Ensemble approaches for estimating congruence between species delimitation and morphological variation: A case study using the Pacific Pond Turtle (*Emys marmorata*)

Peter D Smits1\*, Kenneth D Angielczyk2, Bryan L Stuart3, James F Parham4,  
**1** Department of Integrative Biology, University of California – Berkeley, Berkeley, California, USA.  
**2** Integrative Research Center, Field Museum of Natural History, Chicago, Illinois, USA.  
**3** Section of Research and Collections, North Carolina Museum of Natural Sciences, Raleigh, North Carolina, USA.  
**4** John D. Cooper Archaeological and Paleontological Center, Department of Geological Sciences, California State University – Fullerton, Fullerton, California, USA.

\* 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 used landmark-based morphometric data to quantify plastron shape in 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 seven unambiguously distinct species closely related to *E. marmorata* and an outgroup (*Chrysemys picta*), and the other consisting of two subspecies of *Trachemys scripta*. The analyses of the clear-cut examples produced near perfect classifications, demonstrating that plastron shape typically is a useful marker for differentiating turtle species, and that the methods can recover correct results when an appropriate signal exists. However, there is no clear “best” grouping of *E. marmorata/“E. pallida”* based on plastron shape. Explanations for the lack of grouping in *E. marmorata* include the possibility that genetic differentiation is not associated with plastron shape variation below the species level and/or that local selective pressures (e.g., from hydrological regime) overwhelm morphological differentiation. A reconsideration of the methods used to delimit *“E. pallida,”* the lack of barriers to gene flow, the strong evidence for widespread admixture between lineages, and the fact that plastron shape can be used to delineate other emydid 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 . In attempts to streamline the documentation of biodiversity, several methods of species delimitation that rely almost entirely on genetic data have recently been proposed . Although strong caveats on the utility of these methods have been raised , they are nevertheless being used to name species .

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 . 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 studies focus on using morphometrics to discover differences between taxa that were identified by other means . Additionally, there has been work on automated taxon identification and classification of taxa into groups , as well as the development of models that combine genetic, phenotypic, and geographic data to infer evolutionary units of interest .

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 (ventral shell) shape of seven species of closely related turtles and an outgroup, plastron shape of two subspecies of a single turtle species, and plastron shape of the *Emys marmorata* species complex. In particular, we ask whether it is possible to determine which among a set of classification hypotheses best aligns with the observed morphology, and examine the implications of our results for the *E. marmorata* complex.

## Background and study system

Machine learning is an extension of known statistical methodology that emphasizes predictive accuracy and generality, often at the expense of the interpretability of individual parameters . Basic statistical approaches are supplemented by randomization, sorting, and partitioning algorithms, along with the maximization or minimization of summary statistics, in order to best estimate a general model for all data, both sampled and unsampled . Machine learning approaches have found use in medical research, epidemiology, economics, and automated identification of images such as handwritten zip codes . There are two major classes of machine learning methods: unsupervised and supervised learning. Unsupervised learning methods are used with unlabeled data where the underlying structure is estimated; they are analogous to clustering and density estimation methods . Supervised learning methods are used with labeled data where the final output of data is known and the rules for going from input to output are inferred. These are analogous to classification and regression models . Our application of the supervised learning approaches used in this study illustrates only a sampling of the various methods available for fitting classification models. The specific methods we used were chosen because they are suited for cases with more two or more response classes.

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

Here, we used an ensemble of supervised machine learning methods to compare the congruence of morphological data to different classification hypotheses. Each of these methods provide different advantages for understanding how to classify specimens, as well as the accuracy of the resulting classifications. Machine learning methods have been combined with geometric morphometric data to study shape variation in a variety of contexts, including automated taxon identification and classification of groups . 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 .

We analyzed the problem of whether there are distinct subspecies or cryptic species within the western pond turtle, *Emys marmorata* (formerly *Clemmys marmorata*; see ). *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 . Traditionally, *E. marmorata* was classified into two named subspecies: the northern *E. marmorata marmorata* and the southern *Emys marmorata pallida* , 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 did not formally include the Baja California populations of *E. marmorata* in either taxon, implying the existence of a third distinct but unnamed subspecies.

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

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

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

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

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

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

# Materials and methods

## Specimens, sampling, morphometrics

Three different geometric morphometric datasets describing turtle plastron variation were assembled for this analysis: 1) specimens from seven distinct emydine species and a deirochelyine outgroup (*C. picta*); 2) *T. scripta* specimens from the two main subspecies (*T. scripta elegans* and *T. scripta scripta*); and 3) *E. marmorata* specimens from across the species’ geographic range. The first two datasets are intended to serve as a test of whether machine learning techniques can differentiate species-level groupings of emydid turtles using plastron shape. We expect that the first case represents a low complexity dataset because of the high level of plastron shape disparity that exists among these species , whereas the second dataset should be relatively higher in complexity and more analogous to the *E. marmorata* example. We predict that the *E. marmorata* dataset should be of the highest complexity and our greatest challenge given the previous finding that only very subtle differences existed between geographically-distinct populations . The first dataset we analyzed includes 578 total specimens from the following species: *Chrysemys picta*, *Clemmys guttata*, *Emys blandigii*, *Emys orbicularis*, *Glyptemys insculpta*, *Glyptemys muhlenbergii*, *Terrapene coahuila*, and *Terrapene ornata*. These specimens are a subset of those used in previous studies .

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

The final dataset is of 532 adult *E. marmorata* museum specimens, though not all specimens were able to be assigned a class for all schemes (Fig. [fig:map]). These specimens represent a subset of those included in previous studies . Because previous genetic studies did not use vouchered specimens we were not able to directly sample the individuals in their studies. Instead, our specimen classifications were based solely on the geographic information and not explicit assignment using molecular data. For each taxonomic hypothesis, specimens were assigned to one of the possible classes based on geographic occurrence data recorded in museum collections. In cases where precise latitude and longitude information were not available we estimated them from other locality information. Because the exact barriers between different biogeographic regions are unknown and unclear, we represented each hypothesis with multiple possible realizations representing the classification uncertainty for specimens present at the geographic boundaries. The taxonomic hypotheses and sub-hypotheses for *E. marmorata* used here are presented in Table [tab:hypotheses] and Fig. [fig:map].

We used three main 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. [fig:map]). 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 . We used two versions of the binning scheme based on the results of Spinks et al. (2014): SP14.1 reflects their phylogenetic network analysis and SP14.2 corresponds to their Bayesian species delimitation analysis. SP14.2 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 . Finally, we proposed a conservative morphological scheme (“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.

Table of species delimitation hypotheses for E. marmorata

|  |  |  |
| --- | --- | --- |
| Abbreviation | Number of classes | citation |
| SP10.1 | 3 |  |
| SP10.2 | 3 |  |
| SP10.3 | 4 |  |
| SP14.1 | 2 |  |
| SP14.2 | 4 |  |
| Morph | 2 |  |

[tab:hypotheses]

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 to 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 , we used TpsDig 2.04 to digitize 19 two-dimensional landmarks (Fig. [fig:plastra]). Seventeen of the landmarks are at the endpoints or intersection of the keratinous 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 symmetric 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 , after which we calculated the principal components (PC) of shape using the shapes package for R . 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 on our results by comparing within-specimen (replicated) Procrustes distances to the distances between classification scheme centroids. Ten randomly-selected *E. marmorata* specimens were each digitized four times, with the original set of digitized coordinates serving as a fifth replicate. These 50 landmark configurations were 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 smaller on average 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 plastral concavity in males . 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). This analysis was also done for each of the classification schemes in order to determine if sex is a biasing factor within these groups. As with the total dataset, only a limited number of specimens in each classification scheme have sex information, and the sex ratios within each of the classification schemes are as follows: SP10.1 202 F/199 M, SP10.2 202/199, SP10.3 202/199, SP14.1 177/178, SP14.2 136/125, Morph 112/120.

## Supervised learning approaches

[tab:methods]

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 . The supervised learning methods used here are named in Table [tab:methods]. Each of these methods makes different assumptions, treats data differently, and can produce different classification results depending on the nature of the data . 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, the interaction between scaled centroid size and PC 1, and the interaction between scaled centroid size and PC 2. Additional interaction terms were not considered because of model complexity/sample size concerns. Size and the interaction between size and PCs 1 and 2 were included as predictors to account for known ontogenetic variation in plastron shape as well as potential size differences between classes, even if this is unlikely . 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 the reduction and orthogonality of PCs to improve regression fit . 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 . 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 . The area under the ROC curve (AUC) is the derived estimate of model performance. AUC ranges from 0.5 to 1, which correspond to performance similar to random guesses and perfect classification rates, respectively . Both ROC and AUC are preferable to simple classification accuracy when class membership is unbalanced, as it is in these analyses . 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 .

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 . In a single round of -fold cross-validation, the training data are divided into blocks where the model is fit to blocks and the values of the 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 -fold cross-validation. This process was implemented using the R package caret . 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 . The purpose of this rule is to ameliorate the chances of selecting an overly complex model that will perform poorly when predicting the classes of out-of-sample data.

# Results

## Geometric morphometrics

0.7

0.7

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. [fig:other\_pca]).

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. [fig:emys\_pca]). The first PC axis of shape variation appears to be primarily structured by differences in individual centroid size (Fig. [fig:emys\_pca]); a linear regression of PC 1 vs. log centroid size shows that size explains approximately 81 percent of the variance on this axis. This was the motivation for including centroid size and its interaction with PC1 as predictors in all of the supervised learning models. Size typically explains less than three percent of the variance on the remaining PC axes (e.g., 2.5% on PC2).

Analysis of the differences between sexes of *E. marmorata* indicates that sex does not appear to strongly structure differences in shape (Fig. [fig:sex\_test]). The difference in mean shape between the sexes is very small; the sexes overlap about has much as expected given a null distribution based on permuting the sex-labels. The results form sex comparisons within each classification scheme have similar results with none of the observed differences being significantly different from the permuted distribution, indicating that sex is most likely not a biasing factor in our analyses (Fig. [fig:sex\_test\_group]).

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 |  | 1.06 |  |
| Seven species |  | 0.43 | 0.46 |
| *Trachemys* |  | 0.76 | 0.81 |
| *E. marmorata* | SP10.1 | 0.99 | 1.05 |
|  | SP10.2 | 1.00 | 1.06 |
|  | SP10.3 | 0.95 | 1.01 |
|  | SP14.1 | 1.01 | 1.07 |
|  | SP14.2 | 0.94 | 1.00 |
|  | Morph | 0.99 | 1.05 |

[tab:rep\_res]

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

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 [tab:rep\_res]). In contrast, the different *E. marmorata* classification schemes appear to be barely 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 [tab:rep\_res]).

## Supervised learning

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

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

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

Mean AUC values for the model generalizations, in most cases, are approximately equal to the observed AUC values from the training dataset (Fig. [fig:emys\_sel], [fig:emys\_oos]). The cases in which the AUC from the generalizations is less than the observed indicate poor model fit and a poor classification scheme. Comparison of AUC values from the model generalizations do not indicate a clear “best” classification scheme (Fig. [fig:emys\_sel], [fig:emys\_oos]). Only in the case of the conservative morphological hypothesis (“Morph”) is the mean AUC value potentially distinct from that of other schemes; in this case mean AUC is lower than the average of the other five schemes which indicates that the moprhologically-based scheme performs more poorly than the molecularly-based ones. It is important to note, however, that the training and testing dataset for the “Morph” scheme is the smallest of the six schemes which may lead to poorer performance in in-sample and out-of-sample comparisons.

# Discussion

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

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

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 regime) overwhelm morphological differentiation. Both of these options seem plausible given that shell shape is influenced by selection for both protection and streamlining, but not necessary mate choice , and that shell shape in *E. marmorata* is known to vary among populations inhabiting water bodies with different flow regimes . Plastron shape does not seem to preserve a strong phylogenetic signal at the interspecific level in emydine turtles, at least compared to the effect of the presence or absence of a plastral hinge , and our current results suggest that this may be the case for phylogeographic signal within emydine species as well. A final possibility (explored below) is that the proposed classification schemes themselves do not represent significant evolutionary lineages.

Despite the negative result for *E. marmorata*, it is important to note that plastron shape is an extremely effective method for differentiating classes in the additional datasets we investigated. The magnitude of shape differences between the eight species (measured as Procrustes distance between the eight species’ mean shapes) is approximately an order of magnitude greater than the differences between the *E. marmorata* subgroups, and not surprisingly the machine learning methods had no trouble classifying the specimens correctly. However, the magnitude of the shape differences between the *T. scripta* subspecies is comparable to those separating the different *E. marmorata* subgroups, yet even in this case the machine learning methods returned an almost perfect classification. These results demonstrate that plastron shape is normally a good marker for differentiating real subgroups in close relatives of *E. marmorata*, and that our lack of results for *E. marmorata* is not simply a shortcoming of the methods we applied. Indeed, it begs the question of what factors have suppressed morphological differentiation of plastron shape in *E. marmorata* and *E. pallida* if they are distinct species. Invoking issues such as the role of the plastron in protection or the need for streamlining are insufficient because the other species are expected to be subject to similar constraints . 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 emydid 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.

# Conclusions

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-diagnosablity, it is essential to evaluate the methods and concepts that led to the initial taxonomic revision.

Spinks et al. elevated *E. pallida* based on a species delimitation analysis of SNP data using BPP . However, Spinks et al. did not heed the caveats about such species delimitation methods raised by Carstens et al. . 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. 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 , Integrative Taxonomy , 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.

The only other range-wide analysis of morphological variation is Seeliger (1945), which was cited by Spinks et al. as being “consistent” with their results. A closer examination of Seeliger (1945) raises some questions about that claim. For one, Spinks et al. claim that Seeliger (1945) suggested that “an area of extensive intergradation was restricted to the San Joaquin Valley.” But Seeliger (1945) actually suggest that her *E. m. pallida* and *E. m. marmorata* intergrade south of the San Francisco bay (Coast Ranges) and lists localities in Alameda and Contra Costa County. Spinks et al. are lacking samples from the Bay Area, but the few they do have show admixture. Thus, the putative distinctiveness of *E. marmorata* and *E. pallida* may be partially due to a sampling artifact. A second issue is that the characters used to distinguish the two taxa are not analyzed in a modern, quantitative context. Seeliger (1945) mentions distinct tendencies in coloration and different percentages of large vs. small inguinal scales. However, its not possible to know how these characters vary across the range because Seeliger (1945) doesn’t provide criteria for large vs. small inguinal scales or assign these states to particular specimens or sites. We encourage further studies of the morphological variation in these characters, especially in the broad intergrade zones.

Further following the advice of Carstens et al. we examine the natural history and geographical distribution of *E. marmorata* and *E. pallida* and find that they make the recognition of these two taxa implausible. The mitochondrial data from Spinks et al. show extensive introgression and admixture in Central California, which is likely not just a mitochondrial sweep because there are no significant barriers to nuclear gene flow in this region. Spinks et al. 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. ) 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. Other emydines show fairly high levels of genetic diversity that is often geographically structured (e.g., Amato et al. 2008 for *G. insculpta*; Davy and Murphy 2014 for *C. guttata*; Sethuraman et al. 2014 for *E. blandingii*) but are still regarded as single species. 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.

Amato, M.L., R.J. Brooks, and J. Fu. 2008. A phylogeographic analysis of populations of the wood turtle (Glyptemys insculpta) throughout its range. Molecular Ecology 17:570-581.

<http://www.uoguelph.ca/~jfu/FuLab/Publication_files/Amato.pdf>

Davy, C.M. and R.W. Murphy. Conservation genetics of the endangered Spotted Turtle (*Clemmys guttata*) illustrate the risks of “bottleneck tests”. Canadian Journal of Zoology 92:149-163.

<http://www.nrcresearchpress.com/doi/pdfplus/10.1139/cjz-2013-0188>

Sethuraman, A., S.E. McGaugh, M.L. Becker, C.H. Chandler, J.L. Christiansen, S. Hayden, A. LeClere, J. Monson-Miller, E.M. Myers, R.T. Paitz, J.M. Refsnider, T.J. VanDeWalle, and F.J. Janzen Population genetics of Blanding’s turtle (Emys blandingii) in the midwestern United States. Conservation Genetics 15:61-73.

http://www.public.iastate.edu/~fjanzen/pdf/14ConservGenet.pdf

We fully agree with Spinks et al. 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. 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 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*. However, tests of both our preferred conclusion (*E. marmorata* as a single species) and that of Spinks et al. should include morphological and molecular analyses of specimens from within the Bay Area intergrade zone, ideally the same set of voucher specimens. We also recommend additional tests of species delimitation using alternative methods and corroborating evidence as suggested by Carstens et al. . From a morphological standpoint, support for the validity of “*E. pallida*” may come from other aspects of morphology, such as carapace shape, or other features. Likewise, further investigation of the phylogeographic utility of plastron shape in other turtle species will help to clarify whether the lack of differentiation seen in *E. marmoarata*, and the strong differentiation among the other emydids, is typical or an unusual case.

# Supporting information

#### S1 Fig.

**Bold the title sentence.** Add descriptive text after the title of the item (optional).

# Acknowledgments

Data collection for this project was supported in part by NSF DBI-0306158 (to KDA). G. Miller assisted with data collection and her participation in this research was supported by NSF REU DBI-0353797 (to R. Mooi of CAS). For access to emydine specimens, 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); 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. We would also like to thank M. Lambruschi (FMNH) for help with figure [fig:map].