

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

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

6 ¹**Department of Integrative Biology, University of California – Berkeley**

7 ²**Integrative Research Center, Field Museum of Natural History**

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

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

11 Corresponding author:

12 Peter D Smits¹

13 Email address: psmits@berkeley.edu

14 **ABSTRACT**

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

31 **INTRODUCTION**

32 Molecular systematics has repeatedly demonstrated the existence of cryptic species that can only be diagnosed using
33 genetic data (Stuart et al., 2006; Bickford et al., 2007; Schilck-Steiner et al., 2007; Pfenniger and Schwenk, 2007;
34 Clare, 2011; Funk et al., 2012). In attempts to streamline the documentation of biodiversity, several methods of
35 species delimitation that rely almost entirely on genetic data have recently been proposed (Pons et al., 2006; Carstens
36 and Dewey, 2010; Hausdorf and Hennig, 2010; O’Meara, 2010; Yang and Rannala, 2010; Huelsenbeck et al., 2011).

37 Although strong caveats on the utility of these methods have been raised (Bauer et al., 2000; Carstens et al., 2013), they
38 are nevertheless being used to name species (Leaché and Fujita, 2010; Spinks et al., 2014).

39 In contrast to those genetically-diagnosed species, the majority of extant taxa, and almost all extinct taxa, are delimited
40 by morphology alone. This disjunction complicates interpretations of variation and diversity in deep time, as apparent
41 morphological stasis may not reflect the true underlying diversity (Eldredge and Gould, 1972; Gould and Eldredge,
42 1977; Van Boeckelaer and Hunt, 2013). It also has serious implications for our records of modern biodiversity: for many
43 museum specimens of extant taxa (e.g. those preserved in formalin), it is difficult to acquire the genetic data needed for
44 non-morphological species delimitation methods.

45 These considerations have sparked interest in whether geometric morphometric analyses can capture fine-scale variation
46 that can be used for identifying cryptic species. This would make the task of identifying and maintaining endangered or
47 conserved groups much easier and could contribute to improved classifications of extinct taxa and populations. Most
48 such studies focus on using morphometrics to discover differences between taxa that were identified by other means
49 (Polly, 2003; Zelditch et al., 2004; Gaubert et al., 2005; Gündüz et al., 2007; Polly, 2007; Demandt and Bergek, 2009;
50 Markolf et al., 2013; Fruciano et al., 2016). Additionally, there has been work on automated taxon identification and
51 classification of taxa into groups (Baylac et al., 2003; Dobigny et al., 2003; MacLeod, 2007; van den Brink and Bokma,
52 2011; Vitek et al., 2017), as well as the development of models that combine genetic, phenotypic, and geographic data
53 to infer evolutionary units of interest (Guillot et al., 2012).

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

60 **Background and study system**

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

67 There are two major classes of machine learning method: unsupervised and supervised learning. Unsupervised learning
68 methods are used with unlabeled data where the underlying structure is estimated; they are analogous to clustering
69 and density estimation methods (Kaufman and Rousseeuw, 1990). Supervised learning methods are used with labeled
70 data where the final output of data is known and the rules for going from input to output are inferred. These are
71 analogous to classification and regression models (Breiman et al., 1984; Hastie et al., 2009). Our application of the
72 supervised learning approaches used in this study illustrates only a sampling of the various methods available for fitting
73 classification models. The specific methods used in this study were chosen because they are suited for cases with more
74 two or more response classes.

75 Geometric morphometric approaches to identifying differences in morphological variation between classes, including
76 cryptic species, have mostly relied on methods like linear discriminant analysis and canonical variates analysis (Polly,
77 2003; Zelditch et al., 2004; Gaubert et al., 2005; Gündüz et al., 2007; Polly, 2007; Francoy et al., 2009; Sztencel-
78 Jabłonka et al., 2009; Edwards et al., 2011; Mitrovski-Bogdanovic et al., 2013; Dillard, 2017). Because of their similarity
79 to multivariate approaches like principal components analysis (PCA), these methods are comparatively straightforward
80 ways of understanding the differences in morphology between classes. They also benefit from producing results
81 that can be easily visualized, which aids in the interpretation and presentation of data and results. Most previous
82 morphometric studies did not assess which amongst a set of alternative classification hypotheses was optimal. For
83 example, studies such as those of Caumul and Polly (2005) and Polly (2007) focused on comparing different aspects of

84 morphology and their fidelity to a classification scheme instead of comparing the fidelity of one aspect of morphology to
85 multiple classification schemes. In this context, the study of Cardini et al. (2009) is noteworthy because they compared
86 morphological variation in marmots at the population, regional, and species level and determined the fidelity of shape to
87 divisions at each of these levels.

88 Here, we used an ensemble of supervised machine learning methods to compare the congruence of morphological
89 data to different classification hypotheses. Each of these methods provide different advantages for understanding how
90 to classify specimens, as well as the accuracy of the resulting classifications. Machine learning methods have been
91 combined with geometric morphometric data to study shape variation in a variety of contexts, including automated taxon
92 identification and classification of groups (Baylac et al., 2003; Dobigny et al., 2003; MacLeod, 2007; Van Boeckelaer and
93 Schultheiß, 2010; van den Brink and Bokma, 2011; Navega et al., 2015). In the current study, we not only consider
94 pure classification accuracy but also use a statistic of classification strength that reflects the rate at which taxa are both
95 accurately and inaccurately classified: the area under the Receiver Operating Characteristic curve (Hastie et al., 2009).

96 We analyzed the problem of whether there are distinct subspecies or cryptic species within the western pond turtle,
97 *Emys marmorata* (Baird and Girard, 1852) (formerly *Clemmys marmorata*; see Feldman and Parham, 2002). *Emys*
98 *marmorata* is distributed from northern Washington State, USA to Baja California, Mexico; populations in western
99 Nevada may have been introduced by recent human activity or they could be a genuine part of the species' range (Bury,
100 2017). Traditionally, *E. marmorata* was classified into two named subspecies: the northern *E. marmorata marmorata*
101 and the southern *Emys marmorata pallida* (Seeliger, 1945), with a central Californian intergrade zone in between. *Emys*
102 *marmorata marmorata* is differentiated from *E. marmorata pallida* by the presence of a pair of triangular inguinal
103 scales and darker neck markings. The triangular inguinal plates can sometimes be present in *E. marmorata pallida*
104 although they are considerably smaller. Seeliger (1945) did not formally include the Baja California populations of *E.*
105 *marmorata* in either taxon, implying the existence of a third distinct but unnamed subspecies.

106 Previous work on morphological variation in *E. marmorata* has focused primarily on differentiation between populations
107 over a portion of the species' total range (Lubcke and Wilson, 2007; Germano and Rathbun, 2008; Germano and Bury,
108 2009; Bury et al., 2010); comparatively few studies have included specimens from across the entire range (Holland,
109 1992). Most of these studies considered how local biotic and abiotic factors may contribute to differences in carapace
110 length, and they found that size can vary greatly between different populations (Lubcke and Wilson, 2007; Germano
111 and Rathbun, 2008; Germano and Bury, 2009). There also has been interest in size-based sexual dimorphism in *E.*
112 *marmorata* (Holland, 1992; Lubcke and Wilson, 2007; Germano and Bury, 2009), with males being on average larger
113 than females based on total carapace length and other linear measurements. However, the quality of size as a classifier
114 of sex can vary greatly between populations (Holland, 1992) because of the magnitude of size differences among
115 populations (Lubcke and Wilson, 2007; Germano and Bury, 2009). The effect of sexual dimorphism on shape, *sensu*
116 Kendall (1977), has not been assessed (Holland, 1992; Lubcke and Wilson, 2007; Germano and Rathbun, 2008).

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

127 More recently, the phylogeography of *E. marmorata* and the possibility of cryptic diversity was investigated using
128 molecular data (Spinks and Shaffer, 2005; Spinks et al., 2010, 2014). Based on mitochondrial DNA, Spinks and Shaffer
129 (2005) recognized four subclades within *E. marmorata*, a northern clade, a San Joaquin Valley clade, a Santa Barbara
130 clade, and a southern clade. Analyses with nuclear DNA (Spinks et al., 2010) and single-nucleotide polymorphism
131 (SNP) data suggest a primarily north-south division in *E. marmorata*, although these datasets differed from that of
132 mitochondrial-based results of Spinks and Shaffer (2005) in the location of the break point (Spinks et al., 2014). All

133 three studies discussed the potential taxonomic implications of their results, with Spinks et al. (2014) going so far as to
134 strongly advocate for the recognition of at least two species (*E. marmorata* and *E. pallida*), and a possible third based
135 on populations in Baja California. However, they did not discuss in detail the morphological characters that would help
136 to diagnose these species beyond those specified by Seeliger (1945). Given that these characters are variable within the
137 proposed species, and that Holland (1992) described shell shape variation that might be consistent with this taxonomy,
138 a geometric morphometric analysis of shell shape might provide a reliable way to diagnose groups (whether species or
139 subspecies) within *E. marmorata*.

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

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

149 In the case of the *E. marmorata* species complex, we hypothesize that if one or more of the proposed classification
150 schemes are consistent with the morphological data then our ensemble approach fit to those hypotheses will have
151 higher out-of-sample predictive performance than the more inconsistent hypotheses. However, if all of the classification
152 schemes lead to equal out-of-sample predictive performance then we would conclude that the proposed hypotheses are
153 inconsistent with whatever information is present in the morphological data. Because of unclear geographic boundaries
154 between subgroups of *E. marmorata*, we compare multiple permutations of the (Spinks et al., 2010) and Spinks et al.
155 (2014) hypotheses.

156 METHODS

157 Specimens, sampling, morphometrics

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

168 The first dataset we analyzed includes 578 total specimens from the following species: *Chrysemys picta*, *Clemmys*
169 *guttata*, *Emys blandigii*, *Emys orbicularis*, *Glyptemys insculpta*, *Glyptemys muhlenbergii*, *Terrapene coahuila*, and
170 *Terrapene ornata*. These specimens are a subset of those used in Angielczyk et al. (2011) and Angielczyk and Feldman
171 (2013).

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

174 The final dataset is of 532 adult *E. marmorata* museum specimens, though not all specimens were able to be assigned
175 a class for all schemes (Fig. 1). These specimens represent a subset of those included in Angielczyk and Sheets
176 (2007), Angielczyk et al. (2011), and Angielczyk and Feldman (2013). Because Spinks and Shaffer (2005), Spinks et al.
177 (2010), and Spinks et al. (2014) did not use voucherized specimens we were not able to directly sample the individuals in

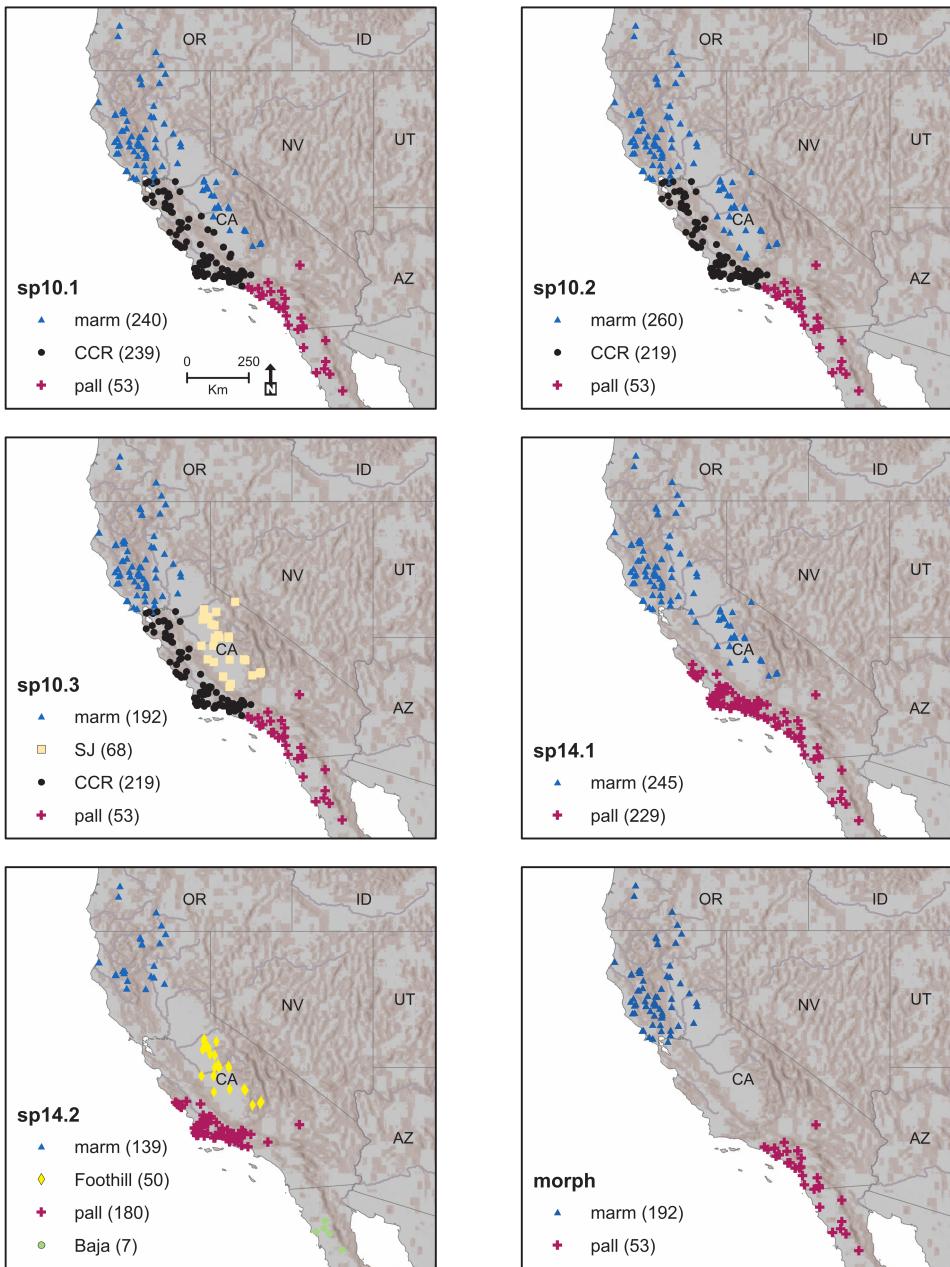


Figure 1. Geographic distribution of specimens sampled for comparing the hypothesized subdivisions of *Emys marmorata*. Each hypothesized scheme has two or more possible classes. Sample size differs between schemes because of our ability to confidently assign museum specimens to the various schemes.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

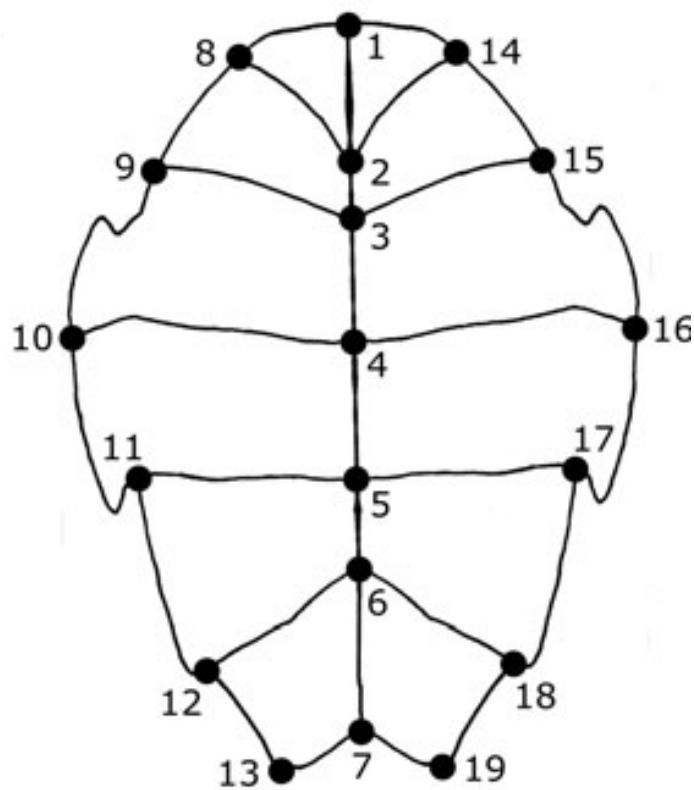


Figure 2. Depiction of general plastral shape of *E. marmorata* and position of the 19 landmarks used in this study. Anterior is towards the top of the figure.

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)

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 Sheets, 2007; Angielczyk et al., 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 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 (R Core Team, 2016; Dryden, 2013). 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; von Cramon-Taubadel et al., 2007; Munoz-Munoz F. and Perpinan D., 2010) on our results by comparing within-specimen (replicated) Procrustes distances to the distances between classification scheme centroids. Ten randomly-selected *E. marmorata* specimens

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)

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 (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

255 the later is used for measuring model performance, a process called model generalization. For each scheme, we limited
256 the model training and testing to only those individuals with class labels for that scheme. In this analysis, we randomly
257 divided 80% of samples into the training set and the remaining 20% into the testing set.

258 In classification studies, such as this one, a common metric of performance is the receiver operating characteristic
259 (ROC) which is the relationship between the false and true positive rates (Hastie et al., 2009). The area under the
260 ROC curve (AUC) is the derived estimate of the model performance; AUC ranges from 0.5 to 1 which correspond to
261 performance similar to random guesses and perfect classification rates, respectively (Hastie et al., 2009). Both ROC and
262 AUC are preferable to simple classification accuracy when class membership is unbalanced, as it is in these analyses
263 (Hastie et al., 2009). The standard ROC and AUC calculations are defined only for binary classifications, which is not
264 the case for our eight species and *Emys* complex datasets. To generalize this approach for situations with multiple
265 response classes, we used an all-against-one strategy where the model AUC is the average of the AUC values from the
266 multiple binary comparisons of one class compared to all others (Hand and Till, 2001).

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

279 RESULTS

280 Geometric morphometrics

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

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

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

290 Comparison of the within to between Procrustes distances of the digitization replicates gives an approximate estimate of
291 the error between distinct groupings (Table 3). The ratio of the average within-individual distance to the average distance
292 between individuals for the replicated datasets is 1.11; this indicates that the grouping is slightly counter-intuitive to the
293 data and is consistent with all shapes being very similar regardless of individual identity. This value also provides a
294 baseline by which to understand how distinct the groupings are, where other ratios are compared to the correction ratio
295 1.11/1.

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

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

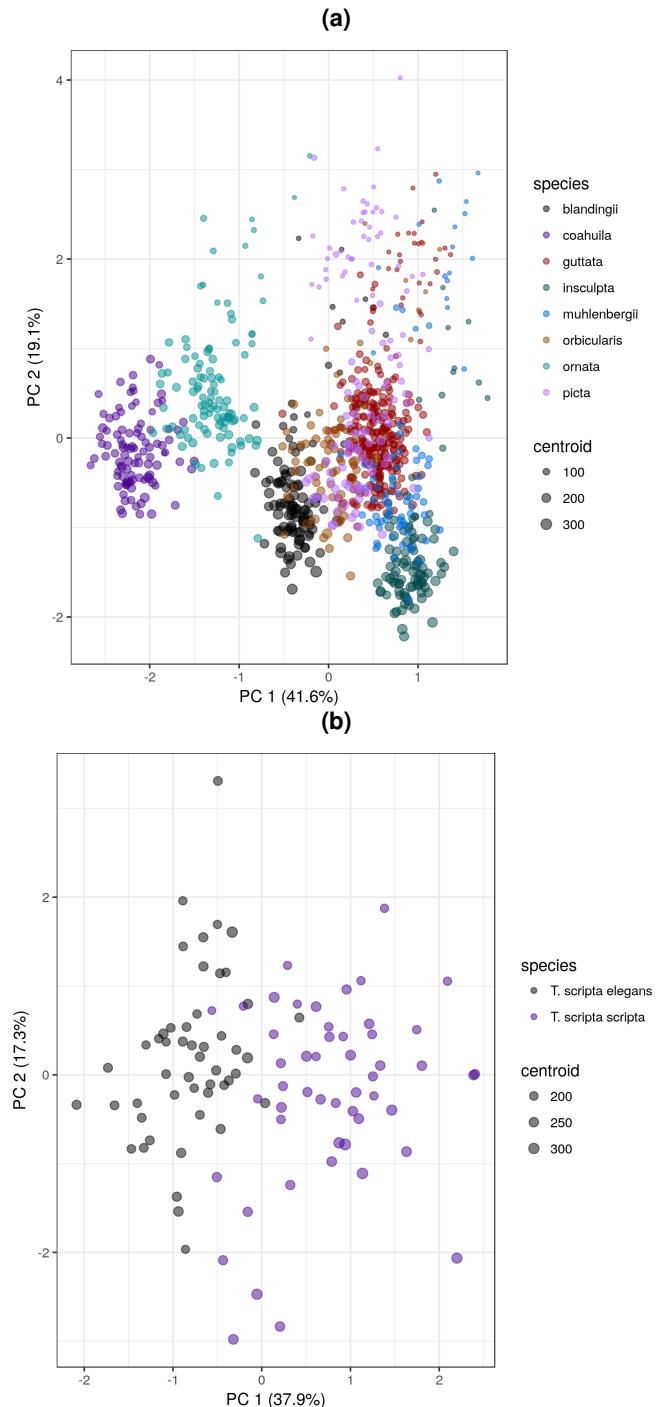


Figure 3. Two scatterplots of morphological differences from two of the three datasets analyzed in this study. (a) Scatterplot of the first two PCA axes from the landmarks from the eight different species dataset, and (b) the first two axes of variation from two subspecies of *Trachemys* dataset. Point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

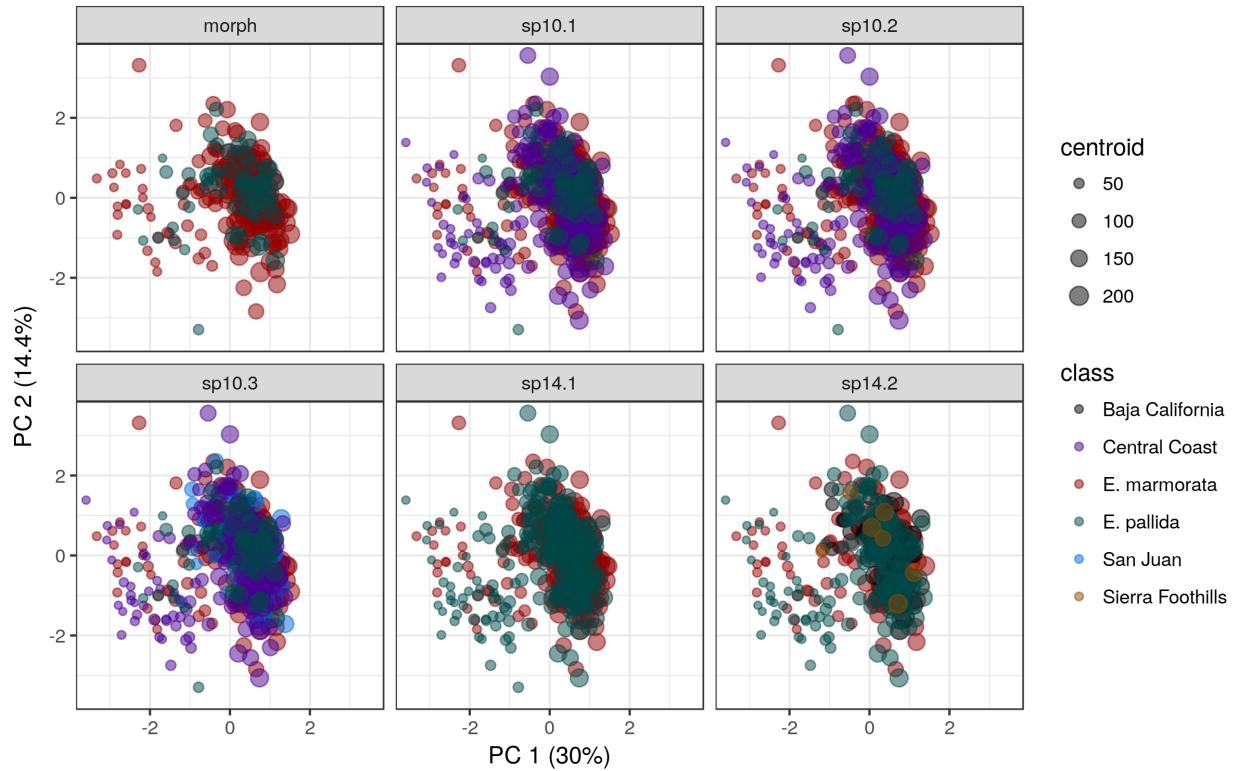


Figure 4. Scatterplot of the first two axes of morphological variation in the *Emys marmorata* dataset. Each panel corresponds to one of the six different classification schemes analyzed as part of this study (Tab. 1). Point color corresponds to the categories within each scheme, and the class names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

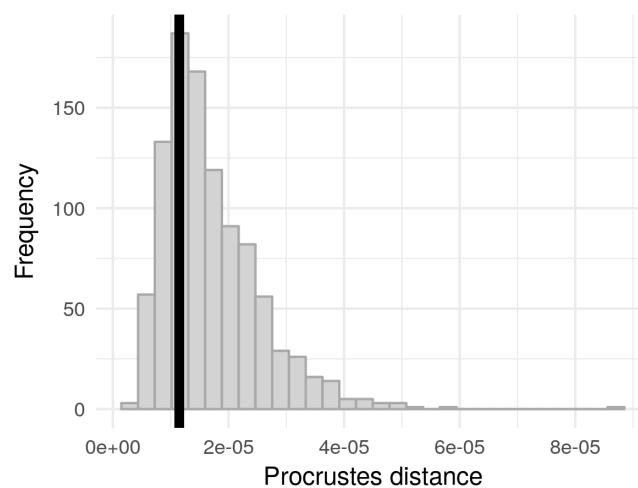


Figure 5. Comparison of observed Procrustes distance between the centroids of each sex (vertical line) to a null distribution generated from 1000 permutations of the sex-labels. This result indicates that the difference between the centroids is as small/smaller than expected by random.

299 that the magnitude of the differences between groupings is approximately the same as the difference between any two
300 random individuals (Table 3).

301 **Supervised learning**

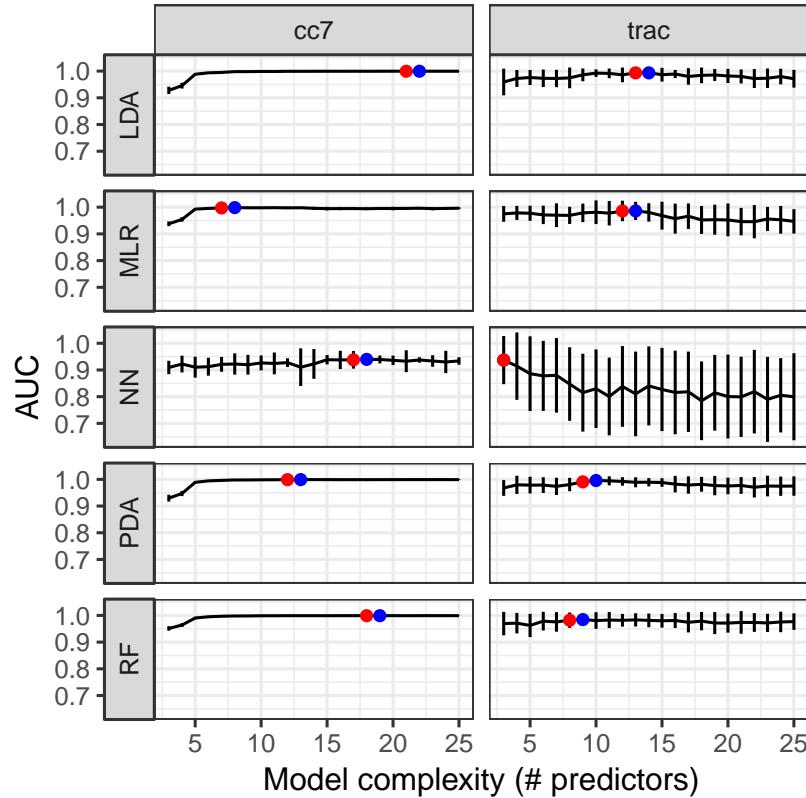


Figure 6. Comparisons of model fit to the training dataset for each of the supervised learning methods applied to the first two datasets; the results from the eight species dataset are presented in the left column, while those from the *Trachemys* dataset are presented in the right column. Models were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

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

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

312 Observed AUC values for all of the optimal models are lower for the *E. marmorata* dataset than for the other two

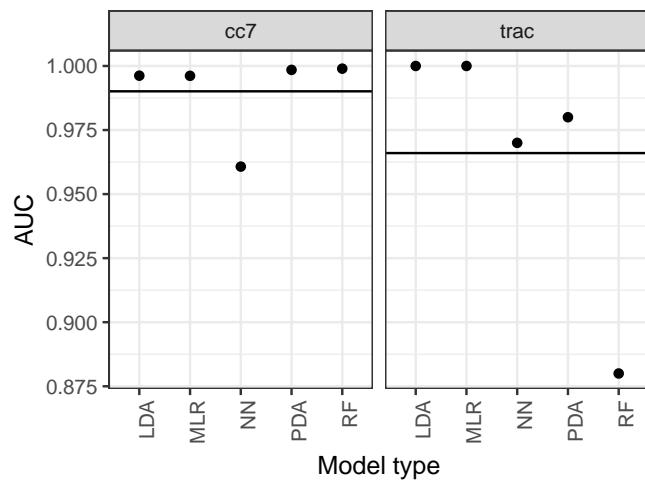


Figure 7. The results of out-of-sample predictive performance of the selected models for both the eight species (left) and *Trachemys* datasets. Predictive performance is measured as the area under the receiver operating characteristic (AUC), which ranges from 0.5 to 1. Points correspond to the individual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

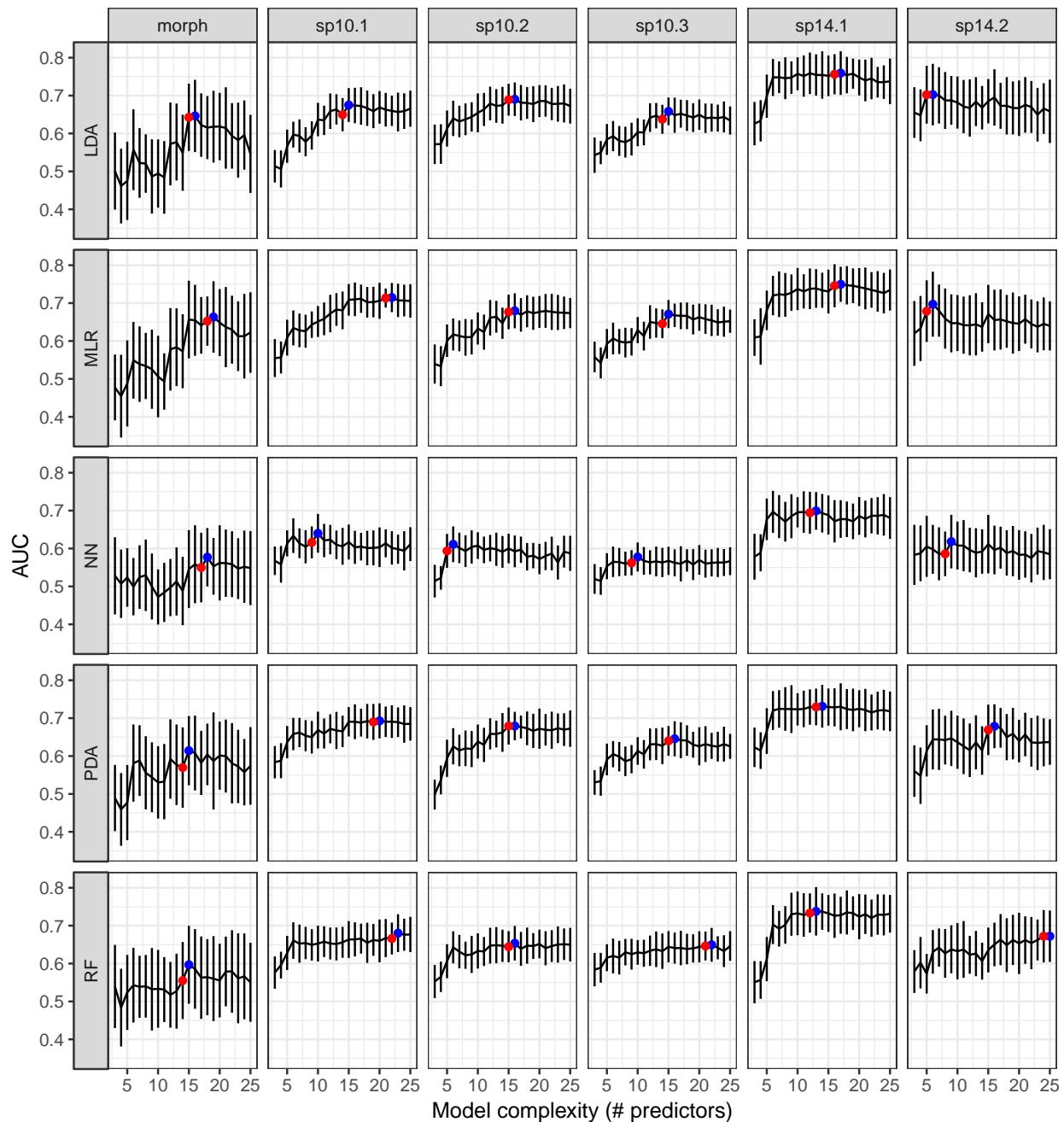


Figure 8. AUC values for models of varying complexity fit to the *Emys marmorata* training datasets for each classification scheme. The x-axis corresponds to the total number of predictors included in each model, while the y-axis corresponds to the AUC value which is a measure of goodness of fit for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

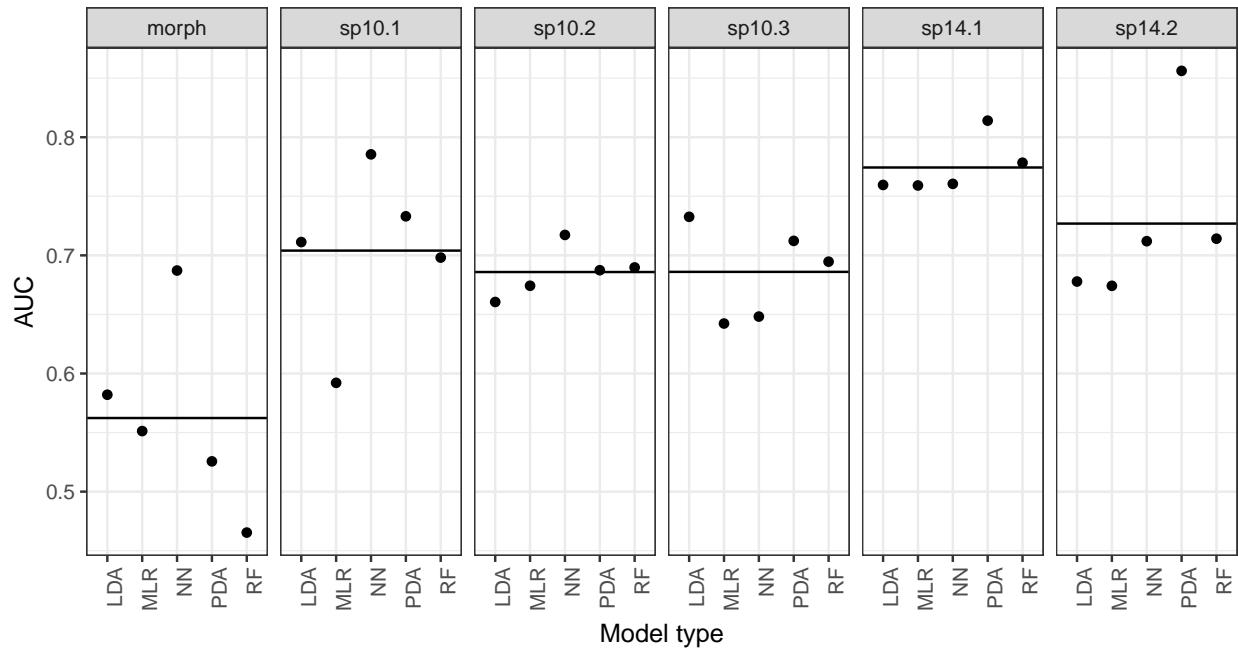


Figure 9. Comparison of out-of-sample AUC estimates from the predictions of selected models (Fig. 8), grouped by classification scheme. The horizontal line in each panel corresponds to the average AUC value across all models of that classification scheme.

313 datasets (Fig. 6, 8). In most cases the different proposed classification schemes are generally poor descriptors of the
314 observed variation. It appears that the dataset is overwhelmed by noise (likely biological and analytical), making any
315 accurate classifications difficult at best. This observation is cemented with the generalizations of the models to the
316 testing dataset (Fig. 9).

317 Mean AUC values for the model generalizations, in most cases, are approximately equal to the observed AUC values
318 from the training dataset (Fig. 8, 9). The cases in which the AUC from the generalizations is less than the observed
319 indicate poor model fit and a poor classification scheme. Comparison of AUC values from the model generalizations
320 do not indicate a clear “best” classification scheme (Fig. 8, 9). Only in the case of the conservative morphological
321 hypothesis (“Morph”) is the mean AUC value potentially distinct from that of other schemes; in this case mean AUC
322 is lower than the average of the other five schemes which indicates that the morphologically-based scheme performs
323 more poorly than the molecularly-based ones. It is important to note, however, that the training and testing dataset
324 for the “Morph” scheme is the smallest of the six schemes which may lead to poorer performance in in-sample and
325 out-of-sample comparisons.

326 DISCUSSION

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

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

339 Biological explanations include the possibility that genetic differentiation is not associated with plastron shape variation
340 and/or that local selective pressures (e.g. from hydrological regime) overwhelm morphological differentiation. Both of
341 these options seem plausible given that shell shape is influenced by selection for both protection and streamlining, but
342 not necessarily mate choice (Rivera, 2008; Rivera and Stayton, 2011; Stayton, 2011; Rivera et al., 2014; Polly et al.,
343 2016), and that shell shape in *E. marmorata* is known to vary among populations inhabiting water bodies with different
344 flow regimes (Holland, 1992; Lubcke and Wilson, 2007; Germano and Bury, 2009). Plastron shape does not seem to
345 preserve a strong phylogenetic signal at the interspecific level in emydine turtles, at least compared to the effect of
346 the presence or absence of a plastral hinge (Angielczyk et al., 2011), and our current results suggest that this may be
347 the case for phylogeographic signal within emydine species as well. A final possibility (explored below) is that the
348 proposed classification schemes themselves do not represent significant evolutionary lineages.

349 Despite the negative result for *E. marmorata*, it is important to note that plastron shape is an extremely effective method
350 for differentiating classes in the additional datasets we investigated. The magnitude of shape differences between
351 the species (measured as Procrustes distance between the eight species’ mean shapes) is approximately an order of
352 magnitude greater than the differences between the *E. marmorata* subgroups, and not surprisingly the machine learning
353 methods had no trouble classifying the specimens correctly. However, the magnitude of the shape differences between
354 the *T. scripta* subspecies is comparable to those separating the different *E. marmorata* subgroups, yet even in this
355 case the machine learning methods returned an almost perfect classification. These results demonstrate that plastron
356 shape is normally a good marker for differentiating real subgroups in close relatives of *E. marmorata*, and that our
357 lack of results for *E. marmorata* is not simply a shortcoming of the methods we applied. Indeed, it begs the question
358 of what factors have suppressed morphological differentiation of plastron shape in *E. marmorata* and *E. pallida* if
359 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 (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 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 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 emydids species leads to plastron shape being a useful tool for species delimitation, even 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?

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 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-diagnosability, it is essential to evaluate the methods and concepts that led to the initial 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 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 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 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 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; Spinks and Shaffer 2009; Parham et al. 2013) 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 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

408 specimens, as well as additional tests of species delimitation using alternative methods and corroborating evidence as
409 suggested by Carstens et al. (2013). From a morphological standpoint, support for the validity of "*E. pallida*" may
410 come from other aspects of morphology, such as carapace shape or other features. Likewise, further investigation of the
411 phylogeographic utility of plastron shape in other turtle species will help to clarify whether the lack of differentiation
412 seen in *E. marmoarata*, and the strong differentiation among the other emydids, is typical or an unusual case.

413 ACKNOWLEDGEMENTS

414 Data collection for this project was supported in part by NSF DBI-0306158 (to KDA). G. Miller assisted with data
415 collection and her participation in this research was supported by NSF REU DBI-0353797 (to R. Mooi of CAS). For
416 access to emydine specimens, we thank: J. Vindum and R. Drewes (CAS); A. Resetar (FMNH); R. Feeney (LACM); C.
417 Austin (LSUMNS); S. Sweet (MSE); J. McGuire and C. Conroy (MVZ); A. Wynn (NMNH); P. Collins (SBMNH); B.
418 Hollingsworth (SDMNH); P. Holroyd (UCMP). We are grateful for S. Sweet for field assistance and the California
419 Department of Fish and Game for permits. We would also like to thank M. Lambruschi (FMNH) for help with figure 1.

420 1 BIBLIOGRAPHY

- 421 Angielczyk, K. D. and Feldman, C. R. (2013). Are diminutive turtles miniaturized? The ontogeny of plastron shape in
422 emydine turtles. *Biological Journal of the Linnean Society*, 108(4):727–755.
- 423 Angielczyk, K. D., Feldman, C. R., and Miller, G. R. (2011). Adaptive evolution of plastron shape in emydine turtles.
424 *Evolution*, 65(2):377–394.
- 425 Angielczyk, K. D. and Sheets, H. D. (2007). Investigation of simulated tectonic deformation in fossils using geometric
426 morphometrics. *Paleobiology*, 33(1):125–148.
- 427 Arnqvist, G. and Mårtensson, T. (1998). Measurement error in geometric morphometrics: Empirical strategies to assess
428 and reduce its impact on measures of shape.
- 429 Baird, S. F. and Girard, C. (1852). Descriptions of new species of reptiles collected by the U.S. Exploring Expedition
430 under the command of Capt. Charles Wilkes. *Proceedings of the National Academy of Sciences Philadelphia*,
431 6:174–177.
- 432 Bauer, A. M., Parham, J. F., Brown, R. M., Stuart, B. L., Grismer, L., Papenfuss, T. J., Bohme, W., Savage, J. M.,
433 Carranza, S., Grismer, J. L., Wagner, P., Schmitz, A., Ananjeva, N. B., and Inger, R. F. (2000). Availability of new
434 Bayesian-delimited gecko names and the importance of character-based species descriptions. *Proceedings of the
435 Royal Society B: Biological Sciences*, 278:490–492.
- 436 Baylac, M., Villemant, C., and Simbolotti, G. (2003). Combining geometric morphometrics with pattern recognition for
437 the investigation of species complexes. *Biological Journal of the Linnean Society*, 80:89–98.
- 438 Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K., and Das, I. (2007).
439 Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution*, 22(3):148–55.
- 440 Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. Wadsworth
441 International Group, Belmont.
- 442 Bury, R. B. (2017). Biogeography of Western Pond Turtles in the western Great Basin: Dispersal Across a Northwest
443 Passage ? *Western Wildlife2*, 4:72–80.
- 444 Bury, R. B., Germano, D. J., and Bury, G. W. (2010). Population Structure and Growth of the Turtle *Actinemys
445 marmorata* from the Klamath–Siskiyou Ecoregion: Age, Not Size, Matters. *Copeia*, 2010(3):443–451.
- 446 Buskirk, S. W., Parham, J. F., and Feldman, C. R. (2005). On the hybridisation between two distantly related Asian
447 turtles (Testudines: Scincidae x Mauremys). *Salamandra*, 41:21–26.
- 448 Cardini, A., Nagorsen, D., O'Higgins, P., Polly, P. D., Thorington Jr, R. W., and Tongiorgi, P. (2009). Detecting
449 biological distinctiveness using geometric morphometrics: an example case from the Vancouver Island marmot.
450 *Ethology Ecology & Evolution*, 21:209–223.
- 451 Carstens, B. C. and Dewey, T. A. (2010). Species Delimitation Using a Combined Coalescent and Information-Theoretic
452 Approach: An Example from North American Myotis Bats. *Systematic Biology*, 59(4):400–414.

- 453 Carstens, B. C., Pelletier, T. a., Reid, N. M., and Satler, J. D. (2013). How to fail at species delimitation. *Molecular*
454 *ecology*, 22(17):4369–83.
- 455 Caumul, R. and Polly, P. D. (2005). Phylogenetic and environmental components of morphological variation: skull,
456 mandible, and molar shape in marmots (Marmota, Rodentia). *Evolution; international journal of organic evolution*,
457 59(11):2460–72.
- 458 Clare, E. L. (2011). Cryptic species? Patterns of maternal and paternal gene flow in eight neotropical bats. *PloS one*,
459 6(7):e21460.
- 460 Claude, J. (2006). Convergence induced by plastral kinesis and geometric morphometric assessment: a geometric
461 morphometric assessment. *Fossil Turtle Research*, 1:34–45.
- 462 Claude, J., Paradis, E., Tong, H., and Auffray, J. C. (2003). A geometric morphometric assessment of the effects
463 of environment and cladogenesis on the evolution of the turtle shell. *Biological Journal of the Linnean Society*,
464 79(December):485–501.
- 465 Dayrat, B. (2005). Towards integrative taxonomy. *Biological Journal of the Linnean Society*, 85:407–415.
- 466 De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic Biology*, 56(6):879–86.
- 467 Demandt, M. H. and Bergek, S. (2009). Identification of cyprinid hybrids by using geometric morphometrics and
468 microsatellites. *Journal of Applied Ichthyology*, 25(6):695–701.
- 469 Dillard, K. C. (2017). *A comparative analysis of geometric morphometrics across two Pseudemys turtle species in east*
470 *central Virginia*. Masters, Virginia Commonwealth University.
- 471 Dobigny, G., Granjon, L., Aniskin, V., Ba, K., and Voloboulev, V. (2003). A new sigling species of Taterillus (Muridae,
472 Gerbillinae) from West Agrica. *Mammalian Biology*, 68:299–316.
- 473 Dryden, I. L. (2013). *shapes: Statistical shape analysis*. R package version 1.1-8.
- 474 Dryden, I. L. and Mardia, K. Y. (1998). *Statistical shape analysis*. Wiley, New York.
- 475 Edwards, S., Claude, J., Van Vuuren, B. J., and Matthee, C. A. (2011). Evolutionary history of the Karoo bush rat,
476 *Myotomys unisulcatus* (Rodentia: Muridae): Disconcordance between morphology and genetics. *Biological Journal*
477 *of the Linnean Society*, 102(3):510–526.
- 478 Eldredge, N. and Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In Schopf, T. J. M.,
479 editor, *Models in Paleobiology*, pages 82–115. Freeman Cooper, San Francisco.
- 480 Feldman, C. R. and Parham, J. F. (2002). Molecular phylogenetics of emydine turtles: taxonomic revision and the
481 evolution of shell kinesis. *Molecular Phylogenetics and Evolution*, 22(3):388–98.
- 482 Francoy, T. M., Silva, R. A. O., Nunes-Silva, P., Menezes, C., and Imperatriz-Fonseca, V. L. (2009). Gender identification
483 of five genera of stingless bees (Apidae, Meliponini) based on wing morphology. *Genetics and molecular research*,
484 8(1):207–214.
- 485 Fruciano, C., Franchini, P., Raffini, F., Fan, S., and Meyer, A. (2016). Are sympatrically speciating Midas cichlid fish
486 special? Patterns of morphological and genetic variation in the closely related species *Archocentrus centrarchus*.
487 *Ecology and Evolution*, 6(12):4102–4114.
- 488 Funk, W. C., Caminer, M., and Ron, S. R. (2012). High levels of cryptic species diversity uncovered in Amazonian
489 frogs. *Proceedings of the Royal Society B: Biological Sciences*, 279(1734):1806–14.
- 490 Gaubert, P., Taylor, P. J., Fernandes, C. a., Bruford, M. W., and Veron, G. (2005). Patterns of cryptic hybridization
491 revealed using an integrative approach: a case study on genets (Carnivora, Viverridae, Genetta spp.) from the southern
492 African subregion. *Biological Journal of the Linnean Society*, 86(1):11–33.
- 493 Germano, D. J. and Bury, R. B. (2009). Variation in body size, growth, and population structure of *Actinemys marmorata*
494 from lentic and lotic habitats in Southern Oregon. *Journal of Herpetology*, 43(3):510–520.
- 495 Germano, D. J. and Rathbun, G. B. (2008). Growth, population structure, and reproduction of western pond turtles
496 (*Actinemys marmorata*) on the Central Coast of California. *Chelonian Conservation and Biology*, 7(2):188–194.
- 497 Gould, S. J. and Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*,
498 3(2):115–151.
- 499 Guillot, G., Renaud, S., Ledevin, R., Michaux, J., and Claude, J. (2012). A unifying model for the analysis of phenotypic,
500 genetic, and geographic data. *Systematic Biology*, 61(6):897–911.
- 501 Gündüz, I., Jaarola, M., Tez, C., Yeniyurt, C., Polly, P. D., and Searle, J. B. (2007). Multigenic and morphometric
502 differentiation of ground squirrels (*Spermophilus*, Sciuridae, Rodentia) in Turkey, with a description of a new species.
503 *Molecular phylogenetics and evolution*, 43(3):916–35.

- 504 Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class
 505 Classification Problems. *Machine Learning*, 45:171–186.
- 506 Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and
 507 prediction*. Springer, New York, 2nd edition.
- 508 Hastie, T., Tibshirani, R., Leisch, F., Hornik, K., and Ripley., B. D. (2015). *mda: Mixture and Flexible Discriminant
 509 Analysis*. R package version 0.4-8.
- 510 Hausdorf, B. and Hennig, C. (2010). Species delimitation using dominant and codominant multilocus markers.
 511 *Systematic biology*, 59(5):491–503.
- 512 Holland, D. C. (1992). *Level and pattern in morphological variation: a phylogeographic study of the western pond
 513 turtle (Clemmys marmorata)*. PhD thesis, University of Southwestern Louisiana.
- 514 Huelsenbeck, J. P., Andolfatto, P., and Huelsenbeck, E. T. (2011). Structurama: bayesian inference of population
 515 structure. *Evolutionary bioinformatics online*, 7:55–9.
- 516 Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data : an introduction to cluster analysis*. Wiley, New
 517 York.
- 518 Kendall, D. G. (1977). The diffusion of shape. *Advances in Applied Probability*, 9(3):428–430.
- 519 Kuhn, M. (2013). *caret: Classification and Regression Training*. R package version 5.15-61.
- 520 Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY.
- 521 Leaché, A. D. and Fujita, M. K. (2010). Bayesian species delimitation in West African forest geckos (Hemidactylus
 522 fasciatus). *Proceedings. Biological sciences / The Royal Society*, 277(1697):3071–7.
- 523 Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
- 524 Lubcke, G. M. and Wilson, D. S. (2007). Variation in shell morphology of the Western Pond Turtle (Actinemys
 525 marmorata Baird and Giarard) from three aquativ habitats in Northern California. *Journal of Herpetology*, 41(1):107–
 526 114.
- 527 MacLeod, N. (2007). *Automated taxon identification in systematics: theory, approaches and applications*. CRC Press,
 528 Boca Raton.
- 529 Markolf, M., Rakotonirina, H., Fichtel, C., von Grumbkow, P., Bräuer, M., and Kappeler, P. M. (2013). True
 530 lemurs... true species - species delimitation using multiple data sources in the brown lemur complex. *BMC
 531 Evolutionary Biology*, 13(1):233.
- 532 Mitrovska-Bogdanovic, A., Petrovic, A., Mitrovic, M., Ivanovic, A., Žikic, V., Starý, P., Vorburger, C., and Tomanovic,
 533 Ž. (2013). Identification of two cryptic species within the Praon abjectum group (Hymenoptera: Braconidae:
 534 Aphidiinae) using molecular markers and geometric morphometrics. *Annals of the entomological society of America*,
 535 106(2):170–180.
- 536 Munoz-Munoz F. and Perpinan D. (2010). Measurement error in morphometric studies: comparison between manual
 537 and computerized methods. *Ann. Zool.*, 47(1):46–56.
- 538 Navega, D., Vicente, R., Vieira, D. N., Ross, A. H., and Cunha, E. (2015). Sex estimation from the tarsal bones in a
 539 Portuguese sample: a machine learning approach. *International Journal of Legal Medicine*, 129(3):651–659.
- 540 O'Meara, B. C. (2010). New heuristic methods for joint species delimitation and species tree inference. *Systematic
 541 biology*, 59(1):59–73.
- 542 Padial, J. M., Miralles, A., De la Riva, I., and Vences, M. (2010). The integrative future of taxonomy. *Frontiers in
 543 Zoology*, 7(16):1–14.
- 544 Parham, J. F., Papenfuss, T. J., Dijk, P. P. V., Wilson, B. S., Marte, C., Schettino, L. R., and Brian Simison, W. (2013).
 545 Genetic introgression and hybridization in Antillean freshwater turtles (*Trachemys*) revealed by coalescent analyses
 546 of mitochondrial and cloned nuclear markers. *Molecular phylogenetics and evolution*, 67(1):176–87.
- 547 Pfenniger, M. and Schwenk, K. (2007). Cryptic animal species are homogeneously distributed among taxa and
 548 biogeographical regions. *BMC evolutionary biology*, 7:121.
- 549 Polly, P. D. (2003). Paleophylogeography of *Sorex araneus*: molar shape as a morphological marker for fossil shrews.
 550 *Mammalia*, 68(2):233–243.
- 551 Polly, P. D. (2007). Phylogeographic differentiation in *Sorex araneus*: morphology in relation to geography and
 552 karyotype. *Russian Journal of Theriology*, 6(1):73–84.
- 553 Polly, P. D., Stayton, C. T., Dumont, E. R., Pierce, S. E., Rayfield, E. J., and Angielczyk, K. D. (2016). Combining
 554 geometric morphometrics and finite element analysis with evolutionary modeling: towards a synthesis. *Journal of*

- 555 *Vertebrate Paleontology*, 4634(March).
- 556 Pons, J., Barracough, T., Gomez-Zurita, J., Cardoso, A., Duran, D., Hazell, S., Kamoun, S., Sumlin, W., and Vogler, A.
557 (2006). Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology*,
558 55(4):595–609.
- 559 R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical
560 Computing, Vienna, Austria.
- 561 Rivera, G. (2008). Ecomorphological variation in shell shape of the freshwater turtle *Pseudemys concinna* inhabiting
562 different aquatic flow regimes. *Integrative and comparative biology*, 48(6):769–87.
- 563 Rivera, G., Davis, J. N., Godwin, J. C., and Adams, D. C. (2014). Repeatability of Habitat-Associated Divergence in
564 Shell Shape of Turtles. *Evolutionary Biology*, pages 29–37.
- 565 Rivera, G. and Stayton, C. T. (2011). Finite element modeling of shell shape in the freshwater turtle *Pseudemys*
566 *concinna* reveals a trade-off between mechanical strength and hydrodynamic efficiency. *Journal of morphology*,
567 272(10):1192–203.
- 568 Rohlf, F. J. (2005). TpsDig 2.04.
- 569 Schilck-Steiner, B. C., Seifert, B., Stauffer, C., Christian, E., Crozier, R. H., and Steiner, F. M. (2007). Without
570 morphology, cryptic species stay in taxonomic crypsis following discovery. *Trends in ecology & evolution*, 22(8):391–
571 392.
- 572 Seeliger, L. M. (1945). Variation in the Pacific Mud Turtle. *Copeia*, 1945(3):150–159.
- 573 Spinks, P. Q. and Shaffer, H. B. (2005). Range-wide molecular analysis of the western pond turtle (*Emys marmorata*):
574 cryptic variation, isolation by distance, and their conservation implications. *Molecular ecology*, 14(7):2047–64.
- 575 Spinks, P. Q. and Shaffer, H. B. (2009). Conflicting mitochondrial and nuclear phylogenies for the widely disjunct *Emys*
576 (Testudines: Emydidae) species complex, and what they tell us about biogeography and hybridization. *Systematic
577 biology*, 58(1):1–20.
- 578 Spinks, P. Q., Thomson, R. C., and Bradley Shaffer, H. (2014). The advantages of going large: genome wide SNPs
579 clarify the complex population history and systematics of the threatened western pond turtle. *Molecular Ecology*,
580 pages n/a–n/a.
- 581 Spinks, P. Q., Thomson, R. C., and Shaffer, H. B. (2010). Nuclear gene phylogeography reveals the historical legacy
582 of an ancient inland sea on lineages of the western pond turtle, *Emys marmorata* in California. *Molecular ecology*,
583 19(3):542–56.
- 584 Stayton, C. T. (2011). Biomechanics on the half shell: functional performance influences patterns of morphological
585 variation in the emydid turtle carapace. *Zoology (Jena, Germany)*, 114(4):213–23.
- 586 Stuart, B. L., Inger, R. F., and Voris, H. K. (2006). High level of cryptic species diversity revealed by sympatric lineages
587 of Southeast Asian forest frogs. *Biology letters*, 2(3):470–4.
- 588 Sztencel-Jablonka, A., Jones, G., and Bogdanowicz, W. (2009). Skull Morphology of Two Cryptic Bat Species: *Pipistrellus*
589 *pipistrellus* and *P. pygmaeus* — A 3D Geometric Morphometrics Approach with Landmark Reconstruction.
590 *Acta Chiropterologica*, 11(1):113–126.
- 591 Van Bocxlaer, B. and Hunt, G. (2013). Morphological stasis in an ongoing gastropod radiation from Lake Malawi.
592 *Proceedings of the National Academy of Sciences*.
- 593 Van Bocxlaer, B. and Schultheiß, R. (2010). Comparison of morphometric techniques for shapes with few homologous
594 landmarks based on machine-learning approaches to biological discrimination. *Paleobiology*, 36(3):497–515.
- 595 van den Brink, V. and Bokma, F. (2011). Morphometric shape analysis using learning vector quantization neural
596 networks — an example distinguishing two microtine vole species. *Annales Zoologici Fennici*, 48:359–364.
- 597 Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.
- 598 Vitek, N. S., Manz, C. L., Gao, T., Bloch, J. I., Strait, S. G., and Boyer, D. M. (2017). Semi-supervised determination of
599 pseudocryptic morphotypes using observer-free characterizations of anatomical alignment and shape. *Ecology and
600 Evolution*, 7(14):5041–5055.
- 601 von Cramon-Taubadel, N., Frazier, B. C., and Lahr, M. M. (2007). The problem of assessing landmark error in geometric
602 morphometrics: theory, methods, and modifications. *American journal of physical anthropology*, 132(4):535–544.
- 603 Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the
604 National Academy of Sciences*, 107(20):9264–9.
- 605 Zelditch, M. L., Swiderski, D. L., and Sheets, H. D. (2004). *Geometric morphometrics for biologists: a primer*. Elsevier

