

Supplementary analysis 1: KDE-based Applicability Domain Analysis

Methods

A formal applicability domain (AD) analysis was performed to analyze the molecular descriptor space covered by the training compounds used to develop the final BBB permeability classifier. The analysis follows the rationale of the density-based framework proposed by Schultz et al. [1], in which kernel density estimation (KDE) is used to estimate the density of training samples in feature space and to derive a dissimilarity measure for new predictions.

The AD was defined in the effective feature space of the final selected model. Specifically, the analysis used the ten descriptors retained by the final Random Forest model, composed of seven molecular descriptors and three MACCS fingerprint.

KDE was fitted using an Epanechnikov kernel. The bandwidth was selected by five-fold grid search on the training compounds. Let $KDE(x)$ denote the estimated density for a compound x in the transformed model feature space, and let $\max KDE(X_{train})$ denote the maximum KDE value observed among training compounds. The corresponding AD dissimilarity score (d) was then defined as:

$$d(x) = 1 - \frac{KDE(x)}{\max(KDE(X_{train}))}$$

Thus, d values close to 0 identify compounds located in densely sampled regions of the training feature space, whereas d values close to 1 identify compounds located in sparse or poorly represented regions. The AD threshold was set to the 95th percentile of the training-set dissimilarity distribution: compounds with d values less than or equal to this threshold were classified as in-domain, while compounds exceeding the threshold were classified as out-of-domain.

The final model was then applied to the external test set, and predictive performance was evaluated both on the full test set and after stratification according to the KDE-based AD label. Balanced accuracy, F1 score, precision, sensitivity, specificity, and confusion-matrix counts were reported.

Results

The KDE-based AD model was fitted on 8,159 training compounds in the transformed ten-feature model space, through Epanechnikov kernel the bandwidth selected is equal to 7.278. The 95th percentile of the training dissimilarity distribution was 0.560 and was used as the AD threshold. By construction, 7,752 training compounds (95.01%) were classified as in-domain and 407 compounds (4.99%) as out-of-domain (Figure 1a).

When the same AD model and threshold were applied to the 27 external test compounds, 26 compounds (96.30%) were classified as in-domain and one compound (3.70%) was classified as out-of-domain. The distribution of KDE-based dissimilarity scores showed that most external compounds overlapped with the dense region of the training distribution, while one compound displayed a markedly higher dissimilarity score above the AD threshold.

On the full external test set, the final Random Forest model reached a balanced accuracy of 0.773, F1 score of 0.720, sensitivity of 0.900, and specificity of 0.647. When the evaluation was restricted to the 26 compounds classified as in-domain, balanced accuracy increased to 0.823 and sensitivity increased to 1, while specificity remained 0.647. The in-domain subset therefore contained no false negatives, although false positives remained present. This indicates that AD filtering identified a less reliable prediction but should not be interpreted as guaranteeing correct classification for all in-domain compounds.

Table 1. External test-set performance before and after KDE-based AD stratification.

Subset	n	%	BA	F1	Prec.	Sens.	Spec.	TN	FP	FN	TP
All	27	100.00	0.7735	0.7200	0.6000	0.9000	0.6471	11	6	1	9
In-domain	26	96.30	0.8235	0.7500	0.6000	1.0000	0.6471	11	6	0	9
Out-of-domain	1	3.70	-	-	-	-	-	0	0	1	0

Abbreviations: BA, balanced accuracy; Prec., precision; Sens., sensitivity; Spec., specificity.

The only out-of-domain external compound was magnesium citrate, which had a dissimilarity score of 0.861, exceeding the AD threshold. Magnesium citrate was labelled as BBB permeable in the external test set but was predicted as non-permeable by the model. Thus, the only out-of-domain compound corresponded to the only false negative in the external test set.

For visualization only, the transformed ten-feature model space was projected onto the first two principal components (Figure 1b). The PCA projection showed that most external test compounds were embedded within or close to regions populated by training compounds, while the single out-of-domain compound was located in a more peripheral region.

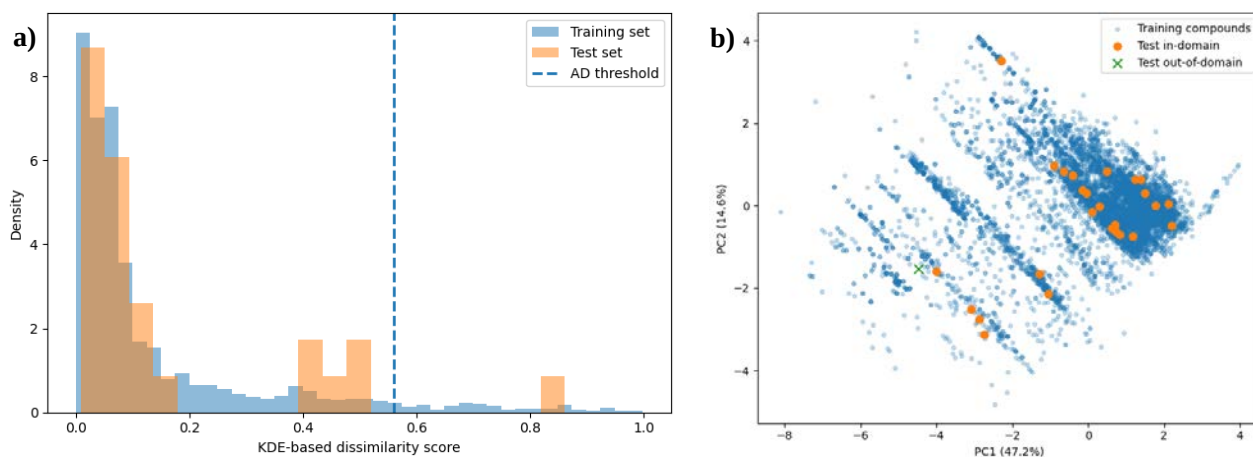


Figure 1. Results of KDE-based Applicability Domain Analysis. a) KDE-based dissimilarity score distribution for the training and external test sets. The dashed vertical line indicates the AD threshold defined as the 95th percentile of the training dissimilarity distribution ($d = 0.560$). b) PCA visualization of the final model feature space. Training compounds are shown in blue, in-domain external test compounds in orange, and the out-of-domain compound in green.

Overall, the KDE-based AD analysis indicates that the external test set largely falls within the applicability domain of the final model. The improvement in balanced accuracy from 0.773 on the full external set to 0.823 on the in-domain subset suggests that predictions within the KDE-defined AD are associated with higher reliability.

References

- Schultz, L.E., Wang, Y., Jacobs, R. et al. A general approach for determining applicability domain of machine learning models. *npj Comput Mater* 11, 95 (2025). <https://doi.org/10.1038/s41524-025-01573-x>