J Mol Model (2008) 14:547–558
DOI 10.1007/s00894-008-0302-3

ORIGINAL PAPER

ParaFrag—an approach for surface-based similarity comparison of molecular fragments

Arjen-Joachim Jakobi · Harald Mauser · Timothy Clark

Received: 30 November 2007 / Accepted: 12 March 2008 / Published online: 1 May 2008

Abstract A frequent task in computer-aided drug design is to identify novel chemotypes similar in activity but structurally different to a given reference structure. Here we report the development of a novel method for atom-independent similarity comparison of molecular fragments (substructures of drug-like molecules). The fragments are characterized by their local surface properties coded in the form of 3D pharmacophores. As surface properties, we used the electrostatic potential (MEP), the local ionization energy (IE_L), local electron affinity (EA_L) and local polarizability (POL) calculated on isodensity surfaces. A molecular fragment can then be represented by a minimal set of extremes for each surface property. We defined a tolerance sphere for each of these extremes, thus allowing us to assess the similarity of fragments in an analogous manner to classical pharmacophore comparison. As a first application of this method we focused on comparing rigid fragments suitable for scaffold hopping. A retrospective analysis of successful scaffold hopping reported for Factor Xa inhibitors [Wood MR et al (2006) J Med Chem 49:1231] showed that our method performs well where atom-based similarity metrics fail.

A.-J. Jakobi · H. Mauser (✉)
Discovery Chemistry, F. Hoffmann–La Roche AG,
4070 Basel, Switzerland
e-mail: harald.mauser@roche.com

A.-J. Jakobi · T. Clark
Computer Chemie Centrum and Interdisciplary
Center for Molecular Materials,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Nägelsbachstrasse 25,
91052 Erlangen, Germany

Keywords Fragment · Molecular surface · Similarity searching · Scaffold hopping · Pharmacophore matching

Introduction

The design and synthesis of novel molecules with a desired biological profile is the key task in medicinal chemistry. Much endeavor in attaining this goal has been driven by the concept of molecular similarity, which states that similar molecules tend to exhibit similar biological responses. Hence, the prerequisite of automated similarity searching is to define descriptors for molecules that relate similarities between observed properties in molecules to similarities in the descriptor space. The dilemma is now to develop descriptors that are not only capable of identifying similarly active molecules, but also of retrieving structurally distinct chemotypes as they may offer new entry points for identification of lead compounds with improved selectivity and/or pharmacokinetic properties [1]. Detailed studies by Brown and Martin [2, 3] showed that topological searches (based on atom connectivity) are well suited for identifying bioactive molecules, indicating that structurally similar molecules also have a similar biological profile. However, several cases have also been reported where structurally different molecules act on the same biological target [4]. This supports the concept of molecular recognition occurring via electronic properties near the molecular surface. Intermolecular interaction could then be described by the distribution of local interaction properties on the molecular surface [5, 6]. This atom-independent description of whole molecules or of discrete substructures could be of advan-

tage for identifying novel scaffolds. There has been considerable interest in this area in recent years, which has led to a number of computational tools that employ varying approaches such as feature trees [7], reduced graph representations [8], topological descriptors [9], molecular field points [10] or molecular interaction fields [11, 12]. Complementary to graph-based methods, quantum mechanical (QM) techniques are independent of topology and use the wavefunction or electron density to obtain information about molecular interaction properties such as polarizability, ionization potentials, electron affinities and multipole moments. Politzer [13] and Clark [14, 15] have developed a set of local properties that describe molecular interaction features on molecular surfaces. These local properties can be obtained readily from semiempirical molecular orbital (MO) calculations. In contrast to classical approaches, these properties are able to describe the surface anisotropy of local properties accurately, which ultimately allows us to account for non-classical interactions such as σ-hole bonding [16–20] or Lewis donor–acceptor interactions [14, 21, 22], which are often neglected in the approximate atomistic interaction potentials of standard force fields. The concept of σ-hole bonding [18–20] was recently proposed as a generalization of the concept of halogen bonding, and describes non-covalent but highly directional interactions of Group V–VII atoms.

One of the most daunting hurdles in identifying new lead structures is the structural diversity of chemical space to be explored during virtual screening. However, the amount of chemical space that can be explored by virtual high-throughput screening approaches is small compared to the accessible number of compounds, estimated to be up to 10^{60} molecules [23]. Alternative approaches are therefore required. One alternative to sampling whole molecules can be the analysis of the fragment space spanned by smaller substructures that occur repeatedly in molecules of interest (e.g., drug-like molecules). Fragment spaces are a combination of molecular fragments and connection rules [24]. They offer an attractive alternative to virtual compound libraries for several reasons. Firstly, they are considerably smaller than conventional virtual screening libraries, but can cover the same area of chemical space. Secondly, they contain small molecular entities, making computational resources less demanding. Last, and most importantly, the treatment of fragments greatly reduces the challenge posed by combinatorial flexibility. This inspired us to exploit the advantages of fragment spaces and to combine them with the non-atomistic representation of local properties projected onto molecular surfaces. Here we describe the design of the method in detail, followed by a validation study using Factor Xa inhibitors as a recently published example for scaffold hopping [25]. In the drug design context, the term scaffold hopping describes the discovery of structurally novel compounds starting from known actives by significantly changing the central core structure (for recent reviews, see [1, 26].

Methods

We use ParaSurf [27] to calculate the local surface properties and the resulting numerical descriptors for fragments whose open valences (attachment points) must be saturated appropriately. In-house tools were developed as extensions to ParaSurf that allow the treatment of fragments and fragment surfaces: All triangulation points of the isodensity surface and corresponding local properties relevant to the saturation valences are omitted in the program used subsequently. In addition, this program recalculates the standard set of descriptors based on the remaining surface points. The resulting data is processed by a Python routine that identifies critical points of the local property surfaces based on a statistical analysis of the local property distributions on the isodensity surface. The critical points represent a set of pharmacophore-like features that correspond to local property extremes. These feature pharmacophores are then stored as a pseudo-molecule, consisting of a standard set of eight atom types that each characterize one of the local property features (MEPmin, MEPmax, IE_Lmin, IE_Lmax, EA_Lmin, EA_Lmax, POLmin, POLmax) described in more detail below.

To account for the fact that calculating an accurate description of the molecular surfaces is computationally quite expensive, we decided to work with static databases, in which the feature pharmacophores of all template fragments are stored. In the query step, the feature pharmacophore of a query fragment is used as reference for calculating the similarity score of the pre-aligned property features for each suitable fragment in the database. As we are dealing with scaffolds that have per definition two or more exit vectors, we used an exit vector matching and alignment procedure (1) to identify suitable fragments in the database, and (2) to superimpose these fragments together with their feature pharmacophores onto the query fragment. Once aligned, it is straightforward to calculate the feature similarities as described below.

Calculation of molecular surface local properties

Chemical structures were coded in the SMILES notation [28] and 3D structures were generated by CORINA [29]. Exit vector valences were saturated with a methyl group. Single-point calculations were performed on the CORINA structures using the semiempirical Austin Model 1 (AM1) Hamiltonian [30] as implemented in VAMP [31]. The resulting SD files were used as input to the calculation of

the molecular surface properties by ParaSurf [27]. ParaSurf uses the results of semiempirical MO calculations to create isodensity surfaces that may fit to a spherical harmonic expansion [32]. Isodensity surfaces are defined as molecular surface representations for which the contour of the surfaces is established by a constant cutoff value for the electron density. The surfaces can be created by marching-cube [33] or shrink-wrap [34–36] algorithms. In this study, marching cube surfaces corresponding to the 0.003 e $Å^{-3}$ isodensity contour were used as default. The surfaces thus created are tessellated and the molecular electrostatic potential (MEP) and the three local properties local ionization energy (IE_L [13]), local electron affinity (EA_L [14]) and local polarizability (POL [14]) are calculated at each triangulation point. The physical significance of the local properties with respect to intermolecular interactions has been discussed in detail and the properties have since been used to generate a set of 40 descriptors appropriate for QSPR studies [15]. In the course of this study, however, we developed a different approach by identifying critical points of these four local property surfaces. To this end, we used the surface triangulation points (vertices) and the corresponding local properties as calculated by ParaSurf.

Identification of surface property extremes and generation of the local property pharmacophore

The pharmacophore generation and similarity search engine (see below) were programmed using the Python language and the OEChem library toolkit [37]. Using triangles is a common way to depict molecular surfaces. In ParaSurf, the vertices of every triangle are associated with the value of each local property at this point on the molecular surface. In addition, every vertex is assigned to the explicit atom that contributes most to the electron density at the surface point under investigation (based on the linear combination of atomic orbitals, LCAO, approach). As local extremes of electronic properties on the molecular surfaces are considered to encode hot spots for intermolecular interactions, a filtering procedure was applied that uses the median and standard deviation (σ) of each property distribution to define the cutoffs for the corresponding property value. Since for some molecules we observed an unbalanced distribution of the local properties, we tried to compensate for this by a pragmatic outlier correction leading to a standard deviation (σ_{corr}) that is less influenced by local property extremes. To reduce the influence of strongly positive or negative properties on the statistical analysis, we extracted the vertices with local property values inside the 2σ region around the mean value and re-calculated the mean and the standard deviation (σ_{corr}) for those vertices. In the ideal case, this region corresponds to a subset of the molecular surface without specific interaction features, so it was used as a reference for hot spot detection. The value of σ_{corr} gives an indication of the distribution of the properties on the molecular surface and was used together with the global extreme values (P_{min}, P_{max}), to define lower and upper threshold values according to Eqs. 1 and 2.

$$\text{Lower threshod}: \quad T^L = P_{min} + \sigma_{corr} \tag{1}$$

$$\text{Upper threshod}: \quad T^U = P_{min} + \sigma_{corr} \tag{2}$$

The surface areas within these ranges were found to be adequate representations of positive or negative hot spots on the surface and are directly linked to the molecular interaction pattern (see Validation). This identification of hot spots on the surface in itself is valuable for visual comparison. However, when dealing, with a large number of fragments, one needs a fast and robust method for numerical comparison of the presence or absence of these hot spots on the molecular surface.

According to our definition of hot spots, for each property we keep only those vertices with values below T^L and above T^U. These critical surface vertices correspond to either minima or maxima of the individual local properties. Generally, we aim at having one critical surface vertex per local extreme on the surface. However, in cases where hot spots correspond to an extended segment on the surface, the vertex selection procedure results in multiple extremes for this segment. Therefore, we used a distance comparison among the critical vertices to identify one extreme out of the neighboring vertices (default: rmsd <=0.2 Å). Separate comparisons are performed for the vertices that correspond to positive or negative property extremes, hereafter termed "unique surface extremes". The resulting unique surface extremes proved to be insensitive to increasing the ranges for the statistical vertex selection (i.e., using larger values for σ_{corr} in Eqs. 1 and 2). In addition, these statistically derived unique surface extremes were found to agree well with those given by a computationally more demanding analytical approach, where local extremes together with the saddle points were obtained from a spherical harmonics expansion of the local property distributions (Fig. 1).

The unique surface extremes for all properties are stored together in the form of a pseudo molecule that contains the extreme vertices as atoms whose atomic numbers correspond to a feature type (Fig. 2). Eight atom types are defined that correspond to the four local properties MEP, IE_L, EA_L and POL, each subdivided into the classes maxima and minima. This form of storing the information gives us easy access to visualizing the surface extremes via standard modeling packages like MOE [38] or PyMol [39]. In addition, we can use standard methods for superimposing pre-calculated surface extremes, thus avoiding redun-

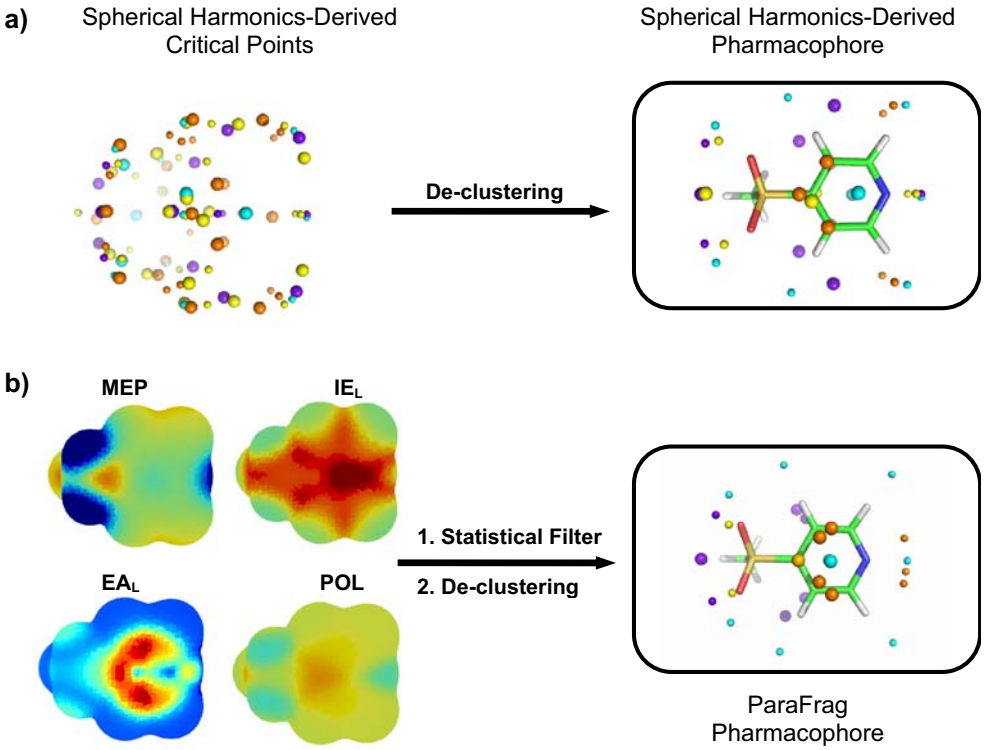

Fig. 1 **a**, **b** Comparison of analytically and statistically derived local property pharmacophores. **a** The *left panel* shows the critical points derived analytically from a tenth order spherical harmonics expansion. Selection of the top extremes and de-clustering results in the spherical harmonics-derived pharmacophore (*upper box*). Note that the position of the retrieved critical points strongly depends on the order of the spherical harmonics expansion employed, which determines the accuracy of the surface approximation. The ParaFrag pharmacophore derived from exclusively filtering surface local property distributions on marching cube surfaces is shown for comparison (*lower box*). **b** The local property surface distributions are depicted for comparative reasons. A concise description of the individual pharmacophore features is given in Fig. 2

dant QM calculations (see section on exit vector matching below).

Conformer generation

As for any 3D-technique, care must be taken in identifying representative conformers. For smaller molecules, or even fragments, the conformational space is significantly reduced, allowing us to use a very accurate but CPU-intensive method for describing the fragment's properties. In our validation study, we started with one input geometry obtained by CORINA from the SMILES depiction of our target fragment *scaffold*-**4** (see Validation) and used Omega [40] to generate a conformer library. As for the semiempirical calculation, we saturated the exit vector valences with methyl groups. The Omega parameters were optimized starting from the settings for high-quality screening of standard molecules as suggested by Kirchmair et al. [41]. The best parameters for fragments were found to be *ewindow*=25, *maxconfs*=250, *rms*=0.3 with electrostatics turned off (*mmff94s_NoEstat*). The comparatively low rms cutoff was used to allow for an adequate number of conformers, including variations of the exit vector orientations that are very important in the context of scaffold hopping. However, these settings can only be seen as provisional since we have not yet used Omega on a larger scale for fragment-based conformer generation. In the case of scaffold-**4**, we obtained 17 conformers with a good coverage of space including the conformation obtained by manual alignment (see below).

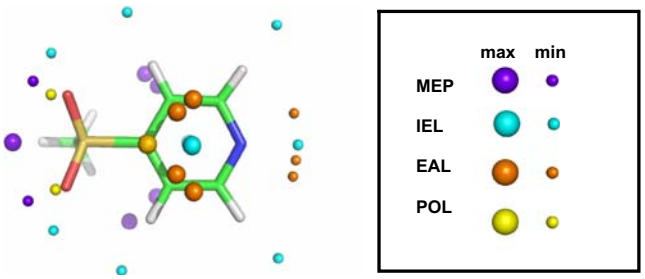

Fig. 2 Example of a local property pharmacophore. Pharmacophore features derived from local surface properties are shown as *spheres*. The 3D molecular structure is shown for clarity. Maxima and minima are represented by *large* and *small spheres*, respectively. Color coding used throughout this paper for individual local properties is summarized in the framed box: *purple-blue* MEP,*cyan* IE_L, *orange* EA_L, *yellow* POL. Individual features are highlighted for each of the local properties and the 3D molecular structure is shown for clarity

Exit vector matching

A successful replacement of central elements from known active structures often crucially depends on a similar spatial arrangement of exit vectors, when comparing template and putative new scaffold. In the context of fragment database queries, this implies that a computationally inexpensive filtering procedure that is capable of identifying fragments with suitable exit vector geometry will significantly reduce the number of fragments to be considered in the subsequent similarity calculation. Moreover, the exit vectors offer an elegant way for a simple and computationally efficient alignment of fragments and their molecular surfaces, reducing the number of resource-intensive QM calculations to the absolute minimum. We describe the exit vector geometry with respect to the fragment scaffold by defining vectors that originate at the attachment points (the heavy atom containing the exit vector valence) and point towards the topologically connected atom that is part of the corresponding saturating exit vector (CH_3 in this study). To facilitate database mining for fragments of suitable exit vector geometry, a second pseudo molecule is created that consists of all exit vectors encoded as Li (for *Li*nker) and the corresponding attachment points in α-position with their original element symbol. By defining bonds between all exit vectors and the corresponding attachment points and connecting all individual attachment points pairwise, the resulting pseudo molecule can serve as a template for a substructure search based on SMARTS strings (Fig. 3).

Starting from the query SMARTS string, the algorithm identifies all fragment conformers in the fragment database that have exit vector geometries similar to the query fragment. Firstly, the database is mined for fragments that contain the query SMARTS string. Subsequently, all possible alignment structures are assessed to determine whether they lie within a user-defined rmsd cutoff with respect to the query structure. Database fragments that do not match this requirement are discarded for subsequent pharmacophore alignment and comparison. For fragments that do match the exit vector geometry of the query, alignments are performed in all possible orientations that satisfy the rmsd criterion. The transformation matrices that result from the substructure alignments are also applied to the corresponding local property pharmacophores in order to fit the rearranged orientation of the respective database fragment.

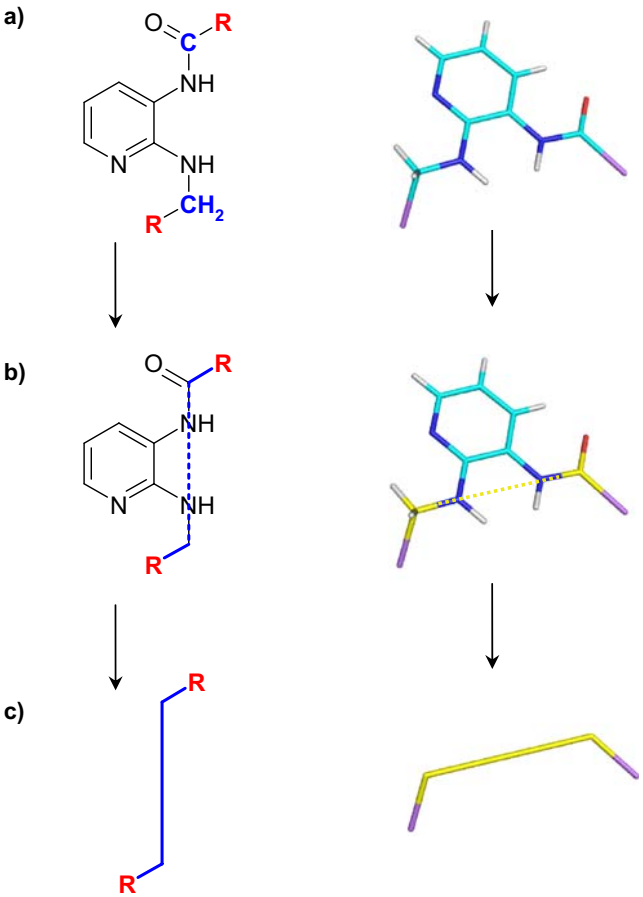

Fig. 3 **a**–**c** Schematic illustration of the exit vector matching and alignment procedure. Generation of the SMARTS exit vector substructure. **a** Attachment points (*blue*) and exit vectors (*red*) are identified. **b** Attachment points are subsequently connected pair wise (*dashed lines*) to generate the substructure pattern. **c** A pseudo-molecule containing the exit vector substructure is created for subsequent SMARTS pattern matching

Feature point comparison and scoring

The similarity metric for comparing local property pharmacophores is based on evaluating distances between the local property features of query and target and resembles a classical pharmacophore alignment. The algorithm implemented in the similarity routine determines distances between local property features of query and target. Only distances between features of equal properties are evaluated in the subsequent procedure. The spatial distance of two features *i* (of query A) and *j* (of target B) is calculated as the modulus of the Euclidian distance vector. As the distance is evaluated with respect to the query, $d_{A,B}$ resembles the feature radius of the conventional pharmacophore model. The algorithm creates a distance matrix$D^{(NxM)}$, whose rows (N) and columns (M) are defined by the local property features of query and target, respectively. The elements of the matrix $(D)_{ij}$ represent the distance between the feature *i* of the query A and the feature *j* of the target B. The distance matrix is rearranged to a binary scheme with matrix elements being set to 1 or 0 for distances of matching features that range within or exceed a predefined distance cutoff, respectively. Matrix elements that correspond to combinations of different local property features are set to 0. The similarity score $S_{A,B}$ is now defined as the

sum over all matrix elements, divided by the minimum number of features to match (Eq. 3).

$$S_{A,B} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\mathbf{D}_{ij}}{\min(N,M)} \quad (3)$$

The algorithm also allows us to exclusively score match features of predefined local properties in a procedure analogous to commonly used pharmacophore comparison methods. Moreover, in order to train the scoring scheme for a specific family (or biologically active cluster) of lead structures, sub-matrices corresponding to individual property features can be evaluated separately and the resulting scores, s_k, can be weighted by coefficients c_k in an additive scheme (Eqs. 4, 5).

$$S_{A,B} = \sum_{k=1}^{k=8} C_k \mathbf{S}_k \quad (4)$$

Where

$$\mathbf{S}_k = \frac{\sum_{i}^{N_k}\sum_{j}^{M_k}\mathbf{D}_{ij}(k)}{\min(N,M)_k} \quad (5)$$

Here, $D_{ij}(k)$ denote the matrix elements contained in the sub-matrix that corresponds to property feature k, while N_k and M_k correspond to the rows and columns of the sub-matrix, respectively.

Validation

The central hypothesis behind our study is that structurally diverse fragments with a similar arrangement of interaction features should possess a similar distribution of local properties on their molecular surface. To corroborate this assumption, the research presented in this study followed two major objectives: to develop a method for the atom-independent description of a molecular fragment's interaction features that is based on molecular surface properties and to validate the application of this strategy with regard to its ability to explain a literature example for scaffold hopping retrospectively.

Method development

In contrast to many surface-based similarity techniques described in the literature, the rationale behind our approach was to extend the portfolio of these methods with a technique based on quantum-mechanical principles that is known for high accuracy and also well suited for the description of non-classical effects. The basic strategy was to define a set of critical points of local properties on the fragment surface that should serve as a means to create a pharmacophore-like model that encodes hot spots for intermolecular interactions with a protein target.

For a first validation of our approach, we chose a pair of structurally similar fragments and used ParaFrag to assess the similarity in terms of surface properties (Fig. 4). 1-Chloro-4-sulfonyl-benzene(**1**) and 4-sulfonylpyridine (**2**) will be used to illustrate the key concept of our method. Usually, the global extremes of a fragment's local property surface will represent the most probable sites for interactions with other molecules. Reducing the information encoded on the property surfaces to just their global maxima and minima, however, neglects important sites of possible intermolecular interactions that might be less pronounced but are still important. To account for these limitations, we developed an alternative strategy that relies on a statistical treatment of the local property distributions on isodensity surfaces, as described in Methods. A shortcoming of this strategy is that it tends to detect clusters of extreme vertices rather than individual extremes. These clusters consist of the local extremes and adjacent vertices with similar absolute property values. To unambiguously select relevant features, the local extreme of each cluster is identified and, within a cluster, only vertices above a certain rmsd threshold with respect to the local extreme are retained (Fig. 5). This shows that our method succeeds in reducing the information from the surface local property distributions significantly to those points critical to intermolecular interactions.

At this point, it becomes important to see whether the methodology developed is able to detect similarities (and differences) between local property isodensity surfaces of

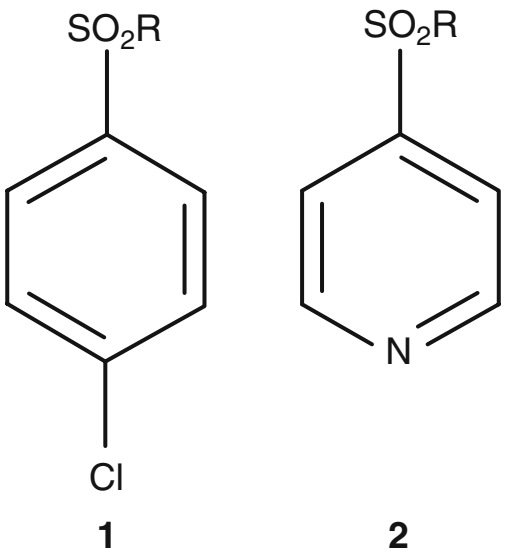

Fig. 4 Comparison of two structurally very similar fragments with subtle electronic differences: 1-chloro-4-sulfonyl-benzene (**1**) and 4-sulfonylpyridine (**2**). Exit vectors are labeled with *R*

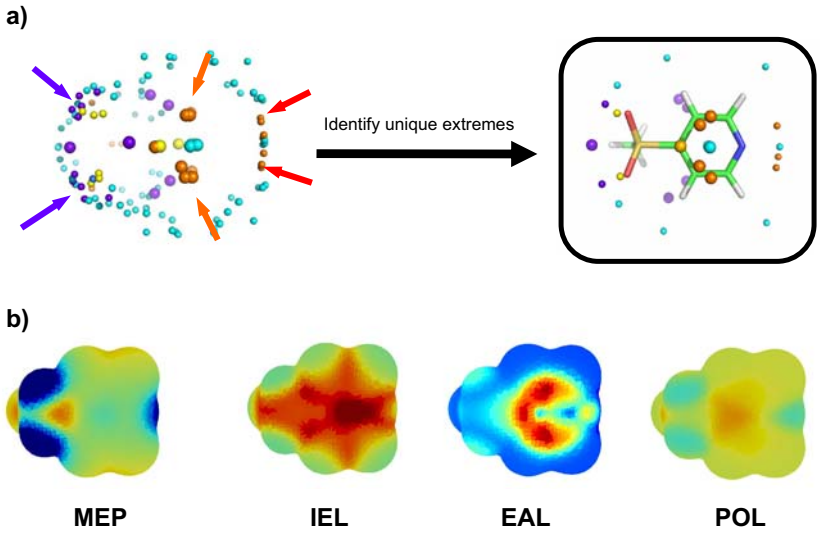

Fig. 5 **a**, **b** Local property pharmacophore model. **a** Generation of the local property pharmacophore. *Left panel* Depiction of critical points on the local property surfaces of 4-sulfonylpyridine. *Large spheres* correspond to local maxima while *small spheres* represent local minima with color coding as described in Fig. 2. Note the clusters around critical spots (*bold arrows*). *Right panel* Resulting local property pharmacophore after application of the rmsd filter to attenuate clustering (*boxed*). **b** Local property surfaces of 4-sulfonylpyridine **2**. Comparison of the surface distribution of local properties with the pharmacophore model reveals that all critical spots are recovered in the pharmacophore

bioisosteric fragments. Using ParaFrag, we calculated the total similarity of the local property pharmacophores (sphere radius=1) of **1** and **2** to be 0.47 (MEP=0.5, IE_L=0.5, EA_L=0.4, POL=0.5). The visualization of the individual feature points using the example of fragment **1** is shown in Fig. 6. Coulomb interactions are usually the most prominent intermolecular interactions, and features of the MEP surface are therefore of crucial importance. Regarding the MEP surface, minima at both the sulfonyl oxygens and the chlorine atom would be expected by classical approaches. Whereas the minima at the sulfonyl oxygens are recovered, it can be observed that a maximum is detected at the surface at the extension of the chlorine σ-bond. Although this may appear counterintuitive at first, it may be regarded as indicative of σ-hole bonding. In this particular case the non-covalent interaction of a halogen atom with a Lewis base suggests the presence of an electron deficiency in distinct regions around the halogen surface [17, 42]. In atomic monopole-based force fields, electrostatic interactions are usually treated by assigning fictitious

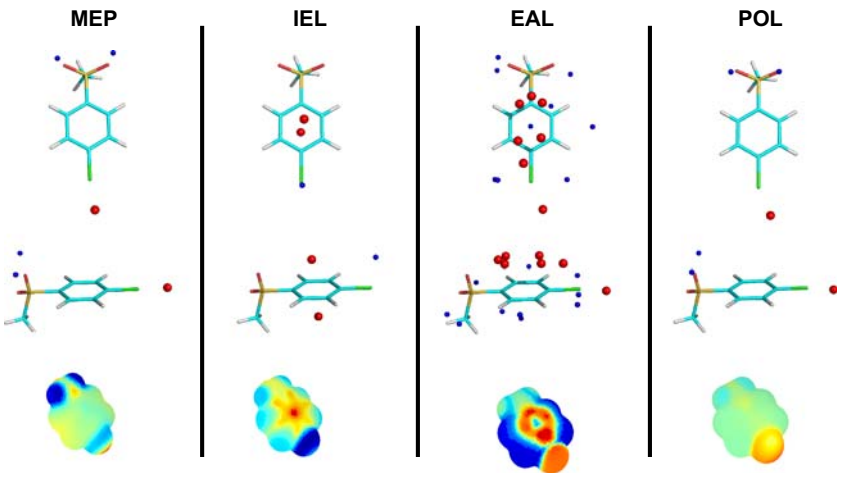

Fig. 6 Individual local property features of 1-chloro-4-sulfonylbenzene. Critical points of local property surfaces are depicted separately for the individual local properties. Maxima are shown as *large red spheres*; minima are represented as *small blue spheres*. The 3D molecular structure is shown in *cyan*. Top views, side views and the corresponding local property isodensity surface are depicted in the *top panel*, *middle panel* and *lower panel*, respectively. Although the 3D molecular structure is shown for clarity, it should be stressed here that the derived properties are not ascribed to functional groups but represent atom-independent features that are governed by the electronic structure of the entire molecule

partial charges to individual atoms followed by calculating the interaction potential by Coulomb's law. In the framework of semiempirical MO theory, the MEP can be calculated efficiently using a zero-differential overlap-based atomic multipole model [43]. Whereas the MEP derived from classical approaches reveals isotropic distributions from an atomistic view, MEPs derived from quantum mechanics encode information on anisotropic electron density distributions around atoms. The presence of anisotropic MEP distributions with a positive MEP along the extension of the halogen σ-bond (the *σ-hole*) for certain types of halogenalkanes was reported only recently by Clark et al. [17], who also presented a theoretical foundation based on density functional theory (DFT) studies. Initiated by the crystallographic studies of Hassel [44], σ-hole bonding has recently gained increased attention due to its potential use as a non-classical effector in intermolecular recognition events and examples much resembling classical hydrogen bonds and cation-π interactions have also been found to occur in protein–ligand complexes [45]. Clearly, the presence or absence of σ-holes on halogen atoms depends on the molecular environment and quantum mechanical methods are required to recover it. To examine whether the observed σ-hole on the chlorine of **1** is merely an artifact of the semiempirical approximation, higher level DFT calculations (Gaussian03 [46]) were performed, employing the B3LYP [47–49] hybrid density functional with the 6–311+G(d) basis set [50–56]. Figure 7 compares the MEP surfaces derived from the AM1 and B3LYP/6–311+G(d) calculations. Although the electropositive potential crown is markedly more pronounced in the AM1-derived MEP surface, the σ-hole is recovered also at the B3LYP/6–311 G(d) level of theory. The example of 1-chloro-4-sulfonyl-benzene thus shows that the semiempirical approach used throughout our method is accurate enough to detect molecular details that might be responsible for intermolecular interactions and that are only partially covered, if at all, in most of the currently available pharmacophore tools.

Using the example of 1-chloro-4-sulfonyl-benzene, we will briefly describe the interactions represented by the local property pharmacophore. The strength of dispersive interactions (London forces) depends on the polarizability of the interacting molecules, which is usually unevenly distributed throughout a molecule. Preferred sites of dispersive interactions will be those parts of the electron density distribution that can be easily polarized. Figure 6 reveals local polarizability minima for the sulfonyl oxygens and a local maximum close to the chlorine. The corresponding surface representation shows that these are the major determinants of the polarizability distribution. Polarizability near the sulfonyl oxygens indicate the compactness of the electron density around this functional group, while the maximum near the chlorine shows up the polarizable nature of the electron density at this site of the molecule. Sites near to the chlorine are thus most likely expected to contribute to dispersive interactions with other molecules.

Another important class of molecular interactions are donor–acceptor interactions, which are described qualitatively by the Lewis acid–base concept. Electron donor–acceptor interactions may lead to electron redistribution processes that favor the formation of induced dipoles. As with the local polarizabilities, preferred sites for such interactions can be expected at the surface extremes of these properties. For the IE_L, Fig. 6 reveals pronounced maxima above and beneath the ring plane, and a minimum in the vicinity of the chlorine, quantitatively encoding the fragment's donor properties. The characteristic maxima on the ring plane surfaces are commonly observed features for aromatic systems. They serve in this sense as a pharmacophoric equivalent for the ring centroids traditionally used for aromatic systems. While the IE_L maxima reflect the energetic preference for maintaining integral π-systems, the minimum close to the chlorine represents the pronounced donor properties at this site of the molecular surface. The EA_L panel shows that the method efficiently recovers maxima *ortho* to ring carbons substituted with electron withdrawing groups and the *ipso* carbon itself. These sites represent areas most likely to interact with a donor site of another molecule. In summary, the set of semiempirically derived local property extremes accurately describes the fragment in terms of important molecular interaction features.

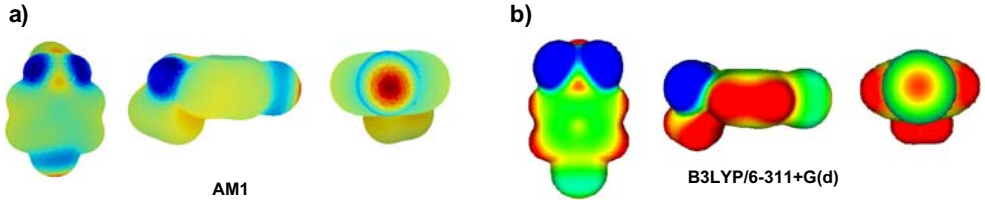

Fig. 7 **a**, **b** MEP surfaces derived from AM1 and B3LYP/6–311+G(d) calculations. Computed MEP on the 0.003 e $Å^{-3}$ isodensity surface of **1** is shown as derived from AM1 (**a**) and B3LYP/6–311+G(d) calculations (**b**). The rear view is directed along the Cl–C axis. A positive electrostatic potential (the σ-hole) is visible along the extension of the Cl-C σ-bond for both **a** and **b**. Images of the B3LYP electrostatic potential surface were created with gOpenMol [57–59]

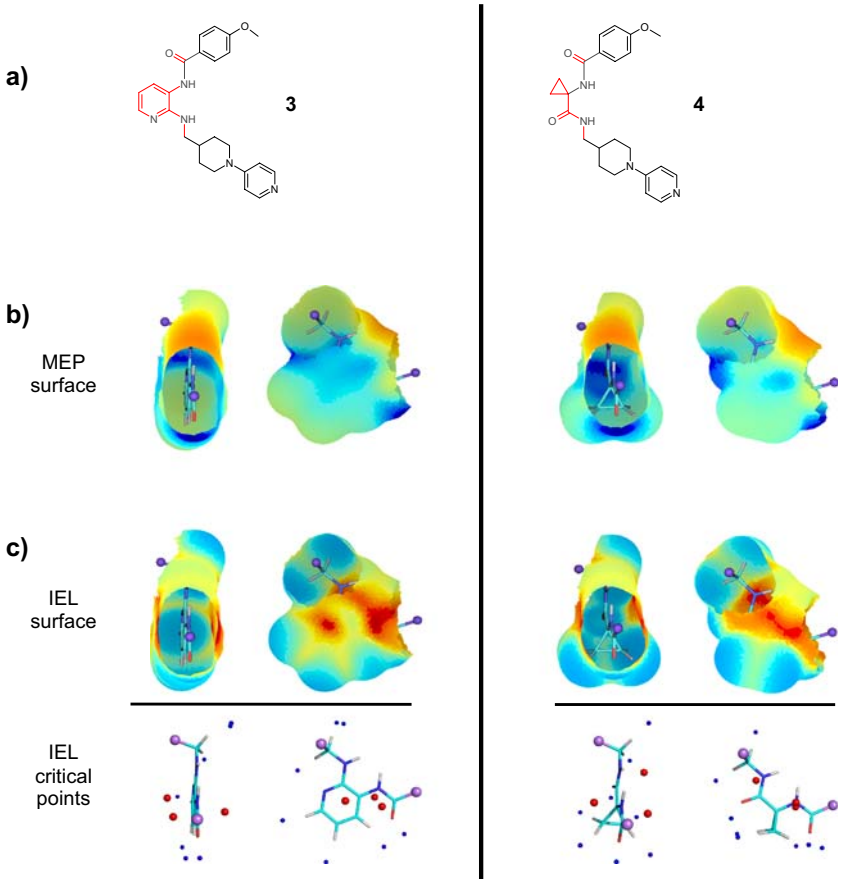

Fig. 8 **a**–**c** Comparison of the MEP and IEL surfaces for the scaffolds of inhibitors **3** and **4**. **a** Structures of inhibitors **3** and **4** with the respective scaffolds highlighted in red. **b**, **c** MEP and IE_L surfaces (IE_L local property pharmacophores are shown)

Validation study—inhibitors of factor Xa

As a proof-of-concept study for similarity identification, the inhibitors *FXa* **3** and *FXa* **4** of the blood coagulation protease Factor Xa [25] with structurally different scaffolds were compared (Fig. 8; the core fragments highlighted in red will be referred to as scaffold-**3** and scaffold-**4**, respectively). The same pharmacophoric replacement of 2,3-diaminopyridine by cyclopropylamino acid amide was also observed for bradykinin B1 receptor antagonists [25].

The validation itself consisted of two parts: firstly, we wanted to determine whether our concept of exit vector matching is suitable for identifying the appropriate fragment conformation and for automatically aligning the coordinates of the surface property extremes onto the query fragment. Secondly, we were interested in comparing our

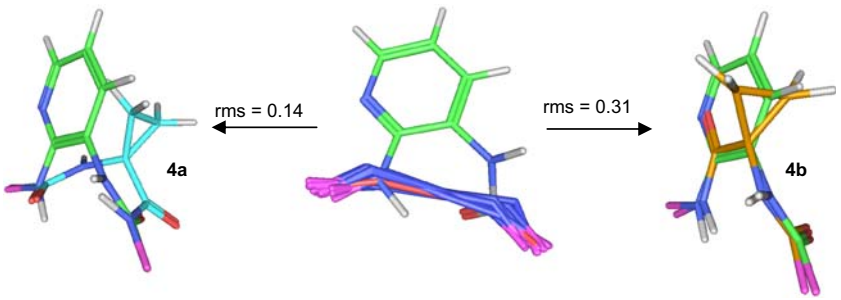

Fig. 9 Exit vector matching used to align the fragment's conformer library of scaffold **4** onto the reference geometry of scaffold-**3**. The binding conformation of scaffold-**3** was obtained by modeling the whole inhibitor **3** into X-Ray structures of related factor Xa inhibitors (data not shown). The template conformer library for scaffold-**3** was generated by Omega and subjected to our exit vector matching and alignment procedure (see Methods for details). *Centre* Aligned exit vectors with rms cutoff set to 0.5 (16 solutions including symmetrical matches for some fragments; the reference exit vector is highlighted in red). *Left* Best exit vector match (scaffold conformer **4a**); however, shows only low overall similarity. *Right* Good exit vector similarity and best overall similarity (scaffold conformer **4b**) for scaffolds **3** and **4.** Despite their structural difference, scaffolds **3** and **4** possess a similar pattern of local property extremes at the respective surfaces

Table 1 Comparison of ParaFrag to alternative similarity metrics. *MCSS* Maximum common substructure search

Query: *FXa* 3	Daylight	MCSS	ROCS (ShapeTanimoto)	Feature Trees	ParaFrag[a]
Target: FXa 4	0.22	0.08	0.88	0.85	0.59 (0.47)

[a] Value obtained for manual alignment (best solution). The result in brackets corresponds to the best result for the automated alignment by exit vector matching performed on a conformer library of scaffold-4 (corresponding to conformer 4b in Fig. 9)

method of pharmacophore-type similarity scoring to other methods that are frequently used for similarity comparison and virtual screening.

Figure 9 shows an example for the exit vector alignment of the conformer library for the target scaffold-**4**. As reference geometry, we modeled the inhibitor **3** into the active site of factor Xa co-crystallized with a structurally related inhibitor (data not shown). After cleaving off the terminal substituents, we kept the geometry of scaffold-**3** fixed and extracted the exit vector as described in Methods. Then we applied our SMART-based exit vector matching and alignment to identify suitable conformers out of the conformer library for scaffold-**4**. Best results were obtained with an rms cutoff of 0.5, which gives 16 solutions for nine conformers, i.e., the other six conformers were rejected due to high rms distances. As the exit vector SMARTS string "LiAALi", Li for the exit vector, allowed symmetrical matches, we obtained two solutions for seven of the conformers in the library. In this example, we identified the choice of the rms cutoff as a potential source of fuzziness in the exit-vector-based alignment. Large rms values also include solutions that are not suitable in terms of their overall alignment, thus increasing the risk of false positives in the subsequent local property pharmacophore comparison. Too low an rms cutoff, however, increases the risk of losing the best matches in terms of overall similarity, as depicted in the two alignment comparisons in Fig. 9. Nevertheless, this should not be seen as the limiting step of the similarity analysis, as the quality of the results can be assessed easily either by visual inspection or by further profiling (e.g., shape filtering). In our hands, this method worked well and allowed the automated alignment of the pre-calculated local surface property extremes in a very accurate manner (no difference was obtained on re-calculating the property pharmacophores of the aligned conformer).

Having identified suitable conformers of the target fragment scaffold-**4**, we compared the local property pharmacophores as described above in detail. Not surprisingly, the radius of the pharmacophore spheres turned out to be an important parameter. Sphere radii larger than 1.3 Å resulted in fuzzy solutions; on the other hand radii smaller than 0.5 Å turned out to be too restrictive and identified matches only of closely related fragments. We obtained the best balance for a sphere radius of 1.0 Å, which was set as the default parameter in subsequent analysis.

For our selection of nine conformers obtained in 16 different poses by exit vector alignment, we found pharmacophore scores ranging from 0.00 to 0.47. In total, the method differentiates well among the different solutions from exit vector matching. For example, we obtained a score of 0.06 for the alignment scaffold-**4a** (Fig. 9). In contrast, scaffold-**4b** was found to be the best solution with a total score of 0.47. This agrees reasonably well with the manual alignment, where we obtained a total score of 0.59. In the bound conformation, modeled in a manner analogous to that described for inhibitor **3**, the total score is reduced to 0.41. This is caused mainly by slight distortions of the geometry in the active site environment.

Valuable information obtained using ParaFrag might include individual contributions of the total score obtained in the property pharmacophore comparison. In our present example, the best score of 0.59 can be partitioned into the

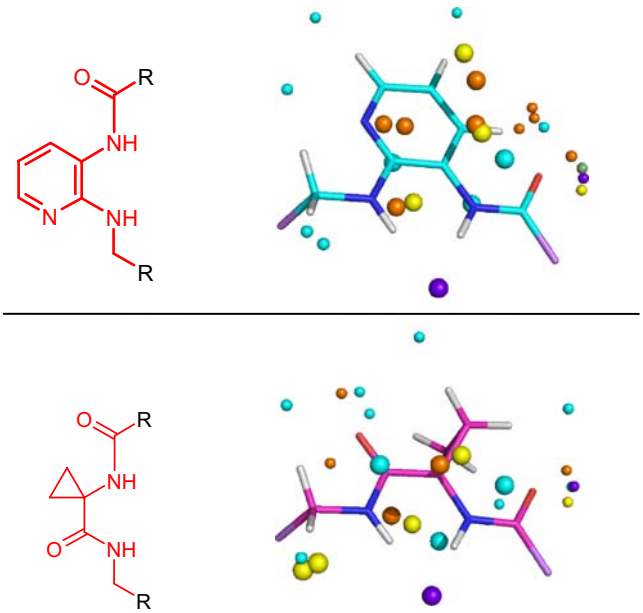

Fig. 10 Local property pharmacophore representations of scaffolds **3** and **4**. A comparison of the entire local property pharmacophores of scaffolds **3** and **4** reveals conservation of essential interaction features despite the markedly different scaffolds. Local property features are coded by color and size as described in Fig. 2

individual terms of MEP=1.0; IE_L=0.13; EA_L=0.75; POL=0.5. This might help to characterize important features of a given fragment to be used for biased searches.

Finally, we compared the results of our method with those obtained using established similarity metrics (see results in Table 1). Not surprisingly, atom-based methods such as Daylight or Maximum common substructure search (MCSS) fail to recognize the similarity of the two scaffolds. In contrast, pure shape matching (ROCS [60]) or feature comparison (Feature Trees [61]) show high similarity scores but for complementary reasons. Feature Trees captures the conserved donor/acceptor patterns of the two scaffolds without giving details of their electronic nature. ROCS, however, provides an insight into the shape similarity of the two scaffolds. Thus, the results of the two methods can be considered complementary. However, in our experience, ROCS is quite tolerant with respect to subtle differences.

ParaFrag seems to lie between the two extremes. The distribution of extreme vertices on the surface provides an approximate description of the molecular shape together with a highly accurate description of the local surface properties (Fig. 8) in summary leading to a more specific scoring scheme. In addition to the numerical similarity value, the method also offers the possibility of inspecting the results in more detail by visual comparison of the positions of the unique surface extremes (Fig. 10).

In summary, scaffold hopping from scaffold-**3** to scaffold-**4** can be explained by a similar pattern of interaction features consistent with the distribution of local properties on the molecular surface, thus corroborating the central hypothesis underlying this work.

Summary and outlook

We have presented a novel approach for the surface-based similarity comparison of rigid molecular fragments. Extremes of semiempirically derived electronic properties on the molecular surface were demonstrated to describe a molecule in terms of intermolecular interaction features independent of chemical structure. Although the method is based on a statistical scrutiny of the local property surfaces, the critical positions identified correlate well with an analytical validation model. It was shown that semiempirical MO theory may be used to explore non-classical interaction features that are not covered in most force-field based approaches. The accuracy and relevance of these features was validated by DFT calculations at the B3LYP level of theory. A retrospective study of a known example for scaffold hopping revealed that our method can reproduce the similarity of structurally different scaffolds with similar biological profiles. The automated exit vector matching and alignment offers a fast and accurate way of conformer selection, reducing the number of QM calculations to the absolute minimum. These findings encourage us to develop our approach further towards an independent tool for scaffold hopping. We emphasize here, however, that the study presented in this work focused on the comparison of rigid fragments. Thus, the critical aspect of conformational flexibility still hampers the wider application of this method.

Acknowledgments We thank David Whitley and Brian Hudson, Center for Molecular Design, University of Portsmouth, UK for helpful discussions and for providing the results of the spherical harmonics calculation. This work would not have been possible without support of our colleagues in the Cheminformatics & Molecular Modeling group at Roche, Basel. In particular we thank Wolfgang Guba, Daniel Stoffler, Olivier Roche and Martin Stahl. We thank Wolfram Altenhofen and Guido Kirsten, Chemical Computing Group, for providing an interface for visualizing the local property surfaces in MOE.

References

1. Böhm HJ, Flohr A, Stahl M (2004) Drug Discov Today: Technologies 1:217–224 DOI 10.1016/j.ddtec.2004.10.009
2. Brown RD, Martin YC (1996) J Chem Inf Comp Sci 36:572–584 DOI 10.1021/ci9501047
3. Brown RD, Martin YC (1997) J Chem Inf Comp Sci 37:1–9 DOI 10.1021/ci960373c
4. Zhao H (2007) Drug Discov Today 12:149–155 DOI 10.1016/j.drudis.2006.12.003
5. Clark T (2006) Proceedings of the International Beilstein Workhop. Bolzano, Italy
6. Stone AJ (1996) The theory of intermolecular interactions. Clarendon, Oxford
7. Maass P, Schulz-Gasch T, Stahl M, Rarey M (2007) J Chem Inf Model 47:390–399 DOI 10.1021/ci060094h
8. Barker EJ, Buttar D, Cosgrove DA, Gardiner EJ, Kitts P, Willett P, Gillet VJ (2006) J Chem Inf Model 46:503–511 DOI 10.1021/ci050347r
9. Schneider G, Neidhart W, Giller T, Schmid G (1999) Angw Chem Int Ed 38:2894–2896 DOI 10.1002/(SICI)1521–3773(19991004)38:19
10. Low CMR, Buck IM, Cooke T, Cushnir JR, Kalindjian SB, Kotecha A, Pether MJ, Shankley NP, Vinter JG, Wright L (2005) J Med Chem 48:6790–6802 DOI 10.1021/jm049069y
11. Ahlström MM, Ridderström M, Luthmann K, Zamora I (2005) J Chem Inf Model 45:1313–1323 DOI 10.1021/ci049626p
12. Bergmann R, Linusson A, Zamora I (2007) J Med Chem 50:2708–2717 DOI 10.1021/jm061259g
13. Sjoberg P, Murray JS, Brinck T, Politzer P (1990) Can J Chem 68:1440–1443
14. Ehresmann B, Horn AHC, Clark T (2003) J Mol Model 9:342–347 DOI 10.1007/s00894–003–0153-x
15. Ehresmann B, de Groot MJ, Alex A, Clark T (2004) J Chem Inf Comp Sci 44:658–668 DOI 10.1021/ci034215e
16. Politzer P, Lane P, Concha MC, Ma Y, Murray JS (2007) J Mol Model 13:305–311 DOI 10.1007/s00894–006–0154–7
17. Clark T, Hennemann M, Murray JS, Politzer P (2007) J Mol Model 13:291–296 DOI 10.1007/s00894–006–0130–2

18. Murray JS, Lane P, Clark T, Politzer P (2007) J Mol Model 13:1033–1038 DOI 10.1007/s00894–007–0225–4
19. Murray JS, Lane P, Politzer P (2007) Int J Quantum Chem 107:2286–2292 DOI 10.1002/qua.21352
20. Politzer P, Murray JS, Lane P (2007) Int J Quantum Chem 107:3046–3052 DOI 10.1002/qua.21419
21. Politzer P, Murray JS, Concha MC (2002) Int J Quant Chem 88:19–27 DOI 10.1002/qua.10109
22. Ping J, Murray JS, Politzer P (2004) Int J Quant Chem 96:394–401 DOI 10.1002/qua.10717
23. Dobson CM (2004) Nature 432:824–828 DOI 10.1038/nature03192
24. Mauser H, Stahl M (2007) J Chem Inf Model 47:318–324 DOI 10.1021/ci6003652
25. Wood MR, Schirripa KM, Kim JJ, Wan B-L, Murphy KL, Ransom RW, Chang RSL, Tang C, Prueksaritanont T, Detwiler TJ, Hettrick LA, Landis ER, Leonard YM, Krueger JA, Lewis SD, Pettibone DJ, Freidinger RM, Boc MG (2006) J Med Chem 49:1231–1234 DOI 10.1021/jm0511280
26. Zhao H (2007) Drug Discov Today 12:149–155 DOI 10.1016/j.drudis.2006.12.003
27. Clark T (2006) ParaSurf, Cepos Insilico, Erlangen, Germany (http://www.ceposinsilico.com)
28. Daylight Toolkit 4.7, Daylight Chemical Information Systems, Aliso Viejo, CA (http://www.daylight.com)
29. Gasteiger J, Rudolph C, Sadowski J (1990) Tetrahedron Comp Method 3:537–547
30. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909 DOI 10.1021/ja00299a024
31. VAMP (Version 9.0), Accelrys, San Diego, CA (http://www.accelrys.com)
32. Ritchie DW, Kemp GJL (1999) J Comput Chem 20:383–395 DOI 10.1002/(SICI)1096–987X(199903)20:4
33. Cai W, Zhang M, Maigret B (1998) J Comput Chem 19:1805–1815 DOI 10.1002/(SICI)1096–987X(199812)19:16
34. van Vrie JH (1997) J Chem Inf Comp Sci 37:38–41 DOI 10.1021/ci960464
35. van Vrie JH, Nugent RA (1998) SAR and QSAR in Environ Res 9:1–21
36. Erikson J, Neidhart DJ, van Vrie JH, Kempf DJ, Wang XC, Norbeck DW, Plattner JJ, Rittenhouse JW, Turon M, Wideburg N et al (1990) Science 249:527–533 DOI 10.1126/science.2200122
37. OEChem Version 1.3.3, OpenEye Scientific, Santa Fe, NM (http://www.eyesopen.com)
38. MOE, Chemical Computing Group, Montréal, Canada (http://www.chemcomp.com)
39. DeLano WL (2002) PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA (http://www.pymol.org)
40. Omega Version 2.0, OpenEye Scientific, Santa Fe, NM (http://www.eyesopen.com)
41. Kirchmaier J, Wolber G, Laggner C, Langer T (2006) J Chem Inf Mod 46:1848–1861 DOI 10.1021/ci060084g
42. Politzer P, Murray J, Concha M (2007) J Mol Model 13:643–650 DOI 10.1007/s00894-007-0176-9
43. Rauhut G, Clark T (1993) J Comput Chem 14:503–509 DOI 10.1002/jcc.540140502
44. Hassel O (1970) Science 170:497–502 DOI 10.1126/science.170.3957.497
45. Auffinger P, Hays FA, Westhof E, Hing Ho P (2004) Proc Natl Acad Sci USA 101:16789–16794 DOI 10.1073/pnas.0407607101
46. Frisch MJ et al (2004) Gaussian 03, Revision C.02. Gaussian, Wallingford, CT
47. Becke AD (1993) J Chem Phys 98:5648–5652 DOI 10.1063/1.464913
48. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785–789
49. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) J Phys Chem 98:11623–11627 DOI 10.1021/j100096a001
50. McLean D, Chandler GS (1980) J Chem Phys 72:5639–5648 DOI 10.1063/1.438980
51. Krishnan R, Binkley JS, Seeger R, Pople JA (1980) J Chem Phys 72:650–654 DOI 10.1063/1.438955
52. Binning RC Jr, Curtiss LA (1990) J Comp Chem 11:1206–1216 DOI 10.1002/jcc.540111013
53. Curtiss LA, McGrath MP, Blaudeau J-P, Davis NE, Binning RC Jr, Radom L (1995) J Chem Phys 103:6104–6113 DOI 10.1063/1.470438
54. McGrath MP, Radom L (1991) J Chem Phys 94:511–516 DOI 10.1063/1.460367
55. Clark T, Chandrasekhar J, Spitznagel GW, Schleyer PvR (1983) J Comp Chem 4:294 DOI 10.1002/jcc.540040303
56. Frisch MJ, Pople JA, Binkley JS (1984) J Chem Phys 80:3265–3269 DOI 10.1063/1.447079
57. Laaksonen L (1992) J Mol Graph 10:33–34
58. Bergman DL, Laaksonen L, Laaksonen A (1997) J Mol Graph Model 15:301–306 DOI 10.1016/S1093–3263(98)00003–5
59. gOpenMol, CSC, Espoo, Finland (http://www.csc.fi/gopenmol)
60. Rocs Version 2.3.1, OpenEye Scientific, Santa Fe NM (http://www.eyesopen.com)
61. Rarey M, Zimmermann M, Hindle S, Feature Trees version 1.5.2 (Biosolve It, http://www.biosolveit.de)