

CypScoreMOE'09

Table of Contents

1	Introduction.....	2
2	Prerequisites.....	3
2.1	ParaSurf.....	3
2.2	VAMP or MOPAC.....	3
2.3	MOE	3
3	Obtaining and Installing CypScoreMOE.....	3
4	Using the software.....	4
4.1	Input	4
4.2	Output.....	4
4.3	Interactive Mode.....	5
4.4	Browsing Results	8
4.5	Batch Mode	9
5	Model Assignment.....	11
6	CypScore Scaling.....	12
7	Support.....	14
7.1	Contact	14
7.2	CAChe Research LLC.....	14
8	References.....	14

1 Introduction

CypScoreMOE predicts positions in drug-like molecules that are likely sites for metabolism by cytochromes P450. It implements the CypScore method devised by Hennemann *et al.* [1] within the Chemical Computing Group's Molecular Operating Environment (MOE) [2]. CypScoreMOE processes molecules stored in a MOE molecular database and may be run interactively from MOE's Database Viewer or in batch mode via the moebatch program. Reactivity scores calculated by CypScoreMOE are added to the molecular database and the results can be viewed graphically in the MOE Database Browser.

Based on a hypothetical "P450 super-enzyme", Hennemann *et al.* [1] derived six models for the oxidation reactions listed in Table 1. Each model is a linear equation $y = a_0 + \sum_{i=1}^4 a_i x_i$ predicting an atom's reactivity y from the values of up to four descriptors x_i calculated by ParaSurf [3]. The descriptors involved in each model and the coefficients a_i are defined in the original publication.

Table 1: CypScore reaction models

Model	Reaction
1	aliphatic hydroxylation, N-dealkylation, O-dealkylation
2	aromatic hydroxylation
3	double-bond epoxidation/oxidation
4a	N-oxidation of amines
4b	N-oxidation of imines
5	S-oxidation

CypScoreMOE assigns a model to each C, N and S atom visible on the molecular surface, then calculates the value of that model using the appropriate ParaSurf descriptors. The model assignment, described in section 5, uses atom types from the mmff94x force field.

Each of the CypScore models produces reactivity predictions on an individual scale. Hennemann *et al.* [1] described a procedure to establish a common reactivity scale for all the models, providing a CypScore value S for each atom in the range from 0 (stable) to 100 (reactive), and implemented this procedure for a proprietary data set. Unfortunately, Hennemann *et al.* [1] do not include sufficient detail to define completely the relationship between the original model values y and the scaled scores S , and it is not possible to reproduce the missing information as the data from which it was derived is not available publicly. To overcome this, CypScoreMOE employs a common reactivity scale that is a close approximation to the one used by Hennemann *et al.* [1], derived from secondary information in the original paper through a process described in section 6. Consequently, the results obtained by CypScoreMOE will approximate closely, but not match exactly, those in the original publication.

2 Prerequisites

2.1 ParaSurf

The models of cytochrome P450-mediated reactions used by CypScoreMOE are based on molecular surface descriptors calculated by ParaSurf [3], which is available from CACHE Research LLC (www.CacheResearch.com). ParaSurf'09 or a later version is required.

2.2 VAMP or MOPAC

ParaSurf requires the results of semi-empirical molecular orbital calculations. These can be provided either by VAMP [4] or by CEPOS MOPAC 6.

VAMP is available from Accelrys Software Inc. (www.accelrys.com).

CEPOS MOPAC 6 is a modified, public-domain version of MOPAC [5] available for free download from CEPOS InSilico Ltd. (www.ceposinsilico.com).

Note that the version of MOPAC distributed with MOE is *not* sufficient for CypScoreMOE.

2.3 MOE

MOE version 2008.10 or later is required.

3 Obtaining and Installing CypScoreMOE

CypScoreMOE consists of a single file `cypscore.svl` containing a script written in MOE's Scientific Vector Language (SVL). It is available for free download from CACHE Research LLC (www.CacheResearch.com).

The `cypscore.svl` script must be loaded into MOE before use. It may be installed in a directory from which MOE automatically loads SVL scripts on start-up, as described in the MOE documentation, or it may be loaded as required from any location using the 'Modules & Tasks' panel on the MOE 'Windows' menu.

The script provides a global command 'cypscore []' that can be executed interactively from the MOE Database Viewer, or in batch mode, as described in section 4. For convenience the stanza

```
MENU APPEND "dbv:Compute"  
    "CypScore..." exec 'cypscore []'  
ENDMENU
```

may be added to an appropriate `moe-menus` file (e.g. `~/moe-menus` or `$MOE/lib/svl/custom/moe-menus`, see the MOE documentation). This will append a 'CypScore...' entry to the 'Compute' menu in the MOE Database Viewer that will launch the interactive CypScoreMOE panel.

Correct functioning of CypScoreMOE requires that the environment variable `PARASURF_ROOT` is set to the directory containing the parameter file `Vhamil.par` distributed with ParaSurf. For the Windows version of ParaSurf this is done automatically by the ParaSurf installer so no further action is required. For Linux versions of ParaSurf this is normally done by sourcing

either `$CEPOS/cepos09/etc/cepos.sh` or `$CEPOS/cepos09/etc/cepos.csh` (depending on the shell in use), where `$CEPOS` is the base directory of the CEPOS installation (typically `/usr/cepos`), as described in the ParaSurf installation instructions. Alternatively, `PARASURF_ROOT` may be set directly by the command

```
export PARASURF_ROOT=/usr/cepos/cepos09/etc/
```

in a Bourne-type shell (`sh`, `bash`, `ksh`), or

```
setenv PARASURF_ROOT /usr/cepos/cepos09/etc/
```

in a C-type shell (`csh`, `tcsh`), adjusting the path to the directory containing `Vhamil.par` in your installation. Note that the trailing slash *is* required.

If CEPOS MOPAC 6 is used for the molecular orbital calculations it is recommended that the `mopac6` executable is installed in the same directory as the ParaSurf executable.

4 Using the software

4.1 Input

CypScoreMOE processes molecules stored in MOE molecular databases. If the database has not already been processed by ParaSurf then it must contain a molecule field where each entry has 3D coordinates and hydrogen atoms added to complete valences. In this case, for each molecule the program will run either MOPAC or VAMP, followed by ParaSurf, in order to calculate the CypScore values. If a set of molecules has already been processed by ParaSurf, then the ParaSurf SD output files `<molecule>_p.sdf` may be opened in a molecular database, importing the block tagged `<CYPSCORE>` in the ParaSurf SD file. In this case the CypScore values will be calculated directly from the information in the `<CYPSCORE>` field and the MOPAC/VAMP and ParaSurf runs will not be repeated. In order to pre-process a set of molecules by ParaSurf, note that the CypScore models were derived using a specific set of ParaSurf options, and ParaSurf only calculates the `<CYPSCORE>` block for this combination of options:

```
surf=cube estat=multi contour=isoden iso=0.0003
```

These values for `estat`, `contour` and `iso` are the default values if `surf=cube` is specified, so it is sufficient to run ParaSurf with `surf=cube`.

4.2 Output

CypScoreMOE adds two fields labelled 'Atom CypScore' and 'Bond CypScore' to the database being processed. Each cell in the 'Atom CypScore' column has a row for each heavy atom in the molecule containing three items: the atom number, the model type assigned to the atom (1, 2, 3, 4a, 4b or 5; with 0 indicating that no model was assigned), and the CypScore value for the atom in the range from 0 (stable) to 100 (reactive). Each cell in the 'Bond CypScore' column has a row for each pair of bonded heavy atoms containing the atom numbers and the mean scores of the bonded atoms.

4.3 Interactive Mode

To run CypScoreMOE in interactive mode follow these steps:

- Open a molecular database containing the fields described in section 4.1 in the MOE Database Viewer.
- Ensure that the MOE current working directory (CWD) is writeable (intermediate files from MOPAC, VAMP and ParaSurf will be created here temporarily).
- Launch the CypScoreMOE panel by selecting 'CypScore...' from the 'Compute' menu of the Database Viewer, or by entering 'cypscore []' in the Database Viewer command window.
- Select the options required in the CypScoreMOE panel, which is shown in Figure 1.
- Select OK to start the calculation.
- Monitor the SVL commands window for progress and error messages. Progress messages are also displayed in the main MOE window.

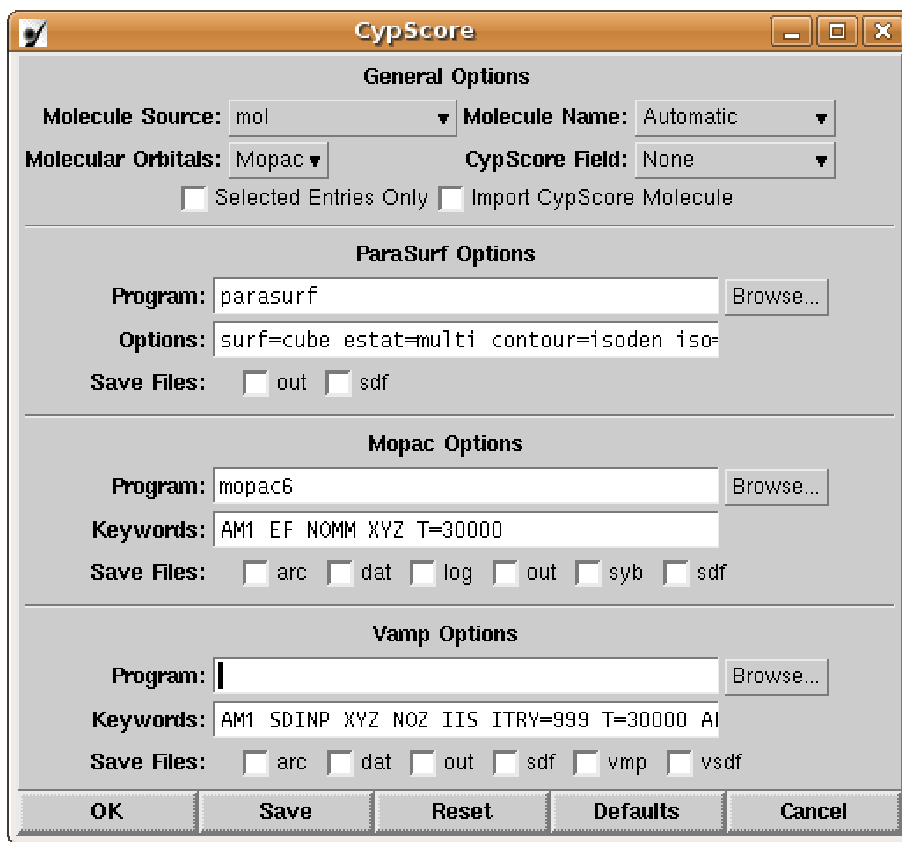


Figure 1. The CypScoreMOE panel.

The options in the CypScoreMOE panel are described in Tables 2 to 5. Pop-up help is also provided for each option. It is usually necessary to specify the 'Molecule Source' and either the 'Molecular Orbitals' program or the 'CypScore Field', depending on whether or not the database has been pre-processed by ParaSurf. The default settings for the remaining options are generally sufficient. As the MOPAC/VAMP and ParaSurf calculations are (relatively) time consuming it may be worthwhile saving the ParaSurf SD output files. The buttons at the foot of the CypScoreMOE panel allow the settings to be saved and restored, as described in Table 6.

Table 2: General Options

Molecule Source	Select the molecule field to be used as input from a drop-down list containing all the molecule fields in the database. The chosen column of the database should contain a 3D structure with hydrogen atoms added.
Molecular Orbitals	The program to be used to generate molecular orbitals (either MOPAC or VAMP).
Molecule Name	A field in the database to be used as a prefix in naming intermediate files generated by ParaSurf and MOPAC or VAMP. If Automatic is chosen intermediate files are prefixed with molecule_1, molecule_2, molecule_3, etc. This is only relevant if the intermediate files are saved. Take care that the entries in this field are valid filenames on the operating system in use.
CypScore Field	If the database includes a field containing the CYPSCORE block imported from a ParaSurf SD file, select it here to use the CYPSCORE block directly without running ParaSurf again. If None is selected then ParaSurf will be run on each molecule.
Selected Only	If this box is checked, only entries that are selected in the Database Viewer will be processed; otherwise all entries in the database are processed.
Import CypScore Molecule	If this box is checked and a MOPAC or VAMP run is performed, the 3D structure produced by MOPAC or VAMP, on which the CypScore calculation is based, is imported into the database in a field named 'CypScore Molecule'.

Table 3: ParaSurf Options

Program	Specify the ParaSurf program. Enter the filename if it lies on the PATH, otherwise enter the full pathname. The Browse button may be used to locate the ParaSurf executable.
Options	The command line options for the ParaSurf program. Note that the options required by CypScoreMOE are quite specific, so do not change the default options provided. However, it is generally safe to add options that produce additional output (e.g. table=<filename>).
Save Files	<p>If these boxes are checked, the corresponding ParaSurf output files will be saved. Otherwise they are deleted after each molecule has been processed.</p> <p>out Save the ParaSurf output file <molecule>_p.out. sdf Save the ParaSurf SD output file <molecule>_p.sdf.</p>

Table 4: MOPAC Options

Program	Specify the MOPAC program. Enter the filename if it lies on the PATH, otherwise enter the full pathname. The Browse button may be used to locate the MOPAC executable.
Keywords	<p>The keywords to control the MOPAC program. The default keywords are: AM1 EF NOMM XYZ T=30000</p>
Save Files	<p>If these boxes are checked, the corresponding MOPAC input and output files will be saved. Otherwise they are deleted after each molecule has been processed.</p> <p>out Save the MOPAC output file <molecule>.out dat Save the MOPAC input file <molecule>.dat log Save the MOPAC log file <molecule>.log arc Save the MOPAC archive file <molecule>.arc syb Save the MOPAC Sybyl output file <molecule>.syb sdf Save the MOPAC SD output file <molecule>_m.sdf</p>

Table 5: VAMP Options

Program	The full pathname of the VAMP program. The Browse button may be used to locate the VAMP executable.
Keywords	The keywords to control the VAMP program. The default keywords are : AM1 SDINP XYZ NOZ IIS ITRY=999 T=30000 ALLVECT
Save Files	If these boxes are checked, the corresponding VAMP input and output files will be saved. Otherwise they are deleted after each molecule has been processed. arc Save the VAMP archive file <molecule>.arc dat Save the VAMP input file <molecule>.dat out Save the VAMP output file <molecule>.out sdf Save the VAMP SD input file <molecule>.sdf vmp Save the VAMP output file <molecule>.vmp vsdf Save the VAMP SD output file <molecule>_v.sdf

Table 6: CypScoreMOE panel buttons.

OK	Run CypScoreMOE with the selected options.
Save	Save the current settings in the user's resource configuration file ~/.moe-rc.
Reset	Restore the most recently saved settings.
Defaults	Restore the default settings.
Cancel	Close the CypScoreMOE panel.

4.4 Browsing Results

A visual representation of the CypScoreMOE results is provided in the MOE Database Browser, launched by selecting 'Browser...' from the 'File' menu in the Database Viewer. For databases that have been processed by CypScoreMOE the Database Browser includes an entry 'CypScore Viewer' in the 'Subject' drop-down. This displays a 2D view of each molecule annotated by the CypScoreMOE results. The CypScore values for each atom are shown in green. Following Hennemann *et al.* [1] atoms with CypScore values of 38 and above are considered reactive and are labelled in red, those with values in the range from 22 to 37 are considered to have medium reactivity and are labelled in blue and those with values below 22 are considered stable and

coloured black. The CypScore Viewer also provides 'Print' and 'Export' buttons for the 2D diagrams.

4.5 Batch Mode

CypScoreMOE can also be run in batch mode via the moebatch program. This may be more convenient for large databases that take a substantial amount of time to process. The command syntax is

```
moebatch -exec "cypscore [[option:value, option:value,...]]"
```

with optional parameters passed as a tagged vector of option:value pairs. The possible options are listed in Table 7 along with their default values. These correspond directly to the settings in the CypScoreMOE panel (Figure 1). The 'save_*' options are Boolean flags taking the values 0 or 1. The remaining options are strings that should be entered within single quotes.

Table 7: Batch mode options.

Option	Default Value
mol_field	'mol'
mo_prog	'Mopac'
cypscore_field	'None'
molname_field	'Automatic'
Import_cypscore_mol	0
parasurf_exe	'parasurf'
parasurf_opts	'surf=cube estat=multi contour=isoden iso=0.0003'
mopac_exe	'mopac6'
mopac_keywords	'AM1 EF NOMM XYZ T=30000'
vamp_exe	"
vamp_keywords	'AM1 SDINP XYZ NOZ IIS ITRY=999 T=30000 ALLVECT'
save_parasurf_out	0
save_parasurf_sdf	0

save_mopac_arc	0
save_mopac_dat	0
save_mopac_log	0
save_mopac_out	0
save_mopac_syb	0
save_mopac_sdf	0
save_vamp_arc	0
save_vamp_dat	0
save_vamp_out	0
save_vamp_sdf	0
save_vamp_vmp	0
save_vamp_vsdf	0

In addition to the options listed in Table 7 the 'mdb' option is required to specify the name of the molecular database to be processed. For example, the command

```
moebatch -exec "cypscore [[mdb:'set0.mdb',
                        mol_field:'mol3D',
                        save_parasurf_sdf:1]]"
```

runs CypScoreMOE on the database 'set0.mdb' using a molecule field named 'mol3D', saving the ParaSurf SD output files. (The command should be entered as one line; it is broken here for clarity.)

Note that the ~/.moe-rc file is read by moebatch, so any options saved there from interactive runs of CypScoreMOE will be used in batch mode unless over-ridden on the command line.

5 Model Assignment

CypScoreMOE assigns one of the six CypScore models to each C, N or S atom with non-zero area on the solvent accessible surface. For carbon atoms this assignment is shown in Table 8, where the atom types are defined by the mmff94x force field in the file \$MOE/lib/mmff94x.ff.

Table 8: Assignment of carbon atoms to CypScore models

Atom Type	Model
C	1
CR3R	1
CR4R	1
Car	2
C5A	2
C5B	2
C5	2
Csp2	3
CE4R	3

For model 1, hydrogen atoms with force field type HC are first identified and their scores are found. If the HC atom is connected to a type C, CR3R or CR4R carbon atom then the HC score is transferred to the carbon atom. If more than one HC atom is connected to a type C, CR3R or CR4R carbon atom then the maximum HC score is assigned to the carbon atom.

For model 3 (double bond epoxidation/oxidation), the carbon atoms are checked to see if they are double bonded to another carbon atom. If both carbon atoms are visible on the molecular surface then they are assigned the mean of their individual scores. If only one of a double bonded pair is visible (*i.e.* its partner is buried) then it retains its individual score.

For the N-oxidation models, model 4a is assigned to the nitrogen atom in R1-N(-R2)-R3 groups and model 4b is assigned to the nitrogen atom in R1=N-R2 groups.

Model 5 is used for all sulphur atoms.

6 CypScore Scaling

To express the reactivity predictions of the CypScore models on a common scale, Hennemann *et al.* [1] employ an empirical procedure based on the enrichment curves obtained for the dataset that was used to train the models. The cross-over value where the false metabolic rate equals the false stable rate is taken as a common reference point for each model. The cross-over points for the individual models are aligned and the range of values for each model is scaled and truncated to lie between 0 (representing stable atoms) and 100 (representing highly reactive atoms).

Specifically, the following sequence of transformations is applied. For a particular model, let y denote the raw score given by the corresponding equation in the paper. First, the cross-over point C for the false metabolic and false stable rates is translated to the origin, defining $y_1 = y - C$. Next, the range $[\max y_1, \min y_1]$ is scaled linearly to fit the interval $[-10, 10]$. Thus define y_2 by

$$y_2 = \begin{cases} -10y_1/\max y_1, & \max y - C \geq C - \min y \\ 10y_1/\max y_1, & \max y - C < C - \min y \end{cases}$$

Finally, the range $[-7, 3]$ for y_2 is mapped linearly to $[0, 100]$ and values outside this range are mapped to the endpoints. Hence the scaled score S is defined by

$$S = \begin{cases} 0, & y_2 \geq 3 \\ -10y_2, & -7 < y_2 < 3 \\ 100, & y_2 \leq -7 \end{cases}$$

For each model, if the transformation from y to y_2 is written as $y_2 = a(y - b)$ then the map from y to S can be written

$$S = \begin{cases} 0, & y \leq y_{min} \\ Ay + B, & y_{min} < y < y_{max} \\ 100, & y \geq y_{max} \end{cases} \quad (*)$$

where $A = -10a$, $B = 10ab + 30$, $y_{min} = b + (3/a)$ and $y_{max} = b - (7/a)$.

Hennemann *et al.* [1] followed this procedure using the false positive and false negative rates for a proprietary dataset and the values of a and b (or their equivalents) obtained for this dataset are not revealed. Hence the precise scaling used in the paper is not known.

However, an approximation to the common reactivity scale can be recovered from a careful examination of Figure 3 in Hennemann *et al.* [1]. This figure contains four plots, each showing the scores for one of the four validation datasets used in the paper: three public domain datasets provided as supporting information (set0, set1 and set2), and one proprietary dataset (set3). Each column in these plots depicts the scores for the atoms in a single molecule as coloured squares, with reactive atoms in blue and stable atoms in red. The unscaled model scores for the public domain datasets can be found by following the protocols described in the

paper. By comparing the plots for the public domain datasets in Figure 3 with the unscaled model scores, it is then possible to detect, approximately, the scaling parameters in equation (*).

The difficulty in doing this is that while each column in Figure 3 contains the set of scores for a single molecule, there is no information to show which model was used for a particular data point; the scores from the six models are interleaved down each column in the figure. The task is therefore to match the data points in the figure with specific atoms. Provided sufficient matches can be made for atoms assigned to a particular model, a simple linear regression of the scaled against the unscaled scores will then reveal the scaling transformation for that model.

To accomplish this, numerical scores were extracted from Figure 3 for set0 and set2 by matching the data points in the figure, by hand, to the nearest point on a scale with intervals of length 0.5. (Set1 was not used as its plot is so dense that individual data points cannot be distinguished reliably; in fact, it is difficult even to detect which columns many data points lie in.)

An iterative strategy was then followed to match data points to atoms. First, reactive atoms in set0 and set2 (those with CYP entries in the SD files supplied as supporting information) were matched, as far as possible, to blue points in the figure. Next, the higher scoring atoms were examined for cases where there was a clear correspondence between atoms and data points. Some molecules, for instance, have models of type 1 or type 2 only. After a first pass through set0 and set2, very rough relationships between the raw and scaled scores emerged for models 1 and 2, which form the majority of the cases. A second pass through set0 and set2 was then made to reassign the worst outliers in the first pass model 1 and model 2 results. This in turn allowed more atoms to be assigned to data points and the process was repeated until a sufficient number of matches were established for atoms assigned to each model.

This laborious process is prone to various types of possible error: differences in atom typing and model assignment, the mismatch in the numbers of atoms with non-zero scores in the validation sets and the number of visible points in Figure 3 (some points are evidently plotted on top of one another), and inaccuracies in assigning numerical scores to the data points in the figure. Nevertheless, convincing solutions were found for all the models, and the resulting scaling parameters, using the notation of equation (*), are shown in Table 9, along with the number of atoms assigned for each model (N) and the R^2 value for the linear fit of S against y . Here y_{min} and y_{max} are given by $y_{min} = -B/A$ and $y_{max} = (100 - B)/A$.

Table9: CypScore scaling parameters

Model	A	B	y_{min}	y_{max}	N	R^2
1	236	-19.6	0.083	0.51	279	0.98
2	497	-12.7	0.026	0.23	194	0.92
3	268	0.19	-0.001	0.37	14	0.99
4a	840	-26.3	0.031	0.15	24	0.90

4b	778	-13.6	0.017	0.15	11	0.96
5	107	-7.65	0.071	1.01	8	0.93

7 Support

7.1 Contact

Questions regarding CypScoreMOE should be addressed to:

CACHE Research LLC (info@CacheResearch.com)

7.2 CAChe Research LLC.

Americas

CAChe Research LLC

Oregon, USA

Email: Info@CACheResearch.com

Tel: +1 503 830 2772

Fax: +1 206 203 4405

Europe

CAChe Research LLC

Somerset, UK

Email: Info@CACheResearch.com

Tel: +44 2081 444080

Web: www.CACheResearch.com

8 References

- [1] M. Hennemann, A. Friedl, M. Lobell, J. Keldenich, A. Hillisch, T. Clark and A. H. Goller. *CypScore: quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory*. ChemMedChem, 2009, 4, 657-669.
- [2] MOE (Molecular Operating Environment), Chemical Computing Group Inc. (www.chemcomp.com), Montreal, Quebec, Canada. 2008.
- [3] A. H. C. Horn, J.-H. Lin and T. Clark. *A multipole electrostatic model for NDDO-based semiempirical molecular orbital methods*. Theor. Chem. Accts., 2005, 113, 159-168. Erratum: Theor. Chem. Accts., 2007, 117, 461-465.
- [4] T. Clark, A. Alex, B. Beck, F. Burkhardt, J. Chandrasekhar, P. Gedeck, A.H.C. Horn, M. Hutter, B. Martin, G. Rauhut, W. Sauer, T. Schindler and T. Steinke. *VAMP 11.0*. Erlangen 2008.

Available from Accelrys Inc., San Diego, USA. (www.accelrys.com/products/materials-studio/modules/VAMP.html)

- [5] J. J. P. Stewart. *MOPAC2000*, 1999. Fujitsu Ltd., Tokyo, Japan. MOPAC 6.0 was once available as: J. J. P. Stewart, QCPE # 455, Quantum Chemistry Program Exchange, Bloomsville, Indiana, USA, 1990.