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

Andrey A. Toropov · Alla P. Toropova · Valentin O. Kudyshkin

Received: 8 October 2021 / Accepted: 28 December 2021 / Published online: 7 January 2022 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

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 is 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. The statistical characteristics of the best model are the following: $n = 57$, $R^2 = 0.7764$, $RMSE = 0.039$ (active training set) and $n = 57$, $R^2 = 0.9028$, $RMSE = 0.019$ (validation set).

Keywords Refractive index of polymers · Self-consistent models · Index of ideality of correlation ($IIC$) · Monte Carlo method · CORAL software

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].

The QSPR/QSAR analysis of polymers is an intensive developing field of theoretical chemistry [8]. The search for new drugs also can be carried out with molecular simulation studies of polymers [9]. The design of relevant monomeric units can be a tool to select polymers with desired properties [10]. There are examples of the intersection of polymer science with searching nanomaterials for medicine [11, 12]. Photoelectrical and optic-electronic properties of polymers [13] are studied using molecular dynamics [14]. The random forest [15, 16] and artificial neural networks [17] also are used for the QSPR/QSAR analysis of the polymer systems. The simplified molecular input-line entry system (SMILES) is a widely used format to represent the molecular structure [18]. Recently, high–refractive index polymers have captured considerable attention of the scientific community due to various applications, aimed to improve advanced optic-electronic devices [19].

The present study aims to build up and validate the QSPR-model for the refractive index of polymers. The assessment of the predictive potential of these models was carried out via so-called system of self-consistent models of the refractive index of polymers. The index of ideality of correlation ($IIC$) also can serve as a criterion of the predictive potential. The $IIC$ demonstrates significant ability to improve the predictive potential of the QSPR-model being applied as additional component of the Monte Carlo optimization aimed to model an arbitrary endpoint.

The QSPR analysis of the refractive index of polymers can be based on the traditional physicochemical descriptors together with topological indices [19]; also in this end, the so-called expert-in-the-loop approach [20] based on automatic feature selection (e.g., descriptors of aromaticity) can serve as the basis to develop corresponding models.
The Monte Carlo technique was used to develop models for the refractive index of polymers with the representation of polymer structures by SMILES of monomers [21]. Here similar SMILES-based models are built up by applying the abovementioned IIC. In addition, the new conception of the validation, the so-called system of self-consistent QSPR-models, is used for the QSPR analysis of the refractive index of different polymers. The CORAL software (http://www.insilico.eu/coral) is applied for building up the QSPR-models.

Optimal hybrid descriptor

The CORAL software can be tuned to build up models based on solely SMILES. In addition, the software can be turned up to build up models based on optimal descriptors which are calculated by taking into account both molecular features extracted from SMILES together with molecular features represented by the graph invariants. Here, namely, the hybrid optimal descriptors [22] are used to build up models.

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

\[
RI = C_0 + C_1 \times DCW(T, N)
\]  

(1)

\[
C_0 \text{ and } C_1 \text{ are regression coefficients; the descriptor of the correlation weights (DCW) is calculated as}
\]

\[
DCW(T, N) = \sum CW(APP_k) + \sum CW(S_k) + \sum CW(EC1_k)
\]

(2)

\[
+ \sum CW(EC2_k) + \sum CW(EC3_k) + CW(C5) + CW(C6)
\]

APP_k are the atom pair’s proportions [23], S_k is SMILES attributes, EC1, EC2, and EC3 are Morgan extended connectivity of first, second, and third orders, respectively [21, 22], and C5 and C6 are special codes of rings [24]. The T is thresholds, i.e., an integer to separate SMILES attributes into rare and non-rare [22]. The rare SMILES attributes have correlation weights equal to zero; i.e., these are not involved in building up a model. N is the number of epochs of the Monte Carlo optimization.

The Monte Carlo optimization

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

The first target function (TF_1)

The first target function is calculated as the following:

\[
TF_1 = R_A + R_p - |R_A - R_p| \times 0.1
\]  

(3)

\(R_A\) and \(R_p\) are correlation coefficients between observed and predicted endpoints 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:

\[
TF_2 = TF_1 + IIC_C \times 0.5
\]  

(4)

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.

Springer
Fig. 1 Histories of the Monte Carlo optimizations with different target functions

\[ TF_1 \text{ (without IIC)} \quad \text{and} \quad TF_2 \text{ (with IIC)} \]

Active training set (○), Passive training set (△), Calibration set (×), Validation set (■)
\(IIC_C\) is the index of ideality of correlation calculated with polymers of the calibration set [24]. \(IIC\) is calculated as the following:

\[
IIC_C = r_C \frac{\min(-MAE_C, +MAE_C)}{\max(-MAE_C, +MAE_C)}
\]

\[
\text{min}(x, y) = \begin{cases} 
  x, & \text{if } x < y \\
  y, & \text{otherwise}
\end{cases}
\]

\[
\text{max}(x, y) = \begin{cases} 
  x, & \text{if } x > y \\
  y, & \text{otherwise}
\end{cases}
\]

\[
^-MAE_C = \frac{1}{N} \sum |\Delta_k|,^-N \text{ is the number of } \Delta_k < 0
\]

\[
^+MAE_C = \frac{1}{+N} \sum |\Delta_k|,^+N \text{ is the number of } \Delta_k \geq 0
\]

**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     | Without \(IIC\), Eq. 12 | Active training | 57  | 0.7545  | 0.8901 | 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 |       |      |
| 1     | With \(IIC\), Eq. 17 | 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 |       |      |
| 2     | Without \(IIC\), Eq. 13 | 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 |       |      |
| 2     | With \(IIC\), Eq. 18 | 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 |       |      |
| 3     | Without \(IIC\), Eq. 14 | 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 |       |      |
| 3     | With \(IIC\), Eq. 19 | 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 |       |      |
| 4     | Without \(IIC\), Eq. 15 | Active training | 55  | 0.9791  | 0.9894 | 0.9541 | 0.012 | 0.009| 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 |       |      |
| 4     | With \(IIC\), Eq. 20 | 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 |       |      |
| 5     | Without \(IIC\), Eq. 16 | 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 |       |      |
| 5     | With \(IIC\), Eq. 21 | 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 |       |      |
\[ \Delta_k = \text{observed}_k - \text{calculated}_k \]  

(10)

The observed and calculated are corresponding values of the endpoint.

**The system of self-consistent models**

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

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

\[
\begin{bmatrix}
(M_1 : V_1 \rightarrow R_{v1}^2) & \cdots & (M_5 : V_1 \rightarrow R_{v5}^2) \\
\vdots & \ddots & \vdots \\
(M_1 : V_5 \rightarrow R_{v1}^2) & \cdots & (M_5 : V_5 \rightarrow R_{v5}^2)
\end{bmatrix}
\]  

(11)

\(M_i\) is the i-th model, \(V_j\) is the list of polymers applied as the validation set in the case of the j-th split, and \(R_{vj}^2\) is the correlation coefficient observed for the j-th validation set if the i-th model is applied.

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 \(R_{vj}^2\) 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.

To this end, five random splits were applied to build up models for the RI of different polymers using the abovementioned 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) + 0.0006405(±0.00008) \* DCW(1, 10)  
(20)

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

Table 2 contains the statistical characteristics of models obtained by the Monte Carlo optimization with target functions \(TF_1\), and \(TF_2\).

**Table 3** Three systems of self-consistent models were obtained by the \(TF_{1'}\)– and \(TF_{2'}\)-optimizations for splits 1–5.

The average determination coefficient for the validation set obtained with \(TF_{1'}\)-optimization is 0.849 ± 0.086; the value in the case of the \(TF_{2'}\)-optimization is 0.885 ± 0.061

| 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 | \textbf{0.961} |
|                 | \(M_4\) | 17 | 0.742  | 13 | 0.712  | 21 | 0.836  | 18 | 0.847  |
|                 | \(M_5\) | 20 | 0.909  | 20 | \textbf{0.942} | 16 | 0.903  | 18 | 0.852  |
| \(TF_2\)        | \(M_1\) | 20 | \textbf{0.855} | 18 | \textbf{0.860} | 17 | \textbf{0.799} | 20 | \textbf{0.860} |
|                 | \(M_2\) | 20 | 0.865  | 16 | \textbf{0.966} | 13 | \textbf{0.961} | 20 | \textbf{0.941} |
|                 | \(M_3\) | 18 | 0.852  | 16 | \textbf{0.949} | 21 | \textbf{0.900} | 16 | 0.940  |
|                 | \(M_4\) | 17 | \textbf{0.745} | 13 | \textbf{0.749} | 21 | \textbf{0.914} | 18 | \textbf{0.891} |
|                 | \(M_5\) | 20 | \textbf{0.925} | 20 | 0.921  | 16 | \textbf{0.906} | 18 | \textbf{0.893} |

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

One can see (Table 2) that the best predictive potential is observed for the $TF_2$-optimization since the correlation coefficients for validation sets, in this case, reach maximums in comparison with $TF_1$- and $TF_2$-optimizations.

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

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

$$Quality = \bar{Rv}^2 - \Delta \bar{Rv}^2$$  \hspace{1cm} (22)

It is clear that some of the SMILES in the cases of situations $M_i : V_j \rightarrow R_{ij}^2 (i \neq j)$ 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 the $TF_1$- and $TF_2$-models for “stranger” validation sets (i.e., applying the $i$-th model to the $j$-th split, $i \neq j$). One can see that there are four better $TF_1$-models, whereas the number of the better $TF_2$-models is sixteen. Thus, a convenient measure of quality for an arbitrary QSAR-approach is demonstrated.

The measure of predictive potential could be expressed by

Table 4 The comparison with models suggested in the literature

| $R^2$ training set | $R^2$ validation set | Reference |
|-------------------|----------------------|-----------|
| 0.932             | 0.882                | [19]      |
| 0.842–0.969       | -                    | [20]      |
| 0.96              | 0.95                 | [21]      |

Table 5 The list of molecular features which are promoters for increase or decrease of the refractive index

| SMILES attribute | CW(S) run1 | CW(S) run2 | CW(S) run3 | A | P | C | d   |
|------------------|------------|------------|------------|---|---|---|-----|
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |
| SMILES attribute |            |            |            |   |   |   |     |

A frequency in active training set, $P$ frequency in passive training set, $C$ frequency in calibration set.

$TF_2$-optimizations give improvement of the statistical quality for the calibration and validation sets but in detriment of the active/passive training sets.
a classic scheme, i.e., the average determination coefficient for the validation set plus/minus its dispersion. Table 3 demonstrates that the above-average value for the TF$_2$-optimization is larger and the dispersion is smaller.

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 that of analogical approaches (Table 4).

Having results of several runs of the Monte Carlo optimization, one can obtain molecular features extracted from SMILES or from the molecular graph, which have solely positive correlation weights (promoters of increase for the refractive index) or solely negative correlation weights (promoters for decrease for the refractive index). Table 5 contains a collection of the above molecular features observed for split 1. According to Table 5, promoters of increase are the presence of chlorine atoms (Cl………), six-member aromatic rings (C6…A…1.. and C6…A…2), whereas branching (C…(………) and the presence of fluorine atoms are promoters of decrease for the refractive index. It is to be noted that the influence of fluorine was mentioned in the work [9].

The Monte Carlo technique described here was applied to the QSPR analysis of polymers [22, 25, 26]. In addition, the approach has been applied for QSPR/QSAR analysis of the system, which is not directly connected to polymer phenomenatics [27–33].

The Supplementary information 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 self-consistent models 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 for checking up the predictive potential of QSPR-models. The statistical quality of the Monte Carlo models is comparable with models for the refractive index suggested in the literature (Table 4). The system of self-consistent models gives the additional measure of the predictive potential for QSPR: the average value of the determination coefficient observed for the abovementioned stranger validation sets.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11224-021-01875-y.

Author contribution 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., and 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.

Funding This study was financially supported by the contribution of the project LIFE-VERMEER (LIFE16 ENVIT/000167).

Availability of data and material Data is available within the article or its supplementary materials.

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

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Amata E, Marrazzo A, Dichiara M, Modica MN, Salerno L, Prezzavento O, Nastasi G, Gescifina A, Romeo G, Pittalà V (2017) Comprehensive data on a 2D-QSAR model for heme oxygenase isofrom 1 inhibitors. Data Brief 15:281–299. https://doi.org/10.1016/j.dib.2017.09.036
2. Gescifina A, Floresta G, Marrazzo A, Parenti C, Prezzavento O, Nastasi G, Dichiara M, Amata E (2017) Development of a Sigma-2 receptor affinity filter through a Monte Carlo based QSAR analysis. Eur J Pharm Sci 106:94–101. https://doi.org/10.1016/j.ejps.2017.05.061
3. Gescifina A, Floresta G, Marrazzo A, Parenti C, Prezzavento O, Nastasi G, Dichiara M, Amata E (2017) Sigma-2 receptor ligands QSAR model dataset. Data Brief 13:514–535. https://doi.org/10.1016/j.dib.2017.06.022
4. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) The predictive model for band gap prediction of metal oxide nanoparticles based on quasi-SMILES. Struct Chem 32(5):1893–1905. https://doi.org/10.1007/s11224-021-01748-4
5. Perić V, Golubović M, Lazarević M, Marjanović V, Kostić T, Đorđević M, Milić D, Veselinović AM (2021) Development of potential therapeutics for pain treatment by inducing Sigma 1 receptor antagonism: in silico approach. New J Chem 45(27):12286–12295. https://doi.org/10.1039/d1nj00838b
6. Wang W, Yang B, Jia X (2021) Predicting the melting point of imidazo-bonded ionic liquids using QSPR model based on SMILES optimal descriptors. IOP Conf Ser Earth Environ Sci 859(1):012084. https://doi.org/10.1088/1755-1315/859/1/012084
7. Lotfi S, Ahmadi S, Kumar P (2021) A hybrid descriptor based QSPR model to predict the thermal decomposition temperature of imidazolium ionic liquids using Monte Carlo approach. J Mol Liq 338:116465. https://doi.org/10.1016/j.molliq.2021.116465
8. De Almeida FB, De Abreu HA, Diniz R (2019) Theoretical calculations of a porous coordination polymer formed by isonicotinyl-hydrazine, 1,4-benzenedicarboxyl and Co2+: electronic properties, lithium doping, and H2 adsorption studies. Struct Chem 30(6):2369–2377. https://doi.org/10.1007/s11224-019-01367-0
9. Eslami M, Nikkhah SJ, Eslami E, Hashemianzadeh SM (2020) A new insight into encapsulation process of a drug molecule in the polymer/surfactant system: a molecular simulation study. Struct Chem 31(5):2051–2062. https://doi.org/10.1007/s11224-020-01550-8
10. Ronova IA, Ponomarev II (2019) Design of monomeric units for rigid aromatic polymers. Struct Chem 30(5):1611–1627. https://doi.org/10.1007/s11224-019-01356-3
11. Zhu K, Chen L, Jin X, Qu C (2019) Two Cu(II) coordination polymers based on benzene-1,3,5-tricarboxylate and 1,2,4-triazole ligands: their crystal structures and application of nanoparticles in anti-esophageal cancer activity evaluation. Struct Chem 30(4):1485–1494. https://doi.org/10.1007/s11224-019-01501-4

12. Peng J-H, Wang X-L, Ran L, Song J-L, Zhang X, Li H-Y (2018) Crystal structures and anti-breast cancer activities of two new coordination polymers [Co2(bdc)(2)(bibi):1.5][H2O]2 and [Gd(bdc)(2)(Hbpp)][H2O]3. Struct Chem 29(6):1671–1675. https://doi.org/10.1007/s11224-018-1145-x

13. Luo Y-N, Jiang H-Y, Liu Z-C, Yu L-Y, Yu X-Y (2018) A new zinc coordination polymer constructed from 4-[(8-hydroxy-5-quinolinyl) azo]-benzenesulfonic acid: synthesis, structure, and photoelectrical property. Struct Chem 29(4):977–982. https://doi.org/10.1007/s11224-018-1079-3

14. Shu Y, Zhang S, Shu Y, Liu N, Yi Y, Hao J, Ding X (2019) Interactions and physical properties of energetic poly-(phthalazinone ether sulfone ketones) (PESKs) and e-exanitrohexaazaisowurtzitane (e-CL-20) based polymer bonded explosives: a molecular dynamics simulations. Struct Chem 30(3):1041–1055. https://doi.org/10.1007/s11224-018-1225-y

15. Röding M, Fager C, Olsson A, von Corswant C, Olsson E, Loren TM (2021) Three-dimensional reconstruction of porous polymer structures and physical properties of energetic poly-(phthalazinone ether sulfone ketones) (PPESKs) and e-exanitrohexaazaisowurtzitane (e-CL-20) based polymer bonded explosives: a molecular dynamics simulations. Struct Chem 30(3):1041–1055. https://doi.org/10.1007/s11224-018-1225-y

16. Fazlalit H, Ghatarband M, Mazinani S, Asadi ZA, Shiri ME, Kalae MR (2012) Predicting the mechanical properties of glass fiber reinforced polymers via artificial neural network and adaptive neuro-fuzzy inference system. Comput Mater Sci 58:31–37. https://doi.org/10.1016/j.commatsci.2012.10.012

17. Weinginer 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. https://doi.org/10.1021/ci00057a005

18. 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. https://doi.org/10.1016/j.commatsci.2017.05.022

19. Schustik SA, Craver F, Ponzioli I, Díaz MF (2021) Polymer informatics: expert-in-the-loop in QSPR modeling of refractive index. Comput Mater Sci 194:110460. https://doi.org/10.1016/j.commatsci.2021.110460

20. Duchowicz PR, Fiorelli SE, Bacelo DE, Saavedra LM, Toropova AP, Toropov AA (2015) QSPR studies on refractive indices of structurally heterogeneous polymers. Chemometr Intell Lab Syst 140:86–91. https://doi.org/10.1016/j.chemolab.2014.11.008

21. Toropov AA, Toropova AA, Kudshykin VO, Bozorov NI, Rashidova SSh (2020) Applying of the Monte Carlo technique to build up models of glass transition temperatures of diverse polymers. Struct Chem 31:1739–1743. https://doi.org/10.1007/s11224-020-01588-8

22. 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. https://doi.org/10.1007/s11224-021-01778-y

23. Toropov AA, Toropova AP (2018) Predicting cytotoxicity of 2-phenylindole derivatives against breast cancer cells using index of ideality of correlation. Anticancer Res 38(11):6189–6194. https://doi.org/10.21873/anticancerres.12972

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

25. Kudshykin 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. https://doi.org/10.3390/mol2net06-06945

26. Ding X, Kang D, Sun L, Zhan P, Liu X (2022) Combination of 2D and 3D-QSAR studies on DAPY and DNA derivatives as potent HIV-1 NNRTIs. J Mol Struct 1249:131603. https://doi.org/10.1016/j.molstruc.2021.131603

27. Duhan M, Kumar P, Sindhu J, Singh R, Devi M, Kumar A, Kumar R, Lal S (2021) Exploring biological efficacy of novel benzo-thiazole linked 2,5-disubstituted-1,3,4-oxadiazole hybrids as efficient α-amylase inhibitors: synthesis, characterization, inhibition, molecular docking, molecular dynamics and Monte Carlo based QSAR studies. Comput Biol Med 138:104876. https://doi.org/10.1016/j.compbiomed.2021.104876

28. Kumar A, Kumar P (2021) Identification of good and bad fragments of tricyclic triazine analogues as potential PKC-0 inhibitors through SMILES-based QSAR and molecular docking. Struct Chem 32(1):149–165. https://doi.org/10.1007/s11224-020-01629-2

29. Kumar A, Kumar P (2021) Prediction of power conversion efficiency of phenothiazine-based dye-sensitized solar cells using Monte Carlo method with index of ideality of correlation. SAR QSAR Environ Res 32(10):817–834. https://doi.org/10.1080/1062936X.2021.1973095

30. Ghiassi T, Ahmadi S, Ahmadi E, Talei Bavl Olyai MR, Khodadadi Z (2021) The index of ideality of correlation: QSAR studies of hepatitis C virus NS3/4A protease inhibitors using SMILES descriptors. SAR QSAR Environ Res 32(6):495–520. https://doi.org/10.1080/1062936X.2021.1925344

31. Jafari K, Fatemi MH (2020) Application of nano-quantitative structure–property relationship paradigm to develop predictive models for thermal conductivity of metal oxide-based ethylene glycol nanofluids. J Therm Anal Calorim 142(3):1335–1344. https://doi.org/10.1007/s10973-019-09215-3

32. Duhan M, Singh R, Devi M, Sindhu J, Bhatia R, Kumar A, Kumar P (2021) Synthesis, molecular docking and QSAR study of thia-zole clubbed pyrazole hybrid as α-amylase inhibitor. J Biomol Struct Dyn 39(1):91–107. https://doi.org/10.1080/07391102.2019.1704885

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.