COSMO*sim*: Bioisosteric Similarity Based on COSMO-RS σ Profiles

Michael Thormann,[†] Andreas Klamt,*[,‡] Martin Hornig,[‡] and Michael Almstetter[†]

Origenis GmbH, Gmunder Str. 37A, 81379 München, Germany, and COSMOlogic GmbH and Co. KG,
Burscheider Str. 515, 51381 Leverkusen, Germany

Received October 20, 2005

A novel approach for the quantification of drug similarity is proposed, which makes use of the surface polarities, that is, conductor surface polarization charge densities σ, as defined in the quantum chemically based conductor-like screening model for realistic solvation(COSMO-RS). The histogram of these surface polarities, the so-called σ profiles, have been proven to be the key for the calculation of all kinds of partition and adsorption coefficients and, therefore, of relevant absorption, distribution, metabolism, and excretion parameters such as solubility, pK_a, log BB, and many others. They also carry a large part of the information required for the estimation of desolvation and binding processes responsible for receptor binding and enzyme inhibition of drug molecules. Thus, a large degree of similarity with respect to the σ profiles appears to be a necessary condition for drugs of similar physiological action. Driven by this insight, we propose a σ-profile-based drug similarity measure COSMO*sim* for the detection of new bioisosteric drug candidates. In several examples, we demonstrate its statistical and pharmaceutical plausibility, its practicability for real drug research projects, and its unique independence from the chemical structure, which enables scaffold hopping in a natural way.

INTRODUCTION

Bioisosteric transformation is one of the most frequently used approaches to the design and optimization of compounds with biopharmacological importance. Both scaffold hopping and group interchange converge seamlessly in this approach. Current computational approaches to bioisosteric transformation are, so far, a posteriori classifiers that deduct rules, which are then applied in a project-dependent manner. We describe the development and application of an a priori method for the prediction of bioisosters. We assume that bioisosters must have similar physicochemical properties that rule their interactions with different environments such as solvents, membranes, and ultimately protein receptors. These interactions define their biological effects and biopharmacological properties. We built our method, therefore, upon the conductor-like screening model for realistic solvation (COSMO-RS), a general and fast methodology for the a priori prediction of thermophysical data. Cheap unimolecular quantum chemical calculations combined with exact statistical thermodynamics provide the information necessary for the evaluation of molecular interactions. Thus, we can represent molecules detached from their chemical structures as electronic surfaces. For the rapid comparison of COSMO surfaces, they are projected into 1D σ profiles. Suitable similarity metrics are then developed based on the σ profiles. Empirical data of many thousand bioisosteric transformations are used to validate the method and to further optimize the free parameters of the similarity metrics. Finally, the method is applied to bioisosteric transformations of functional groups and for virtual high-throughput screening (vHTS).

A number of biological targets of pharmaceutical interest are currently beyond the scope of experimental 3D structure determination. Examples of such targets are membrane-standing proteins such as ion channels and G-protein coupled receptors. The design of potent modulators for such targets is, though, not a hopeless task. The ligand-based design of compounds makes use of the structural information of well-characterized compounds that is used at different levels of abstraction for subsequent intermolecular similarity calculations.[1−3] Ideally, the selection of target ligands should follow the concepts of ligand efficiency.[4−6] Two scenarios of ligand-based design strategies can be envisaged. First, ligands with improved potency or improved biophysical properties ultimately related to their adsorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles are searched within the class of ligands described by the scaffold. Second, alternative structural classes are explored for backup or patent-busting purposes, a concept termed scaffold hopping. Both approaches are intimately related to the concept of bioisosterism, that is, the introduction of structural changes into a given active compound leading to a derivative that broadly maintains the bioactivity at the given receptor in question following the well-known active analogue principle.

The search for novel and better bioisosters has been the challenge for medicinal chemistry over the past decades. On the one hand, a lot of experience has been gained via trial and error. A survey of successful bioisosteric transformations is accessible in form of the Bioster database.[7] This is a well-suited data set for method development in the bioisosterism field.[8,9] A posteriori analyses of successful bioisosteric transformations led to the common classical and nonclassical categorizations of bioisosters.[10,11] On the other hand, some a priori concepts tried to rationalize bioisosterism on the basis of physical concepts such as Grimm's hydride displacement law.[12,13] Newer nonclassical definitions focus on certain aspects of the compound structures such as molecular shape, topology, pharmacophoric patterns, and electronic isosters.[14] However, none of the currently available bioisosteric approaches is built on a solid physicochemical basis.

* Corresponding author phone: +49-2171-731680; fax: +49-2171-731689; e-mail: klamt@cosmologic.de.
[†] Origenis GmbH.
[‡] COSMOlogic GmbH and Co. KG.

10.1021/ci050464m CCC: $33.50
Published on Web 03/01/2006

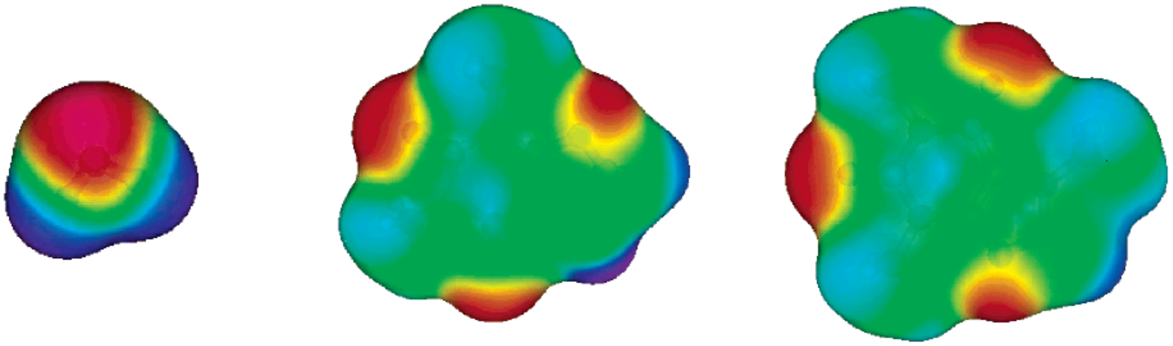

Figure 1. COSMO surfaces color coded by the polarization charge density σ for water, caffeine, and theophylline.

Ultimately, a suitable method must provide an intermolecular similarity measure that ranks proven bioisosteric pairs of compounds high while ranking random pairs low. It can then be used for picking tentative bioisosters for a target ligand from compound databases. Furthermore, intermolecular similarity is used as an objective function for the design and optimization of target-focused compound libraries.

MATERIALS AND METHODS

COSMO-RS. While almost all computational chemistry methods used in drug design are based either on the force-field concept or on decomposition concepts such as group contributions or fingerprints, a rather orthogonal and fundamental approach has been developed over the past 15 years, which is based on quantum chemistry combined with dielectric continuum solvation and statistical thermodynamics. Having been developed in an industrial computational chemistry lab, COSMO-RS[15−17] was originally considered for the quantification of environmental and technical partition behavior, before it was recognized as being a very generally applicable and most predictive model for fluid phase thermodynamics in the chemical engineering community. In light of its applicability in wide ranges of chemistry, its applicability to biophysical properties became apparent, and it has been proven in several papers on solubility, physiological partition properties such as log BB or intestinal absorption, and pK_a during the past years,[18−21] and its extension toward the problem of ligand receptor binding appears attractive and is presently being exploited. While detailed descriptions of the COSMO-RS theory have been given elsewhere, the basic concept of the methods is outlined as follows.

In the first step, all chemical compounds of a liquid (or pseudo-liquid) ensemble are considered as embedded and swimming in an infinite, perfect conductor. This state of the compounds can be very well treated by quantum chemical calculations combined with dielectric continuum solvation models, the most efficient and natural choice for the latter being the conductor-like screening model, COSMO.[22] For quantum chemistry, density functional theory (DFT) has been proven to provide a good compromise of computational efficiency and reliability, because the much faster semiempirical methods suffer from severe deficiencies in solvation electrostatics, especially if hypervalent elements such as sulfur or phosphorus come into play. By such a combination of methods, the self-consistent state of a molecule within such a virtual conductor, that is, its energy, its geometry, and electron distribution, and the surface polarization charges of the conductor on the molecular surface can be calculated at almost the same costs as those in the gas phase.[23]

This state of molecules in a conductor has proven to be a very useful reference for the understanding of molecular behavior in the liquid. It is much better suited than the traditional reference state of an isolated molecule in a vacuum. The conductor surface polarization charge density, σ, is a very good local measure of molecular surface polarity, carrying more information than the often considered electrostatic potential. Examples of COSMO surfaces color coded by σ are given for water, caffeine, and theophylline in Figure 1. The regions of strongly negative molecular polarity are displayed in red. It is important to recognize that the negative molecular regions carry a positive polarization charge density, σ, because σ is just compensating the molecular electrostatic field and has, thus, the opposite sign. The strongly positive molecular regions carrying negative σ are colored blue, while the neutral parts of the molecules with σ close to zero appear green.

A big conceptual advantage of the conductor reference state is the fact that the molecules together with their polarization charges now are electrostatically perfectly non-interacting, because no electric field can escape through the solute−conductor interface. As long as we leave at least a thin film of conductor between the molecules, we can build any geometrical configuration of the molecules without changing the energy of the system. By allowing for small, volume conserving, and energetically irrelevant deformations of the surfaces, we can finally build densely packed, liquidlike systems of the molecules, as schematically shown in Figure 2, but still assuming an infinitely thin film of conductor separating and screening the molecules from each other. If we suppose that the small deformations required for the close packing do not significantly influence the energy and the polarization charge densities, our ensemble now still has the same energy as that in the dilute state in the conductor, but each surface segment μ with polarization σ_μ of a molecule has a nearest neighbor segment ν with σ_ν; that is, we only have direct face-to-face segment pairs.

Because in nature there is no conductor between the molecules, in the next step, the thin film of conductor between the surface segments has to be removed. As shown in the COSMO-RS papers in more detail, we can do this segment pair by segment pair. In this way, we can consider the energy difference going along with the removal of the conductor between a segment pair (μ,n) as a local surface interaction of the neighboring molecules, and we can quantify its electrostatic and hydrogen bond contribu-

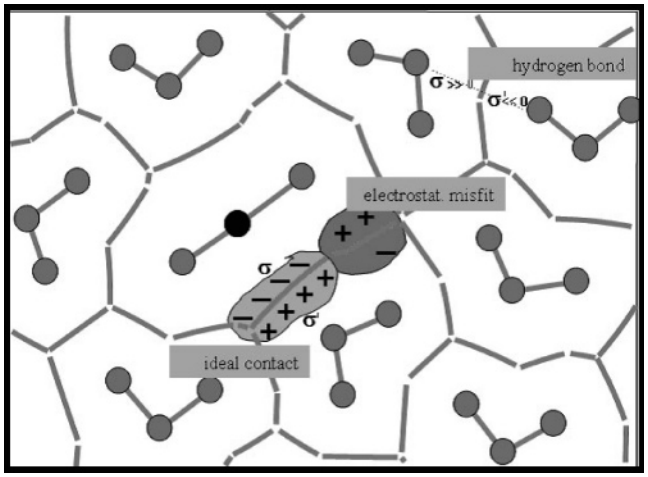

Figure 2. Schematic visualization of COSMO-RS contacts and interactions in the molecular cavity.

tions per unit surface area as

$$e_{\mathrm{misfit}}(\sigma,\sigma') \cong \frac{\alpha'}{2}(\sigma + \sigma')^2 \qquad (1)$$

and

$$e_{\mathrm{hb}}(\sigma,\sigma') \cong c_{\mathrm{hb}}(T)\,\min[0,\sigma\sigma' + \sigma_{\mathrm{hb}}{}^2] \qquad (2)$$

where σ and σ' are the polarization charge densities of the interacting surfaces. While the electrostatic misfit part can be quite accurately derived from theoretical arguments, the hydrogen bond energy term must be considered as an empirical but physically plausible expression. Given the very few parameters in these formulas, the interaction energy of a fixed configuration of our ensemble of molecules relative to its conductor reference state can, thus, be calculated as an integral of all local pairwise surface interactions.

For a liquid system, we have to calculate the thermodynamic averages of the relevant configurations of our ensemble. Normally, this requires the generation of large numbers of such ensembles, as done in Monte Carlo or molecular dynamics simulations. But, because of the local pairwise surface interactions description, the thermodynamics can be reduced to an ensemble of independently interacting surface segments. For this purpose, we just need the surface polarization charge density distribution $p^{\mathrm{X}}(\sigma)$, the so-called σ profile, of each compound X, which tells us how much surface of polarity σ is available on the surface of compound X. Because these σ profiles are of central importance for our new similarity approach, a collection of σ profiles is shown in Figure 3. These σ profiles turned out to be very useful fingerprints of molecules. Next, we denote the normalized σ profile of a solvent S by $p_{\mathrm{S}}(\sigma)$. For a pure solvent, $p_{\mathrm{S}}(\sigma)$ is just the normalized σ profile of the solute molecule, and for a mixture, it is trivially generated from the mole fraction weighted σ profiles of the components. On the basis of $p_{\mathrm{S}}(\sigma)$, the statistical thermodynamics of the interacting surface pairs can be solved efficiently and exactly from the integral equation

$$\mu_{\mathrm{S}}(\sigma) = -kT \ln\left\{\int p_{\mathrm{S}}(\sigma') \exp\left[-\frac{a_{\mathrm{eff}}\, e_{\mathrm{int}}(\sigma,\sigma') - \mu_{\mathrm{S}}(\sigma')}{kT}\right] \mathrm{d}\sigma'\right\} \qquad (3)$$

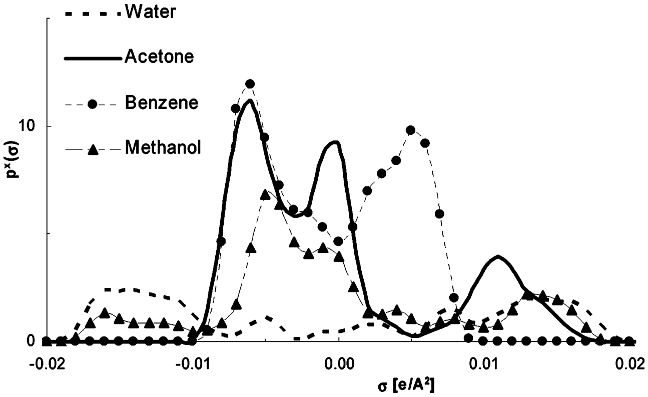

Figure 3. Solvent σ profiles. These profiles show the amount of molecular surface in a given interval of polarization charge density σ.

where $\mu_{\mathrm{S}}(\sigma)$ is the chemical potential of an additional surface segment in the ensemble S and a_{eff} is the size of an effective contact segment. The function $\mu_{\mathrm{S}}(\sigma)$, called the σ potential from here on, has to be derived recursively from eq 3 because of its appearance on both sides of the equation. The σ potential describes how much a solvent S likes additional surface area of polarity σ. As explained in more detail elsewhere, these σ potentials express a wide range of important aspects of the interaction capabilities of the solvent, including electrostatics, hydrogen bonding, and hydrophobicity. Finally, the chemical potential of a compound X in solvent S is calculated from the σ profile of the solute and the σ potential of the solvent as

$$\mu_{\mathrm{S}}^{\mathrm{X}} = \int \mu_{\mathrm{S}}(\sigma)\, p^{\mathrm{X}}(\sigma)\, \mathrm{d}\sigma + \mu_{\mathrm{S,comb}}^{X} \qquad (4)$$

where $\mu_{\mathrm{S,comb}}^{\mathrm{X}}$ is a usually small correction for the size effects of the solute and solvent and, thus, a combinatorial contribution, depending on volumes and surface areas of the solute and solvent. Hence, the important part of the chemical potential of a solute X in a solvent S is expressed as a surface integral of the solvent σ potential over the surface of the solute X, because $p^{\mathrm{X}}(\sigma)\,\mathrm{d}\sigma$ is just the amount of surface area of polarity σ on the surface of X.

In this way, COSMO-RS gives access to the chemical potentials of almost arbitrary compounds in almost arbitrary liquid phases using the σ profile of the solute as the only information. While chemically well-defined phases such as water, octanol, alkanes, and so forth and the corresponding partition coefficients are directly accessible by COSMO-RS theory, partitioning between a physiological phase such as between the blood and brain and other ADME properties can be derived from the σ profiles with a slightly more empirical extension, the σ-moment approach. Finally, it has been shown that even drug solubility can be calculated from the σ profiles, using the chemical potential $\mu_{\mathrm{X}}{}^{\mathrm{X}}$ of the drug in its virtual liquid state as the most important input.

Because solubility and ADME properties can be well-described from the σ profile alone, we can expect that two compounds with similar σ profiles should have similar ADME characteristics. But, we must be aware that the binding of a drug to a receptor involves constraints related to the 3D relations of the various polar, hydrogen bonding, and hydrophobic interaction sites. This information is not included in the σ profile any more. Nevertheless, it is at least

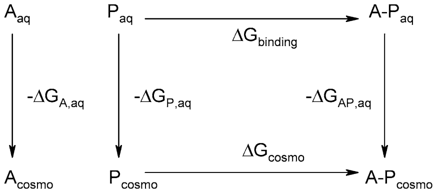

Figure 4. Thermodynamic cycle for the equilibrium binding free energy $\Delta G_{\mathrm{binding}}$ of ligand A binding to receptor P, forming the noncovalent complex AP.

a necessary though not sufficient requirement for strongly binding ligands of a receptor that they have roughly the same amount of surface area available for the various interaction modes. Thus, it is at least plausible that a drug candidate with a σ profile similar to that of a known strongly binding ligand has a good chance to be strongly binding as well. Summarizing the above considerations, we consider it as a plausible assumption that the similarity of σ profiles should be a powerful measure for the assessment of drug similarity.

Application of COSMO-RS to Protein−Ligand Interactions. COSMO−RS provides access to chemical potentials (Gibbs free energies) of almost arbitrary compounds in almost arbitrary phases. This grants direct access to important chemophysical and biophysical properties such as phase partition (e.g., *n*-octanol water partition as expressed in log *P*) and thermodynamic solubility in different solvents (e.g., intrinsic aqueous molar solubility log *S*). COSMO-RS provides, furthermore, a sound basis for the computation of chemical potentials in mixed phases on the basis of statistical thermodynamics (COSMO*therm*). This allows the treatment of more complex compartment partitions including extracellular matrix/cellular membranes, intestinal lumen/blood, and blood/brain.

Noncovalent reversible protein−ligand interactions can well be described by the thermodynamic cycle given in Figure 4 for ligand A binding to receptor P, forming the ligand−receptor complex AP.[24] The complex formation is determined by the binding free energy, $\Delta G_{\mathrm{binding}}$, under thermodynamic equilibrium conditions.

$$\Delta G_{\mathrm{binding}} = \Delta G_{\mathrm{cosmo}} - \Delta G_{\mathrm{A,aq}} - \Delta G_{\mathrm{P,aq}} + \Delta G_{\mathrm{AP,aq}} \quad (5)$$

The binding free energy, $\Delta G_{\mathrm{binding}}$, and complex dissociation constant K_{D} can be interconverted following

$$\Delta G_{\mathrm{binding}} = -RT \ln K_{\mathrm{D}} \quad (6)$$

where *R* is the gas constant and *T* is the temperature.

The rigorous statistical thermodynamic treatment of the ligand, receptor, and ligand−receptor complex as homogeneous pseudo-liquid phases represents the only approximation. Obviously, at the atomic level, protein−ligand interactions can be hampered by steric hindrance due to the anisostropic character of the protein, although electronic complementarity in the COSMO concept is given and would consistently lead to energetically favorable interaction.

Data Sets. A virtual screening database of >3.5 million unique compounds was compiled from various in-house, medchem, and vendor databases. COSMO*frag*[25] was used to calculate approximate σ profiles for all of the compounds. This database is dynamically maintained and updated with new screening compounds. The search for bioisosters of groups and entire molecules was performed with this database.

The Bioster database[7] contains data on successful bioisosteric transformations for various receptors and chemical classes. This data set was previously used for the evaluation of methods to distinguish between bioisosteric pairs and random pairs of molecules.[8,9] These molecular pairs were extracted from the database, and approximate σ profiles were calculated using COSMO*frag*. Obvious prodrugs, ambiguous pairs, and molecules for which no σ profiles could be calculated, like ions, were removed from the set. A list of random pairs was prepared from the same set of molecules by scrambling the right column of the pair table. The final sets contained 6041 bioisosteric pairs and 5823 random pairs. All similarity and energy calculations were run on these two sets.

Biophysical Concept of Bioisosteric Transformations. Medicinal chemistry uses bioisosteric transformation as a tool for the replacement of certain functional groups with alternatives that display higher potency, better specificity, improved safety, and other pharmacological properties or for the design of novel bioactive compounds via scaffold hopping for backup and patent-busting purposes such as therapeutic copies of commercialized drugs.[26] Common concepts deduce rules from data sets that describe previously successful bioisosteric transformations, and these rules are then applied to prioritize the synthesis and characterization of analogues.[27] Our approach attempts to provide a biophysical basis for bioisosteric transformations and will allow, as such, the a priori prediction of bioisosters.

In the ligand-based drug design approach, the receptor is assumed to be constant while the ligand is the variable. The bioisosteric transformation of ligand A to ligand B, both binding to receptor P, can be formulated as a thermodynamic cycle (see Figure 5). For ligand-based drug design purposes, A can be regarded as the target ligand while B is the result of the bioisosteric transformation applied. This approach allows the calculation of the energetic cost of the bioisosteric transformation $\Delta\Delta G_{\mathrm{AP,BP,binding}}$.

$$\Delta\Delta G_{\mathrm{AP,BP,binding}} = \Delta\Delta G_{\mathrm{AP,BP,cosmo}} - \Delta\Delta G_{\mathrm{A,B,aq}} - \Delta\Delta G_{\mathrm{P,aq}} + \Delta\Delta G_{\mathrm{AP,BP,aq}} \quad (7)$$

All terms of this thermodynamic cycle are directly accessible via COSMO-RS and COSMO*therm* under the principle assumption of interacting pseudo-liquid molecules. The protein desolvation term $\Delta\Delta G_{\mathrm{P,aq}}$ cancels out. For compounds A and B with identical σ profiles, the mixed-phase desolvation term $\Delta\Delta G_{\mathrm{AP,BP,aq}}$ becomes zero, as does the protein−ligand interaction term $\Delta\Delta G_{\mathrm{AP,BP,cosmo}}$.

The success of a bioisosteric transformation can, thus, be quantified prior to synthesis by computing $\Delta\Delta G_{\mathrm{AP,BP,binding}}$. For practical purposes, successful bioisosteric transformations should not exceed 1−3 kcal/mol (about 2 log orders of magnitude in the binding constant K_{D}). The final goal is to gain access to the estimated energetic cost of a bioisosteric transformation to be used to rank different transformations in virtual screening.

COSMO*therm* defines the energetics of molecular interactions on the basis of pairwise σ interactions; thus, the unimolecular σ profiles of the ligands in question ultimately define the energetic contributions. Consistently, the more

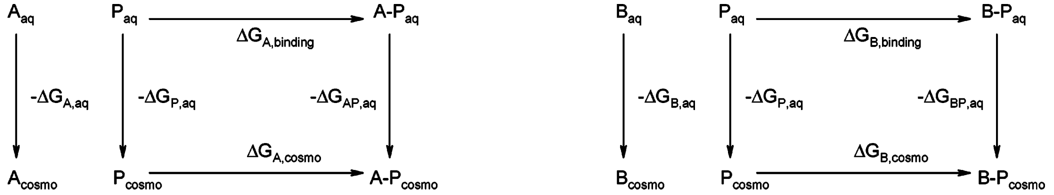

Figure 5. Thermodynamic cycle for the bioisosteric transformation of target ligand A to ligand B, both targeting receptor P.

similar the *σ* profiles of A and B become, the more the ligand-dependent contributing terms cancel out. The next section focuses on different intermolecular similarity methods based on *σ* profiles.

Similarity Coefficients. COSMO*therm* represents the *σ* profiles using 61 real values for the relevant *σ* range from $\sigma = -3$ to $+3$ e/nm^2. We initially started from this 61-bin representation, which represents a one-dimensional structure-free holographic electronic profile. We term *σ*-profile-based similarity methods COSMO*sim*.

A large set of coefficients is available[1,28] to calculate intermolecular distances on the basis of binary, integral, or floating point vectors containing chemical information. One of the widely used coefficients is the Tanimoto coefficient, which has been found to be useful for distance measures of binary and integral vectors encoding molecular descriptors. To compare *σ* profiles, we started with examining the suitability of the Tanimoto coefficient. Usually, the binary variant of the Tanimoto coefficient is used. However, extension to non-negative floating point values is straightforward:

$$T_c = \frac{\sum_{i=1}^{l} 1 - \dfrac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l} \tag{8}$$

The Tanimoto coefficient T_c is the intermolecular similarity, and l is the number of bins, $N_{A,i}$ and $N_{B,i}$ are the surface areas corresponding to bin i in molecules A and B, respectively.

Equation 8 reveals that the usability of the Tanimoto coefficient suffers when many bins are nonzero and few bins differ largely. In such cases, very high similarities are computed that neglect substantial differences. To illustrate these consequences, a congeneric set of *n*-alcohols (*n*-propanol through *n*-hexadecanol) was prepared; the corresponding *σ* profiles were computed for all of the compounds in the set (Figure 6), and T_c was calculated using *n*-propanol as the target ligand. Despite the very different molecular sizes of *n*-propanol and *n*-hexadecanol, high, physically implausible T_c values result (Figure 7).

A straightforward solution to this problem appeared to be the utilization of the relative molecular sizes. The molecular size is here defined as the molecular surface area (COSMO surface) that equals the sum over all bins of the *σ* profile. We call the novel similarity measure the Tanimoto prime coefficient, T_c'.

$$T_c' = T_c \frac{\min(S_A, S_B)}{\max(S_A, S_B)} \tag{9}$$

Where T_c is the Tanimoto coefficient and S_A and S_B are the COSMO surface areas of compounds A and B, respectively.

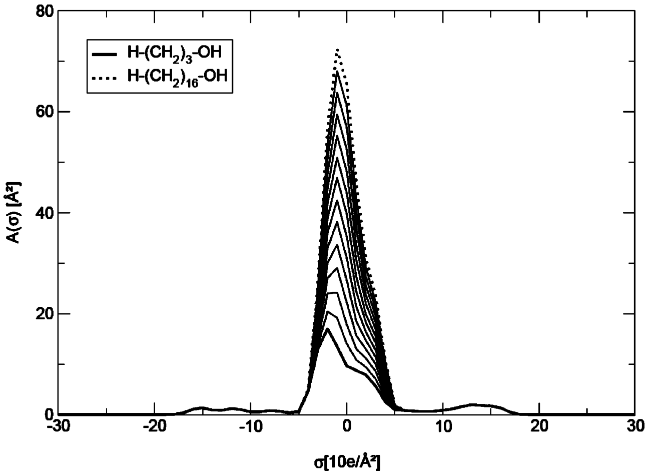

Figure 6. *σ* profiles of a congeneric series of *n*-alcohols.

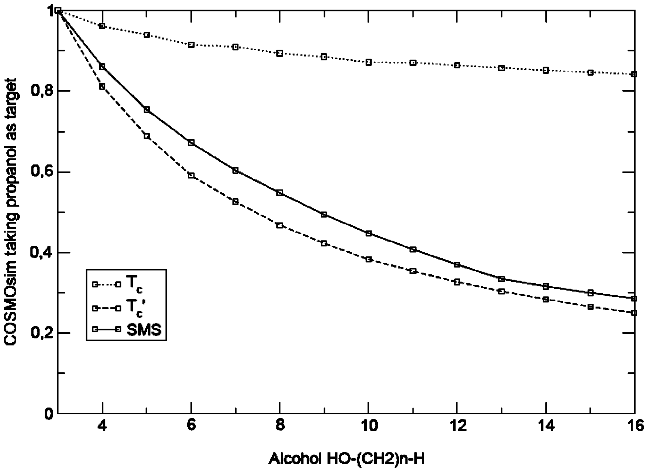

Figure 7. *σ*-profile-based similarities of a congeneric series of *n*-alcohols compared to *n*-propanol using different COSMO*sim* coefficients.

The bin-based Tanimoto similarity of *σ* profiles suffers from a few theoretical weaknesses. First, its values quite strongly depend on the fineness of the discretization of the *σ* range, while a theoretically robust similarity definition should be almost independent of the discretization level. Second, because only the ratio of deviation and sum of the two *σ* values of a bin enters the definition of bin similarity, bins with small values of the similarity may have the same influence on the Tanimoto similarity coefficient as bins with large values, while under physical aspects, bins with small values should be less important. Finally, the bin-based similarity definition totally disregards the physical neighborhood relation of *σ* bins. If one compound has a high intensity in bin i while the other has a high intensity in bin $i + 1$, the Tanimoto coefficient considers this as dissimilarity, while such slightly shifted peaks would still cause rather similar physicochemical behavior.

Table 1. Overview of SMS Parameters and Statistics of Separation of Bioisosteric and Random Molecular Pairs

model	a[a]	b[b]	c[c]	d[d]	g[e]	bioisosters avg SMS ± std dev	random avg SMS ± std dev
worst	1.096	0.009 430	0.001 310	0.993	0.745	0.570 ± 0.155	0.471 ± 0.142
best 3 parameters	2.533	0.000 350	0.009 960	fixed 0	0.424	0.697 ± 0.191	0.382 ± 0.207
best 4 parameters	2.561	0.000 124	0.009 990	0.130	0.424	0.713 ± 0.180	0.416 ± 0.197
T_c					0.406	0.726 ± 0.126	0.528 ± 0.112
T_c'					0.395	0.636 ± 0.165	0.369 ± 0.145

[a] σ tolerance. [b] σ tolerance of polar surface. [c] Weighting of polar surface. [d] Size tolerance. [e] Gray zone overlap.

On the basis of these considerations, we defined another similarity measure that we call σ-match similarity (SMS). The basic idea behind it is the matching of the most similar surface area pairs, starting with the most polar segments present in either of the two σ profiles and proceeding to the least polar ones. In the first step, the most polar piece of the surface available in the two σ profiles is matched with the same area of the most polar surface segment of the same polarity sign of the other compound. Now, the minimum area a_{seg} of the two matched surface segments is subtracted from both σ profiles and a contribution dSMS, as given in eq 10, is added to the raw similarity measure SMS$_0$.

$$\mathrm{dSMS}(a_{\mathrm{seg}},\sigma,\sigma') = a_{\mathrm{seg}}\left[1 + \frac{c}{4}(\sigma + \sigma')^2\right] \exp\left\{\frac{(\sigma - \sigma')^2\left[1 + \frac{b}{4}(\sigma + \sigma')^2\right]}{a^2}\right\} \quad (10)$$

where σ and σ' are the two σ values matched in this step. To better understand this expression, it is useful to consider it first with the two parameters b and c set to zero. In this case, the formula apparently gives a contribution identical to the matched area, if the two matched σ values are identical. Otherwise, the contribution to the raw similarity index is reduced by a Gaussian function. Hence, the similarity strongly decreases with an increasing mismatch of the two σ values. The parameter a is a measure for the σ-mismatch tolerance. By repeating the described procedure until no surface area is left in one of the σ profiles, we finally get a raw similarity coefficient SMS$_0$. At the end of the procedure, we will have a residual surface area a_{res} left in the bigger of the two compounds. The maximum value of SMS$_0$ can be the value of the smaller of the two components. This can only be achieved if the two σ profiles are identical. Using the two parameters b and c, we can introduce two different ways to increase the sensitivity of the σ-similarity measure in the polar σ regions, which might be useful considering the fact that especially the hydrogen bond interactions contribute very strongly in the polar regions thus making them more important for drug similarity than the less polar regions. A positive value of b decreases the σ tolerance in the polar regions, while a positive value of c increases the weight of the surface area of polar segments compared with that of the less polar segments.

If we calculate the maximum achievable values of the raw similarity of the two compounds, that is, their raw self-similarities, and denote them by SMS$_1$ and SMS$_2$, respectively, then we can define our final expression for the σ-match similarity coefficient SMS by

$$\mathrm{SMS} = \frac{\mathrm{SMS}_0 + da_{\mathrm{res}}}{\sqrt{\mathrm{SMS}_1 + \mathrm{SMS}_2}} \quad (11)$$

where d is a parameter responsible for the treatment of the residual surface area a_{res}. For $d = 0$, a_{res} is considered as maximally dissimilar; for $d = 1$, it is treated as maximally similar.

The presented definition of the σ-match similarity fulfills all of the requirements of a similarity metric; that is, it is unity if applied to identical compounds, asymptotically takes a value of 0 for very dissimilar compounds, and is commutative with respect to the compounds. Furthermore, it is rather independent of the bin discretization of the σ range, reasonably weights its contributions according to the surface areas, and introduces a mismatch tolerance with respect to σ. Reasonable values for the four parameters a, b, c, and d of the SMS definition will be presented in the next section.

SMS Parameter Optimization. A genetic algorithm (GA) was used in order to optimize the four free variables a, b, c, and d in the SMS similarity calculation using empirical data of the Bioster database. The target function for the GA optimization was the maximization of the separation of random versus bioisosteric pairs. Therefore, the corresponding similarity values were split into 51 bins covering the SMS similarity range from 0 to 1. The overlapping gray zone, g, was defined as follows, and the GA was used to maximize $1 - g$.

$$g = \sum_{i=0}^{50} \min\left(\frac{n_{\mathrm{bioisoster},i}}{m_{\mathrm{bioisoster}}}, \frac{n_{\mathrm{random},i}}{m_{\mathrm{random}}}\right) \quad (12)$$

A total of 200 generations of GA optimization with 20 unique individual parameter sets per generation were performed starting with random values for a ($1.0 \le a \le 5.0$), b ($0 \le b \le 0.01$), and c ($0 \le c \le 0.01$). For each individual parameter set, the intermolecular similarity values were recalculated for the bioisoster and random data sets, and the individual g values were calculated thereupon. The individual solutions were than sorted by g, and the next generation of individual parameter sets was created by applying crossover and mutation to the binary genomic representations of the better-than-average solutions.

In the first approach, d was held fixed to 0, while in a second independent run, d was also optimized within $0 \le d \le 1$. To check for the possible effects of the four parameters, the GA was rerun but this time maximizing g, thus leading to the worst model possible. The resulting parameters along with the parameter-free metrics T_c and T_c' are shown in Table 1. The optimized parameter sets lead to practically indistinguishable separation values. To limit the number of free parameters, the three-parameter model with d set to zero is used throughout the paper for the COSMO*sim* SMS calculations.

Energetic Cost of Bioisosteric Transformation. A bioisosteric transformation of A to B can hardly be performed

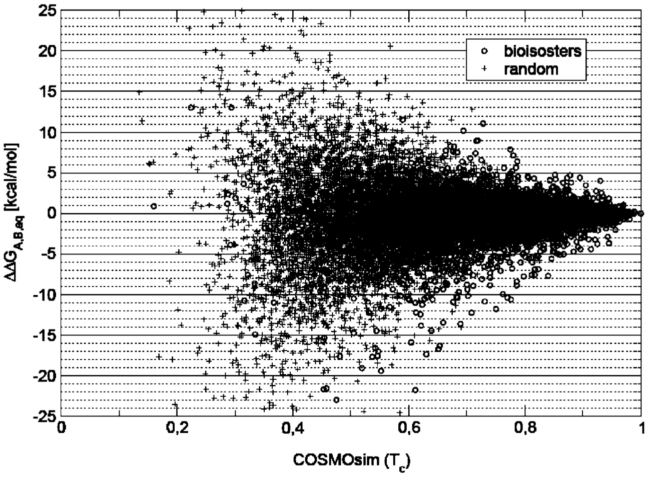

Figure 8. Free energy of solvation changes upon bioisosteric transformation of compound A into B for bioisosteric and random pairs using the COSMO*sim* Tanimoto coefficient.

in practice without affecting the electronic nature of the compound. Such changes will have a more or less drastic effect on the various energetic contributions to the relative binding free energies (eq 5). For practical purposes, a bioisosteric transformation can be regarded as successful if B does not lose more than around 1−2 log orders of magnitude in binding affinity. Such a factor of 10−100 in K_D corresponds to a change in binding free energy by 1.3−2.7 kcal/mol at room temperature.

It is conceptually straightforward to suppose that the change in binding free energies upon bioisosteric transformation becomes ***small*** as the intermolecular similarity approaches unity. The underlying distance metric, however, can have a substantial influence on the meaning of ***small***. To give a qualitative estimation on the energetic cost of bioisosteric transformation, $\Delta\Delta G_{AP,BP,binding}$, in different similarity metrics concepts, one would ideally compute all of the terms of the thermodynamic cycle given in Figure 5. This is currently not possible with reasonable efforts.

To get a rough estimation of the energetic cost, the desolvation term $\Delta\Delta G_{A,B,aq}$ can be rapidly computed with reasonable accuracy (~0.5 kcal/mol), as reported previously.[19] The difference in free energy of ligand desolvation provides an estimate for the energetic cost of the bioisosteric transformation for one of the three terms of eq 8. ΔG_{aq} was calculated for every molecule in the bioisoster and random data sets using COSMO*therm*.[29] Subsequently, $\Delta\Delta G_{A,B,aq}$ was calculated for all pairs.

For a qualitative assessment of the energetic cost, the intermolecular similarity values were plotted against $\Delta\Delta G_{A,B,aq}$. Funnel-shaped curves are obtained for each similarity coefficient (see Figures 8−10), which shows that the solvation free energy differences do become ***small*** as the intermolecular similarity approaches unity. As expected, the funnel shape depends on the metrics. To get a statistically meaningful quantitative estimation of the energetic cost of the bioisosteric transformation, the molecular pairs were binned according to their intermolecular similarities. For each of the 21 bins, the mean averages and standard deviations of $\Delta\Delta G_{A,B,aq}$ were calculated. The results are presented in Figure 11.

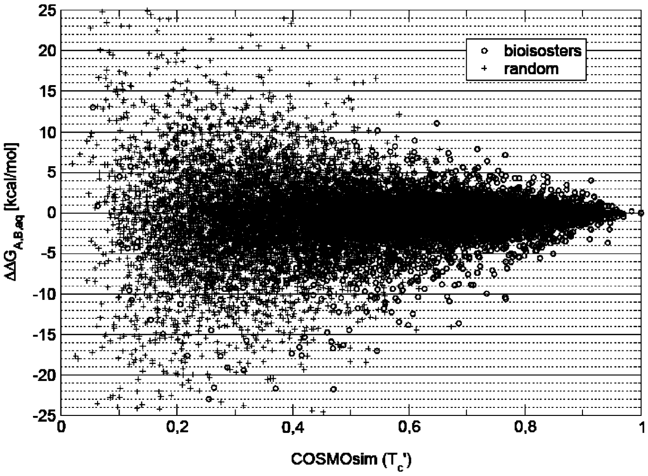

Figure 9. Free energy of solvation changes upon bioisosteric transformation of compound A into B for bioisosteric and random pairs using the COSMO*sim* Tanimoto prime coefficient.

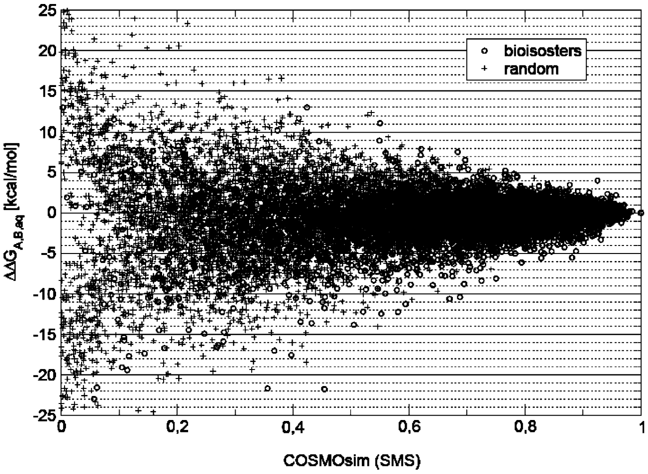

Figure 10. Free energy of solvation changes upon bioisosteric transformation of compound A into B for bioisosteric and random pairs using the COSMO*sim* SMS coefficient.

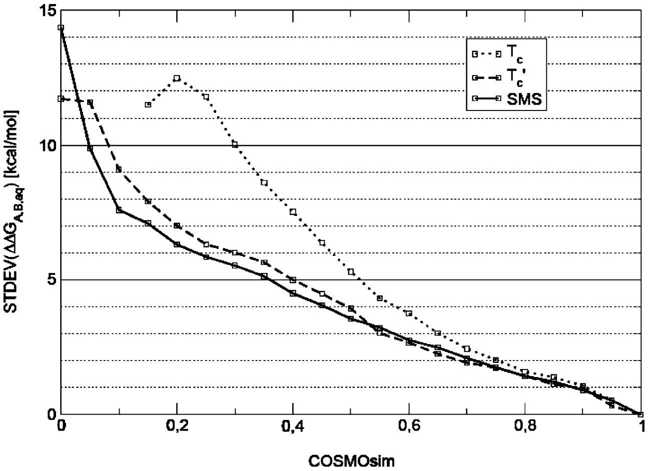

Figure 11. Standard deviation of free energy of solvation changes upon bioisosteric transformation of compound A into B for bioisosteric and random pairs using various COSMO*sim* coefficients.

On the basis of the quantitative energetic assessment, reasonable similarity thresholds were derived for the various distance metrics at the 1, 1.5, and 2 kcal/mol levels. Additionally, the number of bioisosteric and random pairs retrieved at these levels were extracted from the similarity calculations.

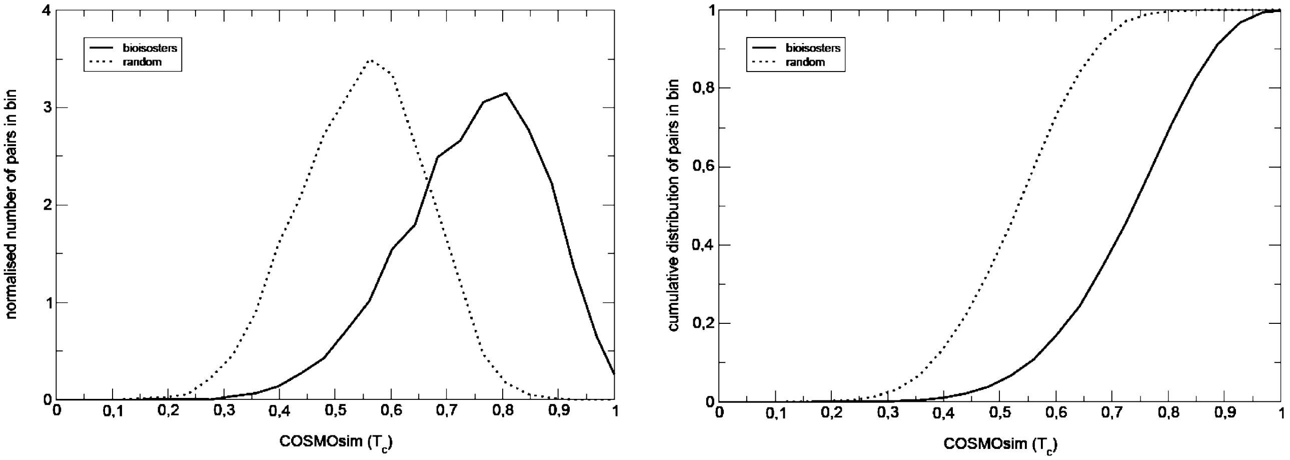

Figure 12. Separation of known bioisosters from random pairs resulting from the application of the COSMO*sim* Tanimoto coefficient to the BioSter data set (left: normalized distribution; right, cumulative normalized distribution).

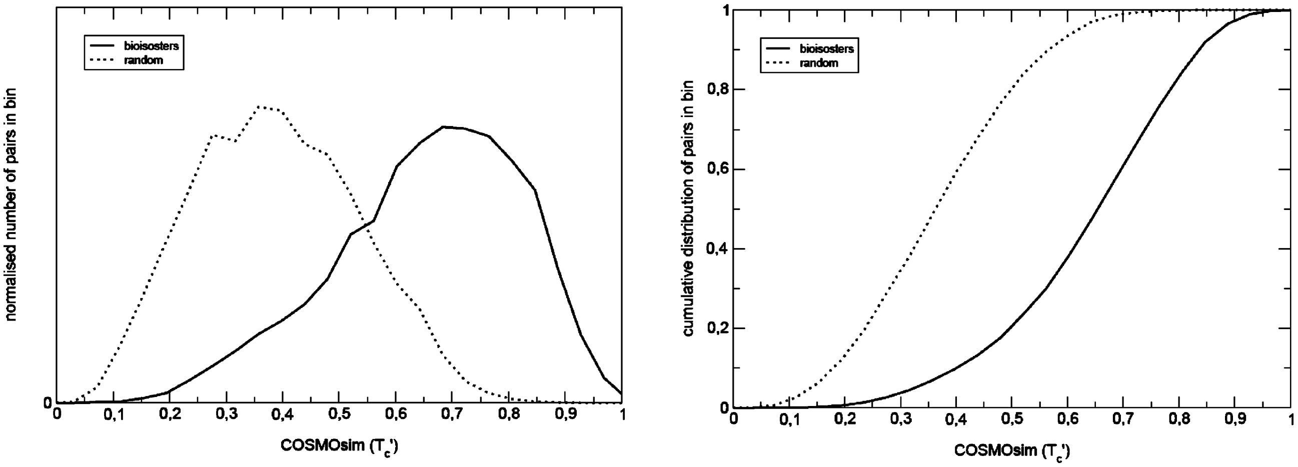

Figure 13. Separation of known bioisosters from random pairs resulting from the application of the COSMO*sim* Tanimoto prime coefficient to the BioSter data set (left, normalized distribution; right, cumulative normalized distribution).

Proper and Improper Bioisosters. Commonly, bioisosters are seen as pairs of compounds that show comparable activity on the target in question. The energetic cost of the corresponding bioisosteric transformation is low for all bioisosters. We have shown that, for compounds with very similar σ profiles, that is, COSMO*sim* approaching unity, the energetic cost of the bioisosteric transformation, $\Delta\Delta G_{\mathrm{AP,BP,binding}}$, becomes small as a result of the different energetic contribution terms, according to eq 7, becoming small. The reverse is, however, not necessarily given: not all fairly equipotent ligands need to have similar σ profiles. Consistently, the COSMO*sim* values of such ligand pairs are low. The reason can be found in the different energetic contribution terms to $\Delta\Delta G_{\mathrm{AP,BP,binding}}$ that sum up to zero, although the single terms do considerably differ from zero and have opposite signs. This distinct energetic behavior gives rise to the definition of two principle classes of bioisosters:

Proper Bioisosters. All contributions to the energetic cost of the bioisosteric transformation, according to eq 7, are small. They have overall similar physicochemical properties, and their similarity is principally receptor-independent. A hypothetical example is given below:

$$\Delta\Delta G_{\mathrm{AP,BP,binding}} = 0\ \mathrm{kcal} = 0 + 0 + 0 + 0\ \mathrm{kcal}$$

Improper Bioisosters. The energetic cost of the bioisosteric transformation is small, but the discrete contributions are not small. They have different biophysical properties, and their similarity is receptor-dependent. A hypothetical example is given below:

$$\Delta\Delta G_{\mathrm{AP,BP,binding}} = 0\ \mathrm{kcal} = 5 - 8 + 15 - 12\ \mathrm{kcal}$$

The application of COSMO*sim* methods in vHTS will, therefore, yield proper bioisosters. They can fail at the target receptor for steric mismatch or repulsion reasons, which are, though, beyond pseudo-liquid treatment. We are, therefore, currently extending COSMO*sim* toward the third dimension, and the first results were recently published.[30]

Separation of Bioisosteric from Random Molecular Pairs. The general applicability of COSMO*sim* was evaluated aiming at the numeric separation of the two sets of molecular pairs, bioisosteric and random. The different coefficients T_c, T_c', and SMS with optimized parameters were applied to each molecular pair. Subsequently, the resulting intermolecular similarities were evaluated in the normalized and cumulative histograms (Figures 12−14).

Group Exchange Transformations with COSMO*sim*. Group exchange applications are useful for the selection of starting materials to be employed in the next generation of compounds maintaining the scaffold (and therefore minimizing the chance of steric repulsion and maximizing the suitability of the pseudo-liquid treatment). We have selected

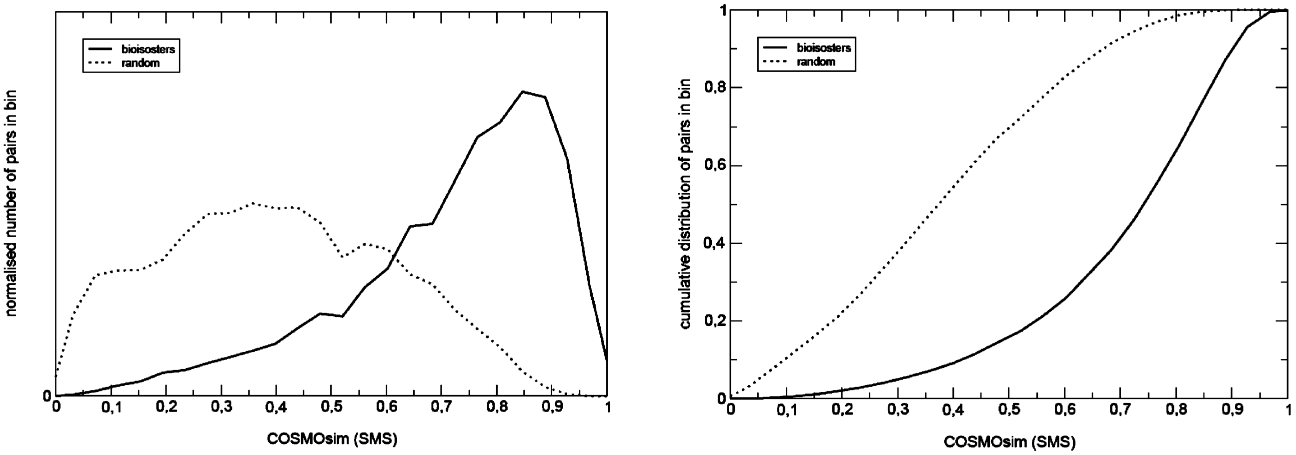

Figure 14. Separation of known bioisosters from random pairs resulting from the application of the COSMO*sim* SMS coefficient to the BioSter data set (left, normalized distribution; right, cumulative normalized distribution).

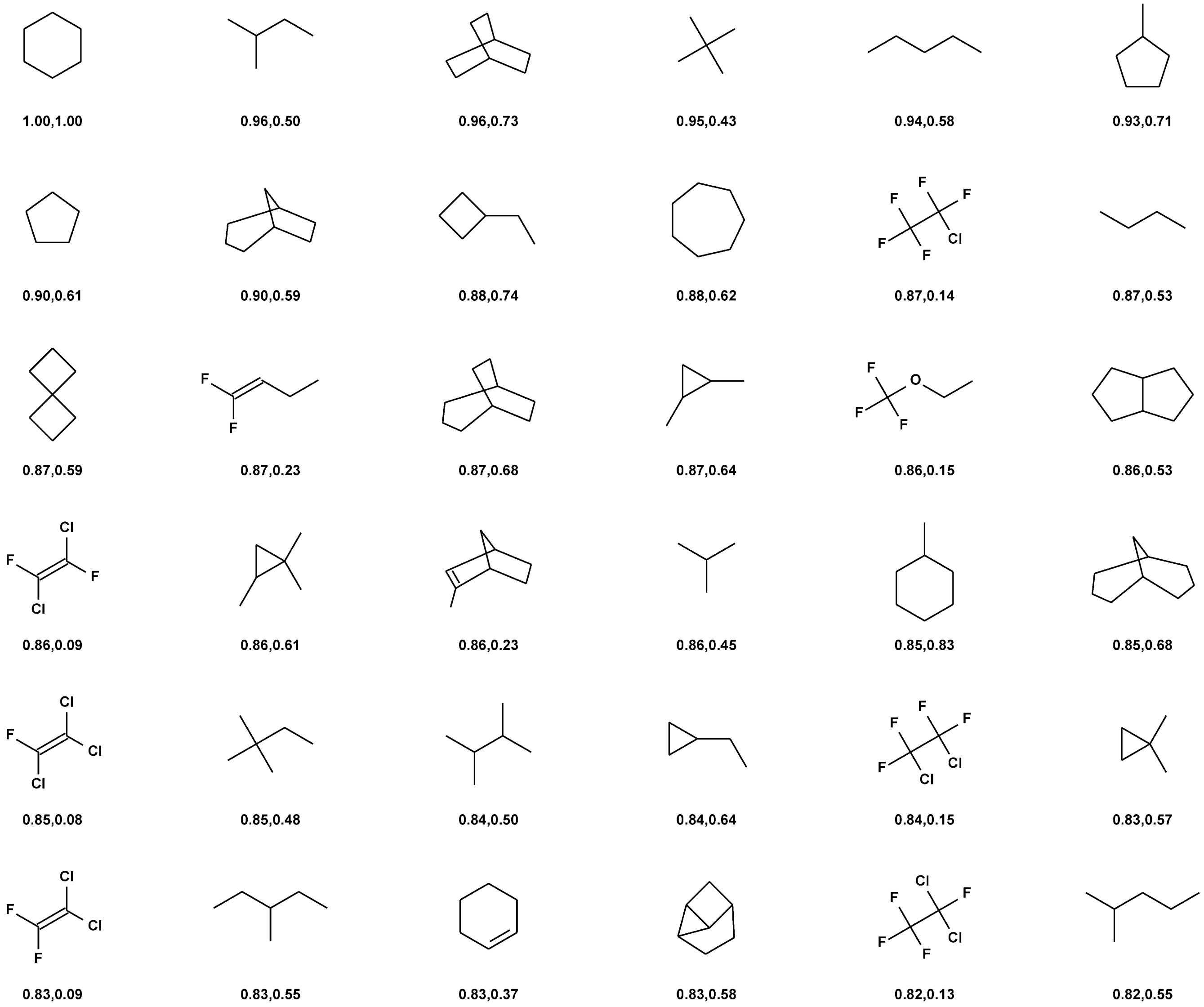

Figure 15. Bioisosters resulting from vHTS using cyclohexane as the target. COSMO*sim* SMS coefficients are given below the structures (left) along with the corresponding Daylight key Tanimoto coefficient (right).

four small molecules that are found incorporated in a plethora of natural compounds and xenobiotics and for which many bioisosteric conformations are known, that is, cyclohexane, naphthalene, thiazole, and propionic acid. The σ profiles of these compounds were calculated, and the σ-profile database was screened for the most similar molecules according to the COSMO*sim* SMS coefficient. To make a comparison to well-established 2D fingerprint similarity techniques possible, we calculated the Tanimoto coefficient on the basis of 1024-bit Daylight keys,[31] often just (wrongly) called Tanimoto

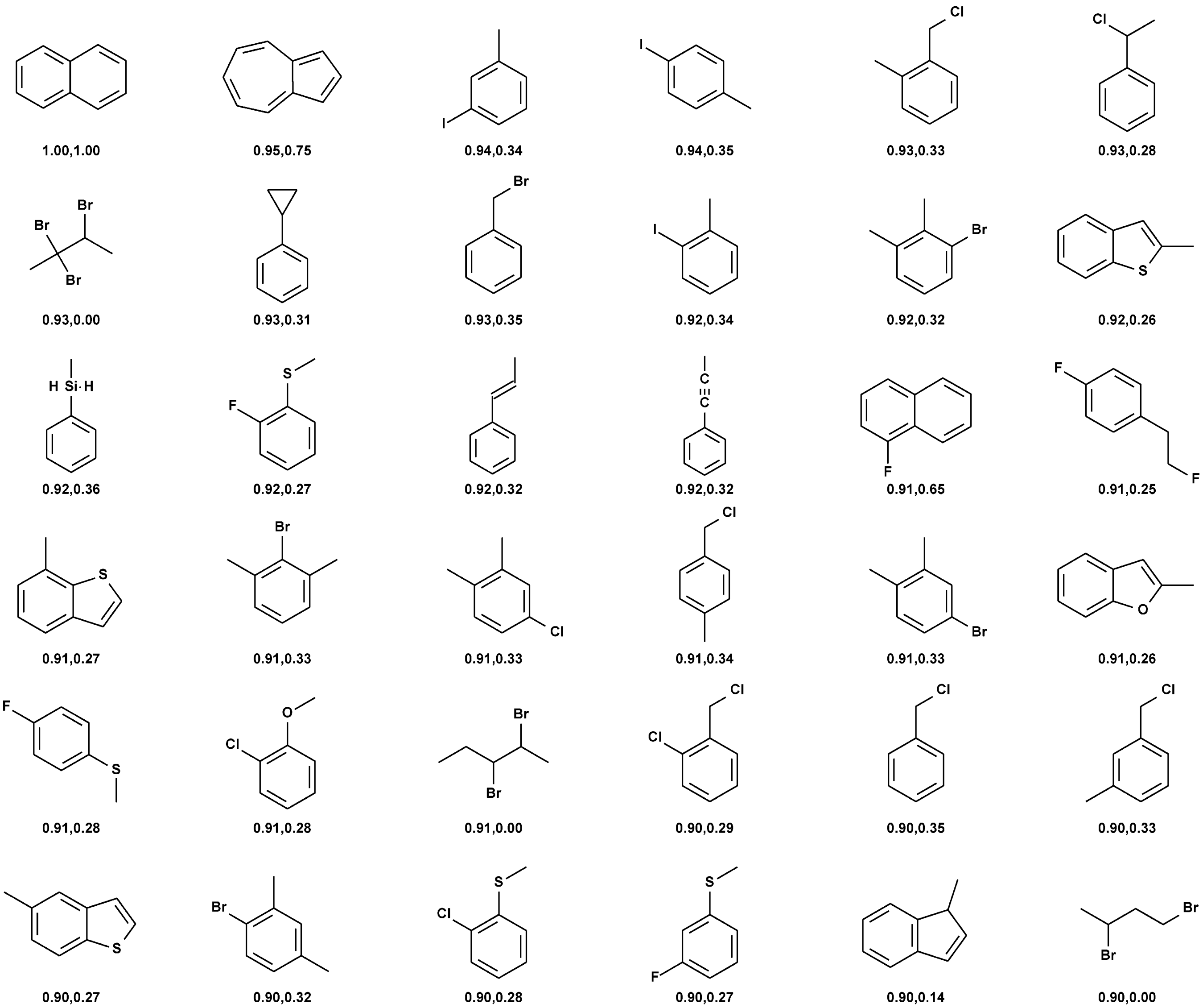

Figure 16. Bioisosters resulting from vHTS using naphthalene as the target. COSMO*sim* SMS coefficients are given below the structures (left) along with the corresponding Daylight key Tanimoto coefficient (right).

similarity, for the top-ranking compounds. The results are given in Figures 15−18.

Whole Molecule Transformations with COSMO*sim*. A virtual screening application of COSMO*sim* was performed. The target ligand, chlorpromazine, is a member of the tricyclic antidepressants class. The most similar compounds of the σ-profile database, in terms of SMS, are given in Figure 19 along with the corresponding Daylight fingerprint T_c values.

RESULTS AND DISCUSSION

COSMO-RS and COSMO*therm* provide the basis for a rigorous treatment of the energetic contributions of protein−ligand interactions. The only necessary assumption is that the ligand, receptor, and noncovalent ligand−receptor complex behave as pseudo-liquids or isotropic phases. This implies that there is no steric repulsion between the ligand and receptor upon complex formation, which is certainly valid for many potent, highly efficient ligands.[4] Steric mismatch and repulsion and their energetic consequences are, thus, beyond the scope of COSMO*sim*.

We have focused our work on ligand-based design, following the concept of bioisosteric transformations, that is, hypothetical reactions linking compounds A and B. Bioisosteric transformations target the same receptor and can, thus, be formulated as the thermodynamic cycle given in Figure 5. This thermodynamic cycle provides the basis for the calculation of the energetic cost, $\Delta\Delta G_{\mathrm{AP,BP,binding}}$, of the bioisosteric transformation on the basis of structure-free 1D σ profiles and can be regarded as the biophysical background of COSMO*sim*.

For practical applications, this energetic cost has to be low, on the order of a few kilocalories per mole, to not lose too much binding affinity. For a large number of successful bioisosteric transformations and random molecular pairs, we could show that bioisosters do indeed have more similar σ profiles than random molecular pairs (see Table 2). The simple Tanimoto similarity coefficient was apparently improved by the introduction of the relative molecular sizes, as outlined in the alcohol example (see Figure 6). The number of bioisosteric pairs retrieved at the same energetic cost (see Table 1) was, though, slightly lower for the

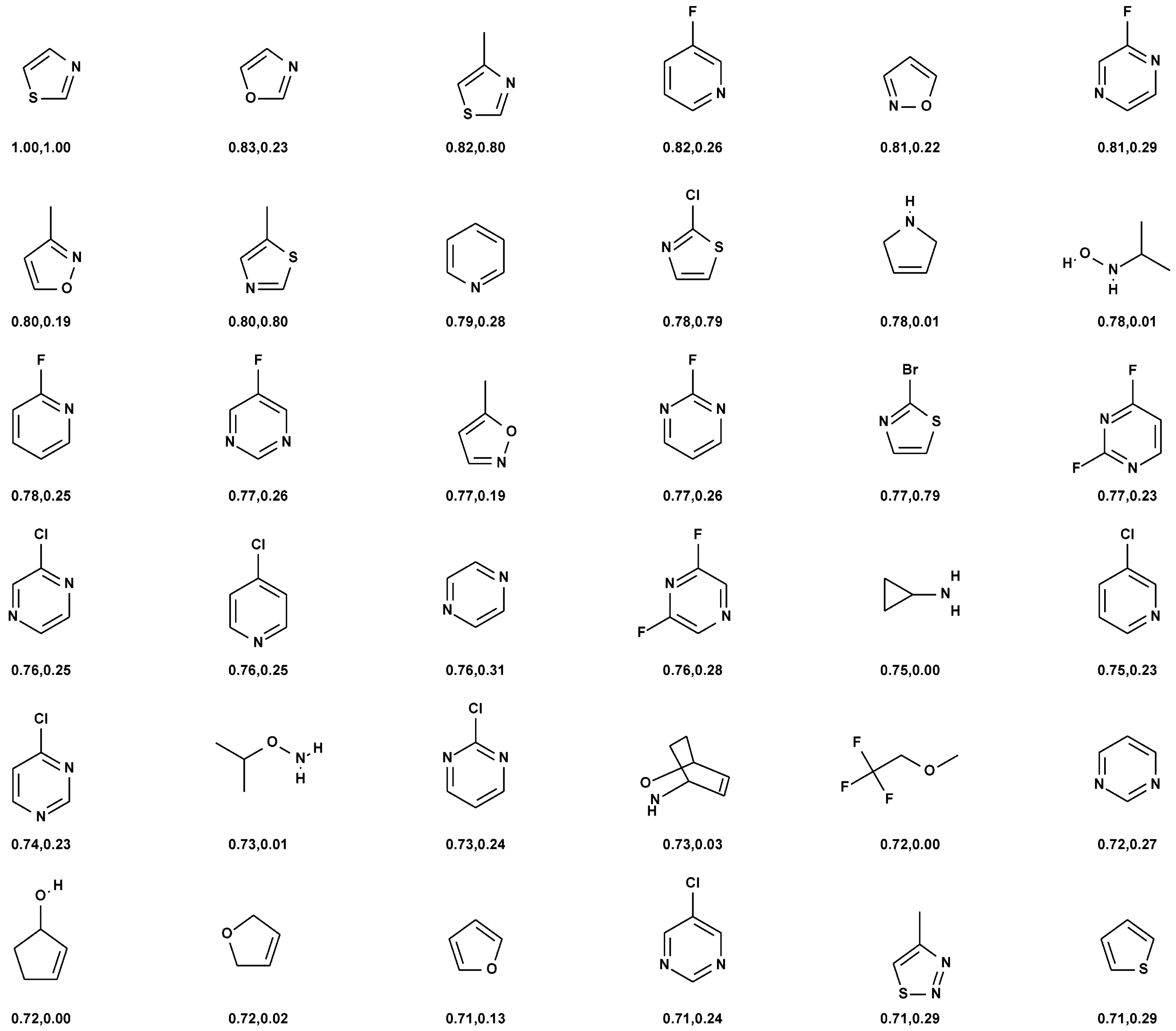

Figure 17. Bioisosters resulting from vHTS using thiazole as the target. COSMO*sim* SMS coefficients are given below the structures (left) along with the corresponding Daylight key Tanimoto coefficient (right).

Tanimoto prime coefficient. The neighborhood of bins in the σ profile and the resulting biophysical impact are completely neglected by both Tanimoto-based coefficients. Therefore, we developed the SMS metric that accounts for the biophysical relevance of neighboring bins, the distinct importance of polar versus apolar regions, and differences in molecular size.

The four free parameters of the SMS were optimized to achieve maximum separation of the bioisosters from random pairs. A closer investigation of the parameter sets obtained from the GA optimization as given in Table 1 reveals the following trend: improved metrics are obtained with a balanced σ tolerance ($a_{\text{best3}} = 2.533$ and $a_{\text{best4}} = 2.561$), while its neglect leads to a bad separation ($a_{\text{worst}} = 1.096$); a harsh decrease of the σ tolerance in the polar region leads to bad separation ($b_{\text{worst}} = 0.009\,430$), while a slight decrease yields a good separation ($b_{\text{best3}} = 0.000\,350$ and $b_{\text{best4}} = 0.000\,124$); the similarity of polar regions must be weighted higher than that of apolar regions ($c_{\text{best3}} = 0.009\,960$ and $c_{\text{best4}} = 0.009\,990$) otherwise the separation suffers ($c_{\text{worst}} = 0.001\,310$); and molecular size should not change for a bioisosteric transformation otherwise the separation suffers ($d_{\text{worst}} = 0.993$).

These findings are in accordance with previous findings concerning the size of bioisosteric groups[10] and the general importance of polar groups for binding[6] and solvation. The most striking difference in SMS as opposed to the Tanimoto coefficients is, however, the introduction of the σ-tolerance parameter a that does, indeed, lead to a great improvement of the method. The application of the various coefficients to the bioisoster and random pair molecule sets show that all COSMO*sim* coefficients can be used for the prediction of bioisosters. However, the SMS metric is to be preferred over the Tanimoto-coefficient-based measures because it is built upon a sound biophysical concept. Moreover, SMS retrieves, about twice as often, bioisosters at the same energetic cost for the bioisosteric transformation, as expressed in the mean square change in solvation $\Delta\Delta G_{\text{A,B,solv}}$ (see Table 1).

COSMO*sim* is extremely fast (29 689, 28 078, and 9869 compounds per second on a single 3 GHz Pentium 4 CPU

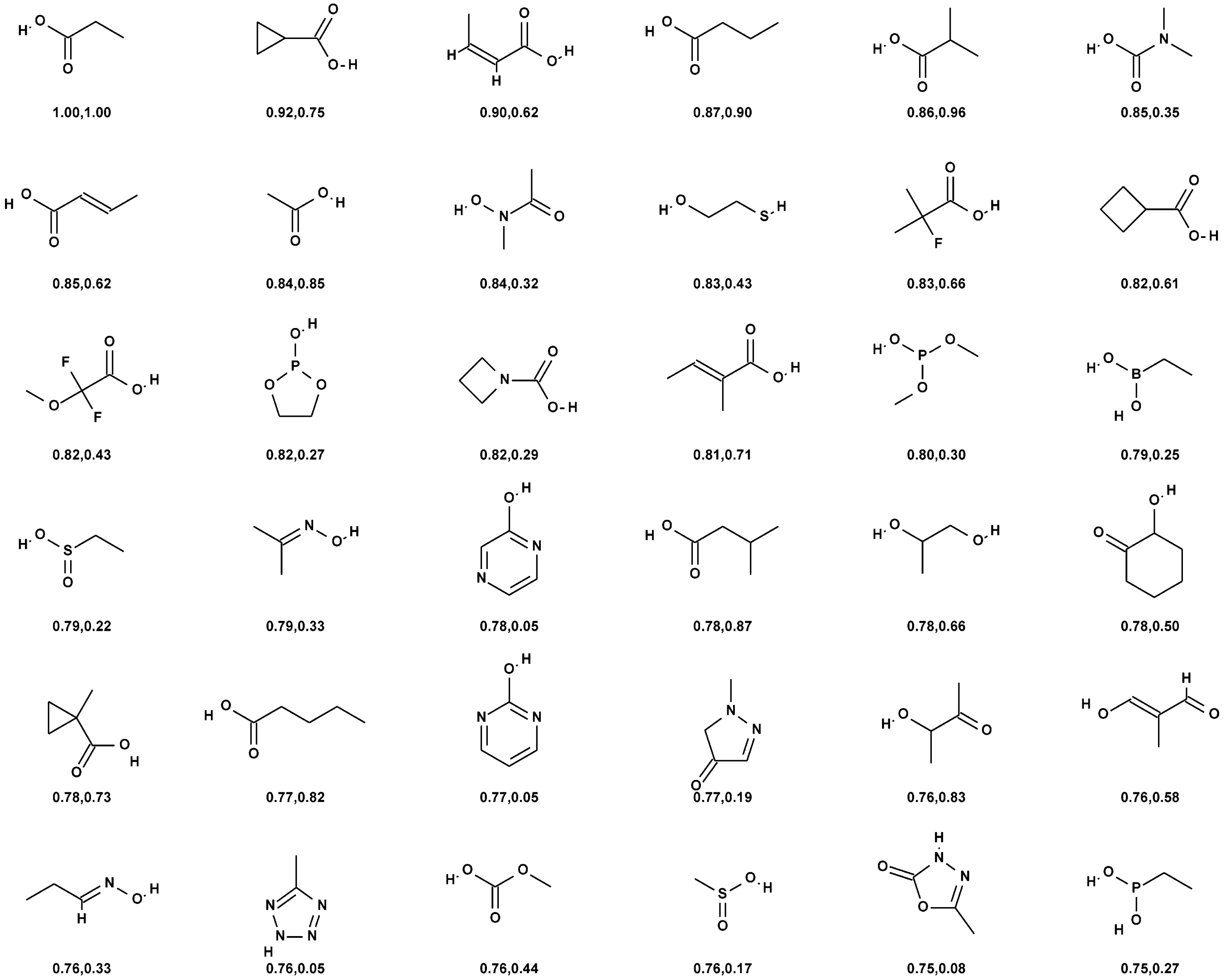

Figure 18. Bioisosters resulting from vHTS using propionic as the target. COSMO*sim* SMS coefficients are given below the structures (left) along with the corresponding Daylight key Tanimoto coefficient (right).

for T_c, T_c', and SMS) and is, therefore, well-suited for the virtual HTS of billions of compounds. We recommend performing explicit calculations at the DFT BP-SVP-COSMO-SP level using a bioactive conformation of the target ligand in order to improve the quality of the target σ profile. Moreover, explicit calculations can be run overnight on the top-ranking compounds because the corresponding DFT-based σ profiles provide a basis for the calculation of biophysically relevant properties such as pK_a, log P, log S, and so forth with higher accuracy.[19,20,32]

The assessment of the group transformation applications reveals clearly that COSMO*sim* perceives molecular similarity in a way that medicinal chemists do. The results given in Figures 15 and 16 underline the apolar aliphatic and aromatic characters and sizes of cyclohexane and naphthalene bioisosters, respectively. These simple bioisosteric transformations are, though, not at all perceived by the Daylight key (substructure) based similarities. Similar results are expected for Unity or MDL keys, but one has to keep in mind that these methods were mainly developed as hash keys and bioisosterism is beyond their scope, although they are widely used for this purpose.[33] The two polar group examples reveal further advantages of the structure-free comparison COSMO*sim*. Well-known bioisosters of thiazole (see Figure 17) such as oxazole and pyridine appear in the top-ranking list as expected, while the structure-based approach would miss them in every virtual screening. It is important to stress that the same structure-based approach would likely fail to rank larger compounds containing thiazole and pyridine as functional groups as well. The propionic acid example (see Figure 18) paints a good picture of how electronics and molecular size are reflected in COSMO*sim*. Obviously, quite a few hits are carboxylic acids of similar size, that is, trivial structure derivatives and as such well-perceived by the Daylight key method. COSMO*sim*, however, retrieves additional well-known bioisosters such as carbamic and hydroxamic acids; enoles; phosphinic, boronic, and sulfenic acids; and acidic heterocycles such as tetrazoles and oxadiazolones. To our knowledge, COSMO*sim* is the only method that retrieves the latter nonclassical bioisosters a priori.

The bioisosteric transformation of chlorpromazine should exemplify the suitability of COSMO*sim* to the discovery and design of therapeutic copies. Chlorpromazine is one of the oldest tricyclic antidepressants, found more than 50 years ago by chance.[34] Apart from its antidopaminergic activity, it antagonizes, furthermore, histamine, 5-hydroxytryptamine, acetylcholin, and cannabinoid receptors. Since its discovery,

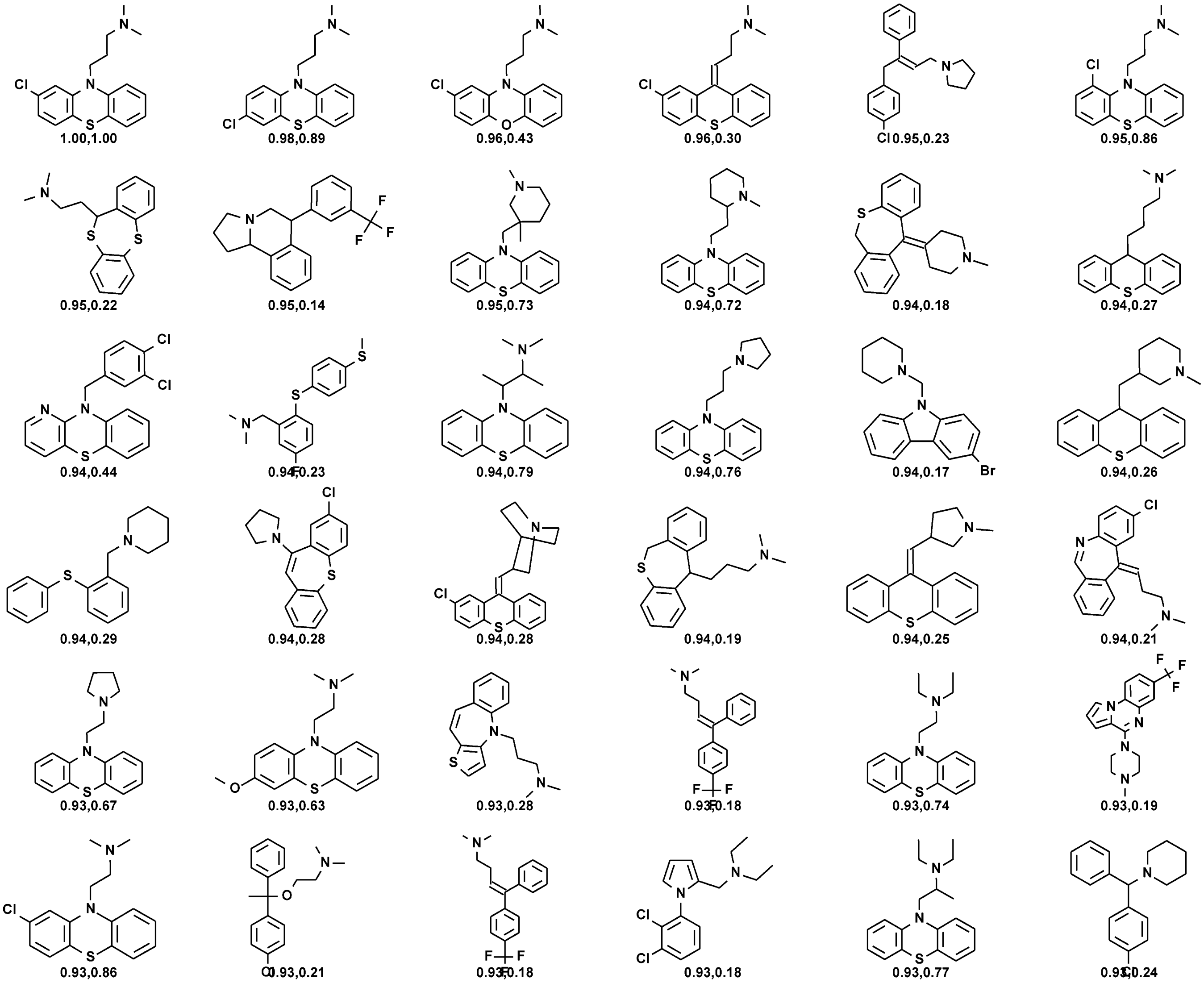

Figure 19. Bioisosters resulting from vHTS using chlorpromazine as the target. COSMO*sim* SMS coefficients are given below the structures (left) along with the corresponding Daylight key Tanimoto coefficient (right).

Table 2. Solvation Free Energy Cost-Based Threshold Similarities t_{sim} and Number of Bioisosters and Random Pairs Found To Be More Similar than t_{sim}

COSMO*sim* coefficient	std dev($\Delta\Delta G_{A,B,solv}$)								
	<1.0 kcal/mol			<1.5 kcal/mol			<2.0 kcal/mol		
	t_{sim}	bioisosters	random	t_{sim}	bioisosters	random	t_{sim}	bioisosters	random
SMS	0.883	846	0	0.789	2380	120	0.715	3379	357
T_c'	0.891	191	0	0.789	1152	6	0.689	2522	72
T_c	0.906	356	0	0.823	1477	2	0.755	2740	83

a plethora of bioisosters have been discovered and designed that focus on some central nervous system activity or another. The top-ranking compounds in Figure 19 clearly show that many of the strategies employed in medicinal chemistry are perceived by COSMO*sim*, that is, small variations around the same scaffold, ring substitutions, side-chain-length variations, and various approaches to rigidify the molecules.[26] Again, the structure-based screening would have omitted most of the chlorpromazine analogues found by COSMO*sim*. This example shows impressively that bioisosteric transformations introducing only small electronic changes can yield a huge variety of distinct proper bioisosters.

The energetic cost of a bioisosteric transformation can be measured experimentally in a biological assay after chemical synthesis. In virtual screening, target-focused solutions are ranked according to the corresponding fitness values, for example, COSMO*sim* values. It is important to know what the energetic cost of a bioisosteric transformation can be and how this cost depends on the intermolecular COSMO*sim* similarity. Unfortunately, the Bioster database does not contain any information on binding or inhibition constants that could serve to calculate the energetic cost of the corresponding bioisosteric transformations. It can be assumed, however, that bioisosteric pairs do not differ by more than 2 log orders of magnitude in their inhibition, which corresponds roughly to 2.7 kcal/mol. This value is on the order of magnitude of the solvation energy changes at relevant similarity levels, according to Table 1. We can, therefore, conclude that the energetic changes in solvation of the protein–ligand complexes $\Delta\Delta G_{\mathrm{AP,BP,aq}}$ and the protein–ligand binding $\Delta\Delta G_{\mathrm{AP,BP,cosmo}}$ are of the same magnitude. Consistently, bioisosteric transformations with high COSMO*sim* values will likely lead to proper bioisosters if no additional steric repulsion is introduced. Also, the energetic cost drops as the proper bioisosteric pair becomes more similar.

It should be underlined that other physicochemical molecular properties (log P, log S, etc.) follow the same thermodynamic cycle for a bioisosteric transformation shown in Figure 5 by substituting the protein receptor P with another pure or mixed isotropic or anisotropic phase of interest (water, octanol, blood, brain, etc.). Successful bioisosteric transformations will, therefore, yield compounds with overall similar properties with the exception of their chemical structure.

For isotropic phases, steric repulsion does not exist. Anisotropic phases such as proteins can, however, display the steric mismatch and repulsion issues, and consistently, a proper bioisosteric transformation will eventually not work for a given receptor (while it might well work for another receptor). The biological assay remains to provide the ultimate proof. COSMO*sim* was successfully employed in a variety of projects for group transformations and target-focused library design where it already delivered a variety of potent inhibitor families for different targets.

REFERENCES AND NOTES

(1) Bender, A.; Glenn, R. C. Molecular Similarity: a Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(2) Jain, A. N. Ligand-Based Structural Hypotheses for Virtual Screening. *J. Med. Chem.* **2004**, *47*, 947–961.

(3) Agrafiotis, D. K.; Myslik, J. C.; Salemme, F. R. Advances in Diversity Profiling and Combinatorial Series Design. *Mol. Diversity* **1999**, *4*, 1–22.

(4) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The Maximal Affinity of Ligands. *PNAS* **1999**, *96*, 9997–10002.

(5) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: a Useful Metric for Lead Selection. *Drug Discovery Today* **2004**, *9*, 431–433.

(6) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional Group Contributions to Drug-Receptor Interactions. *J. Med. Chem.* **1984**, *27*, 1648–1657.

(7) BIOSTER Database, Synopsys Scientific Systems. http://www.synopsys.co.uk (accessed mmm yyyy).

(8) Schuffenhauer, A.; Gillet, V. J.; Willet, P. Similarity Searching of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptor. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.

(9) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386-393.

(10) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147–3176.

(11) Burger, A. Isosterism and Bioisosterism in Drug Design. *Prog. Drug Res.* **1991**, *37*, 287–371.

(12) Grimm, H. G. Structure and Size of the Nonmetallic Hydrides. *Z. Electrochem.* **1925**, *31*, 474–480.

(13) Grimm, H. G. On the Systematic Arrangement of Chemical Compounds from the Perspective of Research on Atomic Composition and on Some Challenges in Experimental Chemistry. *Naturwissenschaften* **1929**, *17*.

(14) Kier, L. B.; Hall, L. H. Bioisosterism: Quantitation of Structure and Property Effects. *Chem. Biodiversity* **2004**, *1*, 138–151.

(15) Klamt, A. *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*; Elsevier: Amsterdam, 2005.

(16) 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.

(17) Klamt, A.; Jonas, V.; Buerger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(18) 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.

(19) 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.

(20) Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. First Principles Calculations of Aqueous pK_a Values for Organic and Inorganic Acids Using COSMO-RS Reveal an Inconsistency in the Slope of the pK_a Scale. *J. Phys. Chem. A* **2003**, *107*, 9380–9386.

(21) 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.

(22) Klamt, A.; Schueuermann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

(23) 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.

(24) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.

(25) Hornig, M.; Klamt, A. COSMOfrag: A Novel Tool for High Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1169–1177.

(26) Wermuth, C. G. *Practice of Medicinal Chemistry*, 2nd ed.; Academic Press Ltd.: New York, 2003.

(27) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.

(28) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.

(29) Eckert, F.; Klamt, A. *COSMOtherm*, version C2.1, revision 01.04; COSMOlogic KG: Leverkusen, Germany, 2004.

(30) Bender, A.; Klamt, A.; Wichmann, K.; Thormann, M.; Glen, R. C. Molecular Similarity Using COSMO Screening Charges (COSMO/3PP). *Lect. Notes Comput. Sci.* **2005**, *3695*, 175–185.

(31) *Daylight*; Daylight Chemical Information Systems: Santa Fe, NM. http://www.daylight.com (accessed Jan 2005).

(32) Klamt, A.; Eckert, F. COSMO-RS: A Novel and Efficient Method for the a priori Prediction of Thermophysical Data of Liquids. *Fluid Phase Equilib.* **2000**, *172*, 43–72.

(33) Martin, Y. C.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

(34) Shen, W. W. A History of Antipsychotic Drug Development. *Compr. Psychiatry* **1999**, *40*, 407–414.

CI050464M