CORINA Symphony Descriptors Community Edition

Algorithms for the Encoding of Molecular Structures

Version 1.0

Brief Description

Molecular Networks GmbH & Altamira LLC

March 2016

mn-am.com

Molecular Networks GmbH
Neumeyerstr. 28
90411 Nuremberg
Germany

Altamira LLC
1455 Candlewood Drive
Columbus, OH 43235-1623
USA

mn-am.com

Content

Introducing CORINA Symphony Descriptors Community Edition

Molecular Networks makes the core set of molecular descriptors provided by **CORINA Symphony** publicly available to the scientific community [1]. This core set, called **CORINA Symphony Descriptors Community Edition (CE)**, includes in total

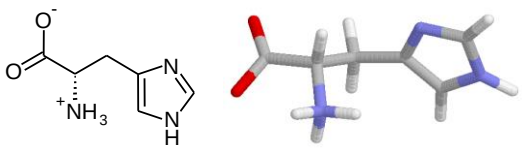
200 molecular descriptors

which are the most versatile and widely applicable descriptors in the areas of drug design, ADME and toxicity prediction, for modeling chemical reactivity and to support the use of computational tools in risk assessment of chemicals.

Objective of CORINA Symphony Descriptors CE

CORINA Symphony Descriptors CE comprises a unique combination of methods for coding molecular structures. It contains a series of methods for the generation of 3D structures, the calculation of physicochemical descriptors based on empirical models for the influences of atoms in molecules and a mathematical transformation technique. The calculated physicochemical descriptors, such as charge distributions or polarizability effects, are used for the representation of molecular structures by means of the mathematical transformation of autocorrelation.

This molecular transform forms a hierarchy of increasing sophistication in representing molecular geometry from the constitution (topological autocorrelation) to the 3D structure (spatial autocorrelation, see Figure 1).



constitution 3D model

Figure 1 A hierarchy of structure representation: from the constitution to the 3D structure.

As a consequence, an entire dataset of molecules can be converted into a uniform representation directly amenable to inductive learning methods such as statistical or pattern recognition methods or neural networks.

The descriptors calculated by **CORINA Symphony Descriptors CE** encode fundamental physicochemical effects and clear-cut geometric features. Thus, relationships between the descriptors from **CORINA Symphony Descriptors CE** and physical, chemical or biological properties are open to a direct interpretation.

Use of Descriptors Provided by CORINA Symphony Descriptors CE

The clear-cut physicochemical basis of molecular descriptors that is realized in **CORINA Symphony Descriptors CE** can serve as a guide in choosing the appropriate descriptors for a given problem.

First, a single molecule can be expressed by a single value that has a clear definition, such as the mean molecular polarizability, the dipole moment or the number of hydrogen bond donor or acceptor atoms. Those descriptors are often called global molecular descriptors (see section "Global Molecular Descriptors" on page 6).

At a second level, descriptors that encode the constitution of a molecule (sometimes also called 2D structure) can be used utilizing a variety of physicochemical atom properties, such as the σ, π or total charges, the effective polarizability, or σ, π or lone pair electronegativities. In order to derive a uniform representation of all molecules in a dataset, the mathematical transformation of autocorrelation is applied. This transformation leads to a vectorial descriptor of fixed length for each molecule independent of its size (see section "Topological or 2D Property-Weighted Autocorrelation" on page 7).

On the next higher level of sophistication, the 3D structure can be used as the basis for the type of descriptor to be calculated. Again, various physicochemical properties, such as the σ, π or total charges, the effective polarizability, or σ, π or lone pair electronegativities, of the atoms of the molecule can be considered. The mathematical transformation of autocorrelation is applied to obtain a vectorial descriptor of fixed length for each molecule independent of its size and orientation in space. These 3D autocorrelation descriptors encode the distribution of atom pair properties in the 3D structure of a molecule (see sections "Spatial or 3D Property-Weighted Autocorrelation" on page 8).

CORINA Symphony Descriptors CE provides a machinery for the representation of molecular structures. From all calculated descriptors, the user can bring into consideration her or his knowledge on the type of effects that are influencing the physical, chemical or biological property to be modeled. The more the user has knowledge about those effects and their influence on the property, the better a guided choice on the types of descriptors that should be used can be made.

Technology

CORINA Symphony Descriptors CE is based on the Chemoinfomatics platform MOSES, developed, maintained and owned by Molecular Networks GmbH [4].

The CORINA Symphony Descriptors CE Web Service

CORINA Symphony Descriptors CE is provided as a web service available on the web server of Molecular Networks at

http://www.molecular-networks.com/services/corinasymphonydescriptors

(see Figure 2).

The **CORINA Symphony Descriptors CE** web service calculates a set of molecular descriptors (in total 200) by processing a file of chemical structures in SDF or SMILES format that is uploaded by the user.

Chemical structure files can be uploaded by typing the full file name of the file into the form next to the **Browse…** button or by clicking the **Browse…** button and selecting a file in the **File Upload** dialog.

The file size of the input file is limited to 1 MB.

After pressing the **Submit** button, the service calculates the molecular descriptors for the molecules in the uploaded file and returns the descriptors in *csv* (comma separated value) file format. The first line in the returned output file is a header line listing the contents of each column (record number, calculated descriptors, error and warning messages).

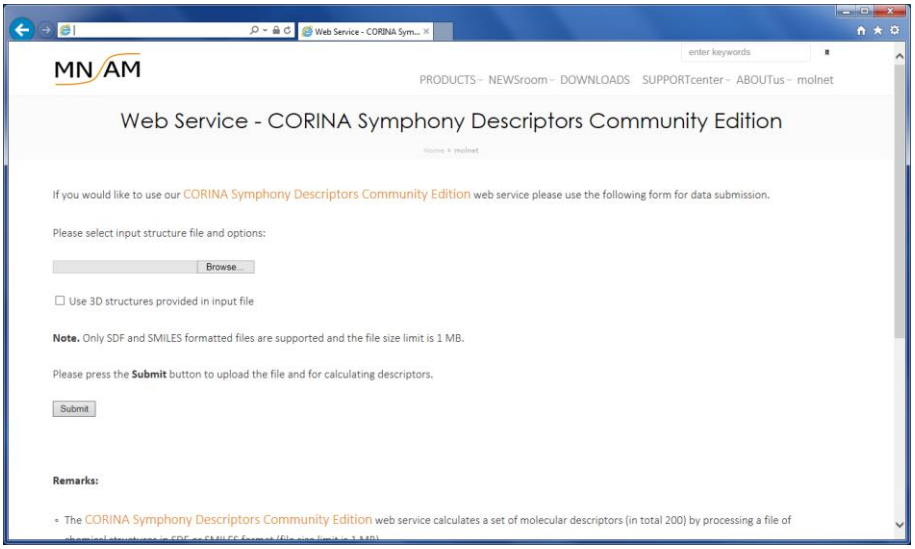


Figure 2 The **CORINA Symphony Descriptors CE** web service.

Supported File Formats

Chemical structures are accepted in the following file formats:

- Accelrys SD file [5]

- Daylight SMILES [6]

The input file format is determined automatically. If an unsupported file format is submitted the following error message will be written out:

```
Error: input file format not supported (SDF or SMILES only)!
```

Submission of 3D SDFiles

3D atom coordinates that are provided in the input file can be used for calculating descriptors that require a valid 3D structure (*e.g.*, dipole moment) by checking the box "`Use 3D structures provided in input file`".

A valid 3D structure requires that

- all atoms are defined including all hydrogen atoms (no implicit hydrogen atoms)
- all atoms have 3D coordinates

If an in-valid input 3D structure is encountered the internal library of the 3D structure generator **CORINA** will automatically generate a new 3D structure [7]-[9]. An appropriate warning message is written to the last column of the descriptor. Such messages are written out individually for each record (molecule) where a warning or an error occurs.

Descriptor Output File

The descriptor output file is *csv* (comma separated value) formatted and consists of a header line and a data block.

The first line is the header line and names the columns in the *csv* file by the abbreviations of the calculated descriptors.

The data block contains the record numbers and the calculated descriptors. Each line in the data block in the *csv* file (data line) represents one molecule of the input structure file. The first column is the record (or molecule) number in the order as they are stored in the structure input file followed by descriptor columns (integers and floating point values). The last three columns contain warning and error messages that may have occurred during the calculation process.

The value "`NULL`" in a field of the descriptor matrix indicates that the calculation of this descriptor failed. In this case, an error (or warning) message should appear in the column "`Error_messages`".

Error and Warning Messages in *csv* Output File

Error and warning messages that are given in *csv* output file in the last three columns correspond to the calculation of the descriptors by **CORINA Symphony Descriptors**

CE. A value of "`none`" indicates that no error or warning message occurred for the respective record (molecule).

Error messages always start with the sequence `@ERROR`.

Warning messages always start with the sequence `@WARNING`.

A special warning message is written to the last column "`3D_warning_messages`" when the box "`Use 3D structures provided in input file`" was checked but an in-valid or no 3D structure was encountered for a specific record (molecule). For this record, a newly generated 3D model by CORINA was used. The warning message is "`@3D_WARNING: 3D model not correct/available: using CORINA`".

Molecular Descriptors Calculated by CORINA Symphony Descriptors CE

In the following sections the molecular descriptors that are calculated by **CORINA Symphony Descriptors CE** are described.

Global Molecular Descriptors

Global molecular descriptors represent a chemical structure by a structural, chemical or physicochemical feature or property of the molecule expressed by a single value. They are derived either from the gross formula, the constitution (2D structure, i.e., the connection table that is the list of atoms and bonds present in a molecule) or the 3D structure of a molecule. Prior to the calculation of the descriptors, implicit hydrogen atoms are added to the input structure. 3D Cartesian coordinates are calculated by the integrated module of the 3D structure generator **CORINA** unless they are not provided in the input file [7]-[9].

The following global molecular descriptors are calculated (see Table A).

Table A Available global molecular descriptors.

Descriptor	Description	Abbreviation
Molecular weight	Molecular weight in [g/mol] derived from the gross formula [10]	Weight
Number of hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule [11]	HAcc
Number of oxygen atom-based hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of oxygen atoms only in the molecule [11]	HAcc_O
Number of nitrogen atom-based hydrogen bonding acceptors	Number of hydrogen bonding acceptors derived from the sum of nitrogen atoms only in the molecule [11]	HAcc_N
Number of hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule [11]	HDon
Number of oxygen atom-based hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of O-H groups only in the molecule [11]	HDon_O
Number of nitrogen atom-based hydrogen bonding donors	Number of hydrogen bonding donors derived from the sum of N-H groups only in the molecule [11]	HDon_N
Octanol/water partition coefficient (log*P*)	Octanol/water partition coefficient in [log units] of the molecule following the XlogP	XlogP

	approach [12]	
Topological polar surface area	Topological polar surface area in [$Å^2$] of the molecule derived from polar 2D fragments [13]	TPSA
Number of rotatable bonds	Number of open-chain, single rotatable bonds [24]	NRotBond
Number of Rule of 5 violations	Number of violations of the Lipinski's rule of 5 (Weight > 500, XlogP > 5, HDon > 5, HAcc > 10) [11]	NViolationsRo5
Number of extended Rule of 5 violations	Number of violations of the extended Lipinski's rule of 5 (additional rule: number of rotatable bonds > 10) [11]	NViolationsExtRo5
Number of atoms	Number of all atoms in the molecule (including hydrogen atoms)	NAtoms
Number of tetrahedral stereo centers	Number of tetrahedral chiral centers in the molecule	NStereo
Molecular complexity	Molecular complexity according to the approach by J. Hendrickson [25]	Complexity
Ring complexity	Ring complexity according to the approach by J. Gasteiger and C. Jochum [26]	RComplexity

Vectorial Molecular Descriptors

The vectorial molecular descriptors that are available in **CORINA Symphony Descriptors CE** represent a chemical structure by a vector of fixed length independent of the size of the molecule. These vectors are derived either from the 2D, the 3D or the molecular surface representation of a chemical compound in combination with physicochemical atom or surface properties. Since only internal coordinates (topological or spatial distances of atom pairs) are taken into account, the resulting descriptors (vectors) are independent of the orientation of the molecules in space (translation and rotation invariant). Therefore, no preprocessing alignment of the molecules of a dataset under investigation is necessary.

Topological or 2D Property-Weighted Autocorrelation

Topological or 2D property-weighted autocorrelation (2D autocorrelation or 2D autocorrelation vectors) uses the 2D structure of a molecule (i.e., the molecular graph as expressed by the connection table) and atom pair properties as a basis to derive vectorial molecular descriptors [1],[27]. The products of atom pair properties are summed up for certain topological distance that is the number of bonds on the shortest path between two atoms. Thus, for each topological distance a single value is obtained that is one coefficient of the resulting 2D autocorrelation vector.

Table B lists the atom pair properties that are used for 2D autocorrelation.

Table B Available atom pair properties for 2D autocorrelation.

Atom Pair Property	Description	Abbreviation
Identity	2D autocorrelation weighted by atom identities, i.e., "1" for an atom	2DACorr_Ident
σ Charge	2D autocorrelation weighted by σ atom charges [18]-[21]	2DACorr_SigChg
π Charge	2D autocorrelation weighted by π atom charges [22]-[23]	2DACorr_PiChg
Total charge	2D autocorrelation weighted by total atom charges (sum of σ and π charges) [18]-[23]	2DACorr_TotChg
σ Electronegativity	2D autocorrelation weighted by σ atom electronegativities [18]-[21]	2DACorr_SigEN
π Electronegativity	2D autocorrelation weighted by π atom electronegativities [22]-[23]	2DACorr_PiEN
Lone pair electronegativity	2D autocorrelation weighted by lone pair electronegativities [22]-[23]	2DACorr_LpEN
Effective polarizability	2D autocorrelation weighted by effective atom polarizabilities [14]-[17]	2DACorr_Polariz

The 2D autocorrelation vectors are calculated using the following parameters that have been proven to be useful for most modeling purposes.

- Hydrogen atoms are ignored and only non-hydrogen (heavy) atoms are taken into account.
- The minimum topological distance taken into account is "0", *i.e.*, the first coefficient (element) of the 2D autocorrelation vector is the sum of the products of atom pair properties of each atom with itself.
- The maximum topological distance taken into account is "10", *i.e.*, up to ten intervening bonds between an atom pair.

Therefore, the dimensionality of a single 2D autocorrelation vector is "11". In total, eight eleven-dimensional 2D autocorrelation vectors using eight different atom pair properties (see Table B) are calculated for a molecule.

Spatial or 3D Property-Weighted Autocorrelation

Spatial or 3D autocorrelation (3D autocorrelation or 3D autocorrelation vectors) uses the 3D structure of a molecule (*i.e.*, the Cartesian atomic coordinates) and atom pair properties as a basis to derive vectorial molecular descriptors [3],[28]. The products of atom pair properties are summed up for certain 3D distance intervals. Thus, for each 3D distance interval a single value is obtained that is one coefficient of the resulting 3D autocorrelation vector.

If no 3D structure is provided in the input file the integrated module of the 3D structure generator **CORINA** is used to generate 3D Cartesian coordinates [7]-[9]. Prior to the calculation of the descriptors implicit given hydrogen atoms are added to the input structure.

Table C lists the atom properties that are available for 3D autocorrelation.

Table C Available atom pair properties for 3D autocorrelation.

Atom Pair Property	Description	Further Details
Identity	3D autocorrelation weighted by atom identities, i.e., "1" for an atom	3DACorr_Ident
σ Charge	3D autocorrelation weighted by σ atom charges [18]-[21]	3DACorr_SigChg
π Charge	3D autocorrelation weighted by π atom charges [22]-[23]	3DACorr_PiChg
Total charge	3D autocorrelation weighted by total atom charges (sum of σ and π charges) [18]-[23]	3DACorr_TotChg
σ Electronegativity	3D autocorrelation weighted by σ atom electronegativities [18]-[21]	3DACorr_SigEN
π Electronegativity	3D autocorrelation weighted by π atom electronegativities [22]-[23]	3DACorr_PiEN
Lone pair electronegativity	3D autocorrelation weighted by lone pair electronegativities [22]-[23]	3DACorr_LpEN
Effective polarizability	3D autocorrelation weighted by effective atom polarizabilities [14]-[17]	3DACorr_Polariz

The 3D autocorrelation vectors are calculated using the following parameters that have been proven to be useful for most modeling purposes.

- Hydrogen atoms are ignored and only non-hydrogen (heavy) atoms are taken into account.
- The minimum spatial distance taken into account is "1 Å", *i.e.*, the first coefficient (element) of the 3D autocorrelation vector is the sum of the products of atom pair properties of atom pairs that are apart from each other between 1 to 2 Å.
- The maximum spatial distance taken into account is "13 Å", *i.e.*, up to ten intervening bonds between an atom pair.
- The number of equal 3D distance intervals is set to "12", *i.e.*, the first interval sums up the atom pair property products from 1 to 2 Å.

Therefore, the dimensionality of a single 3D autocorrelation vector is "12". In total, eight twelve-dimensional 3D autocorrelation vectors using eight different atom pair properties (see Table C) are calculated for a molecule.

References

[1] Chemoinformatics Program Package CORINA Symphony, developed and distributed by Molecular Networks GmbH, Nuremberg, Germany and Altamira LLC, Columbus, OH, USA (www.mn-am.com).

[2] Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205-1213.

[3] Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769-7775.

[4] MOSES is a C++ software library that has been developed in a joint project of the research group of Prof. Johann Gasteiger at the Computer-Chemie-Centrum of the University of Erlangen-Nuremberg (see www2.chemie.uni-erlangen.de) and Molecular Networks GmbH, Nuremberg, Germany. MOSES is taken over by Molecular Networks GmbH for maintenance, further development and commercial distribution of software products based on MOSES (see www.mn-am.com/moses).

[5] 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 CTfile formats (Mol, SDF and RDF) is available on the Internet for download as a PDF document at www.accelrys.com.

[6] 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.

[7] Review Articles: (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.

[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] 3D Structure Generator CORINA, developed and distributed by Molecular Networks GmbH, Nuremberg, Germany (www.mn-am.com).

[10] Atomic weights were taken from www.webelements.com and are currently implemented up to the atomic number of 109 (Mt, Meitnerium).

[11] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3-25.

[12] Wang, R.; Gao, Y; Lai, L. Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47-66.

[13] Ertl, P; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Tansport Properties. *J. Med. Chem.* **2000**, *43*, 3714-3717.

[14] Gasteiger, J.; Hutchings, M.G. Empirical Models of Substituent Polarisability and their Application to Stabilisation Effects in Positively Charged Species. *Tetrahedron Lett.* **1983**, *24*, 2537-2540.

[15] Gasteiger, J.; Hutchings, M.G. Quantitative Models of Gas-Phase Proton Transfer Reactions Involving Alcohols, Ethers and their Thio Analogs. Correlation Analyses Based on Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489-6495.

[16] Kang, Y.K.; Jhon, M.S. *Theor. Chim. Acta* **1982**, *61*, 41.

[17] Miller, K.J. Additivity Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533-8542.

[18] a) Hinze, J.; Jaffe, H.H. *J. Am. Chem. Soc.* **1962**, *84*, 540. b) Hinze, J.; Jaffe, H.H. *J. Am. Chem. Soc.* 1963, 85, 148. c) Hinze, J.; Jaffe, H.H. *J. Phys. Chem.* **1963**, *67*, 1501.

[19] Gasteiger, J; Marsili, M. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett.* **1978**, *34*, 3181-3184.

[20] Gasteiger, J; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219-3228.

[21] Gasteiger, J; Guillen, M.D. Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring Systems. *Tetrahedron* **1983**, *39*, 1331-1335.

[22] Bauerschmidt, S.; Gasteiger J. Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 705-714.

[23] (a) Saller, H.; Gasteiger, J. Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 687-689. (b) Saller, H.; Gasteiger, J. Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes. *Angew. Chem.* **1985**, *97*, 699-701.

[24] Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45 (12)*, 2615-2623.

[25] Hendrickson, J.B.; Huang, P.; Toczko, A.G. Molecular Complexity: A Simplified Formula Adapted to Individual Atoms. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63-67.

[26] Gasteiger, J; Jochum, C. An Algorithm for the Perception of Synthetically Important Rings. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43-48.

[27] Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359-360.

[28] Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem. Chim. Ther.* **1984**, *19*, 66-70.