Supporting Information for Reorganization energies of flexible organic molecules as a challenging test for machine learning models

Ke Chen,[1,2] Christian Kunkel,[1,2] Karsten Reuter,[1,2] and Johannes T. Margraf[1,2,a)]
[1)] *Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany*
[2)] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany*

Molecular transformation	SMARTS pattern
Biphenyl addition	[cH:1]≫[c:1](-c1ccccc1)
6-ring annelation	[cH:1][cH:2]≫[c:1]2C=CC=C[c:2]2
5-ring annelation	[cH:1][cH:2]≫[c:1]2C=CC[c:2]2
Ring contraction	[r6:1]1[r6:2][cH:3][cH:4][r6:5][r6:6]≫[C:1]1=[C:2][CH2:3][C:5]=[C:6]1.[C:4]
Linkage doublebond	[cH:1]≫[c:1](-C=C-c1ccccc1)
Linkage triplebond	[cH:1]≫[c:1](-C#C-c1ccccc1)
Linkage diphenylethene	[cH:1]≫[c:1](-C(=C)-c1ccccc1)
CH2-substitution	[r5:1][CH2r5:2][r5:3]≫[r5:1][C:2](=C)[r5:3]
CH2-functionalization	[r:1][CH2:2][r:3]≫[r:1][CH:2](-c1ccccc1[r:3])

TABLE I: SMARTs patterns for the molecular transformation operations depicted in Figure S1.

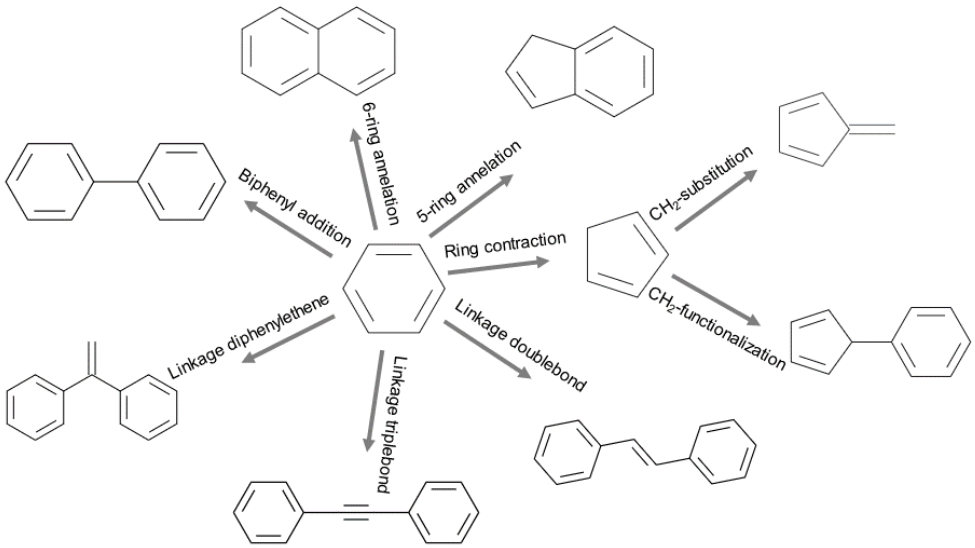

FIG. S1: Illustration of the molecular transformation operations used during the construction of the molecular dataset.

[a)] Electronic mail: Margraf@fhi-berlin.mpg.de

Symbol	Electronic properties
ϵ^{H}_{Nn}	HOMO energy of neutral state with neutral geometry
ϵ^{H}_{Nc}	HOMO energy of cationic state with neutral geometry
ϵ^{H}_{Cn}	HOMO energy of neutral state with cationic geometry
ϵ^{H}_{Cc}	HOMO energy of cationic state with cationic geometry
ϵ^{L}_{Nn}	LUMO energy of neutral state with neutral geometry
ϵ^{L}_{Nc}	LUMO energy of cationic state with neutral geometry
ϵ^{L}_{Cn}	LUMO energy of neutral state with cationic geometry
ϵ^{L}_{Cc}	LUMO energy of cationic state with cationic geometry
ϵ^{G}_{Nn}	HL GAP of neutral state with cationic geometry
ϵ^{G}_{Nc}	HL GAP of cationic state with neutral geometry
ϵ^{G}_{Cn}	HL GAP of neutral state with cationic geometry
ϵ^{G}_{Cc}	HL GAP of cationic state with cationic geometry
ϵ^{F}_{Nn}	Fermi energy of neutral state with neutral geometry
ϵ^{F}_{Nc}	Fermi energy of cationic state with neutral geometry
ϵ^{F}_{Cn}	Fermi energy of neutral state with cationic geometry
ϵ^{F}_{Cc}	Fermi energy of cationic state with cationic geometry
E^{E}_{Nn}	Energy of neutral state with neutral geometry
E^{E}_{Nc}	Energy of cationic state with neutral geometry
E^{E}_{Cn}	Energy of neutral state with cationic geometry
E^{E}_{Cc}	Energy of cationic state with cationic geometry
λ_1	Vertical energy difference $(E_0(R_+) - E_0(R_0))$
λ_2	Vertical energy difference $(E_+(R_0) - E_+(R_+))$

TABLE II: Electronic properties.

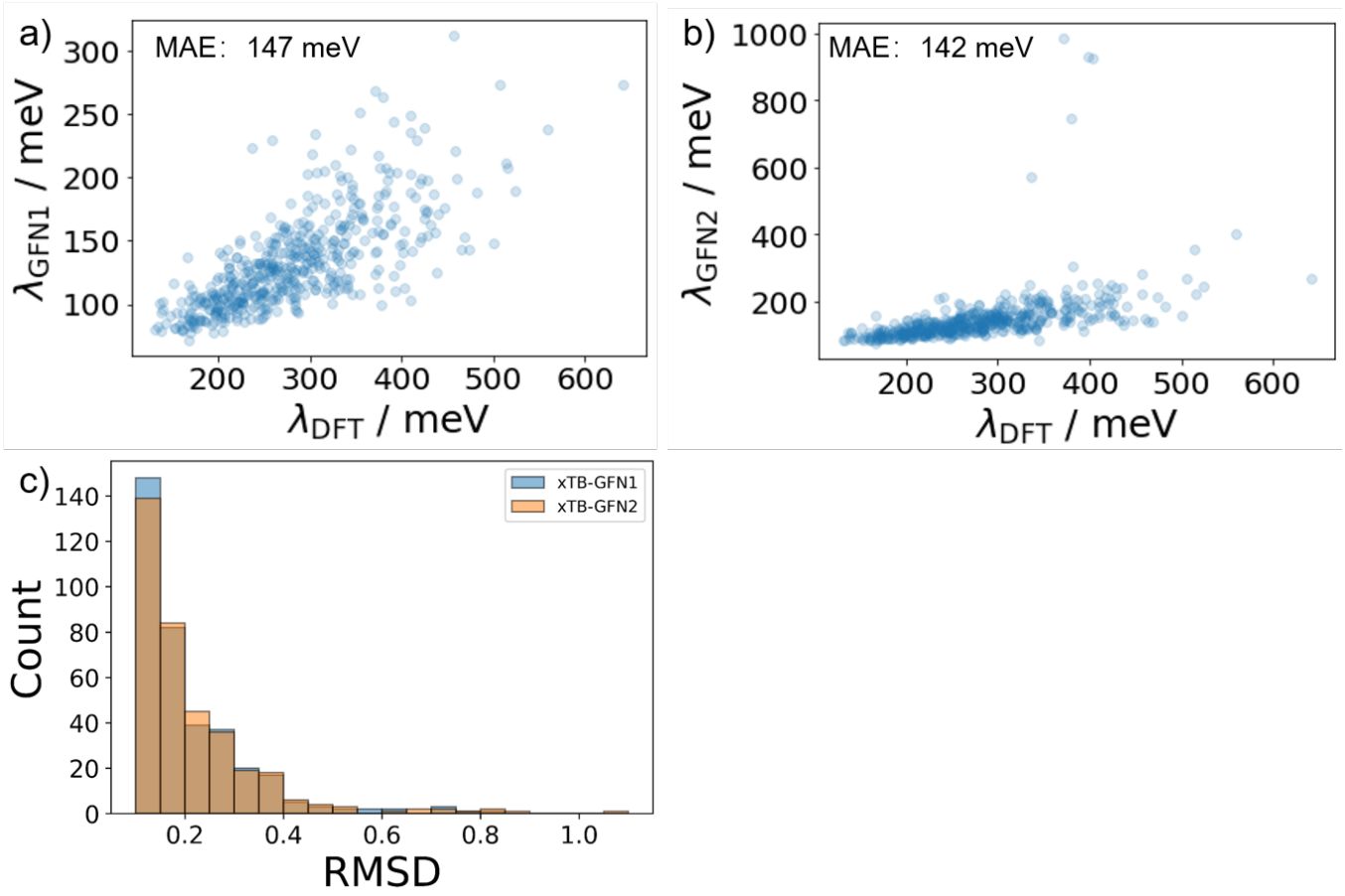

FIG. S2: Comparison of λ predicted with GFN1-xTB a) and GFN2-xTB b) relative to the DFT reference. c) Histogram of root mean squared deviations (RMSD) of GFN1/GFN2-xTB geometries relative to the DFT reference. Both plots are based on 500 randomly selected molecules from the DFT set.

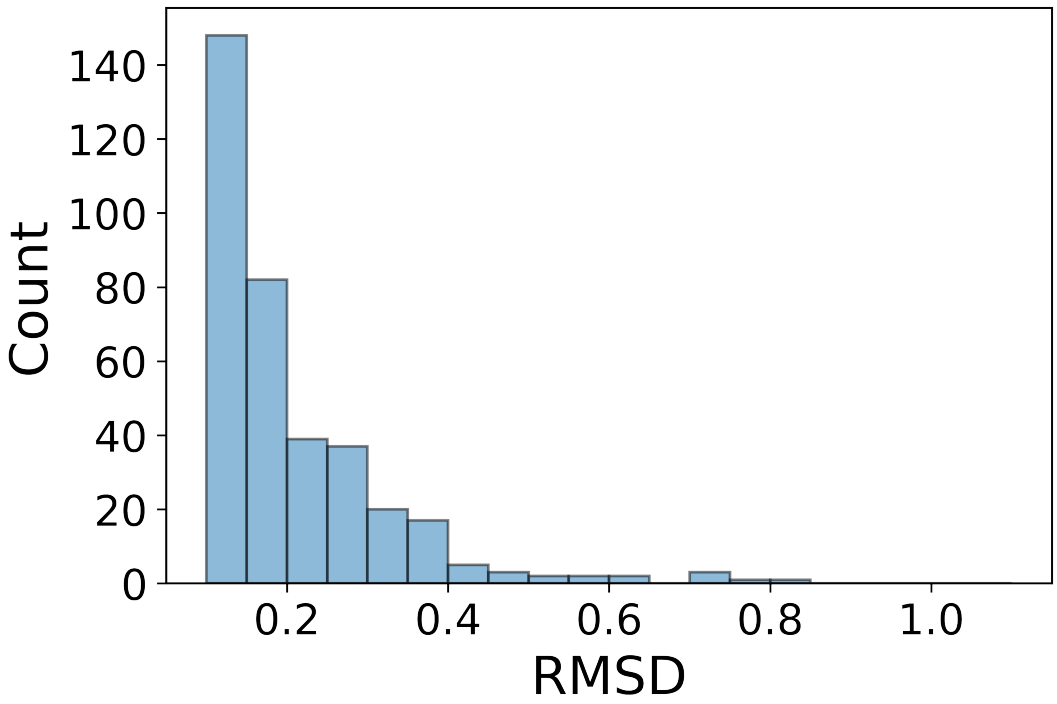

FIG. S3: Histogram of root mean squared deviations (RMSD) of GFN1-xTB geometries relative to the DFT reference with 500 randomly selected molecules from the DFT set.

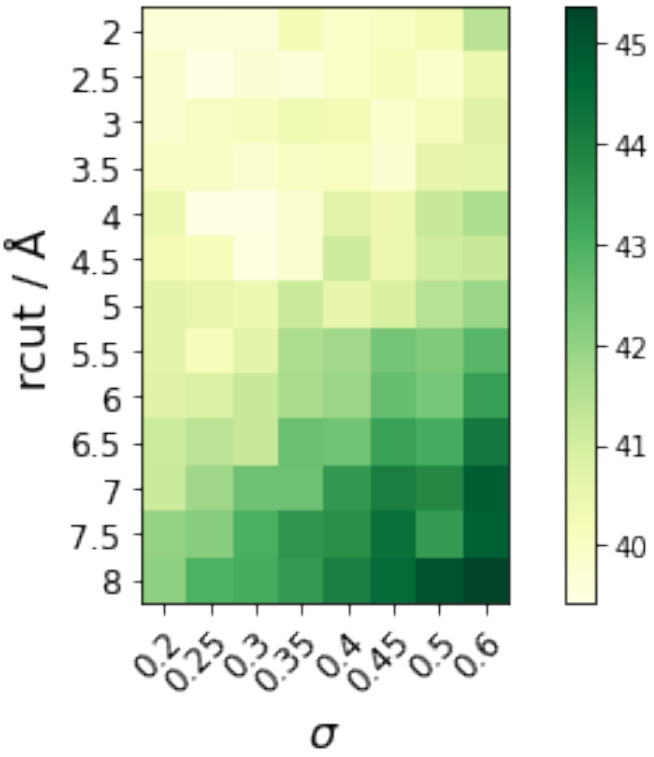

FIG. S4: Hyperparameter optimization for the smooth overlap of atomic orbitals (SOAP) based atomic environment descriptors. We rely on the respective implementation in the DSCRIBE package, using the default gaussian type orbital basis and with $n_{max} = 6$ and $l_{max} = 8$. Based on these converged settings, we only optimize the cutoff r_{cut} and gaussian width σ by grid search with training and test sets of 3000 and 1000 molecules, respectively. Results are averaged over three training sets of random composition. The average MAE for λ_{DFT} (in meV) is shown here for all tested combinations, arriving at a final combination of $r_{cut} = 3.5$ and $\sigma = 0.35$.

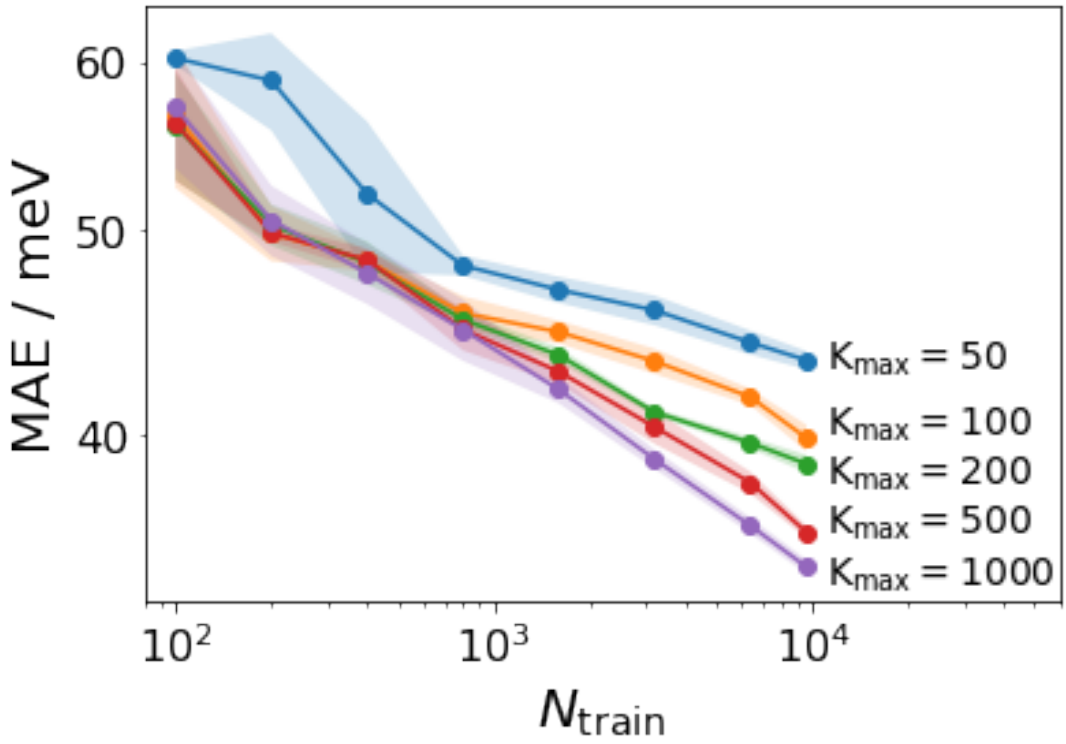

FIG. S5: The learning curve for K_{max} test. The global descriptors are built by autobag method with different K_{max} from geometry-based SOAP local vectors. For each K_{max} test, five GPR models are built with different training set from our DFT data, 1000 data,which are not include in training set, are used as test set. The average MAE (solid line) with variance (shown as shade) is shown here. The shades present prediction errors of the mean as measured over 5 models.

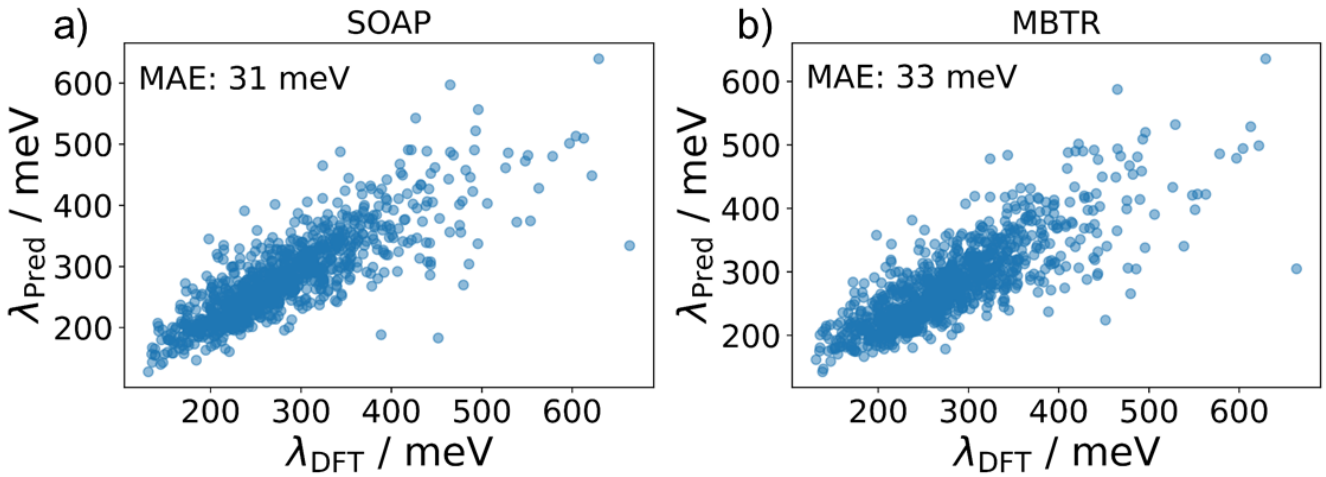

FIG. S6: Correlation plots for GPR predicted λ, using a) the SOAP+autobag representation described in the main manuscript and b) the Many-Body Tensor Representation (MBTR). Both models use structural information only and are performed in the Δ-learning setting with 9600 training points. Shown are the results for 1000 test molecules.

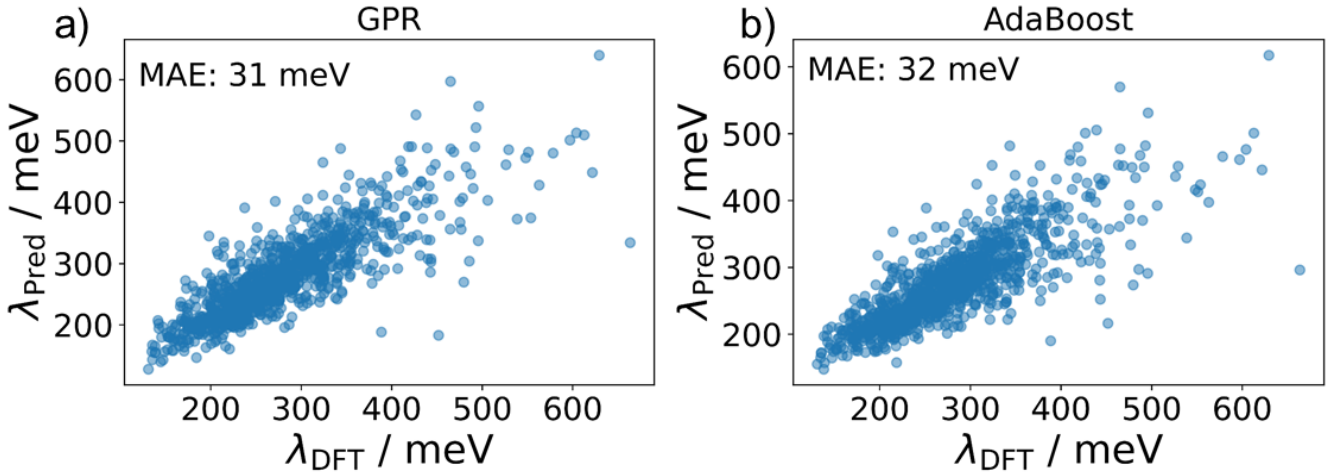

FIG. S7: Correlation plots for predicted λ using the SOAP+autobag representation and a) GPR as described in the main manuscript and b) the AdaBoost regressor (max_depth=100, n_estimators=800). Both models use structural information only and are performed in the Δ-learning setting with 9600 training points. Shown are the results for 1000 test molecules.

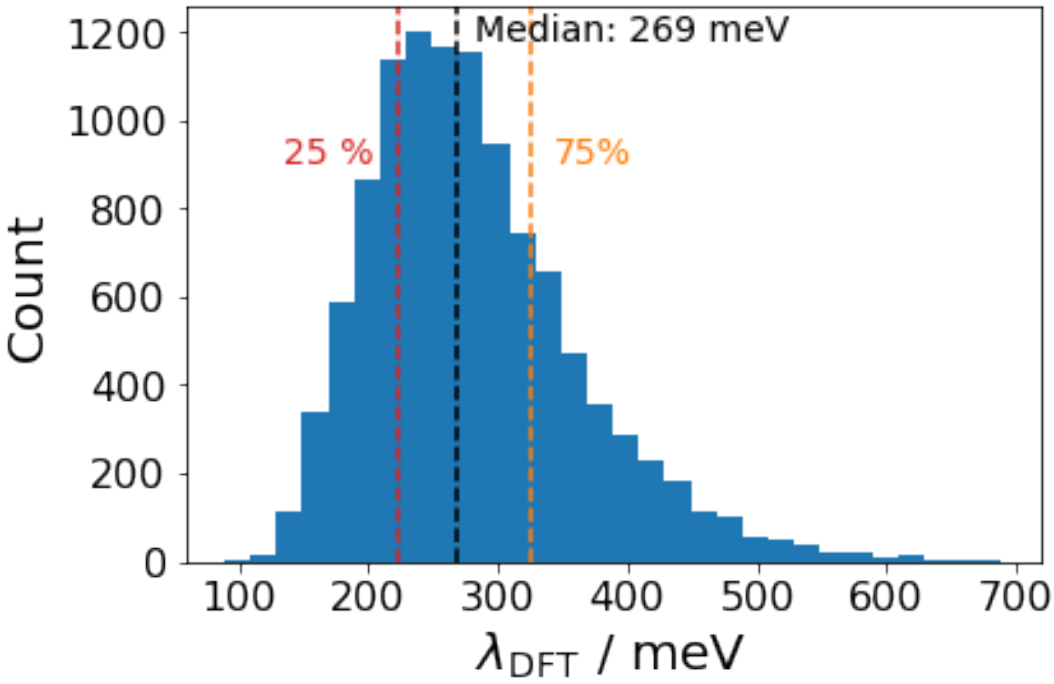

FIG. S8: Distribution of λ_{DFT} for our choosen subset of 10.900 molecules.

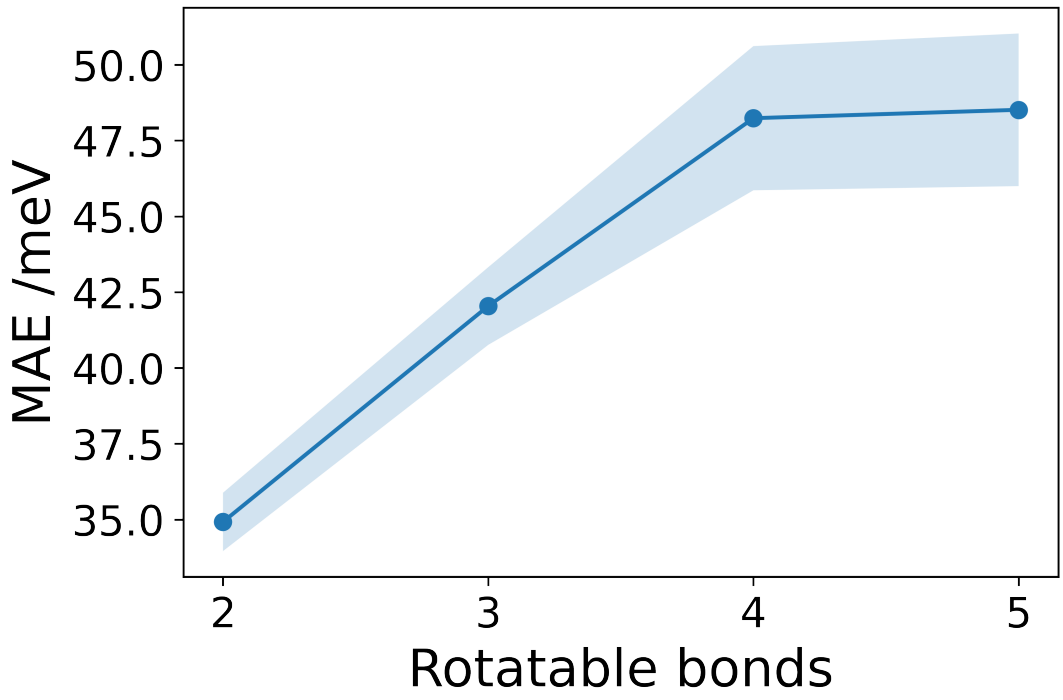

FIG. S9: The performance of ΔK_s models for subsets of the database containing molecules with a fixed number of rotatable bonds (training/test:1000/500).

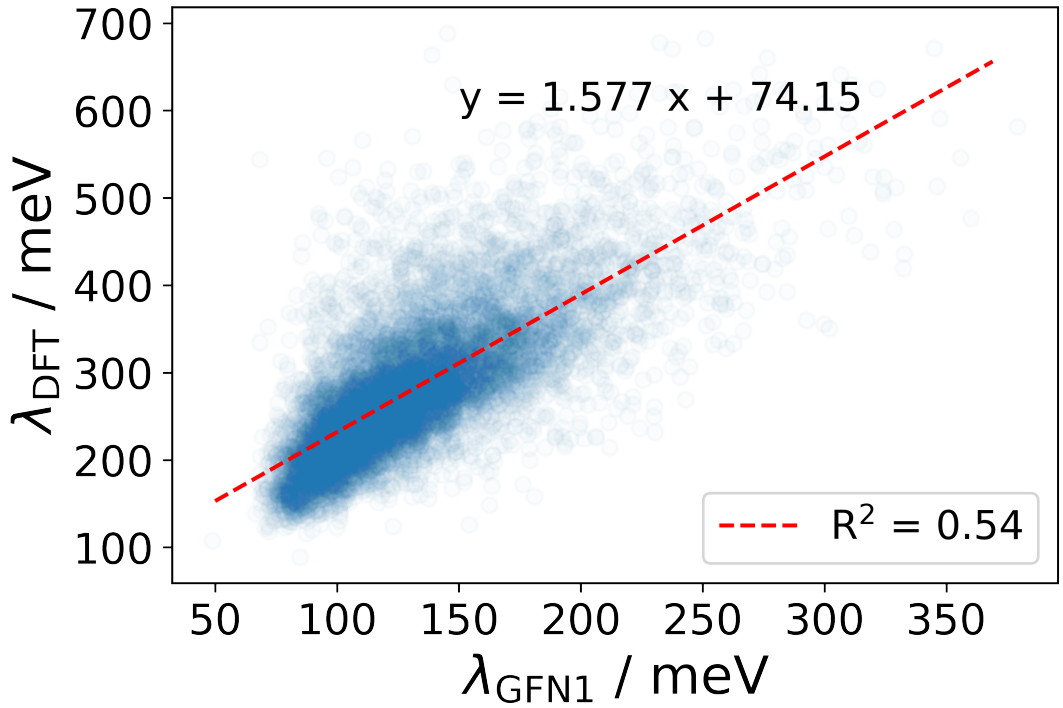

FIG. S10: Correlation between GFN1-xTB and B3LYP based reorganization energies.

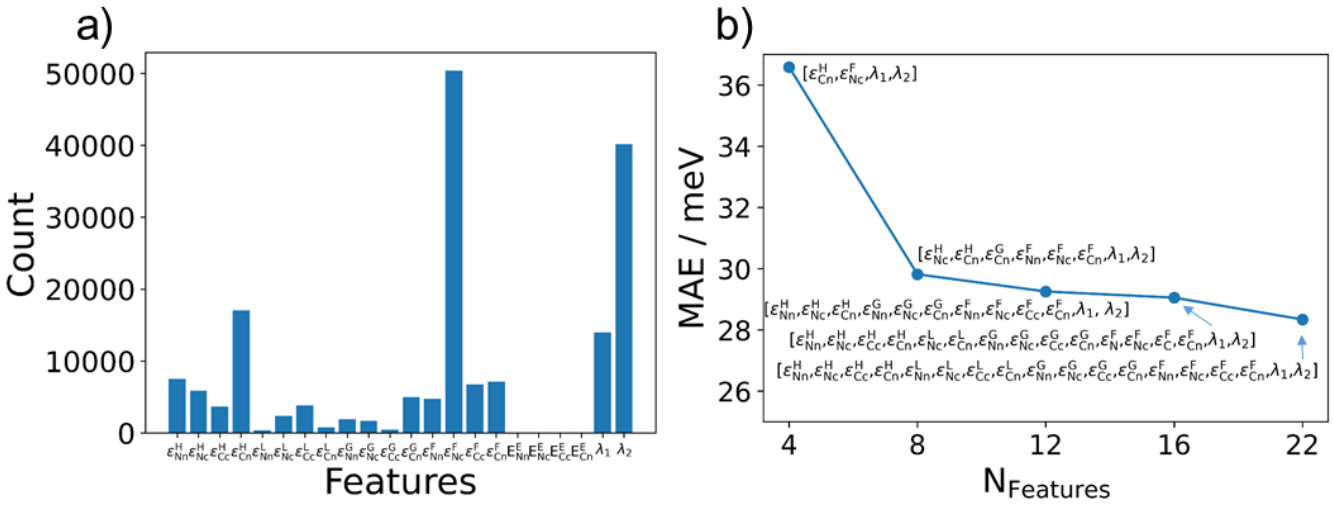

FIG. S11: a) Permutational feature importance for a ΔK_p model trained on 10,900 datapoints. b)The Prediction performance of selected features for ΔK_p (training/test = 9600/1000).

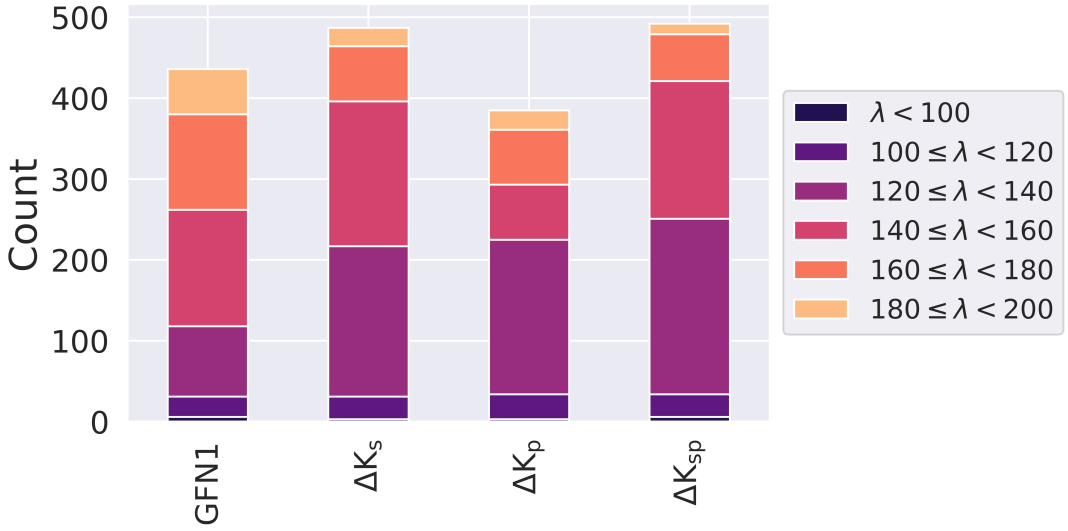

FIG. S12: Analogous to Fig. 5a in the main manuscript, but including the ΔK_p model.

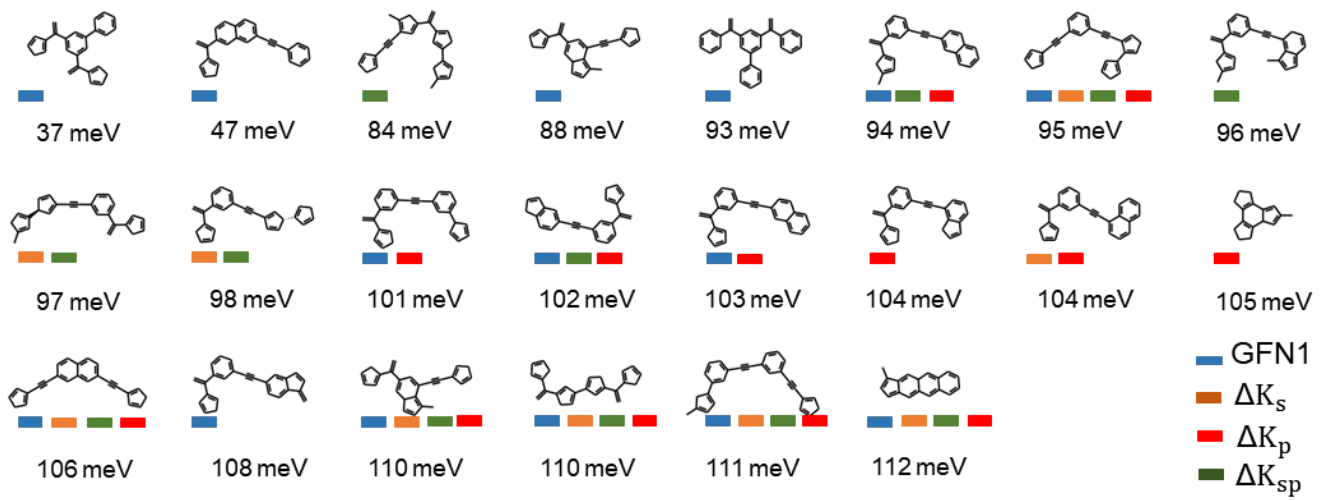

FIG. S13: Analogous to Fig. 6 in the main manuscript, but including the ΔK_p model.

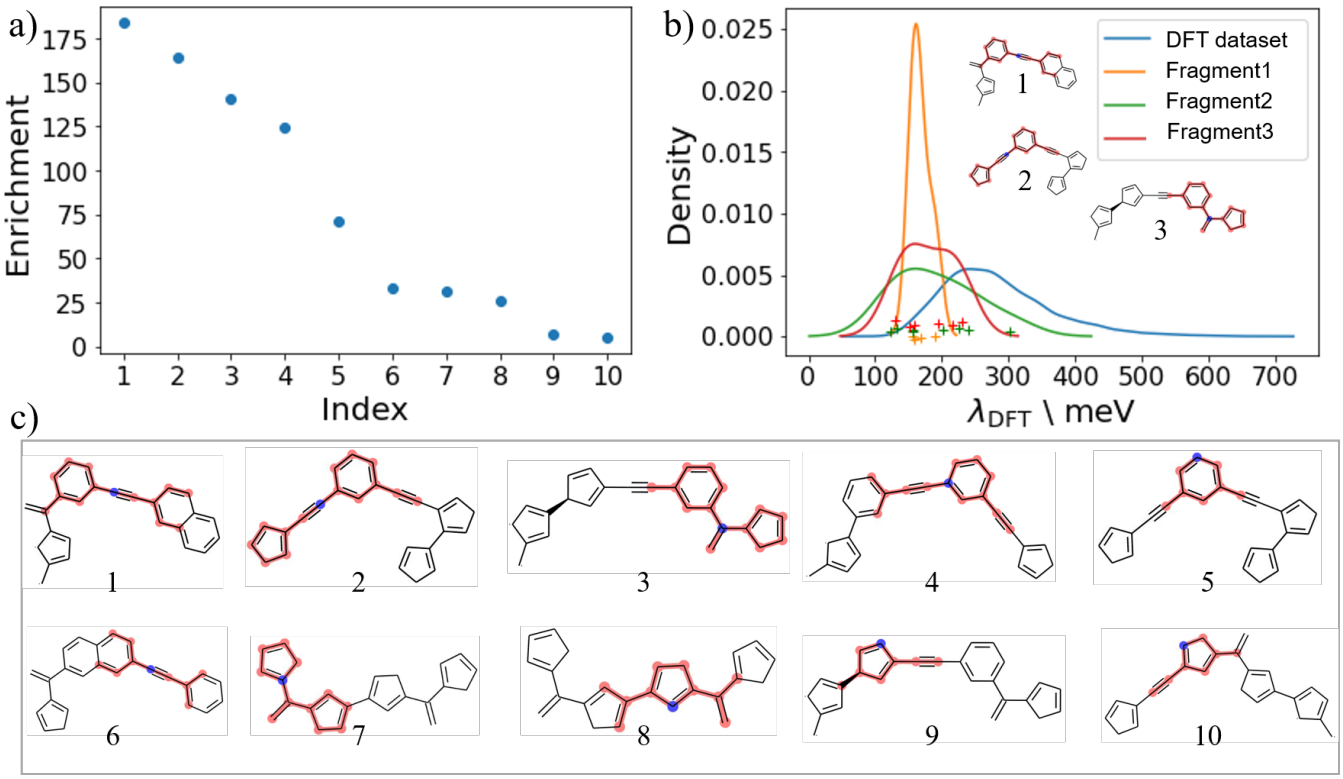

FIG. S14: Top 10 substructures which generated by morgan fingerprint with a bond radius in the set $\{3, 4, 5\}$. a) The enrichment of top 10 highest enrichment of substructures. b) The kde plot of promising substructures containing in training and validation sets (in 10900 λ_DFT data). c) 10 example molecules of 10 corresponding substructures