Electronic Supplementary Information

A Machine Learning Approach for Predicting the Empirical Polarity of Organic Solvents

Vaneet Saini [a*] and Ranjeet Kumar [a]

[a] *Department of Chemistry & Centre for Advanced Studies in Chemistry, Panjab University, Chandigarh 160014, India*, Email: vsaini@pu.ac.in

Contents

General Information

The $E_T(30)$ values were acquired from two sources.[1, 2] Molecules were drawn on Gauss View 6 and computations were performed on Gaussian 16 software.[3] The geometry optimization and frequency calculations were performed on B3LYP/631G++(d,p) for DFT. Minima was confirmed by vibrational frequency calculations showing no imaginary frequency. For machine learning all the calculations were performed on google colab using Python framework. Data analysis and pre-processing was done using Pandas, NumPy and Scikit-learn libraries. Model development was performed using Scikit-learn, Keras and Tensorflow libraries. Visualization of data was done on Seaborn and Matplotlib libraries. Code is available at https://github.com/v-saini/Empirical-Polarity.git

Descriptors

Units

Total energy (E), energies of HOMO and LUMO, electronegativity (χ), chemical hardness (η), chemical potential (μ), electrophilicity index (ω), dipole moment (d), free energy (G), enthalpy (H) (all quantum-mechanical descriptors in a.u., 1a.u.=27.21165 eV, 627.50956 kcal/mol, thermodynamic descriptors in Kcal/mol, dipole moment in Debye calculated with DFT/B3LYP/6-311G(d,p) level of the theory).[4]

Descriptor Info

Table S1. List of calculated quantum chemical descriptors used in this study.

Sr. No	Notation	Definition	Source/Formula
1	E	Total electronic energy-DFT	Gaussian 09
2	E (HOMO)	Energy of highest occupied molecular orbital-DFT	Gaussian 09
3	E (LUMO)	Energy of lowest occupied molecular orbital-DFT	Gaussian 09
4	η (Hardness)	Chemical hardness-DFT	E (LUMO)-E (HOMO)
5	χ (Electronegativity)	Electronegativity-DFT	-(E(HOMO)+E(LUMO))/2
6	μ (Chemical potential)	Chemical potential-DFT	(E(HOMO)+E(LUMO))/2
7	ω (Electrophilicity index)	Electrophilicity index-DFT	$\mu^2/2\eta$
8	d (dipole moment)	Dipole moment-DFT	Gaussian 09
9	H (enthalpy)	Enthalpy-DFT	Gaussian 09
10	G (Gibbs free energy)	Gibbs free energy-DFT	Gaussian 09
11	N-Het	No of heteroatoms	RDkit
12	Solvent type (ST)	ST as a categorical descriptor	Reference 1

Table S2. List of total no. of molecules used in each class with their means $E_T(30)$ values.

Sr. No	Type of Solvent	Total no. of molecules in each class	Mean $E_T(30)$ values
1	Alcohols/Phenols	104	50.03

2	Nitrogen containing compounds	73	41.37
3	Ethers, thioethers and acetals	47	36.92
4	Halo-alkanes/alkenes/alkyne	43	36.91
5	Esters	36	39.64
6	Arenes	29	35.93
7	Organic salts	18	52.47
8	Ketones	18	40.50
9	Alkanes and alkenes	16	31.71
10	Heteroarenes	16	40.11
11	Sulphur compounds	10	42.27
12	Phosphorous compounds	7	40.97
13	Carboxylic acid/anhydrides	4	50.10

Table S3. Descriptor Statistics before and after standardization.

		Before Standardization				**After Standardization**			
	N-Het	E(Energy)	E(HOMO)	E(LUMO)	d	E(Energy)	E(HOMO)	E(LUMO)	d
count	421	421	421	421	421	421	421	421	421
mean	1.985748	-685.272	-0.26117	-0.02619	2.511342	8.86E-17	4.13E-16	2.32E-16	-7.595E-17
std	1.407391	876.0235	0.033164	0.020414	2.215839	1.00119	1.00119	1.00119	1.0011898
min	0	-7753.86	-0.35323	-0.11705	0	-8.07855	-2.77919	-4.45607	-1.1347079
25%	1	-661.584	-0.28163	-0.03299	1.36415	0.027072	-0.61768	-0.33345	-0.5183396
50%	2	-387.536	-0.25871	-0.01768	1.8817	0.340277	0.074241	0.417409	-0.2844933
75%	2	-306.057	-0.23899	-0.01238	3.127199	0.433398	0.669562	0.677341	0.2782646
max	8	-62.0569	-0.1475	0.00697	15.41071	0.71226	3.431523	1.626338	5.8283617

Train:Test Split

A fixed training:test split was used to evaluate the models. The split was random with 90:10 training:test, with proportional data from the different solvent types. Models were built exclusively with the training data and then tested for the test sets. The number of data points in the train and test sets is shown in

Table S4. Dataset size and train and test set sizes

Training:test split ratio	Training set size	Test set size
90:10	378	43
80:20	336	85
70:30	294	127
60:40	252	169

Table S5. Model evaluation metrics for NN model using 10% train test split.

Sr. No.	Regression algorithms	R^2 (cross validation)	RMSE (cross validation)	R^2 (training set)	R^2 (test set)	RMSE (test set)
1	MLR	0.696	3.95	0.723	0.673	3.68
2	PLS	0.687	4.01	0.709	0.656	3.77
3	KNN	0.639	4.28	0.822	0.807	2.83
4	SVR	0.700	3.87	0.799	0.858	2.42
5	ET	0.857	2.72	1.000	0.952	1.40
6	BR	0.792	3.22	0.965	0.897	2.07
7	RF	0.811	3.07	0.975	0.927	1.74
8	NN	0.929	1.79	0.959	0.960	1.29

Table S6. Model evaluation metrics for NN model using 20% train test split

Sr. No.	Regression algorithms	R^2 (cross validation)	RMSE (cross validation)	R^2 (training set)	R^2 (test set)	RMSE (test set)
1	MLR	0.664	4.13	0.714	0.735	3.47
2	PLS	0.659	4.18	0.701	0.721	3.56
3	KNN	0.658	4.20	0.799	0.778	3.17
4	SVR	0.696	3.98	0.799	0.824	2.83
5	ET	0.830	2.93	1.000	0.920	1.91
6	BR	0.746	3.57	0.964	0.837	2.72
7	RF	0.789	3.25	0.973	0.844	2.66
8	NN	0.910	1.96	0.947	0.925	1.85

Table S7. Model evaluation metrics for NN model using 30% train test split

Sr. No.	Regression algorithms	R^2 (cross validation)	RMSE (cross validation)	R^2 (training set)	R^2 (test set)	RMSE (test set)
1	MLR	0.645	4.21	0.713	0.719	3.68
2	PLS	0.645	4.20	0.697	0.703	3.79
3	KNN	0.625	4.41	0.771	0.732	3.60
4	SVR	0.680	4.06	0.764	0.790	3.18
5	ET	0.803	3.11	1.000	0.916	2.01
6	BR	0.733	3.61	0.962	0.877	2.44
7	RF	0.749	3.52	0.968	0.875	2.46
8	NN	0.924	1.96	0.947	0.911	2.07

Table S8.Model evaluation metrics for NN model using 40% train test split

Sr. No.	Regression algorithms	R^2 (cross validation)	RMSE (cross validation)	R^2 (training set)	R^2 (test set)	RMSE (test set)
1	MLR	0.649	4.04	0.717	0.705	3.96
2	PLS	0.656	4.01	0.706	0.685	4.10
3	KNN	0.558	4.56	0.761	0.655	4.29
4	SVR	0.669	3.95	0.770	0.738	3.74
5	ET	0.810	3.02	1.000	0.843	2.90
6	BR	0.676	3.85	0.962	0.814	3.15
7	RF	0.697	3.69	0.965	0.817	3.13
8	NN	0.889	2.24	0.945	0.872	2.60

Table S9. Train and test set metrics with change in the number of neurons in the NN model.

S.No.	Neurons (Hidden Layer-1)	Neurons (Hidden Layer-2)	Neurons (Hidden Layer-3)	R^2 (training set)	R^2 (test set)
1	64	128	128	0.959	0.96
2	64	128	64	0.945	0.946
3	64	64	64	0.938	0.946
4	32	64	64	0.927	0.942
5	16	64	64	0.912	0.925
6	8	64	64	0.888	0.918

Table S10. Actual and predicted ET(30) test set values along with the absolute error in the ascending order for 10% train test split using NN model.

S.No.	**Name of the Molecule**	**Actual**	**Predicted**	**Absolute Error**
1	ethyl benzoate	38	38.0	0
2	dichloromethane	40.7	40.8	0.1
3	methyl cis-9-octadecanoate	34.5	34.6	0.1
4	N,N-Dimethylpropionamide	41.8	41.9	0.1
5	ethyl-2-butynoate	40.1	40.3	0.2
6	bromotrichloromethane	34.9	35.1	0.2
7	1,2-dichlorohexafluorocyclobutane	33.3	33.0	0.3
8	fluorobenzene	37	36.7	0.3
9	diisobutyl ketone	38	38.4	0.4
10	1,2-dibromopropane	39.1	39.5	0.4
11	cyclohexanone	40.1	40.5	0.4
12	1-cyclohexylpyrrolidin-2-one	40.4	40.0	0.4
13	2-phenoxyethanol	51.5	51.9	0.4

14	1,3-difluorobenzene	37.3	36.8	0.5
15	2-butanol	47.1	47.6	0.5
16	1,3-dichloropropane	40.2	40.8	0.6
17	methyl pentanoate	36.6	37.2	0.6
18	diglyme	38.6	38.0	0.6
19	triglyme	38.9	39.5	0.6
20	3,4-xylenol	47.2	46.6	0.6
21	Di-n-propyl ether	34	34.8	0.8
22	1-formylpiperidine	41.7	42.6	0.9
23	tert-butyl methyl ether	34.5	35.5	1
24	dimethyl sulfane	36.8	35.8	1
25	sulfolane	44.1	43.1	1
26	2,5-dibromo-1-methylbenzene	34.7	35.9	1.2
27	tri-n-butyl-n-dodecylphosphonium bromide	44.5	43.3	1.2
28	N-methylformamide	54.1	52.9	1.2
29	N-(tert-butyl)benzylamine	34.9	36.2	1.3
30	quinoline	39.4	38.0	1.4
31	1-heptanol	48.5	47.1	1.4
32	N,N-diethylcyanamide	43.3	41.8	1.5
33	1,1,2-trichlorotriflouroethane	33.2	34.8	1.6
34	2-ethoxyethanol	51	49.4	1.6
35	2-chloropyridine	41.9	40.0	1.9
36	ethylene carbonate	48.6	46.7	1.9
37	1-phenylethanol	46.7	48.7	2
38	n-butylammonium thiocyanate	61.4	59.4	2
39	1,3-propanediol	54.9	52.7	2.2
40	trichloroflouromethane	33.2	35.5	2.3
41	N,N-dimethylthioformamide	44	41.4	2.6
42	N,N-diethylformamide	41.8	44.5	2.7
43	pyrrolidine	39.1	36.3	2.8
	Average			1.0

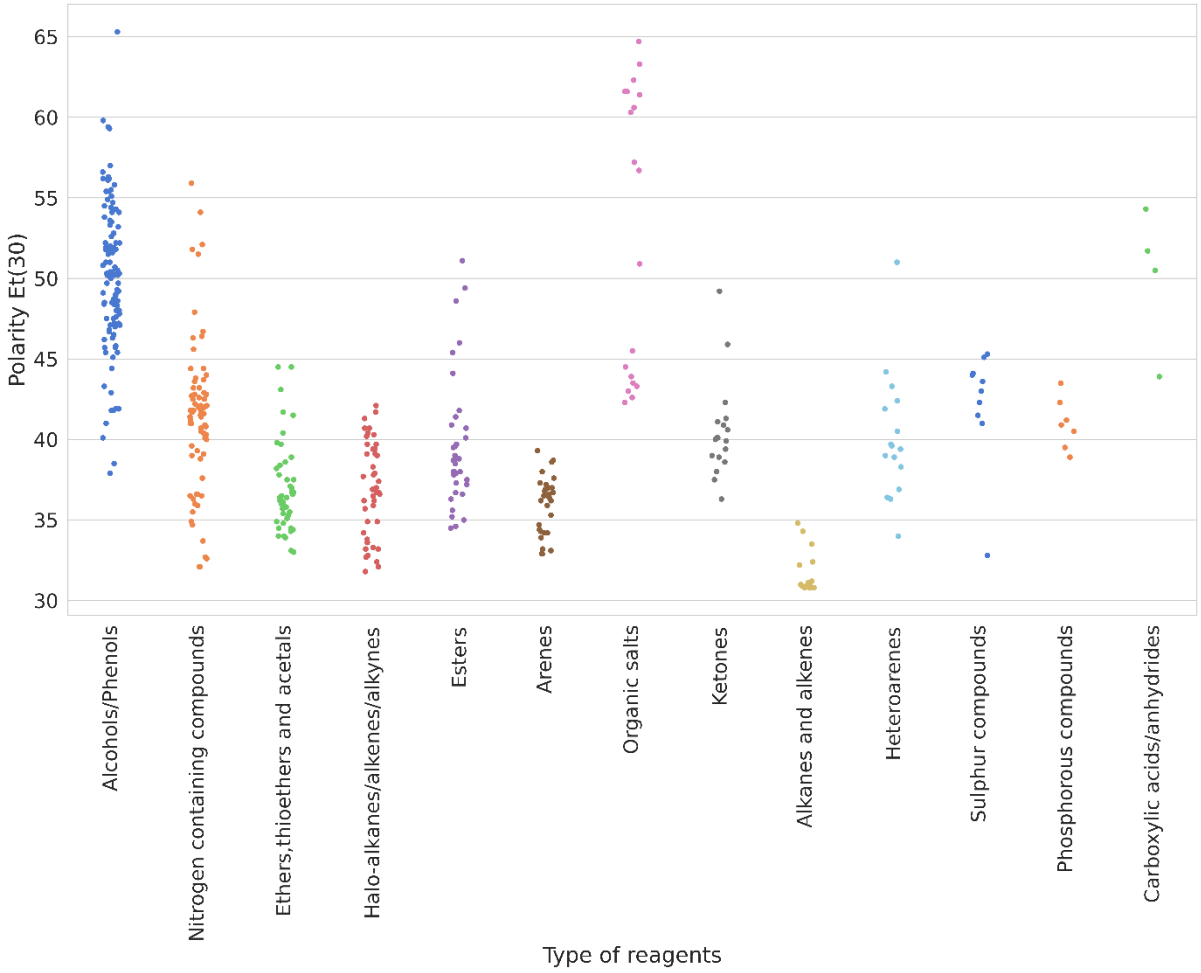

Figure S1. Strip plot depicting range of $E_T(30)$ values across 13 different categories of solvents.

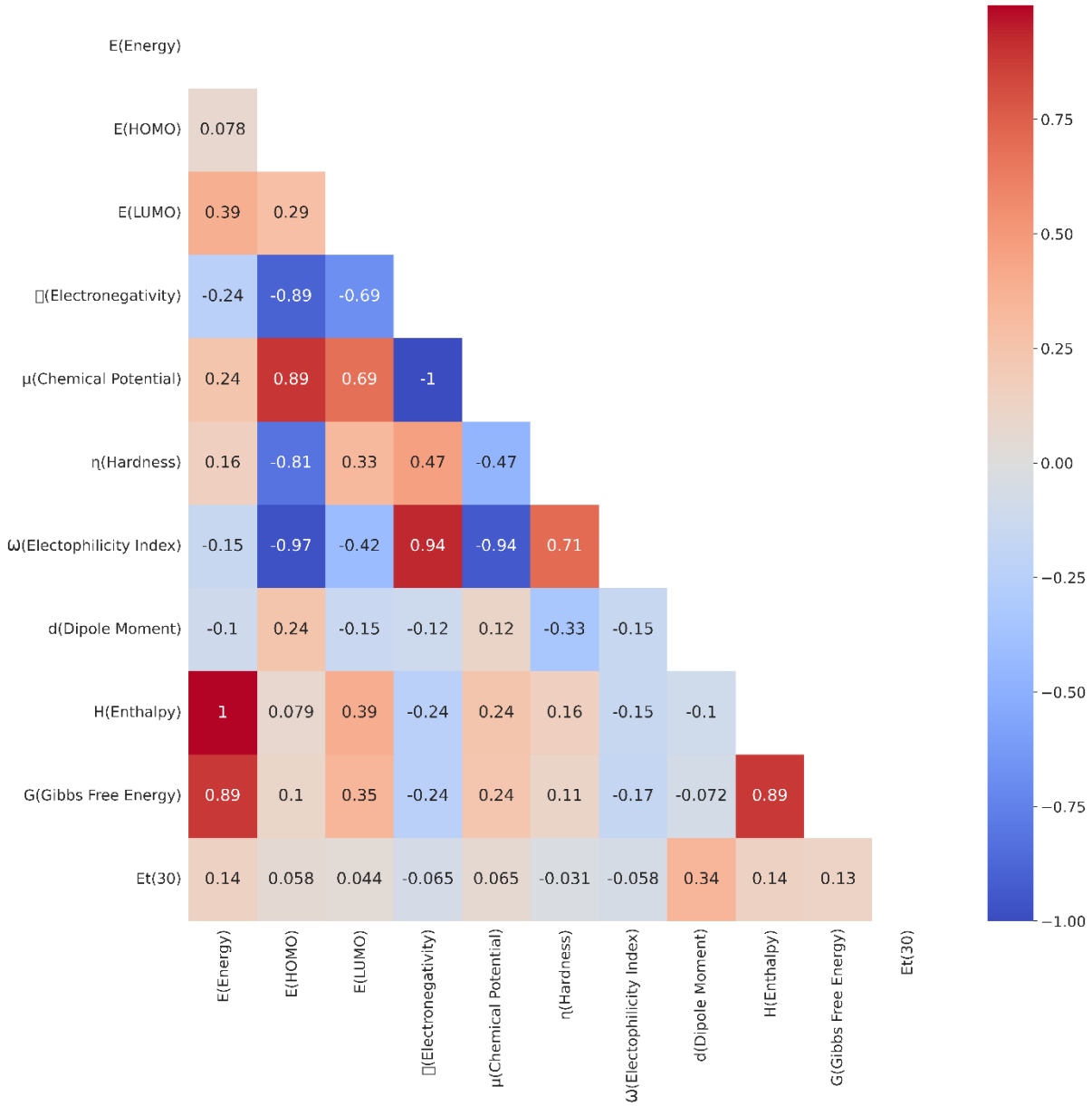

Figure S2. Confusion matrix showing Pearson coefficient (r) for 10 quantum chemical descriptors and $E_T(30)$.

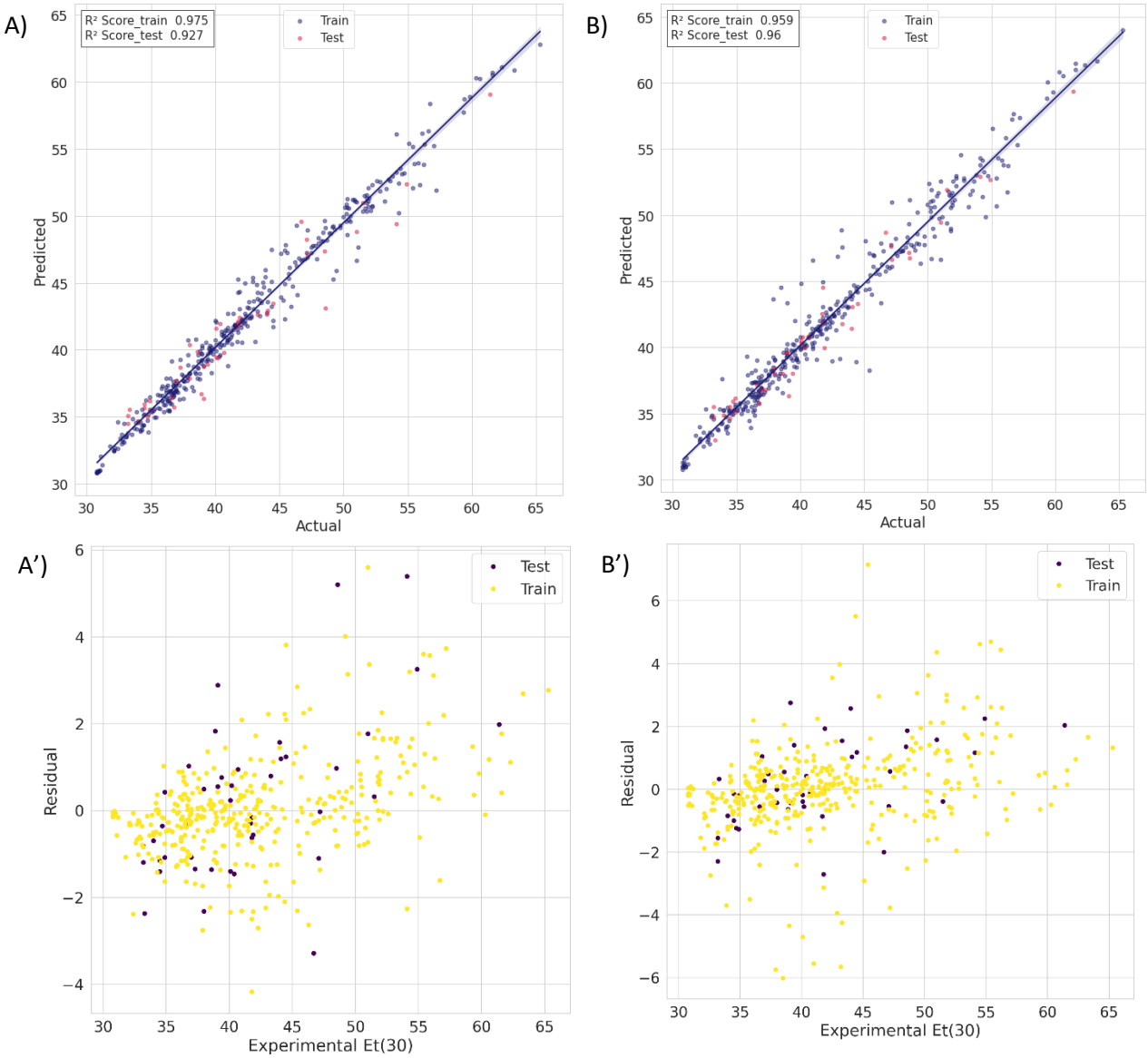

Figure S3. Regression and residual plots showing test and train data points and their R^2 values for; A) RF regression plot, B) NN regression plot, C) RF residual plot, D) NN residual plot.

References:

1. C. Reichardt, Solvatochromic Dyes as Solvent Polarity Indicators, *Chem. Rev.*, 1994, **94**, 2319-2358.
2. J. P. Cerón-Carrasco, D. Jacquemin, C. Laurence, A. Planchat, C. Reichardt and K. Sraïdi, Solvent polarity scales: determination of new ET(30) values for 84 organic solvents, *J. Phys. Org. Chem.*, 2014, **27**, 512-518.
3. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, Gaussian 16 Rev. C.01. Gaussian, Inc.: Wallingford CT, 2016.
4. M. Karelson, V. S. Lobanov and A. R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev.*, 1996, **96**, 1027-1044.