HelmholtzZentrum münchen
German Research Center for Environmental Health

CADASTER

Demonstration of tools developed in Cadaster

Igor V. Tetko and Iurii Sushko

Helmholtz Zentrum München - German Research Center

for Environmental Health (GmbH) and eADMET GmbH

eADMET
THE REFERENCE IN CHEMOINFORMATICS

Program of the session

- Concept of the web site
- Data
- Models
- Upload of models
  - ✧ Tutorial on models upload
- Experimental design

Additional training is available any time during the conference: contact me or Yura

CADASTER tools

- CADASTER web: http://qspr-thesarus.eu
- A customization of http://ochem.eu according to the CADASTER requirements
  - Upload of models
  - QRMF support
  - Web services
  - Integration with OpenTox API (similarity search, AD)
  - Database of 3D structures http://mopac.cadaster.eu

Data

- Properties (user defined)
- Values (as in publication)
- Units (as in the publication) + automatic unit conversion
- Publication source (obligatory)
- Name, CAS RN (validation of names; stereochemistry)
- Evidence for data (experimental, collection of other) + experimental & primary records
- Position in the article (table, page, N)
- Conditions; obligatory conditions
- Introducer/modifier of data + modification time
- Private and public records
- Record and molecule Ids
- Duplicates within an article

Working with data

- Creating properties & condition
- Editing a single record
- Duplication of a record
- Selection of records
- Creation of sets of records
- Working with sets (merging, cloning)
- Batch edit of records
- Batch upload of records
- Export of records

Models

- Supports MLRA & PLS*
- Workflows for model development
- Training and validation sets
- Calculation of AD
- Public and private models
- Export of models and properties
- Accuracy plot
- Integration with QRMF

*More methods on OCHEM http://ochem.eu

Developed models

- Grouped by classes of chemical compounds
- References to original sources of data & publication
- Data, descriptors* and models are downloable
- Identification of AD of models
- Storage of calculated values (to overcome problem of 3D structures)

* With an exception of some commercial descriptors

Predictions of new data

- Draw a molecule
- Upload from an SDF file
- Use previously prepared set or tag
- Export in various formats (Excel, sdf or/and csv)

Upload of models

- Data preparation
  - Upload data using Excel or sdf files
  - Create training and test sets
  - Prepare model coefficients in Excel file
- Upload
  - select datasets and correct units
  - Select pre-processing steps
  - Select descriptors
  - Upload model coefficients (help is provided)
  - Verify mapped coefficients
  - Add AD
  - Start calculations
  - Verify your model by comparing statistics
  - Download your model to verify it
- Add QRMF
- Publish your model

Tutorial

- Detailed steps: http://www.cadaster.eu/maribor

Things to explore:

- Same descriptors but different software versions
  - Dragon 5.4, 5.5, 6.0
- Same software versions but different pre-processing steps
  - No optimization
  - Corina
  - MOPAC AM1

Experimental design

Problem: select compounds for experimental measurements

- Select the design set with BCF measurements (179 compounds)
- **Select 20 compounds** using D-Optimal design
- Build a PLS model on the selected compounds
- **Select 20 more compounds** using PLS-Optimal design and our model
- Build a second PLS model using **all 40 selected compounds**
- *(Optional)* Repeat the PLS-optimal process multiple times, each time selecting more compounds for measurement

Thank you for your attention!