SUPPORTING INFORMATION

Optimizing Oil Detachment from Silica Surfaces Using Gemini Surfactants and Functionalized Silica Nanoparticles: A Combined Molecular Dynamics and Machine Learning Approach.

Gourav Chakraborty[a], *Keka Ojha*[b], *Ajay Mandal*[b], *Niladri Patra*[a*]

[a]Department of Chemistry and Chemical Biology, Indian Institute of Technology (ISM) Dhanbad, Dhanbad - 826004, India

[b]Department of Petroleum Engineering, Indian Institute of Technology (ISM) Dhanbad, Dhanbad - 826004, India

[*]Corresponding author: Niladri Patra, email: npatra2@iitism.ac.in

INDEX

Oil Component	Number of Molecules	Weight (%)
Hexane	56	8.61
Heptane	70	12.51
Octane	80	16.30
Dodecane	104	31.60
Benzene	20	2.79
Toluene	64	10.52
Cyclohexane	36	5.40
Cycloheptane	70	12.26

Table S1: Molecular description of the hydrocarbon components present in Model-2.

van der Waals Parameters:

Atom Type	σ (nm)	ε (kJ mol^{-1})
C	0.350	0.27614
H	0.242	0.12552
Si	0.339	2.44704
O	0.307	0.71128

Atomic partial charges for silanol groups

Atom	Charge (e)
Si (Si-OH)	0.31
O (O-H)	-0.71
H (H-O)	0.40

Table S2: Force Field Parameters for Silica Surface.

For silica surface, the charges on the bulk Si and O atoms of SiO_2 crystal was set to zero. Furthermore, the methyl groups of the hydrophobic silica surfaces were given zero partial charge. This was done to ensure their truly hydrophobic nature. To generate topology for the silica slabs, the inbuilt *"x2top"* command of GROMACS was used, with bond lengths set to their default values, as obtained from Materials Studio database. Non-bonded parameters and partial charges were taken from **Table S2**.

Force-Field Validation:

The validation of the chosen parameters was done by estimating oil-water IFT values for different hydrocarbons used in this work. The protocol developed by *Muller et al*[1] was implemented. Initially a 5x5x4 nm^3 box was generated with a single hydrocarbon placed in its center. Then the entire box was filled completely with similar hydrocarbon molecules. The resulting box was extended along Z-axis to bring the final dimensions to 5x5x10 nm^3. This vacant space on either side was solvated with SPC/E water[2], thereby generating a biphasic system. The resulting assemblies for some representative hydrocarbons from each category (P, N, and A) are shown in Figure S1. Each system was thoroughly minimized in several steps,

followed by 15 ns NVT equilibration. Finally, 15 ns of $NP_{normal}AT$ production was carried out, with the reference pressure set to 1 bar along XY plane using Berendsen Barostat[3].

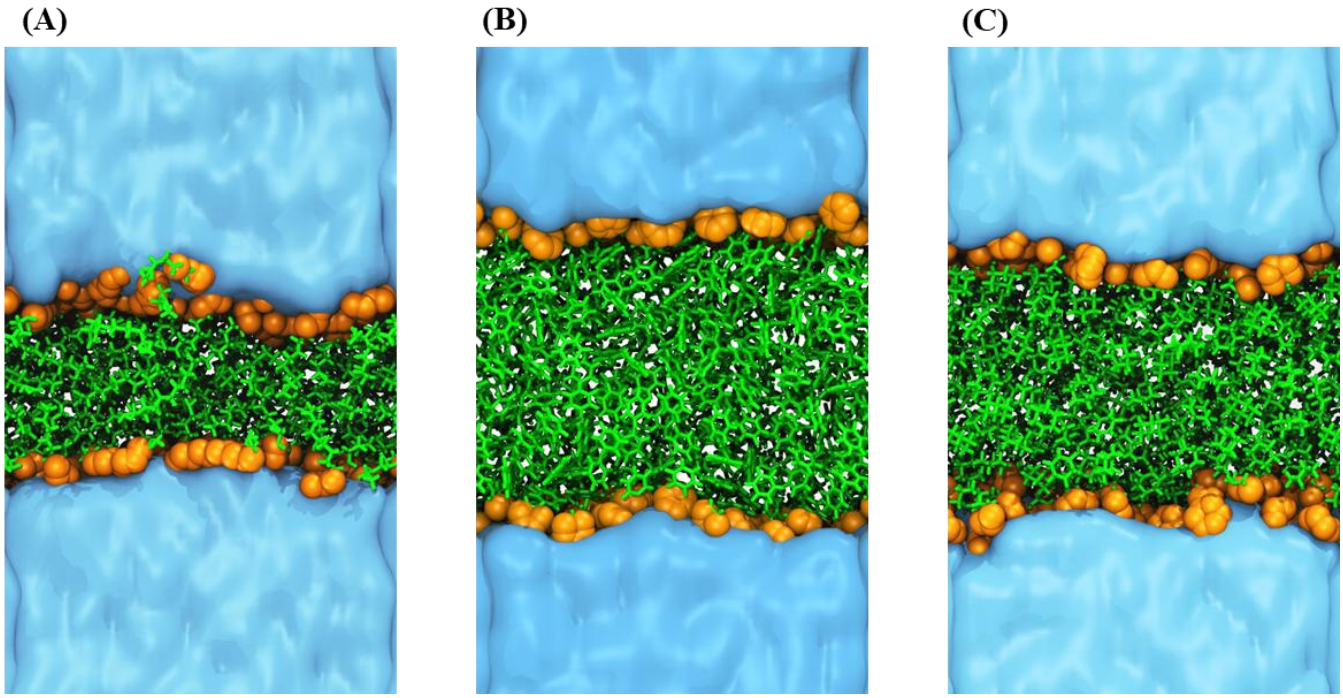

Figure S1: Bilayer sandwiched configuration of **(A)** Dodecane, **(B)** Benzene, and **(C)** Cyclohexane generated for oil-water IFT studies. Colour scheme: **Blue:** water, **Green:** Hydrocarbons, **Orange:** Carbon and within 5 Å of water.

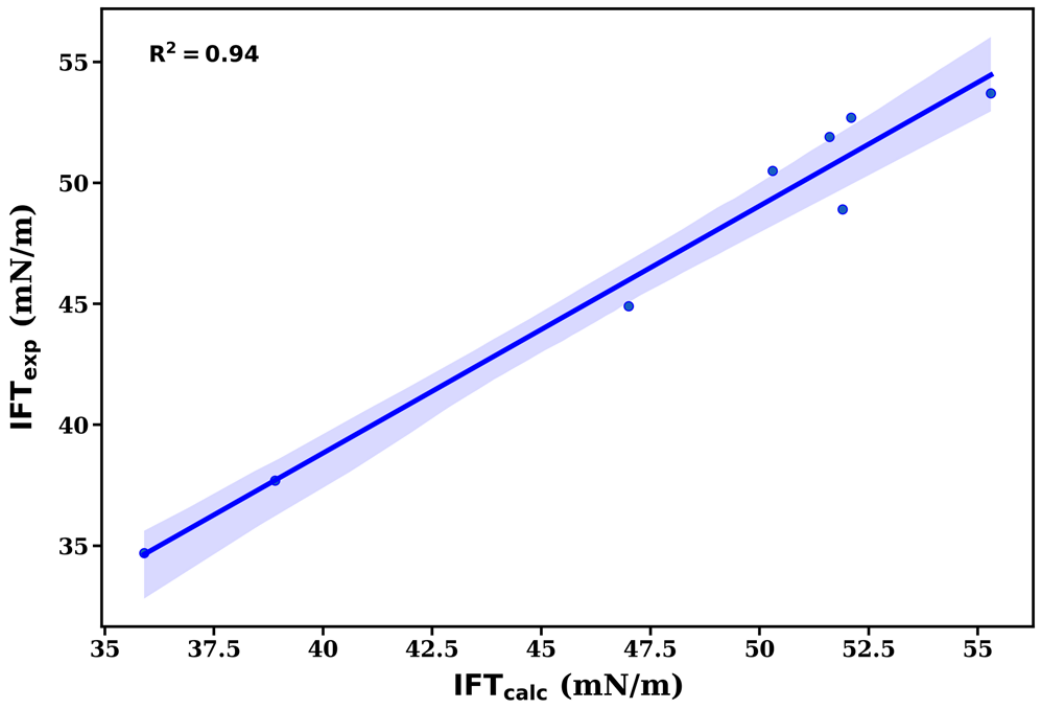

Figure S2: Correlation between calculated and experimental oil-water IFT values for the hydrocarbons of Model-1, and Model-2.

Hydrocarbon	IFT_{calc} (mN/m)	IFT_{exp} (mN/m)	Error (%)
Hexane	50.3	50.5	0.4
Heptane	51.6	51.9	0.6
Octane	52.1	52.7	1.1
Dodecane	55.3	53.7	3.0
Benzene	35.9	34.7	3.5
Toluene	38.9	37.7	3.2
Cyclohexane	51.9	48.9	6.2
Cycloheptane	47.0	44.9	4.7

Table S3: Comparison between calculated and experimental oil-water IFT values for hydrocarbons of Model-1, and Model-2.

Surface Hydrophobicity	Dodecane-Water IE	Dodecane-GS IE	Dodecane-SNP IE
Solvent-Silica IE	Solvent-GS IE	Solvent-SNP IE	Silica-GS IE
Silica-SNP IE	GS-SNP IE	Silica-Dodecane IE	Δ(Silica-Dodecane) IE
Dodecane RDF peak value	Dodecane RDF peak distance	Fraction of dodecane detached	

Table S4: Complete descriptor set tested during machine learning models for oil-model 1. Here IE represents the interaction energy between the given components (in kJ/mol), and Δ(Silica-Dodecane) IE represents $\Delta\left(E_{DOD-SIL,GS+SNP} - E_{DOD-SIL,waterflooding}\right)$.

Surface Hydrophobicity	Oil-Water IE	Oil-GS IE	Oil-SNP IE
Solvent-Silica IE	Solvent-GS IE	Solvent-SNP IE	Silica-GS IE
Silica-SNP IE	GS-SNP IE	Silica-Oil IE	Δ(Silica-Oil) IE
Oil RDF peak value	Oil RDF peak distance	Fraction of Oil detached	

Table S5: Complete descriptor set tested during machine learning models for oil-model 2. Here IE represents the interaction energy between the given components (in kJ/mol), and Δ(Silica-Oil) IE represents $\Delta\left(E_{Oil-SIL,GS+SNP} - E_{Oil-SIL,waterflooding}\right)$.

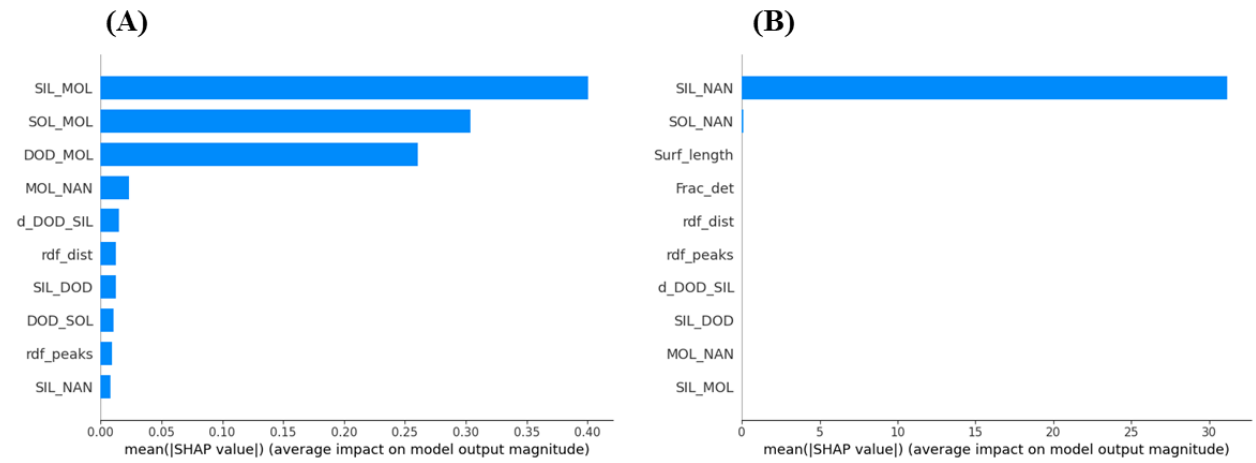

Figure S3: SHAP scores of the 10 best descriptors for **(A)** ML-1, and **(B)** ML-2, based on oil Model-1.

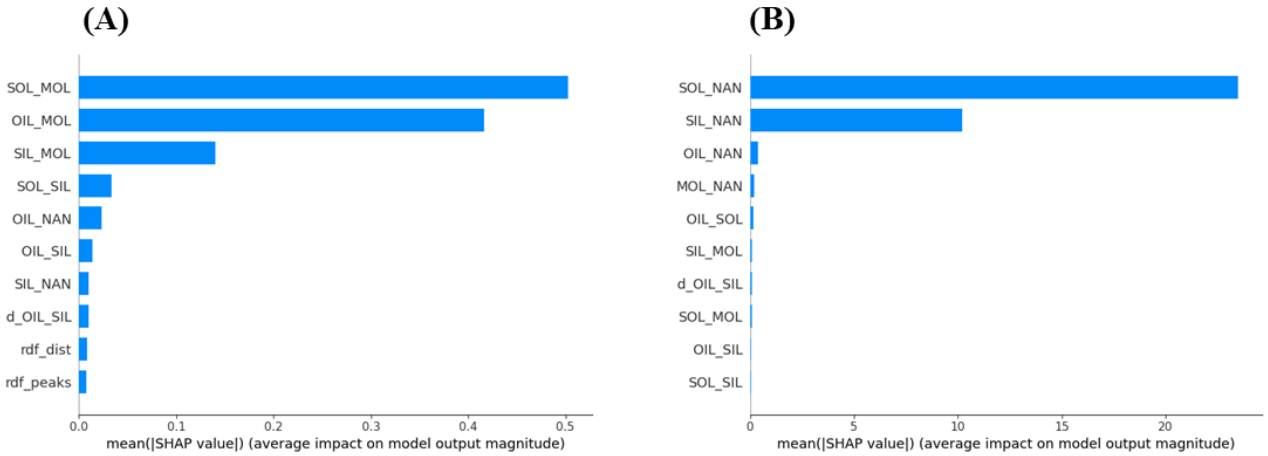

Figure S4: SHAP scores of the 10 best descriptors for **(A)** ML-1, and **(B)** ML-2, based on oil Model-2.

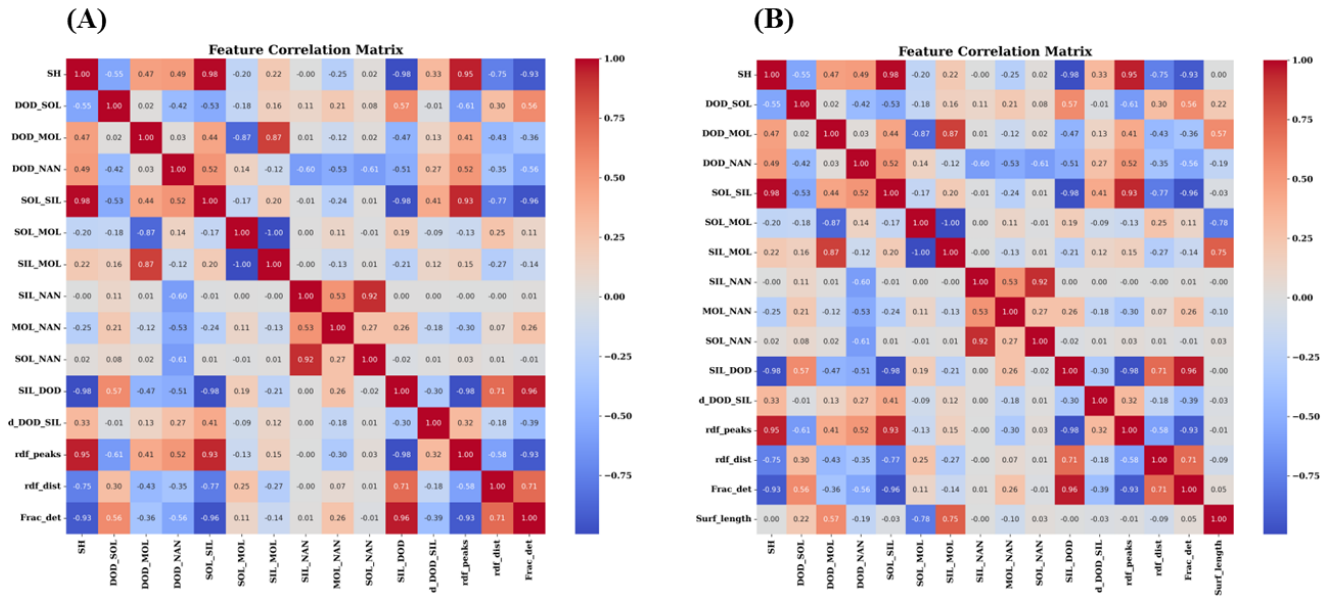

Figure S5: Pairwise Pearson's Correlation Coefficient Matrix for descriptors of **(A)** ML-1, and **(B)** ML-2, based on oil Model-1.

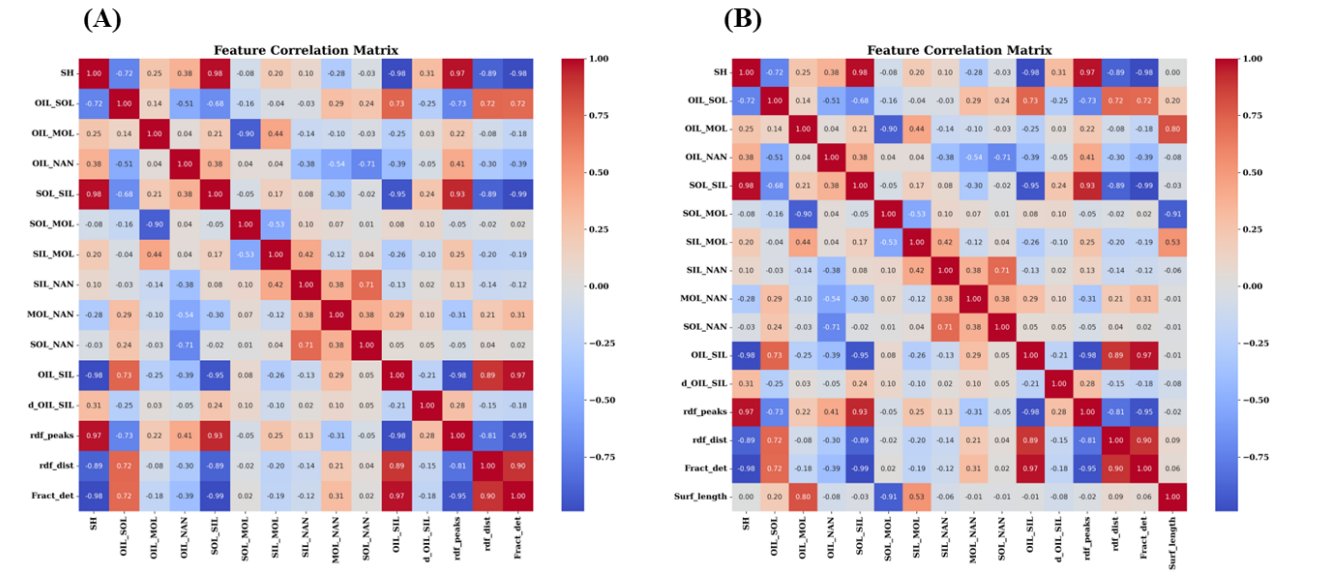

Figure S6: Pairwise Pearson's Correlation Coefficient Matrix for descriptors of **(A)** ML-1, and **(B)** ML-2, based on oil Model-2.

Oil Model	Silica Hydrophobicity	R_g (nm)
1	0	6.46 ± 0.10
1	25	6.50 ± 0.09
1	50	6.18 ± 0.09
1	75	6.43 ± 0.14
1	100	6.30 ± 0.16
2	0	6.57 ± 0.08
2	25	6.42 ± 0.08
2	50	6.48 ± 0.08
2	75	6.45 ± 0.08
2	100	6.46 ± 0.08

Table S6: Average values of R_g along with their associated errors for the two oil models.

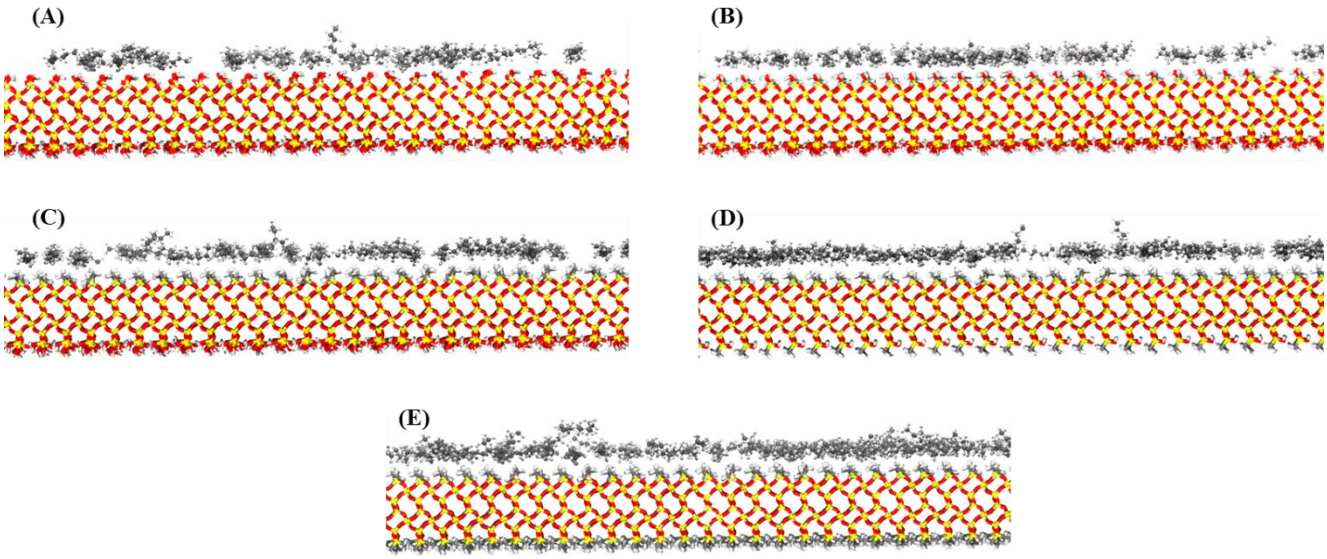

Figure S7: Configuration of adsorbed oil of Model-1 over different silica slabs. The surface hydrophobicity (as %) of the silica surfaces are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100.

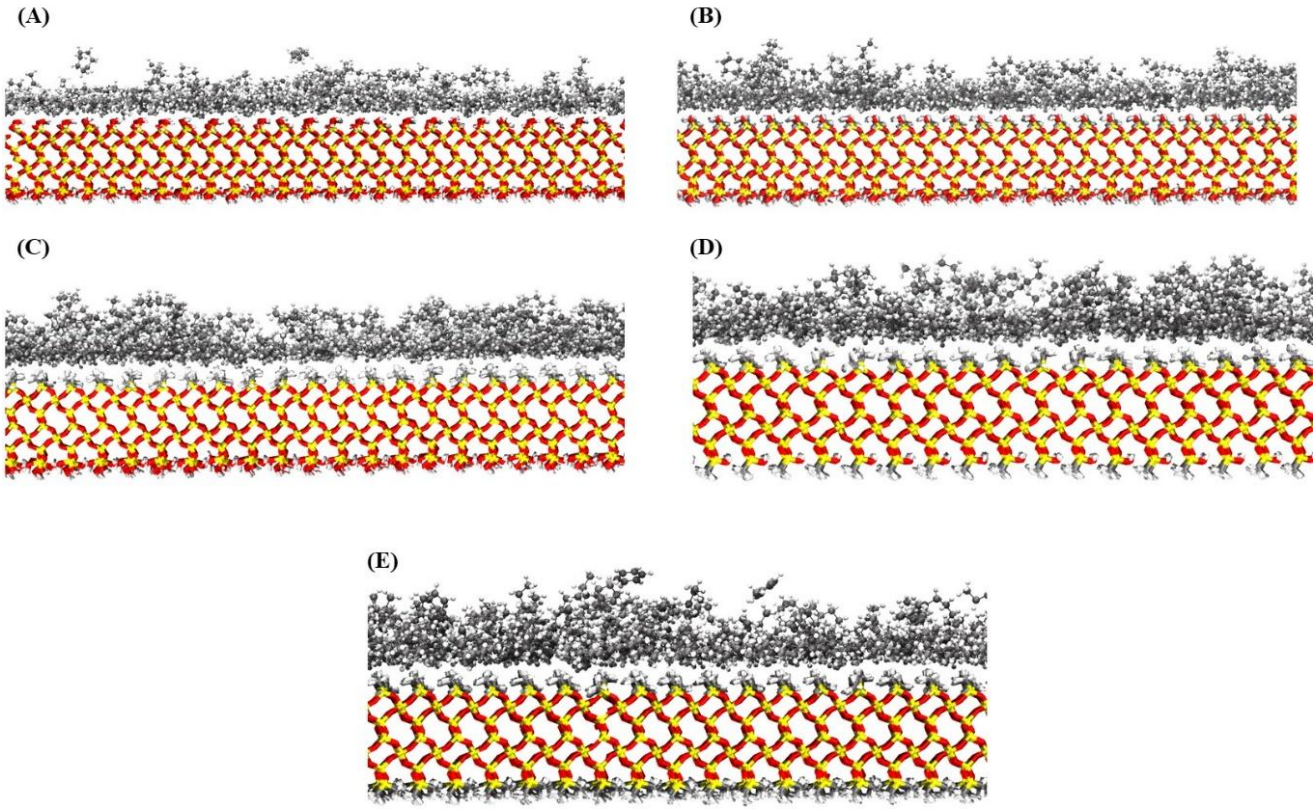

Figure S8: Configuration of adsorbed oil of Model-2 over different silica slabs. The surface hydrophobicity (as %) of the silica surfaces are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100.

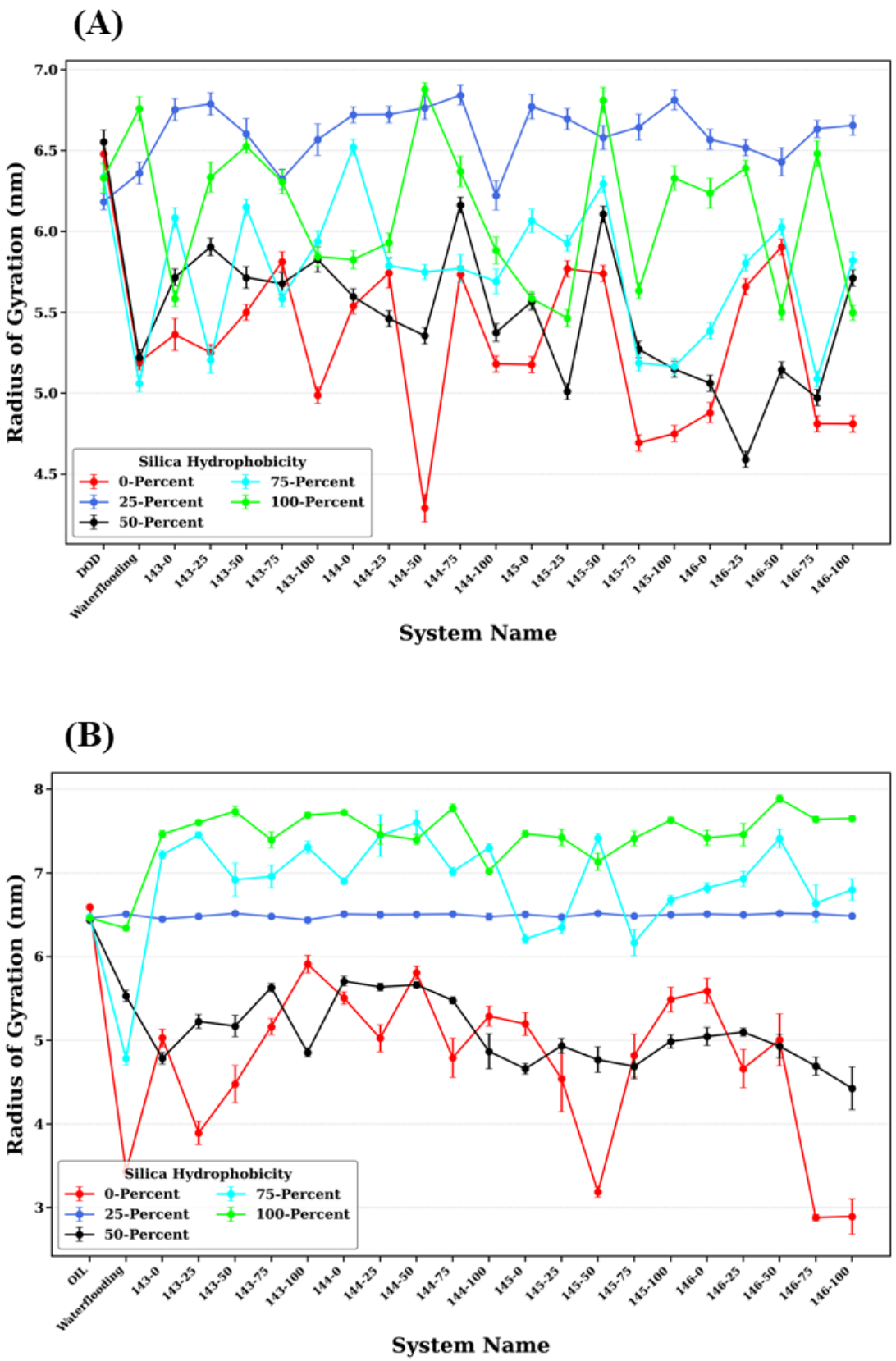

Figure S9: Variation in radius of gyration (R_g) for the oil molecules across various systems of **(A)** Model-1, and **(B)** Model-2

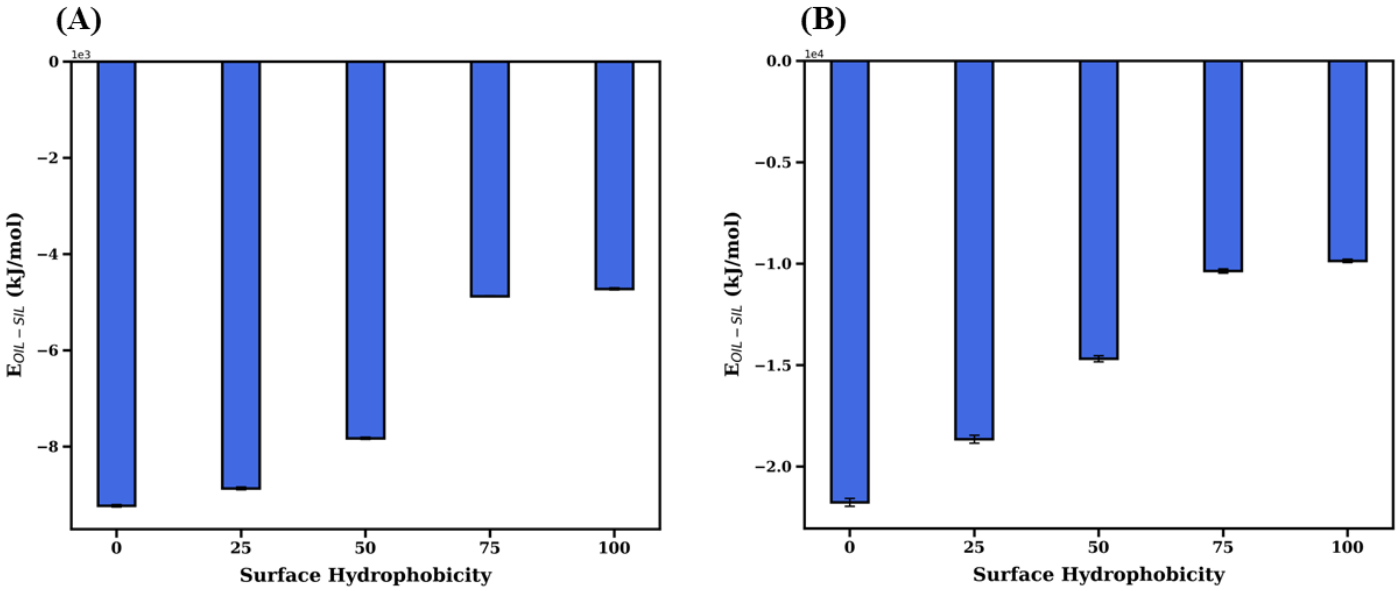

Figure S10: Oil-Silica Interaction energies for **(A)** Model-1, and **(B)** Model-2.

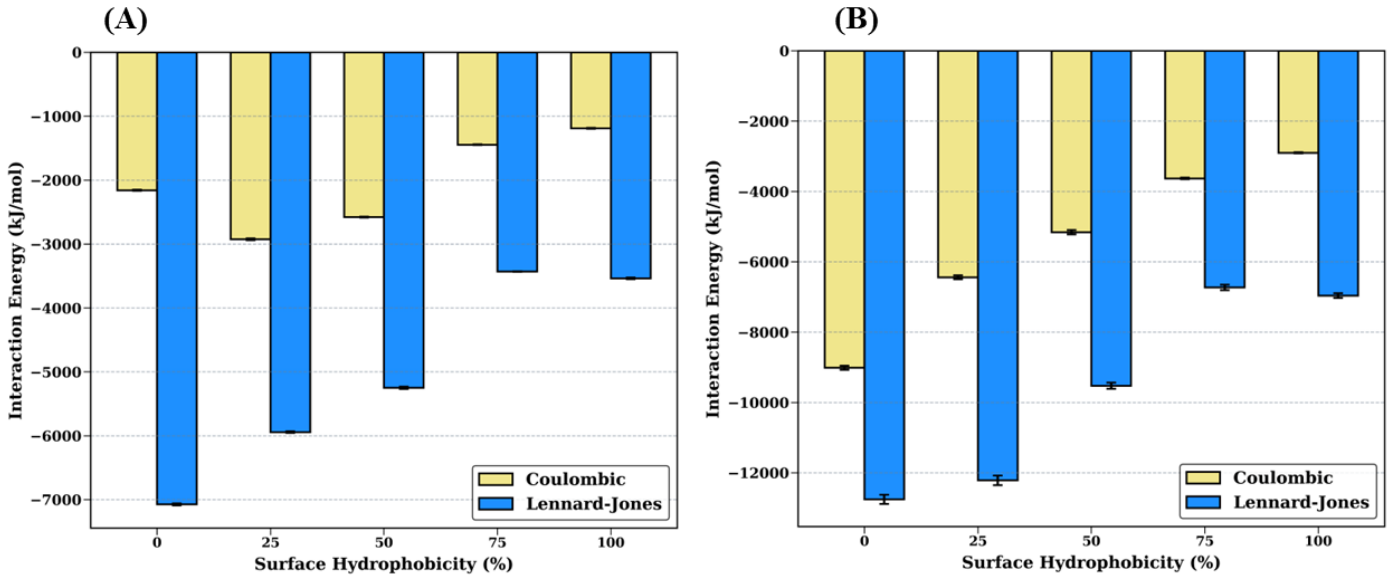

Figure S11: Actual contributions from Coulombic and Lennard-Jones energies toward oil-silica interaction energy values. Here **(A)** Model-1 and **(B)** Model-2.

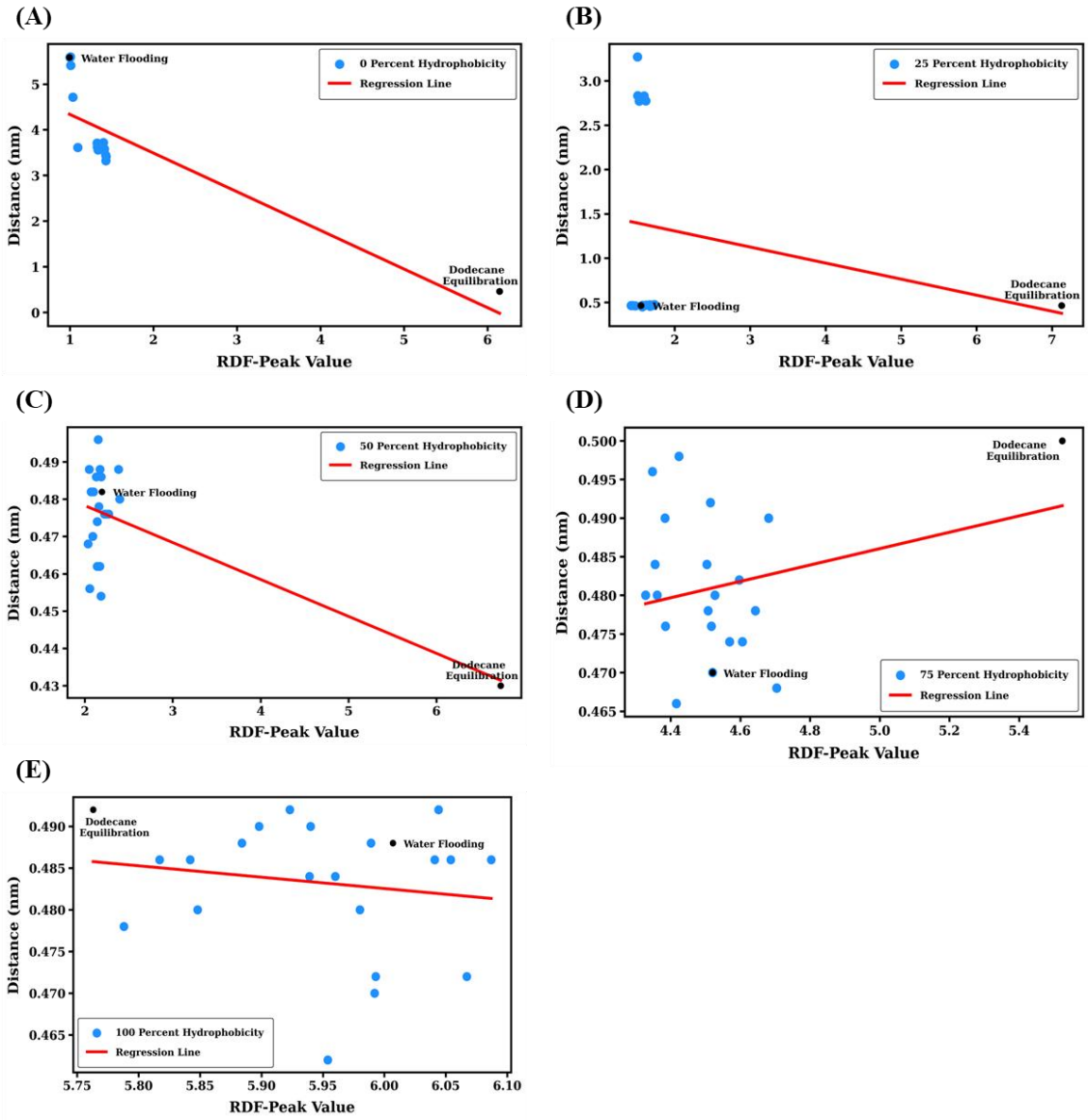

Figure S12: Variation of Model-1 oil RDF peaks with distance for different silica surfaces. The hydrophobicity (%) of the silica slabs are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. Colour scheme: **Black markers:** dodecane adsorbed over silica slab in vacuum, and waterflooding; **Blue markers:** different combinations of surfactants and nanoparticles; **red line**: best fit regression line.

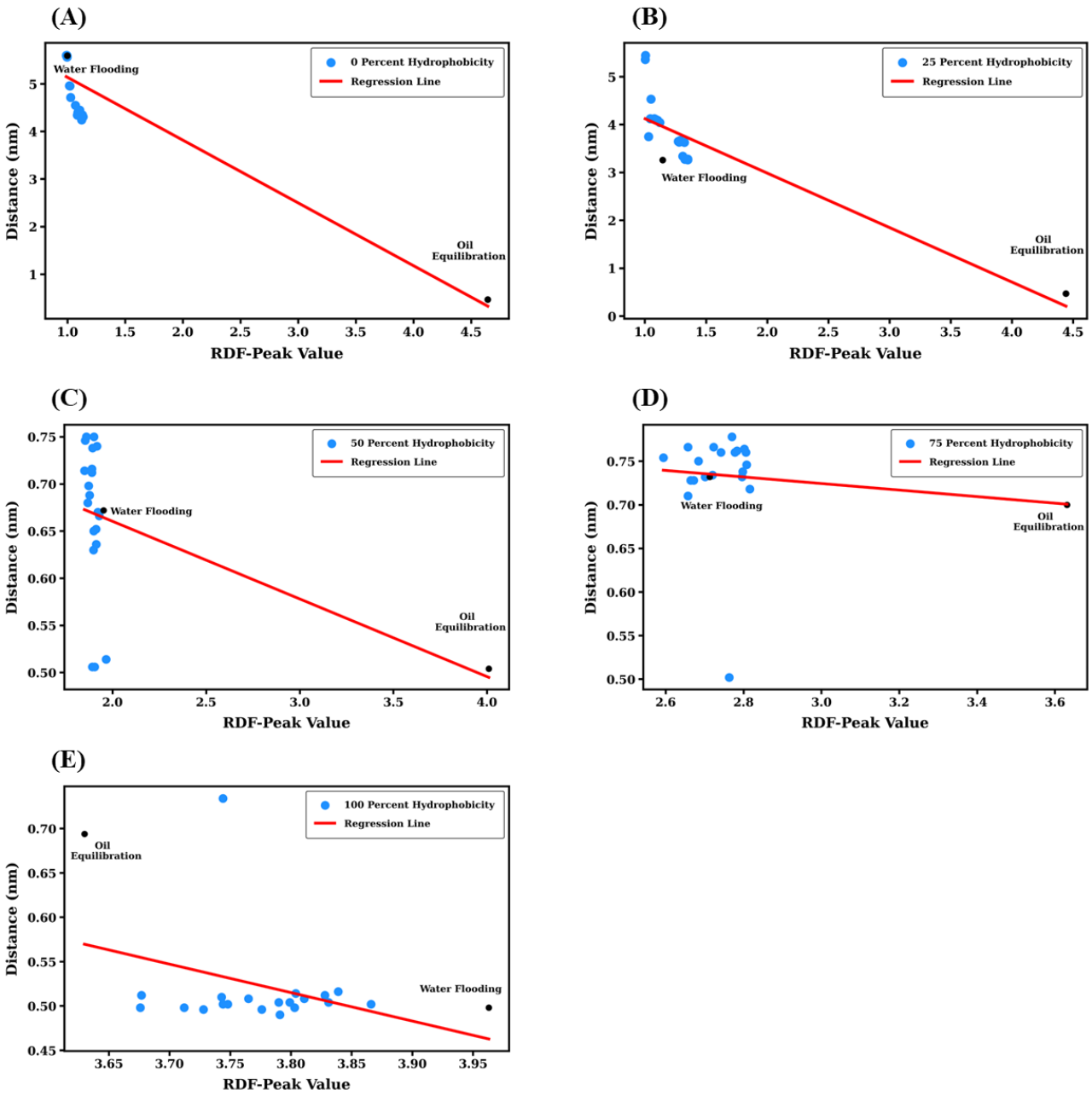

Figure S13: Variation of Model-2 oil RDF peaks with distance for different silica surfaces. The hydrophobicity (%) of the silica slabs are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. Colour scheme: **Black markers:** dodecane adsorbed over silica slab in vacuum, and waterflooding; **Blue markers:** different combinations of surfactants and nanoparticles; **red line**: best fit regression line.

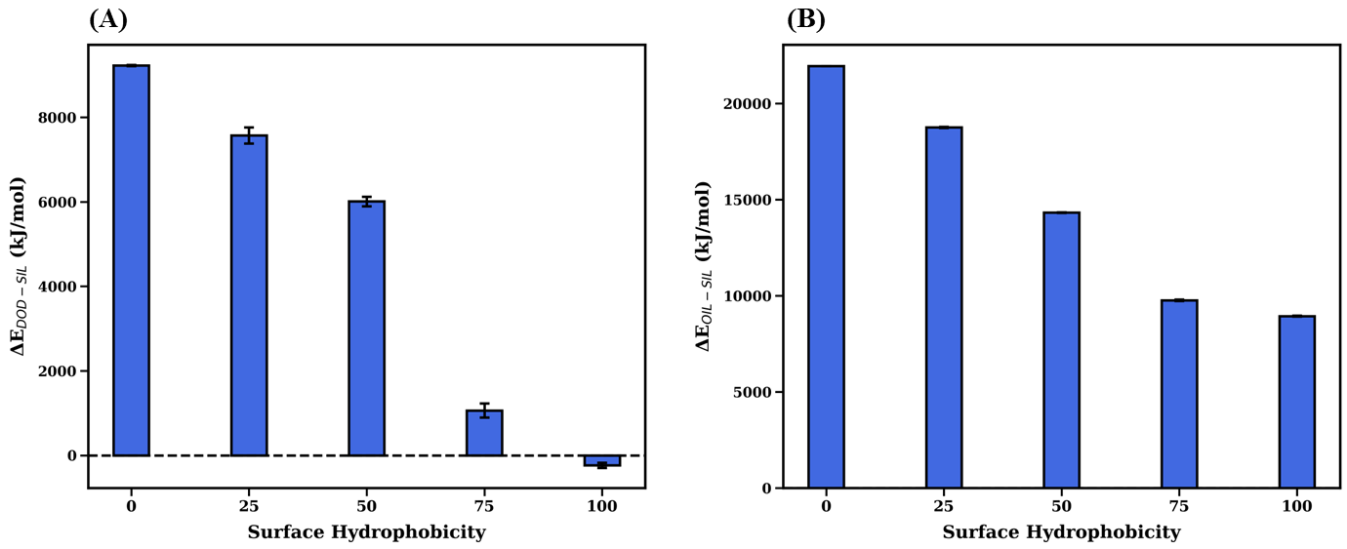

Figure S14: Change in **(A)** DOD-SIL, and **(B)** OIL-SIL interaction energies for silica surfaces with different hydrophobicity. Here, Δ represents the difference in the oil-silica interaction energy for the GS + SNP added system, and the oil adsorbed system in vacuum. Here, averaging is done over all the energy differences as observed for a given silica slab.

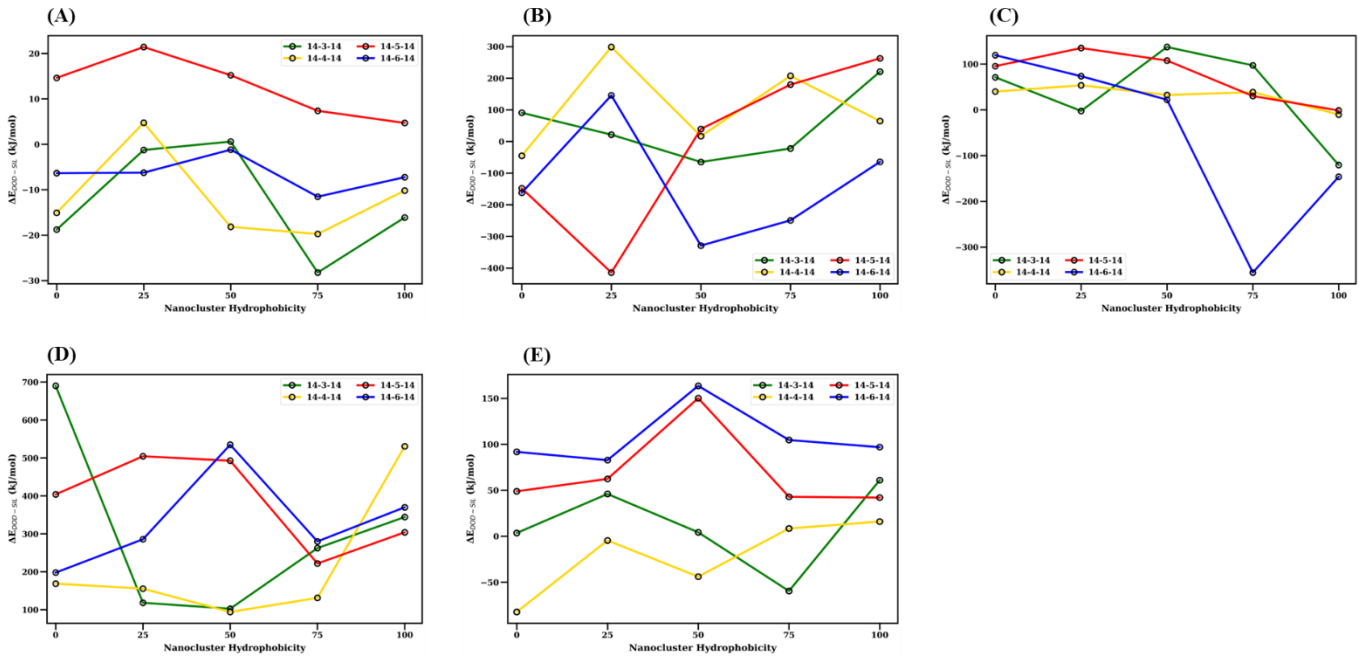

Figure S15: Variation in $\Delta E_{DOD-SIL}$ for Model-1 oil molecules over different silica surfaces. The hydrophobicity (%) of the silica slabs are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100.

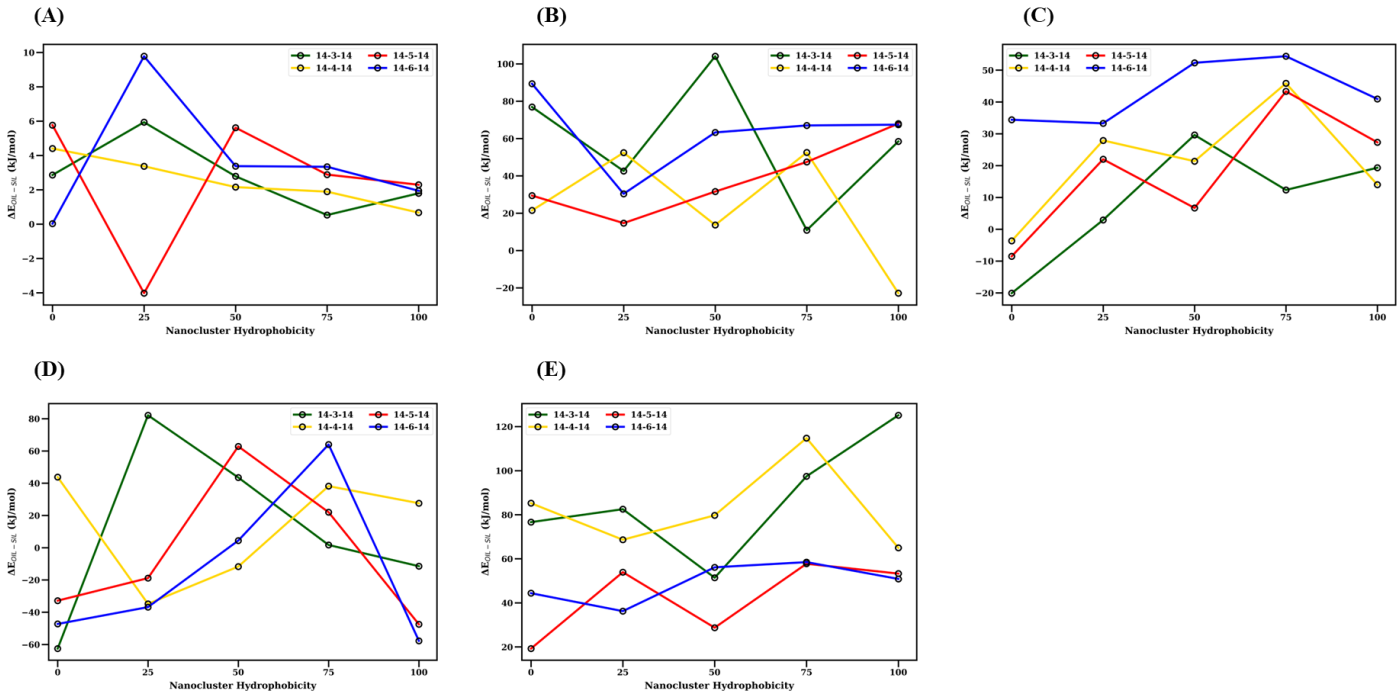

Figure S16: Variation in $\Delta E_{OIL\text{-}SIL}$ for Model-2 oil molecules over different silica surfaces. The hydrophobicity (%) of the silica slabs are **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100.

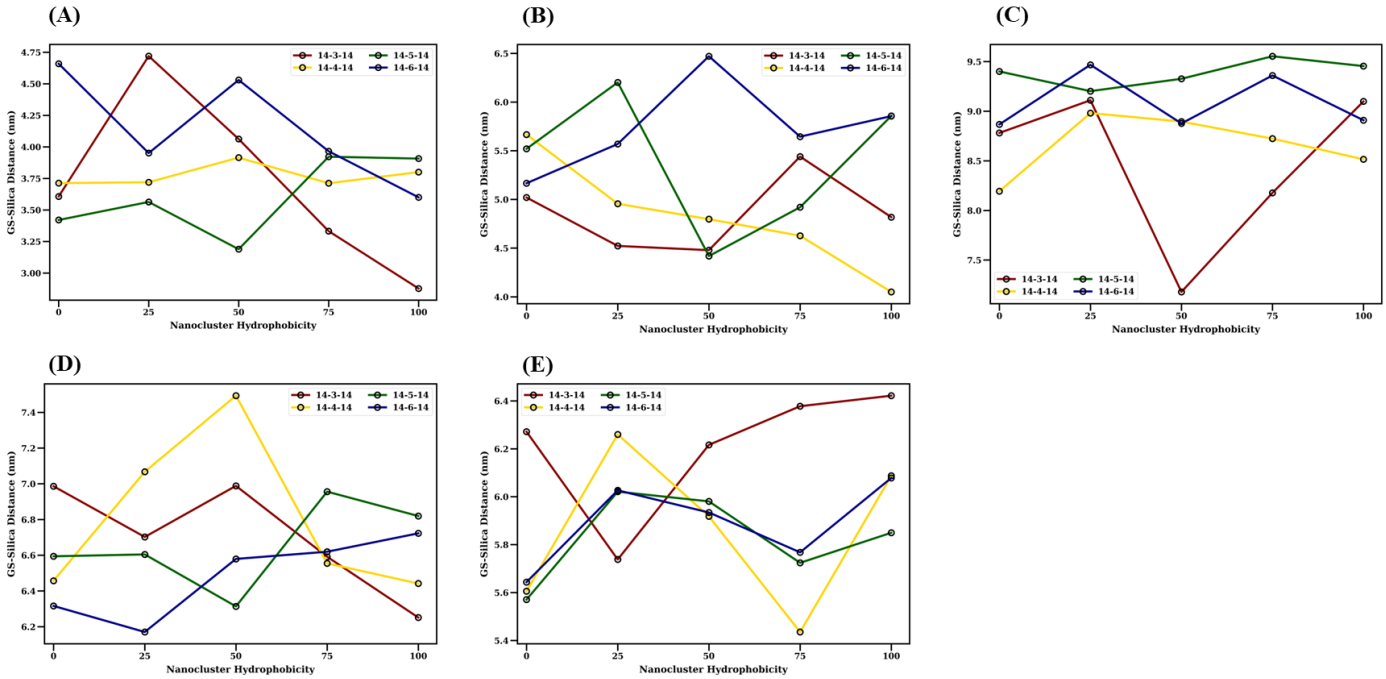

Figure S17: Plots for distance variation between the COM of GS and Silica surfaces, in presence of different SNPs. The hydrophobicity (in %) of the silica slabs are: **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. The variation is for oil Model-1.

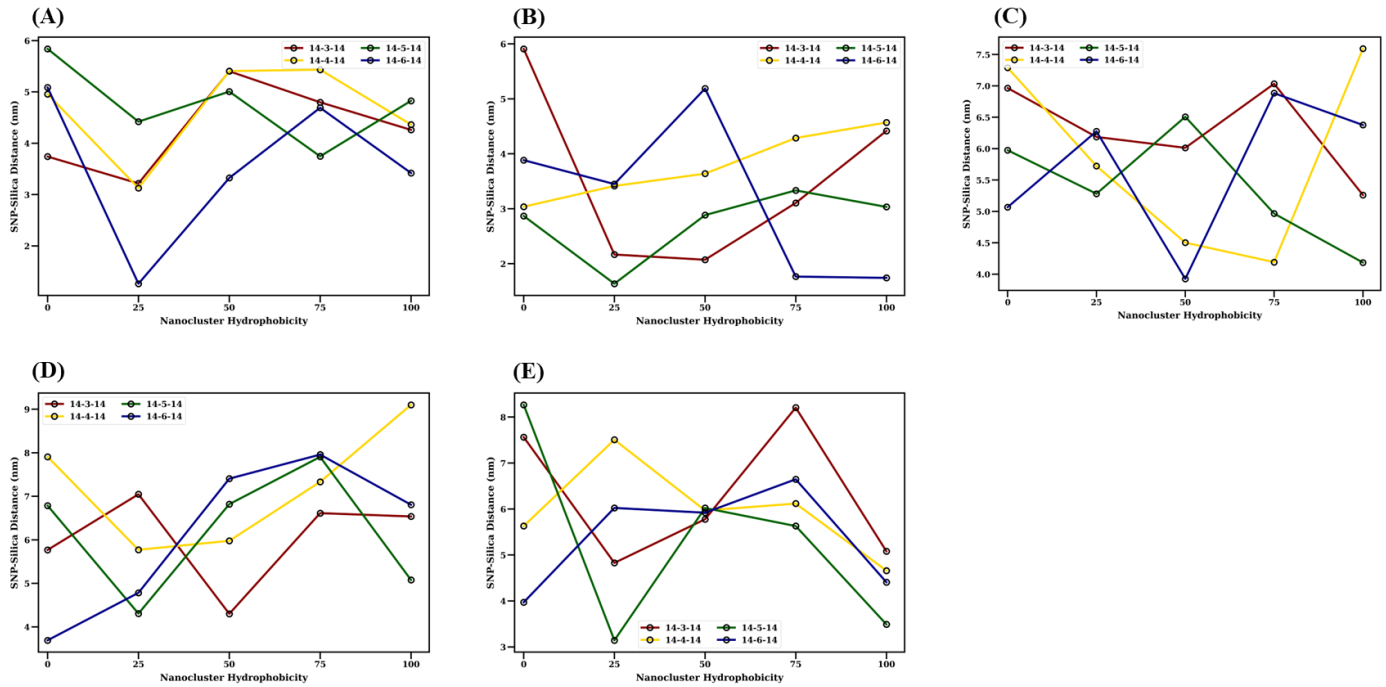

Figure S18: Plots for distance variation between the COM of SNP and Silica surfaces, in presence of different GSs. The hydrophobicity (in %) of the silica slabs are: **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. The variation is for oil Model-1.

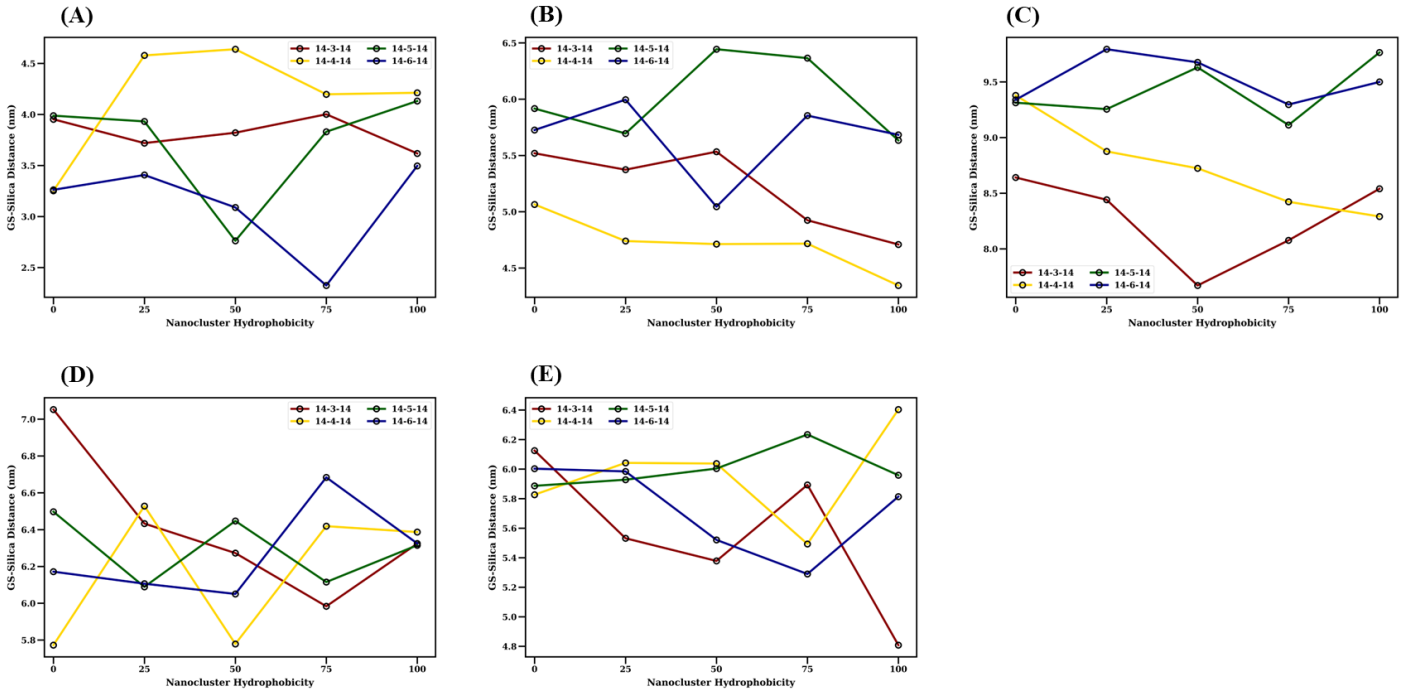

Figure S19: Plots for distance variation between the COM of GS and Silica surfaces, in presence of different SNPs. The hydrophobicity (in %) of the silica slabs are: **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. The variation is for oil Model-2.

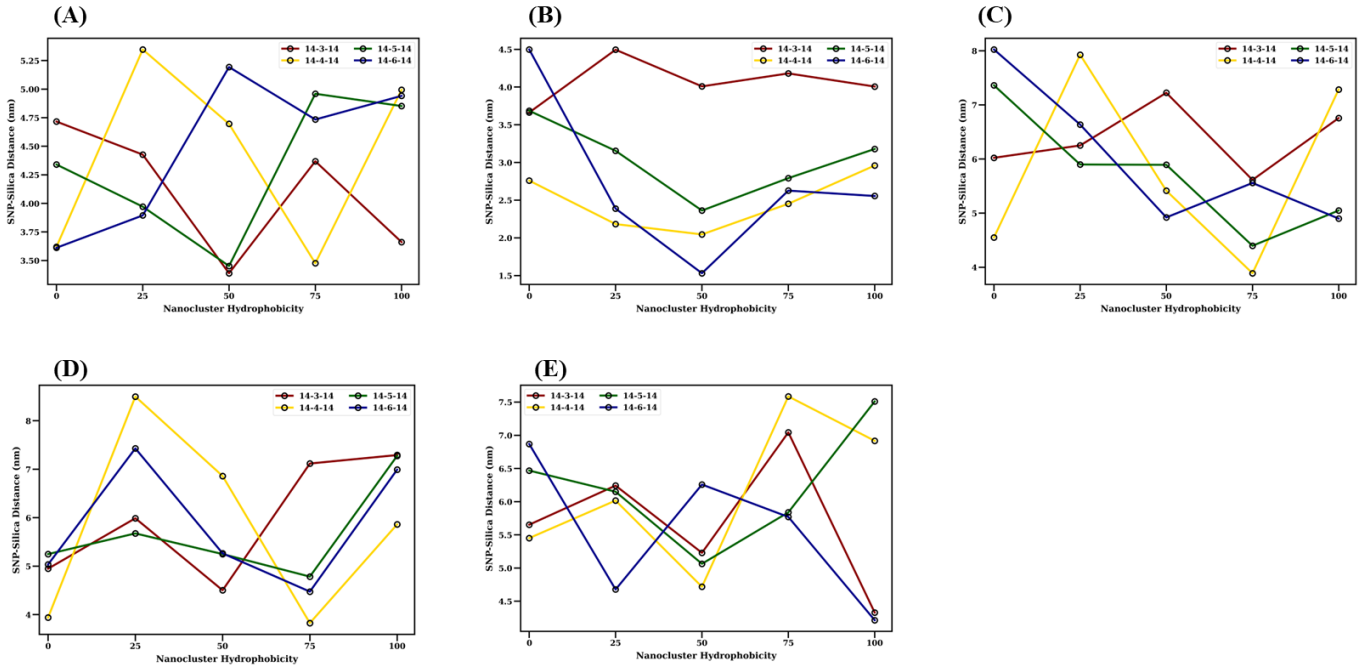

Figure S20: Plots for distance variation between the COM of SNP and Silica surfaces, in presence of different GSs. The hydrophobicity (in %) of the silica slabs are: **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. The variation is for oil Model-2.

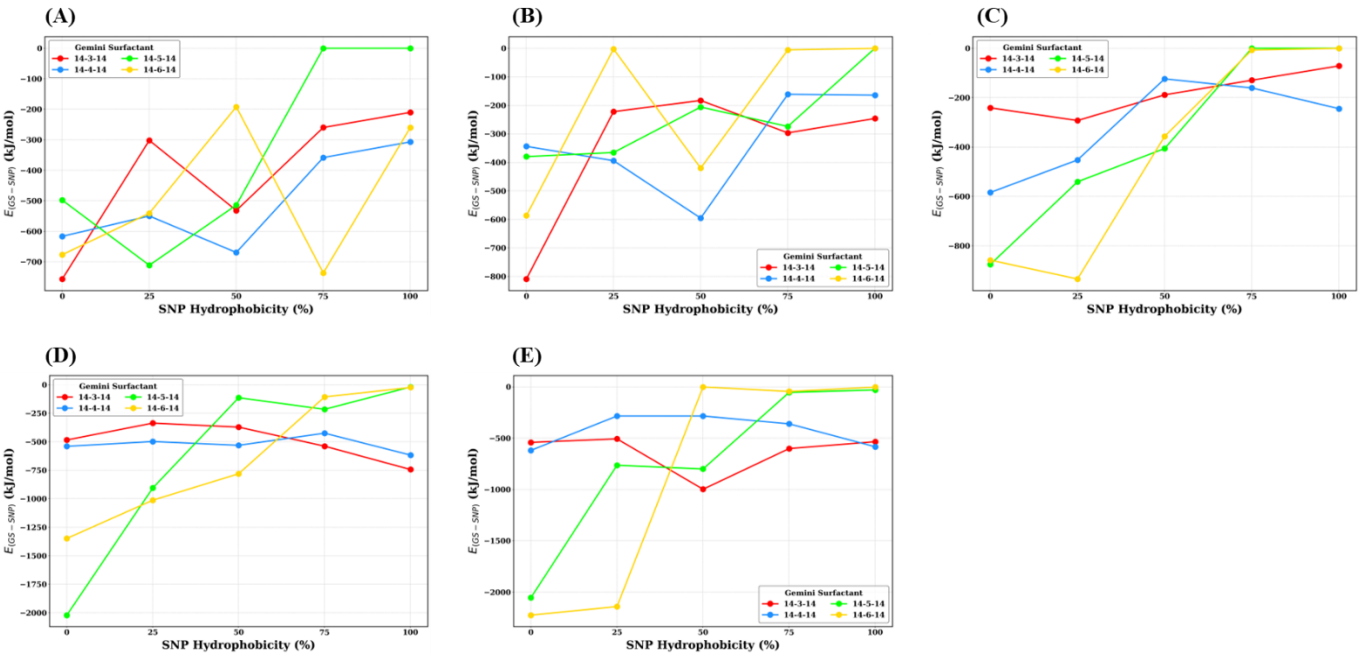

Figure S21: Plots for variation in GS-SNP interaction energy against SNP hydrophobicity. The hydrophobicity (in %) of silica slabs are : **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. Oil Model-1

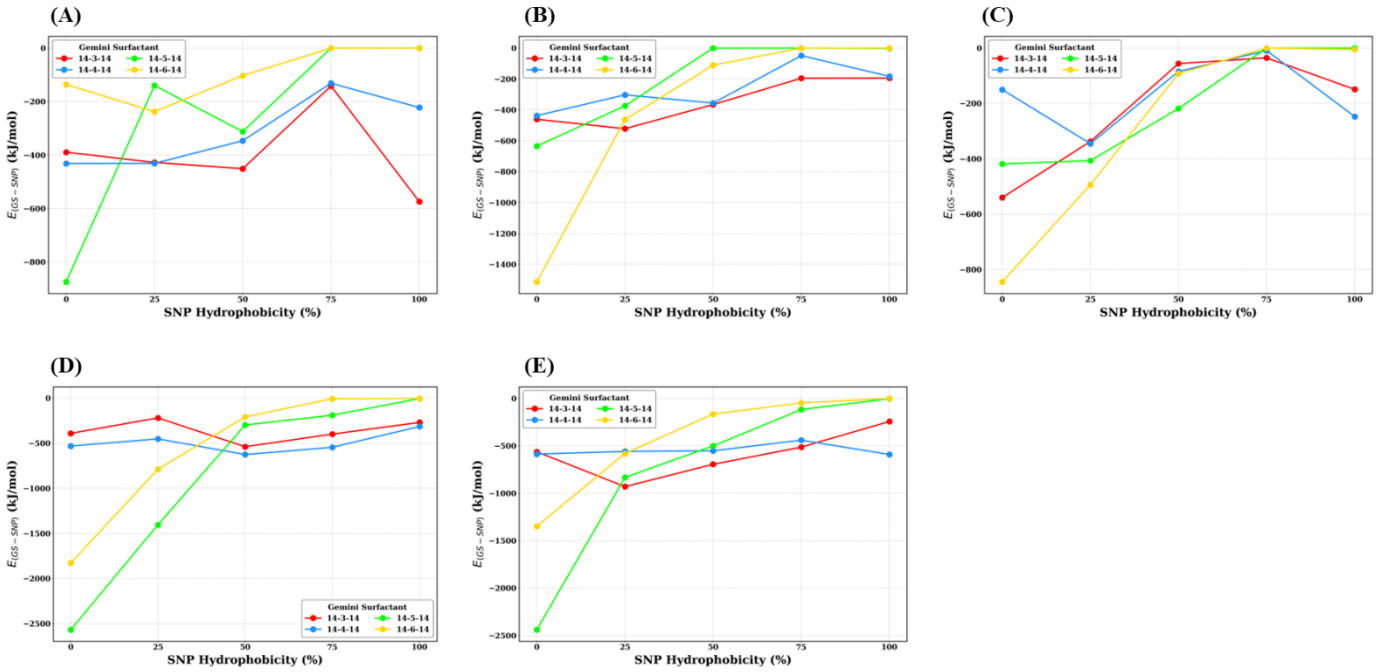

Figure S22: Plots for variation in GS-SNP interaction energy against SNP hydrophobicity. The hydrophobicity (in %) of silica slabs are : **(A)** 0, **(B)** 25, **(C)** 50, **(D)** 75, and **(E)** 100. Oil Model-2.

OIL MODEL/ML Model	ML Algorithm	Hyperparameter Grid	Best Parameters

Model-1, ML-1	Support Vector Regression	param_grid = {
		'C': [0.1, 1, 10, 100],
		'epsilon': [0.01, 0.1, 1],
		'kernel': ['linear', 'poly', 'rbf']
		}
			{'C': 100, 'epsilon': 1, 'kernel': 'linear'}

	Ridge Regression	param_grid = {'alpha': np.logspace(-4, 4, 50)}	{'alpha': 0.8286427728546842}

	XGBoost	param_grid = {
		'n_estimators': [50, 100, 200],
		'max_depth': [3, 5, 7],
		'learning_rate': [0.01, 0.1, 0.3],
		'subsample': [0.7, 0.8, 0.9]
		}
			{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200, 'subsample': 0.7}

	Random Forest	param_grid = {
		'n_estimators': [100, 200, 300],
		'max_depth': [None, 10, 20,50],
		'min_samples_split': [2, 5, 10],
		'min_samples_leaf': [2,4,6,8,10,15],
		'max_features': ['auto', 'sqrt', 'log2'],
		'bootstrap': [True, False]
		}
			{'bootstrap': False, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300}

	AdaBoost	param_grid = {
		'n_estimators': [50, 100, 200, 300],
		'learning_rate': [0.01, 0.1, 1.0, 10.0]
		}
			{'learning_rate': 0.1, 'n_estimators': 100}

Model-1, ML-2	Support Vector Regression	param_grid = {
		'C': [0.1, 1, 10, 100],
		'epsilon': [0.01, 0.1, 1],
		'kernel': ['linear', 'poly', 'rbf']
		}
			{'C': 100, 'epsilon': 1, 'kernel': 'linear'}

	Ridge Regression	param_grid = {'alpha': np.logspace(-4, 4, 50)}	{'alpha': 0.009102981779915217}

	XGBoost	param_grid = {
		'n_estimators': [50, 100, 200],
		'max_depth': [3, 5, 7],
		'learning_rate': [0.01, 0.1, 0.3],
		'subsample': [0.7, 0.8, 0.9]
		}
			{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.7}

	Random Forest	param_grid = {
		'n_estimators': [100, 200, 300],
		'max_depth': [None, 10, 20,50],
		'min_samples_split': [2, 5, 10],
		'min_samples_leaf': [2,4,6,8,10,15],
		'max_features': ['auto', 'sqrt', 'log2'],
		'bootstrap': [True, False]
		}
			{'bootstrap': True, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}

	AdaBoost	param_grid = {
		'n_estimators': [50, 100, 200, 300],
		'learning_rate': [0.01, 0.1, 1.0, 10.0]
		}
			{'learning_rate': 0.01, 'n_estimators': 50}

Model-2, ML-1	Support Vector Regression	param_grid = {
		'C': [0.1, 1, 10, 100],
		'epsilon': [0.01, 0.1, 1],
		'kernel': ['linear', 'poly', 'rbf']
		}
			{'C': 100, 'epsilon': 0.1, 'kernel': 'linear'}

	Ridge Regression	param_grid = {'alpha': np.logspace(-4, 4, 50)}	{'alpha': 0.08685113737513521}

	XGBoost	param_grid = {
		'n_estimators': [50, 100, 200],
		'max_depth': [3, 5, 7],
		'learning_rate': [0.01, 0.1, 0.3],
		'subsample': [0.7, 0.8, 0.9]
		}
			{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.9}

	Random Forest	param_grid = {
		'n_estimators': [100, 200, 300],
		'max_depth': [None, 10, 20,50],
		'min_samples_split': [2, 5, 10],
		'min_samples_leaf': [2,4,6,8,10,15],
		'max_features': ['auto', 'sqrt', 'log2'],
		'bootstrap': [True, False]
		}
			{'bootstrap': False, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}

	AdaBoost	param_grid = {
		'n_estimators': [50, 100, 200, 300],
		'learning_rate': [0.01, 0.1, 1.0, 10.0]
		}
			{'learning_rate': 1.0, 'n_estimators': 100}

	Support Vector Regression	param_grid = {	{'C': 100, 'epsilon': 0.1, 'kernel': 'linear'}

Model-2, ML-2		'C': [0.1, 1, 10, 100], 'epsilon': [0.01, 0.1, 1], 'kernel': ['linear', 'poly', 'rbf'] }	
	Ridge Regression	param_grid = {'alpha': np.logspace(-4, 4, 50)}	{'alpha': 1.2067926406393288}

	XGBoost	param_grid = {
		'n_estimators': [50, 100, 200],
		'max_depth': [3, 5, 7],
		'learning_rate': [0.01, 0.1, 0.3],
		'subsample': [0.7, 0.8, 0.9]
		}
			{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 50, 'subsample': 0.8}

	Random Forest	param_grid = {
		'n_estimators': [100, 200, 300],
		'max_depth': [None, 10, 20,50],
		'min_samples_split': [2, 5, 10],
		'min_samples_leaf': [2,4,6,8,10,15],
		'max_features': ['auto', 'sqrt', 'log2'],
		'bootstrap': [True, False]
		}
			{'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 200}

	AdaBoost	param_grid = {
		'n_estimators': [50, 100, 200, 300],
		'learning_rate': [0.01, 0.1, 1.0, 10.0]
		}
			{'learning_rate': 0.01, 'n_estimators': 200}

Table S7: Hyperparameter grid along with their optimized values, as obtained for each Machine Learning (ML) model.

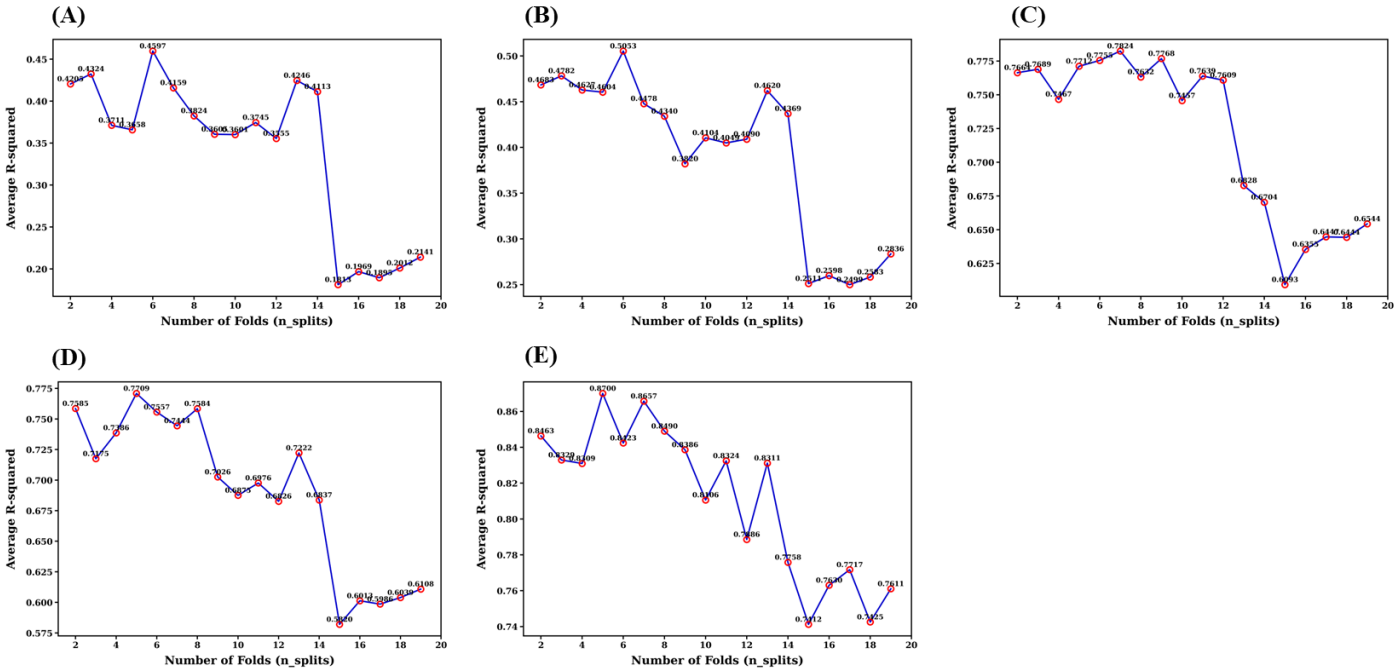

Figure S23: Plots for variation in R^2 vs n-splits during K-fold cross validation of ML-1 models of oil Model-1. The ML algorithms are: **(A)** Support Vector Regression (SVR), **(B)** Ridge regression, **(C)** Extreme Gradient Boosting (XGBoost), **(D)** Random Forest, and **(E)** Adaptive Boosting (Adaboost).

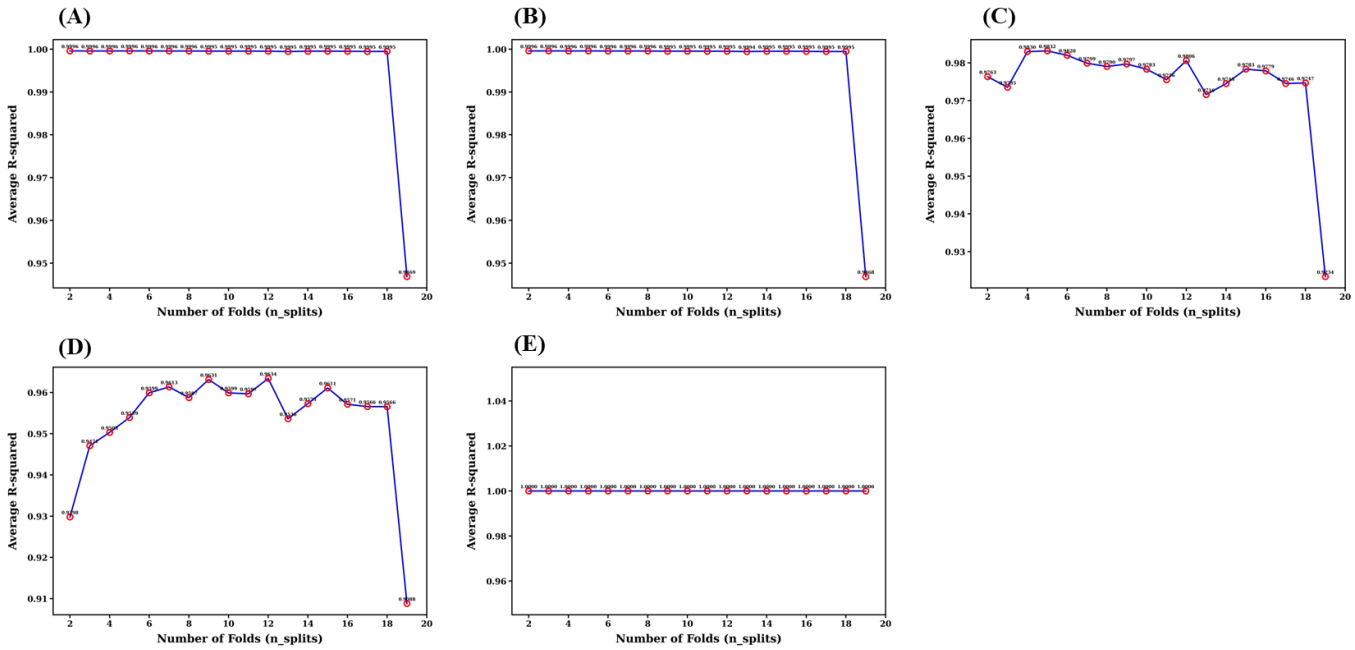

Figure S24: Plots for variation in R^2 vs n-splits during K-fold cross validation of ML-2 models of oil Model-1. The ML algorithms are: **(A)** Support Vector Regression (SVR), **(B)** Ridge regression, **(C)** Extreme Gradient Boosting (XGBoost), **(D)** Random Forest, and **(E)** Adaptive Boosting (Adaboost).

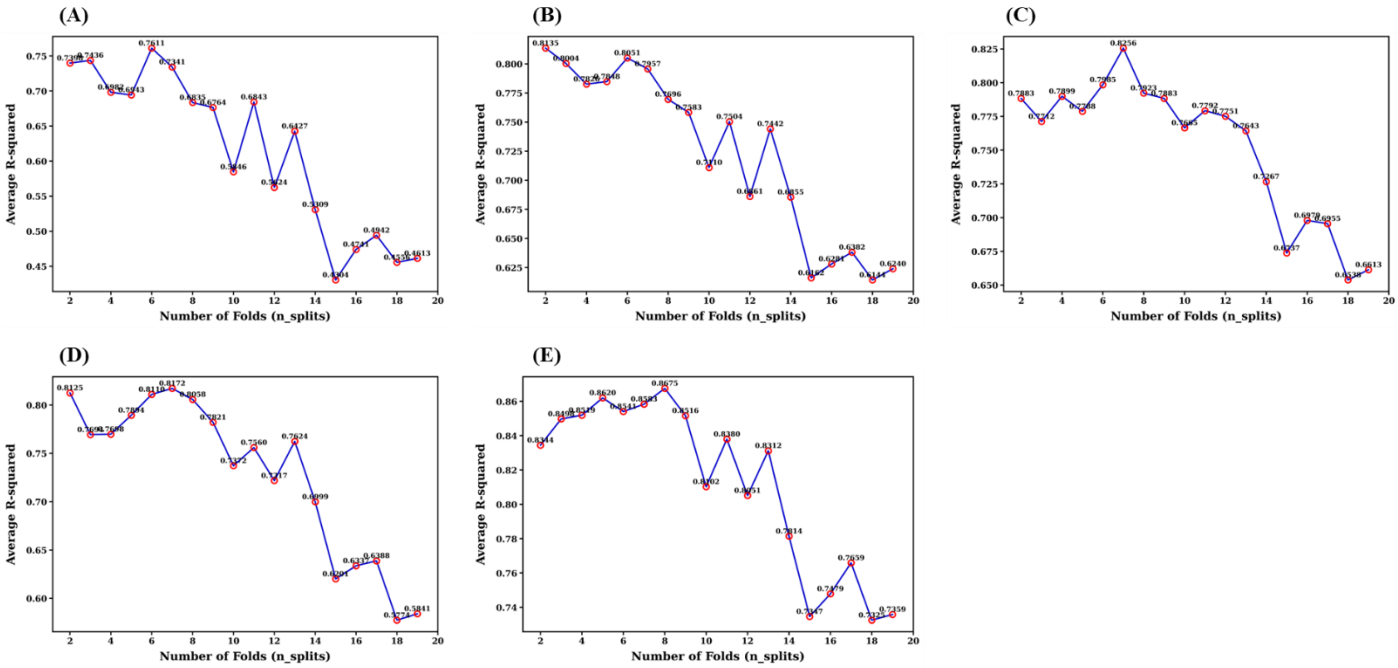

Figure S25: Plots for variation in R^2 vs n-splits during K-fold cross validation of ML-1 models of oil Model-2. The ML algorithms are: **(A)** Support Vector Regression (SVR), **(B)** Ridge regression, **(C)** Extreme Gradient Boosting (XGBoost), **(D)** Random Forest, and **(E)** Adaptive Boosting (Adaboost).

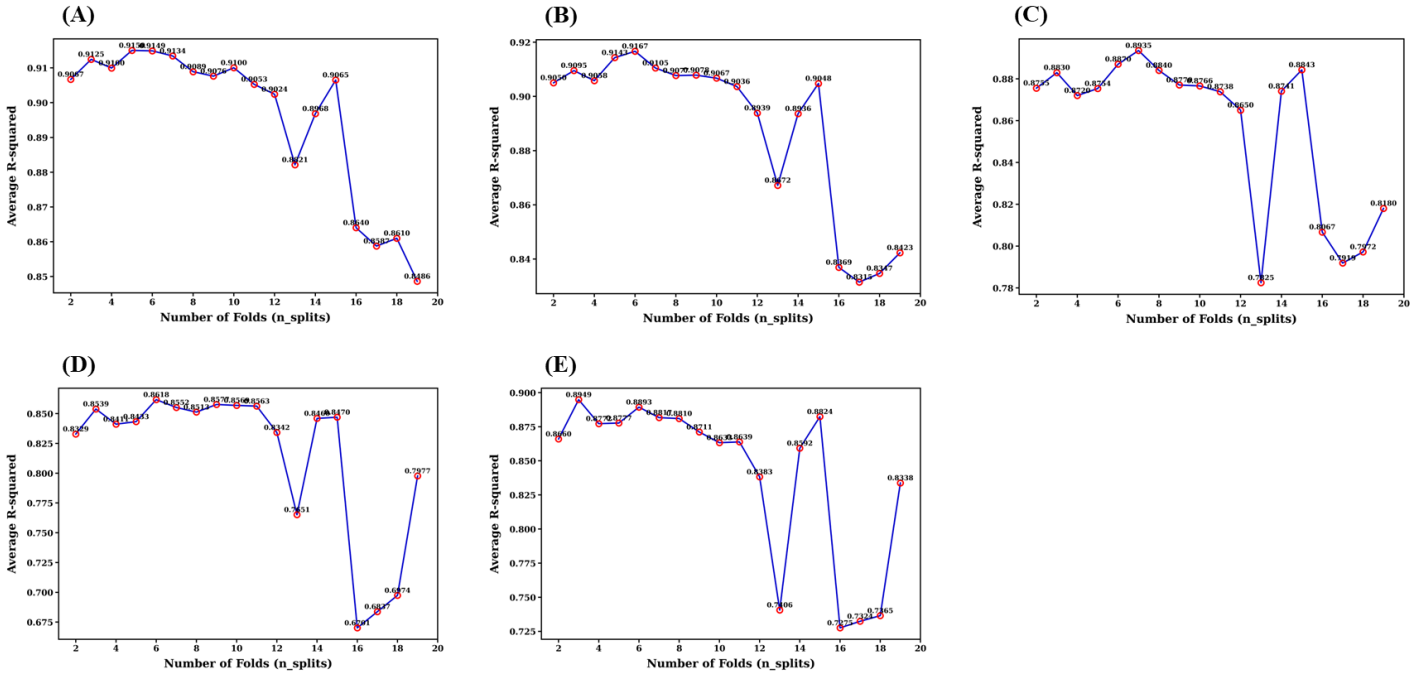

Figure S26: Plots for variation in R^2 vs n-splits during K-fold cross validation of ML-2 models of oil Model-2. The ML algorithms are: **(A)** Support Vector Regression (SVR), **(B)** Ridge regression, **(C)** Extreme Gradient Boosting (XGBoost), **(D)** Random Forest, and **(E)** Adaptive Boosting (Adaboost).

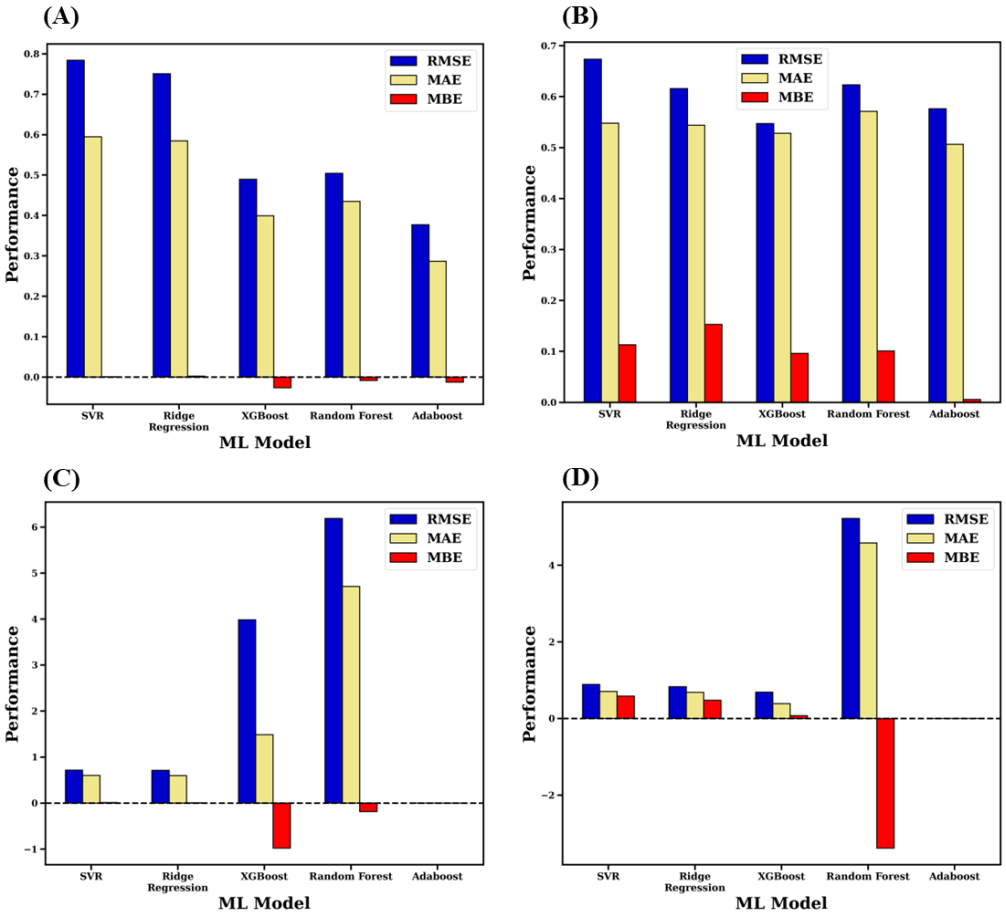

Figure S27: Error parameters associated for K-fold validation **(A, C)**, and test set predictions **(B, D)**. Here, **(A, B)**, and **(C, D)** represents ML-1 and ML-2 models respectively. The data is based on Oil Model-1.

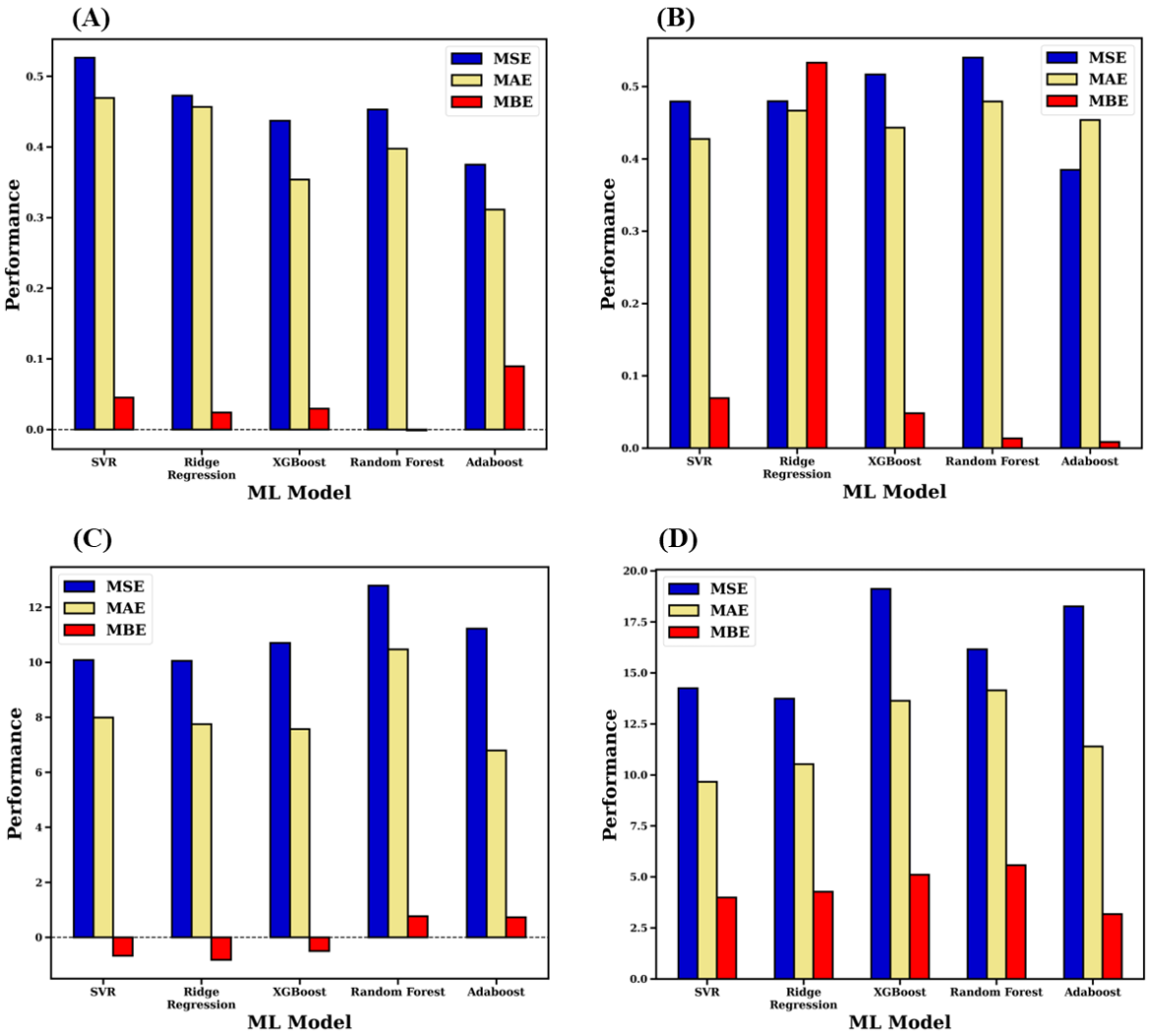

Figure S28: Error parameters associated for K-fold validation **(A, C)**, and test set predictions **(B, D)**. Here, **(A, B)**, and **(C, D)** represents ML-1 and ML-2 models respectively. The data is based on Oil Model-2.

References:

1. E. A. Muller, Å. Ervik and A. Mejía, A Guide to Computing Interfacial Properties of Fluids from Molecular Simulations [Article v1.0], *Living Journal of Computational Molecular Science*, 2020, **2**, 21385–21385.
2. H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, The missing term in effective pair potentials, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
3. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, Molecular dynamics with coupling to an external bath, *The Journal of Chemical Physics*, 1984, **81**, 3684–3690.