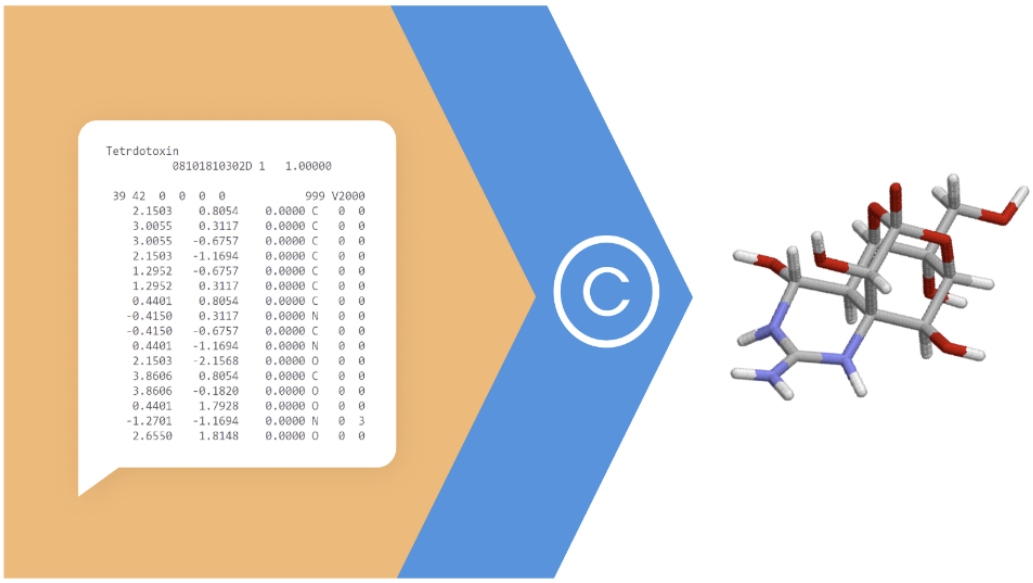
CORINA Classic

Generation of Three-Dimensional Molecular Models

Version 5.0.0

Program Description



Molecular Networks GmbH
August 2023
www.mn-am.com

Molecular Networks GmbH
Neumeyerstraße 28
90411 Nuremberg
Germany

Altamira LLC
470 W Broad St, Unit #5007
Columbus, OH 43215
USA

mn-am.com

(Document version: 5.0.0-2023-08-30)

Content

1 Introducing CORINA Classic

1.1 Objective of CORINA Classic

The three-dimensional structure of a molecule is closely related to a large variety of chemical, physical and biological properties. The need for computer-generated 3D molecular structures has clearly been recognized in drug design and many other areas. Since the number of experimentally determined molecular geometries is limited – over 1,000,000 structures are presently contained in the Cambridge Structural Database (August 2019) [1] compared to more than 96 million of known compounds (*e.g.*, the PubChem database, August 2019) – there is a need for methods to predicting 3D coordinates directly from the constitution of a molecule. Consequently, in the last four decades a number of programs for automatic 2D-to-3D conversion have been reported (for reviews see reference [2]-[3]). Among them, **CORINA** (**CO**o**R**d**INA**tes) **Classic** [2]-[9] automatically generates three-dimensional coordinates for the atoms of a molecule purely from the constitution of a molecule (see Figure 1). The program scope, its reliability, robustness, and speed as well as some special features for handling large rings and metal complexes make it extremely useful for any study or modeling purpose that requires 3D information of the molecules under investigation.

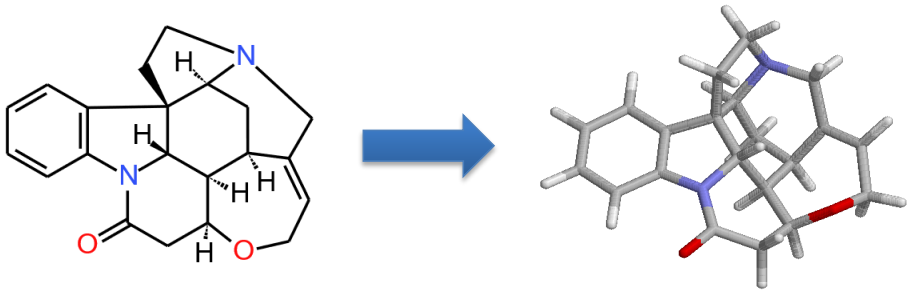


Figure 1 Generating a 3D model from the constitution of a molecule.

1.2 CORINA Classic in Brief

CORINA Classic is a rule- and data-based program system, that automatically generates three-dimensional atom coordinates from the constitution of a molecule as expressed by a connection table or linear code. **CORINA Classic** is powerful, robust, and reliable to convert large databases of millions of compounds.

- **CORINA Classic** is applicable to the entire range of organic chemistry. Structures which can be expressed in a valence bond notation can be processed.

- **CORINA Classic** does not provide any upper limit to the size of the molecule or to the size of ring systems.
- **CORINA Classic** fully considers stereochemical information and generates the defined stereoisomer (if the stereochemistry information is defined properly in the input file).
- **CORINA Classic** processes structures containing atoms with up to six neighbors. Thus, even metal complexes can be processed.
- **CORINA Classic** generates by default one low energy conformation for each input structure. For ring systems consisting of up to nine atoms, multiple conformations can be generated – a useful feature for building flexible 3D databases.
- **CORINA Classic** automatically detects stereocenters (tetrahedral centers, *cis*/*trans* double bonds and stereogenic axes for atropisomerism) and optionally generates all possible stereoisomers. Duplicate isomers, such as *meso* compounds are identified and removed as well as geometrically strained configurations.
- **CORINA Classic** can process a variety of standard file formats for structure input and output such as, *e.g.*, SD/RD V2000/3000 [10], SMILES [11], SYBYL MOL and MOL2 [12], PDB [13], MacroModel [14], Maestro [15],CIF [16], InChI [17] and MOPAC [18] file formats.
- **CORINA Classic** delivers structures of high quality. The RMS deviation of CORINA built models from published X-ray structures is among the best of all commercially available conversion programs.
- **CORINA Classic** is fast (less than 3 ms for small and medium-sized organic molecules on a common x86 Linux workstation), robust and provides excellent conversion rates (99.6%) for the 265,242 structures of the National Cancer Institute (NCI) Open Database [19] without intervention or program crash.
- **CORINA Classic** is general. The PubChem database with more than 91 million compounds has been converted with a conversion rate of 99.7% [20]. Furthermore, the GDB-13 with more than 970 million compounds was converted with a conversion rate of 99.0% [21].
- **CORINA Classic** offers many features to influence the 3D generation process and to clean up and standardize chemical compounds, *e.g.*, addition of lacking or implicitly given hydrogen atoms, neutralization of formal charges, removal of counter ions in salts or small fragments, such as solvents, re-assignment of atom and bond types and many more.
- **CORINA Classic** provides an interface to the ligand-docking program FlexX [22]. During the docking process with FlexX, CORINA generates multiple conformations for the ring systems of the ligand to optimize the steric and electrostatic interactions between the small molecule ligand and the binding site of the protein.

A description of the program scope, range of application and known limitations is given in section 9 on page 108.

2 Release Notes

2.1 CORINA (Full Version)

2.1.1 Version 1.6

CORINA version 1.6 represents a substantial improvement on version 1.5. Both the quality of the results became better, and the program became more flexible. There are five major changes in version 1.6 compared to version 1.5.

1) The input file format SMILES linear notation was added [11].
2) The output file formats SYBYL MOL/MOL2 [12] and Brookhaven Protein Databank PDB [13] were added.
3) The algorithm, which refines atom overlap and close contacts was improved by implementing of a set of rules obtained from a statistical analysis of the conformational preferences of open-chain portions in small molecule crystal structures contained in the Cambridge Structural Database (CSD) [1], [23].
4) A substantial speed-up of almost a factor of 2 was achieved by optimization.
5) The command line options now follow the UNIX command syntax standard.

The quality and speed improvements have been illustrated elsewhere [24]. A side effect of the quality improvements is of course that the resulting 3D structures for a variety of structural classes might have changed.

The changes in the command syntax might cause some portability inconveniences for the user but gave more flexibility for the addition of new options like, *e.g.*, the new input and output file specifications. The old options are no longer valid – the program stops with an error message when recognizing the use of the old syntax.

2.1.2 Version 1.7

CORINA version 1.7 was tailored especially to the database business.

1) Two new driver options **-d flapn** and **sc** were added for generating multiple ring conformations.
2) Two additional PDB output options **-o pdbludi** and **pdbludilabel** allow the generation of fragments for databases interfacing to the *de novo*-design program Ludi [25].

In addition, an exhaustive study on the effect of multiple ring conformations on the performance of flexible 3D pharmacophore searches was performed (see section 12.5 on page 149).

2.1.3 Version 2.0

CORINA 2.0 is now able to interact with the ligand-docking program FlexX [22] as a conformer generator for ring systems (see section 6.12 on page 81). Thus, CORINA ring conformations can be used for flexible ligand docking into a receptor pocket. Changes were mainly made to the file format interfaces and to the ring conformation options.

1) Two new input file formats SYBYL MOL/MOL2 [12] (**-i t=mol** and **mol2**) as required by FlexX were added.
2) A variety of new options were introduced for ring conformations (**-d de**, **timeout** and **flexx**) for tailoring the results for FlexX.

2.1.4 Version 2.1

The following changes and improvements have been implemented.

1) The SMILES interface was made more stable.
2) Three new options **-d ow**, **-d amide** and **-i sdfict** related to the handling of stereochemical information in SD files [10] were added (see section 4.2 on page 30).
3) The most important change concerns the handling of the configuration of amide bonds. In earlier versions, the configuration (*cis* or *trans*) was taken from the 2D drawing in the input file. This behavior must now be switched on explicitly. By default, now the most suitable configuration is taken – in most cases *trans*. Thus, cases with unexpected *cis* amides will no longer be generated.

2.1.5 Version 2.3

The following changes and improvements have been implemented.

1) A new option **-d no3d** allows using CORINA as a file format converter for the supported file formats without generating 3D coordinates.
2) The FlexX interface, the SMILES and SD file interpreter were made more stable.
3) Additional ring conformation patterns for cyclo-octa-1,3-diene were added to the template data file `rings.ctx`.

2.1.6 Version 2.4

The following changes and improvements have been implemented.

1) The data files `stdval.ctx` and `rings.ctx` are now inline – easier installation, less mistakes with different versions.
2) A new driver option **-d 3dst** forces the use of a given 3D configuration instead of the stereochemical descriptors. This might be useful if the stereochemical descriptors are not specified properly but the 3D structure is correct.
3) A new driver option **-d neu** neutralizes formal charges at acids, alcohols, and basic nitrogen atoms by adding or removing protons. Often it is useful to have all molecules of a database in the same protonation state. This option can be used with the option **-d rs** to remove counter-ions from salts.
4) A new driver option **-d ori** orients the generated 3D structure according to the moments of inertia. This might be useful when the structure is directly forwarded to a graphical viewer. The molecule then appears more often in an orientation that shows as much of it as possible on one sight.
5) Some minor problems were fixed in the FlexX and the SD file interfaces with no impact on the 3D generation process.

2.1.7 Version 2.6

The following improvements and changes have been implemented.

1) The file format RDfile [10] was added to the read and write functions of CORINA.
2) In order to provide interfaces to the protein crystallographic and NMR program packages CCP4 [26] and X-PLOR [27] the output file formats CCP4 dictionary file format (**-o dic**), X-PLOR topology (**-o top**) and X-PLOR parameter file format (**-o par**) were added. With the additional options **-o resnam, typchr, dicid**, input files for the CCP4 and X-PLOR program suites can be generated.
3) Atoms with isotopic mass are defined for SD, SMILES and ClearText format [28].
4) The SMILES reader and interpreter are more general now: SMILES strings containing hetero-aromatic rings without explicitly defined hydrogen atoms at the hetero atoms are now tolerated. For example, pyrrole compounds can now be input also as the "incorrect" SMILES n1cccc1 according to the SMILES language definition (correct coding: [nH]1cccc1).
5) The SMILES reader now accepts only one SMILES linear code per line. The SMILES code is expected to be the first string in the line. With the input option **-i smilesname**, all following strings are interpreted as compound name and copied into the corresponding field of the output file. Thus, white or blank spaces within the compound name are now allowed.

6) Non-element symbols, dummy atom types or groups like "X", "R", "Du", "Lp", "D", "T" and "*" are defined for the file formats SD file, SMILES linear code and SYBYL MOL/MOL2. For SMILES linear code the interpretation of dummy atom types or groups must be specified explicitly by using the new input option **-i dummies**.

7) With the new input option **-i csdmol2** specific extensions and information in SYBYL MOL/MOL2 input files, which were generated by the Cambridge Structural Database (CSD) software [1], are written to the output file.

8) A new output option **-o m2l** ("mass to label") copies isotopic mass labels given in the input file into the corresponding atom name field in SYBYL MOL/MOL2 files. Atoms without given mass label remain untouched. The atom name has the format <symbol><mass>. If the corresponding atom is a non-element symbol, the atom name has the format R<mass>. This can be used to create extension points for virtual combinatorial libraries, *e.g.*, as input files for FlexX.

9) A new output option **-o mdldb** creates the additional data fields "<MODEL.SOURCE>", containing information about the program version of CORINA and "<MODEL.CCRATIO>", giving the close contact ratio of the CORINA generated 3D molecular model. This option has been added for compatibility reasons with databases distributed by MDL Information Systems, Inc.

10) A new output option **-o noccat** switches off the automatic conversion of the carbon atom in amidinium-like structures ([NH2+]=CN) to the carbo-cation type SYBYL atom type "C.cat" (N[C+]N). The conversion to this atom type, which is the default, is still strongly recommended.

11) The conformational analysis package for small and medium sized ring systems has been improved: CORINA is now able to generate and to output different ring geometries for ring systems consisting of up to nine ring atoms. In lower program versions, this was limited to a ring size up to eight atoms.

12) The conformational analysis package has been extended to a set of over 900 rules to avoid or eliminate close contacts of non-bonded atom pairs in 3D molecular models. These rules have been derived from a statistical analysis of the conformational preferences of open chain portions in small molecule crystal structures contained in the Cambridge Structural Database (CSD) [1],[23],[29].

13) The new driver option **-d sanpyr** allows the generation of pyramidal nitrogen atoms in sulfonamide groups. The default, which is strongly recommended, is the generation of a planar configuration of the nitrogen atom. The sampling of the "out-of-plane" distances of 1,216 sulfonamide nitrogen atoms as found in the Cambridge Structural Database (CSD) [1] has shown, that the in majority of cases (901 of 1,216 sulfonamides – 74%) an "out-of-plane" distance of less than 0.3 Å is exhibited. Thus, the planar configuration is the preferred geometry compared to the pyramidal configuration.

14) The new driver option **-d newtypes** forces CORINA to generate new atom types in the output file by discarding input atom type and aromaticity information. CORINA can be used for, *e.g.*, assigning new atom types from incorrect input records.

2.1.8 Version 3.0

The following improvements, changes and new features have been implemented.

1) The functionality of the stereoisomer generator STERGEN [30] has been integrated in CORINA. The driver option **-d stergen** forces CORINA to determine all stereocenters in an input structure and to enumerate all possible, but unique stereoisomers. Configurational isomers at tetrahedral coordinated centers as well as at double bonds (*cis*/*trans*) are considered. Duplicate configurations, such as *meso*-compounds are identified and removed. By default (if the driver option **-d stergen** is set), a maximum number of 4 stereocenters are processed and a maximum number of 16 stereoisomeric compounds are generated. However, the driver options **-d msc** and **msi** allow to set a user defined number of stereocenters that should be processed (**msc=<value>**) and to restrict the total number of generated stereoisomers (**msi=<value>**). Stereocenters which have a defined stereochemistry are also processed, unless the driver option **-d preserve** is set which prevents from processing those centers which have a defined stereochemistry (tetrahedral centers and *E*/*Z* double bonds), *i.e.*, a stereochemical descriptor is given in the input structure.

2) In order to provide interfaces to the molecular modeling package MacroModel [14], CORINA now supports the uncompressed MacroModel structure file format (input option **-i t=mmod**) as well as the Maestro file format (input option **-i t=mae**) [15] as new input and output file formats.

3) In addition, the file format CIF (Crystallographic Information File, **-o cif**) [16] supported by a variety of crystallographic program packages, the file format ODB (O Database file format, **-o odb**) [31] to interface to the crystallographic modeling tool O and the file format of the NMR structure calculation program DYANA (**-o dyana**) [32],[33] were added.

4) The input option **-i expandapo** forces CORINA to expand attachment points defined in SD files ("`M  APO`" field in the properties block) into 3D space. The attachment points are added as "artificial" atoms to the connection table (both to the atom and bond list) and 3D coordinates are calculated. Dummy atom types are assigned to the "artificial" atoms, *i.e.*, "`Du`" in SYBYL MOL/MOL2 files, "`*`" (first attachment point) and "`**`" (second attachment point), respectively and "`X`" in PDB files. In addition, the atom names of the attachment point atoms are set to "`R1`" (first attachment point) and "`R2`" (second attachment point), respectively, in the output file for formats that support atom names (*e.g.*, SYBYL MOL2).

5) The combined input and output option **-i/-o xelement** only has an impact if dummy atom types ("`Du`") or element symbols which are unknown SYBYL atom types are defined in SYBYL MOL2 input files. The new input option **-i xelement** forces CORINA to try to derive SYBYL atom types either from the atom names or from the element symbol or to interpret element symbols in order to internally set appropriate atom types for the 3D structure generation process. By default, CORINA outputs dummy atom types ("`Du`") for these atoms. In addition, the new

output option **-o xelement** allows writing the derived SYBYL atom types or the element symbols ("artificial" SYBYL atom types) to the output file. Please use these options carefully and manually check the results, since ambiguous definitions in the input file might lead to misinterpretation or false assignment of atom types.

6) A new output option **-o mdlcompact** restricts the number of output fields in the atom lines of the atom block in SD files (RD files) to the x-, y- and z-coordinates, the atom type (symbol), the mass difference, the atom charge and the stereochemical atom parity (columns 1 through 7 of the atom block). All other fields in the atom lines are omitted, since they contain no data that is mandatory for 3D structure information. The goal is to save disk space (up to 40%) in case of large files containing hundreds of thousand compounds.

7) If stereochemistry information is missing in the input file CORINA assumes reasonable stereochemical descriptors following some implemented rules (see section 12.4 on page 126). The new output options **-o mdl3dparity** forces CORINA to output the stereochemical atom parities in SD and RD files which were calculated and used by CORINA during the 3D structure generation process.

8) If the output file format is set to SYBYL MOL2, the new output option **-o gold** forces CORINA to assign the atom types and bond orders according to the atom and bond type conventions of the docking program GOLD [34],[35] for difficult groups, *i.e.*, functional groups which have more than one canonical form (*e.g.*, guanidinium groups).

9) The new output option **-o fcharges** has only an impact if the output file format is set to SYBYL MOL2 format. Formal atom charges which are given in the input structure are then written to the charge column (column 9) in the corresponding "`@<TRIPOS>ATOM`" data lines of the SYBYL MOL2 output file. In addition, the keyword "`USER_CHARGES`" is set in the "`@<TRIPOS>MOLECULE`" field.

10) By default, if the atom name and the atom type differ, CORINA tries to derive a reasonable atom name from the atom type for PDB, MacroModel and Maestro output files. The new output option **-o keepnames** forces CORINA to keep and to output atom names which are defined in the input file.

11) The new driver option **-d names** consecutively numbers the different conformations (**-d rc**) or stereoisomers (**-d stergen**) in multi-record output files. A counter is added to the compound name of each conformation (*compoundname_c00n, n=1,2,3,...N*; where *N* is the total number of generated conformations) or stereoisomer (*compoundname_i00n, n=1,2,3,...N*; where *N* is the total number of generated stereoisomers). Thus, different conformations or stereoisomers of the same input structure are named uniquely and can be easily distinguished by other program systems or any post-processing software.

12) Several problems in the interfaces to the various input and output file formats as well as in the 3D structure generation routines were fixed. The authors thank all CORINA users who made us aware of these insufficiencies, issues and problems in our software and helped us to make CORINA a more reliable and robust product.

2.1.9 Version 3.1

The following improvements, changes and new features have been implemented.

1) CORINA can now add missing hydrogen atoms and their 3D coordinates to 3D input structures while retaining the input 3D geometry. This can be done by simply combining the driver options **wh** and **no3d** (**-d wh,no3d**). This option can be used to add 3D hydrogen atoms if they are missing in a 3D structure, but the 3D structure of all given atoms should be kept, *e.g.*, for an experimentally determined geometry.

2) A new driver option **-d ringatom=<atom label>** can be used to denote a specific ring system for which multiple ring conformations should be generated, whereas "<atom label>" is the atom label of one atom contained in this specific ring system (requires driver option **-d rc**). This is helpful if multiple ring conformations should be generated only for one specific ring system in an input structure that contains more than one flexible ring system (up to nine ring atoms).

3) The stereoisomer generation module of CORINA can now process double bonds in ring systems having more than 10 ring atoms and output *cis* and *trans*-isomers (driver option **-d stergen**). In addition, para-substituted ring systems are identified as *pseudo* stereocenters to generate di-equatorial and equatorial/axial substituted configurational isomers.

4) For UNIX and Linux systems, CORINA is now available as a library version (shared object). This enables software developers to easily include the CORINA functionalities into their chemoinformatics applications and to call CORINA for 3D structure generation purposes in own source code. The API for input is either a single line formatted SD file or a SMILES linear string. The 3D structures are returned as an SD file (again formatted as a single line with "`\n`" as new line character). In addition, all driver options can be triggered. With this version 3.1, the library version is available for SGI workstations (IRIX 6.5), Sun SPARC stations (Solaris 8) and x86 Linux platforms (kernel 2.4). A version for Microsoft Windows (win32) is currently under development and will be available soon.

5) The support of long compound names is secured, even for file formats which only allow a restricted number of characters for the name field (*e.g.*, SD file, 80 characters). However, CORINA prints a warning message to the trace file when encountering lines in the input file that are too long.

6) The algorithm for orienting 3D structures according to their moments of inertia has been made more reliable.

7) Several changes in the read and write interfaces to SD/RD, SMILES, SYBYL MOL/MOL2, MacroModel and Maestro file formats made the input and output more stable.

8) Improvements and bug fixing in the core algorithms for the handling of ring systems, generation of stereoisomers and usage of system routines increased the conversion rate and decreased CPU times.

2.1.10 Version 3.2

The following improvements, changes and new features have been implemented.

1) A new driver option **-d canon** prevents any atom numbering dependent artifacts in the 3D structure generation process. In some special cases, the 3D structure generation process may be sensitive to the atom numbering in an input structure, *i.e.*, for different atom numberings slightly different conformations may be obtained. This option canonizes the connection table of an input structure internally before the 3D model is built and then uses the reordered atoms for the generation of the 3D coordinates. After the 3D model is generated, the connectivity table is renumbered using the original atom numbering scheme.

2) Structures that cannot be converted by CORINA can now be written to a second output file (error file), if the new driver option **-d errorfile=<value>** (<value> = file name) is used. Currently, only SD and SMILES format are supported. Collecting failed structures in a separate file is valuable for the batch processing of large files.

3) A new input option **-i force3d** forces CORINA to output molecules in SYBYL MOL/MOL2 or PDB file format (these file format (by definition) require 3D coordinates!), even if no 3D coordinates can be generated or are available in the input file. This option is especially useful if CORINA is used only as a file converter with the driver option **-d no3d**. The following two combinations are possible.

-i force3d	Output of records (structures) for which no 3D coordinates can be generated in SYBYL MOL/MOL2 or PDB file format
-i force3d -d no3d	File conversion to SYBYL MOL/MOL2 or PDB file format if no 3D coordinates are present in the input file

4) The new driver option **-d symoff** ignores the symmetry filtering during the ring conformation generation process. This option is especially helpful if CORINA is used in the interface mode for the docking program FlexX in order to generate multiple ring conformations of the ligands (see also sections 2.1.13 and 6.12). CORINA will only process the cyclic fragments originating from the larger ligand molecules. As acyclic atoms are removed by FlexX before sending fragments to CORINA, a higher symmetry might be perceived in the fragment compared to the symmetry of the entire molecule. Consequently, fewer conformations may be generated than required.

5) A new driver option **-d ampax** amplifies axial substituent energy contributions during the ring conformation analysis. In some complex cases, the energy differences between axial and equatorial arrangements may become too small with the default parameters, *i.e.*, as for the amide ligand in the benzodiazepine compound "c12ccccc1NC(=O)C(NC(=O)C)N=C2".

6) A new output file option **-o lname** allows the user to write out compound names to the SD name field and to exceed the 80 characters limit in the header block.

7) **Note.** Per definition, the name field in the header block of an SD file may not be larger than 80 characters [10].

8) The default bond angles for oxygen and nitrogen atoms in ethers and amines (SYBYL atom types "O.3" and "N.3") have been changed to new values based on an analysis of drug-like molecules from the CSD. For sp^3 hybridized oxygen atoms, the bond angles range from 114° to 134° depending on the two adjacent atoms. The bond angles of sp^3 hybridized nitrogen atoms in amines are set to 111°.

9) The SMILES interpreter now supports the OpenEye extensions for hybridization states. The following SMILES patterns for atom primitives are allowed:
   ^3 for sp^3 hybridized atoms
   ^2 for sp^2 hybridized atoms
   ^1 or ^ for sp hybridized atoms
   For example, the SMILES "c1ccccc1[NH2^2]" will result in a planar amine group, whereas in "c1ccccc1[NH2^3]" the amine group will exhibit a pyramidal geometry in the generated 3D structure.

10) The algorithm to detect duplicate stereoisomers that are generated using the driver option **-d stergen** has been revised. The detection is now based on an improved hash coding technique. Originally, the 32bit algorithm could produce equivalent hash code for different isomers. The new implementation based on 64bit generates unique hash codes in more cases than with the 32bit implementation.

11) For Microsoft Windows platforms (2000/XP) a static library of CORINA is now available. This enables software developers to easily include the CORINA functionalities into their chemoinformatics applications and to call CORINA for 3D structure generation purposes in their own source code also on Windows platforms. The API guidelines require the input to be either a single line formatted SD file (with the two text characters "\n" representing each new line character) or alternatively, a SMILES string. The 3D structures are returned as an SD file (again formatted as a single line with "\n" as new line character). In addition, all driver options can be triggered.

12) Several improvements and bug fixes in the routines for the handling the MacroModel file format, the generation of stereoisomers, the orientation of 3D structures, the usage of system functions and the interface to the library versions (Windows and Linux/UNIX) were made to CORINA to address general stability and reliability issues.

13) Section 7 "Error Messages" and section 8 "Warning Messages" of this manual were revised. In addition, the new section 5 "Use Cases of CORINA" is introduced that lists some use cases for CORINA. This section of the manual will be updated every time the software is revised to add contributions from new and experienced users. Comments and suggestions can be sent to our support team at "support@mn-am.com" with a subject line header "Use cases for CORINA Classic".

2.1.11 Version 3.4

The following improvements, changes and new features have been implemented.

1) The new input option **-i sdfi2c=<value>** ("sdf item-to-comment") can be used to copy the contents of the SD data field named <value> (data header) in an SD input file to a comment line in the output file, *e.g.*, to the 3rd line in the header block of an SD file. Thus, the information from a data field in an SD file can be transferred to any file format which supports a comment line (such as SD, SYBYL MOL2, PDB).

   For example, the option **-i sdfi2c=MySDdataField** copies the data entry "`Hello World`" of the SD data field

   ```
   >  <MySDFdataField>
   Hello World
   ```

   to the comment line of the SD output file (3rd line in the header block):

   ```
   ethane
   ABC       06270620222D 0   0.00000     0.000004
   Hello World
     8  7  0  0  0  0  0  0  0  0  1 V2000
   ```

   Note. The original SD data field of the input SD file is preserved and copied to the output SD file. The size of the comment line may not exceed 80 characters.

2) The new output option **-o sdfc2i=<value>** ("sdf comment-to-item") can be used to copy the comment line of an input file, *i.e.*, 3rd line in the header block of an SD file, to a newly generated SD data field named <value> (data header) in the SD output file. Thus, a comment line from file formats which support comments (such as SD or SYBYL MOL2) can be copied to a data field of an SD output file.

   For example, the option **-o sdfc2i=myNewSDdataField** copies the comment line "`Hello World`" of the SD input file

   ```
   ethane
   ABC       06270620222D 0   0.00000     0.000004
   Hello World
     8  7  0  0  0  0  0  0  0  0  1 V2000
   ```

   to the data field `myNewSDdataField` in the SDF output file:

   ```
   >  <myNewSDdataField>
   Hello World
   ```

   Note. The original SD file comment line is overwritten by the comment produced by CORINA. The size of the comment line may not exceed 80 characters.

3) The new output option **-o pascom** ("pass comment") passes comment lines between file formats which support comments (*e.g.*, SD, SYBYL MOL2 or PDB). In the case of SD files as input and output this option preserves the comment line in the header block (3rd line) of the input SD file and writes it to the comment line of the SD output file. By default, CORINA overwrites the comment line of the input SD file with information about the program version. This option prevents

overwriting and passes the comment line of the input to the output file.

4) The handling of large molecules was changed. In general, CORINA does not have any predefined limitations regarding the number of atoms or bonds of an input structure that should be converted into 3D (please be aware of limitations due to the system architecture of the machine on which CORINA runs, 32 and 64bit). However, CORINA has been designed to process small to medium sized organic (typically "drug-like") molecules. The larger a molecule gets the more the intra-molecular interactions gain in importance influencing the secondary structure of a molecule. CORINA can model these interactions only to a limited extend and, therefore, is not able to correctly predict 3D structures of polymers and biopolymers such as proteins, enzymes, or nucleic acids. For this reason, the following changes and improvements have been implemented.

   - By default, the maximum number of atoms and bonds of an input structure is now limited to 999. This default limitation can be extended by the new driver option **-d maxat=<value>**. For example, the option **-d maxat=1001** extends this limit to a maximum of 1001 atoms and bonds.

     Note. Some file formats are limited (by definition) to a certain number of atom and bonds, *e.g.*, SD file is limited to 999. This driver option will not circumvent any of these limitations.
   - The functionality of file format conversion (driver option **-d no3d**) is not affected by this limitation. In addition, all file format interfaces that are commonly used for macromolecular structures (*e.g.*, PDB, MacroModel, Maestro, CIF and SYBYL MOL2) have been carefully reviewed to guarantee a proper file format conversion.

5) The stereoisomer module (driver option **-d stergen**) has been improved.

   - Input 3D coordinates which may define the configuration at a stereocenter are now ignored if no stereochemical descriptors (wedged bonds or parity flags) are set. This is also the case if the additional driver option **-d preserve** is set to preserve defined stereocenters.
   - The handling of spiro compounds and the hash-coding algorithm have been further improved for avoiding the generation of duplicate stereoisomers.

6) New structural parameters (standard bond lengths and angles) have been added for the following atom types and substance classes according to some analysis of X-ray structures.

   - New standard bond angles for S(sp^3) (103°), P(sp^3) (102°) and Se(sp^3) (101°) have been added.
   - The standard value for the "C-S-C" bond angle in thiophenes (s1cccc1) has been set to 91°.

7) Several improvements in the file format interfaces to SD/RD (read/write), SMILES (read), SYBYL MOL/MOL2 (read/write), MacroModel (write) and Maestro (write) file formats made the read and write routines more stable and reliable.

8) The new driver option **-d ist** ("ignore stereochemical information") forces CORINA to ignore any stereochemical information given in the input file including parity flags, wedge bond symbols and definitions on *cis*/*trans* double bonds (*e.g.*, via 2D coordinates in SD files or "/ /" and "/ \" definitions in SMILES files). This option is helpful if the user is aware of stereochemical definitions in the input files that are geometrically forbidden (see *e.g.*, *i*,*o*-norbonane in Figure 31 on page 133). Usually, CORINA will reject structures with geometrically forbidden stereochemical definitions. However, if the driver option **-d ist** is set, CORINA will generate a geometrically possible isomer.

9) **Note.** Please use this option with care as all defined stereochemical information will be ignored.

10) The new output file option **-o hlabel** labels hydrogen atoms separately from the heavy atoms in PDB and CIF output files. By default, CORINA uses the same atom counter for heavy and hydrogen atoms. This option forces CORINA to start the counter at "0" for labeling the hydrogen atoms, independently of the counter for the heavy atoms and to allow for numbering of atoms in molecules with more than 99 atoms (including hydrogen atoms).

11) The new driver option **-d planil** forces a planar geometry at anilinic-type ring nitrogen atoms ("c1ccccc1N(C2...C2"), see also section 12.5.2 at page 150).

12) The new output file option **-o flexrta** influences the REFMAC restraints of torsion angles in aliphatic ring systems in CIF output files. With this output option, all torsions are set to "`var`" (variable) with a period of "`3`" and a standard deviation "`esd`" of "`20`" degree in the loop "`chem_comp_tor`".

2.1.12 Version 3.6

The following improvements, changes and new features have been implemented.

1) The driver option **-d i3dst** forces CORINA to ignore any stereochemical information that can be derived from a 3D structure provided in the input file. By default, CORINA uses the stereochemical descriptors (*e.g.*, wedge symbols or parity flags) provided in the input file to generate the correct configuration of a molecule. If no stereochemical descriptors are provided in the input file, but 3D atom coordinates are given, CORINA can calculate the stereochemical information from these 3D coordinates. With this option, the calculation of stereochemical descriptors from input 3D coordinates is suppressed.

2) The driver option **-d sanpyr** forces CORINA to generate pyramidal nitrogen atoms in sulfonamide groups. The default, which is strongly recommended, is the generation of a planar configuration of the nitrogen atom. The sampling of the "out-of-plane" distances of 1,216 sulfonamide nitrogen atoms as found in the Cambridge Structural Database (CSD) [1] has shown, that the in majority of cases (901 of 1,216 sulfonamides – 74%) an "out-of-plane" distance of less than 0.3 Å is

exhibited. Thus, the planar configuration is the preferred geometry compared to the pyramidal configuration.

3) The driver option **-d neu** now neutralizes negatively charged nitrogen atoms in sulfonamide groups, i.e., "S=O(=O)[N-]" is neutralized to "S=O(=O)[N]".

4) The preparation of version 3.6 included extensive code reviewing, testing, and fixing of some issues in order to make CORINA more robust and reliable. The major changes and improvements apply to the following functionalities.

   - Ring generation
   - Torsion angle optimization
   - SD input and output interface (input option **-d sdf**, output option **-o t=sdf**)
   - CIF output interface (output option **-o t=cif**)
   - Orientation of generated 3D structures according to principal moments of inertia (driver option **-d ori**)
   - Interface to library version (API)

   In addition, the changes and improvements secured the 32/64bit compatibility (x86 Linux) of CORINA.

2.1.13 CORINA Classic Version 4.0

The following improvements, changes and new features have been implemented.

1) The SD V3000 file format [10] is now supported as input and output. By default, CORINA reads SD V2000, SD V3000 or mixed files, that contain both V2000 and V3000 formatted records and writes the converted structure records in the same format as provided in the input file.

   - The default behavior as described above is applied with or without the input and output options **-i t=sdf** and **-o t=sdf**, respectively
   - The new output option **-o t=sdf2** forces the output of SD V2000 formatted files and any SD V3000 record in the input file is converted into SD V2000 formatted record.
   - The new output option **-o t=sdf3** forces the output of SD V3000 formatted files and any SD V2000 record in the input file is converted into SD V3000 formatted record.
   - **Note.** SD V2000 supports chemical structures with a maximum of 999 atoms or bonds. If this limit is exceeded during the structure processing (*e.g.*, because of the addition and writing of hydrogen atoms, driver option **-d wh**), no structure record is written to a V2000 SD output file and the next record is processed. With the driver option **-d maxat=<value>**, the maximum number of atoms can be set to "value" (*e.g.*, >1,000). In the case that the number of

atoms is higher than 999, but lower than "value", a V3000 SD record is written to the output file.

2) The RD V3000 file format [10] is now supported as input and output. With the input option **-i t=rdf**, CORINA reads RD V2000, RD V3000 or mixed files, that contain both V2000 and V3000 formatted records.

   - The output option **-o t=rdf** writes the converted structure and reaction records in the same format as provided in the input file, *i.e.*, as V2000, V3000 or mixed records.
   - The new output option **-o t=rdf2** forces the output of RD V2000 formatted files and any RD V3000 record in the input file is converted into RD V2000 formatted record.
   - The new output option **-o t=rdf3** forces the output of RD V3000 formatted files and any RD V2000 record in the input file is converted into RD V3000 formatted record.

3) Related to the SD/RD V3000 file formats, the enhanced stereo-chemical representation and handling provided by these file formats is also supported [10]. The V3000 stereo-chemical identifiers "ABSOLUTE" (absolute stereochemistry, group "STEABS"), "OR" (relative stereochemistry, group "STEREL*n*") and "AND" (racemic representation, group "STERAC*n*") are supported.

   - Without the stereoisomer generation option (driver option **-d stergen**), all stereocenters are processed according to the defined stereochemical descriptor. If no stereochemical descriptor is defined for a chiral center, CORINA has to assume an arbitrary configuration.
   - The driver option **stergen** generates all possible stereoisomers, regardless of if the configuration of a stereocenter ("CFG") is defined or not or if an atom belongs to a stereo-chemical group ("STEABS", "STEREL*n*" and "STERAC*n*").
   - With the additional driver option **preserve** (*i.e.*, **-d stergen,preserve**), all stereocenters that have a defined stereochemical configuration ("CFG" flag) are not permuted, *i.e.*, only a single isomer is generated for the respective stereocenter. Crossed double bonds (SD input, explicit undefined stereochemistry) are always permuted even if the option **preserve** is used.
   - With the additional driver option **v3000** (*i.e.*, **-d stergen,v3000**), all stereocenters that belong to one of the stereo-chemical groups "STEREL*n*" and "STERAC*n*" are permuted according to their definition of their relative or racemic representation. Stereocenters that belong to the group "STERABS" are not permuted.
   - If both options **preserve** and **v3000** are set (*i.e.*, **-d stergen,preserve,v3000**), only those stereocenters are permuted that are undefined (no "CFG" flag) or do belong to any of the stereo-chemical groups "STEREL*n*", or "STERAC*n*".
   - **Note.** The driver option **v3000** prevails over the driver option **preserve**. The combination of these two driver options is useful for input files with

structures that contain stereocenters with and without constraints by V3000 stereo-chemical groups.

4) The new driver option **chiralflag** is available for the stereoisomer generation option (driver option **-d stergen**) when the input file format is SD/RD V2000/V3000 [10]. The chiral flag is specified in the counts line in the header block of the SD/RD file as shown below (numbers underlined and bold).

V2000: `··8··7··0··0··1··0··0··0··0··0··1·V2000`

V3000: `M··V30·COUNTS·8·7·0·0·1`

A set chiral flag (*i.e.*, a value of "1") will enforce the following behavior.

- With the additional driver option **chiralflag** (*i.e.*, **-d stergen,chiralflag**), all stereocenters that have a defined stereochemical configuration are not permuted, *i.e.*, only a single isomer is generated for the respective stereocenter if the chiral flag is set in the input structure (connection table). If the chiral flag is not set, all stereoisomers are enumerated.

  Note. The driver option **preserve** prevails over the driver option **chiralflag**.

- If both options **v3000** and **chiralflag** are set (*i.e.*, **-d stergen,v3000,chiralflag**) and the chiral flag is set in the input structure, the option **chiralflag** applies only to defined stereocenters that are not part of a V3000 stereochemical group.

5) The MOPAC Cartesian file format [18] is now supported as output file format. The new output file option **-o mopacxyz** forces CORINA to write out the converted chemical structure formatted for input into the semi-empirical QM software package MOPAC. The MOPAC output interface of CORINA provides the following additional options and limitations.

- The new output file option **-o mopackeys=<value>** allows for defining a list of MOPAC keywords that are copied to the keyword line (1st line) in the MOPAC Cartesian output file.
- **Note.** The individual keywords have to be separated by a white space (blank) and that the entire list has to be quoted (single – `'` – or double quotes – `"` – , respectively), *e.g.*, "`1SCF BONDS ESP VECTORS`".
- The output file option **-o mopacaddchg** automatically adds the MOPAC keyword "`CHARGE=n`" to the keyword line of the MOPAC Cartesian output file according to the formal charge of the input molecule. If the user provides the keyword "`CHARGE`" using the option **-o mopackeys=<value>** (see above), the user provided value is adjusted to the actual formal charge of the molecule or the entire keyword is removed, if no formal charge is encountered (*e.g.*, because the driver option **-d neu** to neutralize specific formal charges has been switched on by the user).

- The output file option **-o mopacoptflag=<value>** allows for defining the MOPAC optimization flag which is used in the MOPAC Cartesian output file. The given optimization flag "**value**" is used for each Cartesian coordinate.
- All MOPAC Cartesian file format related output file options also work for file format conversion (driver option **-d no3d**).
- **Note.** If no 2D coordinates are given in the input file, the MOPAC Cartesian file cannot be written out (3D only file format).
- **Note.** MOPAC accepts input structures with Cartesian coordinates only for molecules having more than two atoms.

6) The standard InChI (IUPAC International Chemical Identifier) format is now supported as input file format [17]. The new input file option **-i inchi** expects input structures in InChI format. CORINA uses the InChI Software version 1.05 (January 2017, copyright (C) IUPAC and InChI Trust Limited) under the IUPAC/InChI-Trust License No. 1.0 ("www.inchi-trust.org") to support standard InChI (IUPAC International Chemical Identifier) as input file format.

7) The new output option **-o split** allows for splitting a multi-record input file into single-record output files, regardless of the input and output file format. The output files have an automatically inserted, consecutive number in the file name that is provided by the user. For instance, if the output file name provided by the user is "`output.sdf`" the split output file names have the format "`output.001.sdf`", "`output.002.sdf`", …, "`output.999.sdf`".

8) The new output option **-o splitn0=<value>** also allows for splitting a multi-record input file into single-record output files, regardless of the input and output file format. However, with "**value**" the number of leading "`0`" at the automatically inserted, consecutive number in the output file name can be set. For instance, if the output file name provided by the user is "`output.sdf`" and value is set to "`4`" (*i.e.*, **-o splitn0=4**), the split output file names have the format "`output.00001.sdf`", "`output.00002.sdf`", …, "`output.99999.sdf`".

9) The new output option **-o mdlbond4** sets all aromatic bonds in the output file to the bond type "`4`" (SD query option in bond block).

   Note. This option only has an impact on SD or RD formatted output files.

10) Two new input options **-i sc#=<value>** (structure column number) and **-i nc#=<value>** (name column number) can be used for multi-column SMILES and InChI input files to specify the column number in the input file where the chemical structure and chemical name can be found. For instance, for a file containing SMILES exported from the PubChem collection, the options **-i t=smiles,sc#=2,nc#=1** will process the exported SMILES file correctly (SMILES in column 2 and name in column 1).

11) A speed-up of about 15% was achieved by optimizing CORINA on 64bit platforms.

2.1.14 CORINA Classic Version 4.1.0

The following improvements and changes have been implemented in version 4.1.0.

1) The internal canonization of the connection table of an input structure (former driver option **-d canon**) is used by default. Therefore, the driver option **-d canon** is not available anymore. In addition, the algorithm for the canonization has been optimized. Of course, before the final 3D structure is written to the output file, the original atom (and bond) numbering is restored, *i.e.*, the original atom and bond numbering from the input structure is preserved in the output 3D structure.

2) An inconsistency in the stereoisomer generation module has been detected and fixed. The combination of the driver options to write out implicit and added hydrogen atoms (**-d wh**) and the generation of stereoisomers (**-d stergen**) did not enumerate all possible stereoisomers.

3) On Windows platforms, the optional suppression of writing out the trace/log output (trace file option **-t n**) has been fixed.

4) A series of robustness tests have been performed using large and diverse datasets of chemical compounds. These tests revealed some issues in the gradient optimizer (pseudo force field) when processing complex ring systems and in the algorithm for the arrangement of aromatic bonds in complex fused aromatic ring system (nano materials).

   For the robustness tests, two publicly available databases have been used.

   - The PubChem database ("pubchem.ncbi.nlm.nih.gov"), [20]) with 91 million structures was converted with an error rate of 0.4%.
   - The GDB13 database ("http://gdb.unibe.ch/downloads"), [21]), a virtual library of 971 million small organic molecules up to 13 atoms (C, N, O, S, Cl) generated following simple chemical stability and synthetic feasibility rules, was converted with an error rate of 1%.

5) The improvements described above have the following impacts.

   - CORINA Classic is now deterministic for a broader range of chemical compounds. Any atom numbering dependent artifacts in the 3D structure generation process for some special cases should not occur anymore.
   - The conversion rate has increased compared to version 4.0 and version 3.6. For the Open NCI Database (Release 4, May 2012, "https://cactus.nci.nih.gov/download/nci", total of 265,242 compounds [19]), 403 of the 1,492, which could not be converted by version 4.0, can now be converted by version 4.1.0. This corresponds to an increase of the conversion rate of 27%.
   - The CPU time has slightly increased compared to version 4.0. For the Open NCI Database, the average CPU time increased from 2.7 to 3.1 ms/mol (from 710 to 820 s for the full set of 265,2424 compounds on a common Linux workstation, Intel i7, 3.4 GHz).

2.1.15 CORINA Classic Version 4.2.0

The following new features, options and improvements have been implemented in version 4.2.0.

1) The handling of stereochemistry and the stereoisomer generation module have been improved and extended to also consider axial chirality. If the new driver option **-d axchir** is set, potential stereogenic axis which can cause atropisomerism are detected (see also section 12.4.2.3 on page 130). Atropisomers which provide the relevant stereo descriptors in the input connection table will generate the 3D model of the coded isomer. For the stereoisomer generation module the driver option **-d axchir** (**-d stergen,axchir**) will enumerate all possible atropisomers (in addition to other potential isomers originating from tetrahedral centers and *cis*/*trans* double bonds).

   Note. By default, a maximum of four stereogenic centers (including tetrahedral, cis/trans double bonds and axial centers) are processed and enumerated by the stereoisomer generation module. To process more stereogenic centers, the additional options **msc=<value>** and **msi=<value>** have to be used.

2) The original mode of behavior when reading in and processing SMILES files has been restored. In a SMILES input file, versions before CORINA Classic 4 expected the SMILES string in the very first column and interpreted the information in the second column as the name of the chemical compound. This mode of behavior is the default if neither of the two input file options **-i sc#=<value>** or **-i nc#=<value>** is defined at the command line. This default behavior is even extended as any third column in a SMILES or InChI input file is interpreted as a comment and can be copied to the respective comment field in the output file (please see also below).

3) The two new input file options **-i scn=<value>** or **-i ncn=<value>** can be used instead of the options **-i sc#=<value>** or **-i nc#=<value>**. The new options trigger the same functionality and have been introduced to prevent any issues when using these options in a CORINA Classic command line which is executed in a script (the character "#" might be a special character, e.g., as in a shell script).

4) The new input file option **-i cc#=<value>** or **-i ccn=<value>**, respectively, can be used for multi-column SMILES and InChI input files to specify the column number in the input file in which a comment is stored. For instance, for a input SMILES file containing the SMILES in the first column, the compound name in the second and a comment in column number "*n*", the option sequence **-i t=smiles,ccn=*n*** will use the content of column "*n*" as a comment for each structure.

   To write (or copy) the comment to the output file, the output option **-o pascom** needs to be used. Of course, the output file format needs to support comment lines or fields, such as SD, RD, SYBYL MOL2, or PDB.

   For SD (or RD) output files, it is recommended to always use this option in combination with the output file option **-o lname** to not risk truncating SMILES strings after 80 characters (limitation of the length of an SD comment line by the

file format specifications).

Note. For the input options **sc#**/**scn**, **nc#**/**ncn** and **cc#**/**ccn** described above, the values "<value>" needs to be integral numbers. If a value of "0" (zero) is set, the copying of any information (name or comment) from the input to the output file is suppressed.

5) The new input file option **-o sep=<value>** can be used to specify the column separator used in a multi-column SMILES or InChI input file. By default, blank, white or tab spaces tab are considered as column separators. Any items with blanks or spaces that should be considered as a single term, must be double double-quoted (*e.g.*, **"**`Mesidine hydrochloride`**"**). However, this default behavior fails, if an entry in a column contains, *e.g.*, a blank in a compound name such as "`Mesidine hydrochloride`" but is not double-quoted in the input file. This option provides the flexibility to also have items with blank, white and tab spaces, if a different column separator is used, *e.g.*, ";".

   Note. Special characters (Linux and UNIX shell specific) have to be double-quoted (") on the command line, if they are used as column separators in the input file, *e.g.*, **-i t=smiles,sep="\"**.

   Note. A tabulator as a separator can be defined as **"**`\t`**"**, `TAB` or `tab`.

   Note. Only a single character is allowed as a separator.

6) The new output file option **-o sdfs2i=<value>** ("sdf-structure-to-item") can be used to copy the SMILES or InChI string from a SMILES or InChI input file into the SD data field named "<value>" (data header) in an SD output file.

   For example, the option **-o sdfs2i=myStructure** copies the SMILES string of the SMILES input file into the newly generated SD data field "`myStructure`".

7) The new output file option **-o chirvol** sets the chiral volume in CIF output files to "`both`" for pseudo-stereogenic centers which are introduced by flexible ring systems. In case of super stereogenic centers (such as six-membered rings with axial/equatorial substituents) this option sets the values in the "`chiral volume`" loop to "`both`" (instead of either "`positive`" or "`negative`". Real tetrahedral stereogenic centers are not affected by this option.

8) The new output file option **-o coplan** sets stricter values for coplanar methoxy group substituents at aromatic systems in CIF output files. Methoxy group substituents at aromatic systems are usually coplanar to the aromatic ring if there is no steric crowding that forces the methyl group out of the plane (depending on the substitution pattern at the aromatic ring). With this option coplanar methoxy groups are listed in the "`plane`" loop in CIF files and in the "`torsion`" loop an "`esd`" value (tolerance for torsion angle) of only 10º and a "`period`" of 2 is applied. This guarantees a stable coplanarity of the respective methoxy group also during the REFMAC refinement.

9) General improvements, especially in the ring generation algorithms increased the performance in terms of speed, conversion rate and robustness.

10) The improvements described above have the following impacts.
    - The conversion rate has slightly improved compared to version 4.1.0. For the Open NCI Database (Release 4, May 2012, total of 265,242 compounds, [19]), the number of compounds that CORINA Classic version 4.2.0 is not able to convert, decreases from 1,089 to 1,039 (-4.6%).
    - The CPU time has slightly decreased compared to version 4.1.0. For the Open NCI Database, the average CPU time decreased from 3.0 to 2.3 ms/mol (from 790 to 607 s for the full set of 265,242 compounds on a common x86-64 Linux workstation, Intel i7, 3.4 GHz).
    - The PubChem database [20] with 96 million structures is converted with an error rate of 0.3%.

2.1.16 CORINA Classic Version 4.3.0

The following new features, options and improvements have been implemented in version 4.3.0.

1) The stereoisomer generation module has been improved and extended to consider additional cases of relative stereochemistry for racemic mixtures. For the generation of stereoisomers, the new driver option **-d preserverel** (*i.e.*, **-d stergen,preserverel**) enables CORINA to interpret input structures with more than one fully defined stereocenter as racemic mixtures. For SD V2000/3000 input files, it is required that the chiral flag of the record is not true (value of "0") and/or is not considered (driver option **chiralflag** not set). In this case, fully defined stereocenters are not interpreted as absolute stereocenters. This enables the generation of the respective pair of enantiomers, *i.e.*, the racemate. For example, this option is helpful if a compound collection or library contains molecules which are either racemic mixtures or pure enantiomers defined through the respective stereo descriptors and the chiral flag. This also works for SD V2000 input files or for SD V3000 input files even if no stereogenic group ("`STEREL`*n*" and "`STERAC`*n*") is used. Information about axial chirality is processed as well by the new option **preserverel** and used to generate the respective isomers (provided that the additional options **stergen** and **axchir** are set).

   Note. No diastereomeric compounds are generated with this combination of options and settings provided that all stereocenters are fully defined in the input file. Otherwise, no information about the relative stereochemistry can be derived.

   Note. It is recommended to read the section 12.4 "Handling of Stereochemistry" starting on page 126 which provides an overview of the impact on the results when using the different options available for the stereoisomer generation.

2) The driver option **-d chiralflag** (*i.e.*, **-d stergen,chiralflag**) is now also considered when interpreting information about axial chirality in an SD input file. An applied chiral flag of "1" interprets the defined axial chirality as an absolute configuration.

3) The new output file option **-o xlabel** offers an extended atom labeling schema for PDB, CIF, MacroModel and Maestro output files. Like the output file option **hlabel** which labels hydrogen atoms separately from heavy atoms, each atom type (chemical element) is labeled separately with the new option **xlabel** in the respective atom name field (*e.g.*, columns 13 to 16 in the HETATM or ATOM line of a PDB file). In addition, the new schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record. Below is an example extract for a PDB file.

```
HEADER    UNK                                      19-07-17    1UNK
COMPND    My Molecule
REMARK    1 My File
HETATM    1  C1  UNK     1     -12.218  10.417   7.842  1.00 20.00
HETATM    2  C2  UNK     1     -12.388  10.684   6.345  1.00 20.00
…
HETATM   10 BR1  UNK     1     -13.718  11.400   6.102  1.00 20.00
…
HETATM  169  C1G UNK     1      -9.302   0.623  -1.230  1.00 20.00
HETATM  170  O37 UNK     1     -10.451   0.255  -1.115  1.00 20.00
HETATM  171  C1H UNK     1      -8.805   1.139   2.556  1.00 20.00
…
```

4) The new output file option **-o pdbelement** adds the element symbol to column 77 and 78 of each HETATM or ATOM line in a PDB output file. This procedure is helpful to store the correct element symbol information in a PDB file if one of the above-mentioned labeling schemes (output file options **hlabel** and **xlabel**) is used. Element symbols in column 77 and 78 of a HETATM or ATOM line are conforming with the PDB V2.3 specifications from July 1998. Below is an example extract for a PDB file including the element symbols.

```
HEADER    UNK                                      19-07-17    1UNK
COMPND    My Molecule
REMARK    1 My File
HETATM    1  C1  UNK     1     -12.218  10.417   7.842  1.00 20.00           C
HETATM    2  C2  UNK     1     -12.388  10.684   6.345  1.00 20.00           C
…
HETATM   10 BR1  UNK     1     -13.718  11.400   6.102  1.00 20.00          Br
…
HETATM  169  C1G UNK     1      -9.302   0.623  -1.230  1.00 20.00           C
HETATM  170  O37 UNK     1     -10.451   0.255  -1.115  1.00 20.00           O
HETATM  171  C1H UNK     1      -8.805   1.139   2.556  1.00 20.00           C
…
```

5) The output file option **-o coplan** was introduced in version 4.2.0 (see above) and only affected the parameters in the loops "`plane`" and "`torsion`" in CIF output files for methoxy groups which are coplanar substituents to an aromatic ring. The option **coplan** has now been extended to any coplanar alkoxy group and will adjust the relevant parameters in a CIF output file accordingly.

6) Fixing some minor issues, such as an improved output of the trace file when reporting about the stereoisomer generation process, made the product more user-friendly.

2.1.17 CORINA Classic Version 4.4.0

The following new features, options, and improvements have been implemented in version 4.4.0.

- The stereoisomer generation module has been extended to handle defined *E*/*Z* double bonds separately. The new driver option **preserveez** (*i.e.*, **-d stergen,preserveez**, preserve *E*/*Z* double bonds) preserve the configuration at defined *E*/*Z* double bonds during the enumeration of possible stereoisomers. While the option **preserve** keeps both, defined tetrahedral and *cis*/*trans* double bonds, fixed, the option **preserveez** only enumerates stereoisomers for double bonds with an undefined configuration. Tetrahedral stereocenters are not affected at all. This option is helpful, if all stereocenters except defined *E*/*Z* double bonds should be permuted.
- The new output file option **-o esterplane** adds the five atoms "CC(=O)OC" of an ester groups to the loop "`plane`" and adjusts the parameters in the loop "`torsion`" in CIF output files, accordingly. Ester groups "CC(=O)OC" are usually planar, *i.e.*, the five denoted atoms are positioned in a plane, if there is no steric hindrance that forces one or more atoms out of the plane. With this option the five atoms of an ester group are added to the "`plane`" loop in CIF output files. Furthermore, in the "`torsion`" loop a value for the parameter "`value_angle`" of 180°, a value for the parameter "`value_angle_esd_esd`" of 0° and a value for the parameter "`period`" of 0 is applied for the carbon-oxygen (C-O) single bond of the ester group, if the torsion angle pattern "C-C-O-C" is used in the loop "`torsion`". If the torsion angle pattern "(0=)COC " is used in the loop "`torsion`", a value for the parameter "`value_angle`" of 0°, a value for the parameter "`value_angle_esd_esd`" of 0° and a value for the parameter "`period`" of 0 is set for the carbon-oxygen (C-O) single bond of the ester group. This guarantees a stable planarity of an ester group also during the REFMAC refinement.
- CORINA now recognizes prochiral centers if the input structure has 3D coordinates. For instance, this ensures that the hydrogen atoms at prochiral carbon atoms maintain the same relative 3D position as in the 3D structure of the input file. This option is useful for the interpretation of ^{1}H NMR spectra.

  Note. CORINA needs to assume an arbitrary configuration for prochiral substituents if no 3D atomic coordinates are provided in the input structure.
- A native version for x86-64 RHEL7 and CentOS7 Linux platforms is available since this version 4.4.0. The version for x86 RHEL5 platforms is not supported anymore.
- Section 12.4 on "Handling of Stereochemistry" has been updated reflecting the new option(s) for stereoisomer enumeration and to resolve some ambiguities.
- Fixing and correcting some minor issues made the application more reliable and user-friendly.

2.1.18 CORINA Classic Version 5.0.0

The development of version 5.0.0 focused on the integration capabilities into new or existing chemoinformatics and computational chemistry applications and environments. The following new features, options, and improvements have been implemented.

- CORINA Classic is now available as a module for Python environments. The Python module supports x86-64 Linux platforms (RHEL7) and Python environments of versions 3.8, 3.9, 3.10, and 3.11. This empowers developers to integrate CORINA Classic into their own applications, workflows, scripts, and various ecosystems using the familiar Python syntax. This also includes the usage of CORINA Classic directly in Python tool kits for machine learning and artificial intelligence applications in chemoinformatics projects. Interactive computation support for Jupyter notebooks is also provided. For further information, please see section "4.3 Application Programming Interface (API) for Python Programming/Scripting Language" on page 61 of this document.
- The Python module supports the full set of the available command line options of the standalone version of CORINA Classic including all available input and output file formats, besides the following limitations.
    - The size of input file is limited to the size of available memory.
    - The output of trace files is not supported. However, for the Python library, the trace information is written to the Python variable "`buffer.log`" but misses the overall performance and run statistics at the end of the trace output.
    - The output of not converted structures into a separate output file (command line option "**-d errorfile=<value>**") is not supported.
    - Splitting of an output file into separate files with one structure per file (command line options "**-o split**" and "**-o splitn0=<value>**") is not supported.
- To engage with the vibrant and active Python community, MN-AM set up a GitHub repository at https://github.com/mn-am to share examples, such as Python scripts and Jupyter notebooks, on how to integrate the Python module of CORINA Classic into own workflows and projects.
- As there were no changes introduced to the 3D structure generation and manipulation engine compared to the previous version 4.4.0 (see section 2.1.17 CORINA Classic Version 4.4.0 on page 24 of this document), the generated 3D structures and any output in general are identical to the output of version 4.4.0.

2.2 CORINA_F (Restricted FlexX Interface Version)

CORINA_F is a restricted version of CORINA interfacing to the ligand docking program FlexX [22]. The interface functionality for FlexX is also contained in the full version of CORINA. The driver option **-d flexx** switches on all required command line options to interface to FlexX (see section 4.2 on page 30).

During the docking process, FlexX fragments the ligand into cyclic and acyclic parts. The ring systems including their first exocyclic neighbors are sent to CORINA or CORINA_F, respectively, which then generates a set of low-energy conformations for these ring systems and sends them back to FlexX.

The only difference between CORINA and CORINA_F is that the latter runs only if the following restrictions are fulfilled.

1) Only one input structure per program call is allowed.
2) The structure contains at least one but only one ring system (fused, bridged and spiro systems are regarded as one ring system!).
3) Acceptable ring systems are those with no more than nine atoms in the ring.
4) Exocyclic parts that exceed two bonds are not allowed.

The interface between FlexX and CORINA is described in more detail in section 6.12 on page 81. The method implemented in CORINA and CORINA_F, respectively, for generating multiple ring conformations is briefly described in section 12.5 on page 149.

3 Getting Started with CORINA Classic

CORINA is a command line program system (executable file "`corina`" on UNIX/Linux systems and *corina.exe* on Microsoft Windows platforms) and has to be executed in a shell (*e.g.*, csh, tcsh, or bash on UNIX/Linux systems, see Figure 2) or at a Windows command prompt (see Figure 3).

All command line options provided by CORINA are described in detail in section 4 "Using CORINA Classic" of this manual.

Furthermore, some examples of command lines for different uses cases are provided in section 5 "Use Cases of CORINA Classic" on page 64.

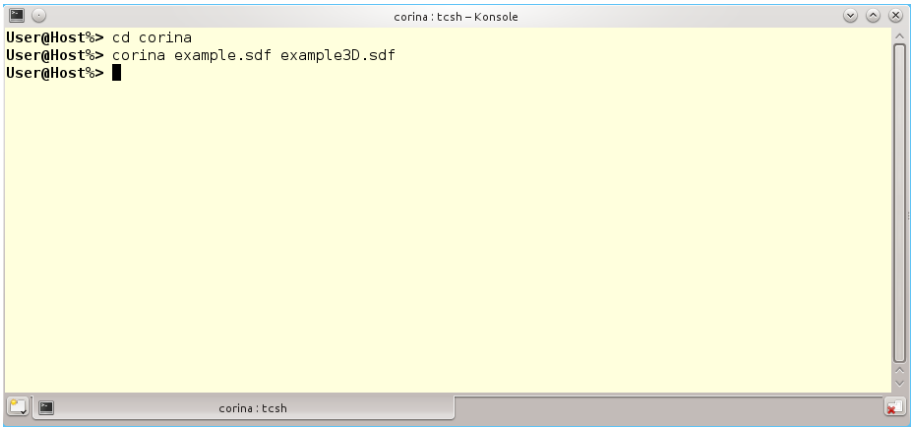


Figure 2 CORINA Classic executed in a UNIX/Linux shell.

Figure 3 CORINA Classic executed at a Windows command line prompt.

The example file "`example.sdf`" provided with the distribution contains the structure information of three molecules in SD file format [10], which is the default file format for structure input and output of CORINA.

Please copy the example file into your working directory and execute the following command at the command line prompt.

```
prompt:>corina example.sdf out.sdf
```

CORINA creates the output file "`out.sdf`" containing the input information and the generated 3D coordinates. Figure 4 shows the generated 3D structures.

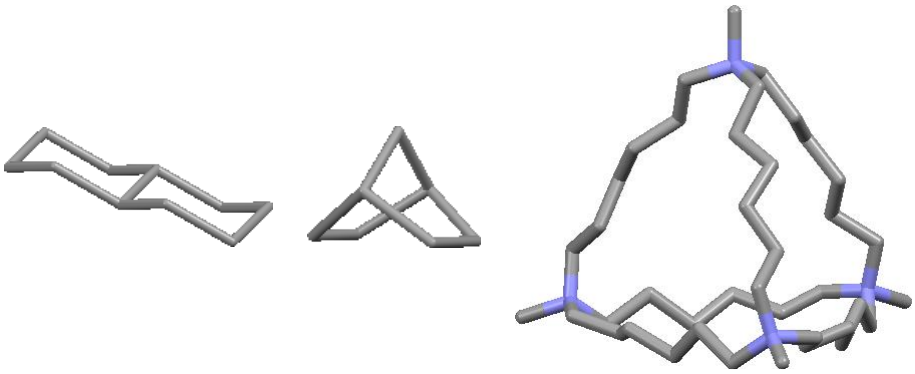


Figure 4 3D models of the structures of the example files.

Furthermore, a trace file (log file) named "`corina.trc`" which contains information on the CORINA run, such as used parameters, defined options, computation times, details on the 3D structure generation process, as well as error and warning messages is created in your working directory.

The additional trace file option **-t s** redirects this trace information to the standard output device (usually the screen). With the command line

```
prompt:>corina -t s example.sdf out.sdf
```

the following output appears on the screen.

```
corina 5.0.0 <serialNumber · compilationDate · user · date · time · host>

Input file type MDL SDFile
Output file type MDL SDFile

Options:
     Canonicalizing molecules before 3D generation
     Max. atoms per fragment: 999

Standard values, version 3.4, January 2006
Ring templates, version 3.0, March 2001
Torsion angle library version December 1999 (1088 patterns) all
rights CCDC, Cambridge, UK
International Chemical Identifier (InChI) Software Version 1.05
(January 2017), copyright (C) IUPAC and InChI Trust Limited

*** RECORD no.:  1 read **********************
   Ident 1
   Name Decaline
   10 atoms
   Elapsed time: 2 ms

*** RECORD no.:  2 read **********************
   Ident 2
   Name Norbornane
   7 atoms
   Elapsed time: 1 ms

*** RECORD no.:  3 read **********************
   Ident 3
   Name Trimacrocyclus
   44 atoms
   Elapsed time: 42 ms

3 record(s) read, 3 converted
   Totally elapsed time: 0 sec
```

4 Using CORINA Classic

4.1 Synopsis

The general synopsis for using CORINA is as follows.

```
corina [-option(s) [sub option[=value],...]] [infile] [outfile]
```

The parameters *infile* and *outfile* are the input and output file names. If no file names are provided, CORINA reads from standard input and writes to standard output. If only one file name is given, this file will be read as input file and the output will be written to standard output. By default, a minimum of trace output is written to the file "`corina.trc`".

Some examples of command lines for different uses cases are provided in section 5 "Use Cases of CORINA Classic" on page 64.

4.2 Options

-i Input file options

t=<value> Set the input file format type to <value>.

Allowed values (input file formats) are as following.

sdf	SD V2000, V3000 and mixed records file [10] (default)
rdf	RD V2000, V3000 and mixed records file [10]
smiles	SMILES linear code [11]
mol	SYBYL MOL file [12]
mol2	SYBYL MOL2 file [12]
mmod	MacroModel structure file [14]
mae	Maestro file [15]
inchi	InChI linear code [17]
ctx	ClearText file [28]

SD input file related options

dummies	Allow the interpretation of dummy atom types in SD files ("A,Q,*").
expandapo	Expand attachment points into 3D. The input option **-i expandapo** forces CORINA to expand attachment points defined in SD files ("M APO" field in the properties block) into 3D space. The attachment points are added as "artificial" atoms to the connection table (both to the atom and bond list) and 3D coordinates are calculated. Dummy atom types are assigned to the "artificial" atoms, *i.e.*, "Du" in SYBYL MOL/MOL2 files, "*" (first attachment point) and "**" (second attachment point), respectively and "X" in PDB files. In addition, the atom names of the attachment point atoms are set to "R1" (first attachment point) and "R2" (second attachment point), respectively, in the output file for formats that support atom names (*e.g.*, SYBYL MOL2).
force3d	Force that SD files having the 2D flag set are processed as if the 3D flag is set. This option is especially useful if CORINA is used only as a file converter with the driver option **-d no3d** (see below). This option allows for the handling of SD files that have the 2D flag set (2nd header line) as if the 3D flag is set. Thus, 2D files can be interconverted into file formats that usually require 3D coordinates (*e.g.*, SYBYL MOL/MOL2, PDB).

sdfict	Ignore *cis*/*trans* configuration of double bonds in SD input files. In SD files, configurations at double bonds are specified by the 2D coordinates of the substituents. This option suppresses the interpretation of the 2D coordinates and generates the most favorable configurations (*E* configuration in most cases). **Note.** The generated isomer might not be the expected one.
sdfi2c=<value>	Copy the content of the data item field <value> in the SD input file to the comment line of the SD output file ("sdf item to comment"). This option copies the contents of the SD data item field <value> (data field header) in the SD input file to the comment line field of the output file (*i.e.*, 3rd line in the header block of an SD output file).
sdfi2n=<value>	Copy the content of the data item field <value> to the name line of the SD output file ("sdf item to name"). This option copies the content of the SD data item field <value> (data field header) specified by the user in the SD input file to the name line in the header block (1st line) of the SD output file.

SYBYL MOL and MOL2 input file related options

csdmol2	Allow the CSD (Cambridge Structural Database) specific extensions in SYBYL MOL/MOL2 input files.
dummies	Allow the interpretation of dummy atom types in SYBYL MOL/MOL2 ("Du").

xelement — Allow extra elements in SYBYL MOL2 input files.

If dummy atom types ("Du") or element symbols which are undefined SYBYL atom types (*e.g.*, "Ni" for a nickel atom) are defined in SYBYL MOL2 input files, this option forces CORINA to derive – if possible – SYBYL atom types either from the atom names or from the element symbol. Furthermore, CORINA tries to interpret element symbols to internally set appropriate atom types for the 3D structure generation process. By default, CORINA then outputs dummy atom types ("Du") for these atoms (unless the output option **-o xelement** is set, see below).

SMILES input file related options

dummies — Allow the interpretation of dummy atom types in SMILES linear code ("[*]") input files.

SMILES and InChI input file related options

sc#=<value>

scn=<value>

Read chemical structure information from column <value> in the input file.

By default, CORINA expects the chemical structure information in the first column of a SMILES or InChI input file.

For a multi-column input file (*e.g.*, values separated by spaces or tabs) containing a SMILES or an InChI string in one column, the value of this option specifies the column number where the chemical structure can be found. The input file parser in CORINA supports only space separated values. Values containing blank characters (space, tab) must be enclosed with "double quotes".

nc#=<value>
ncn=<value>

Read name information from column <value> in the input file.

By default, the second column in the input file is interpreted as the compound name. For multi-column input file (*e.g.*, values separated by spaces or tabs) containing a SMILES or an InChI string in one column, the value of this option specifies the column number where the name of the chemical structure can be found. The input file parser in CORINA supports only space separated values. Values containing blank characters must be enclosed with "double quotes".

The value "<value>" must be an integral number. A value of "0" (zero) suppresses the copying of any name information from the input to the output file.

cc#=<value>
ccn=<value>

Read comment from column <value> in the input file.

By default, the third column in the input file is interpreted as the comment. For multi-column input file (*e.g.*, values separated by spaces or tabs) containing a SMILES or an InChI string in one column, the value of this option specifies the column number where the comment to the chemical structure can be found. The input file parser in CORINA supports only space separated values. Values containing blank characters must be enclosed with "double quotes".

The value "<value>" needs to be an integral number. A value of "0" (zero) suppresses the copying of any name information from the input to the output file.

sep=<value>

Specify a character used as the column separator in a multi-column input file.

By default, blank, white or tab spaces tab are considered as column separators. With this option, different column separators used in the input file can be specified (e.g., ";"). Special characters have to be double-quoted (*e.g.*, **-i t=smiles,sep="\"**).

Only a single character is allowed as a separator.

-o Output file options

t=<value> Set the output file format type to <value>.

Allowed values (output file formats) are as following.

sdf	SD v2000, v3000 and mixed records file as provided in input file (default) [10]
sdf2	SD V2000 file [10]
sdf3	SD V3000 file [10]
rdf	RD V2000, V3000 and mixed records file [10]
rdf2	RD V2000 file [10]
rdf3	RD V3000 file [10]
mol	SYBYL MOL file [12]
mol2	SYBYL MOL2 file [12]
pdb	Brookhaven Protein Data Bank file [13]
dic	CCP4 dictionary file [26]
top	X-PLOR topology file [27]
par	X-PLOR parameter file [27]
mmod	MacroModel structure file [14]
mae	Maestro file [15]
cif	Crystallographic Information File [16]
mopacxyz	MOPAC Cartesian file [18]F
odb	O Database file [31]
dyana	DYANA file [33]
ctx	ClearText file [28]

a Append the output to the input file instead of creating a new output file.

General output file options

split	Split multi-record input files into single-record output files. A consecutive number is automatically inserted into the user-provided output file name, *e.g.*, "`output.001.sdf`", "`output.002.sdf`", …, "`output.999.sdf`".
splitn0=<value>	Split multi-record input files into single-record output files and set the number of leading "0"s to the consecutive number that is automatically inserted into the user-provided output file name. If <value> is set to "`4`" (*i.e.*, -o splitn0=4) and the user-provided output file name is "`output.sdf`", the split output file names have the format "`output.00001.sdf`", "`output.00002.sdf`", …, "`output.99999.sdf`".

SD output file related options

lname	Allow compound names that are longer than 80 characters in the name field in the header block of an SD output file. **Note.** This option may violate the SD file definitions (see also section 6 on 68).
mdlbond4	Set all aromatic bonds in the output file to the bond type "4". **Note.** The bond type 4 is an SD query option in the bond block. Standard SD are alternating single (bond type "1") and double (bond type "2") for aromatic systems.

mdldb

Add the additional data fields "<MODEL.SOURCE>" and "<MODEL.CCRATIO>" to SD output file.

If the output file type is set to SD two additional SD data fields are added to each record in the output file. The field "<MODEL.SOURCE>" provides information about the program version of CORINA which was used. The field "<MODEL.CCRATIO>" contains the smallest close contact ratio (CCR) of the respective CORINA-generated 3D molecular model.

mdlcompact

Write out a compact SD file.

This option restricts the number of output fields in the lines of the atom block in SD files (RD files) to the x-, y- and z-coordinates, the atom type (symbol), the mass difference, the atom charge and the stereo-chemical atom parity (columns 1 through 7 of the atom block). All other fields in the atom lines are omitted, since they contain no data that is mandatory for 3D structure information. The goal is to save disk space (up to 40%) in case of large files.

mdl3dparity

Write out the atom stereochemical parity flags that have been calculated by CORINA for missing stereochemical descriptors to SD (RD) output file.

If stereochemical information is missing in the input file CORINA assumes reasonable stereochemical descriptors following some implemented rules (see section 12.4 on page 126). This option forces CORINA to output the stereochemical parity flags in SD and RD files that were calculated and used by CORINA during the 3D structure generation process.

pascom

Pass the comment line from input to output file.

This option preserves the comment line of the input SD file and writes it to the comment line of the output file. This option works with any file format that supports comments (fields or lines). By default, CORINA overwrites the comment line of the input SD file with information about the program version of CORINA.

sdfc2i=<value>

Copy the content of the comment line in the input file to a SD data item field <value> in the output SD file ("sdf comment to item").

This option copies the content of a comment line in the input file to a user-specified and newly generated SD data field <value> (= data field header). This option works with any input file format that supports comments (fields or lines).

Note. If the comment line of the input file is empty, no additional SD data item field is created.

sdfn2i=<value>

Copy the content of the name field in the input file to the SD data item field <value> ("sdf name to item").

This option copies the content of a name field in the input file to a user-specified and newly generated SD data field <value> (= data field header). This option works with any input file format that supports comments (fields or lines).

Note. If the name field of the input file is empty, no additional SD data item field is created.

sdfs2i=<value>

Copy the line notation of the input structure to the SD data item field <value> ("sdf structure to item").

This option only works with SMILES or InChI input files. This option copies the content of a SMILES or InChI string of the input file to a user-specified and newly generated SD data field <value> (= data field header).

SYBYL MOL and MOL2 output file related options

fcharges

Write formal atom charges into the partial charge column of SYBYL MOL/MOL2 output files.

This option forces CORINA to write formal atom charges which are given in the input structure to the charge column (column 9) in the corresponding "@<TRIPOS>ATOM" data lines of the SYBYL MOL2 output file. In addition, the keyword "USER_CHARGES" is set in the field "@<TRIPOS>MOLECULE".

gold

Force the assignment of atom and bond types in SYBYL MOL/MOL2 output files according to the GOLD conventions for difficult functional groups.

For functional groups that can be expressed by more than one canonical form (*e.g.*, guanidinium groups), the ligand docking program GOLD requires a uniform and defined coding of the atom and bond types in SYBYL MOL/MOL2 files [34],[35]. This option forces CORINA to assign the atom and bond types for those groups following the GOLD conventions which are defined in the GOLD Version 2.0 documentation.

m2l

Copy the given isotopic mass labels in the input file into the corresponding atom name field in the SYBYL MOL/MOL2 output file ("mass to label").

Atoms without a mass label remain untouched. The atom name has the format <symbol><mass>. Non-element symbols are replaced by "R". Thus, an atom type "[8*]" would get the atom name "R8". This can be used for the preparation of combinatorial libraries for FlexX.

noccat — Suppress the automatic conversion of the carbon atom in amidinium-like structures ([NH_2^+]=CN) to the carbocation type SYBYL atom type "`C.cat`" (N[C^+]N) in SYBYL MOL/MOL2 output files.

Note. The conversion to the atom type "`C.cat`", which is the default, is still strongly recommended.

nodummies — Suppress writing of unknown (dummy) atom types in SYBYL MOL/MOL2 output files.

If an unknown SYBYL atom type or a dummy ("`Du`") atom type is encountered the record is discarded from the output file.

xelement — Allow extra elements in SYBYL MOL2 output files.

If the input option **-i xelements** (see above) is set the automatically derived SYBYL atom types or interpreted element symbols ("artificial" SYBYL atom types) are written to the SYBL MOL2 output file.

PDB output file related options

keepnames — Keep any atom name given in the input file.

Usually, if the atom name and the atom type differ, CORINA tries to derive a reasonable atom name from the atom type for PDB output files. This option forces CORINA to keep and to output atom names which are defined in the input file.

pdbatom — Write the keyword "`ATOM`" instead of "`HETATM`" in PDB output files.

pdbludi — Create a PDB input file for a Ludi fragment database.

pdbludilabel	Generate unique three-character labels for a Ludi fragment database in PDB output files.
pdbnoconect	Skip all "CONECT" statements in PDB output files.
pdbelement	Add the atomic element symbol to each "HETATM" or "ATOM" line. Since PDB V2.3 (July 1998), the "HETATM" or "ATOM" lines can store the element symbol in column 77 and 78. This option can be used to add the element symbols as provided in the input file to the PDB output file. This procedure is helpful to store the correct element symbol information in a PDB file if one of the labeling schemes is used (output file options **hlabel** and **xlabel**).
resnam=<value>	Set the residue name to <value> in PDB output files.
xlabel	Label each atom type (chemical element) separately by applying an extended labeling scheme using the character set [0-9;A-Z]. By default, CORINA uses the same counter for hydrogen and non-hydrogen atoms to label the atoms. With this option each atom type (chemical element) is labeled separately within the respective atom name field (*e.g.*, columns 13 to 16 in the "HETATM" or "ATOM" line of a PDB file). In addition, the new schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.
resno=<value>	Set the residue number to <value> in PDB output files.

CCP4 dictionary, X-PLOR topology/parameter, MacroModel, Maestro, Crystallographic Information File (CIF), O Database and DYANA output file related options

coplan

Set strict coplanarity of alkoxy groups at aromatic rings (CIF output files).

Alkoxy group substituents at aromatic systems are usually coplanar to the aromatic ring if there is no steric crowding that forces the substituent out of the plane (depending on the substitution pattern at the aromatic ring). With this option coplanar alkoxy groups are listed in the "`plane`" loop in CIF files and in the "`torsion`" loop an "`esd`" value (tolerance/standard deviation for torsion angle) of only 10° and a "`period`" of 2 is applied. This guarantees a stable coplanarity of the respective alkoxy group also during the REFMAC refinement.

Note. This option was limited to methoxy groups in version 4.2.0, however, has been extended to alkoxy groups since version 4.3.0.

chirvol

Set chiral volume to "`both`" at substituents at flexible ring systems (CIF output files).

In case of flexible ring systems (such as six-membered rings) with axial/equatorial substituents, this option sets the values in the "`chiral volume`" loop to "`both`" (instead of either "`positive`" or "`negative`"). Real tetrahedral stereogenic centers or *E*/*Z* double bonds are not affected by this option.

dicid=<value>

Set the group ID number to <value> in CCP4 dictionary output files.

flexrta

Set all torsion angles in aliphatic ring systems (with more than 4 atoms) to "`var`" (variable) with a period of "3" and a standard deviation "`esd`" of "`20`" degrees in the loop "`chem_comp_tor`".

hlabel	Label hydrogen atoms separately. By default, CORINA uses the same counter for hydrogen and non-hydrogen atoms to label the atoms. This option forces CORINA to (re-)start with the counter at "0" for hydrogen atoms, independently of the counter for the heavy atoms and to allow for labelling of atoms in molecules with more than 99 atoms (including hydrogen atoms).
xlabel	Label each atom type (chemical element) separately by applying an extended labeling scheme using the character set [0-9;A-Z]. By default, CORINA uses the same counter for hydrogen and non-hydrogen atoms to label the atoms. With this option each atom type (chemical element) is labeled separately within the respective atom name field (*e.g.*, the atom ID field in the loop "`atom`" of a CIF file). In addition, the new schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.
keepnames	Keep any atom name given in the input file. Usually, if the atom name and the atom type differ, CORINA tries to derive a reasonable atom name from the atom type for DYANA output files. This option forces CORINA to keep and to output atom names which are defined in the input file.
multor	Write out all possible torsion angle patterns (CIF only). By default, CORINA writes out a single line ("`loop_ chem_comp_tor`" statement) for each torsion angle. This option forces CORINA to write out all possible torsion angle patterns using the loop "`chem_comp_tor`" statement.

novar	Suppress the writing of torsion angle patterns (CIF only). By default, CORINA writes out a single line ("`loop_ chem_comp_tor`" statement) for each torsion angle. This option forces CORINA not to write out any torsion angle information.
esterplane	Add planar ester groups to the plane loop and set specific values in torsion loop in CIF output files. Ester groups "CC(=O)OC" are usually planar, *i.e.*, the five denoted atoms are positioned in a plane, if there is no steric hindrance that forces one or more atoms out of the plane. With this option the five atoms of an ester group are added to the "`plane`" loop in CIF output files. Furthermore, in the "`torsion`" loop a value for the parameter "`value_angle`" of 180°, a value for the parameter "`value_angle_esd_esd`" of 0° and a value for the parameter "`period`" of 0 is applied for the carbon-oxygen (C-O) single bond of the ester group, if the torsion angle pattern "C-C-O-C" is used in the loop "`torsion`". If the torsion angle pattern "(0=)COC " is used in the loop "`torsion`", a value for the parameter "`value_angle`" of 0°, a value for the parameter "`value_angle_esd_esd`" of 0° and a value for the parameter "`period`" of 0 is set for the carbon-oxygen (C-O) single bond of the ester group. This guarantees a stable planarity of an ester group also during the REFMAC refinement.
resnam=<value>	Set the residue name to <value> in CCP4 dictionary, X-PLOR topology, MacroModel, Maestro, Crystallographic Information File, O Database and DYANA output files.
resno=<value>	Set the residue number to <value> in MacroModel, Maestro and DYANA output files.

typchr=<value> Set the atom type character(s) to <value> in X-PLOR topology and parameter output files.

The atom type names in top and par files are defined in the format "<symbol><type character><index>". With this option the field "<type character>" can be assigned to <value>.

MOPAC Cartesian output file related options

mopackeys= <value> Write out the MOPAC keywords provided as list in <value>.

The list must be double-quoted and separated by blanks, *e.g.*, mopackeys="`KEY1 KEY2 KEY3`".

mopacaddchg Add the MOPAC keyword "`CHARGE=`*n*" to the keyword line of the MOPAC Cartesian output file.

The value "*n*" is the formal charge of the molecule.

Note. The driver option **-d neu** can have an impact on the formal charge of a molecule.

mopacoptflag= <value> Add the MOPAC optimization flag <value> to the MOPAC Cartesian output file.

The optimization flag is used and added to each atomic Cartesian coordinate.

-t Trace file options

s	Write trace output to standard error channel (default: "`corina.trc`").
n	Suppress trace output. This option is useful for the conversion of large databases since the trace file "`corina.trc`" might become rather large.
tracefile=<value>	Set trace file name to <value> (default: "`corina.trc`").

-n Record number options

n=<value>	Process only record number <value>.
f=<value>	Process all records from record number <value>.
t=<value>	Process all records to record number <value>.

-d CORINA driver options

General driver options

3dst	Force stereochemical descriptors from the input 3D structure. If this option is switched on and there is a discrepancy between the stereochemical descriptors (*e.g.*, wedge symbols or parity flags) and a 3D structure that is provided in the input file, CORINA takes the configuration derived from the input 3D coordinates (default: usage of stereochemical descriptors, see also option **i3dst** below).
amide	For amide bonds, use the configuration specified in the 2D input file (depiction), rather than the lowest energy configuration (which is the default). By default, CORINA generates the more favorable *trans*-configuration for amide bonds (unless a stereochemical descriptor is set, *e.g.*, in a SMILES code). This option forces CORINA to use the configuration of the amide bond as depicted in the input file (*e.g.*, an SD file), regardless of whether it is a *cis* or *trans* configuration.
ampax	Amplify energy penalties for axial over equatorial substituents. This option tries to direct exocyclic substituents always into the equatorial position unless specified differently or close contacts are detected. **Note.** Using this option may generate geometries that are not the lowest energy conformation CORINA can generate (*e.g.*, 1,2-dimethyl-cyclohexane with two equatorial methyl substituents instead of the lower energy conformation with one in equatorial and the second in axial position).

axchir	Check input structures for potential stereogenic axes which cause axial chirality. By default, axial chirality is ignored, and only tetrahedral chiral centers and *cis*/*trans* (*E*/*Z*) double bonds are checked for their configuration. This option forces to also check for stereogenic axes that cause axial chirality and to generate the correct configuration if the axial chirality is coded in the input structures (see also section 12.4.2.3 on page 130). If the stereoisomer generation module is switched on as well (driver option **stergen**) all possible, but unique isomers are generated also for axial stereoisomers. **Note.** Some of the following options in the sections below for the stereoisomer generation module may limit the total number of generated and output isomers.
errorfile=<value>	Write failed structures to the output file <value>. Structures that couldn't be converted by CORINA can be written to a second output file/error file with the file name <value>. The error file <value> will have the same file format as the input file and only SD, SMILES and InChI format are supported.
i3dst	Ignore any stereochemical information derived by a 3D structure provided in the input file. By default, CORINA uses the stereochemical descriptors (*e.g.*, wedge symbols or parity flags) provided in the input file to generate the correct stereoisomer. If no stereochemical descriptors are provided in the input file, but 3D atom coordinates are given, CORINA can calculate the stereochemical information from these 3D coordinates. With this option, the derivation of stereochemical information from input 3D coordinates is suppressed. **Note.** The generated stereoisomers may not be the expected ones.

ist	Ignore all (or any) stereochemical information that is given in the input file. By default, CORINA reads in and interprets stereochemical information that is given in the input file. This option forces CORINA to ignore any stereochemical information. **Note.** Please use this option with care as ALL/ANY defined stereochemical information will be ignored.
maxat=<value>	Set the maximum allowed number of atoms per molecule to <value>. By default, the number of atoms is restricted to 999 (due to limitations of file formats, *e.g.*, SD v2000 file and to prevent the conversion of macromolecular structures). This option resets this limitation to a user-defined value. **Note.** This option does NOT circumvent any limitations of file formats (*e.g.*, the limitation of SD V2000 to 999 atoms and/or bonds).
no3d	Skip the 3D coordinates generation. This option allows for using CORINA as a file format converter for the supported file formats without generating 3D coordinates. All appropriate options are valid – including the driver options **wh** and **rs**. **Note.** If the input file format is SMILES or InChI, the output file will not contain 2D coordinates of the chemical structures.

neu	Neutralize formal charges at [C,S,P]-[O-], [NH+] and S=O(=O)[N-]. This option can be used to achieve the same protonation states for carboxylic acids, sulfates, phosphates, alcoholates, basic nitrogen atoms and negatively charged nitrogen atoms in sulfonamide groups by adding or removing protons. This option can be used together with the driver option **rs** (see below) to remove counter-ions from salts.
newtypes	Generate new atom types and ignore all atom types and aromaticity information provided in the input file. This option forces CORINA to generate new atom types and aromaticity information instead of using the information provided in the input file (atom and bond types). This can be useful if the input file contains questionable atom or bond types.
ori	Orient the 3D structure according to its principal moments of inertia. This option is useful when the structure is directly forwarded to a graphical viewer. The molecule then appears more often in an orientation that shows as much of it as possible on one sight.
ow	Override wedge symbols. Some input file formats (SD, CTX) support both atom stereochemical descriptors and wedged bonds. When these descriptors are conflicting for the same stereocenter, CORINA by default overrides the atom descriptor and uses the wedged bond symbol for calculating the stereochemistry. This option allows to override the wedged bonds and to use the atomic descriptors instead. In any case, CORINA writes an error message when encountering conflicting stereochemical descriptors for the same stereocenter.

planil	Force anilinic ring nitrogen atoms to a planar geometry. This option forces a planar geometry at anilinic nitrogen atoms that are part of a ring system (default: tetrahedral geometry).
r2d	Remove 2D records from the output. If the input and the output file type are both set to SD file (default), CORINA by default writes the original 2D structure to the output file in cases where no 3D structure could be generated. This option is useful for database purposes in order to obtain consistent input and output files. This option prevents the writing of 2D structures to the output SD file.
rs	Remove small fragments. Remove all but the largest fragments from multi-component records (*e.g.*, counter-ions in salts, solvent molecules).
sanpyr	Force nitrogen atoms in sulfonamide groups to a pyramidal geometry. This option forces a pyramidal (tetrahedral) geometry of nitrogen atoms in sulfonamide groups. The default, which is strongly recommended, is the generation of a planar geometry of the nitrogen atom in sulfonamide groups.
wb	Write bad models. By default, 3D models having close contacts or other deficiencies are not written to the output file. This option enables the output of such models.

wh — Write hydrogen atoms.

Internally, CORINA adds missing or implicitly given hydrogen atoms before the generation of 3D coordinates to obtain structures with higher quality and better resolved close contacts. By default, the added hydrogen atoms are removed from the output file. This option forces CORINA to write out the added hydrogen atoms.

Generation of stereoisomers

stergen — Generate stereoisomeric compounds.

This option forces CORINA to automatically detect tetrahedral chiral centers and double bonds where *cis* and *trans* configuration may appear in an input structure and to generate all possible, but unique stereoisomers, regardless whether correct stereochemical descriptors are given in the input or not (see section 12.4 on page 126). By default, a maximum number of four stereocenters are processed and a maximum number of 16 stereoisomers are generated (see below).

Note. The driver option **axchir** (*vide supra*) needs to be set if axial chirality needs to be considered as well.

Note. It is recommended to read the updated section 12.4 "Handling of Stereochemistry" starting on page 126 which provides an overview of the effects and results by using the different options available for the stereoisomer generation.

chiralflag	Preserve defined stereocenters if the chiral flag is set to "on" (value of "1") in the input structure (driver option **stergen** required). With this option all stereocenters that have a defined stereochemical configuration are not permuted, *i.e.*, only a single isomer is generated for the respective stereocenter if the chiral flag is set to "on" (value of "1") in the input structure (connection table). If the chiral flag is not set, all stereoisomers are enumerated. **Note.** The option **preserve** prevails over the option **chiralflag**. **Note.** This option only effects SD V2000 and V3000 input files (SD definition of chiral flag).
msc=<value>	Set the maximum number of processed stereocenters per molecule to <value> (driver option **stergen** required). By default, the stereoisomer-generating module of CORINA processes a maximum of four stereocenters of an input structure. This option forces CORINA to process the specified number <value> of stereocenters in order to restrict or to increase the number of output isomers.
msi=<value>	Set the maximum number of generated stereoisomers per molecule to <value> (driver option **stergen** required). By default, the stereoisomer-generating module of CORINA generates a maximum of 16 possible but unique stereoisomers. This option forces CORINA to generate the specified number <value> of stereoisomers to restrict or to increase the number of output isomers.

names	Number the generated conformations or stereoisomers consecutively by adding a counter to their names (driver option **stergen** required). If stereoisomers or conformations are generated, this option may be used to consecutively number the different geometries by adding a counter to the compound name.
noflapn	Suppress the flapping (inversion) of nitrogen atoms (driver option **stergen** required). This option suppresses the flapping (inversion) of pyramidal nitrogen atoms during the generation of stereoisomers.
preserve	Preserve defined stereocenters (driver option **stergen** required). If the configuration of an input structure is not specified for all chiral centers and *cis*/*trans* double bonds, this option forces the stereoisomer generating module of CORINA to process only those centers which do not have a defined stereochemistry or configuration. **Note.** The option **preserve** prevails over the option **chiralflag**.

preserverel

Preserve relative stereochemistry of defined chiral centers (driver option **stergen** required).

By default, CORINA enumerates all possible stereoisomers, if the driver option **stergen** is set. This additional option interprets all defined chiral centers as relative and generates the respective racemic mixtures. For SD input files, the additional option **chiralflag** (*vide supra*) and the value of the chiral flag of a chemical record may affect the number of generated stereoisomers. If the value of the chiral flag in the input structure is "1" and the option **chiralflag** is set, all defined chiral centers are interpreted as absolute chiral centers.

Note. The option **preserverel** prevails over the option **preserve** for chiral centers.

preserveez

Preserve configuration of defined *E*/*Z* double bonds (driver option **stergen** required).

By default, CORINA enumerates all possible stereoisomers, if the driver option **stergen** is set. This additional driver option preserves the configuration at defined *E*/*Z* double bonds during the enumeration of possible stereoisomers. While the option **preserve** keeps both, defined chiral centers and *cis*/*trans* (*E*/*Z*) double bonds, fixed, the option **preserveez** only enumerates stereoisomers for double bonds with an undefined configuration. Chiral centers are not affected at all. This option is helpful, if all stereocenters should be permuted, but not defined *E*/*Z* double bonds.

v3000

Permute all stereocenters that belong to one of the stereo-chemical groups "STEREL*n*" and "STERAC*n*" according to their definition of their relative or racemic representation, but do not permute stereocenters that belong to the group "STERABS".

This option can only be used in combination with the driver option **stergen** to generate stereoisomers and only impacts V3000 input files.

This option can be used in combination with the option **preserve** and is useful for input files with structures that contain mixed stereocenters which are constricted or not constricted by V3000 stereo-chemical groups.

Generation of multiple ring conformations (for ring systems of up to 9 ring atoms)

rc

Generate multiple ring conformations.

This option forces the conformational analysis module for small and medium sized ring systems of CORINA to output multiple ring conformations (see section 12.5 on page 149).

This option does not support multi-fragment records (*e.g.*, salts). Work-around: Combine with sub option **rs** (see above). The conformations are written in the order of increasing energy value.

de=<value>

Set an energy window ΔE of <value> kJ/mol for the ring conformations (driver option **rc** required).

This option forces CORINA to output only those conformations which have an energy not higher than <value> (in kJ/mol) with respect to the lowest-energy conformation.

flapn	Flap ring nitrogen atoms to generate multiple ring conformations (driver option **rc** required). This option allows pyramidal ring nitrogen atoms that have one exocyclic neighbor to invert their configuration to obtain all conformations (see Section 12.5 on page 149).
mc=<value>	Set the maximum number of generated ring conformations to <value> (driver option **rc** required). By default, CORINA generates a maximum number of 10 conformations per molecule if the driver option **rc** is set (see above). This option sets the maximum number of output conformations to <value>.
names	Number the generated conformations or stereoisomers consecutively by adding a counter to their names (driver options **rc** or **stergen** required). If stereoisomers or conformations are generated, this option may be used to consecutively number the different geometries by adding a counter to the compound name.
ringatom=<value>	Denote a ring system by the atom with label <value> that is part of the ring system to generate multiple ring conformations (driver option **rc** required). This sub option can be used to denote a specific ring system for which multiple ring conformations should be generated, whereas <value> is the atom label of one atom contained in this specific ring system. It allows generating multiple ring conformations only for one specific ring system in an input structure that contains more than one flexible ring system.

sc	Generate ring conformations simultaneously (driver option **rc** required). By default, when generating multiple ring conformations for compounds having more than one ring system CORINA generates all combinations of all conformations of these ring systems. This option reduces the number of conformations by simultaneously generating conformations for different ring systems (see Section 12.5 on page 149).
symoff	Switch off the symmetry check for multiple ring conformations (driver option **rc** required). To generate different ring conformations that are unique, CORINA checks for symmetries but only in the flexible ring skeletons (including the first exocyclic substituents). This option switches off this check and thus allows for the output of conformations that can be interconverted by symmetry operations (*e.g.*, the two conformations of CCC1=C(CC)CCC1).
timeout=<value>	Restrict the computation time for the ring conformation analysis to <value> milliseconds (driver option **rc** required). For complex fused and bridged ring systems the conformational analysis routine might be quite time consuming if multiple ring geometries should be generated. This option forces to stop the conformational analysis after the set timeout of <value> milliseconds and to output all conformations obtained so far.

Interface to FlexX

flexx	Tailor all CORINA options to interface to the docking program FlexX. CORINA can be used for generating ring conformations during the flexible docking process of FlexX. This option sets the input and output file types and the conformations analysis options to suited values (see section 6.12 on page 81).

-h CORINA on-line help options

i	Print help for CORINA input file options.
o	Print help for CORINA output file options.
t	Print help for CORINA trace file options.
n	Print help for CORINA record number options.
d	Print help for CORINA driver options.
all	Print help for all available CORINA options.

-v Print program version

-m Create a UNIX/Linux on-line reference manual page (man page)

This option can be used to generate a UNIX/Linux on-line reference manual page for the manual pager utility "man", *e.g.*, by typing the command

```
prompt:>corina -m > corina.1
```

The resulting file "corina.1" is the man page of CORINA.

4.3 Application Programming Interface (API) for Python Programming/Scripting Language

Besides the standalone executable files for x86 Linux and Microsoft Windows platforms, CORINA Classic offers an application programming interface (API) for integration of its functionalities into existing or new workflows and platforms in chemoinformatics and computational chemistry projects utilizing the Python programming/scripting language.

The module (shared object file) providing the API for the Python programming language is available for x86-64 RHEL7 platforms. These platforms are compatible with many other Linux distributions, such as Ubuntu 20.04 or 22.04.

The supported Python versions are 3.8, 3.9, 3.10, and 3.11. No special requirements are needed regarding hard disk and RAM space.

The API mimics the command line syntax of the standalone version of CORINA Classic which is provided as an argument of the Python language function "`corinaBuffer()`". The function consumes input structures formatted in one of the supported input file formats. Besides SMILES and InChI formatted chemical structures that provids one structure per line (by definition), other input file formats must be provided in a single line. For this, the Python standard is to use triple single or double quotes (""" or '''', see example below, no "\n" newline escape sequence required).

Most of the command line options provided by the standalone version of CORINA Classic are supported as well by the Python module with the following exceptions.

- The output of trace files is not supported. The trace information is written to the Python variable "`buffer.log`" but misses the overall performance and run statistics at the end of the trace output (see example below).
- The output of not converted structures into a separate output file (command line option "-d errorfile=<value>") is not supported.
- Splitting of an output file into separate files with one structure per file (command line options "-o split" and "-o splitn0=<value>") is not supported.

The code block below shows an example including some comments on how the API for the Python programming environment of CORINA Classic can be used.

```
1    # Lines 3-42 define two "hardcoded" SD records in single line format
2
3    sdRecords = """Benzene
4              01242312572D 1   1.00000     0.00000     0
5
6      6  6  0  0  0  0            999 V2000
7        0.0000    1.5275    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
8       -1.3225    0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
9       -1.3225   -0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
10       0.0000   -1.5275    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
11       1.3225   -0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
12       1.3225    0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
```

```
13     1  6  2  0  0  0  0
14     2  1  0  0  0  0  0
15     2  3  2  0  0  0  0
16     3  4  1  0  0  0  0
17     5  2  0  0  0  0  0
18     6  1  0  0  0  0  0
19   M  END
20   $$$$
21   Phenole
22              01242312572D 1   1.00000     0.00000     0
23
24     7  7  0  0  0  0            999 V2000\n\
25       0.0000    1.5275    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
26      -1.3225    0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
27      -1.3225   -0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
28       0.0000   -1.5275    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
29       1.3225   -0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
30       1.3225    0.7625    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
31       2.7075    1.5575    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
32     1  6  2  0  0  0  0
33     2  1  0  0  0  0  0
34     2  3  2  0  0  0  0
35     3  4  1  0  0  0  0
36     5  2  0  0  0  0  0
37     6  1  0  0  0  0  0
38     7  1  0  0  0  0  0
39   M  END
40   $$$$""";
41
42   # Import Python module of CORINA Classic
43
44   import corina
45
46   # Create instance of function corinaBuffer
47
48   buffer = corina.CorinaBuffer()
49
50   # Set command line and structure input
51
52   buffer.command = "corina -o t=sdf3 -d wh"
53   buffer.input = sdRecords
54
55   # Run 3D structure generation and get success status
56
57   success = buffer.proceed()
58
59   # Get exit code status
60
61   status = buffer.status
62
63   # Print generated 3D structure in SD V3000 format
64
65   print(buffer.output)
66
67   # Print CORINA Classic trace information
68
```

```
69   print(buffer.log)
70
71   # Print CORINA Classic success and exit codes
72
73   print(f"Success: {success}")
74   print(f"CORINA Classic exit code status: {status}")
75   print("")
76
77   # Print CORINA Classic run/performance statistics
78
79   print("CORINA Classic record statistics:")
80   print(f"Number of records read: {buffer.statsRead}")
81   print(f"Number of records converted: {buffer.statsConverted}")
82   print(f"Number of records discarded: {buffer.statsDiscarded}")
83   print(f"Number of CORINA Classic errors: {buffer.statsErrors}")
84
```

To run the above Python script, copy, save, and name it on your Linux computer, *e.g.*, "`corinaClassic.py`", and type the following command at a command prompt in a Linux terminal/shell.

```
python3 corinaClassic.py > output.v3000.sdf
```

More Python scripts that show further functionalities are available at the GitHub repository "CORINA Classic Goes Python" at https://github.com/mn-am.

Important note. The Python module of CORINA Classic **must** have the file name "`corina.so`". Usually, the distributed file has a different name (*e.g.*, "`corina_annual_python-v3-11_x86-64_rhel7_2024-08-31.so`"). Please make a copy of the file and rename it to "`corina.so`" before the first use. You may keep the copy of the distributed, original file for documentation (the file name contains supported Linux version, Linux architecture, Python version, and license timeout) and backup purposes.

5 Use Cases of CORINA Classic

The following section lists some typical use cases of CORINA and shows the corresponding command lines and command line options that have to be used for the specific tasks.

1) Converting a 2D SD file into a 3D SD file

An SD file should be converted into 3D. Implicit hydrogen atoms should be added, small fragments (*e.g.*, counter ions in salts) should be removed and all molecules should be neutralized. In addition, structures that couldn't be converted should be excluded from the 3D output file but written to a separate error file. The output file should also be formatted in SD file format.

Command line:

```
corina -d wh,rs,neu,r2d,errorfile=errors.sdf in.sdf
out3D.sdf
```

2) Using 3D input information to determine the correct stereochemistry

An SD file containing crude 3D structures with missing stereochemical information (no parity flags or wedge symbols) should be converted into 3D. The stereochemical information should be derived from the crude input geometry, implicit hydrogen atoms should be written out and the generated 3D structures should be oriented by their principal moments of inertia. Since the 3D structures should be used in a docking experiment with the ligand docking program GOLD, the output file needs to be in SYBYL MOL2 file format and all atom and bond types should be assigned according to the GOLD conventions for functional groups.

Command line:

```
corina -o t=mol2,gold -d 3dst,wh,ori in.sdf out3D.mol2
```

3) Generating sets of stereoisomers

A file containing several hundreds of SMILES strings should be converted into 3D. For chiral compounds a set of a maximum of 12 stereoisomers should be generated but any defined stereochemistry in the input structures should be preserved. The stereoisomers should be numbered, and implicit hydrogen atoms added. The output file format should be SD file.

Command line:

```
corina -i t=smiles -d stergen,msi=12,preserve,names,wh
in.smi out3D.sdf
```

4) Converting file formats without generating (new) 3D coordinates

An SD file containing X-ray structures should be converted in a SYBYL MOL2 file. The 3D coordinates of the missing hydrogen atoms should be added, however, the coordinates of any atom in the input structures should not be changed at all.

Command line:

```
corina -o t=mol2 -d no3d,wh in3D.sdf out3D.mol2
```

5) Generating sets of multiple ring conformations

For the structures in a SYBYL MOL2 file, new 3D coordinates should be generated. Missing hydrogen atoms should be added, small fragments (*e.g.*, counter ions in salts) and failed structures should be removed from the output file. For each input structure a set of a maximum of ten ring conformations should be generated, nitrogen atoms in flexible rings are allowed to be inverted and the search for different ring geometries should generate a reduced set of conformations (see also Section 12.5 on page 149).

Command line:

```
corina -i t=mol2 -o t=mol2 -d wh,rs,r2d,rc,mc=10,flapn,sc
in3D.mol2 out3D.mol2
```

6) Reading an InChI file with names given in a separate column

An input file that contains chemical structures in InChI format and the compound names in the following column should be read in and written out in PDB format. Hydrogen atoms should be written to the output file and formal charges at [C,S,P]-[O^-] and [NH^+] should be removed. In addition, any aliphatic nitrogen atoms in rings that are bonded to a phenyl ring (anilinic ring nitrogen atoms) should exhibit a planar (not pyramidal) geometry.

Command line:

```
corina -i t=inchi,nc#=2,sc#=1 -o t=pdb -d wh,neu,planil
in.inchi out3D.pdb
```

or

```
corina -i t=inchi,ncn=2,scn=1 -o t=pdb -d wh,neu,planil
in.inchi out3D.pdb
```

7) Generating a MOPAC Cartesian input file from a SMILES file

Chemical structures that are stored in a file in SMILES format should be converted to 3D and written out in MOPAC Cartesian file format. Small fragments (*e.g.*, counter ions in salts) should be removed and hydrogen atoms should be written out. In addition, the MOPAC "CHARGE" keyword should be automatically assigned to each compound record, a list of other MOPAC keywords should be added and the MOPAC coordinate optimization flag "-1" should be used.

Command line:

```
corina -i t=smiles -o t=mopacxyz,mopacaddchg,mopackeys="KW1
KW2 KW3=x" -d rs,wh in.smiles out3D.mopac
```

8) Converting a V2000/V3000 mixed SD file into a 3D V3000-only output file

Chemical structures that are stored in an SD file that contains mixed V2000 and V3000 records should be converted to 3D and written out in an output file that contains V3000 records only. Aromatic bonds should be set to bond type "4" in the output file and hydrogen atoms should be written out.

Command line:

```
corina -o t=sdf3,mdlbond4 -d wh in.sdf out3D.sdf
```

9) Generating stereoisomers from a V3000 Input File

An SD V3000 file should be read in that contains definitions of V3000 stereo-chemical groups of relative and absolute stereo-chemical definitions. Stereoisomers should be generated for a maximum number of 8 stereocenters per molecule (resulting in a maximum number of 64 stereoisomers per molecule, but only for the relative stereo-chemical groups, but not for the absolute stereo-chemical groups. Hydrogen atoms should be added and a SD V3000 file should be written out.

Command line:

```
corina –o t=sdf3 –d wh,stergen,msc=8,msi=64,v3000 in.sdf3
out.sdf3
```

10) Generating a MOPAC Cartesian input file from a SMILES file

Chemical structures that are stored in a file in SMILES format should be converted to 3D and written out in MOPAC Cartesian file format. Small fragments (e.g., counter ions in salts) should be removed and hydrogen atoms should be written out. In addition, the MOPAC "CHARGE" keyword should be automatically assigned to each compound record, a list of other MOPAC keywords should be added and the MOPAC coordinate optimization flag "-1" should be used.

Command line:

```
corina -i t=smiles -o t=mopacxyz,mopacaddchg,mopackeys="KW1
KW2 KW3=x",mopacoptflag=-1 -d rs,wh in.smiles out3D.mopac
```

11) Converting a multi-column SMILES input file into an SD output file

A SMILES file should be converted to 3D in SD file format. The input file is a multi-column (text) file. The columns are separated by semi colons (";"). The SMILES strings are stored in column number 3. The compound names in column number 4 may contain spaces and should be copied to the name field in the SD output file. The SMILES strings should be copied into the comment line of the SD output file (3rd line in

the header block). The output 3D structures should have explicit hydrogen atoms.

Command line:

```
corina -i t=smiles,sep=";",scn=3,ncn=4,ccn=3 –o pascom –d
wh in.smiles out3D.sdf
```

12) Handle atropisomers

An SD file contains structures with atropisomeric centers which are fully defined, should be converted into 3D, and written out in MacroModel format. Implicit hydrogen atoms and the coded atropisomers should be written out. Formal charges (at [C,S,P]-[O-], [NH+] and S=O(=O)[N-]) should be neutralized.

Command line:

```
corina –o t=mmod -d wh,axchir,neu in.sdf out3D.mmod
```

13) Generating all relevant stereoisomers for mixed SD V2000/V3000 structures

An inhouse collection of chemicals is stored in SD file format and contains a mix of V2000 and V3000 records. The chemical structures have either defined relative or absolute stereochemistry including V3000 stereochemical groups and chiral flags or stereocenters are undefined due to the lack of knowledge of the correct isomer or isomeric mixture. The collection should be converted into 3D, implicit hydrogen atoms should be written out and all relevant stereoisomers which are possible within the provided SD definitions should be enumerated.

Command line:

```
corina -d
wh,stergen,v3000,chiralflag,preserverel,preserveez in.sdf
out3D.sdf
```

14) Generating a PDB output file with unique atom labelling and element symbols in HETATM lines

A dataset of molecules should be converted into PDB file format, and a unique atom labelling is required for the HETATM lines as some atom types (chemical elements) may occur more often than 100 times in a molecule. In addition, the element symbols should be written into the HETATM lines and implicit hydrogen atoms should be written out.

Command line:

```
corina -i t=smiles -o t=pdb,xlabel,pdbelement -d wh
in.smiles out3D.pdb
```

6 Supported File Formats and Interfaces

This section describes the file formats that are supported by CORINA and provides additional information on their use in CORINA. Table A provides an overview of the supported input/output file formats and a reference to the file format specification.

Table A Supported input and output file formats.

Format	Input	Output	Reference
SD/RD V2000/V3000	yes	yes	[10]
SMILES	yes	no	[11]
SYBYL MOLFILE	yes	yes	[12]
SYBYL MOL2	yes	yes	[12]
PDB	no	yes	[13]
CTX	yes	yes	[28]
CCP4 dictionary file	no	yes	[26]
X-PLOR topology/parameter file	no	yes	[27]
MacroModel structure file	yes	yes	[14]
Maestro file	yes	yes	[15]
CIF	no	yes	[16]
ODB	no	yes	[31]
DYANA	no	yes	[32],[33]
MOPAC Cartesian	no	yes	[18]
InChI	yes	no	[17]

6.1 V2000 Structure Data File (SD) and Reaction Data File (RD)

The V2000 SD and RD file formats were implemented following the description in the literature [10].

CORINA reads only the name (line 1), the status line (line 2), the atom and bond counts and the chiral flag (line 4), the atom block and the bond block. In addition, the "RAD" and "CHG" atom properties from the properties block are read in. From the atom block, columns 1 through 7 are read in.

From the bond block, columns 1 through 4 are read in. All other information remains unread but is copied unchanged to the output file when the output file type is also set

to SD/RD. This is a great advantage for database purposes since all information except the 3D coordinates remains unchanged.

However, CORINA behaves quite differently if the connection table changed during the 3D structure generation process because of, *e.g.*, adding implicitly given hydrogen atoms, removing small fragments (counter ions), or neutralizing formal charges. In these cases, the program generates the counts line and the atom and bond blocks newly from the changed molecule information and discards all additional, not read-in columns of these blocks.

The individual records are assumed to be closed by the "$$$$" delimiter.

6.1.1 Options to ManipulateV2000 SD and RD Files

The input option **-i sdfi2n=value** copies a one-line data item named <value> to the compound name line (1st line of the header block) in the SD output file. For example, **-i sdfi2n=MY_FIELD** forces CORINA to copy the data line immediately following the data item header <MY_FIELD> into the compound name line. This can be used to export a single special data line into any other output file format which supports a compound name.

The input option **-i sdfi2c=<value>** copies the contents of the SD data field named <value> (data header) in an SD input file to a comment line in the output file, *e.g.*, to the 3rd line in the header block of an SD file. Thus, a piece of information from a data field in an SD file can be transported to any file format which supports a comment (*e.g.*, SD, SYBYL MOL2, PDB).

By default, CORINA considers atom stereochemical descriptors and wedge symbols for chiral centers (see also driver option **-d ow** in section 4.2 on page 30), as well as bond descriptors indicating *cis* or *trans* double bonds that are given in the input structure for generating a 3D structure.

Unfortunately, stereochemical descriptors are quite often even not specified or specified not correctly. Two options specifically designed for SD (RD) input files influence the handling of stereochemistry during the 3D structure generation process. If the input option **-i sdfict** (SD file ignore *cis*/*trans*) is set CORINA ignores all bond descriptors which define *cis*- or *trans*-double bonds in order to convert also those structures with unreasonably defined descriptors, *e.g.*, if a *trans* double bond is specified in a small ring system, or with ambiguous definitions, *e.g.*, contradictory definitions in conjugated systems.

In addition, the output option **-o mdl3dparity** forces CORINA to write out all stereochemical descriptors that were calculated by CORINA for centers with undefined stereochemical information.

The output options **-o mdldb** and **-o mdlcompact** are useful for the conversion of large datasets or databases.

If **-o mdldb** is set the additional data fields "<MODEL.SOURCE>" giving information about the program version of CORINA, which was used to generate the 3D models and

"<MODEL.CCRATIO>" containing the smallest close contact ratio encountered in generated 3D molecular model are added to each record in the output file.

The option **-o mdlcompact** forces CORINA to output only the fields containing the x-, y- and z-coordinates, the atom type (symbol), the mass difference, the atom charge and the atom stereochemical parity (columns 1 through 7 of the atom block) in the atom lines of the atom block. The columns 8 through 16 (in most cases assigned with values of 0) mainly contain information related to chemical reactions and, thus, are not mandatory for 3D structure generation and representation. This option may save disk space of up to 40%. Please always ensure that the information given in the omitted columns is not needed for any other purposes before using this option.

The input option **-i expandapo** only has an impact if attachment points "M APO" are defined in input structures. If this option is set all attachment points are expanded into 3D space. The attachment points are added as "artificial" atoms to the connection table (both to the atom and bond list) and 3D coordinates are calculated. Dummy atom types are assigned to the "artificial" atoms, *i.e.*, "Du" in SYBYL MOL/MOL2 files, "*" (first attachment point) and "**" (second attachment point), respectively and "X" in PDB files. In addition, the names of the attachment point atoms are set to "R1" (first attachment point) and "R2" (second attachment point), respectively, in file formats which support atom names (*e.g.*, SYBYL MOL2).

The output option **-o pascom** passes comment lines between file formats which support comments (*e.g.*, SD, SYBYL MOL2, PDB). If the SD file format is set as input and output file format the comment line in the header block (3rd line) of the input SD file is preserved and passed to the output file. By default, CORINA replaces the comment line given in the input SD file with information about the program version and writes it to the output SD file.

The output option **-o sdfc2i=<value>** ("sdf comment-to-item") copies the comment line of an input file, *e.g.*, 3rd line in the header block of an SD file, to a newly generated SD data field named "value" (data header) in the SD output file. Thus, a comment line from file formats which support comments (*e.g.*, SD, SYBYL MOL2, PDB) can be copied to a data field of an SD output file.

The output file option **-o sdfn2i=<value>** ("sdf name-to-item") copies the content of the name field (1st line) into the SD data item field "value" specified by the user.

The output option **-o sdfs2i=<value>** ("sdf structure-to-item") copies the linear notation of a chemical structure in a SMILES or InChI input file into the SD data item field "value" (data header) specified by the user.

The output file option **-o mdlbond4** automatically assigns to aromatic bonds the bond type of "4" in the bond block of the SD output file. **Note.** A bond type ≥4 is defined as CTAB query features (for substructure searches) and other applications may not be able to assign the correct bond type when reading such a value.

The output file option **-o lname** supports the copying of molecule names (1st line in header block) that have more than 80 characters. By default, the molecule name is cut after 80 characters for compatibility with the V2000 file format.

6.2 V3000 Structure Data File (SD) and Reaction Data File (RD)

The V3000 SD and RD file formats were implemented following the description in the literature including the enhanced stereo-chemical representation [10]. CORINA also accepts mixed records files, *i.e.*, if molecules are stored in the input file in V2000 and V3000 format.

The stereo-chemical identifiers and groups "CFG", "ABSOLUTE" (absolute stereochemistry, group "STEABS"), "OR" (relative stereochemistry, group "STEREL*n*") and "AND" (racemic representation, group "STERAC*n*") are fully supported. When generating stereoisomers, the interpretation of the stereo-chemical identifiers and groups can be manipulated (please see section 12.4.4 on page 137).

Note. The V3000 format does not have any limitations regarding the number of atoms and bonds (999 in V2000).

Note. The V3000 format limits the number of characters per line to 80. If a line has more than 80 characters, the hyphen character "-" can be used as the last character in the line to allow for continuation in the next line.

CORINA does not support the use of templates (useful to represent large structures, such as polypeptides).

6.2.1 Options to Manipulate V3000 SD and RD Files

All command line options that are described in section 6.1.1 on page 69 above for the SD V2000 also apply to the V3000 format. The options to handle V3000 stereochemical enhancements are described in section 12.4.4 "Enumeration of Stereoisomers" on page 137.

6.3 SMILES Linear Notation

The SMILES linear notation was implemented following the literature [11].

In addition, non-standard formal charge qualifiers immediately following the atomic symbol in curly brackets have been implemented (*e.g.*, "CCCC(O{-})CC").

Aromatic atoms coded in lower cases, only the atom types carbon, nitrogen, oxygen and sulfur are allowed. Implicit hydrogen atoms given inside square brackets are expanded and written to the output file, whereas all other missing hydrogen atoms are only written out, if required by the user (driver option **-d wh**).

By default, each line must begin with a SMILES linear code and only one SMILES string per line is allowed. Any additional information in the same line that is separated by a white space (or tab) from the SMILES string is interpreted as following. Any information in the second column is interpreted as the compound name and copied into the respective field (if available) in the output file. The information in the third

column is considered a comment and can be copied to the respective comment field in the output file (output option **-o pascom**).

Note. If the compound name or a comment (or any information) in the SMILES input file contains blanks, white spaces or tabs, the entire expression must be "double quoted". Blanks, white spaces, or tabs are interpreted as column separators.

6.3.1 Options to Manipulate SMILES Files

For multi-column SMILES input files (*e.g.*, columns separated by tabs or blanks), three input file options **-i sc#=<value>**, **-i nc#=<value>** and **-i cn#=<value>** are available to specify the column numbers in the input file in which the chemical structure information (coded as a SMILES string), the compound name and any comment are stored.

Respectively, the three input file options described above can also be used with an "**n**" instead of the character "**#**", *i.e.*, **-i scn=<value>**, **-i ncn=<value>** and **-i cnn=<value>**. These additional options have been introduced to prevent any issues when using these options in a CORINA Classic command line which is executed in a script (the character "#" might be a special character, *e.g.*, as in a UNIX shell script).

With the input option **-i sc#=<value>** (or **-i scn=<value>**), the column number can be specified which contains the chemical structure information (coded as a SMILES string). The option **-i t=smiles,sc#=3** (or **-i t=smiles,scn=3**) reads the SMILES string from column number 3.

With the input option **-i nc#=<value>** (or **-i ncn=<value>**), the column number can be specified which contains the compound name. The option **-i t=smiles,nc#=4** (or **-i t=smiles,ncn=4**) reads the compound name from column number 4.

With the option **-i cc#=<value>** (or **-i ccn=<value>**), the column number can be specified which contains a comment of the compound. The option **-i t=smiles,cc#=5** (or **-i t=smiles,ccn=5**) reads the comment from column number 5. With the additional output option **pascom**, the comment is copied and written out in the respective field of the output file. If the further output option **sdfc2i=value** is set and the output file format is SD (default), the content of the comment is copied to the field data item "<value>" in the output file.

Note. For the input options **sc#**/**scn**, **nc#**/**ncn** and **cc#**/**ccn** described above, the values "<value>" have to be integral numbers. If a value of "0" (zero) is set, the copying of any information (name or comment) from the input to the output file is suppressed.

With the input file option **-o sep=<value>**, column separator used in a multi-column SMILES input file can be specified. By default, blank, white or tab spaces tab are considered as column separators. Any items with blanks or spaces that should be considered as a single term, must be double double-quoted (*e.g.*, "`Mesidine hydrochloride`"). However, this default behavior fails, if an entry in a column contains, *e.g.*, a blank in a compound name such as `Mesidine hydrochloride`, but is not double-quoted in the input file. This option provides the flexibility to also

have items with blank, white and tab spaces, if a different column separator is used, *e.g.*, ";".

Note. Special characters (Linux and UNIX shell specific) have to be double-quoted (") on the command line, if they are used as column separators in the input file, *e.g.*, **-i t=smiles,sep="\"**.

Note. A tabulator as a separator can be defined as "\t", TAB or tab.

Note. Only a single character is allowed as a separator.

The input option **-i dummies** allows for the interpretation of unknown or dummy atom types in SMILES linear code.

Note. The correct definition of a dummy atom according to the SMILES language syntax is " [*] ".

6.4 InChI file format

CORINA Classic uses the InChI Software version 1.05 (January 2017, copyright (C) IUPAC and InChI Trust Limited) under the IUPAC/InChI-Trust License No. 1.0 ("www.inchi-trust.org") to support standard InChI (IUPAC International Chemical Identifier) as input file format.

By default, each line must begin with an InChI linear code and only one InChI string per line is allowed. Any additional information in the same line that is separated by a white space (or tab) from the InChI string is interpreted as following. Any information in the second column is interpreted as the compound name and copied into the respective field (if available) in the output file. The information in the third column is considered a comment and can be copied to the respective comment field in the output file (output option **-o pascom**).

Note. If the compound name or a comment (or any information) in the InChI input file contains blanks, white spaces or tabs, the entire expression has to be "double quoted". Blanks, white spaces, or tabs are interpreted as column separators.

6.4.1 Options to Manipulate InChI Files

For multi-column InChI input files (*e.g.*, columns separated by tabs or blanks), three input file options **-i sc#=<value>**, **-i nc#=<value>** and **-i cn#=<value>** are available to specify the column numbers in the input file in which the chemical structure information (coded as an InChI string), the compound name and any comment are stored.

Respectively, the three input file options described above can also be used with an "**n**" instead of the character "**#**", *i.e.*, **-i scn=<value>**, **-i ncn=<value>** and **-i cnn=<value>**. These additional options have been introduced to prevent any issues when using these options in a CORINA Classic command line which is executed in a script (the character

"#" might be a special character, *e.g.*, as in a UNIX shell script).

With the input option **-i sc#=<value>** (or **-i scn=<value>**), the column number can be specified which contains the chemical structure information (coded as an InChI string). The option **-i t=inchi,sc#=3** (or **-i t=inchi,scn=3**) reads the InChI string from column number 3.

With the input option **-i nc#=<value>** (or **-i ncn=<value>**), the column number can be specified which contains the compound name. The option **-i t=inchi,nc#=4** (or **-i t=inchi,ncn=4**) reads the compound name from column number 4.

With the option **-i cc#=<value>** (or **-i ccn=<value>**), the column number can be specified which contains a comment of the compound. The option **-i t=inchi,cc#=5** (or **-i t=inchi,ccn=5**) reads the comment from column number 5. With the additional output option **pascom**, the comment is copied and written out in the respective field of the output file. If the further output option **sdfc2i=value** is set and the output file format is SD (default), the content of the comment is copied to the field data item "<value>" in the output file.

Note. For the input options **sc#**/**scn**, **nc#**/**ncn** and **cc#**/**ccn** described above, the values "<value>" have to be integral numbers. If a value of "0" (zero) is set, the copying of any information (name or comment) from the input to the output file is suppressed.

Note. The output files generated by the InChI software can be directly read in by CORINA.

With the input file option **-o sep=<value>**, column separator used in a multi-column InChI input file can be specified. By default, blank, white or tab spaces tab are considered as column separators. Any items with blanks or spaces that should be considered as a single term, must be double double-quoted (*e.g.*, "`Mesidine hydrochloride`"). However, this default behavior fails, if an entry in a column contains, *e.g.*, a blank in a compound name such as `Mesidine hydrochloride`, but is not double-quoted in the input file. This option provides the flexibility to also have items with blank, white and tab spaces, if a different column separator is used, *e.g.*, ";".

Note. Special characters (Linux and UNIX shell specific) have to be double-quoted (") on the command line, if they are used as column separators in the input file, *e.g.*, -i t=smiles,sep="\".

Note. A tabulator as a separator can be defined as "\t", TAB or tab.

Note. Only a single character is allowed as a separator.

6.5 SYBYL File Formats

Both SYBYL MOL and MOL2 file formats were implemented following the SYBYL program manual [12].

Since both formats are based on rather special atom types, their applicability as a standard database format is limited and many cases can occur where no meaningful

atom type can be assigned. Dummy atom types are assigned to atoms with lacking atom types. MOL2 files are written by using the keywords "@<TRIPOS>MOLECULE", "@<TRIPOS>ATOM" and "@<TRIPOS>BOND".

6.5.1 Options to Manipulate SYBYL File Files

SYBYL file formats are restricted to a limited number of different atom types, so-called SYBYL atom types, according to the parameterized atom types in the SYBYL force field package. Therefore, SYBYL file formats are rather restricted for general structure representation purposes. However, many program systems and software packages support SYBYL MOL/MOL2 file formats with various extensions in order to overcome the lacking atom type definitions.

By default, CORINA only accepts and interprets atom types in SYBYL files which are properly defined as SYBYL atom types (a straightforward philosophy since the SYBYL interface was implemented accurately following the SYBYL program manual). Furthermore, several input and output options allow CORINA to handle also atom types that cannot be regarded as generic SYBL atom types.

The input option **-i csdmol2** and **-i xelement**, as well as **-i dummies**, force CORINA to interpret CSD specific extensions (*e.g.*, transition metal atom types such as Ni, Zn, or Cu), to internally use atom types which are estimated when encountering element symbols or ambiguous defined SYBYL atom types, or to allow dummy atom types ("Du") in SYBYL input files.

Furthermore, the output option **-o nodummies** suppresses the output of dummy atom types ("Du") in SYBYL files and records that contain dummy atom types or unknown SYBYL atom types are discarded. This option is useful if post-processing software requires or can handle only atom types that are "true" SYBYL atom types.

In contrast to the above, the output option **-o xelement** allows for writing out atom types which are not listed as SYBYL atom types ("artificial" SYBYL atom types, *e.g.*, a SYBYL atom type "Zn" for a zinc atom) or to output atom types which were derived from their element symbol and their chemical environment given in the input file.

The output option **-o nocat** suppresses the automatic conversion of the carbon atom in amidinium-like structures and substructures ([H_2N^+]=**C**N: N.2$^+$=**C.2**-N.pl3) to the SYBYL atom type "C.cat" (N[**C**$^+$]N: N.pl3-**C.cat**-N.pl3).

Note. The conversion to this atom type that is done by default is highly recommended. This option should only be used if the amidinium-like group is actually required with a charged nitrogen atom, *e.g.*, by any post-processing software.

The output option **-o fcharges** creates a charge column (column 9) in a SYBYL MOL/MOL2 output file containing the atom charges (*e.g.*, formal charges) given in the input file. In addition, the charge type contained under the "@<TRIPOS>MOLECULE" is set to "USER_CHARGES".

If CORINA generated 3D molecular models should be used as starting geometries of ligands for docking experiments with the flexible docking program GOLD, the output

option **-o gold** forces the automatic assignment of atom and bond types according to the GOLD conventions for difficult groups (see [35]). The SYBYL MOL2 output file of CORINA can then be directly used as input file for GOLD and ensures a proper and correct atom type assignment in GOLD.

6.6 Brookhaven Protein Data Bank Format (PDB)

The PDB format was implemented following the literature [13].

The following keywords are supported for PDB output files.

"HEADER", "COMPND", "REMARK", "HETAM", "CONECT" and "END"

The compound name is written to the "COMPND" statement.

The atom symbols and the 3D coordinates are written to "HETATM" statements.

The bond graph (connectivity information) is reflected by "CONECT" statements.

6.6.1 Options to Manipulate PDB Files

The output option **-o pdbatom** replaces all "HETATOM" statements which are set by default for the 3D coordinates of non-standard residues (groups) in biological macromolecules in PDB output files by the "ATOM" statements. This is useful if the structures will be post-processed by program systems that need to read in the "ATOM" statement or cannot handle "HETATOM" statements.

The output option **-o pdbnoconect** forces CORINA to skip the "CONECT" statements in PDB output files.

Note. The "CONECT" statements are mandatory for non-standard residues ("HETATOM") but can be neglected for standard groups ("ATOM").

The two output options **-o pdbludi** and **-o pdbludilabel** have been especially designed to generate PDB output files which can be used as input for fragment databases in the *de novo* design program system LUDI, *i.e.*, the "HEADER", "COMPND", "REMARK", "CONECT" and "END" statements are skipped and the "HETATOM" statement is replaced by "ATOM" and, if **-o pdbludilabel** is set, unique fragment labels consisting of a three letter code are generated for each input structure.

The option **-o pdbelement** adds the chemical element symbol to each "HETATM" or "ATOM" line in column 77 and 78. This conforms with the PDB specifications of version 2.3.

The output options **-o resname=<value>** and **-o resno=<value>** allow for the manipulation of the name (three letter code) and the number of the residue items used in the PDB output file.

The output option **-o keepnames** ensures that the atom names that are provided in the input file are copied to the PDB output file.

The output option **-o hlabel** labels the hydrogen atoms independently from the non-hydrogen (or heavy) atoms in the "HETATM" or "ATOM" line (columns 13 to 16). The counter for hydrogen atoms starts at the value "0", independently of the counter of the heavy atoms and allows for unique labeling of the atoms in molecules with up to 99 heavy and 99 hydrogen atoms.

The output option **-o xlabel** labels each atom type (chemical element) separately (and not only hydrogen and non-hydrogen atoms as the option **hlabel**) in the "HETATM" or "ATOM" line (columns 13 to 16). In addition, the labeling schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.

6.7 MacroModel Structure File Format (uncompressed)

The MacroModel structure file format was implemented following the literature [14].

All 58 different atom types that are defined in MacroModel are supported. In addition, three different bond types (single, double and triple bonds) that can be expressed in a valence bond notation (VB method) are supported. The first line of the file or entry contains the number of atoms in the entry and the name of the compound. The atom entries start at line 2, whereas each atom in the entry is described by one single line.

The generated Cartesian coordinates of each atom are added by CORINA in the columns 55 through 87 if the output file type is set to the MacroModel file format.

6.7.1 Options to Manipulate MacroModel Files

The output options **-o resname=<value>** and **-o resno=<value>** allow for the manipulation of the name (three letter code) and the number of the residue items used in the MacroModel output file.

The output option **-o xlabel** labels each atom type (chemical element) separately in the optional ASCII atom name field of each atom entry line. In addition, the labeling schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.

6.8 Maestro File Format

The Maestro file format was implemented following the literature [15].

The following blocks and keywords are supported.

"s_m_m2io_version", "f_m_ct and s_m_title"

"m_atom" (containing: "i_m_mmod_type", "r_m_x_coord", "r_m_y_coord", "r_m_z_coord", "i_m_residue_number", "s_m_insertion_code", "s_m_mmod_res", "s_m_chain_name", "i_m_color", "r_m_charge1", "r_m_charge2", "s_m_pdb_residue_name", "s_m_pdb_atom_name", "s_m_grow_name", "i_m_atomic_number", "i_m_formal_charge" and "s_m_atom_name")

"m_bond" block (containing "i_m_from", "i_m_to" and "i_m_order")

Similar to the MacroModel file format, each atom is described in a single line. The generated Cartesian coordinates of each atom are added by CORINA in the columns 13 through 45 if the output file type is set to the Maestro file format.

6.8.1 Options to Manipulate Maestro Files

The output options **-o resname=<value>** and **-o resno=<value>** allow for the manipulation of the name (three letter code) and the number of the residue items used in the Maestro output file.

The output option **-o xlabel** labels each atom type (chemical element) separately in the atom name field in the full Maestro CT block ("s_m_pdb_atome_name" in the "f_m_ct" block). In addition, the labeling schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.

6.9 Crystallographic Information File Format

The Crystallographic Information File (CIF) format was implemented following the literature [16] (see also "www.iucr.org/resources/cif").

6.9.1 Options to Manipulate CIF Files

The output options **-o resname=<value>** and **-o resno=<value>** allow for the manipulation of the name (three letter code) and the number of the residue items

used in the CIF output file.

With the output option **-o hlabel**, the counter for hydrogen atoms starts at the value "0", independently of the counter of the heavy atoms and, thus, allows for unique numbering of the atoms in molecules with more than 99 atoms (including hydrogen atoms).

The output option **-o novar** suppresses the writing of any torsion angle information to the loop "chem_comp_tor" statement.

With the output option **-o multor** all possible combinations of atoms belonging to a torsion angle are written out to the loop "chem_comp_tor" statement.

The output option **-o flexrta** set all torsion angles in aliphatic ring systems (with more than 4 atoms) to "var" (variable) with a period of "3" and a standard deviation "esd" of "20" degrees in the loop "chem_comp_tor".

The output file option **-o chirvol** sets the chiral volume in CIP output files to "both" for stereogenic centers which are introduced by flexible ring systems. In case of super stereogenic centers (such as six-membered rings with axial/equatorial substituents) this option sets the values in the "chiral volume" loop to "both" (instead of either "positive" or "negative". Real tetrahedral stereogenic centers are not affected by this option.

The output file option **-o coplan** sets stricter values for coplanar alkoxy group substituents at aromatic systems in CIF output files. Alkoxy group substituents at aromatic systems are usually coplanar to the aromatic ring if there is no steric crowding that forces the substituent group out of the plane (depending on the substitution pattern at the aromatic ring). With this option coplanar alkoxy groups are listed in the "plane" loop in CIF files and in the "torsion" loop an "esd" value (tolerance for torsion angle) of only 10º and a "period" of 2 is applied. This guarantees a stable coplanarity of the respective alkoxy group also during a REFMAC refinement.

The output option **-o xlabel** labels each atom type (chemical element) separately in the atom ID field in the loop "atom" of the CIF file output file. In addition, the labeling schema uses the characters set [0-9;A-Z] starting from "1" to "99", going further from "0A" to "9Z" and "A1" to "Z9" and finally from "AA" to "ZZ" to label the individual atoms. Thus, up to 1,295 instances of a specific atom type (chemical element) can be named uniquely in one record.

The output file option **-o esterplane** adds the five atoms "CC(=O)OC" of an ester groups to the "plane" loop in CIF output files. Furthermore, in the "torsion" loop a value for the parameter "value_angle" of 180°, a value for the parameter "value_angle_esd_esd" of 0° and a value for the parameter "period" of 0 is applied for the carbon-oxygen (C-O) single bond of the ester group, if the torsion angle pattern "C-C-O-C" is used in the loop "torsion". If the torsion angle pattern "0=C-O-C" is used in the loop "torsion", a value for the parameter "value_angle" of 0°, a value for the parameter "value_angle_esd_esd" of 0° and a value for the parameter "period" of 0 is set for the carbon-oxygen (C-O) single bond of the ester group. This guarantees a stable planarity of an ester group also during the REFMAC refinement.

6.10 MOPAC Cartesian File Format

The MOPAC Cartesian file format was implemented following the literature [18]. This file format is only supported for output.

6.10.1 Options to Manipulate MOPAC Cartesian Files

The output file option **-o mopackeys=<value>** allows for defining a list of MOPAC keywords that are copied to the keyword line (1st line) in the MOPAC Cartesian output file. **Note.** The individual keywords have to be separated by a white space (blank) and that the entire list has to be quoted (single – `'` – or double quotes – `"` – , respectively), *e.g.*, "`1SCF BONDS ESP VECTORS`".

The output file option **-o mopacaddchg** automatically adds the MOPAC keyword "`CHARGE=n`" to the keyword line of the MOPAC Cartesian output file according to the formal charge of the input molecule. If the user provides the keyword "`CHARGE`" using the option **-o mopackeys=<value>** (see above), the user provided value is adjusted to the actual formal charge of the molecule or the entire keyword is removed, if no formal charge is encountered (*e.g.*, because the driver option **-d neu** to neutralize specific formal charges has been switched on by the user).

The output file option **-o mopacoptflag=<value>** allows for defining the MOPAC optimization flag which is used in the MOPAC Cartesian output file. The given optimization flag "**value**" is used for each Cartesian coordinate.

MOPAC accepts a single structure in an input file. Therefore, the two output file options **-o split** and **-o splitn0=<value>** are useful options to split multi-record input files into single-record MOPAC Cartesian files.

6.11 ClearText File Format (CTX)

CTX is a keyword-oriented ASCII format developed in the research group of Prof. Dr. Johann Gasteiger [28].

The following keywords are read in and interpreted: "`IDENT`", "`NAME`", "`MOLECULS`", "`ATOMS`", "`BONDS`", "`BLABEL`", "`2DCOORD`", "`STEREO`", "`HIGEOM`", "`INTCOORD`" and "`END`". These keywords and all additional information are directly piped to the output if the output file type is also set to CTX. An additional keyword "`3DCOORD`" is written containing the generated 3D coordinates.

6.12 Interface between CORINA Classic and FlexX

The flexible ligand docking program FlexX [22] can use CORINA for the generation of low-energy conformations of ring systems with up to nine atoms per ring. During the docking process FlexX can send the cyclic parts of the ligand to CORINA module that then generates an ensemble of ring conformations. The exchange file format is SYBYL MOL2. To restrict CORINA to the ring systems of a molecule and to provide as much additional information as necessary the molecule is fragmented by FlexX according to the following rules:

1) Every ring system forms a new fragment. Two ring systems are in the same ring system if they have at least one atom in common.
2) Exocyclic substituents of a ring system and their first neighbors are included to provide the information necessary for the correct discrimination between equatorial and axial substituents.
3) All SYBYL atom and bond types of the fragment are retained as in the source molecule.

The option **-d flexx** sets all necessary program parameters to the required values. It is identical to the command line parameter sequence **-i t=mol2 -o t=mol2 -d rc,mc=25,de=30,timeout=30000**.

7 Error Messages

7.1 General Error Messages

ERROR pfopen(): Can’t open file *filename (path)*.

The program cannot open the specified file (path in parentheses).

ERROR ioopen(): Can’t open trace file.

The trace file *corina.trc* cannot be opened.

ERROR ioopen(): Identical input/output files.

Identical file names for the input and output files are not allowed.

ERROR ioopen(): Identical input/output files.

File names of input and output files are identical and CORINA does not overwrite output files.

ERROR ioopen(): Can’t open input file *filename*.

An error occurred while opening the specified input file.

ERROR ioopen(): Can’t open output file *filename*.

An error occurred while opening the specified output file.

ERROR corina(): Too many non-option parameters at *name*.

The command line must not contain more than two non-option parameters (the input and output file names). Options start with a "-". After the first two non-option parameter no more options are allowed.

ERROR corina(): Option invalid in version 1.6 and later.

An old-fashioned version 1.5 option was encountered (see section 4 on page 30).

ERROR corina(): File type not allowed.

A file type that is not allowed for input and/or output files is specified (see section 4 on page 30).

ERROR restrict(): FlexX restrictions violated.

The restrictions of the FlexX interface are violated or not fulfilled (see section 6.12 on page 81).

ERROR writeOk(): File output failed.

Something unexpected happened when trying to write out the output file.

7.2 Input File Format Error Messages

7.2.1 SD File

The error messages given below are completed by the line number in the SD file and the record number where the error occurred.

ERROR rctfile(): Can’t read counts line.

The counts line cannot be read in. The current record is discarded.

ERROR rctfile(): Can’t read atom block.

The atom block cannot be read in. The current record is discarded.

ERROR rctfile(): Can’t read bond block.

The bond block cannot be read in. The current record is discarded.

ERROR rctfile(): Can’t read 2 nd header line.

The 2nd header line cannot be read in. The current record is discarded.

ERROR rctfile(): Dimensional code (2D/3D) not specified.

The 2nd header line did not specify whether the given atomic coordinates are 2D or 3D. The program automatically checks whether z coordinates or up/down bond descriptors are given and thus, which type of coordinates have to be assumed.

ERROR rctfile(): Can't read 3rd header line.

The 3rd header line is incorrectly formatted and cannot be read in. The current record is discarded.

ERROR rctfile(): Unknown element symbol.

An unknown element symbol is encountered. The current record is discarded.

ERROR rctfile(): Bond atoms out of range.

A bond between atoms out of the range 1..N_{Atom} was encountered. The current record is discarded.

ERROR rctfile(): Unknown bond type (set 1).

A bond type not equal to 1, 2, 3, or aromatic is specified. The bond is assigned a bond order of 1.

ERROR rctfile(): Valence error reading a new bond.

The maximum valence state of an atom forming the bond was exceeded or a bond of an atom to itself was encountered. The current record is discarded.

ERROR rctfile(): Can't distribute double bonds over an aromatic system.

The distribution of alternating single and double bonds over an aromatic system failed. This may result from the fact that this distribution requires the introduction of charges. However, the best way to avoid such problems is to specify the correct valence bond notation in the input file. The current record is discarded.

ERROR rctfile(): Non-standard element.

A non-standard element symbol or atom type is encountered and assumed to be a dummy atom type unless dummy atom types are explicitly forbidden.

ERROR rctfile(): Extra character after element symbol.

The atom symbol consists of three characters and only two are allowed. The current record is discarded.

ERROR rctfile(): Dimensions code is 2D in an obvious 3D record.

The dimension flag in the 2nd header line is set to 2D, although the x, y and z coordinates are available for all atoms of the input structure.

Errors with missing columns in SD files

ERROR rctfile(): No mass difference column specified.

The 5th column in the atom block to specify isotopic mass differences is missing.

ERROR rctfile(): No atomic charge column specified.

The 6th column in the atom block to specify charges is missing.

ERROR rctfile(): No atom stereo column specified.

The 7th column in the atom block to specify atom parity stereochemical flags is missing.

ERROR rctfile(): Too many additional lines.

Only a limited number of additional lines (max. 50,000) are allowed in the data section of an SD file.

ERROR rctfile(): Line too long.

Only 80 characters per line are allowed in V2000 SD files.

Errors in charge, radical, isotope and attachment point lines in SD files

ERROR rctfile(): CHG atom out of range.

A charge (M CHG) is specified for an atom with an atom label that does not exist.

ERROR rctfile(): RAD atom out of range.

A radical (M RAD) is specified for an atom with an atom label that does not exist.

ERROR rctfile(): APO atom out of range.

An attachment point (M APO) is specified for an atom with an atom label that does not exist.

ERROR rctfile(): ISO atom out of range.

An isotope (M ISO) is specified for an atom with an atom label that does not exist.

ERROR rctfile(): Invalid chiral flag.

The value of the chiral flag is not 0 or 1.

Errors related to V3000 input SD files

ERROR rctfile(): Expected "M V30" not found.

A line starting with "M V30" was not found.

ERROR rctfile(): Expected "BEGIN CTAB" not found.

A line starting with "M V30 BEGIN CTAB" was not found.

ERROR rctfile(): Expected "END CTAB" not found.

A line starting with "M V30 END CTAB" was not found.

ERROR rctfile(): Expected "BEGIN ATOM" not found.

A line starting with "M V30 BEGIN ATOM" was not found.

ERROR rctfile(): Expected "END ATOM " not found.

A line starting with "M V30 END ATOM" was not found.

ERROR rctfile(): Expected "BEGIN BOND " not found.

A line starting with "M V30 BEGIN BOND" was not found.

ERROR rctfile(): Expected "END BOND " not found.

A line starting with "M V30 END BOND" was not found.

ERROR rctfile(): V3000 non unique atom index.

Duplicated atom index was found in the atom block.

ERROR rctfile(): Invalid atom index in bond.

A bond refers to an atom index that was not defined in the atom block.

ERROR rctfile(): V3000 templates are not supported.

A TEMPLATE definition was found.

ERROR rctfile(): V3000 sgroups are not supported.

A SGROUP definition was found.

ERROR rctfile(): V3000 rgroups are not supported.

A RGROUP definition was found.

ERROR rctfile(): Query bond type detected (type=5,6,7).

Query bonds cannot be used.

Error messages specifically related to use of stereochemical collections in V3000 SD files

Note: Parsing and checking the validity of stereochemical collections is performed only when the driver options "**stergen**" and "**v3000**" are specified.

ERROR rctfile(): Same atom index used in two different internal stereo collections: #.

The same atom index is used for two different stereochemical collections.

ERROR rctfile(): Missing data on collection line #.

Some data important for the definition of a collection in line # is not complete and the V3000 collection cannot be interpreted.

ERROR rctfile(): Expecting "<character> " in collection line #.

An expected character (*e.g.*, "=", ")") required for the correct specification of a V3000 collection is missing in line #.

ERROR rctfile(): Name subname separator not found <string> in collection.

The separator "/" between a name and a sub-name in a V3000 collection is missing.

ERROR rctfile(): Unexpected subname <string> in collection.

The wrong sub-name <string>in a V3000 collection has been encountered.

ERROR rctfile(): Expecting positive number at end of subname in collection.

A sub-name in a collection contains a number that is ≤ 0 (*e.g.*, "STERAC*n*" with $n \leq 0$).

ERROR rctfile(): Invalid field <string> found in collection (should be ATOMS, BONDS, SGROUPS, OBJ3DS, MEMBERS or RGROUPS).

An invalid field name <string> for a V3000 collection has been encountered which cannot be interpreted.

ERROR rctfile(): Number of collection objects not found.

The number that indicates the number of V3000 collection objects was not found (*e.g.*, "STERAC1 ATOMS=(**2** 2 3)" and the first "**2**" is missing). The provided V3000 stereo-chemical representation will not be used.

ERROR rctfile(): Number not found inside collection objects, was expecting <number>,
found only <number>.

The number that indicates the number of V3000 collection objects is not correct (*e.g.*, "STERAC1 ATOMS=(**4** 2 3)" specifies a total of four ("**4**") objects, but only two ("2 3") are provided). The provided V3000 stereochemical representation will not be used.

ERROR rctfile(): first number (<number>) does not match the number of objects found inside collection (<number>).

The number that indicates the number of V3000 collection objects is not correct (*e.g.*, "STERAC1 ATOMS=(**1** 2 3)" specifies a total of one ("**1**") objects, but two ("2 3") are provided).

ERROR rctfile(): Unexpected text at end of collection: <string>.

The characters <string> have been found at the end of a V3000 collection after the closing ")".

ERROR rctfile(): Unexpected subname "<string>" in collection" (should be STERAC, STEREL, HILITE or STEABS).

An invalid sub-name <string> for a V3000 collection has been encountered which cannot be interpreted.

7.2.2 SMILES

The error messages below are completed by indicating the position in the SMILES string where the error occurred.

ERROR smilesct(): Reading branch.

Error while reading a branch in "()" parenthesis. The current record is discarded.

ERROR smilesct(): General error while interpreting this character.

A not interpretable character was encountered. The current record is discarded.

ERROR smilesct(): Unknown element.

An unknown element symbol was found. The current record is discarded.

ERROR smilesct(): Too many ('s.

A closing ")" parenthesis is missing. The current record is discarded.

ERROR smilesct(): Closing ring.

No suited ring bond label. The current record is discarded.

ERROR smilesct(): Valency problem making a new bond.

The standard valency of an atom is exceeded. The current record is discarded.

ERROR smilesct(): Too many ['s.

A closing "]" bracket is missing. The record is discarded.

ERROR smilesct(): Reading bracketed atom.

The atom type in square brackets is incorrect. The current record is discarded.

ERROR smilesct(): Generating Kekule structure for an aromatic system.

A Kekulé structure with alternating single and double bonds cannot be found for an aromatic system. The current record is discarded.

ERROR smilesct(): Duplicate slash at double bond.

Two bonds at one atom of a stereochemical double bond have a slash character as stereochemical descriptor. Only one slash is allowed.

ERROR smilesct(): Not all rings closed.

Not for all ring bonds a second label was defined.

ERROR smilesct(): Reading atomic charge qualifier.

A charge of an atom is defined incorrectly. The current record is discarded.

ERROR smilesct(): Too many {'s.

A closing "}" brace is missing. The current record is discarded.

ERROR smilesct(): Unknown stereo class.

An unknown stereochemical class is specified. The current record is discarded.

ERROR smilesct(): No organic element. Use square brackets.

Inorganic elements must be written in square brackets. The current record is discarded.

ERROR smilesct(): Ring closure label not following immediately the atomic symbol.

Ring bond labels must follow immediately after the atomic symbol. The current record is discarded.

ERROR smilesct(): Conflicting ring closure bond types.

The type of the ring closure bond was defined twice with differing values. The current record is discarded.

ERROR smilesct(): Chiral center has wrong connectivity.

A tetrahedral center with less than four neighbors was encountered. A common error is to forget to specify a hydrogen atom neighboring the stereocenter within the square brackets.

ERROR smilesct(): Expecting atomic symbol.

The first character inside square brackets must be an atomic symbol. The current record is discarded.

ERROR smilesct(): Illegal valence state.

An atom in an illegal valence state was encountered. The current record is discarded.

ERROR smilesct(): Inconsistent EZ specification.

The EZ configuration of a double bond was defined redundantly with different values.

ERROR smilesct(): Isotopic mass error.

The given isotopic mass is out of range. The current record is discarded.

ERROR smilesct(): Stereo permutation not implemented.

The specifications for square-planar, trigonal-bipyramidal and octahedral chirality are not implemented and are therefore ignored.

ERROR smilesct(): Illegal hybridization.

An illegal hybridization state of an atom according to the SMARTS extensions is encountered. The current record is discarded.

ERROR readLineBasedChemicalFile(): Cannot read SMILES from column *n*.

The input option "-i sc#=*n*" or "-i scn=*n*" is set to specify the column *n* that contains the SMILES string in the input SMILES file, however, the column *n* seems to be empty.

ERROR options(): Column separator must be one character long.

A particular column separator has been specified using the option -i sep=<value> and "<value>" contains more than one character.

7.2.3 SYBYL MOL2 File

ERROR rsyb2(): Missing name in line #.

The name of the molecule is missing (1st line after the MOLECULE keyword). The current record is discarded.

ERROR rsyb2(): Missing counts in line #.

The atom and bond counts (number of atoms and bonds within the molecule) are missing (2nd line after the MOLECULE keyword). The current record is discarded.

ERROR rsyb2(): Error reading counts in line #.

A general error occurred when reading the atom lines after the ATOM keyword. The current record is discarded.

ERROR rsyb2(): Missing charge type in line #.

The definition of the type of charges is missing (4th line after the MOLECULE keyword). The current record is discarded.

ERROR rsyb2(): Can't read atoms.

A general error occurred when reading the atom lines after the ATOM keyword. The current record is discarded.

ERROR rsyb2(): Can't read bonds before atoms.

The atom block (after the ATOM keyword) has to be defined before the bond block (after the BOND keyword). The current record is discarded.

ERROR rsyb2(): Can't read bonds.

A general error occurred when reading the bond lines after the BOND keyword. The current record is discarded.

ERROR rsyb2(): Can't read sets before atoms and bonds.

Sets (SET keyword) have to be defined after the atom block (ATOM keyword) and the bond block (BOND keyword). The current record is discarded.

ERROR rsyb2(): Can't read all sets.

A general error occurred when reading the set lines after the SET keyword. The current record is discarded.

ERROR rsyb2(): Can't read rotatable bonds before atoms and bonds.

Rotatable bonds (ROTATABLE_BOND keyword) have to be defined after the atom block (ATOM keyword) and the bond block (BOND keyword). The current record is discarded.

ERROR rsyb2(): Can't convert aromatic system.

The assignment of electrons to an aromatic system failed. Probably, the VB structure is corrupted. The current record is discarded.

ERROR rsyb2(): Can't read atom in line #.

The atom in line # cannot be read. The current record is discarded.

ERROR rsyb2(): Error parsing atom in line #.

The information given for the atom in line # is not sufficient. The current record is discarded. The current record is discarded.

ERROR rsyb2(): Unknown atom type in line #.

The atom type in line # does not comply with definition of allowed SYBYL atom types. The current record is discarded.

ERROR rsyb2(): Can't read bond in line #.

The bond in line # cannot be read. The current record is discarded.

ERROR rsyb2(): Error parsing bond in line #.

The information given for the bond in line # is not sufficient. The current record is discarded.

ERROR rsyb2(): Unknown bond type.

The bond type in line # does not comply with definition of allowed SYBYL bond types. The current record is discarded.

ERROR rsyb2(): Valence problem reading bond #-#.

The assignment of the electrons for the bond between the atoms # and # failed. Probably, the VB structure is corrupted. The current record is discarded.

ERROR rsyb2(): Reading set header.

A general error occurred while reading the header for a set (SET keyword). The current record is discarded.

ERROR rsyb2(): Too few tokens in set header.

The information given in the set header (SET keyword) is not sufficient. The current record is discarded.

7.2.4 InChI file

ERROR readLineBasedChemicalFile(): Cannot read InChI from column *n*.

The input option "-i sc#=*n*" or "-i scn=*n*" is set to specify the column *n* as the column that contains the InChI string in the input InChI file, however, the column n seems to be empty.

ERROR options(): Column separator must be one character long.

A particular column separator has been specified using the option "-i sep=<value>" and "<value>" contains more than one character.

7.3 Error Messages Related to Stereochemistry

Some further information on the handling of stereochemistry by CORINA and for the interpretation of the following error messages is given in section 12.4 on Page 126.

ERROR initster(): Resetting a trans double bond in a small ring: #-#.

A *trans* double bond in a ring with less than eight atoms was encountered. Since this is geometrically impossible it can be corrected by CORINA.

ERROR initster(): More axes than expected at atom #.

The maximum number of six neighboring atoms is exceeded. The current record is discarded.

ERROR initster(): No suitable configuration for atom #.

No suitable configuration for atom # can be derived. The number of axes does not correspond to the number of neighboring atoms. The current record is discarded.

ERROR clcster2(): Ambiguous configuration of ligands.

The specification of a chiral center is ambiguous in the input 2D structure diagram and must be ignored.

ERROR clcster2(): The ligands at atom # don't span up a volume.

The specification of a chiral center is ambiguous in the input 2D structure diagram. CORINA tries to correct it.

ERROR clcster2(): Insufficient number of axes at atom #.

The number of axes for a higher coordinated center # (5 or 6 ligand atoms) that are specified in the input 2D structure diagram is not enough. CORINA tries to correct it.

ERROR clcster3(): Insufficient number of axes at atom #.

The number of axes for atom # that are specified in the input 3D structure is too small. CORINA tries to correct it.

ERROR allcis(): No suited stereo descriptors for bridge #- -#.

The coding of the stereochemistry of the bridgehead atoms #-# is incorrect. No 3D coordinates can be generated.

ERROR stergen(): Maximum number of # stereo centers exceeded.

By default, the stereoisomer generator in CORINA only accepts a maximum of four (4) stereocenters per molecule unless specified differently with the driver option "-d msc=<value>". If the maximum number of stereocenters is exceeded, only one stereoisomer is generated.

7.4 Error Messages Related to Generation of 3D Coordinates

ERROR corina(): Input structure incorrect. No 3D generation.

The input structure is corrupted, and CORINA cannot generate a 3D structure.

ERROR gen3d(): Unable to build a 3D structure/single conformation.

No 3D structure can be generated. A fatal error occurred during the generation process.

ERROR gen3d(): 3D structure didn‘t pass the quality check.

The quality of generated 3D structure is insufficient and not written to the output file.

ERROR gen3d(): No generation of multiple conformations for multiple fragments.

For records containing multiple fragments the output of multiple ring conformations (-d rc) is not supported. Work-around: remove all but the largest fragments (-d rc,rs).

Error gen3d(): To many atoms (max. 999).

The number of atoms in the current molecule exceeded the internal limitation of 999. This limitation can be re-defined by the user with the driver option -d maxat=<value> (*e.g.*, -d maxat=2000).

ERROR quality(): Some internal coordinates are heavily distorted.

A fatal error occurred during the 3D generation. No 3D structure can be generated.

ERROR quality(): Violated cis/trans bond.

A stereogenic double bond changed its configuration during the geometry optimization. No 3D structure is output.

ERROR quality(): Violated stereo atom.

A chiral atom changed its configuration during the geometry optimization. No 3D structure is output.

ERROR quality(): Bad contacts detected.

Unfavorable non-bonded interactions (crowded atoms) are detected in the generated 3D structure and therefore not written to the output file.

ERROR alltempl(): Missing ring template.

For one or more rings no suitable ring template can be found in the list of predefined ring templates. No 3D structure can be generated for the current record.

ERROR ringfrag(): Fragment contains ring(s) > 9.

CORINA handles only rings up to a size of nine ring atoms members by predefined ring templates. Larger rings are reduced to a secondary structure that have less than ten anchor atoms (see section 12.1.2 on page 119). This reduction failed. No 3D structure can be generated for this record.

ERROR ringfrag(): Can’t combine the templates.

The available ring templates cannot be combined to a one single ring conformation. No 3D structure can be generated for this record.

ERROR bigring(): Unable to build the secondary structure.

The secondary structure of a large ring system with ring sizes of more than 9 ring atoms is too complex or the secondary structure contains rings with more than 9 ring atoms. No 3D structure can be generated for this record.

ERROR procfrag(): Can’t process small ring system.

A small ring system cannot be translated into a 3D structure. No 3D structure can be generated for this record.

ERROR bigsys(): Unable to process all fragments.

Some of the fragments of a large ring system cannot be translated into a 3D structure. No 3D structure can be generated for this record.

ERROR getconf(): Can't build ring fragment.

For a small ring system, no suitable ring conformation could be build up and translated into a 3D structure. No 3D structure can be generated for this record.

7.5 Output File Format Error Messages

7.5.1 SD File

ERROR wctfile(): Cannot write more than 999 atoms.

The current molecule has more than 999 atoms which exceed the limitations of a V2000 SD file.

8 Warning Messages

8.1 General Warning Messages

WARNING mnLicenseCheck(): License expired on <date>! Exiting ...

The license time has on <date> expired. The usage of the program is not allowed anymore.

8.2 Warning Messages Related to Input File Format and Processing

8.2.1 SD File

Warnings related to V3000 SD input files

WARNING rctfile(): Any bond type ignored (type=8).

This bond will be ignored to compute the 3D structure.

WARNING rctfile(): Hydrogen bond type ignored (type=10).

This bond will be ignored to compute the 3D structure.

WARNING rctfile(): Aromatic query bond type detected (type=4).

Although the aromatic type is a query feature, the bond is accepted and will be handled as aromatic.

WARNING rctfile(): Coordination bond type (type=9) set to single bond.

CORINA does not support coordination bonds. A bond order of "1" will be used for the 3D structure.

8.2.2 SMILES

WARNING smilesct(): Non-standard formal charge qualifier in curly brackets.

A charge value is defined in "{}" braces (curly brackets) instead of in "[ ]" square brackets. This does not comply with the SMILES language definition. CORINA tries to correct it.

WARNING smilesct(): Unnormal valence state.

An unnormal valence state is encountered. The number of bonds exceeds the free valences of an atom. CORINA tries to correct it.

WARNING smilesct(): Un-paired label--inserting APO.

An unclosed ring system is encountered since the second label for ring closure is missing. An attachment point (APO) is inserted in the internal CT representation to take into account the open valence.

WARNING smilesct(): Duplicate slash at double bond--second one ignored.

A *cis* or *trans* double bond is marked with two consecutive slash characters ("//" or "\\"). The second slash character is ignored to derive a proper definition of the double bond configuration.

WARNING smilesct(): Ignoring chirality at atom with more than one H.

An atom that is marked as chiral has more than one bonded hydrogen atom. Therefore, the chirality is ignored.

WARNING smilesct(): Incomplete EZ specification—ignored.

A *cis* or *trans* double bond is incompletely specified, *i.e.*, one of the two mandatory slash characters is missing. Therefore, the descriptors are ignored.

WARNING smilesct(): Label following branch.

A label indicating a ring closure that is directly placed behind a branch is encountered. This combination might cause problems, but CORINA tries to solve this problem.

WARNING options(): Input name column number is set to 0 and output option sdfn2i is set, no name will be written into data item <data item>.

The option "nc#=0" or "ncn=0" is set to suppress the copying of the compound name from the input to the output file, however, the output option "sdfn2i=<data item>" is set additionally. The two options are contradictory and the output option "sdfn2i=<data item>" is ignored, *i.e.*, no compound name is written into the data item <data item> of the output SD file.

WARNING options(): Input comment column number is set to 0 and output option sdfc2i is set, no comments will be written into data item <data item>.

The option "cc#=0" or "ccn=0" is set to suppress the copying of the comment from the input to the output file, however, the output option "sdfc2i=<data item>" is set additionally. The two options are contradictory and the output option "sdfc2i=<data item>" is ignored, *i.e.*, no comment is written into the data item <data item> of the output SD file.

WARNING options(): Input comment column number is set to 0 and output option pascom is set, no comments will be passed to the output.

The option "cc#=0" or "ccn=0" is set to suppress the copying of the comment from the input to the output file, however, the output option "pascom" is set additionally. The two options are contradictory and the output option "pascom" is ignored, *i.e.*, no comment is written into output file.

WARNING readLineBasedChemicalFile(): Cannot read name from column *n* (SMILES).

The option "nc#= *n* " or "ncn= *n* " is set, but column number *n* does not exist or is empty. No compound name is written to the output file.

WARNING readLineBasedChemicalFile(): Cannot read comment from column *n* (SMILES).

The option "cc#= *n* " or "ccn= *n* " is set, but column number *n* does not exist or is empty. No comment is written to the output file.

8.2.3 InChI File

WARNING options(): Input name column number is set to 0 and output option sdfn2i is set, no name will be written into data item <data item>.

The option "nc#=0" or "ncn=0" is set to suppress the copying of the compound name from the input to the output file, however, the output option "sdfn2i=<data item>" is set additionally. The two options are contradictory and the output option "sdfn2i=<data item>" is ignored, *i.e.*, no compound name is written into the data item <data item> of the output SD file.

WARNING options(): Input comment column number is set to 0 and output option sdfc2i is set, no comments will be written into data item <data item>.

The option "cc#=0" or "ccn=0" is set to suppress the copying of the comment from the input to the output file, however, the output option "sdfc2i=<data item>" is set additionally. The two options are contradictory and the output option "sdfc2i=<data item>" is ignored, *i.e.*, no comment is written into the data item <data item> of the output SD file.

WARNING options(): Input comment column number is set to 0 and output option pascom is set, no comments will be passed to the output.

The option "cc#=0" or "ccn=0" is set to suppress the copying of the comment from the input to the output file, however, the output option "pascom" is set additionally. The two options are contradictory and the output option "pascom" is ignored, *i.e.*, no comment is written into output file.

WARNING readLineBasedChemicalFile(): Cannot read name from column *n* (InChI).

The option "nc#= *n* " or "ncn= *n* " is set, but column number *n* does not exist or is empty. No compound name is written to the output file.

WARNING readLineBasedChemicalFile(): Cannot read comment from column *n* (InChI).

The option "cc#= *n* " or "ccn= *n* " is set, but column number *n* does not exist or is empty. No comment is written to the output file.

8.2.4 SYBYL MOL2 File

WARNING rsyb2(): Discarding record due to problems with aromatic system.

The assignment of electrons to an aromatic system failed. The current record is discarded.

WARNING rsyb2(): Discarding record due to dummy atoms/bonds.

The record has to be discarded due to dummy atom and/or bond types that cannot be interpreted. The current record is discarded.

WARNING rsyb2(): Improper atom and bond types.

Some atom and/or bond types do not comply with the definition of allowed SYBYL atom and bond types. CORINA tries to derive correct types.

WARNING rsyb2(): Dummy atom in line <#> interpreted as <type> from atom name.

The dummy atom type in line # is interpreted as atom symbol <type> derived from the atom name. This message only appears if the input option "xelement" is set.

WARNING rsyb2(): Unknown atom type in line <#> (<type>) interpreted as element symbol.

The unknown SYBYL atom type <type> in line # is interpreted as an element symbol. This message only appears if the input option "xelement" is set.

WARNING sybchkn(): Setting atom <#> from <type> to <type> based on 3D.

Based on the input 3D structure, the SYBYL atom type of atom # does not match to the geometry of the atom. Therefore, the atom type is changed internally (*e.g.*, N.pl3 to N.3).

WARNING sybplaus(): Probably wrong SYBYL type <type> at atom #.

The SYBYL atom type <type> at atom # may be wrong due to geometric reasons. It seems that the atom has more neighbors than its geometry allows. CORINA tries to derive a proper atom type.

8.3 Warning Messages Related to Stereochemistry

Some further information on the handling of stereochemistry by CORINA and for the interpretation of the following warning messages is given in section 12.4 on Page 126.

WARNING initster(): Stereo atom # without stereo descriptor.

No stereochemical descriptor (parity flag or wedge symbol) is given for the chiral atom #. The output 3D structure might have an unexpected configuration since CORINA has to use default rules or assume an arbitrary stereochemistry.

WARNING clcster2(): Possibly stereo problem at atom #.

A general problem while calculating the configuration of a stereocenter from the 2D structure diagram was encountered. CORINA tries to correct it (see below).

WARNING clcster2(): Trying to ignore H-atom at stereo center #.

An ambiguous 2D configuration was encountered. CORINA tries to ignore the 2D coordinates of one hydrogen atom in order to solve the problem since these hydrogen atoms are often automatically added to the 2D structure diagram without regarding the stereochemistry of the central atom.

WARNING clcster2(): Trying to give a direction to bond #-#.

An ambiguous 2D configuration was encountered. The program tries to assign a wedge descriptor to an additional bond to solve the problem.

WARNING clcster2(): Trying to correct by moving the central atom #.

An ambiguous 2D configuration was encountered. The program tries to correct it by moving the central atom #.

WARNING clcster2(): Collision of wedge symbol and stereo descriptor at atom #.

The stereochemistry of an atom was defined by an up/down (wedge) bond descriptor in the 2D structure diagram and by an atom parity descriptor (parity flag) with the two specifications giving opposite configurations. The atom descriptor has the higher priority and overrides the bond descriptor.

WARNING clcster2(): Number of axes at atom # not sufficient.

The number of axes at a coordination center higher than 4 (5 or 6 ligand atoms) in the 2D structure diagram is too small.

WARNING clcster2(): Wedge symbol(s) pointing with the basis to the stereo center.

Wedge symbols that are pointing with the basis (broad end) to the stereocenter lead to an ambiguous definition of stereochemistry. CORINA tries to correct this by finding an alternative coding (wedge symbol) for this center.

WARNING clcster3(): Collision of implicit stereo descriptor and 3D coordinates at atom #.

If 3D coordinates are given in the input structure CORINA checks whether they match to the stereochemical descriptors. By default, the stereochemical descriptors are used, however, if the driver option "-d 3dst" is set the stereochemical information is calculated from the 3D coordinates.

WARNING clcster3(): Number of possible axes at atom # insufficient.

The number of axes for atom # that are specified in the input 3D structure is too small. CORINA tries to correct this.

8.4 Warning Messages for the Generation of 3D Coordinates

The following messages are general warnings which may occur during the process of generating the 3D atomic coordinates including all preparatory steps.

WARNING bondlen(): No bond length #-#.

The system is unable to calculate a bond length from standard atomic parameters.

WARNING initba(): Geometry type for atom # PLANAR --> TETRAEDER changed.

The bridgehead atoms in strained ring systems cannot be planar. This warning occurs in the case of unsaturated ring systems containing rings smaller than 5, *e.g.*, cubene.

WARNING getta(): No TA #-#-#-#.

In case of linear systems (*e.g.*, acetylene) torsion angles (TA) cannot be specified for some bonds.

WARNING hmoboord(): No HMO constant for atom #.

No Coulomb integral parameter is found for a hetero atom. CORINA uses the value for carbon.

WARNING hmoboord(): No HMO constant for bond #-#.

No parameter for a bond resonance integral is found. The value for "C-C" is used instead.

WARNING genconf(): Time out.

The time limit for a complete ring conformation analysis was exceeded. The conformer with lowest energy so far found is not necessarily the global minimum.

WARNING rrefine(): # pair(s) of crowded atoms.

Some pairs of atoms came closer to each other than 75% of their atomic radii.

WARNING canon(): Atom re-numbering failed.

An error occurred during the canonization process of the connection table (CT). The original atom and bond numbering is used for this record.

WARNING gen3d(): Re-ordering CT to build 3D structure (*N*).

No valid 3D structure could be generated with the internally canonized ordering of the connection table (CT). The CT is re-ordered to try to build a 3D structure ("*N*" is the number of trials, maximum of 10).

9 Technical Requirements

9.1 System Requirements

CORINA is a command line tool and has to be executed in a shell (*e.g.*, csh, tcsh, or bash on UNIX/Linux systems) or at a Windows command prompt (see also section 3 on page 27). The following hardware platforms and operating systems are supported.

- x86 platforms with Microsoft Windows 10 or 11, Server 2008 (win64, it is recommended to have installed the latest service pack)
- x86 platforms with Linux distributions RHEL6 (32 and 64bit) and RHEL7 (32 and 64bit)

9.2 Program Scope and Known Limitations

CORINA has been designed to process a broad range of chemistry. The core scope are organic, small to medium-sized, typically drug-like molecules.

In general, there are no limitations concerning the number of atoms or bonds of a molecule that can be processed. For large structures and huge datasets, it is recommended to use the 64bit version of CORINA.

Note. Some structure file formats that are supported as well as hardware and operating system related issues may introduce such limitations regarding the number of atoms and bonds in a molecule (*e.g.*, SD V2000 can only handle molecules up to 999 atoms and/or bonds).

The periodic table is parameterized up to atomic number 103 (Lawrencium).

Organic chemical compounds that can be correctly expressed and described in a correct and valid valence bond (VB) notation can be processed.

Stereo-chemical information (tetrahedral centers and *cis*/*trans* double bonds) is fully considered as far as the supported file formats allow for the definition of stereochemistry. Stereo-chemical information may also be derived from 3D input structures.

Optionally, stereoisomeric compounds can be enumerated.

Atoms with up to six neighbors can be processed.

Multi-fragment structures (*e.g.*, salts, records with solvent molecules,…) can be processed.

By default, a single (low-energy) conformation is generated for each input structure. For ring systems that have less than ten ring atoms, a limited number of reasonable multiple conformations can be generated optionally.

Perspective 2D structure diagrams and class-specific drawing styles (such as Fisher or Haworth projections, *etc.*) cannot be interpreted in terms of their stereochemistry or

the relative positions of the substituents. IUPAC recommends avoiding such perspective 2D structure diagrams and representations, if the structures should be interpreted by computers [36].

There is no general guarantee that a specific compound or compound class can be processed, and 3D coordinates can be generated for such a compound or compound class.

10 Program Installation

10.1 Download from MN-AM.com

CORINA Classic is available for download from the web server of MN-AM.com (Download Area). An account can be requested at

www.mn-am.com/php/profile.php

An account provides access to licensed software, evaluation copies, program manuals, example files and tutorials of CORINA Classic as well as to test copies of a variety of chemoinformatics applications offered by MN-AM.com.

The software packages are provided as compressed files. The downloaded files can be easily uncompressed with standard software tools for file compressing and archiving, such as WinZip, FileZip ("www.filezip.com"), or gzip ("www.gzip.org").

CORINA is currently available for the following hardware platforms and operating systems.

Platform & architecture	Download file name
x86-64 Linux RHEL7/CentOS7 (gcc 4.8.5, libc 2.17, 64bit)	*corina_<annual|eval>_x86-64_rhel7_<date>.lnx.gz*
x86 Linux, RHEL6/CentOS6 (gcc 4.4.7, libc 2.12, 32bit)	*corina_<annual|eval>_x86-32_rhel6_<date>.lnx.gz*
x86-64 Linux RHEL6/CentOS6 (gcc 4.4.7, libc 2.12, 64bit)	*corina_<annual|eval>_x86-64_rhel6_<date>.lnx.gz*
Microsoft Windows win64, 10/11 (gcc 9.3)	*corina_<annual|eval>_Win64_<date>.exe.gz*

corina_<annual|eval>_<OS>_<date>.lnx|exe.gz

(*annual* = licensed version with annual run time; *eval* = evaluation version)

Additional information such as this program manual in PDF format or examples of structure files (see section 3 on page 27) can be downloaded from the web server of MN-AM ("www.mn-am.com").

10.2 New Installation on Linux Systems (x86)

CORINA is a command line program (executable file *corina*) and executed in a Linux shell (*e.g.*, *csh*, *tcsh*, or *bash*). The usage of CORINA as well as all available command line options are described in detail in the sections 3 "Getting Started with CORINA" and 4 "Using CORINA" of this manual.

CORINA for Linux is available as a 32 bit or 64bit application.

To install the command line version of CORINA (*corina.lnx*) please follow the instructions below.

1) Copy the downloaded file of CORINA Classic to your Linux computer and extract it there.

   Example: *gunzip corina_eval_x86-64_rhel7_2021-12-31.lnx.gz*

   Note. The extracted file *corina_eval_x86-64_rhel7_2021-12-31.lnx.gz* is a binary file.

2) Rename the resulting CORINA Classic executable file to, *e.g.*, *corina*.

   Note. Any file name is possible.

3) Create a subdirectory, *e.g.*, *corina*, (for users with admin access when installing software locally, *e.g.*, */opt/mn-am/corina*).

4) Copy the executable file of CORINA Classic *corina* to the subdirectory *corina*.

5) Add the full name of the subdirectory c*orina* (e.g., */opt/mn-am/corina*) to the environment variable *PATH* in your *.login* or *.cshrc* files (*.profile* or *.bashrc*) to make CORINA Classic available system wide.

10.3 New Installation on Microsoft Windows Platforms (7/8/10)

CORINA is a command line program (executable file *corina.exe*) and is executed at a Windows command prompt. The usage of CORINA as well as all available command line options are described in detail in the sections 3 "Getting Started with CORINA" and 4 "Using CORINA" of this manual.

In the following, the installation procedure is described for the 32bit version of CORINA. The installation for the 64bit version is analogous with the only difference that the installation directory is "*C:\Program Files*" and not "*C:\Program Files (x86)*" as for 32bit applications.

To install the command line version *corina.exe*, please follow the instructions below.

1) Uncompress the CORINA Classic download file.

   Sample: corina_eval_Win32_2024-08-31.exe.gz

   Note. corina_eval_Win32_2024-08-31.exe is a binary file.

2) Rename the resulting CORINA Classic executable file to *corina.exe*.

3) Create a subdirectory *corina* in the Windows program folder, *e.g.*, "*C:\Program Files (x86)\corina*" for the Windows 32bit application, (see Figure 5).

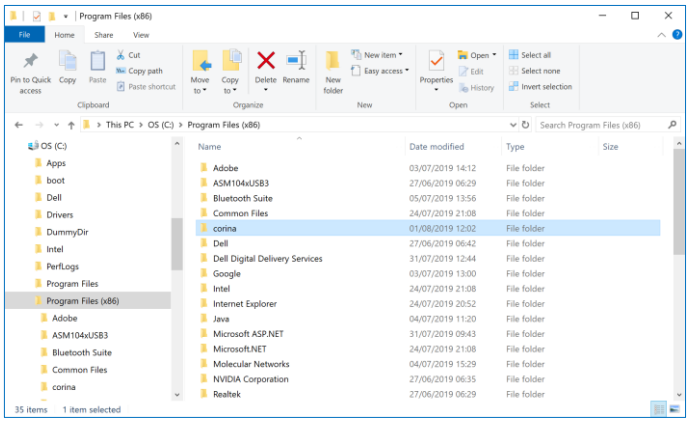


Figure 5 Create a new sub-directory *corina* in the program folder.

4) Copy the file *corina.exe* into the new sub-directory *corina*.

5) In order to execute CORINA from any other directory on the computer, add the sub-directory *corina* where the CORINA executable file is located (*e.g.*, *C:\Program Files (x86)\corina*) to the environment variable *Path* of the system settings as following.

   a) Open the "Start" menu of the Windows system, then select "Control Panel" → "System" → "System" and click on the link "Advanced system settings" in the upper left part of the control panel. The "Systems Properties" dialog appears (see Figure 6).

   b) Select the tab "Advanced" in the "Systems Properties" dialog and press the button "Environment Variables…" (see Figure 6).

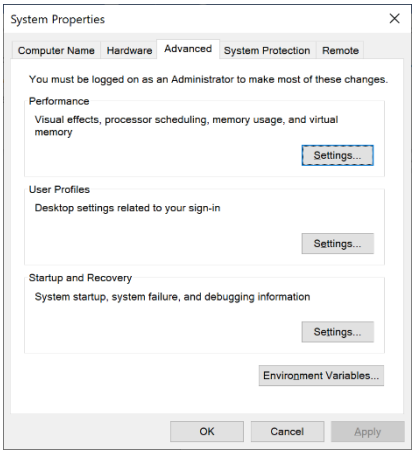


Figure 6 The "System Properties" dialog.

c) The "Environment Variables" dialog appears. Select "Path" in the list of "System variables" and click on the button "Edit" (see Figure 7 left).

d) For Windows 7

Add the full path of the sub-directory *corina* (*e.g., C:\Program Files\corina*) at the end of the field "Variable value" (see Figure 7 center).

Ensure not to delete the current "Variable value" text.

Note. The newly added path variable must be separated by the character ";" (semi colon) from the existing path variables.

e) For Windows 10

Click on the button "New" in the top-right corner of the dialog window and add the full path of the sub-directory "corina" (*e.g., C:\Program Files\corina*) to the empty field (see Figure 7 right).

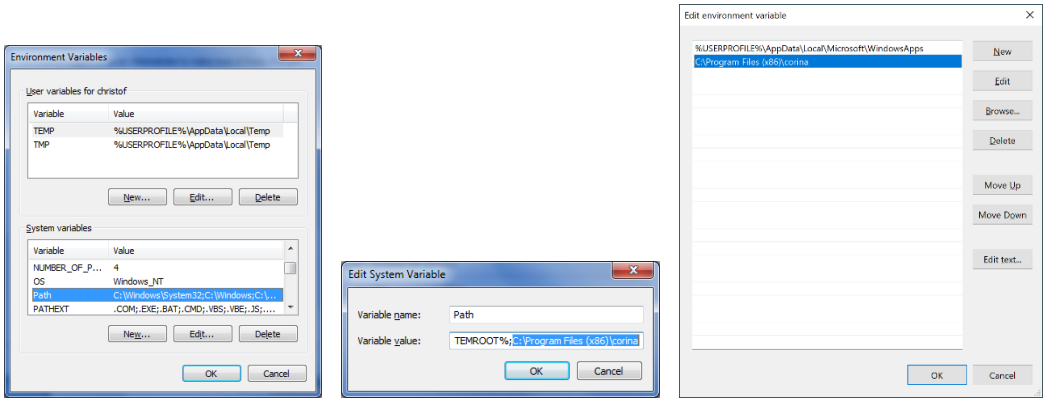


Figure 7 Specifying the "Environment Variable" for CORINA.

e) Confirm all changes by clicking the button "Ok" and close the "Control Panel".

5) To create a short cut on the desktop to open a Windows command prompt in which CORINA can be executed, please follow the instructions below.

a) Right-click on the desktop, select "New" in the context menu and select "Shortcut". Enter the command "%COMSPEC%" in the "Location" field of the "Create Shortcut" wizard and press the button "Next" (see Figure 8).

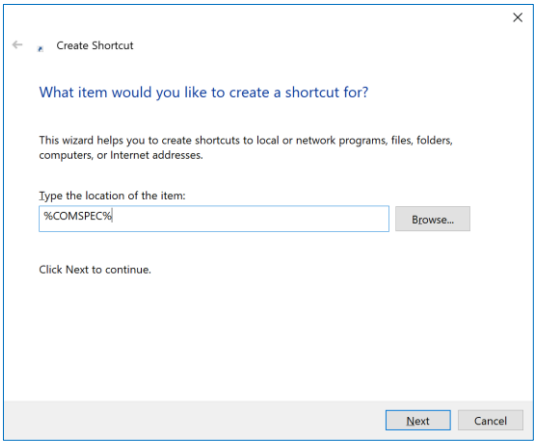


Figure 8 Create a shortcut on the desktop.

b) Enter a name for the shortcut, *e.g.*, "CORINA Classic" and press the button "Finish" (see Figure 9).

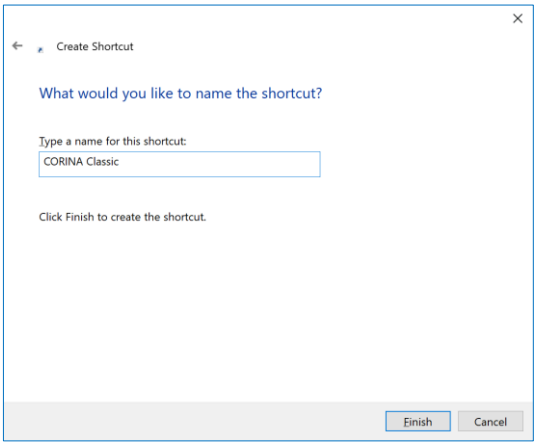


Figure 9 Create a shortcut on the desktop.

c) Double-click the newly created desktop shortcut "CORINA Classic Command Line" and type the command "`corina -v`" at the prompt. The version number of CORINA is printed on the screen (see Figure 10).

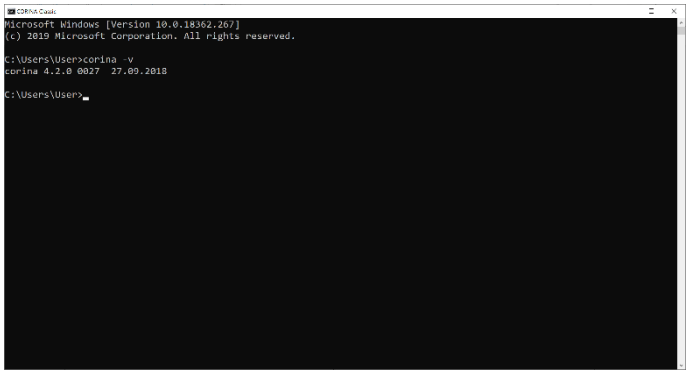


Figure 10 Start command prompt and test CORINA.

d) By right-clicking the desktop shortcut "CORINA Classic" and selecting "Properties" in the context menu, the directory in which the command line prompt should be started (*e.g.*, in the directory "User") can be specified in the field "Start in" in the tab "Shortcut" of the "Properties" dialog (see Figure 11). For other settings that can be specified in this dialog, please consolidate the Windows help.

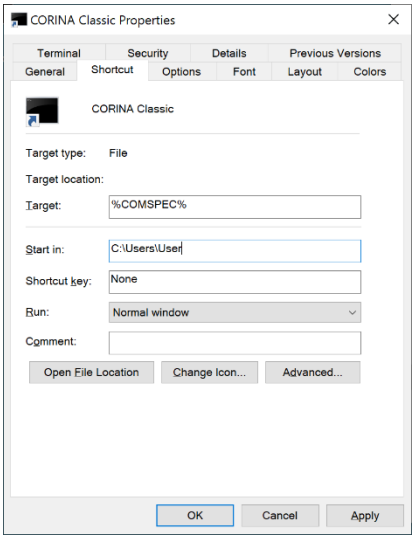


Figure 11 Setting the properties for the CORINA command line shortcut.

10.4 Program Updates

1) Before installing the new version, please copy the old executable and configuration files to a new directory, *e.g.*, *corinaVVV* (*VVV* = *old-version-number*, *e.g.*, *corina410*).

2) According to the hardware platform install the new version following the installation instructions given in section 10.2 on page 111.

   Note. Since CORINA version 2.4, the data files *stdval.ctx* and *rings.ctx* are no longer part of the distribution. All data have been included in the binary file of CORINA (see section "Version 2.6" on page 5).

11 Problems and Help!

If you have any difficulties with the installation of CORINA or if you encounter any problems when running CORINA, please send all your inquiries to the following address:

Molecular Networks GmbH (MN-AM)
Neumeyerstr. 22-34
90411 Nuremberg
Germany

or contact us by email support@mn-am.com,
or by fax +49 911 597 424 09

Please include the input file, the output file, and the CORINA trace file *corina.trc* generated by CORINA to us by email. These files will help us to analyze your problem; if your system displays any error messages, please add them to your report. Thank you!

You can also use the report form in section 15 on page 160 of this manual.

12 Understanding CORINA Classic

12.1 Fast and Efficient Generation of High-Quality 3D Molecular Models

12.1.1 The Core System

CORINA can be regarded as a fully automatic 3D model building kit. By combining mono-centric fragments with standard bond lengths and angles and by using appropriate dihedral angles a 3D model of a molecule is built. Bond lengths and angles possess only one rigid minimum and are taken from tables. Since multiple solutions exist for torsion angles, two major problems arise. First, in ring systems only restricted sets of torsion angles are allowed to ensure proper ring closure. Secondly, non-bonded interactions due to flexible chain portions need to be minimized. Therefore, CORINA handles rings and chains separately.

After reading in the input structure, the connection table (atoms and bonds) are canonized internally. This avoids any atom numbering dependent artifacts in the 3D structure generation process. Before the final 3D structure is written to the output file, the original atom (and bond) ordering and numbering is restored, *i.e.*, the original atom and bond ordering and numbering from the input structure is preserved in the output 3D structure.

Rings of up to a size of nine atoms are processed by using a table of single ring conformations that implicitly ensure ring closure. In the case of fused or bridged systems, a backtracking search procedure finds a contradiction-free set of conformations for each single ring following some geometric and energy restrictions.

Since this strategy works on the torsion angle representations of the ring conformations and uses only logical operations and integer arithmetic, it is extremely fast. The ring systems are then translated into 3D coordinates (including fused and spiro-systems) and a set of different but allowed conformations are considered. These conformations are then further refined applying a simplified *pseudo* force field that contains only special geometric terms for the optimization of ring systems. The major goal of this optimization process is to relax the ring geometries and to identify a low-energy conformation.

For acyclic fragments and molecules, the principle of longest pathways has been implemented (see Figure 12). The main chains are extended as much as possible by setting the torsion angles to *anti* or *trans* configurations unless a *cis* double bond is specified. This method effectively minimizes non-bonding interactions, such as atom overlaps or close contacts.

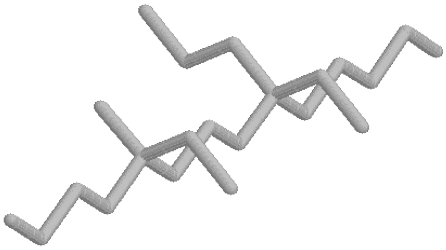


Figure 12 Principle of longest pathways for acyclic structures and fragments.

After the combination of the three-dimensional fragments of the ring systems and of the acyclic parts, the entire 3D model is checked for overlapping atoms and for close contacts. If such situations are detected, CORINA performs a reduced conformational analysis to avoid these interactions.

- First, a strategic rotatable bond within the pathway connecting the two interacting atoms is determined, depending on topological features and double bond character.
- Secondly, the torsion angle of this bond is changed until the non-bonded interactions are eliminated (see Figure 13).

For appropriate torsion angles, CORINA uses a set of rules and data obtained from a statistical analysis of the conformational preferences of open-chain portions in small molecule crystal structures. This knowledge was derived from the Cambridge Structural Database (CSD) and is stored in the Torsion Angle Library [1],[23],[29].

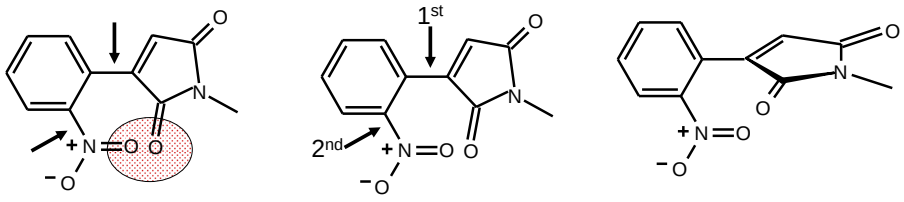


Figure 13 Reduced conformational analysis to avoid non-bonded interactions.

12.1.2 The Challenge: Large Rings

Large rings represent a special challenge and most of the other published 3D structure generators fail to process such systems. The conventional approach of taking small ring conformations from a table will not work for large and flexible rings. The ring table

used by CORINA contains conformations only for ring systems with up to nine ring atoms.

Therefore, for larger rings a different method is necessary. However, polymacrocyclic structures often show a general outline, a superstructure [7]. The porphyrine bridged cyclophane molecule in Figure 14 (left) shows a cage-like superstructure that retains the approximate shape and symmetry of the entire system (Figure 14, right).

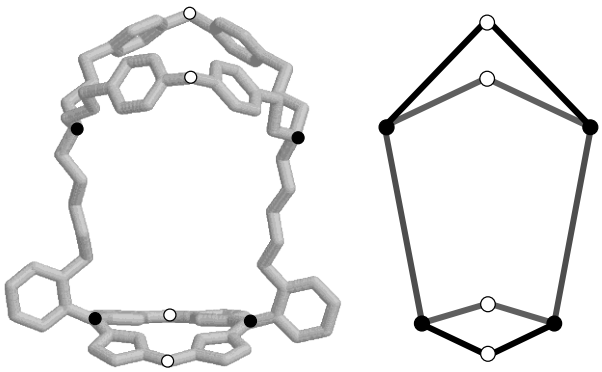
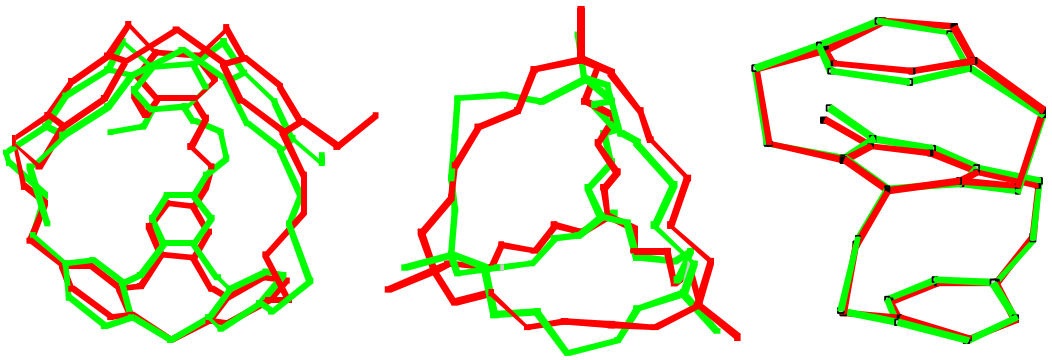


Figure 14 A macrocyclic molecule and the corresponding superstructure.

The procedure for generating a 3D structure for polymacrocycles follows the so-called "principle of superstructure".

- First, anchor atoms are determined (mostly bridge-head atoms, marked as circles in Figure 14) and the ring system is reduced to its superstructure as shown in Figure 14.
- Then, a 3D model for the superstructure that contains only small rings (up to 9 ring atoms) can be generated applying the methods for small rings.
- Finally, the removed atoms are restored, and a complete 3D model of the entire ring system is obtained.

Figure 15 compares the X-ray structures of three polymacrocycles with the corresponding CORINA models and shows the RMS_{XYZ} deviations between them. Although rather large RMS_{XYZ} deviations of 0.14 to 0.95 Å are measured, CORINA succeeds to correctly predict the overall shape and symmetry of the polymacrocyclic structures.



DIJJUB: RMS_{XYZ} = 0.69 Å CISZUZ: RMS_{XYZ} = 0.95 Å FEFZIZ: RMS_{XYZ} = 0.14 Å

Figure 15 Comparison of the X-ray structures of three polymacrocyclic systems with the corresponding CORINA models and their RMS_{XYZ} deviation. (Note. The experimental structure of CISZUZ contains an I_3^- anion inside the ring system that is not modeled by CORINA.)

12.1.3 Another Challenge: Metal Complexes

Another type of structure classes commonly neglected by conventional structure generators are organometallic compounds. CORINA can process compounds containing atoms with up to six neighbors. Thus, metal complexes with up to octahedral centers can be handled. The extensions made are quite simple: First, the input structures must fulfill the restrictions of the valence bond concept. Secondly, appropriate mono-centric geometries are predicted for the metal centers. Third, the lengths of metal-ligand bonds are corrected by specific factors taking into consideration their non-covalent character. The resulting structures correspond quite well to the experimentally determined geometries. Figure 16 shows three examples: a nickel, a ruthenium and a rhodium complex and the *RMS_{XYZ}* deviations from the X-ray structures.

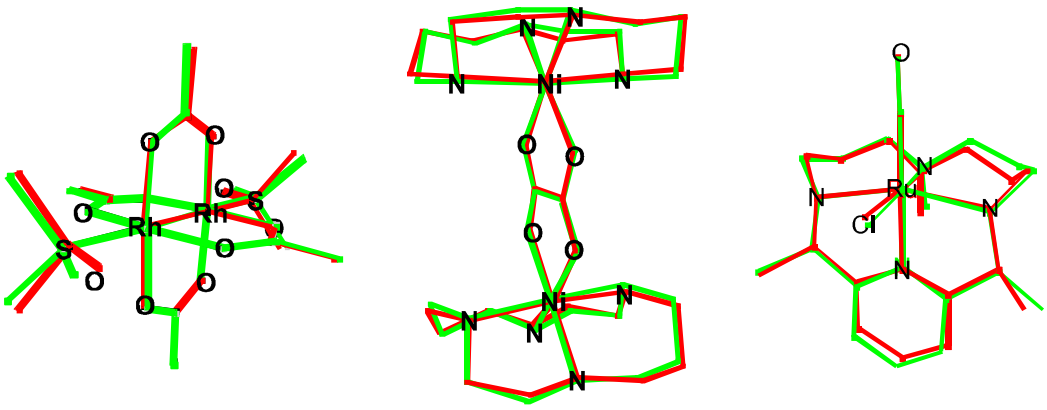


Figure 16 Comparison of the X-ray structures of three metal complexes with the corresponding CORINA models and their RMS deviation.

12.2 Performance in Reproducing X-ray Structures

To test the performance of CORINA in terms of reproducing experimentally determined 3D structures, an evaluation study was performed in year 2001 based on a dataset of 25,017 X-ray structures. In comparison to some previous tests with smaller datasets, the dataset should provide less bias and a more realistic impression of the performance of the program under real-world conditions: CORINA is designed to convert millions of structures as fast as possible while maintaining a good quality.

12.2.1 Dataset

The new dataset was obtained from the Cambridge Structural Database using the retrieval program QUEST in batch mode [1]. The query was a combination of screens which selected error-free organic compounds which had been fully resolved, for which the connection table had been completely assigned and which had an R-factor of less than or equal to 5%.

The compounds were exported in SYBYL MOL2 format. Initially, a dataset of 36,085 compounds was obtained. They were then converted into the SD file format and compounds with obvious errors in the connection tables were removed. This resulted in 35,556 compounds. From these, all purely inorganic compounds not containing any carbon atom, all compounds outside a molecular weight range between 100 and 750, compounds having more than six rotatable bonds and compounds with rings larger

than nine atoms were removed. These criteria should reduce the dataset to reasonably small and moderately flexible compounds resulting in a total of 27,688 compounds.

Finally, in cases with multiple species in the unit cell, all fragments but the largest one was removed (*i.e.*, counter-ions, solvents, *etc.*). In a last filtering step, all duplicate compounds were removed from the dataset.

Finally, a set of 22,768 compounds was obtained. After calculating stereochemical parity flags for stereocenters, this dataset was used for the evaluation study.

12.2.2 Criteria

A set of criteria was defined to assess the quality of the conversion process of each program.

- Conversion rate
- Number of program crashes
- Number of stereochemical errors
- Average computation time per molecule
- Percentage of reproduced X-ray geometries
- Percentage of reproduced ring geometries
- Percentage of reproduced chain geometries
- Percentage of structures without crowded atoms (close contact ratio)

An X-ray geometry is considered as reproduced reasonably well if the RMS_{XYZ} deviation of the atomic positions is less than 0.3 Å. For the ring structures (and ring portions), the percentage of reproduced ring geometries (RMS < 0.3 Å) was restricted to flexible rings and calculated relative to the number of compounds having flexible rings instead of the number of all compounds. This should provide a more realistic insight since it would exclude, *e.g.*, easy cases like phenyl. For acyclic geometries an RMS_{TA} deviation of the torsion angles at rotatable bonds of less than 15° is taken to consider the model compared to the X-ray geometry as well reproduced.

A 3D model is regarded to be free of any non-bonded interactions, if the close contact ratio – the ratio of the smallest non-bonded distance to the smallest acceptable value for this distance – is greater than 0.8.

12.2.3 Program

The program version used for this study was CORINA 3.0.

12.2.4 Results and Discussion

Table B summarizes the results. No program crash or stereochemical errors occurred. The conversion rate was near 100% with an average conversion time of 0.033 seconds per compound. For almost 30% of the structures, the X-ray geometry was reproduced within an RMS deviation of the Cartesian coordinates of all non-hydrogen atoms of 0.3 Å or lower and can be considered as identical conformations. If only the ring systems are considered, almost 80% are identical to the X-ray geometry. For the flexible open-chain portions more than 40% exhibit the same geometry as the X-ray structure measured in torsion angle space.

Table B Performance of CORINA using 22,768 X-ray structures.

	Results
Version of CORINA	3.0
Conversion rate	100
Program crashes	0
Stereochemical errors	0
CPU time [s/mol] [a]	0.033
RMS_{XYZ} < 0.3 Å [%] [b]	28
RMS_{XYZ}^{rings} < 0.3 Å [%] [c]	78
RMS_{TA}^{chains} < 15° [%] [d]	42
Close contact ratio (CCR) > 0.8 [%] [e]	98

12.3 Performance of Speed and Robustness

The number of chemical compounds which are available increases every day. Novel compounds are synthesized, and new natural products are isolated and characterized. On the other hand, computational methods can be used to generate virtual compounds and submit them to *in silico* testing methods before even synthesizing in a laboratory. Obviously, a modern 2D-to-3D structure converter needs to cope with such trends and progresses and another requirement is to be able to process massive volumes of chemical structures in a reliable, robust, and fast manner.

[a] SGI R12000 workstation
[b] Percentage of structures with an RMS deviation of the non-hydrogen atoms of less than 0.3 Å.
[c] Percentage of structures with an RMS deviation of the ring atoms of less than 0.3 Å (flexible rings only).
[d] Percentage of structures with an RMS deviation of the torsion angles in acyclic portions of less than 15°.
[e] Percentage of structures with a close contact ratio of greater than 0.8.

12.3.1 Datasets and Program Version

To test the speed and the robustness of the 2D-to-3D structure conversion process by CORINA to large and freely available compound dataset were used.

- PubChem database (https://pubchem.ncbi.nlm.nih.gov) with 91 million chemical structures [20]
- GDB13 database (http://gdb.unibe.ch) with 971 million organic molecules with up to 13 atoms (atom types C, N, O, S and Cl only) [21]

The two datasets offer some different flavors of chemistry which introduces some interesting aspects in such a robustness test. The PubChem Database mainly consists of "real life" chemical structures which are known and well characterized. The GDB13 is a virtual library of small organic molecules up to 13 atoms (C, N, O, S, Cl) generated following simple chemical stability and synthetic feasibility rules.

For the tests, CORINA Classic version 4.1.0 (August 2017) and version 4.3.0 (September 2019) were used on a x86-64 Linux computer (Intel i7, 3.4 GHz, Ubuntu Linux 14, 64bit). The conversion jobs were run in parallel to use all CPU cores of the processor.

12.3.2 Results and Discussion

Table C summarizes the results of the runs. Both datasets have been converted with conversion rates of ≥99% and without any program crash or intervention by the user. On average, the conversion per structure takes twice as much of computation time for the structures from PubChem than for the virtual compounds from the GDB compounds. While the GDB database contains chemical structures up to only 13 heavy atoms, PubChem also contains much larger and more complex compounds, such as organometallics or natural products.

Table C Conversion of PubChem and GDB13 databases by CORINA Classic.

	PubChem database	GDB13 database
Version of CORINA Classic	**4.3.0**	**4.1.0**
Total number of compounds	96,056,195	971,468,301
Number of converted compounds	95,720,546 (99.7%)	961,454,732 (99%)
Number of **not** converted structures	335,649 (0.3%)	10,013,569 (1%)
Total CPU time	16 hours	112 hours
CPU time per structure	0.6 ms/mol	0.4 ms/mol
Average daily conversion rate	144 million	208 million

12.4 Handling of Stereochemistry

12.4.1 Supported Stereochemistry

CORINA supports stereochemistry originating from atomic chiral centers (asymmetric atoms) and from *E*/*Z* double bond isomerism as well as atropisomerism (special case of axial stereochemistry). The following sections will describe the supported stereochemistry in terms of the handling, interpretation, and performance by CORINA.

12.4.2 Requirements and Coding of Stereochemical Information

CORINA automatically detects chiral centers, interprets them and, if the assignment is correctly defined and geometrically feasible, the respective isomer is generated.

The driver option **-d ist** ("ignore stereo") can be used to fully ignore any stereochemical information provided in the input structure. Of course, CORINA then assumes an arbitrary configuration for each stereocenter, as if no stereo information is provided at all. This option can be useful, if it is known that an input file has a high number of erroneous stereo descriptors.

A similar option is available for input files which contain 3D structures. In this case, CORINA tries to derive the configuration at a stereocenter from the provided 3D coordinates. To avoid this, the driver option **-d i3dst** ("ignore 3D stereo") can be used. Then, CORINA will interpret and process any stereo descriptors, if provided and unless the additional driver option **-d ist** (*vide supra*) is set.

12.4.2.1 Tetrahedral Chiral Centers

The different file formats for chemical structure information support a variety of possibilities for coding the stereochemistry of chemical structures. With the ubiquitous availability of interactive graphical structure editors, the 2D coding of stereochemistry became the most widely used and most convenient method. By using up and down bond symbols (wedge symbols) the local configuration at an atom center is defined as shown for the bridgehead atoms of *cis*- and *trans*-decalin (see Figure 17).

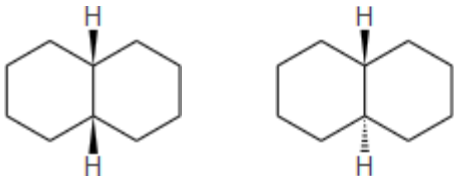


Figure 17 Coding of the stereochemical information of *cis*- and *trans*-decalin by up and down bond (wedge) symbols.

Recipe

Several common errors in specifying up and down bond descriptors often occur. To avoid any issues, the following procedure is recommended.

The atom center in question should be drawn in a quasi-tetrahedral configuration with all four ligand atoms (including hydrogen atoms where appropriate). First, draw three of the four ligand atoms with angles of approximately 120° between the bonds. Then, place the fourth ligand between two of the other ligands and assign the up or down bond symbol (wedge symbol) to this fourth bond.

Figure 18 shows two recommended 2D drawings of a chiral center. Other correct variations are shown in Figure 19.

Figure 18 Recommended input of stereochemistry.

Figure 19 Examples of other correct 2D drawings of a chiral center.

Pitfalls

A variety of ambiguous specifications of stereocenters are possible which are often not easily recognized. Figure 20 shows some examples. The examples all have in common that the ligands of the stereocenter do not span up a suited volume when translating the up and down bond symbols into three dimensions. Thus, the result is ambiguous

and no stereochemical descriptor can be calculated. An error message (see section 7.3 on page 96) is written and the calculation is continued with an arbitrary descriptor.

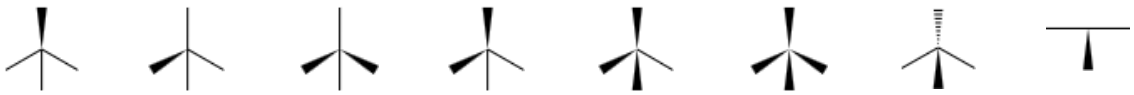


Figure 20 Examples of **in**correct 2D drawings of a chiral center.

Automatic Correction

Often, these errors occur after the automatic addition of hydrogen atoms by the structure editor program. Thus, the specification of all ligand atoms of a stereocenter by hand is strongly recommended. However, in some cases the problem of incorrect coding of stereochemistry can be solved by ignoring one hydrogen atom for the calculation of the stereochemical descriptor (see Figure 21). This is automatically tried and a warning is written (see section 8.3 on page 105) since there is no guaranty that the correction was the intended one.

Figure 21 Correction by ignoring one hydrogen atom.

Another possibility for correcting this type of error is to assume an additional up or down bond descriptor for the bond describing the smallest angles with the other bonds at the center in question (see Figure 22). This is automatically tried, and a warning is written (see section 8.3 on page 105) since there is no guaranty that the correction was the intended one.

Figure 22 Correction by assuming an additional bond descriptor.

The last type of ambiguous coding of stereochemistry is corrected by moving the central atom. Figure 23 gives an example. The three neighbors at the central atom of the fragment on the left-hand side do not span a volume due to the linear position of two of the atoms. This can be corrected by moving the central atom into a direction opposite to the third atom (assuming the hypothetical fourth neighbor in the opposite position). This is automatically tried, and a warning is written (see section 8.3 on page 105) since there is no guaranty that the correction was the intended one.

Figure 23 Correction by moving the central atom.

12.4.2.2 E/Z Double Bonds

The configuration at double bonds is coded by the Cartesian coordinates (either 2D or 3D) in the input file for the most formats. The relative positions of the substituents at a double bond can be calculated from the Cartesian coordinates, but only if the 2D or 3D coordinates provide an unambiguous and clear direction for each substituent. Figure 24 provides some examples of correct drawings for *E*/*Z* double bonds for two-dimensional structure diagrams.

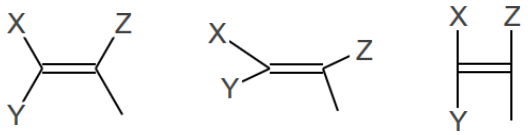


Figure 24 Correct drawings for *E*/*Z* double bonds.

Figure 25 below shows some examples of 2D structure diagrams which do not provide an unambiguous and clear direction for each substituent at the double bond.

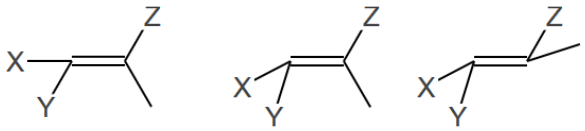


Figure 25 Incorrect drawings for *E*/*Z* double bonds.

The linear SMILES notation offers to code the configuration at a double bond through the characters "/" and "\", which are provided directly at the atoms of the respective double bond. The SMILES "F/C=C/F" and "F\C=C\F" code for the trans isomer of 1,2-di-fluor-ethane (fluorine atoms on "opposite" sides of the double bond), while "F/C=C\F" and "F\C=C/F" code for the cis isomer (fluorine atoms on "same" side).

Double bonds in *E*-configuration (*trans*) in ring systems are only processed for ring systems having more than six (6) ring atoms. For smaller ring systems, *trans* double bonds are automatically reset to the *cis* configuration. A respective warning message is written into the trace file.

The input file option **-i sdfict** ("sdf ignore cis trans") ignores any *cis*/*trans* coding of double bonds provided in the input file and assumes arbitrary configurations, but only for SD input files (V2000/3000). This option is useful, if CORINA fails to process, *e.g.*, a defined *trans* double bond in a highly strained ring system. With the option **-i sdfict**, the *trans* double bond is ignored, and a possibly less strained *cis* bond can be generated. Any other defined stereocenters, such as asymmetric atoms or atropisomeric axes, are not affected by this option.

12.4.2.3 Atropisomerism

Atropisomerism belongs to axial chirality and the reason for chirality does not result from a stereogenic (asymmetric) atom. Atropisomers are caused by a steric effect that hinders the rotation of a single bond (stereogenic axis) between two ring systems. Due to the high rotational barrier of the stereogenic axis, two spatial arrangements (3D structures) can be differentiated which behave like pairs of diastereomers or even pairs of enantiomers. Atropisomerism is sometimes also called "conformational chirality", as it originates from different rotamers.

Figure 26 shows two atropisomers (enantiomers) of the tetra substituted biphenyl compound 6,6'-dinitro-2,2'-diphenic acid.

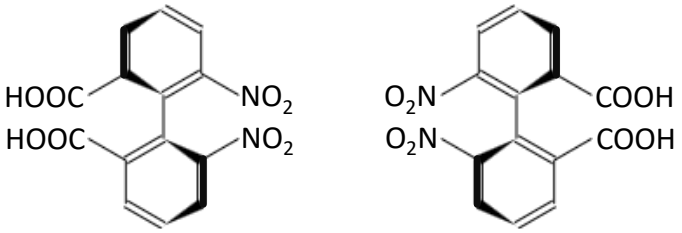


Figure 26 Atropisomers of 6,6'-dinitro-2,2'-diphenic acid.

Since version 4.2.0 of CORINA, atropisomerism is considered and processed appropriately for bicyclic systems which are connected by one single bond. In order to interpret and process information coding a specific atropisomer, the driver option **-d axchir** needs to be set.

Figure 27 lists the mandatory structural requirements which need to be fulfilled by a bicyclic system that CORINA will automatically detect axial chirality.

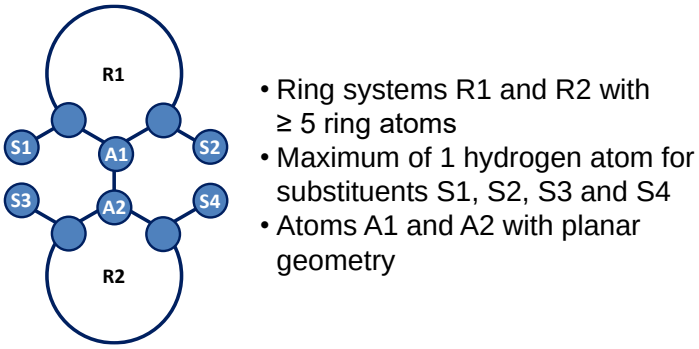
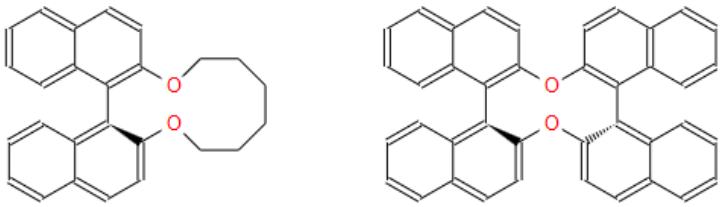


Figure 27 Mandatory structural requirements for atropisomerism.

There are three major structural requirements.

- The two ring systems (R1 and R2) which are connected by a single bond (stereogenic axis) must have at least five ring atoms.
- Three of the four substituents (S1, S2, S3 and S4) which are in *ortho* position to the stereogenic axis must be a non-hydrogen atom.
- The two atoms (A1 and A2) which are connected by the stereogenic axis (single bond) must exhibit a planar geometry.

If two of the four substituents located at the same side of the molecule are part of a third ring system, this third ring system must have at least seven ring atoms. Figure 28 shows two examples which can be processed by CORINA.



2 isomers 3 isomers

Figure 28 Examples of atropisomers with more than two ring systems (and the number of possible isomers)

For coding the stereo information at an atropisomer, at least one wedge symbol has to be defined for a bond in one of the two ring systems which are connected by the stereogenic axis (single bond) and which is connected to this special single bond. Figure 29 illustrates some examples for correct coding and Figure 30 for not correct coding.

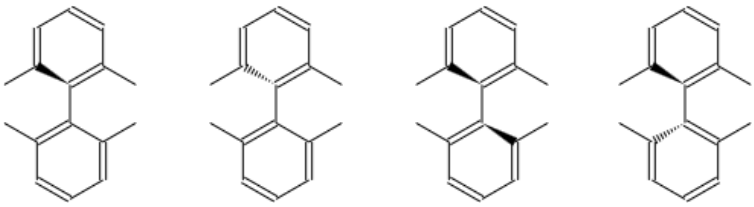


Figure 29 Examples for correct stereo coding of atropisomers.

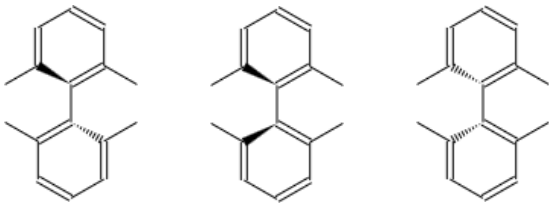


Figure 30 Examples for incorrect stereo coding of atropisomers.

The coding of specific atropisomers with wedge bonds only works with SD files, however, also with input formats which contain 3D coordinates (such as SYBYL MOL2).

12.4.3 Addition of Missing Stereochemical Descriptors

12.4.3.1 Tetrahedral Chiral Centers

The specification of the configuration of chiral centers is essential in generating 3D structures. Thus, the complete definition of stereochemistry of any stereocenter is strongly recommended. CORINA does not seek for the energetically most favorable configuration for undefined centers. However, in most cases a reasonable 3D structure can be generated with arbitrary chosen stereochemical descriptors. Exceptions are ring systems containing chiral atoms. For these systems other than arbitrary values must be found for the stereochemical descriptors of unspecified stereocenters. Geometrical and energy constraints reduce the number of possible stereoisomers and require a careful choice of suited stereochemical descriptors. Some examples shall illustrate this idea.

Bridged Systems

Bridged ring systems as, *e.g.*, norbornane (see Figure 31) require an *o,o*-configuration of the bridgehead atoms. The *i*,*o*-isomer is geometrically forbidden since the bridge cannot be closed as illustrated in Figure 31. CORINA defines stereochemistry at unspecified bridgehead atoms according to this rule. When the input file contains stereochemical descriptors violating this rule, the processing of the molecule is abandoned with an error message (see section 7.3 on page 96).

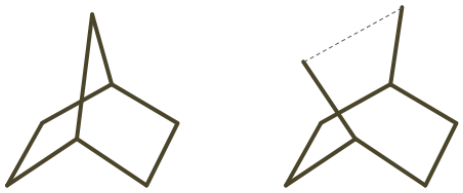


Figure 31 Bridged systems: *o*,*o*- and *i*,*o*-norbornane.

Fused Systems

Fused ring systems as, *e.g.*, decalin (see Figure 32) can occur as different stereoisomers that differ in energy. In the case of decalin, the *trans*-isomer has a lower energy than the *cis* isomer. When the stereochemical information of the input structure is incomplete, an energy criterion is used for setting a default configuration. For the decision whether two fused rings shall prefer *cis*- or *trans*-configuration a set of rules is used depending on the sizes of the two rings.

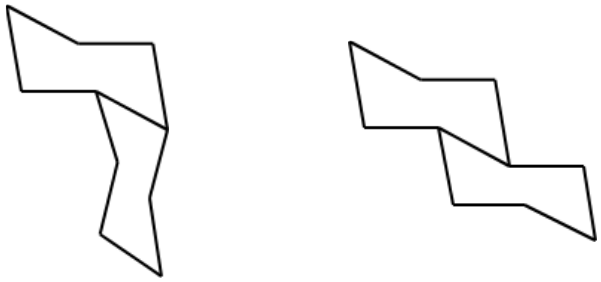


Figure 32 *Cis*- and *trans*-decalin.

Exocyclic Substituents

Exocyclic substituents of ring systems can occupy different spatial positions depending on the stereochemistry and on the conformation. The transition from the equatorial to the axial form of methyl-cyclohexane (see Figure 33) depends on the conformation, whereas the difference between the di-equatorial and the equatorial/axial forms of 1,4-dimethylcyclohexane (see Figure 34) refers to stereochemistry. Thus, only in the second case there is a connection between stereochemistry and steric energy. In case of unspecified stereocenters, CORINA tries to maximize the number of equatorial substituent positions.

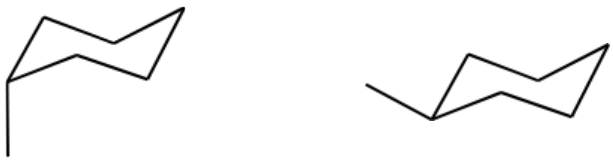


Figure 33 Axial and equatorial geometries of methyl cyclohexane.

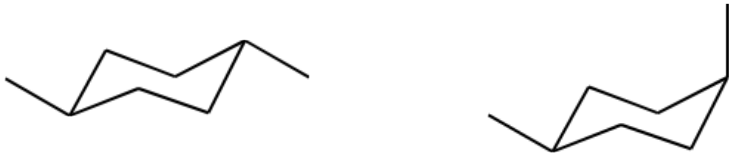


Figure 34 Di-equatorial and equatorial/axial geometries of 1,4-dimethylcyclohexane.

The configuration of the exocyclic substituents 1,4-dimethylcyclohexane can be coded by wedge bonds (see Figure 35).

C1C[N@](C)CC[N@]1C C1C[N@@](C)CC[N@]1C

Figure 35 Coding of stereo chemistry at *para*-substituted rings.

Spiro compounds

Compound with spiro-fused ring systems can also exhibit a stereogenic center(atom) at the spiro atom. Figure 36 illustrates this with the example of a spiro amide compound. The two isomers cannot be superimposed. CORINA must assume an arbitrary configuration if no stereo descriptors are provided. However, CORINA can generate both isomers depending on the configuration coded by, *e.g.*, a wedged bond (see Figure 36).

C1N[C@@]2(C(=O)C1)NCCC2=O C1N[C@]2(C(=O)C1)NCCC2=O

Figure 36 Stereoisomers through a spiro atom.

Note. CORINA cannot interpret 2D sketches of chemical structures which are drawn in a "pseudo 3D" (perspective) representation and which do not provide stereochemical descriptors (such as wedged bonds or parity flags). The IUPAC organization published a paper with recommendations how to draw chemical structures [36]. In this paper, IUPAC recommends avoiding perspective 2D structure diagrams and class-specific drawing styles (such as Fisher or Haworth projections, *etc.*), if the chemical structures should be processed by computers.

Figure 37 shows a cyclohexane derivative in a perspective (or "pseudo 3D") drawing. In this case, the stereochemical configuration at the asymmetric center cannot be derived automatically.

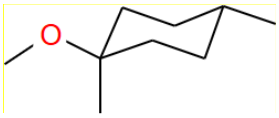


Figure 37 Perspective or "pseudo 3D" drawing of a cyclohexane derivative where stereochemical configuration cannot be derived

Figure 38 shows the same cyclohexane derivative in a perspective drawing, but with defined stereochemical descriptors which force the methoxy group in an equatorial (left) or axial (right) position.

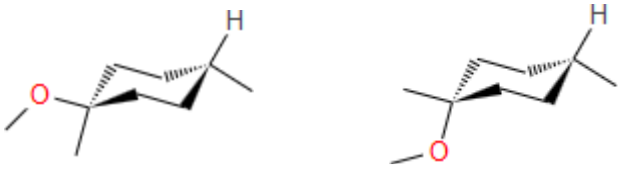


Figure 38 Perspective drawing of a cyclohexane derivative with defined stereochemical descriptors which force the methoxy group in an equatorial (left) or axial (right) position

12.4.3.2 E/Z Double Bonds

As for asymmetric atoms, the correct assignment of the configuration at an *E*/*Z* double bond is essential for the 3D structure generation process. CORINA assumes an arbitrary configuration, if no or an incorrect coding of the configuration at a potential *E*/*Z* double bond is provided.

Double bonds in *E*-configuration (*trans*) in ring systems are only processed for ring systems having more than seven (7) ring atoms. For smaller ring systems, *trans* double bonds are automatically reset to the *cis* configuration.

12.4.3.3 Atropisomerism

In order to generate the desired atropisomer, the appropriate stereochemical information needs to be coded in the input structure either by using wedge bond ssymbols or by, even crude, 3D coordinates. CORINA needs to assume an arbitrary configuration, if the stereochemical information is missing.

12.4.4 Enumeration of Stereoisomers

A substantial step towards the understanding of the physical, chemical or biological properties of a molecule is to study and to analyze its spatial shape. Besides the constitution, a major shape determining feature is the configuration of a molecule, *i.e.*, the stereochemistry.

Furthermore, molecular chirality plays a major role in many areas of chemistry. Enantiomers often exhibit quite different physical, chemical, and biological properties. Therefore, the exploration of the configurational space of a molecule and the analysis of the various isomers a molecule can adopt is of great importance. CORINA provides access to the configurational space of molecules.

As mentioned above, CORINA generates by default a single stereoisomer by taking into consideration the stereochemical information provided in the input connection table and by making reasonable assumptions for missing stereochemical information. The driver option **-d stergen** forces CORINA to automatically identify stereocenters and to generate all possible, but unique and chemically reasonable isomeric compounds which are finally converted into 3D space. Tetrahedral chiral centers as well as *cis*/*trans* isomerism and atropisomerism are considered. In order to include atropisomerism and to enumerate the respective isomers, the additional driver option **axchir** (*i.e.*, **-d stergen,axchir**) has to be set.

Duplicate configurations (*e.g.*, *meso*-compounds) and geometrically strained configurations (*e.g.*, the *i*,*o*-isomer of norbornane, see Figure 31) are rejected. Stereochemical descriptors (parity) are generated and written to the output file. Figure 39 shows some examples.

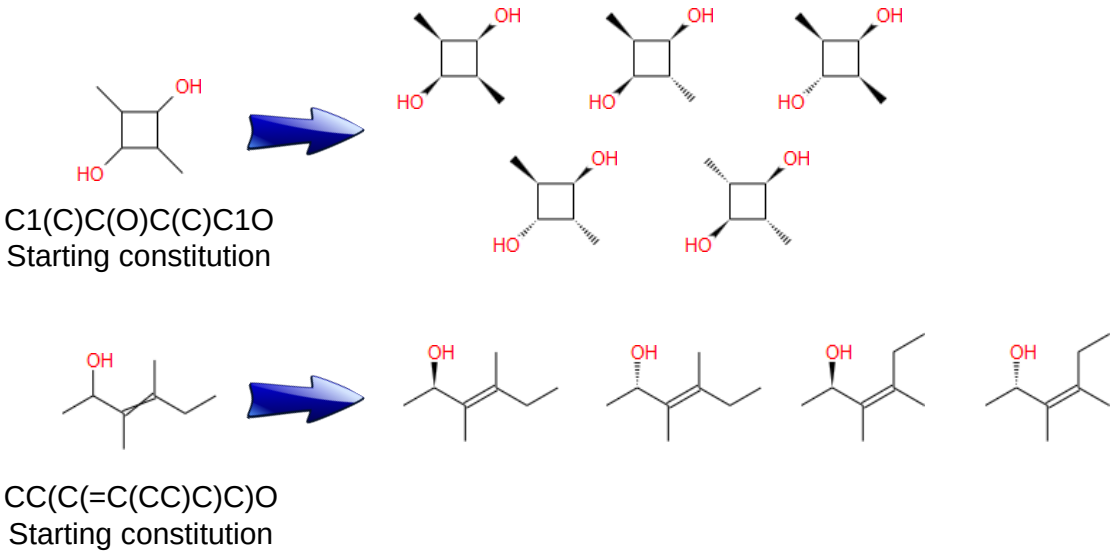


Figure 39 Generated configurations starting from a single initial constitution.

Para-substituted ring systems are identified as *pseudo* stereocenters to generate di-equatorial and equatorial/axial substituted configurational isomers (see Figure 40).

Figure 40 Generation of configurational isomers of *para*-substituted rings.

12.4.4.1 Options to Manipulate the Stereoisomer Enumeration Process

By default (*i.e.*, if the driver option **-d stergen** is set), the stereoisomer enumeration module of CORINA processes a maximum number of four (4) stereocenters and a maximum number of 16 stereoisomers per molecule are generated.

As the number of possible isomers exponentially increases with the number of stereocenters (2^n, where n is the number of stereocenters in a molecule), this limitation was chosen in order not to generate too many isomers and to not provide downstream applications with too many input structures in automated workflows.

Two additional options allow to set the maximum number of output isomers (**msi=<value>**, "maximum number of stereoisomers") or to define a maximum number of stereocenters which should be processed (**msc=<value>**, "maximum number of stereocenters").

If an input structure contains more than the specified number of stereocenters which should be permuted (or four stereocenters, if the default values are used) only **one** isomer is generated. CORINA does not prioritize or rank the stereocenters to select only those for permutation that are the most reasonable. The warning message "Maximum number of # stereocenters exceeded" is printed to the trace file.

The option **preserve** allows for retaining the configuration at atoms that have a defined stereochemistry (*i.e.*, a stereochemical descriptor is given in the input file) or if the input structure already has 3D coordinates. In the latter case, CORINA can derive stereochemical information from the input 3D structure.

The option **preserverel** enables CORINA to interpret input structures with more than one fully defined stereocenter as racemic mixtures. If this option is set, fully defined stereocenters are not interpreted as absolute stereocenters, but as a definition of their relative stereochemistry. For structures with more than a single stereocenter, CORINA will then generate the respective pairs of enantiomers, *i.e.*, the racemic mixtures. This option is helpful if a compound collection or library contains molecules which are

racemic mixtures. Information about axial chirality is processed as well by the option **preserverel** and used to generate the respective isomers (provided that the additional option **axchir** is set).

Note. No diastereomeric compounds are generated with the option **preserverel** provided that all stereocenters are fully defined in the input file.

The option **preserveez** keeps the configuration at defined *E*/*Z* double bonds and only permutes double bonds with an undefined configuration. While the option **preserve** keeps both, chiral centers and *cis*/*trans* (*E*/*Z*) double bonds with defined configurations, fixed, the option **preserveez** only enumerates stereoisomers for double bonds with an undefined configuration. Chiral centers are not affected at all. This option is helpful, if all chiral centers should be permuted, but not defined *E*/*Z* double bonds.

Note. The option **preserveez** does not have an effect if the options **preserve** is set as well (see also Table D on page 141).

The driver option **noflapn** suppresses the flapping (inversion) of pyramidal nitrogen atoms during the generation of stereoisomers (see Figure 41).

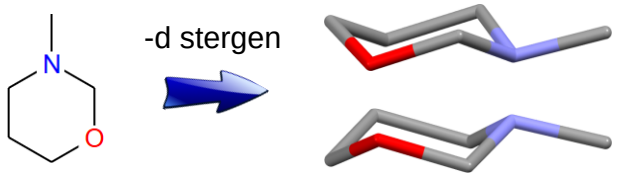


Figure 41 Inversion of pyramidal nitrogen atoms to generate isomers.

Note. CORINA does not identify the lowest-energy configuration (stereoisomer) of a molecule or suggests any ranking of the generated isomers.

Command line example

The following combination of driver options generates a maximum of twenty stereoisomers per molecule (**msi=20**) by processing a maximum of six stereocenters (**msc=6**) and preserves stereocenters that have a defined stereochemistry (*i.e.*, a stereochemical descriptor is given in the input file, **preserve**).

```
-d stergen,msi=20,msc=6,preserve,wh,rs,r2d
```

The stereoisomers are converted to 3D, implicitly given hydrogen atoms in the input structures are added and written to the output file (**wh**), small fragments (*e.g.*, counter ions in salts) are removed from the output file (**rs**) as well as structures which couldn't be converted by CORINA (**r2d**).

12.4.4.2 SD V2000/3000 Specific Performance

Chiral flag

The SD V2000 and V3000 formats provide a chiral flag in the header block which can be set to "on" (value of "1") or set to "off" (value of "0" or missing at all). All defined stereocenters are interpreted as an absolute configuration if the chiral flag is set to "on" (value of "1"). A missing chiral flag or if it is set to "off" (value of "0") is considering all defined stereocenters as a relative configuration.

By default, CORINA does not consider the chiral flag and all defined stereocenters are regarded as is and in an absolute configuration. If the additional driver option **chiralflag** (*i.e.*, **-d stergen,chiralflag**) is set, the chiral flag in the V2000/3000 input file is interpreted according to the V2000/3000 specifications.

The option **chiralflag** also has an impact on axial chirality (atropisomerism) and depending on the status of the chiral flag the respective isomers are generated or not (provided that the additional option **axchir** is set).

The option **chiralflag** is also described in the following section 12.4.4.3 specifically for the V3000 stereochemical extension.

Preserving relative stereochemistry and E/Z double bonds

The option **preserverel** has already been introduced in the previous section and preserves the information about relative stereochemistry provided in an input structure. For V2000/3000 input file, this option has a different impact when used in combination with the option **chiralflag** (*i.e.*, **-d stergen,chiralflag,preserverel**). Depending on the status of the chiral flag of an input structure, different sets of stereoisomers are generated.

If the chiral flag of an input structure is set to "on" (value of "1"), the combination of these two options (**chiralflag,preserverel**) will interpret and process fully defined stereocenters as absolute configurations.

If the chiral flag of an input structure is set to "off" (value of "0") or is missing, the combination of these two options (**chiralflag,preserverel**) will process fully defined stereocenters as relative configurations. In this case, the respective pairs of enantiomers, *i.e.*, the racemic mixtures, will be generated (provided that all stereocenters are defined). This is the same behavior as if the option **preserverel** is not set, as a chiral flag of "off" (value of "0" or missing) implies that only the relative and not the absolute configurations are known.

This combination of options is helpful if a compound collection or library contains molecules which are either racemic mixtures or pure enantiomers defined through the respective stereo descriptors and the chiral flag.

Information about axial chirality (atropisomerism) is processed as well by the option **preserverel** and used to generate the respective isomers (provided that the additional option **axchir** is set).

The option **preserverel** is also described in the following section 12.4.4.3 specifically

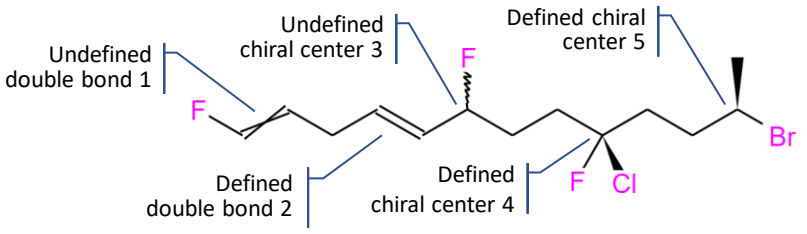
for the V3000 stereochemical extension.

In the SD format, the configuration of double bonds is defined through the 2D or 3D coordinates of the atoms attached to the double bonds. In addition, the configuration of a double bond can be marked as "unknown" in the bond block of the SD record (bond stereo flag with value of "3"). The option **preserveez** enumerates stereoisomers at *E*/*Z* double bonds only for undefined centers, while preserving defined configurations at double bonds.

The following example illustrates the impact of the different command line options when enumerating stereoisomers. The compound in Figure 42 has a total of five stereocenters as following.

- An undefined *E*/*Z* double bond (1)
- A defined *E* double bond (2)
- An undefined chiral center (3)
- Two defined chiral centers (4 and 5, which also defines their relative stereochemistry)

The chiral flag in the input SD file is set to "on" (value of "1").



Chiral flag "on" (1)

Figure 42 Example compound for stereo isomer enumeration options.

Table D lists different combinations of command line options, the number of enumerated isomers and the description which stereocenters contribute to the enumerated stereo isomers.

Table D Enumerated stereo isomers for compound in Figure 42.

Command line options*	Number of isomers	Description
stergen	32	• 2 isomers from undefined double bond 1 • 2 isomers from defined double bond 2 • 2 isomers from undefined chiral center 3 • 2 isomers from defined chiral center 4 • 2 isomers from defined chiral center 5

Command line options*	Number of isomers	Description
stergen,chiralflag	8	• 2 isomers from undefined double bond 1 • 2 isomers from defined double bond 2 • 2 isomers from undefined chiral center 3
stergen,preserve	4	• 2 isomers from undefined double bond 1 • 2 isomers from undefined chiral center 3
stergen,preserverel	16	• 2 isomers from undefined double bonds 1 • 2 isomers from defined double bonds 2 • 2 isomers from undefined chiral center 3 • 2 isomers from defined chiral centers 4 and 5 (relative stereochemistry is defined)
stergen,preserverel, preserve	8	• 2 isomers from undefined double bond 1 • 2 isomers from undefined chiral center 3 • 2 isomers from defined chiral centers 4 and 5 (relative stereochemistry is defined)
stergen,preserverel, preserve,chiralflag	4	• 2 isomers from undefined double bond 1 • 2 isomers from undefined chiral center 3
stergen,preserveez	16	• 2 isomers from undefined double bond 1 • 2 isomers from undefined chiral center 3 • 2 isomers from defined chiral center 4 • 2 isomers from defined chiral center 5
stergen,preserve, preserveez**	4	• 2 isomers from undefined double bond 1 • 2 isomers from undefined chiral center 3

* As the molecule in Figure 42 contains five stereocenters, the additional options **msc** and **msi** have to be set as well (*e.g.*, **msc=5,msi=32**).

** Same results as with driver option **preserve**; **preserveez** has no effect, if option **preserve** is set additionally.

12.4.4.3 V3000 Stereochemical Extension

The SD V3000 format offers some stereochemical extensions for asymmetric atoms (tetrahedral chiral centers). The following Table E provides an overview about the available extensions.

Table E Supported stereochemical groups in SD V3000 format.

Stereochemical group	Value in V3000 format	Description
ABSOLUTE	STEABS	Stereocenter with known absolute stereochemical configuration
OR	STEREL*n*	Stereocenter (minimum of 2 atoms) where only the relative configuration is known, but not the absolute one
AND	STERAC*n*	Mixture of stereoisomers, *e.g.*, a racemate, enantiomers or diastereomers

The V3000 stereochemical extensions are interpreted and processed by CORINA and the stereoisomer generation module. The full consideration and interpretation of the extensions is switched on by the additional option **v3000** (*i.e.*, **-d stergen,v3000**).

Absolute stereochemistry

By default, stereocenters with the stereochemical identifier "STEABS" are fully permuted (*i.e.*, **-d stergen**, all possible stereoisomers are generated). If the additional option **v3000** is set, all defined stereocenters are kept in their corresponding configuration.

As the definition of absolute stereocenters is a strong restriction, the performance described above can also be expected, if the additional option **preserve** is set (*i.e.*, **-d stergen,v3000,preserve**).

The option **v3000** supersedes the option **preserve**. With the combination of the options **v3000** and **preserve** (*i.e.*, **-d stergen,v3000,preserve**), only chiral centers which are not part of a V3000 stereochemical group are permuted.

Relative stereochemistry

By default, stereocenters with the stereochemical identifier "STEREL*n*" or "STERAC*n*" are fully permuted (*i.e.*, with the option **-d stergen**, all possible stereoisomers are generated). If the additional option **v3000** is set, all isomers within the V3000 constraints are generated.

Figure 43 illustrates this process for 3-chlorobutan-2-ol.

The (2R,3S) configuration results in the two enantiomers (2R,3S) and (2S,3R), while the (2R,3R) configuration generates the enantiomeric pair (2R,3R) and (2S,3S).

Note. Due to the definition of the stereochemical identifier "STEREL*n*", no diastereomeric compounds are generated for the example in Figure 43.

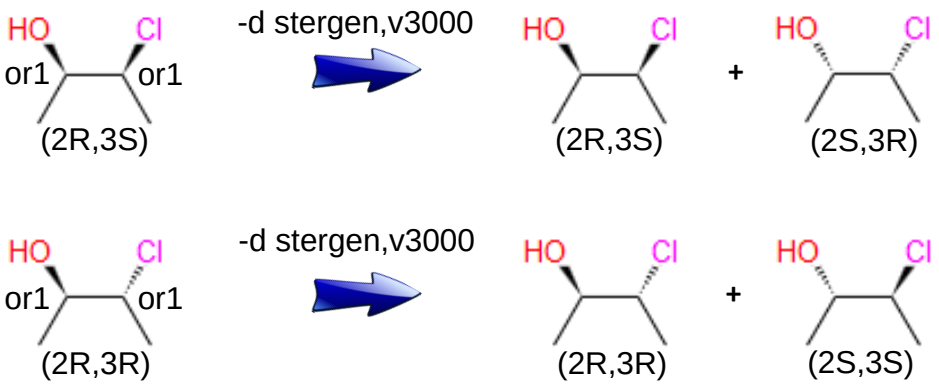


Figure 43 Relative stereocenters with SD V3000 stereochemical constraints.

In the example of Figure 43, the usage of the additional option **preserve** (*i.e.*, **-d stergen,v3000,preserve**) will generate the same two isomers for each starting configuration. The option **v3000** supersedes the option **preserve**.

However, if only the option **preserve** is set (**-d stergen,preserve**), only the defined isomer is generated for each starting configuration in the example of Figure 43.

Chiral flag

The chiral flag in the input structure in V3000 format can have an impact if the option **chiralflag** is set. If the chiral flag in the input structure is set to "off" (value of "0" or missing) and the option **chiralflag** is used (**-d stergen,chiralflag**), any defined stereocenters are not permuted at all. In such cases, the defined stereochemistry as provided in the input structure is preserved. If the additional option v3000 is set (**-d stergen,chiralflag,v3000**), the stereocenters which are part of the stereogenic group "`STEREL`*n*" or "`STERAC`*n*") are permuted within the given V3000 constraints. The option **v3000** supersedes the option **chiralflag** for the stereocenters that are part of a stereogenic group "`STEREL`*n*" (or "`STERAC`*n*").

Preserve relative stereochemistry

The option **preserverel** has already been introduced in the previous section and preserves the information about relative stereochemistry provided in an input structure. For V2000/3000 input file, this option has a different impact when used in combination with the option **chiralflag** (*i.e.*, **-d stergen,chiralflag,preserverel**). Depending on the status of the chiral flag of an input structure, different sets of stereoisomers are generated.

If the chiral flag of an input structure is set to "on" (value of "1"), the combination of these two options will interpret and process fully defined stereocenters as absolute configurations. The option **chiralflag** supersedes the option **preserverel**.

If the chiral flag of an input structure is set to "off" (value of "0") or is missing, the combination of these two options (**chiralflag,preserverel**) will process fully defined stereocenters as relative configurations. In this case, the respective pairs of enantiomers, *i.e.*, the racemic mixtures, will be generated (provided that all stereocenters are defined). This option is helpful if a compound collection or library contains molecules which are either racemic mixtures or pure enantiomers defined through the respective stereo descriptors and the chiral flag.

Information about axial chirality (atropisomerism) is processed as well by the option **preserverel** and used to generate the respective isomers (provided that the additional option **axchir** is set).

Examples

The following example illustrates the performance of CORINA with V3000 stereochemical representations and different driver options when generating stereoisomers. The molecule in Figure 44 has a total number of seven (7) stereocenters. One center is defined as an absolute stereocenter ("STEABS", "abs" in Figure 44). For two stereogenic groups each consisting of two tetrahedral centers, the relative stereochemistry of the two tetrahedral centers is known and, therefore, they are defined as "OR" ("STERELn", "or1" and "or2" in Figure 44). For two further stereocenters, the absolute configuration is not known and they are defined as racemic mixtures ("STERACn", "&1" and "&2" in Figure 44).

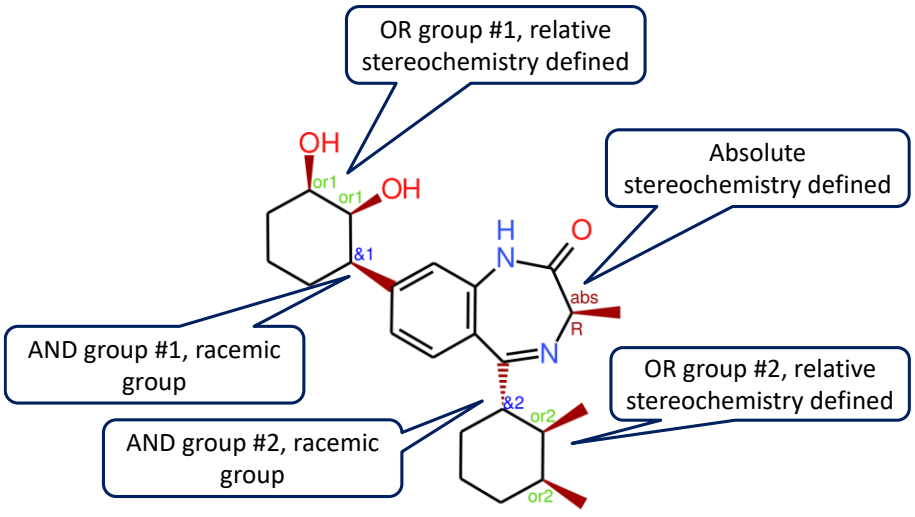


Figure 44 Molecule with SD V3000 stereochemical constraints.

Table F lists the number of stereoisomers that are generated with different driver options for the stereoisomer generation module.

Without any further restrictions (driver option **-d stergen**), all 7 stereocenters are enumerated which results in a total number of 128 (= 2^7) stereoisomers. With the additional driver option **preserve**, only one stereoisomer is generated as all stereocenters are fully defined by wedged bond symbols. With the additional option **preserverel** all relative stereochemistry definition is preserved and only the absolute stereocenter ("abs") is permuted. This results in two (2^1) stereoisomers.

Applying the V3000 stereochemical representations (driver option **v3000**), 16 stereoisomers are generated as follows. The absolute center results in a single isomer. The two stereogenic groups with a defined relative stereochemistry ("STEREL*n*", "or1" and "or2" in Figure 44) result in 2 isomers for each group, *i.e.*, four isomers in total and the two racemic centers ("STERAC*n*", "&1" and "&2" in Figure 44) provide another four isomers. In total, 1 x 4 x 4 = 16 stereoisomers are generated.

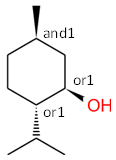
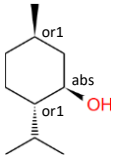
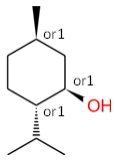
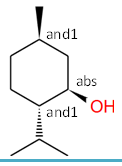
Note. The example in Figure 44 above exceeds the default maximum number of processed stereocenters of four (4). Therefore, the additional option **msc=<value>** and **msi=<value>** need to be used. Possible values for the options are **msc=7,msi=128**.

Table F Generated stereoisomers for molecule in Figure 44.

Command line option	Description	Number of generated stereoisomers
stergen	No restrictions, enumerate all possible, but unique stereoisomers	128 (= 2^7)
stergen,preserve	Preserve all defined stereocenters	1
stergen,preserverel	Preserve all relative stereochemistry	2 (=2^1)
stergen,v3000	Enumerate all stereoisomers according to V3000 stereochemical definitions	16 (= 2^4)

The following example uses different V3000 stereochemical definitions for the compound (-)-menthol and Table G lists the number of generated isomers for different combinations of command line options for the stereoisomer enumeration process.

Table G Number of generated stereoisomers for (-)-menthol with V3000 definitions.

Structure	Command line options (option stergen set)				
	-	preserverel	v3000	v3000, preserverel	v3000, preserve
	8	2	4	4	4
	8	2	8	2	2
	8	2	2	2	2
	8	2	2	2	2

Note. The SD V3000 stereochemical extensions are defined only for tetrahedral stereocenters, but not for stereocenters based on *E*/*Z* double bonds.

Specificity and priority of command line options

The stereoisomer generation module of CORINA offers a variety of command line options to manipulate the process of enumerating stereoisomers. The most specific option is **v3000**, while **preserve** is the less specific option.

In the following, some examples of combinations of command line options of the stereoisomer generation module and their priorities are provided (for chiral centers).

- Options **v3000** and **preserve**: the option **preserve** has no effect on chiral centers with a v3000 stereo identifier
- Options **v3000** and **chiralflag**: the option **chiralflag** has no effect on chiral centers with a v3000 stereo identifier
- Options **v3000** and **preserverel**: the option **preserverel** has no effect on chiral centers with a v3000 stereo identifier
- Options **preserve** and **chiralflag**: the option **chiralflag** has no effect and all defined chiral centers are not changed/permuted (are preserved), even if the value of the chiral flag is "0" or missing in the input structure

- Options **chiralflag** and **preserverel**: the option **preserverel** has no effect and all defined chiral centers are not changed/permuted (are preserved) if the value of the chiral flag is "1" in the input structure (absolute stereocenters)
- Options **preserverel** and **preserve**: the option **preserve** has no effect and all defined chiral centers are changed/permuted, but the relative stereochemistry between the centers is preserved.

Recommendation

As a general recommendation, the following combination of command line options is useful for generating stereoisomers for a collection of chemicals which are stored in SD file format and which contains a mix of V2000 and V3000 records.

```
-d stergen,v3000,chiralflag,preserverel,preserveez
```

The chemical structures can have either defined relative or absolute stereochemistry including V3000 stereochemical groups and chiral flags or stereocenters are undefined due to the lack of knowledge of the correct isomer or isomeric mixture. With this combination, the relevant stereoisomers will be generated. The option **preserveez** will ensure that only undefined *E*/*Z* double bonds will be permuted. Defined *E*/*Z* double bonds will be preserved.

12.5 Conformational Analysis of Ring Systems for Flexible Search Purposes

The method of choice for flexible 3D database searches is to use compound databases that contain single low-energy conformations for each molecule and to solve the flexibility problem on the fly instead of storing conformational ensembles in the database, a rather disk space consuming approach.

A variety of methods such as the directed tweak algorithm exist for flexible searching [38]. These methods are efficient for chain portions of the molecules but run into problems when applied to ring systems (*vide infra*).

One approach is to store 3D models with multiple ring conformations and to apply the flexible search only to the chain portions. CORINA supports this approach by providing methods for generating multiple ring conformations.

12.5.1 Generation of Multiple Ring Conformations

For ring systems consisting of up to nine ring atoms, CORINA performs internally a systematic conformational analysis to find the ring conformation with the lowest energy. Thus, a partial conformational analysis can be easily performed by outputting all conformations that are found in this way. For each side chain only one conformation is generated that remains unchanged unless there are no problems with non-bonded interactions caused by the different ring conformations.

Therefore, the output consists of a set of 3D models having different ring systems and almost the same chain conformations. The philosophy behind this approach is that a conformational analysis for ring systems requires more program intelligence and that an analysis of the side chains can be performed by a much simpler postprocessor (*e.g.*, by systematically permuting all rotatable bonds) using 3D structures with different ring conformations as input. Furthermore, the number of reasonable ring geometries often is orders of magnitude smaller than the number of chain conformations.

The driver option **-d rc** forces CORINA to generate multiple ring conformations. The command **corina -n n=1 -d rc example.sdf out.sdf** produces a conformational ensemble for *trans*-decalin, the first record of the example file *example.sdf*. Figure 45 shows the five conformations obtained. The conformations are output in the order of increasing steric energy. The maximum number of conformations per molecule can be restricted by the driver option **-d mc=<value>** where <value> is the required number of conformations.

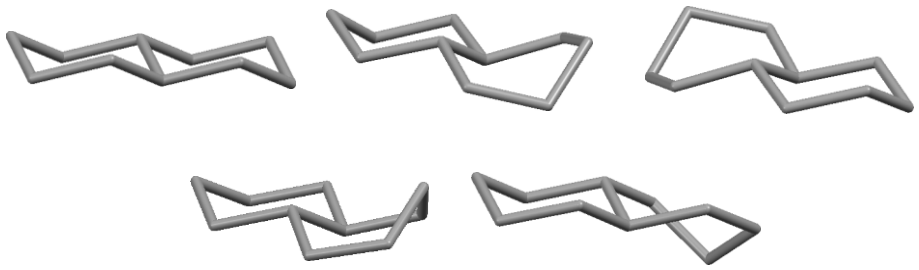


Figure 45 Conformations of *trans*-decalin.

12.5.2 Handling of Pyramidal Ring Nitrogen Atoms

For pyramidal ring nitrogen atoms with one exocyclic substituent, CORINA can generate conformations with both possible configurations at the nitrogen atoms (driver option **-d rc,flapn**). For 1,4-dimethyl-piperidine (SMILESC "N1CCCCC1" without specified stereochemistry), four chair conformations with all combinations of the two substituents in equatorial and axial positions (see Figure 46) are generated.

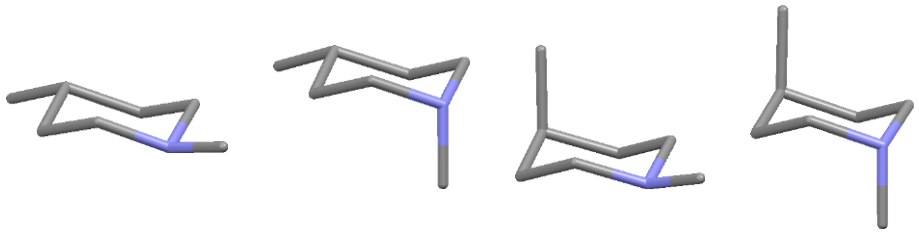


Figure 46 Chair conformations of 1,4-dimethyl-piperidin.

The driver option **-d noflapn** suppresses the flapping of any pyramidal nitrogen atoms only when stereoisomers should be generated (*i.e.*, in combination with the driver option **-d stergen**).

The driver option **-d planil** forces anilinic nitrogen atoms which are part of a ring system into a planar geometry.

12.5.3 Handling of Molecules Having More Than One Ring System

For molecules having more than one ring system connected by chains, CORINA offers the following two techniques.

1) The exhaustive method (default). All combinations of all conformations of the individual ring systems are generated. A possible combinatorial explosion is inherent to this method.
2) The compact method (option **-d sc**). All ring systems simultaneously change their conformations from the low- to the high-energy levels. All conformations of a specific ring system will be seen, but not all combinations of them. A possible loss of bioactive conformations is the price for a significantly smaller number of conformations generated.

These methods are illustrated by using 1-cyclohexyl-2-cyclohex-3-enyl-ethane (SMILES: C1CCCCC1CCC1CC=CCC1) as an example (see Figure 47).

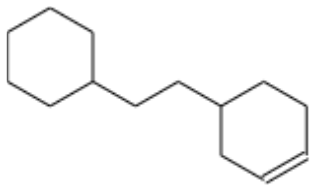


Figure 47 1-Cyclohexyl-2-cyclohex-3-enyl-ethane.

For the cyclohexane and cyclohexene rings each having one exocyclic substituent, CORINA generates 14 and 4 conformations, respectively. Thus, the exhaustive method (1) results in 14 x 4 = 56 conformations. The compact method (2) yields MAX(14, 4) = 14 conformations.

12.5.4 Multiple Ring Conformations in 3D Database Searches

An exhaustive study on the impact of using multiple ring conformations for 3D database searches was performed [39]. A short summary of the results is given in this section.

12.5.4.1 Directed Tweak and Ring Flexibility

One approach to 3D database searching is to address torsion flexibility by a fitting technique called directed tweak [38]. Directed tweak optimizes the torsion angles at

rotatable bonds with respect to a penalty function describing the distance of the actual conformation of a molecule to a given pharmacophore. In principle, this technique is also applicable to ring systems but leads to several problems: First, ring systems are orders of magnitude less flexible than chain fragments and have a limited number of significantly different conformations. Thus, a continuous fit technique will not be as good as for chain fragments. Secondly, the handling of ring closure by opening one bond per ring and introducing additional constraints into the penalty function makes the hyper-surface of the optimization function more complicated and is a potential source for numerical problems.

Convergence problems and questionable hit geometries are the result. In general, the geometries of the hit structures are often rather distorted. Frequently found problems are those of violated sp^2 atoms or stereocenters and atom clashes at the ring closure bonds even when searching with explicit van der Waals check.

12.5.4.2 Hybrid Approach

The following ideas led to a hybrid approach that overcomes the problems sketched above of the directed tweak method.

First, directed tweak performs efficiently for chain portions of molecules but runs into problems when applied to rings.

Secondly, the explicit storage of multiple conformations for addressing flexibility suffers both from the immense requirements of computer resources and from possible losses of bioactive conformations due to a too coarse search grid.

On the other hand, ring systems consisting of small rings (*e.g.*, with up to eight atoms) show only a small number of conformations that represent a coarse grid.

Thus, a hybrid approach combining the two techniques can increase the efficiency of search. The proposed technique splits the handling of flexibility into two routes, one for handling rings and another one for chain portions of the molecules:

1) For chains, the directed tweak method is used.

2) For rings, multiple conformations are stored in a 3D database.

This method was implemented by combining two programs: UNITY [40] for the flexible search (directed tweak) and for the database management and CORINA for the 2D-to-3D conversion and the generation of multiple ring conformations.

12.5.4.3 Evaluation Method

To compare the performance of the hybrid technique with alternative approaches, five reference queries were searched in a public domain database. The public part of the NCI database [19] containing 126,705 compounds was used as an evaluation database.

Two 3D databases were constructed by using CORINA. The first database (NCI) contained only one conformation per molecule and the second database (NCI_FLEX)

contained up to 25 ring conformations per molecule.

CORINA was forced to add missing hydrogen atoms, to remove small fragments, to generate a maximum of 25 conformers per molecule, to invert pyramidal ring nitrogen atoms and to use the compact method for molecules having more than one ring system (command line options **-d wh,rs,rc,mc=25,flapn,sc**).

Test queries for dopamine agonists (derived from the dopamine structure), kinase C agonists [41] histamine agonists 1 and 2 agonists [42] and for antiarrythmic agents [43] where taken from the literature.

These queries were searched in the following three different ways.

1) Search the NCI database only with chain flexibility (reference run).

2) Search the NCI database with both flexible rings and chains (original approach).

3) Search the NCI_FLEX database with flexible chains (hybrid-approach).

To exclude artifacts and unreasonable geometries as mentioned above, initially found hits were relaxed by a robust force field [44] and then searched again without ring flexibility in order to figure out the genuine hits with a relaxed ring conformation.

12.5.4.4 Results and Discussion

Table H shows the results of the three search runs.

Clearly, the hybrid approach (NCI_FLEX, tweak chains) results in a significant additional portion of hits (6-23%) compared to the reference run without ring flexibility. The application of the directed tweak method to rings (original approach) yielded a significantly smaller additional amount of hits.

Moreover, in one case (histamine 2) the tweaking of rings decreased the hit number by 2%. This indicates that a number of "stable" hit structures found without ring flexibility in the reference run may be hidden by the flexible ring search due to an "instable" hit structure falling into a local minimum during relaxation which cannot fulfill the query in the subsequent search run without ring flexibility.

Table H Hit numbers of the three search runs.

Query	NCI tweak chains (reference run)	NCI tweak chains tweak rings (original method)	NCI_FLEX tweak chains (hybrid approach)
Dopamine	117	122 (+4%)	144 (+23%)
Kinase C	490	553 (+13%)	601 (+23%)
Histamine 1	3,736	3,999 (+7%)	4,247 (+14%)
Histamine 2	1,932	1,885 (-2%)	2,050 (+6%)
Antiarrythmic	1,180	1,197 (+1%)	1,441 (+22%)

In addition, the dependence of the hit rate on the maximum number of conformations per molecule was investigated (see Figure 48). In most cases, the hit rates converged to a value of 10 conformations.

At this point, the size of the NCI_FLEX database was increased by a factor of 1.7 compared to the database with one conformer per record (NCI), a rather moderate requirement of additional resources.

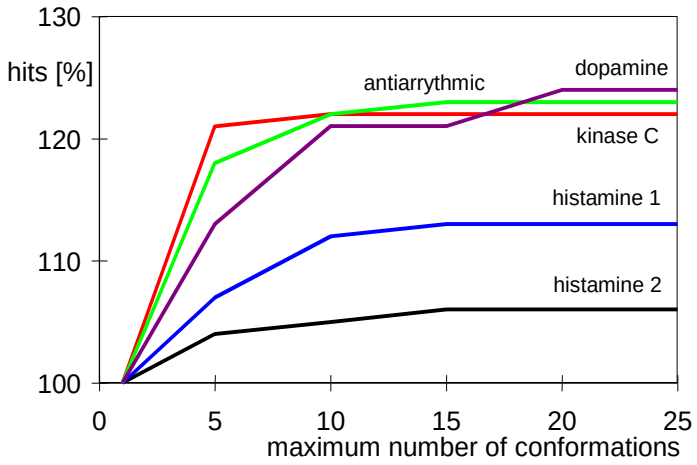


Figure 48 Hit rate vs. maximum number of conformations per molecule.

12.5.4.5 Recommendations

The following CORINA command line parameters are recommended for the generation of 3D databases for pharmacophore search purposes due to the results and performance of the study presented above.

```
corina -d wh,rs,r2d,rc,mc=10,flapn,sc <in> <out>
```

13 Acknowledgements

CORINA was developed in the research group of Prof. Johann Gasteiger at the Technical University of Munich and at the University of Erlangen-Nuremberg since 1985.

The program development was initiated by Dr. Christian Hiller (initial implementation in Fortran programming language) and continued by Dr. Christine Rudolph (treatment of large rings, macrocyclic and polymacrocyclic structures) and Dr. Jens Sadowski (reduced force field, large rings, flexible macrocyclic structures, metal complexes, conformational analysis and refinement of close contacts, reimplementation of CORINA in C programming language).

The authors gratefully acknowledge the permission obtained from the Cambridge Crystallographic Data Centre (Cambridge, UK) to implement in CORINA the more detailed and advanced version of the Torsion Angle Library (915 torsion angle patterns), which was derived from the Cambridge Structural Database (CSD).

Furthermore, the authors thank Prof. Dr. Gerhard Klebe (University of Marburg, Germany), Dipl.-Math. Thomas Mietzner and Dr. Frank Weber (BASF-AG, Ludwigshafen, Germany) for the close cooperation during the implementation of the CSD-derived Torsion Angle Library in CORINA.

CORINA Classic is maintained for general usage by MN-AM, Molecular Networks GmbH, Nuremberg, Germany and Altamira LLC, Columbus, Ohio, USA).

Since version 4.0, CORINA Classic uses the InChI Software version 1.05 (January 2017, copyright (C) IUPAC and InChI Trust Limited) under the IUPAC/InChI-Trust License No. 1.0) to support standard InChI (IUPAC International Chemical Identifier) as input file format.

14 References

[1] Cambridge Structural Database: (a) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, D. S.; Rodgers, J. R.; Watson, D. G. *Acta Cryst. Sect. B: Struct. Crystallogr. Cryst. Chem.* 1979, B35, 2331. (b) Allen, F. H.; Kennard, O. *Acc. Chem. Res.*, 1983, 16, 146. c) Allen, F.H.; Hoy, V.J. Cambridge Structural Database. In *Encyclopedia of Computational Chemistry*, Schleyer, P.v.R.; Allinger, N.L.; Clark, T.; Gasteiger, J.; Kollman, P.A.; Schaefer, III, H.F.; Schreiner, P.R. (Eds.), John Wiley & Sons, Inc., Chichester, UK, 1998; pp. 155-167.

[2] (a) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chemical Reviews* **1993**, *93*, 2567-2581. (b) Sadowski, J., Three-Dimensional Structure Generation: Automation. In *Encyclopedia of Computational Chemistry*, Schleyer, P.v.R.; Allinger, N.L.; Clark, T.; Gasteiger, J.; Kollman, P.A.; Schaefer, III, H.F.; Schreiner, P.R. (Eds.), John Wiley & Sons, Inc., Chichester, UK, 1998; pp.2976-2988. (c) Sadowski, J.; Schwab, C.H.; 3D Structure Generation and Conformational Searching. In *Computational Medicinal Chemistry and Drug Discovery*, Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere J.P., Eds., Dekker Inc., New York, 2004; pp. 151-212. (d) Sadowski, J. 3D Structure Generation. In *Handbook of Chemoinformatics – From Data to Knowledge.* J. Gasteiger, J.; Engel, T., Eds., Wiley-VCH: Weinheim, 2003, pp. 231-261.

[3] Schwab, C.H. Conformations and 3D pharmacophore searching. Drug Discovery Today: Technologies, Volume 7, Issue 4, Winter 2010, e245-e253.

[4] Hiller, C.; Gasteiger, J. Ein automatisierter Molekülbaukasten. In *Software-Entwicklung in der Chemie*, Vol 1; Gasteiger, J., Ed.; Springer: Berlin, 1987; pp 53-66.

[5] Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comp. Method.* **1990**, *3*, 537-547.

[6] Sadowski, J.; Rudolph, C.; Gasteiger, J. The Generation of 3D Models of Host-guest Complexes. *Anal. Chim. Acta* **1992**, *265*, 233-241.

[7] Sadowski, J.; Gasteiger, J. Polygon Patterns for the Generation of Conformations of Large Rings. In *Software Development in Chemistry*, Vol 7; Ziessow, D., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1993; pp 65-76.

[8] Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.

[9] Schönberger, H.; Schwab, C.H.; Hirsch, A; J. Gasteiger, J. Molecular Modelling of Fullerene Dendrimers. *J. Mol. Model*. **2000**, *6*, 379-395.

[10] (a) Dalby, A.; Nourse, J. G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244-255.
(b) A detailed description of the file formats Mol, SD and RD is available online for download as a PDF document at "http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php".
(c) A description of the enhanced stereo-chemical representation supported by the SD V3000 format is available online at "http://accelrys.com/products/pdf/enhanced-stereochemical-representation.pdf".

[11] (a) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36. (b) Daylight Software Manual. Daylight Chemical Information Systems: Santa Fe, NM, USA, 1993, "www.daylight.com".

[12] (a) SYBYL is developed and distributed by Tripos Associates Inc.: St. Louis, MO, USA. (b) A detailed description of SYBYL MOL/MOL2 file formats is available on the Internet for download as a PDF document.

[13] (a) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535-542. (b) Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235-242.

[14] (a) MacroModel is developed and distributed by Schrödinger, LLC, Portland, OR, USA ("www.schroedinger.com"). (b) MacroModel User Manual Version 7.0, "www.schrodinger.com/Support/pdf.html".

[15] (a) Maestro is developed and distributed by Schrödinger, LLC, Portland, OR, USA, www.schroedinger.com. (b) Maestro User Manual Version 5.0, "www.schrodinger.com/Support/pdf.html".

[16] (a) Hall, S.R.; Allen, F.H.; Brown, I.D. The Crystallographic Information File (CIF): a New Standard Archive File. *Acta Cryst.* **1991**, *A47*, 655-685. (b) See also "www.iucr.org".

[17] The IUPAC International Chemical Identifier (InChI), www.inchi-trust.org.

[18] MOPAC 2000, Stewart, J.J.P., Fujitsu Limited, Tokyo, Japan (1999). b) The MOPAC manual is available online at "http://openmopac.net/manual/index.html".

[19] (a) Milne, G.W.A.; Nicklaus, M.C.; Driscoll, J.S.; Wang, S.; Zaharevitz, D.W. The NCI Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219-1224. (b) See also: "http://cactus.nci.nih.gov/ncidb2/download.html".

[20] Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 (Jan 4), 44(D1): D1202-1213. Epub 2015 Sep 22 [PubMed PMID: 26400175]. DOI: 10.1093/nar/gkv951.

[21] Blum L.C.; Reymond J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732-8733.

[22] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470-489.

[23] Klebe, G.; Mietzner, T. A Fast and Efficient Method to Generate Biologically Relevant Conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583-606.

[24] Gasteiger, J.; Sadowski, J. Evaluation of 3D Structure Generators Using 639 X-Ray Structures. Personal communication, 1996.

[25] Böhm, H.-J. The Computer Program Ludi: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61-78.

[26] Collaborative Computational Project, Number 4, 1994: The CCP4 Program Suite: Programs for Protein Crystallography, *Acta Cryst.* **1994***, D50*, 760-763.

[27] X-PLOR, A System for X-ray Crystallography and NMR: Brünger, A.T., The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA.

[28] Gasteiger, J. *et al.* CTX Keyword Reference Manual. University of Erlangen-Nuremberg: 1995, unpublished results.

[29] Klebe, G.; Mietzner, T.; Weber, F. Methodological Developments and Strategies for a Fast Flexible Superposition of Drug-Size Molecules. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 35-49.

[30] STERGEN Version 1.1 is developed and distributed by Molecular Networks GmbH, Nuremberg, Germany, "www.mn-am.com".

[31] Jones, T.A.; Zou, J.Y.; Cowan, S.W.; Kjeldgaard, M. Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in These Models. *Acta Cryst.* **1991**, *A47*, 110-119. b) See also: http://xray.bmc.uu.se/usf.

[32] Güntert, P.; Mumenthaler, C.; Wüthrich, K. Torsion Angle Dynamics for NMR Structure Calculation with the New Program DYANA. *J. Mol. Biol.* **1997**, *273*, 283-298.

[33] (a) DYANA (CYANA) is developed and distributed by P. Güntert Scientific Software, Zürich, Switzerland. (b) DYANA Version 1.5 User's Manual, "www.guentert.com".

[34] (a) Jones, G.; Willett, P.; Glen, R.C. Molecular Recognition of Receptor Sites Using a Genetic Algorithm with a Description of Desolvation. *J. Mol. Biol.* **1995**, *254*, 43-53. (b) Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727-748.

[35] (a) GOLD is distributed by Cambridge Crystallographic Data Centre, Cambridge, UK, "www.ccdc.cam.uk". (b) GOLD 2.0 Documentation (User Manual) "www.ccdc.cam.uk/support/csd_doc/zdoc_main.html".

[36] Brecher, J. Graphical representation of stereochemical configuration (IUPAC Recommendations 2006). *Pure Appl. Chem.* **2006**, *78(10)*, 1897-1970 (DOI:10.1351/pac200678101897).

[37] Pearlman, R.S., Concord, distributed by Certara USA Inc., Princeton, NJ, USA, "www.certara.com".

[38] Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190-196.

[39] Sadowski, J. A Hybrid-Approach for Addressing Ring Flexibility in 3D Database Searching. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 53-60.

[40] UNITY is distributed by Certara USA Inc., Princeton, NJ, USA, "www.certara.com".

[41] Wang, S.; Zaharevitz, D.W.; Sharma, R.; Marquez, V.E.; Lewin, N.E.; Du, L.; Blumberg, P.M.; Milne, G.W.A. The Discovery of Novel, Structurally Diverse Protein Kinase C Agonists through Computer 3D Database Pharmacophore Search. Molecular Modelling Studies. *J. Med. Chem.* **1994**, *37*, 4479-4489.

[42] Good, A.C.; Kuntz, I.D. Investigating the Extension of Pairwise Distance Pharmacophore Measures to Triplet-Based Descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373-379.

[43] Moock, T.E., Henry, D.R., Ozkabak, A.G. and Alamgir, M., Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184-189.

[44] Davies, E.K.; Murrall, N.W. How Accurate Does a Force Field Need To Be? *Computers & Chemistry* **1989**, *13*, 149-156.

15 Report Form

In the case of problems occurring during installation or running CORINA Classic, please complete the following form and send it or fax it to

Molecular Networks GmbH
Neumeyerstraße 28
90411 Nürnberg, Germany
FAX: +49 911 597 424 09

_____

User:

_____

CORINA Classic program and version number ("corina -v"):
Command line to run CORINA Classic:
Error and warning messages by CORINA Classic:

_____

System messages:

_____

Short description:

Please send the input file, output file and trace file (*corina.trc*) generated by CORINA Classic by email to support@mn-am.com These files will help us to analyze your issues. All data will be treated confidentially.