Atomic Local Neighborhood Flexibility Incorporation into a Structured Similarity Measure for QSAR

Nikolas Fechner,* Andreas Jahn, Georg Hinselmann, and Andreas Zell

Center of Bioinformatics (ZBIT), University of Tübingen, Tübingen, Germany

Received September 9, 2008

In this work, we introduce a new method to regard the geometry in a structural similarity measure by approximating the conformational space of a molecule. Our idea is to break down the molecular conformation into the local conformations of neighbor atoms with respect to core atoms. This local geometry can be implicitly accessed by the trajectories of the neighboring atoms, which are emerge by rotatable bonds. In our approach, the physicochemical atomic similarity, which can be used in structured similarity measures, is augmented by a local flexibility similarity, which gives a rough estimate of the similarity of the local conformational space. We incorporated this new type of encoding the flexibility into the optimal assignment molecular similarity approach, which can be used as a pseudokernel in support vector machines. The impact of the local flexibility was evaluated on several published QSAR data sets. This lead to an improvement of the model quality on 9 out of 10 data sets compared to the unmodified optimal assignment kernel.

INTRODUCTION

The prediction of biological activity of small molecules against a specific target protein is an important step in the drug discovery pipeline. Models with good prediction capabilities make it possible to reduce the number of compounds that actually have to be screened. Despite the long history[1,2] beginning with the works of Hansch et al.[3] the field of quantitative structure−activity relationships (QSAR) is still active. The general approach can be divided in two steps: Encoding of the molecule into a computer suitable format and learning the relationship between the compound representation and the property to predict. Because even the best model generation algorithm depends on a suitable representation, the encoding of relevant molecular properties is an important topic in the field of QSAR research. The need to include as much information as possible into the representation has led to a huge number of molecular descriptors.[4]

The first QSAR approaches only regarded the topology of the molecules,[1,2] but it is obvious that the geometry of a ligand molecule is crucial for its interaction with the receptor. Several concepts, which take the 3D information into account, have been developed and can be roughly divided into molecular interaction field approaches (MIF) and geometry or shape encodings. The MIF idea places the molecule in a field of probes, which measure the interaction potential at this position in the space. Popular methods are, for instance, the comparative molecular field analysis[5] (CoMFA) and GRID.[6] Examples for the geometrical or shape encoding are the molecular shape analysis[7] or the shape recognition algorithm by Ballester et al.[8] Most of the three-dimensional QSAR approaches require that the molecule already has 3D coordinates at the beginning of the algorithm. This assumption avoids the questions of how to get these coordinates and whether they are related to the right (i.e., biologically active) conformation. Cramer[5] even motivated his selection of the steroid data as a benchmark problem by the rigid backbone structure that makes consideration of the conformational flexibility unnecessary. Several methods require a precalculation of the 3D structure,[9−13] often done using CORINA.[14] This has the drawback that potentially biologically irrelevant geometries are compared, making the predictions of the 3D QSAR approaches questionable.[15] To overcome this, there is the possibility to enumerate a representative part of the conformational space by a conformational sampling. These approaches, often referred to as 4D approaches, are very time-consuming, because not only many conformations have to be calculated but also compared pairwise to each other.[13]

An alternative way to regard the conformational space of a molecule is to encode it by a measure of its flexibility. This idea lead to the development of molecular descriptors for the structural flexibility. An intuitive way is to investigate which bonds of a molecule are rotatable and use their relative frequency as a measure for flexibility.[16] The flexibility of a bond can be obtained by using a set of rules like the work of Bath et al.[16] or by the development of a quantitative flexibility score.[17] Another approach to encode the molecular flexibility in a single descriptor value is the combination of graph indices to describe topological features, which influence the flexibility like the Φ *molecular flexibility* index proposed by Kier[18] or the *global flexibility index* by von der Lieth.[17] More recently a new flexibility descriptor based on statistical thermodynamics was introduced,[19,20] which describes the conformational flexibility of a chemical structure.

In this paper, we present a method to avoid an explicit conformational sampling by an approximation of the conformational space. Similar to flexibility descriptors, the idea is that the conformational space of a molecule can be encoded by the flexibility of structural elements (e.g., bonds, topological patterns). In contrast to the descriptor approach, our

* To whom correspondence should be addressed. E-mail: nikolas.fechner@uni-tuebingen.de.

10.1021/ci800329r CCC: $40.75
Published on Web 02/18/2009

concept does not lead to a single quantitative value (or a set of values) for the molecules flexibility but to a set of atomic flexibility environments. These environments are single atoms combined with their intramolecular neighborhood and the descriptions of the rotational degrees of freedom of the neighboring atoms. Therefore, the flexibility of a molecule is not reduced to a single numerical value but to a set of substructures annotated with their approximated contribution to the overall molecular flexibility. The approach is conceptually similar to distance geometry and allows to approximate the molecular conformational flexibility by the set of local flexibility patterns. These pattern sets are no numerical molecular descriptors and give therefore no quantitative measure of the flexibility of a structure. But they can be used to augment a topological similarity measure for molecules by an estimation of the similarity of their conformational spaces without needing 3D coordinates, structural alignments, or conformational searches. This procedure does not solve the conformational search problem but gives an indirect access to the problem by estimating the similarity of the conformational spaces of two molecules, broken down to local flexibilities.

Another important development are the kernel methods that emerged during the 1990s. The key difference of kernel methods, like the well-known support vector machines (SVMs), to the classical descriptor-based machine learning algorithms is that a fixed numerical representation of the molecules by descriptors can be avoided by the definition of specific kernel functions. These kernels can be all similarity measures, which fulfill certain requirements. Initially, SVMs were mainly used for training descriptor-based models, and the development of kernels that work directly on the molecular graphs gained attention.[21,13,15,21−26] Alternative kernel methods to support vector machines like Gaussian processes,[27] which represent a Bayesian interpretation of a kernel based algorithm, were used recently in the field of cheminformatics.[28,29]

The goal of our work is to present an atom similarity measure that is capable to describe the similarity of the local structural flexibility and to show that this can be used to augment a topological similarity measure to improve the consideration of the molecular geometry. It is faster than real 3D techniques because no structural alignment and conformational sampling is needed and is therefore suited for the integration into a molecule kernel. This has also the advantage that its effect can be evaluated quantitatively using QSAR techniques. For this purpose, it has been incorporated in the Optimal Assignment Kernel[23] and successfully applied to infer QSAR models with support vector regression on several data sets. The results show significant improvements of the model quality in 9 of 10 cases, compared to the original procedure. In addition, the leave-one-out cross-validation performance was compared to literature results indicating an improvement of the predictive power in most cases of our new approach. The method can be regarded as a new similarity measure; therefore, the similarity ranking was compared to other molecular similarity measures, which also regard the flexibility of molecules.

METHODS

Kernel-Based Machine Learning. Kernel methods,[30] like support vector machines,[31] have become an important machine learning approach in modeling QSAR.[32−35] Although these methods often perform better than classical techniques, like neural networks or decision trees, they cannot bring one of their key advantages into play, if they are used on the same numerical descriptor representation as the classical methods. This advantage, the so-called *kernel trick*, allows to exchange the dot product of the descriptor vectors $k_{\mathrm{dot}}(a, b) = \Sigma_i a_i \cdot b_i = a^T b$, $a, b \in \mathscr{R}^n$, which appears in the basic form of the prediction function $f(x) = \Sigma_i \alpha_i x^T x_i$, with $x, x_i \in \mathscr{R}^n \forall i$, by a novel function, representing an inner product in some not explicitly stated space.

This function does not have to be defined on a numerical space. Any function k: $\chi \times \chi \rightarrow \mathscr{R}$ that is symmetric and positive semidefinite can be used as a *kernel* or, more exactly, a Mercer kernel. The space χ can be adapted to the specific problem like strings[36] or, as in the scope of this work, (molecular) graphs.[12,13,15,21−25]

Thus, the *kernel trick* allows us to apply a kernel similarity that describes the similarity of two molecules in a chemically sensible matter. A kernel based model (e.g., trained by a SVM) then predicts a target value by calculating the weighted similarities of the unknown molecule to the molecules of the training set (or more precisely the support vectors).

The encoding of the local molecular flexibility into an atom kernel similarity measure can be used to incorporate the flexibility information into any structured kernel that can use labeled vertices (atoms). These conditions are met, among others, by the marginalized graph kernel[21] or the optimal assignment kernel,[23] which is examined in this work. Recently, it has been shown[37] that the optimal assignment kernel is not always positive definite (i.e., generates a positive definite similarity matrix). To address this problem, Fröhlich[38] proposed to subtract the smallest negative eigenvalue from the diagonal of the matrix (this is also discussed by Saigo et al.[39]).

Although this is a drawback of the optimal assignment similarity, it does not disqualify this measure for kernel-based machine learning techniques. Indeed there are pseudo-kernels like the sigmoid kernel that are proven to give rise to nonpositive semidefinite matrices but perform very well in practice.[30]

Local Flexibility in an Atom Similarity Measure. The method we present here allows to incorporate the flexibility of the molecular neighborhood of an atom into its numerical descriptor representation. Our idea is to describe the flexibility of the neighboring atoms by the part of the space in which they can be found in relation to a center atom (core) that is considered as stationary. The positions of the direct neighbors are fixed and completely described by the hybridization of the core atom and the bond length. These quantities can be considered as fixed for specific atom and bond types and can be obtained by many chemical expert systems.

When considering possible positions of an atom with a topological distance of two, to which we refer as a second degree neighbor, two cases have to be taken into account. In the case that the bond between the core atom and its neighbor is not rotatable, which corresponds to any bond that is not a single nonring bond in our implementation, only a countable number of positions is possible. In the nonring case, these are completely defined by the hybridization of the neighbor. For instance, if the neighbor is a nonring sp^2 hybridized carbon atom connected to the core atom by a

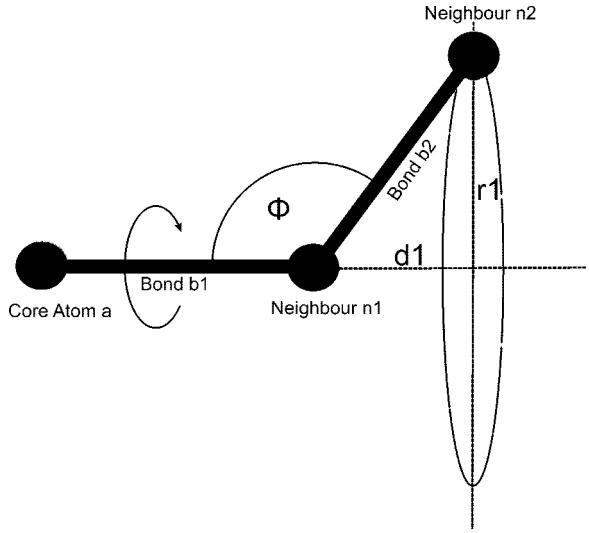


Figure 1. Schematic representation of the second degree neighbor orbit resulting from a rotatable first bond.

double bond its second degree neighbor has to be at one of two possible positions (at least if no further stereochemical information has to be considered).

If the bond is rotatable, the second degree neighbor can be placed anywhere on a circular orbit that is completely specified by the two bond lengths (core-neighbor1, neighbor1-neighbor2) and the angle between them (Figure 1). We refer to this type of local flexibility object as a first-order rotation.

This approach allows to encode the possible spatial positions of the neighbors of one atom as numeric atomic descriptors of the center atom (*core hybridization*, *bond length1*, *rotatability bond1*, *neighbor1 hybridization*, *bond length2*). Estimates of the angles can be taken from the hybridization of the *connecting* atoms. In our formulation, we decided to express the circular orbit of the neighbor atoms not directly by the hybridization and bond lengths but by using a geometric representation of it. This is because different atomic properties can lead to similar orbits without being directly similar. We therefore parametrized the circular orbit by the distance m_1 of the rotation center to the core atom and the radius of the orbit r_1. Note that w.l.o.g., our coordinate system, is chosen such that bond one lies on the x -axis.

$$r_1 = ||b_2|| \sin(\pi - \phi) \tag{1}$$

$$m_1 = ||b_1|| + d_1 \tag{2}$$

$$d_1 = ||b_2|| \cos(\pi - \phi) \tag{3}$$

where $||b||$ is the Euclidian L_2-norm.

This concept can be extended to a second-order rotation by incorporating the third degree neighbors, although the description of the possible locations is getting more complex (Figures 2 and 3). Several cases have to be considered.

In the general case (a) (Figure 2), the first and the second bonds are rotatable (i.e., single nonring bond). It is important to model the possible positions of neighbor n_3 explicitly with regard to the different influences of changes in the parametrization. In principle, n_3 rotates on an orbit around the extension of bond b_2 which also rotates around bound b_1. This link results in a roughly toroidal object (Figure 3). As a parametrization, we choose the distance m_1 between the core and the first rotation center, the radius of the second rotation r_2, the x-value and the y-value of the second rotation center m_2 and h. Note that other formulations are possible as well.

$$m_1 = ||b_1|| + d_1 \tag{4}$$

$$r_2 = \sqrt{(||b_3||^2 - d_2^2)} \tag{5}$$

$$h = (||b_2|| + d_2)\sin(\pi - \phi_1) \tag{6}$$

$$m_2 = m_1 + d_3 \tag{7}$$

with

$$d_1 = ||b_2|| \cos(\pi - \phi_1),$$

$$d_2 = ||b_3|| \cos(\pi - \phi_2)$$

and

$$d_3 = \frac{d_1 d_2}{||b_2||}$$

The probability of observing neighbor three at a specific point is not equally distributed. There are many more possible configurations that finally place n_3 somewhere in the center of the toroidal object than on the edges. Nevertheless, this simplification is acceptable from our point of view because other factors like steric collisions with other parts of the molecule have more impact on the possible positions of n_3 and are also not considered because of the restriction to the local flexibility. Note that it is not claimed that the real intramolecular flexibility is described in this way. This would make it necessary to incorporate interactions between atoms that are many bonds away from each other, which would be computationally infeasible as an atom similarity measure. Therefore, steric collisions, as well as nonequally distributed torsions around a rotation, are not considered. The latter is a simplification because in reality there are hardly any free rotatable bonds.

Some other cases have to be considered as well. In case (b), the first bond is not rotatable, and therefore, the third degree neighbor n_3 can only be placed on a circle in a similar way, but translated and rotated, like the second neighbor in the case of a rotatable first bond. An example for this is shown in Figure 4b.

The last case (c) that has to be taken into account is a rigid second bond. This leads also to a circular position space of neighbor n_3 (Figure 4). The rotation axis is the first bond, the radius, and the rotation center depend on the second and third bonds. Note that whenever the first or the second bond is rigid, cis/trans isomerism has to be considered.

The proposed method generates a flexibility feature set for each atom that can be used for defining a local flexibility similarity measure. But there is a difference to a numeric molecular or atomic descriptors, because the flexibility descriptors lack a unique ordering in the way that the ith value of the descriptor vectors corresponds to the same descriptor. Each atom can have a different number of neighbors with different rotational degrees of freedom, which makes it impossible to use vectorial similarity measures like the dot product or the RBF kernel directly.

One method to solve this problem is to regard these flexibility feature sets as a set of potential rotation possibilities (each of them corresponds to one neighbor) and to compute an optimal mapping of these rotation possibilities

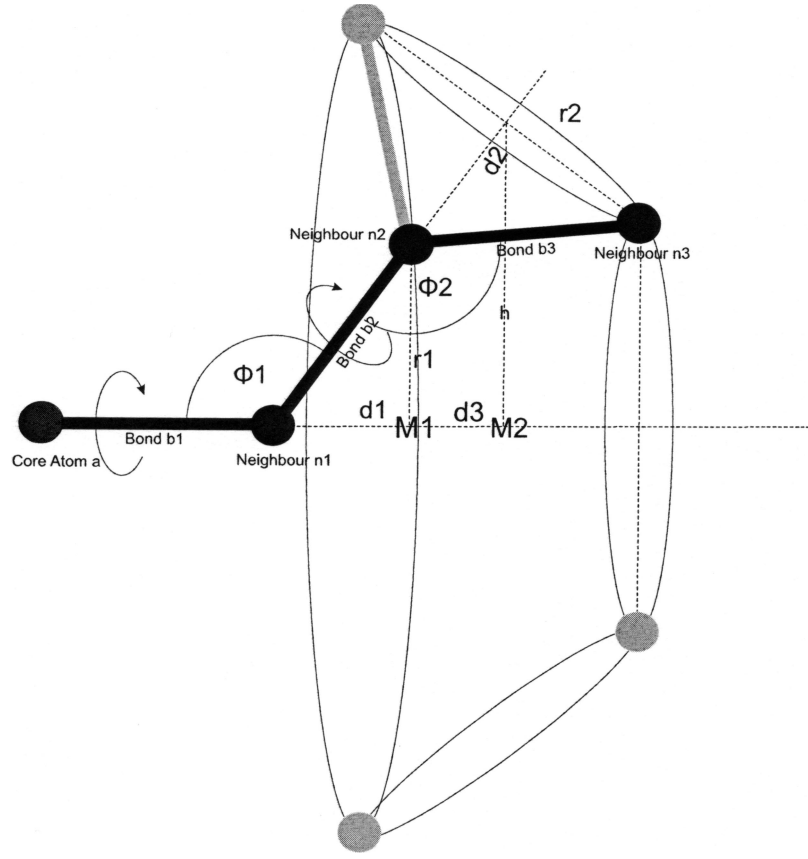


Figure 2. Schematic representation of the third degree neighbor orbit resulting from rotatable first and second bonds (case (a)).

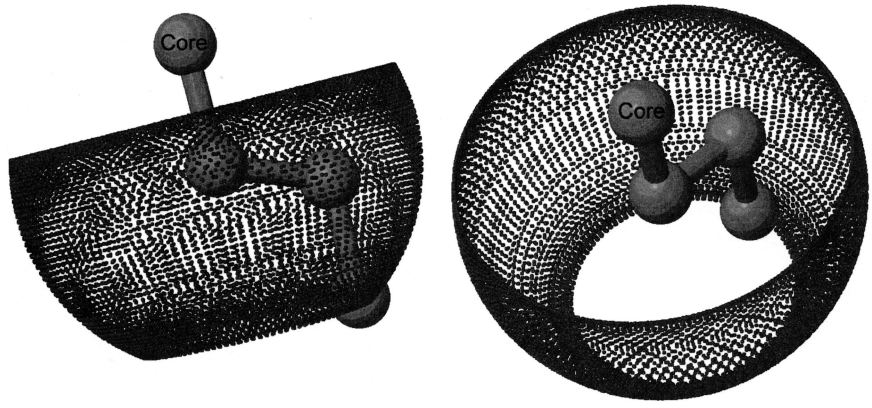


Figure 3. Visualization of the positional space of neighbor n_3.

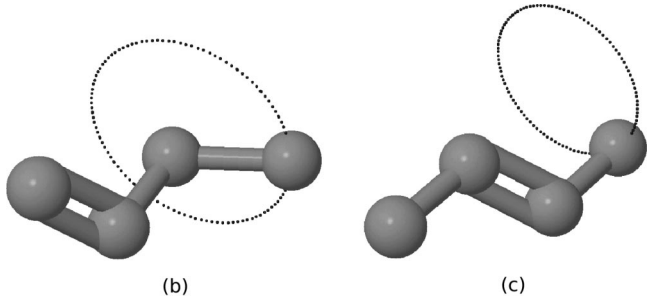
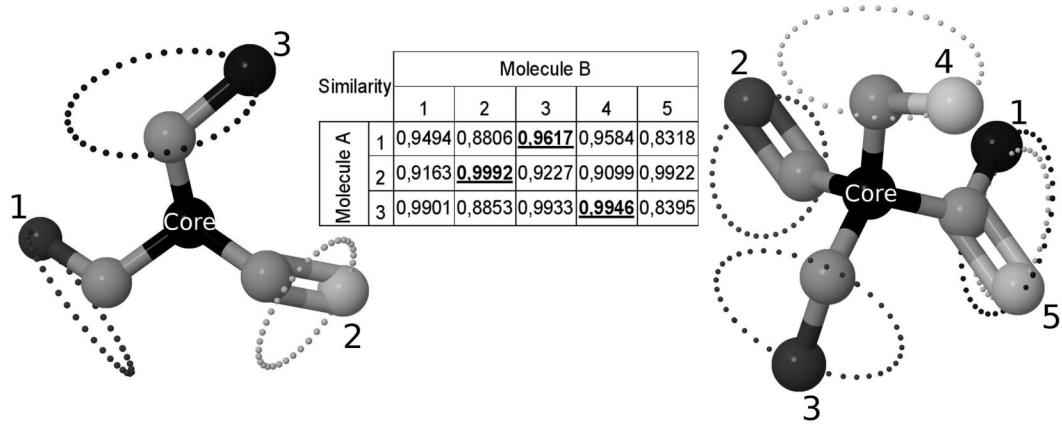


Figure 4. Visualization of the position space of neighbor n_3 with a rigid first (b) and second (c) bond.

of two atoms onto each other, regarding their flexibility relative to the core atoms. This can be done using a weighted bipartite graph matching approach, based on all pairwise similarity values which are calculated using a RBF kernel operating on the parameters of the flexibility objects. At this point, the use of the RBF kernel is possible because each rotation possibility has a predefined number of parameters. Figure 5 shows an example of the enumeration and the mapping of the rotation possibilities of the second degree neighbors. The table contains the calculated pairwise similarity values and acts as the input for the weighted bipartite graph matching algorithm. Because of the computational costs, this is computationally not feasible for each pairwise comparison. In our studies we decided to approximate the local flexibility similarity by solving the assignment problem with a greedy heuristic that always assigns the two most similar (RBF kernel) and yet unassigned rotations (i.e., neighbors) onto each other. This reduces the runtime complexity from $O(\max(|a|,|b|)^3)$ for the exact computation using the Hungarian algorithm[40] to $O(\max(|a|,|b|)^2)$ for the greedy approach, where $|a|$ and $|b|$ denote the number of rotation possibilities of the core atoms a and b. An empirical



Similarity		Molecule B				
		1	2	3	4	5
Molecule A	1	0,9494	0,8806	**0,9617**	0,9584	0,8318
	2	0,9163	**0,9992**	0,9227	0,9099	0,9922
	3	0,9901	0,8853	0,9933	**0,9946**	0,8395

Figure 5. Visualization of the assignment of the rotation possibilities of the second degree neighbors of the core atom. The table shows the calculated similarity values of the rotation objects. The underlined numbers indicate a mapping of the two objects.

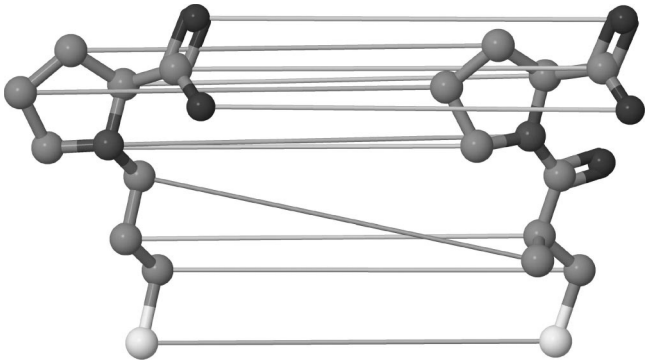


Figure 6. Example of an optimal assignment of the atoms of two molecules.

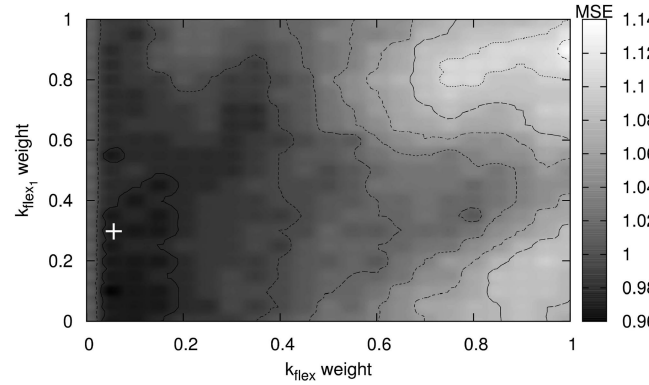


Figure 7. Effect of the parametrization on the performance on the COX2 data set. The white cross indicates the position of the default parametrization I.

comparison of the optimal assignment and the greedy scores on 5 million rotation similarity calculations yielded an average relative difference of about 0.56% for the first and less than 0.01% for the second order rotations, justifying the use of the greedy algorithm here.

The final local flexibility similarity score k_{flex} is obtained by the weighted summation of the assignment ($a_i \rightarrow b_{\pi(i)}$) scores for the rotations up to order two.

$$k_{\text{flex}_r} = \frac{1}{\sqrt{|a||b|}} \sum_{i=1}^{|a|} k_{\text{rot}_r}(a_i, b_{\pi_{\text{rot}_r}(i)}), \quad \text{w.l.o.g.} |a| < |b| \tag{8}$$

k_{rot_r} = similarity of the rth order rotation, k_{flex_r} analogous (9)

$$k_{\text{flex}} = \sum_{r=1}^{R} \rho_r k_{\text{flex}_r}, \quad r = 1, 2 \text{ in our studies, but not limited to} \tag{10}$$

$$\sum_{r=1}^{R} \rho_r = 1.0 \tag{11}$$

The normalization factor $[1/(\sqrt{|a||b|})]$ in the equation k_{flex_r} is needed to avoid a higher score for atoms with more neighbors ($|a|$ denotes the number of rotation elements which contain atom a). The weight coefficients ρ_r adjust the contribution of the rth order rotation to the kernel value. These are data dependent parameters which should be adapted to each QSAR problem. The weights of the different order rotations of a specific atom have to be normalized such that they sum up to one.

Extension of the Optimal Assignment Kernel. In the last section, we have introduced a concept to incorporate local flexibility information into an atom similarity measure. Most atom similarity measures used in chemoinformatics consist of two steps: the encoding of the atom as a vector of numeric descriptors and the comparison of the descriptor representations of two atoms. Since the comparison of two isolated single atoms is of no interest in most cases, the molecular neighborhood should be taken into account as well.

Fröhlich et al.[23] used a weighted recursive formulation of atomic similarity regarding both the atoms and the bonds (i.e., their representation as numerical descriptor vectors x and y for the atoms and $[x, n(x)]$ and $[y, n(y)]$ for the bonds). The local atom similarity kernel $k_{\text{las}}(x, y)$ is represented as a sum of the similarities of the center atom $k_a(x, y)$ and the similarity of the neighborhoods $R_l(x, y)$ of degree l.

$$k_{\text{las}}(x, y) = k_a(x, y) + R_l(x, y) \tag{12}$$

$$R_l(x, y) = R_0(x, y) + \frac{1}{|x||y|} \sum_{i,j} R_{l-1}(n_i(x), n_j(y)) \tag{13}$$

$$R_0(x, y) = \frac{1}{|x|} \max_{\pi} \sum_i k_a(n_i(x), n_{\pi(i)}(y)) \cdot k_b([x, n_i(x)], [y, n_{\pi(i)}(y)]) \tag{14}$$

The two kernel functions $k_a(x, y)$ and $k_b(x, y)$ can be any valid Mercer kernels that work on numeric feature vectors. In this study, we used the RBF kernel with the default variance from the original OAK implementation. The atomic similarity is computed by k_a, while k_b works on the bonds.

The permutation π is chosen such that the sum is maximized. This is accomplished by computing the maximum weight bipartite graph matching with the hungarian method.[40] An example of the mapping resulting from an OAK calculation is shown in Figure 6.

The local flexibility parameters, r_1 and m_1, for a first-order rotation and d_1, r_2, m_2, and h for a second order rotation are integrated into the local atom similarity kernel k_{las} used by the optimal assignment kernel by incorporating the flexibility similarity score k_{flex}. This is done by changing the atom similarity kernel k_a, which primarily uses physicochemical descriptors, to

$$k_a^+ = w_1 k_a + w_2 k_{flex} \quad (15)$$

$$w_2 = 1.0 - w_1 \quad (16)$$

This formulation would yield a valid kernel because of the closure properties if both k_a and k_{flex} were kernels. But because the final optimal assignment step on the matrix k_a^+ is not a valid kernel function, it is necessary to subtract the smallest negative eigenvalue from the diagonal of the final similarity matrix.

We also considered the differentiation of stereoisomers using the neighborhood comparison. Although this does not depend on the flexibility of the local neighborhood it can be incorporated into the similarity calculation by computation of the signed volume of the core atom and its neighborhood. Different chiralities of the same atoms would lead to a different sign of the volume, which can be incorporated into the atom similarity measure. While this leads to a better discrimination of stereoisomers (for a quantitative example, see Supporting Information), which can be very important in pharmaceutical tasks, we excluded this from the final experiments because the consideration of chirality is only sensible if the two core atoms and their direct neighbors are identical. In our test cases, this was very rare and so the chirality consideration only lead to a computational overhead without improving the model quality.

Empirical Determination and Default Values of the Parameters. The parameters ρ and w determine the influence of the different rotation objects (eq 10) and the weight of the original OAK (eq 16), respectively. In the case of the rotation objects, it is sufficient to determine ρ_1 because our experiments only account for the first- and second-order rotations, and so the value of ρ_2 is defined by the value of ρ_1 (eq 11). The same applies to the parameter w with eq 16. Therefore we have to define two parameters ρ and w which implicitly define the values of ρ_1, ρ_2, w_1, and w_2. To obtain these values, we performed a grid search for each data set in the range [0.05, 0.10, ..., 0.95] for ρ and w. Each parametrization in this range is evaluated with a 10-fold cross-validation with 25 multiruns. Visualizations of these grid searches can be seen in Figures 7 and 8 and in the Supporting Information.

Since the execution of a grid search is a time-consuming task, it is necessary to determine a default parametrization that achieves good results over all data sets. This is a challenging task because the best parametrization depends on the nature of the structures of a data set, which can be seen in Figures 7 and 8. On the basis of empirical observations on various data sets, we suggest two different default parametrizations with the objective to incorporate the molecular flexibility in such a way that it is suitable for the structures and makes sense from a chemical point of view. The two different parametrizations are visualized in Figures 7 and 8 by the white cross and show that this approach is a good approximation of the optimal parameters in this cases. Table 1 shows the values for the parameters of both default parametrizations.

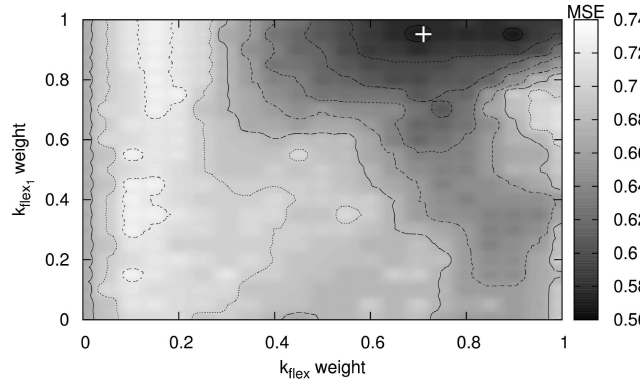


Figure 8. Effect of the parametrization on the performance on the BZR data set. The white cross indicates the position of the default parametrization II.

Table 1. Values of Both Default Parametrizations

	parameters			
name of configuration	w_1	w_2	ρ_1	ρ_2
default I	0.95	0.05	0.3	0.7
default II	0.3	0.7	0.95	0.05

The first default parametrization is suitable for data sets, which are mainly consisting of flexible substructures. The second one is optimized for data sets with molecules containing condensed ring systems. To perform an automatic adaption of the parameters for a data set to the values of one default parametrization, we developed a heuristic based on the frequency of condensed ring atoms. We refer to a condensed ring atom as an atom which belongs to more than one smallest ring system. The heuristic investigates the data set and determines if the molecules are build-up of more or less than 5% condensed ring atoms.

$$\text{default parametrization} = \begin{cases} \text{definition I} & \text{if } \dfrac{\text{no. atoms}_{\text{multiple rings}}}{\text{no. atoms}_{\text{data set}}} < 0.05 \\ \text{definition II} & \text{else} \end{cases} \quad (17)$$

Equation 17 shows the heuristic in which no. $\text{atoms}_{\text{multiple rings}}$ denotes the number of condensed ring atoms and no. $\text{atoms}_{\text{data set}}$ is the total number of atoms in a data set. This approach guarantees a fast method to determine one of the two default parametrizations suitable for the given data set.

RESULTS

Data Sets. The flexibility extension of the optimal assignment kernel (OAK)[23] is evaluated on freely available QSAR data sets. Most of the data sets were compiled for the comparison paper of Sutherland et al.[41] The single data sets were composed in several works.[42−49] The corticosteroid binding globulin data set was introduced in the original CoMFA paper by Cramer et al.[5] and has become a common benchmark for 3D-QSAR methods.[5,9,26,48,50,51] For com-

Table 2. Kendall Rank Correlation between Similarity Rankings of the data sets by Sutherland et al.

data set		OAK_{flex} vs Schröd.	OAK_{flex} vs Kier Flex.	Kier Flex. vs Schröd.	data set		OAK_{flex} vs Schröd.	OAK_{flex} vs Kier Flex.	Kier Flex. vs Schröd.
ACE	1	0.69	0.65	0.58	DHFR	1	0.39	0.28	0.35
	2	0.67	0.44	0.49		2	0.28	0.51	0.22
	3	0.84	0.56	0.58		3	0.32	0.15	0.30
AchE	1	0.27	0.44	0.16	GPB	1	0.48	0.20	0.15
	2	0.37	0.03	0.03		2	0.41	0.27	0.10
	3	0.32	0.43	0.23		3	0.08	0.58	0.07
BZR	1	0.52	0.08	0.04	THERM	1	0.57	0.60	0.55
	2	0.51	0.31	0.27		2	0.74	0.60	0.61
	3	0.41	0.01	0.11		3	0.73	0.64	0.58
COX2	1	0.22	0.17	0.06	THR	1	0.30	0.08	0.35
	2	0.48	0.32	0.10		2	0.40	0.47	0.25
	3	0.37	0.11	0.05		3	0.46	0.31	0.30

Table 3. Runtime Comparison[a]

	ACE	AchE	BZR	COX2	DHFR	GPB	THERM	THR
size	114	111	163	322	397	66	76	88
av no. atoms/molecule	24.5	28.7	23.0	26.9	28.1	21.9	30.5	41.2
av percent ring atoms	35%	71%	77%	66%	58%	36%	24%	51%
av single OAK calcd (ms)	5.3	6.7	4.8	6.2	6.0	4.7	7.2	11.7
av single OAK_{flex} calcd (ms)	7.9	8.5	5.6	8.5	8.2	6.9	12.3	18.0
OAK matrix calcd (s)	34.7	41.5	62.4	322.1	473.6	10.3	21.0	45.8
OAK_{flex} matrix calcd (s)	51.8	53.0	75.4	441.5	645.9	15.2	36.0	70.6
rel. difference (single calcd)	1.49	1.28	1.19	1.37	1.37	1.48	1.71	1.54

[a] Runtime comparison on a dual core P4 3.0 GHz, 2 GB Ram, Scientific Linux 5 (Kernel 2.6.18-92.1.10.el5), Java 1.6.0_04-b12.

parison, we also evaluated on the Prostaglandin F2 data set published by Hopfinger et al.[52] to be able to compare our method to a recently developed flexibility descriptor approach[19,20] Each data set is annotated with activity values to the specified target measured as log IC_{50} or pK_i.

Comparison with other Flexibility-Based Approaches. To clarify the difference between our approach and conventional topological based flexibility descriptors, like the molecular flexibility index Φ introduced by Kier,[18] we performed additional screening experiments with our method. The aim was to show that our approach can be delimited from other approaches that regard the molecular flexibility. For this purpose, we calculated the Kendall τ rank correlation coefficients of the rankings computed by the Kier flexibility index Φ, a sophisticated flexible alignment tool and our method. We used the flexible alignment tool of the Schrödinger toolkit Maestro,[53] which performs a ligand torsional search by varying the dihedral angles using ConfGen.[54] The final similarity value between a structure and the query is the normalized overlap volume of the aligned structures. To get a similarity value of two different Kier indices, we took the absolute value of the difference of both indices, which implies that a small value indicates a high similarity. To achieve a similarity value that is only based on the flexibility of the structures we modified the parameters w_1 and w_2 in such a way that our approach has no information of the physicochemical properties of the structures. For each of the eight data sets, compiled by Sutherland et al., the three most active structures serve as a query to screen the remaining part of the data sets. The correlation coefficients of the three different methods for all data sets and queries is shown in Table 2.

The correlation coefficients show that the rankings of the flexible alignment tool correlate better with the rankings of our flexibility approach than with the rankings of the Kier flexibility index. There also exists a correlation between the rankings of the Kier flexibility index and our method, which in the most cases is lower than the correlation with the flexible alignment tool. These results indicate that the flexibility approach of our method is a tradeoff between the time-consuming flexible alignment tool, based on the evaluation of several conformations, and the fast, but only topological based, Kier flexibility index. This demonstrates the originality of our type of encoding the flexibility of a structure and justifies an increased computation time in comparison to topological based descriptors.

The performance of a molecular similarity method is crucial, especially if it is used as a pseudokernel in support vector machines. To test the suitability of our approach with respect to the computational costs, we benchmarked it on several data sets. Table 3 shows an overview of the computation times for single kernel calculations as well as for the complete similarity matrix generation averaged over 10 runs. Incorporating the flexibility leads to a increased computation time by factor of 1.19−1.71. The difference depends on the relative amount of flexible bonds in the structures. An estimation of the latter is the percentage of atoms which are part of rings (and therefore not considered as flexible in our approach). The relation between a larger fraction of ring atoms and a decreased relative difference in computing time can be seen in Table 3.

Effect of the OAK Extension. The effect of the incorporation of the local flexibility into the optimal assignment kernel was evaluated on several data sets which have been used in previously published three-dimensional QSAR approaches. Table 4 shows the results of the original optimal assignment kernel[23] and the flexibility extended version. Our experiments were carried out using the freely available LibSVM 2.85[1] library[55] on precalculated kernel matrices. The SVM parameter (i.e., the soft margin cost parameter c)

Table 4. Results of 25 Multirun 10-Fold Cross-Validation on the Sutherland, Cramer, and Dervarics Data Sets

data set	size	OAK		OAK_{flex} (optimal)		OAK_{flex} (default)	
		MSE	Q^2	MSE	Q^2	MSE	Q^2
ACE[a]	114	**1.48 ± 0.61**	**0.73 ± 0.13**	1.52 ± 0.63	0.71 ± 0.13	1.56 ± 0.64	0.71 ± 0.15
AchE[b]	111	0.86 ± 0.36	0.48 ± 0.21	**0.80 ± 0.36**	**0.54 ± 0.20**	**0.80 ± 0.36**	**0.54 ± 0.20**
BZR[c]	163	0.67 ± 0.30	0.48 ± 0.19	**0.58 ± 0.26**	**0.53 ± 0.18**	**0.58 ± 0.26**	**0.53 ± 0.18**
COX2[d]	322	1.02 ± 0.31	0.51 ± 0.13	**0.97 ± 0.22**	**0.53 ± 0.12**	**0.97 ± 0.22**	**0.53 ± 0.12**
DHFR[e]	397	0.64 ± 0.19	0.71 ± 0.08	**0.60 ± 0.17**	**0.73 ± 0.08**	0.62 ± 0.18	**0.73 ± 0.07**
GPB[f]	66	**0.55 ± 0.33**	**0.59 ± 0.25**	**0.55 ± 0.35**	0.58 ± 0.26	**0.55 ± 0.35**	0.58 ± 0.26
THER[g]	76	1.64 ± 0.96	0.64 ± 0.21	**1.56 ± 1.00**	0.66 ± 0.22	1.57 ± 0.97	**0.68 ± 0.19**
THR[h]	88	0.47 ± 0.26	0.57 ± 0.25	**0.42 ± 0.24**	0.59 ± 0.24	0.43 ± 0.25	**0.61 ± 0.23**
CBG[i]	31	0.35 ± 0.32	0.78 ± 0.29	**0.22 ± 0.25**	**0.85 ± 0.23**	0.29 ± 0.28	0.80 ± 0.27
$PGF2_\alpha$[j]	38	0.31 ± 0.25	0.74 ± 0.27	**0.28 ± 0.22**	**0.75 ± 0.27**	0.31 ± 0.24	0.69 ± 0.30

[a] Angiotensin converting enzyme (class I). [b] Acetylcholinesterase (class I). [c] Benzodiazepine receptor (class II). [d] Cyclooxygenase II (class I). [e] Dihydrofolate reductase (class II). [f] Glycogen phosphorylase B (class I). [g] Thermolysin (class I). [h] Thrombin (Class I), [i] Corticosteroid binding globulin (class II). [j] Prostaglandin F2 receptor (class I).

Table 5. Significance of the Performance Gain Using the Flexibility Extension

hypothesis	p-values for the different data sets									
	ACE	AchE	BZR	COX2	DHFR	GPB	THERM	THR	CBG	$PGF2_\alpha$
$H_1 = \mu_{orig} < \mu_{flex(o)}$	0.163	0.998	~1.00	0.992	~1.00	0.710	0.962	0.974	~1.00	0.968
$H_1' = \mu_{orig} > \mu_{flex(o)}$	0.837	**0.002**	**<0.001**	**0.008**	**<0.001**	0.290	**0.038**	**0.026**	**<0.001**	**0.032**
$H_1 = \mu_{orig} < \mu_{flex(d)}$	**0.019**	0.998	~1	0.992	0.955	0.710	0.950	0.963	~1	0.487
$H_1' = \mu_{orig} > \mu_{flex(d)}$	0.981	**0.002**	**<0.001**	**0.008**	**0.045**	0.290	**0.050**	**0.037**	**<0.001**	0.513

[a] μ_{orig}, mean of MSE distribution of the unmodified optimal assignment kernel; $\mu_{flex(o)}$, mean of MSE distribution using the flexibility extension with optimal parameters; $\mu_{flex(d)}$, mean of MSE distribution using the flexibility extension with default parameters. The mean squared error is used as performance estimate and in each significance test the null hypothesis is $\mu_{orig} = \mu_{flex(o)}$ and $\mu_{orig} = \mu_{flex(d)}$. The bold values indicate the cases in which the null hypothesis can be rejected to a significance level of 0.05 ($\alpha = 0.05$).

was varied using a grid search and optimized regarding the average mean squared error in a 10-fold cross-validation which is repeated over 25 multiruns. This experimental setup ensures that the parametrization is chosen according to a large number of different splits of the data into a training and a test set and therefore does not adapt to a specific split as it would be the case in a single cross validation or fixed test set (at least if the latter has been chosen at random out of the basic data set). Thus the results should be a better estimate of the generalization power than an external test set if this would be chosen by a random split of the data. The common criticism that cross validation results are no good estimators of the real predictive power does not apply in this case because any overfitting resulting from a certain fold setting is averaged out by the repetition over 25 runs. Nevertheless the bias on the basic data set still remains and questions the real performance on completely new and independent data but this is a general problem of performance estimation and is not considered in this work. The comparison of the results of the different multiruns is ensured by choosing the random numbers that are used for the splitting using the same seed for the pseudorandom number generator. The bias introduced by this fixed seed can be neglected because of the large number of multiruns which are performed.

The incorporation of the flexibility extension was evaluated using two experimental setups. In the first experiment, the flexibility kernel parameters (i.e., the relative weight of the original OAK compared to the extension and the relative weight of the first-order rotation compared to the second-order) was optimized using the same parameter selection protocol like for the SVM parameters. The results of this best-possible kernel adjustment is shown in the middle columns of Table 4. This method has the disadvantage that the parameter choice is data set depended, and it is therefore necessary to conduct this parameter search for every data set on which the method should be applied. This is computationally more expensive than the tuning of the SVM parameter because the kernel parameters influence the kernel matrix which thus have to be recomputed for each setup. To overcome this drawback, we additionally performed experiments using the default parametrization presented in eq 7. The data sets were divided into two classes using the presented heuristic. Class I consists of the data sets, which contain less than 5% condensed ring atoms, Class II of the remaining.

To guarantee that the results, shown in Table 4, are not the consequence of outliers caused by the selection of advantageous folds, we performed a statistical analysis of the results. Using multiruns and a 10-fold cross-validation, we obtained enough observations of the mean squared error for each data set and method to approximate the distribution by a normal distribution.[56] With these normal distributions a one-sided alternative Wilcoxon rank-sum test was computed with R[57] for the different hypotheses shown in Table 5.

The optimal parametrization of OAK_{flex} yielded the best results regarding the MSE on all data sets, except ACE where the flexibility incorporation seems to introduce mainly noise and GPB where no significant difference between the different approaches can be seen. In four cases (AchE, BZR, COX2, GPB), the best parametrization corresponds to the default parametrization. The default parametrization performed comparably well in most cases with a relative deviation of the MSE compared to the optimal parametriza-

Table 6. Results of Comparison of Correlation Coefficients to Literature Results

	Sutherland et al.,[41] PLS with diff. descriptor approaches								
data set	OAK	OAK_{flex}	CoMFA	CoMSIA	CoMSIA (extra)	EVA	HQSAR	2D	2.5D
ACE	**0.72**	0.71	0.68	0.65	0.66	0.70	**0.72**	0.68	**0.72**
AchE	0.48	**0.54**	0.52	0.48	0.49	0.42	0.34	0.32	0.31
BZR	0.49	**0.55**	0.32	0.41	0.45	0.40	0.42	0.36	0.35
COX2	0.51	0.54	0.49	0.43	**0.57**	0.45	0.50	0.49	0.55
DHFR	0.71	**0.74**	0.65	0.63	0.65	0.64	0.69	0.51	0.53
GPB	0.61	0.56	0.42	0.43	0.61	0.58	**0.66**	0.31	0.46
THER	0.61	0.63	0.52	0.54	0.51	0.48	0.49	0.62	**0.66**
THR	0.56	0.59	0.59	0.62	**0.72**	0.47	0.50	0.62	0.52
av rank	2.67	**2.33**	5.00	5.33	3.22	5.77	4.11	5.66	4.66

	Dervarics et al.,[20] chirality sensitive flexibility descriptors			
	OAK	OAK_{flex}	PLS	MCA-PLS
$PGF2_{\alpha}$	0.66	**0.69**	0.36	0.63

	Kubinyi et al.,[9] 3D QSiAR			
	OAK	OAK_{flex}	regression (best)	PLS (best)
CBG	0.74	**0.80**	0.79	0.75

tion below 5% in all but two cases (CBG,$PGF2_{\alpha}$). It also remains better than the original OAK on all data sets in which the optimal parametrized OAK_{flex} improves the OAK except the $PGF2_{\alpha}$ data set where the MSE is the same as for the OAK but the correlation is worse. Using a significance level of 5%, all performance improvements according to the MSE of either the optimal or the default parametrization can be regarded as significantly better than using the original OAK (Table 5). The significance test shows that the results are significantly better than using the original OAK in eight out of ten data sets using an optimal parametrization and seven out of ten with the default setting.

Comparison to Literature Results. The integration of the local flexibility similarity into the optimal assignment kernel improved the prediction accuracy of the models in most cases. Table 6 shows the results of leave-one-out cross-validation (LOO-CV) experiments for both the original and the extended versions of the optimal assignment kernel compared to literature results. In all experiments, the best parameter choices that were obtained in the previous experiments were used for the flexibility integration. The LOO-CV procedure was applied to ensure an completely unbiased setup that can be directly compared to published results. This was not necessary in our previous experiments (Table 4) because we could ensure the comparability by using the same cross-validation seed for both OAK variants.

Most compared published results were obtained by using molecule encodings that regard the molecular geometry and are therefore of comparable complexity to our new flexibility incorporation. Sutherland et al.[41] compared several, mostly three-dimensional, molecular descriptors using partial-least-squares regression. The comparison of the average ranks of the molecular representations on the data set collection shows that the unmodified OAK performs better on average than the descriptors used in the original work. The flexibility extension further improved the average correlation and yields the best Q^2_{LOO} in three cases (AchE, BZR, DHFR), which is more often than any other method.

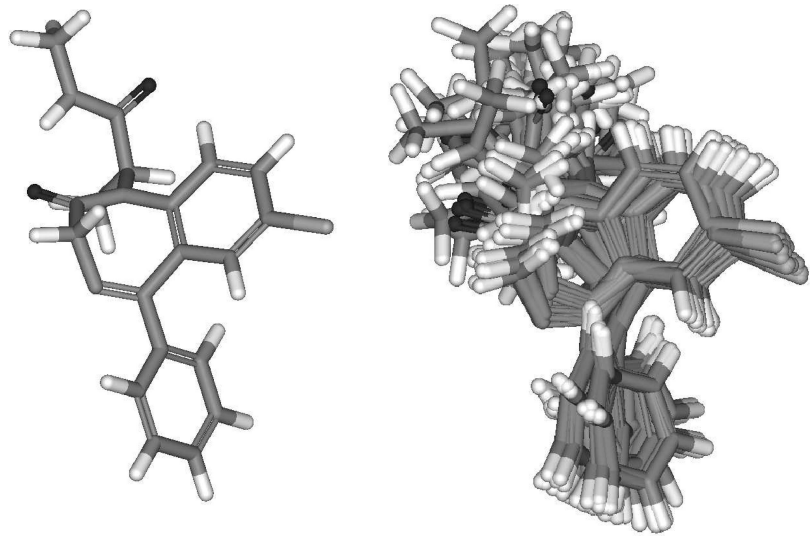


Figure 9. Example of a conformer sampling for a BZR ligand using MacroModel 9.5[54] (OPLS2005 forcefield, MCMM sampling, energy window 50 kJ/mol, 1 Å minimal RMDS). The partition into flexible and rigid structural parts are noticeable.

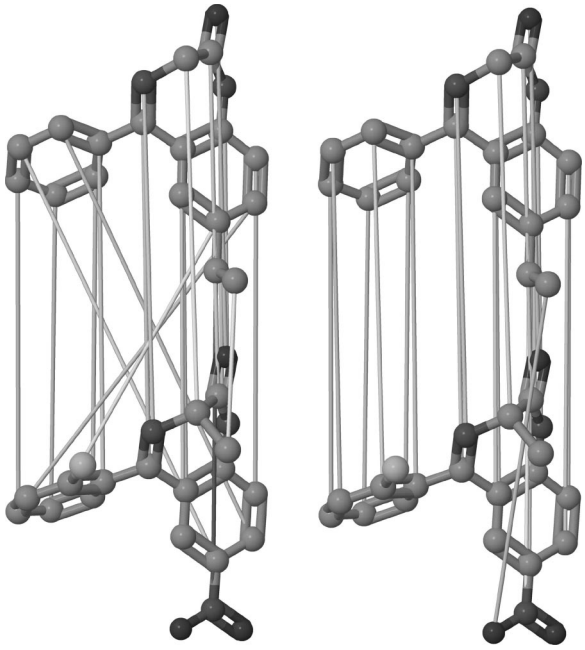


Figure 10. Improved overall topology conservation.

Dervarics and Martinek[20] improved a thermodynamically motivated flexibility descriptor[19] and evaluated it using several PLS variants on an endomorphin analogue and prostaglandin $F2_\alpha$ data set ($PGF2_\alpha$) of which the second one was chosen for the comparison. Both, OAK and its flexibility extension gave higher Q^2_{LOO} than the compared methods.

The steroid data set (CBG) was originally published by Cramer[5] but Kubinyi et al.[9] obtained better results using the 3D-QSiAR[9] approach. This gave a higher LOO correlation coefficient than the original OAK but not than the flexibility extended version of it.

DISCUSSION

The results show that the incorporation of the flexibility information leads to a statistically significant improvement of the predictive performance of the OAK, in most cases independent whether the error or the correlation is regarded. This holds as well for the optimal parametrization as for the default setting of the kernel parameters. Only in the case of ACE the extension introduces mainly noise and performs therefore worse than the OAK. Nevertheless the examination of the significance (Table 5) shows that in this case the null-hypothesis cannot be rejected on a 5% significance level.

Compared to other published studies on the same data sets our results show a very good overall performance. Although it did not obtain the best results in all cases, the flexibility extended OAK was on average the second best solution compared to the approaches of Sutherland.[41] This is about one rank better in average than the best overall method $CoMSIA_{extra}$ proposed in the literature and makes the OAK_{flex} method a suitable default choice. On the remaining data sets, it gave better results than the literature.[9,20] However, the goal of this paper is not to present a completely new method to predict activities toward the targets in the data sets but to show that the proposed method of incorporation of the flexibility information does improve the performance of a structured kernel at a moderate cost. These results also justifies the increased computing time because it is still not as complex as some of the other presented methods that depend on a structural alignment of each molecule to a reference compound (e.g., CoMFA, CoMSiA). These also require a conformational sampling of each molecule to find a conformer which can be aligned in a meaningful matter. These requirements of many alternative 3D QSAR methods justifies the runtime increase by our extension in our opinion. The absolute computing time of less than 10 ms in most cases for a single pairwise similarity remains extremely fast.

A closer inspection of possible causes of the improved performance reveals an unexpected result. The largest impact of the extension is not obtained on the most flexible data sets (e.g., angiotensin converting enzyme (ACE) ligands) but on data sets in which the ligand structures have both rigid parts, like ring systems, and flexible substructures. This is the case for the steroids in the CBG set and the compounds in the BZR set where the OAK_{flex} shows the most significant improvement (see, e.g., Figure 9). The possible cause for this is that the extension allows a more sensitive discrimination of atoms that are part of flexible substructures. This results in a bigger variance in the similarity scores that those parts will receive if mapped onto each other. Therefore, the relative contribution of the flexible parts to the similarity of two molecules is increased compared to the rigid parts. Thus this approach is suitable for describing the functional similarity of two molecules whose functional difference mainly depends on their flexible substructures but which also contains relatively large rigid parts that would influence the similarity more strongly in the original OAK. This is the case if the molecules have a shared or at least similar rigid scaffold like it is the case for the CBG data set. In this case,

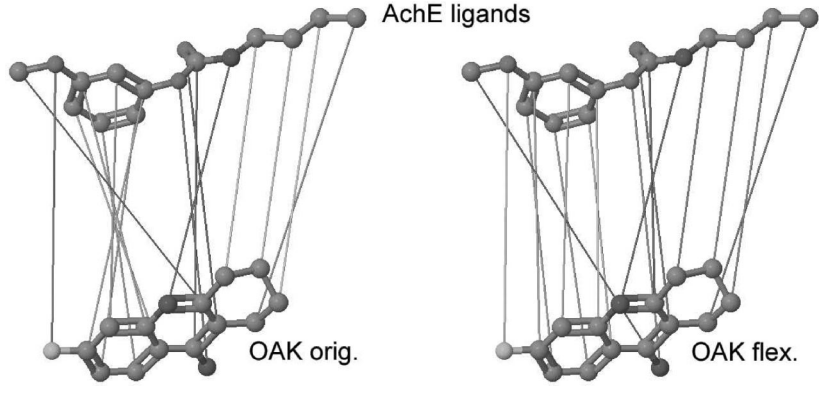


Figure 11. Improved ring topology conservation.

the extension increases the contribution of the flexible side chains and therefore better covers the important structural differences.

Another effect of the flexibility extension can be expected to be useful in most structural classes independent whether they have rigid scaffolds. The visualization of the computed optimal assignments calculated by both the original and the extended method reveals that the extension gives a mapping which conserves the topological order of the molecules much better. In the examples shown in Figures 10 and 11, it can be seen that the extension improves the topological alignment of the molecules (Figure 10) and also prevents the topological distortion of ring systems (e.g., the benzyl group in Figure 11). This effect is suitable if the computed alignments are used for a rigid superposition of the molecules as it is done by the Kabsch algorithm.[58]

CONCLUSION

We have presented a novel procedure of regarding molecular geometry in structured similarity measures by defining a local flexibility atom similarity. This is done by a parametrization of the spatial positions where the neighbors of a specific atom can be placed. These features are further processed to give a measure for the similarity of the flexibility of the local neighborhoods of two atoms. The new method thus results not in a set of descriptors but in a set of atomic pairwise similarities. These local flexibility similarities can easily be integrated into any molecular similarity measure that incorporates pairwise atom similarities. The results indicate that the additional discrimination aspect leads to similarity scores, which in most cases are better suited for modeling quantitative structure−activity relationships. The additionally available information solely depends on the sets of local structural flexibilities and thus the improvement of the similarity measure must be caused by this and the resulting consideration of the implicitly defined conformational space. The extension therefore enables an approximation of the similarity of the conformational spaces of two molecules. This does not claim that our method solves the problem of conformational sampling, but it provides an implicit estimate of the conformational space that can be taken into account for a similarity computation.

To demonstrate the originality of our method compared to other similarity measures, which use molecular flexibility information, we further compared the similarity rankings obtained on our benchmark data sets using a topological flexibility descriptor,[18] a flexible alignment method,[53] and our algorithm. The resulting similarity rankings were compared using the Kendall rank correlation τ. The obtained correlations indicate that each of the three approaches is different enough to be not equivalently replaceable by one of the others.

We quantitatively evaluated our method by incorporating it into an optimal assignment similarity,[23] which can be used as a structured kernel after ensuring a positive definite kernel matrix and applying the eigenvalue fix, if necessary. The resulting kernel is used as an encoding of the molecules for the prediction of activities using support vector machines.

We evaluated the modification on several freely available QSAR data sets that have been used in previous works using 3D QSAR approaches.[5,19,41] The results show that the consideration of the local flexibility leads to a significant improvement of the prediction performance in most cases compared to the original OAK. The method was also compared to published results on this data sets. The extension of the OAK was in most cases better than the literature results and further improved the results obtained by the original OAK. A closer inspection of possible causes for this improvement leads to the assumption that this is the result of a better relative discrimination of the flexible parts of a molecule compared to the rigid parts. This improves the representation of the structural similarity in structural classes where the targeted property depends more on flexible substructures (e.g., side chains) than on rigid scaffolds.

ACKNOWLEDGMENT

This work has been partially funded by Nycomed GmbH, Konstanz, Germany. The authors would like to thank Claude Ostermann from Nycomed GmbH, Konstanz, Germany for the support and fruitful discussions.

Supporting Information Available: The application of the local flexibility similarity to discriminate stereoisomers from each other is shown in detail, and we provide all kernel parametrization contour plots on the Sutherland data sets[41] as a justifcation of our choice for the default parametrization. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Depnath, A. K. Quantitative Structure−Activity Relationship (QSAR) Paradigm−Hansch Era to New Millennium. *Mini-Rev. Med. Chem.* **2001**, *1*, 187–195.
(2) Kubinyi, H. From Narcosis to Hyperspace: The History of QSAR. *Quant. Struct.−Act. Relat.* **2002**, *21*, 348–356.
(3) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammet Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180.
(4) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000; pp 1−514.
(5) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(6) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(7) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines. *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
(8) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28*, 1711–1723.
(9) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity−Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553–2564.
(10) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating Molecular Shape into the Alignment-Free Grid-Independent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805–2815.
(11) 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.
(12) Ceroni, A.; Costa, F.; Frasconi, P. Classification of Small Molecules by Two- and Three-Dimensional Decomposition Kernels. *Bioinformatics* **2007**, *23*, 2038–2045.
(13) Azencott, C.-A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J. Chem. Inf. Model.* **2007**, *47*, 965–974.
(14) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
(15) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for Small Molecules and the Prediction of Mutagenicity,

Toxicity and Anti-Cancer Activity. *Bioinformatics* **2005**, *21*, 359–368.
(16) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 714–716.
(17) von der Lieth, C.-W.; Stumpf-Nothof, K.; Prior, U. A Bond Flexibility Index Derived from the Constitution of Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 711–716.
(18) Kier, L. B. An Index of Flexibility from Molecular Shape Descriptors. *Prog. Clin. Biol. Res.* **1989**, *291*, 105–109.
(19) Martinek, T. A.; Ötvös, F.; Dervarics, M.; Tóth, G.; Fülöp, F. Ligand-Based Prediction of Active Conformation by 3D-QSAR Flexibility Descriptors and Their Application in 3+3D-QSAR Models. *J. Med. Chem.* **2005**, *48*, 3239–3250.
(20) Dervarics, M.; Ötvös, F.; Martinek, T. A. Development of a Chirality-Sensitive Flexibility Descriptor for 3+3D-QSAR. *J Chem Inf Model* **2006**, *46*, 1431–1438.
(21) Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized Kernels Between Labeled Graphs. *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.
(22) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.
(23) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs—A New Similarity Based Approach To ADME Prediction in Classification and Regression. *QSAR Comb. Sci.* **2006**, *25*, 317–326.
(24) Mahé, P.; Ralaivola, L.; Stoven, V.; Vert, J.-P. The Pharmacophore Kernel for Virtual Screening with Support Vector Machines. *J. Chem. Inf. Model.* **2006**, *46*, 2003–2014.
(25) Rupp, M.; Proschak, E.; Schneider, G. Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity. *J. Chem. Inf. Model.* **2007**, *47*, 2280–2286.
(26) Brown, W. M.; Sasson, A.; Bellew, D. R.; Hunsaker, L. A.; Martin, S.; Leitao, A.; Deck, L. M.; Vander Jagt, D. L.; Oprea, T. I. Efficient Calculation of Molecular Properties from Simulation Using Kernel Molecular Dynamics. *J. Chem. Inf. Model.* **2008**, *48*, 1626–1637.
(27) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006; pp 7−79.
(28) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sülzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. *J. Chem. Inf. Model.* **2007**, *47*, 407–424.
(29) Schroeter, T.; Schwaighofer, A.; Mika, S.; Laak, A.; Sülzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Machine Learning Models for Lipophilicity and Their Domain of Applicability. *Mol. Pharm.* **2007**, *21*, 524–538.
(30) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002; pp *25*−*55*, 189−278.
(31) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
(32) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem. (Oxford)* **2001**, *26*, 5–14.
(33) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Model.* **2003**, *43*, 667–673.
(34) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
(35) Li, Q.; Bender, A.; Pei, J.; Lai, L. A Large Descriptor Set and a Probabilistic Kernel-Based Classifier Signicantly Improve Druglikeness Classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776–1786.
(36) Gärtner, T. A Survey of Kernels for Structured Data. *ACM SIGKDD Expl. Newslett.* **2003**, *5*, 49–58.
(37) Vert, J.-P. *The Optimal Assignment Kernel is not Positive Definite*; Technical Report arXiv:0801.4061v1; 2008.
(38) Fröhlich, H. Ph. D. thesis, University of Tübingen, Tübingen, Germany, 2006.
(39) Saigo, H.; Vert, J.-P.; Ueda, N.; Akutsu, T. Protein Homology Detection Using String Alignment Kernels. *Bioinformatics* **2004**, *20*, 1682–1689.
(40) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval Res. Logist.* **1955**, *2*, 83–97.
(41) 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.
(42) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 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.* **1993**, *115*, 5372–5384.
(43) Golbraikh, A.; Bernard, P.; Chretien, J. R. Validation of Protein-Based Alignment in 3D Quantitative Structure−Activity Relationships with CoMFA Models. *Eur. J. Med. Chem.* **2000**, *35*, 123–136.
(44) Maddalena, D. J.; Johnston, G. A. R. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/GABAA Receptors Using Articial Neural Networks. *J. Med. Chem.* **1995**, *38*, 715–724.
(45) Chavatte, P.; Yous, S.; Marot, C.; Baurin, N.; Lesieur, D. Three-Dimensional Quantitative Structure−Activity Relationships of Cyclooxygenase-2 (COX-2) Inhibitors: A Comparative Molecular Field Analysis. *J. Med. Chem.* **2001**, *44*, 3223–3230.
(46) Sutherland, J. J.; Weaver, D. F. Three-dimensional Quantitative Structure−Activity and Structure−Selectivity Relationships of Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309–331.
(47) Gohlke, H.; Klebe, G. 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.* **2002**, *45*, 4153–4170.
(48) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
(49) Böhm, M.; Stürzebecher, J.; Klebe, G. 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.* **1999**, *42*, 458–477.
(50) 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.
(51) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3DQSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
(52) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
(53) *Maestro*, version 8.5; Schrödinger LLC: New York, 2008.
(54) *MacroModel*, version 9.6; Schrödinger, LLC: New York, 2008.
(55) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines; 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Nov 18, 2008).
(56) Montgomery, D. C.; Runger, G. C. *Applied Statistics and Probability for Engineers*, 3rd ed.; John Wiley & Sons, Inc.: New York, 2004.
(57) R Development Core Team, R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
(58) Kabsch, W. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr.* **1976**, *32*, 922.

CI800329R