NAME

cstb - table builder tool

VERSION

1.31, 2014-11-15

SYNOPSYS

cstb [-align *none/x/y*] [-assign propertypairlist] [-autocolumns 0/1] [-cleardirectory 0/1] [-col*n* colspec] [-colblocksize n] [-colnames 0/1] -coretemplate file_list] [-count nrecs] [-defaultformat fmt] [-embedformat *none/cdx/skc*] [-eoltype *mac/pc/unix*] [-fixstereo 0/1] [-flatten 0/1] [-footer text] [-format *arff/cactvs/cdxml/colon/comma/dbase3/dif/dta/excel/fits/google/hdf5/html/htmlpage/json/knime/mtp/pdf/r/rtf/sav/semicolon/space/spss/sql/sqlite/sybyl/sylk/tab/vbar/vtab/weka/xlsx/xpt/xml*] [-header text] [-highlightfile filename] [-highlightsmiles SMARTS] [-hydrogens *stripall/stripstereo/strip/asis/addall/addhetero*] [-imageborder n] [-imagedirectory dirname] [-imagetype *wmf/bmp/pct*] [-imagenote *cdx/cdxml/molfile/none/sdf/smiles/skc/tgf*] [-imageurl urlprefix] [-imgannotationfontsize points] [-imgasymbol *none/symbol/xsymbol/label/index/box/compact*] [-imgatomcolor color/*type*] [-imgbgcolor color] [-imgbondcolor color/*split*] [-imgbonds n] [-imgborder pixel] [-imgcolormode *monochrome/greyscale/color*] [-imgdashes 0/1] [-imgfooter text] [-imgfootercolor color] [-imgfooterproperty property] [-imgheader text] [-imgheadercolor color] [-imgheaderproperty property] [-imghcolor color] [-imghighlightcolor color] [-imgheight npixel] [-imghsymbol *none/special/all*] [-imglinewidth n] [-imglogofile filename] [-imglogoscale factor] [-imgshowcharge 0/1] [-imgshowchirality *none/simple/extended*] [-imgshowisotope 0/1] [-imgshowmapping 0/1] [-imgshowradical 0/1] [-imgshowstereo 0/1] [-imgshowstereoh 0/1] [-imgsymbolfontsize points] [-imgusehatch 0/1] [-imgwedges 0/1] [-imgwidth npixel] [-keepstructures 0/1] [-mapping propertypairlist] [-maxrows n] [-mergecolumns columnlist] [-name text] [-nitrostyle *ionic/asis/penta*] [-offset nrecs] [-outcolumns columnlist] [-outfile filename] [-paperorientation *portrait/landscape*] [-papersize *a4/a3/letter/legal*] [-recalc2D 0/1] [-resolvearo 0/1] [-rotate 0/1] [-security 0/1] [-sheetname text] [-select expression] [-sort sortspec] [-summary 0/1] [-suppress propertylist] [-template SMILES/SMARTS] [-templatealign *none/x/y/diagonal/rotate/redraw*] [-templatefile filename] [-title titlestring] [-version] [-writemode a/w] [writestructures 0/1] ?inputfile?...

DESCRIPTION

cstb is a tool for building tables in various formats from structure data. Besides data, images may be generated and embedded into the tables if the output format supports it.

The program reads structures from one or more files specified on the command line (or from standard input if no file names are given), build a table with data in memory, and writes it out in the selected format, either on the standard output channel, or the file name specified by the *-outfile* option. The program detects and understands a large number of structure exchange formats, for example SDF, SMILES and PDB as input files.

The desired columns in the table are specified by one or more -col*n* options. *n* is a number in the range between 1 and 50. The numerically ordered sequence of the actually used numbers determines the order of the columns. So, for example, if *-col3*, *-col5*, and *-col1* are specified in that order, a table with three columns is generated, with the column defined by *-col1* inserted as first table column, *-col3* as second, and *-col5* as third. No empty filler columns are generated. The order of the *-col* options does not matter, only

their number components.

A -col*n* option has one to three parameter fields. If more than one field is used, it must be specified as proper a Tcl list, for example as'-col1 "E_WEIGHT {Molecular Weight} {fgcolor red format {bold right}}"'. This construct defines a column which stores the property E_WEIGHT, is titled "Molecular Weight", and is written as red, bold, right-aligned text.

The first and mandatory field of a -col*n* option is either a property name, or one of the four magic words *Image*, *Blank*, *Record,* or *Function*. A property name can be either specified in the standard CACTVS syntax (such as E_WEIGHT for the ensemble molecular weight), or in its original name as it is stored in a data file (for example, the name used in the data section of an SD-file). The name lookup is case-sensitive. Note that all property names which use identifiers gathered from the file must be already defined in the first record of the first input file. This record is read ahead of the rest of the input data in order to retrieve these definitions. While it is not required that all file records contain the full set of property data, it is not possible to use property names which will only show up in later records, or later files in case of multiple input file processing. In such case, the synthetic CACTVS naming scheme must be applied, or a **-mapping** option specified. For file data names which do not adhere to known CACTVS property names with or without the *E_* prefix, the auto-generated CACTVS name will be E_**name***, where *name* is the name data field in the input file. These synthetic CACTVS names may be used even if they are not defined in the first input record.

It is possible to add columns with arbitrary CACTVS property data, if this data can be computed by the software from the structure data and other information present in the file. For example, if a structure connection table is contained in the input file, columns such as E_NATOMS (total number of atoms), E_WEIGHT (ensemble molecular weight) or E_COMPLEXITY (structural complexity rating) may be added without problems.

The three magic column property names *Image*, *Blank*, and *Record* insert a structure drawing in a suitable format, a blank column, or the file record number, respectively. These magic columns can be named and assigned display attributes just like normal property data columns.

The magic column name *Function* sets up a function column. A function column contains derived data, using functions and references either to data in other columns in the table, or to structure property data, which might have been read from the file or computed. The set of functions which can be used to manipulate and combine the data is a superset of the standard SQL function set. A full documentation of the functionality of this rather extensive subsystem can be found in the general scripting language documentation of the Cactvs toolkit.

The second field is the table column name. If it is not specified, or empty, the name of the column is the same as the name of the property or magic name.

The third optional parameter field allows the setting of display attributes for the column. This is a list of attribute-value pairs. The following attributes are currently officially supported:

fgcolor - foreground color, specified as a color name in the X11 color database, or in the #rrggbb format

fgcolorfunction - a function which computes the color from column data

bgcolor - background color, specified as a color name in the X11 color database,

or in the #rrggbb format

bgcolorfunction - a function which computes the color from column data

font - a font family name in short notation, i.e. *Arial*, *Symbol, Courier*

fontsize - the font size in points (may be a floating point number)

comment - a free-text comment

description - a free-text column description

headerfgcolor - foreground color, specified as a color name in the X11 color database, or in the #rrggbb format which is used for header lines

headerfont - a font family name in short notation, i.e. *Arial*, *Symbol, Courier* which is used for headers

headerfontsize - the font size in points (may be a floating point number) which is used for header lines

precision - the number of digits after the decimal point for floating-point values. If the column data precision is explicitly specified here, the *useprecision* format flag (see below) will be automatically set.

imgborder - the image border. See **-imageborder** property for a method to set this value globally for all columns.

width - the column width in characters

format - a list of keywords from the set of:

none - no options, default
left - explicitly left-aligned data
center - data is centered
right - data is right-aligned
bold - use bold font for display
highlight - use highlight format for display
histogram - display as histogram bars instead of numeric values
border - surround by border
padding - use extra cell padding
expand - stretch to 100% width
top - data is vertically aligned on top
middle - data is vertically centered in the middle
bottom - data in vertically aligned on bottom
italic - use italic font
underline - text is underlined
mdlnote - attach an MDL Molfile as cell note (Excel only)
smilesnote - attach a SMILES string as cell note (Excel only)
verticaltext - if set, text is written vertically if supported (Excel only)
merge - if set, the column is merged into a multi-line cell beginning with the first column to the left without this attribute, or the column to the absolute left. A sequence of columns to the right of a merge start column may have this attribute. Currently processed only for Excel and HTML output.
useprecision - if set, use numeric precision information which was either decoded or guessed when the data for the table was read, or which was explicitly set for a table column.

headerformat - the same list of keywords as for the *format* keyword, but applied

to column headers.

While all these attributes can be freely set regardless of the selected table output format, their effect and whether they have an effect at all depends on the table output format. A simple comma-separated file simply cannot encode any formatting informations, and thus these attributes are ignored. The formats which support the highest percentage of the various formatting flags are the native CACTVS binary table format (this is the only format which fully preserves all information), the Excel BIFF (.xls) format and HTML.

One or more of the table columns may hold a structure depiction. In case of MS Excel output, the image is a resizable WMF (Windows Metafile) or PICT (Macintosh graphics format) vector drawing. When the table is written as HTML, HTML page, SQL command file or CACTVS native table file, the structure is embedded as a GIF image. The display style of images can be modified by the various *img** options of this program.

As an alternative to displaying numerical data as simple numbers, the Excel output format supports also the generation of histograms. Selected data columns with numerical values may be overlaid by a resizable histogram (using the *fgcolor*, *bgcolor*, *width* attributes) which is scaled between 0% and 100% using the minimum and maximum values of all rows in that data column.

The theoretical maximum size of tables which can be handled by the software is 2^32-1 rows and columns. Since the structures read from the file are not kept in memory after their data content has been copied to the table cells, the memory requirements even for assembling large tables are not excessive. Note however that some table file formats have rather low limitations on the maximum number of rows and columns - for example, the MS Excel file format version generated by this program will not allow more than 65535 rows or 256 columns.

This software supports NULL (missing) values in property cells and will try to preserve this information when writing output files.

Besides having columns with data which is related to the individual file records (a structure ensemble in CACTVS nomenclature, since more than one molecule may be encoded in a structure file record), it is possible to have columns with data on molecules, atoms and bonds. The proper data attachment is automatically determined from the property description. If such columns are present, a file record can result in multiple rows added to the table. For example, if columns with properties M_NATOMS (number of atoms in molecule) and A_SYMBOL (atomic symbol) are added, the number of rows added for each record corresponds to the number of atoms in that record. Data attached to the whole ensemble or molecules is duplicated for each atom row. In case of multiple molecules in a structure ensemble the molecule data of the molecule a specific atom belongs to is entered into the respective cells. The same mechanism is used for bond data. However, it is currently not possible to combine atom and bond data in the same table.

If a property is of an indexible type, it is possible to use property subfields indexed by name or numerical offset as columns. For example, the standard ensemble name (property E_NAME, data type string) is indexible as words, so if a name field in an input file is formatted to contain additional information, a column pair E_NAME(0) and E_NAME(1) can for example be used to extract the first and second words from the name string and enter only that data into the table. The input file name can be accessed as E_FILE(file), which can be useful if multiple input files are read.

As of version 1.12, an attempt is made immediately prior to the output of the result table to optimize the data type of string or un-typed columns. This is done by checking whether

all non-NULL entries in a column can be converted to either an integer, a float, or a date value. If this conversion succeeds, the column type is automatically adjusted.

EXAMPLES

```
cstb -count 20 -imgwidth 200 -imgheight 150 -imgbgcolor black \
    -imgbondcolor split -imgfooter Confidential \
    -imgheaderproperty E_NAME -imgatomcolor type \
    -imgfootercolor yellow -imgheadercolor green \
    -imgasymbol xsymbol -imgbonds 10 -imgsymbolfontsize -1 \
    -name MyExcelTable -header “Drug Design Report” -colnames 1\
    -outfile table.xls -sort “Weight up” \
    -col2 “E_NAME Name {fgcolor blue width 25 format right}” \
    -col3 “E_NATOMS #Atoms {bgcolor yellow format bold fontsize 15}” \
    -col4 Image \
    -col8 “E_WEIGHT {Weight Histogram} {fgcolor orange format histogram
    bgcolor grey width 20}” \
    -col6 “Blank {intentionally left blank} {width 30}” \
    -col10 Record -col1 “E_WEIGHT Weight {precision 1}” \
    -col20 ’Function {concat(“ProjectX_”,E_IDENT)} RegID {fgcolorfunction
    {Activity > 10 ? “red” : “blue”}}’ \
    TESTFILES/data.sdf
```

This will generate an Excel table with the molecular weight (one digit after the decimal point) in column 1, the structure name (25 char column, blue font, aligned to the right) in column 2, the number of atoms (on yellow background, bold, big font) in column 3, a structure plot (200x150 pixels, black background, colored bonds and atoms, structure name as header on the image) in column 4, a blank column, an orange/grey histogram of the molecular weight in column 6, and the record number in column seven. The last column (specified as column sequence number 20, but actually stored in column 8) is a sample function column, which constructs an ID string via a SQL string manipulation function from a string constant and property data, and in addition sets the text color of the individual rows of this column according to the value of the property *Activity* (which, because it is not following the standard property naming scheme, must be found in the input file under this name). The name of the worksheet is *MyExcelTable*, and the header will read “Drug Design Report” when the spreadsheet is printed. The table will contain 20 rows (or less if the input file is shorter), and the table rows are sorted by weight, with the lighter compounds first.

OPTIONS

-align *none/x/y*

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 table columns. Only properties of the same object class can be

assigned. The system will attempt 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. This option is also useful to merge the content of multiple input files with different data record naming conventions into the same column.

-autocolumns 0/1

If this flag is set, data columns will be automatically added for all ensemble-level properties which are found in the original file. These properties may appear anywhere in the file. Additional columns will be added at runtime if necessary. Records which do not contain data for some columns will end up with blank entries in these columns. New columns will always be inserted to the right of the rightmost column. Standard columns with attributes, images, functions, etc. can be freely combined with this flag and will be handled the usual way.

-cleardirectory 0/1

If this flag is set (it is off by default), the external image directory (**-imagedirectory**) is cleared when the program is run. This option has no effect if no image directory is specified.

-colblocksize *n*

Set the column block size value. If it is not one (the default), the columns (in normal mode) of the primary table are formatted as bundled groups of the specified size and not individually. If set, in unrotated output mode a display row does no longer just contain the data of a single underlying table row with its specified data columns, but the data of a block of *n* such primary table rows, each with their own data column set, i.e. if there are two data columns in the primary table, and a block size of two, there will be four display columns in unrotated output, showing the data of two rows of the original table per display row. If the number of output primary table rows is not divisible by the block size, the lower right corner cells remain empty. This option currently has an effect only on the Excel (xls), Excel XML (xlsx), PDF (pdf) and HTML (page and table) format outputs - it generally makes sense only with output designed for human viewing, not for data entry.

-colnames 0/1

Choose whether the column names should be included in the output table or not. By default the flag is set. This flag is ignored in formats which differentiate between cell content and column names, such as the native CACTVS table format or SQL table definitions.

-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 will be 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 will 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 nrecs
Convert a maximum of *nrecs* 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.

-defaultformat fmt
Set default format attributes for columns. The allowed attributes are the same as in the third argument to the **-col*n*** column definition statements. This attribute set is implicitly prefixed to all column format specification statements. Any attribute set explicitly in a column definition statement overrides the attribute with the same name if it is set in the default.

-embedformat *none/cdx/skc*
The format of embedded OLE objects for format which support this (currently only xlsx and rtf). If the value is *none*, the default, structure displays are embedded as passive images in a format dependent on the table file format, for example as WMF or EMF images for MS Excel and MW Word.

-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, such as Excel files.

-fixstereo 0/1
If this option is set, spurious stereochemistry is removed from the read structures. This means, atom and bond stereo descriptors, as well as wiggly bond flags, which are attached to atoms and bonds which cannot possibly be an atomic stereo center or a double bond with stereochemistry are removed. Unfortunately, structure data with an incomplete hydrogen set may be handled incorrectly if missing implicit hydrogens are part of the atom center or double bond neighbor set, because there may be no clear indication whether the potential stereo center or bond has a free electron pair instead. For this reason, hydrogen around a stereo atom or bond should always be encoded explicitly in the input data. By default, this flag is not set.

-flatten 0/1
The CACTVS table object is rather unique in its capability to store vector and multi-field data in its cells. This cannot be represented properly in other table formats. The *-flatten* option, which is set by default except when the output format is the native CACTVS table format, temporarily expands multi-element columns into multiple single-value columns. The columns usually have data types which can be represented

in export formats. The column names of the temporary columns are set to *colname*(0), *colname*(1), etc. If multi-value cells are not flattened, they are output as a string in Tcl list notation.

-footer text
A free-form text for the table footer. The exact output style depends on the output format. Generally, the string is printed below the table, and is not part of the table proper. If it is an empty string (the default), or the output format does not support this type of annotation, no footer is printed.

-format *arff/cactvs/cdxml/colon/comma/dbase3/dif/dta/excel/fits/google/hdf5/html/html page/son/knime/mtp/pdf/r/rtf/sav/semicolon/space/spss/sql/sqlite/sybyl/sylk/tab/vbar/ vtab/weka/xlsx/xpt/xml*
Select a format for the output table file. If this option is not specified, an attempt is made to guess the format from the suffix of the output file name, such as *.csv* or *.xls*. The number of formatting attributes supported by the various formats and also the acceptable data types of columns are dependent on the format. Only the *excel* (embedded WMF image for the MS Windows version, PICT image for Mac output, or BMP pixel images as system-independent format, see **-imagetype** option), *cdxml* (editable structure object in native table object), *sql* (BLOB column with GIF data), *pdf* (integrated PDF plot) *cactvs/html/htmlpage* (embedded or linked GIF image) and *xlsx* or *rtf* (embedded as ChemDraw or ISISDraw Ole objects, or EMF images) formats support embedded structure or reaction images in a form which goes beyond a simple file name reference as a string. *excel* and *xlsx* are the only formats which currently support the display of automatically computed histograms. The *rtf* output is written as portable native RTF tables (not an embeded Excel table, and thus editable in a different fashion), and can these be optionally augmented with embedded OLE structure drawings, just like *xlsx*. The difference between the *html* and *htmlpage* formats is that the former format is just an HTML <table> section, which is intended to be integrated as a building block into an HTML page for display, while the latter is a simple, but complete HTML page. The *colon*, *semicolon*, *comma*, *space*, *tab, vbar* and *vtab* formats are variants of simple ASCII text file dumps with specific field separator characters. If a data value contains the separator character, it will be protected by putting the value into quotes. In these formats, NULL values are represented by empty entries and thus not necessarily recognizable. The *sybyl* table format write a table which is suitable for loading into the Sybyl modeling package. *dif* and *sylk* are universal spreadsheet exchange formats similar to the MS Excel format, but with no method of integrating graphics. They can be read by MS Excel and other spreadsheet programs. *arff* is the native data table format for the Weka machine learning system. *sav* is the native data table format for the SPSS statistics package. *fits* is another universal table data exchange format which was designed by NASA. *hdf5* is somewhat similar, and is one of the preferred formats for the Octave data analysis package. *sqlite* output writes tables in a SQLITE *v3.x* database container file. *xpt* is the SAS statistics suite export file format, and *dta* a comparable format of the STATA package, as are *mtp* for Minitab and *r* for the R statistics packages. The *knime* format selected output as native table files for the KNIME pipelined data analysis package. This format, like the native Cactvs table format, can store structure and reaction data in the table file, if the *writestructures* flag is set. The *google* format is special - this is a virtual access module for spreadsheets hosted on Google Docs. The syntax of a properly formatted pseudo file name is "<google://user:password@gmail.com/spreadsheetname>", where the user, password and spread sheet name components are replaced by the proper

access credentials and spreadsheet name. Google spreadsheets may contain multiple worksheets - by default the first is selected, but the *-sheetname* parameter is also taken into account. The CDXML output format write a (potentially multi-page) document for the ChemDraw package which uses native ChemDraw table objects to arrange the information. Both data and structures are editable within ChemDraw.

-header text

A free-form text for the table header. The exact output style depends on the output format. Generally, the string is printed above the table, and is not part of the table proper. If it is an empty string (the default), or the output format does not support this type of annotation, no header is printed

-highlightfile filename

Specify a file with substructure fragments. All records from this file will be read and matched against the table structures. Implicit hydrogen addition is disabled while reading the file, but no explicit hydrogen stripping is performed, so substructures with extraneous hydrogen atoms may not match. If the table contains image columns, the matched atoms and bonds will be highlighted. Only the first match of the substructure will be shown, but if multiple substructures match the same table compound, instances of all matching substructures will be highlighted.

-highlightsmiles smarts_string

This is an alternative to the **-highlightfile** option. Here, a SMARTS substructure definition is decoded and used as substructure for highlighting purposes. Only a single SMARTS structure may be used.

-hydrogens *stripall/stripstereo/strip/asis/add/addhetero*

Adapt the hydrogen set. Some file formats prefer or require implicit hydrogens. This option can be used to remove all hydrogens (*stripall*), all hydrogens except those on wedge bonds (*stripstereo*), all hydrogens which are usually not drawn in structure plots (*strip*), 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*). The software will try 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. This option has an effect both on the appearance of structure drawings and on some calculated properties such as molecular weight, since CACTVS internally assumes structure encodings with fully specified hydrogens for computations. If the hydrogen status is changed for depiction purposes only, it is preferable to use image attributes to suppress unwanted hydrogen atoms.

-imageborder n

Specify the size of a border around an embedded image. The default value is 0. This option is only used for HTML-style and Excel output. For HTML, it is the thickness of the border in pixels around the GIF image on the HTML page (the BORDER=n field in the <IMG> tag). For Excel, it is the distance between embedded WMF structure drawings or histogram bars and the cell border. In this case, the unit is 1/255 of the cell height and width. Using a non-zero border in Excel files may prevent problems when sorting image cells. They may not move in sync with other sort columns if their size is very close to the cell size. On the other hand, values above 2-3, depending on the image size, prevent resizing of the drawings when the size of the cell is changed. Depending on the application, this can be a feature or a problem.

-imagedirectory dirname

The name of a directory where external images are stored for output formats which cannot store images internally (such as the native CACTVS table format, or Excel files). The directory is created if necessary, and it is cleared if the **-cleardirectory** option is set and an image directory is specified. If no directory is specified, external images are put into the current directory, and the **-cleardirectory** flag is ignored.

-imagenote *cdx/cdxml/inchi/molfile/none/sdf/smiles/skc/stdinchi/tgf*

This option is used to annotate embedded structure images (in Excel, Excel XML, PDF files and other file formats which support embedded images) with a structure record in the selected format. This structure record can for example be used in combination with suitable software for interactive structure export and processing functionality from within a table display tool, or simply for allowing the user to copy out the structure data via the clipboard. An Excel plug-in providing advanced functionality with this data is available from Xemistry as a separate product. The default is *none*, meaning that images are not backed by structure data.

-imagetype *wmf/bmp/pct*

This option is used only for Excel output. By default, embedded images are stored as WMF metafiles, which are small, scale without pixelating, and print well. On Windows computers this is always the preferable image type. However, on Macintosh computers, reading of Excel files with WMF images may be time-consuming and thin lines may disappear in the automatic conversion process. If the output file should be read on Macs, the image type should be set to *pct* (or *pict*), the native Excel image format on that platform. These files are still readable on Windows Excel - but conversion now proceeds in reverse direction. As another alternative, the embedding of the images as Windows bitmaps is also possible. While such Excel files may load well on both Windows and Macintosh platforms, they are much larger than WMF- or PCT-based files, and they do not print well because the bitmap images do not contain vector data and thus do not scale or adapt well to the superior resolution of printers compared to monitors.

-imageurl urlprefix

If this option is specified, the html-based output formats (*html* and *htmlpage*) will use its value as prefix to references to embedded structure GIF images. Note that a possible relative image directory (**-imagedirectory**) is already automatically added to the <IMG SRC> reference path, so it is not necessary to repeat it with this option.

-imgannotationfontsize points

By default (or if you set this number to less than 0) the program chooses suitable font sizes for annotations (charges, etc.) automatically. It is possible to override this choice with this option. The point size may be a floating point number.

-imgasymbol *symbol/xsymbol/label/index/box/compact*

Select the type of symbol to print. For normal atoms, the display types *symbol* and *xsymbol* are equivalent, but *xsymbol* will produce a more detailed text for certain types of query atoms, such as atom lists. With the *label* style, atom labels replace the atomic symbols.These are either taken from the input file (if the files stores this information) or correspond otherwise to the internal atom ordering, which is always the same as in the file. *Index* displays the index of the atom in the atom list, starting with one. In the box style, hetero atoms are depicted as a small rectangles. The *compact* mode displays hydrogen atoms on hetero atoms and other special hydrogens as a common symbol, such as OH or NH_2. Subscripting of hydrogen count numbers only works with Unicode

fonts.

-imgatomcolor colorname/*type*
Select a global color for all atom symbols. If the default color *type* is chosen, the color of atom symbols is determined individually. If the input data contains color information, it is used. Otherwise, a standard element-specific coloring scheme is applied. For hydrogen atoms, the hydrogen color specified with the **-imghcolor** option will override both a global atom color and an individual atom color, if it is not an empty string.

-imgbgcolor colorname/*transparent*
Determines the background color of the image. If the special value *transparent* is used, the background is transparent.

-imgbondcolor colorname/*split*
Select a global color for bond lines. In contrast to atom colors, individual bond colors are currently not supported. The special value *split* splits the bond into two halves. Each half bond is colored in the same color as its associated atom symbol, with the exception of carbon atoms. In case of half bonds to carbon atoms, the split color is either black or white, automatically selected according to the image background.

-imgbonds n
This parameter determines how many standard-length bonds should fit on the image in x-direction. Structures which possess less bonds are centered on the image. Structures which have more bonds are shrunk until they fit tightly into the image. A larger value will let the average structure appear smaller, but more structures will fit in the image without re-scaling, maintaining their correct size relationships. Smaller values will make plots larger, but re-scaling is necessary for more compounds. The default value is 10.

-imgborder pixels
The parameter determines the width of the border from the center of the outmost atoms of a structure which fits tightly into the display area to the outer border of the image. Note that atoms with plotted symbols require a few pixels in all directions around the atom center, so setting this parameter to very small values should be avoided. The default is 12 pixels.

-imgcolormode monochrome/gr[ae]yscale/colo[u]r
Selected the color rendering mode for embedded structure images. This attribute is currently only used for PDF output. The default is *greyscale*, designed for printing on monochrome laser printers.

-imgdashes 0/1
This option controls whether dashed bonds will actually be drawn in this style, or always in solid style. The default is 1, meaning that dash attributes will be rendered. This option interacts with the **-imgwedges** option. Resetting this option alone, without also resetting **-imgshowstereo**, is not useful.

-imgfooter text
This is a free text which is centered on the bottom of every image. Compound data can be automatically inserted into the footer with the **-imgfooterproperty** option.

-imgfootercolor colorname
Selects the color for the text written with the **-imgfooter** option. The default color is black.

-imgfooterproperty propertyname
If no explicit footer is set with the **-imgfooter** option, this option can be used to transfer the data associated with a property into the image footer field. Useful properties are for example E_NAME, E_FORMULA, E_WEIGHT, or E_SMILES, or the name of any table column. If the data is not yet present, but a method is available to compute the data from available information, it is automatically invoked. Currently, the footer property must be of the ensemble property attachment type, and subfield extraction is not yet supported.

-imgheader text
This is a free text which is centered on the top of every image. Compound data can be automatically inserted into the header with the **-imgheaderproperty** option.

-imgheadercolor colorname
Selects the color for the text written with the **-imgheader** option. The default color is black.

-imgheaderproperty propertyname
If no explicit header is set with the **-imgheader** option, this option can be used to transfer the data associated with a property into the image header field. Useful properties are for example E_NAME, E_FORMULA, E_WEIGHT, or E_SMILES, or the name of any table column. If the data is not yet present, but a method is available to compute the data from available information, it is automatically invoked. Currently, the header property must be of the ensemble property attachment type, and subfield extraction is not yet supported.

-imghcolor colorname
Specify an override color for hydrogen atoms. If the color name is not an empty string, it overrides both the global atom color, or an individual hydrogen color taken from file or the standard element color table.

-imgheight pixels
The height of embedded structure drawings in pixels. The default are 150 pixels.

-imghighlightcolor color
Specify a color used for highlighted atoms and bonds. The default is red. The information about the atoms and bonds selected for highlighting can either be present in the read data, or may be added by matched substructures (options **-highlightsmiles**, **-highlightfile**).

-imghsymbol *none/special/all*
Defines how hydrogen atoms are plotted. Mode *none* suppresses them all. In contrast to carbon atoms, the bonds to suppressed hydrogen atoms also vanish. Mode *all* plots them all, and mode *special* displays only hydrogen atoms in a few selected environments, such as on aldehydes, or when bonded to hetero atoms. The default is *special*.

-imglinewidth width
Base value for the computation of line widths of bonds. The default value is 1.4. Depending on the relative scaling of the structure coordinates and bond attributes (single/multiple/fat) this value is modified internally, so this is not a one-to-one mapping to the actual line width on the image, except in the case of single bonds for structures which fit into the drawing area without re-scaling. The parameter is a floating point number.

-imglogofile filename

If this is not an empty parameter, an attempt is made to read the file as GIF or PNG logo. If the image could be read, it will be inserted into the upper left corner of all structure images. The logo image may be scaled by the **-imglogoscale** option.

-imglogoscale factor

This factor can be used to resize a logo file specified by the **-imglogofile** option. The scale factor is a floating point number. Note that this option should not be used on a regular base - a cleanly rendered logo image in the intended final resolution generally has a better graphical quality than a rescaled logo. The default scaling factor is 1.0.

-imgshowcharge *0/1*

If this flag is set to 0, atomic charge symbols are not plotted. The default is 1.

-imgshowchirality *none/simple/extended*

If this option is set to *simple*, the images of compounds with defined chirality will be tagged with a small *chiral* marker in the upper left corner of structure images. In mode *extended*, the chirality status is described in more detail as *unspecified* (compound is potentially chiral, but chirality is not given), *meso* (all stereo centers are paired by an opposite topologically equivalent stereo center), *chiral* (at least one defined stereo center), or *contradictory* (multiple stereo descriptors clash, or stereo descriptors are given for atoms which cannot be stereo centers). If the option is set to *none* (the default), or the compound is not stereogenic, no tag is inserted into the images.

-imgshowisotope 0/1

If this flag is set (the default), isotopically labelled atoms are annotated with their nucleonic number. Heavy hydrogen atoms are displayed as *D* or *T*. If the flag is not set, isotopic information is suppressed.

-imgshowmapping *0/1*

This flag is not set by default. If set, atom mapping information (for example, if the structure was copied from a reaction database) will be displayed as atom annotation in the form <*n*>, where n is the mapping number of the atom.

-imgshowradical *0/1*

If this parameter is set to 0, marks for radical centers will not be plotted. The default is 0.

-imgshowstereo *0/1*

If this flag is set to 0, no stereo descriptors are plotted. Note that this flag has no influence on the display of wedge bonds (see **-imgwedges** and **-imgdashes** options to control their appearance). It only applies to atomic stereo descriptors such as CIP R or S, which might be present in the input file. These descriptors are not computed if not explicitly present in the input data (although CACTVS can do that in principle). They are only plotted if read from file. The default is 0.

-imgshowstereoh 0/1

If this flag is set (the default), hydrogen atoms at stereo centers which are not linked via a wedge bond are explicitly drawn in order to obtain an unambiguous stereocenter display. Hydrogen atoms linked via a wedge bond are always drawn regardless of the value of this flag.

-imgsymbolfontsize points

By default (or if you set this number to less than 0) the program chooses suitable font sizes for atomic symbols automatically. It is possible to override this choice with this

option. If this value is set to 0, no symbols will be printed. Hetero atoms are then marked by small (colored, if in color mode) squares. The desired point size can be a floating point number.

-imgusehatch 0/1

If this flag is set (by default it is not set), dashed wedges in WMF or PICT plots are drawn as a simple triangle and filled with a built-in hatch pattern of the Windows or Macintosh GDI. By default, every sector of such wedges is drawn as an independent filled polygon. In case of very small wedges, the use of the build-in hatch pattern may yield visually more pleasing results than polygon drawings. Larger wedges or printer output in a resolution higher than the screen display generally look better when drawn as polygons. Since the internal coordinate space of PICT images is smaller than that of WMF, this option is more often used with PICT images. In PICT images, dashed wedges drawn as individual segments do not scale well because of the limited coordinate resolution. High-resolution formats such as PDF ignore this option and will always use polygon rendering.

-imgwidth pixels

The width of embedded structure depictions in pixels. The default are 200 pixels.

-imgwedges *0/1*

If this option is set to 0, wedge bonds are not drawn as wedges. If the **-imgdashes** parameter is set, and the bond is a solid wedge, a bold line will be drawn instead, otherwise a dashed line. The default for this parameter is 1.

-keepstructures 0/1

A flag whether to keep the structures used in extracting the table data in memory. If they are kept in memory, and the **-writestructures** flag is set, certain output file formats (such as the CACTVS and Knime native table file formats) can save the structure and reaction data together with the table cell data. For most formats, this flag will have no effect, but can consume a lot of memory if there are many structures. The default is 0.

Keeping structures is also required if a select statement (option **-select**) is referring to property data which is associated with the ensemble that was used to fill the row, not simply existing table column data. For the latter the presence of the row ensemble is no longer required.

-mapping propertypairlist

This option is a method for associating data fields in input files with predefined CACTVS properties. For example, the specification *„E_NAME Catalogname E_WEIGHT Molweight“*, will associate the data field *Catalogname* with the standard property E_NAME and the data field *Molweight* with E_WEIGHT. This kind of mapping is especially useful for correctly setting the data type of yet unspecified properties.

-maxrows n

Limit the maximum number of rows in the table. Only this number of rows will be output at maximum. By default, the maximum is unlimited, which can be explicitly encoded with a negative maximum row count. Regardless of the setting of this parameter, the table is still fully assembled from the data files before this parameter is processed. Sorting and selection statements are executed before the row limitation step, so that this parameter can for example be used to select the records with the top 10 values in a column from a larger input file set.

-mergecolumns columnlist

Specify that certain columns should be merged into multi-line cells on output. This is equivalent to setting the *merge* attribute directly on a column when defining that column. The special value *all* sets the attribute for all but the first column. Otherwise, the argument is expected to be a list of column names or numerical column indices, and these columns are marked. Columns for which the attribute is set are concatenated into multi-line cells on HTML or Excel (xls and xlsx) output. For other output formats, this option is ignored. The combining starts with the first column to the left of a marked column, or the leftmost column, for which the attribute is *not* set. All data from a row in the start column plus subsequent data from columns to the right of that start column in an uninterrupted sequence of marked columns are combined into a single cell, and each data item in such a group is formatted with a forced line-break to guarantee that it appears on a new line within the cell. There may be multiple separate merge groups in a column set. Formatting of merged columns is limited - most of the style attributes are defined exclusively by the start column of that group, and style attributes of the merged columns are ignored. In HTML, text attributes such as font and color (but not the cell background color, etc.) can still be set for data from merged columns, but in Excel output these are currently ignored. Only data and formula columns may be merged, but not depiction or null columns. This option does not change the left-to-right order of the columns as specified by the *-colxx* or generated by the *-autocolumns* options.

-name string

Set the table name. This information will be preserved in output formats which have an internal table name field, but does not influence the name of the output file. The name filed is just used for annotation, with less syntactic constraints than the *sheetname* option.

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

This option controls the encoding of nitro groups and similar functional groups on images in the output file. If the option is not set, or set to *asis* (the default), 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 *x* variants of the option are more aggressive in identifying similar cases to the prototypical nitro group.

-offset nrecs

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.

-outcolumns columnlist

Specify a subset of the columns for output. This can be useful if the content of some columns is, for example, used only for formatting purposes, such as coloring the background of other columns. The default is *all*, meaning that all columns which have been created are also output, to the degree the selected output format supports the data type of these columns. Columns can be specified as a list consisting of individual column identifiers and/or open or closed ranges, using either the column names or numerical column indices to address columns. Column indices begin with 0 for the leftmost column. They usually do not correspond to the arbitrarily numbered *-coln* column specification numbers. In case a column name contains whitespace, it must be

quoted.

Example: -outcolumns ‘0-1 “My Column” 5-’

Above specification will output the leftmost two columns, column “My Column”, and everything after the fifth column. The output order of the selected columns is not changed, nor are multiply selected columns output more than once. The option argument is strictly used to select a subset of the table columns, in the order they are stored in the table.

-outfile filename

This parameter defines the name of the output file. If this option is not specified, output is written to the standard output channel. In case no explicit *-format* option is given, an attempt is made to guess the format of the output file from the suffix. So, if the file is named *table.xls* or *table.csv*, the format will be automatically set to *Excel* or comma-separated text file.

-paperorientation *portrait/landscape*

Specify the orientation of the paper used for printing. This is only used for PDF output. The default is *portrait*. PDF output will be automatically distributed onto multiple pages if necessary, in a grid fashion, without splitting any rows or columns

-papersize *a4/a3/letter/legal*

Specify the size of the paper used for printing. This is only used for PDF output. The default is *a4*. PDF output will be automatically distributed onto multiple pages if necessary, in a grid fashion, without splitting any rows or columns.

-recalc2D 0/1

If set, 2D coordinates found in the input file are discarded and recomputed when required for structure depiction. By default, this option is not set and information will be 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 will resolve such aromatic bonds into a Kekulé system and not interpret the input as ISIS query data. The flag is set by default.

-rotate 0/1

If this format is set, the output is rotated 90 degrees counterclockwise. Every output row consists of a column label (if requested), column header data (if present and supported by the output format), and one extra column per selected original table row. The position of table title, header and footer does not change. The total number of output rows is the number of selected input columns, plus possibly one title and/or one row name row. All row- and column-oriented commands are applied to the original table and their meaning is not exchanged by selecting output rotation. Not all output formats allow rotation. Output formats which currently support this feature include the simple ASCII separator tables, MS Excel (xls and xlsx), HTML (pure table and page) and Sylk.

-security 0/1

This flags controls whether the program may use Internet resources to look up additional information on structures, which generally means that structure data is

transmitted to un-trusted sources. Examples are the look-up of structure names (E_NAME), CAS numbers (E_CAS) and Pubchem CIDs (E_CID). If the security flag is on, structure data will not be transmitted, and requests for computing property data which requires Internet services is disabled. In commercial program versions, this flag is on by default. In academic packages, the default is off.

-select expression

Select only a subset of the table rows for output. By default, the full table is output. The syntax of the select expression is closely related to SQL and supports all standard SQL functions plus most of the MySQL database extensions. The set of SQL functions can be used on all table columns of suitable data types. A simple example is "-select 'E_WEIGHT < 300'", which will restrict table output in any format on the matching row subset. This expression will both work when there is either a weight column in the table, or the structures are retained for processing (see **-keepstructures** option) and the required data can be computed on the fly from them without storage in the table.

-sheetname text

Assign a sheet name to the table. This is mostly useful for tables in container file formats, such as *xls*, *xlsx*, *fits* or *sqlite*. If no sheet name is specified, but a title has been set, the title string is used. If that is also not set, a synthetic name with a syntax dependent on the output file format is generated. Sheet names will need to adhere to the naming syntax of the output file format. If it does not, a limited attempt will be made to rewrite and simplify the name, and if that does not succeed, output will fail.

-sort sortspec

By default, the order of the rows in the table is the same as the structures read from file. With this options, the sequence of the rows may be changed by sorting the table according to the values in one or more columns. The sort specification is a list of column names or name/direction pairs. Column names to the left have higher priority, column names to the right are only used to break ties. The default sort order is *up*, meaning that rows with a lower value of the cells in the sort column show up first. By combining a column name with either *up* or *down*, the sort order of each column may be specified individually. So, a sort specification such as *-sort "{Weight down} {Record up}"* would arrange the rows with the structures of high molecular weight first, with a preference for earlier file records in case of ties. The sort does take the data type of the columns into account, so truly numerical values are sorted differently than the same data encoded as strings if the strings are longer than a single character. Note that it is no always possible to determine the value type of property data when reading from file, and thus numerical data may be unintentionally read as strings. The value types of a properties can be unambiguously defined by providing proper property definition files (generated by the *cspe* program of CACTVS standard toolkit) which is automatically read by the program when matching names of yet undefined properties are encountered.

-summary 0/1

If this flag is set, a short summary of the table structure will be printed on the standard error channel immediately before the table file is written.

-suppress propertylist

This option has an effect only when the **-autocolumns** flag is set. The option argument is a list of properties, separated by commas or whitespace, either in CACTVS nomenclature or using the names of properties as they appear in the input file. Properties listed with this option are not automatically added to the output table.

-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*

Align the layout of the image 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 will allow matching of substructure ring atoms and bonds only on 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 will not match a naphthalene ring. The other match variants do not have this limitation. If the substructure does not possess 2D coordinates and the *rotate* or *redraw* arguments are selected, coordinates will be computed by the standard 2D layout procedure. The first successful match of the substructure is used as template. The *x*, *y*, and *diagonal* parameters will 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 will rotate the structure by multiples of 30 degrees with and without a coordinate flip. From among those 24 orientations, the one with the best overlay to the substructure coordinates is chosen. Finally, the *redraw* variant will regenerate 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 will have exactly the same relative coordinates as the substructure atoms, but the standard bond length will be scaled to one.

-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 will be generated.

-title titlestring

Specify a title string for the table. The exact output style for this data depends on the output format. Generally, this string is output as the first row of the table, in a style spanning all table columns. If the string is empty (the default), or the output format does not support this type of annotation, no title line is output.

-version

Print version and licensing information and then exit.

-writemode a/w

In some table file formats, it is possible to add a table instead of writing a new file. The added table continues to exist as an independent object - this is not a table data concatenation into a single result table. The formats for which this is currently supported are *hdf5*, *sqlite*, *fits* and *xlsx*. For these formats, an “a” mode will add the new table to an existing file. All other formats, or if the selected output file does not exist, will rewrite the file regardless of the mode setting. The default value of the parameter is “w”. When appending tables, it is recommended to set the *sheetname* attribute to assign a proper identifier to the table in the container. If the sheet name collides with the name of a table already existing in the file, the outcome is dependent on the file

format. If duplicate names are allowed, the table is simply added. Otherwise, depending on whether table deletion is supported in the file format, the old table is either removed before the new one is written (*sqlite*, *fits, hdf5*) or the sheet is renamed (*xlsx*).

-**writestructures** 0/1

If set, structure and reaction data is stored in the table file together with the cell data, if the table file format supports this (for example, the native Cactvs and Knime formats). Setting this option implies setting the -**keepstructures** option, since the structures or reactions need to remain accessible until the time the table is written.

COPYRIGHT