Molecule Kernels: A Descriptor- and Alignment-Free Quantitative Structure–Activity Relationship Approach

Johannes A. Mohr,*[,†,‡] Brijnesh J. Jain,[†] and Klaus Obermayer[†]

School for Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany, and Department of Psychiatry and Psychotherapy, CCM, Charité- University Medicine, Berlin, Germany

Received April 27, 2008

Quantitative structure activity relationship (QSAR) analysis is traditionally based on extracting a set of molecular descriptors and using them to build a predictive model. In this work, we propose a QSAR approach based directly on the similarity between the 3D structures of a set of molecules measured by a so-called molecule kernel, which is independent of the spatial prealignment of the compounds. Predictors can be build using the molecule kernel in conjunction with the potential support vector machine (P-SVM), a recently proposed machine learning method for dyadic data. The resulting models make direct use of the structural similarities between the compounds in the test set and a subset of the training set and do not require an explicit descriptor construction. We evaluated the predictive performance of the proposed method on one classification and four regression QSAR datasets and compared its results to the results reported in the literature for several state-of-the-art descriptor-based and 3D QSAR approaches. In this comparison, the proposed molecule kernel method performed better than the other QSAR methods.

1. INTRODUCTION

The 3D structure of a molecule is closely related to its physical, chemical, and biological properties. This is expressed in the similarity principle: “Similar structures have similar physicochemical properties and biological activities”. In quantitative structure−activity relationship (QSAR) analysis, the aim is to predict the biological activity, toxicity, or mutagenicity of a drug. This is useful in drug discovery during the search for lead compounds, where the aim is to maximize the potency or selectiveness of a drug, while at the same time looking for a compound with good pharmacokinetic properties and minimum toxicity. These depend on the local and global electronic, hydrophobic, lipophilic, and steric properties of the compound which are implicitly determined by its 3D structure.

Traditionally, QSAR analysis starts with a representation of the molecular structure, from which a a large number of descriptors are generated, that are concatenated into a descriptor vector. These descriptors replace the initial representation of the molecule. They explicitly encode some aspects of the information which is implicitly contained in the original structure. According to the “dimensionality” they represent, the descriptors are categorized into different classes: The 0D descriptors contain counts of entities like atoms, elements, and bond types, 1D descriptors consist of path and walk counts, 2D descriptors describe the topology of the molecule and are based on the structural formula of a molecule, and 3D descriptors require the reconstruction of the 3D geometry of the compound. These include descriptors obtained by 3D QSAR methods using force-field calculations and physicochemical models. On the basis of this descriptor representation, a predictor is learned on the given training data, which assigns a regression value or a class label to a molecule. Since the number of descriptor vectors is very large, often exceeding the available sample size, these predictors usually involve feature selection or feature construction (e.g., principle component analysis) to reduce the input dimensionality. The most popular choice in chemoinformatics is the partial least-squares method (PLS).[1]

Recent advances in the field of statistical learning theory[2] have led to the development of kernel methods, yielding predictors with very good generalization performance working in high dimensional feature spaces. In general, kernels are functions which take two objects (data points, examples) as input and assign a scalar output value, which is interpreted as a measure of similarity between the objects. The values of the kernel for all pairs of examples in a given dataset are summarized into a matrix called “kernel matrix”. Well-known examples of kernel methods are the support vector machines[3] for classification and regression. Usually, these approaches are applied to vectorial data; however in the past few years, an increasing number of kernel techniques for structured data, like sequences, trees, and graphs, have been developed.[4] In QSAR analysis, support vector machines have been mainly applied to descriptor vectors, although recently approaches based on positive-definite graph kernels[5−7] have been suggested. These graph kernel methods define a similarity function between two molecules by considering them as graphs, in which atoms correspond to vertices and bonds to edges. However, the runtime complexity of these algorithms often grows quickly with the number of atoms in the molecule. The calculation of an all-subgraphs kernel, which calculates the number of all subgraph-isomorphisms, is practically infeasible (NP-hard).[5] Other graph kernel approaches count common walks in two graphs (product

* Corresponding author. Mailing address: FR 2-1, Franklinstrasse 28/29, D-10587 Berlin, Germany. E-mail: johann@cs.tu-berlin.de. Phone: +49 (0)30 31473628. Fax: +49 (0)30 31473121.

[†] Berlin Institute of Technology.

[‡] Charité-University Medicine.

10.1021/ci800144y CCC: $40.75
Published on Web 09/04/2008

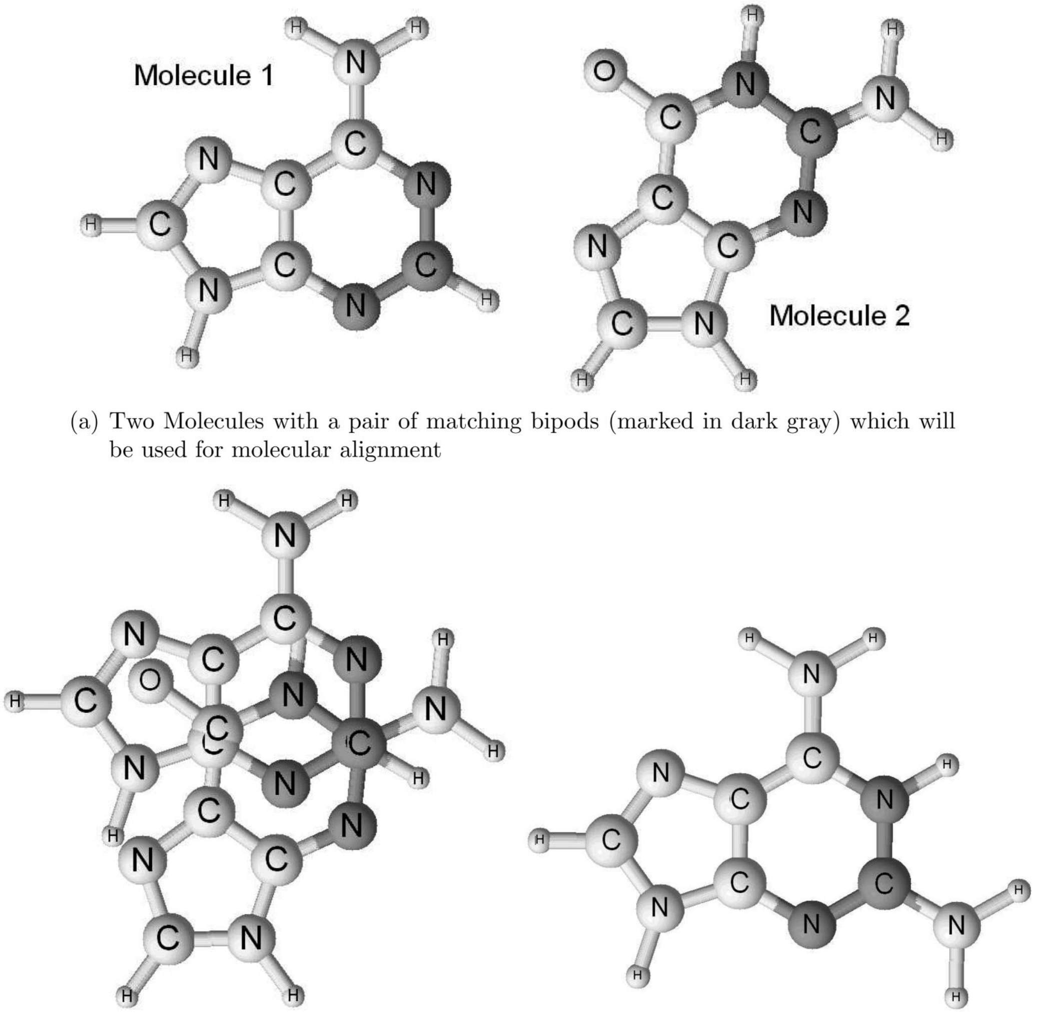
graph kernels[5]) or calculate the expectation of a kernel over all pairs of label sequences in two graphs using random walks (marginalized graph kernels[6]). Nevertheless, the runtime complexity of these approaches still scales with $O(n^6)$, where n is the number of atoms in a molecule, making them impractical for larger-size molecules. Recently, computationally more efficient positive-definite graph kernels based on molecular fingerprints have been proposed.[7] These calculate the similarity of vectors of counts of all labeled paths with a maximum length derived by depth-first searches starting from each vertex of a molecular graph. However, all the above graph kernels make use of path or walk counts, like 1D descriptors, but do not consider the full information contained in the 3D molecular structures.

In this paper, we propose a novel kernel method for QSAR analysis, which is not based on the construction of descriptor vectors but directly evaluates the similarity between two molecular 3D structures. It makes use of the fact that information about the physicochemical properties is already implicitly contained in the 3D structure. So instead of trying to explicitly extract these properties in form of descriptor vectors, the similarity between all pairs of molecular structures can directly be used for prediction. To this end, a new family of structural similarity measures called *molecule kernels* is introduced. In contrast to graph kernels, molecule kernels take both topology and 3D geometry of the molecules into account. However, the resulting kernel matrix is not positive definite, a property required by conventional kernel methods like support vector machines. Therefore, we use the recently proposed potential support vector machine[8] (P-SVM) for dyadic data as predictor, which does not require positive definite kernel matrices. If trained on a molecule kernel matrix, the P-SVM implicitly encodes information about certain structural elements or substructures which are relevant for predicting the desired end point. Unlike methods based on structural libraries, where the presence or absence of elements from a predefined set of structures is encoded explicitly, in our approach the structural elements are encoded implicitly via the parameters of the predictor and the values of the molecule kernel matrix. Like other 3D QSAR approaches, the proposed method requires suitable 3D conformations, which we assume have been determined using geometry optimization techniques or molecular mechanics. However, a spatial prealignment of the compounds with respect to each other or any grid is not necessary.

Suitable kernels for molecules need to fulfill two criteria: First, they need to capture specific aspects of the 3D structure of two molecules, which are expressed in form of a similarity function. Second, they need to be efficiently computable. The problem of maximizing a similarity function which depends on the spatial alignment of two compounds leads to a continuous optimization problem with many local optima. If gradient-based optimization techniques are employed, the solution will most likely correspond to a local optimum. A method based on randomized search heuristics which employs rational function optimization has been proposed by Kearsley et al.;[9] however, it depends on the choice of an alignment parameter and is not guaranteed to identify the global optimum. The molecule kernels we propose in this paper transform the continuous optimization problem into a discrete optimization problem by considering the set of all at least locally matching alignments.[10] The smallest units describing a unique 3D alignment of two compounds consists of a pair of ordered bipods (V-shaped subfragments), one from each compound. This is illustrated in Figure 1, where two molecules are aligned according to the alignment of two matching bipods. Molecule kernels make use of this fact by finding the global optimum in the space of all matching bipod alignments, a problem which can be solved with a runtime complexity of $O(n^4)$.

Moreover, while most other 3D QSAR methods assume the activity of all compounds is mediated by the same interaction mechanism between ligand and receptor protein, the molecular kernel approach does not require this assumption. The reason for this lies in the fact that instead of using a global alignment of all molecules to a common scaffold, receptor, or grid, molecule kernels are based on the optimum pairwise alignments between molecules. The mutual similarities can involve different alignments, which can correspond to different active groups for each pair of molecules. So different mechanisms can be modeled at once, as long as they are all sufficiently represented in the training data.

We compare the predictive performance of the proposed method on publicly available datasets to results from several other QSAR methods which were taken from the literature. In the following, we will review these methods briefly. The technique of hologram QSAR (HQSAR)[11] divides each molecule into a set of overlapping structural fragments and uses their frequency as 2D descriptors. One widely used 3D-QSAR approach is comparative molecular field analysis (CoMFA),[12] which is based on the assumption that similar steric and electrostatic fields of molecules lead to similar activity. CoMFA requires the spatial superimposition of the molecular structures, which needs expert knowledge, is time-consuming and might introduce user bias. After alignment, a 3D grid is generated around the molecule and local potential fields are calculated at the grid points. Steric fields are modeled by the Lennard-Jones potential, and electrostatic fields are computed using the Coulomb potential of the partial atomic charges. Then the interaction energies between certain probes and the molecule are calculated at the grid points and used as descriptor vectors in order to predict the biological activity. A problem with this approach lies in the abrupt change in potential at the Van-der-Waals surface of the molecule, which leads to a critical dependency of the CoMFA results on grid spacing and the relative orientation of molecules and grid. An alternative technique which is less sensitive to these issues is the comparative molecular similarity index analysis (CoMSIA).[13] It makes use of a Gaussian approximation of the force fields which is not subject to sudden potential changes. In addition to steric and electrostatic fields sometimes also hydrophobic fields and hydrogen bond donor/acceptor fields are modeled. Another 3D QSAR approach is the GRIND method,[14] where grid independent descriptors encode the spatial distribution of molecular interaction fields (MIFs) based on an autocorrelation function. An extension of this, the Anchor-GRIND method,[15] allows the inclusion of a priori chemical and biological knowledge. The user defines a specific position of the molecular structure (the "anchor point"), which is used as reference in the comparison of the MIFs of the compounds. This allows a better description of the compounds in the vicinity of a common substructure. It requires less human supervision than the previous 3D QSAR methods but



(a) Two Molecules with a pair of matching bipods (marked in dark gray) which will be used for molecular alignment

(b) The same molecules after a translation of molecule 2 such that the central bipod atoms of the two bipods are superposed.

(c) Molecule alignment corresponding to the matching bipod alignment obtained after the rotation which optimally aligns the bipod atoms in a least squares sense

Figure 1. Illustration of the concept of a molecular alignment using bipod matching. For ease of visualization, this is shown using a simple 2D example. All formal definitions and the details of the alignment procedure will be given in the methods section.

is only applicable for a set of compounds sharing a common scaffold. Finally, QSAR by eigenvalue analysis (EVA)[17] uses a descriptor based on molecular vibrational spectra. It has the advantage of being invariant to the spatial alignment of the molecule.

2. METHODS

2.1. Molecule Kernels. In the following, we introduce the mathematical formalism necessary for defining molecule kernels.

Definition 1 (Molecule). A molecule is an attributed undirected graph $\mathcal{M} = (\mathcal{A}, \mathcal{B}, \epsilon, \tau, \xi)$, *where*

- $\mathcal{A} = \{A_1,..., A_N\}$ *is a set of vertices called atoms*
- $\mathcal{B} \subseteq \mathcal{A}^{[2]} = \{\{A_i, A_j\}|A_i, A_j \in A, i \neq j\}$ *a set of edges corresponding to the bonds between the atoms*
- ϵ: $\mathcal{A} \to \mathcal{E}$ *a mapping of the atoms to a set of labels* $\mathcal{E}$ *corresponding to the chemical elements*
- τ: $\mathcal{B} \to \mathcal{T}$ *a mapping of the bonds to a set of labels* $\mathcal{T}$ *corresponding to bond types*
- ξ: $\mathcal{A} \to \mathbb{R}^3$ *a mapping of the atoms to their 3D coordinates*

The edges of the graph, the bonds, have labels corresponding to the bond types, which can be represented by integer numbers such that $\mathcal{T} = \{1, 2, 3, 4\}$ (1 a single bond, 2 a double bond, 3 a triple bond, 4 an aromatic bond). The vertices of the graph, the atoms, have labels corresponding to the different chemical elements. Therefore we can chose the set $\mathcal{E}$ to consist of all elemental symbols, $\mathcal{E} = \{$H, He, Li, Be, B, C, N, O, etc.$\}$. Moreover, the atoms have real valued, three-dimensional attribute vectors, corresponding to the three-dimensional atomic coordinates in units of angstroms. Note that the rotation and translation of the whole molecule with respect to the global coordinate system is arbitrary.

A molecule contains both topological information, which is determined by the graph structure, as well as geometrical information, which is determined by the coordinate mapping ξ. In a QSAR learning task, we are given a dataset $D = \{(\mathcal{M}_p, t_p), p = 1,..., m\}$ consisting of pairs of molecules and target values. The target values t_p can be either real valued or binary class labels (+1, −1). To distinguish between different molecules from the dataset, the constituents of molecule $\mathscr{M}_p$ are denoted by superscript indices, e.g. the nth atom in molecule $\mathscr{M}_p$ is denoted by A^p_n, while its coordinates are denoted by $\xi^p(A^p_n)$.

If we are given the coordinates of two molecules $\mathscr{M}_p$ and $\mathscr{M}_q$ in the same coordinate system, a *molecular alignment* is a rigid body transformation (involving only translations and rotations) of the coordinates of $\mathscr{M}_q$.

Definition 2 (Bipod). A bipod B^p_{ijk} is an ordered triplet of atoms from the same molecule $\mathscr{M}_p$ connected by two bonds

$$B^p_{ijk} = (A^p_i, A^p_j, A^p_k), \quad \text{with}\{A^p_i, A^p_j\} \in \mathscr{B} \text{ and } \{A^p_j, A^p_k\} \in \mathscr{B} \tag{1}$$

for which the vectors $\xi^p(A^p_i) - \xi^p(A^p_j)$ and $\xi^p(A^p_k) - \xi^p(A^p_j)$ are not colinear.[18]

Thus a bipod is a V-shaped subfragment of a molecule; see Figure 1 for an example. Note that, in general, $B^p_{ijk} \neq B^p_{kji}$, because the triplets are ordered. The middle atom A^p_j in a bipod B^p_{ijk} will be referred to as the *central bipod atom*.

Definition 3 (Matching Bipod Alignment). Let θ be a constant. Assume there exists a pair of bipods (B^p_{ijk}, B^q_{rst}) belonging to molecules $\mathscr{M}_p$ and $\mathscr{M}_q$ such that $\epsilon(A^p_i) = \epsilon(A^q_r)$, $\epsilon(A^p_j) = \epsilon(A^q_s)$ and $\epsilon(A^p_k) = \epsilon(A^q_t)$. Moreover assume there is a transformation $\mathbf{F}$: $\mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{x} \to \tilde{\mathbf{x}} = \mathbf{RT}(\mathbf{x})$, such that

$$\begin{aligned}(\xi^p(A^p_i) - \mathbf{F}\xi^q(A^q_r))^2 &\leq \theta \\ (\xi^p(A^p_j) - \mathbf{F}\xi^q(A^q_s))^2 &\leq \theta \\ (\xi^p(A^p_k) - \mathbf{F}\xi^q(A^q_t))^2 &\leq \theta \end{aligned} \tag{2}$$

where $\mathbf{T}$: $\mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{x} \to \tilde{\mathbf{x}}$ is a translation which superposes the central bipod atoms, i.e.

$$\mathbf{T}\xi^q(A^q_j) = \xi^p(A^p_s) \tag{3}$$

and $\mathbf{R}$: $\mathbb{R}^3 \to \mathbb{R}^3$, $\mathbf{x} \to \tilde{\mathbf{x}}$ is the rotation around the superposed central bipod atoms which optimally aligns the bipods in a least-squares sense. Then, the transformation $\mathbf{F}$ is called a **matching bipod alignment**. *If applied to the coordinates of all atoms in molecule $\mathscr{M}_q$, it uniquely specifies a molecular alignment.*

Thus, for a matching bipod alignment the elements of all corresponding atoms in the bipods have to be the identical and if the bipods are aligned "optimally" (using a translation and a rotation), the squared Euclidean distances between two corresponding atoms must lie below a threshold θ. The introduction of a threshold accounts for small variability in the geometry optimization and numerical inaccuracies. An illustrative example of a matching bipod alignment is given in Figure 1.

Given two ordered bipods, the least-squares 3D rotation matrix is uniquely determined. The following proposition shows how it can be calculated. Let us assume that the two central bipod atoms have already be aligned by a simple translation $\mathbf{T}$. To simplify the equations, we move to a new coordinate system, which has its origin at the position of the aligned central bipod atoms. Let us denote the matrix of the coordinate vectors of the three atoms in bipod B^q_{rst} in this new system by $\mathbf{X}$, a 3 × 3 matrix, where the rows denote the coordinates in 3D space and the columns the 3 atoms. Equivalently, let $\mathbf{Y}$ represent the coordinates of bipod B^p_{ijk}. Further, let $\mathbf{UWV}^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\mathbf{YX}^{\mathrm{T}}$, where $\mathbf{W}$ is a 3 × 3 diagonal matrix containing the (non-negative) singular values of $\mathbf{YX}^{\mathrm{T}}$, and where $\mathbf{U}$ and $\mathbf{V}$ are 3 × 3 orthogonal matrices.

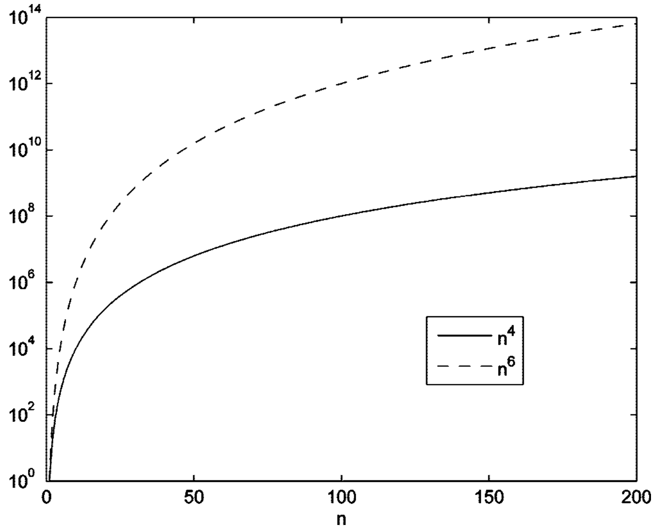


Figure 2. Runtime complexity as a function of molecule size The functions n^6 and n^4, corresponding to the order of the runtime complexity of random walk graph kernels and molecule kernels, respectively, are plotted for growing molecule size n.

Proposition 1 (Optimal Rotation Matrix). The rotation matrix $\mathbf{R}$ which corresponds to the optimal alignment of two given bipods B^p_{ijk} and B^q_{rst} in a least-squares-sense can be calculated by

$$\mathbf{R} = \mathbf{U}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{UV}^{\mathrm{T}}) \end{pmatrix}\mathbf{V}^{\mathrm{T}} \tag{4}$$

The proof can be found in the Appendix.

Definition 4 (Set of All Matching Bipod Alignments). The set of all pairs of matching bipods from two molecules $\mathscr{M}_p$ and $\mathscr{M}_q$ defines a set of molecular alignments, which is called the set of all matching bipod alignments Ω_{pq}.

Ω_{pq} is a finite subset of the (infinite) set of all possible molecular alignments. It is special in that bipod alignments correspond to the most general locally matching alignments which define a molecular alignment. Larger structures than bipods (like tripods, rings, or certain subfragments) would also allow a unique 3D orientation, but the possible number of alignments is a much smaller subset of Ω_{pq}. Smaller structures than bipods (i.e., pairs of atoms connected by bonds) do not allow a unique alignment, as they do not fix the relative rotation of the molecules around such a bond.

Definition 5 (Molecule Kernels). Let $w_i \in \Omega_{pq}$ denote the ith bipod alignment from the set of all possible matching bipod alignments between two molecules $\mathscr{M}_p$ and $\mathscr{M}_q$. Consider a similarity function $s(\mathscr{M}_p, \mathscr{M}_q, w_i)$, which assigns a similarity value to the molecules at a specific bipod alignment. Then a molecule kernel between the two molecules can be calculated as

$$k(\mathscr{M}_p, \mathscr{M}_q) = \max_{w_i \in \Omega_{pq}} s(\mathscr{M}_p, \mathscr{M}_q, w_i) \tag{5}$$

A molecule kernel corresponds to the maximum similarity value under all possible matching bipod alignments. Different molecule kernels can be defined, based on different choices of similarity function.

In this paper, we investigate two different molecule kernels (MK1 and MK2), which will be defined in the following.

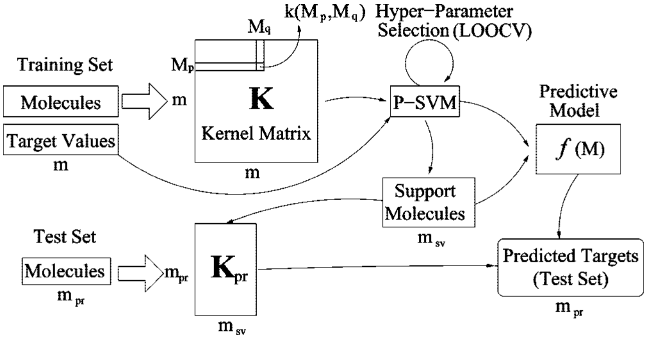


Figure 3. Process of model building and prediction using molecule kernels.

Table 1. Statistics on Datasets

dataset	no. of atoms (training set)	no. of bonds (training set)	no. of atoms (test set)	no. of bonds (test set)
Fontaine	59.7 ± 11.1	62.5 ± 11.6	60.0 ± 11.1	62.9 ± 11.7
ACE	43.6 ± 18.7	44.2 ± 19.6	40.1 ± 17.9	40.8 ± 18.8
AChE	56.4 ± 5.4	58.9 ± 5.7	56.4 ± 6.3	59.2 ± 6.8
BZR	36.3 ± 5.8	38.8 ± 6.2	36.4 ± 5.1	38.8 ± 5.4
DHFR	40.7 ± 6.9	42.8 ± 7.1	40.4 ± 7.7	42.7 ± 7.9

Table 2. Fontaine Dataset: Confusion Matrix of the Molecule Kernel Method Using MK1 on the Test Set

		Actual Label		
		+1	−1	Σ
predicted label	+1	TP: 87	FP: 6	93
	−1	FN: 1	TN: 51	52
	Σ	PC: 88	NC: 57	N:145

Assume we are given two aligned molecules $\mathscr{M}_p$ and $\mathscr{M}_q$, whose alignment is specified by $w_h \in \Omega_{pq}$. Let N_p be the number of atoms in $\mathscr{M}_p$, and N_q be the number of atoms in $\mathscr{M}_q$. Let us further assume that the atomic coordinates of the aligned molecules are given in the common reference frame as $\mathbf{x}_i^p$, $i = 1,..., N_p$ and $\mathbf{x}_j^q$, $j = 1,..., N_q$. Let $\mathscr{N}$ be the set of matching atoms

$$\mathscr{N} = \{A_i^p : \min_{j,\epsilon(A_j^q)=\epsilon(A_i^p)} (\mathbf{x}_i^p - \mathbf{x}_j^q)^2 < \theta, \quad i = 1, ..., N_p, j = 1, ..., N_q\} \tag{6}$$

Let the number of matching atoms N_{pq}^h be defined as the cardinality of the set $\mathscr{N}$ under the matching bipod alignment w_h. Using this number as similarity measure, we obtain a kernel which we denote by MK1:

Definition 6 (Unnormalized Atomwise Correspondence Kernel (MK1)).

$$k(\mathscr{M}_p, \mathscr{M}_q) = \max_{w_h \in \Omega_{pq}} N_{pq}^h \tag{7}$$

This kernel can also be normalized to the range of [0, 1], by using the Jaccard index as a similarity function. The Jaccard index is a similarity measure for two sets, in which the size of the intersection divided by the size of the union of the sets. This normalizes the atomwise correspondence, taking the size of both molecules into account, and yields a second molecule kernel, which we denote by MK2:

Definition 7 (Normalized Atomwise Correspondence Kernel (MK2)).

Table 3. Fontaine Dataset: Confusion Matrix of the Molecule Kernel Method Using MK2 on the Test Set

		Actual Label		
		+1	−1	Σ
predicted label	+1	TP: 87	FP: 7	94
	−1	FN: 1	TN: 50	51
	Σ	PC: 88	NC: 57	N:145

Table 4. Fontaine Dataset: Predictive Performance Measures of the Molecule Kernel Method on the Test Set

	MK1	MK2
sensitivity [TP/PC]	0.987	0.989
specificity [TN/NC]	0.895	0.877
balanced error rate [0.5(FN/PC + FP/NC)]	0.067	0.058
concordance [(TN + TP)/N]	0.945	0.952

$$k(\mathscr{M}_p, \mathscr{M}_q) = \max_{w_h \in \Omega_{pq}} \frac{N_{pq}^h}{N_p + N_q - N_{pq}^h} \tag{8}$$

Note that these kernel functions are symmetric with respect to the interchanging of the two molecules. The above definitions of atomwise correspondence kernels allow for some numerical inaccuracies and a certain variability in the atom's position via the threshold $\theta > 0$. This is the same hyperparameter which was already used in the spatial matching of the bipod atoms. It should be small enough that one atom of molecule $\mathscr{M}_q$ which matches one atom of molecule $\mathscr{M}_p$ is matching only that atom and no others. In our experiments, θ was always fixed at $\theta = 0.25$, which is small enough to ensure a unique assignment of corresponding atoms.

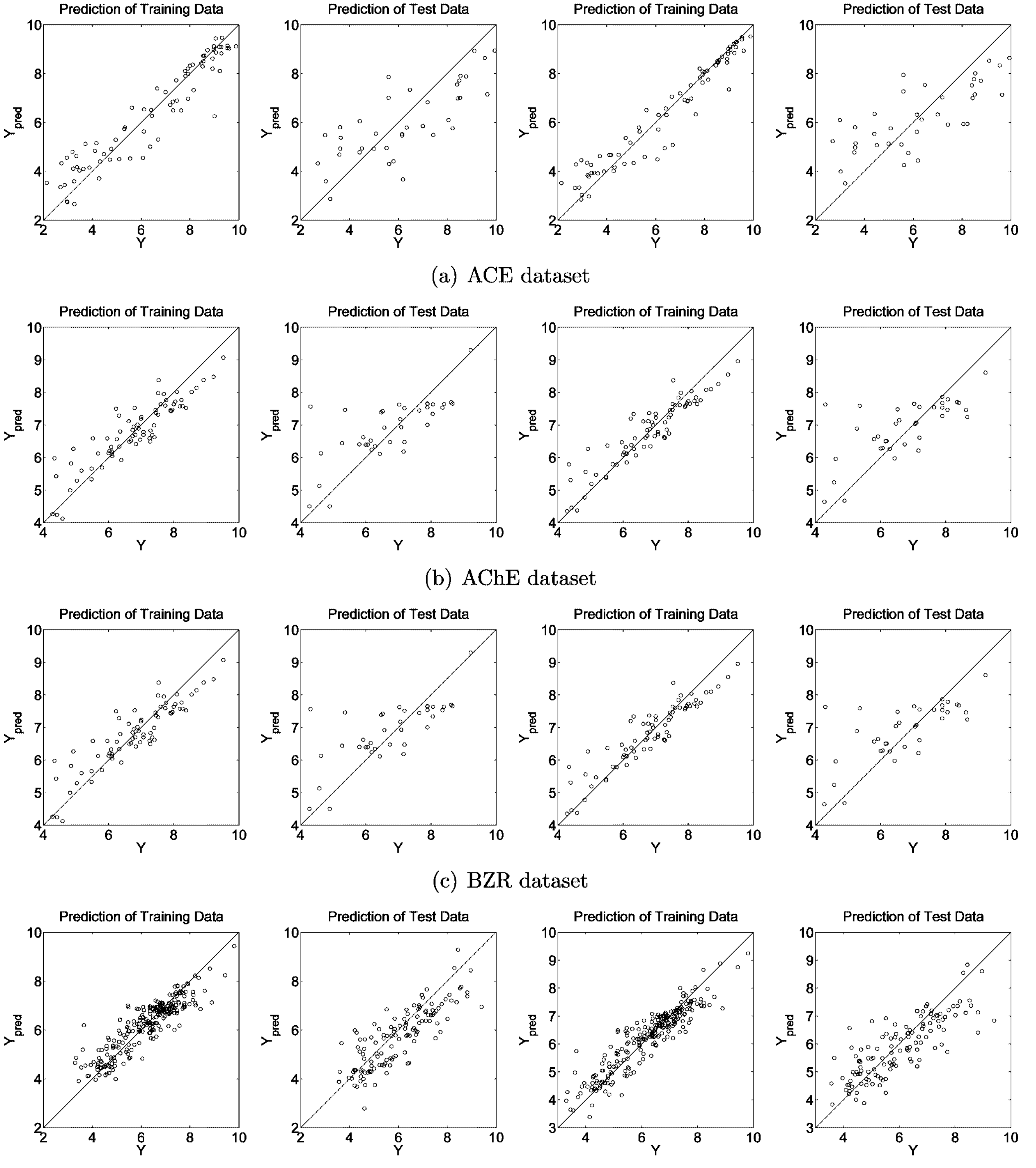
The similarity function determines which general aspects of the molecular representation (e.g., atom types, bonds types, chemophysical properties) are modeled by the kernel. The above-defined atomwise correspondence kernels are rather simple examples of molecule kernels, since they take only the spatial superposition of atoms from the same element into account. However, they do not account for the spatial match of atoms belonging to different elements, nor do the bond types enter the analysis. More elaborate types of molecule kernels can be constructed in the above framework by using similarity functions which take such information into account. In this context we want to mention that the adaptation to the chemical space of a particular class of compounds and a particular end point is not handled by a specific choice of molecule kernel, but by the learning machine used for prediction (see section 2.3).

2.2. Properties of Molecule Kernels. Molecule kernels are symmetric

$$k(B, A) = k(A, B) \tag{9}$$

and indefinite. The normalized atomwise correspondence kernel is bounded to lie between zero and one. Molecule kernels are guaranteed to find the largest similarity value in which at least two bipods match, since the alignments of all matching bipods are checked out. The runtime complexity of the molecule kernel applied to two molecules of the same size n depends on the runtime complexity of the calculation of the similarity function in eq 5.

As an example, for the case of the atomwise correspondence molecule kernels the runtime complexity



(a) ACE dataset

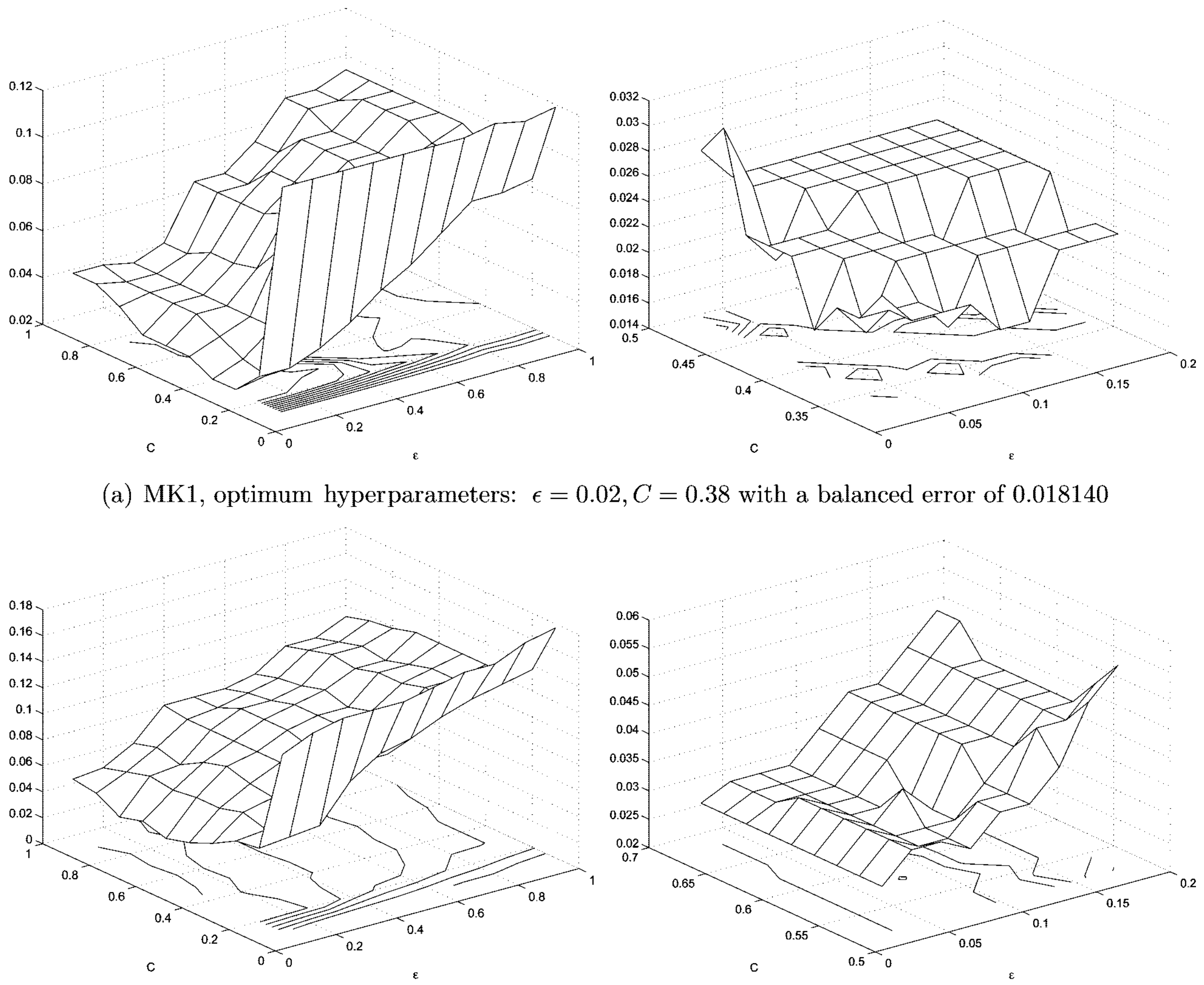
(b) AChE dataset

(c) BZR dataset

(d) DHFR dataset

Figure 4. Application of the molecule kernel method on the regression datasets: scatter plot showing the value of the regression target Y_{pred} predicted by the model against the true value Y. (left) MK1 (training and test set), (right) MK2 (training and test set).

scales with $O(n^4)$. This can be seen as following: Let us assume the worst case scenario, that all atoms in both molecules are from the same element and that each atom has a degree of d (i.e., d topological neighbors). Then each of the n atoms from molecule $\mathcal{M}_q$ can be centered on each of the n atoms of molecule $\mathcal{M}_p$, yielding n^2 combinations. For each of the atoms in such pair, there are $d(d-1)$ combinations of neighboring atoms which can be used to form ordered bipods. Therefore the number of all possible pairs of matching bipods is $(n^2d^2(d-1)^2)$, which is $O(n^2)$. The search for the nearest neighbors would require n^2 operations, leading to a total runtime complexity of $O(n^4)$. This is better than the time complexity of all-subgraph kernels,[5] whose calculation is NP-hard, and of product[5] and marginalized[6] graph kernels, whose runtime complexity scales with n^6. In order to illustrate how the



(a) MK1, optimum hyperparameters: $\epsilon = 0.02, C = 0.38$ with a balanced error of 0.018140

(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 0.52$ with a balanced error of 0.028241

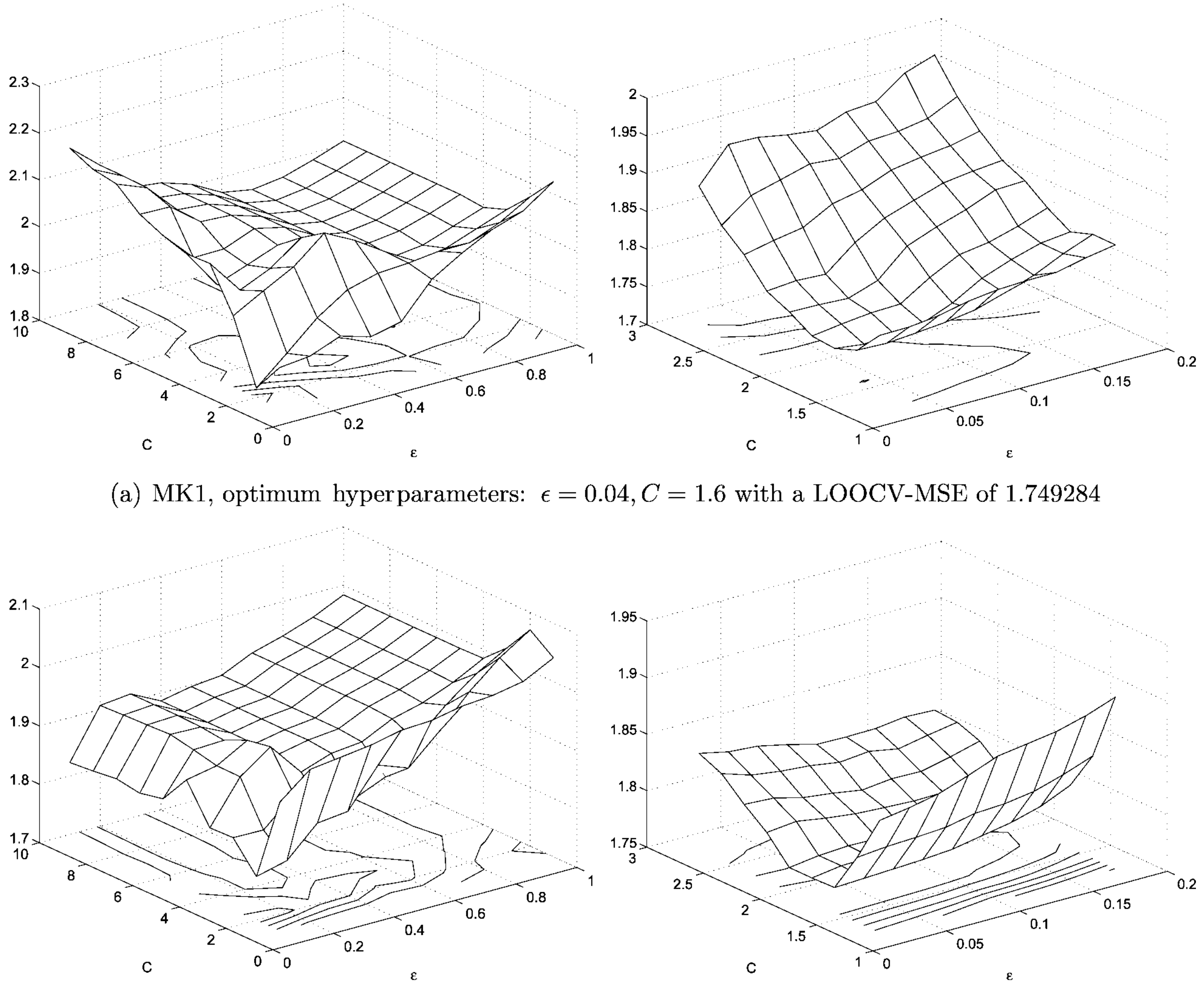
Figure 5. Application of molecule kernel method on the Fontaine dataset: hyperparameter grid search. (left) Phase 1, (right) phase 2. The LOOCV balanced error rate is shown as function of the hyperparameters C and ϵ.

runtime complexity of $O(n^4)$ compares to $O(n^6)$, n^4 an n^6 are plotted as a function of molecule size n in Figure 2.

Note that in real datasets, the two molecules are usually of different size and contain more than one element. Thus, the total number of matching bipods is usually quite small, and the nearest neighbors need only to be evaluated for atoms from the same element. This allows the efficient calculation of this molecule kernel, even for molecules containing several hundreds of atoms.

2.3. Model Building and Prediction. Kernel matrices resulting from molecule kernels are not necessarily positive semidefinite, and even if they are on the training set, they might not be on the test set. Therefore, standard kernel methods requiring positive-semidefinite kernel matrices cannot be employed for prediction. However, for the recently proposed[8] potential support vector machine (P-SVM) this restriction does not hold. The mathematical reason for this is that its (dual) optimization problem only depends on the kernel matrix $\mathbf{K}$ via $\mathbf{K}^T\mathbf{K}$. Therefore, the P-SVM can handle arbitrary kernel matrices, which do not have to be square or positive semidefinite. It can be used for both classification and regression tasks. For the mathematical formulation of the P-SVM see the work of Hochreiter and Obermayer.[8] The P-SVM objective function is a convex optimization problem, therefore a global solution exists, which can be found using an efficient sequential minimal optimization (SMO) algorithm[22] (see the work of Knebel et al.[19] for details). The result of the optimization is a set of so-called Lagrange parameters α_j, one for each training set example, which provide a measure how individual molecules in the training set affect the prediction. The P-SVM usually obtains a sparse solution, which means that many of these α_j values will be zero. Each nonzero α_j corresponds to a molecule, which we will call a "support molecule". The sign of the corresponding α_j serves as class indicator (for classification) or shows whether the respective molecule is associated with increase or decrease in activity (for regression). Its absolute value indicates how relevant a particular molecule is for the prediction.[8]

The process of model building and prediction using molecule kernels and the P-SVM is illustrated in Figure 3. First, the molecule kernel matrix $\mathbf{K}$ on the training set is calculated by evaluating the molecule kernel $k(\mathcal{M}_p, \mathcal{M}_q)$ for all pairs of compounds $\mathcal{M}_p$ and $\mathcal{M}_q$ from the training set. For a training set containing m compounds, this corresponds to an $m \times m$ symmetric matrix with ones on the diagonal, which requires the calculation of $m \cdot (m-1)/2$ scalar kernel values. For the calculation of the kernel matrix, only the



(a) MK1, optimum hyperparameters: $\epsilon = 0.04, C = 1.6$ with a LOOCV-MSE of 1.749284

(b) MK2, optimum hyperparameters: $\epsilon = 0.14, C = 2.0$ with a LOOCV-MSE of 1.769261

Figure 6. Application of the molecule kernel method on the ACE dataset: hyperparameter grid search. (left) Phase 1, (right) phase 2.

molecular structures and not the activity values (or class labels) are needed.

In the next step, a learning machine is trained using both the calculated kernel matrix **K** and the activity values (or class labels) in order to build a model. The P-SVM is used as learning machine, which requires the selection of two hyper-parameters (C and ϵ). This is done by minimizing the leave-one-out cross-validation (LOOCV) prediction error on the training set over a discrete grid of hyperparameter values. Then the P-SVM is trained at the optimal hyperparameters using the full training set, which yields the final model. The prediction function for a molecule $\mathscr{M}$ is specified via the set of α values and an offset b. It takes the form

$$f(\boldsymbol{M}) = \sum_{j=1}^{m} \alpha_j k(\mathscr{M}_j, \mathscr{M}) + b \tag{10}$$

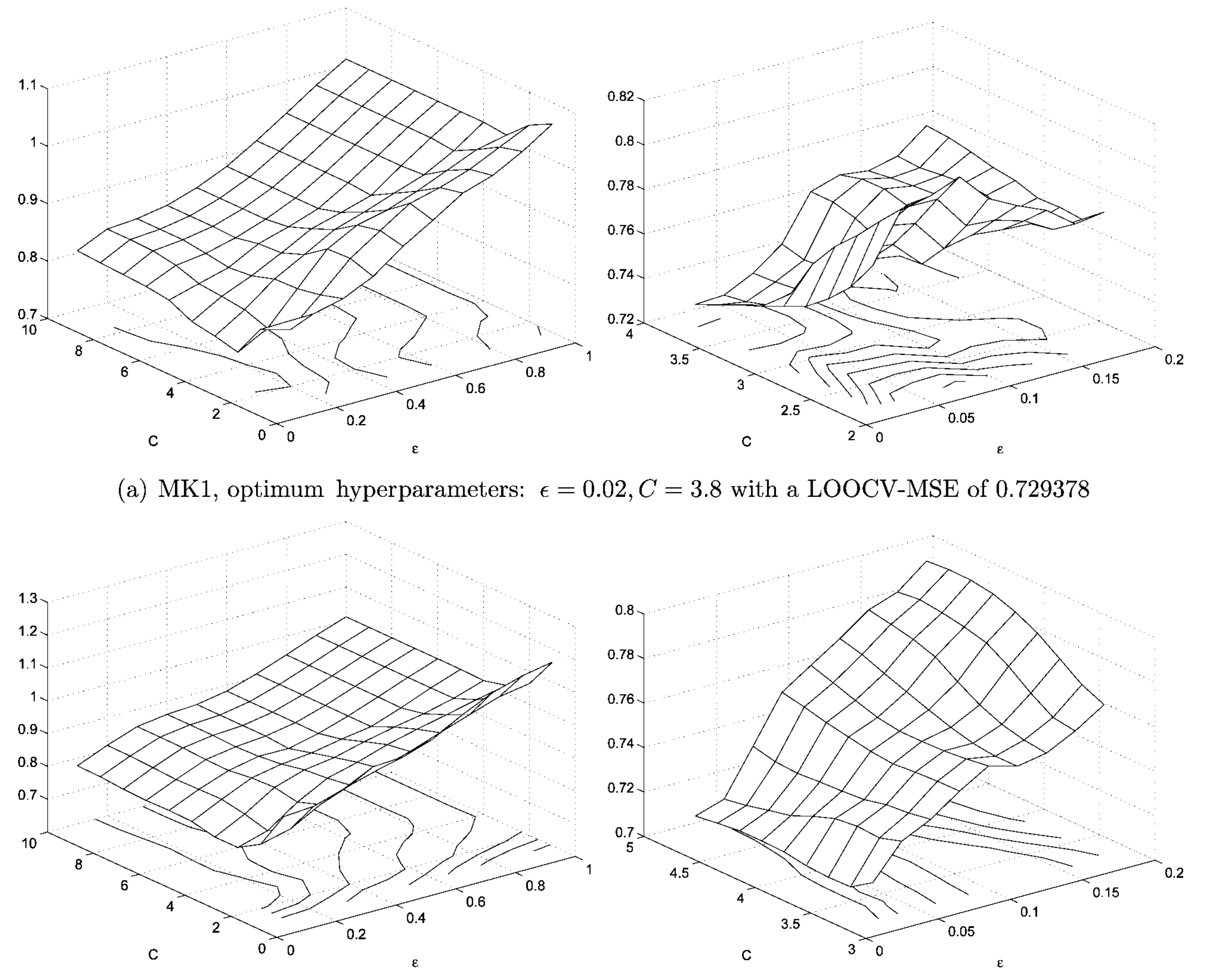
for regression and

$$f(\boldsymbol{M}) = \text{sign}\left(\sum_{j=1}^{m} \alpha_j k(\mathscr{M}_j, \mathscr{M}) + b\right) \tag{11}$$

for classification. Note that only the molecules $\mathscr{M}_j$ corresponding to nonzero α_j, the support molecules, are needed for prediction.

This model can then be used for prediction in the following way: First, the kernel matrix $\mathbf{K}_{pr}$ is calculated between the set of m_{sv} support molecules and the set of m_{pr} molecules for which we wish to predict the activity (or the class label). This is done by evaluating the molecule kernel for all pairs involving a member from each of these sets. The calculation of the $m_{sv} \times m_{pr}$ kernel matrix $\mathbf{K}_{pr}$ thus requires $m_{sv} \cdot m_{pr}$ kernel evaluations. In a regression setting, the P-SVM then yields the m_{pr} unknown activity values via eq 10. In a classification setting, predictions of class labels are obtained via eq 11.

While the choice of kernel influences which general properties of molecules are considered for evaluating the similarity, the adaptation to the chemical space spanned by the compounds in the training set and to a particular end point is handled by the P-SVM. Intuitively, this works as following: By generating a linear combination of the kernel values of all support molecules with the respective test molecule, the presence or absence of certain 3D substructures of the support molecules is used to assign a regression value or class label. However, this is done not explicitly, but implicitly by the support-vector machine based on the calculated kernel matrix.



(a) MK1, optimum hyperparameters: $\epsilon = 0.02, C = 3.8$ with a LOOCV-MSE of 0.729378

(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 3.6$ with a LOOCV-MSE of 0.710225

Figure 7. Application of the molecule kernel method on the AChE dataset: hyperparameter grid search. (left) Phase 1, (right) phase 2.

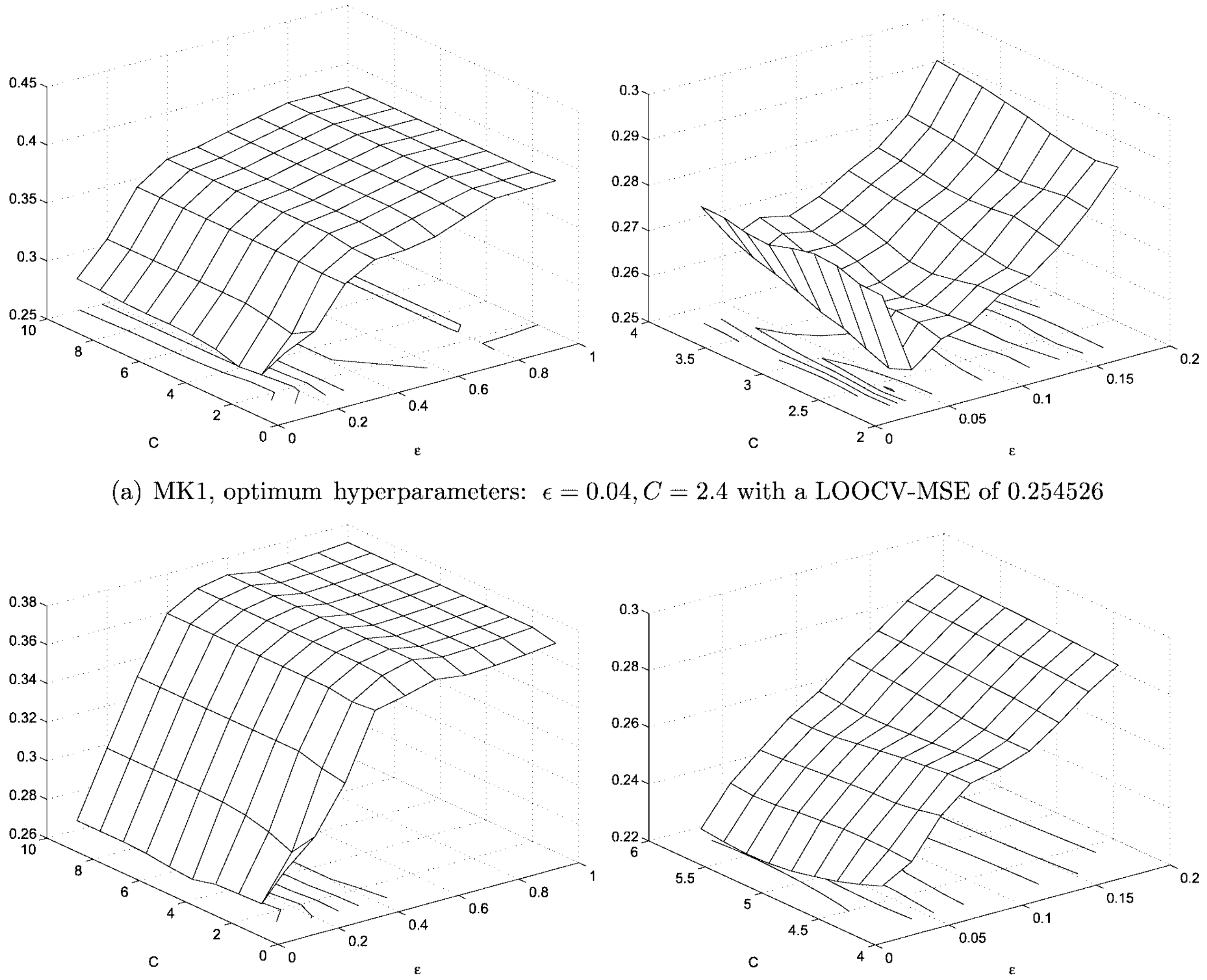
2.4. QSAR Datasets. A crucial requirement for comparing QSAR methods is the availability of enough and suitable data. This makes sure that in the selected training sets the molecular structures and activities do not deviate too far from the test set. This requires also some redundancy in the training data, such that the relevant patterns in the data can be recognized on the training data and used by the QSAR method (learning machine). It is usually assumed that the data points are independent and identically distributed (iid) samples from an underlying population distribution. The dataset must be large enough that both training and test set are likely to be representative for this distribution.

In the following, the QSAR datasets used in the method comparison will be described. These specific datasets have been chosen for the following reasons: (1) They are large enough to allow a comparison of methods, (2) they are publicly available[20] in structural form (sd files), (3) there are results reported in the literature for specific training and test sets, and (4) the activity values of the training and test sets have similar enough distributions to be representative for the population distribution. To check that the latter holds for the activity values of the regression datasets, the histograms of training and test set were compared, and it was verified that the test set did not contain activities in a range which not, or insufficiently, represented by the training set.

Details about the mean and standard deviation of the number of atom and bonds in each training and test set are given in Table 1. The binding affinity for the regression datasets is measured by pIC50 values, which correspond to the negative logarithm of the inhibitor concentration giving a 50% reduction of specific binding.

2.4.1. Fontaine Dataset (Classification). This dataset contains 435 molecules of the benzamidine family which is dichotomized into factor Xa inhibitors of low and high affinity. The dataset contains 156 compounds with low activity and 279 compounds with high activity. It was used[15] for 3D QSAR modeling with the Anchor-GRIND method. There, the dataset was randomly divided into a training set (290 compounds, 99 inactive, 191 active) and a test set (145 compounds, 57 inactive, 88 active).[15]

2.4.2. ACE Dataset (Regression). The ACE dataset comprises a set of 114 angiotensin converting enzyme (ACE) inhibitors was originally published by Depriest et al.[16] The activities range from pIC50 values of 2.1 to values of 9.9. It was used by Sutherland et al.[21] to compare a large variety



(a) MK1, optimum hyperparameters: $\epsilon = 0.04, C = 2.4$ with a LOOCV-MSE of 0.254526

(b) MK2, optimum hyperparameters: $\epsilon = 0.02, C = 5.4$ with a LOOCV-MSE of 0.223651

Figure 8. Application of the molecule kernel method on the BZR dataset: hyperparameter grid search. (left) Phase 1, (right) phase 2.

Table 5. Regression Datasets: Predictive Performance Measures of Molecule Kernel Methods MK1 and MK2 on Test Set

	dataset	ACE	AChE	BZR	DHFR
MK1	MSE	1.906	0.792	0.614	0.687
	PRESS	72.441	29.286	30.062	85.173
	SD	171.878	61.677	45.329	235.561
	r_{pred}^2	0.579	0.525	0.337	0.638
	dataset	ACE	AChE	BZR	DHFR
MK2	MSE	2.049	0.863	0.591	0.658
	PRESS	77.862	31.877	28.965	81.536
	SD	171.878	61.677	45.329	235.561
	r_{pred}^2	0.547	0.483	0.361	0.654

of QSAR methods, where it was split into a trainings set containing 76 compounds and a test set with 38 compound.

2.4.3. AChE Dataset (Regression). The AChE dataset contains 111 acetylcholinesterase (AChE) inhibitors whose pIC50 values lie in the range between 4.3 and 9.5. It has been assembled by Sutherland et al.,[21] who used it in their QSAR method comparison and divided into a trainings set (74 compounds) and a test set (37 compounds).

2.4.4. BZR Dataset (Regression). The BZR dataset consist of 163 benzodiazepine receptor ligands, whose pIC50 values lie in the range between 5.5 and 8.9. A subset of it was used in ref 21 and subdivided into a training set of 98 compounds and a test set with 49 substances.

2.4.5. DHFR Dataset (Regression). The DHFR dataset consists of 397 dihydrofolate reductase inhibitors (DHFR) with pIC50 values for rat liver enzyme ranging from 3.3 to 9.8. It has been compiled by Sutherland et al.[21] From the original data, a training set of 237 compounds and a test set of 124 compounds were generated (and a further set of 36 inactive compounds which we did not use here).

2.5. Assessment of Generalization Performance. All methods used the same training−test set split for evaluating the predictive performance. The data points from the test set were not used at all for model building. This includes the choice of hyperparameters (parameters which are considered fixed during the learning of the other model parameters, e.g. parameters representing the model complexity or the smoothness of a function), feature selection, feature construction, model selection, or parameter fitting. A measure of the generalization performance of the build models was obtained by applying the method to the test set.

The measures of performance for classification problems are based on the numbers of false positives (FP), true positives (TP), false negatives (FN), true negatives (TN), size of positive class (PC), and size of negative class (NC). From these, sensitivity (TP/PC), specificity (TN/NC), balanced error rate [0.5(FN/PC + FP/NC)], and concordance (TN + TP)/(PC + NC) can be calculated. However, for the method comparison on the Fontaine dataset, only concordance can be used, since only this result is stated in the literature.[15]

For regression problems, the mean squared error (MSE) can be used, which is the average PRESS (predictive sum of errors), defined as

$$\mathrm{PRESS} = \sum_{i=1}^{N_\mathrm{test}} (y^{(i)} - y_\mathrm{pred}(x^{(i)}))^2 \tag{12}$$

In QSAR studies, usually $r_\mathrm{pred}{}^2$ is used instead,

$$r_\mathrm{pred}{}^2 = \frac{\mathrm{SD} - \mathrm{PRESS}}{\mathrm{SD}} \tag{13}$$

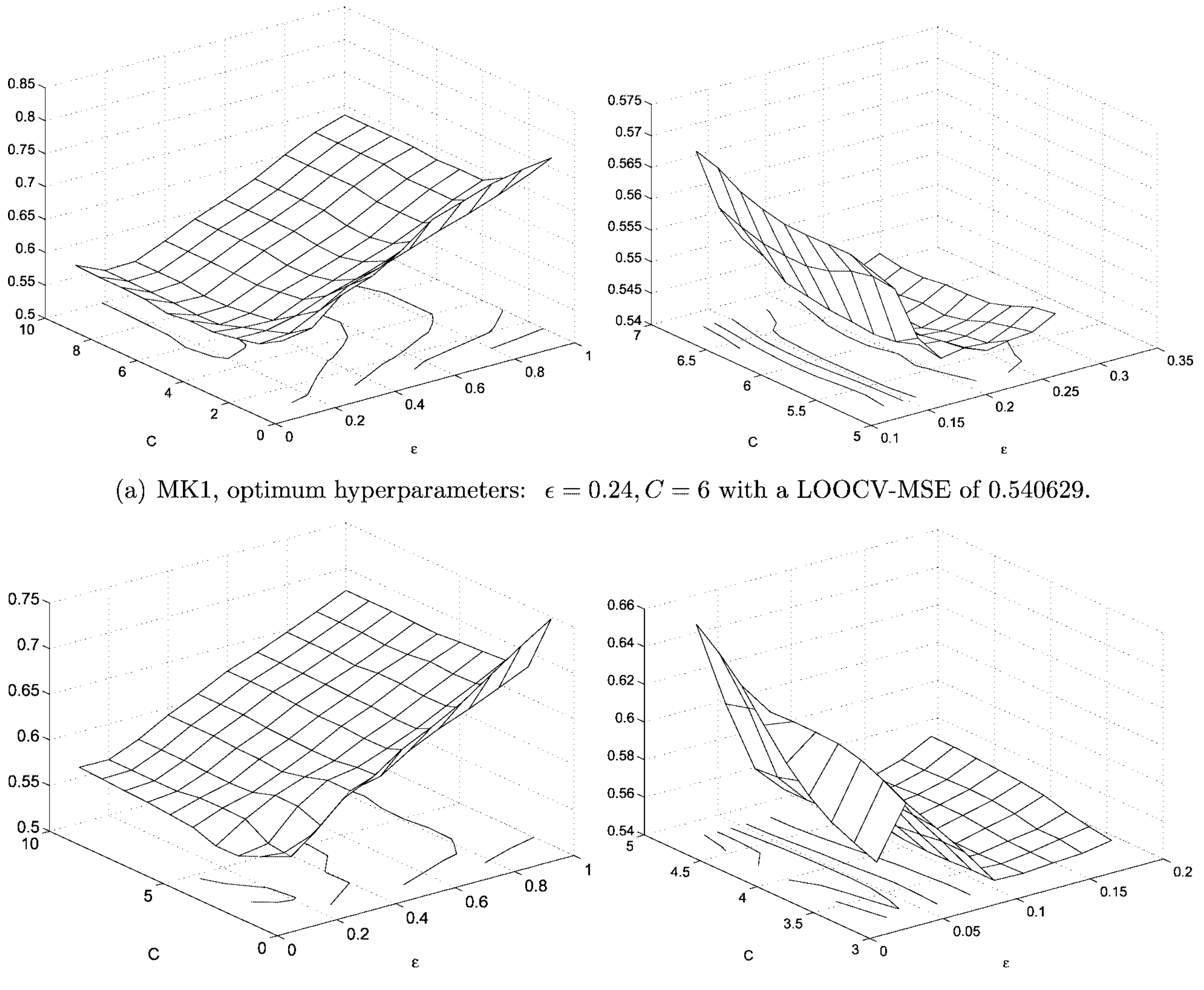
where SD is defined as

$$\mathrm{SD} = \sum_{i=1}^{N_\mathrm{test}} (y^i - \bar{y})^2 \tag{14}$$

with $\bar{y}$ denoting the average value of y. There are two different conventions in the QSAR literature with respect to the data points used in this average. In one, $\bar{y}$ is the average over the target values of the training set. In the other, $\bar{y}$ is the average over the target values of the test set. In order to allow a comparison of results, we here stick to the first convention which was used in ref 21.

3. RESULTS

All datasets were analyzed using the two molecule kernels defined in eqs 7 (MK1) and 8 (MK2). The model building process was carried out as described in section 2.3, using the P-SVM as predictor for both classification and regression.



(a) MK1, optimum hyperparameters: $\epsilon = 0.24, C = 6$ with a LOOCV-MSE of 0.540629.

(b) MK2, optimum hyperparameters: $\epsilon = 0.12, C = 3.6$ with a LOOCV-MSE of 0.543014

Figure 9. Application of the molecule kernel method on the DHFR dataset: hyperparameter grid search. (left) Phase 1, (right) phase 2.

Table 6. Classification Dataset[a]

dataset Fontaine	molecule kernel MK1	molecule kernel MK2	Anchor-GRIND (two-block model)	Anchor-GRIND (one-block model)
concordance	**0.95**	**0.95**	0.84	0.88

[a] The correct classification rate (concordance) is given for different methods. The results for the Anchor-GRIND method are taken from the literature.[15] The best result is printed bold font.

Table 7. Regression Datasets: r_{pred}^2 for Different Datasets and Methods[a]

dataset	MK1	MK2	CoMFA	CoMSIA basic	CoMSIA extra	EVA	HQSAR	2D	2.5D
ACE	**0.58**	*0.55*	0.49	0.52	0.49	0.36	0.30	0.47	0.51
AChE	**0.50**	*0.48*	0.47	0.44	0.44	0.28	0.37	0.16	0.16
BZR	*0.34*	**0.36**	0.00	0.08	0.12	0.16	0.17	0.14	0.20
DHFR	*0.64*	**0.65**	0.59	0.52	0.53	0.57	0.63	0.47	0.49

[a] All results except for the two molecule kernels are taken from the literature.[21] The best result for each dataset is printed bold font, the second best in italics.

The P-SVM implementation by Knebel et al.[19] was used in the analysis, which is available under the GNU General Public License from the Neural Information Processing Group at the Berlin Institute of Technology.[22] For the proposed method, there were two hyperparameters of the P-SVM which needed adjustment. The hyperparameter search was carried out systematically via a two-step grid search, using leave-one-out cross-validation on the training set. This process is documented in Appendix B. In the following, we refer to this whole model building method simply as the "molecule kernel method".

The results of all the other QSAR methods were taken from the literature. Exactly the same training and test sets were used by us and in the respective publications. For the classification dataset (Fontaine dataset), results were available for two variants of the Anchor-GRIND method, the one-block version (using Anchor-MIF descriptors) and the two-block version (using both Anchor-MIF and MIF-MIF descriptors). For details on the methods and their application to the dataset, see the publication by Fontaine et al.[15] and the references therein. For the regression datasets (ACE, AChE, DHFR) results were available for descriptors calculated with CoMFA, CoMSIA basic (with steric and electrostatic fields), CoMSIA extra (with additional hydrogen-bonding fields, hydrophobic fields, or both), EVA, HQSAR, and traditional 2D and 2.5D descriptors[23] using PLS as predictor. For details on the various methods and their application to the datasets, see the paper by Sutherland at al.[21] and the references therein.

First, the results for the molecule kernel method are presented. For the classification dataset, the confusion matrix for the prediction of the trained model on the test set is shown in Table 2 (MK1) and Table 3 (MK2). Table 4 lists the resulting performance statistics. For the regression datasets, the fitted predictions for both training and test set are plotted versus the actual activities as scatter diagrams in Figure 4. As performance statistics, the values for MSE, PRESS, SD, and r_{pred}^2 are listed in Table 5.

The above results were compared to the results achieved with the other QSAR methods. For the comparison of methods on the classification dataset, the correct classification rate (concordance) on the test set is given for each method in Table 6. The molecule kernel method achieves 5% prediction error rate which is much lower than the prediction error rates of the Anchor-GRIND methods (16% for the two-block model and 12% for the one-block model).

For the regression datasets, the value of r_{pred}^2 is given for all compared methods in Table 7. On the ACE dataset, the molecule kernel method MK1 performs best ($r_{\mathrm{pred}}^2 = 0.58$), followed by MK2 ($r_{\mathrm{pred}}^2 = 0.55$). All other QSAR methods except EVA and HQSAR reach r_{pred}^2 values around 0.5. On the AChE dataset, MK1 gives the best result (r_{pred}^2 of 0.50), and MK1 scores second best ($r_{\mathrm{pred}}^2 = 0.48$). Also good results are achieved by CoMFA ($r_{\mathrm{pred}}^2 = 0.47$), and the two CoMSIA approaches ($r_{\mathrm{pred}}^2 = 0.44$). Also on the BZR dataset, the two molecule methods show the best performance, $r_{\mathrm{pred}}^2 = 0.36$ (MK2) and $r_{\mathrm{pred}}^2 = 0.34$ (MK1), while all other method yield values of $r_{\mathrm{pred}}^2 \leq 0.2$. On the DHFR dataset, the molecule kernel methods perform best with an r_{pred}^2 of 0.65 (MK2) and 0.64 (MK1). They are followed by HQSAR ($r_{\mathrm{pred}}^2 = 0.63$), CoMFA ($r_{\mathrm{pred}}^2 = 0.59$), and EVA ($r_{\mathrm{pred}}^2 = 0.57$).

4. DISCUSSION

We have introduced a new kernel method for QSAR analysis, which is based on a novel similarity measure (the molecule kernel) and the use of the P-SVM as predictor. Instead of using descriptor vectors, the proposed molecule kernel method implicitly employs similarities in the 3D structures for prediction. In contrast to graph kernel QSAR approaches, which are based on counting paths and walks in the molecular graph, the molecule kernel method takes the 3D geometry of the molecular structure into account. The molecule kernels represent a measure of structural similarity between two given compounds. The P-SVM uses a linear combination of the pairwise similarities of the molecules in the training set to build a predictive activity model. This model will implicitly extract 3D substructures relevant for the prediction of a given end point. A necessary precondition for this to work is that the respective substructures are present often enough in the training set, so that the underlying pattern can be learned.

In this paper, two different molecule kernels where investigated. Both use similarity functions based on the spatial match of atoms belonging to the same element, which are, however, of different functional form. The method was compared on four regression and one classification dataset to several state-of-the art descriptor-based QSAR methods. These included approaches based on traditional 2D and 2.5D descriptor vectors as well as a variety of 3D QSAR methods (CoMFA, CoMSIA, HQSAR, EVA). In these experiments, the results of the two proposed molecule kernels using the P-SVM as predictor were consistently better than the results reported in the literature for the other QSAR methods included in the comparison. The empirical evidence suggests that the proposed descriptor-free method offers a promising alternative to existing descriptor-based approaches.

The two kernels we investigated are based on a different similarity measure and, therefore, yield different numerical values for the kernel matrix. However, their predictive performance was very similar on all datasets. This could be due to the fact that both kernels use the same level of molecular representation, the atomwise correspondence.

These simple kernels were already able to capture enough of the relevant structural similarities to yield excellent prediction performance, which even outperformed much more complex, force-field based approaches on several datasets. On the basis of the given framework, other molecule kernels can be constructed using more sophisticated similarity measures, which also model bond similarities or electrostatic and steric properties.

A problem of descriptor-based QSAR models is that the prediction function is expressed in the space of the descriptor variables; therefore, the mechanisms underlying the activity of a molecule are hard to investigate. In contrast to this, the prediction function of the molecule kernel method is expressed as a linear combination of structural similarities to a set of support molecules. The molecule kernel between a molecule and the set of support molecules can be used together with the respective Lagrange multipliers to gain insight into the structural properties which influence the activity of the molecule.

As an alignment-free method, molecule kernels do not require any user alignment of the molecules, like CoMFA and CoMSIA, nor the selection of anchor points, as in the Anchor-GRIND method. This saves time, makes the method usable by nonexperts and eliminates potential user bias. In contrast to other 3D QSAR methods, molecule kernels do not require the assumption that there is a single interaction mechanism at the same active site of a macromolecule. We expect that this feature of molecule kernels allows successful application of the molecule kernel method also in areas like toxicity prediction, where a large number of mechanisms is involved and where other 3D QSAR methods like CoMFA cannot be applied.

Abbreviations: P-SVM, potential support vector machine; CV, cross-validation, LOOCV, leave-one-out cross-validation; MSE, mean squared error.

ACKNOWLEDGMENT

This work was funded by the Bernstein Center for Computational Neuroscience Berlin (BMBF grant 01GQ0411).

APPENDIX A: PROOF OF PROPOSITION 1

Proof. The goal is to find the rotation matrix that minimizes the average squared distances between corresponding atoms in the two bipods. This means we would like to minimize the following cost function.

$$\begin{aligned} S &= \mathrm{Tr}((\mathbf{R}\mathbf{X}-\mathbf{Y})^T(\mathbf{R}\mathbf{X}-\mathbf{Y})) \\ &= \mathrm{Tr}(\mathbf{X}^T\underbrace{\mathbf{R}^T\mathbf{R}}_{=\mathbf{I}}\mathbf{X} - \mathbf{X}^T\mathbf{R}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{R}\mathbf{X} + \mathbf{Y}^T\mathbf{Y}) \\ &= \mathrm{Tr}(\mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T\mathbf{R}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) \end{aligned} \tag{15}$$

As a rotation matrix, **R** must be a special orthogonal matrix, which has the properties

$$\mathbf{R}^T\mathbf{R} = \mathbf{I} = \mathbf{R}^{-1}\mathbf{R} = \mathbf{R}\mathbf{R}^T \tag{16}$$

$$\det \mathbf{R} = +1 \tag{17}$$

Equation 16 gives rise to the set of constraints

$$\mathbf{R}\mathbf{R}^T - \mathbf{I} = \mathbf{0} \tag{18}$$

where **0** is a 3 × 3 matrix of zeros.

The cost function eq 15 together with the constraints (eq 18) yields the following Lagrangian,

$$\mathcal{L} = \mathrm{Tr}(\mathbf{X}^T\mathbf{X} - 2\mathbf{X}^T\mathbf{R}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}) + \mathrm{Tr}(\boldsymbol{\Lambda}(\mathbf{R}\mathbf{R}^T - \mathbf{I})) \tag{19}$$

where $\boldsymbol{\Lambda}$ is the matrix of (unknown) Lagrange multipliers. Note that $\boldsymbol{\Lambda}$ must be symmetric, since $\mathbf{R}^T\mathbf{R} - \mathbf{I}$ is symmetric.

The Lagrangian $\mathcal{L}$ should be minimized with respect to the elements of **R**

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathbf{R}} &= -2\,\frac{\partial\mathrm{Tr}(\mathbf{X}^T\mathbf{R}^T\mathbf{Y})}{\partial\mathbf{R}} + \frac{\partial\mathrm{Tr}(\boldsymbol{\Lambda}\mathbf{R}\mathbf{R}^T)}{\partial\mathbf{R}} \\ &= 2\mathbf{Y}\mathbf{X}^T + (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T)\mathbf{R} \\ &= -2\mathbf{Y}\mathbf{X}^T + 2\boldsymbol{\Lambda}\mathbf{R} = 0 \end{aligned} \tag{20}$$

where we made use of the fact that $\boldsymbol{\Lambda}$ is symmetric. It follows that

$$\boldsymbol{\Lambda}\mathbf{R} = \mathbf{Y}\mathbf{X}^T \tag{21}$$

This can be solved by singular value decomposition,

$$\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{22}$$

where **W** is a 3 × 3 diagonal matrix containing the (non-negative) singular values of $\mathbf{Y}\mathbf{X}^T$, and where **U** and **V** are 3 × 3 orthogonal matrices. Now we can determine $\boldsymbol{\Lambda}$ from eq 21 by using the fact that **R** must be orthogonal:

$$\begin{aligned} & (\boldsymbol{\Lambda}\mathbf{R})(\boldsymbol{\Lambda}\mathbf{R})^T = (\mathbf{Y}\mathbf{X}^T)(\mathbf{Y}\mathbf{X}^T)^T \\ \Rightarrow\ & \boldsymbol{\Lambda}\mathbf{R}\mathbf{R}^T\boldsymbol{\Lambda}^T = (\mathbf{U}\mathbf{W}\mathbf{V}^T)(\mathbf{U}\mathbf{W}\mathbf{V}^T)^T \\ \Rightarrow\ & \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T = \mathbf{U}\mathbf{W}\mathbf{V}^T\mathbf{V}\mathbf{W}\mathbf{U}^T \\ \Rightarrow\ & \boldsymbol{\Lambda}^2 = \mathbf{U}\mathbf{W}^2\mathbf{U}^T \\ \Rightarrow\ & \boldsymbol{\Lambda} = \mathbf{U}\mathbf{W}\mathbf{U}^T \end{aligned} \tag{23}$$

Inserting this result in eq 21 yields

$$\begin{aligned} & \boldsymbol{\Lambda}\mathbf{R} = \mathbf{Y}\mathbf{X}^T \\ \Rightarrow\ & \mathbf{U}\mathbf{W}\mathbf{U}^T\mathbf{R} = \mathbf{U}\mathbf{W}\mathbf{V}^T \\ \Rightarrow\ & \mathbf{U}^T\mathbf{U}\mathbf{W}\mathbf{U}^T\mathbf{R} = \mathbf{U}^T\mathbf{U}\mathbf{W}\mathbf{V}^T \\ \Rightarrow\ & \mathbf{W}^{-1}\mathbf{W}\mathbf{U}^T\mathbf{R} = \mathbf{W}^{-1}\mathbf{W}\mathbf{V}^T \\ \Rightarrow\ & \mathbf{U}^T\mathbf{R} = \mathbf{V}^T \\ \Rightarrow\ & \mathbf{U}\mathbf{U}^T\mathbf{R} = \mathbf{U}\mathbf{V}^T \\ \Rightarrow\ & \mathbf{R} = \mathbf{U}\mathbf{V}^T \end{aligned} \tag{24}$$

With eq 24 we have obtained an expression for optimal rotation matrix. However, so far the constraint eq 17, det **R** = +1, has not been used. Therefore the orthogonal solution could still describe a reflection (for det **R** = −1). In order to make sure that in fact a rotation matrix is obtained, the following modification should be used,[24]

$$\mathbf{R} = \mathbf{U}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{U}\mathbf{V}^T) \end{pmatrix}\mathbf{V}^T \tag{25}$$

which ensures that constraint eq 17 is fulfilled and **R** is indeed a rotation matrix.

APPENDIX B: GRID SEARCH FOR HYPERPARAMETER SELECTION

The hyperparameters C and ϵ of the P-SVM were selected using a two-phase grid search for the parameter values giving minimal leave-one-out cross-validation (LOOCV) error on each training set.

In the first phase, a grid search on a rough initial search grid was done, with grid points lying at $\epsilon \in \{0.1, 0.2,..., 1.0\}$, $C \in \{1, 2,..., 10\}$ for regression and $\epsilon \in \{0.05, 0.1,..., 1.0\}$, $C \in$

{0.05, 0.1,..., 1.0} for classification. In the second phase, a refined grid search was conducted around the minimum found for the first grid search. The new search grid was chosen as an equally spaced 9 × 9 grid centered around the found minimum, such that the start and end points were $^1/_5\times$ the previous grid step size away from the points neighboring the minimum in the previous grid.

The result of the hyperparameter selection via a two-phase grid search on the training set of the classification dataset (Fontaine) is shown in Figure 5. The predictive balanced error rate is shown as a combined surface and contour plot over the hyperparameters C and ϵ of the P-SVM for the first phase (left) and second phase (right) of the grid-search. For the regression datasets (ACE, AChE, BZR, and DHFR), the results of the hyperparameter search, conducted using a two-phase grid search on the respective training sets are shown in Figures 6−9. The combined surface and contour plots depict the mean squared predictive error as a function of the hyper-parameters C and ϵ of the P-SVM.

REFERENCES AND NOTES

(1) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(2) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.

(3) Schoelkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(4) Bakhir, G. H.; Hofmann, T.; Schoelkopf, B.; Smola, A. J.; Taskar, B.; Vishwanathan, S. V. N. *Predicting Structured Data*; MIT Press: Cambridge, MA, 2007.

(5) Gaertner, T.; Flach, P. A.; Wrobel, S. On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and the 7th Kernel Workshop*, Washington, DC, August 2003; Schoelkopf, B., Warmuth, M. K., Eds.; Springer: Berlin, Heidelberg, New York, 2003; pp 129−143.

(6) Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized kernels between labeled graphs. In *Proceedings of the International Conference Machine Learning*, Washington, DC; Morgan Kaufmann: San Francisco, CA., 2003; pp 321−328.

(7) Ralaivola, L.; Swamidass, S.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.

(8) Hochreiter, S.; Obermayer, K. Support vector machines for dyadic data. *Neural Comput.* **2006**, *18*, 1472–1510.

(9) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comp. Methodol.* **1990**, *3*, 615–633.

(10) These terms are used here informally to give an intuition, their rigorous definition will be given in the Methods section.

(11) Heritage, T. W.; Lowis, D. R. Molecular hologram QSAR In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, A. L., Reddy, M. R., Eds.; ACS Symposium Series 719, American Chemical Society: Washington, D.C., 1999.

(12) Cramer, R.; Patterson, D.; Bunce, J. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(13) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indexes in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological-activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.

(14) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.

(15) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687–2694.

(16) Depriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3DQSAR 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.* **1993**, *115*, 5372–5384.

(17) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; et al. A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143–152.

(18) The requirement of non-colinearity makes sure that a pair of matching bipods can later be used to define a unique spatial alignment (otherwise the alignment would allow arbitrary rotations around the colinear bipod axes).

(19) Knebel, T.; Hochreiter, S.; Obermayer, K. An SMO algorithm for the potential support vector machine. *Neural Comput.* **2008**, *20*, 271–287.

(20) All the used datasets are publicly available at http://www.cheminformatics.org/datasets/ (accessed July 7, 2008).

(21) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(22) http://ni.cs.tu-berlin.de/software/psvm/index.html(accessed July 7, 2008).

(23) Sutherland et al.[21] use the term 2.5D descriptors to distinguish descriptors which involve straightforward calculations like molecular volume from descriptors based on force-field calculations.

(24) Challis, J. A Procedure for Determining Rigid Body Transformation Parameters. *J. Biomech.* **1995**, *28*, 733–737.

CI800144Y