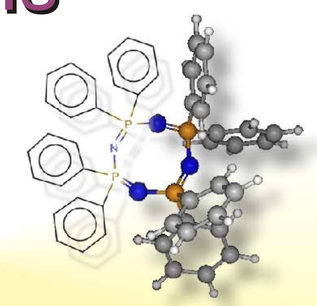
# 3D structure prediction and conformational analysis

G. Imre[1,2] and Ö. Farkas[1]

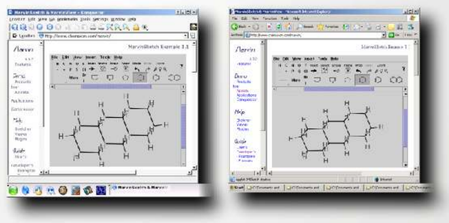
(1) Department of Organic Chemistry, Eötvös Loránd University, 1/A Pázmány Péter sét., Budapest H-1117, Hungary
(2) Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Goldman György sq. 3., Budapest H-1111, Hungary
E-mail: imreg@organ.chem.elte.hu, farkas@organ.chem.elte.hu



## Clean3D

Clean3D is a project at the Department of Organic Chemistry, Eötvös Loránd University, Budapest, Hungary which aims the automated 3D coordinate/conformer generation from coordinate-less structure information for small to medium sized structures.

Clean3D is integrated into the software package Marvin[1] from ChemAxon as a part of GUI and as a stand-alone component. It is also callable from custom applicaton through API.



Two cause initiated the development at the begining:

At first, Marvin is implemented in JAVA: the same compiled code runs on many platforms (without recompilation or reconfiguration) and this is required from all modules.

On the other hand, at the beginning of this project, we had a great idea, which was worth to develop...

[1]`http://www.chemaxon.com/marvin`

## Method

The method determines atom coordinates in a step-by-step manner by fragments. It is aimed to determine multiple energetically favorable structures (conformers) for fragments.

Build steps usually results multiple conformers. If the resulting conformer count is greater than a predefined limit, some of them will be ignored in the further process. With this conformer count limit the coordinate generation process have a scalability parameter which affects the quality of resulting structures.

**Build steps can characterized as below:**

**Single atom fusing**

The base fragment will be extended with a connected atom. Starting from multiple fragment conformers, multiple possible atom orientations will be determined for each of them. Several limiting heuristics are developed to balance conformational diversity and conformer count.

**Multiple atom fusing**

Two connected multiple-atom fragments will be joined together. The connecting bonds possible orientations (for each fragments) can be determined. These overlapping directions can help to align the fragments.

**Fragment database query**

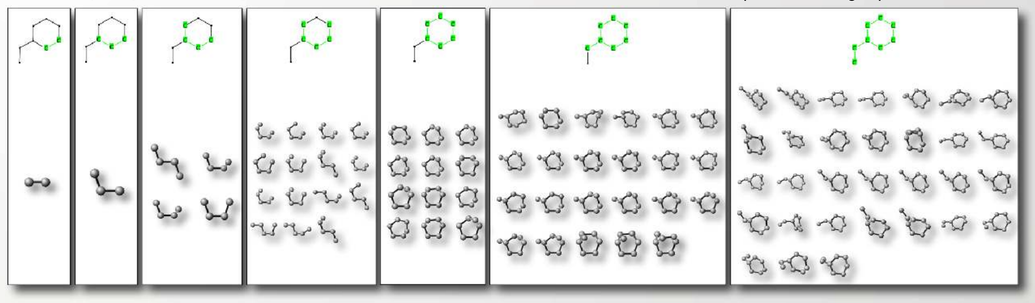
The frequently occurrent fragments with multiple conformations can be stored in a database. During the building process fast database access can be implemented by using structural fingerprints.

**Direct coordinate generation through multidimensional Minkowski-space**

A direct method was developed which allows the efficient 3D coordinate generation for rigid, compact structures with tensions.

**Development status:**

In the current version of Marvin a single-atom fuse based cleaning process is deployed. With the added helper methods the overall efficiency is compelling for small and medium sized structures
.
The Minkowski-based coordinate generation process is implemented and verified, but its deployment must follow the realization of multiple atom fusing capabilities.



## Abstract

Numerous theoretical method in the field of computational chemistry falls back on the availability of 3D structural information about compounds. Determining molecular structure without human interaction is an essential component of several techniques, like QSAR, 3D pharmacophore analysis, reaction prediction, etc. Current computational tools used for structure determination including force-fields and quantum chemical methods, even require a complete set of initial 3D coordinates. The efficiency of 3D structure based HTS tools also can be enhanced by employing conformational analysis to yield multiple valid structures.

Our approach utilize a composition of several methods ranging from pure rule based (as classified in [3]) multi dimensional distance geometry method (described in [1]) to data based stored substructure lookup features in a flexible software framework. The actual implementation is a highly portable JAVA software, which fits a broad scale of applications: it is used in small web drawing applets (available at [2]) as well as standalone database processing component.

The coordinate determination process can be best characterized by the "divide and conquer" approach: the structure is composed of fragments, which are joined together. From the available fragment conformers the conformers of the joined structures can be generated during the fusing step. The fragment conformers are generated either through further fragmentation or with an elemental structure/conformer prediction method, consequently the conformational analysis is an inherent part of the building process (in contrast with methods proceeds from 3D initial structures like [4]). The novelty of our approach lies in the diversity of utilised such elemental methods and the arisen scalability options.

**References**

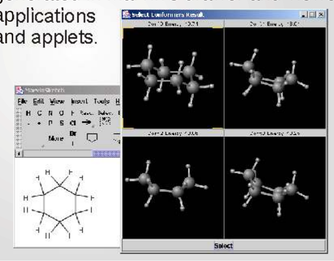
[1] G. Imre, G. Veress, A. Volford and Ö. Farkas, "*Molecules from the Minkowski Space: An approach to building 3D molecular structures*", J. Mol. Struct. (Theochem), *666-667*, 51-59 (**2003**)

[2] http://www.chemaxon.com/marvin

[3] J. Sadowski and J. Gasteiger, "*From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders*", Chem. Rev., *93*, 2567-2581 (**1993**)

[4] J. Weiser, M. C. Holthausen, L. Fitjer, *.HUNTER: "A Conformational Search Program for Acyclic to Polycyclic Molecules with Special Emphasis on Stereochemistry"*, J. Comput. Chem., *18*, 1265-1281 (**1997**)

3D structures and conformers can be generated in Marvin's drawer and viewer applications and applets.



With the functionality given through the public API (Application Programming Interface) the 3D coordinate generation can be easily invoked from custom applications. Fine tuning of the cleaning process can be done by passing parameters.

```
void cleanAndOut( Molecule m ) {
   m.clean(3,"S{conformers}");
   String sdfstr = m.toFormat("sdf");
   System.out.print( sdfstr );
}
```

Batch processing of multiple structures can be automated through provided command line tools, where the fine tuning possibility is also present.

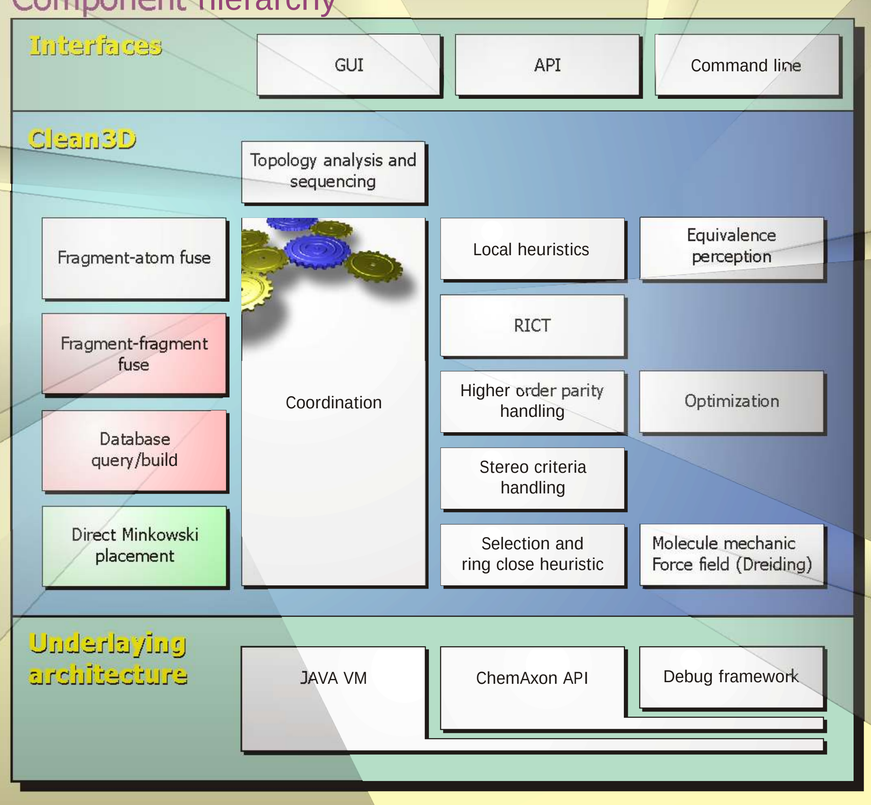
```
>molconvert sdf -3:"S{fine}[timestamp]"
2Dstructures.smiles > 3Dstructures.sdf
```

The processing sequence of atoms has a significant impact on the overall performance of the cleaning process. The most important aspect considering the sequencing is the minimization the built fragments conformational flexibility and the correct handling of critical stereo centers.

By growing the determined fragment, the connected neighbor candidates are ordered by priorities, which are briefly introduced below (highest priority at first):

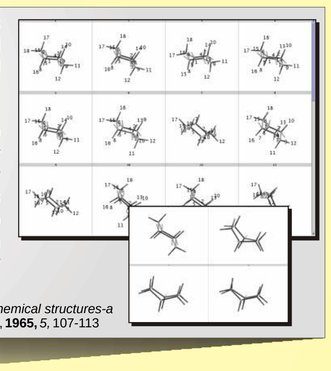
- Neighbor of a placed atom with locked parity or having more than 4 neighbors
- Neighbor of an Sp2 atom
- Atom which will close a ring
- Ring atom
- Other atoms

### Component hierarchy




Diversity of explored conformational space is flawed by the consequences of topological equivalences found in most of the real structures unless they are not recognized.

An algorithm developed which try to map fragment conformers to each other by permutating atoms allowed by their equivalence groups. (Equivalence groups determined using the Morgan-algorithm[1].) To reduce the permutation space to discover in case of differing conformers, a 3D geometry based limiting heuristic utilized.



[1]Morgan, H. L., *.Generation of a unique machine description for chemical structures-a technique developed at Chemical Abstracts Service.*, J. Chem. Doc., **1965,** *5*, 107-113

Iterative redundant internal coordinate transformation series can be used to adjust structures when a goal can be formulated for defined coordinates. After iterations, the defined internal coordinates will be loaded with the same magnitude of error.

At the current version RICT used only at single atom fuse, when important internal coordinate values (bond lengths, angles) can not be satisfied (for example, some ring closures). The method can be utilized for adjust multiple atoms (at ring closure steps, move not only the one closing atom, or at fragment-fragment fusing, align fragment joins).

About the theoretical background[1]:

The *finite* displacement, which is available in internal coordinates ($\Delta\mathbf{q}$) can be transformed into a 3D Cartesian displacement ($\Delta x$) using an *iterative* process, due to the curvilinear nature of internal coordinates: $\mathbf{B}d\mathbf{x} = d\mathbf{q} \qquad \Delta\mathbf{x} \approx \mathbf{B}^{-}\Delta\mathbf{q}$

Where B is the Wilson B-matrix: $B_{i,j} = \frac{\partial q_i}{\partial x_j} \Rightarrow d\mathbf{q} = \mathbf{B}d\mathbf{x}$

[1]Farkas, Ö "*Fast and robust geometry optimization algorithm for large systems*", CESTC **2004**, Tihany, Hungary

In the current deployed version the single atom fusing functionality used for coordinate generation. During a fuse step, multiple possible fused atom orientation will be disclosed for every stored conformer with the following steps:

- Determine possible values of internal coordinates (bond angles, lengths, dihedral angles) by local heuristics
- Try to satisfy most important coordinate variations via triangulation
- If triangulation fails, use iterative redundant internal coordinate transformation steps to construct an initial condition for a geometry optimization (this will load the internal coordinate values with the same magnitude of error)
- Do geometry optimization if required
- Throw identical conformers
- Select conformers to survive for the next step

A flexible interface connects molecular mechanics force fields to the software, which can help to automatize the extension to multidimensional Minkowski space.

Currently the Dreiding[1] force field is implemented, but using others too is planned.

[1]S. L. Mayo, B. D. Olafson, W. A. Goddard III., J. Phys. Chem., **1990,** *94,* 8897-8909

Fragment-fragment fusing is currently under development. In many cases, the stored conformer count can be reduced dramatically (for the same quality; in contrast with the pure single atom fuse build process) if parts are built separated and joined together.

This problem can be solved by means of the following tools:

- The multiple possible orientation of joining bond(s) can be revealed by the machinery deployed for single atom fuses
- A best alignment of the fragments (without changing them) can be calculated
- Aligned fragments can fit using iterative redundant internal coordinate transformation steps or geometry optimization

At the beginning of the Clean3D development a good idea arisen which was worth to develop. The described method[1] was implemented and verified and planned to integrate into the current version.

The method aims the generation of valid 3D coordinates for structures given by topology.

Distance criteria can be established from topology. Atom-atom distance "wishes" mainly comes from estimated or determined internal coordinates (bond lengths, bond angles, dihedral angles). These local assumptions about the 3D geometry may contain inconsistency: 3D coordinates satisfy all of them can not exists.




A non Euclidean space can be defined where all of the internal distance requirements can be satisfied. Define the metric **w** tensor as:

$$\mathbf{W} = \begin{pmatrix} \pm 1 & 0 & \cdots & 0 \\ 0 & \pm 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \pm 1 \end{pmatrix} \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

Accordingly, the norm of a vector (square of "distance", *metrid)* is $d^2(\mathbf{a}) = \mathbf{a}^T\mathbf{W}\mathbf{a}$




$\mathbf{v}_1^T\mathbf{W}\mathbf{v}_1 < 0$

$\mathbf{v}_2^T\mathbf{W}\mathbf{v}_2 = 0$

$\mathbf{v}_3^T\mathbf{W}\mathbf{v}_3 > 0$

This definition induces the presence of singlar directions: in these directions a vector with non-zero coordinates have 0 or negative metrid. The illustration shows a 3D minkowski space with metric tensor (1,1,-1). Vectors lying on the depicted cone surface has 0 metrid values, vectors originating form origo and pointing inside the dual cone has negative metrid values.

For N nodes, any (N-1)*(N-1) internal distance matrix can be satisfied in at most N-1 (not necessity real) dimensions. For this, a straight algorithm is given, which can assign such coordinates for a point which satisfy any distance vector to previously placed points with arbitrary coordinates.

After Minkowski coordinates assigned a geometry optimization used to reduce dimensionality. This optimization step will destroy some of the established distances, but if we choose the used force field properly, the resulting structure will be a low energy, valid conformer.

The main attribute of the used force field is the slight forces pointing from over-3D extra dimensions to zero, which will collapse the structure into 3D. For keeping the structure valid a molecule mechanics or pseudo molecule mechanics is responsible. This can be a classical force field (like Dreiding) extended to multiple dimensions or a pseudo force field based on the original inner distance matrix. Extending a real-world force field is a simple task considering the energy components used can be represented in an at most 3D dimensional subspace.

This method - as expected - can produce valid coordinates for structures with heavy tensions too, but the generation is slow, since te total dimensionality to optimize is proportional to the square of atom count.

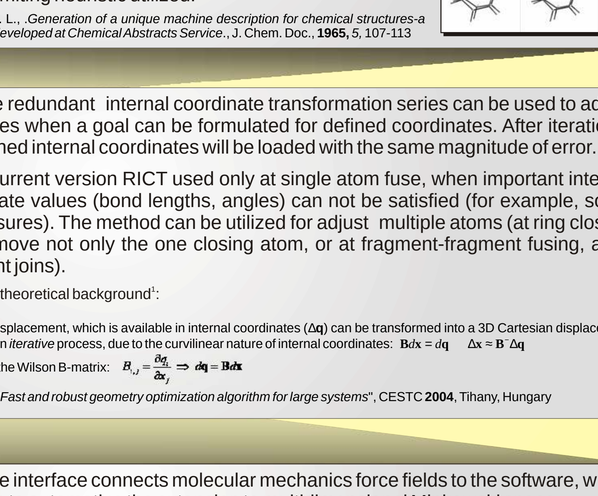
[1]G. Imre, G. Veress, A. Volford and Ö. Farkas, "*Molecules from the Minkowski Space: An approach to building 3D molecular structures*", J. Mol. Struct. (Theochem), *666-667*, 51-59 (**2003**)

## Results

The latest released version of Clean3D was tested on the NCI (National Cancer Institute) database of August 2000 version with ~250.000 molecules.

The coordinate generation failed with the default time limit on less than 0.1 % of input structures.

From them with increased time limit value only 42 structures failed.



## Acknowledgements

Budapest University of Technology and Economics
ChemAxon
InfoPark Budapest