Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final project report 2013

25 May 2013

Scaffold analysis for interpretation of QSAR models

Duration of Short Term fellowship:
February - May 2013

Early stage researcher:
"Milan Voršilák"

Project supervisor:
"Dr. Igor Tetko"

Research Institution:
Helmholtz Centrum Munich

Introduction

Goal

The main point of my work was to identify the most promising method and software for computing chemical scaffolds of compounds and implement it in the On-line Chemical Modeling Environment (OCHEM) to be used as descriptors for QSAR (Quantitative Structure-Activity Relationship) models [1] as well as for the interpretation of QSAR/QSPR models. The next step was to develop workflow for chemical space mapping with use of previously implemented tool(s). This tool should compute coverage of chemical space by a molecule library and visualize it as a map. Maps were used to display differences and similarities between libraries. The comparison of coverage of chemical space for chemical industry, as provided by ECHA preliminary list [2], and pharmaceutical industry, as provided by ChEMBL [3] and DrugBank [4], was done.

What is scaffold?

Chemical scaffold describes the most significant motif of a molecule that is based on ring systems. There are many ways how to create them but main idea is to keep rings and just remove side chains. This method describes compounds with one scaffold as the same. It is used for investigation of common shapes of molecules.

One of the first articles, which proposed chemical frameworks, was written by Bemis and Murcko (BM) [5]. They divide structure of molecule to side chains and framework that consists of ring systems and linkers. Ring system is one or more fused rings. Linker connects two or more ring systems. The authors proposed two levels: simply remove side chains. In a more general approach all bonds were converted to single bonds and all heteroatoms were converted to carbon. First type is called BM scaffold and the other one is known as BM framework. We call scaffold such process that does not change atoms and bonds in a rings in a difference to framework, which does it.

Oprea et al came with even more generalized structure [6]. They preserved only topology. Thus structures, to some extent, were losing chemical meaning and represented only the overall topology of atoms.

Another approach was used by Schuffenhauer et al [7]. From the largest (original) scaffold they hierarchically removed rings by set of rules, in each step the least significant ring was removed.

Software for scaffold analysis

New implementation of scaffold calculation tools from a scratch would take a lot of time and would be redundant. Therefore, at first, we analyzed the existing software, which provide calculation of different types of scaffolds. We found several tools which implements basic BM scaffolds and frameworks, such as RDKit [8] and Indigo [9]. These types of structural features were also offered by Chemaxon kit [10] and Strip-it [11], which we eventually decided to use due to their simplicity of integration as well as their coverages of the provided features.

Strip-it was developed by Silicos-it [12]. It is written in C++ and calculates BM scaffolds and frameworks. It provides several options for the customization of calculations. Strip-it allows keeping double bonded substituent (e.g. keto group) for scaffolds or compressing linkers to unity length for frameworks. This tool can compute topology of a ring system, similar to that proposed by Oprea, and it adds information about H-acceptors and donors to this topology. Strip-it also implements, Schuffenhauer scaffolds, which are based on scaffold tree [7]. Strip-it calculates up to 5 levels of scaffolds for a single molecule.

Chemaxon kit implements BM scaffolds and frameworks only. We decided to implement this tool to have an alternative way to characterize chemical structures. Moreover, the Chemaxon kit was developed in Java thus simplifying its incorporation with OCHEM, which was also developed in Java.

Thus, we mainly selected Strip-it because of diversity of scaffold types and Chemaxon, because of its convenience.

Implementation of scaffold descriptors

We chose Strip-it and Chemaxon kit for their incorporating into OCHEM [1] as descriptors calculation blocks. OCHEM is web-based platform for automating and simplifying steps required for QSAR modeling. The platform consists of two major subsystems: the database of experimental measurements and the modeling framework. The descriptors based on the scaffold representation can be used for data modeling and/or for comparison of datasets.

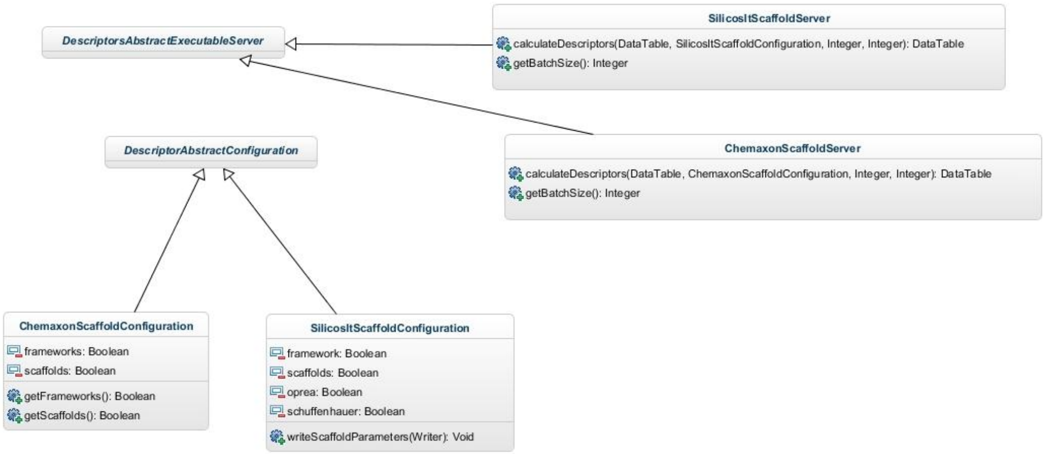

DescriptorsAbstractExecutableServer
SilicosItScaffoldServer
calculateDescriptors(DataTable, SilicosItScaffoldConfiguration, Integer, Integer): DataTable
getBatchSize(): Integer
DescriptorAbstractConfiguration
ChemaxonScaffoldServer
calculateDescriptors(DataTable, ChemaxonScaffoldConfiguration, Integer, Integer): DataTable
getBatchSize(): Integer
ChemaxonScaffoldConfiguration
frameworks: Boolean
scaffolds: Boolean
getFrameworks(): Boolean
getScaffolds(): Boolean
SilicosItScaffoldConfiguration
framework: Boolean
scaffolds: Boolean
oprea: Boolean
schuffenhauer: Boolean
writeScaffoldParameters(Writer): Void

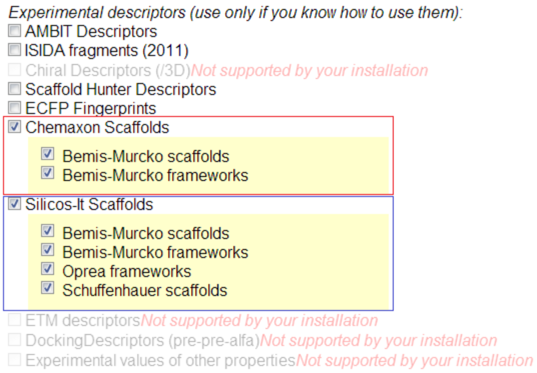
Fig. 1 UML diagram of created classes

Functions calculating descriptors were inherited from *DescriptorsAbstractExecutableServer* and override abstract method *calculateDescriptors(...)* (see Fig. 1). These classes calculate descriptors and allow parallelization of computation on several servers simultaneously.

Program Strip-it is available as a standalone tool [13], which operates by means of input and output files. Input should be in a format readable by OpenBabel [14]. We used SDF files, which are default ones for storing of chemical information in OCHEM. Parameters of the analysis are provided to the program as a file, which is written by a configuration class. Output file is similar to csv file with white spaces as delimiters. Molecules without any ring are transformed into methane as the simplest hydrocarbon.

Chemaxon kit was already used by OCHEM before our development. As it was aforementioned, Chemaxon kit shared the same programming language, Java as OCHEM. Therefore, the API of OCHEM was used to implement the descriptor server. Namely, functionality of Chemaxon's *StructuralFrameworksPlugin*, which splits molecules to scaffolds, frameworks and ring systems, was used to implement the server.

As a result of this part of work, two descriptors blocks, that can be used during a process of creating predictive models and for statistical analysis of datasets (see Fig. 2 and highlighted area) were added and implemented in OCHEM.



Experimental descriptors (use only if you know how to use them):
- □ AMBIT Descriptors
- □ ISIDA fragments (2011)
- □ Chiral Descriptors (/3D)*Not supported by your installation*
- □ Scaffold Hunter Descriptors
- □ ECFP Fingerprints
- ☑ Chemaxon Scaffolds
  - ☑ Bemis-Murcko scaffolds
  - ☑ Bemis-Murcko frameworks
- ☑ Silicos-It Scaffolds
  - ☑ Bemis-Murcko scaffolds
  - ☑ Bemis-Murcko frameworks
  - ☑ Oprea frameworks
  - ☑ Schuffenhauer scaffolds
- □ ETM descriptors*Not supported by your installation*
- □ DockingDescriptors (pre-pre-alfa)*Not supported by your installation*
- □ Experimental values of other properties*Not supported by your installation*

Fig. 2 Experimental descriptors in OCHEM

Silicos-it descriptors were divided into 4 groups: BM scaffolds and frameworks, Oprea frameworks and Schuffenhauer scaffolds. Each option keeps similar parameters together (see Fig. 3 and for more detail information).

Chemaxon scaffolds has 2 options: calculation of BM scaffolds and frameworks. Additionally scaffolds and frameworks are divided to ring systems with removing linkers (see Fig. 4).

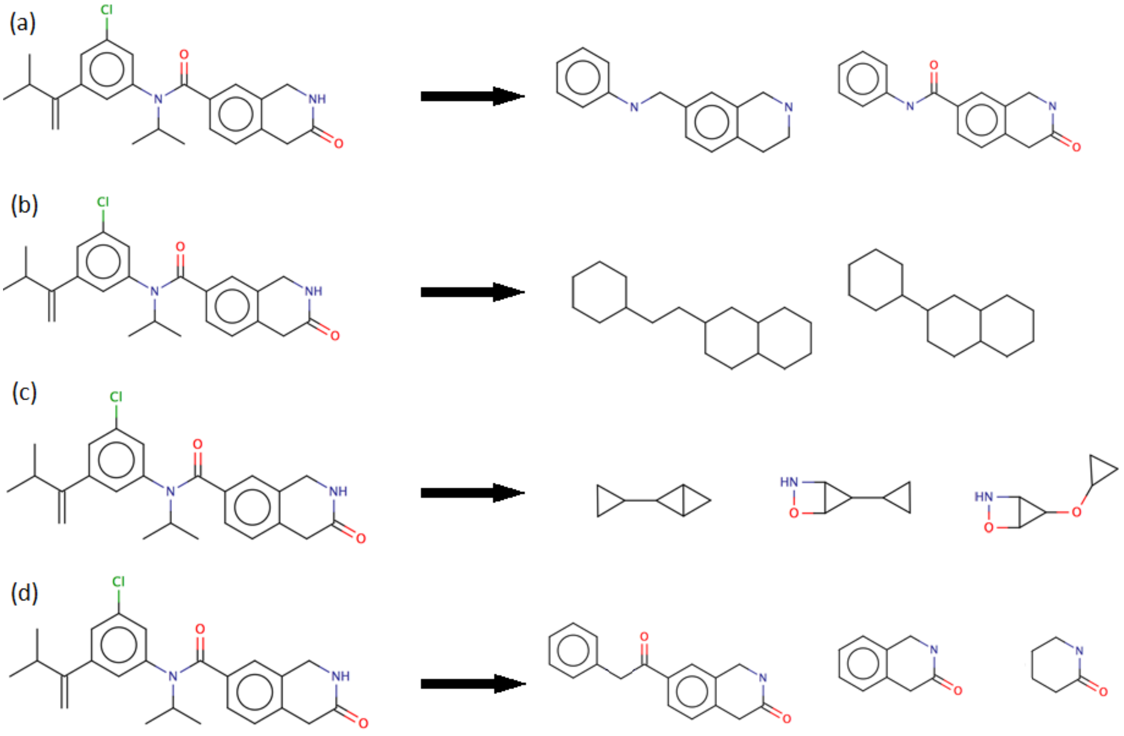


Fig. 3 Silicos-it scaffolds: a) BM scaffolds, b) BM frameworks, c) Oprea frameworks and d) Schuffenhauer scaffolds

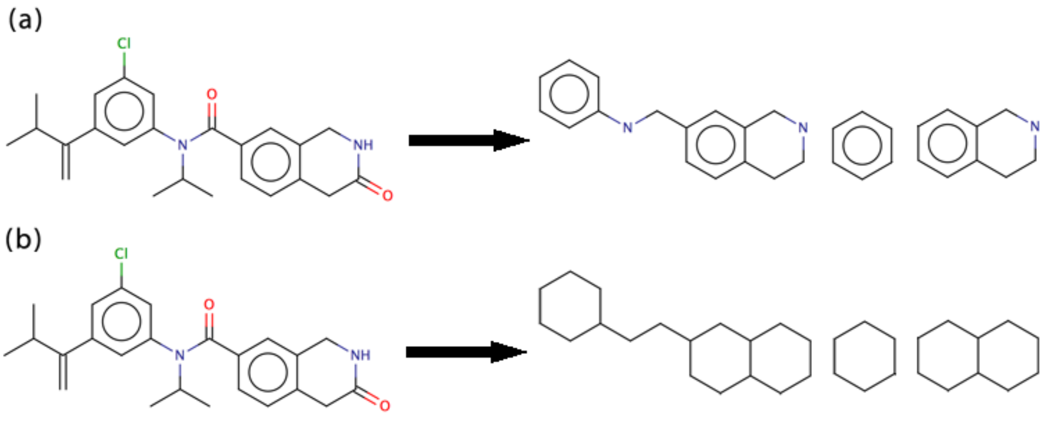


Fig. 4 Chemaxon scaffold descriptors, (a) BM scaffolds, (b) BM framework

Visualizing of chemical (scaffold) space

Data preparation

Chemical space is (almost) infinite (10^{60} compounds lighter then 500 Da [15]) so it can't be possible to have a complete database. One could use artificially created database GDB [16] but then he/she would be limited with 17 atoms per molecule and lost resolution for bigger molecules. Therefore, we used a set of databases with commercially available compounds as the data source. It contains approximately 66 million compounds.

Calculation of Scaffolds

Schuffenhauer scaffolds were computed for all standardized compounds with OCHEM. Due to implementation reasons all 5 levels of scaffolds were computed for each molecule. It means that one molecule could have from 1 to 5 scaffolds depending on the number of rings. It makes smaller scaffolds more frequent ones.

10,75M scaffolds were calculated for all compounds in the database. E-State descriptors [17] were calculated for each of them. For further research we used only scaffolds containing organic subset of atoms as defined in Daylight SMILES documentation [18]. E-State descriptors were chosen for its fastness and simplicity. They were successfully used in many QSAR/QSPR studies and frequently contributed top-performing models [19] [20] [21] [22] [23]. However, these descriptors are rather difficult to interpret. Another sets of more easily interpretable descriptors can be used in further research and for comparison, e.g. those from references [16] [24].

Datasets of Molecules

We decided to compare the chemical coverage of molecules used by pharmaceutical companies with those used by chemical industry. The drug-like molecules included DrugBank set, more precisely all approved drugs [25], and ChEMBL DrugStore set from ZINC inventory [26]. The environmental dataset included molecules from EINECS (European INventory of Existing Commercial chemical Substances) list [27].

Self-Organizing map

For visualization we chose Self-Organizing Map (SOM) first proposed by Kohonen [28]. SOM easily transform high dimensional space to a 2D space. It preserves topological properties of the analyzed datasets. SOM is inspired by organization of the brain cortex and it is one of the most popular artificial neural network methods. We used open source SOM implementation [29]. The application of SOM includes training phase. During this stage the algorithm adjusts the connections of the network and builds the projections of data to SOM neurons. The training is done in cycles (epochs) and its speed is proportional to the number of neurons and training samples as well as to the dimensionality (number) of the descriptors.

Results

Square-shaped SOM with 2,500 neurons was trained during 500 epochs, which is the default proposed number for the algorithm. Training set was 100.000 scaffolds randomly chosen from previously prepared database and thus corresponding to about 0.1% of scaffolds in the database. Each sample was represented with 38 E-State descriptors. All descriptors were normalized into range -1 to 1.

SOM learning phase with these settings took about 26 hours. We were not able to use larger training set or SOM due to the memory constrains.

Training data covered almost the whole map, exactly 95.8% of area as shown on the density map (see Fig. 5). Average intraneural Euclidean distance between scaffolds is 0.223 (i.e., distance between scaffolds which were clustered to the same SOM neuron), but for randomly selected scaffolds the same distance was 0.589 (rounded value for several runs). This is not surprising as the main task of SOM is to cluster similar objects together. Similar results were obtained using Tanimoto similarity and MACCS (Molecular ACCess System) fingerprints computed with OpenBabel. Average intraneural Tanimoto similarity was 0.686, but with random scaffolds similarity decreases to 0.501. Average similarity between scaffolds in training set was 0.474. Therefore the developed map keeps similar scaffolds close together and these scaffolds are related with respect to different similarities measures.

In the same way as for training set we calculated Schuffenhauer scaffolds for two drug sets and one environmental datasets. Each scaffold was represented with the same set of E-State descriptors as previously described. Afterwards we mapped scaffolds into a map. Each scaffold belonged to its closest neuron. Density maps and statistical data can be seen in Fig. 6 and Tab. 1, respectively.

Intersection between sets is shown in Fig. 7. All 3 datasets covered 65.8% of the scaffold space. Drug sets shared 35.4% of space while differed on 24.2% of space. The EINECS set scaffolds covered only 37.1% of space while ChEMBL and DrugBank sets covered 49.8% and 45.1%, respectively. Additionally, only 16.7% scaffolds from this set were not in the scaffold space covered by drug-like molecules from the other two sets. This is an interesting result, since EINECS set was 6 and 10 times larger than ChEMBL and DrugBank sets, respectively. Thus this set consisted from more restricted set of scaffolds, which, to the large degree, were also represented in the sets of pharmaceutically relevant compounds. Therefore, the models developed to cover drug like sets could be also applicable to compounds from the EINECS set too.

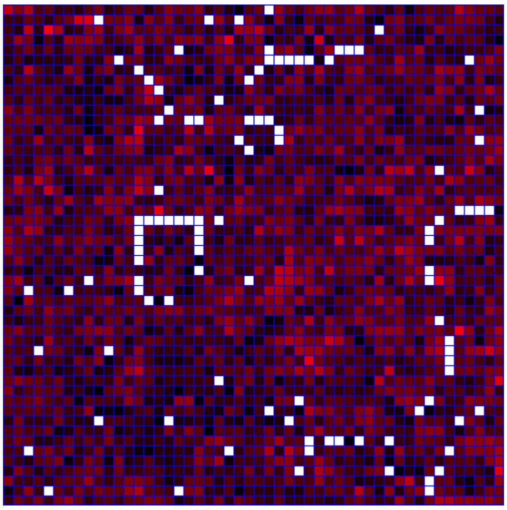


Fig. 5 Density map of the training dataset; white cells are unoccupied, blue or red colors reflect the occupation frequency (red is higher)

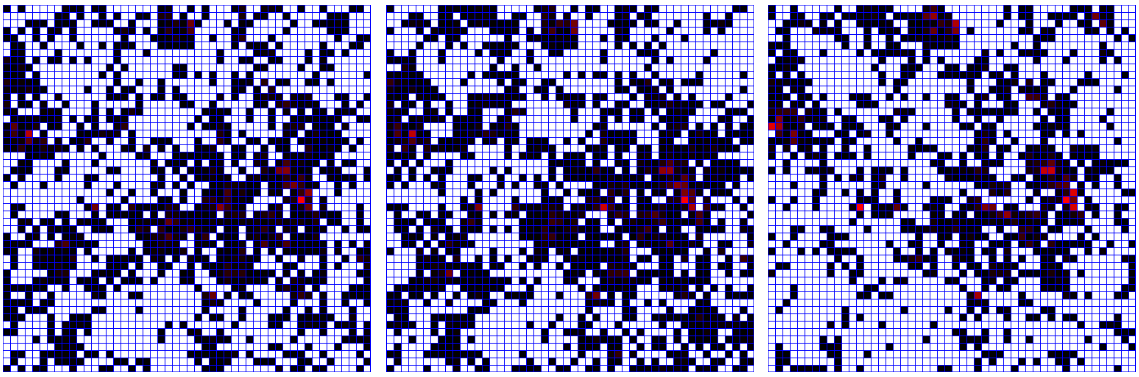


Fig. 6 Density map of analysed databases, a) ChEMBL drugstore, b) DrugBank and c) EINECS set

Tab. 1 Size and coverage of examined datasets

	ChEMBL DrugStore	DrugBank	EINECS
Compounds	11799	6541	70767
Scaffolds	10425	6834	12012
Coverage	49.84%	45.08%	37.08%

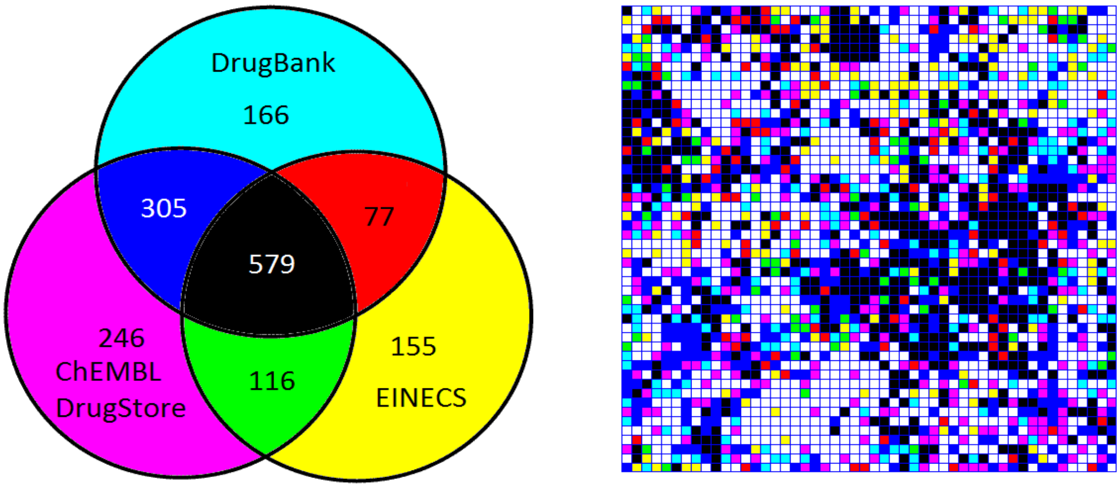


Fig. 7 Venn diagram and map displaying intersections between datasets . The numbers of SOM neurons (the total number was 2500) occupied by scaffolds from different sets are shown.

Conclusion

We implemented into OCHEM an option to calculate chemical scaffolds as molecular descriptors for predictive model building, for analysis of datasets and interpretation of models.

Using the previously developed tool we trained self-organizing map and then used it compare results of mapping of drug and environmental datasets. ChEMBL drugstore set covers nearly 50% of scaffold space and DrugBank set 45%, although it contains around 35% less scaffolds than ChEMBL set. Both drug sets share large part of the map area but the other part is still significant. We didn't find that drugs are concentrated in certain area but all 3 datasets intersect in the central region of the map. This region may contain the most common scaffolds.

EINECS dataset surprisingly covers only 37% of space despite it is almost an order larger than each of the drug-like sets. This could mean that many its compounds have the same scaffold and scaffolds in this set are more similar. Therefore many compounds may have similar physico-chemical and biological properties and their prediction can be done more easily to that of pharmaceutically relevant molecules. Moreover, scaffolds from drug like sets cover the scaffolds from this set. Therefore, the models developed for drug like compounds could be highly predictive and cover the applicability domain for the environmental compounds.

References

[1] I. Sushko, S. Novotarskyi, R. Korner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmare, A. Cherkasov, J. Aires-de-Sousa, Q. Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko and I. V. Tetko, "nline chemical modeling environment (OCHEM): web," *J. Comput. Aided. Mol. Des.,* vol. 25, pp. 533-540, 2011.

[2] "Pre-registered substances," European Chemicals Agency, [Online]. Available: http://echa.europa.eu/information-on-chemicals/pre-registered-substances. [Accessed 1 April 2013].

[3] "ChEMBL," EMBL-EBI, [Online]. Available: https://www.ebi.ac.uk/chembldb/. [Accessed 1 April 2013].

[4] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Guo and D. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.," *Nucleic Acids Res.,* vol. 39, pp. 1035-1041, Jan 2011.

[5] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. Molecular frameworks," *J. Med. Chem.,* vol. 39, pp. 2887-2893, 1996.

[6] S. N. Pollock, E. A. Coutsias, M. J. Wester and T. I. Oprea, "Scaffold topologies. 1. Exhaustive enumeration up to eight rings," *J. Chem. Inf. Model.,* vol. 48, pp. 1304-1310, 2008.

[7] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch and H. Waldmann, "The scaffold tree - visualization of the scaffold universe by hierarchical scaffold classification," *J. Chem. Inf. Model.,* vol. 47, pp. 47-58, 2007.

[8] G. Landrum, "RDKit: Cheminformatics and Machine Learning Software," [Online]. Available: http://www.rdkit.org/.

[9] "GGA Software Services – Indigo Toolkit," GGA Software Services LLC, [Online]. Available: http://www.ggasoftware.com/opensource/indigo.

[10] "ChemAxon – cheminformatics platforms and desktop applications," ChemAxon Ltd., [Online]. Available: http://www.chemaxon.com/. [Accessed 4 March 2013].

[11] "Strip-it™," Silicos-it, [Online]. Available: http://silicos-it.com/software/strip-it/1.0.2/strip-it.html. [Accessed 4 March 2013].

[12] "Silicos-it," Silicos-it, [Online]. Available: http://silicos-it.com/. [Accessed 4 March

2013].

[13] Silicos-it, "Strip-it™," [Online]. Available: http://silicos-it.com/_php/download.php?file=strip-it-1.0.2.tar.gz. [Accessed 4 March 2013].

[14] "Open Babel: The Open Source Chemistry Toolbox," [Online]. Available: http://openbabel.org/wiki/Main_Page. [Accessed 4 March 2013].

[15] R. S. Bohacek, C. McMartin and W. C. Guida, "The art and practice of structure-based drug design: A molecular modeling perspective," *Medicinal Research Reviews,* vol. 16, pp. 3-50, 1996.

[16] J.-L. Reymond and M. Awale, "Ecploring Chemical Space for Drug Discovery Using the Chemical Universe Database," *ACS Chem. Neurosci.,* vol. 3, pp. 649-657, 2012.

[17] H. H. Lowell and K. B. Lemont, "Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information," *Journal of Chemical Information and Computer Sciences,* vol. 35, pp. 1039-1045, 1995.

[18] "Daylight Theory: SMILES," Daylight Chemical Information Systems, Inc., [Online]. Available: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. [Accessed 25 March 2013].

[19] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A. K. Pandey and I. V. Tetko, "Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients.," *J Chem Inf Model.,* vol. 49, pp. 133-144, 2009.

[20] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropshe, E. Papa, T. Oberg, R. Todeschini, D. Fourches and A. Varnek, "Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection," *J Chem Inf Model.,* vol. 48, 2008.

[21] I. V. Tetko, V. P. Solov'ev, A. V. Antonov, X. Yao, J. P. Doucet, B. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachine and A. Varnek, "Benchmarking of linear and nonlinear approaches for quantitative structure-property relationship studies of metal complexation with ionophores.," *J Chem Inf Model.,* vol. 46, pp. 808-819, 2006.

[22] A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin and V. P. Solov'ev, "Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points?," *J Chem Inf Model.,* vol. 47, pp. 1111-1122, 2007.

[23] I. V. Tetko, S. Novotarskyi, I. Sushko, V. Ivanov, A. E. Petrenko, R. Dieden, F. Lebon and B. Mathieu, "Development of dimethyl sulfoxide solubility models using 163,000 molecules: using a domain applicability metric to select more reliable predictions," *J Chem Inf Model.,* vol. 53, 2013.

[24] P. Ertl, S. Jelfs, J. Mühlbacher, A. Schuffenhauer and P. Selzer, "Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroatomatic Scaffolds," *J. Med. Chem.,* vol. 49, pp. 4568-4573, 2006.

[25] "DrugBank: Downloads," DrugBank, [Online]. Available: http://www.drugbank.ca/downloads. [Accessed 1 April 2013].

[26] "ChEMBL Drugstore," ZINC, [Online]. Available: http://zinc.docking.org/catalogs/drugstore. [Accessed 1 April 2013].

[27] "EC Inventory," Institute for Health and Consumer Protection, [Online]. Available: http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/information-sources/ec_inventory. [Accessed 6 May 2013].

[28] T. Kohonen, "Self-Organizing Maps. 3 ed.," *Springer Verlag: New York,* 2000.

[29] Y. Chesnokov, "Kohonen's Self Organizing Maps in C++ with Application in Computer Vision Area," [Online]. Available: http://www.codeproject.com/Articles/21385/Kohonen-s-Self-Organizing-Maps-in-C-with-Applicati. [Accessed 8 May 2013].