SUPPLEMENTARY INFORMATION

for

Revving up ^{13}C NMR Shielding Predictions across Chemical Compound Space: Benchmarks for Atoms-in-Molecules Kernel Machine Learning with new data for 134 Kilo Molecules

Amit Gupta[1], Sabyasachi Chakraborty[1], and Raghunathan Ramakrishnan[1*]

[1] *Tata Institute of Fundamental Research,*

Centre for Interdisciplinary Sciences, Hyderabad 500107, India

(Dated: December 3, 2020)

[*] ramakrishnan@tifrh.res.in

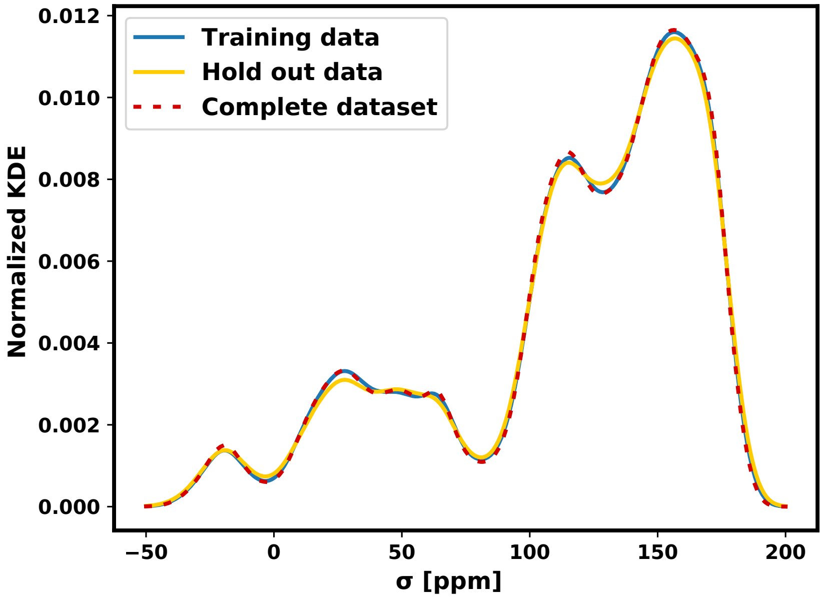


FIG. S1. Normalized distribution of the ^{13}C NMR shielding, σ, in ppm for 100k training, 50k hold-out and 832k full sets. KDE stands for Kernel Density Estimation.

TABLE S1. Mean, max and median values of descriptor differences in a random 10k training set, and the corresponding ω_{opt}. See Eqs. 8–9 in the main text.

Descriptor		Max	Mean	Median
CM	D_{ij}	1247.35	293.05	279.79
	ω_{opt}	1799.54	422.78	403.65
	MAE	5.25	5.03	5.03
SOAP	D_{ij}	37.69	13.07	12.85
	ω_{opt}	54.37	18.85	18.54
	MAE	3.59	3.50	3.50

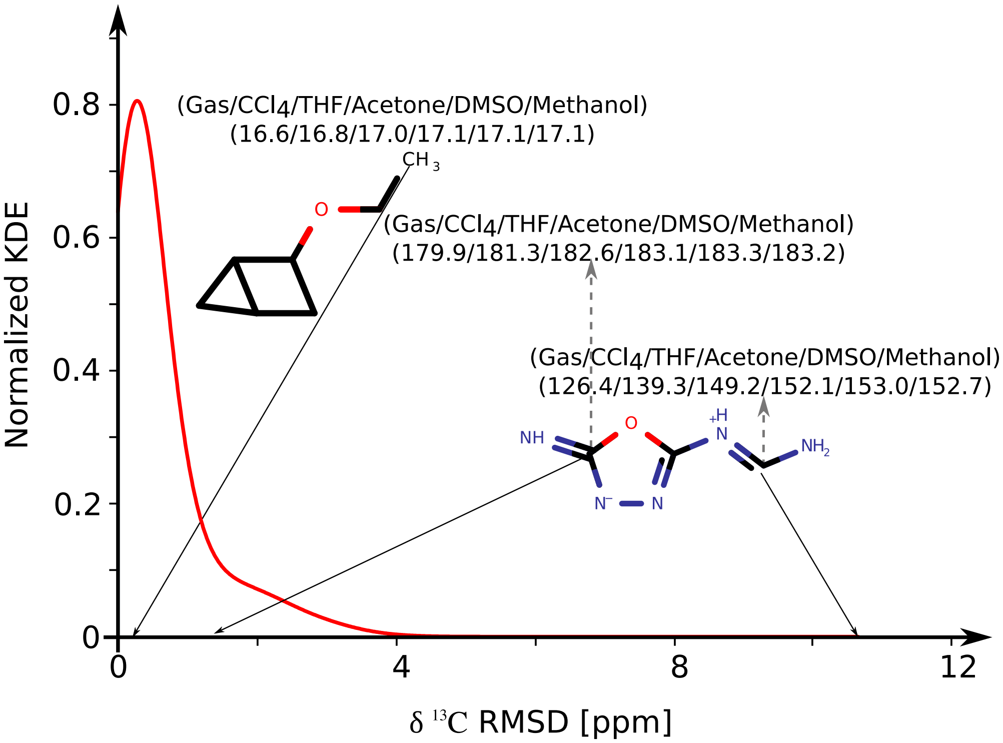


FIG. S2. Spread of mPW1PW91/6-311+G(2d,p) predicted ^{13}C chemical shifts across gas- and solvent phases. For every ^{13}C nucleus, standard deviation from the mean of values from 6 phases is plotted. Representative examples of C atoms least and most influenced by the effect of medium are highlighted along with δ^{13}C values in all 6 phases. KDE stands for Kernel Density Estimation.

TABLE S2. Solver and Prediction times for Machine Learning models with different descriptors–Coulomb Matrix (CM), Smooth Overlap of Atomic Positions (SOAP) and Faber-Christensen-Huang-Lilienfeld (FCHL). As discussed in the Computational Details, each of the descriptors have unique implementation.

N_{tr}	Solver Time(sec)			Prediction Time(sec)		
	CM	SOAP	FCHL	CM	SOAP	FCHL
10	0.00048	0.00060	0.00063	0.00021	0.00032	0.000801
20	0.00061	0.00052	0.00064	0.00026	0.00027	0.000217
50	0.00052	0.00062	0.00067	0.00020	0.00027	0.000244
100	0.00120	0.00159	0.00104	0.00035	0.00047	0.000358
200	0.00535	0.00542	0.00437	0.00073	0.00076	0.000655
500	0.02837	0.02782	0.02578	0.00125	0.00135	0.001076
1000	0.11102	0.10590	0.10682	0.00216	0.00194	0.001877
2000	0.46068	0.43622	0.46964	0.00432	0.00363	0.004443
5000	5.33974	3.94419	5.16217	0.02053	0.01487	0.019488
10000	28.27485	24.46544	26.64362	0.05581	0.05019	0.054018
20000	209.00160	178.20360	200.27772	0.27252	0.18667	0.182466
50000	3100.03850	2660.60884	3198.29773	1.89458	1.10533	1.770686
100000	27838.42770	23625.02418	24459.95798	6.73415	5.49872	6.725433

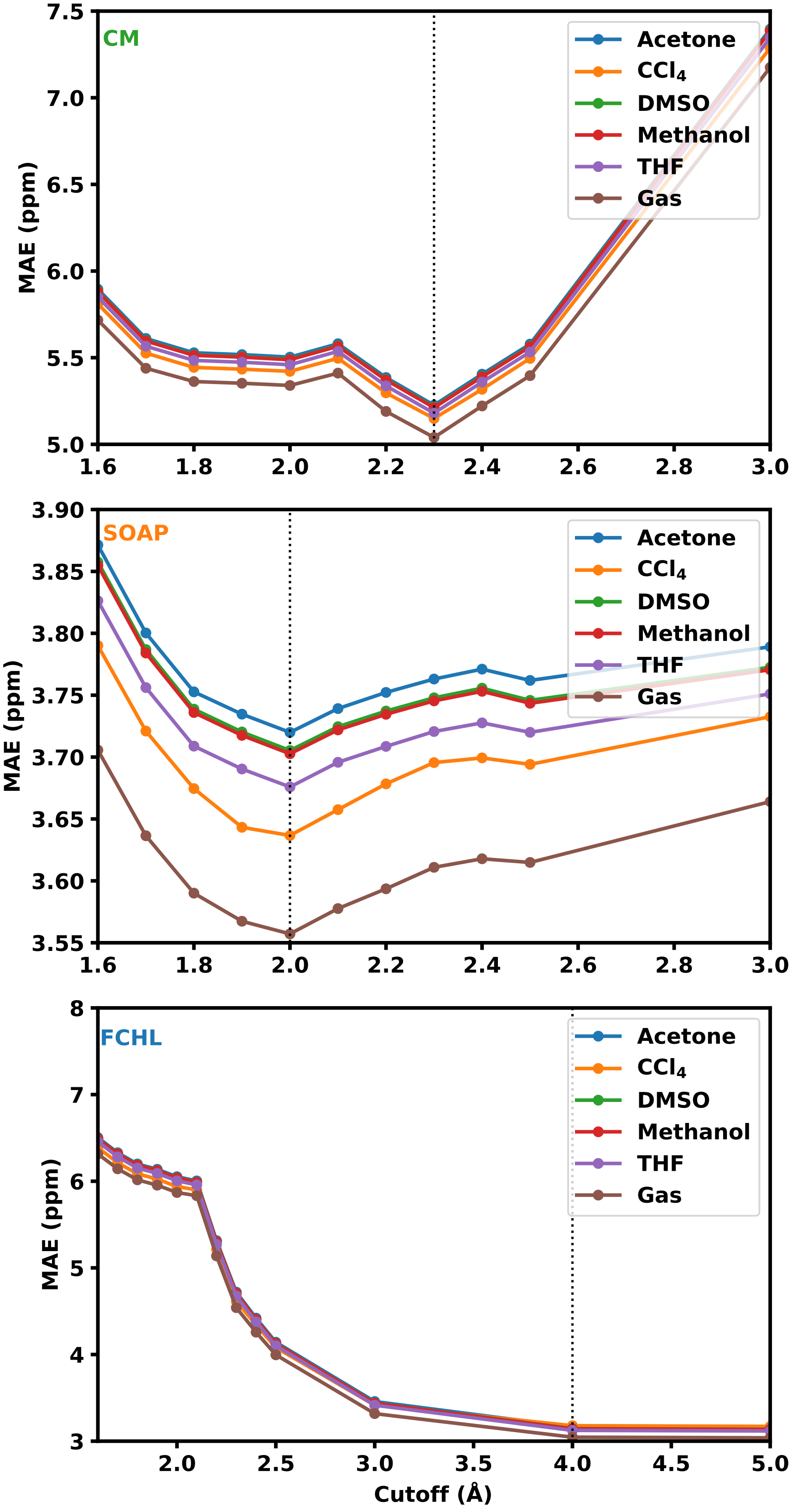


FIG. S3. Optimization of the cutoff values for the local descriptors: CM, SOAP and FCHL. Variation of the mean absolute error (MAE) in the prediction of NMR shielding are reported for gas and solvent phases.

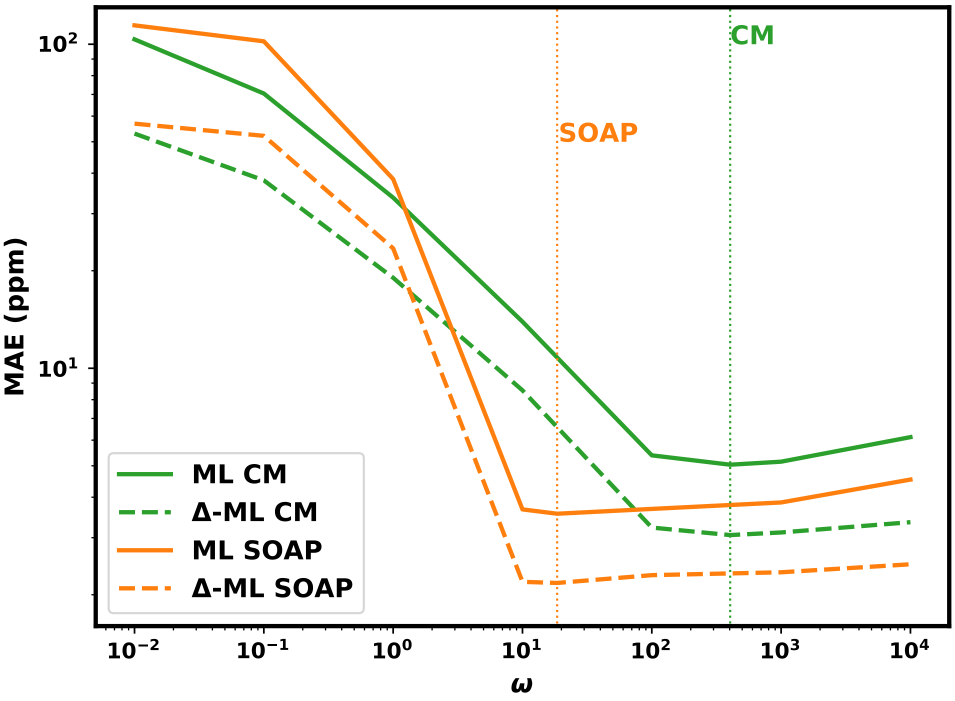


FIG. S4. Effect of the kernel width, ω, on cross-validation errors of ML and Δ-ML models. For CM and SOAP descriptors, mean absolute error (MAE) in the predicted NMR shielding for a hold out set of ^{13}C atoms from the training data pool is shown. Dotted lines denote ω_{opt}.

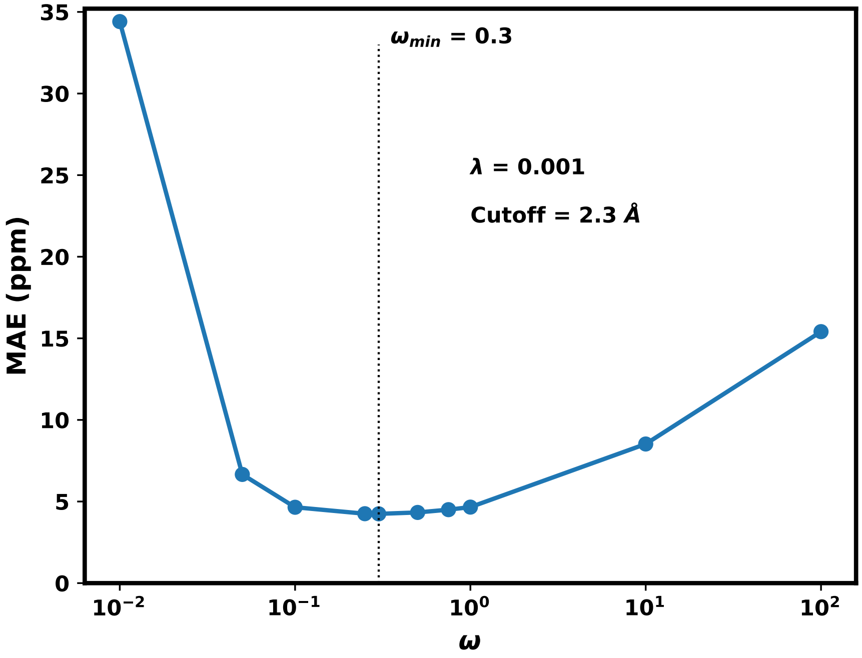


FIG. S5. Kernel width optimization of FCHL-based model via grid-search.

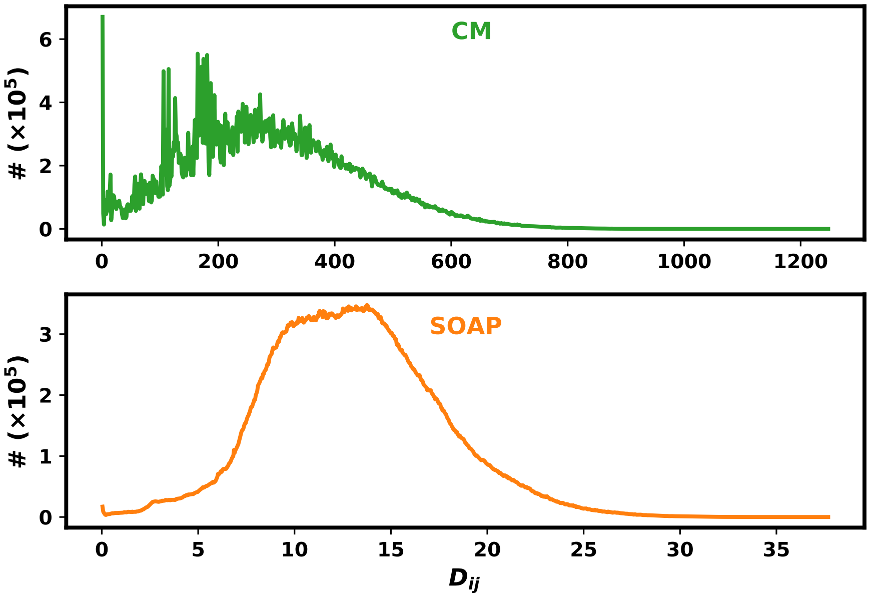


FIG. S6. Distribution of the pairwise differences, D_{ij}, of SOAP and CM descriptor vectors for randomly drawn 10k training samples.

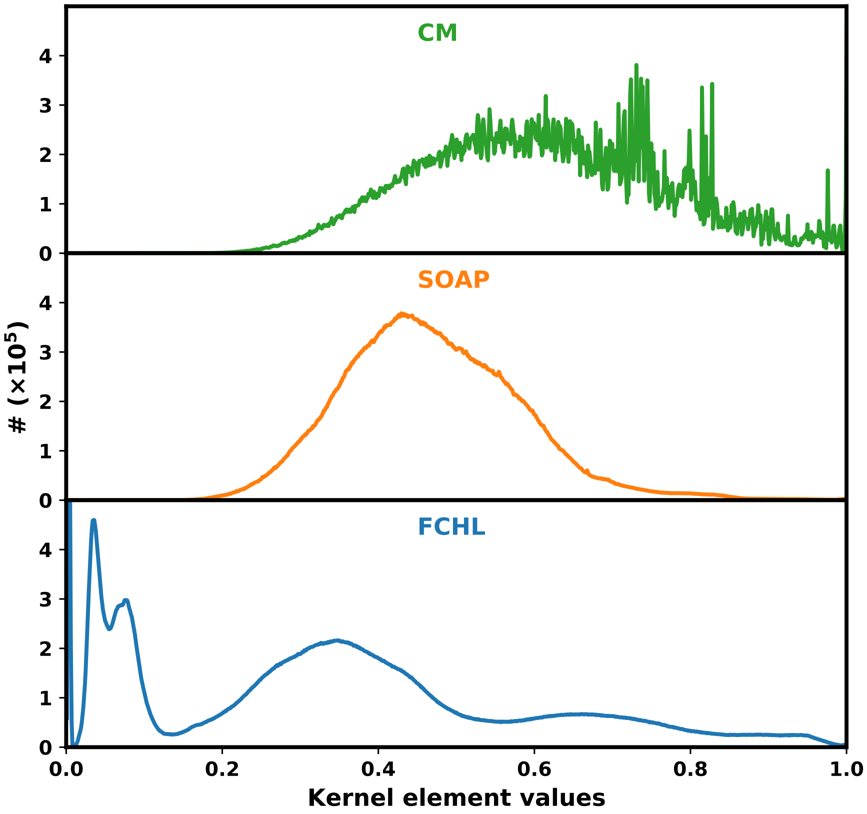


FIG. S7. Distribution of the off-diagonal elements of a 10k×10k kernel matrix for optimal hyperparameters.

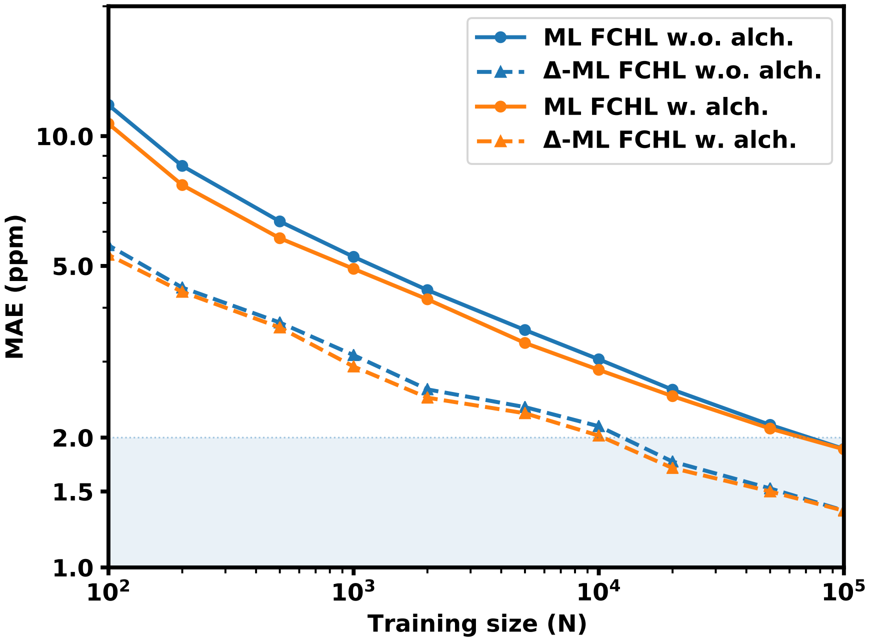


FIG. S8. Effect of alchemical extrapolations on the out-of-sample prediction errors of FCHL-based ML and Δ-ML models. A row cutoff of 1.1 was used for alchemy-enabled FCHL.

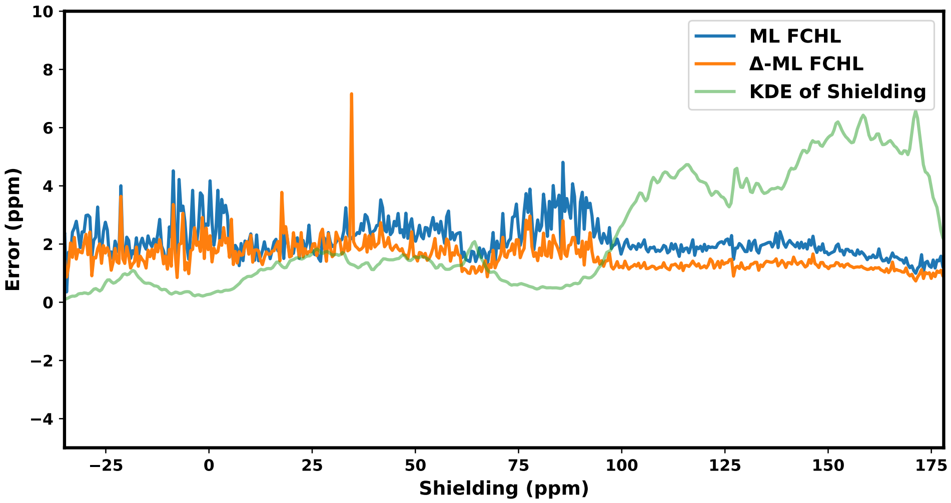


FIG. S9. Variation of ML/Δ-ML prediction errors with distribution of the target ^{13}C NMR shielding values for 50k out-of-sample nuclei. For a bin width of 0.5 ppm, signed mean errors are shown along with the probability density of the target value.

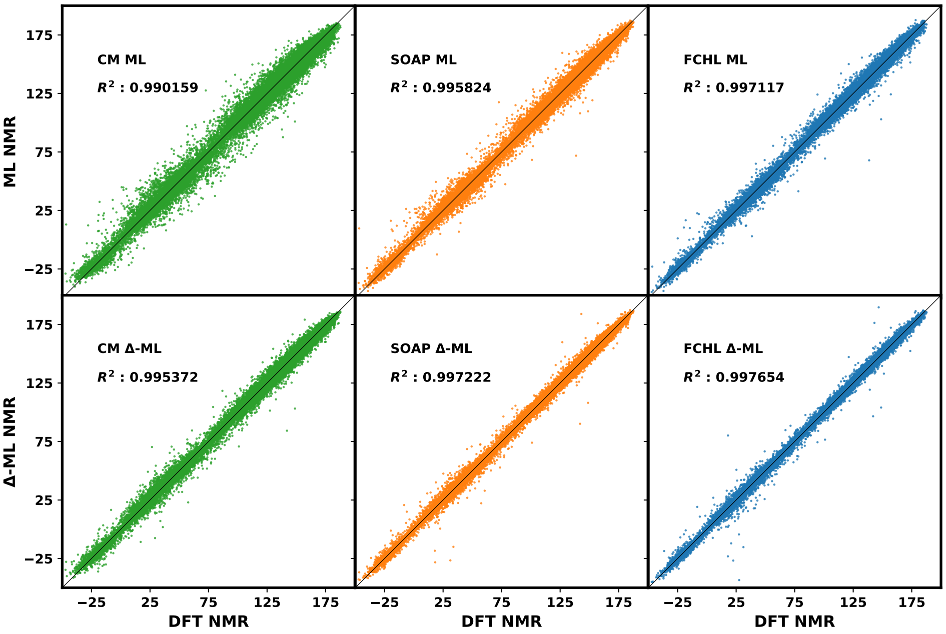


FIG. S10. Predicted and Reference (DFT) shielding tensors in training set across different ML models.

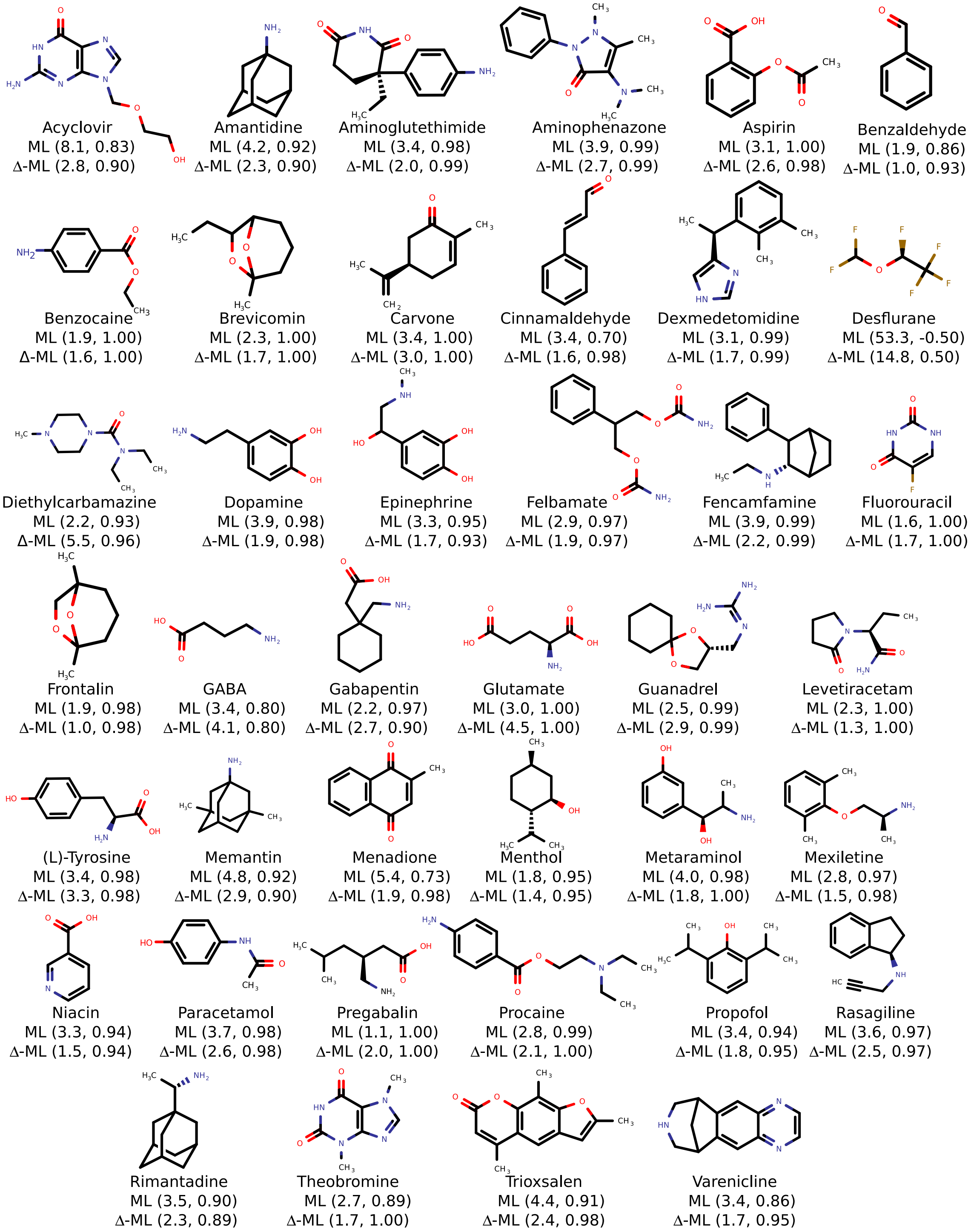
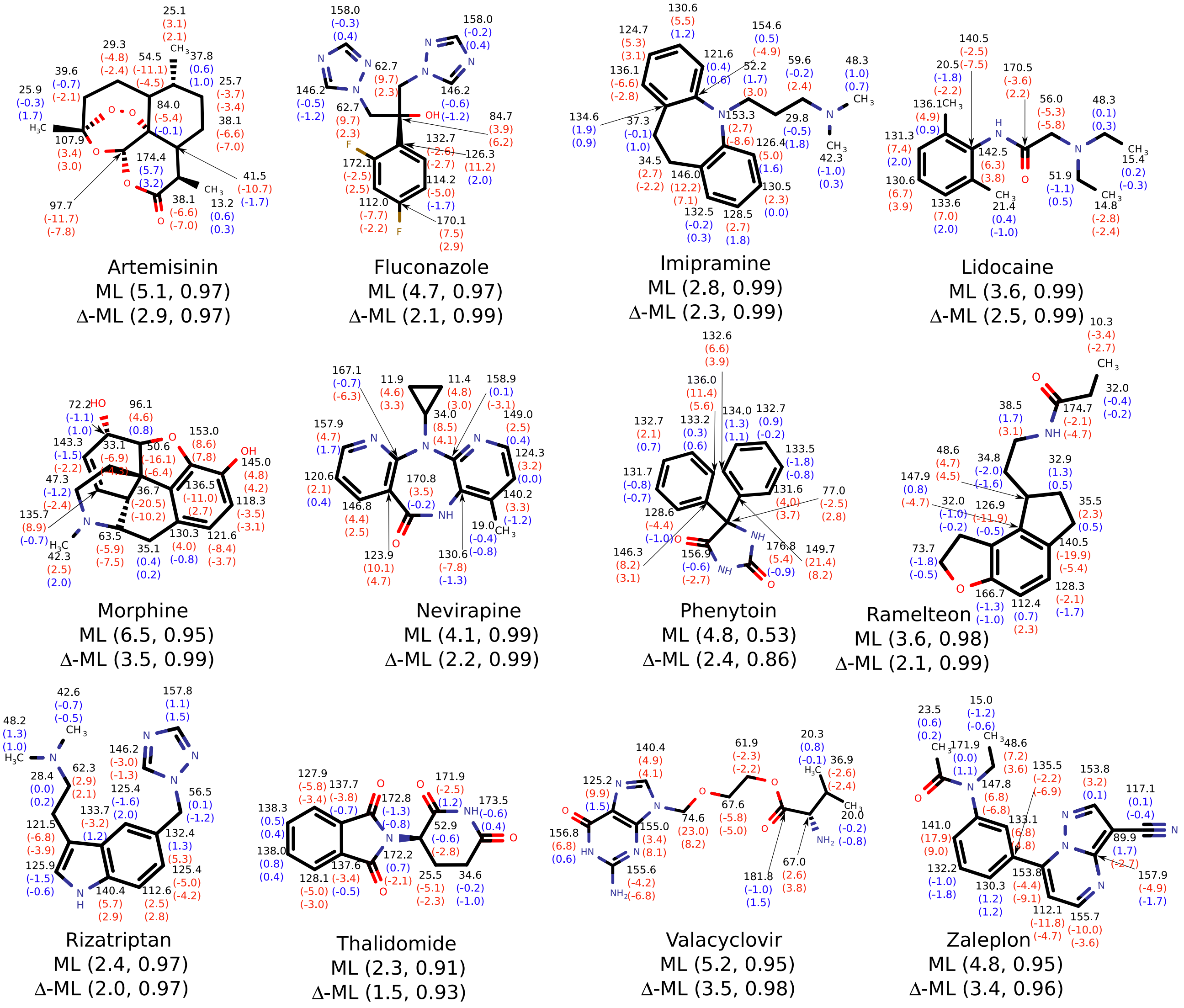


FIG. S11. Accuracies of ML- and Δ-ML-predicted mPW1PW91/6-311+G(2d,p)-level ^{13}C chemical shifts of 40 drug molecules from the GDB17 dataset. Mean absolute error (in ppm) and Spearman rank correlation coefficient (ρ) are collected in parenthesis.



Artemisinin
ML (5.1, 0.97)
Δ-ML (2.9, 0.97)

Fluconazole
ML (4.7, 0.97)
Δ-ML (2.1, 0.99)

Imipramine
ML (2.8, 0.99)
Δ-ML (2.3, 0.99)

Lidocaine
ML (3.6, 0.99)
Δ-ML (2.5, 0.99)

Morphine
ML (6.5, 0.95)
Δ-ML (3.5, 0.99)

Nevirapine
ML (4.1, 0.99)
Δ-ML (2.2, 0.99)

Phenytoin
ML (4.8, 0.53)
Δ-ML (2.4, 0.86)

Ramelteon
ML (3.6, 0.98)
Δ-ML (2.1, 0.99)

Rizatriptan
ML (2.4, 0.97)
Δ-ML (2.0, 0.97)

Thalidomide
ML (2.3, 0.91)
Δ-ML (1.5, 0.93)

Valacyclovir
ML (5.2, 0.95)
Δ-ML (3.5, 0.98)

Zaleplon
ML (4.8, 0.95)
Δ-ML (3.4, 0.96)

FIG. S12. Accuracies of ML and Δ-ML predicted mPW1PW91/6-311+G(2d,p)-level ^{13}C chemical shifts of 12 large drug molecules. Mean absolute error (in ppm) and Spearman rank correlation coefficient (ρ) are collected in parenthesis. Reference DFT results are provided next to C atoms along the deviations of ML and Δ-ML predictions from the DFT values. For clarity, unsigned deviations less than 2 ppm are shown in blue while larger ones are shown in red.

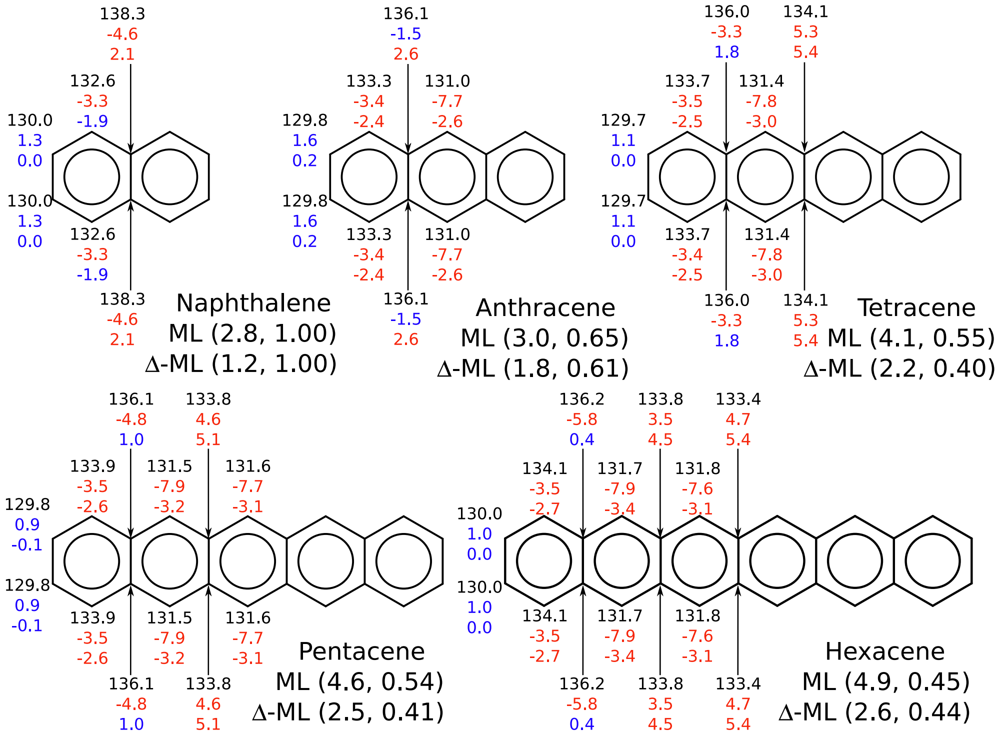


FIG. S13. Challenging case: Accuracies of ML- and Δ-ML-predicted mPW1PW91/6-311+G($2d$,p)-level ^{13}C chemical shifts of linear polycyclic aromatic hydrocarbons. Mean absolute error (in ppm) and Spearman rank correlation coefficient (ρ) are collected in parenthesis. Reference DFT results are provided next to the corresponding C atoms along with the deviations of ML and Δ-ML predictions from the DFT values. For clarity, unsigned deviations less than 2 ppm are shown in blue while larger ones are shown in red. Interstitial atoms show the maximum deviation because of the effect of delocalization.