Szeged Matrix Property Indices as Descriptors to Characterize Fullerenes

Lorentz Jäntschi[1,2], and Sorana D. Bolboacă[3,*]

[1] *Technical University of Cluj-Napoca, Department of Physics and Chemistry, 103-105 Muncii Bldv., RO-400641, Cluj-Napoca, Romania*

[2] *Babeş-Bolyai University, Doctoral Studies - Chemistry, 11 Arany Janos str., RO-400028, Cluj-Napoca, Romania*

[3] *Iuliu Haţieganu University of Medicine and Pharmacy, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur str., RO-400349, Cluj-Napoca, Romania.*

** corresponding author: Phone: +40750774506; E-mail: sbolboaca@umfcluj.ro*

Running title: Szeged Matrix Property Indices on Fullerenes

Abstract

Fullerenes are class of allotropes of carbon organized as closed cages or tubes of carbon atoms. The fullerenes with small number of atoms were not frequently investigated. This paper presents a detailed treatment of total strain energy as function of structural feature extracted from isomers of C_{40} fullerene using Szeged Matrix Property Indices (SMPI). The paper has a two-fold contribution. First, the total strain energy of C_{40} fullerene isomers (40 structures) was linked with SMPI descriptors under two scenarios, one which incorporate just the SMPI descriptors and the other one which contains also five calculated properties (dipole moment, scf-binding-energy, scf-core-energy, scf-electronic-energy, and heat of formation). Second, the performing models identified on C_{40} fullerene family or the descriptors of these models were used to predict the total strain energy on C_{42} fullerene isomers. The obtained results show that the inclusion of properties in the pool of descriptors led to the reduction of accurate linear models. One property, namely scf-binding-energy

proved a significant contribution to total strain energy of C_{40} fullerene isomers. However, the top-three performing models contain just SMPI descriptors. A model with four descriptors proved most accurate model and show fair abilities in prediction of the same property on C_{42} fullerene isomers when the approach considered the descriptors identified on C_{40} as the input descriptors for C_{42} fullerene isomers.

Keywords: *nano structure-property relationship; C_{40} fullerene; C_{42} fullerene; Szeged Matrix Property Indices (SMPI)*

1. Introduction

Fullerenes are class of allotropes of carbon organized as closed cages or tubes of carbon atoms. Fullerenes received attention from the researchers all over the word and led to the synthesis of new compounds [1-4] and identification of different applications due to their hardness, high electron affinity, increased incident light intensity, and biological activities [5-7].

C_{40} is one of the small fullerene and several symmetries of these cages such as D_{5d} [8,9], D_{4h} [10,11], D_{2d} [9], D_{2h} [12] were identified and studied. Further, *ab initio* studies on stability of C_{40} fullerene were performed [13,14] C_{40} fullerene has 40 known isomers. Dinca et al. conducted a theoretical study on the C_{40} isomers and showed that pentagon valence parameter correlates well with heat of formation as a measure of thermodynamic stability [7] Halogenated C_{40} cage has been identified as a good candidate for hydrogen storage [15] while all C_{40} fullerene isomers were found to be highly aromatized at the polyvalent anionic states [16].

A nano-quantitative structure-property relationship modeling on C_{42} fullerene isomers showed the ability of Szeged Matrix Property Indices (SMPI [17]) as structural descriptors to fit the total strain energy [18]. The aim of this study was to assess the estimation degree for total strain energy derived in the context of continuum elasticity theory on a pool of structural descriptors and respectively structural and property descriptors. Furthermore, the prediction abilities of the most accurate models were assessed on C_{42} fullerene isomers in the context of the same property.

2. Materials and Methods

2.1. Data sets

Data and the values of the total strain energy (continuum elasticity) are online available and were taken from the following addresses:

URL: http://nanotube.msu.edu/fullerene/fullerene.php?C=40

URL: http://nanotube.msu.edu/fullerene/fullerene.php?C=42

Forty isomers of C_{40} fullerene and forty-five isomers of C_{42} fullerene were included in the study.

2.2. Regression analysis

The structures of C_{40} and C_{42} fullerene isomers were downloaded as *.xyz files and the molecules were included in the analysis as downloaded. The geometry of the investigated fullerene was based on the geometry of the structures in the Yoshida's Fullerene Library and re-optimized using Dreiding-like force-field [19]). The procedure presented in Table 1 was applied on both investigated sets.

Table 1. Preliminary operations apply to C_{40} and C_{42} isomers.

Step	What?	Input files	Output files	Why? & Program and what does it do
1	Converting files	*.xyz	*.mol	*.xyz files do not contain bond information & Spartan program (automatically detect bonds and connect the atoms)
2	Converting files	*.mol	*.hin	*.mol files do not contain partial charge information & Babel program (http://openbabel.org) was used to convert the *.mol files to HyperChem files (http://www.hyper.com/)
3	Partial charges calculation	*.hin	*.hin	SMPI [**Error! Bookmark not defined.**] needs partial charges in order to provide a full family of structure descriptors & HyperChem/AM1/SinglePoint [20] was used
4	Structure descriptors	*.hin	*.txt	The tool http://l.academicdirect.org/Chemistry/SARs/SMPI/ & fast, simple and provide descriptors for all molecules at once

The pool of structural descriptors (scenario 1) and of structural and property descriptors (scenario 2; properties: dipole moment, scf-binding-energy, scf-core-energy, scf-electronic-energy, and heat of formation) was used as raw data in estimation of most accurate structure-property and structure-property-property models on C_{40} data set, respectively.

The files containing the raw data from both scenarios entered separately into the regression analysis. The analysis was conducted using classical approach of multiple linear regressions, when

sum of squares of residuals from vertical offsets were minimized:

$$Y \sim \hat{Y} = a_0 + \Sigma_{1\leq i\leq m} b_i X_i$$

or

$$Y \sim \hat{Y} = \Sigma_{1\leq i\leq m} b_i X_i \text{ when } a_0 \text{ is not significantly different by 0}$$

where Y is the total strain energy (dependent variable), X_i is the structural or property descriptor $\{X_i;\ 1\leq i\leq m\}$ (independent variables), m is the number of independent variables in the model, a_0 is the intercept of the model, b_i is the slope.

The coefficients of the regression model were obtained by minimizing the residuals:

$$SSr = \Sigma_{1\leq j\leq n}(Y_j - \hat{Y}_j)^2 \rightarrow \min.$$

Systematic search for those descriptors able to explain the investigated continuum elasticity was conducted on simple and multiple regression analysis (up to four descriptors) on C_{40} dataset. The size of descriptors pools for identification of the most accurate models for each scenario is given in Table 2. One program has been developed and implemented to filter the regressors (both structural descriptors and properties) based on their explanatory power (absolute values: $10^{-7} < |X_i| < 10^7$), and association between the property as dependent variable and the regressors (for correlations (for all regressors): $0.001 < r^2(X_i,Y), r^2(X_{i1},X_{i2}) < 0.999$, where r^2=determination coefficient). The number of filtered descriptors is given in Table 2.

Table 2. Size of the pools of regressors in the considered scenarios.

Scenario	Data file	No. of regressors	Qualified descriptors
1	C40_data.txt	1512	232
2	C40_datap.txt	1517 (1512+5)	236 (232+4)

An additional program was developed to systematically search for linear models (LM, with the dependent variable and among regressors). A huge number of regressions were tested (125,991,255 regressions only for the second scenario and only for the case of the search with four variables). Several special features were implemented in this program to assure a fast run and to provide useful information (see Table 3).

Table 3. Features of the program implemented to find linear models

Feature	Explanation
The descriptors were stored into dynamic arrays.	Running the program for different dataset or scenario need the input data of different sizes (see Table 2)
Two scenarios of output were implemented, one in which regressions are listed only if has higher r^2 value, and the other in which the regressions are listed if have the r^2 value higher than a given value.	The scenario for listing only of the regression with higher r^2 works very well in the testing of the association, but r^2 is not the only criterion used to select the most accurate associations
When coefficients of regression are obtained, in the same time are obtained their associated t-values (null hypothesis: the value of coefficient is not significantly different by aero) by calculating the inverse of the matrix of the system.	For two reasons: ÷ First, worth nothing to list a regression that has coefficients with no statistical significance. ÷ Second, if specifically the intercept has no statistical significance, in means that also the model with no intercept should be tested for significance.
A modification of the classical formula for the calculation of the correlation coefficient between observed (Y) and estimated value ($\hat{Y}$) was implemented. This method did not require the calculation of the estimated values for each molecule in the dataset.	After obtaining of the coefficients, the determination coefficient between observed (Y) and estimated value ($\hat{Y}$) should be calculated, to list or not the possible regression. But, the classical formula requires need to calculate first the estimated values, which is time consuming (complexity of O(n·m) order). This task is time and resources consuming. The modification reduced the complexity at O(m).
Prior conducting any regression, the following sums were calculated: S(Y), S(Y^2), S(X_i), S(X_i^2), S(X_iY), S(X_iY_i)	This implementation allows a significant reducing of the complexity of calculations, because these sums are involved for (about) each coefficient of the matrix of the system (of size m·m).

The program to find the SLR (simple linear regression) and MLR (multiple linear regression) models was run in the test mode, namely to list regressions only if an improvement in the determination coefficient exists.

The determination coefficient (r^2) was provided as an estimation parameter and determination coefficient in leave-one-out analysis (Q^2) as a parameter of internal validation of the model [21,22].

2.3. Assessment of the models

The performances in estimation of the top-three models with highest goodness-of-fit were assessed using the measures presented in Table 4 [23-25].

Table 4. Statistics for assessment of the regression models.

Name	Abbreviation	Desired value
Adjusted determination coefficient	r^2_{adj}	high
Ratio of variance explained by the model	F-value	high
Residual mean square	RMS	low
Average prediction variance	APV	low
Average prediction mean squared error	APMSE	low
Mean absolute error	MAE	low
Root mean square error	RMSE	< MAE
Mean absolute percentage error	MAPE	closest to zero
Standard error of prediction	SEP	low
Relative error of prediction	REP%	low
Predictive squares correlation coefficient in training set	Q^2_{F1}	high

2.4. Assessment of prediction power

The prediction power of the most accurate models was assessed in two scenarios:

- The descriptors identified by the most accurate models on C_{40} dataset were used to predict the total strain energy for C_{42} dataset.
- The adjusted most accurate models obtained on C_{40} dataset were applied to C_{42} congeners.

The metrics used to assess the prediction ability [25] are presented in Table 5.

Table 5. Metrics for assessment of the prediction power of the models.

Name	Abbreviation	Desired value
Determination coefficient of the prediction set	r^2_{ext}	high
Predictive square correlation coefficient in external set	Q^2_{F2}	high
External prediction ability	Q^2_{F3}	high
Root mean square error of predicted	RMSEP	low
Mean absolute error of predicted	MAEP	low
Percentage predictive error	%PredErr	low
Concordance correlation coefficient (http://services.niwa.co.nz/services/statistical/concordance)	CCC	high

3. Results and Discussion

3.1. Estimation models

The study conducted to identify the most accurate models able to estimate the total strain energy on the C_{40} isomers showed that 37 models proved accurate when just structural descriptors are considered as independent variables. In addition, when the pool of descriptors contained both structural and property descriptors, just 28 models are identified. The trends in regard of determination coefficient of the identified models obtained on both scenarios are presented in Figure 1. One property descriptor represented by 'scf-binding-energy' is the only property with significant contribution to the total strain energy on C_{40} dataset. Its contribution is observed from the first model until the sixteenth model in the second scenario (Figure 1) while the last twelve models in both scenarios are identical and contain only structural descriptors. Details on all 37 and respectively 28 models are given in *Supplementary Material*.

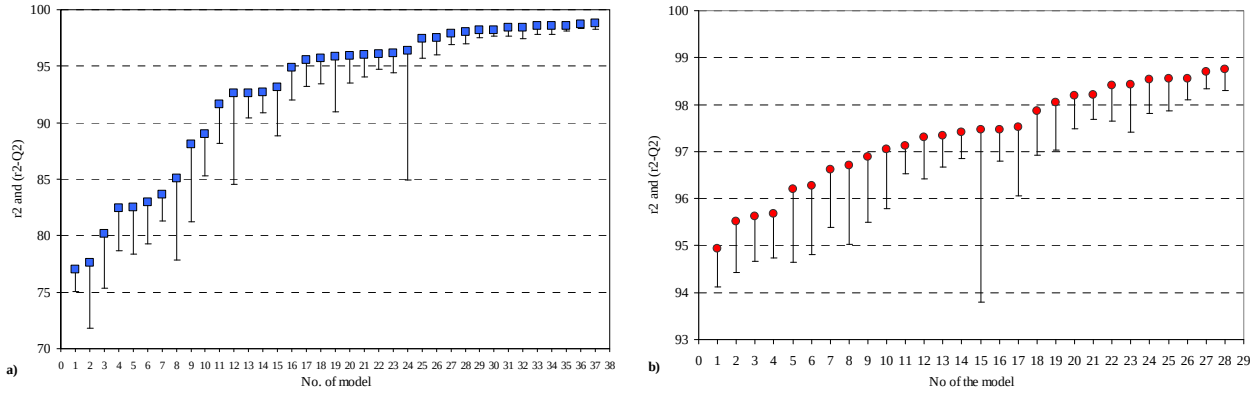

Figure 1. Improvement in regard of r^2 with the increase of the number of descriptors in the models and the distance between r^2 and Q^2 (the ended line): a) first scenario (2 descriptors: model from 2 to 16; 3 descriptors: models from 17 to 29; 4 descriptors: models from 30 to 37), b) second scenario (2 descriptors: model from 2 to 6; 3 descriptors: models from 7 to 20; 4 descriptors: models from 21 to 28). The first model is resulted from simple linear regression analysis and the contribution of property is significant, leading to an improvement of determination coefficient from 77.02% (first scenario) to 94.94% (second scenario).

The investigated C_{40} dataset have some 'advantages' and 'disadvantages'. All atoms are carbon atoms, so it is easy to do the research on such sample. Nevertheless, is not an advantage for SMPI, which operates at the level of the type of the atom too (take into account different atomic properties for different atom types). SMPI produces degenerated descriptors when all atoms are the identical, which reduces its explanatory power, so it is a disadvantage. All molecules have the same number of atoms, and all atoms have the same 'vertex degree' - e.g. number of bonds attached to it; this is another disadvantage, for a method based on topology, but not necessary of SMPI that works also at geometrical level. All bonds are of same type - aromatic bonds - and this is a disadvantage for SMPI (which degenerates again), since SMPI takes two topological approaches - one classical in which the topological distances are counted as the number of bonds, and another one in which the distance is counted as the inverse of the bond order.

Three models with high goodness-of-fit and small difference between goodness-of-fit and determination coefficient in leave-one-out analysis were assessed concerning their prediction

abilities of the set of C_{42} congeners. The top-three models are given in Eq(1)-(3):

$$\hat{Y} = 2.50\times10^{3} + 0.0010*\text{ImUGG} -6.23\times10^{5}*\text{RJUGE}-2.1762*\text{IFEGE}+3.91\times10^{-4}*\text{IIEGA} \quad (1)$$

r^2=0.9856, r^2_{adj}=0.9839, F-value (p) = 598 (1.12×10^{-31}), Q^2=0.9810

$$\hat{Y} = 2.44\times10^{3} -3.1038*\text{LMEGG} -6.20\times10^{5}*\text{RJUGE}-2.0521*\text{IFEGE} +3.90\times10^{-4}*\text{IIEGA} \quad (2)$$

r^2=0.9870, r^2_{adj}=0.9855, F-value (p) = 665 (1.78×10^{-32}), Q^2=0.9833

$$\hat{Y} = 2.5822*\text{IJUGE}+2.7106*\text{IFUGE}-3.0808*\text{IIUGE}-0.5322*\text{IIPTB} \quad (3)$$

r^2=0.9876, r^2_{adj}=0.9587, F-value (p) = 714 (5.15×10^{-33}), Q^2=0.9830

where $\hat{Y}$ = estimated total strain energy; ImUGG, RJUGE, IFEGE, IIEGA, LMEGG, IJUGE, IFUGE, IIUGE, and IIPTB = SMPI structural descriptors; r^2 = determination coefficient; r^2_{adj}= adjusted determination coefficient; F-value = ratio of variance explained by the model; p = p-value associated to F-value Q^2=determination coefficient in leave-one-out analysis.

The models in Eq(1)-Eq(3) considered a total number of nine SMPI descriptors, five of them linking electronegativity [26,27] with the total strain energy, while other considered melting point temperature ('G' as the last letter in the descriptor name), atomic mass ('A') or atomic number ('B'). With one exception represented by IIPTB descriptor, all other descriptors considered the distance matrix calculated using topological distance ('G' as the fourth letter in the descriptors name). The third letter in the descriptors name refers the interaction effects matrix operating on the properties and on the distances matrices. The second letter is related with the value calculated in the interaction effect matrix as minimum or maximum ('m' respectively 'M' letter as second letter in the descriptor name), half-sum($M_{i,j}*M_{j,i}*Ad_{i,j}$, where $M_{i,j}$ = the i^{th} and j^{th} element on matrix, $M_{j,i}$ = the j^{th} and i^{th} element on matrix, and $Ad_{i,j}$ = the i^{th} and j^{th} element on adjacency matrix) ('F' letter), half-sum($M_{i,j}$) ('I' letter), or half-sum($M_{i,j}*M_{j,i}$) ('J' letter). The first letter in the descriptor name is related with the linearization operator.

The analysis of the results revealed that the Eq(2) model is the one with both higher adjusted determination coefficient and higher determination coefficient in leave one-out analysis. Just one measure associated to the residual errors, named mean absolute error indicate that model from

Eq(2) is superior compared with Eq(1) and Eq(3) (see Table 6). All other investigated measures (see Table 6) sustain the model from Eq(3) as the most accurate model in estimation of the total strain energy on C_{40} fullerene congeners.

Table 6. Characteristic of the models from Eq(1)-Eq(3): estimation power.

Parameter (Abbreviation)	**Eq(1)**	**Eq(2)**	**Eq(3)**
Residual Mean Square (RMS)	0.00114	0.00103	0.00098
Average Prediction Variance (APV)	0.00125	0.00113	0.00108
Average Prediction Mean Squared Error (APMSE)	0.00003	0.00003	0.00003
Mean Absolute Error (MAE)	0.1425	0.1409	0.1541
Root Mean Square Error (RMSE)	0.0320	0.0304	0.0297
Mean Absolute Percentage Error (MAPE)	0.0050	0.0050	0.0055
Standard Error of Prediction (SEP)	0.0324	0.0308	0.0301
Relative Error of Prediction (REP%)	0.1132	0.1074	0.1051
Predictive Squares Correlation Coefficient in Training Set (Q^2_{F1})	0.9856	0.9870	0.9876

3.2. Assessment of prediction power

Two different approaches were used to assess the prediction power of the models: the use of the weighted equations obtained on C_{40} fullerene dataset and the use of descriptors from Eq(1)-Eq(3) models to predict the total strain energy using C_{42} dataset.

The same weight of 0.5 proved able to led to best fit of Eq(1) and Eq(2) on C_{42} congeners. The proper weight able to led to best fit of Eq(3) on C_{42} congeners proved equal to 0.1.

The SMPI descriptors from the best performing models identified on C_{40} (n=40) congeners were used as independent variable to predict total strain energy on C_{42} congeners (n=45) and the results are presented in Eq(4)-(6):

$$\hat{Y} = 2.62\times10^{3} + 0.0004*\text{ImUGG} - 6.09\times10^{5}*\text{RJUGE} - 2.3287*\text{IFEGE} + 3.22\times10^{-4}*\text{IIEGA} \quad (4)$$

r^2=0.9591, r^2_{adj}=0.9550, F-value ($p=3.53\times10^{-27}$) = 234, Q^2=0.9448

$$\hat{Y} = 2.63\times10^{3} - 0.1674*\text{LMEGG} - 6.11\times10^{5}*\text{RJUGE} - 2.3259*\text{IFEGE} + 3.12\times10^{-4}*\text{IIEGA} \quad (5)$$

r^2=0.9585, r^2_{adj}=0.9543, F-value ($p=4.69\times10^{-27}$) = 231, Q^2=0.9453

$$\hat{Y} = 2.4422*\text{IJUGE} + 1.4432*\text{IFUGE} - 2.5143*\text{IIUGE} - 0.4643*\text{IIPTB} \quad (6)$$

r^2=0.9785, r^2_{adj}=0.9526, F-value (5.52×10^{-33}) =467, Q^2=0.9745

where $\hat{Y}$ = estimated total strain energy; ImUGG, RJUGE, IFEGE, IIEGA, LMEGG, IJUGE, IFUGE, IIUGE, and IIPTB = SMPI structural descriptors; r^2 = determination coefficient; r^2_{adj}=

adjusted determination coefficient; F-value = ratio of variance explained by the model; p-value = p-value associated to F-value Q^2=determination coefficient in leave-one-out analysis.

The prediction metrics for both approaches are presented in Table 7. An analysis of Table 7 showed that the models that used the SMPI descriptors had better prediction abilities compared with the weighted Eq(1)-(3) models. Even if some of the weighted models had prediction abilities (see Eq(1)*0.5, Table 7), the models from Eq(4)-(6) more accurate prediction powers.

Table 7. Prediction metrics on C_{42} congeners.

Model	r^2_{ext}	Q^2_{F2}	Q^2_{F3}	RMSEP	MAEP	CCC [95%CI]	%PredErr
Eq(1)*0.5	0.5251	0.6249	0.6673	0.9225	0.7744	0.7031 [0.6060–0.7795]	1.12
Eq(4)	0.9591	0.9591	0.9713	0.2708	0.2129	0.9791 [0.9624–0.9884]	0.31
Eq(2)*0.5	NR	0.3457	NR	1.3386	2.7387	0.1991 [0.1296–0.2667]	3.96
Eq(5)	0.9585	0.9585	0.9709	0.2728	0.2172	0.9788 [0.9620–0.9882]	0.31
Eq(3)*0.1	0.6143	NR	NR	1.3386	1.7185	0.1663 [0.1063–0.2251]	2.49
Eq(6)	0.9785	0.9785	0.9850	0.1962	0.1544	0.9891 [0.9803–0.9940]	0.22

r^2_{ext} = determination coefficient of the prediction set; Q^2_{F2} = predictive square correlation coefficient in external set;
Q^2_{F3} = external predictive ability; RMSEP = root means square error of predicted;
MAEP = mean absolute error of predicted; %PredErr = percentage predictive error;
CCC [95%CI] = concordance correlation coefficient [two-sided 95% confidence intervals];
NR=not reliable value

The plots associated with the applied approaches are presented in Figure 2. The analysis of the graphical representations of the models leads to the same conclusion as the analysis of the prediction metrics presented in Table 7. These lead to the conclusion that SMPI descriptors belonging to the most accurate estimation models to fit total strain energy on C_{40} congeners are also able to fit the total strain energy on C_{42} congeners. Similar results are expected to be seen also on other similar sets of C_n congeners.

The model with four descriptors showed abilities in estimation (on C_{40} dataset) and prediction (on C_{42} dataset). The best prediction is obtained when the descriptors identified to belong to the most accurate models on C_{40} congeners are used to predict the same property, namely total strain energy, on C_{42} congeners. This result is similar with the previously reported results [18]. The analysis of Eq(3) and Eq(4) showed lower values of the coefficients in prediction model compared with estimation model but without any change of the sign (as + or -) of the coefficients. According with this model, the total strain energy of C_{40} and C_{42} fullerene congeners is explained by electronegativity and atomic number as atomic property of the compounds, having geometric and

topologic component. One of the fourth descriptors seen in the most accurate model, namely IJUGE, was also identified as descriptor linked with the total strain energy in the previously reported study on C_{42} congeners [18].

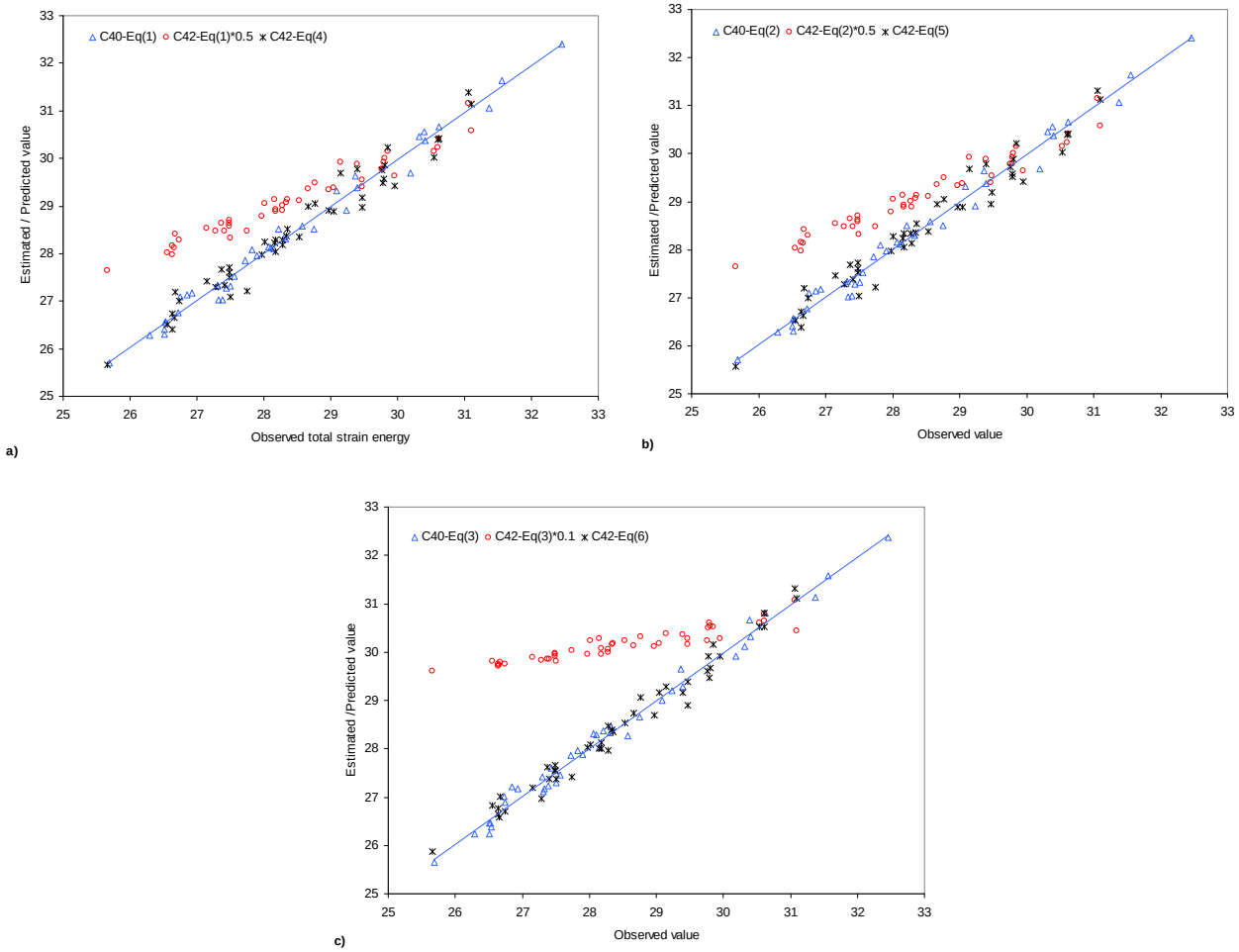

Figure 2. The fit between observed total strain energy and estimated values by models: a) Eq(1) on C_{40} & Eq(1)*0.5 on C_{42} and Eq(4) on C_{42}, b) Eq(2) on C_{40} & Eq(2)*0.5 on C_{42} and Eq(5) on C_{42}, c) Eq(3) on C_{40} & Eq(3)*0.1 on C_{42} and Eq(6) on C_{42}.

3.3. Comparison with other reported models

A regression model is considered to have prediction power if is accurate on compounds not included in the dataset on which the model was obtained. The prediction power of regression models was tested on external data set represented by C_{42} fullerene isomers in this manuscript and the model in Eq(5) and Eq(6) proved accurate models. However, *are the models given by Eq(5) and Eq(6) different by the more accurate models obtained on C_{42} fullerene isomers*? To test this, the

models in Eq(5) and Eq(6) were compared in regard of goodness-of-fit with two previously reported models [18] using Steiger's correlated correlation analysis [28].

The most accurate models reported on C_{42} fullerene isomers able to estimate and predict the total strain energy using SMPI descriptor as previously reported [18] are as follow:

$$\hat{Y} = 838.80-1.41*IFEGE-3.66\times10^{-3}*IIUGF+2.16*IJUGE \quad (7)$$

$r^2=0.9836$, $r^2_{adj}=0.9824$, F-value (p) = 820 (1.30×10^{-36}), $Q^2=0.9809$, %PredErr = 19.76

$$\hat{Y} = -199.61 - 21.63*IFETB+40.90*IFUGB-2.62\times10^{-}3*IIUGF+1.56*IJUGE \quad (8)$$

$r^2=0.9898$, $r^2_{adj}=0.9888$, F-value (2.87×10^{-39}) = 974, $Q^2=0.9768$, %PredErr = 15.95

No significant difference in regard of correlation coefficients was observed when Eq(7) was compared with Eq(8) (Table 8). All other cases showed significant higher correlation coefficients as the number of equation increased (Table 8).

Table 8. Correlated correlation analysis: p-value matrix for comparisons amongst Eq(5) to Eq(7).

	Eq(6)	Eq(7)	Eq(8)
Eq(5)	0.0001	0.0001	$5.69\cdot10^{-8}$
Eq(6)		$3.08\cdot10^{-11}$	$1.12\cdot10^{-14}$
Eq(7)			0.0705

An analysis of models given by Eq(5)-Eq(8) and of the results presented in Table 8 reveal the following:

- At least one SMPI descriptor is the same in Eq(5) and Eq(7), and Eq(6) and Eq(8), respectively
- The model with four descriptors in Eq(8) has an intercept significantly different by zero compared with the model in Eq(6) that proved an intercept not significantly different by zero
- The explanatory power express by r^2 is higher on models from Eq(7) and Eq(8) compared with models on Eq(5) and Eq(6)
- No difference in regard of goodness-of-fit is observed between Eq(7) and Eq(8)

The results of our study showed that the SMPI descriptors accurately fit the total strain energy on C_{40} isomers. Nevertheless, the SMPI descriptors able to explain the total strain energy of C_{40} fullerene isomers provide fair models also on C_{42} fullerene congeners.

Even if fair prediction power was obtained on C_{42} fullerene congeners, the goodness-of-fit is

lower compared with the goodness-of-fit of the most accurate models previously reported on C_{42} fullerene isomers. Furthermore, differences are observed in atomic properties and the contribution of topology and/or geometry to the total strain energy are observed when the model is constructed on the C_{42} fullerene isomers compared with the approach when the model constructed on C_{40} fullerene congeners is used to predict the total strain energy.

4. Conclusions

Estimation of properties with families of descriptors derived from structure is generally superior to estimation of the properties from other properties. In fact, it is a hazard to predict one property from another since the properties are measured in different conditions and/or with different instrumentation, or are calculated using different formulas and/or approaches).

The total strain energy was successfully model on C_{40} fullerene isomers and those structural characteristics able to explain the variation of total strain energy were identified. A model with four descriptors proved most accurate model and show fair abilities in prediction of the same property on C_{42} fullerene isomers when the approach considered the descriptors identified on C_{40} as the input descriptors for C_{42} fullerene isomers.

Conflict of Interests

The authors declare that there is no conflict of interests.

References

[1]. O. Kharlamov, G. Kharlamova, N. Kirillova, O. Khyzhun, and V. Trachevskii, NATO Science for Peace and Security Series A: Chemistry and Biology 245 (2012)

[2]. P. Peng, F.-F. Li, F. L. Bowles, V. S. P. K. Neti, A. J. Metta-Magana, M. M. Olmstead, A. L. Balch, and L. Echegoyen, Chem. Commun. **49**, 3209 (2013).

[3]. J. Pattanayak, T. Kar, and S. Scheiner, J. Phys. Chem. A **108**, 7681 (2004).

[4]. E. E. Maroto, M. Izquierdo, S. Reboredo, J. Marco-Martínez, S. Filippone, and N. Martín, Acc. Chem. Res. **47**, 2660 (2014).

[5]. E. Ulloa, Fullerenes and their Applications in Science and Technology [online] [accessed on 28th of November 2015]. Available from: http://web.eng.fiu.edu/~vlassov/EEE-5425/Ulloa-Fullerenes.pdf

[6]. C. B. Nielsen, S. Holliday, H.-Y. Chen, S. J. Cryer, and I. McCulloch, Acc. Chem. Res. **48**, 2803 (2015).

[7]. W. Fa, S. Chen, S. Pande, and X. Cheng Zeng, J. Phys. Chem. A **119**, 11208 (2015).

[8]. D. Bakowies, and W. Thiel, J. Am. Chem. Soc. **13**, 3704 (1991).

[9]. G. B. Adams, M. O'Keefe, and R. S. Ruoff. J. Phys. Chem. **98**, 9465 (1994).

[10]. G. Ying-Duo, and W. C. Herndon, J. Am. Chem. Soc. **115**, 8459 (1993).

[11]. J. Xiao, M. Li, Y.-N. Chiu, M. Fu, S.-T. Lai, and N. N. Li, J. Mol. Struct. **428**, 149 (1998).

[12]. R. Salcedo, and L. E. Sansores, J. Mol. Struct. **422**, 245 (1998).

[13]. X. Yang, G. Wang, Z. Yang, Z. Shang, Z. Cai, Y. Pan, B. Wu, and X. Zhao, J. Mol. Struct. **579**, 91 (2002).

[14]. M. F. Dinca, S. Ciger, M. Ştefu, F. Gherman, K. Miklos, C. Nagy, O. Ursu, and M. V. Diudea, Carpathian J. Math. **20**, (2004).

[15]. A. A. Hindi, and A. A. El-Barbary, J. Mol. Struct. **1080**, 169 (2015).

[16]. A. Kerim, J. Phys. Org. Chem. **25**, 379 (2012).

[17]. L. Jäntschi [online] 2014 [accessed August 3, 2015] Szeged Matrix Property Indices. URL: http://l.academicdirect.org/Chemistry/SARs/SMPI

[18]. S. D. Bolboacă, and L. Jäntschi, J. Chem. **2016**, Article ID 1791756 (2016).

[19]. S. L. Mayo, B. D. Olafson, and W. A. Goddard, J. Phys. Chem. **94**, 8897 (1990).

[20]. M. J. S. Dewar, E. G. Zoebisch, Eamonn H. F., and J. J. P. Stewart. J. Am. Chem. Soc. **107**, 3902 (1985).

[21]. S. D. Bolboacă, L. Jäntschi, and M. V. Diudea, Curr. Comput. Aided Drug Des. **9**, 195 (2013).

[22]. S. D. Bolboacă, and L. Jäntschi. Environ. Chem. Lett. **6**, 175 (2008).

[23]. S. D. Bolboacă, and L. Jäntschi. BIOMATH **2**, 1309089 (2013).

[24]. S. D. Bolboacă, and L. Jäntschi, Combin. Chem. High Throughput Screen. **16**, 288 (2013).

[25]. N. Chirico, and P. Gramatica, J. Chem. Inf. Model **52**, 2044 (2012).

[26]. L. Pauling, J. Am. Chem. Soc. **54**, 3570 (1932).

[27]. A. L. Allred, J. Inorg. Nucl. Chem. **17**, 215 (1961).

[28]. Steiger J. H. Psychol. Bull. **87**, 245 (1980).

Szeged Matrix Property Indices as Descriptors to Characterize Fullerenes

Lorentz Jäntschi[1,2], and Sorana D. Bolboacă[3,*]

Supplementary material

Schematic flowchart of the applied experimental design is presented in Figure 1.

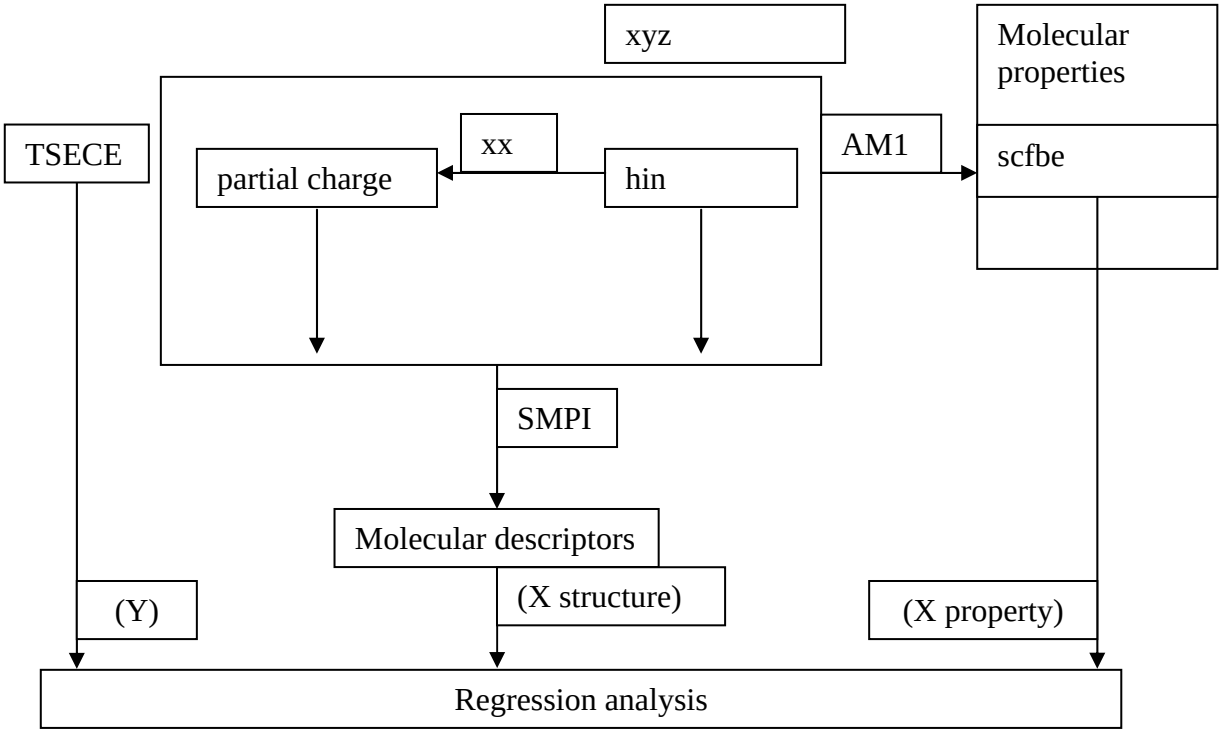

Figure 1. Summary of the applied design (TSECE = total strain energy (continuum elasticity), AM1 = Austin Model 1; scfbe = scf-binding-energy)

The results of each scenario are listed in Table 1 and Table 2, r^2 being the determination coefficient, and Q^2 being the determination coefficient in leave-one-out analysis, both expressed as percentages. The analysis of the results presented in Table 1 and 2 revealed that the total strain energy derived in the context of continuum elasticity theory (our Y) correlates high with a AM1 (Austin Model 1 [1], an accurate semi-empirical SCF method) calculated property, namely 'scfbe' (scf-binding-energy in the HyperChem formalism [2]) - over 94% of the variance explained (model with one variable in scenario 2). Identification of a model with a high goodness-of-fit as the first model in the second scenario could discourage further analysis, which it will be a mistake because the contribution of this property is discharge on model 17 in the second scenario.

[1] Dewar M. J. S., E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. J. Am. Chem. Soc **107**, 3902 (1985).
[2] HyperChem(TM) Professional 7.51, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA.

Table 1. First scenario regressions: structure-property analysis.

Model ($\hat{Y}$ =)	r^2 (%)	Q^2 (%)
1v model		
Y=b0(t0)+b1(t1)*IJUGB	77.02	75.06
2v models		
4.70e+2(t=9.48e+0)+IEUUG*5.35e-5(t=3.09e+0)+IIUTG*-1.25e-4(t=1.12e+1)	77.62	71.84
4.96e+3(t=1.20e+1)+IEUUG* 1.62e-4(t=8.31e+0)+LIUTC*-6.14e+2(t=1.21e+1)	80.19	75.35
4.23e+2(t=1.06e+1)+IEUUG* 1.59e-4(t=8.75e+0)+IIUTC*-1.39e-1(t=1.30e+1)	82.41	78.69
-1.08e+2(t=4.85e+0)+IEUUG* 5.21e-5(t=3.40e+0)+IJUGB*5.26e+0(t=1.30e+1)	82.49	78.39
4.45e+2(t=1.13e+1)+IEUGG* 3.59e-4(t=8.97e+0)+IIUTC*-1.48e-1(t=1.34e+1)	82.97	79.25
-2.86e+3(t=3.93e+0)+LJUGG* 1.18e+2(t=3.89e+0)+IJUGB*7.37e+0(t=1.06e+1)	83.68	81.28
-4.42e+3(t=7.31e+0)+IIUGG*-7.23e-5(t=7.74e+0)+LJUGE*7.42e+2(t=8.00e+0)	85.04	77.84
-4.82e+2(t=4.89e+0)+IIUGG*-7.50e-5(t=9.18e+0)+IJUGE*1.37e+0(t=9.47e+0)	88.07	81.26
-5.29e+3(t=1.10e+1)+LJEGG* 3.88e+1(t=9.71e+0)+LJUGE*6.60e+2(t=8.01e+0)	88.97	85.31
-1.82e+3(t=1.95e+1)+LJEGG* 3.96e+1(t=1.17e+1)+IJUGE*1.22e+0(t=9.84e+0)	91.66	88.13
-3.39e+2(t=4.10e+0)+IIUGF*-2.60e-3(t=1.26e+1)+IJUGE*1.80e+0(t=1.70e+1)	92.59	84.58
-1.81e+3(t=1.87e+1)+RFUGE*-3.77e+5(t=1.15e+1)+IFEGE*4.91e+0(t=1.90e+1)	92.62	90.42
5.68e+3(t=1.63e+1)+RFUGE*-3.91e+5(t=1.17e+1)+RFDGA*-4.31e+5(t=1.91e+1)	92.67	90.86
5.27e+3(t=1.61e+1)+RJUGE*-6.06e+5(t=1.76e+1)+LIUGE*-6.54e+2(t=1.29e+1)	93.17	88.87
1.77e+3(t=2.66e+1)+RJUGE*-6.07e+5(t=2.04e+1)+IIUGE*-1.15e+0(t=1.53e+1)	94.88	92.04
3v models		
2.29e+3(t=2.50e+1)+IEUUG*-2.98e-5(t=3.38e+0)+RJUGE*-4.27e+5(t=1.37e+1)+IJUTE*-1.79e+0(t=1.57e+1)	95.58	93.20
2.04e+3(t=2.63e+1)+IEUUG*-3.81e-5(t=4.25e+0)+RJUGE*-4.51e+5(t=1.50e+1)+IIUTE*-1.76e+0(t=1.60e+1)	95.70	93.44
2.07e+3(t=1.77e+1)+LMUUG*-7.66e+0(t=2.98e+0)+RJUGE*-6.62e+5(t=2.01e+1)+IIUGE*-1.35e+0(t=1.40e+1)	95.89	90.94
2.08e+3(t=1.77e+1)+LMUUG*-7.89e+0(t=3.05e+0)+RJUGE*-6.74e+5(t=2.03e+1)+IIUGD*-1.50e-3(t=1.40e+1)	95.90	93.50
2.49e+3(t=1.69e+1)+LFUGG*-2.54e+1(t=4.66e+0)+RJUGE*-4.42e+5(t=1.50e+1)+IIUTE*-1.67e+0(t=1.72e+1)	95.97	94.08
1.49e+3(t=2.29e+1)+IIUGG*-1.95e-4(t=1.22e+1)+RJUGE*-5.80e+5(t=1.85e+1)+IJPTC*1.17e-3(t=7.85e+0)	96.11	94.71
-2.37e+3(t=6.26e+0)+LJEGG*5.98e+1(t=1.35e+1)+RJUGE*-3.74e+5(t=1.23e+1)+IEUGD*1.35e-2(t=5.48e+0)	96.13	94.42
-1.45e+3(t=7.65e+0)+IEUGF*2.95e-2(t=6.16e+0)+IIUGF*-4.01e-3(t=1.48e+1)+IJUGE*1.95e+0(t=2.47e+1)	96.39	84.96
1.04e+3(t=7.85e+0)+IEUGF*2.27e-2(t=5.91e+0)+RJUGE*-6.40e+5(t=2.88e+1)+IIUGE*-1.61e+0(t=1.69e+1)	97.40	95.73
1.02e+3(t=7.81e+0)+IEUGF*2.37e-2(t=6.24e+0)+RJUGE*-6.53e+5(t=2.98e+1)+IIUGD*-1.80e-3(t=1.73e+1)	97.52	96.04
-9.29e+2(t=4.59e+0)+IEUGF*3.20e-2(t=8.26e+0)+RJUGE*-4.17e+5(t=1.94e+1)+IIEGA*2.83e-4(t=1.89e+1)	97.87	96.93
3.01e+3(t=1.15e+1)+IEEGF*-2.57e-2(t=8.82e+0)+RJUGE*-4.96e+5(t=2.43e+1)+IIEGA*3.24e-4(t=1.78e+1)	98.05	97.01
2.50e+3(t=1.28e+1)+RJUGE*-5.99e+5(t=2.46e+1)+IFEGE*-2.22e+0(t=9.31e+0)+IIEGA*3.73e-4(t=1.68e+1)	98.19	97.48
4v models		
-1.49e+3(t=1.84e+1)+LMUUG*-4.96e+0(t=2.59e+0)+IFUGE*3.34e+0(t=8.58e+0)+IJUGE*1.23e+0(t=1.82e+1)+IIEGA*2.92e-4(t=2.22e+1)	98.21	97.69
2.43e+3(t=1.28e+1)+LMUUG*-3.81e+0(t=2.16e+0)+RJUGE*-6.03e+5(t=0.59e+1)+IFEGE*-2.06e+0(t=8.60e+0)+IIEGA*3.75e-4(t=1.77e+1)	98.40	97.63
3.96e+3(t=1.41e+1)+IIUGG*-5.57e-5(t=9.22e+0)+IEEGF*-1.80e-2(t=7.74e+0)+RJUGE*-4.60e+5(t=2.47e+1)+IJUTE*-1.58e+0(t=1.41e+1)	98.42	97.40
2.58e+3(t=2.33e+1)+IIUGG*-4.87e-5(t=9.13e+0)+RFUGE*-1.02e+5(t=8.21e+0)+RJUGE*-3.95e+5(t=2.12e+1)+IJUTE*-1.44e+0(t=1.43e+1)	98.54	97.81
3.73e+3(t=1.58e+1)+IIUGG*-6.10e-5(t=9.95e+0)+RJUGE*-5.25e+5(t=2.56e+1)+IFEGE*-1.49e+0(t=8.26e+0)+IJUTE*-1.75e+0(t=1.48e+1)	98.55	97.85
2.50e+3(t=1.41e+1)+ImUGG*1.00e-3(t=2.99e+0)+RJUGE*-6.23e+5(t=2.66e+1)+IFEGE*-2.18e+0(t=1.01e+1)+IIEGA*3.91e-4(t=1.86e+1)	98.56	98.10
2.44e+3(t=1.45e+1)+LMEGG*-3.10e+0(t=3.72e+0)+RJUGE*-6.20e+5(t=2.86e+1)+IFEGE*-2.05e+0(t=9.78e+0)+IIEGA*3.90e-4(t=1.99e+1)	98.70	98.33
IFUGE*2.71e+0(t=1.04e+1)+IJUGE*2.58e+0(t=3.04e+1)+IIUGE*-3.08e+0(t=2.48e+1)+IIPTB*-5.32e-1(t=1.46e+1)	98.76	98.30

The most important point related to the results presented in the Supplementary Material is that the SMPI performs better in the estimation of the total strain energy compared with scf-binding-energy on MRL models with three or four variables. Of course, the expected results, which profs the reproducibility, is that the 'best to moment' equations from one scenario to another are changed (different) only at the beginning when the supplementary descriptors - calculated properties - make their room for describing the association with the total strain energy. Adding the calculated properties in structure-property-property analysis decrease the number of equations with improvement in r^2 and converged to identical models for the last four models with three variables and all equations with four variables.

Table 2. Second scenario: structure-property-property analysis.

Model ($\hat{Y}$ =)	r^2 (%)	Q^2 (%)
1v model		
b0(t0)+b1(t1)*scfbe	94.94	94.12
2v models		
2.09e+2(t=7.06e+0)+LMUUG*-5.28e+0(t=2.16e+0)+scfbe*2.33e-2(t=2.06e+1)	95.51	94.42
1.95e+2(t=9.33e+0)+LMEGG*-3.01e+0(t=2.39e+0)+scfbe*2.40e-2(t=1.84e+1)	95.62	94.66
1.64e+2(t=1.97e+1)+IMEGG*-2.91e-5(t=2.52e+0)+scfbe*2.42e-2(t=1.83e+1)	95.68	94.74
-3.84e+2(t=2.54e+0)+IEUGF*1.15e-2(t=3.50e+0)+scfbe*2.35e-2(t=2.59e+1)	96.20	94.65
4.06e+2(t=5.66e+0)+RFUGE*-5.06e+4(t=3.63e+0)+scfbe*2.32e-2(t=2.74e+1)	96.27	94.80
3v models		
LJUGG*-2.28e+1(t=3.44e+0)+IEUGF*1.48e-2(t=4.38e+0)+scfbe*2.29e-2(t=3.19e+1)	96.62	95.39
1.03e+3(t=3.53e+0)+LJUGG*-2.35e+1(t=2.20e+0)+RFUGE*-6.52e+4(t=4.40e+0)+scfbe*2.25e-2(t=2.59e+1)	96.71	95.03
-6.10e+2(t=3.79e+0)+IIUGG*-2.26e-5(t=2.80e+0)+IEUGF*1.79e-2(t=4.73e+0)+scfbe*2.16e-2(t=2.04e+1)	96.88	95.49
6.37e+2(t=6.44e+0)+IIUGG*-2.47e-5(t=3.09e+0)+RFUGE*-8.13e+4(t=5.07e+0)+scfbe*2.10e-2(t=2.03e+1)	97.05	95.78
LJEGG*2.17e+1(t=7.24e+0)+RFUGE*-1.01e+5(t=5.65e+0)+scfbe*1.84e-2(t=2.86e+1)	97.12	96.53
-1.55e+3(t=4.70e+0)+LIEGG*4.30e+1(t=3.84e+0)+IEUGF*1.98e-2(t=5.59e+0)+scfbe*2.12e-2(t=2.17e+1)	97.30	96.42
6.88e+2(t=3.53e+0)+LIEGG*5.31e+1(t=4.33e+0)+IEEGF*-1.63e-2(t=5.67e+0)+scfbe*2.30e-2(t=2.24e+1)	97.34	96.67
LIEGG*3.20e+1(t=7.90e+0)+RFUGE*-8.48e+4(t=5.92e+0)+scfbe*2.16e-2(t=3.69e+1)	97.41	96.85
3.32e+2(t=2.03e+0)+LIEGG*4.69e+1(t=4.20e+0)+REUGE*-1.12e+5(t=5.94e+0)+scfbe*2.12e-2(t=2.24e+1)	97.46	93.80
-1.69e+3(t=5.09e+0)+LIEGG*4.69e+1(t=4.22e+0)+IEUGD*1.03e-2(t=5.97e+0)+scfbe*2.11e-2(t=2.23e+1)	97.47	96.80
1.02e+3(t=7.81e+0)+IEUGF*2.37e-2(t=6.24e+0)+RJUGE*-6.53e+5(t=2.98e+1)+IIUGD*-1.80e-3(t=1.73e+1)	97.52	96.06
-9.29e+2(t=4.59e+0)+IEUGF*3.20e-2(t=8.26e+0)+RJUGE*-4.17e+5(t=1.94e+1)+IIEGA*2.83e-4(t=1.89e+1)	97.87	96.93
3.01e+3(t=1.15e+1)+IEEGF*-2.57e-2(t=8.82e+0)+RJUGE*-4.96e+5(t=2.43e+1)+IIEGA*3.24e-4(t=1.78e+1)	98.05	97.03
2.50e+3(t=1.28e+1)+RJUGE*-5.99e+5(t=2.46e+1)+IFEGE*-2.22e+0(t=9.31e+0)+IIEGA*3.73e-4(t=1.68e+1)	98.19	97.49
4v models		
-1.49e+3(t=1.84e+1)+LMUUG*-4.96e+0(t=2.59e+0)+IFUGE*3.34e+0(t=8.58e+0)+IJUGE*1.23e+0(t=1.82e+1)+IIEGA*2.92e-4(t=2.22e+1)	98.21	97.69
2.43e+3(t=1.28e+1)+LMUUG*-3.81e+0(t=2.16e+0)+RJUGE*-6.03e+5(t=2.59e+1)+IFEGE*-2.06e+0(t=8.60e+0)+IIEGA*3.75e-4(t=1.77e+1)	98.40	97.64
3.96e+3(t=1.41e+1)+IIUGG*-5.57e-5(t=9.22e+0)+IEEGF*-1.80e-2(t=7.74e+0)+RJUGE*-4.60e+5(t=2.47e+1)+IJUTE*-1.58e+0(t=1.41e+1)	98.42	97.42
2.58e+3(t=2.33e+1)+IIUGG*-4.87e-5(t=9.13e+0)+RFUGE*-1.02e+5(t=8.21e+0)+RJUGE*-3.95e+5(t=2.12e+1)+IJUTE*-1.44e+0(t=1.43e+1)	98.54	97.82
3.73e+3(t=1.58e+1)+IIUGG*-6.10e-5(t=9.95e+0)+RJUGE*-5.25e+5(t=2.56e+1)+IFEGE*-1.49e+0(t=8.26e+0)+IJUTE*-1.75e+0(t= .48e+1)	98.55	97.86
2.50e+3(t=1.41e+1)+ImUGG* 1.00e-3(t=2.99e+0)+RJUGE*-6.23e+5(t=2.66e+1)IFEGE*-2.18e+0(t=1.01e+1)+IIEGA*3.91e-4(t=1.86e+1)	98.56	98.10
2.44e+3(t=1.45e+1)+LMEGG*-3.10e+0(t=3.72e+0)+RJUGE*-6.20e+5(t=2.86e+1)+IFEGE*-2.05e+0(t=9.78e+0)+IIEGA*3.90e-4(t=1.99e+1)	98.70	98.33
IFUGE*2.71e+0(t=1.04e+1)+IJUGE*2.58e+0(t=3.04e+1)+IIUGE*-3.08e+0(t=2.48e+1)+IIPTB*-5.32e-1(t=1.46e+1)	98.76	98.30