Supporting information for: Enhancing Predictive Modeling with Molecular Fingerprint Fusion Strategies

Turkina Viktoriia[1*], Messih Melanie R.W.[1], Kant Etienne[1], Gringhuis Jelle T.[1], Petrignani Annemieke[1], Corthals Garry[1], O'Brien Jake W.[1,2], Samanipour Saer[1,2,3*]

[1*]Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam, 1090 GD, The Netherlands.
[2]Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, 20 Cornwall Street, Woolloongabba, QLD 4102, Australia.
[3]UvA Data Science Center, University of Amsterdam, Amsterdam, The Netherlands.

*Corresponding author(s). E-mail(s): v.turkina@uva.nl; s.samanipour@uva.nl;

Contents

List of Figures

List of Tables

S1 Selection of FPs

Table S1: Description of selected non-hashed molecular fingerprints employed in this study.

Fingerprint	Abbreviation	Software	Length	Description
Atom Pair 2D Count	AP2DC	PaDEL	780	The count vector of the atom pairs containing specific elements (C, N, O, Cl, I, Br, F, P, S, Si, B, and X representing all halogens) within a maximum distance of 10.
E-state fingerprint	E-state	RDKit	79	The count vector encoding electrotopological state (E-state) indices, which incorporate details regarding functional groups, graph topology, and the electronegativity of each atom according to the Kier-Hall method.
Klekotha-Roth Count	KRC	PaDEL	4860	The count vector of substructures related to certain bioactivity.
Molecular Access Systems	MACCS	PaDEL	166	The binary string representing the presence/absence of certain atom types, bond types, atom environments, groups, and properties.
PubChem fingerprint	PubChem	PaDEL	881	The binary string, covering a wide range of different substructures and features divided into several sections: hierarchic element counts, rings, simple atom pairs, simple atom nearest neighbors, detailed atom neighborhoods, simple SMART patterns, complex SMART patterns.
Substructure Keys Count	SSC	PaDEL	307	The count vector of SMARTS patterns for functional group classification designed by Christian Laggner.

S2 Preprocessing summary

Table S2: Summary of SMILES removal during fingerprint generation.

Dataset	Number of SMILES	Removed (FP failure)	Retained	% removed
FishTox	907	0	907	0.00
KOC	824	11	813	1.33
RIs	3018	10	3008	0.33
ApisTox	1035	0	1035	0.00
BBBP	2053	5	2048	0.24
BACE	1547	0	1547	0.00

S3 Evaluation Metrics

The regression tasks were evaluated using the root-mean-square error (RMSE) and coefficient of determination (R^2), while classification tasks were assessed using the area under the receiver operating characteristic curve (ROC-AUC) and the F1-score. The formulas for each metric are as follows:

Root-Mean-Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{S1}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{S2}$$

F1-Score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{S3}$$

where

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{S4}$$

Area Under the ROC Curve (ROC-AUC)

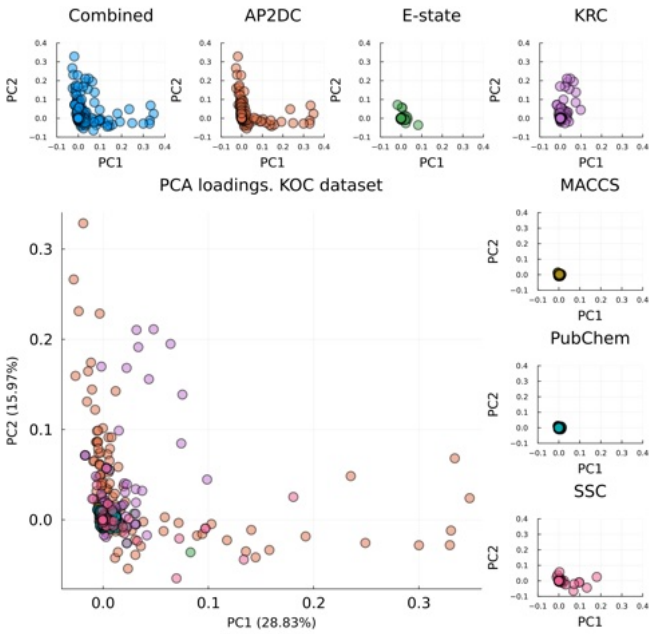
ROC-AUC is computed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, where TPR and FPR are given by:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{S5}$$

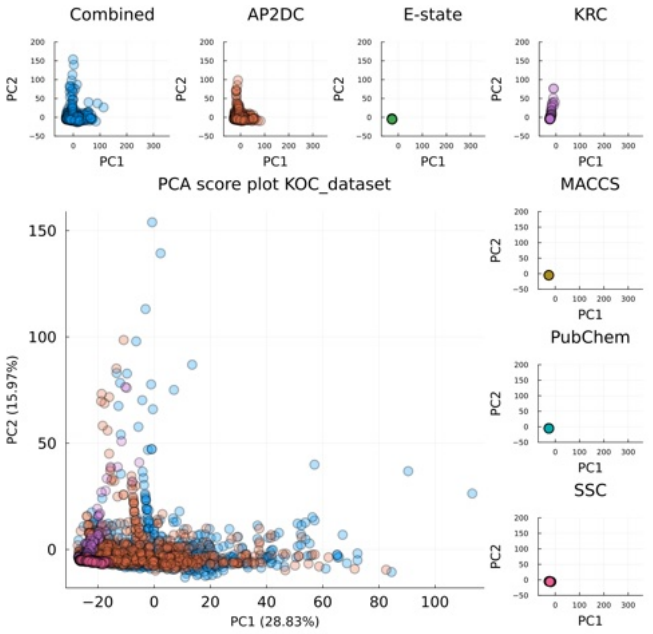
S4 Individual FPs on PCA-space

PCA loadings (a) and scores (b) plots for the KOC, RIs, ApisTox, BACE, and BBBP datasets using each of the six individual fingerprints and their horizontal concatenation. PCA was performed on the combined fingerprint set. In the loadings plots (a),

the contributions from each individual fingerprint are shown separately to illustrate their structural coverage. In the scores plots (b), the projection of the dataset using each individual fingerprint is visualized in the PCA space derived from the combined fingerprints (Figure S1-S5)

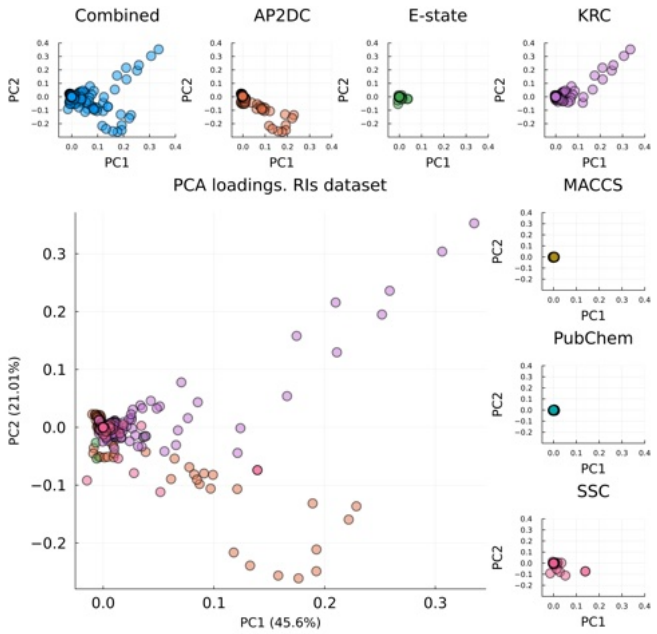


(a)

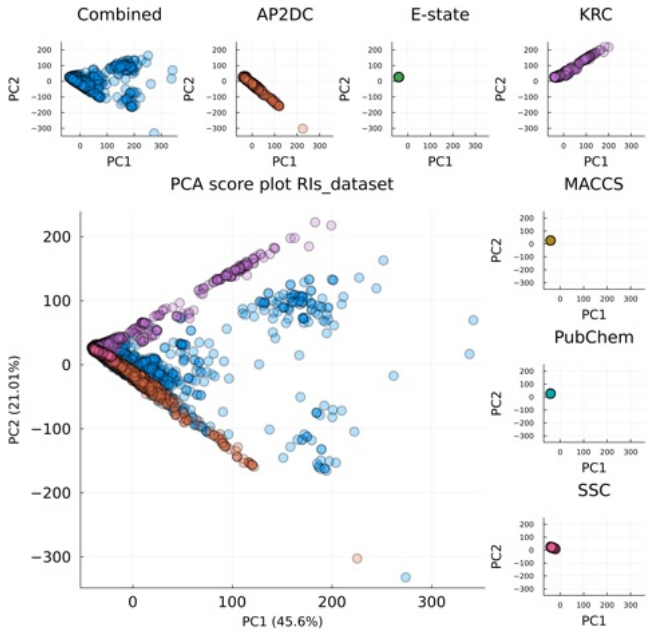


(b)

Fig. S1: PCA loadings (a) and scores (b) plots for the KOC dataset using each of the six individual fingerprints and their horizontal concatenation.

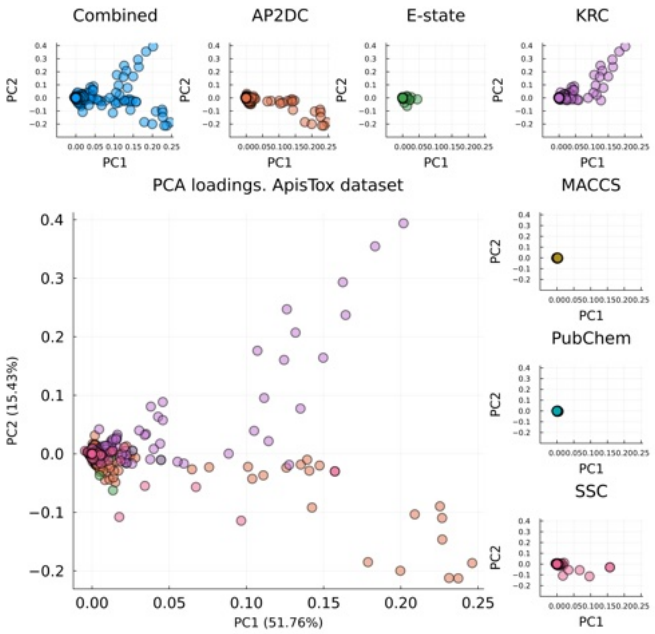


(a)

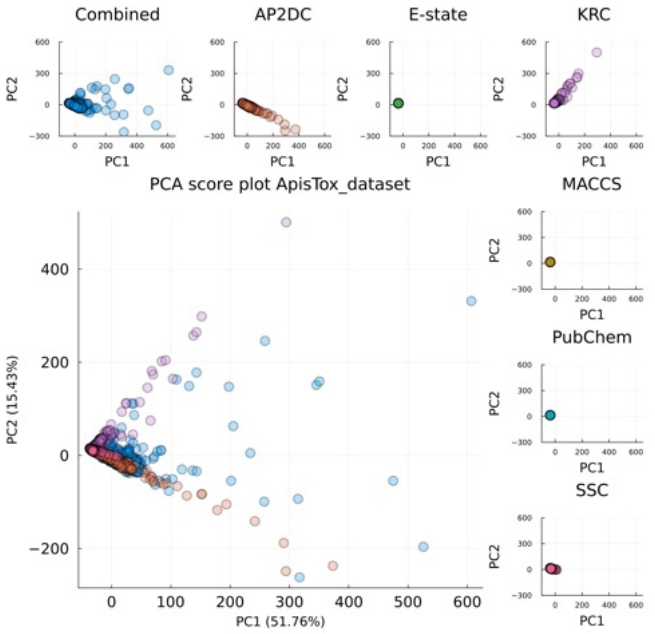


(b)

Fig. S2: PCA loadings (a) and scores (b) plots for the RIs dataset using each of the six individual fingerprints and their horizontal concatenation.

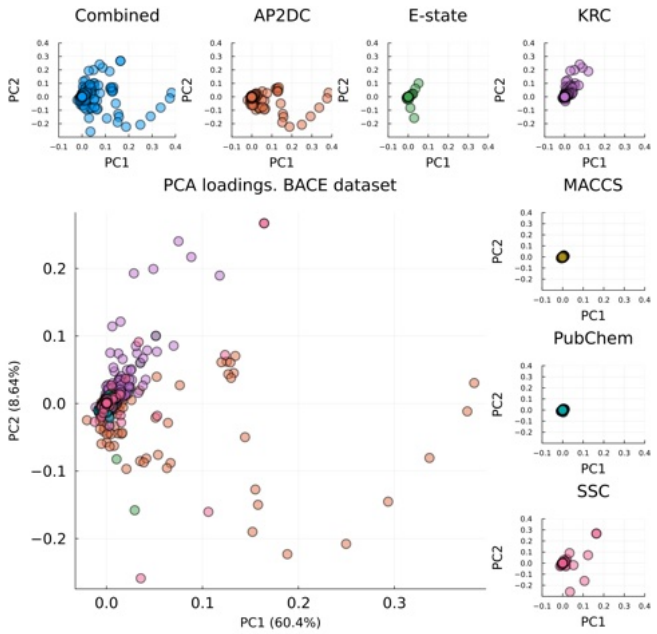


(a)

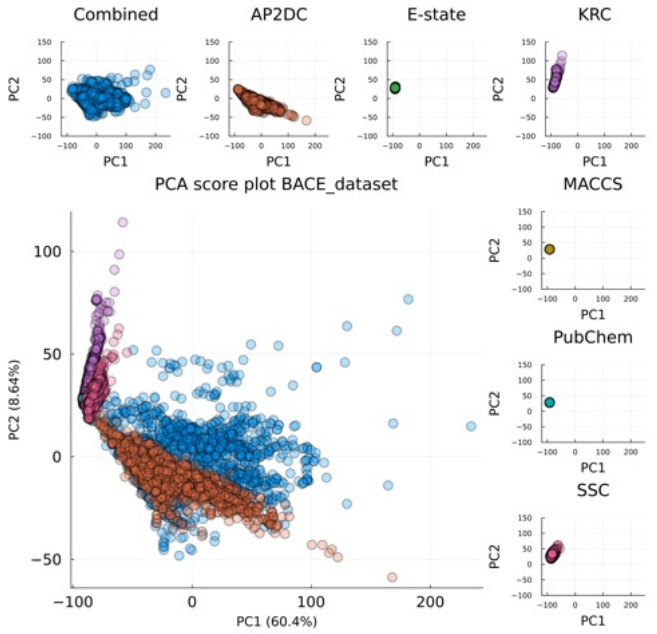


(b)

Fig. S3: PCA loadings (a) and scores (b) plots for the ApisTox dataset using each of the six individual fingerprints and their horizontal concatenation.

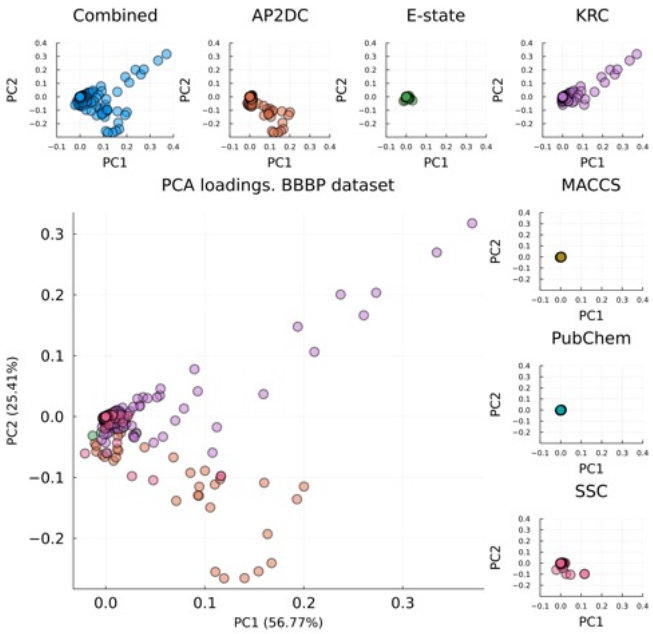


(a)

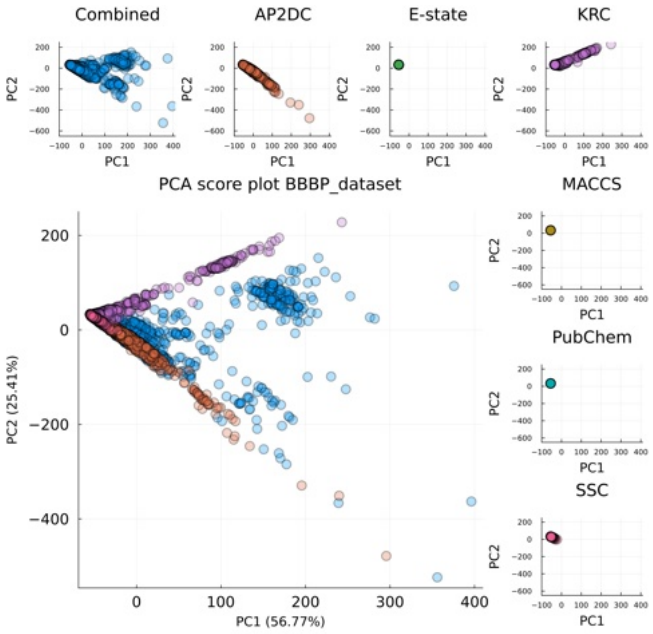


(b)

Fig. S4: PCA loadings (a) and scores (b) plots for the BACE dataset using each of the six individual fingerprints and their horizontal concatenation.



(a)



(b)

Fig. S5: PCA loadings (a) and scores (b) plots for the BBBP dataset using each of the six individual fingerprints and their horizontal concatenation.

S5 Fingerprints fusion performance

	FishTox	KOC	RIs	ApisTox	BACE	BBBP
Common bits	0.587	0.811	0.785	0.696	0.869	0.939
Optimized 95	0.609	0.806	0.783	0.809	0.869	0.942

Table S3: Comparison of test set performance between models trained on optimized 95 FP (95% cumulative importance per dataset) and models trained using only the common features of optimized 95 FP shared across all six datasets.