COSMO*frag*: A Novel Tool for High Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry

Martin Hornig, Andreas Klamt*

COSMOlogic GmbH and Co. KG, Burscheider Str. 515, 51381 Leverkusen, Germany

CORRESPONDING AUTHOR FOOTNOTE: +49-2171-731683; hornig@cosmologic.de

ABSTRACT: The COSMO-RS method has proven its broad applicability for the accurate prediction of thermodynamic, environmental or physiological properties. Basing on quantum chemical calculations with the continuum solvation model COSMO, COSMO-RS calculations were unavoidably restricted to small to medium sized compound sets, due to the time demand of the COSMO calculations. The COSMO*frag* method, presented here, overcomes this restriction by replacing the costly quantum chemistry step by a selection of suitable fragments from a database of presently 40,000 DFT/COSMO pre-calculated molecules. Since in the COSMO-RS picture any molecular information is gathered in the so-called σ-profiles, COSMO*frag* replaces the single σ-profile by a composition of partial σ-profiles, selected by the use of extensive similarity searching algorithms. On five representative datasets the accuracy loss of COSMO*frag* vs. full COSMO-RS calculations has been shown to be only in the range of 0.05 log-units. From the performance point of view it is now possible to carry out COSMO-RS property calculations for more than 100,000 compounds a day per standard PC CPU.

Introduction

The virtual screening of compound libraries is well-established in modern drug discovery and design. Calculations of physicochemical properties of the drug candidates are essential for the estimation of their pharmacokinetics. To evaluate the so-called ADME (absorption, distribution, metabolism, and elimination) parameters basically aqueous solubility and lipophilicity are in demand[1], though for the consideration of acidic or basic compounds, partitioning and solubility become pH-dependent and pK_A or rather log D are needed additionally[2]. Lipophilicity is most commonly assessed in form of partition coefficients, usually n-octanol/water (P_{OW}). However cyclohexane/water or 1,2-dichloroethane/water[3] are partly considered as more appropriate measure for lipophilicity with regard to membrane permeability. Due to the smaller water fraction in cyclohexane or 1,2-dichloroethane these partition coefficients better account for hydrogen-bond desolvation.

The importance of the molecular electrostatics and the related hydrogen bonding and hydrophobic interactions is widely accepted in different areas of drug design. For example models like CoMFA[4] in 3D-QSAR or molecular polar surface area descriptors (PSA)[5-7] for QSPR demonstrate the broad acceptance of models describing the electrostatics of molecules.

The COSMO-RS method, a combination of the quantum chemical continuum solvation model COSMO and a statistical thermodynamics treatment for more realistic solvents (RS) simulations, is a novel, widely applicable tool for accurate predictions of many kinds of thermodynamic as well as physiological properties[8-10] . In this approach all information about solutes and solvents is gathered from initial density functional (DFT) COSMO calculations. On the basis of this very fundamental and broad knowledge of structure and electrostatics of the molecule in solution, a large set of physicochemical properties is accessible by means of the polarization (or screening) charge density σ on the molecular surface. The COSMO-RS theory has introduced this surface polarization charge density as a novel and highly significant description of the surface electrostatics. It turned out to be more local and better transferable than the electrostatic potential (ESP) itself. Since in COSMO-RS properties as $\log P_{OW}$ are calculated as surface integrals of σ-functions, this theory and its σ-perspective provide a qualitative and quantitative understanding of the widely recognized relation between such properties and surface electrostatics.

A straightforward and logical extension of the COSMO-RS methodology is the calculation of similarity coefficients based on σ[11]. This approach allows for the comparison of similarities of molecular surfaces and their electrostatics independent of the structure of the molecules, enabling scaffold hopping in a natural way.

COSMO-RS property calculations using the COSMO*therm* program[12] only require fractions of a second per compound. The overall speed of the COSMO-RS method is mainly limited by the time demand of the underlying quantum chemical calculations for the molecules. On the high quality level BP-TZVP (geometry optimaization with BP functional[13-15] and TZVP basis set[16]) such calculations take about 4 hours on average for molecules with up to 40 heavy atoms on a 3 GHz CPU, using the TURBOMOLE program package (University of Karlsruhe, Karlsruhe, Germany)[17,18]. This is acceptable for chemical engineering applications where normally only a few new molecules are considered, besides many common compounds which can be taken from a database. A database of carefully prepared BP-TZVP-COSMO files for 3,000 common compounds and solvents is available (COSMO*base*)[19].

This differs strongly in the area of drug design. Here often up to hundreds of thousands or even millions of potential drug candidates have to be pre-screened regarding their physiochemical properties or biological activities, each of them being typically in the range of a molecular weight of 300 - 500, i.e. having about 25 - 40 heavy atoms. Therefore we have introduced a slightly more approximated “drug calculation level”, which uses BP-SVP[20,21] single point DFT/COSMO calculations on semi-empirical MOPAC[22] AM1/COSMO geometries. This level reduces the computation time of typical drug molecules by approximately a factor of 30, i.e. roughly to 8 minutes per drug. Still on this level a pre-screening of compound numbers as large as that is unfeasible even on large parallel computer clusters. For these applications a very fast bypass for the demanding DFT/COSMO calculations called COSMO*frag*[23] has been developed. This is described in the present paper. The basic idea is to avoid the time consuming DFT/COSMO calculation of the screening charge densities (σ-profiles) for each individual molecule and to replace it by a composition of partial σ-profiles taken from locally most similar fragments of molecules whose DFT/COSMO files are stored in a database. It should be noted that the database does not consist of molecule fragments but of entire molecules and the fragmentation is individually composed from these molecules respectively. These fragment based σ-profiles can then be used as starting point for any COSMO-RS calculation, i.e. physicochemical, physiological or environmental[24] property calculations, similarity searching or even receptor binding approaches.

General COSMO-RS Theory

COSMO-RS is a model combining quantum theory, dielectric continuum models, surface interactions and statistical thermodynamics. The theory of COSMO-RS has been described in detail in several articles[25-27]. Therefore we will only give a short survey of the basic concept here and refer the interested reader to these articles for details.

COSMO-RS considers a liquid system as an ensemble of molecules of different kinds, thus solvent or solvent mixture and solutes. Precondition is a density functional (DFT) calculation with the dielectric continuum solvation model COSMO[28] for each kind of molecule X, in order to get the total energy E^{X}_{COSMO} and the polarization (or screening) charge density (SCD) σ on its molecular surface. The COSMO calculation has to be carried out only once per compound and thus COSMO files can be stored for future use. The σ value is a good local descriptor of molecular surface polarity[29].

For the purpose of an efficient statistical thermodynamics calculation the liquid ensemble of molecules now is considered as an ensemble of pair-wise interacting molecular surfaces. The most important parts of the specific interaction between molecular surfaces, i.e. electrostatics (es) and hydrogen bonding (hb), are expressed by the SCDs σ and σ' of the contacting surface pieces:

$$E_{es}(\sigma,\sigma') = \frac{\alpha'}{2}(\sigma+\sigma')^2 \qquad (1)$$

and

$$E_{hb}(\sigma,\sigma') = c_{hb} \min\{0, \sigma\sigma'+\sigma_{hb}{}^2\} \qquad (2)$$

The three parameters α', c_{hb}, and σ_{hb} have been adjusted to a large number of thermodynamic data. Since all relevant interactions depend on σ, the distribution functions (histograms) $p^X(\sigma)$ are required for the statistical thermodynamics.

$p^X(\sigma)$, in the following called σ-profile, displays the composition of the ensemble of surface pieces with respect to σ. The σ-profile of a special molecule has characteristic shape and provides a vivid picture of the molecular polarity (see Fig. 1, and Klamt et al.[26,28]). Furthermore, we need the σ-profile $p^S(\sigma)$ of the ensemble S, which is simply calculated as a sum of the molecular σ-profiles weighted by mol-fractions.

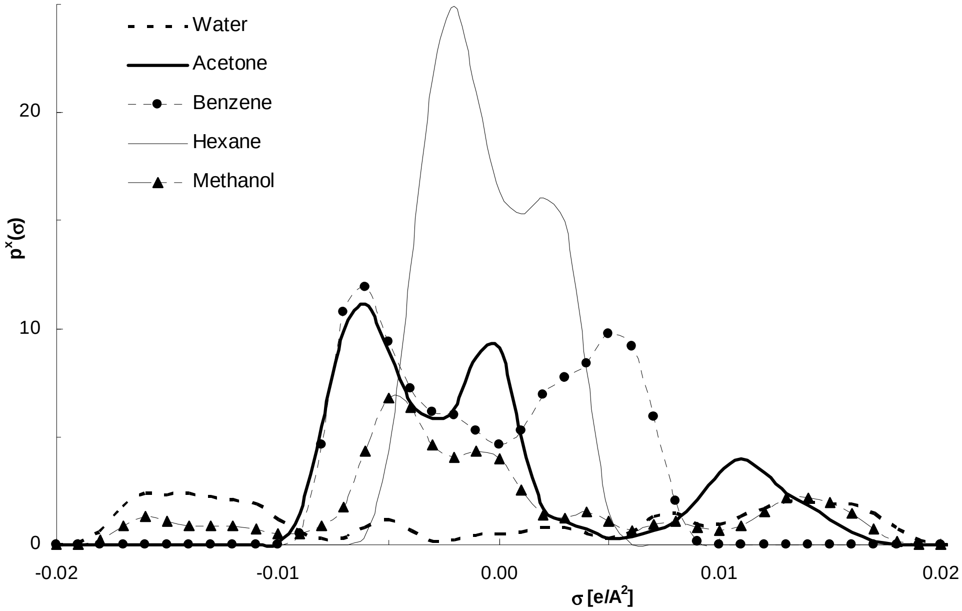


Figure 1. Solvent σ-Profiles. These profiles show the amount of molecular surface in a given interval of polarization charge density σ.

Now the chemical potentials of the compounds in the solvent are calculated by a novel, exact, and very efficient statistical thermodynamics procedure. The first step is the iterative solution of the equation

$$\mu_S(\sigma) = -\frac{RT}{a_{eff}} \ln\left\{ \int d\sigma' \, p_S(\sigma') \exp\left(\frac{a_{eff}}{RT} (\mu_S(\sigma') - E(\sigma, \sigma')) \right) \right\} \qquad (3)$$

This implicit equation, in which a_{eff} denotes an effectively independent piece of molecular area and E(σ,σ’) the sum of the energy contributions of eq. 1 and 2, can be solved by iteration within milliseconds on a PC.

The resulting function $\mu_S(\sigma)$, the σ-potential, describes the solvent behavior regarding electrostatics, HB-affinity and hydrophobicity. In a second step the σ-potential is integrated over the surface of each compound X, yielding the chemical potential of X in S:

$$\mu_S^X = \int p^X(\sigma) \mu_S(\sigma) d\sigma + \mu_{combS}^X \qquad (4)$$

In this equation the surface integral is evaluated as σ-integral, making use of the σ-profile of solute X. The combinatorial contribution $\mu^X_{comb,S}$ to μ takes into account size and shape effects of solute and solvent[27]. Usually it is small compared to the first term in eq. 4 which results from the surface interactions. It is sufficient to consider it as a solvent specific constant, here.

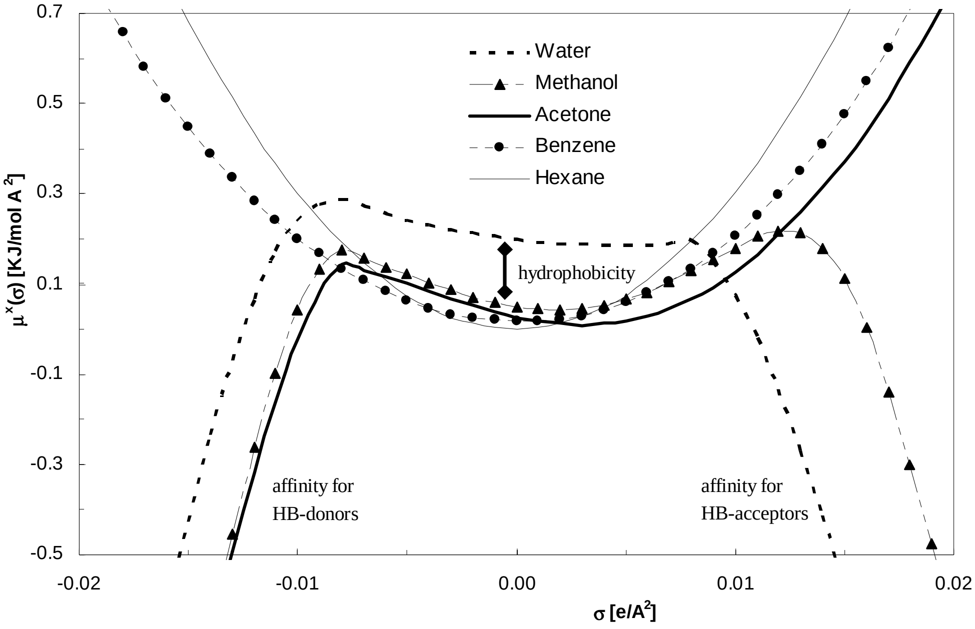


Figure 2. σ-Potentials of Solvents. These curves show the chemical potential of surface pieces of polarization charge density σ in a solvent. Thus they quantify the affinity of a solvent to surface of polarity σ.

Starting from a quantum chemical calculation for each compound, we found as a result of a few statistical thermodynamical steps an expression for the pseudo-chemical potential of an almost arbitrary chemical compound X in an almost arbitrary solvent S, which may be a pure compound or a mixture. This allows for the calculation of any partition coefficient as well as solubility. The few adjustable parameters required in COSMO-RS have been fitted to a large set of experimental data[25].

COSMO*frag* Methodology

The polarization charge density σ is a rather local feature of the molecular surface. Therefore it reasonably can be assumed that structural similar regions of molecules give similar contributions to the σ-profile. As a simple example the contribution of the sp^3-oxygens in water and methanol can be considered which exhibit an almost identical contribution to the σ-profiles, as can be seen in figure 1. Thus it is plausible to assume that the σ-profiles of larger new molecules can well be approximated by contributions taken from other, locally most similar molecules. Since for most COSMO-RS applications only the σ-profile and some information about the area and volume of the molecule is needed, the basic idea of COSMO*frag* is the composition of the σ-profile of new molecules from existing σ-profiles of molecules that have already been pre-calculated and are stored in a database. For this purpose a database of presently 40,000 COSMO files of highly diverse, smaller basic- and larger drug-like compounds has been prepared.

Conformers. A full conformational analysis of such large numbers of compounds is hardly feasible. In spite of that, in some cases a single conformation of a molecule is insufficient for property predictions of highest accuracy, and even depending on the solvent the favourable conformer may differ. Non high throughput calculations with COSMO*therm* allow for the utilization of different conformer COSMO files by weighting them using the COSMO energies and their chemical potentials. Proceeding as differentiated like that is impracticable within a high throughput application. Therefore the COSMO*frag* database (CFDB) consists of one single conformation for each database compound.

Attempts have been made to compose COSMO*frag* databases from sets of conformers generated by two different quantum chemical procedures. One CFDB was built up from lowest energy conformers optimized on the MOPAC AM1 gas phase level, the other with

MOPAC AM1 COSMO optimized structures, representing the opposite cases of polar and nonpolar surrounding. The results of any calculation listed in table 1 on both of these special CFDBs without exception were slightly worse (1 – 10 % of rms), when compared with the standard CFDB. It therefore can be concluded that the influence of the conformational selection on the prediction results on average is rather small. Even a better suitability of the polar versus the nonpolar CFDB, e.g. for water solubility predictions, could not be found.

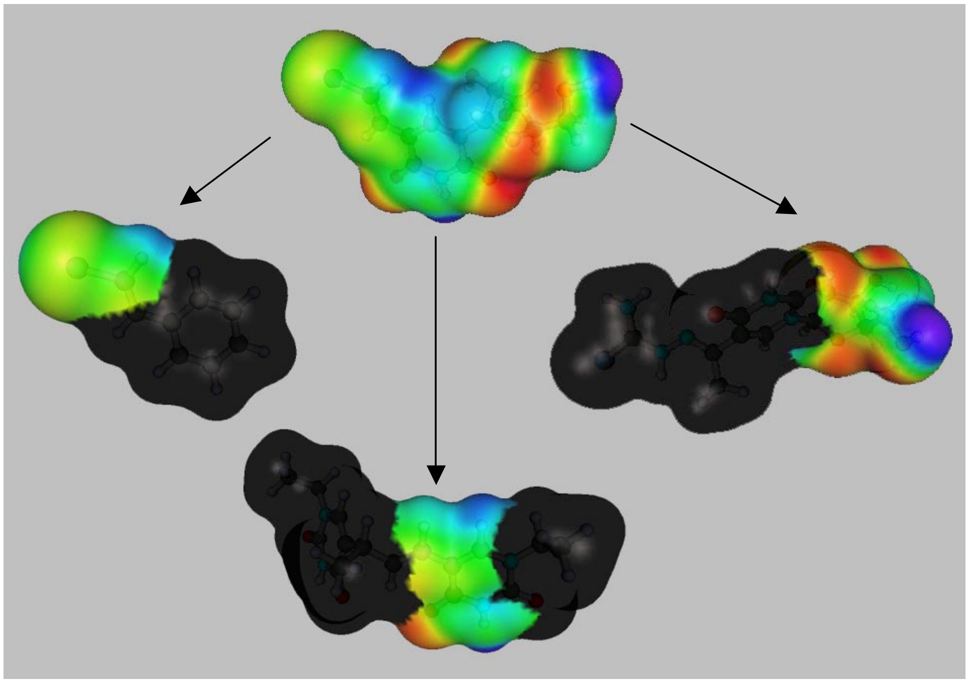


Figure 3. COSMO*frag* decomposition of the drug Sorivudine into 3 fragments. The parts of the COSMO surface not used for the decomposition are colored in black.

Molecules entering the CFDB are subject to a standard procedure regarding their geometrical optimization. This heuristic procedure has been developed to gain reliable quantum chemically optimized geometries from 2D structures. As a first step a 3D starting structure is generated using CORINA[30] , followed by an AM1/COSMO geometry optimization using a modified MOPAC7[31] version, which is customized to produce better geometries especially for amines, sulphonamides and phenols. Finally a single point

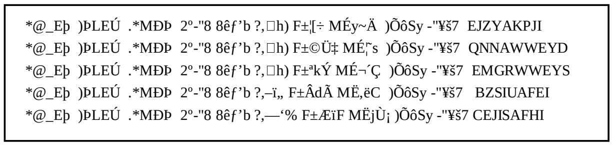
DFT/COSMO calculation (BP-SVP) with TURBOMOLE is performed. In particular cases the MOPAC geometry optimizations, likewise MOPAC7 or an up-to-date version, result in wrong structures in comparison to the given 3D structure. This is typically the case for compounds containing multiple sulphur or phosphorus atoms. The optimization for these molecules is alternatively carried out with a DFT geometry optimization (BP-SVP) with TURBOMOLE.

COSMO*frag* Similarity Algorithms. As core functionality COSMO*frag* possesses a molecular perception routine, which analyses a molecule with respect to the hybridization states of atoms, bond orders, rings, and ring properties such as aromaticity, and even their stereochemical classification. This perception is able to start from most different common electronic file formats, such as SDF files, SMILES code, XYZ files, and others. Some of them include bond tables with bond orders, others include elements and geometry only, some have explicit hydrogens, and other implicit hydrogen atoms have to be analyzed. It is most important that such a perception routine is able to end with a unique internal representation of the molecule, independent of all the different ways the chemical structure can be represented in the original input. Furthermore, care must be taken that equivalent atoms in a molecule also have an equivalent description. As an example the nitro group may be considered. Normally represented by a four-valent nitrogen atom with formal charge +1, one single bonded oxygen with formal charge –1, and a double-bonded oxygen atom, it is as well frequently given with neutral five-valent nitrogen with two neutral double bonded oxygens. Although the former description may be closer to chemical correctness, in COSMO*frag* the second convention is applied, as it describes the two oxygen atoms as chemically equivalent. In the same way, all cases of partially ionic bond descriptions had to be reduced to a neutral multiple bond representation. A unique representation of aromatic bonds is also of crucial importance, however the usual Kekule description of alternating single and double bonds leads to non-unique representations.

Once a unique representation of all atoms, bonds, and rings in the molecule is achieved, the second major step is the definition of the most useful measure for local similarity of atoms and atomic environment. For COSMO*frag* atoms should be considered as most similar, if their partial molecular surfaces and surface polarities, i.e., SCDs σ, are most similar. But since the latter is not known, at least for the new molecule under consideration, we have to ensure that the local geometries and the electronic effects of the surrounding atoms are most similar. Obviously, two similar atoms should at least be identical with respect to their element and their hybridization. By the usage of hashing algorithms this information is turned into a unique real number for each atom, a similarity index of lowest order (0th order). Since hydrogen atoms are not considered explicitly, also the number of implicit hydrogen atoms is included in the 0th order similarity index. In a next step a similarity index of 1th order can be defined by propagation of the 0th order similarity indices of the neighbor atoms to the central atoms and the addition of this new information to the 0th order similarity index of the central atom. Bond orders of the bonds used for the propagation are explicitly taken into account. By doing so 1th order identity of two atoms ensures 0th order identity of all neighboring atoms. In the same way we can now generate the next higher similarity indices out of the similarity indices of all neighbors and continue this up to any level we require. Identity of the ith order similarity index will ensure chemical identity up to the ith order neighbor spheres of the atoms. Additionally detailed information on ring sizes, cis- and trans- isomers and hydrogen-bond donor or acceptor atoms is incorporated into the similarity indices. Since for example a carbon atom of cyclohexane must be distinguished from a central atom of a long-chain alkane, the information about the minimum ring size to which an atom belongs is integrated into the similarity indices, starting at 0th order for 3- and 4-membered rings and continuing in this way for the higher similarity indices with the next higher ring sizes. Information about the cis- or trans- position with respect to double bonds is taken into account at the 2th order and the information whether a typical hydrogen-bond donor or acceptor atom can make a favorable intramolecular hydrogen bond, by forming a 5- or 6-ring, is included in the 2nd order similarity indices as well. Especially the ability to form intramolecular H-bonds may strongly change the properties of atoms in a molecule and are thus important for finding similar atoms in other molecules in the sense of COSMO-RS.

COSMO*frag* makes use of these similarity indices in two ways. First, it calculates the sum of the similarity indices of order 7 for all atoms of a molecule. The resulting molecular similarity index is essentially an identity index, because to our best knowledge identical indices imply identity of the molecular structures. In COSMO*frag* this index is called “unique name” and is used to detect the identity of molecular structures in the database. However more important is the use of the atomic similarity indices for database screenings regarding the highest similarities of atoms. For that purpose each of the eight similarity indices of an atom (0th to 7th order) is converted into a five digit ASCII word and afterwards combined to a 40 digit string. Then the identity of the atomic ASCII similarity-strings up to the 5th digit ensures 0th order similarity, identity up to 10th ensures 1th order similarity, etc.

The atom- similarity strings of all atoms of the COSMO*frag* database are stored as a sorted ASCII list. For the 40,000 molecules with an average of about 17 heavy atoms this results in a list of more than 700,000 entries. For an atom out of a new molecule the most similar atom in the database can now be found by a very efficient binary search. Obviously, in many cases there will be more than one atom in the database having the same maximum similarity level. All these atoms are considered as candidates for fragment formation for the new molecule under consideration. In a final step the fragments are built

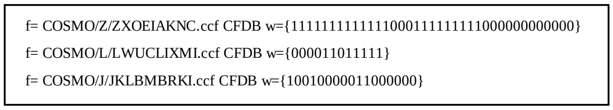


```
*@_Eþ  )ÞLEÚ  .*MÐÞ  2º-"8 8êƒ’b ?,□h) F±¦[÷ MÉy~Ä  )ÕôSy -"¥š7  EJZYAKPJI
*@_Eþ  )ÞLEÚ  .*MÐÞ  2º-"8 8êƒ’b ?,□h) F±©Ü‡ MÉ¦˜s  )ÕôSy -"¥š7  QNNAWWEYD
*@_Eþ  )ÞLEÚ  .*MÐÞ  2º-"8 8êƒ’b ?,□h) F±ªkÝ MÉ¬´Ç  )ÕôSy -"¥š7  EMGRWWEYS
*@_Eþ  )ÞLEÚ  .*MÐÞ  2º-"8 8êƒ’b ?,–ï„ F±ÂdÃ MË‚ëC  )ÕôSy -"¥š7   BZSIUAFEI
*@_Eþ  )ÞLEÚ  .*MÐÞ  2º-"8 8êƒ’b ?,—‘% F±ÆïF MËjÙ¡ )ÕôSy -"¥š7 CEJISAFHI
```

Figure 4. Random section of the COSMO*frag* database file: five digit atom words are separated by blanks. The first 8 words are the atom similarity codes of 0th to 7th order,

followed by 2 by-codes containing additional information. At the end of each line the molecule of each atom is marked, using a 9-letter unique name constructed from the molecule identity index, followed by information on the neighbor atoms and implicit hydrogens. In this section all atoms are identical up to the 4th order, while atoms 1 and 2 are most similar (up to 5th order)

from all the candidate atoms in such a way that a small number of fragments is ensured. The result of this extensive selection process is written in a COSMO metafile, displaying the chosen fragment molecules and their selected atoms respectively.



```
f= COSMO/Z/ZXOEIAKNC.ccf CFDB w={1111111111111000111111111000000000000}
f= COSMO/L/LWUCLIXMI.ccf CFDB w={000011011111}
f= COSMO/J/JKLBMBRKI.ccf CFDB w={10010000011000000}
```

Figure 5. COSMO metafile coding the parts of database molecules to be used as pictures for the construction of a new molecule. Database molecules are named with their unique name

Results and Discussion

Results. A selection of five datasets has been chosen to evaluate the accuracy of the prediction of different physicochemical, physiological or environmental properties. Table 1 displays the statistics of the COSMO*frag* calculations versus experiment on the one hand and the calculations on the full COSMO files on the other. Owing to the individual fragmentation based on the described concept of maximum similar substructures the accuracy loss of COSMO*frag* is always below 0.05 log-units compared to direct DFT/COSMO calculations. These results demonstrate the ability of the COSMO*frag* algorithms to compose sets of reasonable fragments as substitute for a molecule under consideration and support the approach of partial σ-profiles.

Since the CFDB has been constructed and extended under the aspect of maximum structural diversity in optimal representation of typical basic and life science compounds, it meanwhile has achieved a status which ensures a good representation of most compounds appearing in life science or drug design projects. The CFDB will be further extended in future by parsing additional datasets for less well represented compounds and adding these wherever possible. Presently for a very small portion of typical datasets a reasonable fragmentation is not possible due to missing fragments (see table 1). Beyond such rare fragmentation failures, bad fragmentations may rarely occur in other cases. Due to wrong or incompatible conformations of the fragment molecules in the database or weakness of the similarity algorithms, unreasonable fragmentations may occur. Such unfavorable metafiles often can be detected in the COSMO*frag* results by their noticeable total COSMO charge. Altogether 2 – 4 % of the molecules of a dataset in average cannot be satisfactorily processed by the present COSMO*frag* release.

Table 1. Results statistics for calculations of different physicochemical, physiological, and environmental properties with COSMO*frag* and on full COSMO files

Dataset	**N***	**Property**	**COSMO*frag***		**(full) COSMO**	
		($\log_{10}$ units)	**RMS**	**MUE**	**RMS**	**MUE**
Pesticides[a]	107/105	Water Solubility	0.62	0.50	0.60	0.44
Pesticides[b]	53/50	Soil Sorption	0.75	0.60	0.72	0.62
BOSS[a]	150/147	Water Solubility	0.70	0.56	0.66	0.52
PHYSPROP[c]	2570	Pow	0.62	0.50	0.59	0.47
Abraham[d]	170/166	Intestinal Absorption	15.24	9.78	14.86	10.22

a published in ref.[9], *b* published in ref.[24], *c* selected from PHYSPROP[32], *d* data from[33], * number of compounds, for COSMO*frag* calculations mostly smaller, due to missing appropriate or selected inappropriate fragment molecules in the database

Nevertheless, it must be pointed out that COSMO*frag* is only applicable to neutral compounds, because ionic compounds can be much less well fragmentized and are not represented in the CFDB for this reason. As a consequence, pK_a prediction which involves ionic species is not feasible with COSMO*frag*. Also some other COSMO*therm* features which involve total energy differences of molecular species or conformations are out of the scope of COSMO*frag*.

Performance Aspects. Apart from accuracy, computing time is the most important aspect when evaluating COSMO*frag*. In relation to the time demand of the quantum chemical calculations, the property calculation with COSMO*therm* is extremely fast. However in connection with COSMO*frag* the COSMO*therm* percentage of the overall runtime lies between 30 and 80 %, depending on the property to be computed and the performance of the metafile generation on the special dataset (see table 2). Basically the computation of QSPR properties like intestinal absorption or soil sorption with COSMO*therm* is significantly faster than the calculation of partition coefficients or solubilities. The time demand of the COSMO*frag* metafile generation on the other hand strongly depends on the representation of the given structures within the CFDB. If no long ranging similarity can be found in the database, the number of fragments increases and similarly the computing time. It can be stated that the overall performance typically lies below 1 second per compound, therefore allowing for the calculation of 100 – 150,000 compounds a day on a 3 GHz standard PC.

Table 2. Performance statistics for the calculations listed in Table 1

Dataset	**N***	**Property**	**CPU* Time**	**Avg. [s/comp.]**
Pesticides	107	Water Solubility	70 s	0.63
Pesticides	53/50	Soil Sorption	40 s	0.75
BOSS	150/147	Water Solubility	150 s	1.0
PHYSPROP	2570	logPow	1750 s	0.59
Abraham	170/166	Intestinal Absorption	90 s	0.53

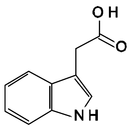
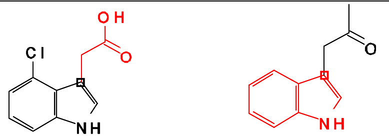
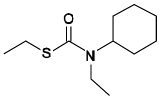
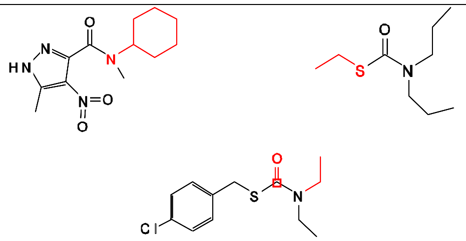
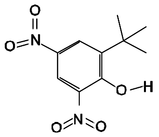
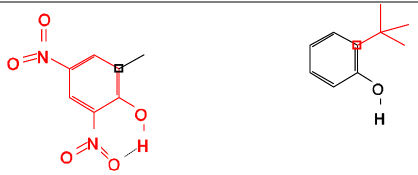
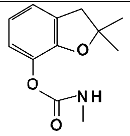
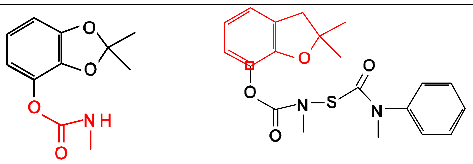
*Calculations carried out on a 3 GHz standard PC

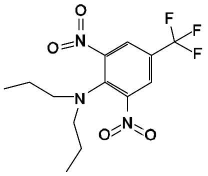
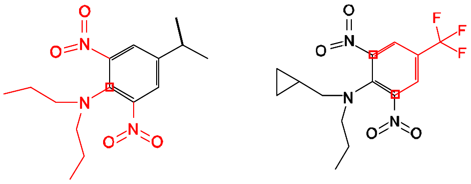
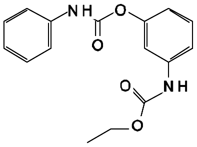
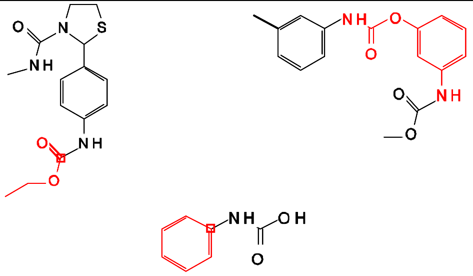
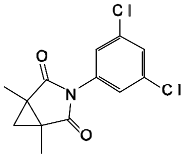
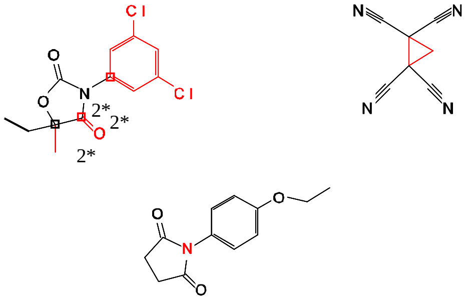
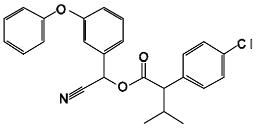
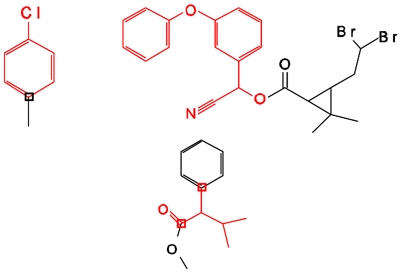
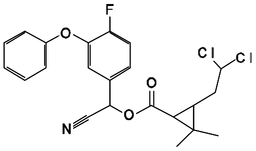
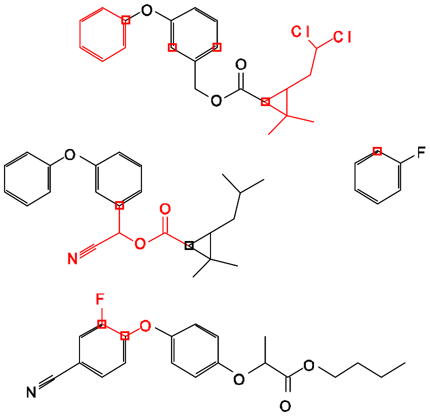
Fragmentations. Table 3 shows a number of exemplary fragmentations generated with COSMO*frag* and calculated water solubilities respectively. Basically it can be stated that highest similarity of local polarity is crucial for a good property prediction. For most of the example cases the algorithms were able to identify database molecules to be applied as fragments whose local surrounding were similar enough to assume similar electronic conditions. Compound 5 demonstrates the case of an aromatic system with strongly pulling substituents. Naturally a fragmentation of such conjugated systems requires the selection of atoms from molecules with either the identical substituent pattern or an electronically comparable one. Subdividing of such an aromatic ring in multiple fragments is justifiable if each single fragment exhibits the special electronical conditions. The generation of many thousands of metafiles has shown that in very most cases a suitable fragment molecule could be found in the database that meets the substituent pattern in demand. However, in particular cases exactly fitting aromatic fragments are missing (e.g. fluorinated benzene ring of compound 9). Even in such cases an acceptable fragmentation can be achieved by superposition of aromatic fragments that only partly meet the substituent pattern. The completion of the COSMO*frag* database concerning substituted aromatic or hetero aromatic fragments is an important goal for the future. Besides that, a few fragmentations (see compound 7) provide indications of weaknesses in the current fragmentation algorithms. For the joining atoms of cyclopropane and the succinimide ring in this example, the algorithm does not demand comparable fragments also possessing a similar condensed ring system. Therefore, the single cyclopropane fragment is chosen which exhibits a completely different polarity, especially due to the 4 nitrile substituents. Indication of such poor fragmentations is given by COSMO*frag* by means of the total COSMO charge, which is then above 0.5 in such cases.

Conformers. Conformational aspects also may have strong influence on the prediction quality for the single molecule. This is of course not only the case for the fragment molecules.

Compound 3 displays a case where a suboptimal conformer has been chosen on the side of the full COSMO calculations. The dinoterb geometry, in this case generated by a different 3D builder, exhibits no internal hydrogen bond which would be favourable here. This may be the reason for the large deviation of the COSMO*therm* calculated and experimental values here. The database fragment on the other hand shows the mentioned hydrogen bond and the solubility prediction therefore is much closer to the experimental value. For the sake of completeness, the prediction for the dinoterb, as optimized by the heuristic standard optimization procedure, yields -4.85, close to the COSMO*frag* result.

Table 3. Example fragmentations and solubilities of a handpicked number of compounds from the pesticides dataset[9]

	Compound name/ **CAS No.**	**Compound** **structure**	**Fragment** **structures**	**Log(XH2O)** **COSMO**	**Log(XH2O)** **Meta**	**Log(XH2O)** **Exp.**
1	**Indole-3-** **acetic acid** [87-51-4]			-3.62	-3.53	-3.81
2	**Cycloate** [1134-23-2]			-5.10	-5.50	-5.20
3	**Dinoterb** [1420-07-1]			-4.43	-5.12	-6.47
4	**Carbofuran** [1563-66-2]			-4.50	-4.74	-4.58

5	**Trifluralin** [1582-09-8]			-7.70	-7.72	-8.00
6	**Desmedipham** [13684-56-5]			-6.29	-6.32	-6.38
7	**Procymidone** [32809-16-8]			-6.14	-	-6.54
8	**Fenvalerate** [51630-58-1]			-9.49	-8.96	-9.37
9	Cyfluthrin **[68359-37-5]**			-9.05	-9.40	-10.04

* multiple weighted fragment atoms

Summary and outlook

The COSMO*frag* method has been introduced as high quality shortcut for almost any kind of COSMO-RS calculation. It therefore makes the COSMO-RS method applicable for high throughput tasks especially in life science. Properties of different application areas, e.g. water solubility or intestinal absorption, mostly published earlier, have been calculated with an almost negligible loss of accuracy. In the same way COSMO*frag* enables similarity screenings basing on σ-profiles for large numbers of compounds.

Though the COSMO*frag* database of presently 40,000 molecules allows for the reliable calculation of properties for almost any class of compounds in life science or drug design, it nevertheless will be extended further on, especially what the electronically complicated substituted aromats and hetero-aromats are concerned. Similarly the refinement of the COSMO*frag* hashing methodology and the systematic optimization of the CFDB molecule geometries will be carried on.

References and Notes

(1) Bergstrom, C. A.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K. et al. Absorption Classification of Oral Drugs Based on Molecular Surface Properties. *J. Med. Chem.* **2003**, *46*, 558 - 570.

(2) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to Predict log D Distribution Coefficient for Pfizer Proprietary Compounds. *J. Med. Chem.* **2004**, *47*, 5601 - 5604.

(3) Bouchard, G.; Carrupt, P.-A.; Testa, B.; Gobry, V.; Girault, H. H. Lipophilicity and Solvation of Anionic Drugs. *Chem. Eur. J.* **2002**, *8*, 3478-3484.

(4) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.

(5) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714-3717.

(6) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480-490.

(7) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; P., A. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32-39.

(8) Diedenhofen, M.; Eckert, F.; Klamt, A. Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Ionic Liquids Using COSMO-RS. *J. Chem. Eng. Data* **2003**, *48*, 475-479.

(9) Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Burger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comp. Chem.* **2002**, *23*, 275-281.

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

(11) Hornig, M.; Klamt, A. in preparation. **2005**.

(12) Eckert, F.; Klamt, A. COSMOtherm Ver C2.1-Revision 01.04:COSMOlogic KG, Leverkusen, Germany. **2004**.

(13) 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.

(14) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098-3100.

(15) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*.

(16) Schaefer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829-5835.

(17) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic structure calculations on workstation computers: The program system Turbomole. *Chem. Phys. Letters* **1989**, *162*, 165-169.

(18) Schafer, 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.

(19) Eckert, F.; Klamt, A. COSMObase Ver C2.1-Revision 01.04:COSMOlogic KG, Leverkusen, Germany. **2004**.

(20) Schaefer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms lithium to krypton. *J. Chem. Phys.* **1992**, *97*, 2571-2577.

(21) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(22) Fujitsu MOPAC2000 Program. Tokyo, Japan. **2002**.

(23) Hornig, M.; Klamt, A. COSMOfrag Version 2.1:COSMOlogic KG, Leverkusen, Germany. **2005**.

(24) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of soil sorption coefficients with a conductor-like screening model for real solvents. *Env. Tox. Chem.* **2002**, *21*, 2562-2566.

(25) 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.

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

(27) 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.

(28) 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.

(29) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: a novel view to physiological solvation and partition questions. *J. Comp. Aided Mol. Design* **2001**, *15*, 355-365.

(30) Gasteiger, J.; Rudolph, C.; Sadowski, J. CORINA program: Molecular Networks GmbH, Erlangen, Germany. *Tetrahedron Comp. Method.* **1990**, *3*, 537-547.

(31) Stewart, J. J. P. MOPAC program package (MOPAC7). *QCPE* **1993**, *455*.

(32) Howard, P.; Meylan, W. PHYSPROP DATABASE;Syracuse Research Corp.: Syracuse, NY. **2000**.

(33) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J. et al. Evaluation of human intestinal absorption data and subsequent derivation of a QSAR with the Abraham descriptors. *J. Pharm. Sci.* **2000**, *90*, 749.