Fortgeschrittenen Praktikum Bioinformatik

Implementation of a Generic Fragment Based Method for Automated Calculation of the Octanol-Water Partition Coefficient

Thomas Binsl
Universität des Saarlandes

supervised by
Dr. Andreas Kämper and Prof. Dr. Thomas Lengauer, Ph.D.
Computational Biology and Applied Algorithmics Group
MPI für Informatik
Saarbrücken

22nd November 2004

Contents

1 Introduction

Even today, in times of technological and medicinal progress, one of the most fatal scourges is the uncountable number of diseases humanity has to deal with. Although the treasure of drugs is already pretty large, there are still uncurable diseases like AIDS and SARS, for which drugs have to be discovered for. Another major problem is the rising resistance of bacterias against existing drugs. Therefore the development of drugs has to be improved and speeded up continuously.

One of the major limiting factors in speeding up this development is the enormous search space of molecules which can be possible drugs. This number has been estimated from 10^{60} to 10^{200} [1]. Since this search space is nearly infinite, many approaches have been developed to reduce it significantly. A successful method is filtering out substances with undesired properties. These filters should exclude those compounds, which are not drug-like or have a poor bioavailability even before they are used in a more time consuming method later in the virtual screening process [2].

Now the question arises, which molecular properties determine a molecules bioavailability. Many researchers have attempted to find rules describing the bioavailability of a molecule [3]. The most famous set of rules in this direction is safe to say Lipinski's Rule of 5 [4]. These rules imply that a compound shows poor bioavailability if two of the following rules are fulfilled: hydrogen bond donors ≥ 5, hydrogen bond acceptors ≥ 10, molecular weight ≥ 500, and $\log P \geq 5$.

Here, $\log P$ denotes the logarithm base ten of the octanol-water partition coefficient K_{OW}, which represents the ratio of a chemicals molar concentration in the 1-octanol phase C_{Oi} to its molar concentration in the aqueous phase C_{Wi} of an octanol-water system in equilibrium [5].

$$\log P = \log K_{OW} = \log \frac{C_{Oi}}{C_{Wi}} \tag{1}$$

It describes the lipophilicity of a compound and is an important indicator for bioavailability as a drug should not be too lipophilic because of its aqueous solubility behavior and not too lipophobic. Therefore $\log P$ is an important descriptor in several applications as e. g. solubility and dissolution prediction or QSAR (Quantitive Structure Activity Relationship) studies.

Since experimental methods for $\log P$ determination take a lot of time or are not accurate enough and thus, cannot be applied to a large dataset, computational methods have been developed. Today, a large number of different approaches for $\log P$ computation were elaborated [6], which fall into the group of QSPR (Quantitive Structure Property Relationship) methods. Every single member of this group presents a mathematical model that uses fragment contributions and structural descriptors to relate the property of interest to molecular composition and geometry [5, 15].

The first methodology introduced and widely accepted was developed by Hansch *et. al.* [7] and is based on substitution. Here, $\log P$ was considered as the sum of a components $\log P$ value substituted with a hydrogen atom and a π-term representing the contribution of the substituent itself.

$$\log P_{R-X} = \log P_{R-H} + \pi_X \tag{2}$$

An example for an alternative method was published in 1997 by Wang *et. al.* [8]. Their algorithm, also known as XLOGP rested upon summation of atomic contributions including correction factors e.g. for intermolecular interactions

$$\log P = \sum_i a_i A_i + \sum_j b_j B_j \tag{3}$$

where a_i = contribution of the ith atom
A_i = number of occurrences of the ith atom
b_j = contribution of the jth correction factor
B_j = number of occurrences of the jth correction factor

Other methods as e.g. published by Rogers and Cammarata use molecular properties [9] or solvatochromic parameters as first proposed by Kamlet *et. al.* [10, 11], for their calculation.
The method used in the present study was those introduced by Hansch and Leo in 1979 [5, 6, 12, 13]. It uses a summation over all fragment and structural feature contributions belonging to a query molecule, whereas fragments are atoms or atom groups with a final $\log_P$ contribution on the one hand and structural descriptors are additional contributions considering molecular flexibility, multiple halogenation, unsaturation, branching and H bond interactions.

$$\log P = \sum_i a_i f_i + \sum_j b_j F_j \tag{4}$$

where a_i = number of fragments i within the molecule
f_i = contribution of fragment i to $\log P$
b_j = number of structural features j within the molecule
F_j = contribution of structural feature j to $\log P$

It is a well established and accepted method for $\log P$ computation and reached best results in a test set on 19 drugs between other four computational methods as found out by Wang *et. al.* [8].
The first main goal of the *Forschungspraktikum* was the full implementation of the first partial sum given in equation 4. Therefore we set our focus on the fragments presented in [5] and fully implemented an algorithm computing all contributions maintained by this fragments. Additionally, we chose two structural descriptors for polyhalogenation also mentioned in [5] and implemented them. This lead us to the following full implemented equation

$$\log P = \sum_i a_i f_i + (b_1 F_1 + b_2 F_2) \tag{5}$$

where a_i = number of fragments i within the molecule
f_i = contribution of fragment i to $\log P$
b_1 = number of polyhalogenation at same carbon atoms within the molecule
F_1 = contribution of polyhalogenation at same carbon atoms to $\log P$
b_2 = number of polyhalogenation at adjacent carbon atoms within the molecule
F_2 = contribution of polyhalogenation at adjacent carbon atoms to $\log P$

which represents a subset of equation 4.
After the implementation we tested our algorithm on two different datasets of compounds. The first dataset was a subset of 8 compounds presented in [13], whereas the second dataset includes all 19 compounds used by Wang *et. al.* For matching the fragments we used the chemical subgraph matching code from the FlexX project [17, 18]. In order to extend its capabilities, the second main goal of the *Forschungspraktikum* was the implementation of a parser for SYBYL Line Notation (SLN), [16]. SLN is a language for storing compounds, structures, fragments and query patterns as a text string. We set our focus on the SLN features for describing fragments. With SLN we now provide between the already existing implementations for SMARTS [20] and chemical substructures a third possibility of compound representation.

2 Methodology

In the following we describe the overall strategy applied, the datasets used for testing and the two different implementations for the $\log P$ calculation and the SLN parser.

2.1 Strategy

We created two different datasets of query molecules and started to convert the rules given by Hansch and Leo [5] into SYBYL Line Notation [16]. Afterwords we implemented the SLN parser which parses the SLN text string into a subgraph-entry-pointer, a data structure used in FlexX and necessary for using the subgraph matcher already implemented in FlexX. Finally we implemented the main program which uses the parsed rules to compute the total $\log P$ of all query molecules.

2.2 Data

In the following we describe the creation of the two different datasets used for testing our implementation. Furthermore we explain the preparation of the fragments used and the additional structural descriptors used for polyhalogenation.

2.2.1 Datasets S1 and S2

We created a first dataset S1 consisting of 8 small non drug molecules taken out of the "Handbook of Chemical Property Estimation Methods" [13] to check if our algorithm matches all fragments presented by the query molecule. All molecules were drawn with "ChemDraw Ultra" [19] and saved as "MDL Molfile" for a later use.
The second dataset S2 was completely taken from the paper of Wang *et. al.* [8]. It consists of 19 molecules, all characterized to be drugs and was used to check the correlation of our results, computed without some structural descriptors and the results expected with a complete implementation. All molecules with exception of diphenhydramine were taken from the electronic version of the "Merck-Index", pasted into "ChemDraw Ultra" and saved as "MDL Molfile". Diphenhydramine which was not found within the "Merck-Index" was drawn

with "ChemDraw Ultra" by hand and also saved as "MDL Molfile".
Complete lists of the molecules used are summarized in Appendices A1 and A2.

2.2.2 Fragments

All fragments introduced by Hansch and Leo were taken from Baum [5] except the fragment of trifluormethyl (–CF_3) since its $\log P$ contribution was apparently wrong. Nevertheless, if a trifluormethyl group is present within the query molecule its contribution is calculated by the fragments of fluorine and carbon what seemed to be better than using the specified value.
Additionally, the following two fragments referred to in Figures 1 and 2 were taken from [13] since they seemed to be missing in [5] and were needed within our tests. Here R^1 indicates any rest and R^2 indicates any aromatic ring system. Furthermore, two additional fragments for carbon and one additional for

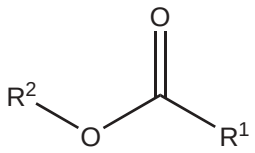


Figure 1: First additional fragment

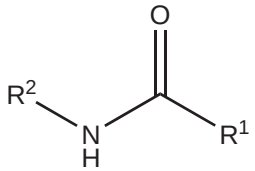


Figure 2: Second additional fragment

nitrogen were created since they were used in example calculations of [5, 13] and were helpful while calculating the $\log P$ values. They are given in Figures 3, 4 and 5 where R^1, R^2, R^3 and R^4 indicate any rest. After completing the

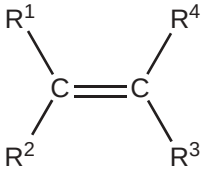


Figure 3: Third additional fragment

number of fragments all fragments were splitted off into three different levels. The need for this splitting off is described now in a simple example.
Imagine you have a molecule like R-$C(O)NH_2$ and three different fragments H, NH_2 and $C(O)NH_2$ used for the $\log P$ calculation. While doing the calculation

Figure 4: Fourth additional fragment

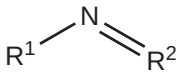


Figure 5: Fifth additional fragment

it might now be possible that the algorithm reaches one of the molecules hydrogens first until it reaches the nitrogen or carbon atom. Then the only fragment found for a possible matching is the H fragment and the others are not considered, although they are more specific and therefore have to be preferred. This is also the case if the nitrogen is reached first and the $C(O)NH_2$ fragment is not considered. Thus, we performed the following segmentation.

Level 1 Elemental fragments e.g. N and H

Level 2 Fragments that contain elemental fragments within their declaration and that are enclosed within other fragments e.g. NH_2 contains N and H and is enclosed within $C(O)NH_2$

Level 3 All remaining fragments e.g. $C(O)NH_2$

Afterwords each fragment of each level received a special priority in view of being more specific than another closely related fragment e.g. NH_2 received a higher priority than NH and NH between two aromatic rings received a higher priority than NH attached to one aromatic ring or attached to non aromatic carbons only. Additionally the fragments obtained an integer value indicating the number of atoms they consist of. Finally all fragments were translated into SYBYL Line Notation and embodied within the following notation:

```
@subgraph 0 "priority" "name"
sln "string"
data
"number of fragments atoms"
"log P contribution"
end
```

In the event of $C(O)NH_2$ this looks like the following:

```
@subgraph 0 1 C(O)NH2
sln C(=O)(-N(-H)(-H))(-Any)
data
5
-2.18
end
```

2.2.3 Special rules for polyhalogenation

In the next step we created some rules for polyhalogenation states. Polyhalogenation is a structural descriptor and could easily be covered by creating additional fragments. Within the rules given by Hansch and Leo this feature is subdivided into polyhalogenation at the same and at adjacent carbon atoms. Therefore two datasets were created as both polyhalogenation states are regarded independently. The first dataset contains all possible polyhalogenation states at one carbon atom whereas the second contains all possibilities of polyhalogenation at adjacent carbon atoms. The following example corresponding to Figure 6 explains the calculation. The first two fragments found represent

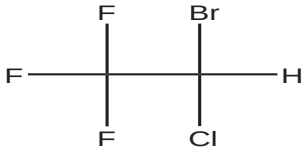


Figure 6: 2-Bromo-2-chloro-1,1,1-trifluoro-ethane

the polyhalogenation at the same carbon atom referred to in Figures 7 and 8, whereas the third represents polyhalogenation at adjacent carbon atoms referred to in Figure 9. And actually all of them have to be considered.

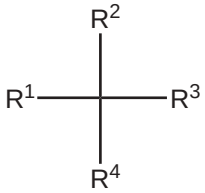


Figure 7: Polyhalogenetion at the same carbon atom where R^1, R^2 and R^3 are any halogens and R^4 is any atom except a halogen.

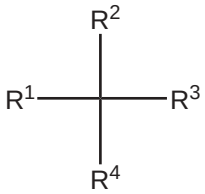


Figure 8: Polyhalogenetion at the same carbon atom where R^1 and R^3 are any halogens and R^2 and R^4 are any atoms except a halogen.

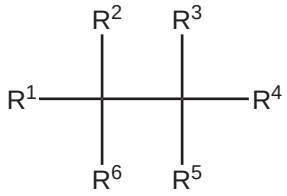


Figure 9: Polyhalogenetion at adjacent carbon atoms where R^1, R^2, R^3, R^5 and R^6 are any halogens and R^4 is any atom except a halogen.

2.2.4 Final databases

In the end we built five databases (DB1-DB5) for fragmental and structural datasets which are referred to in Appendix A3.

DB 1 Fragments indicating polyhalogenation at the same carbon atom

DB 2 Fragments indicating polyhalogenation at adjacent carbon atoms

DB 3 Fragments from level 3

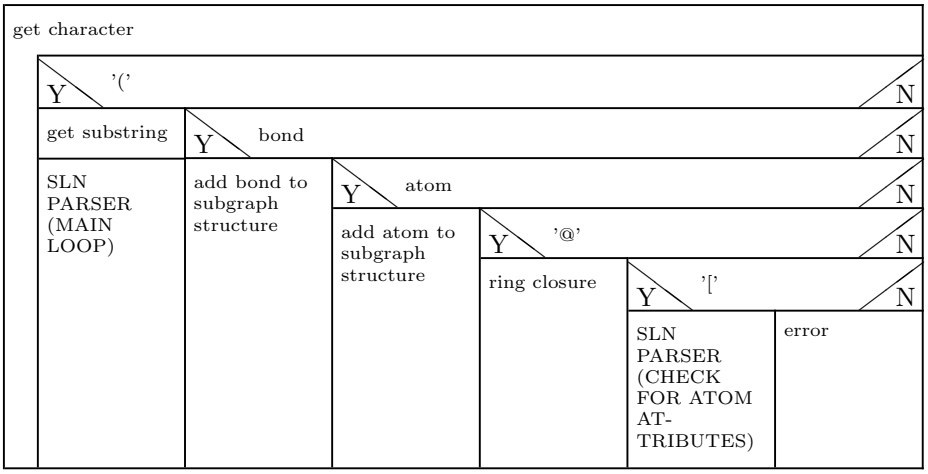
DB 4 Fragments from level 2

DB 5 Fragments from level 1

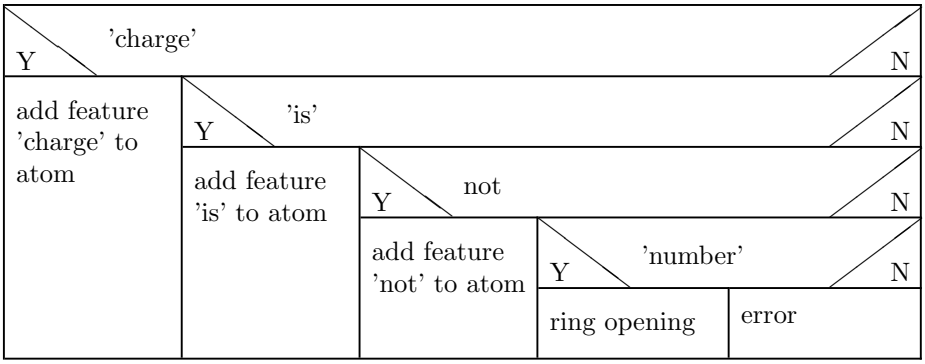
2.3 Implementation: SLN Parser

In cooperation with the *Fortgeschrittenen Praktikum Bioinformatik* of Andrea Volkamer "Implementation of a Method to Filter out Compounds with Toxic, Unsuitable or Reactive Functional Groups", we implemented a parser that parses a given SLN string into a subgraph-entry-pointer data structure. Every SLN string is processed from left to right until the strings right end is reached. During the parsing procedure all declared atoms, atom attributes, bonds etc. are recognized and saved within the subgraph structure. Two flowcharts of the parsing algorithm are shown below whereas the first contains the main loop looking at branches ('('), bonds ('-','=','#',':'), atoms, ring closures ('@') and atom attributes. The second flow chart presents the parsing of the different atom attributes. Additional an example for an SLN string and its dedicated subgraph entry pointer is referred to in Appendix B1.

SLN PARSER — MAIN LOOP



SLN PARSER — CHECK FOR ATOM ATTRIBUTES

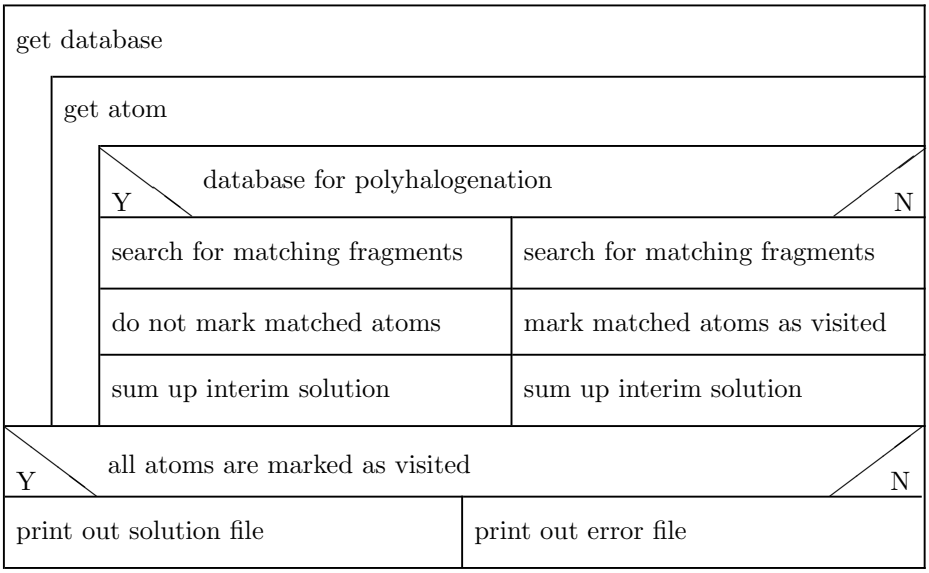


2.4 Implementation: Main Program

For automatic calculating $\log P$ values we finally implemented the main program. It receives a molecule from the user and computes the value incrementally. For each database it runs over all atoms of the molecule. We started with the two databases containing the fragments for polyhalogenation (DB1, DB2), whereas labeling for polyhalogenation was not necessary since it is a matter of structural descriptors.. For each fragment that matches within the molecule the $\log P$ contribution is summed up and saved for later use. In the next three steps the third, fourth and fifth databases (DB3, DB4, DB5) are passed through. For each matching fragment the $\log P$ contribution is added to the interim solution and the matching atoms within the molecule are labeled as visited. This had to be done since no atom has to be regarded twice. Finally the program maintained a value and starts to test if all atoms were visited. If this is the case a solution file is printed out containing the computed $\log P$, a list of all matched fragments and the contribution for polyhalogenation, if there is any. If any atom was not

visited, an error file is printed out which contains a list of all these atoms.
A flowchart of the algorithm is shown below. Further details concerning implementation and an example of a solution file on the one hand and an error file on the other hand are given in Appendices B2 and B3.

MAIN PROGRAM — LOGP CALCULATION



3 Results

In the following part we describe and analyse all results maintained with our implementation and compare them to other calculated and observed values.

3.1 Results within dataset S1

In the first step we tested all molecules of S1 and compared them to the reference values taken from [13]. This comparison was done for testing whether there is a great deviation within the values of small molecules induced by the missing structural descriptors and is referred to in Figure 10.
In a second step we compared our values and the reference values from Hansch and Leo with the observed experimental values also taken from [13] to get an impression of the correlation to measured $\log P$ values. Additionally we computed the $\log P$ values using the CLOGP4 program available at `www.daylight.com`, which is an improvement of Hansch and Leo's method using generic data, additional fragments and additional structural descriptors, and compared it to the observed experimental values, too. This can be looked up at Figures 11, 12 and 13. In addition to that a list of all calculated values is shown in Table 1.

Table 1: Calculated $\log P$ values of S1

Name	experimental	CLOGP4	Hansch & Leo	this work
Acetic-acid-methyl-ester	0.180	0.182	0.170	0.290
p-Xylene	3.150	3.140	3.450	3.460
1-Chloro-4-nitro-benzene	2.390	2.598	2.580	2.590
Benzoic-acid-methyl-ester	2.120	2.111	2.110	2.235
Acetic-acid-phenyl-ester	1.490	1.491	1.490	1.615
Diphenyl-amine	3.500	3.620	3.510	3.720
Dichloro-fluoro-methane	1.550	1.512	1.520	1.760
2-Bromo-2-chloro-1,1,1-trifluoro-ethane	2.300	2.272	2.460	3.060

Figure 10 shows that there is a large correlation of 0. 9718 between our calculated values and the reference values from Hansch and Leo, indicating that the influence of the structural descriptors is not that large for small molecules. Nevertheless a detailed analysis of the deviations is presented in the following.

Acetic-acid-methyl-ester: The missing structural descriptor of conformational flexibility within hydrocarbon chains leads to a deviation of $1 * (-0.12) = -0.12$.

p-Xylene: Within the reference value the C_6H_4 fragment is calculated by subtracting the value of one H from the fragment value of C_6H_5 which leads to a fragment value of 1.67. Our program calculates the C_6H_4 fragment by summing up four times the value of CH and two times the value of C what ends up in a value of 1.68.

1-Chloro-4-nitro-benzene: See p-Xylene.

Benzoic-acid-methyl-ester: The missing structural descriptor of conformational flexibility within hydrocarbon chains leads to a deviation of $1 * (-0.12) = -0.12$. In addition to that the C_6H_5 fragment value for calculating the reference value is a rounded value of 1.90 whereas our value is the exact one of 1.905.

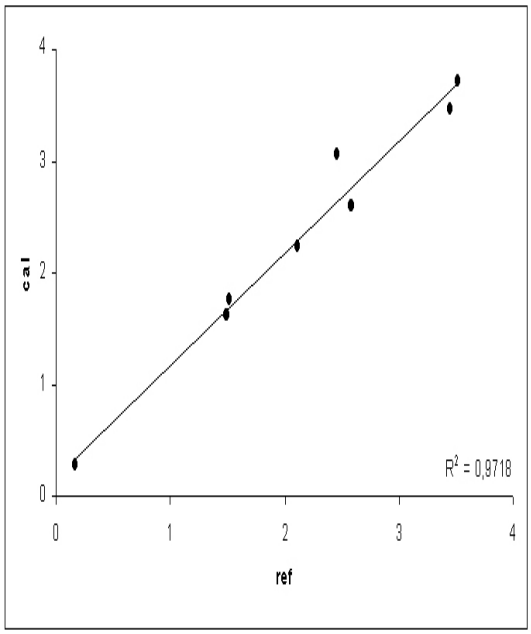


Figure 10: Correlation of reference values with the calculated values of S1

Acetic-acid-phenyl-ester: See Benzoic-acid-methyl-ester.

Diphenyl-amine: The calculation of the reference value includes two times the value of C_6H_5, see Benzoic-acid-methyl-ester. In addition to that a special value for branching including amine $1 * (-0.20) = -0.20$ is used which was not mentioned in [13].

Dichloro-fluoro-methane: The missing structural descriptor of conformational flexibility within hydrocarbon chains leads to a deviation of $2 * (-0.12) = -0.24$.

2-Bromo-2-chloro-1,1,1-trifluoro-ethane: The missing structural descriptor of conformational flexibility within hydrocarbon chains leads to a deviation of $6 * (-0.12) = -0.60$. Thus, 2-Bromo-2-chloro-1,1,1-trifluoro-ethane is the largest outlier since the missing structural contribution is larger then in all other examples.

The results presented in Figure 11, 12 and 13 show that the best method used for calculating the $\log P$ value is the CLOGP4 program receiving a correlation of 0.9944 with the measured values, followed by the reference values from Hansch and Leo with a correlation of 0.9909 and our values with a correlation of 0.9669. This is not a surprise thus the CLOGP4 method is an improvement of the method used for calculating the reference values and the missing structural descriptors in our implementation.

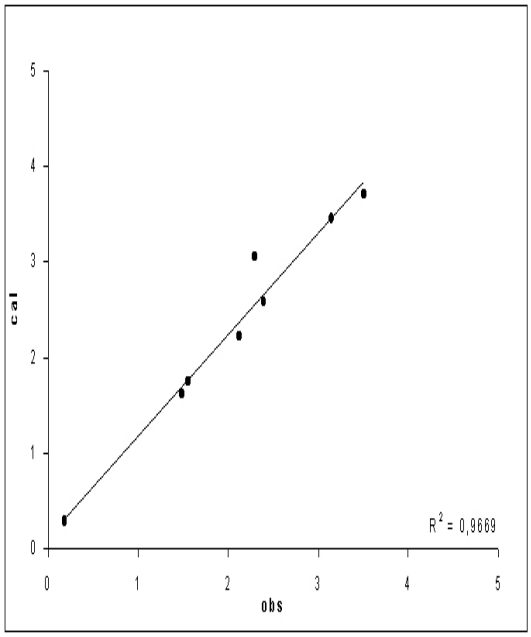


Figure 11: Correlation of observed values with the calculated values of S1

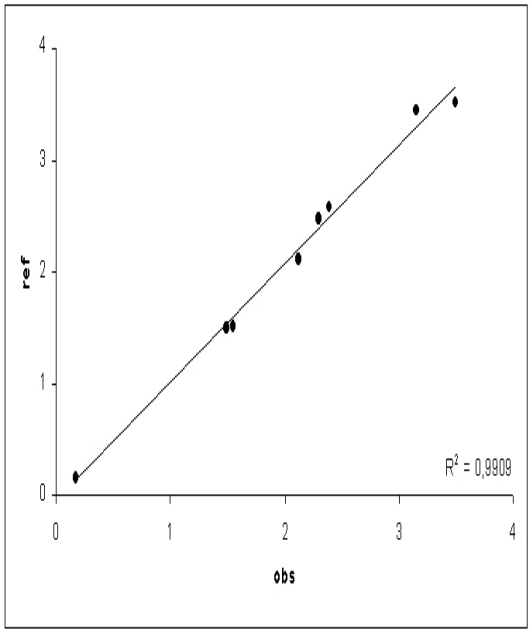


Figure 12: Correlation of observed values with the reference values of S1

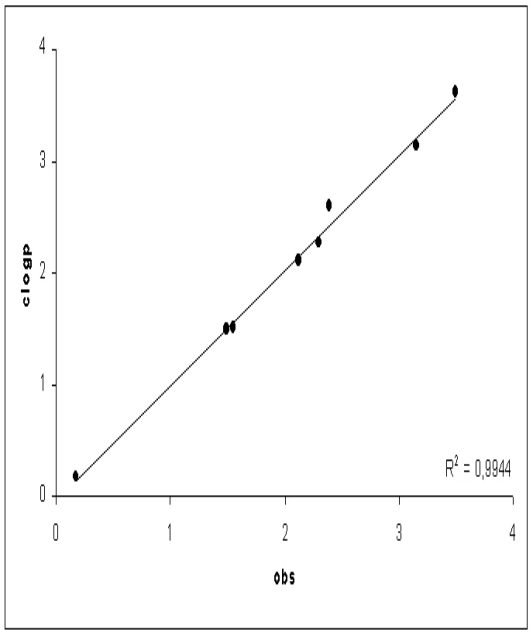


Figure 13: Correlation of observed values with the CLOGP4 values of S1

3.2 Results within dataset S2

In our second testing phase we considered all compounds belonging to dataset S2. We compared them to the reference values taken from [8] for checking if there is an increase of deviation between the values of larger molecules induced by the missing structural descriptors. The result can be looked up in Figure 14. In a second step we made additional comparisons between our values, the reference values from [8] and those values calculated by CLOGP4 with the observed experimental values, as we did in section 3.1. The results are referred to in Figures 15, 16 and 17 and an overall list of all calculated and measured values is available within Table 2.

Table 2: Calculated $\log P$ values of S1

Name	**experimental**	**CLOGP4**	**Hansch & Leo**	**this work**
Atropine	1.830	1.299	1.320	2.505
Chloramphenicol	1.140	1.283	0.690	−1.130
Chlorothiazide	−0.240	−0.294	−1.240	−1.670
Chlorpromazine	5.190	5.300	5.200	5.925
Cimetidine	0.400	−0.044	0.210	−8.080

continued on next page

Table 2 – concluded from previous page

Name	experimental	CLOGP4	Hansch & Leo	this work
Diazepam	2.990	2.961	3.320	1.410
Diltiazem	2.700	3.647	3.550	2.330
Diphenhydramine	3.270	3.452	2.930	3.340
Flufenamic acid	5.250	5.526	5.580	3.720
Haloperidol	4.300	3.977	3.520	4.580
Imipramine	4.800	5.037	4.410	5.760
Lidocaine	2.260	1.954	1.360	3.305
Phenobarbital	1.470	1.365	1.370	−2.585
Phenytoin	2.470	2.085	2.090	−0.070
Procainamide	0.880	1.423	1.110	1.110
Propanolol	2.980	2.753	2.750	2.625
Tetracaine	3.730	3.834	3.650	3.880
Trimethoprim	0.910	0.981	0.660	−0.765
Verapamil	3.790	4.466	3.530	7.180

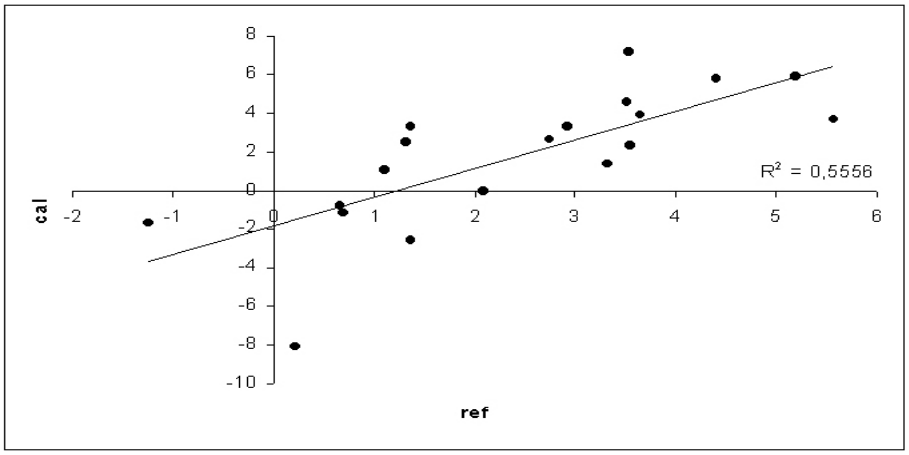


Figure 14: Correlation of reference values with the calculated values of S2

Figure 14 shows that there is a low correlation of 0.5556 between our calculated values and the reference values taken from [8], indicating that the influence of the structural descriptors increases and have to take into account within larger molecules as the fragments were all considered. The results presented in Figures 15, 16 and 17 show again that the best method used for calculating the $\log P$ value is the CLOGP4 program being an improvement of Hansch and Leo's method and receiving a correlation of 0.9536 with the measured values. This is followed by the reference values calculated with Hansch and Leo's method which received a correlation of 0.9407. The correlation of 0.6500 with our values show that within larger compounds the presence of structural descriptors is a must since with growing compounds structural features become much more important.

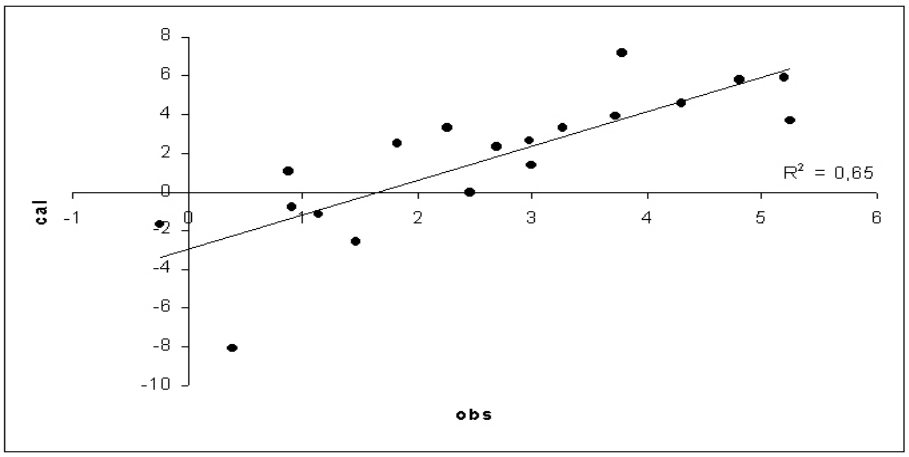


Figure 15: Correlation of observed values with the calculated values of S2

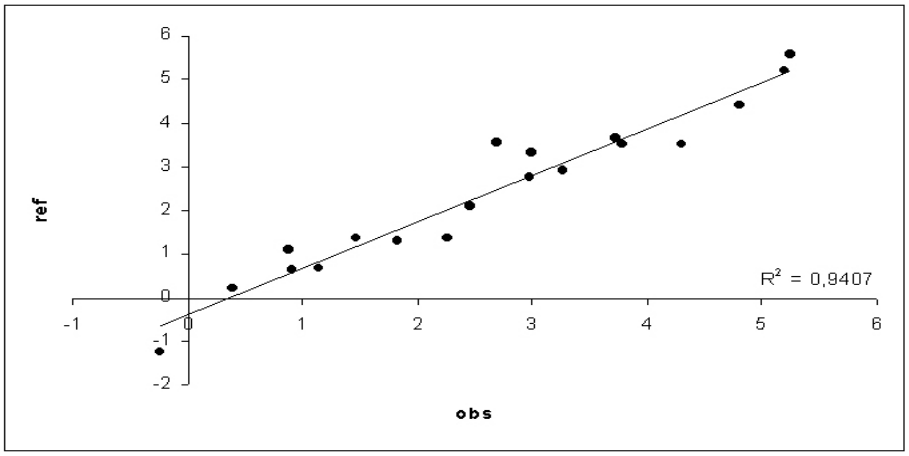


Figure 16: Correlation of observed values with the reference values of S2

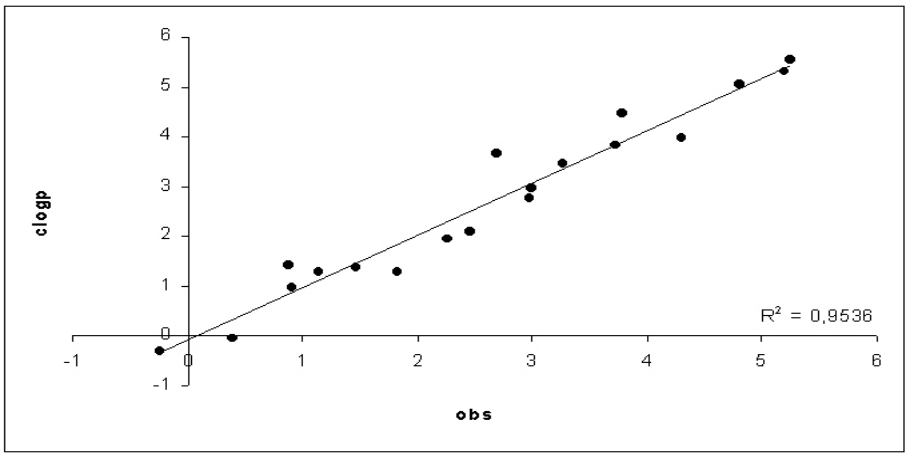


Figure 17: Correlation of observed values with the CLOGP4 values of S2

4 Conclusions

We implemented a method for calculating $\log P$ values from fragments. Additionally, we implemented a parser for SLN. In the presented results according to dataset S1 we showed that the usage of our fragments as well as the additionally implemented structural descriptors for polyhalogenation work correctly. The solutions we maintained had a total correlation of 0.9718 with the reference values taken from [13] whereas the deviation to a full correlation depends on the missing structural descriptors, rounding errors and different calculation approaches.

Since our implementation is not complete because of some missing structural descriptors we had to test our implementation on a second data set. In the light of the results presented according to dataset S2 we saw that the missing of the remaining structural descriptors not implemented yet effects the results in a way that a reliable $\log P$ estimation is not advisable for larger compounds. This is shown by the low correlation of 0.5556 between our values and the reference values taken from [8]. Therefore it would be a necessary improvement if the remaining structural descriptors, which are referred to in Appendix C, are going to be implemented as well.

Finally a further extension would be the embedding of the stand alone program within the ECLIPPSE (Environment for Chemical Library Inspection Pre- and Postprocessing, and Screening Evaluation) program developed by Andreas Kämper. This is a graphical tool containing all steps in ligand preprocessing e.g. structure generation, duplicate elimination and filtering on the one hand and all steps of ligand postprocessing e.g. evaluation and visualization on the other hand [14], whereas the here presented $\log P$ calculation method could be used as a filter criterion.

The program developed works in generic way and allows valid SLN as input. This makes the program extendable for calculating other fragment additive properties, like molar refractivity or polar surface area.

References

[1] R. S. Bohacek, C. Martin, W. C. Guida; The art and practice of structure-based drug design: a molecular modeling persepctive; *Med. Res. Rev.*, **1996**, *3*, 16.

[2] W. P. Walters, M. T. Stahl, M. A. Murcko; Virtual screening - an overview; *Drug. Discov. Technol.*, **1998**, *3*, 160-178.

[3] W. P. Walters, M. A. Murcko; Prediction of 'drug-likeness'; *Adv. Drug Deliv. Rev.*, **2002**, *54*, 255-271.

[4] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney; Experimental and computational approaches to estimate solubility and permeability in drug discovery; *Adv. Drug Deliv. Rev.*, **1997**, *23*, 3-25.

[5] E. J. Baum; Chemical Property Estimation, Lewis Publishers (CRC Press), Boca Raton; **1997**, 135-164.

[6] A. J. Leo; Calculating log P_{oct} from Structures; *Chem. Rev.*, **1993**, *93(4)*, 1281-1306.

[7] T. Fujita, J. Iwasa, C. J. Hansch; A new substituent constant, π, derived from partition coefficients; *Am. Chem. Soc.*, **1964**, *86*, 5175.

[8] R. Wang, Y. Fu, L. Lai; A New Atom-Additive Method for Calculating Partition Coefficients; *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 615-621.

[9] K. S. Rogers, A. Cammarata; A molecular orbital description of the partitioning of aromatic compounds between polar and nonpolar phases; *Biochem. Biophys. Acta*, **1969**, *193*, 22.

[10] M. Kamlet, J. L. Abboud, M. Abraham, R. Taft; Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π^*, α and β, and some methods for simplifying the generalized solvatochromic equation; *J. Org. Chem.*, **1993**, *48*, 2877.

[11] M. Kamlet, J. L. Abboud, R. Taft; The solvatochromic comparison method. 6. The π^* scale of solvent polarities; *J. Am. Chem. Soc.*, **1977**, *99*, 6027.

[12] C. Hansch, A. Leo; Substituent Constants for Correlation Analysis in Chemistry and Biology, John Whiley & Sons, New York; **1979**, 18-43.

[13] W. J. Lyman; Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington D. C. ; **1990**, chapter 1, 1-54.

[14] A. Kämper; Computer Aided Drug Design; lecture 8, Universität des Saarlandes, **2004**.

[15] A. Kämper; Computer Aided Drug Design; lecture 10, Universität des Saarlandes, **2004**.

[16] Tripos; SLN; **2002**.

[17] M. Rarey, B. Kramer, T. Lengauer, G. Klebe; A fast flexible docking method using an incremental construction algorithm; *J. Mol. Biol.*, **1996**, *261*, 470-489.

[18] M. Rarey, B. Kramer, T. Lengauer; Multiple automatic base selection: Protein-ligand docking based on incremental construction without manual intervention; *J. Comput. -Aided Mol. Des.*, **1997**, *11*, 369-384.

[19] ChemDraw Ultra 7, CambridgeSoft, Cambridge.

[20] SMARTS, Daylight Chemical Information Systems Inc., Los Altos.

A Datasets

A.1 First dataset of query molecules S1

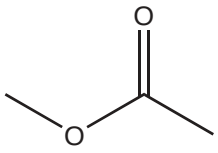


Figure 18: Acetic acid methyl ester

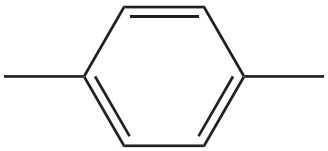


Figure 19: p-Xylene

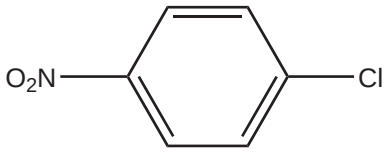


Figure 20: 1-Chloro-4-nitro-benzene

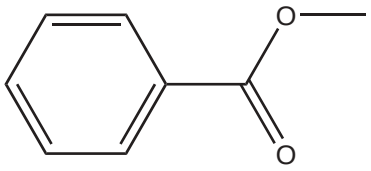


Figure 21: Benzoic acid methyl ester

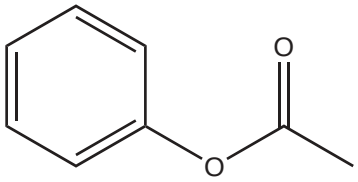


Figure 22: Acetic acid phenyl ester

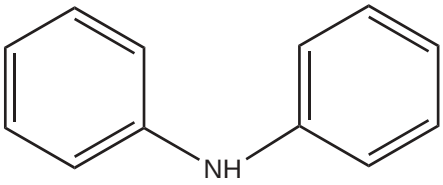


Figure 23: Diphenyl-amine

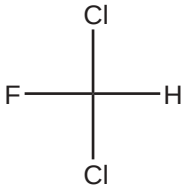


Figure 24: Dichloro-fluoro-methane

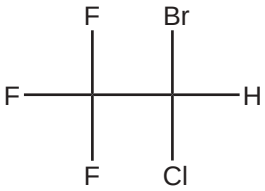


Figure 25: 2-Bromo-2-chloro-1,1,1-trifluoro-ethane

A.2 Second dataset of query molecules S2

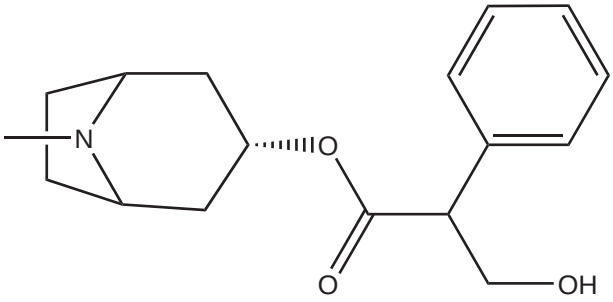


Figure 26: Atropine

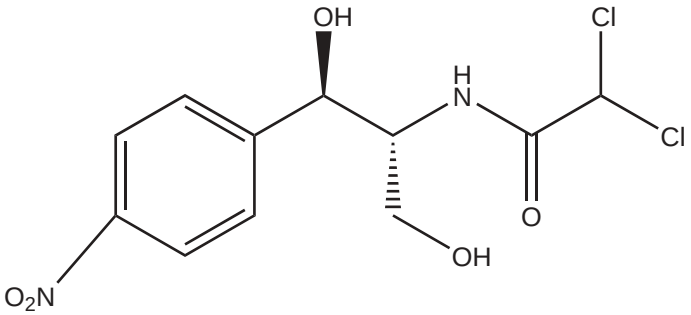


Figure 27: Chloramphenicol

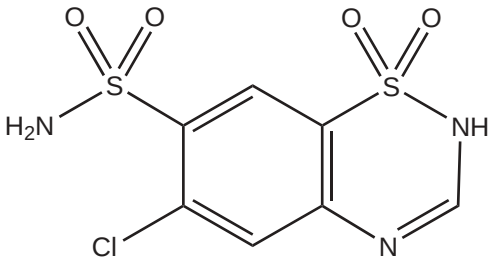


Figure 28: Chlorothiazide

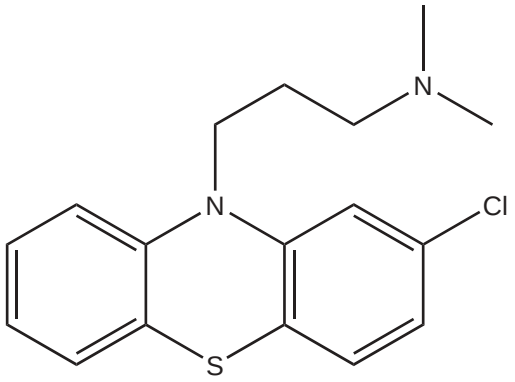


Figure 29: Chlorpromazine

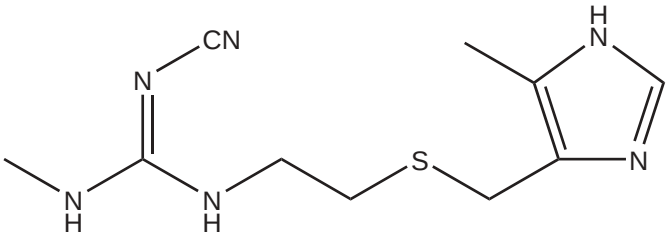


Figure 30: Cimetidine

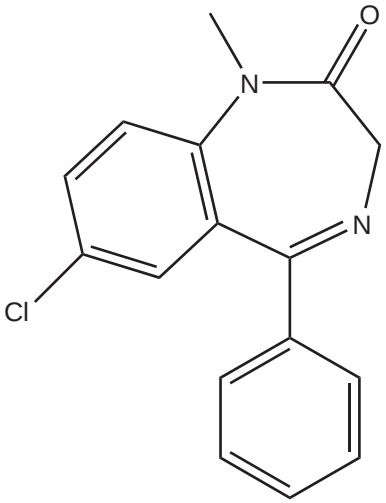


Figure 31: Diazepam

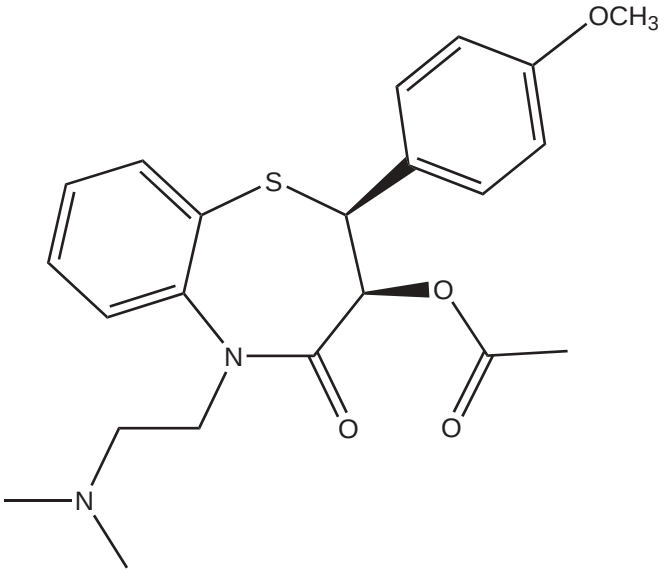


Figure 32: Diltiazem

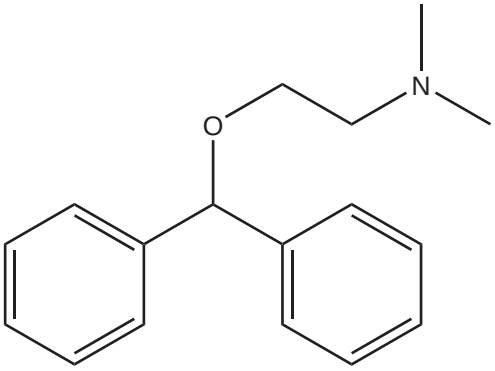


Figure 33: Diphenhydramine

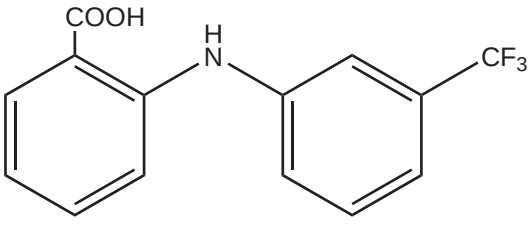


Figure 34: Flufenamic Acid

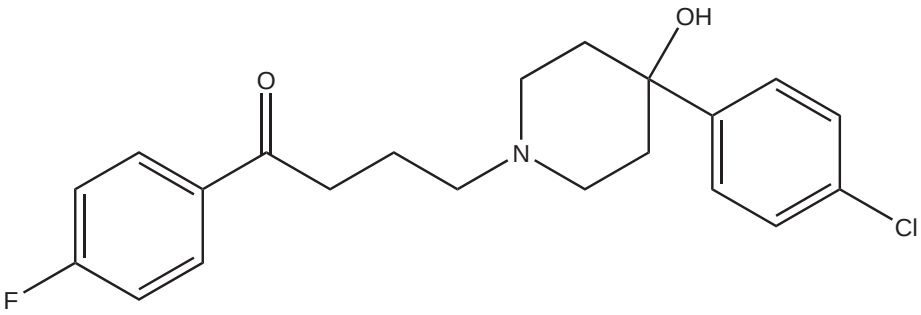


Figure 35: Haloperidol

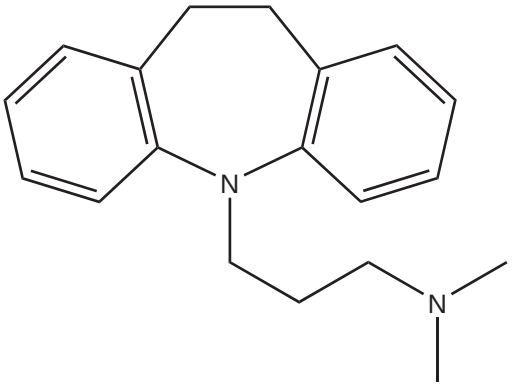


Figure 36: Imipramine

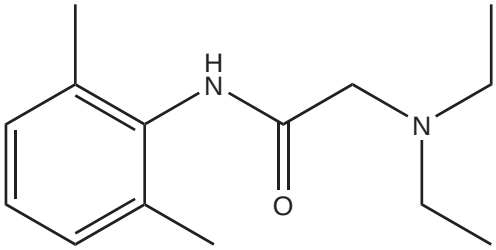


Figure 37: Lidocaine

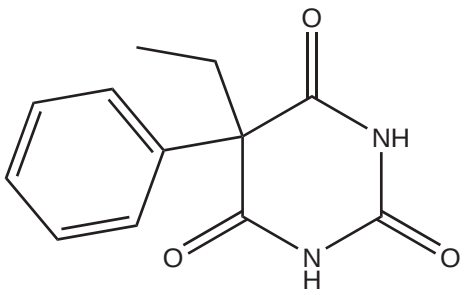


Figure 38: Phenobarbital

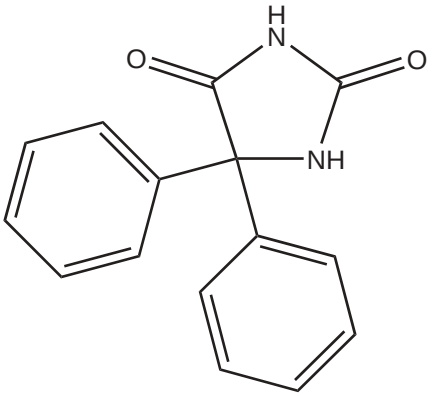


Figure 39: Phenytoin

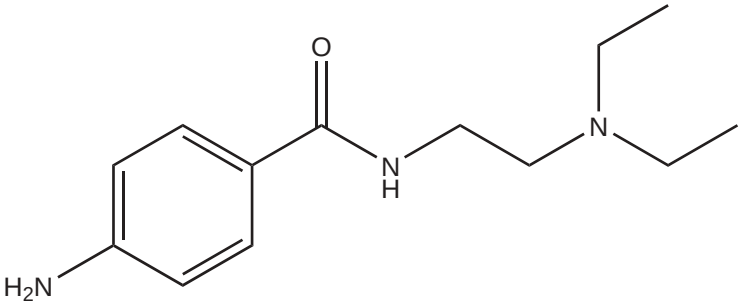


Figure 40: Procainamide

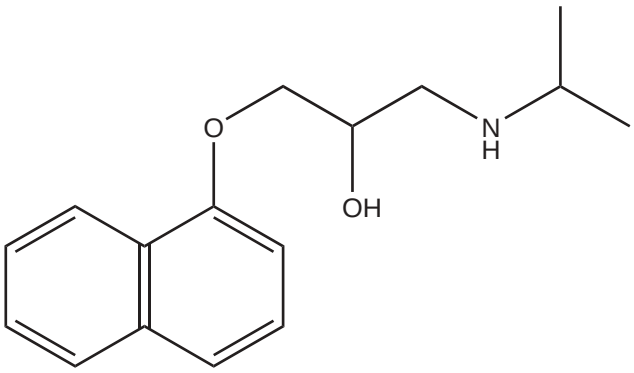


Figure 41: Propranolol

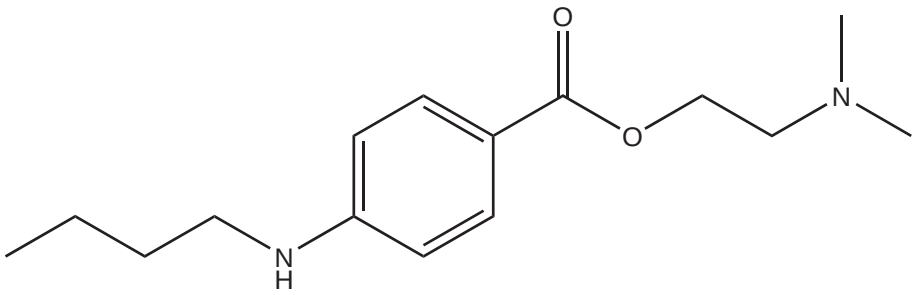


Figure 42: Tetracaine

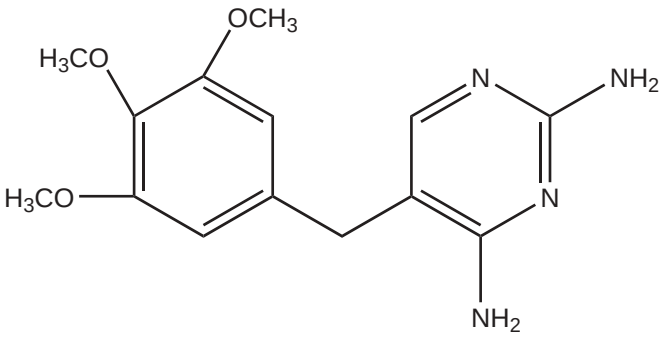


Figure 43: Trimethoprim

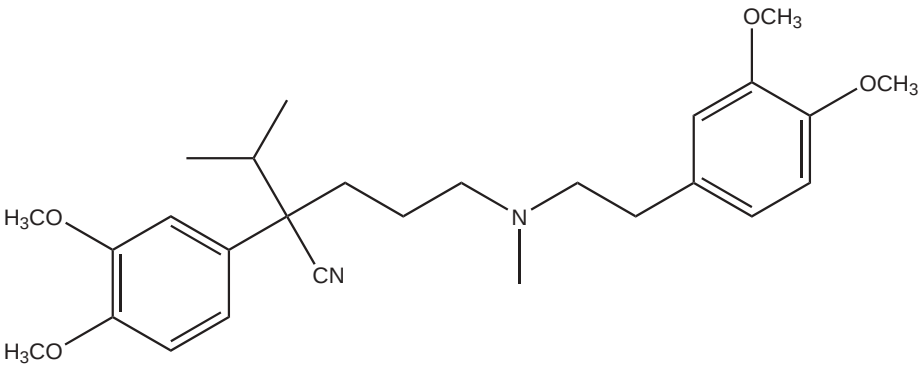


Figure 44: Verapamil

A.3 Fragments

Table 3: Fragments in level 1

Fragment	Priority	# atoms	LogP
H(-Any)	0	1	0.23
H(-Any(:Any)(:Any))	1	1	0.23
C(-Any)(-Any)(-Any)(-Any)	0	1	0.20
C(-Any)(-Any)(-Any)((-Any(:Any)(:Any))	1	1	0.20
O(-Any)(-Any)	0	1	−1.82
O(-Any)(-Any(:Any)(:Any))	1	1	−0.61
O(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	1	0.53
O(-C(=C(-Any)(-Any))(-Any))(-Any)	3	1	−1.21
N(-Any)(-Any)(-Any)	0	1	−2.18
N(=Any)(-Any)	0	1	−2.18
N(-Any(:Any)(:Any))(-Any)(-Any)	1	1	−0.93
N(-Any(:Any)(:Any))(=Any)	1	1	−0.93
N(-Any(:Any)(:Any))(-Any(:Any)(:Any))(-Any)	2	1	−0.50
S(-Any)(-Any)	0	1	−0.79
S(-Any(:Any)(:Any))(-Any)	1	1	−0.03
S(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	1	0.77
F(-Any)	0	1	−0.38
F(-Any(:Any)(:Any))	1	1	0.37
F(-C(=C(-Any)(-Any))(-Any))	3	1	0.00
Cl(-Any)	0	1	0.06
Cl(-Any(:Any)(:Any))	1	1	0.94
Cl(-C(=C(-Any)(-Any))(-Any))	4	1	0.50
Br(-Any)	0	1	0.20
Br(-Any(:Any)(:Any))	1	1	1.09
Br(-C(=C(-Any)(-Any))(-Any))	3	1	0.64
Br(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	1	0.48
I(-Any)	0	1	0.59
I(-Any(:Any)(:Any))	1	1	1.35

continued on next page

Table 3 – concluded from previous page

Fragment	Priority	# atoms	LogP
I(-C(=C(-Any)(-Any))(-Any))	3	1	0.97
Se(-Any)(-Any)	0	1	0.45
Si(-Any)(-Any)(-Any)(-Any)	0	1	−0.09
Si(-Any(:Any)(:Any))(-Any)(-Any)(-Any)	1	1	0.65
Si(-Any)(-Any)(-Any)(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	3	1	−0.38

Table 4: Fragments in level 2

Fragment	Priority	# atoms	LogP
O(-H)(-Any)	0	2	−1.64
O(-H)(-Any(:Any)(:Any))	1	2	−0.44
O(-H)(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	3	2	−1.34
N(-Any)(-Any)(-Any(:Any)(:Any))	0	1	−0.56
N(-Any)(-Any(:Any)(:Any))(-Any(:Any)(:Any))	1	1	−0.50
N(-H)(-Any)(-Any)	1	2	−2.15
N(-H)(-Any(:Any)(:Any))(-Any)	2	2	−1.03
N(-H)(-Any(:Any)(:Any))(-Any(:Any)(:Any))	3	2	−0.09
N(-H)(-H)(-Any)	4	3	−1.54
N(-H)(-H)(-Any(:Any)(:Any))	5	3	−1.00
N(-H)(-H)(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	3	−1.35
C(=O)(-Any)(-Any)	0	2	−1.90
C(=O)(-Any)(-Any(:Any)(:Any))	1	2	−1.09
C(=O)(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	2	−0.50
C(=O)(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))(-Any)	3	2	−1.77
S(-H)(-Any)	0	2	−0.23
S(-H)(-Any(:Any)(:Any))	1	2	0.62
S(-C(#N))(-Any)	1	3	−0.48
S(-C(#N))(-Any(:Any)(:Any))	2	3	0.64
S(-C(#N))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	3	−0.45

continued on next page

Table 4 – concluded from previous page

Fragment	Priority	# atoms	LogP
S(=O)(-Any)(-Any)	0	2	−3.01
S(=O)(-Any(:Any)(:Any))(-Any)	1	2	−2.12
S(=O)(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	2	−1.62
S(=O)(=O)(-Any)(-Any)	0	3	−2.67
S(=O)(=O)(-Any(:Any)(:Any))(-Any)	1	3	−2.17
S(=O)(=O)(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	3	−1.28
C(=Any)(-Any)(-Any)	0	1	0.20
C(=Any)(=C)	1	1	0.20
C(#Any)(-Any)	0	1	0.20

Table 5: Fragments in level 3

Fragment	Priority	# atoms	LogP
C(-H)(-H)(-H)(-Any)	0	4	0.89
C(-H)(-H)(-H)(-Any(:Any)(:Any))	1	4	0.89
C(=O)(-O(-Any))(-Any)	0	3	−1.49
C(=O)(-O(-Any))(-Any(:Any)(:Any))	1	3	−0.56
C(=O)(-O(-Any(:Any)(:Any)))(-Any)	1	3	−1.18
C(=O)(-O(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	3	−0.09
C(=O)(-O(-Any))(-C(=C(-Any)(-Any))(-Any))	3	3	−1.18
C(=O)(-O(-Any))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	3	−1.38
C(=O)(-O(-H))(-Any)	1	4	−1.11
C(=O)(-O(-H))(-Any(:Any)(:Any))	2	4	−0.03
C(=O)(-O(-H))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	4	−1.03
O(-C(-H)(=O))(-Any)	0	4	−1.14
O(-C(-H)(=O))(-Any(:Any)(:Any))	1	4	−0.64
O(-C(=O)(-N(-H)(-Any)))(-Any)	0	5	−1.79
O(-C(=O)(-N(-H)(-Any)))(-Any(:Any)(:Any))	1	5	−1.45
O(-C(=O)(-N(-H)(-Any)))(-C(=C(-Any)(-Any))(-Any))	2	5	−0.91

continued on next page

Table 5 – continued from previous page

Fragment	Priority	# atoms	LogP
O(-C(=O)(-N(-H)(-H)))(-Any)	1	6	−1.58
O(-C(=O)(-N(-H)(-H)))(-Any(:Any)(:Any))	2	6	−0.82
O(-C(=O)(-N(-H)(-H)))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	6	−1.24
O(-N[charge=+1](-O[charge=-1])(=O))(-Any)	0	4	−0.36
O(-N(=O)(=O)(-Any)	0	4	−0.36
C(-H)(=N(-O(-H)))(-Any)	0	5	−1.02
C(-H)(=N(-O(-H)))(-Any(:Any)(:Any))	1	5	−0.15
S(=O)(=O)(-O(-Any))(-Any)	0	4	−2.11
S(=O)(=O)(-O(-Any))(-Any(:Any)(:Any))	1	4	−2.06
S(=O)(=O)(-O(-Any(:Any)(:Any)))(-Any)	1	4	−1.42
S(=O)(=O)(-O(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	2	4	−0.62
C(=S)(-O(-Any))(-Any)	0	3	−1.11
O(-P(=O)(-O[1])(-O[2])(-Any(:Any)(:Any))).(Any-@1).(Any-@2)	0	5	−2.33
O(-P(=O)(-O[1])(-O[2]))(-Any).(Any-@1).(Any-@2)	0	5	−2.29
O(-P(=O)(-O[1])(-O[2]))(-Any(:Any)(:Any)).(Any-@1).(Any-@2)	1	5	−1.71
S(-P(=S)(-O[1])(-O[2]))(-Any).(Any-@1).(Any-@2)	0	5	−2.89
S(-P(=O)(-O[1])(-N(-H)(-Any)))(-Any).(Any-@1)	0	6	−2.18
S(-P(=O)(-N(-H)(-H))(-O(-Any)))(-Any)	0	7	−2.50
As(-O(-H))(-O(-H))(-O(-Any))(-Any(:Any)(:Any))	0	6	−1.84
As(=O)(-O(-H))(-O(-H))(-Any(:Any)(:Any))	0	6	−1.90
B(-O(-H))(-O(-H))(-Any(:Any)(:Any))	0	5	−0.32
N(-H)(-O(-H))(-Any(:Any)(:Any))	1	4	−1.11

continued on next page

Table 5 – continued from previous page

Fragment	Priority	# atoms	LogP
N(-H)(-N(-H)(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	2	4	−0.74
N(-H)(-N(-H)(-Any))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	3	4	−2.84
N(-H)(-N(-H)(-H))(-Any(:Any)(:Any))	1	5	−0.65
N(=N(-N(-Any)(-Any)))(-Any(:Any)(:Any))	1	3	−0.85
N(-N(=O))(-Any)(-Any)	0	3	−2.40
N(-N(=O))(-Any(:Any)(:Any))(-Any)	1	3	−0.84
N(-H)(-C(#N))(-Any(:Any)(:Any))	1	4	−0.03
C(-H)(=N(-N(-Any)(-Any)))(-Any(:Any)(:Any))	1	4	−1.71
S(=O)(=O)(-N(-Any)(-Any))(-Any(:Any)(:Any))	1	4	−2.09
S(=O)(=O)(-N(-H)(-Any))(-Any(:Any)(:Any))	2	5	−1.75
S(=O)(=O)(-N(-H)(-Any(:Any)(:Any)))(-Any)	2	5	−1.72
S(=O)(=O)(-N(-H)(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	3	5	−1.10
S(=O)(=O)(-N(-H)(-H))(-Any(:Any)(:Any))	3	6	−1.59
S(=O)(=O)(-N(-H)(-N(-H)(-H)))(-Any)	3	8	−2.04
N(-H)(-S(=O)(=O)(-N(-H)(-H)))(-Any(:Any)(:Any))	1	8	−1.50
C(=S)(-N(-H)(-Any))(-Any)	0	4	−2.00
C(=S)(-N(-H)(-Any(:Any)(:Any)))	3	4	−0.96
C(=S)(-N(-H)(-H))(-Any(:Any)(:Any))	1	5	−0.41
N(-H)(-C(=S)(-N(-H)(-Any)))(-Any(:Any)(:Any))	1	6	−1.79
N(-H)(-C(=S)(-N(-H)(-H)))(-Any)	0	7	−1.29
N(-H)(-C(=S)(-N(-H)(-H)))(-Any(:Any)(:Any))	2	7	−1.17
N(-P(=S)(-N[1])(-N[2]))(-Any)(-Any).(Any-@1).(Any-@1).(Any-@2).(Any-@2)	0	5	−3.37

continued on next page

Table 5 – continued from previous page

Fragment	Priority	# atoms	LogP
C(=O)(-H)(-Any)	0	3	−1.10
C(=O)(-H)(-Any(:Any)(:Any))	1	3	−0.42
C(=O)(-N(-H)(-Any))(-Any)	1	4	−2.17
C(=O)(-N(-H)(-Any))(-Any(:Any)(:Any))	2	4	−1.81
C(=O)(-N(-H)(-Any(:Any)(:Any)))(-Any) 2	4	-1.51	
C(=O)(-N(-H)(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	3	4	−1.06
C(=O)(-N(-H)(-Any))(-C(=C(-Any)(-Any))(-Any))	4	4	−1.51
C(=O)(-N(-H)(-H))(-Any)	1	5	−2.18
C(=O)(-N(-H)(-H))(-Any(:Any)(:Any))	2	5	−1.26
C(=O)(-N(-H)(-H))(-C(-H)(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	4	5	−1.99
C(=O)(-C(=O)(-Any))(-Any)	0	4	−3.00
C(=O)(-C(=O)(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	2	4	−0.30
N(-H)(-C(=O)(-N(-H)(-Any)))(-Any)	0	6	−2.18
N(-H)(-C(=O)(-N(-H)(-Any)))(-Any(:Any)(:Any))	1	6	−1.57
N(-H)(-C(=O)(-N(-H)(-Any(:Any)(:Any))))(-Any(:Any)(:Any))	2	6	−0.82
N(-H)(-C(=O)(-N(-H)(-H)))(-Any)	1	7	−2.18
N(-H)(-C(=O)(-N(-H)(-H)))(-Any(:Any)(:Any))	2	7	−1.07
N(-C(=O)(-N(-H)(-H)))(-Any(:Any)(:Any))(-Any)	2	6	−2.25
N(-C(=O)(-N(-H)(-H)))(-Any(:Any)(:Any))(-Any(:Any)(:Any))	3	6	−2.15
N(-C(-H)(=O))(-Any)(-Any)	0	4	−2.67
N(-C(-H)(=O))(-Any(:Any)(:Any))(-Any)	1	4	−1.59
N(-H)(-C(=O)(-N(-Any)(-Any)))(-Any(:Any)(:Any))	1	5	−2.29

continued on next page

Table 5 – continued from previous page

Fragment	Priority	# atoms	LogP
N(-H)(-C(=O)(-N(-Any(:Any)(:Any))(-Any)))(-Any)	3	5	−2.42
N(-H)(-C(-H)(=O))(-Any(:Any)(:Any))	2	5	−0.64
S(=O)(=O)(-F)(-Any(:Any)(:Any))	0	4	0.30
S(-F)(-F)(-F)(-F)(-F)(-Any(:Any)(:Any))	1	6	1.45
N(=C(-Cl)(-Cl))(-Any(:Any)(:Any))	1	4	0.64
O[charge=-1](-Any(:Any)(:Any))	0	1	−3.64
C(=O)(-O[charge=-1])(-Any)	0	3	−5.19
C(=O)(-O[charge=-1])(-Any(:Any)(:Any))	1	3	−4.13
S(=O)(=O)(-O[charge=-1])(-Any)	0	4	−5.87
S(=O)(=O)(-O[charge=-1])(-Any(:Any)(:Any))	1	4	−4.53
N[charge=+1](-O[charge=-1])(=O)(-Any)	0	3	−1.16
N(=O)(=O)(-Any)	0	3	−1.16
N[charge=+1](-O[charge=-1])(=O)(-Any(:Any)(:Any))	1	3	−0.03
N(=O)(=O)(-Any(:Any)(:Any))	1	3	−0.03
O(-S(=O)(=O)(-O[charge=-1]))(-Any)	0	5	−5.23
C(-Any)(:Any)(:Any)	0	1	0.13
C(:C)(:C)(:C)	0	1	0.225
C(:Any[!C])(:C)(:C)	0	1	0.44
C(:Any[!C])(:Any[!C])(:C)	0	1	0.44
C(:Any[!C])(:Any[!C])(:Any[!C])	0	1	0.44
C(-H)(:Any)(:Any)	1	2	0.355
O(:Any)(:Any)	0	1	−0.08
C(=O)(:Any)(:Any)	0	2	−0.59
O(-C(=O)(:Any))(:Any)	0	3	−1.40
O(:C(=O)(:Any))(:Any)	0	3	−1.40
N(-Any)(:Any)(:Any)	0	1	−1.10
N(-H)(:Any)(:Any)	1	2	−0.65
N(:Any)(:Any)	0	1	−1.12
N(=N(:Any))(:Any)	0	2	−2.14
N(:N(:Any))(:Any)	0	2	−2.14
N(-H)(-N(=N(:Any)))(:Any)	0	4	−0.86

continued on next page

Table 5 – concluded from previous page

Fragment	Priority	# atoms	LogP
N(-H)(:N(:N(:Any)))(:Any)	0	4	−0.86
C(-H)(=N(-N(-H)(:Any)))(:Any)	0	5	−0.47
C(-H)(:N(:N(-H)(:Any)))(:Any)	0	5	−0.47
N(=C(-H)(-N(-H)(:Any)))(:Any)	0	5	−0.79
N(:C(-H)(:N(-H)(:Any)))(:Any)	0	5	−0.79
N(=C(-H)(-O(:Any)))(:Any)	0	4	−0.71
N(:C(-H)(:O(:Any)))(:Any)	0	4	−0.71
C(-H)(=N(-O(:Any)))(:Any)	0	4	−0.63
C(-H)(:N(:O(:Any)))(:Any)	0	4	−0.63
N(=C(-H)(-N(:Any)))(:Any)	0	4	−1.46
N(:C(-H)(:N(:Any)))(:Any)	0	4	−1.46
S(:Any)(:Any)	0	1	0.36
N(=C(-H)(-S(:Any)))(:Any)	0	4	−0.29
N(:C(-H)(:S(:Any)))(:Any)	0	4	−0.29
C[1]:C(-H):C(-H):C(-H):C(-H):C(-H):@1(-Any)	0	11	1.90
N(=N(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	2	2	0.14
N(=N[charge=+1](=N[charge=-1]))(-Any(:Any)(:Any))	1	3	0.69
N(=O)(-Any(:Any)(:Any))	1	2	0.11
C(#N)(-Any)	0	2	−1.27
C(#N)(-Any(:Any)(:Any))	1	2	−0.34
C(#N)(-C(-H)(-C[1]:C(-H):C(-H):C(-H):C(-H):C(-H)@1))	3	2	−0.88
C(-H)(=N(-Any))(-Any(:Any)(:Any))	1	3	−1.03
C(-H)(=N(-Any(:Any)(:Any)))(-Any(:Any)(:Any))	2	3	0.08
C(=N(-H))(-Any(:Any)(:Any))(-Any(:Any)(:Any))	2	3	−1.29
I(=O)(=O)(-Any(:Any)(:Any))	1	3	−3.23
P(=O)(-Any(:Any)(:Any))(-Any(:Any)(:Any))(-Any(:Any)(:Any))	3	2	−2.45

Table 6: Polyhalogenation at the same carbon atom

Fragment	Priority	# atoms	LogP
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])	0	4	0.60

continued on next page

Table 6 – concluded from previous page

Fragment	Priority	# atoms	LogP
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])	1	4	1.59
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])	2	4	2.88

Table 7: Polyhalogenation at adjacent carbon atoms

Fragment	Priority	# atoms	LogP
C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	0	4	0.28
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	1	4	0.56
C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	1	4	0.56
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	2	4	0.84

continued on next page

Table 7 – concluded from previous page

Fragment	Priority	# atoms	LogP
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	2	4	0.84
C(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I]))	2	4	0.84
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I]))	3	4	1.12
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[not=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I]))	3	4	1.12
C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-C(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I])(-Any[is=F,Br,Cl,I]))	4	4	1.40

B Implementation

B.1 Example of a SLN string and its dedicated subgraph structure

This is an example of subgraph structure for a given SLN string.

```
Name             : N(-P(=S)(-N[1])(-N[2]))(-Any)(-Any)
                     .(Any-@1).(Any-@1).(Any-@2).(Any-@2)
------------------------------------------------------
Group            :    1
Priority         :    0
Vertex nr.       :    1
 Elements       :   -7:(N)
 Charge         :not specified
 Adjacency list :    7(1)   6(1)   2(1)
Vertex nr.       :    2
  Elements       :  -15:(P)
  Charge         :not specified
  Adjacency list :    5(1)   4(1)   3(2)   1(1)
Vertex nr.       :    3
  Elements       :  -16:(S)
  Charge         :not specified
  Adjacency list :    2(2)
Vertex nr.       :    4
  Elements       :   -7:(N)
  Charge         :not specified
  Adjacency list :    9(1)   8(1)   2(1)
Vertex nr.       :    5
  Elements       :   -7:(N)
  Charge         :not specified
  Adjacency list :   11(1)  10(1)   2(1)
Vertex nr.       :    6
  Elements       : all
  Charge         :not specified
  Adjacency list :    1(1)
Vertex nr.       :    7
  Elements       : all
  Charge         :not specified
  Adjacency list :    1(1)
Vertex nr.       :    8
  Elements       : all
  Charge         :not specified
  Adjacency list :    4(1)
Vertex nr.       :    9
  Elements       : all
  Charge         :not specified
  Adjacency list :    4(1)
Vertex nr.       :   10
  Elements       : all
```

```
  Charge         :not specified
  Adjacency list :    5(1)
Vertex nr.       :   11
  Elements       : all
  Charge         :not specified
  Adjacency list :    5(1)
```

B.2 Example of program output

In this section we present typical program output files, using Diazepam as an example.

```
1.410000

# Name: Diazepam

# Fragment                          |LogP
# ----------------------------------------
# C(-H)(-H)(-H)(-Any)               |0.89
# C(-Any)(:Any)(:Any)               |0.13
# C(-Any)(:Any)(:Any)               |0.13
# C(-Any)(:Any)(:Any)               |0.13
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# C(-Any)(:Any)(:Any)               |0.13
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# C(-H)(:Any)(:Any)                 |0.355
# N(-Any)(-Any)(-Any(:Any)(:Any))   |-0.56
# C(=O)(-Any)(-Any)                 |-1.90
# C(=Any)(-Any)(-Any)               |0.20
# N(=Any)(-Any)                     |-2.18
# C(-Any)(-Any)(-Any)(-Any)         |0.20
# Cl(-Any(:Any)(:Any))              |0.94
# H(-Any)                           |0.23
# H(-Any)                           |0.23
# Polyhalogenation                  |0.000000

# Total LogP: 1.410000
```

The next file shows a typical output, if an error occured. This file was artificially generated by removing all chlorine containing fragments from the databases and performing a calculation with Diazepam. As expected, the chlorine atom is not visited and thus this error file is created.

```
NULL

#"LogP for Diazepam could not be calculated
#because of the following not considered atoms!

# ID                                |Atom
# ----------------------------------------
# 20                                |Cl
```

B.3 Usage and command line options

The program is fully automated and can be started within a shell using the program name "logp" and three additional necessary options:

- -i string
  set the full path of the query molecule
- -o string
  set the full path of the solution/error file
- -m string
  select the method(only "logp" available yet)

C Structural descriptors not implemented

In the following we give a list of the most important structural descriptors not implemented yet:

- Multiple bonding within carbon chains
  - Double bond (-0.09)
  - Triple bond (-0.50)
- Fragment attachments to an aromatic ring with a second electron withdrawing substituent.
  - Hammet σ_I of second substituent exceeds 0.50 and two halogens are also attached
  - Hammet σ_I of second substituent exceeds 0.75
- Conformational flexibility
  - Hydrocarbon chains $(n-1)(-0.12)$
  - Alicyclic rings $(n-1)(-0.09)$

- Chain branching
  - Nonpolar chain (-0.13)
  - Polar chain (-0.22)
- Nonhalogen polar fragments
  - On same carbon atom
    - In chain $(-0.42)(f_1 + f_2)$
  - On adjacent carbon atoms
    - In chain $(-0.26)(f_1 + f_2)$
    - In alicyclic ring $(-0.32)(f_1 + f_2)$
    - In aromatic ring $(-0.16)(f_1 + f_2)$
  - On carbon atoms separated by one carbon atom
    - In chain $(-0.10)(f_1 + f_2)$
    - In alicyclic ring $(-0.20)(f_1 + f_2)$
    - In aromatic ring $(-0.08)(f_1 + f_2)$
- Intramolecular hydrogen bonding
  - With –OH (1.0)
  - With –NH (0.6)
- Subdividing of nonhalogen fragments into three classes according to the strength of interactions with α-halogens.
  - –S(=O)– and –SO_2–: Add 0.9 for each α-halogen.
  - –CONH–R, –O–R, –S–R and –NHR: Add 0.9 for the first α-halogen and half that for each of the next two α-halogens.
  - All remaining fragments: Add 0.9 for the first α-halogen only.