The System of Self-Consistent QSPR-Models for Refractive Index of Polymers

Andrey A. Toropov (andrey.toropov@marionegri.it)
Istituto di Ricerche Farmacologiche Mario Negri https://orcid.org/0000-0001-6864-6340
Alla P. Toropova
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Valentin O. Kudyshkin
Institute of Polymer Chemistry and Physics

Research Article

Keywords: refractive index of polymers, self-consistent models, Index of Ideality of correlation (IIC), Monte Carlo method, CORAL software

Posted Date: October 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1005707/v1

Version of Record: A version of this preprint was published at Structural Chemistry on January 7th, 2022.
See the published version at https://doi.org/10.1007/s11224-021-01875-y.

The system of self-consistent QSPR-models for refractive index of polymers

Andrey A. Toropov[1*], Alla P. Toropova[1], Valentin O. Kudyshkin[2]

[1] *Istituto di Ricerche Farmacologiche Mario Negri IRCCS,*

Via Mario Negri 2, 20156 Milano, Italy

[2] *Institute of Polymer Chemistry and Physics, Academy of Sciences of the Republic of Uzbekistan,*

Kodyri street 7b, 100128, Tashkent, Uzbekistan

Abstract

Quantitative structure-property/activity relationships (QSPRs/QSARs) are a component of modern natural science. The system of self-consistent models is a specific approach to build up QSPR/QSAR. A group of models of refractive index for different distributions in training and test sets compared. This comparison is a basis to formulate the system of self-consistent models. The so-called index of ideality of correlation (*IIC*) has been used to improve the predictive potential of models of the refractive index of different polymers (n=255). The predictive potential of the suggested models is high since the average value of the determination coefficient for the validation set is 0.885. In addition, the system of self-consistent models may be applied as a tool to assess the predictive potential of an arbitrary QSPR-approach.

Keywords: refractive index of polymers; self-consistent models; Index of Ideality of correlation (*IIC*); Monte Carlo method; CORAL software

[*)] Corresponding author:

Andrey A. Toropov, PhD,

Email: andrey.toropov@marionegri.it

Istituto di Ricerche Farmacologiche Mario Negri IRCCS

Via Mario Negri 2, 20156 Milano, Italy

Tel: +39 02 3901 4595; Fax: +39 02 3901 4735

Introduction

Quantitative structure-property/activity relationships (QSPRs/QSARs) are a tool to assess various endpoints via analysis of available databases on experimental values of the endpoint of interest [1-7]. Simplified molecular input-line entry system (SMILES) is a widely used format to represent the molecular structure [8]. Recently, high refractive index polymers have captured considerable attention of the scientific community due to various applications, aimed to improve advanced optic-electronic devices [9]. The present study aims to build up and validate of QSPR model for the refractive index of polymers. The assessment of the predictive potential of these models carried out via so-called the system of self-consistent models of the refractive index of polymers. The index of ideality of correlation (*IIC*) also can be serve as a criterion of the predictive potential. The *IIC* demonstrates significant ability to improve the predictive potential of QSPR model being applied as add component of the Monte Carlo optimization aimed to model an arbitrary endpoint.

Method

Dataset

The experimental data on the refractive index (*RI*) of different polymers were taken in the literature [10]. Two duplicates were removed. The remaining set list of polymers (n=255) has been distributed randomly in four special subsets: the active training set (25%), passive training set (25%), calibration set (25%), and validation set (25%). Table 1 confirms that the five described above random distributions are not identical.

[Table 1 around here]

Each of the above subsets has its task. The task for the active training set is to calculate correlation weights, which give as large as the possible correlation between experimental and predicted endpoint for the active training set. The task for the passive training set is inspection: whether these data give a reasonable correlation coefficient for the similar compounds in the passive training set. The task of

the calibration set is to detect overtraining. The task for the validation set is the final estimation of the predictive potential of the model.

Optimal quasi-SMILES-based descriptor

The optimal SMILES-based descriptor *DCW(T,N)* is applied for a predictive model of *RI* via the equation:

$$RI = C_0 + C_1 \times DCW(T,N) \quad (1)$$

The C_0 and C_1 are regression coefficients, the descriptor of the correlation weights (DCW) is calculated as

$$DCW(T,N) = \sum CW(APP_k) + \sum CW(A_k) + CW(C5) + CW(C6) \quad (2)$$

The APP_k are atoms pair's proportions [11]; A_k is SMILES attributes [6,7]; C5 and C6 are special codes of rings [12]. The *T* is thresholds, i.e. an integer to separate SMILES attributes into rare and non-rare [6,7,12]. The rare SMILES attributes have correlation weights equal to zero, i.e. these are not involved in building up a model. The *N* is the number of epochs of the Monte Carlo optimization.

The Monte Carlo optimization

Eq. 2 needs the numerical data on the above correlation weights. The Monte Carlo optimization is a tool to calculate those correlation weights. Here three target functions for the Monte Carlo optimization are examined.

The first target function (TF_1)

The first target function is calculated as the following:

$$TF_1 = R_A + R_P - |R_A - R_P| \times 0.1 \quad (3)$$

The R_A and R_P are correlation coefficients between observed and predicted endpoint for the active training set and passive training set, respectively.

The second target function (TF_2)

The second target function is calculated as the following:

$$TF_2 = TF_1 + IIC_C \times 0.5 \quad (4)$$

The IIC_C is the index of ideality of correlation calculated with polymers of the calibration set [11,12]. The *IIC* is calculated as the following:

$$IIC_C = r_C \frac{\min(^{-}MAE_C, ^{+}MAE_C)}{\max(^{-}MAE_C, ^{+}MAE_C)} \quad (5)$$

$$\min(x, y) = \begin{cases} x, if\ x < y \\ y, otherwise \end{cases} \quad (6)$$

$$\max(x, y) = \begin{cases} x, if\ x > y \\ y, otherwise \end{cases} \quad (7)$$

$$^{-}MAE_C = \frac{1}{^{-}N} \sum |\Delta_k|, ^{-}N\ is\ the\ number\ of\ \Delta_k < 0 \quad (8)$$

$$^{+}MAE_C = \frac{1}{^{+}N} \sum |\Delta_k|, ^{+}N\ is\ the\ number\ of\ \Delta_k \geq 0 \quad (9)$$

$$\Delta_k = observed_k - calculated_k \quad (10)$$

The observed and calculated are corresponding values of the endpoint.

The system of self-consistent models

Each i-th model has i-th validation set. As it is demonstrated, (Table 1) the validation sets are far from identical. It is important whether the arbitrary model can be used for an arbitrary validation set? If answer yes, these different models should be considered as self-consistent ones.

The measure of self-consistency is average and dispersion of the correlation coefficient on different validation sets. The corresponding computational experiments are represented by the matrix:

$$\begin{bmatrix} (M_1: V_1 \rightarrow Rv_{11}^2) & \cdots & (M_5: V_1 \rightarrow Rv_{51}^2) \\ \vdots & & \vdots \\ (M_1: V_5 \rightarrow Rv_{15}^2) & \cdots & (M_5: V_5 \rightarrow Rv_{55}^2) \end{bmatrix} \quad (11)$$

the M_i is i-th model; the V_j is the list of polymers applied as the validation set in the case of *j-th* split; the Rv_{ij}^2 is the correlation coefficient observed for *j-th* validation set if applied *i-th* model.

The main quality of an approach is the ability to provide good statistics for the external validation set. Consequently, different approaches should be assessed by the corresponding correlation coefficient for the validation set. In the situation where five models are built up with different splits, the Rv_{ij}^2 estimation could be the clear basis to compare the suitability of different approaches (i.e. optimizations with target functions TF_1, or TF_2). Figure 1 gives histories of the Monte Carlo optimizations with different target functions.

[Figure 1 around here]

To this end, five random splits were applied to build up models for the RI of different polymers using the above-mentioned three target functions. These models are listed below.

TF1-optimization

RI = 1.4389 (± 0.0009) + 0.003810(± 0.00004) * DCW(1,1) (12)

RI = 1.4722(± 0.0003) + 0.006397(± 0.00003) * DCW(1,2) (13)

RI = 1.4523(± 0.0009) + 0.005811(± 0.00007) * DCW(1,1) (14)

RI = 1.4833(± 0.0003) + 0.009368(± 0.00002) * DCW(1,2) (15)

RI = 1.4349(± 0.0006) + 0.007874(± 0.00006) * DCW(1,1) (16)

TF2-optimization

RI = 1.4543(± 0.0009) + 0.003772(± 0.00004) * DCW(1,10) (17)

RI = 1.4801(± 0.0006) + 0.004271(± 0.00003) * DCW(1,10) (18)

RI = 1.4698(± 0.0007) + 0.003889(± 0.00005) * DCW(1,10) (19)

RI = 1.4606(± 0.0009) + 0.006405(± 0.00008) * DCW(1,10) (20)

RI = 1.4807(± 0.0008) + 0.004384(± 0.00005) * DCW(1,10) (21)

Results and Discussion

Table 2 contains the statistical characteristics of models obtained by the Monte Carlo optimization with target functions TF_1, and TF_2. One can see, that the best predictive potential observes for the

TF_2-optimization since the correlation coefficients for validation sets, in this case, reach maximums in comparison with TF_1- and TF_2-optimisations.

It is to be noted that TF_1-optimization with a large number of epochs gives overtraining (Figure 1), whereas TF_2-optimizations give the improvement of the statistical quality for the calibration and validation sets but in detriment the active/passive training sets.

[Table 2 around here]

Three different approaches based on different target functions can be compared with characteristics calculated as

$$Quality = \overline{Rv^2} - \Delta Rv^2 \quad (22)$$

It is clear that some of the quasi-SMILES in the cases of situations $M_i: V_j \rightarrow Rv_{ij}^2$ $(i \neq j)$ are presented in both training and validation sets. However, the general conditions of building up for the groups of models must be quite different.

Table 3 contains data on applying of the TF_1- and TF_2-models for "stranger" validation sets (i.e. applying i-th Model to j-th Split, i≠j). One can see, there are four better TF_1-models, whereas the number of the better TF_2-models is sexteen. Thus, a convenient measure of quality for an arbitrary QSPR-approach is demonstrated.

[Table 3 around here]

The comparison of the models examined here with RI models described in the literature confirms that the predictive potential of the suggested here models is comparable with analogical approaches (Table 4).

[Table 4 around here]

It is to be noted, the Monte Carlo technique was applied to QSPR analysis of polymers [13-15]. Probably, the approach able to be basis for new researches dedicated to polymer sciences.

Supplementary Materials section contains the technical details of described computational experiments.

Conclusions

Applying the *IIC* improves the statistical characteristics of a model for the validation set but to the detriment of the active/passive training sets. The system of the self-consistent model gives the possibility of assessment of different approaches in an aspect of the predictive potential of corresponding models. Factually, the system of self-consistent models is a new tool of checking up the predictive potential of QSPR-models.

Supplementary Materials: *Supplementary materials* section contains details on the model calculated with Eq. 17 i.e. Table S1 contains experimental and calculated values of RI; Table S2 contains the numerical data on the SMILES attributes and corresponding correlation weights. The similar data on split #2 - #5 available on request.

Funding: The authors are grateful for the contribution of the project LIFE-VERMEER (LIFE16 ENV/IT/000167) for the financial support.

Conflicts of Interest /Competing interests: The authors declare no conflict of interest.

Availability of data and material: Data available with in the article or its supplementary materials.

Code availability: CORAL software (http://www.insilico.eu/coral)

Author Contributions: Conceptualization, A.P.T., A.A.T., and V.O.K.; methodology, A.P.T., A.A.T., and V.O.K.; software, A.A.T.; validation, A.P.T., A.A.T., and V.O.K.; data curation, A.P.T., A.A.T., V.O.K.; writing—original draft preparation, A.P.T., A.A.T., V.O.K.; writing—review and editing, A.P.T., A.A.T., and V.O.K. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors are grateful for the contribution of the project LIFE-VERMEER (LIFE16 ENV/IT/000167) for the support.

References

1. Benfenati E, Toropov AA, Toropova AP, Manganaro A, Gonella Diaza R (2011) CORAL software: QSAR for anticancer agents. Chem Biol Drug Des 77 (6): 471-476. DOI: 10.1111/j.1747-0285.2011.01117.x

2. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) CORAL: QSPR models for solubility of [C60] and [C70] fullerene derivatives. Mol Divers 15(1): 249-256. DOI: 10.1007/s11030-010-9245-6

3. Toropov AA, Toropova AP, Ismailov T, Bonchev D (1998) 3D weighting of molecular descriptors for QSPR/QSAR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes. J Mol Struct THEOCHEM 424(3): 237-247. DOI: 10.1016/S0166-1280(97)00151-6

4. Mercader A, Castro EA, Toropov AA (2000) QSPR modeling of the enthalpy of formation from elements by means of correlation weighting of local invariants of atomic orbital molecular graphs. Chem Phys Lett 330(5-6): 612-623. DOI: 10.1016/S0009-2614(00)01126-X

5. Toropova AP, Toropov AA, Benfenati E, Gini G (2011) Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: An unexpected good prediction based on a model that seems untrustworthy. Chemom Intell Lab Syst 105(2): 215-219. DOI: 10.1016/j.chemolab.2010.12.007

6. Toropov AA, Toropova AP, Benfenati E (2010) SMILES-based optimal descriptors: QSAR modeling of carcinogenicity by balance of correlations with ideal slopes. Eur J Med Chem 45(9): 3581-3587. DOI: 10.1016/j.ejmech.2010.05.002

7. Toropov AA, Rasulev BF, Leszczynska D, Leszczynski J (2007) Additive SMILES based optimal descriptors: QSPR modeling of fullerene C60 solubility in organic solvents. Chem Phys Lett 444(1-3): 209-214. DOI: 10.1016/j.cplett.2007.07.024

8. Weininger D (1988) SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci 28(1): 31-36. DOI: 10.1021/ci00057a005

9. Jabeen F, Chen M, Rasulev B, Ossowski M, Boudjouk P (2017) Refractive indices of diverse data set of polymers: A computational QSPR based study. Comput Mater Sci 137: 215-224. DOI: 10.1016/j.commatsci.2017.05.022

10. Schustik SA, Cravero F, Ponzoni I, Díaz MF (2021) Polymer informatics: Expert-in-the-loop in QSPR modeling of refractive index. Comput Mater Sci 194: 110460. DOI: 10.1016/j.commatsci.2021.110460

11. Toropova AP, Toropov AA, Benfenati E (2021) The self-organizing vector of atom-pairs proportions: use to develop models for melting points. Struct Chem 32(3): 967-971. DOI: 10.1007/s11224-021-01778-y

12. Toropov AA, Toropova AP (2018) Predicting cytotoxicity of 2-phenylindole derivatives against breast cancer cells using index of ideality of correlation. Anticancer Research, 38 (11), pp. 6189-6194. DOI: 10.21873/anticanres.12972

13. Toropova AP, Toropov AA, Kudyshkin VO, Rallo R (2015) Prediction of the Q-e parameters from structures of transfer chain agents. J Polym Res 22: 128. https://doi.org/10.1007/s10965-015-0778-3

14. Toropov AA, Toropova AP, Kudyshkin VO, Bozorov NI, Rashidova SSh (2020) Applying of the Monte Carlo technique to build up models of glass transition temperatures of diverse polymers. Struct Chem31: 1739–1743. DOI: 10.1007/s11224-020-01588-8

15. Kudyshkin VO, Toropov AA, Rashidova SSh (2020) Constants of chain transmission in the radical polymerization as a mathematical function of the molecular structure of monomers and regulators, which are presented by SMILES. MDPI AG in MOL2NET 2020, International Conference on Multidisciplinary Sciences, 6th edition session CHEMINFOUNC-02: Chemoinformatics Workshop, UNC Chape Hill, USA, Published: 09 October 2020. DOI: 10.3390/mol2net-06-06945

Table 1

The percentage of identic splits to the active training set and the validation set.

	S_1	S_2	S_3	S_4	S_5
S_1	0	27.8	37.5	26.8	31.6
S_2	35.7	0	26.5	28.3	31.3
S_3	31.6	28.1	0	20.0	21.4
S_4	30.4	23.2	36.8	0	35.7
S_5	35.4	35.4	27.8	31.9	0

Matrix Element [i,j], if i>j is the measure of identity of the active training sets

Matrix Element [i,j], if i<j is the measure of identity of the validation sets

Table 2

The statistical characteristics of models built up with target function TF_1 (without *IIC*), and TF_2 (with *IIC*).

Split	Target function	Set	*n*	R^2	*CCC*	*IIC*	*RMSE*	*MAE*	*F*
1	**With *IIC***	Active training	57	0.7813	0.8772	0.6906	0.045	0.036	196
		Passive training	57	0.7744	0.8790	0.8442	0.040	0.032	189
		Calibration	55	0.8955	0.9377	0.9462	0.019	0.014	454
		Validation	56	0.8647	0.9155		0.021	0.017	
	Without *IIC*	Active training	57	0.7545	0.8601	0.6786	0.047	0.037	169
		Passive training	57	0.6870	0.8255	0.7855	0.049	0.039	121
		Calibration	55	0.7311	0.8491	0.6508	0.031	0.024	144
		Validation	56	0.7645	0.8692		0.028	0.022	
2	**With *IIC***	Active training	58	0.8436	0.9152	0.7462	0.028	0.021	302
		Passive training	55	0.8874	0.9411	0.8797	0.025	0.018	418
		Calibration	56	0.8180	0.8958	0.9044	0.029	0.022	243
		Validation	56	0.8852	0.9330		0.025	0.019	
	Without *IIC*	Active training	58	0.9581	0.9786	0.9135	0.014	0.011	1281
		Passive training	55	0.9363	0.9656	0.8926	0.019	0.015	779
		Calibration	56	0.7489	0.8619	0.6582	0.034	0.027	161
		Validation	56	0.8933	0.9420		0.023	0.019	
3	**With *IIC***	Active training	55	0.7790	0.8758	0.6344	0.031	0.024	187
		Passive training	55	0.8408	0.8837	0.7473	0.038	0.028	280
		Calibration	57	0.9087	0.9437	0.9532	0.023	0.017	547
		Validation	58	0.8714	0.9285		0.024	0.019	
	Without *IIC*	Active training	55	0.8153	0.8982	0.7524	0.028	0.021	234
		Passive training	55	0.7962	0.8801	0.7796	0.041	0.031	207
		Calibration	57	0.7753	0.8780	0.6408	0.034	0.025	190
		Validation	58	0.8450	0.9143		0.029	0.022	
4	**With *IIC***	Active training	55	0.8380	0.9118	0.8827	0.034	0.027	274
		Passive training	57	0.8152	0.8596	0.3915	0.038	0.027	243
		Calibration	57	0.8273	0.9073	0.9092	0.025	0.021	263
		Validation	56	0.8671	0.9236		0.025	0.020	
	Without *IIC*	Active training	55	0.9791	0.9894	0.9541	0.012	0.00901	2482
		Passive training	57	0.9470	0.9091	0.4151	0.030	0.024	982
		Calibration	57	0.7539	0.8573	0.5337	0.032	0.023	168
		Validation	56	0.7678	0.8628		0.036	0.028	
5	**With *IIC***	Active training	57	0.7764	0.8741	0.7930	0.039	0.030	191
		Passive training	55	0.8526	0.7670	0.6037	0.046	0.039	307
		Calibration	56	0.9321	0.9607	0.9654	0.018	0.014	741
		Validation	57	0.9028	0.9373		0.019	0.015	
	Without *IIC*	Active training	57	0.8824	0.9375	0.8454	0.029	0.021	413
		Passive training	55	0.8869	0.8747	0.3654	0.037	0.031	416
		Calibration	56	0.8726	0.9164	0.6635	0.030	0.021	370
		Validation	57	0.8896	0.9343		0.021	0.016	

*) A = active training set; P = passive training set; C = calibration set; V = validation set

Table 3

Three systems of self-consistent models were obtained by the TF_1-, and TF_2-optimization for splits 1-5.

Target function	**Model**	**n**	**V_1**	**n**	**V_2**	**n**	**V_3**	**n**	**V_4**	**n**	**V_5**
TF_1	M_1			20	0.828	18	0.599	17	0.797	20	0.827
	M_2	20	**0.890**			16	0.926	13	0.880	20	0.922
	M_3	18	**0.858**	16	0.943			21	0.815	16	**0.961**
	M_4	17	0.742	13	0.712	21	0.836			18	0.847
	M_5	20	0.909	20	**0.942**	16	0.903	18	0.852		
TF_2	M_1			20	**0.855**	18	**0.860**	17	**0.799**	20	**0.860**
	M_2	20	0.865			16	**0.966**	13	**0.961**	20	**0.941**
	M_3	18	0.852	16	**0.949**			21	**0.900**	16	0.940
	M_4	17	**0.745**	13	**0.749**	21	**0.914**			18	**0.891**
	M_5	20	**0.925**	20	0.921	16	**0.906**	18	**0.893**		

*) V_k is the validation set related to k-*th* split; the preferable predictive potential of models (obtained using TF_1 or TF_2) indicated by bold.

Table 4

The comparison with models suggested in the literature

R^2 training set	**R^2 validation set**	**Reference**
0.932	0.882	[9]
0.842 – 0.969	-	[10]
TF_1-Optimization 0.864 (average) TF_2-Optimization 0.804 (average)	TF_1-Optimization 0.849 (average) TF_2- Optimization 0.885 (average)	In this work

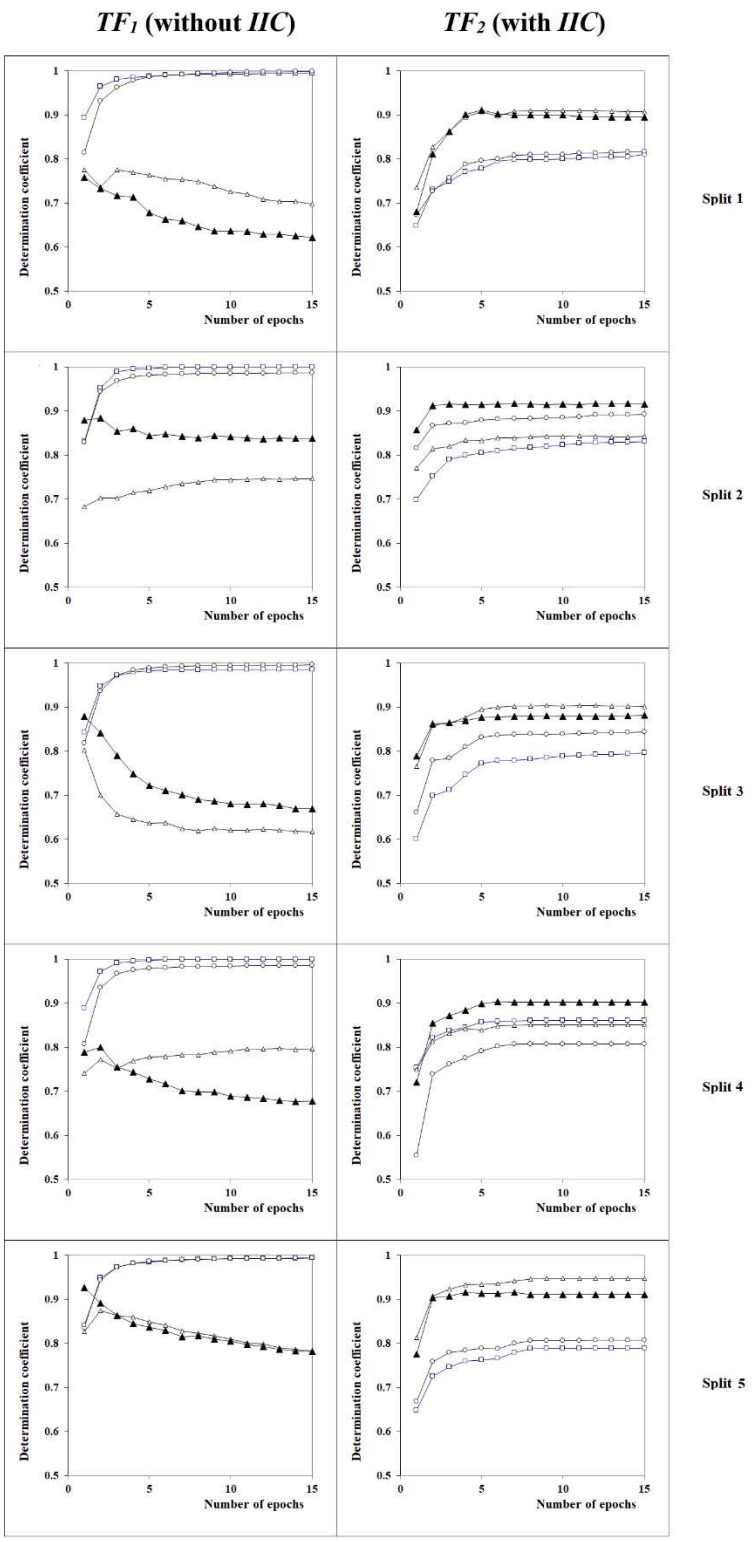

Figure 1

Histories of the Monte Carlo optimizations with different target functions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.xlsx