Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Electronic Supplementary Information

Δ-Machine Learning for Quantum Chemistry Prediction of Solution-phase Molecular Properties at Ground and Excited States

Xu Chen, Pinyuan Li, Eugen Hruska, Fang Liu

Department of Chemistry, Emory University, Atlanta, Georgia, 30322

Electronic mail: fang.liu@emory.edu, xu.chen@emory.edu

Text S1. Procedures to obtain the EC descriptor.

To obtain the energy component descriptors through Terachem, four sets of single-point energy calculations were carried out. (1) The gas phase calculation was first conducted to produce the gas-phase single-point energy (index 1, index as in Table S2). (2) The converged molecular orbitals (MOs) of the gas phase calculations were treated as the initial guess of a new round of PCM calculation. By collecting the energy of the first step of the self-consistent-field (SCF) calculation, which we assumed to be the total energy of a solute under an unrelaxed potential field potential field, the index 2 component can be achieved. With the PCM calculation finished, we can obtain the final energy as the index 4 component and a relaxed potential field. (3) The relaxed potential field combined with the MO initial guess extracted from the gas phase calculation was then treated as the input files in another round of PCM calculation. In this calculation, the index 5 component can be attained by reading the total energy of the first step of the SCF calculation. (4) Similarly, the index 3 component was obtained by incorporating the MO initial guess from the (2) step and reading the energy of the first step of the SCF calculation. The other components can then be obtained by directly extracting from previous output files or be calculated based on the first 5 components.

Text S2. Feature engineering by adding SS or EC features to cFP.

Here, we decided not to use RF-RFA to determine features to be added, because the initial RF feature importance ranking may be unfair when the few features of SS or EC are combined with the high-dimensional cFP. Hence, we used the sequential feature selection (SFS)[1] approach on the ROAS dataset to avoid RF rank bias, because SFS does not need an initial rank. With the raw E_{abs} and Morgan FP set as the initial feature, components of SS, EC, and five solvent constant descriptors from cFP are added incrementally. In each SFS cycle, each of the m unselected candidate features will be tentatively combined with the n previously selected features to retrain a GB model, resulting in m new models in total. The candidate feature used in the best performing model among the m models will be officially added to the feature set. This is drastically different from RF-RFA, where each cycle only retrains the one model using the existing features plus the candidate feature with the highest initial RF rank if this one model makes even tiny MAE improvement. The feature addition was stopped at the fourth round (SI Table S23) due to the increase of MAE after adding SA, C-PCM cavity surface, and ET30 as the best among candidates, but we still tested a few rounds of addition manually, observing only small fluctuations occurred near the best performance.

Table S1. Hyperparameter tuning search space for each type of ML model. To avoid overtraining ML models, hyperparameter tuning was performed for all models using hyperopt[2]. Adaptive Tree Parzen Estimators (TPE) was used as the searching algorithm with a maximum of 200 trials. Hyperparameters are named as in scikit-learn and searching space as in hyperopt.[2]

ML model	hyperparameter	searching space
linear	alpha	uniform (50,200)
	solver	auto, svd, cholesky, lsqr, sparse_cg, sag, saga)
KRR	alpha	loguniform (-10, 1)
	gamma	loguniform (-10, 1)
	kernel	rbf, linear, laplacian
GB	n_estimators	randint (2000)
	learning_rate	loguniform (-1, 0.5)
	max_depth	quniform (1, 10, 2)
	min_samples_leaf	quniform (1, 20, 2)
	min_samples_split	quniform (1, 10, 2)
RF	n_estimators	randint (2000)
	max_depth	choice (None, [2, 30]
	min_samples_leaf	quniform (1, 20, 2)
	min_samples_split	quniform (1, 10, 2)
	max_features	None, sqrt, log2
ANN	learning_rate_init	loguniform (-5, 0.2)
	hidden_layer_sizes	[5], [10], [20], [50], [5] *2, [10] *2, [20]*2, [50]*2, [5]*3, [10]*3, [20]*3, [50]*3
	activation	logistic, tanh, relu
	alpha	loguniform (-5, 0.2)

Table S2. The 10 energy components of the energy component descriptor.

index	Energy Component
1	$\langle\Psi^{(0)}\|H\|\Psi^{(0)}\rangle$
2	$\langle\Psi^{(0)}\|H+V^{(0)}/2\|\Psi^{(0)}\rangle$
3	$\langle\Psi^{(1)}\|H\|\Psi^{(1)}\rangle$
4	$\langle\Psi^{(1)}\|H+V^{(1)}/2\|\Psi^{(1)}\rangle$
5	$\langle\Psi^{(0)}\|H+V^{(1)}/2\|\Psi^{(0)}\rangle$
6	$\langle\Psi^{(0)}\|H+V^{(1)}/2\|\Psi^{(0)}\rangle-\langle\Psi^{(0)}\|H\|\Psi^{(0)}\rangle$
7	$\langle\Psi^{(1)}\|H\|\Psi^{(1)}\rangle-\langle\Psi^{(0)}\|H\|\Psi^{(0)}\rangle$
8	$\langle\Psi^{(1)}\|H+V^{(1)}/2\|\Psi^{(1)}\rangle-\langle\Psi^{(1)}\|T\|\Psi^{(1)}\rangle$
9	$\langle\Psi^{(1)}\|T\|\Psi^{(1)}\rangle$
10	$\langle\Psi^{(1)}\|H\|\Psi^{(1)}\rangle-\langle\Psi^{(1)}\|T\|\Psi^{(1)}\rangle$

Table S3. The best Δ-ML MAE of all combinations of ML models and energy component descriptors predicting on three respective ωB97(-D3) functional calculated datasets. The 'removed' column means the linear dependent components were removed from the original 10 energy components. The 'origin' column means using the 10 energy components as mentioned in Table S2. There is no change of the performance after removing the redundant components.

dataset	removed	origin
OROP	RF/0.166 V	RF/0.166 V
OMROP	ANN/0.516 V	ANN/0.516 V
ROAS	RF/0.458 eV	RF/0.458 eV

Table S4-S6. Δ-ML model performance for each descriptor on various datasets without adding the raw QM calculated results as features.

Table S4. OROP data set, Δ-ML redox potential, ωB97-D3 functional.

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
SS		0.541	0.508	0.684	0.505	0.493	RF/0.505
CM		0.525	0.511	0.581	0.430	0.489	RF/0.430
EC	0.618	0.538	0.486	0.505	0.472	0.493	RF/0.472
FP		0.483	0.465	0.457	0.442	0.444	RF/0.442

Table S5. OMROP data set, Δ-ML redox potential, ωB97-D3 functional.

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
SS		1.390	1.300	1.510	1.285	1.336	RF/1.285
CM		1.435	1.434	1.506	1.465	1.423	ANN/1.423
EC	1.342	1.377	1.378	1.531	1.337	1.423	RF/1.337
FP		1.381	1.380	1.407	1.264	1.266	RF/1.264

Table S6. ROAS data set, Δ-ML absorption energy, ωB97 functional.

MAE (eV)	raw	linear	KRR	GB	RF	ANN	best ML
SS		0.689	0.535	0.585	0.544	0.542	KRR/0.535
CM		0.655	0.541	0.522	0.505	0.633	RF/0.505
EC	0.822	0.723	0.607	0.616	0.608	0.616	KRR/0.607
FP		0.542	0.450	0.424	0.416	0.450	RF/0.416

Table S7-S18. Δ-ML model performance for each descriptor on various datasets.

Table S7. SS descriptor, ROAS data set, Δ-ML absorption energy

MAE (eV)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP	0.787	0.590	0.466	0.453	0.443	0.483	RF/0.443
ωB97	0.822	0.576	0.427	0.388	0.386	0.465	RF/0.386
ωB97X	1.016	0.564	0.414	0.437	0.409	0.452	RF/0.409
ωPBEh	1.119	0.556	0.422	0.436	0.417	0.438	RF/0.417
PBE0	0.844	0.586	0.460	0.460	0.437	0.518	RF/0.437
CAM-B3LYP	1.099	0.563	0.453	0.455	0.436	0.503	RF/0.436
functional sensitivity	0.332	0.034	0.052	0.072	0.057	0.080	linear/0.031
best functional	B3LYP/ 0.787	ωPBEh/ 0.556	ωPBEh/ 0.414	ωB97/ 0.388	ωB97/ 0.386	ωPBEh/ 0.438	ωB97 RF/0.386

Table S8. CM descriptor, ROAS data set, Δ-ML absorption energy

MAE (eV)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP	0.787	0.600	0.455	0.444	0.466	0.522	GB/0.444
ωB97	0.822	0.572	0.455	0.415	0.407	0.476	RF/0.407
ωB97X	1.016	0.557	0.441	0.416	0.421	0.471	GB/0.416
ωPBEh	1.119	0.555	0.435	0.430	0.425	0.489	RF/0.425
PBE0	0.844	0.596	0.458	0.477	0.462	0.570	KRR/0.458
CAM-B3LYP	1.099	0.565	0.447	0.450	0.444	0.510	RF/0.444
functional sensitivity	0.332	0.045	0.023	0.062	0.059	0.099	KRR/0.023
best functional	B3LYP/ 0.787	ωPBEh/ 0.555	ωPBEh/ 0.435	ωB97/ 0.415	ωB97/ 0.407	ωB97X/ 0.471	ωB97 RF/0.407

Table S9. EC descriptor, ROAS data set, Δ-ML absorption energy

MAE (eV)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP	0.787	0.607	0.542	0.550	0.507	0.528	RF/0.507
ωB97	0.822	0.581	0.474	0.480	0.458	0.485	RF/0.458
ωB97X	1.016	0.574	0.505	0.489	0.480	0.478	RF/0.480
ωPBEh	1.119	0.569	0.481	0.497	0.493	0.474	ANN/0.474
PBE0	0.844	0.604	0.521	0.501	0.510	0.527	GB/0.501
CAM-B3LYP	1.099	0.574	0.489	0.511	0.493	0.496	KRR/0.489
functional sensitivity	0.332	0.038	0.068	0.070	0.052	0.054	linear/0.038
best functional	B3LYP/ 0.787	ωPBEh/ 0.569	ωB97/ 0.474	ωB97/ 0.480	ωB97/ 0.458	ωPBEh/ 0.474	ωB97 RF/0.458

Table S10. FP descriptor, ROAS data set, Δ-ML absorption energy

MAE (eV)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP	0.787	0.442	0.389	0.324	0.333	0.394	GB/0.324
ωB97	0.822	0.418	0.394	0.311	0.317	0.372	GB/0.311
ωB97X	1.016	0.456	0.404	0.318	0.324	0.415	GB/0.318
ωPBEh	1.119	0.441	0.389	0.306	0.321	0.415	GB/0.306
PBE0	0.844	0.463	0.394	0.318	0.330	0.442	GB/0.318
CAM-B3LYP	1.099	0.454	0.393	0.318	0.335	0.430	GB/0.318
functional sensitivity	0.332	0.045	0.015	0.018	0.018	0.070	KRR/0.015
best functional	B3LYP/ 0.787	ωB97/ 0.418	ωPBEh/ 0.389	ωPBEh/ 0.306	ωB97/ 0.317	ωB97/ 0.372	ωPBEh GB/0.306

Table S11. SS descriptor, OROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.410	0.208	0.151	0.152	0.161	0.265	KRR/0.151
ωB97-D3	0.618	0.375	0.160	0.131	0.152	0.253	GB/0.131
ωB97X-D3	0.480	0.370	0.198	0.141	0.160	0.171	GB/0.141
ωPBEh-D3	0.413	0.313	0.190	0.207	0.215	0.264	KRR/0.190
PBE0-D3	0.263	0.193	0.195	0.174	0.161	0.191	RF/0.161
CAM-B3LYP-D3	0.364	0.266	0.211	0.171	0.208	0.224	GB/0.171
functional sensitivity	0.355	0.182	0.060	0.076	0.063	0.094	RF/0.063
best functional	PBE0-D3/0.263	PBE0-D3/0.193	B3LYP-D3/0.151	ωB97-D3/0.131	ωB97-D3/0.152	ωB97X-D3/0.171	ωB97-D3 GB/0.131

Table S12. CM descriptor, OROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.410	0.227	0.201	0.245	0.204	0.222	KRR/0.201
wb97-D3	0.618	0.435	0.186	0.219	0.183	0.246	RF/0.183
wb97X-D3	0.480	0.435	0.264	0.266	0.222	0.284	RF/0.222
wPBEh-D3	0.413	0.349	0.228	0.232	0.247	0.268	KRR/0.228
PBE0-D3	0.263	0.194	0.163	0.235	0.191	0.249	KRR/0.163
CAM-B3LYP-D3	0.364	0.298	0.215	0.231	0.247	0.207	ANN/0.207
functional sensitivity	0.355	0.241	0.101	0.047	0.064	0.077	GB/0.047
best functional	PBE0-D3/0.263	PBE0-D3/0.194	PBE0-D3 /0.163	ωB97-D3/0.219	ωB97-D3/0.183	CAM-B3LYP-D3/0.207	PBE0-D3 KRR/0.163

Table S13. EC descriptor, OROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.410	0.226	0.181	0.184	0.174	0.233	GB/0.174
ωB97-D3	0.618	0.438	0.225	0.196	0.166	0.217	RF/0.166
ωB97X-D3	0.480	0.417	0.220	0.192	0.179	0.212	RF/0.179
ωPBEh-D3	0.413	0.340	0.302	0.248	0.230	0.190	ANN/0.190
PBE0-D3	0.263	0.199	0.175	0.177	0.166	0.209	RF/0.166
CAM-B3LYP-D3	0.364	0.286	0.247	0.207	0.230	0.199	ANN/0.199
functional sensitivity	0.355	0.239	0.127	0.071	0.064	0.043	ANN/0.043
best functional	PBE0-D3/0.263	PBE0-D3/0.199	PBE0-D3/0.175	PBE0-D3/0.177	ωB97-D3/0.166	ωPBEh-D3/0.190	PBE0-D3 RF/0.166

Table S14. FP descriptor, OROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.410	0.205	0.260	0.244	0.227	0.261	RF/0.227
ωB97-D3	0.618	0.457	0.449	0.235	0.216	0.472	RF/0.216
ωB97X-D3	0.480	0.415	0.427	0.332	0.232	0.392	RF/0.232
ωPBEh-D3	0.413	0.322	0.340	0.263	0.261	0.388	RF/0.261
PBE0-D3	0.263	0.177	0.200	0.255	0.221	0.213	linear/0.177
CAM-B3LYP-D3	0.364	0.269	0.328	0.313	0.274	0.296	RF/0.274
functional sensitivity	0.355	0.280	0.249	0.097	0.058	0.213	RF/0.058
best functional	PBE0-D3/0.263	PBE0-D3 /0.177	PBE0-D3 /0.200	ωB97-D3/0.235	PBE0-D3/0.216	PBE0-D3/0.213	PBE0-D3 linear/0.177

Table S15. SS descriptor, OMROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.817	0.735	0.509	0.478	0.498	0.602	GB/0.478
ωB97-D3	1.342	0.672	0.493	0.381	0.423	0.554	GB/0.381
ωB97X-D3	1.222	0.607	0.446	0.408	0.435	0.644	GB/0.408
ωPBEh-D3	1.573	0.743	0.490	0.508	0.395	0.649	RF/0.395
PBE0-D3	0.817	0.710	0.537	0.537	0.460	0.577	RF/0.460
CAM-B3LYP-D3	1.358	0.639	0.476	0.480	0.498	0.552	GB/0.476
functional sensitivity	0.756	0.136	0.091	0.156	0.103	0.097	KRR/0.091
best functional	PBE0-D3 /0.817	ωB97X-D3/0.607	ωB97X-D3/0.446	ωB97-D3/0.381	ωPBEh-D3/0.395	CAM-B3LYP-D3/0.552	ωB97-D3 GB/0.381

Table S16. CM descriptor, OMROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.817	0.786	0.588	0.560	0.591	0.629	GB/0.560
ωB97-D3	1.342	0.677	0.443	0.381	0.504	0.471	GB/0.381
ωB97X-D3	1.222	0.523	0.417	0.414	0.401	0.462	RF/0.401
ωPBEh-D3	1.573	0.769	0.537	0.482	0.451	0.533	RF/0.451
PBE0-D3	0.817	0.730	0.561	0.489	0.577	0.616	GB/0.489
CAM-B3LYP-D3	1.358	0.796	0.594	0.489	0.496	0.541	GB/0.489
functional sensitivity	0.756	0.273	0.177	0.179	0.190	0.167	ANN/0.167
best functional	PBE0-D3 /0.817	ωB97X-D3/0.523	ωB97X-D3/0.417	ωB97-D3/0.381	ωB97X-D3/0.401	ωB97X-D3/0.462	ωB97-D3 GB/0.381

Table S17. EC descriptor, OMROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.817	0.774	0.624	0.558	0.605	0.620	GB/0.558
ωB97-D3	1.342	0.631	0.518	0.697	0.543	0.516	ANN/0.516
ωB97X-D3	1.222	0.519	0.574	0.672	0.522	0.597	linear/0.519
ωPBEh-D3	1.573	0.759	0.584	0.525	0.493	0.589	RF/0.493
PBE0-D3	0.817	0.707	0.567	0.594	0.574	0.558	ANN/0.558
CAM-B3LYP-D3	1.358	0.795	0.577	0.706	0.555	0.587	RF/0.555
functional sensitivity	0.756	0.276	0.106	0.181	0.112	0.104	ANN/0.104
best functional	PBE0-D3 /0.817	ωB97X-D3/0.519	ωB97-D3/0.518	ωPBEh-D3/0.525	ωPBEh-D3/0.493	ωB97-D3/0.516	ωPBEh-D3 RF/0.493

Table S18. FP descriptor, OMROP data set, Δ-ML redox potential

MAE (V)	raw	linear	KRR	GB	RF	ANN	best ML
B3LYP-D3	0.817	0.801	0.804	0.631	0.518	0.740	RF/0.518
ωB97-D3	1.342	0.948	0.938	0.530	0.529	1.108	RF/0.529
ωB97X-D3	1.222	0.880	0.837	0.667	0.496	0.898	RF/0.496
ωPBEh-D3	1.573	0.983	0.976	0.552	0.484	0.853	RF/0.484
PBE0-D3	0.817	0.752	0.750	0.531	0.507	0.829	RF/0.507
CAM-B3LYP-D3	1.358	0.836	0.815	0.661	0.593	0.785	RF/0.593
functional sensitivity	0.756	0.234	0.172	0.137	0.109	0.368	RF/0.109
best functional	PBE0-D3 /0.817	PBE0-D3/0.751	PBE0-D3/0.750	PBE0-D3/0.531	ωPBEh-D3/0.484	B3LYP-D3/0.740	ωPBEh-D3 RF/0.484

Table S19. Δ-ML redox potential performance (ωB97-D3) before and after RF-RFA on OROP data set predicting redox potential.

MAE (V)	linear	KRR	GB	RF	ANN
before	0.375	0.160	0.131	0.152	0.253
after	0.371	0.177	0.170	0.151	0.155

Table S20. ML performance (ωB97-D3) before and after RF-RFA on OMROP data set predicting redox potential.

MAE (V)	linear	KRR	GB	RF	ANN
before	0.673	0.493	0.381	0.423	0.554
after	0.635	0.533	0.422	0.429	0.553

Table S21. Δ-ML absorption energy performance (ωB97) of different bits of Morgan FP with the 5 solvent constant descriptors and $E_{calculated}$.

Bits	linear	KRR	GB	RF	ANN
1024	0.418	0.394	0.311	0.317	0.372
512	0.422	0.386	0.326	0.305	0.394
256	0.425	0.359	0.330	0.316	0.374
128	0.459	0.350	0.322	0.321	0.363

Table S22. A comparison of FP ML performance (ωB97/GB) between direct generation of 1024, 512… bits Morgan fingerprints from RDKit, and generating 3072 bits Morgan fingerprints from RDKit and then using "SelectKBest" reducing bits to 1024, 512… on ROAS predicting absorption energy.

MAE (eV)	1024	512	256	128
directly generate	0.311	0.326	0.330	0.321
"SelectKBest"	0.540	0.572	0.549	0.570

Table S23. ML performance (ωB97/GB) of each step of SFS on ROAS data set predicting absorption energy. The initial features are set as Morgan fingerprints (1024 bits) and the calculated absorption energy. The test set MAEs are lower than the validation set here because the model was trained on the whole training set (80% of all data) before predicting on the test set. In contrast, the MAEs of validation set are the averaged 5-fold validation score by training only 64% of all data. SA, ET are two of the solvent constants. 6 is the index 6 energy component of EC descriptors. C-PCM cavity surface and spin are two features of SS descriptors.

Feature	validation set MAE (eV)	test set MAE (eV)
FP + $E_{calculated}$ + SA	0.3232	0.299
FP + $E_{calculated}$ + SA + C-PCM cavity surface	0.3221	0.295
FP + $E_{calculated}$ + SA + C-PCM cavity surface + ET30	0.3208	0.301
FP + $E_{calculated}$ + SA + C-PCM cavity surface + ET30 + 6	0.3222	0.296
FP + $E_{calculated}$ + SA + C-PCM cavity surface + ET30 + 6 + spin	0.3221	0.306

References

1. L. Breiman, *Machine Learning*, 2001, **45**, 5-32.
2. J. Bergstra, D. Yamins and D. Cox, presented in part at the Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research2013.