COSMO*sim3D*: 3D-similarity and alignment based on COSMO polarization charge densities

Michael Thormann,[1] *Andreas Klamt,*[*,2,3] *Karin Wichmann,*[2]

Abstract

COSMO σ-surfaces resulting from quantum chemical calculations of molecules in a simulated conductor, and their histograms, the so-called σ-profiles, are widely proven to provide a very suitable and almost complete basis for the description of molecular interactions in condensed systems. The COSMO*sim* method therefore introduced a global measure of molecular similarity based on similarity of σ-profiles, but it had the disadvantage of neglecting the 3D distribution of molecular polarities which is crucially determining all ligand-receptor binding. This disadvantage is now overcome by COSMO*sim3D*, which is a logical and physically sound extension of the COSMO*sim* method, which uses local σ-profiles on a spatial grid. This new method is used to measure intermolecular similarities based on the 3D representation of the surface polarization charge densities σ of the target and the probe molecule. The probe molecule is translated and rotated in space in order to maximize the sum of local σ-profile similarities between target and probe. This sum, the COSMO*sim3D* similarity, is a powerful descriptor of ligand similarity and allows for a good discrimination between bioisosters and random pairs. Validation experiments using about 600 pharmacological activity classes in the MDDR database are given. Furthermore, COSMO*sim3D* represents a unique and very robust method for a field-based ligand-ligand alignment.

Introduction

In the framework of the Conductor-like Screening Method for Realistic Solvation (COSMO-RS[1-3]), which allows for the prediction of a broad range of fluid phase thermodynamic properties based on quantum chemical calculations for solutes and solvents, the COSMO[4] surface polarization charge density σ, and its molecular surface histograms, the so-called σ-profiles, have been proven to be excellent descriptors for the quantification of the most important kinds of molecular interactions in the liquid phase, such as polar interactions, hydrogen bonding and hydrophobicity.[5-8] For the reader not familiar with the concept of σ-profiles, detailed introductions can be found in any of the COSMO-RS publications,[1-3] and in special detail in ref. 1 (pages 85 and following). The exceptional suitability of σ for the quantification of hydrogen bond interactions has been further confirmed in a recent quantum chemical study.[9] Since the same intermolecular interaction modes, which govern fluid phase thermodynamics, are also responsible for ligand-receptor-interactions, it is most plausible that a σ-based description of ligand-ligand similarity or ligand-receptor-interactions should be very promising. Based on these considerations we presented COSMO*sim*,[10,11] which uses the molecule-specific global σ-profile disregarding the spatial distribution of the polarization charge density to measure ligand-ligand similarities. This method was shown to provide useful discrimination of bioisosteric and random ligand pairs, especially for smaller molecules. Besides speed, one of the major advantages of COSMO*sim* is that it naturally supports scaffold hopping by using the molecular COSMO-RS σ-surface instead of the molecular structure. Furthermore, analogy-based QSPR based on COSMO*sim* delivers powerful models for properties that are mainly governed by isotropic interactions, like logS, logP, logBB, etc.

However, the selective binding of ligands to receptors is known to be based on multiple strong interactions, and the 3D arrangement of the interaction sites of a ligand thus plays a crucial role. Such information is not included in ordinary σ-profiles. As a result, in anisotropic protein-ligand interactions, COSMO*sim* tends to retrieve false-positives along with true-positives. This clearly results from the fact that the global σ-profiles do not contain any information about the spatial distribution of the polarization charge densities on the molecular surface. The use of a grid of local σ-profiles should overcome this deficiency. In order to generate 3D σ-profiles, we project the COSMO-σ-surface onto a 3D grid of a certain resolution and thus generate local σ-profiles (LSPs) for the grid points (see Figure 1 and Methods section). While the target molecule is fixed on the grid, the probe molecule is represented in different orientations and is translated and rotated to achieve the maximum overlap of the local σ-profile similarities. In this way, the optimal alignment and the COSMO*sim3D* similarity are obtained.

COSMO*sim3D* requires COSMO-σ-surfaces of target and probe molecules which can be obtained from a DFT calculation using different quantum chemical programs with suitable COSMO implementations. The TURBOMOLE/COSMO implementation[7,8] is used throughout this paper. It is advisable to take multiple conformers of each ligand into account, as their spatial σ-profiles are much more sensitive to conformational changes than the global σ-profiles used in the original COSMO*sim* method.

DFT calculations of large scale databases can still be very time-consuming, in particular with geometry optimization. We found that proper conformation analysis at MM level or at semiempirical QM-level followed by a single point COSMO calculation at BP-SVP-COSMO-SP level delivers good-quality σ-surfaces in reasonable time, i.e. within a few minutes per conformer. By a recent update, the COSMO*frag* program[12,13] is able to provide approximate COSMO σ-surfaces (CF-COSMO-files) based on just a 3D-geometry of the ligand and the COSMO*frag* database of precalculated σ-profiles. This shortcut reduces the computational time for the COSMO file generation to less than a second per conformer, and thus makes the COSMO*sim3D* method presented here suitable for high-throughput screening applications.

[*] Tel: +49-2171-731681; fax: +49-2171-731689. E-mail: klamt@cosmologic.de

[1] Origenis GmbH, Am Klopferspitz 19A, 82152 Martinsried, Germany

[2] COSMO*logic* GmbH and Co. KG, Burscheider Str. 515, 51381 Leverkusen, Germany.

[3] Institute of Physical and Theoretical Chemistry, University of Regensburg, 93053 Regensburg, Germany

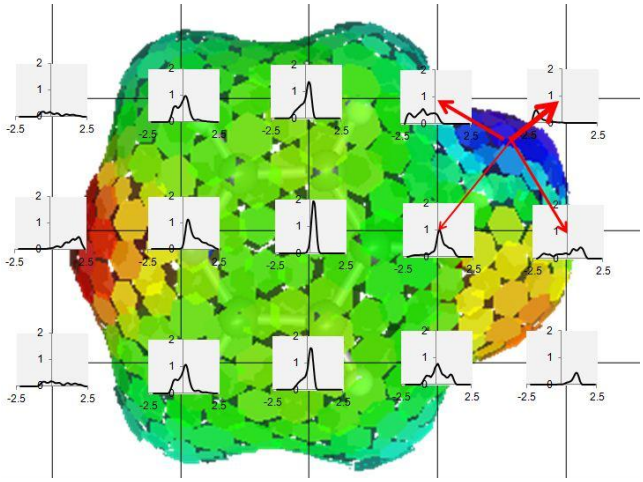

Figure 1. Schematic visualization of the construction of local σ-profiles, in 2D for clarity. Each surface segment gets assigned to the local σ-profiles of the neighboring grid points.

In the following we will first present the COSMO*sim3D* method in detail. Then we will describe the computational methods used throughout three computational experiments. Then we discuss the results of a broad validation study of the COSMO*sim3D* method on the discrimination of true and random bioisosteric pairs from the BioSter database.[14] This is followed by a large scale study on identification of compounds with identical activity classification in the MDDR database.[15] As a final validation we present the results of an automated alignment of ligands using COSMO*sim3D*, followed by a summary and outlook.

Methods

COSMO*sim3D* method description. The starting point of the COSMO*sim3D* method are COSMO files generated by quantum-chemical calculation combined with the continuum solvation model COSMO (with ε = infinity, i.e. in the conductor limit). These files contain the information about the position (x_i, y_i, z_i), the areas a_i, and the COSMO polarization charges q_i for all COSMO surface segments. For each segment i a locally averaged polarization charge density σ_i is calculated according to the standard procedure,[1] using an averaging radius of 0.5 Å. For standard σ-profiles, as used in the COSMO*sim* method, a histogram with σ-bin width of δ^σ = 0.1 e/nm^2 is generated from all segments of a molecule. For the generation of a σ-profile the area of each segment i is associated to the two neighboring σ-grid centers of the actual value σ_i, denoted as σ_{i+} and σ_{i-}, according to σ-distance weights (eq. 1).

$$w_{i+}^{\sigma} = (\sigma_{i+} - \sigma_i)/\delta^{\sigma}$$
$$w_{i-}^{\sigma} = (\sigma_i - \sigma_{i-})/\delta^{\sigma} \qquad (1)$$

This weighting generates a smooth and charge conserving assignment of the COSMO surface segments to the histogram, ensuring that the integral of the σ-profile is the total surface of the molecule, and that the σ-weighted integral, i.e. the first moment of the histogram, is the sum of the original COSMO charges, i.e. the negative of the total charge of molecule.

Instead of generating just one such one-dimensional σ-profile for the entire molecules, in COSMO*sim3D* we generate a local one dimensional σ-profile at each position of a regular 3D grid, i.e. altogether a 4-dimensional histogram, with three Cartesian dimensions (x,y,z) and σ as fourth dimension. The grid size in space is set to $\delta^x = \delta^y = \delta^z = 1$ Å by default and used this way throughout this paper. For the 16 neighboring grid points of a segment i with coordinates (x_i,y_i,z_i,σ_i) in the 4D space, the weights w^x, w^y, w^z, w^σ are computed in complete analogy to eq. 1, and the segment area is assigned to the 16 neighbor grid points according to the product of the four weights. This weighting ensures a smooth linear interpolation of the segment assignment if a surface segment is moved between the grid points.

In COSMO*sim3D* the first molecule, called target further on, is initially shifted in a way that its COSMO surface center is located on the center of the cubic grid, followed by the calculation of the 3D-σ-profile (sp3d1) on the grid points. Then the same is done for the second molecule, called probe further on, resulting in a 3D-σ-profile sp3d2. Next the 3D-σ-similarity COSMO*sim3D* is calculated as a weighted sum over the σ-similarities of the local σ-profiles according to eq. 2,

$$\text{COSMO}sim3d(sp3d1,sp3d2) = \frac{\sum_{ix,iy,iz} (a1(ix,iy,iz) + a2(ix,iy,iz))sms(sp3d1(ix,iy,iz),sp3d1(ix,iy,iz))}{\sum_{ix,iy,iz} (a1(ix,iy,iz) + a2(ix,iy,iz))} \qquad (2)$$

where $a1(ix,iy,iz)$ and $a2(ix,iy,iz)$ are the total intensities of the local target and probe σ-profiles at the grid point, respectively. The denominator is identical to the sum of the total surface areas of target and probe. The SMS (sigma-match similarity) calculations are performed according to reference 10, with the default parameters derived in that paper.[12]

After the initial evaluation of COSMO*sim3D*, the position and orientation of the probe is optimized in order to maximize COSMO*sim3D*. This is done by a trial-and-error line search in the direction of each of the 3 unit-translation and unit-rotations, with minimum steps of 0.01 Å and 0.1°, respectively. After each translational or rotational step COSMO*sim3D* is re-evaluated and the step is accepted if it leads to an increase of COSMO*sim3D*. After convergence of the optimization more optimizations are performed, each starting from another start position and start orientation of the probe. During our test we found that 25 reasonably chosen rotations were sufficient to find the optimal superposition and the maximum COSMO*sim3D* value in essentially all cases. After tuning the optimization algorithm (for more details see Appendix A), the typical time demand for the evaluation of the optimal COSMO*sim3D* similarity for a pair of drug-like molecules thus is in the order of 5 s on a single 2.5 GHz CPU.

COSMO file generation. COSMO files were generated in two ways. In the conventional way of COSMO file generation we used the TURBOMOLE program[16,17] for a single-point BP-SVP-COSMO density functional calculation in combination with the COSMO solvation model implemented in TURBOMOLE. These calculations typically take about 3-5 min. for one drug molecule on a single 2.5 GHz CPU. Alternatively we used the novel

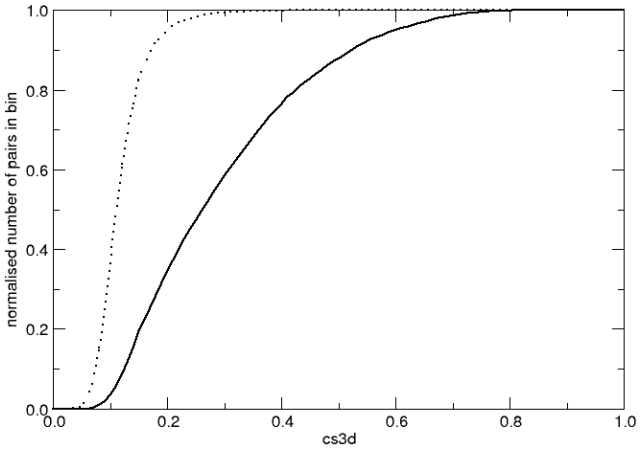

Figure 2. cs3d separates random pairs (dotted line) from bioisosteric pairs (solid line).

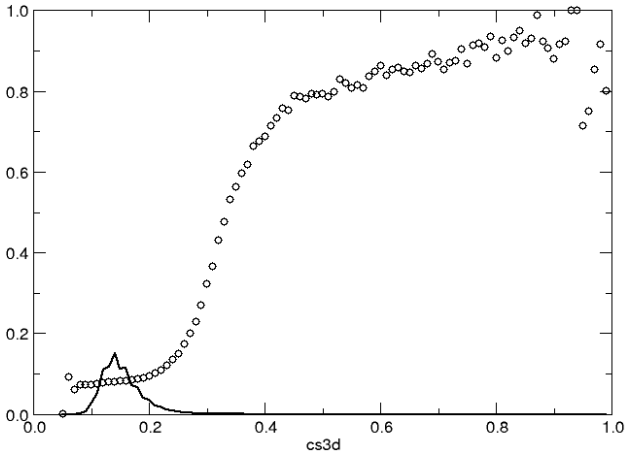

Figure 3. Fraction of pairs sharing at least one MDDR activity class as function of cs3d after filtering with the limit COSMOsim > 0.9.

CF-COSMO capability of COSMO*frag*,[13,14] which generates approximate COSMO files from fragments taken from the COSMO*frag* database (CFDB) of precalculated COSMO files. Technically this is done in the same way as COSMO*frag* usually generates approximate σ-profiles, but with the difference that for the CF-COSMO file generation a 3D-input geometry is required. The CF-COSMO module takes for each atom of the input structure the COSMO surface segments from the most suitable precalculated compound stored in the CFDB and translates and rotates this into the local coordinate system of the target atom. In this way COSMO files are generated which do not represent a perfectly closed molecular surface, but which have roughly the right surface segments and polarization charge densities in roughly the right spatial position. For more details of the CF-COSMO method and a comparison of CF-COSMO files with original COSMO files, see Appendix A.

Tautomer and conformer generation. SDF versions of the BioSter and MDDR databases were converted into SMILES using OpenBabel.[18] The chemical structures were desalted and neutralized, if possible. Tautomers were generated for each compound based on the canonical SMILES using an in-house program. For each tautomer up to three main conformers were generated using the Msmab and Mcnf options of the MOLOC program.[19]

COSMO*sim* calculations. The COSMO*sim* module in the COSMO*frag* program[10,13] was used to compute the complete similarity matrix of 214,513×214,513 compounds. For each compound, a list with the potential bioisosters was generated using a cutoff for the COSMO*sim* value of > 0.9, which corresponds to a reasonable cutoff derived from our previous work.[10] This preselection was then used to run COSMO*sim3D* on these tentative bioisosteric pairs of compounds.

COSMO*sim3D* calculations. COSMO*sim3D* was used as described above to calculate the σ-surface-based similarity of the tentative bioisosteric pairs, whereby all conformers of all tautomers of the target compound where compared with all conformers of all tautomers of the probe compound. The highest COSMO*sim3D* value obtained from this procedure was assigned as COSMO*sim3D* value to the corresponding compound pair. These calculations were done in two ways, based on the original COSMO files and with the approximate CF-COSMO files.

Results and discussion

Validation study on bioisosteric pairs

A screening experiment on the separation of true and false (random) bioisosteric pairs was performed using the 5,089 bioisosteric pairs and the 5,589 random pairs provided in the BioSter database.[14] In this test only the main conformation was used for each compound. It should be noted, that the random pairs are not proven to be non-bioisosteric. Hence a small percentage of true bioisosteric pairs may be contained in this subset. The COSMO*sim3D* similarity (cs3d) was calculated for all pairs of the two subsets. Figure 2 shows the accumulated fraction of pairs as a function of cs3d. For the random pairs the maximum slope is at ~0.1, and 95% of the bioisosteric pairs have cs3d < 0.2. The curve for the true bioisosteric pairs is clearly shifted towards higher cs3d-values. Its maximum slope is at ~0.25. A separation at a value of 0.2 would leave ~5% false positives in the set of random pairs, at the cost of ~30% false negatives in the bioisoster pair subset.

Validation study on recognition of activity classes

The MDDR database[15] was used to extract 214,513 chemical structures and their assignment to 580 different activity classes (ACTIV_INDEX). Each compound may be assigned to more than one class. Canonical SMILES strings were generated using OpenBabel. Approximate σ-profiles, one per compound, were generated using COSMO*frag* based on the canonical SMILES. In addition, tautomers and conformers were generated for each compound along with their CF-COSMO files, as described in the Methods section.

As a first step, for all 214,513×214,512/2 pairs, i.e. for $23 \cdot 10^9$ pairs, the COSMO*sim* similarity was calculated. Pairs with COSMO*sim* < 0.9 were skipped. 59,144,819 of the pairs passed this threshold. In the original dataset the probability that a pair shares at least one activity class was 6.4%. In the remaining dataset this probability had increased to 24.2%, which corresponds to an overall enrichment factor of 3.8 resulting from the COSMO*sim* > 0.9 filter.

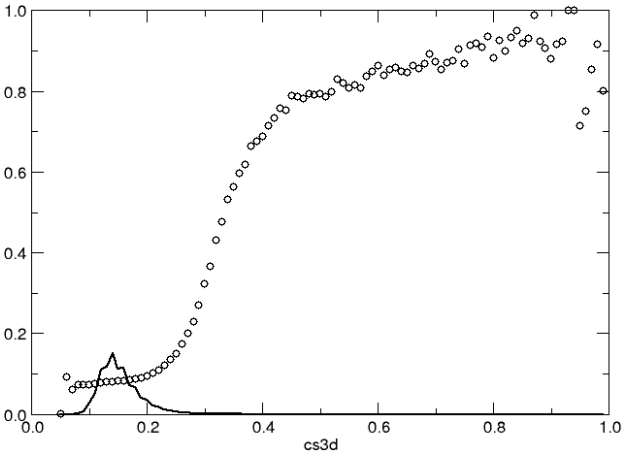

Figure 4. Enrichment factors of pairs belonging to the same MDDR activity class as function of cs3d for the 26 most populated classes, after filtering with the limit COSMOsim > 0.9. The MDDR activity class code is given in the inset. Explanations of the MDDR activity class codes are given in Table S1 in the supporting information.

Originally we planned to calculate the COSMO*sim3D* similarity (cs3d) for all of the 59,144,819 remaining pairs. Since in average 2.8 tautomeric/conformeric forms of each compound, i.e. ~8 combinations per compound pair, had to be considered, this would have corresponded to ~$480 \cdot 10^6$ cs3d evaluations. Finally this turned out to be too time-consuming. Observing no significant changes in the fraction of activity pairs anymore (see Figure 3), and having spent already 180 days on ~50 CPUs, we stopped the experiment after the evaluation of ~30% of these pairs. It is worth noting that the average time demand per cs3d-evaluation for a pair thus was only ~5 s on one CPU.

Figure 3 shows the overall statistics from this experiment. The solid line shows the distribution of cs3d values within all tested pairs. Although the COSMO*sim* similarity for all pairs was > 0.9, the cs3d for most pairs is only in the range of 0.1-0.2. The dots mark the percentage of pairs sharing at least one activity class at each binned value of cs3d. At low cs3d values, i.e. for the largest part of the pairs, this value is only in the range of 7-10%, i.e. essentially not much higher than the value of 6.4% for random pairs from the MDDR database. A strong increase can be observed at cs3d ~0.3. 50% chance for sharing at least one activity class is reached at cs3d = 0.33 and 80% are reached at cs3d = 0.45, further increasing to 100% at higher values of cs3d, about with increasing statistical noise, since the number of pairs having such large values of cs3d becomes very small.

Figure 4 shows the cs3d-dependence of the individual enrichment factors for the 26 most populated activity classes in MDDR compared to the population of these pairs in the complete MDDR. All enrichment factors strongly increase in the range of 0.1-0.3, and mainly reach a plateau at higher values. In detail, interesting differences between the different classes can be observed with respect to the initial value, i.e. the enrichment factor already achieved through the COSMO*sim* > 0.9 pre-selection, with respect to the exact point of maximum increase and the maximum achievable enrichment factor, but an investigation and discussion of these details is beyond the scope of this paper.

COSMO*sim3D* alignment study

Obviously COSMO*sim3D* can be used not only for pairwise alignment, but also for the alignment of larger data sets. For such we have additionally implemented a self-consistent alignment mode, in which the first *m* of a list of *n* compounds can be defined as targets. In this mode COSMO*sim3D* starts with the alignment of compound C2 *vs* compound C1. Then it simultaneously aligns compound C3 *vs* (C1+C2)/2, which means that on each grid point the average of the similarities to the two previously aligned compounds is used. This is continued until to the last target, i.e. C*m*, which is aligned *vs* the average of the *m*-1 previous targets. Then we start again at C1, align it to the other *m*-1 targets, and continue until no significant increase in the overall similarity is observed throughout an entire loop. Finally, the remaining *m*-*n* compounds get aligned *vs* the average of the *n* target compounds. Furthermore, in order to avoid any potential bias resulting from starting with the C1-C2 pair, a super-self-consistent alignment mode was introduced, in which the procedure of the self-consistent alignment is repeated, starting with all subsequent pairs instead of starting just with C2 *vs* C1 alignment.

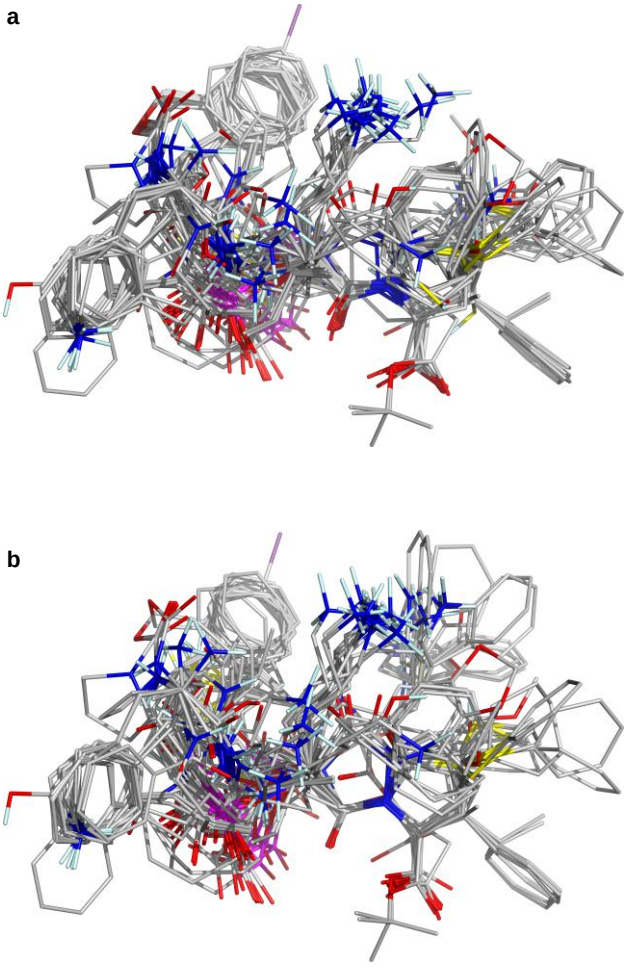

Figure 5. Alignment of the 114 ACE ligands from the Sutherland dataset: a) generated by COSMOsim3D with super-self-consistent treatment of the 3 target ligands; b) alignment as generated by Sutherland.

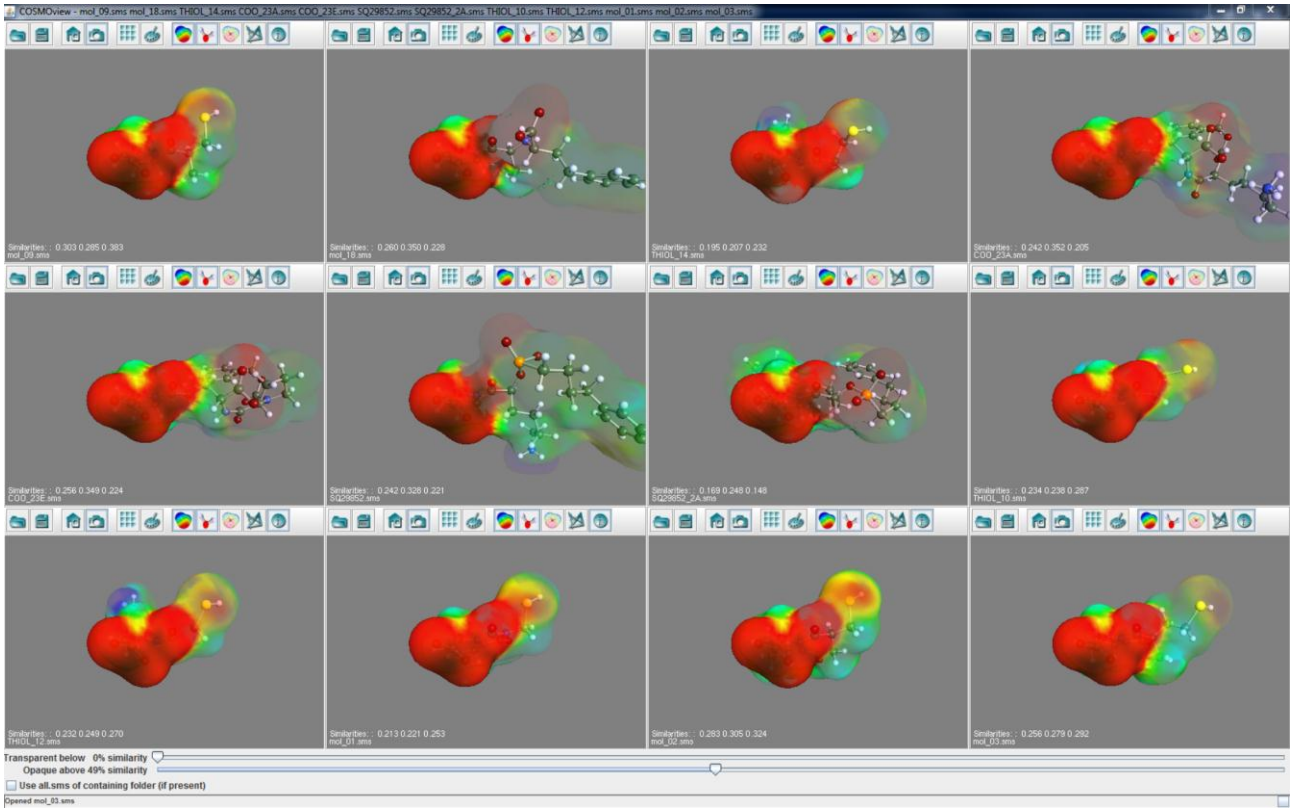

Figure 6. Screenshot of the cs3d-multiview mode in COSMO*view*. The same 3D-rotation and zoom operations apply to all visualized molecules; the degree of surface transparency depends on the local values of the cs3d similarity.

The best of the resulting self-consistent alignments of the targets is then used in order to align the remaining compounds. As an example we applied this automated unsupervised alignment to the ACE dataset from the Sutherland data collection.[20] For all 114 molecules, COSMO files were generated by single-point DFT/COSMO calculations, conserving the geometries given by Sutherland. After randomization of the positions and orientations of all compounds, the alignment was carried out as a super-self-consistent alignment with respect to the three molecules considered as targets by Sutherland on this dataset, i.e. with MOL_09, MOL_18, THIOL_14. We used a grid size of 1.0 Å and 80 starting orientations per molecule.

The resulting alignment is shown in Figure 5a, with the Sutherland alignment given for comparison in Figure 5b. At least visually, the cs3d alignment appears to be more consistent than the alignment provided by Sutherland. The performance of these alignments in the context of 3D-QSAR will be described in a forthcoming paper.[21]

On the example of 12 compounds from this cs3d-aligned ACE data set, starting with the 3 targets and completed with representatives of the different compound classes, we show in Figure 6 a novel visualization mode of the 3D similarity of molecules, called cs3d-multiview mode further on. The COSMO surfaces with embedded ball-and-stick structures get visualized for *k* compounds in *k* separate windows, which are subject to the same 3D-rotation and zoom operations. The local values of the cs3d similarity are used in order to control the surface transparency. By default regions with high similarity are shown with low transparency and thus have intensive colors, while areas with low similarity get high transparency, so that in these regions mainly the ball-and-stick structure can be seen. We consider this visualization mode of molecular similarity as an interesting auxiliary tool for investigating sets of aligned compounds, although we are aware that the number of compounds which can be visualized simultaneously in this way is limited by the considerable memory requirements.

Conclusions

By the presented COSMO*sim3D* method, the rich and consistent information content of the COSMO surface polarization charge densities with respect to all physiologically relevant intermolecular interaction modes can be used for a 3D-alignment and 3D-similarity measure, which allows for a very good separation of true bioisosteric pairs from random pairs. Hence COSMO*sim3D* can be used as a powerful novel tool to search for bioisosteric analogues of known biologically active compounds. A special strength is the sound theoretical and almost ab initio foundation of the underlying COSMO σ-surfaces, which allow for their application to very different and novel chemical situations. Another advantage is the exclusive dependence of COSMO*sim* on the surface polarization charges, allowing for the detection of physiological similarity of chemically very different structures and hence for scaffold hopping.

Outlook

In this article we have restricted ourselves to report the application of COSMO*sim3D* to the alignment and similarity measure of pairs of potential ligand molecules. Based on the simple rule, that the best interaction partner for a molecular surface segment with polarization charge density σ is a segment with the opposite polarization charge density, or simply the rule "σ likes $-\sigma$", it is straightforward to extend COSMO*sim3D* to the alignment and similarity measure of ligand candidates to the inverted sigma profiles of enzyme receptor areas. The achieved COSMO*sim3D* similarity can in this way be considered as a measure of the potential interaction free energy, which corresponds to the pK_i value for the enzyme inhibition. Such applications will be investigated in future. Another straightforward extension to the COSMO*sim3D* method is the usage of 3D-σ-profiles of a set of aligned ligands as basis for a 3D-QSAR analysis, e.g. by the molecular field analysis approach, in order to generate a ligand-based model for the prediction of pK_i, where the alignment may be performed with the COSMO*sim3D* method. Details and benchmark results for this method, which we call COSMO*sar3D*, will be reported soon in a forthcoming paper.[21]

Supporting Information

More detailed information about the 26 MDDR activity classes considered in Figure 4 are available as Table S1.

Appendix A: COSMO*frag*

The COSMO*frag* method[11] has been described as a fast shortcut for COSMO-σ-profile based calculations, which replaces the relatively time-consuming quantum chemical DFT/COSMO calculation by a fast generation of an approximate σ-profile, composed of partial, atom-wise σ-profiles taken from structurally most similar atoms, which are retrieved from a database of meanwhile ~60,000 chemically most diverse drug-like compounds with pre-calculated DFT-COSMO σ-profiles, the so-called COSMO*frag* database (CFDB). Based on just a chemical structure input like SMILES, such COSMO*frag* σ-profile generation takes only a fraction of a second.

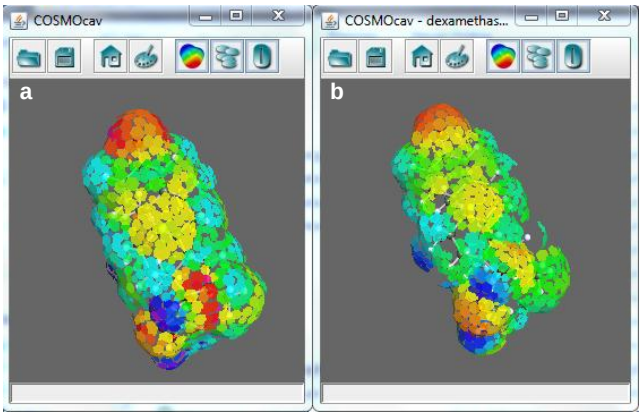

Figure A1. Comparison between a true COSMO surface (a) and a *fcos* surface generated by COSMO*frag* (b) for dexamethasone. While the two surfaces are overall remarkably similar, a few small open regions can be observed in the *fcos* surface.

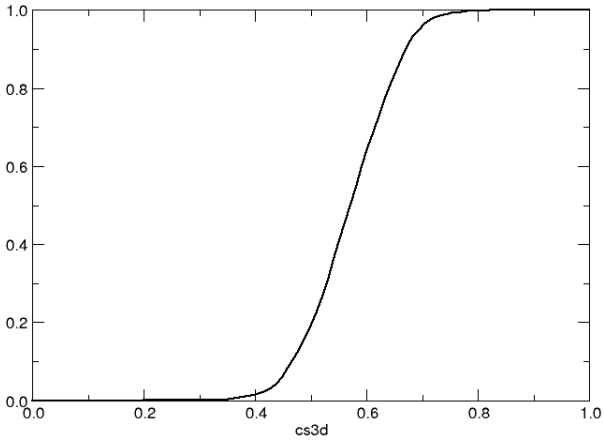

Figure A2. Self-similarity test of CF-COSMO files *vs* the corresponding original COSMO files for the 11,280 compounds of the BioSter study (see text for details).

For the purpose of COSMO*sim3D* calculations we have extended the COSMO*frag* technique towards the generation of approximate 3D-COSMO files, marked by the file extension *fcos*. Based on an input 3D-structure of a new molecule *M*, this CF-COSMO extension of COSMO*frag* for each atom *i* of *M* selects the most appropriate representation *j* from the database in the usual way of COSMO*frag*. It opens the compressed COSMO file (*ccf*) of the respective molecule containing atom *j*, which is in the CFDB, and takes the surface segments belonging to *j* out of the local coordinate system of atom *j* and moves them into the corresponding local coordinate system of the target atom *i* by a linear transformation, consisting of a translation and rotation. Since the *i* and *j* have a similar chemical neighborhood, the two local coordinate systems of *i* and *j* have a meaningful relation to each other, if they are based on the local directions to most similar nearest neighbors. As a result, the surface segments taken from the reference atom *j* end up approximately at the correct position in the surroundings of the target atom *i*. In this way the surface of molecule *M* will at the end be filled with COSMO-σ-surface segments which have roughly the correct σ-values at roughly the right positions. In contrast to original COSMO files, these segments will not be strictly surface filling, but leave some smaller open regions, while in other regions segments may be overlapping (See Figure A1). Therefore, it is important that the local σ-averaging is performed on the reference molecules, because σ-averaging on the CF-COSMO files might yield artifacts.

In Figure A2 we show the results of a self-similarity test of CF-COSMO files *vs* the corresponding original COSMO files for the 11,280 compounds considered in the BioSter study described above. Almost all cs3d values are in the range 0.4-0.65, with the maximum at about cs3d = 0.55. This is clearly separated from the critical cs3d range of 0.2-0.3, which according to the examples considered above appears to be essential for the identification of bioisosterism. Therefore, bioisosterism can almost equally well be identified based on CF-COSMO files as on original COSMO files.

We confirmed this finding by repeating the bioisosteric pair experiment described above with CF-COSMO files instead of the original COSMO files. The corresponding curves for the accumulated fractions of pairs are very close (within 2%) to those obtained with the full COSMO files.

References

1. Klamt, A. *COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design*; Elsevier: Amsterdam, 2005.
2. Klamt, A. Conductor-Like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224-2235; doi.
3. Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and parameterization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074-5085; doi.

4. Klamt, A.; Schüürmann, G. COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perk. T. 2* **1993**, 799-805; doi.
5. Klamt, A.; Eckert, F.; Arlt, W. COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures. *Ann. Rev. Chem. Biomol. Eng.* **2010**, *1*, 101-122; doi.
6. Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: a novel view to physiological solvation and partition questions. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 355-365; doi.
7. Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **2002**, *23*, 275-281; doi.
8. Klamt, A.; Diedenhofen, M.; Jones, R.; Connolly, P. C. The use of surface charges from DFT calculations to predict intestinal absorption. *J. Chem. Inf. Model.* **2005**, *45*, 1337-1342; doi.
9. Klamt, A.; Reinisch, J.; Eckert, F.; Hellweg, A.; Diedenhofen, M., Polarization charge densities provide a predictive quantification of hydrogen bond energies. *Phys. Chem. Chem. Phys.* **2011**, *14*, 955-963; doi.
10. Thormann, M.; Klamt, A.; Hornig M.; Almstetter M. COSMO*sim*: bioisosteric similarity based on COSMO-RS σ-profiles. *J. Chem. Inf. Model.* **2006**, *64*, 1040-1053; doi.
11. Please note that in eq. 11 of ref. 10 the denominator was erroneously given as square-root of the self-similarities SMS1 and SMS2, while it should have been the maximum of SMS1 and SMS2.
12. Hornig, M.; Klamt, A. COSMO*frag*: A novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1169-1177; doi.
13. COSMO*frag*, version 3.3; COSMOlogic GmbH&CoKG: Leverkusen, Germany, 2011.
14. *BIOSTER database*, version 2010.1, Accelrys: San Diego, CA, 2010.
15. *MDDR database*, version 2010.1, Accelrys: San Diego, CA, 2010.
16. *TURBOMOLE*, version 6.3, COSMOlogic GmbH&CoKG: Leverkusen, Germany, 2011; TURBOMOLE is a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; see also URL: http://www.turbomole.com.
17. Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. COSMO implementation in TURBOMOLE: extension of an efficient quantum chemical code towards liquid systems. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2187-2193; doi.
18. OpenBabel, version 2.2.3; O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33; doi.
19. *MOLOC*, version 2011, Gerber Molecular Design: Amden, Swizerland, 2011.
20. 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; doi.
21. Klamt, A.; Thormann, M.; Wichmann, K.; Tosco, P. COSMO*sar3D*: molecular field analysis based on local COSMO σ-profiles. *J. Chem. Inf. Model.* **2012**; doi.