Electronic Supplementary Material (ESI) for RSC Advances.

Electronic Supplementary Information

Predicting Emission Wavelength of Organic Molecules Using a Combinatorial QSAR and Machine Learning Approach

Zong-Rong Ye,[1] I-Shou Huang,[1,2] Yu-Te Chan,[1,3] Zhong-Ji Li, Chen-Cheng Liao,[1] Hao-Rong Tsai,[1] Meng-Chi Hsieh,[1] Chun-Chih Chang,[1,4] Ming-Kang Tsai[1,*]

[1] Department of Chemistry, National Taiwan Normal University, Taipei, 11677, Taiwan
[2] Department of Chemistry, The University of Chicago, Chicago, IL 60637, US
[3] Theoretical Chemistry and Catalysis Research Center, Technical University Munich, Lichtenbergstr 4, 85747 Garching, Germany
[4] Department of Chemical and Materials Engineering, Chinese Culture University, Taipei, 11114, Taiwan

Email: mktsai@ntnu.edu.tw

Details for the Dataset Generation

The 11460 fluorescent organic molecules were extracted from Reaxys database on June 1, 2018. There were more than 75000 compounds accompanied with the fluorescence spectra in the database without applying any filtering, however, only 11000-plus out of 75K compounds were documented with the numerical values for the emission wavelength (the wavelengths for the rest of the compounds could be determined manually with visualizing the corresponding spectrum). With the removal of the compounds containing the wavelength less than 100 nm, the current 11460 dataset was finally achieved.

Determination of the Elbow Point (Figure 5a)

The elbow point at $k = 7$ is selected due to the consistent variation subject to the increment of k than the neighboring data points.

k	I(k)	A = 0.5*[I(k-1) – I(k+1)]	B = I(k) – I(k-1)	B/A	C = I(k) – I(k+1)	C/B
2	1.11E+16					
3	3.57E+15	4.57E+15	7.53E+15	1.65	1.61E+15	0.35
4	1.96E+15	1.18E+15	1.61E+15	1.36	7.50E+14	0.64
5	1.21E+15	5.98E+14	7.50E+14	1.25	4.46E+14	0.75
6	7.64E+14	3.22E+14	4.46E+14	1.39	1.97E+14	0.61
7	**5.67E+14**	**1.88E+14**	**1.97E+14**	**1.05**	**1.79E+14**	**0.95**
8	3.88E+14	1.39E+14	1.79E+14	1.29	9.80E+13	0.71
9	2.90E+14	7.50E+13	9.80E+13	1.31	5.20E+13	0.69

1. I(k) denotes the inertia score with the k partition.

Summary of the Regression Models

Along the progress of descriptor space reduction, we have generated four descriptor ensembles, i.e. VTS, VTSsel, VTSfp, and Lasso. Five fitting models using these descriptor ensembles are summarized in Table S3 in addition to the statistical indicators, i.e. R^2, adjR^2, Akaike's information criterion (AIC), the number of descriptors with large p-value (p-value > 0.03), mean absolute error against the experimental values (MAE), for the justification of regression results. Four models give good fitting results with $R^2 \geq 0.86$ and adj$R^2 \geq 0.80$ except the Lasso-LR case. However, the VTS series are found to contain high uncertainty caused by overfitting being indicated by AIC and high percentile of descriptors with the p-value > 0.03. Two Lasso-based

models are confirmed to be statistically meaningful results while the nonlinear RF regression method gives the best results. Despite the Lasso-LR model gives less predictivity than its RF counterpart, the descriptor contribution of LR is more interpretable.

Table S1. The complete list of the categories of the descriptors generated by PaDEL	
Acidic group count	Largest Pi System Descriptor
ALOGP	Longest Aliphatic Chain Descriptor
Amino Acid Count Descriptor	Mannhold LogP Descriptor
APol	McGowan Volume Descriptor
Aromatic atoms count	MDE Descriptor
Aromatic bonds count	MLFER Descriptor
Atom count	Path Count Descriptor
Autocorrelation Descriptor	Petitjean Number Descriptor
BaryszMatrix Descriptor	Ring Count Descriptor
Basic Group Count Descriptor	PaDEL Rotatable Bonds Count Descriptor
BCUT Descriptor	Rule Of Five Descriptor
PaDEL Bond Count Descriptor	Topological Descriptor
BPol Descriptor	Topological Charge Descriptor
Burden Modified Eigenvalues Descriptor	Topological Distance Matrix Descriptor
PaDEL Carbon Types Descriptor	TPSA Descriptor
Chi Chain Descriptor	VABC Descriptor
Chi Cluster Descriptor	VAdjMa Descriptor
Chi Path Cluster Descriptor	Walk Count Descriptor
PaDEL Chi Path Descriptor	PaDEL Weight Descriptor
Constitutional Descriptor	PaDEL Weighted Path Descriptor
Crippen Descriptor	Wiener Numbers Descriptor
Detour Matrix Descriptor	XLogP Descriptor
Eccentric Connectivity Index Descriptor	Zagreb Index Descriptor
Electrotopological State Atom Type Descriptor	*CDK fingerprint*
Extended Topochemical Atom Descriptor	*CDK extended fingerprint*
FMF Descriptor	*Estate fingerprint*
PaDEL Fragment Complexity Descriptor	*CDK graph only fingerprint*
PaDEL HBond Acceptor Count Descriptor	*MACCS fingerprint*
PaDEL HBond Donor Count Descriptor	*Pubchem fingerprint*
Hybridization Ratio Descriptor	*Substructure fingerprint*
Information Content Descriptor	*Substructure fingerprint count*
IP Molecular Learning Descriptor	*Klekota-Roth fingerprint*
Kappa Shape Indices Descriptor	*Klekota-Roth fingerprint count*
Kier Hall Smarts Descriptor	*2D atom pairs*
Largest Chain Descriptor	*2D atom pairs count*
1. The red-italic categories denote the fingerprint types of descriptors.	

Table S2. The regression results (R^2) of each partition in the cross-validation (Q^2) tests for each model.					
Model	VTS-MLR	VTSsel-MLR	VTSfp-MLR	Lasso-LR	Lasso-RF
1	-5725.51	-136.73	-2.39E+15	0.59	0.63
2	-27.53	-1.74	-2.69E+14	0.56	0.75
3	-134.34	-42.07	-1.47E+15	0.66	0.66
4	-324.33	-1.92	-3.10E+14	0.51	0.65
5	-4252.55	-9.21	-2.50E+16	0.66	0.67
6	-287.15	-14.18	-1.24E+15	0.42	0.58
7	-2.91	0.27	-1.42E+13	0.60	0.67
8	-4661.87	-353.20	-5.16E+14	0.59	0.57
9	-4.38	-14.51	-4.91E+13	0.61	0.47
10	-36016.58	-497331.26	-2.52E+16	0.60	0.56
Avg (Q^2)	-5143.72	-49790.45	-5.64E+15	0.58	0.62

Table S3. Descriptor ensemble comparison by the statistical indicators

Trials	VTS-MLR	VTSsel-MLR	VTSfp-MLR	Lasso-LR	Lasso-RF
# of descriptor	6208	4300	5158	480	480
[1] R^2	0.92	0.86	0.89	0.66	0.92
adjR^2	0.84	0.80	0.81	0.65	0.92
AIC	12384.28	7093.08	12383.59	925.28	926.99
[2] # of p-value >0.03	1349(22%)	922(21%)	1214(24%)	55(11%)	55(11%)
[3] MAE (nm)	18.41	20.88	22.13	39.99	24.42

1. R^2 is calculated using all 11,460 samples.
2. The number of descriptors with p-value > 0.03 and the corresponding ratio of those descriptors.
3. MAE: mean absolute error in comparison with the experimental values using all 11,460 samples.

Table S4. The predicted emission wavelength by DFT vs. Lasso-RF for the selected 20 fluorescence compounds.

ID	SMILE	Exp (nm)	DFT (nm)	Lasso-RF (nm)
1	O=C1C2=C(C=CC=C2)C(=O)C2=C1C=CC=C2	350	387	536
2	CC1=CC=C(C=C1)C1=NN=C(O1)C1=CC=C(C=C1)C(O)=O	376	371	410
3	COC(=O)C1=CC=C(C=C1)C1=NN=C(O1)C1=C(O)C=CC(Cl)=C1	386	366	472
4	C1=CC=C(C=C1)C#CC1=CC=C(C=C1)N1C2=C(C=CC=C2)C2=C1C=CC=C2	390	361	434
5	CC1=C(C(=O)OC2=CC=CC=C12)C1=CC=CC=C1	406	396	432
6	C(CCC)C1=C(SC=C1CCCC)C#CC=1C=CC(=NC1)C1=NC=C(C=C1)C#CC=1SC=C(C1CCCC)CCCC	426	428	450
7	COC1=CC=2C(=NON2)C=C1	430	394	368
8	C(CCC)C1=C(SC(=C1CCCC)C#CC=1C=CC(=NC1)C1=NC=C(C=C1)C#C)C#CC=1C=CC(=NC1)C1=NC=C(C=C1)C#C	449	475	438
9	C(C)(=O)NC1=CC=2C(=NON2)C=C1	455	376	429
10	CC(=O)OC1=CC=C2C(OC(=O)C(C3=CC=CC=C3)=C2C)=C1	431	394	437
11	COC1=C(C=CC=C1)N1CCN(CCN2C(=O)C3=CC=CC4=C(N)C=CC(C2=O)=C34)CC1	500	418	524
12	COC1=C(C=CC=C1)N1CCN(CCNS(=O)(=O)C2=C3C=CC=C(N(C)C)C3=CC=C2)CC1	508	423	520
13	CSC1=CC=2C(=NON2)C=C1	512	405	387
14	COC1=C(C=CC=C1)N1CCN(CCCN2C(=O)C3=CC=CC4=C(N)C=CC(C2=O)=C34)CC1	520	418	526
15	NC1=CC=2C(=NON2)C=C1	525	416	438
16	CCC1=C(C)C2=CC3=NC(=CC4=C(CC)C(=C([NH]4)C=C5N=C(C=C1[NH]2)C(=C5C)CC)C)C(=C3C)CC	624	595	716
17	CCC1=C2NC(\C=C3/N=C(C(CC)=C/3CC)/C(C)=C3NC(=C\C4=N\C(=C/2)\C(CC)=C4CC)/C(CC)=C\3CC)=C1CC	640	607	656

18	CCC1=C(C)/C2=C/C3=C(CC)C(C)=C(N3)\C=C3/N=C(/C=C4NC(=C\C1=N\2)/C(C)=C\4)C(CCC(O)=O)=C/3C	690	594	689
19	CCC1=C(C)/C2=C/C3=N/C(=C\C4=C(CC)C(C)=C(N4)/C=C4/C=C(C)C(\C=C\1N2)=N/4)/C(C)=C3	750	591	716
20	CC(=O)OC1=CC=C2C=C(N3C=C(N=N3)C3=CC=C(C)C=C3)C(=O)OC2=C1	494	380	442
	Mean absolute error (MAE)		60	48

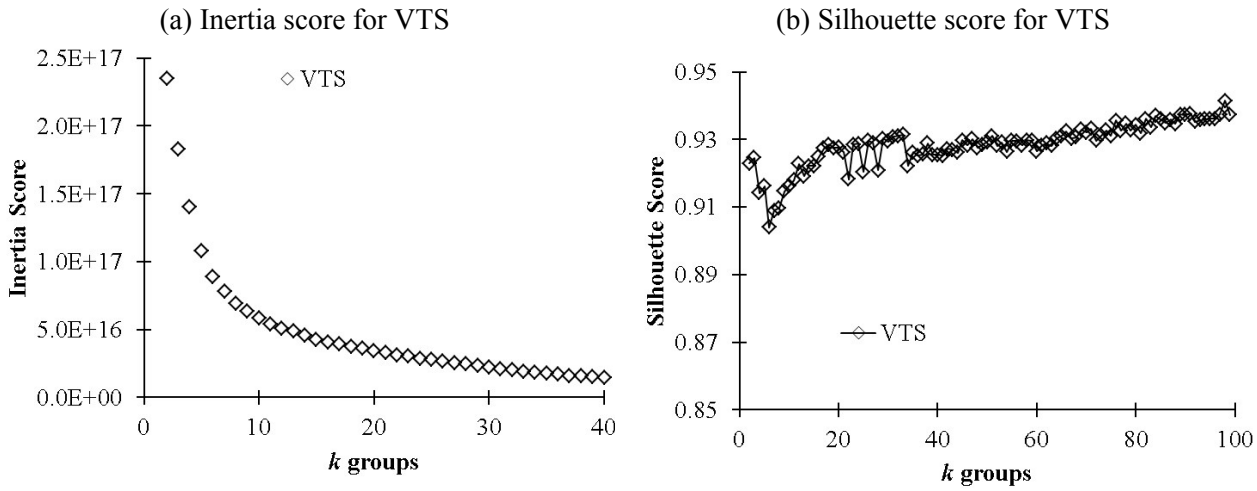


Figure S1. The inertia and Silhouette scores by the direct K-means clustering using VTS descriptor assemble.

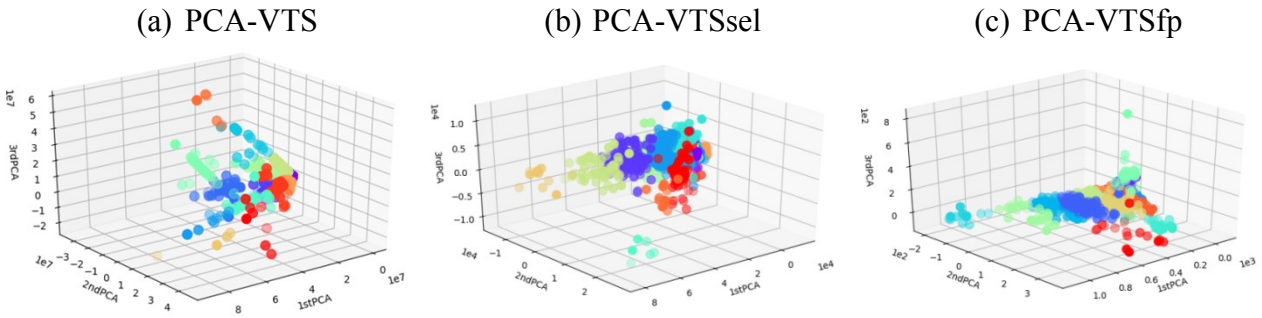


Figure S2. The PCA-transformed three dimensional visualization of VTS series descriptor assembles (*k* groups = 15).

(a) Lasso (*k* groups = 7)

(b) Lasso (*k* groups = 13)

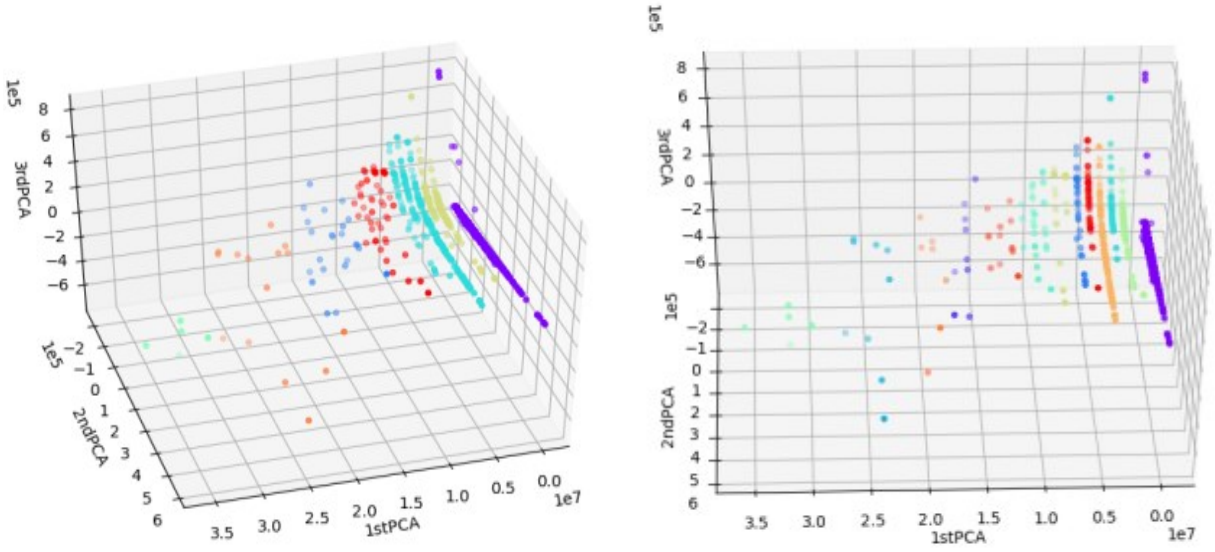
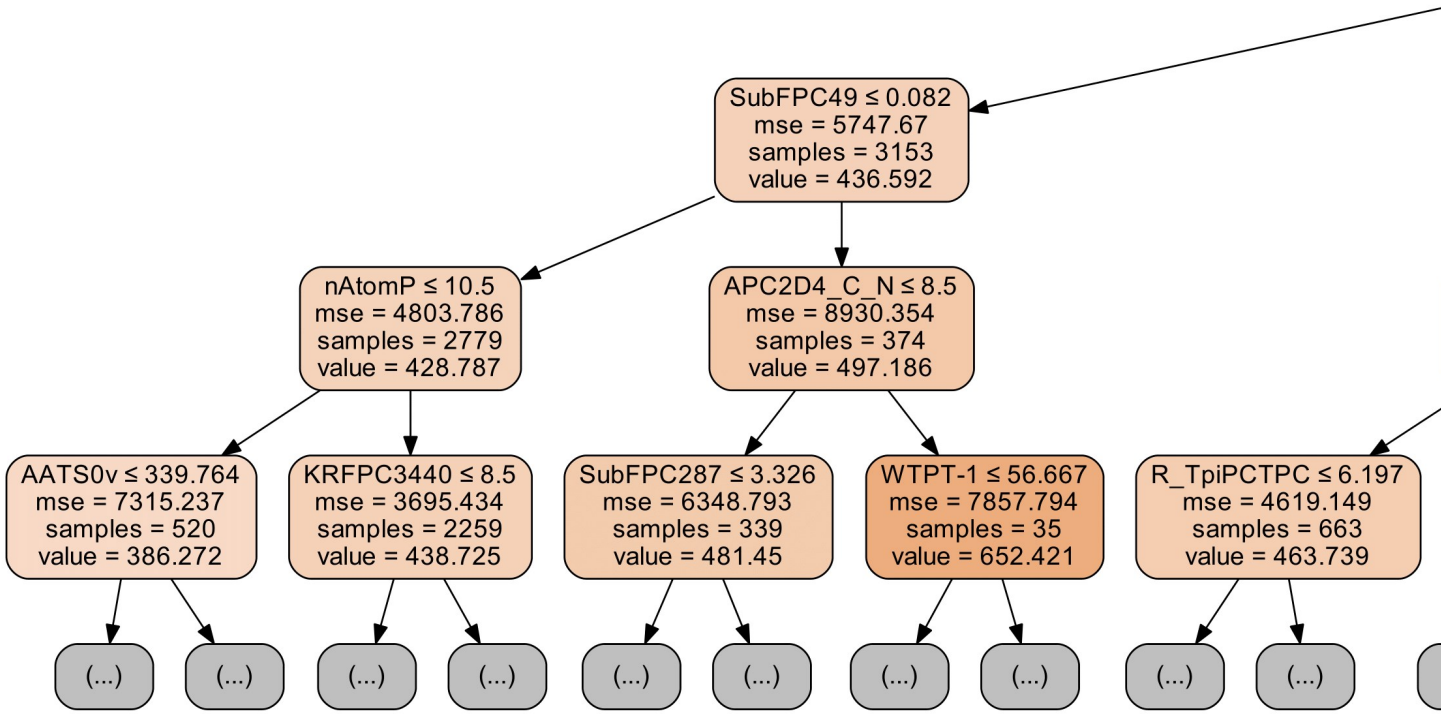


Figure S3. The 3D visualization of the PCA-transformed Lasso descriptor assembles (*k* groups = 7 and 13).


SubFPC49 ≤ 0.082
mse = 5747.67
samples = 3153
value = 436.592
nAtomP ≤ 10.5
mse = 4803.786
samples = 2779
value = 428.787
APC2D4_C_N ≤ 8.5
mse = 8930.354
samples = 374
value = 497.186
AATS0v ≤ 339.764
mse = 7315.237
samples = 520
value = 386.272
KRFPC3440 ≤ 8.5
mse = 3695.434
samples = 2259
value = 438.725
SubFPC287 ≤ 3.326
mse = 6348.793
samples = 339
value = 481.45
WTPT-1 ≤ 56.667
mse = 7857.794
samples = 35
value = 652.421
R_TpiPCTPC ≤ 6.197
mse = 4619.149
samples = 663
value = 463.739
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

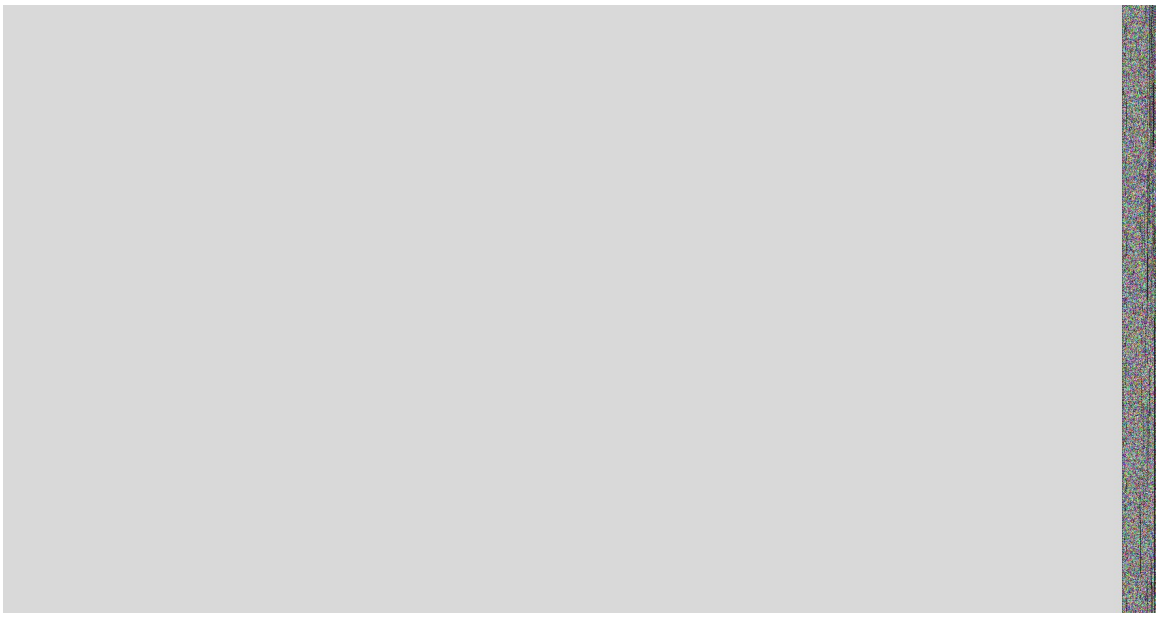


Figure S4. The schematic representation of the top 4 layers of Lasso-RF model.

Definitions of the statistical terms

(1) Inertia score

$$\text{Inertia} = \sum_{i=1}^{N} \min\left(\|X_i - \mu_j\|^2\right),\ \mu_j \in C$$

The *k*-means algorithm divides a set of N samples (set X) into *k* disjoint clusters (set C) in terms of the descriptor space, each cluster described by the mean μ_j (j = 1 – k), of the sample in the cluster. The inertia score is the minimized results of the within-cluster sum-of-squares criterion by adjusting the mean position of each cluster.

(2) Silhouette score

$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$ where *d(i,j)* is the distance between data *i* and data j, and $a(i)$ is the mean distance between data *i* and data j in the same cluster C_i. $b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$ where $b(i)$ is the smallest mean distance of data *i* to all points *j* in any other cluster C_i.

$$s(i) = {}^{b(i) - a(i)}\!/_{max\{a(i),b(i)\}},\ if\ |C_i| > 1$$

$$s(i) = 0,\ if\ |C_i| = 1$$

(3) Terms for conducting Table 3

$y = expriment\ value$ and $\tilde{y} = predicted\ value$

For regression through the origin (intercept set to 0)

$y^{r^0} = k\tilde{y}$ and $\tilde{y}^{r^0} = k'y$

Slopes of the regression through the origin can be defined as

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad \text{and} \quad k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}$$

Correlation coefficients can be defined as

$$R_0^2 = 1 - \frac{\sum\left(\tilde{y}_i - y^{r^0}\right)^2}{\sum\left(\tilde{y}_i - \bar{\tilde{y}}\right)^2} \quad \text{and} \quad R_0^{'2} = 1 - \frac{\sum\left(y_i - \tilde{y}^{r^0}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$

A converged model should have a high predictive ability: both R^2 and either R_0^2 or $R_0^{'2}$ should give similar values. Therefore, $R^2 \geq \max\left(R_0^2, R_0^{'2}\right)$ can be interpreted as follow:

$$\frac{\left(R^2 - R_0^2\right)}{R^2} < 0.1 \quad \text{or} \quad \frac{\left(R^2 - R_0^{'2}\right)}{R^2} < 0.1$$