COSMO*sar3D*: Molecular Field Analysis based on local COSMO σ-profiles

Andreas Klamt,[*,1,2] *Michael Thormann,*[3] *Karin Wichmann,*[1] *Paolo Tosco,*[4]

Abstract

The COSMO surface polarization charge density σ resulting from quantum chemical calculations combined with a virtual conductor embedding has been widely proven to be a very suitable descriptor for the quantification of interactions of molecules in liquids. In a preceding paper, grid-based local histograms of σ have been introduced in the COSMO*sim3D* method, resulting in a novel 3D-molecular similarity measure and going along with a novel property-based molecular alignment method. In this paper we introduce under the name COSMO*sar3D* the usage of the resulting array of local σ-profiles as a novel set of molecular interaction fields for 3D-QSAR, containing all information required for quantifying the virtual ligand-receptor interactions, including desolvation. In contrast to currently used molecular interaction fields, we provide a theoretical rationale that the logarithmic binding constants of ligands should be a linear function of the array of local σ-profiles. This makes them especially suitable for linear regression analysis methods such as PLS. We demonstrate that the usage of local σ-profiles in molecular field analysis inverts the role of ligands and receptor: while conventional 3D-QSAR considers the virtual receptor in potential energy fields provided by the ligands, our COSMO*sar3D* approach corresponds to the calculation of the free energy of the ligands in a virtual free energy field provided by the receptor. First applications of the COSMO*sar3D* method are presented which demonstrate its ability to yield robust and predictive models, which seem to be superior to the models generated based on conventionally used molecular fields.

Introduction

Since the development and introduction of the comparative molecular field analysis technique (CoMFA) by Cramer et al. in the years 1979-1988, the molecular interaction field (MIF)-based statistical analysis approaches (MFA) have become an important branch of computational drug design.[1-3] Even though due to the increasing availability of good quality crystallographic structures of protein targets in the past decade structure-based methods gained more attention, ligand-based statistical approaches are still of considerable importance in lead optimization, taking advantage of binding affinities determined by in vitro assays. The chances of success of a MFA project strongly depend on various factors, among them the amount and quality of the binding information used as input, the statistical analysis method employed, and the quality, completeness and balance of the MIFs used for the quantification of the interactions of the ligands with the virtual receptor. Conventional MIFs are based on force-fields and take into account steric and electrostatic interactions, usually described by the Lennard-Jones potential and the molecular electrostatic potential (MEP), as resulting from the Coulombic field generated by atom-centered point charges. Besides, MIFs have also been derived from more heuristic descriptions of hydrophobic interactions and hydrogen bonding, often based on probe atom or probe group potentials. Good reviews of the CoMFA technique and its variations have been given by Kubinyi[3] and more recently by Verma et al.[4]

In this paper we suggest a novel set of MIFs, the local grid-based COSMO σ-profiles (LSPs), as a promising alternative to force-field based MIFs. Our starting point is the quantum chemistry-based COSMO-RS method,[5,6] which during the past decade has become an important method for the prediction of fluid phase equilibrium constants such as partition coefficients, solubilities, and vapor pressures in many areas of chemistry, biochemistry, and chemical engineering.[7] In COSMO-RS all molecular interactions are quantified as local contact interactions of the surface polarization charge densities σ which arise on a molecular surface, if it is virtually embedded in a conducting medium. These polarization charge densities can nowadays be calculated at moderate costs by a combination of quantum chemical methods with the continuum solvation model COSMO.[8] The COSMO-RS method initially and mostly has been applied to the prediction of free energies of molecules in homogeneous bulk liquids, which are calculated as surface integrals of solvent specific σ-potentials over the σ-surfaces of the solute molecules. The σ-potentials express the affinity of a solvent for molecular surfaces of certain polarity σ, which originates from electrostatic interactions, hydrogen bonding and hydrophobic effects, if one likes to consider the latter as a separate class of interactions, as often done in medicinal chemistry. More recently, straightforward extensions of COSMO-RS to inhomogeneous situations such as interfaces, micelles, and bio-membranes have been reported,[9] which are based on inhomogeneous σ-potentials.

In a preceding paper[10] we have introduced local, grid-based σ-profiles (LSPs), which are four-dimensional histograms describing the amount of ligand surface area within a certain σ-interval and space interval. These LSPs have been shown to be valuable descriptors for the assessment of molecular similarity and alignment. If we now consider a protein receptor together with its aqueous embedding as a locally slightly flexible, and thus locally pseudo-liquid, matrix with locally varying preference for certain surface polarity, i.e. with locally varying σ-potential, then the free energy of a ligand in this receptor matrix should be a surface integral of the locally varying receptor σ-potential over the σ-surface of the ligand. Approximating the surface integration by a grid summation, the free energy of the ligand in the receptor matrix turns into a linear function of the LSPs. Since also the free energy of the ligand in water is of such form, the same must be true for the difference of the free energy of the ligand between the receptor-bound state and its state in

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

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

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

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

[4] Department of Drug Science and Technology, University of Turin, 10125 Torino, Italy.

aqueous solution. Hence, by these minimal assumptions we have come to the result that the free energies of binding, and thus the pK_i values of ligands to a receptor, should be describable as a linear model with respect to the LSPs of the ligands. Therefore, the LSPs should provide an optimally suited set of descriptors for a linear regression analysis, e.g. with partial least squares (PLS),[11] of pK_i values, according to the 3D-QSAR paradigm. To the best of our knowledge, no other set of molecular fields used so far in MFA can claim such a sound theoretical justification for the expectation of a linear pK_i model.

Our approach has additional attractive features:

- As shown in a recent paper,[12] the polarization charge densities σ used in our approach are better suited for the description of hydrogen bonding than the electrostatic potential, which is usually employed in MFA.
- The polarization charge densities σ of neutral and charged species are in the same range, which enables the inclusion of compounds of varying charge states in the same model.
- For the first time a histogram of a property is used as input for MFA, while in all previous approaches the properties are used directly as fields. The usage of smooth spatial histograms introduces an increased robustness of the models against small geometrical shifts of the ligands relative to the grid even at larger grid spacing. If a property has a local hotspot, the latter may be initially located close to a grid point and thus be well represented, but after a small shift it may fall in between the grid points and loose importance. In our histograms, such hotspot would be well described in any case, and its representation would just be smoothly partitioned over the neighboring grid points.
- The recently published COSMO*sim3D* method[10] enables ligand alignment based on the same fields as used for the MFA, i.e. the LSPs.
- These theoretical considerations gave us sufficient confidence to expect that LSP-based MFA might be superior to currently employed MFA approaches.[13] This confidence meanwhile has been supported by the assessment of COSMO*sar3D* performance on a number of publicly available test cases[14] for 3D-QSAR.

In the next paragraphs we provide a formal derivation of the underlying theory, followed by a methods section describing the datasets and the statistical tools employed. Next, we report the performance of different variants of the COSMO*sar3D* method compared to publicly available results of other MFA approaches for the same datasets. Finally, we report the results of several robustness tests.

Theory

Within the COSMO-RS theory[5-7] the free energy of a solute *X* in a homogeneous liquid solvent *S* is given as a surface integral of a homogeneous solvent-specific σ-potential $\mu_S(\sigma)$ over the COSMO σ-surface S^X_{COSMO} of the solute *X*. The latter arises from a quantum chemical calculation, usually at DFT level, for the solute *X* in a virtual conductor embedding, simulated by the conductor-like screening model COSMO:

$$\begin{aligned}\mu_S^X &= \oiint_{S^X_{COSMO}} \mu_S\left(\sigma(\underline{r})\right) d^2\underline{r} + kT\ln\gamma_{comb}(X,S) \\ &= c(S) + \sum_{\nu\in X} a_\nu \mu_S(\sigma_\nu)\end{aligned} \tag{1}$$

On the right side of eq. 1 we have replaced the surface integral by a summation over all surface segments ν of the COSMO σ-surface, with a_ν and σ_ν being the surface area and polarization charge density of a surface segment ν. For a fixed solvent *S* and for solutes of about the same size the last term, the so-called combinatorial contribution, can be safely regarded as a constant $c(S)$. Within the COSMO-RS theory the σ-potential of the pure or mixed solvent can be derived from its σ-profile, which is a histogram of the solvent surface area with respect to the polarization charge density σ. The σ-potential is calculated by a statistical treatment of all possible contacts of the solvent *S* with a surface segment of polarity σ, taking into account the electrostatic and hydrogen bond interactions between molecular surfaces. As a result, the σ-potentials include all information about polar, hydrogen-bond, and hydrophobic interactions of the solvent *S*. Further details can be found in previously published introductions to the COSMO-RS theory.[6,7]

So far we have assumed a homogeneous solvent and hence a homogeneous σ-potential $\mu_S(\sigma)$. Extensions of COSMO-RS have been developed which allow its application to inhomogeneous systems as phase boundaries, micelles, and biomembranes[9] by the introduction of a spatially varying σ-potential $\mu_S(\underline{r},\sigma)$. Due to the known inhomogeneous composition of these systems the local σ-potentials can be derived in a similar way as in the homogeneous case. If we now assume that a protein receptor *R* provides a position-dependent σ-potential which is evaluated on the nodes of a regular grid and on an equidistant set of σ-values, then we have for the free energy of a ligand *L* in the receptor:

$$\begin{aligned}\mu_R^L &= c(R) + \oiint_{S^L_{COSMO}} \mu_R\left(\underline{r},\sigma(\underline{r})\right) d^2\underline{r} \\ &= c(R) + \sum_{\nu\in L} a_\nu \mu_R(\underline{r}_\nu,\sigma_\nu) \cong c(R) + \sum_{\nu\in L} a_\nu \hat{\mu}_R(\underline{r}_\nu,\sigma_\nu)\end{aligned} \tag{2}$$

where $\hat{\mu}_R(\underline{r}_\nu,\sigma_\nu)$ denotes the linear interpolation of the σ-potential at position $\underline{r}_\nu$ and polarity σ_ν from the σ-potential values at the 2^4 = 16 neighboring grid points. On a one-dimensional iso-spaced grid of step size δ_x, the weight of a grid point *ix* at position x_{ix} for the linear interpolation of a function at position *x* is

$$w_x(ix,x) = \max\left\{0, 1 - \left|x - x_{ix}\right| / \delta_x\right\} \tag{3}$$

Using such weights in all four dimensions we get for the interpolation of the σ-potential $\hat{\mu}_R(\underline{r}_\nu,\sigma_\nu)$

$$\begin{aligned}&\hat{\mu}_R(\underline{r}_\nu,\sigma_\nu) \\ &= \sum_{ix,iy,iz,i\sigma} w(ix,r_{\nu x}) w(iy,r_{\nu y}) w(iz,r_{\nu z}) w(i\sigma,\sigma_\nu) \mu_R(ix,iy,iz,i\sigma)\end{aligned} \tag{4}$$

Inserting this into eq. 2 we find

$$\begin{aligned}\mu_R^L &\cong c(R) + \sum_{\nu\in L} a_\nu \hat{\mu}_R(\underline{r}_\nu,\sigma_\nu) \\ &= \sum_{ix,iy,iz,i\sigma} \mu_R(ix,iy,iz,i\sigma) \sum_{\nu\in L} a_\nu w(ix,r_{\nu x}) w(iy,r_{\nu y}) w(iz,r_{\nu z}) w(i\sigma,\sigma_\nu) \\ &= \sum_{ix,iy,iz,i\sigma} \mu_R(ix,iy,iz,i\sigma)\,\mathrm{LSP}^L(ix,iy,iz,i\sigma)\end{aligned} \tag{5}$$

where $LSP^L(ix,iy,iz,i\sigma)$ is the local σ-profile of ligand L, i.e. the local histogram of the COSMO surface area of ligand L with respect to polarity σ in the vicinity of grid point (ix,iy,iz), as it was recently introduced within the COSMO*sim3D* method.[10] Since the same derivation holds for the free energy of ligand L in water (W), even with a position-independent σ-potential, we find for the logarithmic binding constant pK_i of the ligand L from water to receptor R

$$\begin{aligned} pK_i(L,R) &\equiv \frac{\mu_W^L - \mu_R^L}{RT\ln 10} \\ &\cong \frac{c(W)-c(R)}{RT\ln 10} + \sum_{ix,iy,iz,i\sigma} \frac{\mu_W(i\sigma)-\mu_R(ix,iy,iz,i\sigma)}{RT\ln 10} LSP^L(ix,iy,iz,i\sigma) \quad (6) \\ &= c_R^0 + \sum_{ix,iy,iz,i\sigma} c_{WR}(ix,iy,iz,i\sigma)\, LSP^L(ix,iy,iz,i\sigma) \end{aligned}$$

i.e., the pK_i is a linear functional of the LSPs. This means that we have exactly the kind of relation between the pK_i values and the descriptors as it is assumed by most statistical methods applied in MFA, especially by the most widely employed PLS technique. This is an absolutely unique result for MFA, since for no other set of descriptors such a proof for the existence of a linear relationship between the set of descriptors and the pK_i values has ever been derived. We even have a reasonable guess of the highest achievable accuracy: a model based on the last part of eq. 6 should be ideally able to describe logarithmic binding constants with about the same accuracy as standard COSMO-RS does for logarithmic partition coefficients, i.e. with an accuracy of about 0.35 log-units. Obviously, such high precision will hardly be achievable within MFA.

In conventional MFA a number of potential energy fields are evaluated on a regular grid, large enough to embed an aligned set of ligands. For each ligand, each of the fields describes the potential energy of a probe particle of certain specification, e.g. a positive probe atom, a hydrogen bond donor/acceptor, a hydrophobic group, etc., in the field provided by the ligand. This energy is commonly evaluated by force field methods. Then, a relationship (usually linear) between the logarithmic binding constant $pK_i(L,R)$ of the ligand L in the virtual receptor R is assumed, as

$$\begin{aligned} pK_i(L,R) &\equiv \frac{\mu_W^L - \mu_R^L}{RT\ln 10} \\ &\approx \hat{c}_R^0 + \sum_{ix,iy,iz,ip} \hat{c}_{RW}(ix,iy,iz,ip)\, PEF^L(ix,iy,iz,ip) \end{aligned} \quad (7)$$

where the index ip shall denote the various potential energy fields (PEFs). By that, the coefficients $\hat{c}_R(ix,iy,iz,if)$, apart from a factor $-RT\ln 10$, represent a population probability of the grid points with respect to receptor atoms of the various probe types; alternatively, taking into account the desolvation from the bulk water, they need to describe a differential population probability between receptor and water embedding of the ligands. In other words, the coefficients, which finally have to be regressed by a statistical method such as PLS, represent a differential 3D composition histogram. The existence and the right choices of proper potential energy fields allowing for a linear description of the pK_i values are much less obvious than the existence of an effective position-dependent σ-potential as it is assumed in the COSMO*sar3D* methodology.

	Ligand property (provided to PLS)	Virtual receptor property (fitted by PLS)
Conventional MFA	**potential**(s) (often called fields)	**3D-composition** variables
COSMO*sar3D*	**3D-composition** variables (LSPs)	local (i.e., 3D) **σ-potential**

Figure 1. Schematic diagram illustrating the reversed roles of ligands and receptor in conventional MFA and in COSMO*sar3D*.

It is worthwhile to note that there is an inversion of the roles of ligands and receptor as well as of molecular fields and composition variables between conventional 3D-QSAR (eq. 7) and the novel COSMO*sar3D* approach (eq. 6). In fact, as illustrated in Figure 1, in conventional MFA the ligands provide potential energy fields, and the differential receptor/water composition coefficients are regressed as coefficients, i.e. the virtual receptor is considered in the field of the ligands. Conversely, in COSMO*sar3D* composition histograms of the ligands are given as descriptors, and the coefficients of a differential receptor/water free energy functional are regressed, which means that the ligands are considered in the field of the receptor. Therefore, COSMO*sar3D* may literally be considered as a “re-volution” of 3D-QSAR, which may shed new light on the old problem from a very different perspective.

Methods

Coordinates for the datasets used in this work were taken from literature;[14] the same partitioning between training and test set as chosen by the original authors was maintained. PLS and statistical analysis were performed with the Open3DQSAR software,[15] which was adapted in order to read the LSP file format. O3QMFA descriptors (see below) were computed with Open3DQSAR, while LSP-based descriptors where computed with the COSMO*sim3D* program.[10] DFT-level COSMO σ-surfaces were obtained by single point calculations with the quantum chemical software suite TURBOMOLE.[16,17] The BP-SVP-COSMO[18-22] level of theory was used, applying infinity for the value of the dielectric constant. Approximate CF-COSMO σ-surfaces were generated with the COSMO*frag* program.[23,24]

Results and discussion

In order to demonstrate the performance of our new approach we took the eight datasets collected by Sutherland and co-workers,[14] and we applied the above described procedure using the same conformations and alignments as in the original literature. As a starting point we calculated the LSP descriptor matrix based on σ-surfaces resulting from BP-SVP-COSMO calculations, using the same grid spacing of 2.0 Å as used by Sutherland to generate 3D-QSAR models, and our default σ-interval (DELSIG) of 0.1 e/nm^2. The latter leads to partioning of the σ range in 61 bins, and therefore results in 61 MIFs per

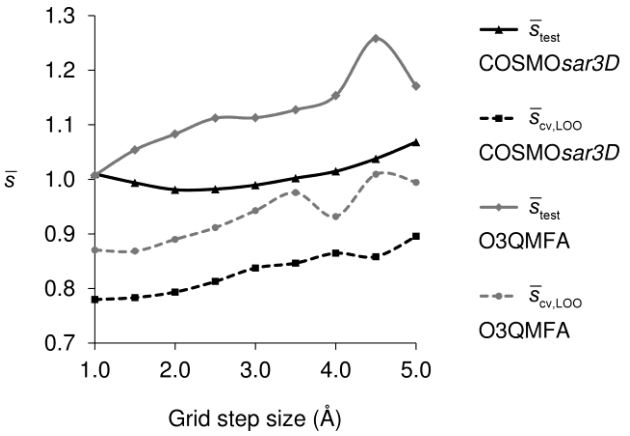

Figure 2. Dependence of the standard deviation of the training (dotted lines) and test set (plain lines) residuals for COSMO*sar3D* (black) and O3QMFA (gray) models upon systematic variation of the grid step size.

grid node. We soon realized that such a high resolution is not required for 3D-QSAR purposes and that DELSIG may be safely increased up to 0.6 e/nm^2 without impacting on model performance, while reducing the size of the descriptor arrays by roughly a factor 6; therefore, a σ-grid spacing of 0.6 e/nm^2 has been set as the default value and used for all further analyses. As usual, the optimal number of PLS components was determined as the one yielding the minimum standard deviation of the error of prediction during leave-one-out (LOO) cross-validation.[2]

For the assessment of the performance of the COSMO*sar3D* method relative to the seven conventional 2D/3D-QSAR methods considered by Sutherland, we will consider as indicators the average standard deviations of the error of internal ($\bar{s}_{cv,LOO}$) and external ($\bar{s}_{test}$) predictions over the eight datasets, and the respective average correlation coefficients ($\bar{r}^2_{test}$ and $\bar{q}^2_{LOO}$). Since external prediction is the ultimate purpose of QSAR models, we consider $\bar{s}_{test}$ as the most important among the four indicators.

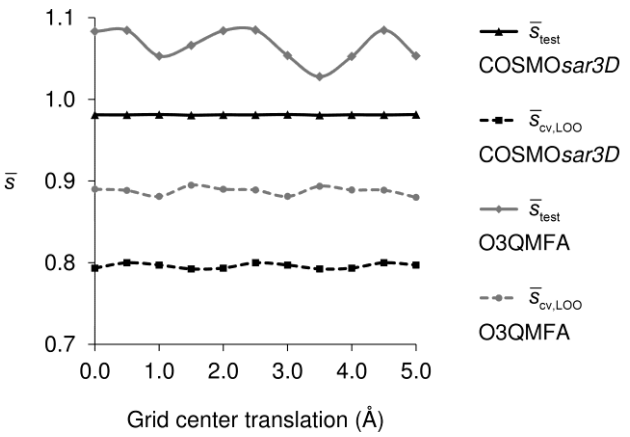

Figure 3. Dependence of the standard deviation of the training (dotted lines) and test set (plain lines) residuals for COSMO*sar3D* (black) and O3QMFA (gray) models upon systematic translation of the grid center position (0.5-5.0 Å on the three Cartesian axes).

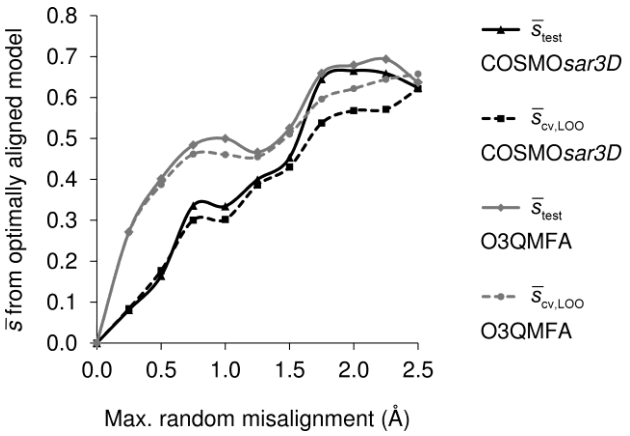

Figure 4. Sensitivity of the internal (dotted lines) and external (plain lines) predictive quality of COSMO*sar3D* (black) and O3QMFA (gray) models with respect to translational misalignment of individual compounds of increasing amplitude (randomly chosen between 0.0 and 0.25-2.5 Å on the three Cartesian axes). Standard deviations of the errors of prediction are computed with respect to the optimally aligned model.

For comparison purposes, we also report the results obtained from a reimplementation of CoMFA within the Open3DQSAR software (O3QMFA). Some minor differences with respect to original CoMFA occur due to the usage of the Merck force-field parameters and charges,[25] while the Tripos force field parameters[26] coupled with scaled MNDO ESP-fit (Gasteiger-Marsili for the THER dataset) charges were used by Sutherland; nevertheless, it overall nicely reflects the same behavior as CoMFA on the eight datasets. All individual and average statistics for the models are collected in Table 1, together with the corresponding data for O3QMFA and the seven QSAR methods considered by Sutherland. Based on $\bar{s}_{test}$ the COSMO*sar3D* models are better than all non-LSP models. The average advantage over the best two models obtained by Sutherland, i.e. over CoMSIA-extra and standard CoMFA, is 0.10 and 0.12 log-units, respectively, and larger than the advantage of the latter over the other QSAR models. Hence, we may conclude that the LSP-based models instanter have a significantly higher predictivity than standard 3D-QSAR models. Although we consider $\bar{s}_{test}$ as the most relevant indicator for the quality of the models, it is worth noting that the LSP based models also are best with respect to the other three overall quality measures, i.e. with respect to the external $\bar{r}^2_{test}$ and the internal $\bar{s}_{cv,LOO}$ and $\bar{q}^2_{LOO}$. Furthermore, for six of the eight datasets the standard deviation of test set residuals achieved by the LSP-based models is smaller than those of all QSAR models considered by Sutherland. Only in two cases, i.e. for the BZR and DHFR datasets, HQSAR has the smallest standard deviation on the test set residuals.

After having confirmed the predictive ability of LSP-based COSMO*sar3D*, we have investigated its robustness with respect to systematic variation of the grid resolution from 1.0 to 5.0 Å in 0.5 Å increments. For comparison, all experiments have been carried out with O3QMFA as well; results are reported in Figure 2. It can clearly be seen that O3QMFA models, in addition to being less predictive according to both internal and external indicators, show much larger fluctuations in $\bar{s}_{test}$; i.e., the COSMO*sar3D* models are much more robust with respect to the variation of the grid step size.

Table 1. Performance of the LSP-based COSMO*sar3D* models on the eight Sutherland datasets. The data for the O3QMFA reference model and for seven 2D/3D-QSAR methods evaluated by Sutherland et al.[14] are reported for comparison.

		COSMO*sar3D*[a]	O3QMFA[b]	CoMFA[c]	CoMSIA basic[c]	CoMSIA extra[c]	EVA[c]	HQSAR[c]	2D[c]	2.5D[c]
ACE	r^2_{train}	0.93	0.75	0.80	0.76	0.73	0.84	0.84	0.76	0.82
	s_{train}	0.60	1.16	1.04	1.15	1.22	0.93	0.95	1.15	1.00
	q^2_{LOO}	0.71	0.65	0.68	0.65	0.66	0.70	0.72	0.68	0.72
	$s_{cv,LOO}$	1.26	1.39	1.32	1.38	1.36	1.28	1.24	1.32	1.24
	r^2_{test}	0.62	0.45	0.49	0.52	0.49	0.36	0.30	0.47	0.51
	s_{test}	1.30	1.57	1.54	1.48	1.53	1.72	1.80	1.57	1.50
AChE	r^2_{train}	0.88	0.72	0.88	0.86	0.86	0.96	0.72	0.40	0.38
	s_{train}	0.41	0.64	0.41	0.45	0.45	0.23	0.64	0.94	0.95
	q^2_{LOO}	0.53	0.41	0.52	0.48	0.49	0.42	0.34	0.32	0.31
	$s_{cv,LOO}$	0.83	0.93	0.84	0.87	0.86	0.92	0.98	1.00	1.00
	r^2_{test}	0.61	0.61	0.47	0.44	0.44	0.28	0.37	0.16	0.16
	s_{test}	0.81	0.81	0.95	0.98	0.98	1.11	1.01	1.20	1.20
BZR	r^2_{train}	0.59	0.53	0.61	0.62	0.62	0.51	0.64	0.51	0.52
	s_{train}	0.42	0.45	0.41	0.41	0.41	0.47	0.40	0.46	0.46
	q^2_{LOO}	0.45	0.41	0.32	0.41	0.45	0.40	0.42	0.36	0.35
	$s_{cv,LOO}$	0.49	0.51	0.54	0.51	0.49	0.51	0.50	0.53	0.53
	r^2_{test}	0.13	0.13	0.00	0.08	0.12	0.16	0.17	0.14	0.20
	s_{test}	0.90	0.90	0.97	0.93	0.91	0.89	0.88	0.90	0.87
COX2	r^2_{train}	0.75	0.68	0.70	0.69	0.69	0.68	0.70	0.62	0.68
	s_{train}	0.50	0.57	0.56	0.56	0.57	0.58	0.55	0.63	0.58
	q^2_{LOO}	0.54	0.43	0.49	0.43	0.57	0.45	0.50	0.49	0.55
	$s_{cv,LOO}$	0.69	0.77	0.73	0.77	0.67	0.75	0.72	0.73	0.68
	r^2_{test}	0.43	0.37	0.29	0.03	0.37	0.17	0.27	0.25	0.27
	s_{test}	1.10	1.16	1.24	1.44	1.17	1.33	1.26	1.27	1.25
DHFR	r^2_{train}	0.80	0.80	0.79	0.76	0.75	0.81	0.81	0.61	0.65
	s_{train}	0.56	0.57	0.59	0.62	0.63	0.55	0.55	0.79	0.75
	q^2_{LOO}	0.69	0.69	0.65	0.63	0.65	0.64	0.69	0.51	0.53
	$s_{cv,LOO}$	0.71	0.71	0.75	0.77	0.75	0.76	0.70	0.89	0.87
	r^2_{test}	0.58	0.59	0.59	0.52	0.53	0.57	0.63	0.47	0.49
	s_{test}	0.89	0.88	0.89	0.96	0.95	0.90	0.84	1.00	0.99
GPB	r^2_{train}	0.95	0.76	0.84	0.78	0.92	0.89	0.77	0.55	0.70
	s_{train}	0.25	0.52	0.43	0.50	0.30	0.36	0.52	0.72	0.59
	q^2_{LOO}	0.61	0.30	0.42	0.43	0.61	0.58	0.66	0.31	0.46
	$s_{cv,LOO}$	0.67	0.89	0.81	0.81	0.67	0.69	0.62	0.89	0.78
	r^2_{test}	0.63	0.29	0.42	0.46	0.59	0.49	0.58	-0.06	0.04
	s_{test}	0.73	1.02	0.94	0.90	0.79	0.88	0.80	1.27	1.20
THER	r^2_{train}	0.90	0.78	0.85	0.85	0.77	0.86	0.81	0.79	0.85
	s_{train}	0.58	0.89	0.73	0.73	0.91	0.72	0.82	0.86	0.73
	q^2_{LOO}	0.58	0.47	0.52	0.54	0.51	0.48	0.49	0.62	0.66
	$s_{cv,LOO}$	1.21	1.37	1.30	1.27	1.31	1.35	1.34	1.16	1.09
	r^2_{test}	0.59	0.49	0.54	0.36	0.53	0.36	0.53	0.14	0.07
	s_{test}	1.47	1.63	1.59	1.87	1.60	1.87	1.59	2.16	2.24
THR	r^2_{train}	0.90	0.85	0.86	0.88	0.89	0.83	0.87	0.79	0.75
	s_{train}	0.30	0.37	0.36	0.34	0.32	0.39	0.35	0.43	0.47
	q^2_{LOO}	0.74	0.65	0.59	0.62	0.72	0.47	0.50	0.62	0.52
	$s_{cv,LOO}$	0.49	0.56	0.61	0.58	0.50	0.69	0.67	0.58	0.66
	r^2_{test}	0.66	0.60	0.63	0.55	0.63	0.11	-0.25	0.04	0.28
	s_{test}	0.65	0.70	0.70	0.76	0.69	1.08	1.27	1.12	0.96
AVERAGES OVER THE EIGHT DATASETS	r^2_{train}	0.84	0.73	0.79	0.78	0.78	0.80	0.77	0.63	0.67
	s_{train}	0.45	0.65	0.57	0.60	0.60	0.53	0.60	0.75	0.69
	q^2_{LOO}	0.60	0.50	0.52	0.52	0.58	0.52	0.54	0.49	0.51
	$s_{cv,LOO}$	0.79	0.89	0.86	0.87	0.83	0.87	0.85	0.89	0.86
	r^2_{test}	0.53	0.44	0.43	0.37	0.46	0.31	0.33	0.20	0.25
	s_{test}	0.98	1.08	1.10	1.17	1.08	1.22	1.18	1.31	1.28

[a] LSP-based COSMO*sar3D* model (grid step size 2.0 Å, DELSIG 0.6 e/nm^2). [b] Model computed with Open3DQSAR CoMFA-like descriptors (grid step size 2.0 Å). [c] Different 2D/3D-QSAR methods reported in the original literature.[14]

Then, we analyzed the sensitivity to grid positioning, while making sure that the grid consistently exceeded the largest compound by at least 5.0 Å in all directions; although this is not required for COSMO*sar3D* models (see below), it is necessary for a fair comparison of O3QMFA results. In Figure 3 we see the dependence of $\bar{s}_{test}$ on the position of the grid center; while the CoMFA-type method O3QMFA displays the well-known strong fluctuations depending on the grid position,[27,28] LSP-based COSMO*sar3D* appears to be essentially invariant.

As the next experiment, we applied random translations of increasing amplitude to each molecule, thus assessing the robustness of the models with respect to misalignment. The results are shown in Figure 4. As expected, the standard deviations of the errors of prediction increase with increasing noise amplitude for both COSMO*sar3D* and O3QMFA, but the latter appears to be much more sensitive to misalignment than the LSP-based COSMO*sar3D*.

Subsequently, we have tested the sensitivity of LSP-based COSMO*sar3D* models upon the quality of the underlying COSMO σ-sufaces. Firstly, in the DFT calculations we replaced the standard BP-SVP level of theory with the higher level BP-TZVP method, which generally yields more accurate results than BP-SVP in COSMO-RS applications. We found a negligible performance increase of 0.003 for $\bar{s}_{test}$, which clearly does not justify the larger computational demands compared to the BP-SVP level. Next, we explored the replacement of COSMO σ-surfaces explicitly computed by DFT/COSMO calculations for each of the considered molecules by approximate CF-COSMO σ-surfaces, which are generated within ~ 0.5 s per compound with the COSMO*frag* program.[23] This replacement of the explicit quantum chemical calculations by COSMO*frag* has been shown to be very efficient and almost equivalently accurate within COSMO*sim3D* similarity and alignment applications.[10] Since unfortunately this approach currently is not applicable to compounds with nonzero net charge, we restricted this test to the four datasets BZR, COX2, DHFR and GPB, which involve mostly neutral compounds; the few ionic compounds were represented by the BP-SVP-COSMO files. The performance of the BP-SVP LSP models on these four datasets was $\bar{s}_{test}$ = 0.91, while the models based on CF-COSMO files yield $\bar{s}_{test}$ = 0.94. Hence we may conclude that, at least for screening purposes, the costly DFT/COSMO calculations can be safely replaced by quickly generated CF-COSMO files with negligible loss of accuracy. By that the computational demands of the descriptor generation for COSMO*sar3D* become on par with force-field based methods. It might be worth noting that also the sizes of the descriptor matrix and hence the PLS demands are not far from standard MFA approaches, because the larger number of ~ 11 σ bins compared to the two CoMFA fields (electrostatic and steric) is partially compensated by the fact that the spatial grid of CoMFA type approaches must extend ~ 5 Å outside the molecules, while the LSP grid just needs to include all COSMO cavities, i.e. extends 1.9 Å at the most for organic molecules.

As a final experiment, we tested COSMO*sar3D* with a consistent field-based alignment generated with COSMO*sim3D*. For that purpose, we applied random translations and rotations to the molecular geometries of all compounds of the eight datasets. Subsequently, we used COSMO*sim3D* to re-align each dataset, using the same superposition templates as used by Sutherland. The alignment was done in a completely automated fashion, using the super-self-consistent alignment mode of COSMO*sim3D*.[10] The resulting alignments and detailed COSMO*sar3D* results are shown in the Supporting Information. The overall performance of these COSMO*sim3D*-aligned models is almost identical to the performance of the models based on the alignment given by Sutherland. The overall standard deviation of test set residuals only increases by 0.02, while the other 3 indicators differ by 0.01 from those achieved on the original alignment; hence, there is no significant loss of performance if a fully automated COSMO*sim3D* alignment is used instead of the supervised alignment reported by Sutherland. Nevertheless, it must be acknowledged that our alignment is based on the conformational selection made by Sutherland and hence is not completely independent of Sutherland's work.

Conclusions

Starting from the framework of COSMO-RS theory, the local σ-profiles have been introduced as a natural and essentially complete set of descriptors for 3D-QSAR and molecular field analysis. Application of this COSMO*sar3D* concept to the eight reference MFA datasets published by Sutherland et al. demonstrates a significant increase of the predictive accuracy of the resulting models compared to standard 3D-QSAR methods. Furthermore, the COSMO*sar3D* models turn out to be exceptionally robust with respect to grid step size, grid positioning and random misalignment. Furthermore, no cutoff parameters are required, neither with respect to the descriptors, nor with respect to the field extension outside the molecules. Although being originally based on quantum chemically calculated and hence computationally more expensive COSMO σ-surfaces, the COSMO*sar3D* methodology is shown to perform almost equally well on approximate σ-surfaces quickly generated from a database of precalculated COSMO files. Finally, it is shown that COSMO*sar3D* also performs well using a consistent σ-surface alignment with COSMO*sim3D*. Summarizing, the COSMO*sar3D* method introduced herein appears to be a promising novel tool for 3D-QSAR, overcoming many of the problems inherent in commonly used methods.

As a next step, we are planning to develop an iterative workflow for the selection of the relevant ligand conformations, making use of the ability of the COSMO-RS method to quantify the free energies in aqueous solution, and of the capability of COSMO*sar3D* to predict the free energy differences of ligands between the aqueous state and the receptor bound state.

Supporting Information

The following material is available as supporting information: Table S1: Statistics of the PLS models obtained from COSMOsim3D alignments; Figure S1: Original Sutherland alignment compared to the unsupervised COSMOsim3D alignment for ACE, AChE, BZR and COX2 datasets; and Figure S2. Original Sutherland alignment compared to the unsupervised COSMO*sim*3D alignment for DHFR, GPB, THERM and THR datasets.

References

1. Cramer, R. D. III; Milne, M. Abstracts of Papers of the Am. Chem. Soc. M., April 1979, Computer Chemistry Section, no. 44.
2. Cramer, R. D. III; 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; doi.
3. Kubinyi, H. *Comparative Molecular Field Analysis (CoMFA)*. In Encyclopedia of Computational Chemistry; Wiley: New York, 1998; Vol. 1, pp 448-460.
4. Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in drug design - a review. *Curr. Top. Med. Chem.* **2010**, *10*, 95-115; doi.
5. 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.
6. Klamt, A. COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design; Elsevier: Amsterdam, **2005**.
7. 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.
8. 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.
9. Klamt, A.; Huniar, U.; Spycher, S.; Keldenich, J. COSMO*mic*: a mechanistic approach to the calculation of membrane-water partition coefficients and internal distributions within membranes and micelles. *J. Phys. Chem. B* **2008**, *112*, 12148-12157; doi.
10. Thormann, M.; Klamt, A.; Wichmann, K. COSMO*sim3D*, 3D-similarity and alignment based on COSMO polarization charge densities. *J. Chem. Inf. Model.* **2012**; doi.
11. Wold, S. Partial Least Squares (PLS) in chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P. A.; Schaefer, H. F. III; Schreiner, P. R., Eds.; Wiley: Chichester, 1998; pp 2006-2021.
12. 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.* **2012**, *14*, 955-963; doi.
13. Klamt, A.; Thormann, M. Molecular field analysis based on COSMO polarization charge densities, European Patent Application No. 10014109.2, 2011.
14. 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.
15. Tosco, P.; Balle, T. Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *J. Mol. Model.* **2011**, *17*, 201-208; doi; http://open3dqsar.org.
16. Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic structure calculations on workstation computers: the program system TURBOMOLE. *Chem. Phys. Lett.* **1989**, *162*, 165-169; doi.
17. *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.
18. Vosko, S. H.; Wilk, L.; Nussair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200-1211; doi.
19. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098-3100; doi.
20. Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822-8824; doi.
21. Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms lithium to krypton. *J. Chem. Phys.* **1992**, *97*, 2571-2577; doi.
22. 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.
23. 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.
24. For the COSMO*frag* program see http://www.cosmologic.de/cosmofrag.html.
25. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490-519; doi.
26. Clark, M.; Cramer, R. D. III; Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982-1012; doi.
27. Cho, S. J.; Tropsha, A. Cross-validated r^2-guided region selection for Comparative Molecular Field Analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060-1066; doi.
28. Wang, R.; Gao, Y.; Liu, L.; Lai, L. All-orientation search and all-placement search in Comparative Molecular Field Analysis. *J. Mol. Model.* **1998**, *4*, 276-283; doi.