COSMOlogic
Predicting Solutions

COSMO*frag*

Release 3.5

User's Manual

by COSMOlogic GmbH & Co. KG

A. Klamt

Imbacher Weg 46, D-51379 Leverkusen, Germany

Phone +49-2171-731-681

Fax +49-2171-731-689

E-mail info@cosmologic.de

Web http://www.cosmologic.de

Contents

1. Introduction

"COSMO*frag*, a fast shortcut for high-throughput COSMO-RS calculations."

The COSMO-RS method has become an efficient and versatile tool for the prediction of a large variety of physicochemical properties, especially in its efficient implementation within the COSMOtherm program. Based on quantum chemical (DFT/COSMO) calculations for the individual molecules it allows for physically most sound estimations of general vapour-liquid and liquid-liquid equilibria and of related properties like solubilities and partition coefficients. In addition it has been extended to properties like drug- and pesticide solubility, blood-brain partition coefficients, intestinal absorption, soil sorption coefficients, etc. which are of importance in the design and development of drugs, pesticides and other physiological agents.

Since thermodynamic calculations only require fractions of a second, the overall timing of a COSMO-RS investigation is mainly determined by the time demand of the underlying quantum chemical calculations for the molecules. In the field of the design of drugs, pesticides and other agents, shortly called drug design below, often a large variety of hundreds to several ten thousands of potential drug candidates has to be pre-screened regarding their physiochemical properties, each of them being typically in the range of a molecular weight (MW) of 300 - 500, i.e. having about 25 - 40 non-hydrogen atoms. A pre-screening for large numbers of compounds using quantum chemical methods is almost unfeasible even on large parallel computer clusters. In order to make COSMO-RS applicable to high-throughput calculations COSMOfrag has been developed.

The basic idea of COSMO*frag* is to avoid the time consuming calculation of the screening charge densities (σ-profiles) by DFT/COSMO calculations for each individual molecule and to replace it by a composition of the σ-profile out of partial σ-profiles taken from locally most similar fragments of molecules whose DFT/COSMO calculations are stored in a database. If for all parts of the molecule images with a sufficient degree of local similarity can be found in the database, this fragmentation causes only a minimal loss of accuracy. Since the fragmentation – even for a large database of several ten thousands of molecules - requires only parts of a second, the total property calculation including the subsequent COSMOtherm calculation takes half a second per compound on average. Thus property calculations for about 100000 - 150000 compounds are feasible within a day on a 3 GHz computer.

Probably, the most essential part of COSMOfrag is a thorough molecular perception of the input molecule which can be given in different molecular formats, including SMILES notation. In this perception eventually missing hydrogens or bond orders are added. Bonds and rings are analyzed with respect to conjugation, E-Z-substitution, and aromaticity. Finally, for each atom hash coefficients are calculated taking into account all the local information about the atom itself and about bonds and neighbour atoms. Step by step higher order hash codes are calculated which include the information about an increasing number of neighbouring spheres of the atom. Thus, two atoms that have an equivalent molecular environment up to the n^{th} neighbouring sphere have identical hash codes up to the n^{th} order. Indeed even some more distant information may be included in the n^{th} sphere since information is preferentially propagated along conjugated bonds assuming similarity along such bonds to

be more important than similarity along single bonds. The highest order similarity code taken into account presently is 7. Including the 0th order we thus have 8 coefficients per atom. Some additional less local information is gathered in 2 by-coefficients per atom, resulting in a total number of 10 coefficients per atom.

1.1. CFDB – the COSMO*frag* Database

After the molecular perception the similarity coefficients are converted into 5-character ASCII codes and combined to a 50-character atom code. For all atoms in each molecule being added to the COSMO*frag* database (CFDB), these strings are stored in the CFDB.txt file together with a link to the molecule itself. Thus, the search for most similar atoms can be done by a simple search for most similar atom codes. Keeping the CFDB ordered alphanumerically the search for the most similar atom can be done by recursive interval splitting, which is very efficient in a large database. The time required for finding the most similar atom only increases with the logarithm of the database size. In this way, for all atoms of a new molecule image atoms of maximum local similarity can be efficiently found in the database. Finally those molecules out of the CFDB are chosen for partial images, which have most similar image atoms for a large number of atoms of the molecule to be fragmented. Based on our present database of presently more than 100000 molecules and drug-like compounds most molecules get fragmented into 2 – 4 fragments.

Based on the 7th order similarity coefficients of all atoms a unique coefficient is generated for each molecule. This is converted into a 10-letter ASCII code being unique for each molecule apart from cis/trans isomerism and stereo-chemical differences. Cis/trans isomerism is translated into a 11th letter and stereochemistry into a 12th letter. Thus, a 12-letter unique name results for each molecule. This name is used for the identification of molecules in the CFDB database and as name for the compressed COSMO files. The simple unique molecular names may be useful even in contexts different from COSMOtherm property prediction.

If favored compounds should not be sufficiently represented in the CFDB, you can run DFT/COSMO calculations for a small number of representative compounds, add them to the database and consecutively improve the fragmentations. The COSMOfrag software will assist you in finding appropriate candidates for database enhancement. COSMOfrag efficiently performs the fragmentation of a new molecule by molecules of the CFDB. The result of the fragmentation is written to a COSMOtherm meta-file (`.mcos`), which then can be used as substitute for the full COSMO file in the COSMOtherm input. Thus, almost any kind of calculation which can be done with COSMO files in COSMOtherm can be done similarly using COSMOfrag meta-files.

Additionally, COSMOfrag can be used to generate new COSMOfrag databases, add molecules to an existing database or compress COSMO files by almost a factor of 20 in order to allow for an efficient storage of the CFDB.

1.2. COSMO*sim* – Extension to Bioisoster Screening

The success of the COSMO-RS theory has proven, that σ-profiles hold the crucial information for most ADME properties as solubility, blood-brain-partition coefficients, intestinal absorption, and even for many adsorption phenomena. Considering the fundamental importance of the σ-profiles for surface interactions of molecules in liquid states, they obviously also carry a large part of information required for the estimation of desolvation and binding processes, which are responsible for the inhibition of enzyme receptors by drug molecules. Thus, a high similarity with respect to the σ-profiles appears to be an important condition for drugs of similar physiological activity. Driven by this insight, we have developed a σ-profile based drug similarity measure, denoted σ-match similatity or SMS, for the detection of new bioisosteric drug candidates. The program extension COSMOsim enables the efficient calculation of this similarity for large libraries, making use of the COSMOfrag technology. In several examples COSMOsim has demonstrated its statistical and pharmaceutical plausibility, its practicability for real drug research projects, and its unique independence from the chemical structure. The latter can be regarded as scaffold hopping in a natural way. COSMOsim is integrated in the COSMOfrag program and it can be activated via a special license.

2. Installation

The COSMOfrag distribution contains the executable program, the COSMOfrag database CFDB and some examples. We recommend setting up such a file directory tree:

/software/COSMOfrag/	binLinux/cf	Linux executable
	binWindows/cf.exe	Windows executable
	binMac/cf	MacOS executable
	CFDB/	COSMO*frag* data base, including subdirectories

Extract the COSMOfrag database CFDB.zip to the folder /software/COSMOfrag. All subdirectories are automatically created.

To install COSMOfrag on a Linux, Mac, or Windows system, copy the appropriate executable to a local directory and set the path.

If using Bourne Again Shell (bash) on a Linux system, add to your .bashrc-file

```
export PATH=$PATH:/software/COSMOfrag/binLinux
```

In a Windows environment, set the path from the system control. From the Windows "Start" menue, select "Control Panel", then "System". Select the "Advanced" tab. With the "Environment Variables" button at the bottom

of the window you get a dialogue where you can edit several variables. Edit the "Path" variable from the lower part and add the path to the directory where the COSMOfrag executable cf.exe is located.

If the path is set correctly, you should be able to call COSMOfrag on the dos shell from any directory by typing cf.exe name.inp (where name.inp is the name of the input file which should be located in the working directory).

If the path for the COSMOfrag executable is not set, you can still run a COSMOfrag calculation by typing in the absolute path for the executable, e.g.

C:\Programme\COSMOfrag\cf.exe name.inp.

3. Running COSMO*frag*

3.1. Program execution

COSMOfrag is started from a Linux shell with the following command:

```
cf <input_file_name.inp>
```

And from a Windows shell with:

```
cf.exe <input_file_name.inp>
```

COSMO*frag* requires an input file with the extension `.inp` (e. g. `filename.inp`). The output of each COSMO*frag* run is stored in `filename.out` and, in a more comprehensive way, in `filename.tab`. Results of COSMO*therm* calculations are written to a tabulated output file `filename.res`.

3.2. Input File

The COSMO*frag* program requires a formatted input file consisting of the following parts:

- Three global command lines.
- Optional: COSMOtherm template input.
- List of structure files or smiles strings.

3.2.1. Global command lines

All commands are given in the form

`command` or `command=argument` or `command={arg1 arg2}`

i.e. if several arguments are given for a command, the arguments have to be included into curved brackets and separated by blank spaces. Commands are not case sensitive. (Of course, paths and filenames on Unix or Linux systems are case sensitive.)

ACTION= This keyword determines the mode in which COSMOfrag is run:

- 1 Operations on listed molecules without database access (very fast):
  - Parsing of the molecule and unique name determination
  - File format conversions possible (WCAR, WXYZ, …)
  - Various checks of the validity of the geometry possible (GEOCHECK)
  - Determination of descriptors possible (WRTDESCRIPT)
  - Splitting of multiple (SMILES /SD) files with USENAM
  - Generation of SMILES strings for all possible EZ-isomers possible (W_EZSMI)
  - Operations on charged compounds and compounds with alkali atom are possible (ALLOW_ALKALI and ACCEPT_CHARGE)
  - instead of structure files a list of unique names can be given
- 6 Search database for molecules:
  - covers ACTION=1 functionality
  - representation of the molecule within the database is checked (identity, minimum/maximum similarity listed in statistics ouput)
  - printout of MAXSTRING possible
  - instead of structure files a list of unique names can be given
- 7 Fragmentation run
  - covers ACTION=1+6 functionality
  - COSMO*therm* calculations possible (ctcalc)
  - COSMO*sim* calculations possible (COSMOsim={})
  - generates metafiles (.mcos)
  - usable keywords: NO_META_IDENT, META_UNIQUENAME
  - for named metafiles for the structure file list the =named option can be added to any keyword
- 8 Add virtually to database; for CFDB enlargement screenings
  - covers ACTION=1+6 functionality
  - molecules are added virtually to the database taking set keywords into account (MINSADD, MIN/MAXNUMHEV). Virtually added molecules are taken into consideration for the database representation of molecules following on the list
  - no cosmo files necessary
- 9 Add really to CFDB
  - covers ACTION=1+6 functionality
  - molecules are added to the database taking set keywords into account (MINSADD, MIN/MAXNUMHEV)
  - cosmo files have to be provided and will be stored as compressed cosmo file (.ccf) in the database given with the CFDBDIR keyword
- 10 Remove molecules from database
  - covers ACTION=1+6 functionality

- database molecules corresponding to the molecules on the list will be removed
- instead of structure files a list of unique names can be given

CFDBDIR=	relative or absolute path to the CFDB
STRDIR=	relative or absolute path to the structure files directory
LICENSEDIR=	path to license file license.ctd. If not specified COSMOfrag searches in "CFDBDIR"
ADDSIM	add to CFDB also similar molecules (stereo isomers)
ACCEPT_CHARGES	allows for charges in the structure, ACTION=1 functionality only
ALLOW_ALKALI	allows alkali atoms in the compounds, ACTION=1 functionality. Restricted to W_EZSMI and USMI only (No Warnings and Errors are written to the tab-fie).
ALLOW_ZW	switches ZWITTERION detection to warning
ALLOW_RARE	allows rare elements in the compounds. The rare elements available are: Sn, Sb, Te, Hg, Bi.
CHECKN=X	checks for C=N and N=N doublebonds which are not in a ring system and prints the configuration into the out-file
CCFCOM	read a comment for the header of ccf-files from line 4 of the input and places it in the ccf-header
CFDB_LIST=nnn	calculation of the first nnn compounds of the CFDB, the complete CFDB is calculated by CFDB_LIST. Keyword has to be in the structure file list.
CFCOSMO[=named]	triggers writing of temp_cf.cosmo. Use CFCOSMO=NAMED for getting <name>_cf.cosmo
CT_EXT	run an external COSMO*therm* program
CTCALC[=named]	do a COSMOtherm calculation for each compound in the structure file list. The fourth line of the input must be $start CT_input, and the template input for the COSMOtherm calculation is closed by a line $end CT_input. The input, output, and and table files of the COSMOtherm calculation are named temp.inp, temp.out, and temp.tab, and the metafile is named temp.mcos. If the CTCALC=named option is used, for each structure file, e.g. molecule.sdf, the COSMOtherm calculation files will be named with the name of the structure file molecule.inp, molecule.out, and molecule.tab, and the metafile will be named molecule.mcos. Note that this option will create four files per file in your structure file list.

CTTABCOMP=n	triggers that the nth compound line will be read from the temp.tab file.
CTtotCH	output of the total COSMO charge. This is useful for the evaluation of the fragmentation quality
FULLTAB	Writes some useful additional information to the tab-file
GEOCHECK	information on special geometrical characteristics is written to the out file (possible internal H-bonds, lone-pair conflicts, planar NH2, trans_O=COX_group…)
GET_CCFCHARGE	reads molecular charge from .ccf file, default when using CFDB_LIST
LARGERING=nnn	sets the limit for the larger ring detection. Cyclic compounds with a ring size below the threshold are treated as a normal compound in COSMO*frag*. If the threshold is exceeded in case ACTION=6+7 calculations the compound is omitted due to the large-ring error (error code 41). In case of ACTION=1 in combination with WRTDESCRIPT a string "LARGERING" is written to the tab file. The default is LARGERING=13. Set nnn=999 to deactivate.
MAXNUMHEV=	specifies the maximum number of non-hydrogen atoms in a molecule to be accepted for CFDB addition
MAXSTRING	output a string with maximum similarities of each atom in order of the atom numbering
META_UNIQUENAME	indicates that the 12-letter unique name shall be used for meta file labelling (automatically set for SMILES structures)
MINNUMHEV=	specifies the minimum number of non-hydrogen atoms in a molecule to be accepted for CFDB addition (default is 1)
MINSADD=	specifies the threshold value of the minimum similarity in a molecule for CFDB addition (default is 2). Values can range from 1 to 7.
NEZCHIR	writes to the tab file the numbers of double bonds with possible cis-trans isomerism and numbers of stereo atoms. In combination with WRTDESCRIPT
NMETA=	specifies the maximum number of meta files which are generated per compound (instead of 1). Values can range from 1 to 9. Only fragmentations of equivalent quality to the best possible fragmentation are written into meta files. Thus, it is possible that less meta files per compound are written than indicated by the keyword.
NO_META_IDENT	indicates not to write identity meta files (where possible), but to use an alternative fragmentation (leave-one-out principle)

POLYMER	marks a polymer repeat unit by Iodine cap atoms. If Iodine is used otherwise, you may specify another halogen by "=F", "=Cl" or "=Br".
STREZCHIR	writes a string to the out-file, which specifies for 3D-compounds R/S and E/Z for stero centres and double bonds, respectively. “F” indicates a failed stero analysis.
TRY_CHARGE	tries total charges of +/-1 if no neutral structure can be assigned.
USESMINAM	use the name after the SMILES code as molecular name
USENAM=<identifier>	tag in multiple SD files to use for naming when splitting up single structure files, metafiles, …
USEOLDNAM	indicates that the original COSMO files shall be referenced in meta files, not the .ccf files from CFDB
USMILES	generate a unique SMILES string for each file on the list
USMILES_CHIR	generate a unique SMILES string with chirality information for each file on the list
USMILES_EZ	generate a unique SMILES string with EZ information for each file on the list
W_EZSMI	trigges writing of SMILES for all possible EZ-isomers to the out-file
WRTDESCRIPT	write descriptors to the tab-file: no. of atoms, hydrogens, bonds, ringbonds, conjugated atoms, internal hydrogen bonds, rotatable bonds, no. of terminal alkyl carbons, no. of terminal alkyl groups, molecular weight, no. of atoms in of the largest ring (LargeRing)
WCAR	generate .car for each (3D) file on the list
WXYZ	generate .xyz for each (3D) file on the list
WSDF	generate .sdf for each (3D) file on the list
WCCF	generate .ccf for each .cosmo file on the list
WMCOSCAR	generate a "_mcos.car" file for each successfully fragmented file on the list that visualizes the fragmentation. 3D input structures will be displayed central.

ACTION and CFDBDIR are indispensable for COSMOfrag runs. If STRDIR is omitted, the absolute path for the compounds can/should be specified.

3.2.2. Structure file list

COSMOfrag supports the following structure file formats:

1. SMILES:
    a. Direct input via list. Notation: “smi:CCC=OCC”
    b. SMILES in .smi file(s). Multiple smiles strings in one file supported. Names after Smiles supported.
2. SDF (MDL Isis):
    a. Single or multiple .sdf files
3. CAR (MSI Biosym)
4. XYZ
5. COSMO (Turbomole), CCF (compressed COSMO), COS (Mopac COSMO)
6. ML2 (Sybyl)
7. MOPAC .dat, .mop, .arc (or .arcxx) files
8. MOPAC .out files, also of multiple jobs

CFDB_LIST=nnn calculation of the first nnn compounds of the CFDB, the complete CFDB is calculated by CFDB_LIST. Keyword has to be in the structure file list.

Lists containing files of different formats are permitted. Empty rows terminate processing of the list.

3.3. COSMO*therm* calculations

COSMO*frag* is a shortcut for high-throughput COSMO-RS calculations. COSMO*frag* calculates the σ-profiles by a composition of partial σ-profiles taken from locally most similar fragments from the CFDB. The time consuming quantum chemical calculation is avoided.

Generally, the whole functionality of COSMO*therm* is available and only restricted by the limitations of COSMO*frag* (e.g. restricted to neutral molecules). However, we recommend predicting only those properties which can be well be described on the drug like level.

3.3.1. COSMO*therm* input section

The COSMO*therm* calculation is enabled by the keyword CTCALC in the COSMO*frag* global command lines. The COSMO*therm* input template is similar to the COSMO*therm* input file. Inside of the COSMO*frag* input, it is enclosed by the two lines

$start CT_input as the fourth line and
$end CT_input as the line closing the COSMOtherm template input section.

The following parameters have to be specified in the input template:

ctd=[named.cdt] Use the file name.ctd as COSMOtherm parameter file

cdir= Sets the directory where to search for the COSMOtherm parameter file

fdir= Sets the directory where to search for the .cosmo, .cos or .ccf files of the COSMO calculations.

In the COSMO*therm* compound list, the compound indicated by f=CFcomp is taken from the structure file list following the COSMO*therm* input template.

An example input file for the calculation of the 1-octanol-water partition coefficient is shown in tabular 1. Detailed instruction about COSMO*therm* calculations can be found in the COSMO*therm* documentation.

Tabular 1 COSMOfrag input file for the calculation of the 1-octanol-water partition coefficient.

Line	Keyword	Comment
1	`Action=7, CFDBDIR= STRDIR=`	These COSMO*frag* Keywords **must be** in the first three lines. STRDIR can be omitted if compounds are specified with their absolute file paths
2	`LICENSEDIR=`	
3	`CTCALC CTtotCH NMETA=3`	
4	`$start ct_input`	Start of COSMO*therm* Inputs. This command has to be in line 4.
5	`cdir=, ctd=, fdir=`	This keywords must be in lines 5 and 6
6		
7	`!! logP !!`	Comment or empty line
8	`f = h2o.cosmo`	The following lines should be like these. Here, the partition between water and (wet) 1-octanol is calculated at a temperature of 25 °C
9	`f = 1-octanol.cosmo`	
10	`f = CFcomp`	
11	`tc=25 logp={1 2} xl1={1 0} xl2={0.24 0.76} vq=0.11415 wcomp={3} logp_amine_corr={-0.6 -1.1 -1.5}`	
12	`$end ct_input`	End of COSMOtherm input template
13	`smi:CC(=O)c1ccc(F)cc1`	Start the compound list
OR 13	`list`	"list" referes to a text file with the name of the compounds.

The hash character '#' is used to identify comments in the input file. If the hash character is the first character of a line, the complete line will be ignored. In other positions, any text after the hash character will be ignored. This applies only to the COSMO*therm* template.

3.3.2. Property prediction by COSMO*therm*

The COSMO*frag* program can be used for COSMO*therm* batch processing. Basically there are two application scenarios:

- A COSMO*frag* run in fragmentation mode (ACTION = 7) and subsequent COSMO*therm* calculation on the produced metafiles. A path to the input structures must be given with the keyword STRDIR.
- COSMO*therm* batch processing on existing .cosmo/.ccf/.mcos files without COSMO*frag* database operations (ACTION = 1). A path to the .cosmo/.ccf/.mcos files must be given with the keyword STRDIR.

For the following list of properties a fully automated batch calculation can be performed. Here the calculated properties are written in tabulated form to a file with the extension .res, ready for post-processing in a spreadsheet program.

1. Water solubilities (or other solvents)
2. Partition coefficients (Octanol-Water or any two phases)
3. Activity coefficients
4. QSPR properties
    a. Intestinal Absorption Coefficient (logKIA)
    b. Blood-Brain Partitioning Coefficient (logBB)
    c. Organic Carbon (Soil)-Water Partition Coefficient (logKOC)
    d. logKHSA (Human Serum Albumin)
    e. Octanol-Water Partition Coefficient (logPOW)

For solubilities or activity coefficients the SCREEN flag should be set in the COSMOtherm template input section (not documented in the COSMOtherm manual), which speeds up the calculation of these properties.

The 1-octanol water partition coefficient (logKOW) is of special interest for the drug design process. Based on the statistical analysis of several validation studies the prediction of the logKOW value is improved by the command:

logp_amine_corr={-0.6 -1.1 -1.5}

The NMETA=n (n= 1 – 9) option can be used to generate several fragmentations of a compound. Each of the fragmentations is stored into a separate metafile, and the property calculation is done for each metafile of the compound. Note that if another fragmentation of equivalent quality (in terms of atom-wise similarity as indicated by the MAXSTRING) is not available, the number of fragmentations written to metafiles can be smaller than the number indicated by the NMETA keyword. In the results file filename.res, the results of the property calculation are tabulated for each fragmentation, together with the average value and the standard deviation of the calculated values.

3.4. Multiple jobs

Multiple COSMO*frag* jobs may be combined within a single input file. The input for each successive job is separated from that of the preceding job step by a line of the form:

```
$$$$
```

Note that an empty row terminates processing of the input.

3.5. COSMO*frag* output

The output of each COSMOfrag run is stored in `filename.out` and `filename.tab` (where `filename.inp` is the name of the input file). For named metafiles for the structure file list the `=named` option can be added to any keyword (ACTION ≥ 7). Results of COSMOtherm calcultations for all compounds in the structure file list are written to a tabulated output file `filename.res`.

3.5.1. Metafile quality

Even though in general an evaluation of the quality of a specific fragmentation is not possible, the total COSMO charge can provide an indication of low quality metafiles. By setting the `CTtotCH` keyword the total COSMO charge will be tabulated in the output file `filename.res` file. Values > 0.4 or < - 0.4 indicate possibly bad fragmentations and should be separated out, although in some cases even those metafiles can produce good property predictions due to error compensation.

3.5.2. Errors and warnings

Due to its functional range COSMO*frag* offers a rather long and detailed list of error codes. These errors predominantly occur within the structure parsing part of the program and are listed in `filename.tab` separately for each molecule on the list. Some of the error codes are of interest only for the development of the program. In the following list, we have grouped the error codes corresponding to their causes. They should help the user to filter wrong structures or structures not supported by COSMO*frag*.

ERRORS

- Filesystem operations:
  - 1 file open error
  - 2 file read error
  - 3 file format not known
  - 4 name assignment failed
  - 5 file open error occured
  - 6 filename or SMILES too long (max=299 letters)

- Errors in SMILES code:
  - 16 SMILES contains more than one molecule
  - 20 Error reading SMILES string
  - 31 unresolved atom in SMILES
- Structures possibly valid but presently not supported by COSMOfrag:
  - 10 number of rings exceeds maximum
  - 12 more than 50 intramolecular donor-acceptor pair found
  - 17 element unknown or not allowed
  - 18 residual charges in molecule
  - 19 molecule charged
  - 21 molecule too large
  - 22 parts of the molecule not connected
  - 30 isotope specified
  - 32 biradical specified
  - 39 no kekule structure found after ion loop
  - 40 compound most likely will be zwitterionic in water. The keyword "ALLOW_ZW" triggers on the calculation of zwitterions.
  - 41 molecule has a large ring which cannot be well handled by COSOM*frag* The keyword "LARGERING=999" activates the calculation of larger rings.
- CFDB offers no acceptable fragment for atoms/parts of the molecule:
  - 38 no image for one atom
- Invalid structures:
  - 11 aromaticity of some rings could not be defined
  - 13 invalid hybridization
  - 23 no atoms found
  - 24 only hydrogens in molecule
  - 25 single aromatic bond
  - 26 error in hydrogen addition
  - 27 invalid valence
  - 28 bond order trace back failed
  - 29 traceback failed in ring analysis
  - 33 wrong stereo definition (or stereo centre with more than 4 neighbors)
  - 36 strange hybridization
- Erros in metafile generation:
  - 14 too many loops in generation of metafile
  - 15 generation of metafile failed
  - 35 atom not assigned in makemeta
- Errors due to set filters:
  - 34 number of heavy atoms larger than MAXNUMHEV
  - 37 number of heavy atoms smaller than MINNUMHEV
- Output errors:
  - 64 Error in writing ccf-file

WARNINGS:

- 1 The structure has no z-coordinates. Make sure that it is a valid 3D structure.
- 2 The structure has unrealistically short bonds. This geometry not considered as valid 3D structure
- 10 A trans conformation was found for a carboxylic acid or an ester-like fragment. This is unlikely
- 12 Isotope information is ignored.
- 13 Due to the lack of cis/trans information trans was assumed for a double bond.
- 20 Extreme sigma value beyond +/- 3e/nm2 found.
- 30 Compound from PRF-line processed without structural details.
- 40 Compound most likely will be zwitterionic in water.
- 51 Conflict in SMILES EZ assignment.
- 52 No single bonds in ring. Write_ SMILES failed.
- 53 No valid ring assignment found within 20 attempts.
- 54 More than 99 rings. Write_SMILES failed.
- 60 Erroneously identical UNIQUE NAME found.
- 87-93 Molecules has charge iwrn-90. iwrn=90 means zwitterion.

4. The COSMO*frag* database

The COSMOfrag database (CFDB) of the current release consists of 112371 common compounds and drug-like structures, collected from databases like Physprop*, NCI**, NCBI*** , PUBCHEM*** and various others.

The quality of the COSMO-RS property predictions resulting from COSMOfrag metafiles depends on the quality and size of the CFDB database. The database molecules have been calculated with TURBOMOLE on the BP-SVP-AM1 level, i.e. by a single-point DFT calculation with SVP basis set, on a molecular geometry optimized with AM1/COSMO with a modified MOPAC 7 program. In the majority of the cases the initial 3D structures are generated by the CORINA program by Molecular Networks. It is recommended to use compounds calculated by a comparable procedure for addition to the CFDB.

* Physical Properties Database by Syracuse Research Corporation

** National Cancer Institute Database

*** National Center for Biotechnology Information

5. Applications

5.1. Visualization of metafiles

To visualize a fragmentation corresponding to a certain metafile, setting the keyword `WMCOSCAR` will generate a structure file in CAR format, which contains each CFDB molecule occurring in the metafile. For a 3D input structure the input molecule is placed in the centre of the visualization. Opening this `xx_mcos.car` file in a molecular viewer, the fragmentation is comfortably visualized and by displaying the name-field (first column) the weights of the metafile can be displayed.

5.2. Expansion of the COSMO*frag* Database (CFDB)

Although the CFDB now consists of more than 100.000 drug-like compounds, it is possible that a few molecules out of a larger set or groups of certain functionality are not represented reasonably. To identify those compounds, the `MAXSTRING` keyword can be used. A string in which the maximum similarity of the atoms is given in input order is then written to the output and table files. From our point of view, a similarity value ≥ 1 can always be regarded as adequate. '0' similarities on the other hand should be replaced in either case. COSMO*frag* therefore denotes these molecules with error code 38.

Concerning '1' similarities no definite advice can be given. For fast screenings, predictions with a moderate loss of accuracy may be acceptable and the number of compounds with '1' similarity can be quite high. For high quality predictions with a small number of '1' similarities one could think of replacing them by molecules covering the missing fragments.

A pre-screening for compounds with low similarity can be done by the following procedure:

- Run COSMO*frag* with ACTION=8 on your set of molecules. In this run molecules which cannot be fragmented adequately with molecules from the CFDB are added virtually. That means the hash coefficients for the molecule are generated and added to the database in the memory at this time, but no COSMO file is needed. Consecutively, the molecule just added will be taken into account for the representation of the other molecules on the list. So if one molecule of a structurally similar group is added virtually, the rest of the group can be represented by a fragment of this very molecule.
- To perform an operation like this another flag has to be set in the input file: If MINSADD=1 is set in a virtual run, only those molecules are virtually added to the CFDB whose smallest similarity of an atom is ≤1. Thus, the first molecule of a group of molecules with a certain structural feature which is not yet represented in the CFDB is added. For the following molecules of the group the minimum similarity is > 1 for this special atom or group, due to the molecule added before. Then the MINSADD flag prevents those molecules from being virtually added as well.

In summary, a run on a complete set of molecules with ACTION=8 and MINSADD=1 should be done (virtual runs are very fast). Afterwards all compounds virtually added to the CFDB have to be selected (there is a flag in the table file for 'added') and calculated quantum chemically.

One thing can be done additionally to reduce computer time: In case there are compounds in the set which are rather large, i.e. more than 40 - 50 non-hydrogen atoms, the list of molecules can be ordered from smaller to larger ones before the virtual run. COSMOfrag with ACTION=1 and WRTDESCRIPT can be used for that task; the number of atoms is available then from the table file. The effect would be that smaller molecules are processed first and a higher probability for smaller molecules being selected for COSMO calculations results.

When the COSMO calculations are done, the COSMO files should be added to the CFDB (ACTION=9).

6. COSMO*sim*

The calculation of compound similarities based on the similarity of COSMO σ-profiles is triggered in the COSMOfrag input file by the keyword

COSMOsim or COSMOsim={ntarget,nbest}

Here, ntarget defines the number of compounds to be used as target. By default, only the first compound in the structure file list will be considered as target. However, it may be useful to search for the maximum average similarity of compounds to a set of ntarget target compounds, which must be listed at the head of the structure file list.

For each compound in the list, the average similarity as well as the individual similarities will be written to an output list (jobname.prf). There is no limitation on the number of compounds to be processed in one job. nbest is the number of most similar compounds which will be written to a sorted list (jobname.prf2). The maximum possible value for nbest is 10000.

By default the similarity is calculated as the "sigma-match similarity index" (SMS) with the default parameters as described in the COSMOsim publication. With the keyword

SMS={a,b,c,d}

The COSMOsim parameters can be altered, but we do not recommend doing so. By supplying a value of a < -1.d0, the usage of the Tanimoto' index as suggested by Thormann can be triggered.

The keywords

wprf[=named] causes the program to write the σ-profiles of all compounds in the structure file list to jobname.prf without calculating the SMS. If wprf=named is used, the metafiles generated

during fragmentation are named with the compound name instead of temporary metafiles, i.e. this option produces one .mcos file per compound in the structure file list.

NOPRF No jobname.prf file is written, only the sorted list jobname.prf2 is written. Useful if the σ-profile database already exists.

All formats supported by COSMOfrag can be used in combination with COSMOsim. Usually COSMOfrag will generate a fragmentation of the compound (i.e. a mcos-file) and will generate a σ-profile from this fragmentation. If the file format is a COSMO file or a compressed COSMO file (ccf) then the σ-profile is generated directly from this file without fragmentation.

In a COSMOsim run, files with the extension .prf or .prf2 can also be within the compound list. In that case the σ-profile required for the calculation of the SMS coefficient is not newly generated, but read from the .prf or .prf2 files, which is much faster. Thus, it is possible to re-use previously generated σ-profiles from one or more COSMOsim jobs in later similarity searches. This will be extremely fast because for the similarity search only the similarity coefficient has to be calculated.

A database of precalculated σ-profiles of more than 8.8 million public available compounds is available from COSMO*logic*. If you have access to this database we recommend installing it to the folder `/software/COSMOfrag/CS_screeningDB/` and using the NOPRF keyword.

7. References

COSMO*frag*: A Novel Tool for High Throughput ADME Property Prediction and Similarity Screening Based on Quantum Chemistry
Martin Hornig, Andreas Klamt
J. Chem. Inf. Model. 45, 1169-1177 (2005).

Prediction of aqueous solubility of drugs and pesticides with COSMO-RS
Andreas Klamt, Frank Eckert, Martin Hornig, Michael E. Beck and Thorsten Bürger
J. Comp. Chem. 23, 275-281 (2002).

COSMO-RS: a novel view to physiological solvation and partition questions
Andreas Klamt, Frank Eckert and Martin Hornig
Journal of Computer-Aided Molecular Design 15, 355-365 (2001).

Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach
Frank Eckert and Andreas Klamt
AICHE Journal 48, 369 -385 (2002).

Prediction of Blood/Brain Partitioning and Human Serum Albumin Binding Based on COSMO-RS σ-Moments
Karin Wichmann, Michael Diedenhofen and Andreas Klamt
J. Chem. Inf. Model. 47, 228 -233 (2007)

Use of Surface Charges from DFT Calculation to Predict Intestinal Absorption
Ron Jones, Paul C. Connolly, Andreas Klamt and Michael Diedenhofen
J. Chem. Inf. Model. 45, 1337-1342 (2005).

COSMO*sim*: Bioisosteric Similarity Based on COSMO-RS Profiles
Michael Thormann, Andreas Klamt, Martin Hornig, and Michael Almstetter
J. Chem. Inf. Model 46, 1040-1053 (2006).

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

8. Troubleshooting and Support

The present COSMO*frag* Version is subject to permanent methodical development and software optimization. We explicitly request for sending us bug reports and feature suggestions.

Contact: Prof. Dr. Andreas Klamt
COSMOlogic GmbH & Co. KG
Imbacher Weg 46
51379 Leverkusen
Germany

klamt@cosmologic.de
and / or
info@cosmologic.de

+49(0)2171 731681