Method of the 'Area of Applicability' (AOA) explained in figures

Supplement to the paper 'Predicting into unknown space? Estimating the area of applicability of spatial prediction models' (Methods in Ecology and Evolution)

Hanna Meyer, Edzer Pebesma

17/04/2021

Contents

1 Introduction

This document is a supplement to the paper 'Predicting into unknown space? Estimating the area of applicability of spatial prediction models'. It visualizes the idea and workflow of the method to estimate the area of applicability (AOA) of prediction models. Note that the purpose of this document is to visualize the idea of the method. Hence, the dataset is small and not a scientifically meaningful prediction task. For application in more realistic applications, as well as for detailed explanations please see the manuscript. For reproducing this document, please use the Rmd file available at https://github.com/HannaMeyer/MEE_AOA.

2 General problem of random forests in “unknown environments”

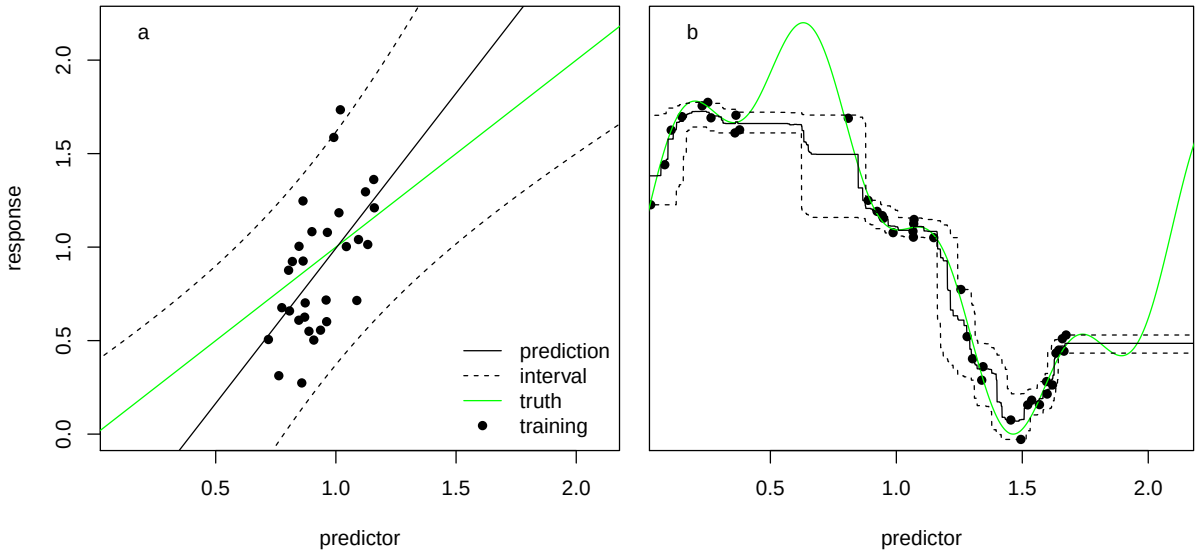


Figure 1: Problem of predicting beyond the training data and behaviour of prediction intervals for different models. Left: linear regression prediction interval width increases with distance from the center of the training data, right: a more complex relationship fitted with Random Forest. Prediction beyond the training data becomes highly unreliable, although prediction interval width outside the data range is constant. Random Forest prediction intervals were obtained by computing quantiles over the predictions from individual trees.

3 Method of the DI and AOA estimation

3.1 Create Sample data, scale and weight them

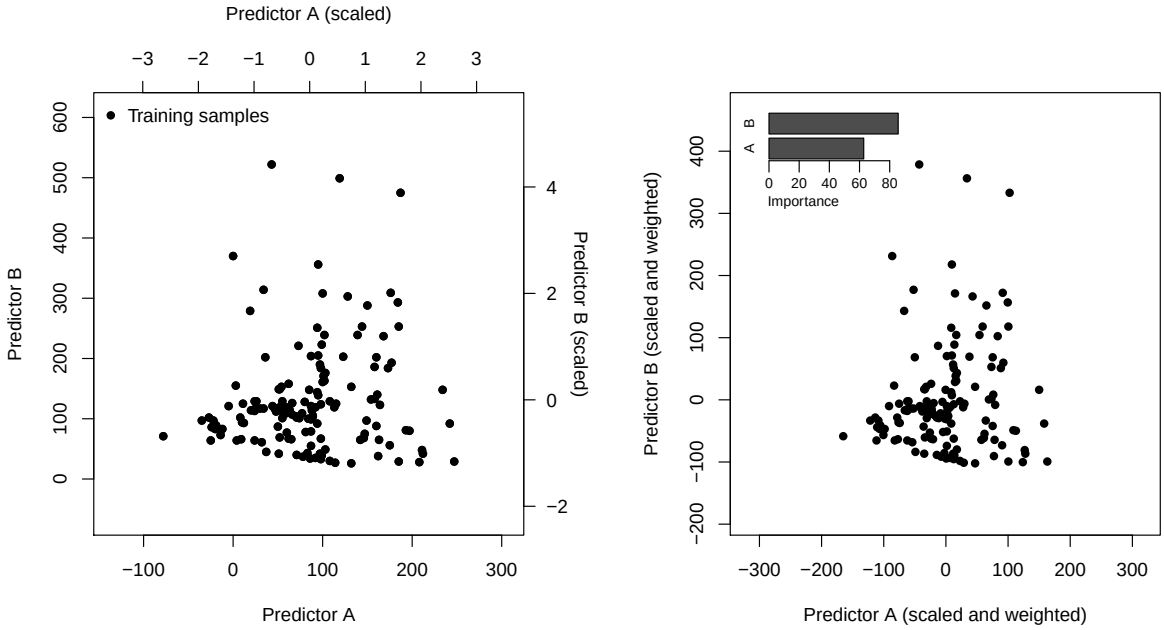


Figure 2: Initial situation: Training samples in a multidimensional (here 2) predictor space (a). First, the predictor space is scaled (second x and y axis in the left plot) and then weighted (plot right) according to the estimated variable importance shown in the topleft corner.

3.2 Calculate the DI within the training data to get the threshold for the AOA

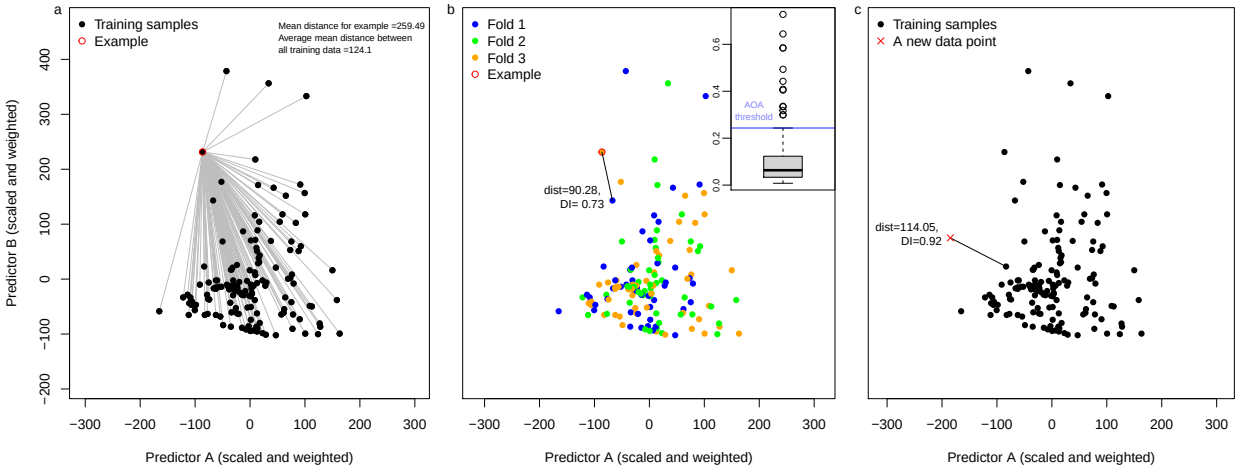


Figure 3: Training samples in a multi-dimensional (here 2-dimensional) predictor space that has been scaled and weighted. First, the average of the mean distances between all training data is calculated (a). Next, the DI of the training data is calculated. For each training data point (shown here for one example), the distance to the nearest training data point not located in the same cross-validation fold is calculated (here visualized assuming a 3-fold cross-validation)(b). This distance is divided by the average of the mean distances between all training data (a) to derive the DI. The DI is calculated for each training data point (boxplot in b) and the threshold for the AOA is then derived from the upper whisker of the DI values. For a new data point, the DI is calculated accordingly (c). In this example, the DI is larger than the DI-threshold, indicating that this new data point falls outside the AOA.

3.3 Estimate the AOA for each new potential data point

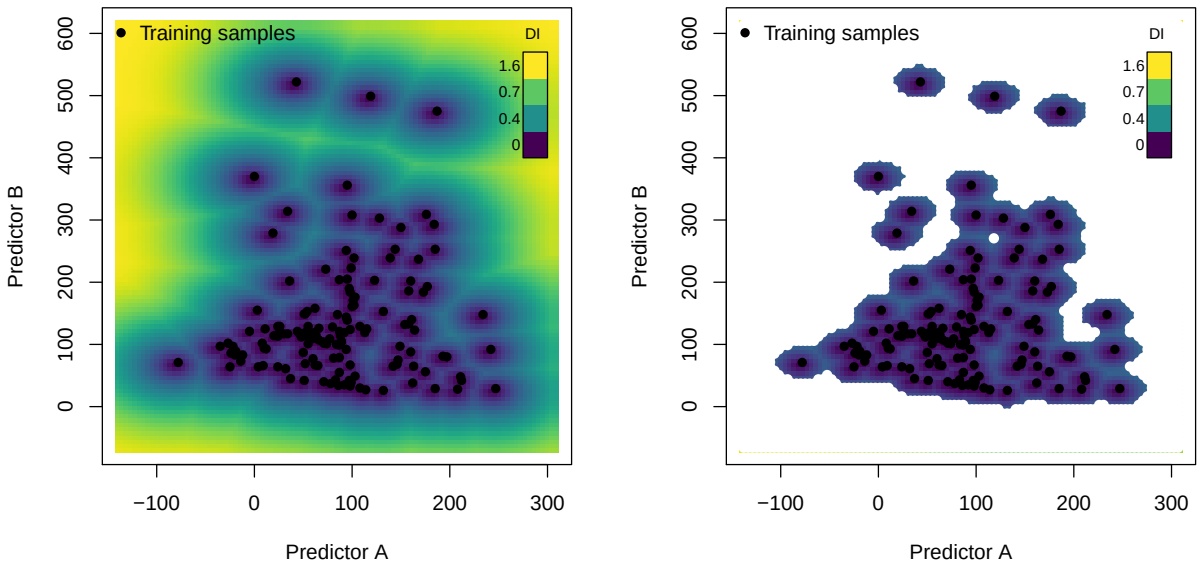


Figure 4: The AOA threshold can be applied to the predictor space of the entire area of interest to derive the AOA (b) from the DI (a) of each new data point which is to be predicted. Areas shown in white in b are outside the AOA.

3.4 Using the DI to quantitatively express prediction uncertainty

3.4.1 Estimating the relationship using the cross-validated data

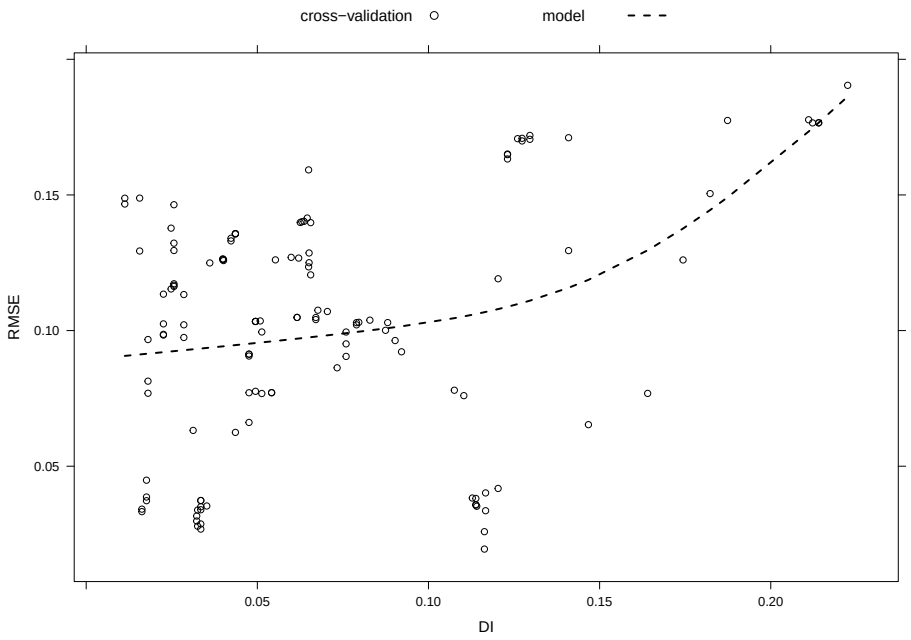


Figure 5: Relationship between the DI and RMSE within the AOA based on single cross-validation. RMSE was calculated in a sliding window (here: size of 10) along the DI. To estimate the RMSE based on the DI, a shape constrained additive model was used. Note that in this small example there is no strong relationship between DI and RMSE. Please see the paper for a more meaningful example.

3.4.2 Using Multi-purpose CV

The results shown above are based on a single-purpose CV. We can also estimate the relationship between performance and DI using multiple CVs. Here, we split the data several times into folds using clusters in the predictor space with the number of clusters ranging between 3 and the number of data points (=LOOCV). The following figure shows how data splitting can look like for different numbers of k.

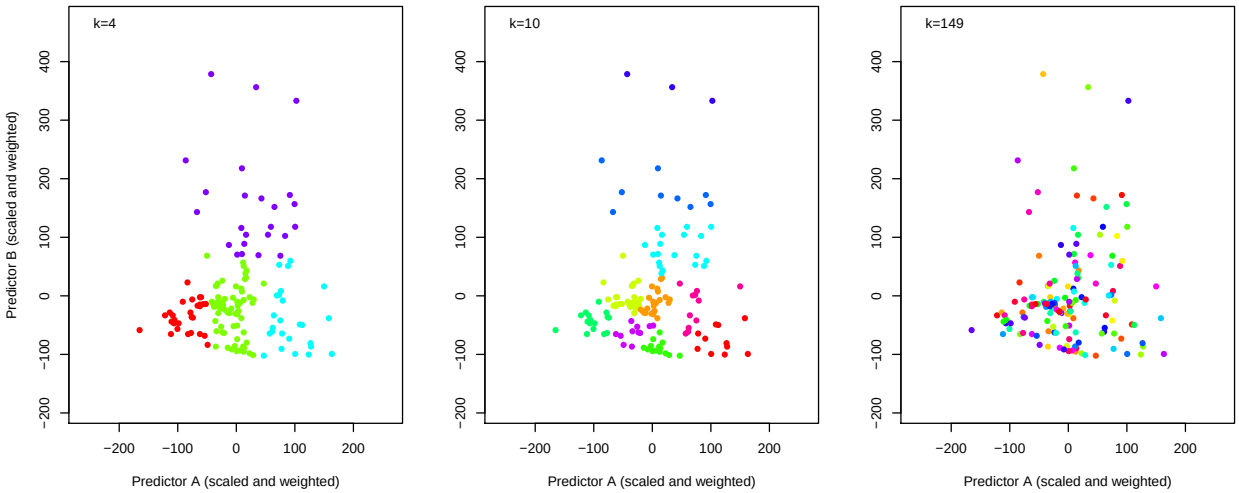


Figure 6: Example of multi-purpose cross-validation folds where data are split into folds (colors) based on clusters in the predictor space. This is done for different number of clusters, here shown for k=4, k=10, and k=149.

Using multi-purpose CVs allows expanding the AOA and also allows giving a more robust estimate for the relationship. The relationship can be used to estimate the DI-dependent prediction performance for each new data point based on its DI.

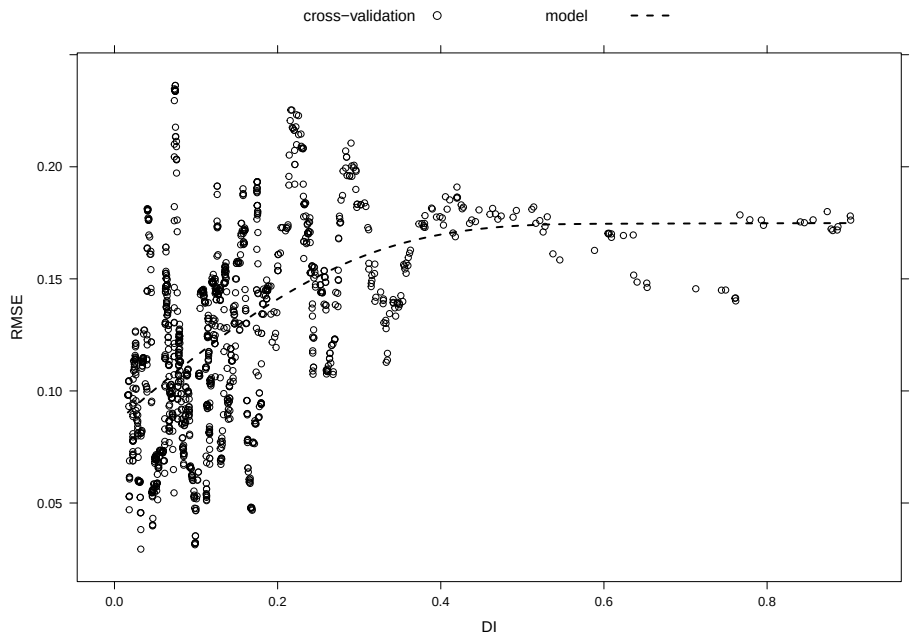
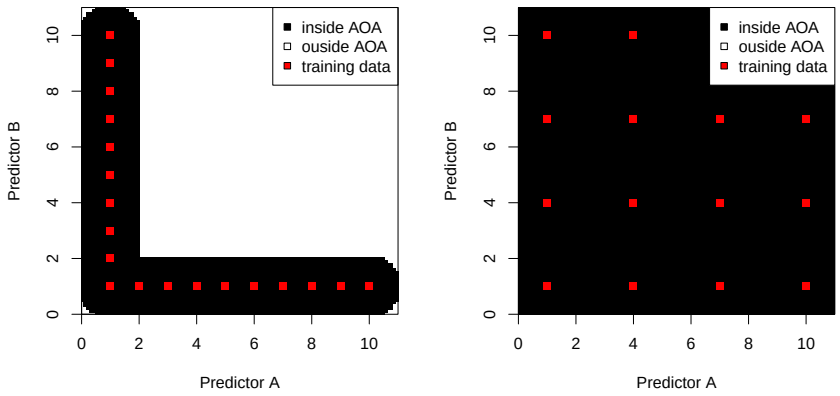


Figure 7: Relationship between the DI and RMSE within the AOA based on multi-purpose cross-validation. RMSE was calculated in a sliding window (here size of 25. Due to multiple cross-validations more data points are available which allows for a larger window size) along the DI. To estimate the RMSE based on the DI, a shape constrained additive model was used. Note that in this small example there is no strong relationship between DI and RMSE. Please see the paper for a more meaningful example.

4 FAQ explained in figures

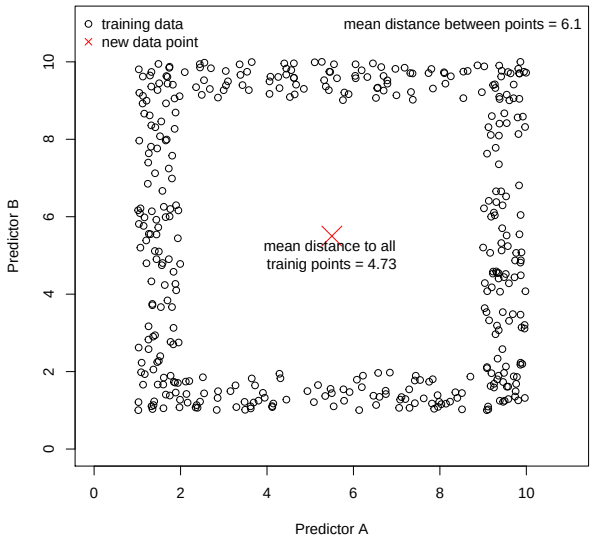
4.1 What happens if the full range of predictors is covered by training data?

When the full range of environmental conditions is covered the result is that the model can be applied to the entire study area. However, covering the full range of each variable is not enough (figure left) but the combined ranges need to be covered if the model should be applicable to the entire study area (figure right).



4.2 Why can't we use the distance to all data points, or hulls, instead of distance to the nearest neighbor?

The distance to all training points is not helpful here because gaps could not be considered. But gaps are equally problematic (see our first figure in the manuscript) and using the average distance to training points cannot account for that. In the example below the new data point even has a lower average distance to training data although it is clearly in an area not covered by training data. The problem is similar for hulls, that do not take gaps in the predictor space into account.


training data
new data point
mean distance between points = 6.1
mean distance to all
trainig points = 4.73
Predictor B
Predictor A
0
2
4
6
8
10

4.3 How would the AOA look like for Fig. 1 of the manuscript?

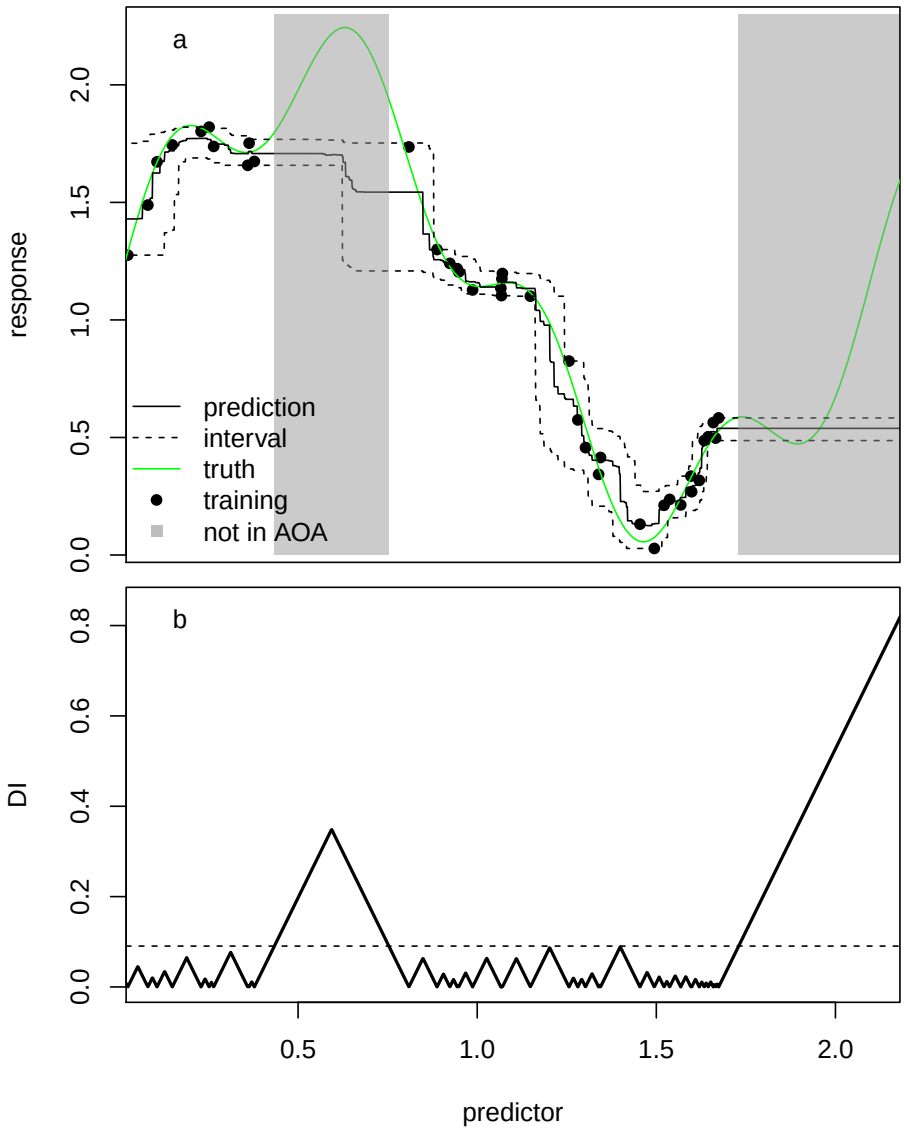


Figure 8: Hypothetical example of a relationship between a virtual predictor and response variable as well as the predictions made by Random Forest already presented in Figure 1. Low values in the dissimilarity index (DI) do not necessarily mean that the prediction error is high. As moving away from the last training data point (x= 1.7), the value of the DI increases (b). However, the error does not necessarily increases in the same way (comparing the predictions with the truth in a). The uncertainty must still be considered as very high because this area of the predictor space is unknown to the model. The area of applicability (AOA) is derived using the outlier-removed maximum DI observed in the cross-validated training data as a threshold (dashed line in b) and is used to exclude predictions in areas where dissimilarity is too high (grey area in a).