Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of physicochemical and biological activities of small molecules

Iurii Sushko[1,2], Sergii Novotarskyi[1,2,], Anil Kumar Pandey[1], Robert Körner[1,2], Matthias Rupp[1], Wolfram Teetz[1,2], Stefan Brandmaier[1], Ahmed Abdelaziz[1,2], Volodymyr V. Prokopenko[3], Vsevolod Y. Tanchuk[3], Roberto Todeschini[3], Alexander Varnek[3], Gilles Marcou[3], Peter Ertl[3], Vladimir Potemkin[3], Maria Grishina[3], Johann Gasteiger[3], Igor I. Baskin[3], Vladimir A. Palyulin[3], Eugene V. Radchenko[3], William J. Welsh[3], Vladyslav Kholodovych[3], Dmitriy Chekmarev[3], Artem Cherkasov[3], Joao Aires-de-Sousa[3], Qing-You Zhang[3], Andreas Bender[3], Florian Nigsch[3], Luc Patiny[3], Antony Williams[3], Valery Tkachenko[3], Igor V. Tetko[1,2]*

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[2]eADMET GmbH, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany

[3]OCHEM consortium
**Presenting author, sergii.novotarskyi@helmholtz-muenchen.de*

The Online Chemical Modeling Environment is a web-based platform that aims to automate and simplify the typical steps required for QSAR modeling. The platform consists of two major subsystems: the database of experimental measurements of chemical, physical and biological activities of small molecules and the modeling framework.

A user-contributed database contains a set of tools for easy input, search and modification of thousands of records. The OCHEM database is based on the wiki principle and focuses primarily on the quality and verifiability of the data. This is facilitated by storing the links to the articles where the data was initially published. The database also gives the possibility to store additional information about the datapoint in the form of experimental conditions - pH, temperature, pressure, etc. Tight integration with PubChem, PubMed and other publicly available services allows mass import and verification of data. Import and export of date to and from popular formats (CSV, SDF and Excel) is also possible.

The database is tightly integrated with the modeling framework, which supports all the steps required to create a predictive QSAR model. Advanced filters allow to select appropriate data and group the records to training and validation sets. The modeling framework provides a selection of popular machine learning methods, meta-learning methods and validation techniques. Both classification and regression tasks are supported, and several methods support multi-learning techniques. A rich and constantly expanding selection of molecular and atomic descriptors allows a thorough modeling of most physicochemical and biological properties. Applicability domain techniques are also implemented and provide estimation of model accuracy for every predicted compound.

The system is designed in a way to make every model reproducible. The models, model creation protocols and the original datasets can be made publicly available.

As compared to other similar systems, OCHEM is not intended to re-implement the existing tools or models but rather to invite the original authors to contribute their results, share them with other users and to become members of the growing research community. Our intention is to make OCHEM a widely used platform to perform the QSPR/QSAR studies online and share it with other users on the Web. The ultimate goal of OCHEM is collecting all possible chemoinformatics tools within one simple, reliable and user-friendly resource.

The OCHEM is free for web users and it is available online at http://www.ochem.eu .