NAME

csfc - batch chemistry file format converter

VERSION

1.52, 2017-2-21

SYNOPSYS

csfc [-align *none*/*x*/*y*] [-assign propertypairlist] [-cleardirectory] [-chargebalance 0/1] [-chargecombine 0/1] [-coretemplates file_list] [-computestereo 0/1] [-count maxrecs] [-desalt 0/1] [-dimension *0/2/2.5/3*] [-directory dirname] [-eoltype *mac*/*pc*/*unix*] [-expand 0/1] [-feedback n] [-fixbridgeheadstereo 0/1] [-format fmt] [-help] [-hydrogens *stripall*/*stripstereo*/*strip*/ *asis*/*addall*/*addhetero*] [-ignoreempty 0/1] [-imgparameters parameterlist] [-implicith 0/1] [-info] [-inputformat *fmt*] [-join 0/1] [-lock propertylist] [-mapping propertypairlist] [-miniheader 0/1] [-nitrostyle *ionic*/*asis*/*penta/xionic/xpenta*] [-noradicals 0/1] [-offset n] [-outfile filename] [-parent 0/1] [-pedantic 0/1] [-pseudo3d 0/1] [-properties propertylist/*all*] [-purgeisotope 0/1] [-purgestereo 0/1] [-recalc 0/1] [-resolvearo 0/1] [-saltfile filename] [-scope *mol*/*ens*/*reaction*] [-separate 0/1] [-showcrossedbonds 0/1] [-skiperrors] [-split 0/1] [-suffixfmt 0/1] [-suppress propertylist] [-tablefile filename] [-tablekey columnname/index] [-tablekeytype *single/multi/strict*] [-tautomercoordinates *discard/preserve*] [-tautomerrules *full restricted*] -tautomers *none/all/unique/canonic/2/3/4...*] [-template SMILES/SMARTS] [-templatealign *none/x/y/diagonal/rotate/redraw*/*besteffort*] [-templatefile file] [-templatematch *strict/relaxed*] [-timeout2d *secs*] [-timeout3d *secs*] [-structurekey property] [-uncharge 0/1] [-unique *none/default/ stereo/isotope/tautomer/parent*] [-usearo 0/1] [-usename 0/1/2/3/4/5] [-usestereo 0/1] [-version] [-wedgepairs *no/all/hydrogen/ unconnected*] [-wedgestyle *default/opposite*] [-writemode *a*/*w*] ?files?

DESCRIPTION

csfc is an universal batch-mode chemistry data exchange file converter. It reads the specified input files, perform selected operations on the input data, and output it in a variety of chemistry exchange formats. If no input file is specified, or the file name „-“ is used, the program reads from standard input.

The file type of the input files is automatically recognized. It may change from file to file. In the full toolkit version, the tool loads all locally available structure file I/O extensions before commencing its operation to cope with the maximum number of file formats. Standalone variants contain an exhaustive set of compiled-in I/O modules. Input files can be processed in compressed, gzip-ed or bzip2-ed form without prior unpacking. In addition, ZIP-compressed or TAR-packed (the later only on Unix/Linux versions of the software) files may also be opened directly for reading. All files in such archives are read as one big virtual multi-record file.

Each input file name argument may independently be a local file, an URL (http, ftp, gopher, file) or an e-mail message file containing the structure data in the mail body or as one or more attachments. URL retrieval and compression processing can be combined. If a file name argument is a directory, all readable files in that directory are processed as a single virtual input file. If a file argument is not a readable file, but the expansion of the name as a shell match pattern results in a set of readable files, these are processed instead.

If no output file is specified, the output has the same name (but with an updated suffix) and resides in the same directory as the input file. If the output file would have exactly the same name as the source file, a plus character is appended. If the *-directory* parameter is used, the output file is put into that directory and inherit only the non-directory part of the

input file name. If the output file is specified explicitly with the *-outfile* parameter, that file name, including the chosen suffix, is used. However, even in this case the *-directory* parameter overrides the directory part of the specified file name. If the program operates in split mode (either by setting the *-split* option, or choosing an output format which is limited to single-record files in combination with a multi-record input file), a record count string in the form *_n* is inserted before the suffix to indicate the original record number. Such record numbers begin with 1, but are adjusted by file offsets (*-offset* option). If the program operates in join mode (option *-join*), all data is stored in the file explicitly set by the *-outfile* parameter or constructed from the first input file name.

The special filename *stdout* can be used to direct output to the standard output channel. However, this is only possible if a single output file is obtained - otherwise, the file name is used literally to construct the output file name according to the described rules.

The software can process both structure files and reaction files (in RXN, RDF, CTX, JME, CDX, CDXML, or CACTVS formats).

The supported file formats are listed in the table below.

In case of custom development, additional file formats may be locally available. File formats handler are typically implemented as dynamically loaded modules.

All csfc versions include a 2D layout coordinate generator, which computes layout coordinates from input without such information. For output to 3D files from input without 3D coordinates, the atomic coordinates can either be computed locally (if you are running a special csfc version with a built-in Molecular Networks CORINA module), or an attempt is made to find a 3D conformation of the structure in PubChem. The latter implies structure export over the Internet and should not be used with confidential data.

Format	Read	Write	Comment
441	Yes	Yes	
Alchemy	Yes	Yes	
ASN.1 binary	Yes	Yes	NCBI PubChem native ASN.1 binary format
ASN.1 text	Yes	Yes	NCBI PubChem native ASN.1 ASCII format. Support for this format requires that the NCBI *datatool* application is found in the program search path.
BDB	*Yes*	*Yes*	This database-like format, designed for effective updates and structure queries, is not part of the standard version, because it requires a license to a 3rd party storage manager.
Cactvs/Ascii	Yes	Yes	Structures, reactions and datasets
Cactvs/Scan (CBS)	Yes	Yes	Structures, reactions and datasets
Cactvs/Binary (CBIN)	Yes	Yes	Structures, reactions and datasets
CAR	Yes	Yes	

Format	Read	Write	Comment
CAS	Yes	Yes	This is a CAS number look-up via the PubChem database. Internet access is required, and the validity of the results cannot be guaranteed because of the nature of the PubChem database.
CDX (ChemDraw)	Yes	Yes	Structures, reactions and simple queries (no 3D, no Markush)
CDXML (Chem-Draw)	Yes	Yes	Structures, reactions and simple queries (no 3D, no Markush)
Cerius II	Yes	Yes	
CHAI	Yes	Yes	
Charmm	Yes	Yes	
ChEMBL	Yes	Yes	ChEMBL XML format
Chiron	Yes	Yes	
CID	Yes	Yes	PubChem CID numbers. Decoding or encoding these requires Internet access.
CIF	Yes	Yes	Recently improved release, can now read crystal-lographic files with fractional coordinates and convert these.
CMF	Yes	Yes	Compressed MDL Molfile, used for example in MDL Internet display plug-ins.
CML	Yes	Yes	CML 2.0, support not complete. Does read NIH DTP CML structure files with special bond types.
Compass	Yes	Yes	
Cosmo	Yes	No	Understands COSMO output from Gaussian and Turbomole, but not DMOL.
CTX	Yes	Yes	Structures and reactions
EMF/WMF	Yes	Yes	Retrieves only structures from embedded hidden structure data on input - no chemical OCR! Indi-rectly configurable via E_EMF_IMAGE prop-erty attributes. Output format may actually be EMF, WMF or placeable WMF.
FPS	Yes	Yes	Dalke fingerprint exchange format
FIG	No	Yes	
Gaussian Archives	Yes	No	
Gaussian Cube	Yes	No	

Format	Read	Write	Comment
Gaussian Input	Yes	Yes	
Ghemical	Yes	Yes	
GIF, PNG	Yes	Yes	Retrieves only structures from images with hidden embedded structure data on input - no chemical OCR! Indirectly configurable via E_GIF property attributes. Output format may actually be GIF, PNG or BMP in various color depths. Supports both structures and reactions.
GROMACS	Yes	Yes	
Hitlist	Yes	Yes	
Hyperchem HIN	Yes	Yes	
IFF	Yes	Yes	The format of the Vega modelling package. The RIFF format variant is also supported.
InChI	Yes	Yes	The IUPAC chemical structure identifier. Output is formatted with localized hydrogens. See also STDInChI format.
Index	Yes	Yes	This is a special indirect file type which can be used to select subsets of larger files in arbitrary formats
JCAMP	Yes	Yes	Supports JCAMP-DX and JCAMP-CS
JME	Yes	Yes	The format of the JME Java editor applet. Supports structures, query specifications and reactions
KCF	Yes	Yes	KEGG database structure format
KNIME	Yes	Yes	KNIME native binary table format
M3D	Yes	Yes	Molecules3D company format.
Maestro	Yes	Yes	The format of MacroModel.
MMD	Yes	Yes	Older MacroModel format
Molconn-Z	Yes	Yes	The native structure format of a popular structure descriptor computation package.
MDL Molfile	Yes	Yes	This format covers both plain Molfile and SD-file. Most ISIS query attributes, including 3D search and R-groups, are also supported.
Molgen	Yes	Yes	
Mopac Input	Yes	Yes	
Mopac Output	Yes	No	

Format	Read	Write	Comment
MRV	Yes	Yes	Marvin document format. Limited support for R-groups and Markush structures.
NETCDF	Yes	No	Does not support all conventions for spectral data exchange
PDB	Yes	Yes	Supports multi-record files. Does attempt to generate connectivity including guessed bond orders if no CONECT records are found or these are identified as incomplete.
PDF	Yes	Yes	Retrieves only structures from images with hidden embedded structure data on input - no chemical OCR! Indirectly configurable via E_PDF_IMAGE property attributes.
PDBcode	Yes	No	A file with standard PDB codes. The program attempts to download PDB data from the official PDB repository over the Internet.
PICT/PCT	Yes	Yes	Macintosh PICT/PCT image format. The program reads PICT images with embedded structure information, for example from ChemDraw or ISISDraw.
PostScript	No	Yes	Indirectly configurable by attributes of properties E_EPS_IMAGE and D_PS_PAGES
RDF	Yes	Yes	Both structures and reactions are supported.
RIFF	Yes	Yes	See IFF
RTF	Yes	Yes	Produces RTF tables with embeded OLE structure objects for ChemDraw or ISISDraw. On input, decodes OLE object content of these types, plus WMF/EMF with embedded structure data, but not other types of graphics.
RXN	Yes	Yes	MDL reaction data
SCF	Yes	Yes	
SDF	Yes	Yes	See MDL Molfile
SDF3000	Yes	Yes	MDL V3000 format variant. Supports structures, data, collections and simple query constructs, but no R- and S-groups or 3D queries.
SDDATA	Yes	Yes	SD-style data file without structure block
Sharc	Yes	Yes	QM computed NMR spectra DB format.
Shel-X	Yes	No	X-ray structures.

Format	Read	Write	Comment
SKC	Yes	Yes	From ISIS/Draw. Supports molecules, query structures (but no 3D queries yet) and reactions.
SK2	Yes	No	From ACD/Labs ChemSketch. Input only. Supports molecules, query structures, reactions, and multi-page documents
SID	Yes	No	PubChem SID numbers. Decoding or encoding these requires Internet access.
SLN	Yes	Yes	Sybyl linear notation. Also supports and translates a number of basic SLN query attributes, but no complete support for these.
SMARTS	Yes	Yes	As far as input is concerned, a joint module is used both for SMILES and SMARTS. The SMARTS output format uses explicit hydrogen enumeration. Example: [CH3][CH3] in SMARTS, vs. CC in SMILES for ethane.
SMIRKS	Yes	Yes	Variant of Reaction SMILES, same relationship as SMILES/SMARTS.
SMD 4	Yes	Yes	
SMD 5	Yes	Yes	Not a complete implementation
SMILES	Yes	Yes	Includes full SMARTS and Recursive SMARTS support, as well as reactions. Explicit SMARTS output is provided via the SMARTS output-only module.
STDInChI	Yes	Yes	The IUPAC InChI identifier, in its standard configuration (see also InChI format)
STF	Yes	Yes	
Stigmata/ Thor Data File	Yes	Yes	Daylight application format
SVG	Yes	Yes	Vector graphics format, retrieves only hidden embedded structure information on input, no chemical OCR
SWF	Yes	Yes	Flash vector graphics, retrieves only hidden embedded structure information on input, no chemical OCR
Sybyl	Yes	Yes	
Sybyl II	Yes	Yes	The .mol2 format

Format	Read	Write	Comment
TAR	Yes	Yes	The program can read files which contain multiple structure record files in other supported formats within a TAR archive, without the need to unpack the archive
TGF	Yes	Yes	Supports molecules and structure queries (with the exception of 3D queries), but not yet reactions.
Table	Yes	Yes	The program tries to automatically detect table layout and look for SMILES or other line notatio columns to decode as structure data.
Tinker	Yes	Yes	The structure file format of the Tinker modelling package
USMILES	Yes	Yes	Unique SMILES according to original publication. Not compatible with current Daylight software, and incomplete when stereochemistry or isotopes are involved.
Vamp	Yes	No	Format of the VAMP semi-empirical QM software.
VRML	Yes	Yes	Retrieves only structure data from embedded information on input, not a scene analysis! Indirectly configurable by E_VRML property attributes.
WMF	Yes	Yes	See EMF
XBSA	Yes	Yes	
XDF	Yes	Yes	MDL Pseudo-XML format
Xtelplot	Yes	Yes	
XYZ	Yes	Yes	Trajectories are supported.
XYZR	Yes	Yes	Since this file format only contains atomic radii and no element information, input is reliable only if the file uses the same VDW radii table as CACTVS, or the radius table has been adapted by a script modification.
ZIP	Yes	No	The program can read files which contain multiple structure record files in other supported formats within a ZIP archive, without the need to unpack the archive.

EXAMPLES

csfc -count 10 -directory . -hydrogens hetero -chargecombine 1 -feedback -format xyz test1.sdf test2.sdf

csfc -format sdf -desalt 1 -directory conf -properties ’E_NAME E_WEIGHT’ -cleardirectory -resolvearo 1 \

-mapping 'E_NAME Name E_WEIGHT Molw' -offset 5 -split 1 -recalc 1 -ignoreempty 0 <test2.ctx

OPTIONS

-align *none/x/y/diagonal*

Change the alignment of the 2D structure layout. By default, structure coordinates are generated in a layout where common ring systems are in their familiar orientations. In case of rectangular image sizes, a rotation of the structure so that the largest coordinate extent is aligned with the x or y axis can sometimes improve the visual appearance. This option can be used both for newly computed 2D plot coordinates or coordinates read from file. Diagonal alignment is along a 30 degrees angle. Structures can also be aligned to a substructure template. This procedure is accessible through the **-template** set of options.

-assign propertypairlist

Assign one or more properties read from the input files to other properties, which are, for example, used in the output file. Only properties of the same object class can be assigned. The software attempts to convert the data type, if the property data types of source and destination property are not the same. Property names can be given either in CACTVS syntax or with the name they appeared in the original file. Case is important. An application example is the assignment of an SD file property to the CACTVS core property E_NAME, which is output in prominent locations in a number of important file format.

-chargebalance 0/1

If set, the program attempts to produce a neutral, uncharged form of the structure. This option does not have an effect on data files which either contain explicit atom charges, or a global ensemble charge.

-chargecombine 0/1

If set, the program attempts to merge adjacent charges (such as in a zwitterion) without violating valence restrictions.

-cleardirectory

If a destination directory has been specified with the *-directory* option, and this parameter is set, the directory is completely cleared before the conversion begins. Subdirectories are deleted with all their content. This option should be used with great care.

-computestereo 0/1

If set, the program tries to compute stereo descriptors from the input data. Secondary, indirect sources of stereo information can be 3D coordinates, various stereo descriptors already present in the input (such as parity), wedge attributes on bonds, and 2D coordinates for stereo bonds. If the computation fails, for example because no suitable secondary data source was present, no error is reported.

-coretemplates file_list

A list of files with structure fragments to augment the built-in set of level 2 2D ring system templates. These are not the same as the templates used for aligning sequences of compounds in a common fashion (**-templatealign** option) which are used at a higher level of processing. The core templates are used directly in the low-level layout of complex ring systems. Multiple files can be listed with this parameter, and files can be multi-record. All recognized file formats which contain basic structure data and 2D coordinates are acceptable. A maximum of 100 user-defined core templates in all files is currently supported. Additional files or records are ignored.

The core templates are simple structure fragments with specified 2D coordinates. The coordinates are automatically scaled and do not need to adhere to specific value ranges and scaling. Elements are ignored in matching the templates, so typically only an all-carbon structure framework is supplied. Single bonds in that pattern match any bond in the processed structures, including multiple and aromatic bonds. Other bond orders need to match exactly. This is useful to ensure, for example, that a specific double bond in a macrocycle is always placed in the same position. Level 2 templates must consist of a single fragment and must contain only ring atoms. They can only match complete ring systems of the structures being processed. This is more restrictive than for high-level alignment templates. Level 2 templates override the more elementary built-in level 1 templates but have lower precedence than user-specified alignment templates. In case the processed structures contain multiple ring systems, more than one template may be applied to different sections of the molecule, and even if a high-level alignment template matches, other parts of the processed structures may still be drawn using these templates.

-count n

Convert a maximum of *n* records from the files. This count applies to each individual input file. The *-offset* option can be used to position the file before the count begins and thus convert only a region of a large file.

-desalt 0/1

If set, counterions are removed from the input. The default algorithm splits each input record into molecules and discards all but the largest molecule, as determined by the atom count, including hydrogens. Specific larger counterions can be added for selective removal with the **-saltfile** option. Molecules found in that file have precedence for deletion from the input record, even if they are larger than their counterparts.

-dimension *0/2/2.5/3*

Select the output format dimension. The use of this option is only necessary for file formats which can either store 3D or 2D data (e.g. MDL Molfiles). If a file format can only use one type of coordinate information, the correct one is automatically chosen. The program contains a sophisticated 2D structure layout generator, so 2D plot coordinates can be generated from input data without or with only 3D coordinates. The writing of 3D coordinate data is only possible, if either the input file already contained this information, or an auxiliary CORINA module has been licensed. Some file formats (such as the MDL family) also support 0D output, which is pure connectivity without any 2D or 3D coordinates. These are output as zero coordinates if they cannot be completely omitted from the output. The dimension 2.5 is a special compatibility option for writing 3D formats without access to 3D coordinates (from a CORINA module or the input file). In that dimension, if no 3D coordinates are present, 2D coordinates (with all the 2D coordinate generation options) are computed or used instead. These are copied to the 3D coordinates, with all zero for the atomic Z coordinates. By this means, exclusive 3D format files can be written with a flat 2D structure representation, although care must be taken to avoid subsequent interpretation of the coordinates as real 3D data.

This option replaces the *-3D* option in older versions of this program. It is still supported, but considered deprecated.

-directory dirname

This option sets the target directory. If this option is not used, or an empty string is

passed, the directory of the output files is the same as of the corresponding input files, or the current directory, if the input file names do not contain directory information. If the directory does not yet exist, it is created.

-eoltype *mac/pc/unix*
Chose between different end-of-line characters. For Unix (the default), lines are terminated by an NL character. Macs use a CR character, and PCs a CR/NL pair. This option is ignored if the output file format is binary.

-expand 0/1
If this option is activated, an attempt is made to expand identified superatoms in the input data, such as "COOH" or "CO_2Et". Currently, a built-in table with about 200 fragments is used, which cannot be extended by user-defined fragments. Expanded fragments are laid out in the opposite direction of the bond connecting them to the framework. This only works reasonably well in case of fragments that are connected only by a single bond to the rest of the structure. Recomputation of layout coordinates is recommended in these cases. Superatoms which cannot be resolved are passed through unchanged and output as such in files which allow this. In case of output formats without superatom support, unexpanded superatoms are discarded. Some file formats, such as the ISIS SKC format, may internally store selected (but generally not all, even within a single drawing there may be closed and pre-expanded fragments side by side) fragments already in expanded form, even if they do not show up expanded in native applications such as ISIS/Draw. If such a pre-expanded fragment is encountered, its expanded form is automatically used in this application without need to use this option. However, even in such cases, this option can be helpful to expand those fragments which are not pre-expanded.

-feedback n
If this option is set to a value larger than zero (the default), a control message is printed after processing a block of *n* items (structures or reactions). The current record number and the object name is printed on the standard error channel.

-fixbridgeheadstereo 0/1
If this flag is set, the program tries to detect problematic stereo center displays at bridgehead atoms. In case of a bicyclic ring system, the bridgehead atom to the front always has a solid up wedge, and the rear atom a dashed wedge. If necessary, the structure is flipped vertically - which may result in a drawing which appears to be drawn with a flipped template in case templates are used. This flag has an effect only if new 2D coordinates are computed. By default it is not set. In any case, the local stereochemistry at all atoms is correctly depicted - but if these atoms are part of a complex ring system with bridges, the overall geometry may be questionable.

-format
Set the output file format. If this option is not used, an attempt is made to guess the output file format from its suffix. Please refer to the introductory section for more information.

-help
This options prints a short help text explaining the most common options. Afterwards, the program exits.

-hydrogens *stripall/stripstereo/strip/asis/add/hetero*
Adapt the hydrogen set. Some file formats prefer or require implicit hydrogens. This option can be used to remove all hydrogens (*stripall*), all hydrogens which are usually

not drawn in structure plots (*strip*), strip all hydrogens except those which define a stereo center (*stripstereo*), keep them as they are (the default, *asis*), or add all hydrogens required by the valence rules (*add*) or just to those positions where they are usually drawn, but remove them from other locations (*hetero*). Beginning with version 1.16, the software tries to preserve stereo information which is encoded as wedges to hydrogens if these hydrogens are removed by shifting the wedge with proper modification of the wedge type to an adjacent preserved bond.

-ignoreempty 0/1

If this flag is set, empty records without any remaining atoms after processing are not written to file. Sometimes programs have problems with this kind of input data, so there may be a need to filter out such records. This flag is active by default.

-imgparameters parameterlist

The parameters are a list of keyword/value pairs. They are transferred directly to the E_GIF, E_EMF_IMAGE and other image property computation settings and thus influence the display attributes of image output formats.

-implicith 0/1

This flag controls whether hydrogens which are implicitly defined by the syntax of a file format are added during reading or not. The flag is *on* by default. Note that it does not have any influence on the interpretation of file formats where structures can simply be written with or without hydrogens, such as MDL Molfiles. In contrast, SMILES defines an implicit valence for certain elements which is satisfied by adding a suitable number of implicit hydrogens if explicit hydrogens are not present. Reading a SMILES file with a cleared flag suppresses the implicit hydrogen addition step, effectively reading SMILES as SMARTS. This flag is interpreted directly within the basic file input routines. Hydrogen processing selected by the **-hydrogens** flag takes place at a later stage, and can be used to add or remove hydrogens regardless of the setting of this flag.

-info

If this flag is set, statistical information about the number of successfully converted records and conversion failures is written to the standard error channel.

-inputformat fmt

This option may be used to specify the input format explicitly. The recognized format names are the same as for the output format (**-format**). In most circumstances, this option should be omitted, because the input format is reliably auto-detected from file contents. The suffix of the name of the input file is also used as additional information to break ties. The danger of format name mismatches is avoided if the default automatic mechanism is used. However, in certain circumstances, it is possible that the input format is ambiguous. The most common example is when single SMILES strings are read via standard input. In this case, there is no file name, and certain SMILES strings are also syntactically correct SLN strings, albeit sometimes representing a different structure. In such cases, the input format (“smiles” or “sln”) could be explicitly set to enforce one interpretation.

Setting this option has the side effect that the automatic format detection is not performed and data processing immediately commences, without waiting for sufficient data to accumulate to allow format determination. If used on standard input, the automatic detection of *gzip* compression is also disabled in order to allow the instantaneous processing of data. Records read via standard input must be un-compressed if this option is used.

-**join** 0/1

This flag controls whether the input data is joined into one multi-record output file or not (the default, one output file for each input file is produced). This option can be used only with output file formats which support multi-record data (for example, SDF). Some formats, such as PDB, do support this feature in principle, but hardly any software exists to read such files. We recommend to use this function only on well.established multi-record file formats, such as SDF, RDF, SMILES, Cactvs/Binary and Cactvs/CBS.

-lock propertylist

The CACTVS system, which was used to implement this program, contains an automatic data consistency manager. Properties with known dependencies are invalidated if the structure changes in specific ways. For example, deleting a counterion implicitly invalidates a present CAS number. If this behavior is not desired, a list of properties which should remain valid under all circumstances can be specified here.

-**mapping** propertypairlist

This option can be used to specify the name of data fields in formats such as SD-files. The parameter is a list of pairs of property names (either in the CACTVS nomenclature, or as read from the input sources) and the desired name in the output file. For example, the specification „E_NAME Catalogname E_WEIGHT Molweight“, in combination with the parameter *-properties „E_NAME E_WEIGHT“* adds two data fields to an SD-file output, one with the label *Catalogname*, containing the structure name, the other with the label *Molweight*, containing the molecular weight.

-**maxtautomers** n

Set a limit to the number of tautomers generated, if the tautomer generation or canonization mode are active (see -**tautomers** option). The default is 100 tautomers per file record.

-miniheader 0/1

If set, some file formats are written with only minimal header information. This option is intended to be used for software which cannot process the full header specifications of formats such as MDL Molfiles.

-**nitrostyle** *ionic/asis/penta/xionic/xpenta*

This option controls the encoding of nitro groups and similar functional groups in the output file. If the option is not set, or set to *asis*, no processing takes place. Otherwise, all nitro groups and similar functional units are re-coded as charge pairs (with a tetravalent, positively charged nitro etc. atom, and a negatively charged ligand) or alternatively as the uncharged variant with an octet expansion on the nitrogen (or similar atom). The *xionic* and *xpenta* options process a broader class of functional groups than the standard options - as an example, they include azides.

-noradicals 0/1

If this flag is set, any information from the input files which indicates the presence of radicals (such as M RAD lines in MDL files) is suppressed. Erroneous radical information may be found from time to time for example because in an earlier I/O step implicit hydrogens were not taken into account and thus spurious radical centers were thought to be present by processing software. This software does not create spurious radicals by changes in the hydrogen status, so usually this flag should not be set (which is also the default). In case the flag is set, information about actually present proper

radical centers is deleted, and in case hydrogen is added, these centers are hydrogenated, so this option can have undesired side effects.

-offset n
This parameter specifies a record offset into each individual file which is processed. The default offset is 0. If used in combination with the *-count* option, sections of larger files can be processed.

-outfile filename
This parameter defines the name of the output file, or a part of the name if multiple files are generated. Please refer to the introductory section for a discussion of the file naming algorithm. The output file name can be an anonymous ftp URL.

-parent 0/1
If this flag is set, the parent structure(s) if the input objects are generated and output, instead of the original input data. A parent structure is generated from the original structures by removing counterions, and removing charges by protonation and deprotonation, where possible. The effect is comparable to using the *-uncharge* and *-desalt* options simultaneously.

-pedantic 0/1
If set, error messages are produced for various non-critical errors in input files. By default this option is off.

-properties propertylist/*all*
This parameter determines additional data fields that are written out if the output file format supports this feature. It is usually used for SD-files or Cactvs/Binary files. The properties can be specified in the normalized CACTVS nomenclature, or as they appear in the input file. Note that the property data does not need to be already present in the input files as long as CACTVS can compute the data from the available information. For example, a molecular weight field (E_WEIGHT) can easily be added, regardless whether this number was read from file or computed from the structure data. If a requested property cannot be found or computed, it is simply omitted from the output file without raising an error. This option is often used in combination with the *-mapping* option to set the data field name in the output files. Usually, only ensemble properties are used - while CACTVS also outputs data of other chemistry objects such as atoms, many simple SD-file readers choke on such data. If the special property name *all* is used, all data fields from the source file are copied.

-pseudo3d 0/1
This is a parameter which controls the layout algorithms for 2D structure plot coordinates. It has no effect if 2D data is already present in the file. Essentially, if set, the longest central carbon chain in the input structure is depicted in a pseudo-3D projection. It does not have an effect on compounds without central chains of at least three carbon atoms.

-purgeisotopes 0/1
If this flag is set, any isotope information read from the input is purged before any further structure processing commences. By default, isotope information is retained.

-purgestereo 0/1
If this flag is set, any stereo information read from the input is purged before any further structure processing commences. By default, stereo information is retained.

-**recalc** 0/1
If set, 2D or 3D coordinates found in the input file are discarded and recomputed when required during file writing. By default, this option is not active and information is preserved where possible.

-**resolvearo** 0/1
A number of well-known chemistry software packages do not implement the MDL structure exchange formats correctly. According to the original specifications, an aromatic bond in these files can only be used as a query attribute, and it is read as such by CACTVS and therefore does not have a bond order, electron count, etc. However, if this flag is activated, the program resolves such aromatic bonds into a Kekulé system and not interpret the input as ISIS query data.

-**saltfile** filename
The name of a multi-record file in arbitrary format (usually Cactvs/Binary or MDL SD-file) with fragments that should be discarded when removing counterions. If the *-desalt* option is not set, this parameter has no effect.

-scope *mol/ens/reaction*
This parameter is intended for expert use. It controls the basic units read from the input file (normally, molecules or molecular ensembles from structure files and reactions from reaction files). However, if the scope is set to *ens*, it is for example possible to read reagent and product ensembles as two separate records from a reaction data file. The precise meaning of *mol* vs. *ens* data input depends on the file format, and only a few formats make a distinction at all. It is not possible to use a mode higher in the hierarchy than the file contents actually present (i.e. it is not possible to read reactions from a simple structure file).

-**separate** 0/1
If set, multi-molecule input records are split into multiple output records with one molecule each. If the output file cannot store multiple records, more than one output file may be generated. For MDL reaction files, a split produces reaction records with multiple reagent or product sub-records instead of one multi-molecule reagent and product record each.

-**showcrossedbonds** 0/1
If set, which is the default, mark stereogenic double bonds which do not possess expicit stereochemistry as stereochemically undefined, e.g. draw them as crossed bonds in depictions or attach an explicit ‘stereochemically undefined’ tag in formats which support this concept. If the flag is unset, such annotations are suppressed.

-skiperrors
If this flag is set, records which could not be read due to syntax errors in the input file are skipped and processing resumes with the next record after the defective entry. By default, the processing of such a file is stopped when an error occurs. Note that it is difficult in some file formats to re-synchronize to the beginning of a new record if an error was encountered, so sometimes correct records following the corrupted entry may still not be readable. Also, in extreme cases a faulty record among the first few records may prevent processing altogether, if the problem prevents the file format recognition routines from working properly. Simple ASCII file formats such as SMILES or SDF are the most robust to use with this feature.

-**split** 0/1
If this option is set, the output is split into files with a single record each, even if the

chosen output format supports multi-record data. Please refer to the introduction section to learn about the mechanisms behind the naming of output files.

-structurekey property

This option is used in combination with the *-tablefile* and *-tablekey* parameters. If this parameter is not specified, it defaults to the name of the table column identified by the *-tablekey* parameter. If both are unspecified, both default to *record*. If the *-tablekey* parameter is a table column index, the structure key is not identical to the table key - it is the name of the indexed column, or *coln* (n being the index plus 1) in case the table columns are not named. This structure key is interpreted as a name of a property, either in toolkit nomenclature or as used in the input file, which is used to find rows corresponding to the current structure. If it is the special value *record*, the current global record count (ignoring any offsets, and not reset in case multiple input files are processed) is used as row number. Otherwise, the property value is extracted from the record (it may be computed in case of computable properties) and used to find the corresponding table row with data to be added to the structure.

Example: csfc -tablefile data.tab -tablekey 1 -structurekey E_NAME data.smi

This example associates the second column in the input table (index begins with 0) with the structure name (whitespace-separated part to the right of the SMILES string).

-suffixfmt 0/1

If this option is given, the format of the input files is extracted from the suffix of their name. By default, the software instead looks at a couple of lines at the beginning of a file and determine its file format automatically, ignoring the suffix except as a an additional hint. However, there are cases where already the first lines of a file are corrupted, and thus the format detection routine may fail. If the file still should be processed with the **-skiperrors** option in order to salvage data from it, it can be useful to override the default format detection algorithm with this flag and name the file with a standard distinctive suffix, such as *.smi* or *.mol2*. The suffix *.mol* should be avoided because it is not unique.

-suppress propertylist

By default this property list is empty. It can be used to specify properties (in CACTVS nomenclature or the original name in the input file) which should not be copied to the output files, even if they could encode this data. Please refer also to the *-mapping* and *-properties* options for additional information.

-tablefile filename

This option specifies the name of a table file which contains additional structure information. Data from this table is merged to the current structure. By default, if the parameter *-tablekey* is not used, the data in the table is assumed to be in the same sequence as the structures in the input file(s). Data assignment takes place before any operations which could change the number of structures coming from the input sources, i.e. before any filtering, tautomer generation, etc. If the table contains explicit column names, the output data fields, in file formats which support this, bear the same name as the table column. In case the columns are not named, synthetic names in the form *coln* is used, with *n* being the index of the column, starting with one. The format of the table file is automatically determined. Supported table formats include dBase3 files, Sylk, Sybyl tables, and simple text tables with fields separated by a separator character like tab, vertical tab, whitespace or semicolon. In case of text tables, data fields with embedded separator characters are allowed if the data is enclosed in quotes, and quotes in the quoted text are allowed if escaped with a backslash. An attempt is

made to identify and optimize the data type of columns automatically. If this fails, data is assumed to be a string. In case any columns resolve to a non-string data type if the first row is ignored, it is assumed that the first row represents column names in formats which do not provide explicit column naming. The input of Excel files, a commonly requested feature, is not yet supported.

Example: csfc -tablefile data.tab data.sdf

This simple example assumes that there is one table row for every record in the input file, and both input sources follow the same sequence. Both the structure and table keys default to *record*. Data from one table row is merged to every structure record.

-tablekey columnname/index

This option is used to set the name of the key column when a structure file is merged with a data table (see option -*tablefile*). If no merging takes place, the option is ignored. The default value for this parameter is *record*, indicating that the sequence of structures in the input file(s) is the same as in the table. Alternatively, a name of a column in the table, or its numerical index (starting with 0) may be specified. The special value *record* always works, even if the table does not contain an explicit record column, because it is added when the table is read and filled with the row number in case it is not present as an explicit column. The column name of the index column is by default used as structure key, but this may be overridden by the -*structurekey* option. The structure key is interpreted as the name of a property (in toolkit nomenclature, or as used in the structure input file) of the current structure or reaction. It is possible to use structure keys which are computed on the fly from more elementary structure data.

Example: csfc -tablefile data.tab -tablekey REGID data.sdf

This program invocation merges data from the input table with the structures read from the SD file. The table must have a column named REGID, and the SD file likewise.

-tablekeytype *single/multi/strict*

This parameter defines the processing of cases where a structure key matches more than one table row, or no row. The default value is *single*. In this case, if there are multiple matching rows, only the first is used, and it is no error to have structure keys without corresponding table rows. Mode *strict* is nearly the same as *single*, but it is an error to find no matching table row for a structure record. In mode *multi*, multiple table rows are assigned to distinct property instances and output as such in file formats which support this.

-tautomercoordinates *discard/preserve*

The option decides on the fate of existing 2D and 3D coordinates when generating tautomers (option **-tautomers**). The default is *discard*, meaning that any existing atomic coordinates are discarded. The alternative *preserve* keeps the core fragment coordinates. Coordinates of hydrogens are recomputed.

-tautomerrules *full/restricted*

Select the set of structure transformation rules to be used in tautomer generation (option **-tautomers**). The default is *full*, a set primarily designed to be comprehensive and to virtually guarantee that all low-energy tautomer forms are included even if starting from a high-energy input compound. This rule set is especially useful for registration purposes. The *restricted* rule set generates a smaller, less comprehensive set of tautomers and is intended to be used, for example, for modelling applications.

-tautomers *none/all/unique/canonic/2/3/4..100*

Select tautomer processing mode. By default, tautomer forms of structures are output as they are encoded. This corresponds to tautomer mode *none*. In mode *all*, a comprehensive set of tautomers is generated, and these are all written if the output file format supports it. Additional processing options, such as hydrogen status, uniqueness, etc. are applied to all molecules of the tautomer set. Tautomer generation mode *unique* (or its alias *canonic*) normalizes the structure to a canonic low-energy form. All structures within a tautomer set collapse to the same unique structure. This canonic structure is not guaranteed to be the one lowest in energy, but usually is a sensible choice. However, tautomer forms with double bonds which could potentially exhibit stereochemistry are artificially downgraded in their energy rating in order to sidestep the problem of whether these forms have defined double bond stereochemistry or not. It is also possible to specify a positive number as argument to this option instead of one of the listed keywords. In that case, the generated tautomer forms are sorted and the n best-scoring tautomer forms are output in the order of their score, with the highest scoring structure coming first. It is no error if an input structure does not generate a sufficient number of tautomers to match the specified number. In that case, all found tautomers are output. Specifying 1 as argument is equivalent to the *unique/canonic* keyword. The maximum numerical tautomer count for this option is 100.

Tautomer processing is currently disabled when working with reactions instead of structure records.

-template SMILES/SMARTS

A substructure template in SMILES or SMARTS notation. This option is used in combination with the **-templatealign** option.

-templatealign *none/x/y/diagonal/rotate/redraw/besteffort*

Align the layout of the structures according to a substructure template, which was specified by the **-template** or **-templatefile** options. If no substructure is present, this parameter is ignored. The substructure is matched on all structures. If is does not match, no error is generated and processing continues as if this parameter had not been specified or set to *none*. The *redraw* option implicitly sets additional substructure flags which restrict matching of substructure *ring* atoms and bonds to corresponding structure atoms and bonds which are in the same class of ring system. With this option, a ring system must be matched completely, so for example a phenyl ring in the template does not match a naphthalene ring in a structure. Non-ring parts of the structure may be matched in any style. The only exception is that it is also allowable for a terminal chain atom in the template substructure to match a ring atom in the structure. The other template match variants do not have the ring system match limitations. If the template substructure does not possess 2D coordinates of its own and the *rotate* or *redraw* modes are selected, coordinates for the template are computed by the standard 2D layout procedure. The first of potentially multiple different matches of the template substructure is used a the starting point for structure coordinate updates. The simple *x*, *y*, and *diagonal* modes align the major axis of the matched atoms of the structure to the x and y axis or on a 30 degrees angle to the x axis, respectively. For these options, no substructure 2D layout coordinates are used. The *rotate* variant rotates the structure by multiples of 30 degrees with and without a coordinate flip. From among those 24 orientations, the one with the best similarity to the substructure coordinates is chosen. Finally, the *redraw* variant regenerates the 2D layout coordinates, using the matched fragment with its coordinates transferred from the substructure as the nucleus for the layout. In this style, all matched structure

coordinates have exactly the same relative coordinates as the substructure atoms, but the standard bond length is scaled to one. The *besteffort* (or *combined*) combined mode first attempts to match the template with the redraw option. If the template does not match in this mode, a second attempt is made with a *rotate* mode.

-templatefile filename
The name a file which contains a substructure template. This option is used in combination with the **-templatealign** option. Only the first record of the file is read. If the file does not contain 2D coordinates, and these are needed for the selected **-templatealign** option, coordinates are generated by the standard layout algorithm.

-templatematch *strict/relaxed*
This flag influences the match operation of fragments on structures. In mode *strict*, aliphatic fragment atoms do not match aromatic structure parts. In *relaxed* mode, the default, aliphatic fragment atoms also match aromatic systems. This setting applies only to atoms without further query attributes. If any atom bears an explicit *aromatic* or *aliphatic* query attribute, this attribute has precedence. If a template is specified as SMARTS strings, uppercase atoms are decoded without an explicit *aliphatic* query attribute and do thus match aromatic systems. This option can be used to counter that convention.

-timeout2d secs
Specify a per0-record timeout for 2D atomic layout coordinate generation. The argument can be a floating point number to set the timeout with sub-second precision.

-timeout3d secs
Specify a per0-record timeout for 3D atomic coordinate generation. The argument can be a floating point number to set the timeout with sub-second precision. Note that this parameter has a direct effect only on program versions with a built-in CORINA 3D coordinate generator module which compute 3D coordinates locally. If 3D coordinates are fetched from other locations, such as PubChem, the Web access attempt is not timed.

-uncharge 0/1
If set (by default this parameter is not active), the program attempts to neutralize a compound, for example by removing excess protons or adding them to negatively charged atoms.

-unique *none/default/isotope/parent/stereo/tautomer*
This option is syntactically unique in that it accepts a combination of any of the terms listed above, as a whitespace-separated list (which probably needs quoting on the command line in order to avoid separation by the shell). If this flag is set to anything but *none*, records which were already encountered in any file processed during the current run are discarded. The *default* mode activates the default duplication check mechanism, which ignores differences in isotope labelling and stereochemistry, and considers tautomer forms to be distinct compounds. Adding a single flag, or any combination of the flags *isotope*, *stereo* or *tautomer* activates the recognition of isotopic label isomers and/or stereo isomers as distinct compounds, and/or the recognition of tautomer forms as a common structure, respectively. In case any of the modifier flags are used, the *default* mode flag may be omitted because it is implied in the presence of a modifier flag. If the *parent* flag is added, duplicate detection is performed on the parent structure (see option *-parent*), not the original input structure. Note that modifier flags such as *parent* or *tautomer* do not directly influence the output of structures which pass the duplicate filter. The encoding of the output structure is still

preserved as far as possible. If an actual change of the output structure to a standard form is desired, the editing options *-tautomers* and/or *-parent* should be used.
If reactions are filtered, both the reagent and product side must be identical to a previous record in order to qualify for removal.

-usearo 0/1
This flag is deactivated by default. If set, aromaticity information is used to format the output. For SMILES, this means that the π-center variant of atom encoding is used. For example, benzene in the Kekulé-style default encoding is C1=CC=CC=C1, but c1ccccc1 in aromatic encoding. For MDL formats, aromatic structure bonds can be output as aromatic query bonds, if necessary. The latter is not correct use outside ISIS structure query specifications according to the original format definition, but there are a number of programs which expect their input encoded this way (see also *-resolvearo* flag).

-usename 0/1/2/3/4
The exact effect of this flag depends on the output format. For formats which support the optional addition of a name field (for example, SMILES or InChI), any value larger than 0 activates this output option. Values of two or more also change the methods how names are derived. If the input structure or reaction does not already possess a name, value 2 tries to look up a suitable name from the NCI resolver, value 3 from PubChem, and value 4 from both Resolver and PubChem. Note that these remote lookup options imply Internet structure export. They should not be used with confidential data. The default *usename* value is 1.

-usestereo 0/1
This flag controls whether stereochemistry information should be contained in the output. By default the flag is active. Note that by default no attempt is made to *compute* stereo descriptors from secondary information (see *-computestereo* option for a way to achieve this). This option only controls whether the output file should contain stereo information (such as wedge bonds, atom parity codes) which was either already present, or requested with the *-computestereo* option.

-version
Print version and licensing information and then exit.

-wedgepairs *no/all/hydrogen/unconnected*
Add extra wedges when generating 2D coordinates. By default (option value *no*), only a single wedge is used to determine the stereochemistry at atomic centers. In modes *all, unconnected* and *hydrogen*, a second wedge is attached to a bond extending from the stereo center to further disambiguate the stereochemical relationships. The second wedge is preferably attached to a bond leading to a hydrogen ligand, if such neighbor exists. In mode *hydrogen*, only bonds to hydrogen are used, if no such bond exists, no second wedge is drawn. Mode *unconnected* is the same as *all*, with the exception that the extra wedge to the preferred ligand is not drawn in case this ligand already participates in a wedge bond, regardless whether it is located at the tip or the bottom of such a wedge bond. In any case, the wedge is never drawn to a ring atom - if no chain neighbor without a wedge bond exists, the second wedge is omitted in all modes.

-wedgestyle *default/opposite*
If the mode *opposite* is selected, the drawing of the environment of atoms which are located at the tip of two or more wedge bonds of the same class (solid or dashed) is changed. If any of the offending ligands is a single atom, it is moved into another gap of the neighbor sphere in such a way that the wedge style is changed to the opposite,

and the stereochemistry of the center retained. While this feature is useful to meet specific drawing style demands, the graphical quality of the structure layout usually suffers.

-writemode a/w

The default write mode is *w*, meaning that existing files are overwritten if sufficient permissions exist. The alternative mode *a* is used to append to existing files. Please note that many chemistry file formats are inherently single-record formats and cannot be used with this option.

COPYRIGHT