Supplement to: Mixture-Based Machine Learning Analysis to Predict Fouling Release Using Insights from Newly Developed Mixture Descriptors

Rahil Ashtari Mahini,[†,‡] Maryam Safaripour,[‡] Achiya Khanam,[‡] Gerardo M. Casanola-Martin,[‡] Dean C. Webster,[‡] Simone A. Ludwig,[*,†] and Bakhtiyor Rasulev[*,‡]

†*Department of Computer Science, North Dakota State University, 1320 Albrecht Boulevard, Fargo, North Dakota 58105*

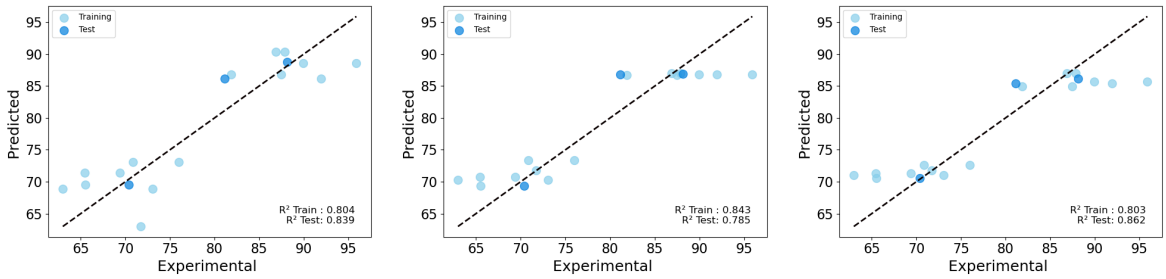
‡*Department of Coatings and Polymeric Materials, North Dakota State University, 1735 NDSU Research Park Drive N, Fargo, North Dakota 58102*

E-mail: simone.ludwig@ndsu.edu; bakhtiyor.rasulev@ndsu.edu

Table S1: Performance metrics for different models and mixture descriptors

Mixture Descriptors	Models	Variables/Equation	R^2_{Train}	R^2_{Test}	MAE $_{Train}$	MAE $_{Test}$	RMSE $_{Train}$	RMSE $_{Test}$	Q^2_{LOO}
combinatorial	LR[a]	$Y = 79.066 + 178.47362 \times$ CIC1-SIC0-S0K $- 350.90637 \times$ Mv-MWC06-MDDD $- 163.33612 \times$ WiA_H2-PW4-Psi_e_A	0.852	0.883	3.507	2.238	4.06	2.495	0.727
	DT[b]	PW5-TIC1-MWC08	0.987	0.987	0.86	0.826	1.192	0.836	0.791
	RF[c]	TIC2-X0A-SMTIV	0.965	0.925	1.557	1.968	1.975	1.999	0.747
	Linear SVR[d]	AMW-X1A-max_conj_path, AMW-Mi-nBM	0.813	0.875	3.224	2.426	4.554	2.585	0.781
	LASSO[e]	$Y = 78.999 - 8.626 \times$ X3sol-MAXDN-AVS_Dz(Z)	0.821	0.737	3.508	2.462	4.457	3.746	0.751
centroid	LR	$Y = 78.33 + 2.703 \times$ E3v $- 10.853 \times$ RDF015v	0.801	0.795	4.112	2.372	4.702	3.304	0.683
	DT	Wi_G/D	0.926	0.561	2.567	3.868	2.868	4.841	0.703
	RF	E3v	0.885	0.694	2.639	3.382	3.579	4.04	0.686
	Linear SVR	Wi_G/D	0.708	0.827	4.432	2.898	5.699	3.041	0.652
	LASSO	$Y = 78.31 - 5.483 \times$ Wi_G/D $+ 0.958 \times$ E3v $- 3.89 \times$ RDF015v	0.8	0.786	4.071	2.217	4.714	3.38	0.662
sqr-diff	LR	$Y = 78.499 - 8.761 \times$ H_Dz(v)	0.799	0.665	4.042	2.539	4.72	4.231	0.733
	DT	H_Dz(v)	0.927	0.54	2.534	4.032	2.847	4.953	0.684
	RF	RDF015v	0.9	0.638	2.6	3.331	3.34	4.392	0.793
	Linear SVR	Mor02u	0.843	0.785	3.316	2.641	4.173	3.389	0.746
	LASSO	$Y = 78.351 - 7.757 \times$ H_Dz(v)	0.791	0.741	3.984	2.778	4.823	3.716	0.725
abs-diff	LR	$Y = 78.439 - 9.314 \times$ Mor02u	0.867	0.632	3.207	3.596	3.844	4.43	0.823
	DT	H_Dz(v)	0.927	0.54	2.534	4.032	2.847	4.953	0.684
	RF	Mor02u	0.91	0.458	2.631	4.495	3.164	5.379	0.81
	Linear SVR	Mor02u	0.803	0.862	3.718	2.152	4.684	2.714	0.728
	LASSO	$Y = 78.303 - 8.284 \times$ Mor02u	0.858	0.751	3.3	2.526	3.974	3.646	0.812
fmol-sum	LR	$Y = 78.36 - 8.905 \times$ Mor09m	0.818	0.662	3.695	3.072	4.493	4.244	0.76
	DT	Mor09m, RDF095i	0.981	0.616	1.104	3.145	1.437	4.527	0.604
	RF	RDF115i, Mor09m	0.964	0.942	1.485	1.535	1.994	1.756	0.741
	Linear SVR	RDF095i	0.79	0.81	3.864	2.532	4.833	3.182	0.735
	LASSO	$Y = 78.23 - 7.9 \times$ Mor09m	0.809	0.746	3.625	2.713	4.603	3.678	0.751
fmol-diff	LR	$Y = 78.348 - 81.614 \times$ Mor02u $+ 72.624 \times$ RDF015v	0.832	0.648	3.632	3.045	4.315	4.335	0.761
	DT	RDF015v	0.986	0.923	0.852	1.724	1.241	2.024	0.77
	RF	RDF015v	0.964	0.895	1.559	2.35	2.011	2.364	0.789
	Linear SVR	Mor02u	0.764	0.753	3.917	3.08	5.115	3.631	0.74
	LASSO	$Y = 78.311 - 8.013 \times$ Mor02u	0.822	0.738	3.505	2.504	4.45	3.738	0.766
sqr-fmol	LR	$Y = 78.199 - 8.59 \times$ Mor09m	0.777	0.688	4.073	3.405	4.982	4.081	0.706
	DT	RDF115i, Mor06e	0.983	0.621	0.981	2.909	1.362	4.495	0.63
	RF	RDF115i	0.966	0.904	1.578	1.993	1.9477	2.257	0.754
	Linear SVR	RDF115i, Mor06e, Mor09m	0.781	0.749	3.895	3.001	4.929	3.657	0.666
	LASSO	$Y = 78.144 - 7.687 \times$ Mor06e	0.778	0.756	3.932	2.862	4.967	3.61	0.711
root-fmol	LR	$Y = 78.439 - 9.008 \times$ Mor06e	0.833	0.648	3.562	2.852	4.31	4.332	0.778
	DT	RDF115i	0.981	0.653	1.157	2.952	1.467	4.305	0.704
	RF	RDF095i	0.973	0.935	1.486	1.548	1.725	1.869	0.77
	Linear SVR	RDF115i, RDF095i	0.773	0.839	3.927	2.287	5.017	2.933	0.711
	LASSO	$Y = 78.301 - 7.998 \times$ Mor06e	0.824	0.739	3.479	2.519	4.424	3.731	0.768
sqr-fmol-sum	LR	$Y = 78.368 - 8.842 \times$ Mor09m	0.803	0.664	3.841	3.001	4.675	4.237	0.74
	DT	RDF095i, Mor09m	0.981	0.616	1.104	3.145	1.437	4.527	0.611
	RF	RDF115i, Mor09m	0.965	0.942	1.475	1.535	1.983	1.756	0.745
	Linear SVR	RDF095i	0.808	0.792	3.674	2.602	4.624	3.332	0.752
	LASSO	$Y = 78.235 - 7.828 \times$ Mor09m	0.794	0.752	3.804	2.645	4.78	3.634	0.732
norm-cont	LR	$Y = 78.392 - 8.94 \times$ Mor06e	0.822	0.656	3.665	2.987	4.453	4.283	0.764
	DT	RDF115i	0.983	0.616	1.052	3.112	1.369	4.528	0.707
	RF	Mor06e, RDF095i	0.91	0.718	2.476	2.84	3.166	3.877	0.816
	Linear SVR	RDF095i	0.754	0.745	3.998	3.188	5.227	3.691	0.734
	LASSO	$Y = 78.259 - 7.930 \times$ Mor06e	0.813	0.745	3.594	2.638	4.563	3.691	0.755
mol-dev	LR	$Y = 78.575 + 5.923 \times$ RDF015v $- 3.059 \times$ Mor02u	0.854	0.673	3.003	3.06	4.024	4.179	0.785
	DT	RDF015v, Mor02u	0.965	0.798	1.442	2.574	1.975	3.28	0.778
	RF	RDF015v	0.974	0.865	1.358	2.627	1.694	2.685	0.852
	Linear SVR	Mor02u	0.806	0.832	3.563	2.808	4.646	2.99	0.764
	LASSO	$Y = 78.404 - 7.981 \times$ Mor02u	0.837	0.773	3.219	2.651	4.261	3.482	0.794
sqr-mol-dev	LR	$Y = 77.884 + 4.003 \times$ RARS $- 5.421 \times$ Mor15e	0.817	0.716	3.547	3.362	4.509	3.889	0.687
	DT	Mor15e	0.962	0.589	1.479	4.198	2.053	4.684	0.541
	RF	Mor15e	0.895	0.784	2.674	3.087	3.422	3.397	0.622
	Linear SVR	ZM2Mad, RARS, Mor15e	0.808	0.73	3.394	3.171	4.618	3.797	0.707
	LASSO	$Y = 77.79 + 3.621 \times$ RARS $- 4.738 \times$ Mor15e	0.807	0.742	3.669	3.5	4.63	3.707	0.692
mol-dev-sqr	LR	$Y = 77.778 + 72.1 \times$ B04[O-Si] $+ 85.431 \times$ F08[O-S] $+ 81.683 \times$ HATS5u	0.887	0.606	2.773	4.156	3.543	4.586	0.807
	DT	B04[O-Si], F08[O-S]	0.899	0.651	2.527	3.191	3.346	4.317	0.622
	RF	B04[O-Si], HATS5u	0.967	0.707	1.418	3.036	1.922	3.952	0.725
	Linear SVR	B04[O-Si]	0.751	0.649	4.113	4.198	5.263	4.33	0.573
	LASSO	$Y = 78.336 - 7.326 \times$ B04[O-Si]	0.757	0.609	4.025	4.323	5.198	4.569	0.693

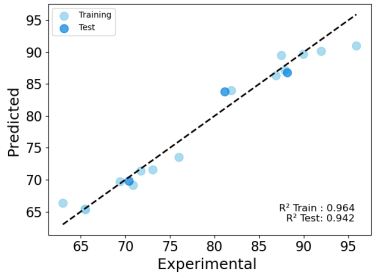
[a]Linear regression, [b]Decision tree, [c]Random forest, [d]Support vector regression, [e] Least absolute shrinkage and selection operator



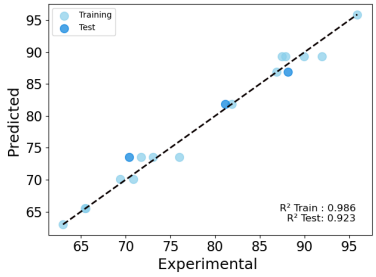
(a) centroid + LR

(b) sqr-diff + SVR

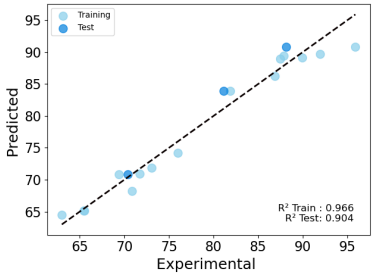
(c) abs-diff + SVR



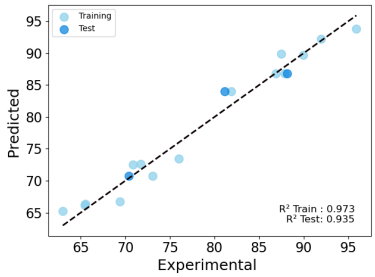
(d) fmol-sum + RF



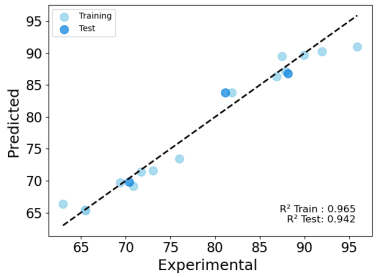
(e) fmol-diff + DT



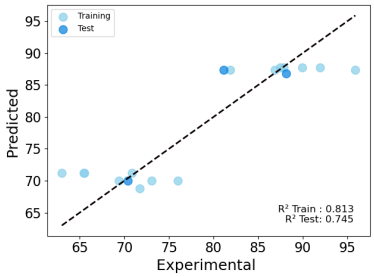
(f) sqr-fmol + RF



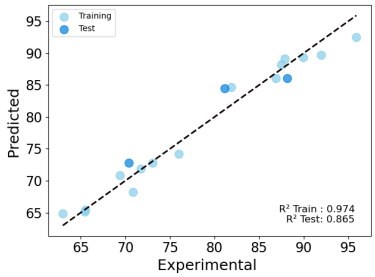
(g) root-fmol + RF



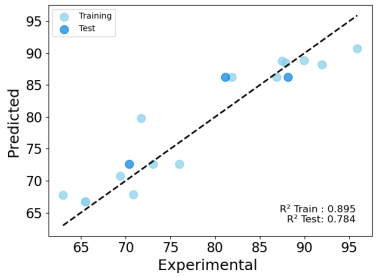
(h) sqr-fmol-sum + RF



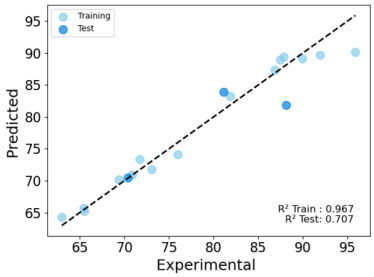
(i) norm-cont + LASSO



(j) mol-dev + RF

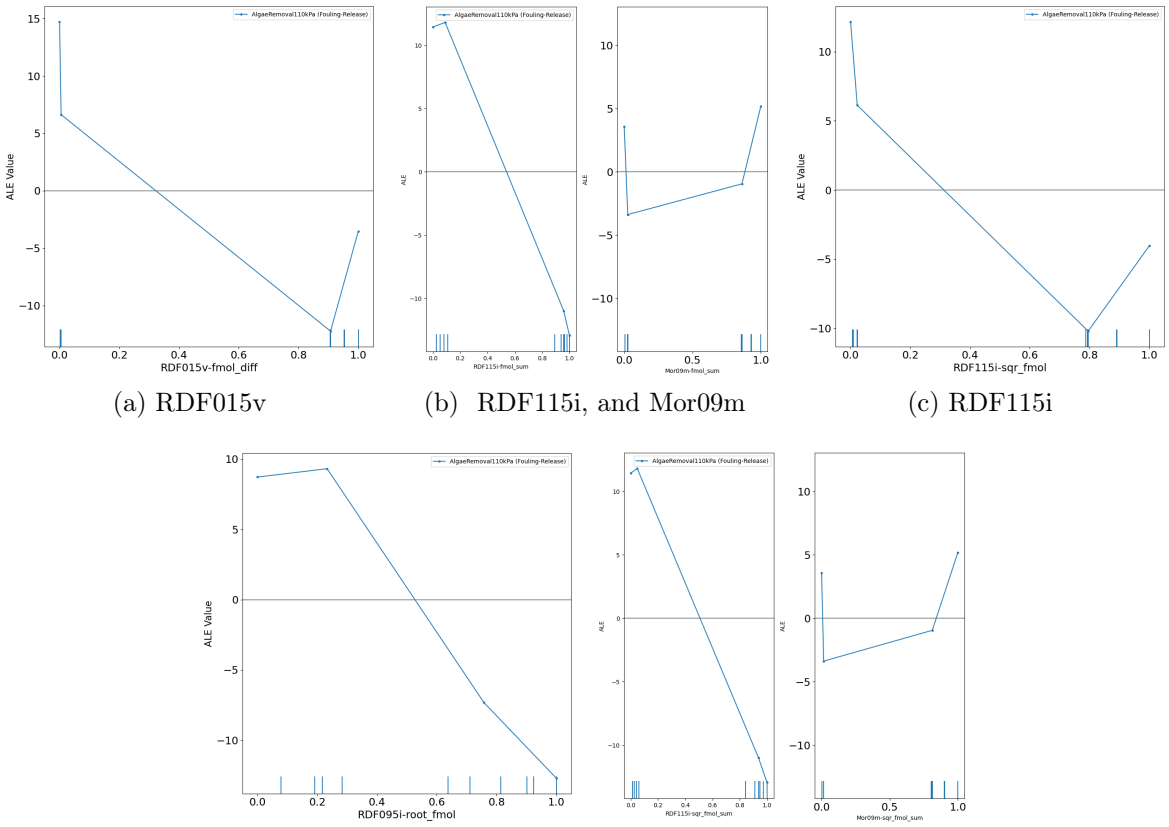


(k) sqr-mol-dev + RF



(l) mol-dev-sqr + RF

Figure S1: **Correlation plots of the best models of each additive mixture descriptor.** Observed vs predicted correlation plots of the best models of each additive mixture descriptor. Linear regression (LR), least absolute shrinkage and selection operator (LASSO), decision tree (DT), random forest (RF), support vector regression (SVR)



(a) RDF015v (b) RDF115i, and Mor09m (c) RDF115i

(d) RDF095i (e) RDF115i, and Mor09m

Figure S2: **ALE plots for the decision tree and random forest models showing the contribution of RDF015v, RDF115i, RDF095i and Mor09m mixture descriptors**. The ontribution of additive mixture descriptors, calculated using a) fmol-diff, b) fmol-sum, c) sqr-fmol d) root-fmol, and e) sqr-fmol-sum mixing rules, to FR activity, as represented by ALE plots for the decision tree and random forest models.