Early ADME/Tox predictions: toy or tool?

Igor V. Tetko

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)
Institute of Bioinformatics & Systems Biology

Obernai, France, 21 June 2010
Helmholtz Zentrum München statistics

- Previous name (before 2008): GSF (Forschungszentrum für Umwelt und Gesundheit GmbH)
- Part of Helmholtz Network (2.35 Milliards Euro, 26500 people, 15 centers)
- Leading center for Environmental Health in Germany
- 25 institutes (1797 people, ca 700 scientists & 300 PhD students)*
- 70 contracts with EU
- Strong IPR and management support

- Institute for Bioinformatics & Systems Biology
  - 50 peoples, strong expertise in \textit{in silico} data analysis, machine learning methods, software development, data dissemination (Web, Internet)

*January 2008
Layout of presentation

Productivity of R&D companies
Importance of ADMETox parameters
Overview of eADMETox properties/data
Applicability Domain challenges
  • LogP benchmarking study
  • AD for qualitative models – AMES test
Data integration
OCHEM – On-line CHEmical database & Modeling environment
Conclusions
Pharma R&D Cost and Productivity: Fewer drugs, more expenditure

Dwindling R&D Productivity in the Pharmaceutical Industry

2008 – 24(3*); 2009 – 25(6*)

*Biological license applications
Potential ADME/T market (US $ billions)$^1$

- In vitro Toxicology ($0.2$)
- In vivo Toxicology ($1.3$)
- ADME ($1.5$)

It will grow up to US$ 4.4 billion up to 2012$^2$

2) http://www.researchandmarkets.com/reports/c84850
Pharma R&D: Cost and Productivity issues
Compound numbers

1,000,000 Cpd $\leq$ 5000 Cpd $\leq$ 500 Cpd $\leq$ 5 Cpd

Up to 15 Years:
1 drug

In vitro and in vivo ADME/T property determination:
Millions of screens for solubility, stability, absorption, metabolism, transport, reactive products, drug interactions, etc etc

Preclinics Costs: > $300m PER COMPOUND to reach approval
ADME/T

Absorption
enters organism (by oral administration)

Distribution
distributed between blood/plasma/tissues (e.g. brain)

Metabolism
bio-converting

Elimination
mechanisms and pathways for excretion of drugs

Toxicity
undesired interactions of drug or its metabolites

Size, lipophilicity, solubility, ionization, permeability, active transport

Affinity to different tissues, permeability, active transport

Affinity to different enzymes

Active transport, size, lipophilicity, ionization, permeability (also for metabolites)

Presence of toxicological pharmacophores, lipophilicity
Interplay of physico-chemical properties with *in vivo* pharmacological activities/data

Wang & Shkolnik, Chem. & Biodiversity, 2009, 6, 1887.
Interest in Phys-Chem properties

Wang & Shkolnik, Chem. & Biodiversity, 2009, 6, 1887.
Number of molecules processed at the Abbot site through the various algorithms available on the property calculation web page

# Properties Used to Define Drug-Likeness

<table>
<thead>
<tr>
<th>Property</th>
<th>Drugs</th>
<th>CNS-Drugs</th>
<th>Leads</th>
<th>Fragments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MW</td>
<td>&lt;500</td>
<td>&lt;450</td>
<td>&lt;400</td>
<td>&lt;300</td>
</tr>
<tr>
<td>logP</td>
<td>&lt;5</td>
<td>0-4</td>
<td>&lt;4</td>
<td>&lt;3</td>
</tr>
<tr>
<td>HA</td>
<td>&lt;10</td>
<td></td>
<td>&lt;8</td>
<td>&lt;3</td>
</tr>
<tr>
<td>HD</td>
<td>&lt;5</td>
<td></td>
<td>&lt;4</td>
<td>&lt;3</td>
</tr>
<tr>
<td>logD\textsubscript{7.4}</td>
<td>1-3</td>
<td>1-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSA</td>
<td>&lt;140</td>
<td>&lt;80</td>
<td>&lt;120</td>
<td>&lt;90</td>
</tr>
</tbody>
</table>

### Profiling of chemical compounds (Optibrium Ltd)

<table>
<thead>
<tr>
<th>Property</th>
<th>Desired Value</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>logS</td>
<td>&gt; 1</td>
<td></td>
</tr>
<tr>
<td>HIA category</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>logP</td>
<td>&lt;= 3.5</td>
<td></td>
</tr>
<tr>
<td>P-gp category</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>PPB category</td>
<td>low</td>
<td></td>
</tr>
<tr>
<td>2C9 pKi</td>
<td>&lt;= 6</td>
<td></td>
</tr>
<tr>
<td>2D6 affinity category</td>
<td>low medium</td>
<td></td>
</tr>
<tr>
<td>hERG pIC50</td>
<td>&lt;= 5</td>
<td></td>
</tr>
<tr>
<td>BBB log([brain]:[blood])</td>
<td>&lt;= -0.5</td>
<td></td>
</tr>
<tr>
<td>BBB category</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Profiling of chemical compounds (Optibrium Ltd)

ADMET parameters

**Absorption**
Caco-2 (colorectal carcinoma cells)
MDCK (Madin-Darby Canin Kidney)
PAMPA (parallel artificial membrane permeability assay)
Human Intestinal Absorption
%FA (% of absorbed drug mass)
Active transporters (P-glycoprotein, multi drug resistance protein – efflux, peptide, amino-acid transporters – absorption)

**Metabolism**
CYP450
Aldehyde/Alcohol dehydrogenases
Hydrolases, oxidases, esterases
Microsomes, hepatocytes
Genetically modified cell lines to express single CYP450

**Distribution**
BBB (Blood-Brain Barrier)
PPB (Plasma protein Binding)
HSA (Human Serine Albumin) binding
Tissue partitioning
Volume of distribution=dose/C₀ (C₀ is initial concentration of a drug in plasma)

**Elimination**
Route of elimination (renal, liver (bile->faeces, metabolism)

**Toxicity**
AMES test
DDI (inhibition/activation of CYP450)
hERG (human ether-a-go-go-related gene) potassium channel inhibition
Toxicity alerts
Physico-chemical properties

In vitro:
logP
logD
Solubility in water
Solubility in DMSO
pKa
Solubility in simulated intestinal fluid
Absorption:

In vitro:
CaCo-2
MDCK
PAMPA

In vivo:
%FA
Distribution:

In vitro:  In vivo:
PPB          BBB – animal models
HSA          Tissue partitioning
BBB
Metabolism:

In vitro:
CYP450
Genetically engineered cell lines to study individual CYP
Microsomes
Hepatocytes

In vivo:
MS analysis
Toxicity

In vitro:
AMES mutagenicity
hERG toxicity

In vivo:
Animal models (LD50)
### ADMETTox properties

#### Physico-chemical

- **Lipophilicity (logP/logD)**
  - $\sim 20k$ ($>100k$)

- **Aqueous solubility**
  - $\sim 10k$ ($\sim 100k$)

- **pKa**
  - $\sim 10k$ ($\sim 100k$)

- **Solubility in DMSO**
  - $\sim 1k$ ($>100k$)

#### Biological properties

- **%FA (Fraction Absorbed)** $\sim 1k$
- **Blood-Brain barrier** $\sim 1k$
- **CYP450 affinities** $\sim 10k$
- **Transporters (PgP)** $\sim 10k$
- **Ion channels (hERG)** $\sim 10k$
- **Microsomes** $\sim 100k$
- **Hepatocytes** $\sim 10k$
- **VD** $\sim 300$

Available sources: WOMBAT, Symyx, CHeMBL, PHYSPROP, ChemSpider, OCHEM
Additional readings:


What are the goals of modeling?

Decrease number of experimental measurements by substitution of them with computational predictions.

This can be achieved when computational accuracy of models is similar (or better!) to that of experimental measurements.

Can we achieve it?
"One can not embrace the unembraceable."

Possible: $10^{60} - 10^{100}$ molecules theoretically exist

Achievable: $10^{20} - 10^{24}$ can be synthesized now by companies (weight of the Moon is ca $10^{23}$ kg)

Available: $2 \times 10^7$ molecules are on the market

Measured: $10^2 - 10^5$ molecules with ADME/T data

Problem: To predict ADME/T properties of just molecules on the market we must extrapolate data from one to 1,000 - 100,000 molecules!

There is a need for methods which can estimate the accuracy of predictions!
Current dogma about prediction of physico-chemical properties

- Prediction of physico-chemical properties, in particular $\log P$, is simple
- There is no need to measure them
- We have enough number of good computational methods

Is this true?
Statistics of logP benchmarking

30 (18) methods - major commercial providers and public software

Public dataset:
$N=266$ molecules

_in house_ data:
$N=95809$ molecules from Prizer
$N=889$ molecules from Nycomed

**Arithmetic Average Model (AAM):**
mean log$P$ was used as a prediction (one value for all molecules)

**Rank III:** models with errors ($RMSE \geq AAM$), i.e. non-predictive
**Rank I:** models with $RMSE$ identical or close to the best method
**Rank II:** remaining models

### Performance of algorithms for the public dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Star set (N = 223)</th>
<th>Non-Star set (N = 43)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>rank</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rank</td>
</tr>
<tr>
<td>AB/LogP</td>
<td>0.41</td>
<td>I</td>
</tr>
<tr>
<td>S+logP</td>
<td>0.45</td>
<td>II</td>
</tr>
<tr>
<td>ACD/LogP</td>
<td>0.50</td>
<td>I</td>
</tr>
<tr>
<td>Consensus log P</td>
<td>0.50</td>
<td>I</td>
</tr>
<tr>
<td><strong>CLOGP</strong></td>
<td>0.52</td>
<td>II</td>
</tr>
<tr>
<td><strong>VLOGP</strong></td>
<td>0.52</td>
<td>II</td>
</tr>
<tr>
<td>ALOGPS</td>
<td>0.53</td>
<td>II</td>
</tr>
<tr>
<td>MiLogP</td>
<td>0.57</td>
<td>II</td>
</tr>
<tr>
<td>XLOGP</td>
<td>0.62</td>
<td>II</td>
</tr>
<tr>
<td>KowWIN</td>
<td>0.64</td>
<td>II</td>
</tr>
<tr>
<td>CSlogP</td>
<td>0.65</td>
<td>II</td>
</tr>
<tr>
<td>ALOGP (Dragon)</td>
<td>0.69</td>
<td>II</td>
</tr>
<tr>
<td>MolLogP</td>
<td>0.69</td>
<td>II</td>
</tr>
<tr>
<td>ALOGP98</td>
<td>0.70</td>
<td>II</td>
</tr>
<tr>
<td>OsirisP</td>
<td>0.71</td>
<td>II</td>
</tr>
<tr>
<td>VLOGP</td>
<td>0.72</td>
<td>II</td>
</tr>
<tr>
<td>TLOGP</td>
<td>0.74</td>
<td>II</td>
</tr>
<tr>
<td>ABSOLV</td>
<td>0.75</td>
<td>II</td>
</tr>
<tr>
<td>QikProp</td>
<td>0.77</td>
<td>II</td>
</tr>
<tr>
<td>QuantlogP</td>
<td>0.80</td>
<td>II</td>
</tr>
<tr>
<td>SLIPPER-2002</td>
<td>0.80</td>
<td>II</td>
</tr>
<tr>
<td>COSMOFrag</td>
<td>0.84</td>
<td>II</td>
</tr>
<tr>
<td>XLOGP2</td>
<td>0.87</td>
<td>II</td>
</tr>
<tr>
<td>QLOGP</td>
<td>0.96</td>
<td>II</td>
</tr>
<tr>
<td>VEGA</td>
<td>1.04</td>
<td>II</td>
</tr>
<tr>
<td>CLIP</td>
<td>1.05</td>
<td>II</td>
</tr>
<tr>
<td>LSER</td>
<td>1.07</td>
<td>II</td>
</tr>
<tr>
<td>MLOGP (Sim+)</td>
<td>1.26</td>
<td>II</td>
</tr>
<tr>
<td><strong>NC+NHET</strong></td>
<td>1.35</td>
<td>III</td>
</tr>
<tr>
<td>SPARC</td>
<td>1.36</td>
<td>III</td>
</tr>
<tr>
<td>MLOGP(Dragon)</td>
<td>1.52</td>
<td>III</td>
</tr>
<tr>
<td>LSER UFZ</td>
<td>1.60</td>
<td>III</td>
</tr>
<tr>
<td><strong>AAM</strong></td>
<td>1.62</td>
<td>III</td>
</tr>
<tr>
<td>VLOGP-NOPS</td>
<td>1.76</td>
<td>III</td>
</tr>
<tr>
<td>HINT</td>
<td>1.80</td>
<td>III</td>
</tr>
<tr>
<td>GBLOGP</td>
<td>1.98</td>
<td>III</td>
</tr>
</tbody>
</table>

**AAM = base ("no model") model, \(R^2=0\), it used just one logP value as predicted value for all 95809 or 882 molecules, respectively.**
Our methodology is top-ranked in a recent benchmarking

- Benchmarking was done by Pfizer and Nycomed – no data were available to participants
- Our ALOGPS algorithm was top-ranked (according to the lowest RMSE errors)
- Several methods performed worse than making no prediction,

AAM base (“no model”) model, $R^2 = 0$, it used just one logP value as predicted value for all 95809 or 882 molecules, respectively.

ALOGPS decreases errors about twice using local corrections for $N=95809$ *in house* Pfizer molecules.

![ALOGPS Blind prediction vs ALLOGPS LIBRARY graphs](image)

RMSE=1.02 \[\rightarrow\] RMSE=0.59

ca 30 minutes of calculations on a notebook!

The descriptor space challenge

We need to know the target property and select correct descriptors!
Property-based space similarity illustration

*Do they agree in their votes (STD)*?
*Do they have the same pattern of votes (CORREL)*?
Associative Neural Network Property-Based DMs

logP=3.11

logP=3.48

Morphinan-3-ol, 17-methyl-

Levallorphan

CORREL - correlation between vectors of predictions

STD - standard deviation of ensemble predictions

Illustration of local correction using nearest neighbors

Real activity

Predicted activity
ALOGPS decreases errors about twice using local corrections for \( N=95809 \) \textit{in house} Pfizer molecules

ca 30 minutes of calculations on a notebook!

Illustration of local correction using nearest neighbors

Real activity

Predicted activity
ALOGPS distinguishes reliable vs. non-reliable predictions in property-based space (CORREL)

CORREL identifies 60% of molecules predicted with average accuracy of 0.3 log units

The use of ALOGPS advanced features dramatically increase prediction accuracy of the predictions. The experimental measurements accuracy was achieved for >60,000 Pfizer compounds.

Estimation of toxicity against *T. pyriformis*

The overall goal is to predict and to assess the reliability of predictions toxicity against *T. pyriformis* for chemicals directly from their structure.

Dataset: 1093 molecules

Analyzed QSARs (Quantitative Structure Activity Relationship) and distances to models (DM)

<table>
<thead>
<tr>
<th>country</th>
<th>modeling techniques</th>
<th>descriptors</th>
<th>abbreviation</th>
<th>distances to models (in space)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ensemble of 192 kNN models</td>
<td>MolconnZ</td>
<td>kNN-MZ</td>
<td>EUCLID, STD</td>
</tr>
<tr>
<td></td>
<td>ensemble of 542 kNN models</td>
<td>Dragon</td>
<td>kNN-DR</td>
<td>EUCLID, STD</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>MolconnZ</td>
<td>SVM-MZ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>Dragon</td>
<td>SVM-DR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>Fragments</td>
<td>SVM-FR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>Fragments</td>
<td>kNN-FR</td>
<td>EUCLID, TANIMOTO</td>
</tr>
<tr>
<td></td>
<td>MLR</td>
<td>Fragments</td>
<td>MLR-FR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLR</td>
<td>Molec. properties (CODESSA-Pro)</td>
<td>MLR-COD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OLS</td>
<td>Dragon</td>
<td>OLS-DR</td>
<td>LEVERAGE</td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>Dragon</td>
<td>PLS-DR</td>
<td>LEVERAGE, PLSEU</td>
</tr>
<tr>
<td></td>
<td>ensemble of 100 neural networks</td>
<td>E-state indices</td>
<td>ASNN-ESTATE</td>
<td>CORREL, STD</td>
</tr>
<tr>
<td></td>
<td>All consensus model</td>
<td>-</td>
<td>CONS</td>
<td>STD</td>
</tr>
</tbody>
</table>

# Overview of analyzed distances to models (DMs)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EUCLID</strong></td>
<td>$EU_m = \frac{\sum_{i=1}^{k} d_j}{k}$</td>
<td>$k$ is number of nearest neighbors, $m$ index of model</td>
</tr>
<tr>
<td></td>
<td>$EUCLID = E\bar{U}_m$</td>
<td></td>
</tr>
<tr>
<td><strong>TANIMOTO</strong></td>
<td>$Tanimoto(a,b) = \sum x_{a,i}x_{b,i} + \sum x_{b,i}^2 - \sum x_{a,i}^2$</td>
<td>$x_{a,i}$ and $x_{b,i}$ are fragment counts</td>
</tr>
<tr>
<td><strong>LEVERAGE</strong></td>
<td>$LEVERAGE=x^T(X^TX)^{-1}x$</td>
<td></td>
</tr>
<tr>
<td><strong>PLSEU</strong> (DModX)</td>
<td>Error in approximation (restoration) of the vector of input variables from the latent variables and PLS weights.</td>
<td></td>
</tr>
<tr>
<td><strong>STD</strong></td>
<td>$STD = \frac{1}{N-1} \sum (y_i - \bar{y})^2$</td>
<td>$y_i$ is value calculated with model $i$ and $\bar{y}$ is average value</td>
</tr>
<tr>
<td><strong>CORREL</strong></td>
<td>$CORREL(a) = \max_j CORREL(a,j)=R^2(Y_u^\text{calc}, Y_j^\text{calc})$</td>
<td>$Y^a=(y_1,\ldots,y_N)$ is vector of predictions of molecule $i$</td>
</tr>
</tbody>
</table>
Property-based, ASNN model: DM does work!

STD

Descriptor space, ASNN model: DM does not work

## Ranking of Distance to Models (DM)

<table>
<thead>
<tr>
<th>DM</th>
<th>average rank</th>
<th>highest rank$^1$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOO</td>
<td>5-CV</td>
<td>Valid.*</td>
<td>LOO</td>
<td>5-CV</td>
</tr>
<tr>
<td>STD-CONS</td>
<td>1</td>
<td>1.8</td>
<td><strong>1.1</strong></td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>STD-ASNN</td>
<td>2</td>
<td>1.2</td>
<td><strong>2.5</strong></td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>STD-kNN-DR</td>
<td>6.6</td>
<td>4.3</td>
<td><strong>4.1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STD-kNN-MZ</td>
<td>9.2</td>
<td>8.3</td>
<td><strong>5.3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-DR</td>
<td>7.1</td>
<td>4.9</td>
<td><strong>5.4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEVERAGE-PLS</td>
<td>8.4</td>
<td>5</td>
<td><strong>6.3</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-MZ</td>
<td>7.5</td>
<td>7.1</td>
<td><strong>6.4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TANIMOTO-kNN-FR</td>
<td>7</td>
<td>6.1</td>
<td><strong>6.8</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TANIMOTO-MLR-FR</td>
<td>8.3</td>
<td>8.3</td>
<td><strong>9</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORREL-ASNN</td>
<td>10.7</td>
<td>10.8</td>
<td><strong>9.4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEVERAGE-OLS-DR</td>
<td>12.3</td>
<td>12.6</td>
<td><strong>11.1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUCLID-MLR-FR</td>
<td>7</td>
<td>9.3</td>
<td><strong>11.5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLSEU-PLS</td>
<td>11.1</td>
<td>11.8</td>
<td><strong>11.5</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-FR</td>
<td>12.1</td>
<td>13.3</td>
<td><strong>12.1</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Ordered by performance of the DMs on the validation dataset

---

Analysis of DMs for a linear model

Log(IGC_{50}^{-1}) =
-18(\pm 0.7) + 0.065(\pm 0.002) \text{AMR}-0.50
(0.04) O_{56}-0.30(0.03) O_{58}
-0.29(0.02) nH_{\text{Acc}}+0.046(0.005)
 H-046+16(0.7) \text{Me}

The results of DM performance are consistent across different models

Classification task distance measures

A. CLASS-LAG

B. CONS-STD

C. PROB-CONS-STD

Helmholtz Zentrum München
German Research Center for Environmental Health
Binary classification

\[ d_{\text{PROB-STD}}(x) = \begin{cases} \int_{x}^{+\infty} N(q, p(x), d_{\text{STD}}(x)) \text{ if } c(x) = 1 \\ \int_{-\infty}^{0} N(q, p(x), d_{\text{STD}}(x)) \text{ if } c(x) = 0 \end{cases} \]
Prediction of Ames Mutagenicity set

http://ml.cs.tu-berlin.de/toxbenchmark
Toxicity against *Salmonella typhimurium*

Training dataset: 4361 molecules
“Blind” test dataset: 2181 molecules
54% with mutagenic effect

Large international collaboration effort of
>10 labs from USA, Canada, EU, Russia,
the Ukraine & China (see also poster P-22)

Prof. Bruce N. Ames
Inventor of the test (1975)

Accuracy of a AMES consensus model as function of two Distances to Models

Accuracy = CORRECT / ALL
Averaged ranking of DMs according to the percentage of compounds with 90% accuracy for training and test sets.

<table>
<thead>
<tr>
<th>Distance to model</th>
<th>Average rank - training set</th>
<th>Average rank – test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONS-STD-QUAL-PROB</td>
<td>2.17</td>
<td>1.83</td>
</tr>
<tr>
<td>CONCORDANCE</td>
<td>1.62</td>
<td>2.1</td>
</tr>
<tr>
<td>CONS-STD-PROB</td>
<td>3.43</td>
<td>3.05</td>
</tr>
<tr>
<td>CONS-STD-QUAL</td>
<td>3.67</td>
<td>4.9</td>
</tr>
<tr>
<td>ASNN-STD-PROB</td>
<td>6.52</td>
<td>5.48</td>
</tr>
<tr>
<td>CONS-STD</td>
<td>4.83</td>
<td>5.6</td>
</tr>
<tr>
<td>CLASS-LAG</td>
<td>7.1</td>
<td>6.24</td>
</tr>
<tr>
<td>ASNN-STD</td>
<td>8.14</td>
<td>7.67</td>
</tr>
<tr>
<td>AD_MEAN1*</td>
<td>10.71</td>
<td>9.07</td>
</tr>
<tr>
<td>CORREL</td>
<td>9.26</td>
<td>10.26</td>
</tr>
<tr>
<td>AD_MEAN2*</td>
<td>9.71</td>
<td>10.86</td>
</tr>
<tr>
<td>LEVERAGE*</td>
<td>10.83</td>
<td>10.95</td>
</tr>
</tbody>
</table>

CONCORDANCE is the number of models that give the same prediction, as the current model does.
Accuracy of different AMES model as function of a Distance to Models
Multi-task learning
Multi-task learning: unequal number of data

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

Multi-task learning can improve models for small sets

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

ADMETox *in silico* challenges

ADMETox models should allow navigation in space of molecules with a confidence and:

- should reliably estimate which compounds can/can’t be reliably predicted.
- provide experimental design and to minimize costs of new measurements.
- be easily interpretable for chemists
Online CHEmical Modeling environment (OCHEM)

http://ochem.eu
Motivation

Properties of molecules

- Data are lost after publication of an article
- The original sources of data are difficult to track
- The conditions of experiments are frequently not provided
- The conversion between different units is error prone
- Current databases do not allow community correction of errors
- The tracking of changes (by users) is required

Models

- Most published models are never used
- Implementation can be as difficult as new model development
- Different implementations can produce different results*
Database schema
Simplified overview

Evidences

Properties

Conditions

Units

Molecules

Names

Articles

Users

log((GC50-1) = 2.02 -log (mmol/L)  Temperature = 25.0
Zhu, H
Combinatorial QSAR modeling of chemical toxicants tested again...
N: 445
Journal of chemical information and modeling 2008: 48 (4) 766-84
2579-22-8, phenylpropargyl aldehyde
midnight / letoko

log(GC50-1) (concentration) 1093 records
LogPsuv (dimensionless) 21 records
LogPsuv(ion) (dimensionless) 21 records
LogPi (dimensionless) 35 records

Species (dimensionless)
Temperature (temperature)
Dose (concentration)
Concentration (concentration)

log(mmol/L) (concentration)
-log(mg/l) (concentration)
nM (concentration)
-log (mmol/L) (concentration)

Helmholtz Zentrum München
German Research Center for Environmental Health

Helmholtz Association
Data structure: behind the scene
Acknowledgements

My group
Iurii Sushko
Sergii Novotarskyi
Anil K. Pandey
Robert Körner
Stefan Brandmaier
Matthias Rupp
Vijyant Srivastava
Wolfram Teetz

Funding
GO-Bio BMBF http://qspr.eu
Germany-Ukraine grant UKR 08/006
FP7 Marie Curie ITN ECO http://eco-itn.eu
FP7 CADASTER http://www.cadaster.eu

Thanks to:
Dr. G. Poda (Pfizer)
Dr. C. Ostermann (Nycomed)
Dr. C. Höfer (DMPKore)
Prof. A. Tropsha (NC, USA)
Prof. T. Oprea (New Mexico, USA)
Prof. A. Varnek (Strasbourg, France)
Prof. R. Mannhold (Düsseldorf, Germany)
Prof. R. Todeschini (Milano, Italy)
+all coauthors of AMES
+many other colleagues
Interested in Short-Term stays (3-12 months)? Apply!!

http://www.eco-itn.eu