3D-MoRSE descriptors explained

Oleg Devinyak [a,*], Dmytro Havrylyuk [b], Roman Lesyk [b]

[a] Department of Pharmaceutical Disciplines, Uzhgorod National University, 88000 Uzhgorod, Ukraine
[b] Department of Pharmaceutical, Organic and Bioorganic Chemistry, Danylo Halytsky Lviv National Medical University, 79010 Lviv, Ukraine

ARTICLE INFO

ABSTRACT

3D-MoRSE is a very flexible 3D structure encoding framework for chemoinformatics and QSAR purposes due to the range of scattering parameter values and variety of weighting schemes used. While arising in many QSAR studies, up to this time they were considered as hardly interpreted and were treated like a "black box". This study is intended to lift the veil of mystery, providing a comprehensible way to the interpretation of 3D-MoRSE descriptors in QSAR/QSPR studies.

The values of these descriptors are calculated with rather simple equation, but may vary when using differing starting geometries as optimization input. This variation increases with scattering parameter and also is higher for electronegativity weighted and unweighted descriptors. Though each 3D-MoRSE descriptor incorporates the information about the whole molecule structure, its final value is derived mostly from short-distance (up to 3 Å) atomic pairs. And, if a QSAR study covers structurally similar set of compounds, then the role of 3D-MoRSE descriptor in a model can be interpreted using just several pairs of neighbor atoms. The guide to interpretation process is discussed and illustrated with a case study. Realizing the mathematical concept behind 3D-descriptors and knowing their properties it is easy not only to interpret, but also to predict the importance of 3D-MoRSE descriptors in a QSAR study. The process of prediction is described on the practical example and its accuracy is confirmed with further QSAR modeling.

1. Introduction

QSAR/QSPR modeling is a widely recognized and successful approach to the prediction of compounds' physical and biological properties. Its major goal is to represent the studied attribute as some function of other numerical molecular properties and features called molecular descriptors. And one of the crucial steps in QSAR modeling is the interpretation of obtained models in terms of understandable and clear relationships between structure and activity. The importance of QSAR models interpretation is widely recognized [1,2]. Dearden et al. [3] defined the lack of mechanistic interpretation as an error in QSAR development. OECD guideline makes this point less strict, stating that "a (Q)SAR should be associated with a mechanistic interpretation, if possible" [4]. According to this document, mechanistic interpretation "is the basis for discovery of underlying causal relationships" and its consistency "with other knowledge of fundamental processes in chemistry and toxicology adds to credibility and acceptance of the predictions from the model". The knowledge extracted during interpretation of QSAR

model also can be used to design new potent compounds or to describe molecular regions that are involved into interaction with (often unknown) biotarget [5–9]. Citing the OECD guideline further, "mechanistic interpretations of (Q)SARs begin with the number and the nature of the molecular descriptors used in the model". That is why understanding of the meaning of descriptors is so important during QSAR interpretation step. A lot of descriptors such as partition coefficient log P, HOMO energy, molecular weight, functional group counts and so on are easy interpretable, since their meaning is clear and straightforward. But a major part is much more sophisticated and need sometimes significant efforts during interpretation. One class of such descriptors is represented with 3D-MoRSE. 3D-MoRSE denotes 3D molecular representations of structure based on electron diffraction descriptors and has been introduced in 1996 by Schuur, Gasteiger and coauthors in two seminal papers [10,11]. These descriptors have found a broad application and have been shown as predominant in a number of QSAR/QSPR studies [12–22]. The majority of these papers describe the effect of 3D-MoRSE values on activity but lack of interpretation how the values of used 3D-MoRSE descriptors relate to the molecular structure. Actually, the inventors of 3D-MoRSE comprehensively describe computational method used to obtain these descriptors but did not give a key to their interpretation. Two studies of Saíz-Urra et al. [14,15] were

focused primarily to 3D-MoRSE descriptors giving the conclusion that "the distances among the different atoms will be the principal means to separate the molecules according to their structural features when the other parameters remain constant" and "deeper analysis is necessary to interpret the application of these kinds of descriptors". This paper is intended to evaluate the properties and to disclose the chemical meaning of 3D-MoRSE descriptors.

2. Theory and intuition behind 3D-MoRSE descriptors

2.1. Description and calculation of 3D-MoRSE descriptors

3D-MoRSE (Molecular Representation of Structures based on Electronic diffraction) descriptors were introduced in 1996 by Schuur, Selzer and Gasteiger with the motivation for encoding 3D structure of a molecule by a fixed number of variables [10,11]. Indeed, the most obvious way to present 3D structure is its representation within cartesian or internal coordinates. But the statistical methods used in computational chemistry cannot handle such objects; these methods are requiring data with fixed number of features instead. Simplifying equations used in electron diffraction studies, the authors have got the function (1):

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin s r_{ij}}{s r_{ij}} \tag{1}$$

where s is the scattering parameter, r_{ij} is the euclidean distance between ith and jth atoms, N is the total number of atoms and A_i and A_j are different atomic properties used as weights. Each term of this function depends on distance and thus may be viewed as a radial basis function itself. Further we will use the words "term", "summand" or "radial basis function" interchangeably. Assigning to s integer values in the range of 0–31 Å^{-1}, 32 values of function 1 can be calculated. These values constitute the 3D-MoRSE code of a molecule. There are few programs that provide possibility to calculate 3D-MoRSE descriptors [23–25], and we extend this set with our free "3dmorse" program which is written in C++ and currently supports only MOPAC2012 output files as input [26]. 3D-MoRSE descriptors usually are calculated with different weights. For example, widely used DRAGON program (version 5.5) for descriptor calculations proposes five kinds of the descriptors: unweighted 3D-MoRSE and weighted with atomic mass, atomic van der Waals volume, atomic Sanderson electronegativity and atomic polarizability [24]. It is worth to note, that in this program the numeration of 3D-MoRSE descriptors starts from 1, so, for example, Mor01u denotes unweighted descriptor with scattering parameter $s = 0$ Å^{-1} (since scattering parameter starts from zero), Mor02u denotes descriptor with $s = 1$ Å^{-1} and so on. This point often is confusing and, due to wide usage of DRAGON in QSAR studies, became the reason for large number of misinterpretations.

For better understanding of 3D-MoRSE descriptors nature, let us calculate several simple examples by hands, starting with methane molecule. So, the simplest descriptor among the studied class is Mor01u. Putting zero instead of s is giving

$$I(0) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(0 \times r_{ij})}{0 \times r_{ij}} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(0)}{0} \tag{2}$$

Actually, the value of sinus zero divided by zero is undefined, however, the limit (3) is one of the most important limits in trigonometry and is equal to 1:

$$\lim_{\theta \to 0} \frac{\sin(\theta)}{\theta} = 1 \tag{3}$$

Since the descriptor is unweighted, all $A_i = A_j = 1$, and we get simply a number of possible atom pairwise combinations, which is

equal to corresponding binomial coefficient and can be calculated with factorials:

$$I(0) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} 1 = \binom{N}{2} = \frac{N!}{2!(N-2)!} \tag{4}$$

Clearly, this descriptor is a function of number of atoms only, moreover, 3D-MoRSE descriptors with scattering parameter $s = 0$ Å^{-1} are always positive for all positive weightings schemes (like atomic mass, van der Waals atom volume, electronegativity, polarizability, but not partial charge).

For methane, which has 5 atoms, Mor01u is equal to

$$\text{Mor01u} = \frac{5!}{2!(5-2)!} = 10 \tag{5}$$

For the second descriptor, Mor02u with $s = 1$, we should consider another case (keeping in mind, that $A_i = A_j = 1$ for unweighted 3D-MoRSE):

$$I(1) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \frac{\sin(1 \times r_{ij})}{1 \times r_{ij}} \tag{6}$$

To calculate this, the pairwise distances between all atoms in the molecule should be given. Thus the geometry of methane has been modeled with PM7 semiempirical method in MOPAC2012 [27]. The results show the same interatomic distance for 4 C–H pairs (1.085 Å) and slightly varying distances for 6 H–H pairs (range 1.771–1.773 Å, mean 1.772 Å). Now, since the methane molecule is highly symmetric and has repeating interatomic distance values, we can write the expression for Mor02u as:

$$\text{Mor02u} = 4\frac{\sin(1.085)}{1.085} + 6\frac{\sin(1.772)}{1.772} = 3.260 + 3.318 = 6.578 \tag{7}$$

Going to more complicated molecule like acetic acid, we still can measure all interatomic distances with some chemical computation or visualization tool and then provide necessary calculations. Using information from Table 1 (geometry has been optimized with PM7 method), we can write the weighted by atomic electronegativities Mor02e as a series of summands. Since the DRAGON program uses carbon-scaled weights, we will also use such convention. Anyway, the difference between the results with original and carbon-scaled weights lays in the constant multiplier, so this point has no effect in QSAR studies. The first term in a sum corresponds to O–O atom pair:

$$1.327 \times 1.327 \times \frac{\sin(1 \times 2.191)}{1 \times 2.191} = 0.654 \tag{8}$$

Similarly, all other terms can be calculated and placed in an upper diagonal matrix, where the diagonal elements will display the contribution of corresponding atom (Table 2). The contribution is represented as a half-sum of all terms where studied atom participates (in such a way diagonal sum is equal to Mor02e value).

This way, a value of Mor02e is set at 11.431, and exactly the same value is given by DRAGON program. Atoms contributions can be further used to identify the most and the less influential atoms or groups in the molecule.

2.2. Theoretical studies of 3D-MoRSE descriptors

Let us consider the roles of all three parameters (pairwise interatomic distance, weighting and scattering) in 3D-MoRSE framework.

Table 1
Parameters of acetic acid 3D structure required to calculate Mor02e.

No.	Atom pair	Interatomic distance, Å	No.	Atom pair	Interatomic distance, Å	No.	Atom	Carbon-scaled Sanderson electronegativity
1	O(1)–O(2)	2.191	15	C(3)–H(5)	1.993	1	O	1.327
2	O(1)–C(3)	1.198	16	C(3)–H(6)	2.179	2	C	1
3	O(1)–C(4)	2.411	17	C(3)–H(7)	2.146	3	H	0.942
4	O(1)–H(5)	3.041	18	C(3)–H(8)	2.147			
5	O(1)–H(6)	3.256	19	C(4)–H(5)	2.484			
6	O(1)–H(7)	2.953	20	C(4)–H(6)	1.096			
7	O(1)–H(8)	2.645	21	C(4)–H(7)	1.101			
8	O(2)–C(3)	1.366	22	C(4)–H(8)	1.104			
9	O(2)–C(4)	2.426	23	H(5)–H(6)	2.301			
10	O(2)–H(5)	0.978	24	H(5)–H(7)	2.917			
11	O(2)–H(6)	2.615	25	H(5)–H(8)	3.493			
12	O(2)–H(7)	2.953	26	H(6)–H(7)	1.780			
13	O(2)–H(8)	3.313	27	H(6)–H(8)	1.769			
14	C(3)–C(4)	1.494	28	H(7)–H(8)	1.765			
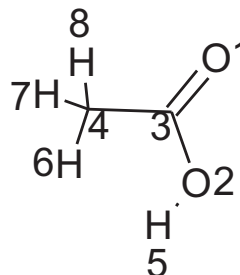
Table 2
Summands and atom contributions of Mor02e descriptor for acetic acid.

	O(1)	O(2)	C(3)	C(4)	H(5)	H(6)	H(7)	H(8)
O(1)	1.178	0.654	1.032	0.367	0.041	−0.044	0.079	0.225
O(2)		1.640	0.951	0.359	1.06	0.24	0.079	−0.064
C(3)			2.086	0.667	0.431	0.355	0.368	0.368
C(4)				1.957	0.232	0.764	0.763	0.762
H(5)					1.016	0.287	0.068	−0.087
H(6)						1.291	0.487	0.492
H(7)							1.169	0.493
H(8)								1.094

2.2.1. Pairwise interatomic distance

Interatomic distance participates both in the numerator and denumerator of radial basis function in Eq. (1). Thus the result is a periodical function with decreasing amplitude (except Mor01 descriptors where the distance has no any effect). To describe the dependence of 3D-MoRSE summand values on interatomic distance we may visualize the radial basis function of Mor02u (Fig. 1). Since distance is always positive, the x-axis should be presented with its positive half only (unlike the plot of radial basis function in Saíz-Urra et al. [15]).

When s is fixed at $1\,\text{Å}^{-1}$, its period is 2π. So radial basis function reaches zero at π, has the largest negative value at 1.5π (this corresponds to interatomic distance $4.712\,\text{Å}$), and then begin to increase, crossing zero at 2π and so on. Since amplitude drops down quickly, the influential summands of Mor02u are represented by atoms with pairwise distance $<2\,\text{Å}$. And the effect of atoms located at more than $5\,\text{Å}$ apart is tiny.
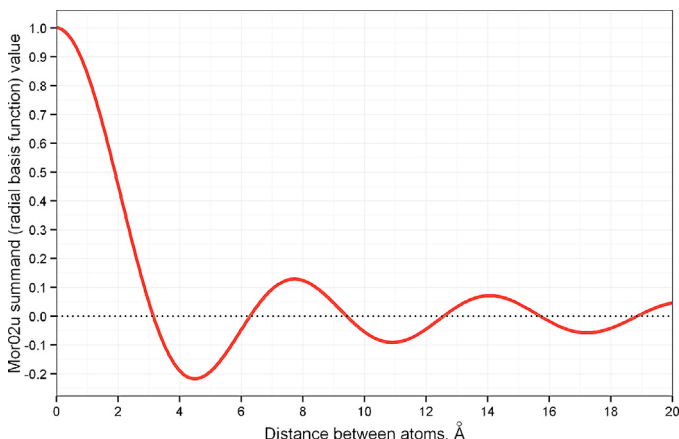

Fig. 1. Mor02u radial basis function.

2.2.2. Weighting scheme

Being a multiplicator at periodic function, weight changes the amplitude of radial basis function. Unweighted descriptors treat each atom equally. Weighting is intended to bring discrimination among atoms, and thus weighted descriptors are sensitive to the presence of specific molecular fragments. The properties of different weighting schemes can be easily deduced from weighting tables (e.g. table of atomic van der Waals volumes or radii, table of atomic polarizabilities, electronegativities, etc.), thus only brief summary is presented here:

- *Atomic mass* – practically eliminates the role of hydrogen atoms, while significantly increases the effect of phosphorus, sulfur and chlorine and greatly increases the effect of heavy atoms like bromine and iodine on the values of 3D-MoRSE descriptors.
- *van der Waals volume* – significantly decreases the effect of hydrogen, diminishes the roles of nitrogen, oxygen and fluorine, while giving more influence to silicium, phosphorus, bromine and iodine.
- *Polarizability* – acts similarly to van der Waals volume. The major difference is greatly enhanced effect of metals, but metalorganic compounds are rare in medicinal chemistry.
- *Electronegativity* – increases mainly the contributions of fluorine, oxygen and chlorine.
- *Atomic partial charge* – unlike the other weighting schemes, this can take both positive and negative values. The weights are not constant across atoms and depend on the surroundings. This weighting may be the most relevant, since it reflects distances between atoms with excessive or deficient electronic density. Unfortunately, it is not implemented in DRAGON, but it is available in ChemoPy [25] and 3dmorse [26].

Also, since each summand represents some atomic pair, the weights of two atoms are multiplying, so the distance between two influential atoms may define the major part of the descriptor (like
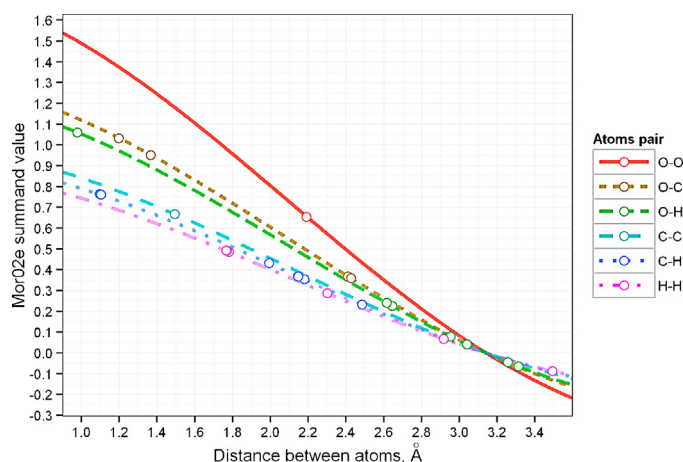
Fig. 2. The dependence of Mor02e summand value on interatomic distance and corresponding values for acetic acid (hollow circles).

Fig. 4. Distribution of Mor02u values for GABA calculated with slightly varying starting atomic coordinates.

two Bromines in a single molecule for weighted by atomic mass 3D-MoRSE descriptors). To understand the effect of weighting, we may look at the plot providing Mor02e summands values on the ordinate axis and corresponding interatomic distance on the abscissa using acetic acid example (Fig. 2).

While electronegativity of Oxygen is only 1.33 times higher the Carbon electronegativity, the effect of O–O pair is 1.77 times higher than that of C–C. However only amplitude is changing, while period remains the same (all curves are crossing at distance that corresponds to π Å).

2.2.3. Scattering parameter

Like pairwise interatomic distance, the scattering parameter occurs both in numerator and denumerator of radial basis function in Eq. (1). While treating interatomic distance as independent variable, the scattering parameter in numerator can be viewed as angle frequency. So the period of radial basis function decreases with scattering parameter growth. In the same time, the amplitude is also decreasing due to denominator augmentation (Fig. 3).

3. Empirical properties of 3D-MoRSE descriptors

In order to test the properties of studied descriptor class we need slightly more complicated model molecule than acetic acid. That molecule should be closer to the structures of modern drugs but also should be as simple as possible to preserve the understanding of the relationship between molecular geometry and descriptors
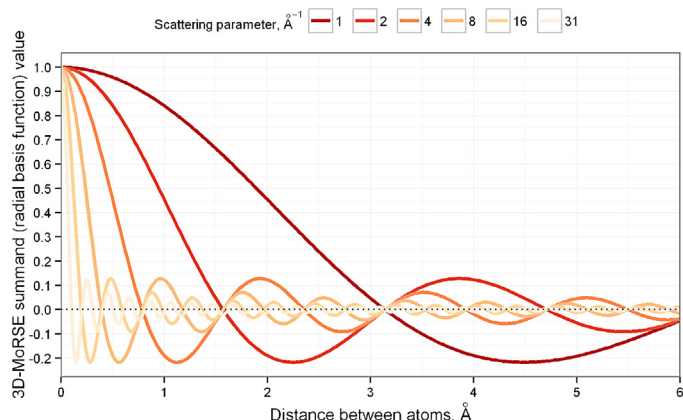

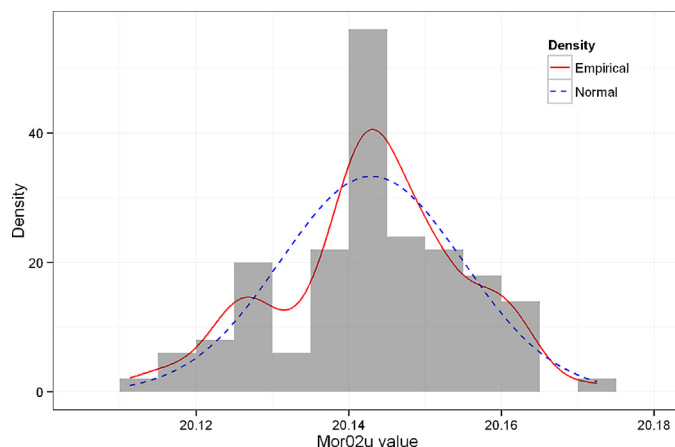**Fig. 3.** 3D-MoRSE radial basis function at different scattering parameters.

values. So we have chosen γ-aminobutyric acid (GABA) as such model molecule. The next text is organized as follows: for each property the theoretical suggestion are formulated firstly, and then the results of empirical testing are given and discussed.

3.1. 3D-MoRSE descriptors are random variables

The variable in probability and statistics is called random if its value is subject to variations due to chance. 3D-MoRSE descriptors depend on interatomic distances (except the cases when scattering parameter equals zero). Interatomic distances are subject to random variations both in nature (molecular vibration) and in computational studies (due to stochastic methods of molecular geometry optimization). Consequently, though 3D-MoRSE descriptors are calculated with exact equation (1), indeed their values are drawn from some probability distribution. According to Central Limit Theorem, we can expect this distribution to be normal, since 3D-MoRSE descriptor is represented with sum of terms, each depending on distance between two atoms, which in turn is affected by multiple factors.

To study this point empirically, we have calculated the values of Mor02u for 100 cases of GABA geometry optimization (with PM7 method) using starting preoptimized atom coordinates altered with normally distributed $N(\mu = 0, \sigma^2 = 0.01)$ random shifts (Fig. 4).

Standard deviation of the sample of Mor02u values is 0.012. But from practical point of view we are interested not in absolute values of 3D-MoRSE descriptor variation, but in its ability to preserve the discrimination among different structures. That is why the comparison of standard deviation of 3D-MoRSE values for single compound with the standard deviation for some decoy set of compounds is highly relevant. Clearly, when within-compound standard deviation (noise) comes close to between-compounds standard deviation (signal, i.e. standard deviation of a sample of descriptor values, each corresponding to single compound from some decoy set), it becomes impossible to identify any compound from a set by its 3D-MoRSE descriptor value. For naturally occurring 20 α-amino acids as a decoy set, standard deviation of the sample of 20 Mor02u values is 3.834. In order to score relative variation of Mor02u, we can use the ratio between two standard deviations $SD_{GABA}/SD_{\alpha\text{-amino acids}}$ (equivalent to noise-to-signal ratio). When the value of ratio equals to 1 that means complete absence of discriminative power of descriptor. In the studied case the ratio $SD_{GABA}/SD_{\alpha\text{-amino acids}}$ is 0.00312 which is small. The visual comparison of between-compounds Mor02u values variation of α-amino acids and within-compound Mor02u values variation of GABA is provided as supplementary information. Additionally, the shape
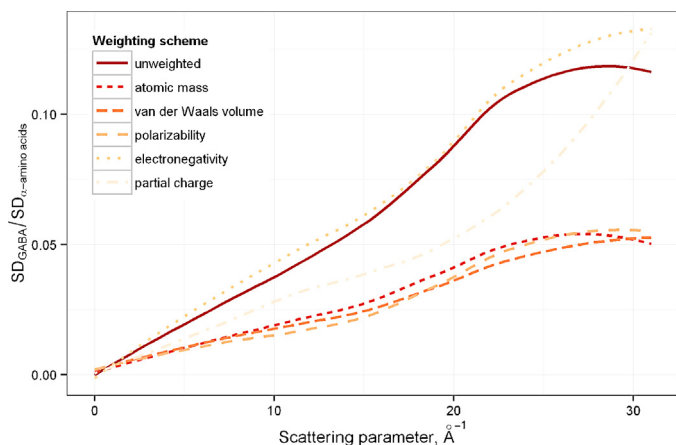
Fig. 5. Relationships between relative variation of 3D-MoRSE descriptors and scattering parameter for different weighting schemes.

of distribution (Fig. 4) is not smooth, showing two spikes which may correspond to two close but distinct energetically favorable geometries.

3.2. The relative variation of 3D-MoRSE descriptors increases together with scattering parameter

As it has been already shown, the increase in scattering parameter leads to shorter period of radial basis function. When period is rather short, small perturbations in geometry lead to significant changes in the value of radial basis function, which in turn become the source of variation. Since the positions of H-atoms are constrained the least, they contribute much to the variation and thus 3D-MoRSE descriptor weighted with schemes where the role of Hydrogen is diminished should exhibit lower variation. The validity of these theoretical suggestions is clearly demonstrated on the GABA-vs-amino acids example by the polynomially smoothed curves of relative variation as a function of scattering parameter (Fig. 5). All weightings show positive correlation between relative variation and scattering parameter with Spearman's ρ values in 0.77–0.91 range. The lowest relative variation is observed for atomic mass, van der Waals volume and polarizability weightings as it has been predicted. The variation of 3D-MoRSE descriptors with high scattering parameter causes significant decrease in discriminatory ability. For example, the range of GABA Mor32u descriptor values obtained with slightly modified starting geometry covers the values of three different amino acids, making the differentiation between GABA and those amino acids impossible. The comparison of within-compound and between-compound variations of 3D-MoRSE values at scattering parameters 1 and 31 Å^{-1} is provided as supplementary information.

3.3. The effect of distant atom pairs on 3D-MoRSE values is small and can be neglected

Since interatomic distance participates in the denominator of radial basis function, the greater distances correspond to smaller summand values. This inference is valuable for the interpretation of 3D-MoRSE descriptors: when looking for atoms and groups responsible for certain 3D-MoRSE value, we can narrow the searching scope to those atomic pairs in which the distance is small. To inspect the impact of different interatomic distances in GABA molecule, we can look at the dynamics of cumulative sum of 3D-MoRSE terms ordered by interatomic distance (Fig. 6). On the given example, the terms that correspond to atoms located more than 5 Å away practically has no influence. Moreover, it can be seen that
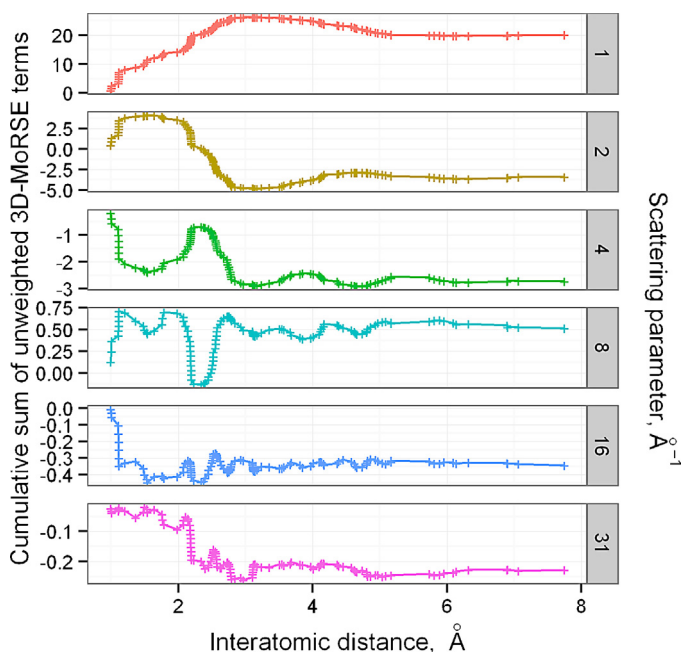

Fig. 6. The role of interatomic distance in unweighted 3D-MoRSE descriptors at different scattering parameters.
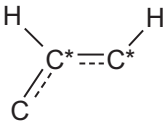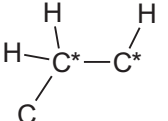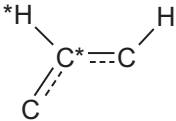
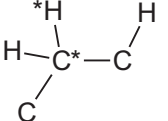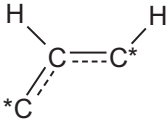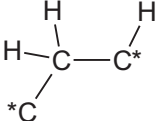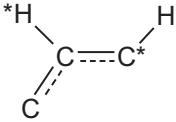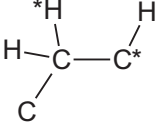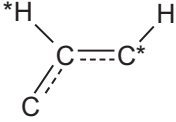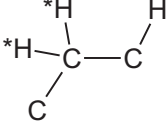all presented descriptors reach their approximate values before 3 Å, and the changes of 3D-MoRSE values after this threshold is small. Now how this fact can be utilized? For example, if increasing values of some 3D-MoRSE descriptor lead to increase in biological activity, then atomic groups that contribute to the 3D-MoRSE descriptor mostly are preferential for activity. So searching for these groups we may check only those pairs that are closer than 3 Å (in most cases such interval corresponds to α and β atoms relative to the each studied). That is a key point for translating differences in 3D-MoRSE values into structural differences. We have no need to wade through all possible atomic binary combinations: checking the closest neighbors is sufficient, since 3D-MoRSE descriptors are defined predominantly by short-distance atom pairs.

4. 3D-MoRSE descriptors in action

4.1. Prediction of 3D-MoRSE descriptors effects in a QSAR model

One of the key points of scientific method tells that a good theory should make accurate predictions. Let us show how the knowledge about 3D-MoRSE descriptors can help to make reasonable prior suggestions about their effect in a QSAR model. The classification of 20 naturally occurring α-amino acids into ring-containing (cyclic) and acyclic structures has been utilized as a case study. This classification task is not trivial, since there are four aromatic α-amino acids (histidine, phenylalanine, tryptophan and tyrosine) and one alicyclic heterocycle (proline). Additionally, only three aromatic α-amino acids contain phenyl ring, while the fourth (histidine) has imidazole cycle in its structure. Our goal is to predict which 3D-MoRSE descriptors (i.e. which weightings and scattering parameter values) will discriminate ring-containing and acyclic α-amino acids better. To make a prediction, the first step will be to answer: what are the atomic pairs that can differentiate ring-containing and acyclic structures in the given context? First of all, these are two carbons linked with an aromatic bond, since they are present in aromatic α-amino acids while in acyclic acids carbons are linked with single bonds. However, not the bond nature is primary for 3D-MoRSE, but bond length. The lengths of aromatic bond (about 1.40 Å) and C(sp^3)–C(sp^3) bond (about 1.54 Å) differ, and

Table 3
Closest atomic pairs that could differentiate cyclic α-amino acids from acyclic ones.

Atomic pair in cyclic (aromatic) α-amino acid	Distance, Å	The typical counterpart in acyclic (aliphatic-chain) α-amino acid	Distance, Å	The difference has a tangible effect on 3D-MoRSE descriptors values
	1.40		1.54	Yes
	1.09		1.11	No
	2.40		2.50	Yes
	2.15		2.18	No
	2.15		1.76	Yes

that contributes to the possibility of distinguishing between the two amino acids classes. The second pair could be $C(sp^2)$–H with bond length 1.09 Å, but it is very close to $C(sp^3)$–H, which is about 1.11 Å. The difference in 0.02 Å is too small to has some effect. Even the most sensitive to small changes 3D-MoRSE descriptors with the highest scattering parameter 31 Å^{-1}, have period $2\pi/31 = 0.20$ Å. So to obtain effect of amplitude size we need at least 0.05 Å difference. Going to further neighbors, we may select few more atomic pairs that are relevant for classification purposes (Table 3).

Now, to predict the values of scattering parameter that provide the best discrimination, we can sum up the values of radial basis function for the three selected atomic pairs in cyclic amino acids and subtract the values of radial basis function for the corresponding three atomic pairs in acyclic amino acids:

$$\frac{\sin 1.40s}{1.40s} + \frac{\sin 2.40s}{2.40s} + \frac{\sin 2.15s}{2.15s}$$
$$- \frac{\sin 1.54s}{1.54s} - \frac{\sin 2.50s}{2.50s} - \frac{\sin 1.76s}{1.76s} \tag{9}$$

Carrying out this procedure for each integer value of scattering parameter in the range 1–31 (Fig. 7), we may assess the relative discriminative power of 3D-MoRSE descriptors (clearly, the scattering parameter values at which radial basis functions for aromatic acid atomic pairs fall into one phase and radial basis functions for aliphatic acid atomic pairs fall into corresponding counterphase should provide excellent discriminatory ability).

Looking into the obtained results, we may predict that 3D-MoRSE descriptors with scattering parameters 5, 6 and 8 Å^{-1} are the most relevant to discriminate cyclic α-amino acids from acyclic ones. The lowest scattering parameter 0 Å^{-1} was not taken into consideration, since corresponding 3D-MoRSE descriptor Mor01u

depends only on total number of atoms (see discussion above, in the theoretical part of article). In turn, there is no significant difference in total number of atoms between cyclic and acyclic α-amino acids. However, ring-containing acids have more carbons, while acyclic have more hydrogens. Thus, when applying weighting schemes which provide carbon atoms with much more influence than hydrogen atoms (atomic mass, polarizability, van der Waals volume), we may expect the rise of discriminative ability. Speaking further about impact of different weightings on relative discriminative ability of 3D-MoRSE descriptors, we may note, that those schemes that give larger role for carbon atoms should lead to better performance. The weighting by atomic masses greatly diminishes
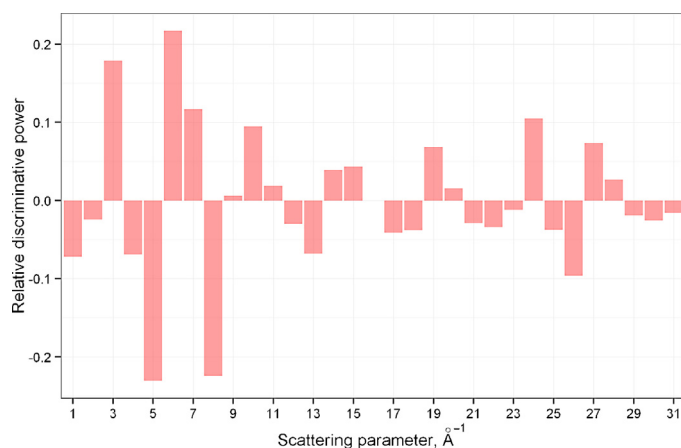

Fig. 7. The relative discriminative power of unweighted 3D-MoRSE descriptors predicted for the case classification study.
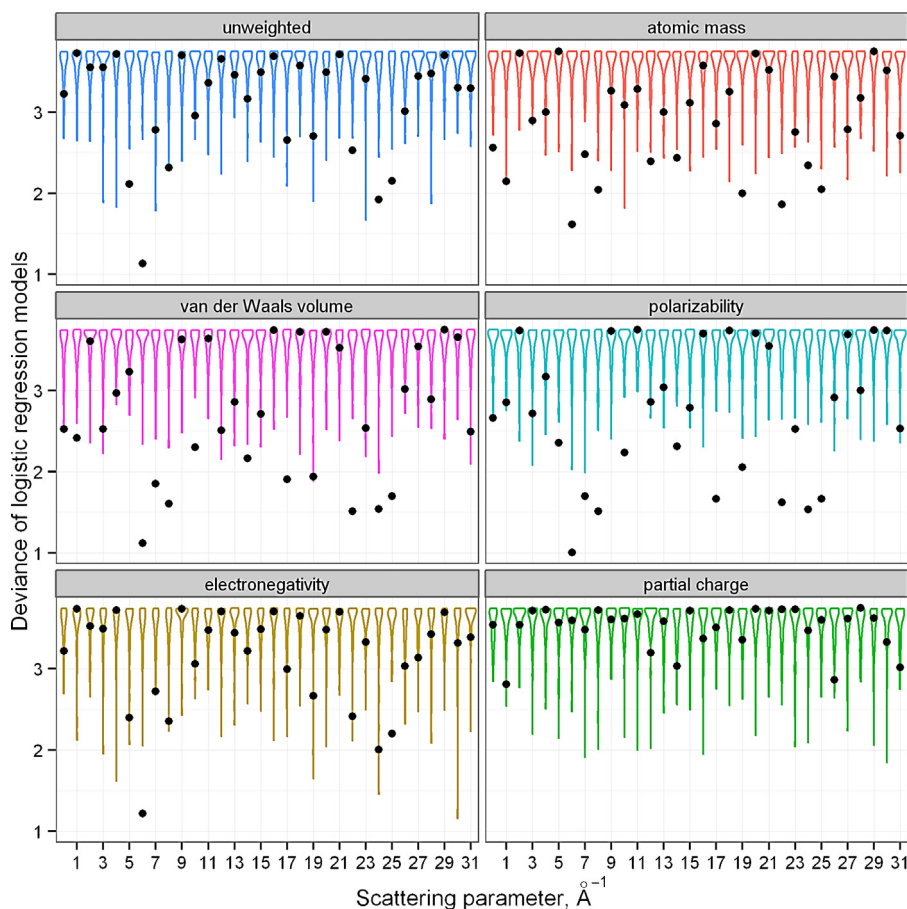
Fig. 8. Deviance of logistic regression models of the α-amino acids classification study developed with different 3D-MoRSE descriptors. The results of response permutation test are illustrated with violin plots.

influence of hydrogens, but increase the influence of heteroatoms, the last unlikely contribute to the classification task. In the other hand, weightings by van der Waals volumes and by polarizabilities decrease roles of nitrogen and oxygen atoms as well as hydrogen. Since the major difference between cyclic and acyclic α-amino acids is presented with carbon atoms, these two weightings are expected to be optimal for the given case study.

In order to test the validity of predictions made, the classification models each using single 3D-MoRSE descriptor have been fitted with logistic regression. Goodness-of-fit has been evaluated as model deviance (lower value of deviance corresponds to better accuracy of model). To see how likely each deviance value can be reached by chance, response permutation test has been carried out for each model with the number of iterations equal to 100.

The results showed good agreement with the prediction (Fig. 8). The best classification performance for unweighted 3D-MoRSE descriptors is observed for Mor07u, that corresponds to scattering parameter $s = 6 \text{ Å}^{-1}$. The values of $s = 5$ and 8 Å^{-1} also lead to reasonable and statistically significant models. As it has been predicted, when $s = 0 \text{ Å}^{-1}$ unweighted descriptor Mor01u has no discriminative ability, but Mor01m, Mor01v and Mor01p do have. The suggestion that van der Waals volume and polarizability weighting schemes are favorable is confirmed as well. For sure, there are some other scattering parameter values with discriminative ability, not covered by prediction (especially for van der Waals volume and polarizability weightings). That is because our analysis was rather shallow. It should be emphasized, that all predictions were based entirely on the theory behind 3D-MoRSE and analysis of distances in just few atomic pairs. This case study also illustrates the claim

that analysis of the closest atoms (and here we used only the atoms connected directly with covalent bond or through the third atom) is sufficient to obtain the overall picture in 3D-MoRSE framework.

4.2. Interpretation of 3D-MoRSE descriptors in a QSAR model

In the presented above case study the structural features that distinguish two classes were known and the task was to find the most favorable 3D-MoRSE descriptors without using QSAR models. In the typical QSAR analysis the problem is reverse: the developed QSAR model contains the most favorable 3D-MoRSE descriptors and we should interpretate them. That means we should identify those structural features that are responsible for the desired class (or level) of biological activity.

The road of 3D-MoRSE descriptors interpretation can be passed with the next steps:

1. Estimate the favorable values of 3D-MoRSE descriptors (whether these are high or low values, or the best activity is observed in some optimal range of 3D-MoRSE values).
2. Plot radial basis function with corresponding to studied 3D-MoRSE descriptor scattering parameter and consider weights for different atomic pairs if the descriptor is weighted.
3. Identify on the plot ranges of interatomic distance that lead to desired 3D-MoRSE values.
4. Search in your active compounds the pairs of atoms located at previously found distance. Check whether these pairs are absent in non-active compounds. Those atomic pairs that are present in active compounds and absent in non-active and
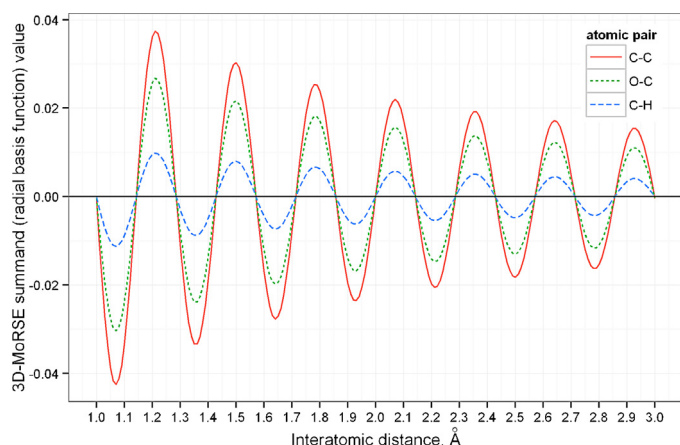
Fig. 9. Radial basis functions of Mor23v descriptor corresponding to different atomic pairs.
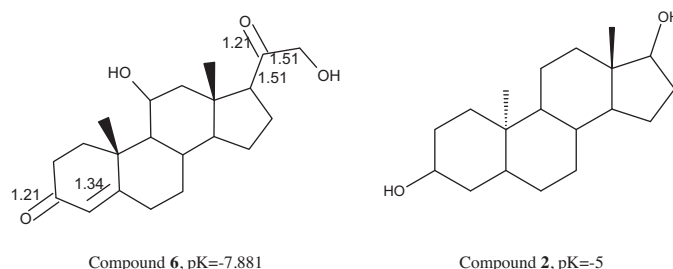

Fig. 10. Representatives of the most and the least active steroid compounds with the most relevant for Mor23v descriptor interatomic distances. *The compounds are depicted in 2D representation just for easy understanding, while the distances are measured in 3D for sure.

having interatomic distance in the favorable ranges promote biological activity.

Cramer's steroid data has been used to produce the case QSAR study with 3D-MoRSE interpretation [28]. The data has been downloaded from web resource [29] and the molecular geometries were left as is. The dependent variable was the negative logarithm of corticosteroid-binding globulin (CBG) binding constant. Iterating through the set of 3D-MoRSE descriptors treated as independent variables, simple monoparametric linear regression models were fitted. The whole dataset was employed in model deriving and none validation was carried out since the primary aim is to show 3D-MoRSE descriptor interpretation process, while just simulating the QSAR study to be close to a real one. In such a way, the best monoparametric model (10) has been obtained:

$$pK = -5.98 - 4.75 \times Mor23v \tag{10}$$

The determination coefficient of this model is $R^2 = 0.7385$, and this value is rather high considering the level of parsimony. The Mor23v descriptor is the 3D-MoRSE descriptor calculated with scattering parameter $s = 22\,\text{Å}^{-1}$ and weighted by atomic van der Waals volume. Now to find how this descriptor is related to molecular structure, let us pass the required steps.

Step 1. Since high activity in steroid data is expressed with negative numbers (i.e. the less pK values mean the better binding) and β coefficient at the Mor23v descriptors is negative too, the higher Mor23v values correspond to higher activity levels.

Step 2. In this stage the plots of corresponding radial basis functions (11) should be produced:

$$f(r) = A_1 A_2 \frac{\sin 22r}{22r} \tag{11}$$

where A_1 and A_2 are corresponding carbon-scaled atomic van der Waals volumes, r is the interatomic distance.

The steroid dataset include molecules with common atoms C, O, and H (additionally, the compound **31** contains flourine). The carbon-scaled van der Waals volumes for these atoms are 1, 0.715 and 0.263 respectively. The low weight for hydrogen atom makes the contribution of hydrogen-containing atomic pairs negligible (and we would plot radial basis function for the most influential hydrogen-containing atomic pair C–H just for visual support of this claim). Since it is reasonable to search the favorable atomic pairs restricting maximum interatomic distance to 3 Å, and there are no C–C or C–O bonds shorter than 1 Å, we specify the proper limits for the x-axis. The result is presented in Fig. 9.

Step 3. Looking at Fig. 9, the most favorable atomic pairs are located at the distances about 1.21, 1.50, 1.78, 2.07, 2.36, 2.64

and 2.93 Å, while the most detrimental pairs are located at 1.07, 1.36, 1.64, 1.93, 2.21, 2.50 and 2.78 Å. Not only these exact interatomic distances should be searched in studied molecules, but the deviations up to 0.02 Å are acceptable also (for smaller scattering parameter values the acceptable region is larger). The favorable and detrimental distances also can be obtained more accurately with direct calculations. Indeed, the coordinates of extremums for $\sin sr$ function (other components do not have impact on the period) are

$$\frac{((\pi/2) + k\pi)}{s} \tag{12}$$

where s is scattering parameter and k takes all non-negative integer values.

Step 4. To search the prespecified distances in studied molecules, we can safely take into consideration only few the most and the least active compounds. Here only two compounds are utilized: compound **6** and compound **2** that are one of the most and the least active compounds respectively (Fig. 10). First of all, the single C–C bonds lengths in cyclopentanoperhydrophenanthrene fragment can be omitted since that part is common for all steroid molecules (ignoring the influence of substituents on the steroid core geometry). In the compound **6** C(3)=O, C(20)=O, C(17)–C(20) and C(20)–C(22) bonds have favorable lengths, while C(4)=C(5) bond has unfavorable length (Fig. 10). Considering the different amplitude for C–C and O–C atomic pairs, the positive contribution of C–C is higher, despite the double bond in C=O is shorter. C(10)–C(19) and C(13)–C(18) are located at 1.53 Å in both molecules. Such distance is not so far from positive peak at 1.50 Å, so the methyl groups in 10th and 13th positions of steroid core have small positive effect on the activity predicted. Single C–O bond has length 1.43 Å, which, according to Fig. 9, has zero impact. Compound **2** do not show specific favorable or unfavorable close pairs of atoms. Summarizing, the interpretation of Mor23v descriptor in the current case study can be formulated as follows: the presence of carbonyl groups and carbon chain in the 17th position have positive impact on the CBG binding affinity, and the presence of double C=C bonds is detrimental.

What about improving the monoparametric model with one more 3D-MoRSE descriptor? Making an exhaustive search, the best two-parametric QSAR model ($R^2 = 0.793$, the improvement is not so high) is:

$$pK = -7.61 - 5.06 \times Mor23v - 1.36 \times Mor13m \tag{13}$$

Let us find what is the sense of Mor13m in this QSAR model. Using the known properties of 3D-MoRSE descriptors and making some model diagnostics it is possible to interpret the descriptor in a more elegant way than four-steps distance analysis. To know which atomic pairs have the greatest influence on Mor13m, the identification of those compounds that have decreased their prediction error significantly when moving from monoparametric to two-parametric model may be used due to the causal chain (Fig. 11).
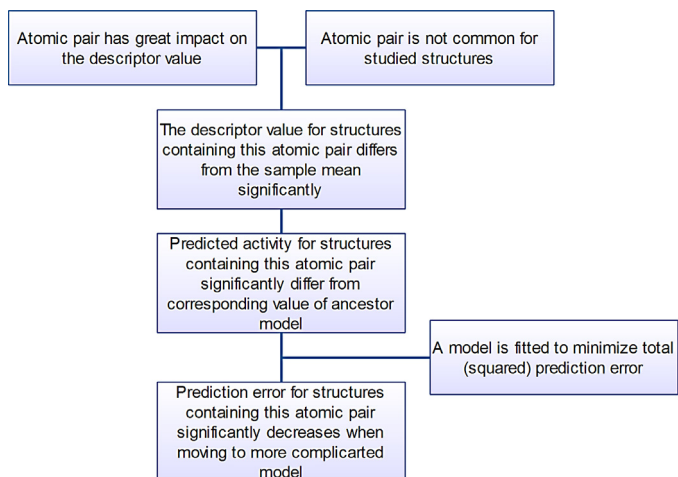
Fig. 11. The relationship between the presence of influential atomic pairs in molecular structure and the decrease of prediction error for this structure.

Fig. 13. Radial basis functions of Mor13m descriptor corresponding to different atomic pairs.

The plot of squared residual dynamics (Fig. 12) shows that compound **31** has the greatest decrease of prediction error. The term that corresponds to the compound **31** holds 68% of total decrease in the sum of squared errors (*SSE*). Taking into account, that compound **31** contains fluorine, which is the heaviest atom in the studied dataset, and the chosen second descriptor is weighted accordingly, i.e. with atomic masses, it may be argued that one of Mor13m roles is to penalize the compound **31** for the fluorine atom. Indeed, the C–F distance (1.40 Å) is close to one of the local minimums on the corresponding radial basis function plot (Fig. 13). Additionally, single C–O bonds (each having the length of 1.43 Å) created by three attached hydroxyl groups also fall into the local minimum, thus penalizing the predicted activity of **31** further (Fig. 13).

The interpretation of Mor13m descriptor in this case study can be formulated as follows: the presence of fluorine atoms and hydroxyl groups in steroid structure is unfavorable and decreases CBG binding affinity. However, we would like to note, that this inference is caused mostly by the single compound **31**, and thus much credits should not be given to it. Apparently, when a model begin to optimize the prediction error just for single compounds, that is stopping criterion and further complication of such model is inappropriate.
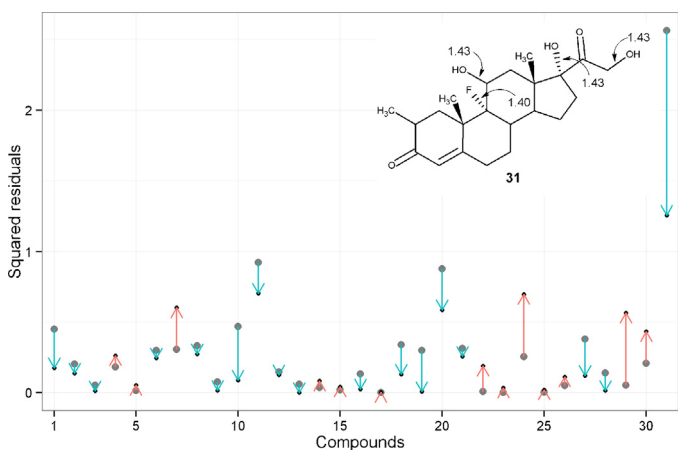
5. Methods

All structures used in the theoretical analysis and case studies were downloaded in three-dimensional representation and further optimized in MOPAC2012 with PM7 method [27] (except the steroids, which were used as-is for the consistency purposes). To calculate 3D-MoRSE descriptors from MOPAC2012 output files the short program "3dmorse" has been written in C++. The possibility to return 3D-MoRSE summands in a separate file is implemented as a specific feature of the program. The source code and the executable file compiled for Windows environment can be freely downloaded from the web [26]. All other calculations has been carried out in R 3.0.1 environment [30] with additional package for graphics preparation "ggplot2" [31].

6. Conclusions

Lifting the veil of 3D-MoRSE descriptors, we have showed both their strength and weakness. The range of scattering parameter values and variety of weighting schemes provide 3D-MoRSE with significant flexibility, ensuring good discriminatory power even for similar structures. The values of the descriptors are calculated with rather simple equation, but when using differing starting geometries as optimization input the values were unstable and exhibited variation. This variation increased with scattering parameter and also was higher for electronegativity weighted and unweighted descriptors. Though each 3D-MoRSE descriptor incorporates the information about the whole molecule structure, it has been shown that its final value is derived mostly from short-distance (up to 3 Å) atomic pairs. Thus, if a QSAR study covers structurally similar set of compounds, then the role of 3D-MoRSE descriptor in a model can be interpreted using just several pairs of neighbor atoms. As a brief synopsis, to explain the effect of 3D-MoRSE descriptor its radial basis function should be plotted and then atoms that are located at distances that coincide with maxima or minima of the function should be identified. In turn, the presence or absence of the found atomic pairs in a molecule determine its biological activity value (at least this is suggested by the interpreted descriptor). Realizing the mathematical concept behind 3D-descriptors and knowing their properties it is easy not only to interpret, but also to predict the importance of 3D-MoRSE descriptors in a QSAR study.

Finishing the paper, we would like to compare any descriptor class with a human language. We, the medicinal chemists, speak about compounds in terms of their structure: atoms, bonds, groups, templates, scaffolds and substituents. But machine learning methods do not understand such language. Thus we have to project

Fig. 12. The dynamics of squared residuals when moving from monoparametric to two-parametric QSAR model.

chemical structures into a space of numbers following the pre-defined rules and equations. Each molecular descriptor is a word, and QSAR models choose the best words to explain activity phenomenon. Thus, the ability to translate these words into the familiar language of structural entities is crucial for successful application of QSAR method. And, like there are no better or worse languages, there are no better or worse descriptor classes. The languages of Alaska Natives can perfectly describe snow, but are poor to describe, for example, technology. In such a way some descriptors are suitable in one case and not relevant in another. 3D-MoRSE descriptors cannot describe complex atomic groups or regions with high or low electron density or some quantum-chemical properties etc., but result in a good model performance when activity variation coincides with variation in interatomic distances due to changes of bonds order and the introduction of new atoms.

Conflict of interest

The authors declare no competing financial interest.

Acknowledgment

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2014.10.006.

References

[1] D.T. Stanton, On the physical interpretation of QSAR models, J. Chem. Inf. Comput. Sci. 43 (2003) 1423–1433.
[2] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000, pp. 688.
[3] J.C. Dearden, M.T. Cronin, K.L. Kaiser, How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR), SAR QSAR Environ. Res. 20 (2009) 241–266.
[4] Organization for Economic Co-operation and Development, Guidance document on the validation of (quantitative) structure–activity relationship ((Q)SAR) models, in: OECD Series on Testing and Assessment 69, OECD Document ENV/JM/MONO, 2007, pp. 154.
[5] R. Todeschini, P. Gramatica, SD-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors, Quant. Struct.-Act. Relat. 16 (1997) 113–119.
[6] M. Randić, J. Zupan, On interpretation of well-known topological indices, J. Chem. Inf. Comput. Sci. 41 (2001) 550–560.
[7] Q.-N. Hu, Y.-Z. Liang, H. Yin, X.-L. Peng, K.-T. Fang, Structural interpretation of the topological index. 2. The molecular connectivity index, the kappa index, and the atom-type E-state index, J. Chem. Inf. Comput. Sci. 44 (2004) 1193–1201.
[8] O. Devinyak, M. Slivka, M. Slivka, V. Vais, V. Lendel, Quantitative structure–activity relationship study and directed synthesis of thieno[2,3-d]pyrimidine-2,4-diones as monocarboxylate transporter 1 inhibitors, Med. Chem. Res. 21 (2012) 2263–2272.
[9] O. Devinyak, B. Zimenkovsky, R. Lesyk, Biologically active 4-thiazolidinones: a review of QSAR studies and QSAR modeling of antitumor activity, Curr. Top. Med. Chem. 12 (2012) 2763–2784.
[10] J.H. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.
[11] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3d space, J. Chem. Inf. Comput. Sci. 36 (1996) 1030–1037.
[12] H. Kušić, B. Rasulev, D. Leszczynska, J. Leszczynski, N. Koprivanac, Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study, Chemosphere 75 (2009) 1128–1134.
[13] J. Schuur, J. Gasteiger, Infrared spectra simulation of substituted benzene derivatives on the basis of a 3d structure representation, Anal. Chem. 69 (1997) 2398–2405.
[14] L. Saíz-Urra, Y. Pérez-Castillo, M. Pérez González, R. Molina Ruiz, M. Cordeiro, J.E. Rodríguez-Borges, X. García-Mera, Theoretical prediction of antiproliferative activity against murine leukemia tumor cell line (L1210). 3D MoRSE descriptor and its application in computational chemistry, QSAR Comb. Sci. 28 (2009) 98–110.
[15] L. Saíz-Urra, M.P. González, M. Teijeira, QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases i and ii: 3d-morse descriptors and statistical considerations about variable selection, Biorg. Med. Chem. 14 (2006) 7347–7358.
[16] P.R. Duchowicz, M.G. Vitale, E.A. Castro, M. Fernández, J. Caballero, QSAR analysis for heterocyclic antifungals, Biorg. Med. Chem. 15 (2007) 2680–2689.
[17] Z. Cheng, Y. Zhang, W. Fu, QSAR study of carboxylic acid derivatives as HIV-1 integrase inhibitors, Eur. J. Med. Chem. 45 (2010) 3970–3980.
[18] M. Arab Chamjangali, M. Beglari, G. Bagherian, Prediction of cytotoxicity data (CC50) of anti-HIV 5-pheny-l-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg–Marquardt algorithm, J. Mol. Graph. Model. 26 (2007) 360–367.
[19] J. Caballero, M. Fernández, Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks, J. Mol. Model. 12 (2006) 168–181.
[20] J. Caballero, A. Vergara-Jaque, M. Fernández, D. Coll, Docking and quantitative structure–activity relationship studies for sulfonyl hydrazides as inhibitors of cytosolic human branched-chain amino acid aminotransferase, Mol. Divers. 13 (2009) 493–500.
[21] M. Sun, J. Chen, J. Cai, M. Cao, S. Yin, M. Ji, Simultaneously optimized support vector regression combined with genetic algorithm for QSAR analysis of KDR/VEGFR-2 inhibitors, Chem. Biol. Drug. Des. 75 (2010) 494–505.
[22] M. Di Tullio, C. Maccallini, A. Ammazzalorso, L. Giampietro, R. Amoroso, B. De Filippis, M. Fantacuzzi, P. Wiczling, R. Kaliszan, QSAR, QSPR and QSRR in terms of 3D-MoRSE descriptors for in silico screening of clofibric acid analogues, Mol. Inf. 31 (2012) 453–458.
[23] I. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk, V. Prokopenko, Virtual computational chemistry laboratory – design and description, J. Comput. Aided Mol. Des. 19 (2005) 453–463.
[24] Dragon Program. http://www.talete.mi.it (accessed July 2014).
[25] D.-S. Cao, Q.-S. Xu, Q.-N. Hu, Y.-Z. Liang, ChemoPy: freely available python package for computational biology and chemoinformatics, Bioinformatics 29 (8) (2013) 1092–1094.
[26] 3dmorse Program. http://github.com/devinyak/3dmorse (accessed July 2014).
[27] J.J. Stewart, Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, J. Mol. Model. 19 (2013) 1–32.
[28] E.A. Coats, The COMFA steroids as a benchmark dataset for development of 3D QSAR methods, in: 3D QSAR in Drug Design, Springer, Dordrecht, 1998, pp. 199–213.
[29] 31 Steroids Binding to the Corticosteroid Binding Globulin (CBG) Receptor. http://www2.chemie.uni-erlangen.de/services/steroids/ (accessed July 2014).
[30] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012.
[31] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009, pp. 213.