J Comput Aided Mol Des (2013) 27:427–442
DOI 10.1007/s10822-013-9656-4

The continuous molecular fields approach to building 3D-QSAR models

Igor I. Baskin · Nelly I. Zhokhova

Received: 31 January 2013 / Accepted: 22 May 2013 / Published online: 30 May 2013

Abstract The continuous molecular fields (CMF) approach is based on the application of continuous functions for the description of molecular fields instead of finite sets of molecular descriptors (such as interaction energies computed at grid nodes) commonly used for this purpose. These functions can be encapsulated into kernels and combined with kernel-based machine learning algorithms to provide a variety of novel methods for building classification and regression structure–activity models, visualizing chemical datasets and conducting virtual screening. In this article, the CMF approach is applied to building 3D-QSAR models for 8 datasets through the use of five types of molecular fields (the electrostatic, steric, hydrophobic, hydrogen-bond acceptor and donor ones), the linear convolution molecular kernel with the contribution of each atom approximated with a single isotropic Gaussian function, and the kernel ridge regression data analysis technique. It is shown that the CMF approach even in this simplest form provides either comparable or enhanced predictive performance in comparison with state-of-the-art 3D-QSAR methods.

Keywords Continuous molecular fields · 3D-QSAR · Kernel ridge regression · Functional data analysis · R program

Introduction

Various 3D-QSAR approaches, in which the spatial structure of molecules is taken into account explicitly, play an important role in modern studies in the field of chemoinformatics and computer-assisted drug design [1–4]. Most of them are based on the use of molecular fields reflecting different types of intermolecular interactions in which molecules under study are involved. The most widely used 3D-QSAR method is comparative molecular field analysis (CoMFA) [5], in which electrostatic interactions are approximated by means of the Coulomb law with point partial charges computed for each atom, whereas steric interactions are expressed using the Lennard-Jones potentials with standard force field parameters. Some other types of molecular fields, such as the hydrophobic [6], hydrogen bond donor and acceptor fields [7], molecular orbital fields [8], E-state fields [9], fields of atom-based indicator variables [10], are also used in the framework of the CoMFA method. In another popular 3D-QSAR method, the comparative molecular similarity indices analysis (CoMSIA) approach [11], the same types of molecular fields are approximated using the Gaussian radial basis functions. In the GRID method [12] molecular fields are computed as interaction energies of certain probe atoms and group of atoms with molecules under study.

In all these approaches, molecular fields are computed at the nodes of an imaginary grid surrounding the set of aligned molecules. The advantage of applying such lattices is the ability to use the values of molecular fields calculated at grid nodes as vectors of descriptors, which can further be processed using the partial least squares (PLS) [13] regression method in order to build 3D-QSAR models. Another appealing feature of this approach is the ability to visualize the regression coefficients of such models by means of easily interpretable isosurfaces (usually colored according to the sign of coefficients and the type of molecular field) surrounding the molecules. Meanwhile, in order to build 3D-QSAR models, it is necessary to: (1)

I. I. Baskin (✉) · N. I. Zhokhova
Faculty of Physics, M. V. Lomonosov Moscow State University, Leninskie Gory, 119991 Moscow, Russia
e-mail: igbaskin@gmail.com

choose biologically active conformation for each molecule; (2) align in space the training set of molecules; (3) build a lattice around such set of molecules; (4) choose molecular fields and compute their values at grid points; (5) build predictive statistical models. The problems associated with each of these stages are well known and present a challenge for the current stage of the development of the 3D-QSAR methodology [1–3].

Consider the problems caused by the necessity to define an imaginary grid of points. It is known that the quality of 3D-QSAR models sharply depends on the spatial orientation and the extent of such grid, as well as on the step size (i.e. the distance between the closest points) in it [3]. Another problem caused by the use of grids is very high dimensionality of the regression task caused by the big number of grid points, which precludes the use of many efficient statistical methods. Unfortunately, it is impossible to decrease the number of grid points by increasing the step size or decreasing the extent of the grid, because this leads to the loss of important information. Recently suggested methodology to compress such descriptor vectors by means of wavelets presents only a partial solution to this problem [14], because it is still based on the use of a lattice and therefore does not eliminate the dependence on lattice parameters and the loss of information in consequence of grid approximation. Another approach, the volume learning algorithm, uses combination of supervised back-propagation neural network and unsupervised self-organizing map of Kohonen in order to cluster field variables computed at grid points and to use a small number of the most relevant parameters in order to build 3D-QSAR models [15]. In an alternative approach, called kernel molecular dynamics [16], interaction of probes with molecules is described by means of probability density functions resulting from molecular dynamics simulations. In the latter case, the kernels comparing such probability densities are applied for building 3D-QSAR models. This approach, however, requires rather sophisticated practical implementation, highly depends on the accuracy of the force fields used for simulations, and involves intensive computations. Another partial solution is to use molecular field extrema in order to generate finite sets of three-dimensional molecular field descriptors [17].

So, calculating the values of molecular fields at a finite discrete set of points, followed by application of the PLS regression procedure, constitutes the basis of current 3D-QSAR methodology. Nonetheless, there exists an alternative to this paradigm. Instead of computing descriptor values at a discrete set of points, one can apply statistical analysis to molecular fields represented as continuous functions of spatial coordinates (continuous molecular fields). R. Carbó-Dorca and co-authors pioneered in this area by introducing molecular quantum similarity measures, which compare electron density functions of molecules [18]. Matrices composed of such indices computed for all pairs of compounds in datasets can further be processed using the PLS regression procedure to provide 3D-QSAR models [19, 20]. Later, this approach has been extended to the use of so-called conceptual DFT molecular fields [21]. Meanwhile, to our opinion, the emphasis on exclusively quantum similarity measures and the use of only traditional multivariate data analysis approaches (such as the PLS regression) did not allow to fully disclose the potential inherent in the use of continuous molecular fields.

The aim of this paper is to present conceptually novel approach consisting in encapsulating continuous molecular fields into specially constructed kernels. It is based on the application of continuous functions for the description of molecular fields instead of finite sets of molecular descriptors (such as interaction energies computed at grid nodes) commonly used for this purpose. *We refer to this methodology as continuous molecular fields* (*CMF*) *approach.* The feasibility of using molecular-field kernels in combination with the support vector regression (SVR) machine learning method to build 3D-QSAR models has been demonstrated by us earlier (see preliminary short communication [22]). We have also demonstrated that actually the same kernels can be used in conjunction with the one-class Support Vector Machines (1-SVM) for conducting virtual screening based on similarity of molecular fields [23, 24]. Moreover, by combining different types of molecular fields and methods of their approximation, different types of kernels with different types of kernel-based machine learning methods, it is possible not only to present lots of existing methods in chemoinformatics and medicinal chemistry as particular cases within a single methodology, but also to develop new approaches aimed at solving new problems.

This is the first full-fledged paper in which the CMF approach is described in detail by the example of a simple method of constructing 3D-QSAR models. This particular application of the CMF approach is based on the use of a linear combination of several molecular-field kernels in conjunction with the kernel ridged regression (KRR) data analysis algorithm. Since, for the sake of simplicity and computational efficiency, each molecular field is approximated with atom-centered isotropic Gaussian functions, one function for each atom and each type of molecular field, there is a similarity between some computational formulas in this particular implementation of the CMF approach and the corresponding expressions in the CoMSIA method. In this paper, the CMF approach is applied to building 3D-QSAR models for 8 datasets through the use of five types of molecular fields (the electrostatic, steric, hydrophobic, hydrogen-bond acceptor and donor ones). We assess the predictive ability of models built with CMF

using the external five-fold cross-validation procedure. We also compare them with models built using CoMFA and CoMSIA on the same data sets with the same alignments of molecules with the same choice of their conformations [25]. Addressing the problems of spatial alignment of molecules and their conformational flexibility is beyond the scope of this publication.

In this article, we first consider the construction of kernels for continuous molecular fields, and then we discuss the notion of the fields of regression coefficients and the ways to find them. Following this, we consider parameterization of molecular fields, the use of kernel regression methods to build statistical model, statistical parameters to assess predictive performance of models and the issue of the "model selection bias". After that we assess and compare the predictive ability of models built with CMF and models built using CoMFA and CoMSIA. In the Conclusions section, we describe the prospects of further development of the CMF approach.

The CMF approach to constructing 3D-QSAR models

Procedure of kernel calculation

The central element of the CMF approach is the procedure for calculating the kernels of molecular fields. Joint kernel $K(M_i, M_j)$, which describes the similarity between all types of molecular fields of molecules M_i and M_j, is calculated as a linear combination of kernels corresponding to each of the types of molecular fields:

$$K(M_i, M_j) = \sum_f h_f K_f(M_i, M_j), \tag{1}$$

where h_f is mixing coefficient for the fth type of molecular field; $K_f(M_i, M_j)$ is kernel describing similarity between the molecular field of the fth type of the ith and jth molecules. The function $K(M_i, M_j)$ is a valid kernel because linear combination of kernels is the kernel.

In the CMF approach, the kernel $K_f(M_i, M_j)$ for the molecular field of the fth type is calculated by summing up the kernels for all pairs of atoms of the ith and jth molecules:

$$K_f(M_i, M_j) = \sum_{l,m} k_f(A_{il}, A_{jm}), \tag{2}$$

where $k_f(A_{il}, A_{jm})$ is the kernel describing the similarity between the molecular field of the fth type of the lth atom in the ith molecule and mth atom in the jth molecule. According to Eq. (2), $K_f(M_i, M_j)$ may be considered as the convolution kernel corresponding to the decomposition of molecules into atoms. The value of the kernel $k_f(A_{il}, A_{jm})$ can be calculated by integrating the product of molecular fields of the fth type for all pairs of atoms over the whole physical space $\Re^3$:

$$k_f(A_{il}, A_{jm}) = \int_{\Re^3} \rho_{fil}(\mathbf{r})\rho_{fjm}(\mathbf{r})d^3\mathbf{r}, \tag{3}$$

where $\rho_{fil}(\mathbf{r})$is the value of the molecular field of the fth type induced by the lth atom of the ith molecule at the point $\mathbf{r}$ of the physical space; $\rho_{fjm}(\mathbf{r})$ is the same for the mth atom of the jth molecule. To simplify the integration, one can expand any molecular field as a linear combination of Gaussian basis functions. We have found empirically that in many cases it is sufficient to use a single Gaussian function to represent any kind of molecular fields produced by a single atom, exactly like in the CoMSIA method:

$$\rho_{fil}(\mathbf{r}) = w_{fil} \exp\left(-\frac{1}{2}\alpha_f \|\mathbf{r} - \mathbf{r}_{il}\|^2\right), \tag{4}$$

where $\mathbf{r}_{il}$ is the location of the lth atom of the ith molecule in the physical space; α_f is the attenuation factor for the molecular field of the fth type; w_{fil} is the weight of the contribution of lth atom of the ith molecule to the molecular field of the fth type. For example, for the electrostatic field the w_{fil} is the partial charge on the lth atom of the ith molecule. Obviously, different sets of values for w_f define different parametrizations for CMF. Due to this approximation, the foregoing integral can be calculated analytically:

$$\begin{aligned} k_f(A_{il}, A_{jm}) &= \int_{\Re^3} \rho_{fil}(r)\rho_{fjm}(r)d^3r \\ &= w_{fil}w_{fjm}\sqrt{\frac{\pi^3}{(\alpha_f)^3}} \cdot \exp\left(-\frac{\alpha_f}{4}\|r_{il} - r_{jm}\|^2\right) \end{aligned} \tag{5}$$

Kernel $K(M_i, M_j)$, or its normalized version, can be plugged in any kernel-based machine learning method, such as support vector machine (SVM) [26], support vector regression (SVR) [27], kernel ridge regression (KRR) [28], kernel partial least squares (KPLS) [29], Gaussian processes (GP) [30], etc., in order to build regression, classification, clustering, ranking, dimensionality reduction, density estimation or novelty detection models. In the case of regression models, the value of the predicted property y_t for a new molecule M_t can be calculated using the following expression:

$$y_t = \sum_j a_j K(M_t, M_j) + b. \tag{6}$$

If SVR is used for deriving the values of a_j and b, then the vector a_j appears to be sparse, with non-zero values

corresponding to a certain subset of compounds from the training set. In contrast, contributions of all molecules from the training set are needed to make predictions based on the regression models built using the KRR, KPLS and GP machine learning methods. In the latter case the value of b is assumed to be zero.

In addition to the set of adjustable coefficients a_j and b included in Eq. (6), CMF also requires a certain number of additional adjustable parameters to be computed. Among them are the parameter v for the method v-SVR and the regularization parameter γ for KRR. Their values should be optimized with the aim to improve the predictive ability of the model under construction. In addition, for each type of molecular field one can adjust the values of up to two parameters, α_f (attenuation factor, which is related to the width of the corresponding Gaussian function) and h_f (mixing coefficient for this type of molecular field). The latter parameters are sometimes called hyper-parameters in order to distinguish them from parameters a_j and b, which are computed by kernel-based machine learning methods and therefore treated differently.

The fields of regression coefficients

Equation (6) represents the dual form of the regression model, since in it the activity y_t is predicted by considering similarity measures of a test compound M_t in relation to the training set compounds M_j. In order to obtain the traditional primal form of the 3D-QSAR model, which involves an explicit consideration of molecular descriptors and regression coefficients, one can make the substitution of Eqs. (1–5) to Eq. (6) to obtain:

$$y_t = \sum_f h_f \int_{\Re^3} C_f(\mathbf{r}) X_{ft}(\mathbf{r}) d^3\mathbf{r} + b \tag{7}$$

$$X_{ft}(\mathbf{r}) = \sum_l \rho_{ftl}(\mathbf{r}) = \sum_l w_{ftl} \exp\left(-\frac{\alpha_f}{2}\|\mathbf{r} - \mathbf{r}_{tl}\|^2\right) \tag{8}$$

$$C_f(\mathbf{r}) = \sum_j a_j \sum_m \rho_{fjm}(\mathbf{r}) = \sum_j a_j \sum_m w_{fjm} \exp\left(-\frac{\alpha_f}{2}\left\|\mathbf{r} - \mathbf{r}_{jm}\right\|^2\right) \tag{9}$$

The primal form of the regression model is expressed by Eq. (7), in which the molecular field $X_{ft}(\mathbf{r})$ plays the role of the continuous field of molecular descriptors for the test molecule t, whereas $C_f(\mathbf{r})$ is a continuous field of the corresponding regression coefficients. One can give the following physical interpretation of the notion of the field of regression coefficients: *the value of a field of regression coefficients at a given point is equal to the weight with which the corresponding molecular field at that point contributes to activity value*. Therefore, the higher overlap between continuous molecular fields (which characterize molecule) and the corresponding fields of regression coefficients (which characterize 3D-QSAR model), the higher is the value of activity. This can be regarded as the principle of using the fields of regression coefficients in molecular design.

Figure 1 shows that the CMF approach can be considered as a functional generalization of the traditional approach to building 3D-QSAR models. The picture depicts 50 %-level isosurfaces for the field of regression coefficients $C(\mathbf{r})$ and the steric molecular field $X(\mathbf{r})$, along with a spatial model of the most active chemical compound from the training set. The yellow isosurfaces [marked with the minus sign (−)] correspond to the negative values, whereas the green isosurfaces [marked with the plus sign (+)] correspond to the positive areas of the fields. The high value of biological activity for this compound is the consequence of the positive overlap between $C(\mathbf{r})$, which characterizes the 3D-QSAR model, and $X(\mathbf{r})$, which characterizes the chemical compound.

The principal distinction of the CMF Eq. (7) from that of an ordinary 3D-QSAR linear model lies in the infinite number of descriptors. As a consequence, (1) continuous molecular field is used in CMF instead of several thousands of descriptors computed in CoMFA or CoMSIA at lattice points, (2) continuous field of regression coefficients is used in CMF instead of several thousands of regression coefficients obtained by means of the PLS regression, and (3) integration over the whole physical space substitutes summation over the grid points. It also follows from this analysis that isosurfaces of the fields of regression coefficients $C_f(\mathbf{r})$ could be used in the same manner and for the same purposes as contour maps in CoMFA and CoMSIA. Note, however, an important difference between

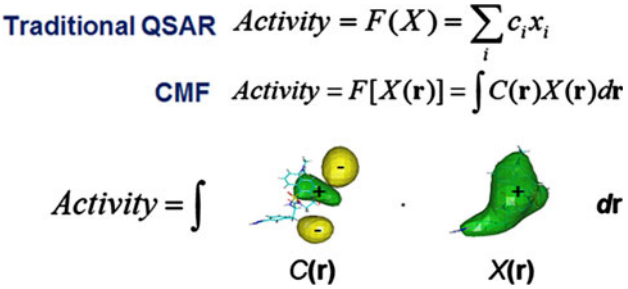

Fig. 1 The CMF approach as a functional generalization of the traditional approach to building 3D-QSAR models. The picture depicts 50 %-level isosurfaces for the field of regression coefficients $C(\mathbf{r})$ and the steric molecular field $X(\mathbf{r})$. *Each positive area* is marked with a *plus sign* (+), while *each negative area* is marked with a *minus sign* (−)

isosurfaces in CMF, from one side, and CoMFA and GRID contour maps, from the other: the former are centered on atoms [as follows from formula (9)], whereas the latter are situated around molecules (as a consequence of the impossibility to place probe atoms and groups inside atoms because of steric hindrance). Although such location of isosurfaces might seem unusual from the first glace, but it offers more direct answer to the question as to what changes should be introduced in order to increase biological activity of chemical compound. This also allows building and analyzing 3D-QSAR models based on continuous functions on coordinates (i.e. molecular fields) that describe molecular properties essential for intermolecular interactions but cannot be represented as interaction energy with some probes. Examples of such functions are electron density, most of conceptual DFT molecular fields [21], molecular orbital fields [8], E-state fields [9], fields of atom-based indicator variables [10], etc. This allows using CMF in areas other than biomedical applications, such as material science, catalysis, etc.

One of the most important features of the fields of regression coefficients stems from their ability to be represented as a weighted sum of molecular fields of training set compounds according to Eq. (9). This results in the ability to describe fields of regression coefficients in terms of molecular fields of several reference compounds which correspond to maximum absolute values of coefficients a_i in Eq. (6). This provides the means of linking the field of regression coefficients to receptor by docking the reference compounds to receptor and analyzing their interactions with it as well as giving clear interpretation to each zone of the field of regression coefficients from the point of view of ligand-receptor interactions. The detailed consideration of different ways provided by the CMF approach to give qualitative interpretation to structure–activity models is, however, beyond the scope of this paper and will be published elsewhere.

It should also be mentioned that in the framework of the CMF approach it is possible to compute continuous fields of model coefficients not only for regression, but also for: (a) classification (in the case of SVM models they describe the angular orientation of the hyperplane separating active from inactive compounds in the infinite-dimensional feature space); (b) novelty detection (or one-class classification [31], which can be used for virtual screening, see our preliminary communication [23]); (c) clustering.; (d) dimensionality reduction; etc. We hope that additional methods of analyzing and visualizing chemical databases and SAR/QSAR/QSPR models offered by the use of continuous molecular fields could deepen insights into the nature of structure–activity relationships and facilitate drug design. This is one of the major directions of our future studies.

Functional versus multivariate data analysis

Since molecular fields are treated in the CMF approach as continuous functions with respect to spatial coordinates, they can always be differentiated and integrated (even analytically, due to the use of Gaussian basis function). Easy differentiability of molecular field functions allows to apply powerful although still unexplored in chemistry apparatus of data analysis and visualization offered by functional data analysis [32].

Functional data analysis (FDA) is a newly emerged branch of mathematical statistics dealing with the analysis of functional data [32]. It can be positioned at the interface between the machine learning and functional analysis. Loosely speaking, FDA is a statistical analysis dealing with the infinite number of descriptors represented as continuous functions. FDA operates with functions (in functional Hilbert space) instead of data vectors (in Euclidean space) in multivariate data analysis, linear operators instead of matrices, integration instead of summation, etc.

Although functional data can always be discretized (like in grid-based 3D-QSAR methods) and processed using multivariate data analysis tools (such as PLS, PCA, etc.), the use of FDA offers great advantages over multivariate statistics for processing functional data. The first origin of these advantages is very valuable information contained in function derivatives, which disappears after discretization. This information is of prime importance for analyzing data, because it helps in uncovering such patterns in data that are invisible when only function values are analyzed. The most prominent example of the usefulness of derivatives is Bader's analysis of gradients (first partial derivatives) and the Laplacians (second partial derivatives) of electron density functions leading to detection of molecular graphs in molecules, which are invisible on maps and isosurfaces built for original electron density functions [33]. FDA provides powerful tools (such as Principal Differential Analysis) to analyze derivatives of function data systematically. Information on derivatives makes it possible to apply additional means (in comparison with multivariate analysis) to control smoothness/complexity of statistical models, leading to the increase in predictive performance. Indeed, the requirement for regression coefficients to change smoothly from point to point in physical space is completely ignored by multivariate methods acting on grid points, but can efficiently be used by FDA methods. Finally, the second origin of advantages of FDA methods stems from the ability to apply powerful methods of functional analysis, such as integral transforms, to data processing.

One of the main distinctions of chemical data is that they are functional by their nature. Molecules can be much more naturally described in terms of electron density functions and

molecular fields than by means of a huge number of various scalar molecular descriptors. We show in this paper that continuous functions describing molecular fields can easily be used to build 3D-QSAR models. Our assumption is that continuous functions can be considered as the third major data type in chemoinformatics, in addition to commonly used descriptor vectors and graph kernels.

Method

Parametrization of continuous molecular fields

The values of partial charges on atoms are used in this study as parameters w_{fil} in Eq. (4) for defining electrostatic molecular field. Steric molecular fields are defined as follows:

$$\rho_{fil}(\mathbf{r}) = E_{il}^{VDW} \exp\left(-\frac{1}{2}\alpha_f \left\|\frac{\mathbf{r}-\mathbf{r}_{il}}{R_{il}^{VDW}}\right\|^2\right), \tag{10}$$

where E_{il}^{VDW} (van-der-Waals energy) and R_{il}^{VDW} (van-der-Waals radius) are empirical parameters (for atom l in molecule i) takenfrom the Tripos force field [34]. This particular form of molecular fields however requires a special way of computing atomic kernels instead of formula (5):

$$k_f(A_{il},A_{jm}) = \int_{\Re^3} \rho_{fil}(\mathbf{r})\rho_{fjm}(\mathbf{r})d^3\mathbf{r} = E_{il}^{VDW}E_{jm}^{VDW} \sqrt{\frac{8\pi^3}{(\beta_{il}+\beta_{jm})^3}} \cdot \exp\left(-\frac{\beta_{il}\beta_{jm}}{2(\beta_{il}+\beta_{jm})}\left\|\mathbf{r}_{il}-\mathbf{r}_{jm}\right\|^2\right),$$
$$\beta_{il} = \frac{\alpha_f}{R_{il}^{VDW}}, \quad \beta_{jm} = \frac{\alpha_f}{R_{jm}^{VDW}} \tag{11}$$

Hydrophobic molecular fields are defined in accordance with Eq. (4) with parameters w_{fil} being equal to regression coefficients of a QSPR linear regression model based on the use of 1-atom fragment descriptors, see [35, 36]). The value of w_{fil} reflects in this case the contribution of atom l to hydrophobicity of molecule i. Two other types of molecular fields, hydrogen-bond acceptor and donor ones, are parameterized analogously, using the data on Abraham's A and B constants for 457 organic compounds, respectively [37]. All the above QSPR models were built using the fast stage-wise multiple linear regression method [38] implemented in the NASAWIN software [39]. Statistical parameters of these models were reported earlier [40].

Machine learning method

In this paper, we apply the algorithm of kernel ridge regression (KRR) to obtain an array of coefficients $A = (a_i)$, which is necessary for making predictions in accordance with formula (6) as well as for visualizing the corresponding fields of regression coefficients computed using Eq. (9), by solving the following equation:

$$\begin{bmatrix} K+\lambda I_n & 1_n \\ 1_n^T & 0 \end{bmatrix}\begin{bmatrix} A \\ b \end{bmatrix} = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \tag{12}$$

where n is the number of compounds in the training set; K is the $n \times n$ square matrix containing the values of molecular field kernels computed using expression (1) for all pairs of molecules in the training set; I_n is the $n \times n$ square identity matrix with ones on the main diagonal and zeros elsewhere; 1_n is a column vector consisting on n ones; Y is the array containing activity values for all compounds from the training set; γ is an adjustable regularization coefficient. The optimal value of the regularization parameter is determined by minimizing $RMSE_{CV}$, the root-mean-square error in cross-validation, see description below.

It should be mentioned that the first implementation of the CMF approach described in the preliminary communication [22] was based on the use of the support vector regression (SVR) method in combination with several optimization techniques to adjust the values of parameters γ, α_f and h_f. However, the first experience of practical application of this approach with numerous data sets has revealed some shortcomings. First, the application of stochastic optimization procedures did not provide good reproducibility of models. Second, some regression models built on small data sets appeared to be significantly over-fitted and did not provide acceptable predictions on external data sets. Third, the increase in the number of the types of molecular fields makes the first two problems particularly relevant. In addition, it has been shown in paper [41] that the method of support vector machines becomes very close to ridge regression when applied to functional data. In connection with all these, we have completely redesigned the CMF approach to building regression models and started to use the method of KRR instead of SVR. Through this the whole process of building regression models has become much more simple, easily controllable and reproducible.

Assessing statistical performance of 3D-QSAR models and the "model selection bias"

In this paper, we use three approaches to the assessment of the predictive ability of 3D-QSAR models. One of them is based on the use of a ten-fold cross-validation procedure [42, 43], in which the quality of models is controlled by

means of two interrelated statistical parameters, q^2 and $RMSE_{CV}$, which are calculated according to the following equations:

$$q^2 = 1 - \frac{PRESS_{CV}}{SS} = 1 - \frac{\sum_{i=1}^{n} (y_i - y_i^{pred})^2}{\sum_{i=1}^{n} (y_i - y^{mean})^2} \quad (13)$$

where $PRESS_{CV}$ is the predictive sum of squares of the differences between the experimental (y_i) and predicted (y_i^{pred}) values of biological activity over all n compounds taking part in the cross-validation procedure; SS is the sum of squared deviations of experimental values (y_i) from their arithmetic mean (y^{mean}),

$$RMSE_{CV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^{pred})^2}{n}} \quad (14)$$

where y_i is the experimental value for biological activity of the i_{th} compound, y_i^{pred} is the predicted biological activity of the same compound, and n is the number of compounds in the data set.

Although cross-validation has mathematically been proven to provide nearly unbiased estimation of the predictive performance of statistical models (e.g., see [43]), this however requires its proper use. In particular, property values of test compounds should not be involved in any stage of model construction and selection. Otherwise, cross-validation yields optimistically-biased assessment of predictive performance, see [44]. In the current version of the CMF approach, the optimal values of hyper-parameters γ, α_f and, optionally, h_f are determined by minimizing the value of $RMSE_{CV}$, which plays here the role of objective function for optimization. As can be seen from Eq. (14), experimental values of biological activity of test compounds are involved in the calculation of $RMSE_{CV}$. This means that the information concerning property values of test compounds is partially involved in model selection. In the case of small data sets this can lead to the phenomenon called "model selection bias" [45]. It is generally considered that the participation of cross-validation to set up the values of one or two parameters does not yield significant bias, and the "model selection bias" can be neglected. Indeed, it is usually ignored when q^2 is used to select the optimal number of latent variables for the PLS regression, or when cross-validation classification accuracy is used to find the best values for a couple of hyper-parameters (e.g., C and γ) for support vector machines (SVM). However, with further increase in the number of adjustable hyper-parameters involved in model selection, the "model selection bias" cannot be neglected, because it can lead to severe over-fitting.

In this connection, to assess the predictive ability of models, in addition to the internal predictive performance of 3D-QSAR models evaluated using the above-mentioned parameters q^2 and $RMSE_{CV}$, we use so-called external predictive performance assessed by making predictions on the external data sets, which are in no way involved in model construction and selection. In the latter case we use statistical parameters R^2_{pred} and $RMSE_{pred}$, which are calculated according to Eqs. (15, 16):

$$R^2_{pred} = 1 - \frac{PRESS_{TS}}{SS_{TS}} = 1 - \frac{\sum_{i=1}^{m} (y_i - y_i^{pred})^2}{\sum_{i=1}^{m} (y_i - y^{mean})^2}, \quad (15)$$

where $PRESS_{TS}$ is the sum of squares of the differences between the experimental (y_i) and predicted (y_i^{pred}) values of biological activity over all m compounds in the test set; SS_{TS} is the sum of squared deviations of experimental values (y_i) from their arithmetic mean (y^{mean}) over the test set,

$$RMSE_{pred} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^{pred})^2}{m}}, \quad (16)$$

where y_i is the experimental value for biological activity of the i_{th} compound in the test set, y_i^{pred} is the predicted biological activity of the same compound in the test set, and m is the number of compounds in the test set.

As a consequence of the above-mentioned "model selection bias", parameters q^2 and $RMSE_{CV}$ are expected to be optimistically biased. One can guess that this bias could be rather small for the case of two adjustable hyper-parameters, γ and α_f, but should become significant whenever mixing coefficients h_f are also included in the list of parameters to be optimized. From the other hand, the variance of these statistical parameters should be relatively low, since they are based on cross-validation. One can also guess that R^2_{pred} and $RMSE_{pred}$ could be nearly unbiased due to the lack of the "model selection bias". From the other hand, the latter statistical parameters should be characterized by a high variance, because they are based on a single test set.

The third approach is based on the procedure of external five-fold cross-validation (e.g., see [46]), which ensures both low bias and variance in estimating statistical parameters. In this case, the data are split 5 times into a working and an external validation set, so that each compound is selected into the latter exactly once. Each working set is split 10 times into the training and test sets in the inner ten-fold cross-validation cycle, statistical parameters of which are used to optimize adjustable hyper-parameters γ, α_f and, optionally, h_f. For each splitting, the predicted values y_i^{pred} are estimated for all compounds from the external validation set. After that statistical parameters q^2_{ex} and $RMSE_{cvex}$ are estimated for the external cross-validation using formulas analogous to (13) and (14), respectively.

So, in this paper we consider three sets of statistical parameters. The first one, q^2 and $RMSE_{cv}$, are assessed in

the internal cross-validation procedure and characterize the robustness of models. The second one, R^2_{pred} and $RMSE_{pred}$, are estimated using a single external validation set. We use them the only to compare the predictive performance of our models with literature data for the same data sets. The third one, q^2_{ex} and $RMSE_{cvex}$, evaluated using the external five-fold cross-validation procedure are used by us to provide correct estimation of the predictive performance of models.

Programming environment

The CMF approach is implemented as a set of scripts operating under the R environment for statistical computing and graphics [47]. The scripts allow loading structure–activity data sets, performing spatial alignment of molecules, building 3D-QSAR models, performing predictions for external data sets, analyzing their predictive performance, visualizing molecular fields and fields of regression coefficients, etc. A version of the software, along with the data sets needed to reproduce all 3D-QSAR models considered in this article, is available via Internet at http://sites.google.com/site/conmolfields/files (download the archive file *supplementary_material_jcamd.zip*). This set of scripts is self-contained and can be used to build 3D-QSAR models with other data sets.

Results and discussions

Modelling biological activity

Eight data sets were used in this study for testing the performance of the CMF approach in building 3D-QSAR regression models: 114 angiotensin converting enzyme (ACE) inhibitors [48], 111 acetylcholinesterase (AChE) inhibitors [25], 163 ligands for benzodiazepine receptors (BZR) [25], 322 cyclooxygenase-2 (COX-2) inhibitors [25], 397 dihydrofolatereductase (DHFR) inhibitors [25], 66 glycogen phosphorylase b (GPB) inhibitors [49], 76 thermolysin (THER) inhibitors [49], and 88 thrombine (THR) inhibitors [50]. Statistical characteristics of the obtained models were compared with the same characteristics built earlier for the same data sets using the most popular 3D-QSAR methods, comparative molecular fields analysis (CoMFA) [5] and comparative molecular similarity index analysis (CoMSIA) [11], based on the use of molecular fields. All data, including chemical structures with activity values, their splitting into the training and test sets, ionization states and conformations for all molecules, their spatial alignment and partial charges on atoms, were taken from the supplementary materials to Sutherland's paper [25]. As it was stated in the Sutherland's article [25], ionization states of molecules had been prepared by deprotonating carbocyclic acids and phosphates and protonating non-aryl basic amines (except NH_2 groups that coordinate Zn in the ACE set), atomic coordinates had been obtained by energy-minimizing the aligned molecules with MMFF94S force field in Sybyl, scaled MNDO ESP-fit partial charges [51] had been calculated for all atoms with MOPAC 6.0, except that for the THER set all partial charges on atoms had been computed using the Gasteiger-Marsili method [52] as also implemented in Sybyl. Several important characteristics of these data sets are presented in Table 1.

Statistical parameters of 3D-QSAR models obtained for these data sets with the CMF, CoMFA and CoMSIA methods are shown in Table 2, which includes the values of six statistical parameters: q^2 and $RMSE_{cv}$ characterizing internal predictive performance, R^2_{pred} and $RMSE_{pred}$—external predictive performance estimated using a single external validation set, q^2_{ex} and $RMSE_{cvex}$—external predictive performance estimated using five-fold cross-validation procedure. Both CoMSIA1 and CoMSIA2 models are based on electrostatic and steric fields molecular fields, whereas CoMSIA2 models also involve contributions from the hydrophobic and two hydrogen-bonding molecular fields. All CMF models are based on the use of all afore-mentioned five types of molecular fields. All CoMFA and CoMSIA models were obtained by using a lattice with 2 Å spacing expanding at least 4 Å in each direction beyond aligned molecules. Only the most predictive CoMFA and CoMSIA models are included in the Table 2. In the course of building CMF models only two hyper-parameters, attenuation factor α_f (which was kept the same for all types of molecular fields in this study) and regularization coefficient γ, were optimized. In this study we used the same fixed value for all mixing coefficients, $h_f = 1$. The use of only two adjustable hyper-parameters provided satisfactory external predictive performance for all data sets. Although optimization of mixing coefficients h_f always leads to sharp increase of q^2, in several cases parameter q^2_{ex} for the external predictive performance becomes lower. This might happen because of the "model selection bias".

In order to compare predictive performance of CMF models built with results published in literature, all the data sets were split into the training and the test sets with sizes specified in Table 1, then the training sets were used for building 3D-QSAR models and for assessing the internal predictive performance using the ten-fold cross-validation procedure, while the test sets were used for assessing the external predictive performance of the models.

As it is clearly seen from Table 2, models built for all data sets by using the CMF approach almost in all cases

Table 1 QSARDataSets

Ligand data set	Training set	Test set	Activity ranging
Angiotensin converting enzyme (ACE) inhibitors	76	38	pI_{50} 2.1–9.9
Acetylcholinesterase (AchE) inhibitors	74	37	pI_{50} 4.3–9.5
Ligands for benzodiazepine receptors (BZR)	98	49	pI_{50} 5.5–8.9
Cyclooxygenase-2 (COX-2) inhibitors	188	94	pI_{50} 4.0–9.0
Dihydrofolatereductase (DHFR) inhibitors	237	124	pI_{50} 3.3–9.8
Glycogen phosphorylase b (GPB) inhibitors	44	22	p*K*i 1.3–6.8
Thermolysin (THER) inhibitors	51	25	p*K*i 0.5–10.2
Thrombine (THR) inhibitors	59	29	p*K*i 4.4–8.5

In the BZR, Cox-2, and DHFR data sets several compounds (16, 40, 36, respectively) were considered as inactive and not included in the training and the test sets [25]

show better internal (cross-validation) predictive performance (i.e., higher q^2 and lower $RMSE_{cv}$) then the corresponding models obtained by the CoMFA, CoMSIA1, and CoMSIA2 methods. The only exception is BZR for CoMSIA2. The q^2 values of the CMF models obtained for the COX2 and BZR data sets are, respectively, equal and slightly lower as compared to the corresponding CoMSIA2 models.

There is also a moderate advantage in external predictive performance (estimated on external test sets using the parameters R^2_{pred} and $RMSE_{pred}$) of the CMF models over the CoMFA, models for 5 data sets (ACE, AChE, BZR, DHFR, and GPB), CoMSIA1 models for 7 data sets (ACE, AChE, BZR, COX2, DHFR, GPB, and THR), and CoMSIA2 models for 4 data sets (ACE, AChE, BZR and DHFR).

One can notice that the performance of CMF is closer to that of the CoMSIA2 approach in comparison with CoMFA and CoMSIA1. This could be attributed to the fact that the mathematical form of Eq. (4) resembles expressions for similarity indices in CoMSIA. So, in spite of absolutely different underlying ideas, CoMSIA can formally be regarded as a discretized approximation of the current version of CMF, or, vice versa, CMF—as a continuous functional extension of CoMSIA. Therefore, the difference between the models produced by these methods might result from the effect of field discretization, different statistical procedure and parameterization of molecular fields.

Figure 2 shows the scatter plots obtained for three data sets, ACE, AChE and DHFR. In the left column, i.e. in plots (a), (c) and (e), all predicted activity values were computed in the cross-validation mode. Plots (b), (d) and (f) in the right column are based on predictions obtained for external test sets. Examples of the fields of regression coefficients for the most influential types of molecular field for the same data sets are depicted in Fig. 3.

Thus, the 3D-QSAR models obtained by CMF are comparable by the predictive ability with models built by means of such popular state-of-the-art approaches as CoMFA and CoMSIA. Moreover, in some cases, e.g. for data sets ACE, AChE, BZR and DHFR, the CMF approach is clearly advantageous.

In all cases the value q^2_{ex}, which characterizes the external predictive performance, is lower than the value q^2 computed using the internal cross-validation. This means that the use of only two adjustable hyper-parameters may cause the "model selection bias". Almost in all cases the value q^2_{ex} lies between R^2_{pred} and q^2. It is interesting to note that q^2_{ex} for CMF models are usually higher than R^2_{pred} for CoMFA and CoMSIA models. The predictive performance assessed using the external five-fold cross-validation procedure is especially high for ACE, DHFR and THR.

All the above-mentioned CMF models and all results presented in Tables 2 and 3 can be reproduced using a set of R scripts available via Internet at http://sites.google.com/site/conmolfields/files (find "R scripts for CMF—publication edition" and click on the link "View" to download the archive file supplementary_material_jcamd.zip).

The main directions of further development of the CMF approach

In this article, we have considered a particular implementation of the CMF approach aimed at building 3D-QSAR models. This implementation is being actively developed, and its future version will surely be better than the current one. This section, however, concerns strategic directions of further development of the whole CMF approach.

The CMF approach is not confined to the simplest approximation scheme introduced by Eqs. (2) and (4), which lead to CoMSIA-like Eq. (5). Any number of Gaussian functions, (both isotropic, i.e. spherically symmetrical, and non-isotropic) as well as any other set of basic functions (such as splines, wavelets, etc.) can be used for approximating continuous molecular fields. This provides the ability to work with complex types of molecular fields, including those derived from the electron density function. Especially promising is the use of conceptual DFT molecular fields [21] based on conceptual DFT [53].

Although all methods of molecular alignment useful for building traditional lattice-based 3D-QSAR models can also be applied in the framework of the CMF approach, it offers additional possibilities. In particular, this approach provides a consistent criterion for the pairwise alignment of molecules i and j: maximization of the kernel $K(M_i, M_j)$.

Table 2 Statistical parameters of 3D-QSAR CMF, CoMFA CoMSIA1, and CoMSIA2 models

	CMF	CoMFA	CoMSIA1	CoMSIA2
ACE				
PLS components/additional field		3	3	2/hydro
q^2	0.72	0.68	0.65	0.66
$RMSE_{cv}$	1.24			
R^2_{pred}	0.65	0.49	0.52	0.49
$RMSE_{pred}$	1.24	1.54	1.48	1.53
q^2_{ex}	0.67			
$RMSE_{cvex}$	1.31			
AChE				
PLS components/additional field		5	6	4/hydro +H-bonding
q^2	0.58	0.52	0.48	0.49
$RMSE_{cv}$	0.79			
R^2_{pred}	0.64	0.47	0.44	0.44
$RMSE_{pred}$	0.77	0.95	0.98	0.98
q^2_{ex}	0.54			
$RMSE_{cvex}$	0.84			
BZR				
PLS components/additional field		3	3	3/hydro
q^2	0.40	0.32	0.41	0.45
$RMSE_{cv}$	0.51			
R^2_{pred}	0.20	0.00	0.08	0.12
$RMSE_{pred}$	0.79	0.97	0.93	0.91
q^2_{ex}	0.27			
$RMSE_{cvex}$	0.65			
COX2				
PLS components/additional field		5	6	4/hydro +H-bonding
q^2	0.57	0.49	0.43	0.57
$RMSE_{cv}$	0.67			
R^2_{pred}	0.14	0.29	0.03	0.37
$RMSE_{pred}$	1.23	1.24	1.44	1.17
q^2_{ex}	0.41			
$RMSE_{cvex}$	0.89			
DHFR				
PLS components/additional field		5	5	4/hydro +H-bonding
q^2	0.67	0.49	0.53	0.57
$RMSE_{cv}$	0.73			
R^2_{pred}	0.65	0.59	0.52	0.53
$RMSE_{pred}$	0.80	0.89	0.96	0.95
q^2_{ex}	0.67			
$RMSE_{cvex}$	0.75			
GPB				
PLS components/additional field		4	4	4/hydro
q^2	0.69	0.42	0.43	0.61
$RMSE_{cv}$	0.60			
R^2_{pred}	0.51	0.42	0.46	0.59

Table 2 continued

	CMF	CoMFA	CoMSIA1	CoMSIA2
$RMSE_{pred}$	0.84	0.94	0.90	0.79
q^2_{ex}	0.59			
$RMSE_{cvex}$	0.71			
THER				
PLS components/additional field		4	6	3/hydro
q^2	0.60	0.52	0.54	0.51
$RMSE_{cv}$	1.18			
R^2_{pred}	0.31	0.54	0.36	0.53
$RMSE_{pred}$	1.86	1.59	1.87	1.60
q^2_{ex}	0.40			
$RMSE_{cvex}$	1.56			
THR				
PLS components/additional field		4	5	4/hydro +H-bonding
q^2	0.73	0.59	0.65	0.72
$RMSE_{cv}$	0.50			
R^2_{pred}	0.63	0.63	0.55	0.63
$RMSE_{pred}$	0.66	0.70	0.76	0.69
q^2_{ex}	0.71			
$RMSE_{cvex}$	0.54			

Moreover, the CMF approach offers an additional criterion for the multiple alignment of molecules—the "compressibility" of molecular fields, which can be assessed using unsupervised dimensionality reduction approaches, such as the kernel (functional) principal component analysis. Both criteria could also be applied to choose molecular conformations for tackling the problem of molecular flexibility. Furthermore, the CMF approach can provide a new way of building alignment-free 3D-QSAR models through the use of 3D-rotation invariant kernels [54, 55] in the frame of the concept of invariant pattern recognition [56].

The CMF approach can be extended to the case of the existence of several tautomers, ionization (protonation) states and conformers by replacing the Eq. (4) with its more general form (17):

$$\rho_{fil}(\mathbf{r}) = \sum_s v_s \sum_t v_{st} \sum_c v_{stc} w_{fil} \exp\left(-\frac{1}{2}\alpha_f \|\mathbf{r} - \mathbf{r}_{il}\|^2\right), \quad (17)$$

where index s counts different ionization (protonation) states, index t counts tautomers, index c counts conformers, v_s is the population of the ionization state s, v_{st} is the relative population of tautomer t in ionization state s, while v_{stc} is the relative population of the conformer c of the molecule in ionization state s and in tautomeric state t. Populations v_s, v_{st} and v_{stc} can be assessed using molecular modeling simulations.

One of the most promising ways to develop further the CMF approach is its further extension to the description of the binding sites of biomolecules and their interactions with ligands. This seems to be feasible, because molecular fields of biological targets are identical by their nature to those of small ligands. One can suggest several ways of conducting research in this direction. First, kernels for comparing molecular fields of biological targets (or their binding sites) could be constructed in exactly the same way as it has been done small molecules and described in this paper. Second, kernels for protein–ligand pairs could be constructed by combining kernels for small molecules (ligands) and kernels for macromolecules (proteins), as it was done by Erhan et al [57],Faulon et al [58], Jacob and Vert [59], and Geppert et al. [60]. Third, kernels for protein–ligand interactions could be constructed in the frame of the CMF approach by encapsulating products of molecular fields of ligand and protein into kernels.

The CMF approach is easily extensible thanks to its modularity. By combining different types of molecular fields, different types of kernels with different types of kernel-based machine learning methods, one can obtain various methods for building SAR/QSAR/QSPR models and conducting virtual screening. Table 3 lists different

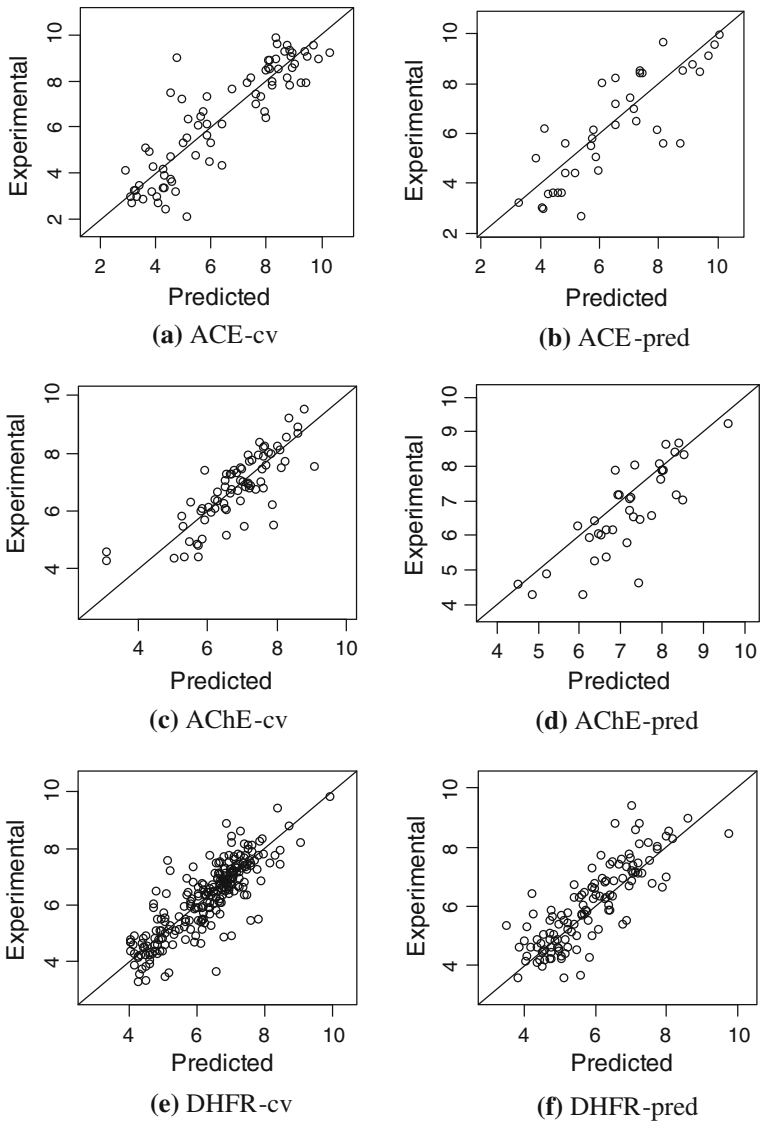

Fig. 2 *Scatter plots* for data sets ACE (**a**, **b**), AChE (**c**, **d**) and DHFR (**e**, **f**) made for cross-validation (**a**, **c**, **e**) and external test set prediction (**b**, **d**, **f**)

tasks being solved by kernel-based machine learning methods, the names of such methods, and the roles that the CMF approach could play in conjunction with them in chemoinformatics. The first row in this table deals with the regression task considered in this paper. The second row concerns the use of molecular kernels in combination with one-class classification kernel-based methods for conducting virtual screening based on the similarity of molecular fields. The feasibility of this approach has already been proved by us, see preliminary communication [23]. The rest of the table shows the promising directions for further development of the CMF approach.

Like any new approach, CMF is not free from certain disadvantages and limitations. As it has already been discussed, one of the drawbacks of the CMF approach in its present state is the danger of over-fitting because of the model selection bias [45]. Several ways to prevent this phenomenon have been suggested in literature, including Bayesian regularization of hyper-parameters [61] and hyper-parameter averaging [62]. Algorithms of multiple kernel learning [63] might also be useful in this case. For small datasets the preferred solution would be to apply the full Bayesian approach [64], where the hyper-parameters are integrated out rather than optimized.

One of the most serious limitations of the CMF approach, at least in its present form, comes from the mere nature of kernel-based machine learning methods. The amount of computational resources needed to calculate a kernel matrix scales as a square of the number of compounds in the training set. The mean amount of computational resources needed to calculate each element of a kernel matrix also scales as a square of the average number of atoms in molecules. As a result, it becomes impractical to build 3D-QSAR models using a training set with more than 300 medium sized compounds. Obviously, without finding at least partial solution to this problem it will be

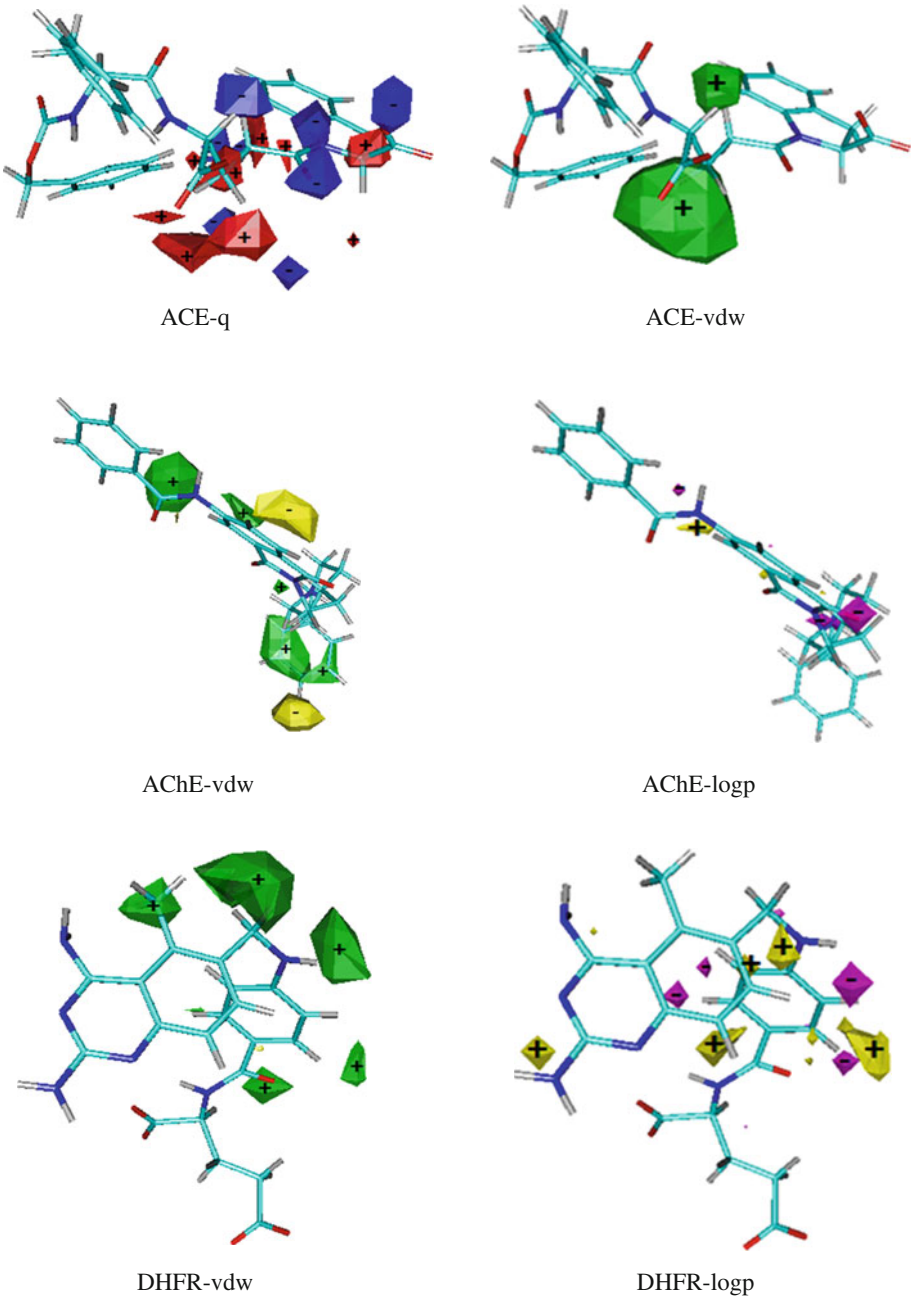

Fig. 3 Fields of regression coefficients (*q*—electrostatic field, *vdw*—steric field, *logp*—hydrophobic field). *Each positive area* is marked with a *plus sign* (+), while *each negative area* is marked with a *minus sign* (−)

impossible to process big chemical databases. Recent achievements in developing kernel based algorithms for processing huge data sets (see [65, 66]) suggest that such a solution may well be found.

Conclusions

The CMF approach, developed in this paper, describes molecules by ensemble of continuous functions (molecular fields), instead of finite sets of molecular descriptors (such as interaction energies computed at grid nodes). The potential advantages of this approach results from the ability to approximate electronic molecular structures with any desirable accuracy level, the ability to leverage the valuable information contained in partial derivatives of molecular fields (otherwise lost upon discretization) to analyze models and enhance their predictive performance, the ability to apply integral transforms to molecular fields and models, etc.

The most attractive features of the CMF approach are its versatility and universality. By combining different types of molecular fields and methods of their approximation, different types of kernels with different types of kernel-based machine learning methods, it is possible to present lots of existing methods in chemoinformatics and

Table 3 The use of the CMF approach in conjunction with different kernel-based machine learning methods

Machine learning task	Machine learning methods	Role in chemoinformatics
Regression	SVR, KRR, KPLS, KQR, GP-R	QSAR/QSPR
One-class classification (novelty detection)	1-SVM, SVDD	Virtual screening based of similarity of molecular fields
Binary and multi-class classification	SVM, GP-C	Classification of chemical compounds (active/inactive), predicting profiles of biological activity for chemical compounds
Rank correlation	K-ranking	Ranking chemical compounds by property, virtual screening
Dimensionality reduction and data visualization	K-PCA, KFA	Drawing maps of chemical space
Cluster analysis	SPECC	Classification of chemical compounds by mechanism of action (including binding mode)
Canonical correlation	KCCA	Establishing relationships between molecular fields of ligands and molecular fields of the corresponding binding sites in proteins

medicinal chemistry as particular cases within a universal methodology. The CMF methodology can easily be extended to building classification and novelty detection models, visualizing them, performing virtual screening, processing diverse datasets. The feasibility of this has recently been demonstrated [23, 24].

This article deals with the particular application of the CMF approach to building 3D-QSAR models through the use of five types of molecular fields (the electrostatic, steric, hydrophobic, hydrogen-bond acceptor and donor ones), the simplest linear convolution molecular kernel with the contribution of each atom to each type of fields approximated with a single isotropic Gaussian function of the same width, and the kernel ridge regression data analysis technique. The resulting procedure may be regarded as a functional extension of the CoMSIA method. As follows from the results presented in this paper, this particular implementation of the CMF approach provides an appealing alternative to the traditional lattice-based methodology. This method provides either comparable or enhanced predictive performance in comparison with state-of-the-art 3D-QSAR methods, such as CoMFA and CoMSIA. Not using lattices, it completely eliminates the problem of the choice of the lattice parameters, i.e. its origin, orientation, extent and step size. Moreover, the CMF approach eliminates the loss of important information resulting from the use of coarse lattice. It can easily be implemented and built into statistical data analysis software or molecular modeling environment. A reference open-source implementation using scripts for the R environment for statistical computing and graphics is available via Internet at http://sites.google.com/site/conmolfields/files.

We see the following main directions for further development of the CMF approach: introduction of additional types of molecular fields as well as improvement of the existing ones; finding efficient solutions the problem of over-fitting resulting from the model selection bias; finding the ways to work with big sets of chemical compounds; tackling the issue of molecular alignment and flexibility; taking into account different ionization states of molecules, their tautomers and conformers; extension of the approach to work with biological macromolecules and supramolecular complexes; combining with different machine learning methods aimed at solving various tasks.

Acknowledgments The authors thank Prof. Yu. A. Ustynyuk for stimulating discussion and advice. The authors also thank Prof. A. Varnek and Dr. G. Marcou for valuable comments regarding the developed approach. This work was supported by Russian Foundation for Basic Research (Grant 13-07-00511).

References

1. Kubinyi H (ed) (2000) 3D QSAR in drug design: vol 1: theory methods and applications (three-dimensional quantitative structure activity relationships). Kluwer/Escom, Dordrecht
2. Kubinyi H, Folkers G, Martin YC (eds) (2002) 3D QSAR in drug design. Vol 2: ligand-protein interactions and molecular similarity, vol 2. Kluwer Academic Publishers, Dordrecht
3. Kubinyi H, Folkers G, Martin YC (eds) (2002) 3D QSAR in drug design. Vol 3: recent advances. Kluwer Academic Publishers, Dordrecht
4. Cruciani G (ed) (2006) Molecular interaction fields; application to drug discovery and ADME prediction. Wiley-VCH, Weinheim
5. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA) 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110(18):5959–5967. doi:10.1021/ja00226a005
6. Testa B, Carrupt PA, Gaillard P, Billois F, Weber P (1996) Lipophilicity in molecular modeling. Pharm Res 13(3):335–343. doi:10.1023/a:1016024005429
7. Kim KH, Greco G, Novellino E, Silipo C, Vittoria A (1993) Use of the hydrogen bond potential function in a comparative molecular field analysis (CoMFA) on a set of benzodiazepines. J Comput Aided Mol Des 7(3):263–280
8. Waller CL, Marshall GR (1993) Three-dimensional quantitative structure-activity relationship of angiotesin-converting enzyme and thermolysin inhibitors. II. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free

energies based on active-analog and complementary-receptor-field alignment rules. J Med Chem 36(16):2390–2403
9. Kellogg GE (1996) E-state fields: applications to 3D QSAR. J Comput Aided Mol Des 10(6):513–520
10. Kroemer RT, Hecht P (1995) Replacement of steric 6–12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency. J Comput Aided Mol Des 9(3):205–212
11. Klebe G, Abraham U (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. J Comput Aided Mol Des 13(1):1–10
12. Goodford P (2006) The basic principles of GRID. In: Cruciani G (ed) Molecular interaction fields. Applications in drug discovery and ADME prediction. Methods and principles in medicinal chemistry, vol 27. Wiley-VCH, Weinheim, pp 3–26
13. Höskuldsson A (1988) PLS regression methods. J Chemom 2(3): 211–228
14. Martin RL, Gardiner E, Gillet VJ, Muñoz-Muriedas J, Senger S (2010) Wavelet approximation of GRID fields: application to quantitative structure-activity relationships. Mol Inform 29(8–9):603–620. doi:10.1002/minf.201000066
15. Tetko IV, Kovalishyn VV, Livingstone DJ (2001) Volume learning algorithm artificial neural networks for 3D QSAR studies. J Med Chem 44(15):2411–2420
16. Brown WM, Sasson A, Bellew DR, Hunsaker LA, Martin S, Leitao A, Deck LM, Vander Jagt DL, Oprea TI (2008) Efficient calculation of molecular properties from simulation using kernel molecular dynamics. J Chem Inf Model 48(8):1626–1637. doi:10.1021/ci8001233
17. Cheeseright T, Mackey M, Rose S, Vinter A (2006) Molecular field extrema as descriptors of biological activity: definition and validation. J Chem Inf Model 46(2):665–676. doi:10.1021/ci050357s
18. Carbo-Dorca R, Robert D, Amat L, Girones X, Besalu E (2000) Molecular quantum similarity in QSAR and drug design. Lecture notes in chemistry. Springer, Berlin
19. Fradera X, Amat L, Besalu E, Carbo-Dorca R (1997) Application of molecular quantum similarity to QSAR. Quant Struct Act Relat 16(1):25–32
20. Besalu E, Girones X, Amat L, Carbo-Dorca R (2002) Molecular quantum similarity and the fundamentals of QSAR. Acc Chem Res 35(5):289–295
21. Van Damme S, Bultinck P (2009) 3D QSAR based on conceptual DFT molecular fields: antituberculotic activity. J Mol Struct THEOCHEM 943(1–3):83–89. doi:10.1016/j.theochem.2009.10.031
22. Zhokhova NI, Baskin II, Bakhronov DK, Palyulin VA, Zefirov NS (2009) Method of continuous molecular fields in the search for quantitative structure-activity relationships. Dokl Chem 429(1):273–276
23. Karpov PV, Baskin II, Zhokhova NI, Zefirov NS (2011) Method of continuous molecular fields in the one-class classification task. Dokl Chem 440(2):263–265
24. Karpov PV, Baskin II, Zhokhova NI, Nawrozkij MB, Zefirov AN, Yablokov AS, Novakov IA, Zefirov NS (2011) One-class approach: models for virtual screening of non-nucleoside HIV-1 reverse transcriptase inhibitors based on the concept of continuous molecular fields. Russ Chem Bull 60(11):2418–2424
25. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure-activity relationships. J Med Chem 47(22):5541–5554
26. Vapnik VN (1995) The nature of statistical learning theory. Springer, Berlin
27. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
28. Smola AJ, Scholkopf B, Muller KR (1998) The connection between regularization operators and support vector kernels. Neural Netw 11(4):637–649. doi:10.1016/s0893-6080(98)00032-x
29. Bennett KP, Embrechts MJ (2003) An optimization perspective on kernel partial least squares regression. In: Suykens JAK, Horvath G, Basu S, Micchelli C, Vandewalle J (eds) Advances in learning theory: methods, models and applications. NATO science series III: computer and systems sciences, vol 190. IOS Press, Amsterdam, pp 227–250
30. Rasmussen CE, Williams CKI (2006) Gaussian processes in machine learning. Adaptive computation and machine learning. The MIT Press, Cambridge
31. Baskin II, Kireeva N, Varnek A (2010) The one-class classification approach to data description and to models applicability domain. Mol Inform 29(8–9):581–587. doi:10.1002/minf.201000063
32. Ramsay JO, Silverman BW (2005) Functional data analysis. Springer series in statistics, 2nd edn. Springer, New York
33. Bader RFW (1985) Atoms in molecules. Acc Chem Res 18(1):9–15
34. Tripos Inc., St. Louis, MO. http://www.tripos.com
35. Artemenko NV, Baskin II, Palyulin VA, Zefirov NS (2001) Prediction of physical properties of organic compounds using artificial neural networks within the substructure approach. Dokl Chem 381(1):317–320
36. Artemenko NV, Baskin II, Palyulin VA, Zefirov NS (2003) Artificial neural network and fragmental approach in prediction of physicochemical properties of organic compounds. Russ Chem Bull 52(1):20–29
37. Jover J, Bosque R, Sales J (2004) Determination of Abraham solute parameters from molecular structure. J Chem Inf Comput Sci 44(3):1098–1106
38. Zhokhova NI, Baskin II, Palyulin VA, Zefirov AN, Zefirov NS (2007) Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies. Dokl Chem 417(2):282–284
39. Baskin II, Halberstam NM, Artemenko NV, Palyulin VA, Zefirov NS (2003) NASAWIN—a universal software for QSPR/QSAR studies. In: Ford M (ed) EuroQSAR 2002 designing drugs and crop protectants: processes, problems and solutions. Blackwell Publishing, Massachusetts, pp 260–263
40. Baskin II, Zhokhova NI, Palyulin VA, Zefirov AN, Zefirov NS (2009) Multilevel approach to the prediction of properties of organic compounds in the framework of the QSAR/QSPR methodology. Dokl Chem 427(1):172–175
41. Rossi F, Villa N (2006) Support vector machine for functional data classification. Neurocomputing 69(7–9):730–742
42. Geisser S (1993) Predictive inference. Chapman and Hall, New York
43. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc 36:111–147
44. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20(4):269–276
45. Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection Bias in performance evaluation. J Mach Learn Res 11:2079–2107
46. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A (2008) Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. J Chem Inf Model 48(9):1733–1746. doi:10.1021/ci800151m
47. R: A Language and Environment for Statistical Computing (2012). http://www.R-project.org/
48. DePriest SA, Mayer D, Naylor CB, Marshall GR (1993) 3D-QSAR of angiotensin-converting enzyme and thermolysin inhibitors: a comparison of CoMFA models based on deduced

and experimentally determined active site geometries. J Am Chem Soc 115(13):5372–5384. doi:10.1021/ja00066a004
49. Gohlke H, Klebe G (2002) DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. J Med Chem 45(19):4153–4170. doi:10.1021/jm020808p
50. Böhm M, Stüjrzebecher J, Klebe G (1999) Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J Med Chem 42(3):458–477. doi:10.1021/jm981062r
51. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11(4): 431–439. doi:10.1002/jcc.540110404
52. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. Tetrahedron 36(22):3219–3228
53. Geerlings P, De Proft F, Langenaeker W (2003) Conceptual density functional theory. Chem Rev 103(5):1793–1874. doi: 10.1021/cr990029p
54. Hamsici OC, Martinez AM (2009) Rotation invariant kernels and their application to shape analysis. IEEE Trans Pattern Anal 31(11):1985–1999. doi:10.1109/tpami.2008.234
55. Haasdonk B, Burkhardt H (2007) Invariant kernel functions for pattern analysis and machine learning. Mach Learn 68(1):35–61. doi:10.1007/s10994-007-5009-7
56. Wood J (1996) Invariant pattern recognition: a review. Pattern Recognit 29(1):1–17. doi:10.1016/0031-3203(95)00069-0
57. Erhan D, L'Heureux P-J, Yue SY, Bengio Y (2006) Collaborative filtering on a family of biological targets. J Chem Inf Model 46(2):626–635
58. Faulon J-L, Misra M, Martin S, Sale K, Sapra R (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. Bioinformatics 24(2):225–233. doi:10.1093/bioinformatics/btm580
59. Jacob L, Vert JP (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24(19): 2149–2156
60. Geppert H, Humrich J, Stumpfe D, Gaertner T, Bajorath J (2009) Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. J Chem Inf Model 49(4):767–779. doi:10.1021/ci900004a
61. Cawley GC, Talbot NLC (2007) Preventing over-fitting during model selection via bayesian regularisation of the hyper-parameters. J Mach Learn Res 8:841–861
62. Hall P, Robinson AP (2009) Reducing variability of cross validation for smoothing-parameter choice. Biometrika 96(1): 175–186. doi:10.1093/biomet/asn068
63. Gönen M, Alpaydin E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
64. Bishop CM (2006) Pattern recognition and machine learning. Information Science and Statistics, Springer
65. Varnek A, Baskin I (2012) Machine learning methods for property prediction in chemoinformatics: quo vadis? J Chem Inf Model 52(6):1413–1437. doi:10.1021/ci200409x
66. Huang T-M, Kecman V, Kopriva I (2006) Kernel based algorithms for mining huge data sets. Supervised, semi-supervised, and unsupervised learning. Springer, Berlin