Evaluating Molecular Representations for Predicting Cyclodextrin-PFAS Binding Energy with Machine Learning: Domain Transfer and Data Limitations

Cole Brzakala[1], Othonas A. Moultos[2], Jan Peter van der Hoek[1], Riccardo Taormina[1]

[1]Water Management Department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628CN Delft, Netherlands

[2]Engineering Thermodynamics, Process and Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Leeghwaterstraat 39, 2628CB Delft, Netherlands

Supporting Information for Publication

External Test Dataset Compositions

Table S1 and S2 provide the complete list of host-guest pairs used in the external test datasets, showing the diversity of cyclodextrin derivatives and PFAS compounds evaluated.

Table S1: Host-guest pairs in E-PFAS$_{\text{test}}$ (21 complexes).

Host	**Guest**
α-CD	PFMOPrA
β-CD	PFMOPrA
α-CD	PFMOBA
β-CD	PFMOBA
α-CD	PFPrOPrA
β-CD	PFPrOPrA
β-CD	PFDMMOBA
γ-CD	PFDMMOBA
α-CD	PFO2HpA
β-CD	PFO2HpA
γ-CD	PFO2HpA
β-CD	PFO2DA
γ-CD	PFO2DA
α-CD	PFO3DA
β-CD	PFO3DA
γ-CD	PFO3DA
β-CD	PFO3TDA
γ-CD	PFO3TDA
β-CD	PFPA
β-CD	PFHxA
β-CD	PFHpA

Table S2: Host-guest pairs in E-β-CD_{test} (42 complexes).

Host	**Guest**
β-CD	PFOA
β-CD	PFOS
β-CD	HFPO-DA
β-CD	PFPeA
β-CD	PFOA
β-CD	PFOS
β-CD	HFPO-DA
β-CD	PFPeA
BnNH-β-CD	PFOA
BnNH-β-CD	PFOS
BnNH-β-CD	HFPO-DA
BnNH-β-CD	PFPeA
BnNH-β-CD	PFOA
BnNH-β-CD	PFOS
BnNH-β-CD	HFPO-DA
BnNH-β-CD	PFPeA
(3-OH)BnNH-β-CD	PFOA
(3-OH)BnNH-β-CD	PFOS
(3-OH)BnNH-β-CD	HFPO-DA
(3-OH)BnNH-β-CD	PFPeA
(3-OH)BnNH-β-CD	PFOA
(3-OH)BnNH-β-CD	PFOS
(3-OH)BnNH-β-CD	HFPO-DA
(3-OH)BnNH-β-CD	PFPeA
SH-β-CD	PFOA
SH-β-CD	PFOS
SH-β-CD	HFPO-DA
SH-β-CD	PFPeA
SH-β-CD	PFOA
SH-β-CD	PFOS
SH-β-CD	HFPO-DA
SH-β-CD	PFPeA
$(SH)_7$-β-CD	PFOA
$(SH)_7$-β-CD	PFOA
NH_2-β-CD	PFOA
NH_2-β-CD	PFOS
NH_2-β-CD	HFPO-DA
NH_2-β-CD	PFPeA
NH_2-β-CD	PFOA
NH_2-β-CD	PFOS
NH_2-β-CD	HFPO-DA
NH_2-β-CD	PFPeA

Hyperparameter Tuning Configurations

All models were tuned using Bayesian optimization with 100 iterations each. The hyperparameter search spaces are detailed in Table S3. All models restore best weights from the epoch with optimal validation performance. AlignedUMAP visualization parameters were fixed at n_neighbors=15, min_dist=0.15, and metric=euclidean.

Table S3: Hyperparameter search spaces used for model optimization.

Model	**Parameter**	**Type**	**Min**	**Max**
LightGBM	num_leaves	int	10	300
	max_depth	int	10	250
	learning_rate	float	0.0001	0.2
	n_estimators	int	25	2500
	reg_alpha	float	0.0	1.0
	reg_lambda	float	0.0	1.0
	bagging_fraction	float	0.2	1.0
	feature_fraction	float	0.2	1.0
FNN	hidden_dim	int	100	1500
	num_layers	int	2	20
	activation	fixed	ReLU	
	dropout	float	0.0	0.4
	learning_rate	float	1×10^{-4}	1×10^{-2}
	regularization	float	0.0	0.1
ChemBERTa Finetuned	unfreeze_layers	categorical	1, 2, or 3 layers	
	transformer_lr	log-float	1×10^{-8}	1×10^{-5}
	fnn_lr	log-float	1×10^{-8}	1×10^{-3}
	batch_size	categorical	16, 32, 64	
Early Stopping & Learning Rate	early_stopping_patience	int	50	100
	early_stopping_min_delta	float	0.001	0.01
	lr_scheduler_patience	int	15	35
	lr_scheduler_factor	float	0.1	0.8
	lr_scheduler_min_lr	log-float	1×10^{-8}	1×10^{-6}

Additional AlignUMAP Plots

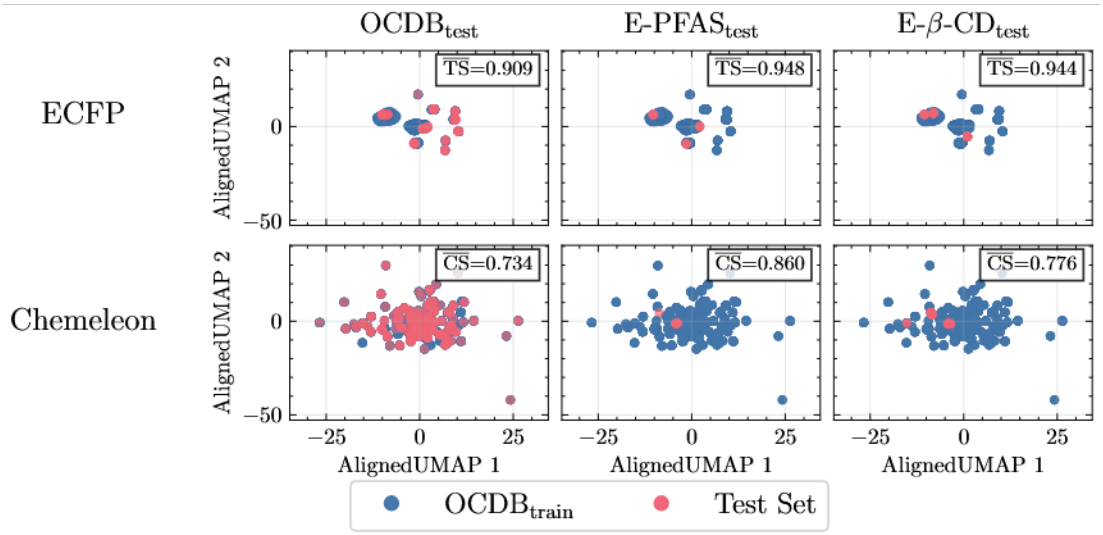

Figure S1: AlignedUMAP visualization of host embeddings across different test sets. The blue points represent the OpenCycloDB training data (OCDB$_{train}$), and the pink points represent the test data for OCDB$_{test}$, E-PFAS$_{test}$, and E-β-CD$_{test}$. All embeddings were jointly reduced into a shared 2D space to enable direct comparison across datasets.

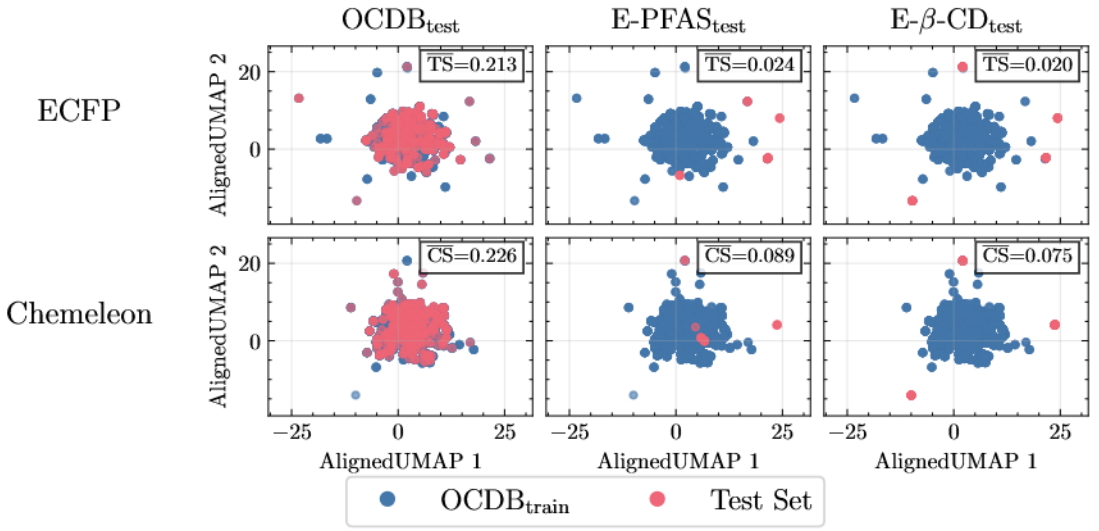

Figure S2: AlignedUMAP visualization of guest embeddings across different test sets. The blue points represent the OpenCycloDB training data (OCDB$_{train}$), and the pink points represent the test data for OCDB$_{test}$, E-PFAS$_{test}$, and E-β-CD$_{test}$. All embeddings were jointly reduced into a shared 2D space to enable direct comparison across datasets.

Additional Neighbor Analysis AlignUMAP Plots

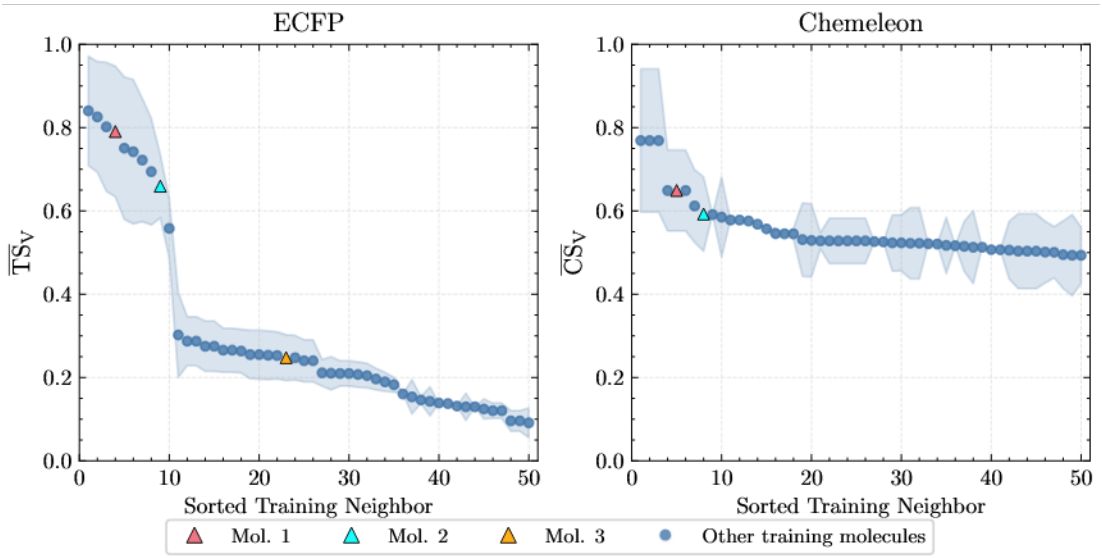

Figure S3: Plots of neighbor analysis for ECFP and Chemeleon guest embeddings. The top 50 nearest neighbors of each test point in the E-PFAS$_{\text{test}}$ and E-β-CD$_{\text{test}}$ datasets were identified based on Euclidean distance in the original embedding space. These neighbors were then visualized to assess how well the embeddings capture chemical similarity relevant to the test points.

UniMol2 Embedding Distribution

Figure S4 shows the distribution of UniMol2 guest embedding features across the $OCDB_{train}$ set. The ranges of the mean and standard deviation of each embedding dimension indicate the variability of learned features used in model training. Notably, there are a few features with significantly larger values than the remaining features. This could lead to these features dominating distance calculations in nearest neighbor analyses, reducing information from similarity assessments. However, these dominant features may also capture important chemical characteristics that differentiate molecules for predicting binding interactions.

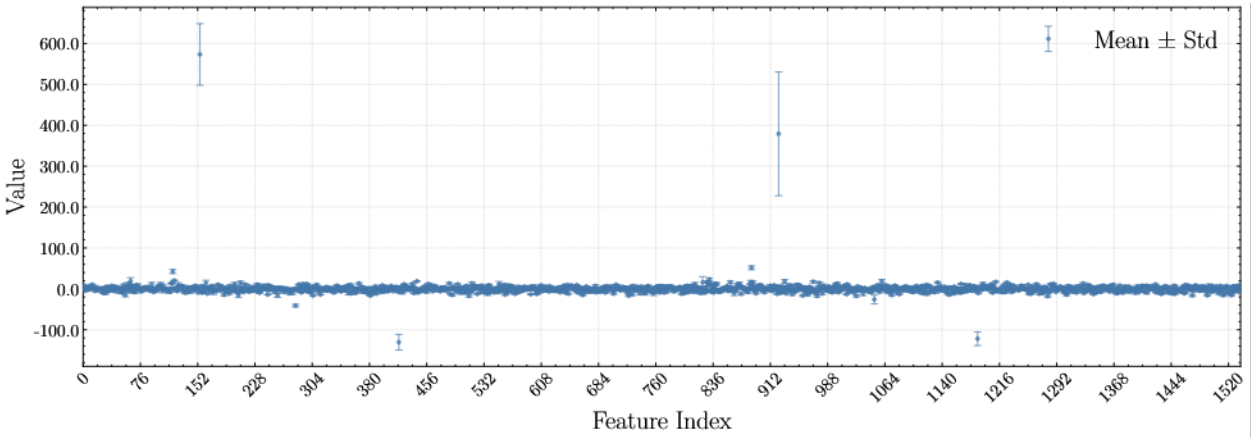

Figure S4: Distribution of UniMol2 guest embedding features across the $OCDB_{train}$ set. Violin plots show the distribution of values for each of the dimensions in the UniMol2 molecular embeddings, revealing the range and variability of learned features used in model training.