(Big) Data analysis using On-line Chemical database and Modelling platform

Dr. Igor V. Tetko

Institute of Structural Biology, Helmholtz Zentrum München & BIGCHEM GmbH

September 14, 2018, EPFL, Lausanne
Data storage and model development: http://ochem.eu
OCHEM Database schema

Molecules

Names

Articles

Properties

- log(I(GC50-1)) = 2.02 - log (mmol/L)
  Zhu, H
  Combinatorial QSAR modeling of chemical toxicants tested against N. crassa
  Journal of chemical information and modeling 2008, 48 (4) 766-84
  2579-22-0, phenylpropargyl aldehyde

Conditions

- Temperature = 25.0

Units

- log(mmol/L) (concentration)
- log(mg/L) (concentration)
- nmol (concentration)

Users

Units

- log(I(GC50-1)) (concentration)
- LogPsuv (dimensionless)
- LogPsuv(ion) (dimensionless)
- LogPI (dimensionless)

Species

- Temperature (temperature)
- Dose (concentration)
- Concentration (concentration)

CC-BY license for the uploaded data
### Support of mixtures

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Mixture Formula</th>
<th>Molecular Structure</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>R32620625</td>
<td>CCl(\text{Cl})ccCl</td>
<td><img src="image1" alt="Molecule Image" /></td>
<td>Public and freely downloadable record (awaiting approval)</td>
</tr>
<tr>
<td>R32620624</td>
<td>CCCO(\text{CC})C</td>
<td><img src="image2" alt="Molecule Image" /></td>
<td>Public and freely downloadable record (awaiting approval)</td>
</tr>
<tr>
<td>R32620623</td>
<td>CO(\text{C}(\text{Cl}))OCC</td>
<td><img src="image3" alt="Molecule Image" /></td>
<td>Public and freely downloadable record (awaiting approval)</td>
</tr>
<tr>
<td>R32620622</td>
<td>CCC(\text{Cl})C(\text{C})</td>
<td><img src="image4" alt="Molecule Image" /></td>
<td>Public and freely downloadable record (awaiting approval)</td>
</tr>
</tbody>
</table>
QSPR/QSAR modelling in OCHEM

Select the training and validation sets:

- Training set (required): hERG blockage training.xls [details]
- Add a validation set

The model will predict this property:
- hERG K+ Channel Blocking using unit: CLASS

Choose the learning method:

- ASNN: ASSociative Neural Networks
- CHEMCHAINER: Chainer Chemistry models (GPU)
- Consensus model (based on models developed for the same set)
- DEEPCHEM: several methods from DeepChem (GPU)
- DNN: Deep Neural Network (GPU)
- FSMLR: Fast Stagewise Multiple Linear Regression
- KNN: k - Nearest Neighbors
- Library model (A local bias correction model based on another ASNN model)
- LibSVM: grid-search parameter optimisation
- LSSVMG: Least Squares Support Vector Machine (GPU)
- MLR: Multiple Linear Regression
- NNF2T: Tensor flow version of NNF2N: another Neural Network Fingerprint (GPU)
- PLS: Partial Least Squares
- RFR: Random Forest regression and classification
- WEKA-J48: Weka C4.5 decision trees, only classification - use with bagging
- WEKA-RF: Random Forest, only classification
- XGBoost: Scalable and Flexible Gradient Boosting

Methods under development:

- N-Fold cross-validation
- Stratified cross-validation (classification only)

You can create a model from template: import an XML model template or use another model

Select the molecular descriptors

Recommended descriptor types:

- E-state
- ALogPS (2)
- GSFragment (1138)
- Dragon v. 7 (6270/3D)
- ISIDA fragments
- CDK 2.0 descriptors (306/3D)
- 'Inductive' descriptors (54/3D)
- MERA descriptors (529/3D)
- MERSY descriptors (42/3D)
- Chemaxon descriptors (499/3D)
- QNPR
- Spectrophores (144/3D)
- Structural alerts (ToxAlerts)

Predictions by OCHEM’s featured models

- Ames leavenberg
- Toxicity against T. Pyriformis
- ALogPS 3.0
- CYP1A2 Estate+ALogPS
- CYP2C9 Estate+ALogPS
- CYP2D6 Estate+ALogPS
- CYP3A4 Estate+ALogPS
- Pyrolysis point prediction (best Estate)
- Melting Point prediction (best Estate)
- Water solubility model based on logP and Melting Point
- ALogPS 2.1 logP
- ALogPS 2.1 logS
- Outputs of other OCHEM models

Obsolete/Additional descriptor types

- CDK 1.4.11 descriptors (274/3D)
- OESState
- Dragon v. 5.4 (1630/3D)
- Dragon v. 5.5 (3190/3D)
- Dragon v. 6 (485/3D)
- MOPAC 7.1 descriptors (25/3D)

Special descriptors (scaffolds, fingerprints):

- Chemaxon Scaffolds
- Silicon-II Scaffolds
- ECFP Fingerprints
- MolPrint Fingerprints

Conditions of experiments

- Test duration default value: 72h
- Target default value: Pseudomonas aeruginosa
- Material Nanoparticles of Elements default value: Silver
- APS default value: 10 nanometer
- Surface coating
- Exposure concentration
- Shape of nano particles default value: Spherical

Under development: can change anytime and backward compatibility is not guaranteed.
## Comprehensive Modeling

**Training set (required):** ALOGPS 3.01 [details]
Add a validation set

The model will predict these properties:
- logPow using unit: Log unit
- Aqueous Solubility using unit: log(mol/L)

Select the methods you want to use for the modeling:

### Method
- [ ] ANN
- [x] ASNN (bias correction)
- [ ] KNN
- [ ] LibSVM
- [x] FSMLR
- [x] MLRA
- [x] PLS
- [ ] WEKA-RF (classification only)
- [ ] WEKA-J48 (classification only)
- [ ] LSSVM (Least-Squares SVM)
- [ ] DNN (Deep Neural Network)
- [ ] DEEPCHIMM DAG
- [x] DEEPCHIMM GRAPH_CONV
- [ ] DEEPCHIMM TEXTCNN
- [ ] DEEPCHIMM WEAVE
- [ ] DEEPCHIMM MULTITASK
- [ ] DEEPCHIMM IRV (classification only)
- [ ] DEEPCHIMM ROBUST_MTNN (classification only)
- [ ] XGBOOST
- [ ] RFR
- [ ] CHEMCHAINER GGN
- [ ] CHEMCHAINER NFP
- [ ] NNF2N Neural Network Fingerprint
- [ ] MACAU (only for model with several properties)

### Descriptors
- [ ] CDK 2.0 (3D)
- [ ] Dragon v6 (all blocks; 3D)
- [ ] ALogPS, OEstage
- [ ] ISIDA Fragments (Length 2 - 4)
- [ ] GSFrag
- [ ] Mera and Merzy (3D)
- [ ] Chemaxon descriptors (3D)
- [ ] Inductive Descriptors (3D)
- [ ] Spectrophores (3D)
- [ ] QNPR (SMILES - length 1 - 3)
- [x] StructuralAlerts (EFG)
- [ ] SIRMS
- [ ] MW + # of carbons: (baseline model)
- [ ] PyDescriptor (3D)
- [ ] no descriptors (CHEMCHAINER, DEEPCHIMM, NNF)

### Descriptor selection
- [ ] Unsupervised forward selection
- [x] Pairwise de-correlation (R < 0.95)

+add a custom template

### Model validation
- [ ] 5-fold cross-validation
- [ ] 5-fold cross-validation (stratified - classification only)
- [ ] Bagging with 64 models
- [ ] Bagging with 64 models (stratified - classification only)

+add a custom template
Comprehensive View

Predicted property: Melting Point
Training set: `meltingpoint.xlsx` (2 different versions detected)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>RMSE - Root Mean Square Error</th>
<th>for</th>
<th>Training set</th>
<th>Validation:</th>
<th>All valid</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>DNN</th>
<th>GGNN (tr. set. 2)</th>
<th>NNF2N (tr. set. 2)</th>
<th>NNF2T (tr. set. 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2 (constitutional, topological, geometrical, electronic, ...)</td>
<td>39.5</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>ALogPS, OState</td>
<td>40.8</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Fragmentor (Length 2 - 4)</td>
<td>42.3</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SIRMS (LABELING = CHARGE;LOGP;HB;REFRACTIVITY noH (1-4))</td>
<td>43.4</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PyDescriptor (PyDescriptor)</td>
<td>41.6</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RDKit (blocks: 1-11 15-16)</td>
<td>40.6</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dragon6 (blocks: 1-29)</td>
<td>39</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dragon7 (blocks: 1-30)</td>
<td>39</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dragon6 (blocks: 15-19)</td>
<td>42.2</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>GSFrag (GSFrag GSFragL)</td>
<td>43.8</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Structural Alerts</td>
<td>44</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>SMILES</td>
<td></td>
<td></td>
<td>45.6</td>
<td>49.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consensus</th>
<th>Misc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.5</td>
<td>38</td>
</tr>
</tbody>
</table>

bigchem
275k Melting Point Datasets (Big Data)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bergström</td>
<td>277</td>
</tr>
<tr>
<td>Bradley</td>
<td>2886</td>
</tr>
<tr>
<td>OCHEM</td>
<td>22404</td>
</tr>
<tr>
<td>Enamine</td>
<td>21883</td>
</tr>
<tr>
<td>PATENTS</td>
<td>228079</td>
</tr>
</tbody>
</table>

COMBINED: OCHEM + Enamine + Bradley + Bergström

Tetko et al. *J. Chemoinformatics, 2016, 8, 2.*
Extraction of MP information from patents
## Modeling of MP data

<table>
<thead>
<tr>
<th>Package name</th>
<th>Type of descriptors</th>
<th>Number of descriptors</th>
<th>Matrix size, billions</th>
<th>Non zero values, millions</th>
<th>Sparseness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional Groups</td>
<td>integer</td>
<td>595</td>
<td>0.18</td>
<td>3.1</td>
<td>33</td>
</tr>
<tr>
<td>QNPR</td>
<td>integer</td>
<td>1502</td>
<td>0.45</td>
<td>6.3</td>
<td>49</td>
</tr>
<tr>
<td>MolPrint</td>
<td>binary</td>
<td>688634</td>
<td>205</td>
<td>8.1</td>
<td>7200</td>
</tr>
<tr>
<td>Estate count</td>
<td>float</td>
<td>631</td>
<td>0.19</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Inductive</td>
<td>float</td>
<td>54</td>
<td>0.02</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>ECFP4</td>
<td>binary</td>
<td>1024</td>
<td>0.31</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>Isida</td>
<td>integer</td>
<td>5886</td>
<td>1.75</td>
<td>18</td>
<td>37</td>
</tr>
<tr>
<td>ChemAxon</td>
<td>float</td>
<td>498</td>
<td>0.15</td>
<td>23</td>
<td>1.5</td>
</tr>
<tr>
<td>GSfrag</td>
<td>integer</td>
<td>1138</td>
<td>0.34</td>
<td>24</td>
<td>5.7</td>
</tr>
<tr>
<td>CDK</td>
<td>float</td>
<td>239</td>
<td>0.07</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Adriana</td>
<td>float</td>
<td>200</td>
<td>0.06</td>
<td>32</td>
<td>1.3</td>
</tr>
<tr>
<td>Mera, Mersy</td>
<td>float</td>
<td>571</td>
<td>0.17</td>
<td>61</td>
<td>1.1</td>
</tr>
<tr>
<td>Dragon</td>
<td>float</td>
<td>1647</td>
<td>0.49</td>
<td>183</td>
<td>1.5</td>
</tr>
</tbody>
</table>
Prediction and experimental errors for consensus model based on the PATENTS set

Experimental accuracy was based on $N = 18058$ duplicated measurements

$\sigma \approx 32^\circ C$
Prediction of Huuskonen set using ALOGPS logP and MP based on 230k measurements

$$\log S = 0.5 - 0.01(MP-25) - \log Kow \,*$$

**Predicted property:** Aqueous Solubility modeled in log(mol/L)

Training method: MLRA

<table>
<thead>
<tr>
<th>Data Set</th>
<th>#</th>
<th>R²</th>
<th>q²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set: logS set</td>
<td>1311 records</td>
<td>0.842 ± 0.009</td>
<td>0.83 ± 0.01</td>
<td>0.84 ± 0.02</td>
<td>0.64 ± 0.02</td>
</tr>
</tbody>
</table>

*Feature net model: uses other models as descriptors*
Transfer Learning

Inductive Transfer Learning
- Labeled data are available in a target domain
- Labeled data are available in a source domain

Self-taught Learning
- No labeled data in a source domain

Multi-task Learning
- Labeled data are available in a source domain
- Different domains but single task

Transductive Learning
- Different domains but single task
- Single domain and single task

Unsupervised Transfer Learning
- Labeled data are available only in a source domain
- No labeled data

Domain adaptation
- Sample selection bias

Multi-task learning
Problem:

- prediction of tissue-air partition coefficients
- small datasets 30-100 molecules (human & rat data)

Results:

simultaneous prediction of several properties increased the accuracy of models

Prediction of toxicity of chemical compounds: 
REGISTRY OF TOXIC EFFECTS OF CHEMICAL SUBSTANCES (RTECS®)

Different species
- Rat
- Mouse
- Rabbit
- ... 
- Human

~ 129k records
~ 87k compounds
29 properties

- Different toxicities
  - LD50
  - TDL
  - NOEL
  - LDLo

- Administration
  - Oral
  - IPR (intraperitoneal)
  - IVR (intravenous)

Comparison of different models to predict toxicity (RMSE)

<table>
<thead>
<tr>
<th></th>
<th>single</th>
<th>multi</th>
<th>single</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DNN</td>
<td>DNN(2)</td>
<td>XGBOOST</td>
</tr>
<tr>
<td>CDK2 (constitutional, topological, geometrical, electronic, ...)</td>
<td>0.9 0.56 1.33 0.474 0.56 1.1</td>
<td>0.76 0.47 1.22 0.472 0.51 0.93</td>
<td>0.8 0.47 1.29 0.454 0.5 1.02</td>
</tr>
<tr>
<td></td>
<td>0.478 0.477 0.66 1.05 0.623</td>
<td>0.471 0.459 0.54 0.96 0.576</td>
<td>0.439 0.56 1.04 0.584 0.75</td>
</tr>
<tr>
<td></td>
<td>0.78 0.68 0.7 0.63 0.99 0.724</td>
<td>0.68 0.59 0.591 0.47 0.91</td>
<td>0.65 0.59 0.95 0.66 1.33 0.91</td>
</tr>
<tr>
<td></td>
<td>1.41 0.63 0.86 1.1 0.85 1.31</td>
<td>0.577 1.25 0.581 0.66 1.02</td>
<td>0.75 1.08 0.764 1.3 0.67 0.81</td>
</tr>
<tr>
<td></td>
<td>0.72 0.85 1.01 0.8 0.66 1.27</td>
<td>0.69 1.21 0.65 0.66 0.76 0.63</td>
<td>0.76 0.63 1.2 (0.779)</td>
</tr>
<tr>
<td>(0.834)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dragon6 (blocks: 1-29)</td>
<td>0.89 0.58 1.3 0.458 0.56 1.06</td>
<td>0.78 0.44 1.31 0.445 0.474</td>
<td>0.8 0.49 1.3 0.454 0.523 1.01</td>
</tr>
<tr>
<td></td>
<td>0.481 0.472 0.6 1.06 0.63 0.74</td>
<td>0.96 0.461 0.446 0.52 1.055</td>
<td>0.439 0.59 1.02 0.588 0.73</td>
</tr>
<tr>
<td></td>
<td>0.66 0.686 0.63 0.97 0.69 1.32</td>
<td>0.68 0.55 0.581 0.47 0.95 0.57</td>
<td>0.66 0.602 0.94 0.67 1.33 0.91</td>
</tr>
<tr>
<td></td>
<td>0.622 0.82 1.09 0.83 1.33 0.76</td>
<td>1.31 0.574 0.65 1.08 0.68 1.2</td>
<td>0.76 1.09 0.77 1.38 0.68 0.82</td>
</tr>
<tr>
<td></td>
<td>0.83 0.98 0.8 0.7 1.24 (0.82)</td>
<td>0.68 0.67 0.74 0.64 0.59 1.22</td>
<td>0.74 0.63 1.24 (0.786)</td>
</tr>
<tr>
<td>ALogPS, OEstate</td>
<td>0.91 0.61 1.32 0.461 0.54 1.1</td>
<td>0.79 0.44 1.23 0.447 0.49 0.94</td>
<td>0.84 0.5 1.42 0.456 0.519 1.0</td>
</tr>
<tr>
<td></td>
<td>0.478 0.469 0.6 1.1 0.617 0.75</td>
<td>0.467 0.444 0.53 0.99 0.554</td>
<td>0.44 0.56 1.03 0.58 0.73</td>
</tr>
<tr>
<td></td>
<td>0.7 0.652 0.64 1 0.69 1.36</td>
<td>0.66 0.55 0.59 0.49 0.9 0.58</td>
<td>0.65 0.61 0.95 0.64 1.34 0.59</td>
</tr>
<tr>
<td></td>
<td>0.617 0.84 1.1 0.87 1.43 0.76</td>
<td>1.21 0.571 0.65 1.05 0.69 1.22</td>
<td>1.11 0.79 1.33 0.69 0.8 0.81</td>
</tr>
<tr>
<td></td>
<td>0.85 0.95 0.8 0.71 1.2 (0.832)</td>
<td>0.65 0.7 0.74 0.64 0.6 1.17</td>
<td>0.63 1.21 (0.786)</td>
</tr>
<tr>
<td>Fragmentor (Length 2 - 4)</td>
<td>0.96 0.61 1.43 0.463 0.542</td>
<td>0.73 0.45 1.25 0.44 0.48 0.95</td>
<td>0.78 0.45 1.38 0.447 0.52 1</td>
</tr>
<tr>
<td></td>
<td>1.14 0.491 0.484 0.62 1.1</td>
<td>0.465 0.448 0.502 0.99 0.554</td>
<td>0.476 0.436 0.58 1.09 0.592</td>
</tr>
<tr>
<td></td>
<td>0.647 0.81 0.71 0.71 0.64 1.04</td>
<td>0.65 0.55 0.56 0.46 0.92 0.575</td>
<td>0.61 0.67 0.59 0.94 0.67 1.31</td>
</tr>
<tr>
<td></td>
<td>0.74 1.38 0.643 0.79 1.14 0.86</td>
<td>1.28 0.564 0.63 1.07 0.69 1.24</td>
<td>0.77 1.14 0.79 1.43 0.69 0.83</td>
</tr>
<tr>
<td></td>
<td>1.33 0.82 0.86 0.94 0.84 0.66</td>
<td>0.7 0.66 0.73 0.63 0.62 1.2</td>
<td>0.77 0.64 1.29 (0.797)</td>
</tr>
<tr>
<td>1.22 (0.849)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Profile-like QSAR

Non-neural network approaches to multi-learning:
Least Squares Support Vector Regression (LSSVM)


**Chainer Chemistry (“ChemChainer”)**

- Chainer – one of popular frameworks for Deep Learning
- Algorithms provided by Chainer developers
- Can be installed using Python tools
- [https://github.com/pfnet-research/chainer-chemistry](https://github.com/pfnet-research/chainer-chemistry)
DEEPCHEM

- Based on TensorFlow (google)
- Available as part of Python (Anaconda) or as a Docker
- Supports multiple MTL and STL approaches
- https://github.com/deepchem/deepchem

### Summary of “readily” available methods

<table>
<thead>
<tr>
<th>Package</th>
<th>Examples of supported algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chainer Chemistry</td>
<td>NFP, GGNN, RSGCN, WeaveNet, SchNet</td>
</tr>
<tr>
<td>DeepChem</td>
<td>DAG, NNF, MPNN, TEXTCNN, WEAVE, IRV</td>
</tr>
<tr>
<td>OCHEM</td>
<td>Above methods + DNN, LSSVM, Macau, feature net as well as use of tasks classes as descriptors</td>
</tr>
</tbody>
</table>

NFP/NNF - Neural Fingerprint; GGNN - Gated Graph Neural Network; MPNN - Message Passing Neural Networks; SchNet - continuous-filter convolutional neural network for modeling quantum interactions; DAG - Directed Acyclic Graphs; IRV - Influence Relevance Voters; LSSVM – Least Squares Support Vector Machines
## Comparison of MTL and STL

### Multiple models overview

**Predicted property:** Cblood/Cair(Human)  
**Training set:** tissue/air set

<table>
<thead>
<tr>
<th>Metrics</th>
<th>RMSE - Root Mean Square Error for Training set</th>
<th>Validation: Cross-Validation (16 models)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2 (constitutional, topological, geometrical, electronic,...)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASNN</td>
<td>MTL</td>
<td>DNN</td>
</tr>
<tr>
<td>0.45 0.28 0.21 0.29 0.39 0.33 0.28 0.32 0.4 0.33 0.4 (0.335)</td>
<td>0.54 0.33 0.38 0.35 0.4 0.45 0.321 0.43 0.44 0.49 0.52 (0.423)</td>
<td>0.41 0.41 0.45 0.42 0.44 0.56 0.279 0.5 0.39 0.37 0.44 (0.424)</td>
</tr>
<tr>
<td>0.44 0.35 0.31 0.33 0.4 0.44 0.32 0.33 0.33 0.31 0.36 (0.356)</td>
<td>0.42 0.29 0.31 0.32 0.38 0.41 0.31 0.33 0.41 0.37 0.4 (0.359)</td>
<td>0.41 0.47 0.44 0.51 0.66 0.6 0.37 0.57 0.5 0.39 0.48 (0.491)</td>
</tr>
<tr>
<td>OEstate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DAG</th>
<th>GRAPH_CONV</th>
<th>TEXTCNN</th>
<th>WEAVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTL</td>
<td>0.75 0.55 0.6 0.35 0.94 0.67 0.44 0.64 0.58 0.57 0.92 (0.637)</td>
<td>0.93 0.64 0.8 0.58 1 1 0.6 0.79 0.85 0.89 0.8 (0.807)</td>
<td>0.53 0.4 0.43 0.33 0.48 0.53 0.35 0.53 0.47 0.48 0.5 (0.457)</td>
<td>0.7 0.69 0.8 0.61 0.9 0.64 0.41 0.74 0.57 0.61 0.7 (0.67)</td>
</tr>
<tr>
<td>STL</td>
<td>0.63 0.52 0.9 0.47 1.1 1 0.38 0.8 0.62 0.62 1 (0.731)</td>
<td>0.8 0.61 0.9 0.7 0.9 0.78 0.65 0.8 0.86 0.92 0.9 (0.802)</td>
<td>0.58 0.54 0.57 0.51 0.7 0.63 0.39 0.66 0.51 0.62 0.48 (0.563)</td>
<td>0.62 0.52 0.7 0.59 0.8 1.1 0.48 0.71 0.72 0.72 0.8 (0.705)</td>
</tr>
</tbody>
</table>
big data in chemistry + informatics = chemoinformatics

The increasing volume of biomedical data in chemistry and life sciences requires development of new methods and approaches for their analysis.

The BIGCHEM project will provide innovative education in large chemical data analysis. The innovative research program will be implemented with the target users, large pharma companies and SMEs, which generate and analyze large chemical data as well as will promote technology transfer from academy to industrial applications.

Marie Skłodowska-Curie European Industrial Doctorate (EID)
Application of Generative Autoencoder in de Novo Molecular Design

Thomas Blaschke,*[a, b] Marcus Olivecrona,[a] Ola Engkvist,[a] Jürgen Bajorath,[b] and Hongming Chen*[a]

Abstract: A major challenge in computational chemistry is the generation of novel molecular structures with desirable pharmacological and physicochemical properties. In this work, we investigate the potential use of autoencoder, a deep learning methodology, for de novo molecular design. Various generative autoencoders were used to map molecule structures into a continuous latent space and vice versa and their performance as structure generator was assessed. Our results show that the latent space preserves chemical similarity principle and thus can be used for the generation of analogue structures. Furthermore, the latent space created by autoencoders were searched systematically to generate novel compounds with predicted activity against dopamine receptor type 2 and compounds similar to known active compounds not included in the trainings set were identified.

Keywords: Autoencoder • chemoinformatics • de novo molecular design • deep learning • inverse QSAR
Summary

• OCHEM is powerful extendable platform for data storage
• Works with millions of datapoints
• Provide an integrated support of various (multi-learning) algorithms
• Very useful for ADMETox and (Q)SAR studies
Acknowledgements

Pavel Karpov
Dipan Ghosh
Michael Withnall
Allison Keys
Zhonghua Xia
Barbara Gasset
Monica Campillos

Guillaume Godin
Sergey Sosnin
Daniel Lowe
Maxim Fedorov
Anthony Williams

Yura Sushko
Sergey Novotarskyi
Robert Körner

Michael Sattler