# The Self-organizing Vector of Atom-pairs Proportions: Use to Develop Models for Melting Points

**Alla P. Toropova** ( alla.toropova@marionegri.it )
Istituto di Ricerche Farmacologiche Mario Negri IRCCS https://orcid.org/0000-0002-4194-9963
**Andrey A. Toropov**
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
**Emilio Benfenati**
Istituto di Ricerche Farmacologiche Mario Negri IRCCS

**Research Article**

**Keywords:** QSPR model, melting point, large database, Monte Carlo method, CORAL software

**Posted Date:** March 22nd, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-309701/v1



**Version of Record:** A version of this preprint was published at Structural Chemistry on April 5th, 2021. See the published version at https://doi.org/10.1007/s11224-021-01778-y.

# The self-organizing vector of atom-pairs proportions: use to develop models for melting points

Alla P. Toropova*, Andrey A. Toropov, Emilio Benfenati

*Department of Environmental Health Science, Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche Mario Negri, Via Mario Negri 2, 20156 Milano, Italy*

**Abstract**

Atom-pairs proportions are the transparent quality of a molecule: if a molecule has two atoms of oxygen and three atoms of nitrogen, the atom-pair atom1-atom2 can be expressed as a code 'atom1-atom2-n1-n2', indicating the different atoms and their numbers. These codes for a group of atoms (nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, bromine, as well as, double and triple covalent bonds) are applied to build up the so-called optimal molecular descriptor calculated with special coefficients named correlation weights of corresponding pairs. The numerical data on the correlation weights are calculated by the Monte Carlo technique using the CORAL software (http://www.insilico.eu/coral). The one-variable model for melting points of 8653 various organic compounds is characterized by the following statistical quality: n=6483, r2=0.6452; RMSE=61.9'C; n=2170, r2=0.7941, RMSE=39.2'C.

**Keywords:** QSPR model; melting point; large database; Monte Carlo method; CORAL software

*) Corresponding author: Alla P. Toropova, Email: alla.toropova@marionegri.it

Laboratory of Environmental Chemistry and Toxicology,

Istituto di Ricerche Farmacologiche Mario Negri

Via Mario Negri 2, 20156 Milano, Italy

Tel: +39 02 3901 4595

Fax: +39 02 3901 4735

## Introduction

The knowledge of the physical and chemical properties of a compound is required for understanding and modeling the action of a compound for various possible applications. Many physicochemical properties of compounds are strongly interconnected. The relationships between physicochemical properties can be established by the theoretical analysis or found empirically [1-3]. Physicochemical properties are also very important in the case of toxicokinetic models, to describe partitioning between the different organs, and processes as skin permeation [4].

The numerical data on melting points has a diversity of applications [5, 6]. For example, information on the melting point is applicable to check up the purity of an experimental sample. In addition, the information is applicable to establish the existence of various conformations of the molecular structure for a compound under consideration. Data on melting point may be useful information for both studies of pure substances and studies of mixtures. However, the experimental definition of this endpoint is impossible for all substances applying in science and everyday life [7-9].

Similar to most quantitative structure-property/activity relationships (QSPRs/QSARs), the models for melting points are based on a well-known group of mathematical approaches, e.g. linear regression and neural networks [10], random forest [9], comparative molecular field analysis (CoMFA) [11], partial least squares [12], k-nearest neighbor approach [13], and Monte Carlo approach [14]. Thus, modeling of melting points is not a simple task and previously melting point models for relatively simple and small compound sets have been developed [15].

Nonetheless, relationships between molecular structure and biological activity or molecular structure and physical properties can be investigated for large databases of organic compounds using the newest computer-assisted conceptions aimed to derive quantitative relationships between a property and a structure via modelling [1,2]. Sometimes, a simple Monte Carlo approach provides reliable models for large datasets of complex molecules [16].

The Monte Carlo approach based on so-called optimal descriptors was studied as a tool to build up models for various endpoints such as bioactivity of anticonvulsant agents [17], the biological activity of various drug-like substances [18-22], and biological activity of nanomaterials [23]. However, the approach was not applied to build up models for melting points.

The aim of the present study is the assessment of the approach as a tool to build up models for melting points for a large set of organic compounds that contains more than eight and a half thousand various organic compounds.

## Method

### *Data*

The numerical data on the melting point expressed in Celsius has been taken in the literature [24]. The data has been randomly distributed into four sub-set an equivalent percentage. The active training set (25%), passive training set (25%), and calibration set (25%) are a special group of training for the model. The external validation set (25%) is used to estimate the predictive potential of the model. The range of melting point for united active training set, passive training set, and calibration set is min = -196$^0$C and max = 492.5$^0$C; the range for validation set is min = -134$^0$C and max = 376$^0$C.

### *Model*

The model for the melting point is the following one-variable generalized formula

$$T_m^0 = C_0 + C_1 \times DCW(T,N) \quad (1)$$

The $T_m^0$ is expressed in Celsius; $C_0, C_1$ are regression coefficients;

*DCW(T,N)* is the so-called optimal descriptor (*D*) that is calculated with correlation weights (*CW*).

### *Descriptor of Correlation weights (DCW)*

The descriptor is calculated as the following:

$$DCW(T,N) = \sum CW(APP_j) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (2)$$

The $AAP_j$ is a vector of atom-pairs proportions. These are special SMILES attributes encoded as the following: (atom1, atom2).n1.n2. The atom1, atom2 can be N, O, S, P, F, Cl, Br, and addition SMILES-atoms: '=' (double covalent bond), and '#' (triple covalent bond). Each compound (SMILES) has a self-organized $AAP_j$ vector where includes atom-pairs present in the corresponding molecule, whereas atom-pairs for atoms absent in the molecule is not appearing.

It is expected that the correlation weights of $AAP_j$ can improve the predictive potential of the model for melting points. Table 1 contains an example of the vector of atom-pairs proportions.

Table 1.

An example of building a self-organized vector of atom-pairs proportions ($APP_j$) for the following SMILES: CCC(Br)(CC)C(=O)NC(N)=O

| Atom1 | Atom2 | n1 | n2 | Code applied to build up the model |
|---|---|---|---|---|
| Br | N | 1 | 2 | (Br.N)..1.2. |
| Br | O | 1 | 2 | (Br.O)..1.2. |
| Br | Double bond | 1 | 2 | (Br.=)..1.2. |
| N | O | 2 | 2 | (N..O)..2.2. |
| N | Double bond | 2 | 2 | (N..=)..2.2. |
| O | Double bond | 2 | 2 | (O..=)..2.2. |

The $S_k$ is the "SMILES-atom" i.e. one symbol or two symbols (e.g. 'C', 'N', 'O', etc.) which cannot be examined separately (e.g. 'Cl', 'Si', etc.); the $SS_k$ is a combination of two SMILES-atoms; the $SSS_k$ is a combination of three SMILES-atoms; the $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ are so-called correlation weights of the above-mentioned attributes of SMILES. The numerical data on the $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ are calculated with the Monte Carlo method, i.e. the optimization procedure which gives a maximal value of a special target function (*TF*).

Table 2 contains an interpretation of SMILES attributes. The correlation weights for the SMILES attributes are calculated by the Monte Carlo technique using the CORAL software (http://www.insilico.eu/coral).

Table 2

The list of SMILES attributes for the following SMILES: CCC(Br)(CC)C(=O)NC(N)=O

| $S_k$ | $SS_k$ | $SSS_k$ |
|---|---|---|
| C..... | | |
| C..... | C...C..... | |
| C..... | C...C..... | C...C...C... |
| (..... | C...(..... | C...C...(... |
| Br..... | Br..(..... | Br..(...C... |
| (..... | Br..(..... | (...Br..(... |
| (..... | (...(..... | Br..(...(... |
| C..... | C...(..... | C...(...(... |
| C..... | C...C..... | C...C...(... |
| (..... | C...(..... | C...C...(... |

| | | |
|---|---|---|
| C..... | C...(..... | C...(...C... |
| (..... | C...(..... | (...C...(... |
| =..... | =...(..... | C...(...=... |
| O..... | O...=..... | O...=...(... |
| (..... | O...(..... | =...O...(... |
| N..... | N...(..... | O...(...N... |
| C..... | N...C..... | C...N...(... |
| (..... | C...(..... | N...C...(... |
| N..... | N...(..... | N...(...C... |
| (..... | N...(..... | (...N...(... |
| =..... | =...(..... | N...(...=... |
| O..... | O...=..... | O...=...(... |

***Monte Carlo optimization of the correlation weights***

The Monte Carlo method applied here is based on two different target functions $TF_1$ and $TF_2$:

$$TF_1 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| * 0.5 \quad (3)$$

$$TF_2 = TF_1 + IIC_{CLB} * 0.5 \quad (4)$$

The $r_{AT}$ and $r_{PT}$ are the correlation coefficient between the observed and predicted values of the endpoint for the active training and passive training sets, respectively.

The index of ideality of correlation (*IIC*) is special characteristic able to improve the predictive potential of a model [25,26].

The $IIC_{CLB}$ is calculated with data on the calibration set as the following:

$$IIC_{CLB} = r_{CLB} \frac{\min(^{-}MAE_{CLB}, ^{+}MAE_{CLB})}{max(^{-}MAE_{CLB}, ^{+}MAE_{CLB})} \quad (5)$$

$$^{-}MAE_{CLB} = \frac{1}{^{-}N} \sum_{k=1}^{^{-}N} |\Delta_k|, \qquad \Delta_k < 0; \; ^{-}N \; is \; the \; number \; of \; \Delta_k < 0 \quad (6)$$

$$^{+}MAE_{CLB} = \frac{1}{^{+}N} \sum_{k=1}^{^{+}N} |\Delta_k|, \qquad \Delta_k \geqslant 0; \; ^{+}N \; is \; the \; number \; of \; \Delta_k \geqslant 0 \quad (7)$$

$$\Delta_k = observed_k - calculated_k \quad (8)$$

The observed and calculated are the corresponding values of the endpoint.

***Domain of applicability***

Domain of applicability of the CORAL model is defined according to the distribution of SMILES attributes in the active training and calibration sets ($A_k = S_k$, $SS_k$, $SSS_k$, and $APP_j$): the defect of SMILES-atom calculated as

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) + N(A_k)} \quad (9)$$

where $P(A_k)$ and $P'(A_k)$ are the probability of $A_k$ in the training and calibration sets, respectively; $N(A_k)$ and $N'(A_k)$ are frequencies of $A_k$ in the training and calibration sets, respectively.

The defect of SMILES in whole (SMILES-defect) calculated as:

$$D_j = \sum_{k=1}^{NA} d_k \quad (10)$$

where NA is the number of non-blocked SMILES attributes in the SMILES.

A substance falls in the domain of applicability if

$$Dj < 2 * \overline{D} \quad (11)$$

where $\overline{D}$ is an average of the statistical SMILES-defect for the training set.

It is to be noted, the sum of SMILES-defect on the active training set is a measure of the statistical quality of the selected split.

**Results and Discussion**

Two new elements of the CORAL model are studied here: (1) Self-organized vector of atom-pairs proportions (*APP*); and (2) index of ideality of correlation (*IIC)*. Table 3 contains the statistical quality of models obtained with applying and without applying these elements.

Table 3

The statistical characteristics of models for melting points obtained with various combinations of applying (non-applying) of *APP* and *IIC*

| ***APP*** | ***IIC*** | **Set** | ***n*** | $r^2$ | ***RMSE($^0$C)*** |
|---|---|---|---|---|---|
| **Non-applied** | **Applied** | Active training | 2155 | 0.4352 | 83.2 |

| | | | | | |
|---|---|---|---|---|---|
| | | Passive training | 2159 | 0.4408 | 83.5 |
| | | Calibration | 2169 | 0.6490 | 51.5 |
| | | Validation | 2170 | 0.6564 | 50.7 |
| **Applied** | **Non-applied** | Active training | 2155 | 0.7405 | 57.2 |
| | | Passive training | 2159 | 0.7404 | 57.7 |
| | | Calibration | 2169 | 0.6114 | 61.6 |
| | | Validation | 2170 | 0.6070 | 62.3 |
| **Applied** | **Applied** | Active training | 2155 | 0.5661 | 72.9 |
| | | Passive training | 2159 | 0.5733 | 72.9 |
| | | Calibration | 2169 | 0.7961 | 39.8 |
| | | Validation | 2170 | 0.7941 | 39.2 |

A comparison of these models (Table 3) has shown that the best model was obtained by applying both $APP_j$ and *IIC*. Applying $APP_j$ without *IIC* gives an improvement of the statistical characteristics for the group of training (active and passive), but it accompanied a decrease of the determination coefficient for the validation set.

The best model for melting point according to the statistical quality for the validation set is the following

$$T_m^0 = -34.99 + 8.132 \times DCW(1,15) \qquad (12)$$

Table 4 contains a comparison of the model calculated with Eq. 12 with models for melting points suggested in the literature. It should be noted that the first pioneer works dedicated to models of melting points were oriented to limited datasets, but models for melting points for large datasets gradually become more popular [1].

The comparison confirms that the CORAL model calculated with $APP_j$ and *IIC* is quite comparable with other models for melting points. However, it is important to notice that the models in the literature have a higher complexity at least regarding the characterization of the chemical information and the associated descriptors. The model we present here only uses very simple chemical information, such as atom type, their

number, and the information derived directly from the SMILES string, without further steps for the calculation of molecular descriptors.

Table 4.

The statistical quality of models for melting point taken from the literature

| | | **Training set** | | | **Validation set** | | | |
|---|---|---|---|---|---|---|---|---|
| | ***n*** | ***r²*** | ***RMSE*(⁰C)** | ***MAE*(⁰C)** | ***n*** | ***r²*** | ***RMSE*(⁰C)** | ***MAE*(⁰C)** |
| [1] | 38.167 | - | 37.1 | - | - | - | - | - |
| [5] | 42 | 0.941 | 24.6 | - | - | - | - | - |
| [6] | 979 | 0.808 | 35.5 | - | - | - | - | - |
| [9] | - | 0.67 | 44 | - | - | 0.66 | 46 | - |
| [10] | 62 | 0.97 | - | - | - | 0.85 | - | - |
| [11] | 51 | 0.978 | - | - | - | - | - | - |
| [12] | 3.000 | 0.81 | - | - | 1.173 | 0.80 | - | - |
| [15] | - | - | - | - | 277 | 0.662 | | 32.6 |
| [24] | 6.486 | 0.75 | 49.1 | - | 2.167 | 0.74 | 52.2 | - |
| [27] | 432 | 0.96 | 13.2 | - | 105 | 0.76 | 27 | - |
| Eq. 12 | 6.483* | 0.645** | 61.9** | 49.1** | 2.170 | 0.794 | 39.2 | 31.2 |

*) The 6483 is the sum of compounds in the active training set (n=2155), passive training set (n=2159), and calibration set (n=2169)

**) average values

**Conclusions**

The here described approach allows building up the model for melting point comparable with related models described in the literature [24]. The suggested self-organizing vector of Atom-Pairs Proportions together with the Index of Ideality of Correlation can serve as a tool to improve the predictive potential of QSPR/QSAR models. The CORAL software gives the possibility to define and study various hypotheses on the aspect of improving the predictive potential of these models by the use of the Index of Ideality of Correlation. This criterion improves the statistical quality of a model for the calibration set to the detriment of the training set. Likely, the above index deserves further study both for melting points modelling as well as for modelling of any endpoints.

**Supplementary Materials:** *Supplementary materials* section contains details on the model calculated with Eq. 12. Table S1 contains experimental and calculated melting points in Celsius (validation set); Table S2

contains the numerical data on the correlation weights; Table S3 contains the comparison of learning curves observed in the case optimization with $CW(APP_j)$ and optimization without $CW(APP_j)$.

**Funding:** The authors are grateful for the contribution of the project LIFE-CONCERT (LIFE17 GIE/IT/000461) for the financial support.

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

**Availability of data and material:** Data available within 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 E.B..; methodology, A.P.T., A.A.T., and E.B.; software, A.A.T..; validation, A.P.T., A.A.T., and E.B.; formal analysis, A.P.T.; data curation, A.P.T., A.A.T.; writing—original draft preparation, A.P.T., A.A.T.; writing—review and editing, A.P.T., A.A.T., and E.B.; supervision, E.B. 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-CONCERT (LIFE17 GIE/IT/000461) for support.

**References**

1. Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, Charochkina L, Asiri AM (2014) How Accurately Can We Predict the Melting Points of Drug-like Compounds? J Chem Inf Model 54(12): 3320–3329. https://doi.org/10.1021/ci5005288

2. Tetko IV, Yan A, Gasteiger J (2018) Prediction of Physicochemical Properties of Compounds. In: Engel T, Gasteiger J (Eds) Applied Chemoinformatics, Chapter 3, pp. 53-81. doi:10.1002/9783527806539.ch3

3. Yan F, Shi Y, Wang Y, Jia Q, Wang Q, Xia S (2020) QSPR models for the properties of ionic liquids at variable temperatures based on norm descriptors. Chem Eng Sci 217: 115540. https://doi.org/10.1016/j.ces.2020.115540

4. ten Berge W (2009) A simple dermal absorption model: derivation and application. Chemosphere 75(11): 1440-1445. DOI: 10.1016/j.chemosphere.2009.02.043

5. Dearden JC (1991) The QSAR prediction of melting point, a property of environmental relevance. Sci Total Environ 109–110: 59-68. https://doi.org/10.1016/0048-9697(91)90170-J

6. Dearden JC (2003) Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. Environ Toxicol Chem 22: 1696-1709. DOI:10.1897/01-363

7. Brown TN, Armitage JM, Arnot JA (2019) Application of an Iterative Fragment Selection (IFS) Method to Estimate Entropies of Fusion and Melting Points of Organic Chemicals. Mol Inform 38(8-9): 1800160. DOI: 10.1002/minf.201800160

8. McDonagh JL, Van Mourik T, Mitchell JBO (2015) Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation. Mol Inform 34(11-12): 715-724. DOI: 10.1002/minf.201500052

9. Venkatraman V, Evjen S, Knuutila HK, Fiksdahl A, Alsberg BK (2018) Predicting ionic liquid melting points using machine learning. J Mol Liq 264: 318-326. DOI: 10.1016/j.molliq.2018.03.090

10. Fatemi MH, Izadian P (2012) In silico prediction of melting points of ionic liquids by using multilayer perceptron neural networks. J Theor Comput Chem 11(1): 127-141. DOI: 10.1142/S0219633612500083

11. Park HY, Li J, Park B-H, Kim CK (2015) MSEP and CoMFA studies on the melting points of nitroaromatic compounds. Bull Korean Chem Soc 36(7): 1838-1847. DOI: 10.1002/bkcs.10356

12. Hemmateenejad B, Shamsipur M, Zare-Shahabadi V, Akhond M (2011) Building optimal regression tree by ant colony system-genetic algorithm: Application to modeling of melting points. Anal Chim Acta 704(1-2): 57-62. DOI: 10.1016/j.aca.2011.08.010

13. Bhat AU, Merchant SS, Bhagwat SS (2008) Prediction of melting points of organic compounds using extreme learning machines. Ind Eng Chem Res 47(3): 920-925. DOI: 10.1021/ie0704647

14. Kang J-W, Kwon OK, Lee S, Lee SH, Kim DH, Hwang H-J (2010) Kinetic lattice Monte Carlo simulations of vacancy diffusion in silicon below the melting point. J Comput Theor Nanosci 7(3): 604-611. DOI: 0.1166/jctn.2010.1401

15. Karthikeyan M, Glen RC, Bender A (2005) General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. Chem Inf Model 45(3): 581–590. https://doi.org/10.1021/ci0500132

16. Toropova AP, Toropov AA (2014) CORAL software: Prediction of carcinogenicity of drugs by means of the Monte Carlo method. Eur J Pharm Sci 52(1): 21-25. DOI: 10.1016/j.ejps.2013.10.005

17. Garro Martinez JC, Duchowicz PR, Estrada MR, Zamarbide GN, Castro EA (2011) QSAR study and molecular design of open-chain enaminones as anticonvulsant agents. Int J Mol Sci 12(12): 9354-9368. DOI: 10.3390/ijms12129354

18. Achary PGR (2014) Simplified molecular input line entry system-based optimal descriptors: QSAR modelling for voltage-gated potassium channel subunit Kv7.2. SAR QSAR Environ Res 25(1): 73-90. DOI: 10.1080/1062936X.2013.842930

19. Toropov AA, Toropova AP, Carnesecchi E, Benfenati E, Dorne JL (2020) The Index of Ideality of Correlation and the variety of molecular rings as a base to improve model of HIV-1 protease inhibitors activity. Struct Chem 31: 1441–1448. https://doi.org/10.1007/s11224-020-01525-9

20. Toropov AA, Toropova AP, Veselinović AM, Leszczynska D, Leszczynski J (2020) SARS-CoV Mpro inhibitory activity of aromatic disulfide compounds: QSAR model. J Biomol Struct Dyn Published online: 09 Sep 2020. DOI: 10.1080/07391102.2020.1818627

21. Rescifina 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. DOI: 10.1016/j.ejps.2017.05.061

22. Veselinović AM, Veselinović JB, Živković JV, Nikolić GM (2015) Application of smiles notation based optimal descriptors in drug discovery and design. Curr Top Med Chem 15(18): 1768-1779. DOI: 10.2174/1568026615666150506151533

23. Toropov AA, Toropova AP (2021) Quasi-SMILES as a basis for the development of models for the toxicity of ZnO nanoparticles. Sci Total Environ 772: 145532. https://doi.org/10.1016/j.scitotenv.2021.145532

24. Mansouri K, Grulke CM, Judson RS, Williams AJ (2018) OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminformatics 10(1): 10. DOI: 10.1186/s13321-018-0263-1

25. Toropov AA, Toropova AP (2017) The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models? Mutat Res-Gen Tox En 819: 31-37. DOI: 10.1016/j.mrgentox.2017.05.008

26. Toropova AP, Toropov AA (2019) Does the Index of Ideality of Correlation Detect the Better Model Correctly? Mol Inform 38(8-9): 1800157. DOI: 10.1002/minf.201800157

27. Watkins M, Sizochenko N, Rasulev B, Leszczynski J (2016) Estimation of melting points of large set of persistent organic pollutants utilizing QSPR approach. J Mol Model 22(3): 1-14. https://doi.org/10.1007/s00894-016-2917-0

# Supplementary Files

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

- SupplemMaterials.xlsx