Electronic Supplementary Information

Quantitative structure-property relationship of the photoelectrochemical oxidation of phenolic pollutants at modified nanoporous titanium oxide using supervised machine learning

*Jesse S. Dondapati and Aicheng Chen**

Electrochemical Technology Center, Department of Chemistry, University of Guelph, Guelph, Ontario N1G 2W1, Canada.

* Corresponding Author. Email: aicheng@uoguelph.ca

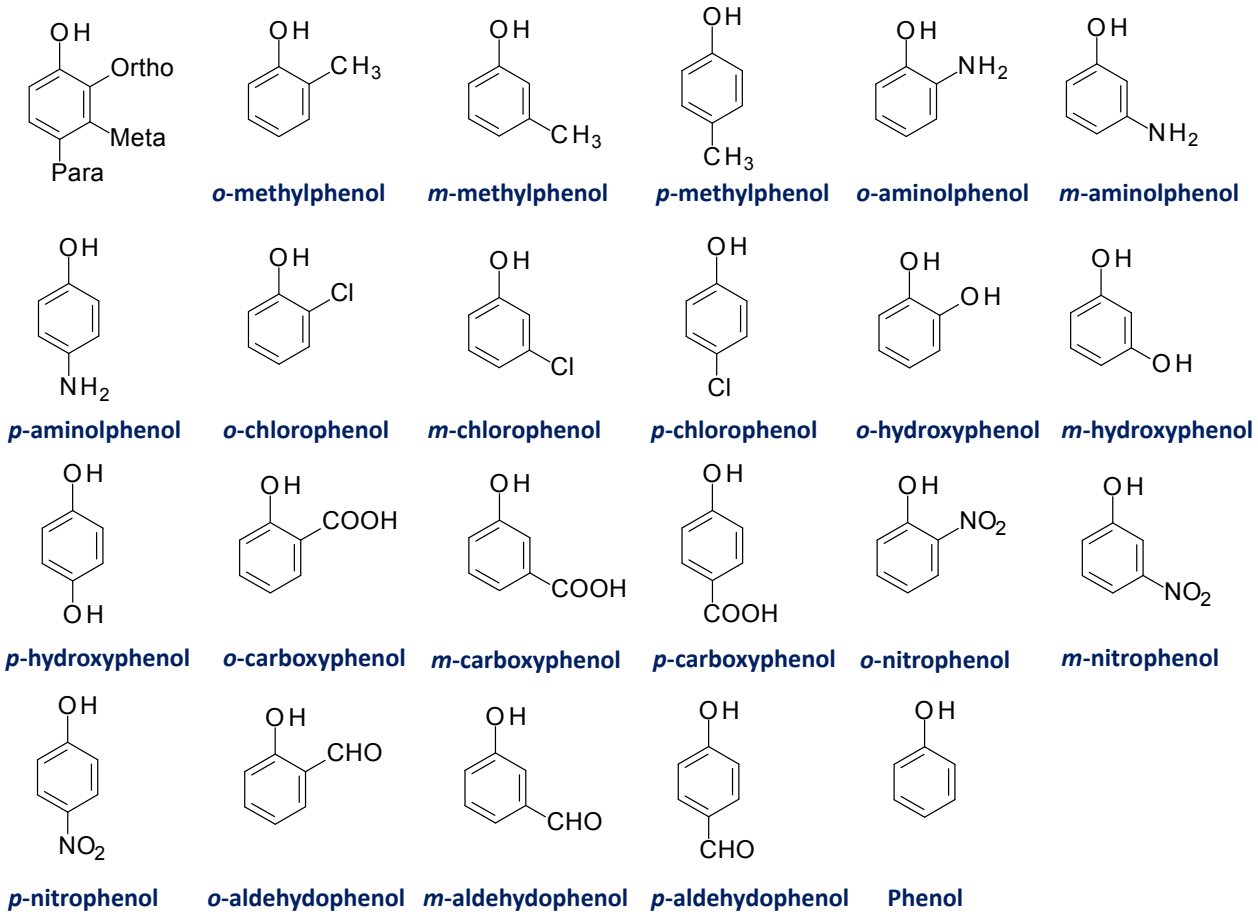


Figure S1: The set of 22 (o-, m-, p-) substituted phenolic compounds used for the experimental degradation rate constants and computational QSPR modelling study

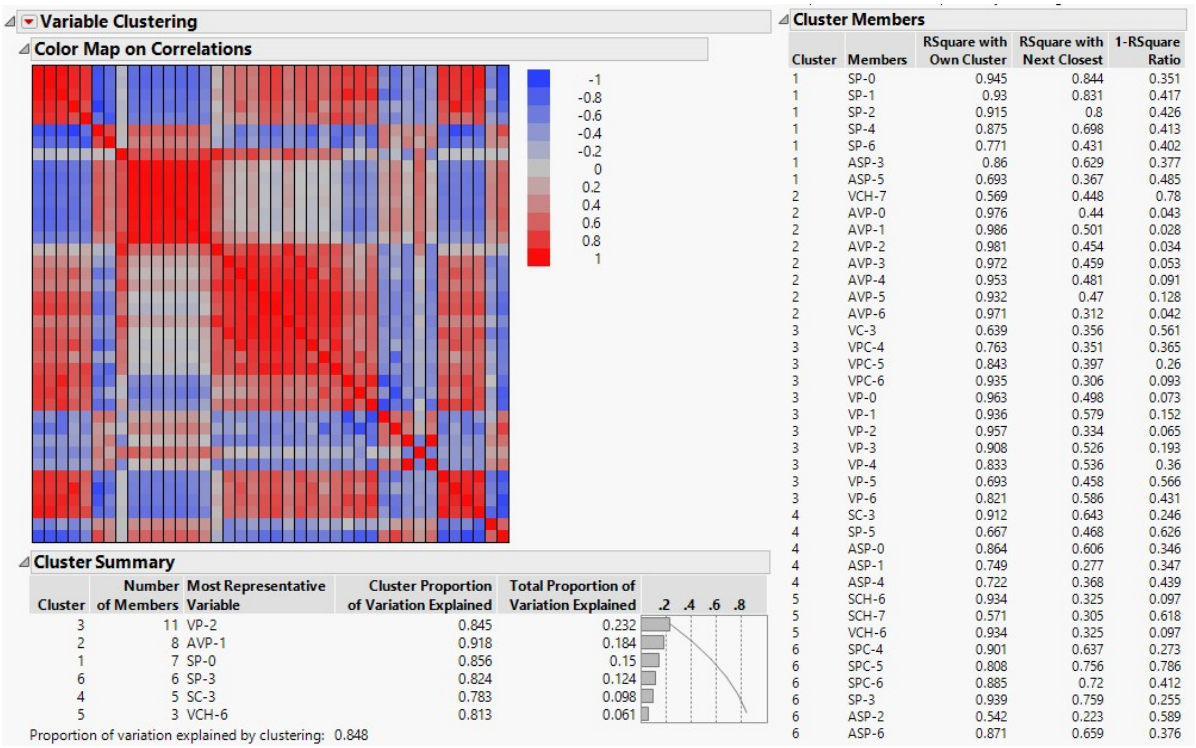


Figure S2: The variable clustering analysis of the connectivity descriptors (Kier and Hall Chi Path Cluster Indices) shows a high correlation (Red) among some of the descriptors.

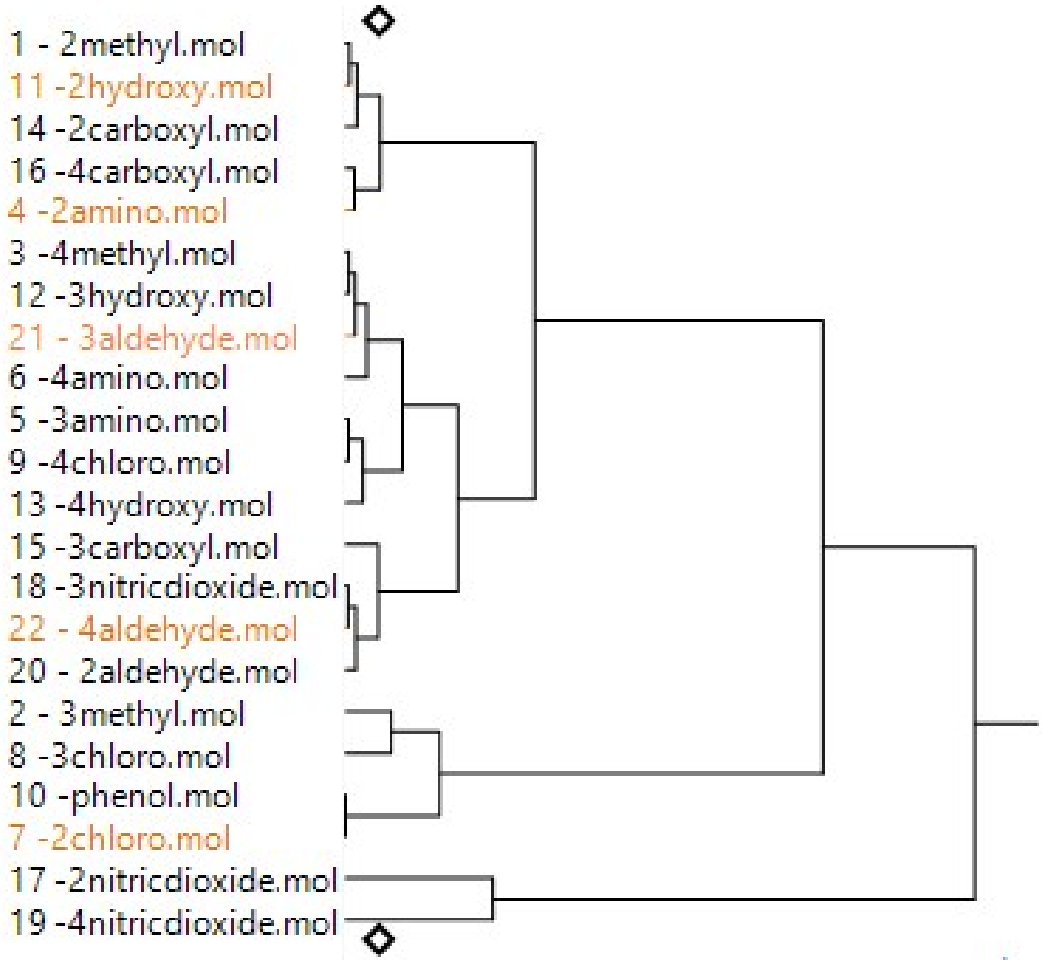


Figure S3: A dendrogram of the experimental rate constants of 22 phenolic compounds used in the study.

Note: The dendrogram was generated using the hierarchical clustering analysis (HCA) tool available in the JMP v13 statistical software. The HCA dendrogram is generated based on the distance or closeness of the degradation rate constants to each other. The dendrogram spans the range of slowest to the fastest degrading phenols. This tool is useful for selection of test set spanning the entire range as shown above.

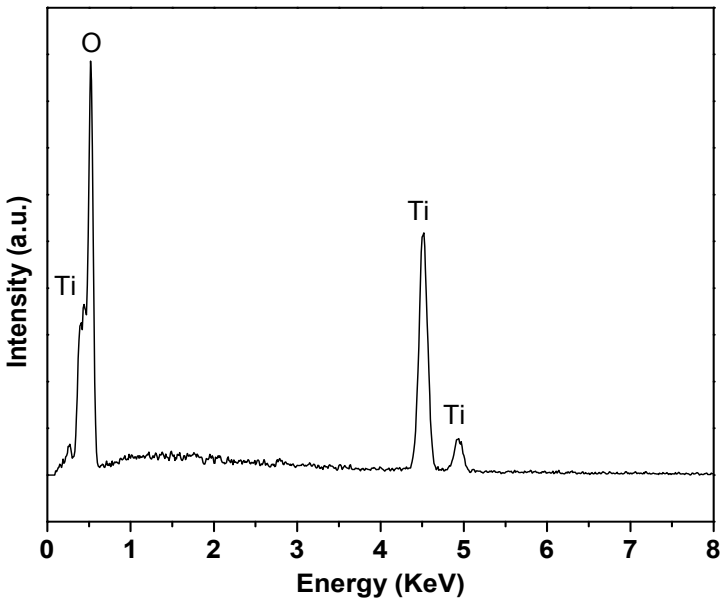


Figure S4: The energy dispersive X-ray spectroscopy (EDS) spectrum of the formed nanoporous TiO_2 electrode.

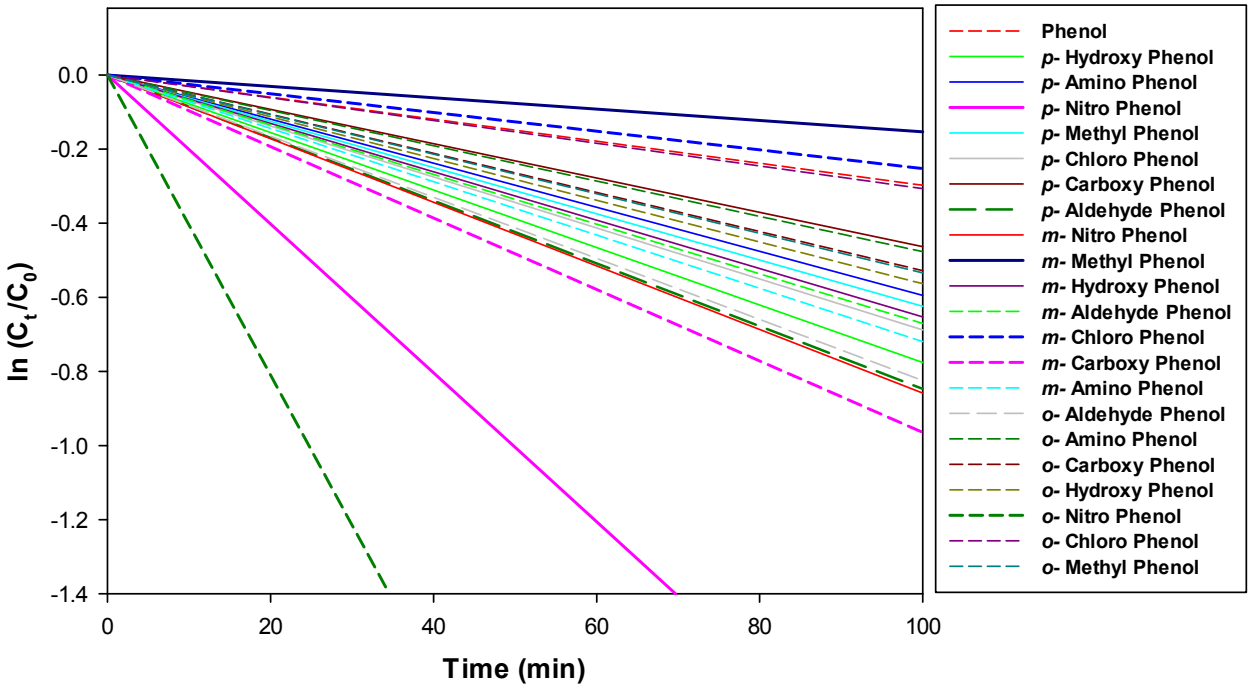


Figure S5: Typical kinetic plots representing the degradation rate, plotted as ln (C_t /C_0) vs. time of degradation for all the 22 phenolic compounds in the dataset.

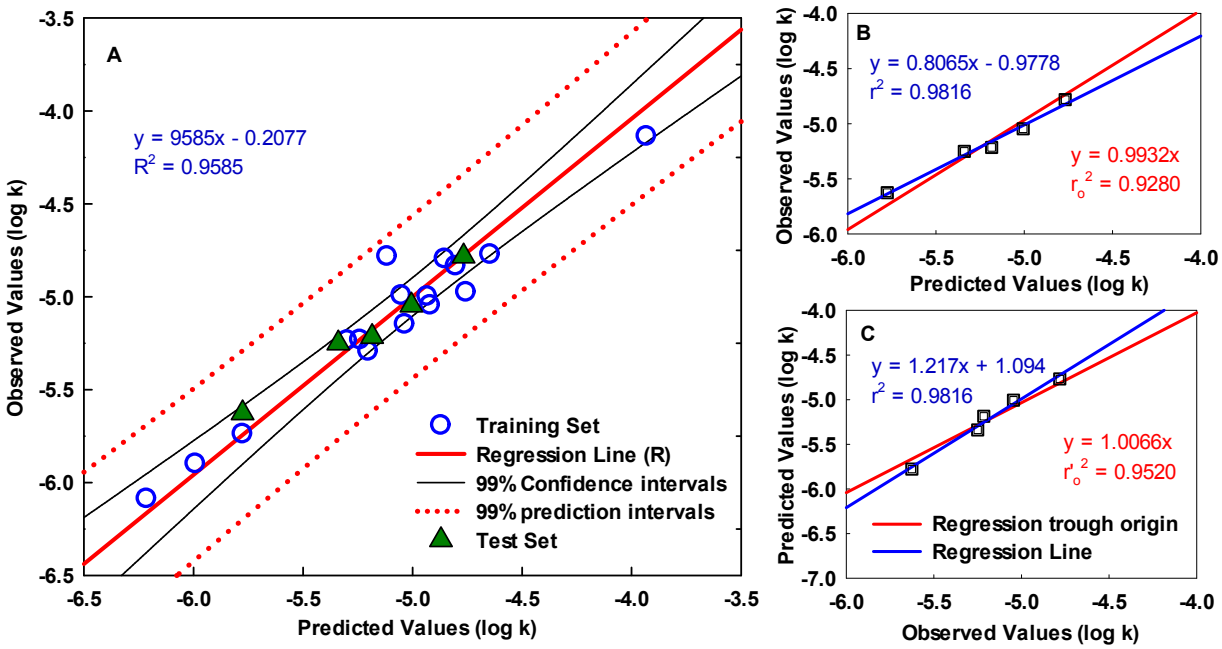


Figure S6: Collinearity of the data in QSPR Model I (A) Observed vs predicted values of training set overlaid with test set values; (B) Predicted vs observed plot with r^2 and r_o^2; (C) Observed vs predicted plot showing r^2 and r'^2_o corresponding to regression lines drawn from intercept and through origin, respectively.

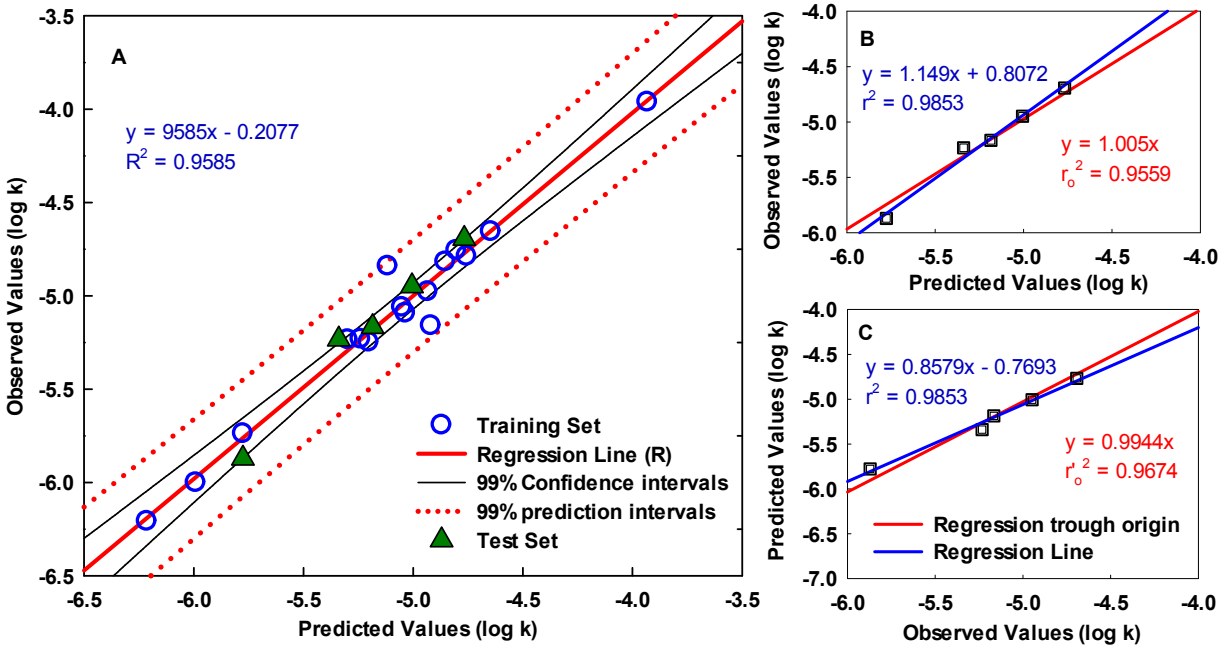


Figure S7: Collinearity of the data in QSPR Model II (A) Observed vs predicted values of training set overlaid with test set values; (B) Predicted vs observed plot with r^2 and r_o^2; (C) Observed vs predicted plot showing r^2 and r'^2_o corresponding to regression lines drawn from intercept and through origin, respectively.

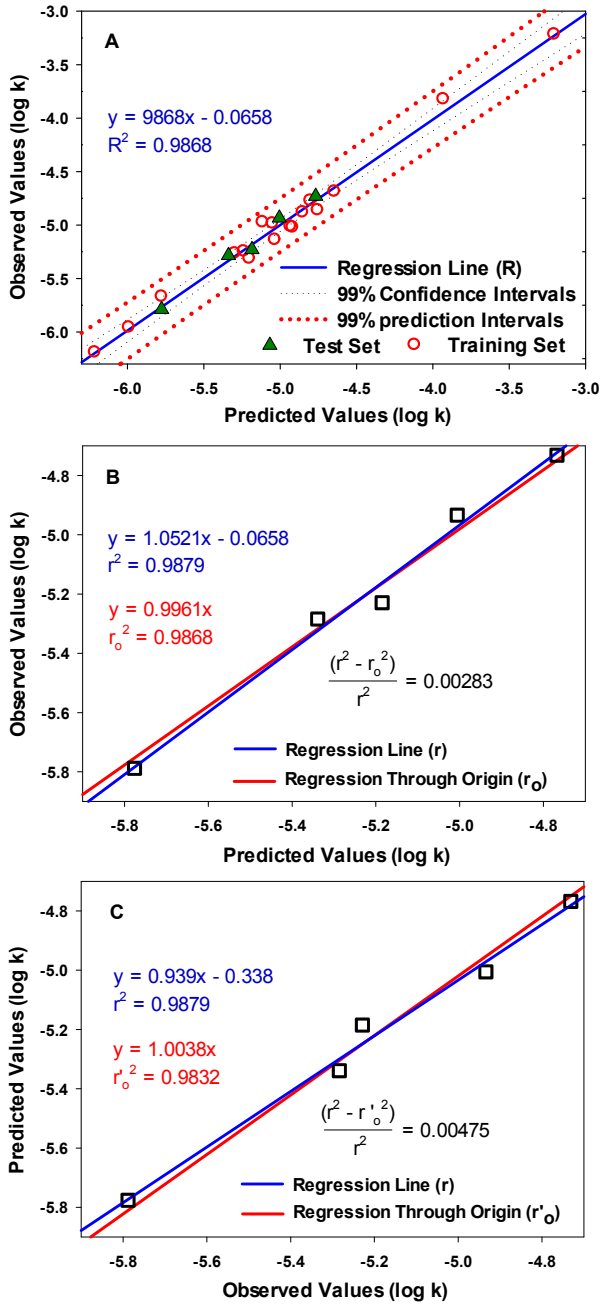
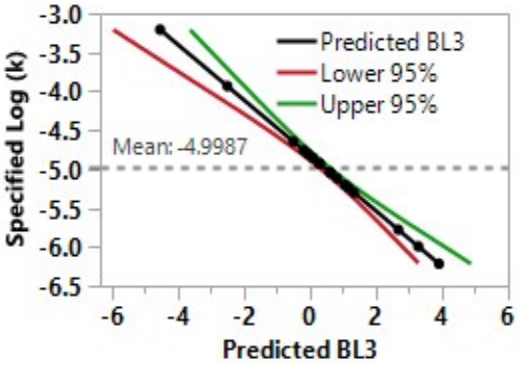
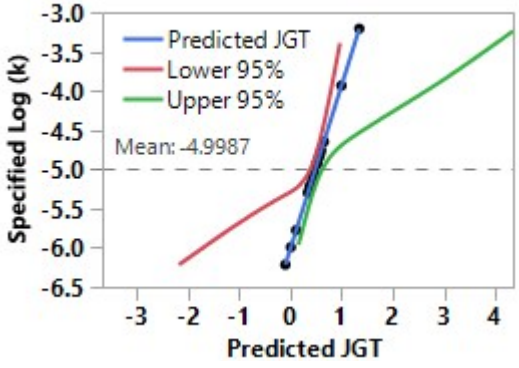
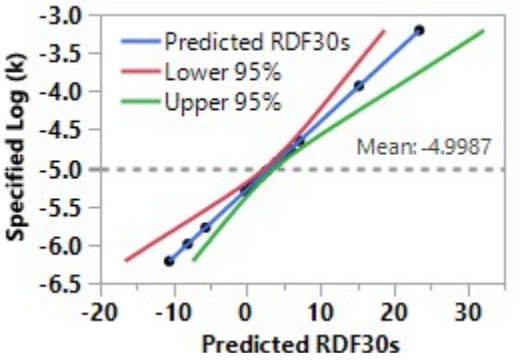
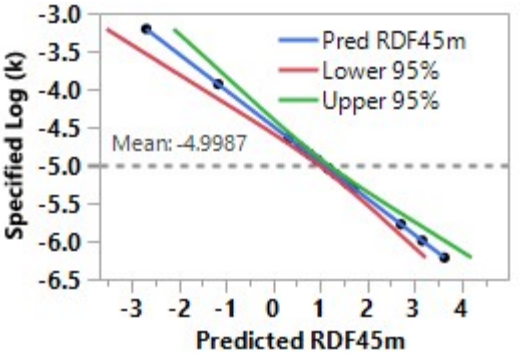


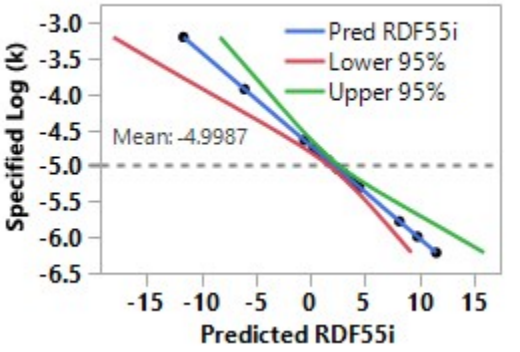
Figure S8: Collinearity of the data in QSPR Model III (A) Observed vs predicted values of training set overlaid with test set values; (B) Predicted vs observed plot with r^2 and r_o^2; (C) Observed vs predicted plot showing r^2 and r'^2_o corresponding to regression lines drawn from intercept and through origin, respectively.

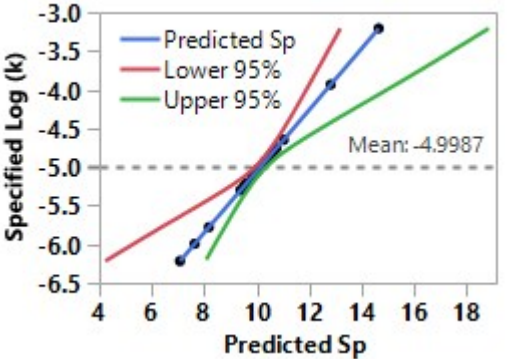

Predicted BL3
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted BL3

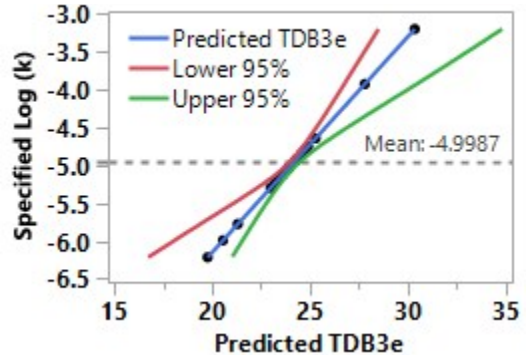

Predicted JGT
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted JGT


Predicted RDF30s
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted RDF30s


Pred RDF45m
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted RDF45m


Pred RDF55i
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted RDF55i


Predicted Sp
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted Sp


Predicted TDB3e
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted TDB3e

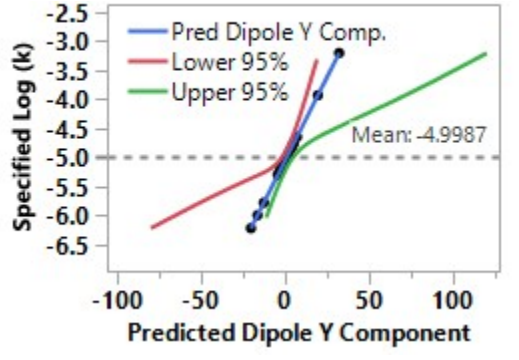

Pred Dipole Y Comp.
Lower 95%
Upper 95%
Mean: -4.9987
Specified Log (k)
Predicted Dipole Y Component

Figure S9: Lower and upper confidence intervals for the individual significant descriptors inversely predicted from their observed variables (Log k) and its mean value (-4.9987) intersection shows the least error on the predicted descriptors.

Table S1: The list of connectivity descriptor type clusters such as Chi Chain, Chi cluster, Chi path cluster and Chi path, that were included in the data prior to dimension reduction process.

Descriptor Type	No. of descriptors	Individual descriptors
Chi chain	10	SCH-3, SCH-4, SCH-5, SCH-6, SCH-7, VCH-3, VCH-4, VCH-5, VCH-6, VCH-7
Chi cluster	8	SC-3, SC-4, SC-5, SC-6, VC-3, VC-4, VC-5, VC-6
Chi path cluster	6	SPC-4, SPC-5, SPC-6, VPC-4, VPC-5, VPC-6
Chi path	32	SP-0, SP-1, SP-2, SP-3, SP-4, SP-5, SP-6, SP-7, ASP-0, ASP-1, ASP-2, ASP-3, ASP-4, ASP-5, ASP-6, ASP-7, VP-0, VP-1, VP-2, VP-3, VP-4, VP-5, VP-6, VP-7, AVP-0, AVP-1, AVP-2, AVP-3, AVP-4, AVP-5, AVP-6, AVP-7

Note: In the table S1, an example of untreated data relating to the molecular connectivity descriptor types in four clusters are presented. In our study, since the isomers of substituted phenols are structurally comparable to an extent, such clusters tend to generate highly correlated descriptors. Therefore, a high correlation filter method is adopted as one of the dimensionality reduction methods, which was employed using variable clustering tool of JMP v13. Most of the variables used in the study were originally calculated as groups or clusters as shown above. The list of the available descriptors type cluster that can be calculated with PaDEL Descriptors v2.0 software can be found at:
http://www.yapcwsoft.com/dd/padeldescriptor/Descriptors.xls

Table S2: An example of intercorrelations between the representative descriptor types selected from the variable clustering analysis is provided below.

	SC-3	**SP-1**	**VP-2**	**AVP-6**	**Sp**
SC-3	1				
SP-1	0.737	1			
VP-2	0.563	0.471	1		
AVP-6	-0.368	-0.594	0.393	1	
Sp	0.472	0.591	0.689	0.001	1

NOTE: The Variable clustering analysis identifies clusters of highly correlated descriptors as shown in Figure S2. We were able to select individual representative descriptors from each cluster to represent each cluster type and the overall intercorrelation is maintained below 0.75.

Table S3: List of descriptors used in building the QSPR Models I to IV.

S.No	**Descriptor Types**	**Number**	**Descriptors used in this study**
1	Bond Lipophilicity (Subst. x)	4	BL (S1), BL (S2), BL (S3), and BL (S4)

2	Dipole	4	Dipole components (x, y and z), Total Dipole
3	Bond Dipole (Subst. x)	4	BD (S1), BD (S2), BD (S3), and BD (S4)
4	Bond Molar Refractivity	4	BMR (S1), BMR (S2), BMR (S3), and BMR (S4)
5	Quantum descriptors	4	HOMO, LUMO, Electronic energy, Mean polarizability
6	Quantum derived descriptors	3	Chemical potential (μ), hardness (η), Electrophilicity index (ω)
7	3D autocorrelation	4	TDB1e, TDB2e, TDB3e, and TDB4e
8	Charged partial surface area	3	RPCG, WNSA-2 and WNSA-3
9	Molecular linear free energy relation	3	MLFER_A, MLFER_E, MLFER_L
10	Length over breadth	2	LOBMIN, LOBMAX
11	RDF	4	RDF45m, RDF20e, RDF55i, RDF30s
12	WHIM	7	Ti, Ai, Ki, Di, L1s, L2s, and L3s
13	Chi path cluster	2	SPC-4, SPC-5,
14	Chi path	3	SP-1, SP-5, ASP-6
15	Chi Cluster	3	SC-3, SC-4, SC-5
16	Constitutional	1	Sp
17	Crippen logP and MR	2	CrippenLogP, CrippenMR
18	Detour matrix	1	SpMAD_Dt
19	Information content	3	SIC0, SIC1, SIC2
20	Topological charge	1	JGT
21	Topological distance matrix	1	SpMAD_D

Table S4 (a): The complete set of descriptors and 22-member dataset comprising phenolic compounds and experimental the rate constants (Log k) used for the development of Model I and II.

Substituted Phenols	Log (k)	BL1	BL2	BL3	BL4	TDN3e	RPCG	LOBMIN	RDF45m	RDF20e	RDF55i	RDF30s	Ti	Ki	Di	L2s
o-CH_3	-5.20301	0.10505	1.48521	0.36253	0.362736	22.9	0.34701	1.07162	0.41398	3.21731	1.93146	0.2338	5.50297	0.46919	1.32577	1.87669
m-CH_3	-6.21461	0.108225	0.362003	1.4847	0.362578	22.9203	0.34703	1.43382	0.28771	3.13776	7.38264	0.14948	5.78351	0.47069	1.30471	2.08197
p-CH_3	-5.05146	0.108135	0.36227	0.362533	1.48455	22.9202	0.34712	1.27242	0.45179	3.13722	2.45102	0.14957	5.9786	0.55018	1.21035	1.56327
o-NH_2	-5.33914	0.150888	-0.80134	0.3621	0.362606	23.3764	0.25602	1.10092	0.56208	4.53568	1.69361	0.68491	5.19918	0.48844	1.32033	1.94029
m-NH_2	-4.93367	0.109288	0.361585	-0.79758	0.362173	23.3362	0.26853	1.39658	0.4071	4.01727	2.51527	0.6232	5.39532	0.48957	1.30238	2.15584
p-NH_2	-5.116	0.108804	0.362176	0.36214	-0.79786	23.3361	0.26975	1.2229	0.59555	4.01674	0.72863	0.62334	5.5591	0.52701	1.17385	1.60045
o-Cl	-5.77635	0.114324	1.68388	0.362427	0.362816	24.9215	0.32525	1.11165	2.28996	2.13941	1.93128	2.33847	4.88372	0.49999	1.60893	1.89603
m-Cl	-5.99146	0.106971	0.361983	1.68282	0.362472	24.7482	0.34881	1.4732	1.30792	2.46152	2.13714	0.23344	4.90263	0.5	1.47488	1.99518
p-Cl	-4.91988	0.106583	0.362264	0.362432	1.68275	24.7482	0.35106	1.28392	1.41478	2.46097	0.00261	0.23335	4.92005	0.49999	1.42357	1.7685
-H	-5.77635	0.10808	0.362165	0.362727	0.36273	22.7875	0.35981	1.20759	0.24993	2.76878	0.75947	0.0801	4.63175	0.49999	1.53852	1.76405
o-OH	-5.18499	0.126617	0.109771	0.362198	0.362765	24.2802	0.24403	1.20695	0.49113	2.53381	1.19883	0.39129	5.02629	0.49999	1.7268	2.02493
m-OH	-5.03595	0.106023	0.361651	0.106024	0.362321	24.1634	0.25825	1.34714	0.32526	2.81634	1.57542	0.13554	5.10516	0.5	1.66157	2.24982
p-OH	-4.85363	0.107208	0.362205	0.362205	0.107207	24.1634	0.25964	1.22363	0.51912	2.8158	0.41855	0.13554	5.22875	0.49999	1.65067	1.62881
o-COOH	-5.24005	0.102785	-0.58883	0.362779	0.362808	24.3135	0.256	1.04066	2.72253	4.12489	3.76535	6.60291	5.68471	0.44381	1.42122	1.87313
m-COOH	-4.64599	0.106764	0.362304	-0.58401	0.362758	24.3432	0.26271	1.47453	2.0119	3.13533	3.28196	5.7106	5.93536	0.44462	1.36603	2.05147
p-COOH	-5.29832	0.106089	0.362227	0.362758	-0.58492	24.3472	0.26336	1.3828	3.15428	3.12154	2.19847	4.03765	6.14505	0.5339	1.30814	1.52987
o-NO_2	-3.20645	0.100699	2.63074	0.362868	0.362946	25.3895	0.27629	1.0488	1.43063	2.12751	1.53086	16.7132	5.65317	0.4537	1.33242	1.8755
m-NO_2	-4.75599	0.105367	0.362431	2.6019	0.362903	25.3493	0.29613	1.55436	0.80684	2.44632	2.73528	6.08372	5.98548	0.45612	1.2773	1.87878
p-NO_2	-3.93223	0.104488	0.362394	0.362885	2.59969	25.3491	0.29791	1.60773	0.78667	2.44563	2.44495	6.09304	6.06514	0.53151	1.197	1.56303
o-CHO	-4.80362	0.093176	-0.50968	0.36293	0.362816	23.6502	0.25402	1.06189	1.46844	4.41069	0.94638	5.90671	5.38256	0.47291	1.3548	1.86311
m-CHO	-5.00565	0.107083	0.362342	-0.50737	0.362657	23.6766	0.26523	1.47189	1.28852	3.7977	3.35894	4.29159	5.61877	0.47383	1.32579	2.08518
p-CHO	-4.76769	0.105248	0.3622	0.362605	-0.50807	23.6765	0.26621	1.11116	1.47247	3.79701	0.44234	4.30604	5.75496	0.52338	1.19806	1.56202

Table S4 (b): The complete set of descriptors and 22-member dataset comprising phenolic compounds and experimental the rate constants (Log k) used for the development of Model I and II (continuation to table S4a).

Substituted Phenols	Log (k)	SPC-4	SPC-5	SP-1	SP-5	Sp	SpMAD_Dt	SIC0	SIC1	SIC2	JGT	SpMAD_D
o-CH_3	-5.20301	1.13807	1.04044	3.80453	1.02438	10.6745	7.78845	0.31795	0.67571	0.76946	0.4566	3.82666
m-CH_3	-6.21461	0.74158	1.09341	3.78769	0.97141	10.6745	7.72933	0.31795	0.67571	0.80071	0.46217	3.89267
p-CH_3	-5.05146	0.8165	0.76008	3.78769	1.30474	10.6745	7.67538	0.31795	0.67571	0.73821	0.52571	3.97205
o-NH_2	-5.33914	1.13807	1.04044	3.80453	1.02438	9.93386	7.78845	0.40001	0.69285	0.79523	0.4566	3.82666
m-NH_2	-4.93367	0.74158	1.09341	3.78769	0.97141	9.93386	7.72933	0.40001	0.69285	0.82936	0.46217	3.89267
p-NH_2	-5.116	0.8165	0.76008	3.78769	1.30474	9.93386	7.67538	0.40001	0.69285	0.76111	0.52571	3.97205
o-Cl	-5.77635	1.13807	1.04044	3.80453	1.02438	9.78202	7.78845	0.43625	0.6674	0.75055	0.4566	3.82666
m-Cl	-5.99146	0.74158	1.09341	3.78769	0.97141	9.78202	7.72933	0.43625	0.6674	0.79212	0.46217	3.89267
p-Cl	-4.91988	0.8165	0.76008	3.78769	1.30474	9.78202	7.67538	0.43625	0.6674	0.75055	0.52571	3.97205
-H	-5.77635	0.40825	0.43301	3.39385	0.90105	8.8759	7.37476	0.35518	0.51733	0.61825	0.34702	3.47727
o-OH	-5.18499	1.13807	1.04044	3.80453	1.02438	9.35614	7.78845	0.38053	0.58727	0.66231	0.4566	3.82666
m-OH	-5.03595	0.74158	1.09341	3.78769	0.97141	9.35614	7.72933	0.38053	0.58727	0.69983	0.46217	3.89267
p-OH	-4.85363	0.8165	0.76008	3.78769	1.30474	9.35614	7.67538	0.38053	0.58727	0.58727	0.52571	3.97205
o-COOH	-5.24005	1.49827	2.17456	4.71521	1.34172	10.8364	8.49743	0.37631	0.6875	0.8125	0.47576	4.68875
m-COOH	-4.64599	1.38961	1.52281	4.69838	1.5721	10.8364	8.38223	0.37631	0.6875	0.84375	0.48957	4.81311
p-COOH	-5.29832	1.44869	1.40419	4.69838	1.44914	10.8364	8.27711	0.37631	0.6875	0.78125	0.54043	4.95891
o-NO_2	-3.20645	1.49827	2.17456	4.71521	1.34172	10.0958	8.49743	0.4561	0.72698	0.82936	0.47576	4.68875
m-NO_2	-4.75599	1.38961	1.52281	4.69838	1.5721	10.0958	8.38223	0.4561	0.72698	0.86349	0.48957	4.81311
p-NO_2	-3.93223	1.44869	1.40419	4.69838	1.44914	10.0958	8.27711	0.4561	0.72698	0.79523	0.54043	4.95891
o-CHO	-4.80362	1.06904	1.40829	4.34254	1.20977	10.3561	8.27027	0.36589	0.66538	0.82936	0.43262	4.37183
m-CHO	-5.00565	0.8371	1.14666	4.3257	1.33122	10.3561	8.17583	0.36589	0.66538	0.86349	0.43808	4.4755
p-CHO	-4.76769	0.90105	0.93744	4.3257	1.37377	10.3561	8.08958	0.36589	0.66538	0.79523	0.47539	4.59788

Table S5: The additional set of descriptors along with the descriptors presented in table S7 that were used for building Model III and IV

Substituted Phenols	LUMO	HOMO	Total Dipole Moment (WM)	Dipole x component (WM)	Dipole Y Component (WM)	Dipole Z Component (WM)
o-CH_3	0.37027	-8.99693	1.60388	0.408545	-1.55097	-0.0048502
m-CH_3	0.52257	-8.5298	1.81381	-0.878612	-1.58637	-0.0372399
p-CH_3	0.439251	-8.26984	1.5766	0.404532	-1.52358	0.0270806
o-NH_2	0.406952	-8.64184	1.3284	0.75396	-1.07911	-0.178087
m-NH_2	0.018625	-9.33501	2.41256	-0.942763	-1.77468	1.33498
p-NH_2	0.392482	-9.02094	2.2283	-1.5815	-0.512377	1.48379
o-Cl	0.029431	-9.25957	1.18905	0.603481	1.02452	5.67203E-05
m-Cl	0.430239	-8.88123	2.2783	2.01121	-1.07037	0.000960285
p-Cl	0.09462	-9.12453	1.32995	-1.3279	-0.073789	0.000092587
-H	-1.01333	-9.95381	1.30016	-0.663059	1.11837	-0.000366983
o-OH	0.29703	-8.885	2.20221	1.92212	-1.0748	-0.000709248
m-OH	-1.15932	-9.94672	1.27095	-1.27091	-0.00981126	-6.31183E-05
p-OH	-1.06522	-10.0725	2.20999	-1.92527	1.08507	-0.00018432
o-COOH	0.397365	-9.11437	2.63543	2.57548	0.55893	-0.000358723
m-COOH	0.274362	-8.98203	1.61243	1.56308	0.395874	-0.000195338
p-COOH	0.232847	-8.73425	3.03018	-0.857647	2.90627	-0.000429832
o-NO_2	-0.59046	-9.46227	6.75076	2.9346	6.07954	0.00126674
m-NO_2	-0.57358	-9.5175	6.34853	4.04968	4.88917	-0.00102094
p-NO_2	-0.48087	-9.60834	5.42925	1.6224	5.18118	-7.59699E-05
o-CHO	-0.57035	-9.49494	1.70204	-0.0223051	1.70181	0.0174728
m-CHO	-0.5396	-9.4456	1.95347	1.9438	-0.194156	0.000456167
p-CHO	-0.44573	-9.49295	3.42383	-1.67525	2.986	0.00190441

JMP Script for building Model I.

The data required for the script to run in JMPv13 is included in table S4(a) and table S4(b) (includes column names as well).

```
Fit Group(
	Fit Model(
		Y( :Name( "Log (k)" ) ),
		Effects( :BL3, :RDF45m, :RDF55i, :RDF30s, :JGT ),
		Personality( "Standard Least Squares" ),
		Emphasis( "Minimal Report" ),
		Run(
			:Name( "Log (k)" ) << {Summary of Fit( 1 ), Analysis of
Variance( 1 ),
			Parameter Estimates( 1 ), Lack of Fit( 0 ),
			Plot Actual by Predicted( 0 ), Plot Regression( 0 ),
			Plot Residual by Predicted( 0 ), Plot Studentized Residuals( 0
),
			Plot Effect Leverage( 0 ), Box Cox Y Transformation( 0 )}
		)
	)

);
```

Table S6: Significant descriptors of the selected QSPR models with high predictive ability.

	Descriptors	Abbreviated	QSPR Model IV	Rh B	*o*- NO_2	*p*- NO_2	*o*- COOH
1.	Bond Lipophilicity on substituent 3	BL S3	- 0.3567	-0.6746	0.3629	0.3629	0.3627
	Radial distribution function weighted by mass	RDF45m	- 0.4758	14.34	1.430	0.78667	2.723
2.	Radial distribution function weighted by first ionization potential	RDF55i	- 0.1299	24.57	1.530	2.445	3.765
3.	Radial distribution function weighted by ionization state	RDF30s	+ 0.0833	17.49	16.71	6.093	6.603
	Global Topological index	JGT	+ 2.089	0.4308	0.4758	0.5404	0.4758
4.	Simple path, order 1	SP-1	+ 0.3973	16.02	4.715	4.698	4.715
5.	Bond Lipophilicity on substituent 4	BL S4	-	2.808	0.3629	2.599	0.3628
6.	Spectral mean absolute deviation from topological distance matrix	SpMAD_D	-	11.27	8.497	8.277	8.497
7.	3D topological distance-based autocorrelation - lag 3 / weighted by Sanderson electronegativities	TDB3e	+ 0.2843	23.59	25.39	25.35	24.31
8.	Dipole Y - component	Dipole (y)	+ 0.0575	-20.99	6.079	5.181	0.5589

Table S7: Predictive ability evaluation of the selected QSPR models via conditions delineated in Eqs. 5 to 8.

Model	v*	Eq. 5	Eq. 6		Eq. 7		Eq. 8	
		R^2_{adj}	$Q^2_{cv\text{-}tr}$	$Q^2_{cv\text{-}test}$	$\frac{\left(r^2 - r_o^2\right)}{r^2}$	$\frac{\left(r^2 - r_o'^2\right)}{r^2}$	k	k'
I	5	0.9396	0.9680	0.9780	0.0537	0.0299	1.006	0.994
II	6	0.9692	0.9936	0.9759	0.0298	0.0182	0.994	1.006
III	7	0.9766	0.9991	0.9896	0.0028	0.0048	1.004	0.996
IV	8	0.9868	0.9993	0.8741	0.0735	0.0278	1.014	0.985

Table S8: The k-fold cross validation (k=5, 10, 15, 20) and Leave-One-Out (LOO) method by applying k=22 for the entire dataset of 22 phenolic compounds.

	Model I	**Model II**	**Model III**	**Model IV**
k=5	$R^2_{\text{k-fold}}$ =0.8805	$R^2_{\text{k-fold}}$ =0.9513	$R^2_{\text{k-fold}}$ =0.9538	$R^2_{\text{k-fold}}$ =0.9432
	RMSE = 0.1492	RMSE=0.1073	RMSE=0.0898	RMSE=0.0898
k=10	$R^2_{\text{k-fold}}$ =0.8580	$R^2_{\text{k-fold}}$ =0.9651	$R^2_{\text{k-fold}}$ =0.9339	$R^2_{\text{k-fold}}$ =0.9637
	RMSE = 0.1492	RMSE=0.1073	RMSE=0.0898	RMSE=0.0898
k=15	$R^2_{\text{k-fold}}$ =0.8602	$R^2_{\text{k-fold}}$ =0.9617	$R^2_{\text{k-fold}}$ =0.9519	$R^2_{\text{k-fold}}$ =0.9570
	RMSE = 0.1492	RMSE=0.1073	RMSE=0.0898	RMSE=0.0898
k=20	$R^2_{\text{k-fold}}$ =0.8541	$R^2_{\text{k-fold}}$ =0.9638	$R^2_{\text{k-fold}}$ =0.9425	$R^2_{\text{k-fold}}$ =0.9553
	RMSE = 0.1492	RMSE=0.1073	RMSE=0.0898	RMSE=0.0898
k=22	$R^2_{\text{k-fold}}$ =0.9456	$R^2_{\text{k-fold}}$ =0.9625	$R^2_{\text{k-fold}}$ =0.9541	$R^2_{\text{k-fold}}$ =0.9369
	RMSE = 0.1492	RMSE=0.1073	RMSE=0.0898	RMSE=0.1405

Note: Leave-One-Out (LOO) methodology was carried out using the k-fold cross validation (when k=22) for the entire dataset with 22 members. The k-fold cross validation is available at JMPv13 statistical software.