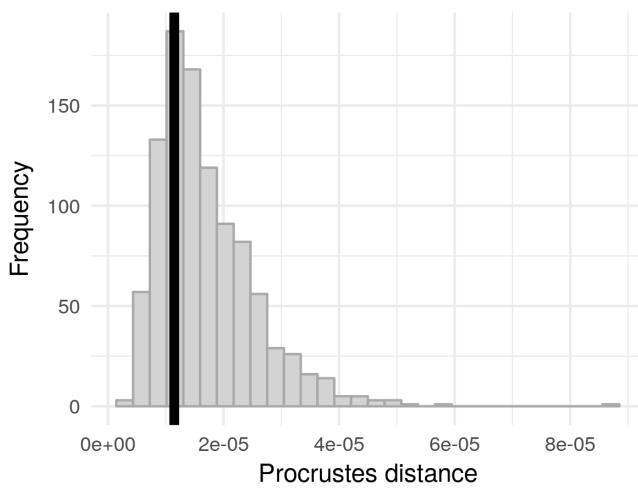


Figure 5: Comparison of observed Procrustes distance between the centroids of each sex (vertical line) to a null distribution generated from 1000 permutations of the sex-labels. This result indicates that the difference between the centroids is as small/smaller than expected by random.

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

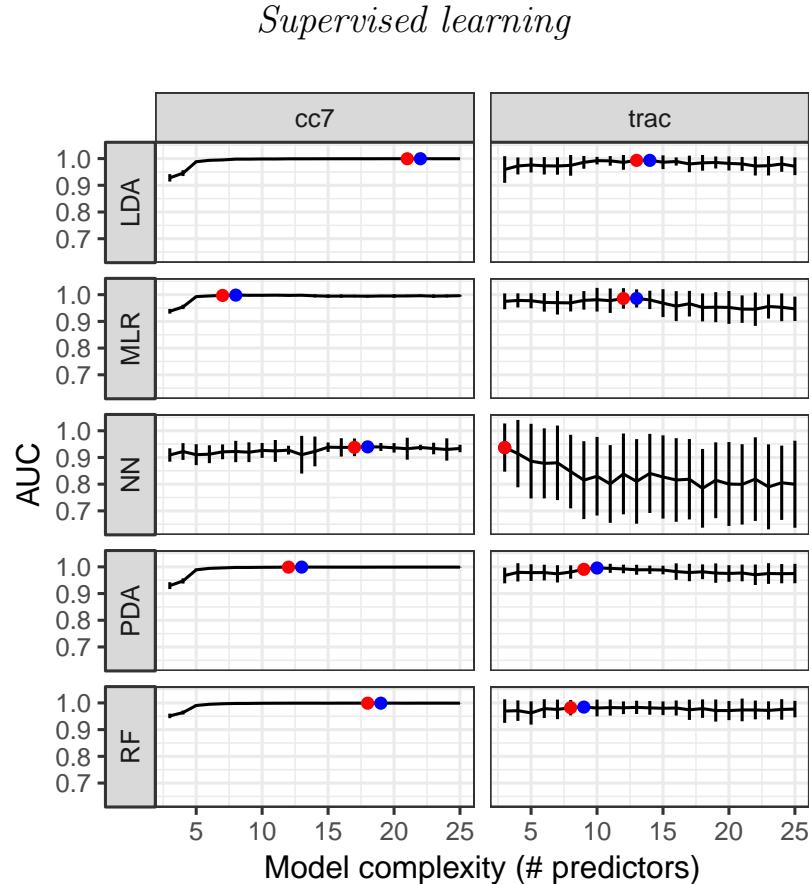


Figure 6: Comparisons of model fit to the training dataset for each of the supervised learning methods applied to the first two datasets; the results from the eight species dataset are presented in the left column, while those from the *Trachemys* dataset are presented in the right column. Models were 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.

Analysis of the eight morphologically- and genetically-distinct species and the *T. 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

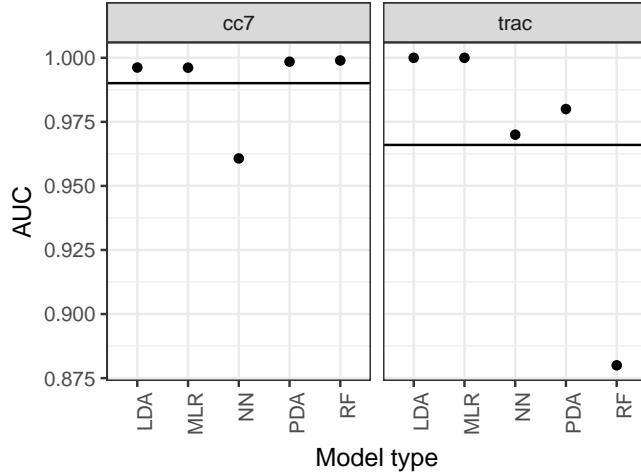


Figure 7: The results of out-of-sample predictive performance of the selected models for both the eight species (left) and *Trachemys* datasets. Predictive performance is measured as the area under the receiver operating characteristic (AUC), which ranges from 0.5 to 1. Points correspond to the individual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

- 362 classification rates (Fig. 6, 7). For both datasets, the ROC scores from testing datasets are
 363 tightly clustered near $AUC = 1$ (Fig. 7). These results demonstrate that when there are
 364 distinctions between the states of the classification schemes (i.e., differences in plastron shape
 365 that correlate with the different taxonomic groups), the methods used here can recover them.
 366 AUC-based model selection revealed some important patterns of variation and congruence
 367 between the classification schemes and the actual data. Generally, the best performing models
 368 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
 370 than for the other two datasets (Fig. 6, 8). In most cases the different proposed classification
 371 schemes are generally poor descriptors of the observed variation. It appears that the dataset
 372 is overwhelmed by noise (likely biological and analytical), making any accurate classifications
 373 difficult at best. This observation is cemented with the generalizations of the models to the
 374 testing dataset (Fig. 9).

Mean AUC values for the model generalizations, in most cases, are approximately equal to the
 376 observed AUC values from the training dataset (Fig. 8, 9). The cases in which the AUC from
 377 the generalizations is less than the observed indicate poor model fit and a poor classification
 378 scheme. Comparison of AUC values from the model generalizations do not indicate a clear
 379 “best” classification scheme (Fig. 8, 9). Only in the case of the conservative morphological
 380 hypothesis (“Morph”) is the mean AUC value potentially distinct from that of other schemes;
 381 in this case mean AUC is lower than the average of the other five schemes which indicates
 382 that the morphologically-based scheme performs more poorly than the molecularly-based

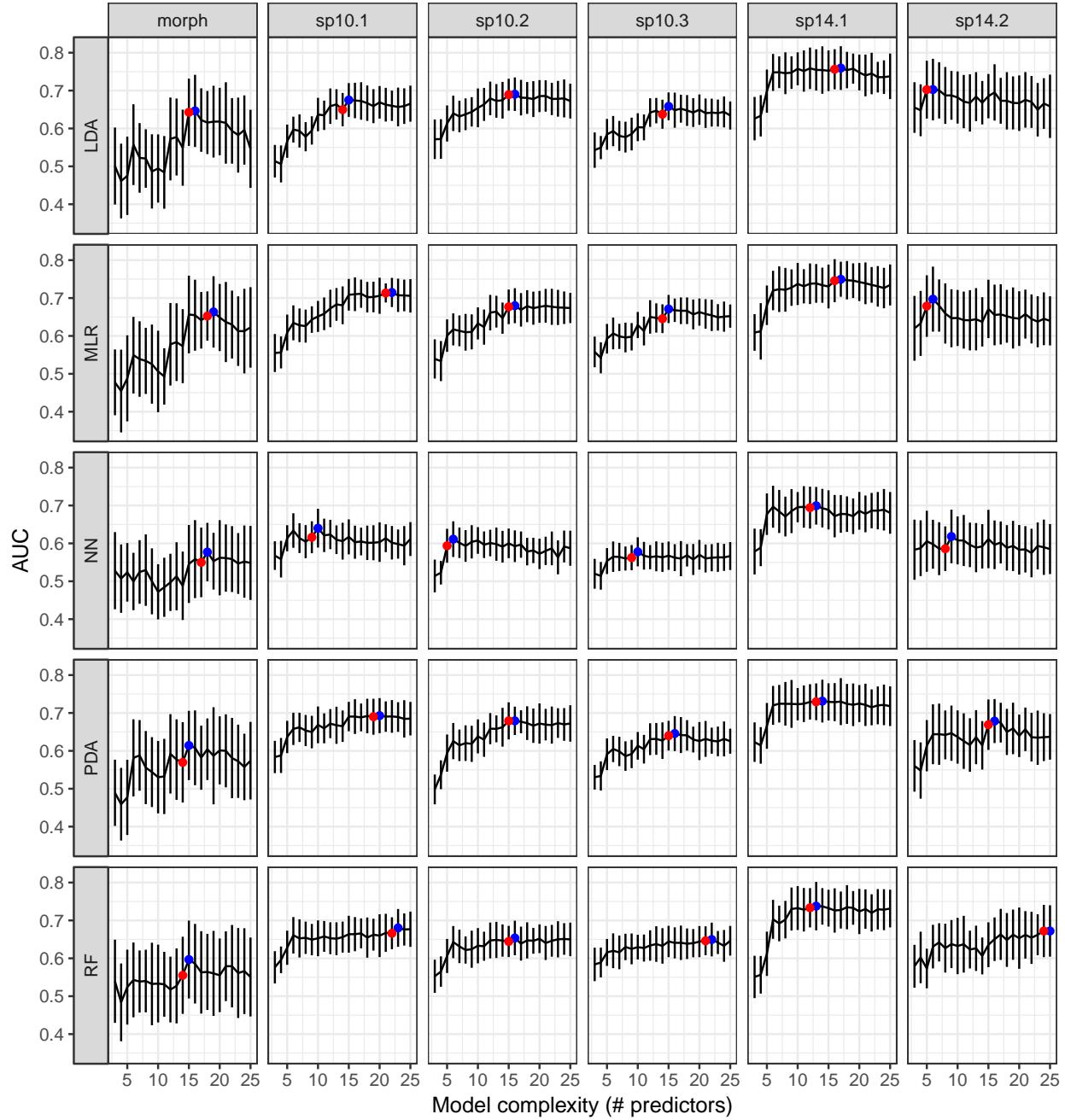


Figure 8: AUC values for models of varying complexity fit to the *Emys marmorata* training datasets for each classification scheme. The x-axis corresponds to the total number of predictors included in each model, while the y-axis corresponds to the AUC value which is a measure of goodness of fit for 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.

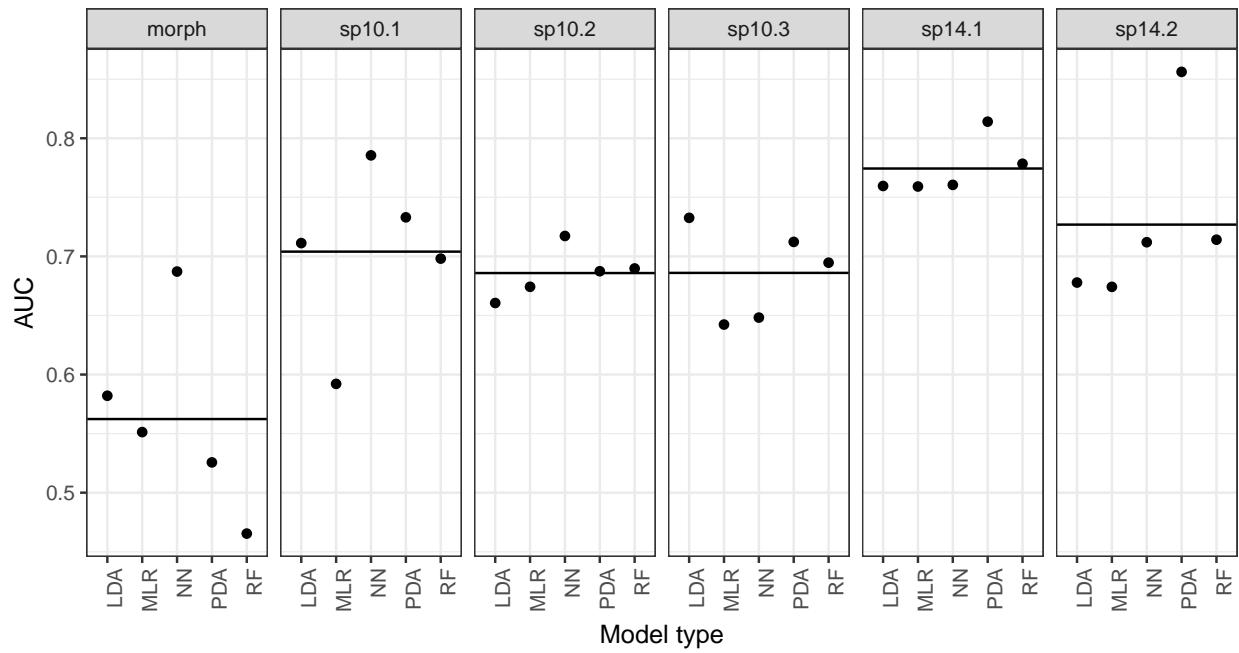


Figure 9: Comparison of out-of-sample AUC estimates from the predictions of selected models (Fig. 8), grouped by classification scheme. The horizontal line in each panel corresponds to the average AUC value across all models of that classification scheme.

ones. It is important to note, however, that the training and testing dataset for the “Morph”
384 scheme is the smallest of the six schemes which may lead to poorer performance in in-sample
and out-of-sample comparisons.

386

DISCUSSION

As expected, our ensemble approach yields high out-of-sample classification performance for
388 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.

390 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
392 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
394 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
396 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
398 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.*
400 *marmorata* that would make it significantly more difficult to accurately digitize than the
plastra of the other species.

402 Biological explanations include the possibility that genetic differentiation is not associated
with plastron shape variation and/or that local selective pressures (e.g. from hydrological
404 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
406 necessary mate choice (Rivera 2008; Rivera and Stayton 2011; Stayton 2011; Rivera et al.
2014; Polly et al. 2016), and that shell shape in *E. marmorata* is known to vary among
408 populations inhabiting water bodies with different flow regimes (Holland 1992; Lubcke and
Wilson 2007; Germano and Bury 2009). Plastron shape does not seem to preserve a strong
410 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
412 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
414 not represent significant evolutionary lineages.

Despite the negative result for *E. marmorata*, it is important to note that plastron shape
416 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
418 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
420 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

422 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
424 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
426 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
428 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
430 constraints (Stayton 2011; Polly et al. 2016). Although it may seem counterintuitive that
plastron shape is both useful for species delimitation but has weak or absent phylogenetic
432 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
434 *Terrapene* to the closely related spotted turtle *Clemmys guttata*), the variation within a
species typically is much less than the variation between species. Therefore, the consistent
436 plastron shapes that characterize different emydids species leads to plastron shape being
a useful tool for species delimitation, even when other selective factors have overprinted
438 similarities stemming from patterns of descent from common ancestors.

Is there more than one species of Western Pond Turtle?

440 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
442 prompt a reexamination of the methods and concepts that led to that taxonomic revision,
especially considering that plastron shape is demonstrably capable of differentiating species
444 and subspecies among other emydids. In other words, before we can assess the significance of
the morphological non-diagnosability, it is essential to evaluate the methods and concepts
446 that led to the initial taxonomic revision.

Spinks et al. (2014) elevated *E. pallida* based on a species delimitation analysis of SNP data
448 using BPP (Yang and Rannala 2010). However, Spinks et al. (2014) did not heed the caveats
about such species delimitation methods raised by Carstens et al. (2013). In addition to
450 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
452 emphasized that “Inferences regarding species boundaries based on genetic data alone are
likely inadequate, and species delimitation should be conducted with consideration of the life
454 history, geographical distribution, morphology and behaviour (where applicable) of the focal
system...” These caveats evoke the development of the Unified Species Concept (Dayrat
456 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
458 *E. marmorata* system until now, and in doing so we find the proposal that *E. pallida* is a
distinct species to be lacking.

460 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

462 implausible. The mitochondrial data from Spinks et al. (2014) show extensive introgression
463 and admixture in Central California, which is expected because there are no significant
464 barriers to gene flow in this region. They also lack sampling from the populations between the
465 two putative species in the San Francisco Bay Area, which we predict would likely show even
466 more genetic mixing. Combined with the well-demonstrated ability for testudinoid turtles,
467 including emydids and even *Emys*, to hybridize (e.g. Buskirk et al. 2005; Spinks and Shaffer
468 2009; Parham et al. 2013) it is hard to imagine how *E. marmorata* and *E. pallida* could
469 maintain their integrity in the face of such admixture. Any argument for the validity of *E.*
470 *pallida* as a distinct species needs to address these points. Because the geography, natural
471 history, limited sampling from key areas, demonstrated genetic admixture of *E. marmorata*,
472 and comparisons with other morphologically diagnosable species and subspecies conflict with
473 the recognition of *E. pallida*, we hypothesize that *E. pallida* is not a distinct species.
474 We fully agree with Spinks et al. (2014) that *E. marmorata* (*sensu lato*) is a species deserving
475 of strong conservation efforts, and we do not wish to trivialize this need. Moreover, the genetic
476 diversity uncovered by the analysis of Spinks et al. (2014) should be accounted for explicitly
477 in any conservation plan. Given the apparent lack of morphological distinction combined with
478 the broad range of intergradation and other problems with the species hypothesis outlined
479 above, we recommend that the populations elevated to *E. pallida* by Spinks et al. (2014) are
480 best considered Evolutionary Significant Units or Distinct Population Segments instead of
481 distinct species.
482 Finally, it is important to note that the data and analyses we present do not let us definitively
483 say whether the apparent lack of morphological divergence within *E. marmorata* truly
484 reflects the presence of a single species, or if it is an artifact of plastron shape being a
485 poor morphological marker for phylogenetic and phylogeographic divergences in the case
486 of *E. marmorata*. This is because we could not carry out our morphometric analyses on
487 the specimens from which the genetic data were obtained. The comparisons with the other
488 emydid taxa suggest that our negative result is because *E. marmorata* is a single species.
489 However, tests of both our preferred conclusion (*E. marmorata* as a single species) and
490 that of Spinks et al. (2014) should include morphological and molecular analyses of the
491 same set of voucher specimens, as well as additional tests of species delimitation using
492 alternative methods and corroborating evidence as suggested by Carstens et al. (2013). From
493 a morphological standpoint, support for the validity of “*E. pallida*” may come from other
494 aspects of morphology, such as carapace shape or other features. Likewise, further investigation
495 of the phylogeographic utility of plastron shape in other turtle species will help to clarify
496 whether the lack of differentiation seen in *E. marmoarata*, and the strong differentiation
497 among the other emydids, is typical or an unusual case.

*

498 Acknowledgements Data collection for this project was supported in part by NSF DBI-0306158
499 (to KDA). G. Miller assisted with data collection and her participation in this research was
500

supported by NSF REU DBI-0353797 (to R. Mooi of CAS). For access to emydine specimens,
502 we thank: J. Vindum and R. Drewes (CAS); A. Resetar (FMNH); R. Feeney (LACM); C.
Austin (LSUMNS); S. Sweet (MSE); J. McGuire and C. Conroy (MVZ); A. Wynn (NMNH);
504 P. Collins (SBMNH); B. Hollingsworth (SDMNH); P. Holroyd (UCMP). We are grateful for
S. Sweet for field assistance and the California Department of Fish and Game for permits.
506 We would also like to thank M. Lambruschi (FMNH) for help with figure 1.

BIBLIOGRAPHY

- 508 K. D. Angielczyk and C. R. Feldman. Are diminutive turtles miniaturized? The ontogeny
of plastron shape in emydine turtles. *Biological Journal of the Linnean Society*, 108(4):
510 727–755, apr 2013. ISSN 00244066. doi: 10.1111/bij.12010. URL <http://doi.wiley.com/10.1111/bij.12010>.
- 512 K. D. Angielczyk and H. D. Sheets. Investigation of simulated tectonic deformation in fossils
using geometric morphometrics. *Paleobiology*, 33(1):125–148, 2007.
- 514 K. D. Angielczyk, C. R. Feldman, and G. R. Miller. Adaptive evolution of plastron shape
in emydine turtles. *Evolution*, 65(2):377–394, feb 2011. ISSN 1558-5646. doi: 10.1111/j.
516 1558-5646.2010.01118.x.
- 518 G. Arnqvist and T. Mårtensson. Measurement error in geometric morphometrics: Empirical
strategies to assess and reduce its impact on measures of shape, 1998. ISSN 12178837.
- 520 S. F. Baird and C. Girard. 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, 1852.
- 522 A. M. Bauer, 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.
524 Ananjeva, and R. F. Inger. Availability of new Bayesian-delimited gecko names and the
importance of character-based species descriptions. *Proceedings of the Royal Society B:
526 Biological Sciences*, 278:490–492, 2000. ISSN 07331347. doi: 10.2307/1467045. URL
<http://www.jstor.org/stable/1467045?origin=crossref>.
- 528 M. Baylac, C. Villemant, and G. Simbolotti. Combining geometric morphometrics with
pattern recognition for the investigation of species complexes. *Biological Journal of the
530 Linnean Society*, 80:89–98, 2003.
- 532 D. Bickford, D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram,
and I. Das. Cryptic species as a window on diversity and conservation. *Trends in ecology &
534 evolution*, 22(3):148–55, mar 2007. ISSN 0169-5347. doi: 10.1016/j.tree.2006.11.004. URL
<http://www.ncbi.nlm.nih.gov/pubmed/17129636>.
- 536 L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*.
Wadsworth International Group, Belmont, 1984.

- R. B. Bury. Biogeography of Western Pond Turtles in the western Great Basin: Dispersal
538 Across a Northwest Passage ? *Western Wildlife*2, 4:72–80, 2017.
- R. B. Bury, D. J. Germano, and G. W. Bury. Population Structure and Growth of the
540 Turtle *Actinemys marmorata* from the KlamathSiskiyou Ecoregion: Age, Not Size, Matters.
Copeia, 2010(3):443–451, sep 2010. ISSN 0045-8511. doi: 10.1643/CH-08-096. URL
542 <http://www.bioone.org/doi/abs/10.1643/CH-08-096>.
- S. W. Buskirk, J. F. Parham, and C. R. Feldman. On the hybridisation between two distantly
544 related Asian turtles (Testudines: *Scalia* x *Mauremys*). *Salamandra*, 41:21–26, 2005.
- A. Cardini, D. Nagorsen, P. O'Higgins, P. D. Polly, R. W. Thorington Jr, and P. Tongiorgi.
546 Detecting biological distinctiveness using geometric morphometrics: an example case from
the Vancouver Island marmot. *Ethology Ecology & Evolution*, 21:209–223, 2009.
- B. C. Carstens and T. A. Dewey. Species Delimitation Using a Combined Coalescent
548 and Information-Theoretic Approach: An Example from North American Myotis Bats.
Systematic Biology, 59(4):400–414, 2010. URL [papers2://publication/doi/10.1093/sysbio/syq024](http://publications/doi/10.1093/sysbio/syq024).
- B. C. Carstens, T. a. Pelletier, N. M. Reid, and J. D. Satler. How to fail at species delimitation.
552 *Molecular ecology*, 22(17):4369–83, sep 2013. ISSN 1365-294X. doi: 10.1111/mec.12413.
554 URL <http://www.ncbi.nlm.nih.gov/pubmed/23855767>.
- R. Caumul and P. D. Polly. Phylogenetic and environmental components of morphological
556 variation: skull, mandible, and molar shape in marmots (*Marmota*, Rodentia). *Evolution; international journal of organic evolution*, 59(11):2460–72, nov 2005. ISSN 0014-3820. URL
558 <http://www.ncbi.nlm.nih.gov/pubmed/16396186>.
- E. L. Clare. Cryptic species? Patterns of maternal and paternal gene flow in eight neotropical
560 bats. *PLoS one*, 6(7):e21460, jan 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0021460.
- J. Claude. Convergence induced by plastral kinesis and geometric morphometric assessment:
562 a geometric morphometric assessment. *Fossil Turtle Research*, 1:34–45, 2006.
- J. Claude, E. Paradis, H. Tong, and J. C. Auffray. A geometric morphometric assessment
564 of the effects of environment and cladogenesis on the evolution of the turtle shell.
Biological Journal of the Linnean Society, 79(December):485–501, 2003. ISSN 00244066.
566 doi: 10.1046/j.1095-8312.2003.00198.x.
- B. Dayrat. Towards integrative taxonomy. *Biological Journal of the Linnean Society*, 85:
568 407–415, 2005.
- K. De Queiroz. Species concepts and species delimitation. *Systematic Biology*, 56(6):879–86,
570 dec 2007. ISSN 1063-5157. doi: 10.1080/10635150701701083. URL <http://www.ncbi.nlm.nih.gov/pubmed/18027281>.
- M. H. Demandt and S. Bergek. Identification of cyprinid hybrids by using geometric
572

- 574 morphometrics and microsatellites. *Journal of Applied Ichthyology*, 25(6):695–701, dec
2009. ISSN 01758659. doi: 10.1111/j.1439-0426.2009.01329.x. URL <http://doi.wiley.com/10.1111/j.1439-0426.2009.01329.x>.
- 576 K. C. Dillard. *A comparative analysis of geometric morphometrics across two Pseudemys turtle species in east central Virginia*. Masters, Virginia Commonwealth University, 2017.
- 578 G. Dobigny, L. Granjon, V. Aniskin, K. Ba, and V. Voloboulev. A new sigling species of Taterillus (Muridae, Gerbillinae) from West Agrica. *Mammalian Biology*, 68:299–316, 2003.
- 580 I. L. Dryden. *shapes: Statistical shape analysis*, 2013. URL <http://CRAN.R-project.org/package=shapes>. R package version 1.1-8.
- 582 I. L. Dryden and K. Y. Mardia. *Statistical shape analysis*. Wiley, New York, 1998.
- 584 S. Edwards, J. Claude, B. J. Van Vuuren, and C. A. Matthee. Evolutionary history of the Karoo bush rat, Myotomys unisulcatus (Rodentia: Muridae): Disconcordance between morphology and genetics. *Biological Journal of the Linnean Society*, 102(3):510–526, 2011.
586 ISSN 00244066. doi: 10.1111/j.1095-8312.2010.01583.x.
- 588 N. Eldredge and S. J. Gould. Punctuated equilibria: an alternative to phyletic gradualism. In T. J. M. Schopf, editor, *Models in Paleobiology*, pages 82–115. Freeman Cooper, San Francisco, 1972.
- 590 C. R. Feldman and J. F. Parham. Molecular phylogenetics of emydine turtles: taxonomic revision and the evolution of shell kinesis. *Molecular Phylogenetics and Evolution*, 22(3):388–98, mar 2002. ISSN 1055-7903. doi: 10.1006/mpev.2001.1070. URL <http://www.ncbi.nlm.nih.gov/pubmed/11884163>.
- 594 T. M. Franco, R. A. O. Silva, P. Nunes-Silva, C. Menezes, and V. L. Imperatriz-Fonseca. Gender identification of five genera of stingless bees (Apidae, Meliponini) based on wing morphology. *Genetics and molecular research*, 8(1):207–214, 2009.
- 598 C. Fruciano, P. Franchini, F. Raffini, S. Fan, and A. Meyer. Are sympatrically speciating Midas cichlid fish special? Patterns of morphological and genetic variation in the closely related species Archocentrus centrarchus. *Ecology and Evolution*, 6(12):4102–4114, 2016.
600 ISSN 20457758. doi: 10.1002/ece3.2184.
- 602 W. C. Funk, M. Caminer, and S. R. Ron. High levels of cryptic species diversity uncovered in Amazonian frogs. *Proceedings of the Royal Society B: Biological Sciences*, 279(1734):1806–14, may 2012. ISSN 1471-2954. doi: 10.1098/rspb.2011.1653.
- 604 P. Gaubert, P. J. Taylor, C. a. Fernandes, M. W. Bruford, and G. Veron. 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(1):11–33, aug 2005. ISSN 00244066. doi: 10.1111/j.1095-8312.2005.00518.x. URL <http://doi.wiley.com/10.1111/j.1095-8312.2005.00518.x>.

- D. J. Germano and R. B. Bury. Variation in body size, growth, and population structure of *Actinemys marmorata* from lentic and lotic habitats in Southern Oregon. *Journal of Herpetology*, 43(3):510–520, 2009.
- D. J. Germano and G. B. Rathbun. Growth, population structure, and reproduction of western pond turtles (*Actinemys marmorata*) on the Central Coast of California. *Chelonian Conservation and Biology*, 7(2):188–194, 2008.
- S. J. Gould and N. Eldredge. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, 3(2):115–151, 1977.
- G. Guillot, S. Renaud, R. Ledevin, J. Michaux, and J. Claude. A unifying model for the analysis of phenotypic, genetic, and geographic data. *Systematic Biology*, 61(6):897–911, 2012. ISSN 10635157. doi: 10.1093/sysbio/sys038.
- I. Gündüz, M. Jaarola, C. Tez, C. Yeniyurt, P. D. Polly, and J. B. Searle. Multigenic and morphometric differentiation of ground squirrels (*Spermophilus*, *Sciuridae*, *Rodentia*) in Turkey, with a description of a new species. *Molecular phylogenetics and evolution*, 43(3):916–35, jun 2007. ISSN 1055-7903. doi: 10.1016/j.ympev.2007.02.021. URL <http://www.ncbi.nlm.nih.gov/pubmed/17500011>.
- D. J. Hand and R. J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45:171–186, 2001.
- T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, 2nd edition, 2009.
- T. Hastie, R. Tibshirani, F. Leisch, K. Hornik, and B. D. Ripley. *mda: Mixture and Flexible Discriminant Analysis*, 2015. URL <https://CRAN.R-project.org/package=mda>. R package version 0.4-8.
- B. Hausdorf and C. Hennig. Species delimitation using dominant and codominant multilocus markers. *Systematic biology*, 59(5):491–503, oct 2010. ISSN 1076-836X. doi: 10.1093/sysbio/syq039. URL <http://www.ncbi.nlm.nih.gov/pubmed/20693311>.
- D. C. Holland. *Level and pattern in morphological variation: a phylogeographic study of the western pond turtle (*Clemmys marmorata*)*. PhD thesis, University of Southwestern Louisiana, 1992.
- J. P. Huelsenbeck, P. Andolfatto, and E. T. Huelsenbeck. Structurama: bayesian inference of population structure. *Evolutionary bioinformatics online*, 7:55–9, jan 2011. ISSN 1176-9343. doi: 10.4137/EBO.S6761. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118697/>?tool=pmcentrez&rendertype=abstract.
- L. Kaufman and P. J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley, New York, 1990.
- D. G. Kendall. The diffusion of shape. *Advances in Applied Probability*, 9(3):428–430, 1977.

- M. Kuhn. *caret: Classification and Regression Training*, 2013. URL <http://CRAN.R-project.org/package=caret>. R package version 5.15-61.
- M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer, New York, NY, 2013.
- A. D. Leaché and M. K. Fujita. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings. Biological sciences / The Royal Society*, 277(1697):3071–7, oct 2010. ISSN 1471-2954. doi: 10.1098/rspb.2010.0662. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2982061&tool=pmcentrez&rendertype=abstract>.
- A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL <http://CRAN.R-project.org/doc/Rnews/>.
- G. M. Lubcke and D. S. Wilson. 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(1):107–114, 2007.
- N. MacLeod. *Automated taxon identification in systematics: theory, approaches and applications*. CRC Press, Boca Raton, 2007.
- M. Markolf, H. Rakotonirina, C. Fichtel, P. von Grumbkow, M. Brameier, and P. M. Kappeler. True lemurs... true species - species delimitation using multiple data sources in the brown lemur complex. *BMC Evolutionary Biology*, 13(1):233, 2013. ISSN 1471-2148. doi: 10.1186/1471-2148-13-233. URL <http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-13-233>.
- A. Mitrovski-Bogdanovic, A. Petrovic, M. Mitrovic, A. Ivanovic, V. Žikic, P. Starý, C. Vorburger, and Ž. Tomanovic. 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(2):170–180, 2013.
- Munoz-Munoz F. and Perpinan D. Measurement error in morphometric studies: comparison between manual and computerized methods. *Ann. Zool.*, 47(1):46–56, 2010.
- D. Navega, R. Vicente, D. N. Vieira, A. H. Ross, and E. Cunha. Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. *International Journal of Legal Medicine*, 129(3):651–659, 2015. ISSN 14371596. doi: 10.1007/s00414-014-1070-5.
- B. C. O’Meara. New heuristic methods for joint species delimitation and species tree inference. *Systematic biology*, 59(1):59–73, jan 2010. ISSN 1076-836X. doi: 10.1093/sysbio/syp077. URL <http://www.ncbi.nlm.nih.gov/pubmed/20525620>.
- J. M. Padial, A. Miralles, I. De la Riva, and M. Vences. The integrative future of taxonomy. *Frontiers in Zoology*, 7(16):1–14, 2010.
- J. F. Parham, T. J. Papenfuss, P. P. V. Dijk, B. S. Wilson, C. Marte, L. R. Schettino, and W. Brian Simison. Genetic introgression and hybridization in Antillean freshwater turtles

- (*Trachemys*) revealed by coalescent analyses of mitochondrial and cloned nuclear markers.
 682 *Molecular phylogenetics and evolution*, 67(1):176–87, apr 2013. ISSN 1095-9513. doi:
 10.1016/j.ympev.2013.01.004. URL <http://www.ncbi.nlm.nih.gov/pubmed/23353072>.
- 684 M. Pfenninger and K. Schwenk. Cryptic animal species are homogeneously distributed
 among taxa and biogeographical regions. *BMC evolutionary biology*, 7:121, jan 2007. ISSN
 686 1471-2148. doi: 10.1186/1471-2148-7-121. URL <http://www.pubmedcentral.nih.gov/articleRender.fcgi?artid=1939701&tool=pmcentrez&rendertype=abstract>.
- 688 P. D. Polly. Paleophylogeography of *Sorex araneus*: molar shape as a morphological marker
 for fossil shrews. *Mammalia*, 68(2):233–243, 2003.
- 690 P. D. Polly. Phylogeographic differentiation in *Sorex araneus*: morphology in relation to
 geography and karyotype. *Russian Journal of Theriology*, 6(1):73–84, 2007.
- 692 P. D. Polly, C. T. Stayton, E. R. Dumont, S. E. Pierce, E. J. Rayfield, and K. D. Angielczyk.
 Combining geometric morphometrics and finite element analysis with evolutionary modeling:
 694 towards a synthesis. *Journal of Vertebrate Paleontology*, 4634(March), 2016. ISSN 0272-4634.
 doi: 10.1080/02724634.2016.1111225.
- 696 J. Pons, T. Barraclough, J. Gomez-Zurita, A. Cardoso, D. Duran, S. Hazell, S. Kamoun,
 W. Sumlin, and A. Vogler. Sequence-Based Species Delimitation for the DNA Taxonomy
 698 of Undescribed Insects. *Systematic Biology*, 55(4):595–609, aug 2006. ISSN 1063-5157.
 doi: 10.1080/10635150600852011. URL <http://sysbio.oxfordjournals.org/cgi/doi/10.1080/10635150600852011>.
- 700 R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for
 Statistical Computing, Vienna, Austria, 2016. URL <http://www.R-project.org/>.
- 704 G. Rivera. Ecomorphological variation in shell shape of the freshwater turtle *Pseudemys*
 706 *concinna* inhabiting different aquatic flow regimes. *Integrative and comparative biology*, 48
 (6):769–87, dec 2008. ISSN 1540-7063. doi: 10.1093/icb/icn088. URL <http://www.ncbi.nlm.nih.gov/pubmed/21669831>.
- 708 G. Rivera and C. T. Stayton. Finite element modeling of shell shape in the freshwater turtle
 710 *Pseudemys concinna* reveals a trade-off between mechanical strength and hydrodynamic
 efficiency. *Journal of morphology*, 272(10):1192–203, oct 2011. ISSN 1097-4687. doi:
 10.1002/jmor.10974. URL <http://www.ncbi.nlm.nih.gov/pubmed/21630321>.
- 712 G. Rivera, J. N. Davis, J. C. Godwin, and D. C. Adams. Repeatability of Habitat-Associated
 714 Divergence in Shell Shape of Turtles. *Evolutionary Biology*, pages 29–37, jul 2014. ISSN
 0071-3260. doi: 10.1007/s11692-013-9243-6. URL <http://link.springer.com/10.1007/s11692-013-9243-6>.
- F. J. Rohlf. TpsDig 2.04, 2005.
- 716 B. C. Schilck-Steiner, B. Seifert, C. Stauffer, E. Christian, R. H. Crozier, and F. M. Steiner.
 Without morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends*

- 718 in *ecology & evolution*, 22(8):391–392, aug 2007. ISSN 0169-5347. doi: 10.1016/j.tree.2007.
05.003. URL <http://www.ncbi.nlm.nih.gov/pubmed/17573147>.
- 720 L. M. Seeliger. Variation in the Pacific Mud Turtle. *Copeia*, 1945(3):150–159, 1945.
- 722 P. Q. Spinks and H. B. Shaffer. Range-wide molecular analysis of the western pond turtle
724 (*Emys marmorata*): cryptic variation, isolation by distance, and their conservation im-
plications. *Molecular ecology*, 14(7):2047–64, jun 2005. ISSN 0962-1083. doi: 10.1111/j.
1365-294X.2005.02564.x. URL <http://www.ncbi.nlm.nih.gov/pubmed/15910326>.
- 726 P. Q. Spinks and H. B. Shaffer. Conflicting mitochondrial and nuclear phylogenies for the
728 widely disjunct Emys (Testudines: Emydidae) species complex, and what they tell us about
biogeography and hybridization. *Systematic biology*, 58(1):1–20, feb 2009. ISSN 1076-836X.
doi: 10.1093/sysbio/syp005. URL <http://www.ncbi.nlm.nih.gov/pubmed/20525565>.
- 730 P. Q. Spinks, R. C. Thomson, and H. B. Shaffer. Nuclear gene phylogeography reveals the
732 historical legacy of an ancient inland sea on lineages of the western pond turtle, *Emys
marmorata* in California. *Molecular ecology*, 19(3):542–56, feb 2010. ISSN 1365-294X.
doi: 10.1111/j.1365-294X.2009.04451.x. URL <http://www.ncbi.nlm.nih.gov/pubmed/20051011>.
- 734 P. Q. Spinks, R. C. Thomson, and H. Bradley Shaffer. The advantages of going large: genome
736 wide SNPs clarify the complex population history and systematics of the threatened
738 western pond turtle. *Molecular Ecology*, pages n/a–n/a, mar 2014. ISSN 09621083. doi:
10.1111/mec.12736. URL <http://doi.wiley.com/10.1111/mec.12736>.
- 740 C. T. Stayton. Biomechanics on the half shell: functional performance influences patterns
742 of morphological variation in the emydid turtle carapace. *Zoology (Jena, Germany)*,
744 114(4):213–23, sep 2011. ISSN 1873-2720. doi: 10.1016/j.zool.2011.03.002. URL <http://www.ncbi.nlm.nih.gov/pubmed/21820295>.
- 746 B. L. Stuart, R. F. Inger, and H. K. Voris. High level of cryptic species diversity revealed by
748 sympatric lineages of Southeast Asian forest frogs. *Biology letters*, 2(3):470–4, sep 2006.
ISSN 1744-9561. doi: 10.1098/rsbl.2006.0505. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/1686201/>{&}tool=pmcentrez{&}rendertype=abstract.
- 750 A. Sztencel-Jabłonka, G. Jones, and W. Bogdanowicz. Skull Morphology of Two Cryptic
752 Bat Species: *Pipistrellus pipistrellus* and *P. pygmaeus* A 3D Geometric Morphometrics
754 Approach with Landmark Reconstruction. *Acta Chiropterologica*, 11(1):113–126, jun 2009.
ISSN 1508-1109. doi: 10.3161/150811009X465730. URL <http://www.bioone.org/doi/abs/10.3161/150811009X465730>.
- B. Van Bocxlaer and G. Hunt. Morphological stasis in an ongoing gastropod radiation from
Lake Malawi. *Proceedings of the National Academy of Sciences*, aug 2013. ISSN 0027-
8424. doi: 10.1073/pnas.1308588110. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/3754130/>{&}tool=pmcentrez{&}rendertype=abstract.

- B. Van Bocxlaer and R. Schultheiß. Comparison of morphometric techniques for shapes with few homologous landmarks based on machine-learning approaches to biological discrimination. *Paleobiology*, 36(3):497–515, 2010.
- V. van den Brink and F. Bokma. Morphometric shape analysis using learning vector quantization neural networks an example distinguishing two microtine vole species. *Annales Zoologici Fennici*, 48:359–364, 2011.
- W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL <http://www.stats.ox.ac.uk/pub/MASS4>.
- N. S. Vitek, C. L. Manz, T. Gao, J. I. Bloch, S. G. Strait, and D. M. Boyer. Semi-supervised determination of pseudocryptic morphotypes using observer-free characterizations of anatomical alignment and shape. *Ecology and Evolution*, 7(14):5041–5055, 2017. ISSN 20457758. doi: 10.1002/ece3.3058. URL <http://doi.wiley.com/10.1002/ece3.3058>.
- N. von Cramon-Taubadel, B. C. Frazier, and M. M. Lahr. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American journal of physical anthropology*, 132(4):535–544, 2007. ISSN 00029483. doi: 10.1002/ajpa.
- Z. Yang and B. Rannala. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20):9264–9, may 2010. ISSN 1091-6490. doi: 10.1073/pnas.0913022107. URL <http://www.ncbi.nlm.nih.gov/articlerender.fcgi?artid=2889046&tool=pmcentrez&rendertype=abstract>.
- M. L. Zelditch, D. L. Swiderski, and H. D. Sheets. *Geometric morphometrics for biologists: a primer*. Elsevier Academic Press, Amsterdam, 2004.