How cryptic is cryptic diversity? Machine learning approaches to classifying morphological variation in the Pacific Pond Turtle (*Emys marmorata*)

Peter D Smits[1], Kenneth D Angielczyk[1,2], James F Parham[3], and Bryan Stuart[4]

[1]Committee on Evolutionary Biology, University of Chicago
[2]Integrative Research Center, Field Museum of Natural History
[3]Department of Geological Sciences, California State University – Fullerton
[4]Section of Research and Collections, North Carolina Museum of Sciences

May 4, 2016

Corresponding author: Peter D Smits, Committee on Evolutionary Biology, University of Chicago, 1025 E. 57th Street, Culver Hall 402, Chicago, IL, 60637, USA; E-mail: psmits@uchicago.edu

Abstract

Introduction

Materials and Methods

Specimens, sampling, morphometrics

Three different landmark-based morphometric datasets describing plastron variation were assembled for this analysis: specimens from seven distinct emydine species, *T. scripta* specimens from both subspecies, and *E. marmorata* specimens from across its geographic range. We chose to focus on adults because significant changes in plastron shape occur over the course of ontogeny in *E. marmorata* and other emydines (Angielczyk and Feldman 2013).

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

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

The final dataset dataset included 354 adult *E. marmorata* museum specimens; a subset of those included in Angielczyk and Sheets (2007), Angielczyk et al. (2011), and Angielczyk and Feldman (2013). We assigned a classification to each specimen for the different binning schemes based on geographic occurrence data recorded in museum collection archives. When precise latitude and longitude information were not available we estimated them from locality information. Because Spinks and Shaffer (2005), Spinks et al. (2010), and Spinks et al. (2014) did not use vouchered specimens we were not able to directly sample the individuals in their studies. Therefore our specimen classifications were based solely on the geographic information, not explicit assignment using molecular data. Because the exact barriers between different biogeographic regions are unknown and unclear, we represented some hypothesis with two schemes for a total of six different schemes. These schemes differed based on where geographic boundaries were assigned. This changes the classification of certain individuals near the boundaries between groups, providing a test of the robustness of the classification schemes. Sex information was only know for a subset of the total dataset and was not included as a predictor of classification. Sex information was used to determine if observations cluster by sex or not. The scheme names are as follows: Mito 1 and 2 correspond to Spinks and Shaffer (2005), Mito 3 corresponds to Spinks et al. (2010), Morph 1 and Morph 2 correspond to Holland (1992), and Nuclear corresponds to Spinks et al. (2014).

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 landmarks (Fig. 1). Seventeen of the landmarks are at the endpoints or intersection of the keratinous plastral scutes that cover the plastron. Twelve of the landmarks were symmetrical across the axis of symmetry. Because damage prevented the digitization of all the symmetric landmarks in some specimens, we reflected landmarks across the axis of symmetry (i.e. midline) prior to analysis and used the average position of each symmetrical pair. In cases where damage or incompleteness prevented symmetric landmarks from being determined, we used only the single member of the pair. We conducted all subsequent analyses on the resulting "half" plastra. We superimposed the plastral landmark configurations using generalized Procrustes analysis (Dryden and Mardia 1998), after which, we calculated the principal components (PC) of shape using the `shapes` package for R (**?**Dryden 2013).

Biasing effects

We estimated the possible effect of digitization error CITATIONS on our results by comparing within (replicated) specimen Procrustes distances to the distances between classification scheme centroids. 10 randomly selected specimen both for this study and an additional four times. These 50 landmark configurations were then Procrustes superimposed. A range of four Procrustes distances were then calculated as the average of the pairwise distances between

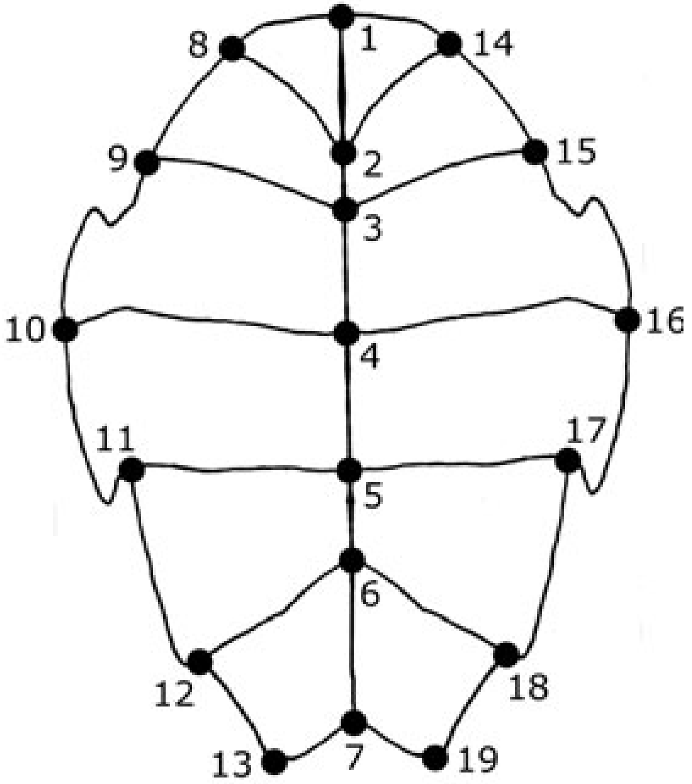

Figure 1: Depiction of general plastral shape of *E. marmorata* and position of the 19 landmark used in this study. Anterior is towards the top of the figure.

each of the replicate configurations. These distances were then compared to the average pairwise Procrustes distance between the class centroids for each classification schemes of *E. marmorata* being analyzed.

Supervised learning approaches

The maximum set of possible predictors or features used for any model are the first 25 principal components (PCs), scaled centroid size, and the interaction between scaled centroid size and PC 1. Size and the interaction between size and PC 1 were included as predictors in order to account for a possible interaction between size and shape over the duration of an individual as well as potential size differences between classes, even if this is unlikely (Seeliger 1945; Holland 1992). We say "maximum set" because the best or selected models based on 5-fold cross-validation does need not to, nor will they likely, include all predictors possible (see below).

This approach is in many ways analogous to PCA regresion. PCA regression takes advantage of two aspects of PCA for improving regression fit (Hastie et al. 2009). Because the PCs of shape are by definition orthagonal, allowing them to easily as independent predictors or features of class membership without fear of colinearity.

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

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 is split into a training dataset and a testing dataset. The former is used for fitting the model where as the later is used for measuring model performance, a process called model generalization. In this analysis, we used 80% of samples as the training set while the remaining 20% were used as the testing set.

For a given supervised learning method, we compared the fit of 27 models as the average AUC from 10 rounds of 5-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 data itself (Hastie et al. 2009). In a single round of k-fold cross-validation, the training data is divided into k blocks where the model is fit to $k - 1$ blocks and the values of the kth block are predicted; this is then repeated for all combinations of blocks. Within each round, the predictive performance metrics is averaged across all folds. Finally, the predictive performance metric is the averaged across all rounds of k-fold cross-validation. This process was implemented using the R package `caret` CITATION.

For a given supervised learning method, the “best” trained model is that the highest mean AUC as estimated from 5-fold cross-validation. The selected or final model, however, is the next most parsimonious model that is within one standard error of the best model; this is a variant on the “one-standard error” rule from Hastie et al. (2009).

Instead of relying on a single supervised learning method, we chose to use an ensemble of mulitple approaches so that the congruence bewteen the them could be used as a means of “support” for one conclusion or another. The supervised learning methods used here are described in Table 1. Each of these methods makes different assumptions, treat data differently, and can produce very different classification results (Hastie et al. 2009). The

Method name	abbreviation	R package	citation
multinomial logistic regression	MLR	nnet	citation
linear discriminate analysis	LDA	MASS	citation
penalized discrminiate analysis	PDA	mda	citation
single-hidden-layer neural network	NN	nnet	citation
random forests	RF	randomForest	citation

Table 1: table of the methods

common assumption of all of these methods is that the predictors or features are independent and/or have additive effects on prediction (Hastie et al. 2009).

Results

Biasing effects

Supervised learning

Discussion

*

Acknowledgements 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.

Bibliography

K. D. Angielczyk and C. R. Feldman. Are diminutive turtles miniaturized? The ontogeny of plastron shape in emydine turtles. *Biological Journal of the Linnean Society*, 108(4): 727–755, apr 2013. ISSN 00244066. doi: 10.1111/bij.12010. URL `http://doi.wiley.com/10.1111/bij.12010`.

K. D. Angielczyk and H. D. Sheets. Investigation of simulated tectonic deformation in fossils using geometric morphometrics. *Paleobiology*, 33(1):125–148, 2007.

K. D. Angielczyk, C. R. Feldman, and G. R. Miller. Adaptive evolution of plastron shape in emydine turtles. *Evolution*, 65(2):377–394, feb 2011. ISSN 1558-5646. doi: 10.1111/j.1558-5646.2010.01118.x.

I. L. Dryden. *shapes: Statistical shape analysis*, 2013. URL `http://CRAN.R-project.org/package=shapes`. R package version 1.1-8.

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

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

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

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

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

F. J. Rohlf. TpsDig 2.04, 2005.

L. M. Seeliger. Variation in the Pacific Mud Turtle. *Copeia*, 1945(3):150–159, 1945.

P. Q. Spinks and H. B. Shaffer. Range-wide molecular analysis of the western pond turtle (Emys marmorata): cryptic variation, isolation by distance, and their conservation implications. *Molecular ecology*, 14(7):2047–64, jun 2005. ISSN 0962-1083. doi: 10.1111/j.1365-294X.2005.02564.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/15910326`.

P. Q. Spinks, R. C. Thomson, and H. B. Shaffer. Nuclear gene phylogeography reveals the historical legacy of an ancient inland sea on lineages of the western pond turtle, Emys marmorata in California. *Molecular ecology*, 19(3):542–56, feb 2010. ISSN 1365-294X. doi: 10.1111/j.1365-294X.2009.04451.x. URL `http://www.ncbi.nlm.nih.gov/pubmed/20051011`.

P. Q. Spinks, R. C. Thomson, and H. Bradley Shaffer. The advantages of going large: genome wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. *Molecular Ecology*, pages n/a–n/a, mar 2014. ISSN 09621083. doi: 10.1111/mec.12736. URL `http://doi.wiley.com/10.1111/mec.12736`.

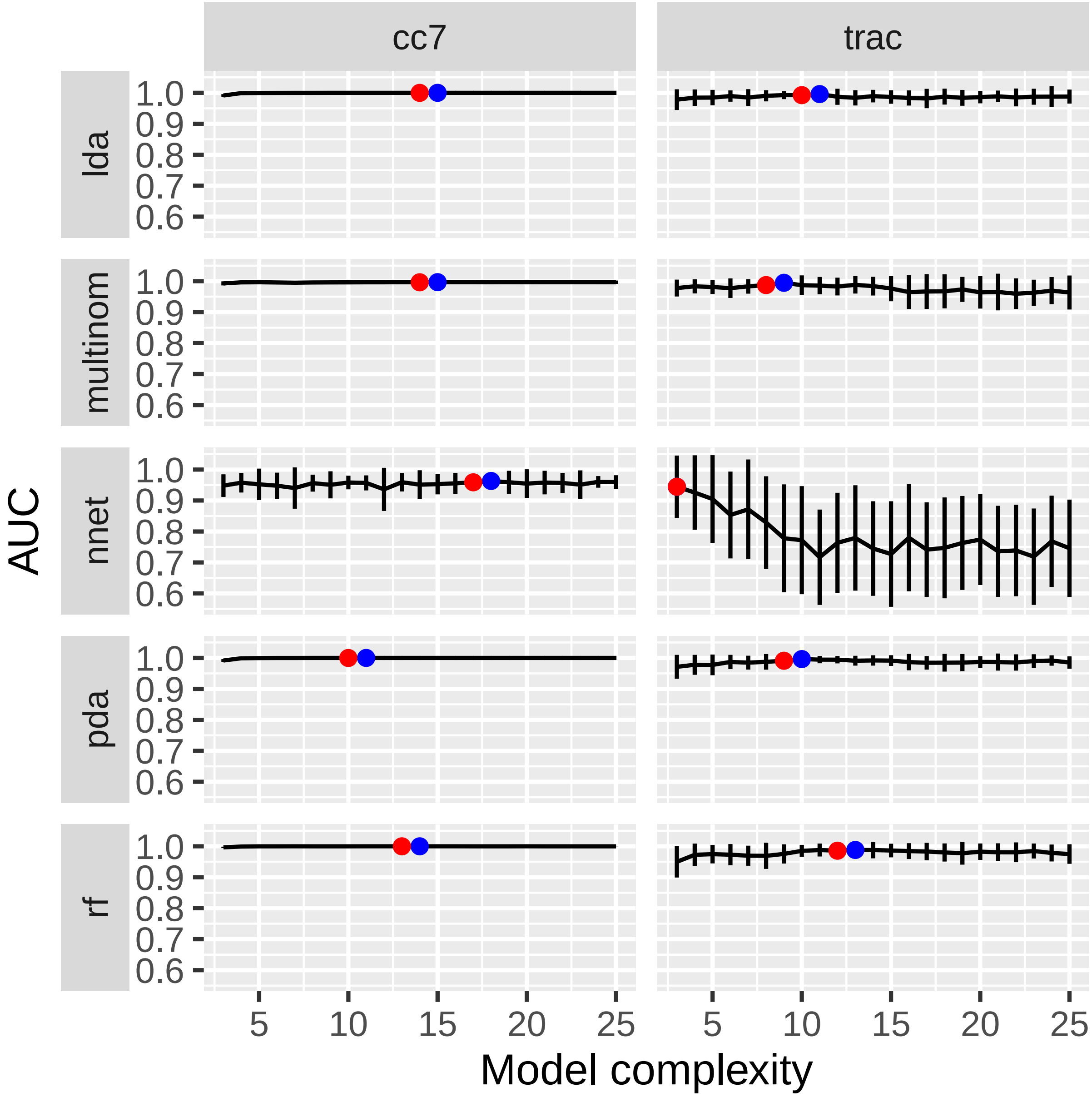

Figure 2: CAPTION

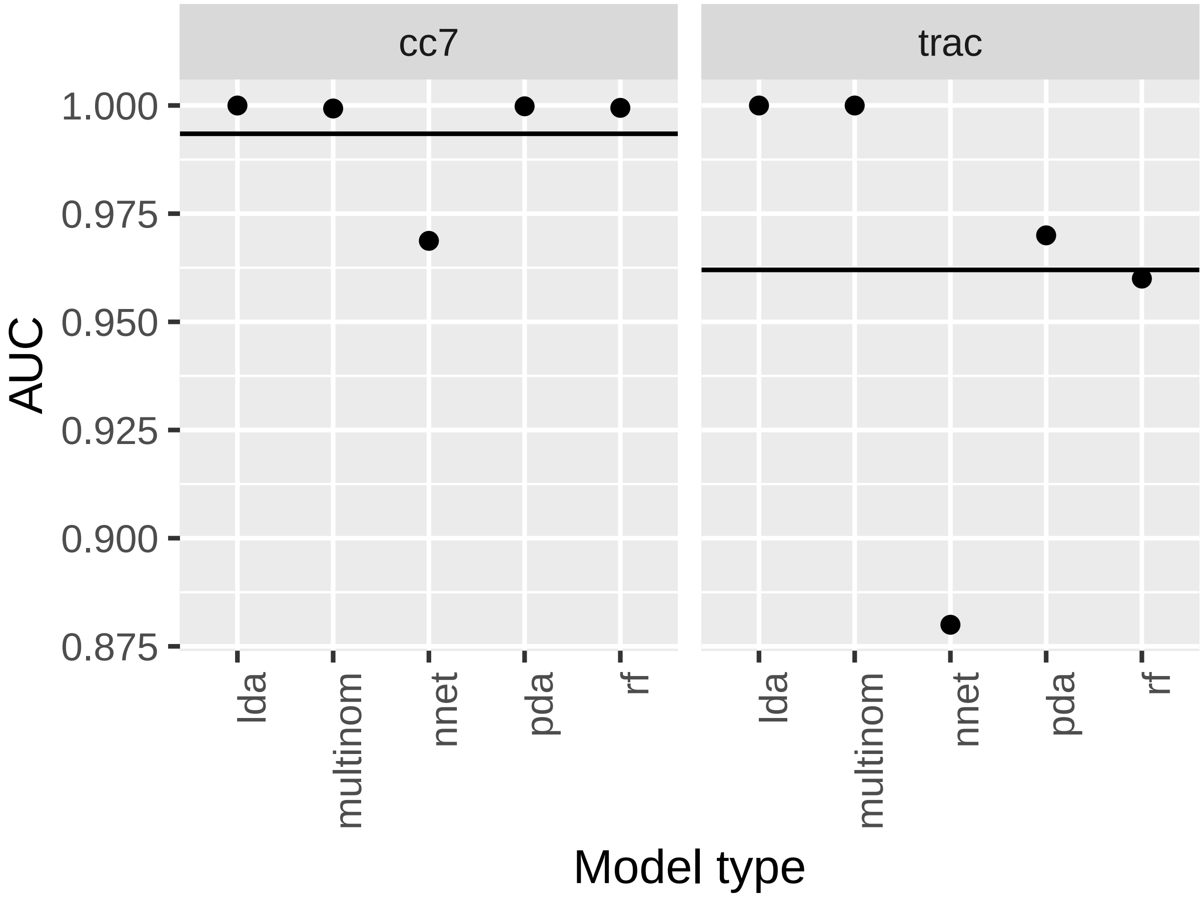

Figure 3: CAPTION

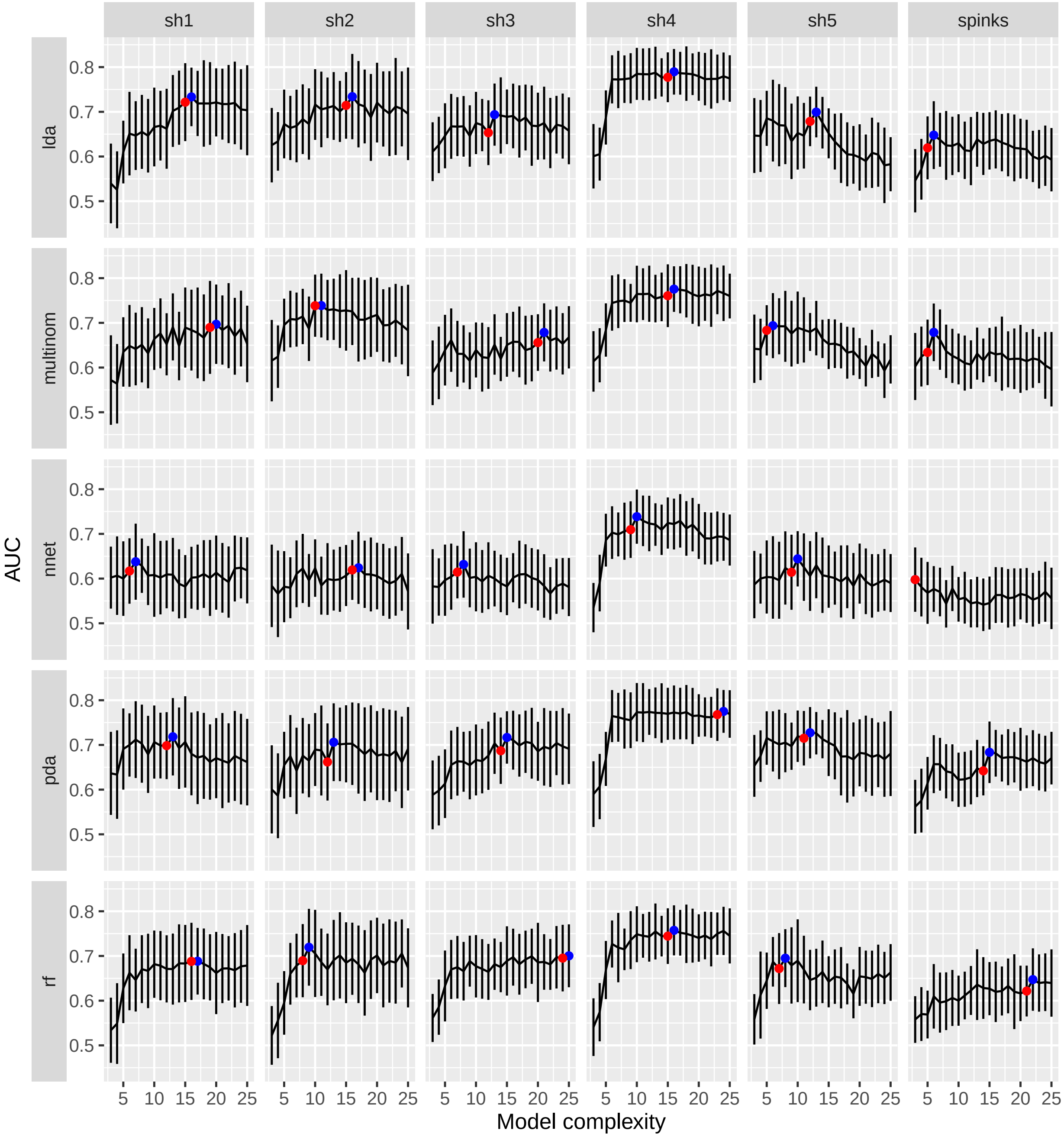

Figure 4: CAPTION

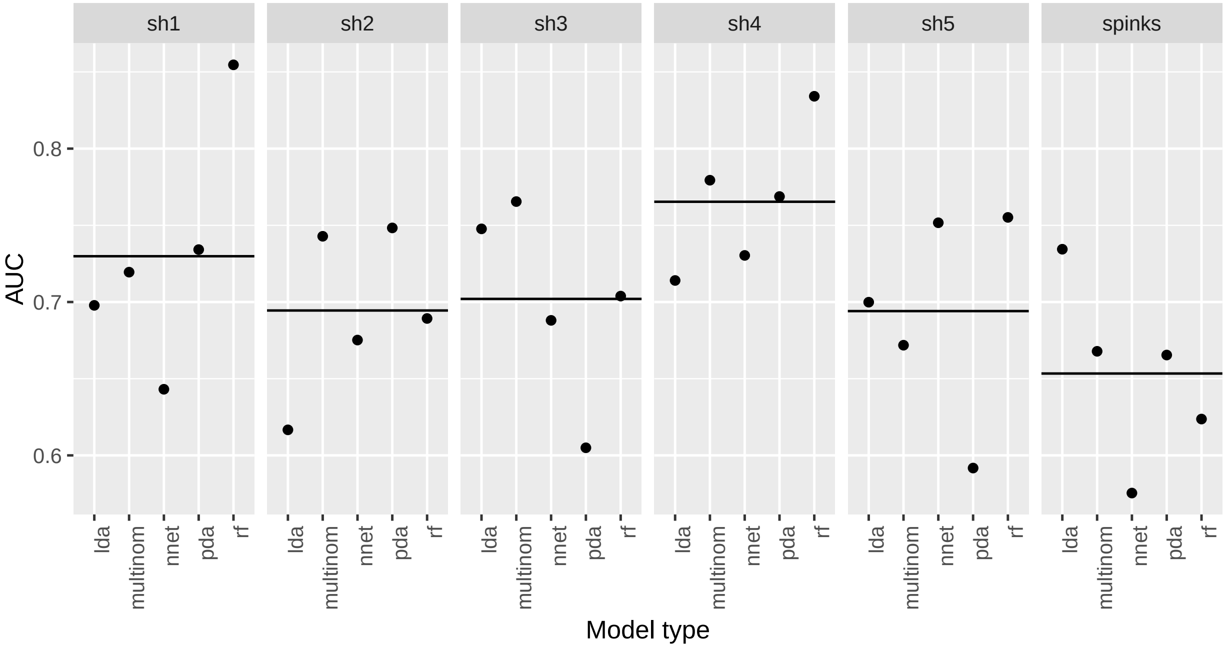

Figure 5: CAPTION