Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Electronic Supplementary Information for

SurfPro - A curated database and predictive model of experimental properties of surfactants

Stefan L. Hödl,[a] Luc Hermans,[a] Pim F.J. Dankloff,[a] Aigars Piruska,[a] Wilhelm T.S. Huck[a‡] and William E. Robinson[a‡]

[a] *Physical Organic Chemistry, Radboud University, Heyendaalseweg 135, 6525AJ Nijmegen, The Netherlands*

[‡]*Corresponding authors: W.T.S. Huck (w.huck@science.ru.nl), W. E. Robinson (william.robinson@ru.nl)*

Electronic Supplementary Information

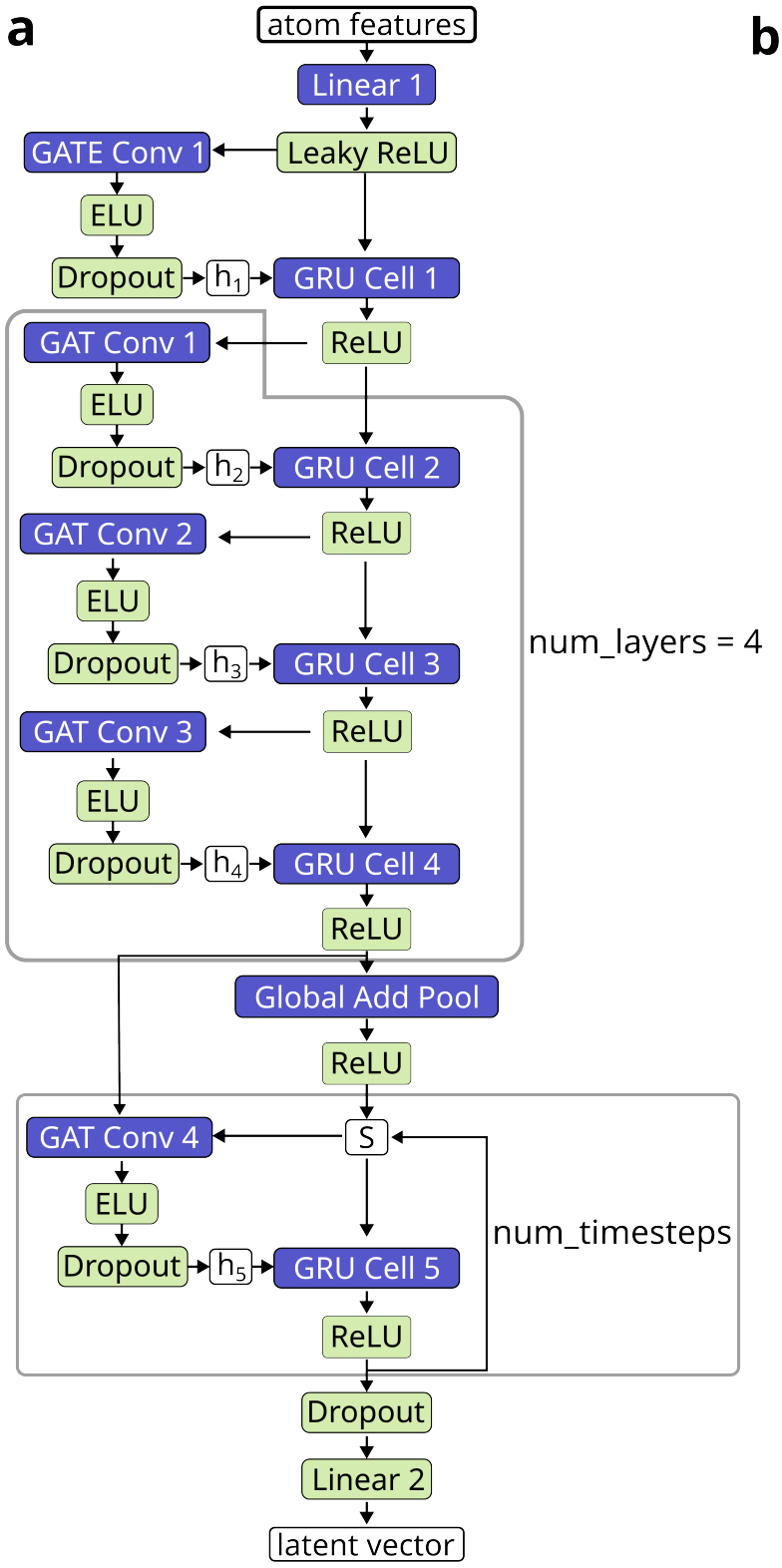
1 Input graph featurization

Atom Feature	Size	Type	Description
Atom Symbol	16	One-hot	[B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, I, At, metal]
Degree	6	One-hot	Number of covalent bonds [0,1,2,3,4,5]
Formal Charge	1	Integer	Electrical charge
Radical Electrons	1	Integer	Number of radical electrons
Hybridization	6	One-hot	[sp, sp^2, sp^3, sp^3d, sp^3d^2, other]
Aromaticity	1	One-hot	Whether the atom is part of an aromatic system [0/1]
Hydrogens	5	One-hot	Number of connected hydrogens [0,1,2,3,4]
Chirality	1	One-hot	Whether the atom is chiral [0/1]
Chirality Type	2	One-hot	[R, S]

Table S 1 Initial atom features used as input to the AttentiveFP encoder, following Xiong *et. al.*. For each atom in the molecular graph, 39 atom features are calculated using RDKit and encoded as a vector, corresponding to the AttentiveFP parameter "in_channels".

Bond Feature	Size	Type	Description
Bond Type	4	One-hot	[single, double, triple, aromatic]
Conjugation	1	One-hot	Whether the bond is conjugated [0/1]
Ring	1	One-hot	Whether the bond is in a ring [0/1]
Stereo	4	One-hot	[StereoNone, StereoAny, StereoZ, StereoE]
Self-loop	1	One-hot	Whether the bond is a self-loop [0/1]

Table S 2 Initial bond features used as input to the AttentiveFP encoder. Compared to Xiong *et. al.*, the self-loop bond feature has been added to capture self-loops. For each bond in the molecular graph, 11 bond features are calculated using RDKit and encoded as a vector, corresponding to the AttentiveFP parameter "edge_dim".



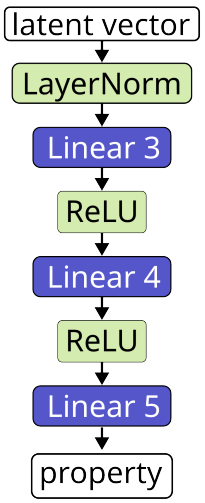


Fig. S 1 Architecture of the graph neural network used in this work. a: The architecture of the AttentiveFP network with num_layers=4 according to the implementation in PyTorch Geometric 2.6.1. Linear: linear layers, ReLU: rectified linear unit, ELU: exponential linear unit, Dropout: dropout layer, GRU Cell: gated recurrent unit cell, GAT Conv: Graph attention unit, Global Add Pool: global addition pooling layer, LayerNorm: layer normalisation layer. Vectors containing data are depicted with white backgrounds. b: Architecture of the regressor layer, which accepts the latent vector produced by the AttentiveFP encoder and produces a vector of property prediction outputs.

2 Average predictive results

	pCMC		γ_{CMC}		$\Gamma_{max} \cdot 10^6$		pC_{20}	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
$\mathrm{AttentiveFP}_{32d}^{single}$	0.326	0.488	3.085	4.364	0.570	1.033	0.514	0.702
$\mathrm{AttentiveFP}_{64d}^{single}$	0.311	0.463	2.864	4.143	0.519	0.937	0.400	0.570
$\mathrm{AttentiveFP}_{96d}^{single}$	0.296	0.442	2.951	4.149	0.485	0.880	0.347	0.496
$\mathrm{AttentiveFP}_{32d}^{multi}$	0.351	0.496	3.301	4.282	0.510	0.963	–	–
$\mathrm{AttentiveFP}_{64d}^{multi}$	0.308	0.442	3.159	4.192	0.439	0.777	–	–
$\mathrm{AttentiveFP}_{96d}^{multi}$	0.307	0.442	2.878	3.949	0.412	0.684	–	–
$\mathrm{AttentiveFP}_{32d}^{all}$	0.346	0.495	3.283	4.310	0.550	1.064	0.431	0.619
$\mathrm{AttentiveFP}_{64d}^{all}$	0.316	0.446	2.924	3.959	0.461	0.914	0.345	0.491
$\mathrm{AttentiveFP}_{96d}^{all}$	0.316	0.443	3.058	4.135	0.419	0.792	0.340	0.474
RDKFP – Ridge	0.697	0.934	3.505	4.756	0.475	0.828	0.532	0.793
RDKFP – RF	0.643	0.854	3.015	4.341	0.469	0.803	0.588	0.760
ECFP – Ridge	0.775	1.029	4.037	5.271	0.499	0.903	0.604	0.787
ECFP – RF	0.745	1.010	4.183	5.550	0.477	0.814	0.598	0.763

Table S 3 Average predictive errors for all model variants and properties under investigation. For each model, we reported the "average" prediction errors of all 10 models on the test set. We report the mean absolute error (MAE) and root mean squared error (RMSE) for each property individually, specifically for the pCMC, γ_{CMC}, $\Gamma_{\mathrm{max}} \cdot 10^6$ and pC_{20}. See also Figure S2-S5 for boxplot visualizations of the MAE (top) and RMSE (bottom) for all four properties.

3 Model comparison plots.

3.1 Boxplot visualizations for pCMC

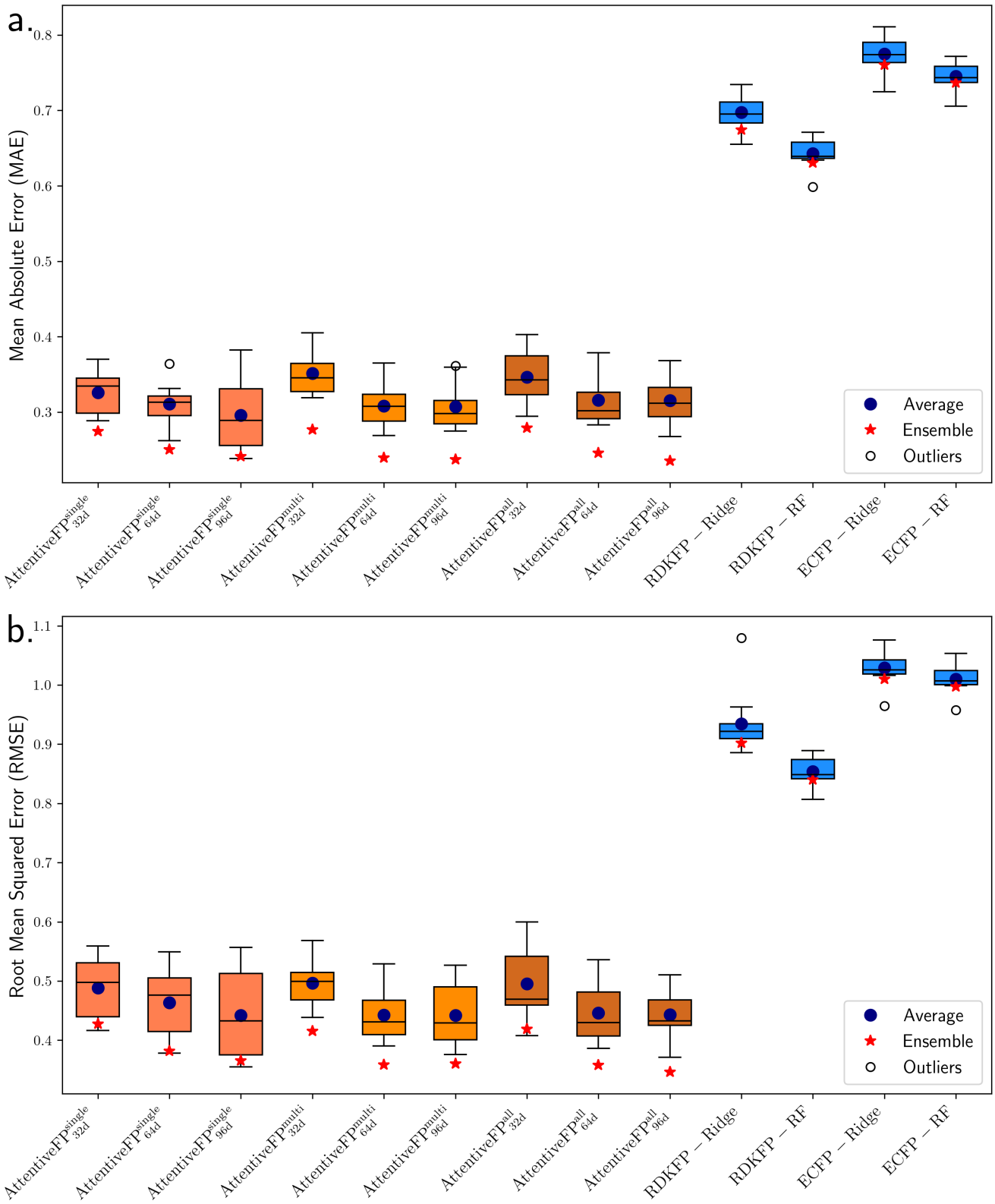


Fig. S 2 Boxplot visualization of the test set MAE prediction error of pCMC for all AttentiveFP model sizes (32d, 64d, 96d), settings (single-, multi-, all-property) and baselines models. For each model, the boxplot shows the 10 test set prediction MAEs (top) or RMSEs (bottom), specifically the first to third quartiles (colored box), its median (black line) and outliers (whiskers & circles). Additionally, the average (dark blue dot) and ensemble (red star) MAE/RMSE are visualized.

3.2 Boxplot visualizations for γ_{CMC}

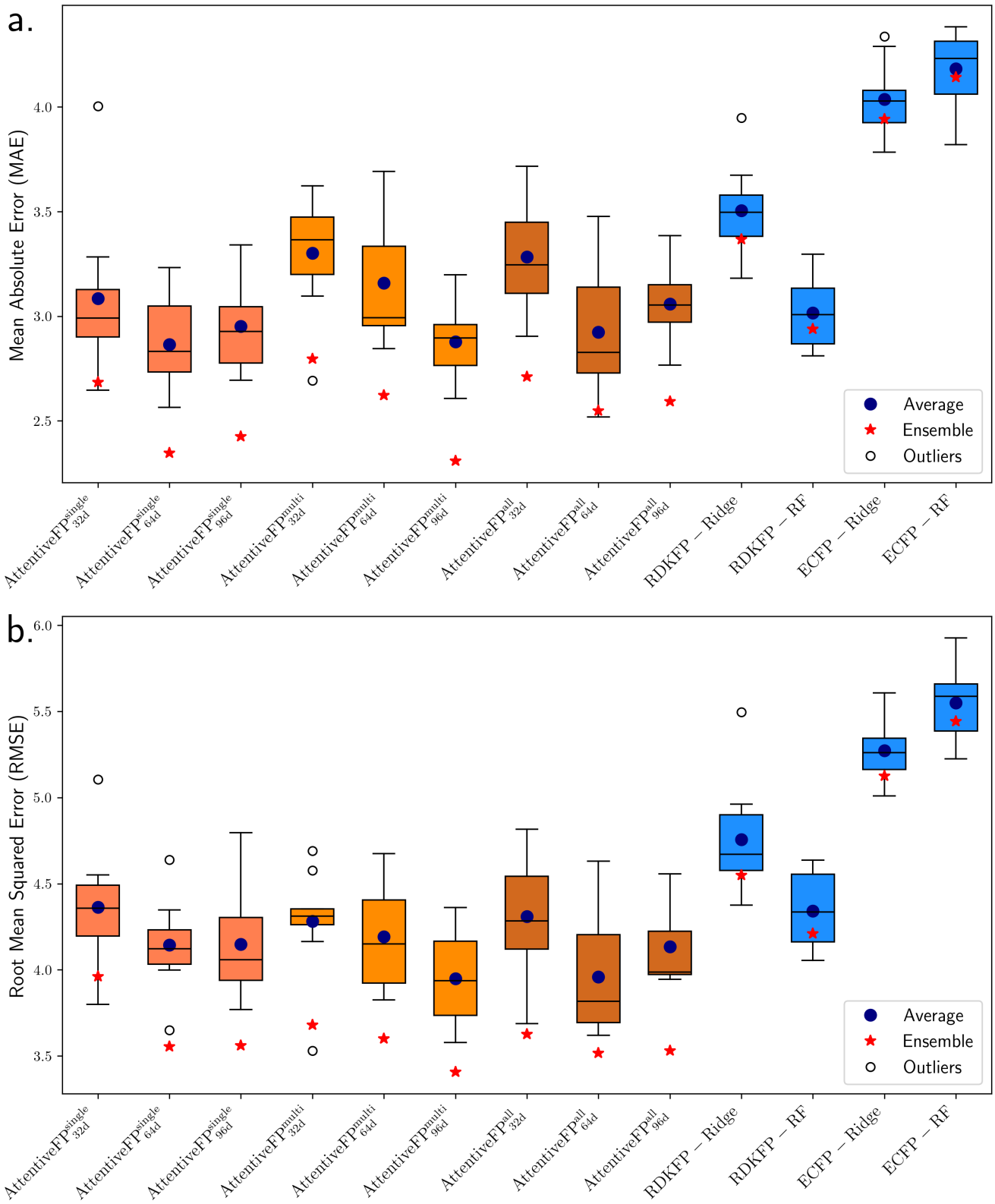


Fig. S 3 Boxplot visualization of the test set MAE prediction error of γ_{CMC} (mN/m) for all AttentiveFP model sizes (32d, 64d, 96d), settings (single-, multi-, all-property) and baselines models. For each model, the boxplot shows the 10 test set prediction MAEs (top) or RMSEs (bottom), specifically the first to third quartiles (colored box), its median (black line) and outliers (whiskers & circles). Additionally, the average (dark blue dot) and ensemble (red star) MAE/RMSE are visualized.

3.3 Boxplot visualizations for Γ_{max}

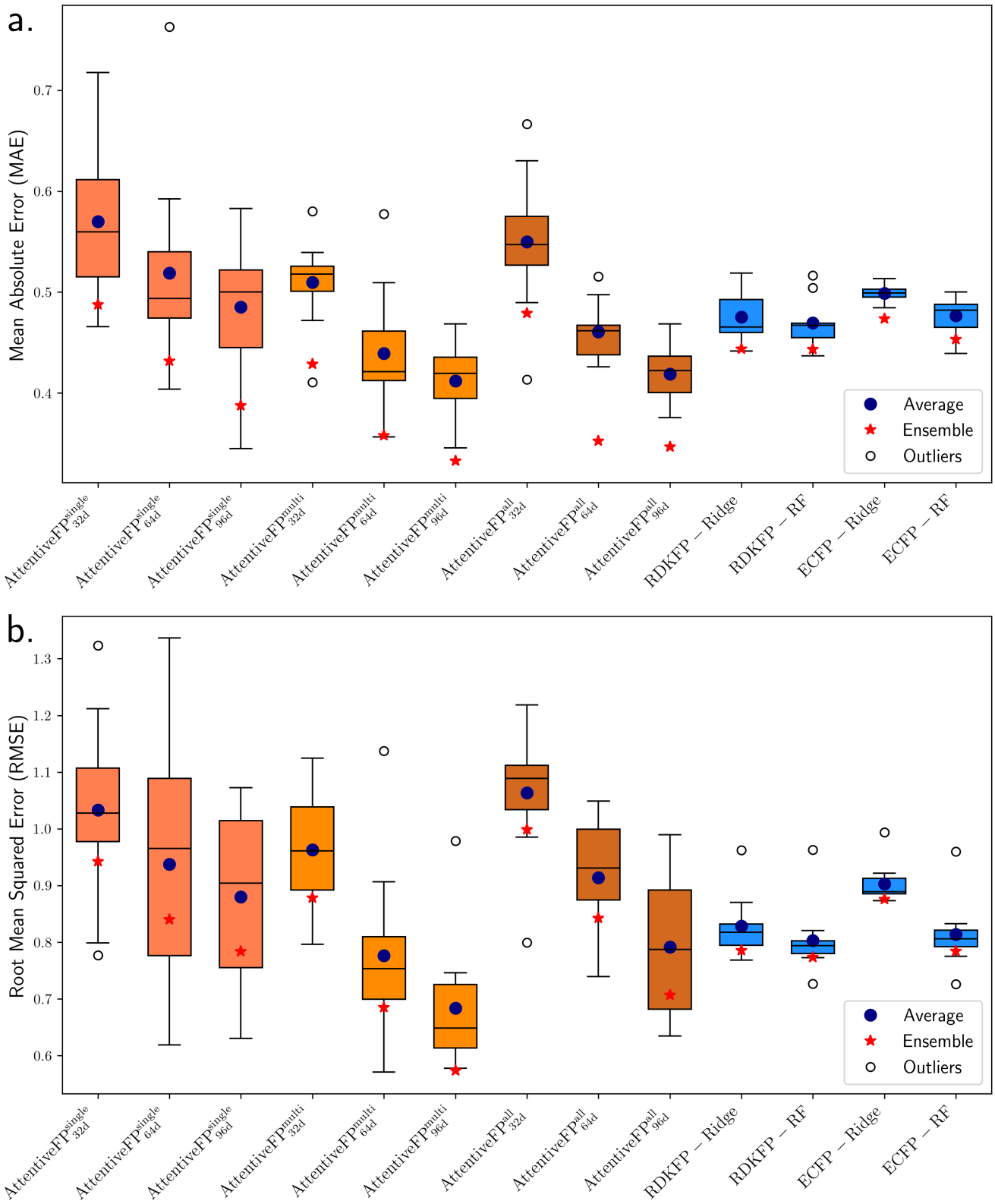


Fig. S 4 Boxplot visualization of the test set MAE prediction error of Γ_{max} ($mol/m^2 \cdot 10^6$) for all AttentiveFP model sizes (32d, 64d, 96d), settings (single-, multi-, all-property) and baselines models. For each model, the boxplot shows the 10 test set prediction MAEs (top) or RMSEs (bottom), specifically the first to third quartiles (colored box), its median (black line) and outliers (whiskers & circles). Additionally, the average (dark blue dot) and ensemble (red star) MAE/RMSE are visualized.

3.4 Boxplot visualizations for pC_{20}

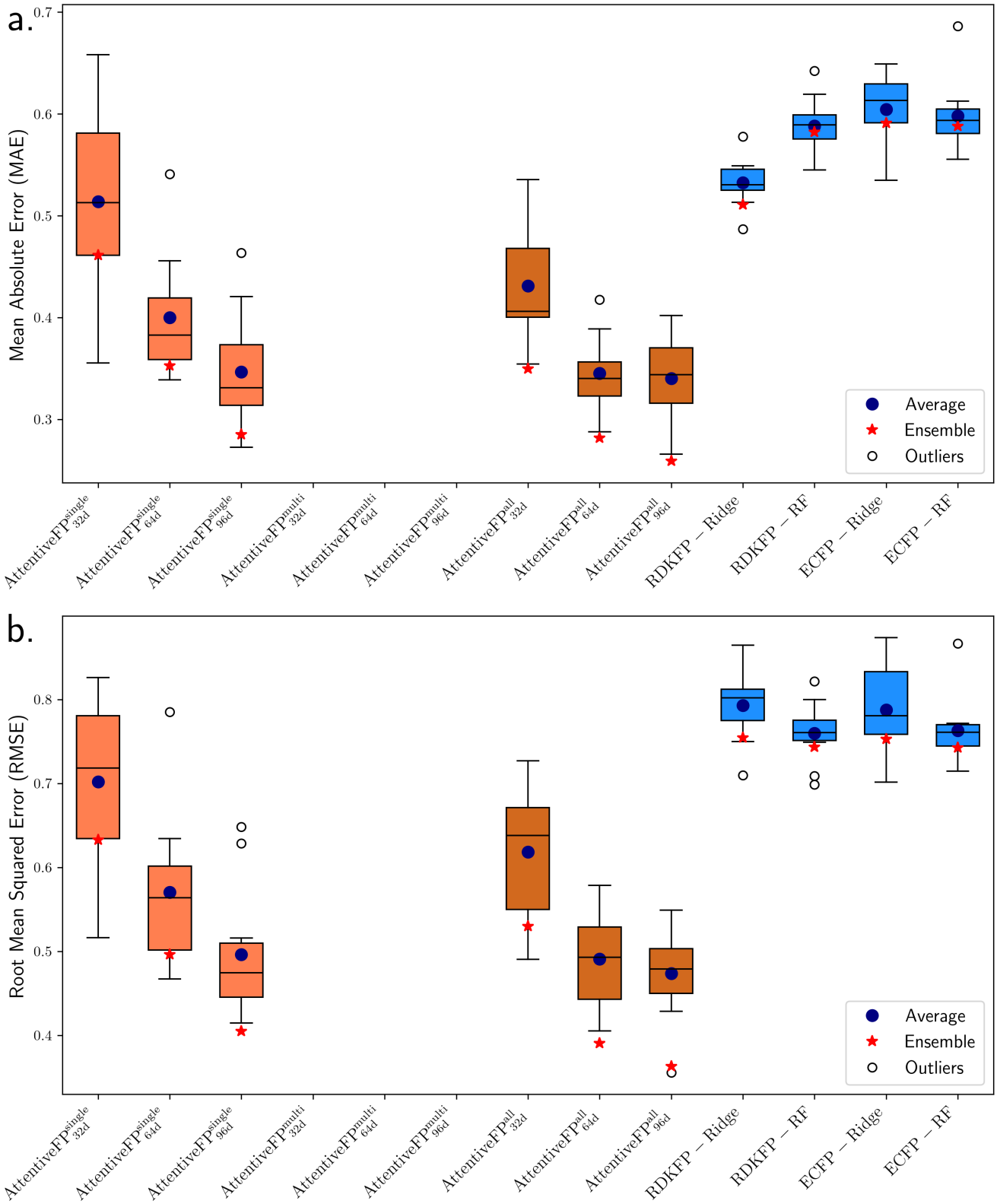


Fig. S 5 Boxplot visualization of the test set MAE prediction error of pC_{20} for all AttentiveFP model sizes (32d, 64d, 96d), settings (single-, multi-, all-property) and baselines models. For each model, the boxplot shows the 10 test set prediction MAEs (top) or RMSEs (bottom), specifically the first to third quartiles (colored box), its median (black line) and outliers (whiskers & circles). Additionally, the average (dark blue dot) and ensemble (red star) MAE/RMSE are visualized. pC_{20} is not directly predicted in the multi-property setting.

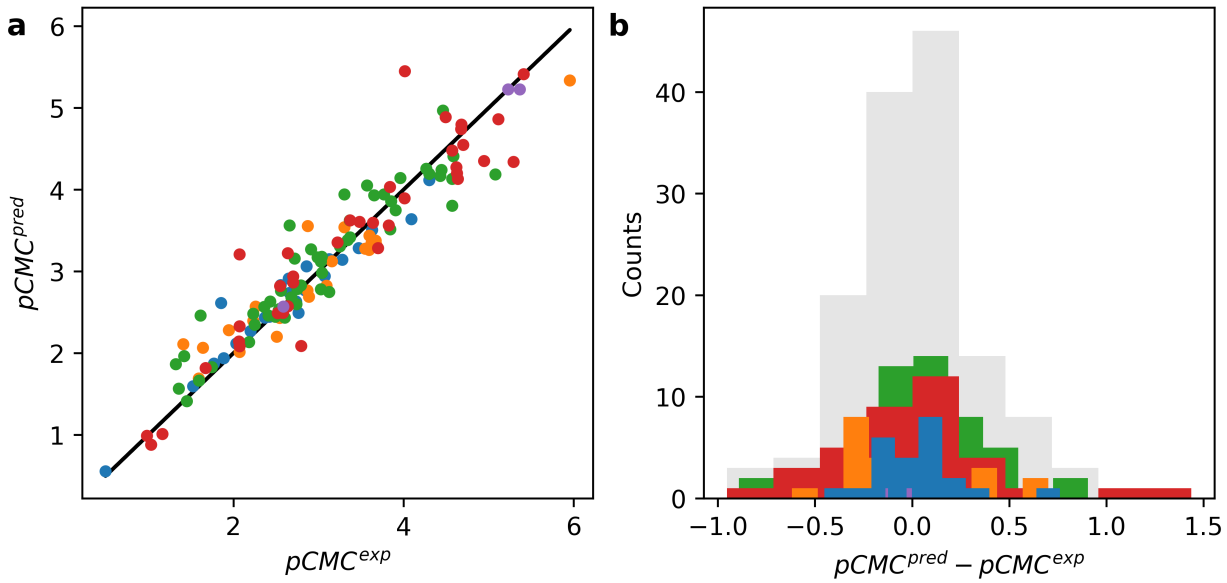


Fig. S 6 a. Parity plots of experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for pCMC. Blue dots: anionic surfactants, orange dots: cationic surfactants, green dots: gemini cationic surfactants, red dots: non-ionic surfactants, purple dots: zwitterionic surfactants. The black line indicates where $\text{pCMC}^{exp} = \text{pCMC}^{pred}$. b. Histograms of the error between experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for *pCMC*. Blue: anionic surfactants, orange: cationic surfactants, green: gemini cationic surfactants, red: non-ionic surfactants, purple: zwitterionic surfactants.

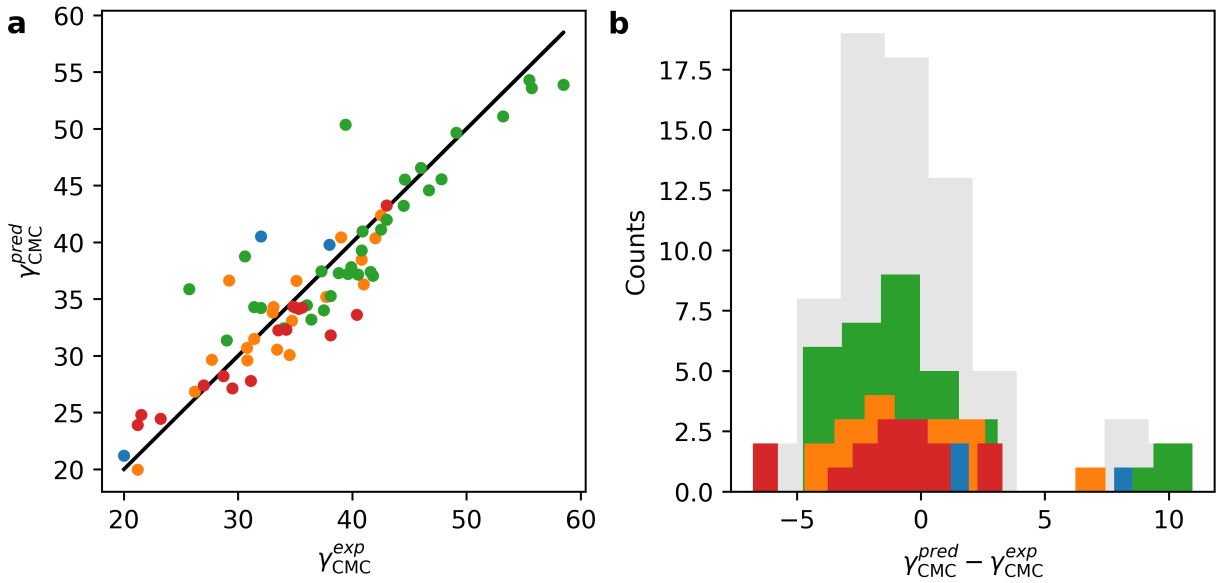


Fig. S 7 a. Parity plots of experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for γ_{CMC}. Blue dots: anionic surfactants, orange dots: cationic surfactants, green dots: gemini cationic surfactants, red dots: non-ionic surfactants. The black line indicates where $\gamma^{exp}_{\text{CMC}} = \gamma^{pred}_{\text{CMC}}$.
b. Histograms of the error between experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for γ_{CMC}. Blue: anionic surfactants, orange: cationic surfactants, green: gemini cationic surfactants, red: non-ionic surfactants.

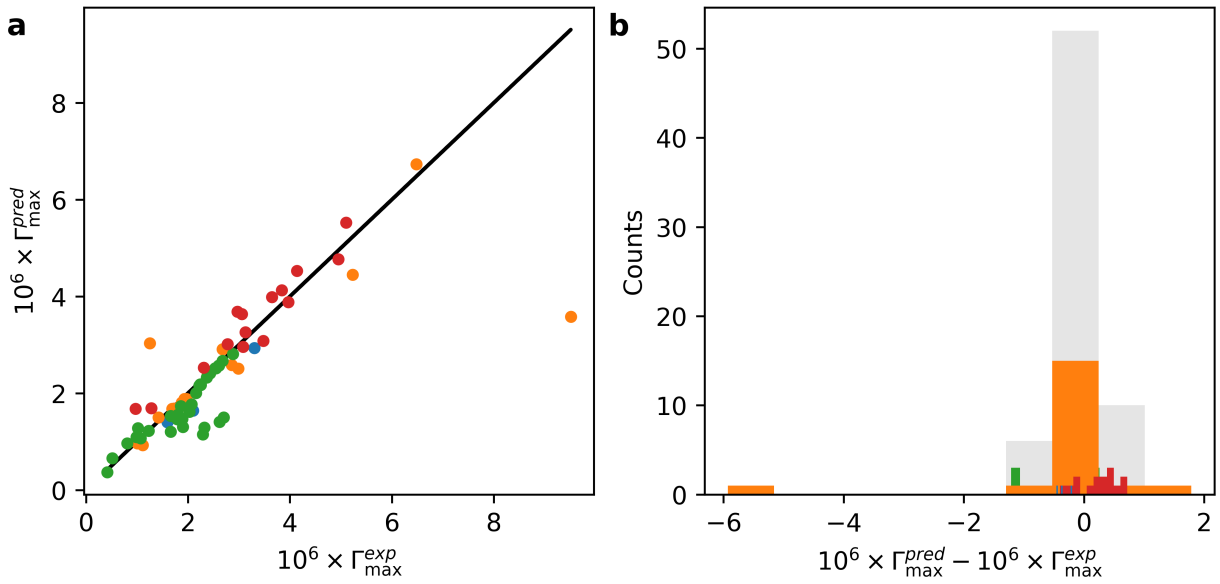


Fig. S 8 a. Parity plots of experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for Γ_{max}. Blue dots: anionic surfactants, orange dots: cationic surfactants, green dots: gemini cationic surfactants, red dots: non-ionic surfactants. The black line indicates where $\Gamma^{exp}_{max} = \Gamma^{pred}_{max}$.
b. Histograms of the error between experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for Γ_{max}. Blue: anionic surfactants, orange: cationic surfactants, green: gemini cationic surfactants, red: non-ionic surfactants.

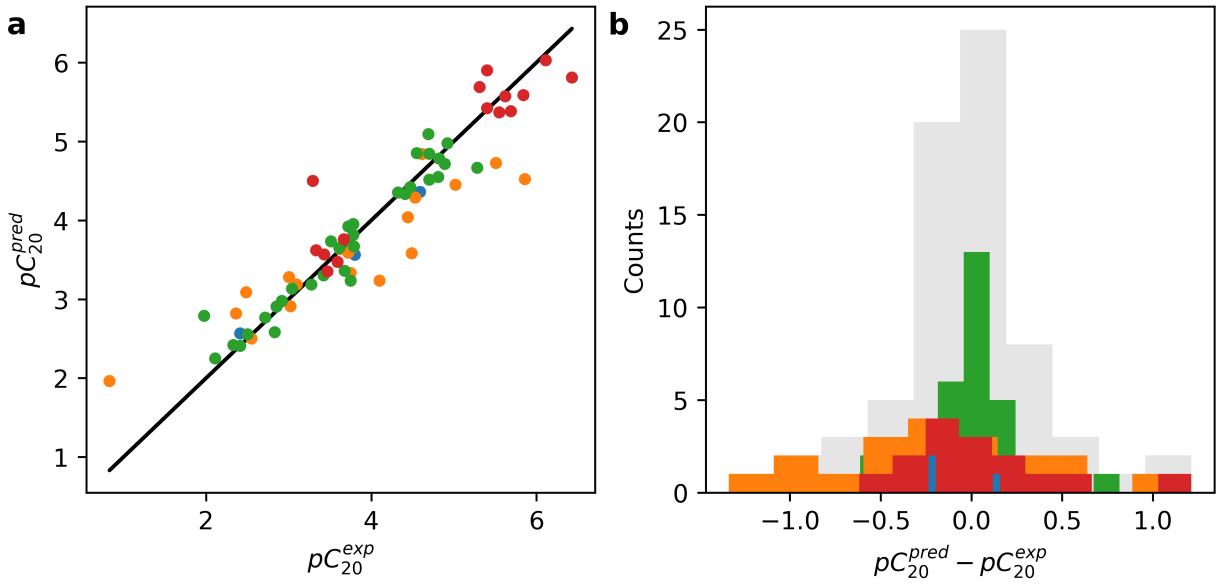


Fig. S 9 a. Parity plots of experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for pC_{20}. Blue dots: anionic surfactants, orange dots: cationic surfactants, green dots: gemini cationic surfactants, red dots: non-ionic surfactants. The black line indicates where $pC^{exp}_{20} = pC^{pred}_{20}$.
b. Histograms of the error between experimental vs. predicted ($\text{AttentiveFP}^{all}_{64d}$) values from the test set for pC_{20}. Blue: anionic surfactants, orange: cationic surfactants, green: gemini cationic surfactants, red: non-ionic surfactants.