

Supporting Information

The First Frustrated Lewis Pairs Database: Machine Learning and Cheminformatics Aided Prediction of Small Molecule Activation

Jingyun Ye^{1, 2 ‡, *}, Chathura Wijethunga^{2, ‡}, Megan McEwen²

¹Department of Chemistry and Biochemistry, Duquesne University, Pittsburgh, Pennsylvania, 15282, USA

²Department of Chemistry and Biomolecular Science, Clarkson University, Potsdam, New York 13699, USA

Corresponding author: yej1@duq.edu

1. Features

d_{A-B} is the distance between LA and LB sites, which could affect the distortion energy of the substrate in the transition state and influence the catalyst-substrate geometric match, thus tuning the reaction mechanism.^{50,51} Therefore, to take advantage of the cooperative effect of the LA and the LB, designing FLPs with suitable LA-LB distances would ultimately enhance the FLP activity.

ΔG_{H^-} and ΔG_{H^+} are the hydride and proton attachment energy, respectively. They represent the acidity and basicity of FLPs, which are important factors that influence the FLP chemistry.^{46,52}

η and S are chemical hardness and chemical softness derived from hard/soft acid/base (HSAB) principle, which could be used to determine whether a reagent's reactivity is dominated by electron transfer or by electrostatic effects.⁵³ Electron-transfer effects favor soft/soft interactions, while electrostatic effects favor hard/hard interactions or acid-base exchange reactions. The SM activation is in the category of electron transfer reactions. $\eta \approx \frac{1}{2}(E_{\text{LUMO}} - E_{\text{HOMO}})$ and $S \approx 1/(E_{\text{LUMO}} - E_{\text{HOMO}})$, where E_{LUMO} and E_{HOMO} are the HOMO and LUMO energies calculated from DFT.⁵⁴⁻⁵⁶

χ and ω are the electronegativity and electrophilicity of a FLP molecule, respectively. χ is defined as $\chi \approx -\frac{1}{2}(E_{\text{LUMO}} + E_{\text{HOMO}})$, which measures the power of FLPs attracting electron, and ω is defined as $\omega \approx \chi^2/2\eta$, which measures the reactivity of FLPs toward attracting electrons from a nucleophile, so that they form a bond.⁵⁴⁻⁵⁶

E_g is the HOMO-LUMO gap energy, which is extensively used to analyze the activity of molecules, $E_g = E_{\text{LUMO}} - E_{\text{HOMO}}$.

E_{prep} is the preparation energy, which is extensively used to analyze the activity of molecules, $E_{\text{prep}} = E_{\text{FLP}'} - E_{\text{FLP}}$.

q_A and q_B are the atomic charge of LA and LB sites, respectively, which could affect the charge separation and transfer in the binding and reactions of the SMs.^{57,58}

f_A^+/f_B^- are Fukui functions of the LA or LB site of a FLP molecule, which governs the nucleo/electrophilic attacking an atom k in a molecule^{56,59} and allows us to probe the reactive sites LA and LB within a FLP; the maximum value of f_k is usually associated with the most reactive site.⁶⁰⁻⁶³ f_A^+/f_B^- are defined as $f_A^+ = q_A(N+1) - q_A(N)$ and (f_B^-) : $f_B^- = q_B(N) - q_B(N-1)$,

where $q_k(N)$, $q_k(N + 1)$ and $q_k(N-1)$ are the electronic population of atom k evaluated on the N, (N + 1) and (N-1) electron systems with the ground state geometry of the N-electron species.

Δf is the difference between nucleophilic and electrophilic Fukui function ($\Delta f_k = f_k^+ - f_k^-$), which provides useful information about the nucleo/electrophilic behavior of a specific site.^{64,65} The positive value of Δf indicates the site is favored for a nucleophilic attack, conversely, the negative value of Δf means the site is favored for an electrophilic attack.⁶³

ω_A^+/ω_B^- stands for the local electrophilicity of LA and LB site, respectively, which is a DFT-based reactivity descriptor, defined as ($\omega_k^+ = \omega f_k^+$ or $\omega_k^- = \omega f_k^-$).

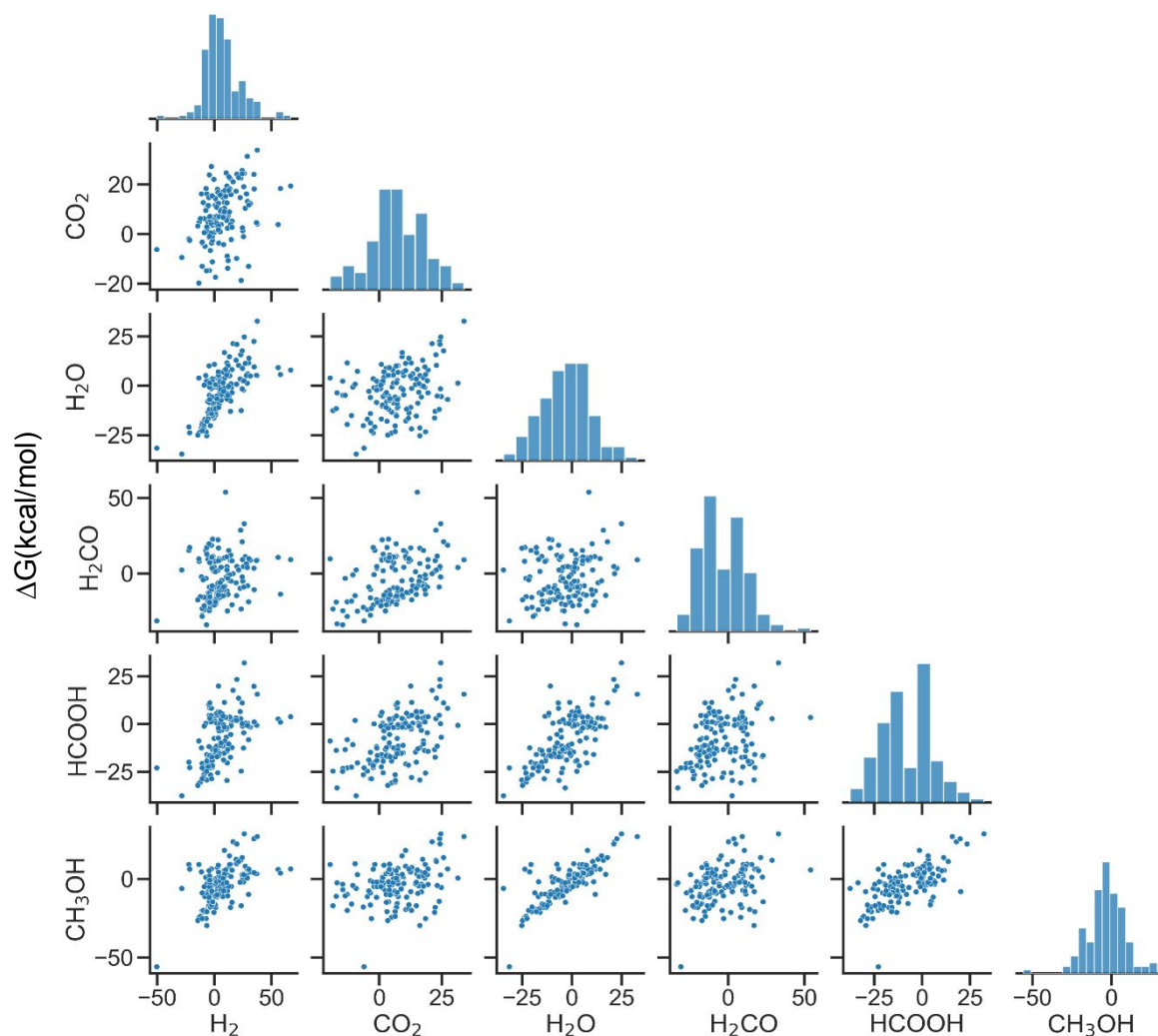


Figure S1. The binding free energies of one small molecule against another one at FLPs.

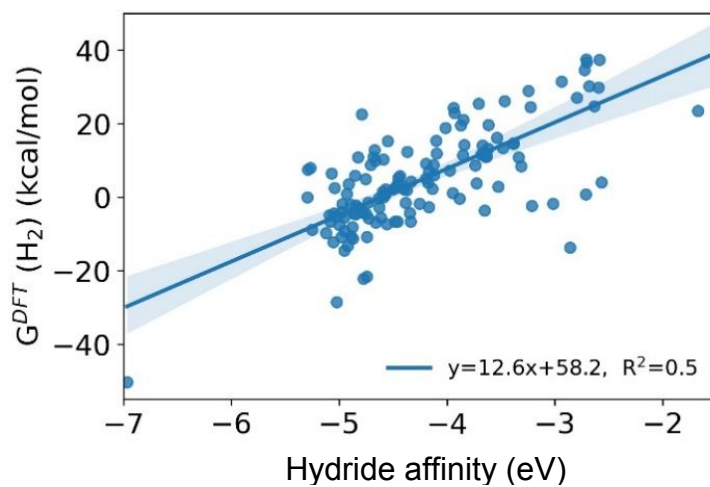


Figure S2. The binding free energies of one small molecule at FLPs against the hydride affinity.

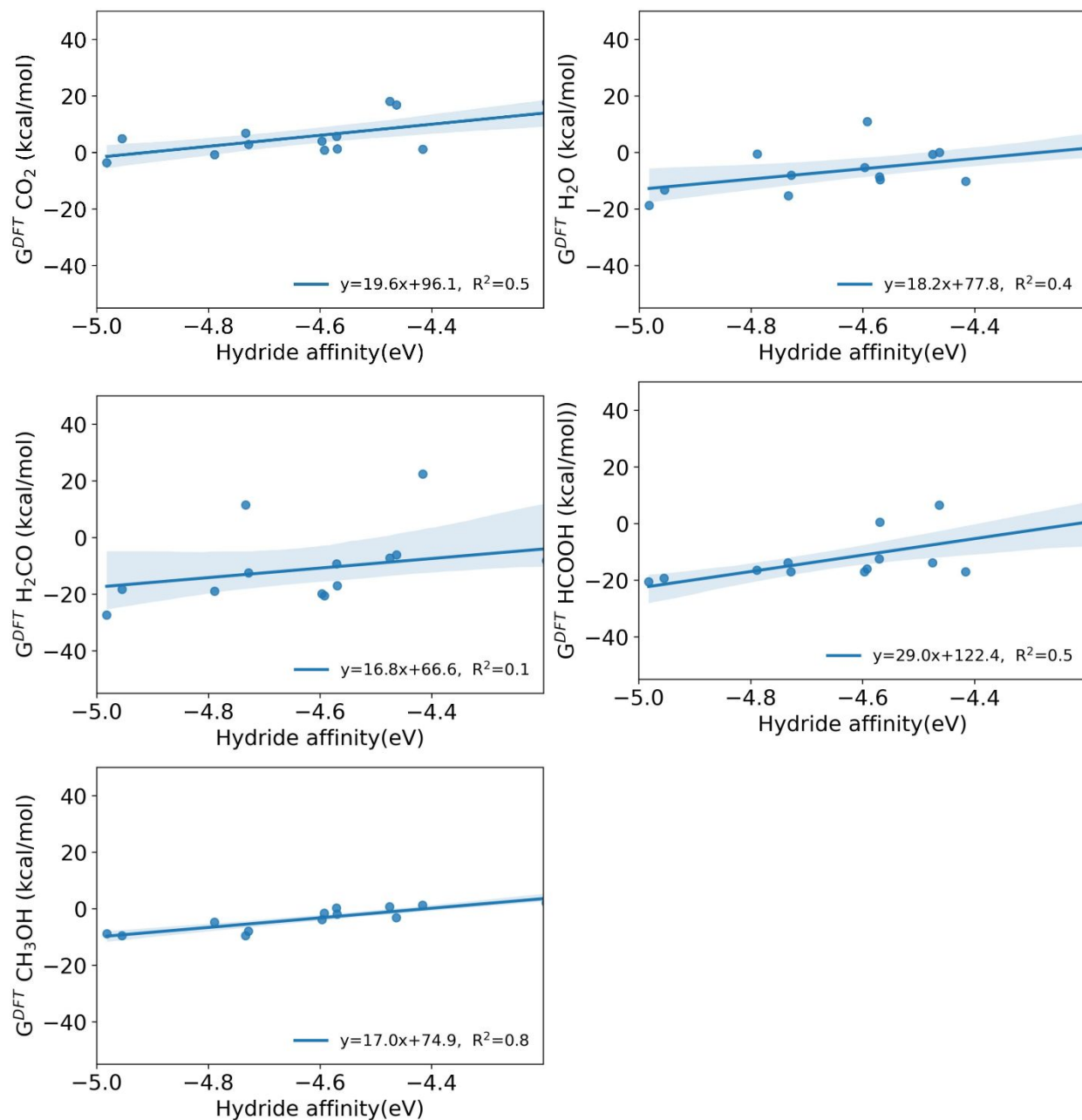


Figure S3. The binding free energies of CO₂, H₂O, H₂CO, HCOOH, and CH₃OH as a function of hydride affinity for FLPs with similarity score ≥ 0.55 compared to **0BP07011**.

Table S1. Best hyper-parameters via GridSearchCV with five-fold for H₂ binding free energies prediction with different ML models, and the root mean squared error (rmse) and R².

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	4.261	0.855	6.474	0.786	{'alpha': 1}
Lasso	4.548	0.835	6.286	0.798	{'alpha': 0.1}
KernelRidge	1.070	0.991	7.232	0.732	{'alpha': 0.1, 'kernel': 'poly'}
DecisionTree	0.030	1.000	11.071	0.373	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5}
RandomForest	2.728	0.941	9.556	0.533	{'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 10}
GradientBoosting	0.000	1.000	10.369	0.450	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50}
NN	0.257	0.999	6.199	0.803	

Table S2. Best hyper-parameters via GridSearchCV with five-fold for CO₂ binding free energies prediction with different ML models, and the root mean squared error (rmse) and R².

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	5.804	0.683	8.063	0.335	{'alpha': 1}
Lasso	5.563	0.709	8.821	0.204	{'alpha': 0.001}
KernelRidge	3.063	0.912	7.595	0.410	{'alpha': 0.1, 'kernel': 'rbf'}
DecisionTree	5.361	0.729	8.773	0.212	{'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 10}
RandomForest	3.503	0.884	7.754	0.385	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 10}
GradientBoosting	0.000	1.000	7.857	0.368	{'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 200}
NN	0.228	0.999	9.040	0.163	

Table S3. Best hyper-parameters via GridSearchCV with five-fold for H₂O binding free energies prediction with different ML models, and the root mean squared error (rmse) and R².

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	6.146	0.730	7.425	0.674	{'alpha': 1}
Lasso	6.320	0.715	7.697	0.650	{'alpha': 0.1}
KernelRidge	3.096	0.932	8.582	0.565	{'alpha': 0.1, 'kernel': 'rbf'}
DecisionTree	2.971	0.937	11.423	0.229	{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 2}
RandomForest	3.655	0.905	9.770	0.436	{'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
GradientBoosting	0.014	1.000	10.362	0.366	{'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 50}
NN	0.167	1.000	7.999	0.622	

Table S4. Best hyper-parameters via GridSearchCV with five-fold for H₂CO binding free energies prediction with different ML models, and the root mean squared error (rmse) and R².

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	11.105	0.387	16.800	-0.122	{'alpha': 1}
Lasso	13.306	0.120	16.305	-0.057	{'alpha': 1}
KernelRidge	10.886	0.411	15.951	-0.011	{'alpha': 1, 'kernel': 'rbf'}
DecisionTree	9.217	0.578	21.051	-0.761	{'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5}
RandomForest	6.134	0.813	17.43	-0.207	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}
GradientBoosting	3.735	0.931	17.012	-0.150	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200}
NN	0.316	1.000	18.069	-0.298	

Table S5. Best hyper-parameters via GridSearchCV with five-fold for HCOOH binding free energies prediction with different ML models, and the root mean squared error (rmse) and R^2 .

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	10.424	0.433	9.281	0.509	{'alpha': 1}
Lasso	12.218	0.221	11.159	0.290	{'alpha': 1}
KernelRidge	5.516	0.841	11.181	0.287	{'alpha': 0.1, 'kernel': 'rbf'}
DecisionTree	7.986	0.667	16.668	-0.584	{'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 10}
RandomForest	7.228	0.727	10.988	0.312	{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 10}
GradientBoosting	9.016	0.576	13.211	0.005	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50}
NN	0.357	0.999	13.997	-0.117	

Table S6. Best hyper-parameters via GridSearchCV with five-fold for CH₃OH binding free energies prediction with different ML models, and the root mean squared error (rmse) and R^2 .

Model	Train_rmse	Train_r2	Test_rmse	Test_r2	Best_hyper-parameters
Ridge	5.386	0.722	9.175	0.601	{'alpha': 0.1}
Lasso	5.767	0.681	8.285	0.675	{'alpha': 0.1}
KernelRidge	1.977	0.963	10.807	0.447	{'alpha': 0.05, 'kernel': 'rbf'}
DecisionTree	4.624	0.795	12.499	0.260	{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
RandomForest	2.470	0.941	10.663	0.461	{'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
GradientBoosting	0.012	1.000	12.011	0.316	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
NN	0.193	1.000	8.986	0.617	

Table S7. Regularization test for H₂ binding free energies prediction with MLP model, and the root mean squared error (rmse) and R^2 .

Alpha	r2_train	rmse_train	r2_test	rmse_test
0.0001	0.999473	0.257115	0.803443	6.199311
0.1	0.999827	0.14748	0.810375	6.089021