Recent Cheminformatics development at NCCT applied to ER, AR and physicochemical properties of chemicals

Kamel Mansouri

National Center for Computational Toxicology, U.S. EPA, RTP, NC, USA

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA
Problem Statement

Too many chemicals to test with standard animal-based methods

- Cost (~$1,000,000/chemical), time, animal welfare
- 10,000 chemicals to be tested for EDSP
- Fill the data gaps and bridge the lack of knowledge

Alternative

(Q)SAR
= (Quantitative) Structure-Activity Relationship

IN SILICO
Recent Cheminformatics development at NCCT

• We are building a new cheminformatics architecture
• PUBLIC dashboard gives access to curated chemistry
• Focus on integrating EPA and external resources
• Aggregating and curating data, visualization elements and “services” to underpin other efforts
  • RapidTox
  • Read-across
  • Predictive modeling
  • Non-targeted screening
**Quantitative Structure Activity/Property Relationships (QSAR/QSPR)**

*Congenericity principle*: QSARs correlate, within congeneric series of compounds, their chemical or biological activities, either with certain structural features or with atomic, group or molecular descriptors.


\[ Y = f(b_i, X) \]

- **X** - *descriptors (selected variables)*
- **b_i** - *fitted parameters*
QSARs validity, reliability, applicability and adequacy for regulatory purposes

The conditions for the validity of QSARs

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) A defined endpoint</td>
<td>Any <strong>physicochemical, biological or environmental</strong> effect that can be measured and therefore modelled.</td>
</tr>
<tr>
<td>2) An unambiguous algorithm</td>
<td><strong>Ensure transparency</strong> in the description of the model algorithm.</td>
</tr>
<tr>
<td>3) A defined domain of applicability</td>
<td><strong>Define limitations</strong> in terms of the types of <strong>chemical structures</strong>, physicochemical properties and mechanisms of action for which the models can generate <strong>reliable predictions</strong>.</td>
</tr>
</tbody>
</table>
| 4) Appropriate measures of goodness-of-fit, robustness and predictivity | a) The internal **fitting** performance of a model  
   b) the **predictivity** of a model, determined by using an appropriate **external test set**. |
| 5) Mechanistic interpretation, if possible                     | Mechanistic **associations** between the **descriptors** used in a model and the **endpoint being predicted**.                            |
Development of a QSAR model

- Curation of the data
  » Flagged and curated files available for sharing
- Preparation of training and test sets
  » Inserted as a field in SDFiles and csv data files
- Calculation of an initial set of descriptors
  » PaDEL 2D descriptors and fingerprints generated and shared
- Selection of a mathematical method
  » Several approaches tested: KNN, PLS, SVM…
- Variable selection technique
  » Genetic algorithm
- Validation of the model’s predictive ability
  » 5-fold cross validation and external test set
- Define the Applicability Domain
  » Local (nearest neighbors) and global (leverage) approaches
Public domain data sources

[Images of various public domain data sources websites such as ChemSpider, chemexper.com, ChemIDplus Advanced, ECOTOX Database, PubChem, and OpenTox.]
Structure curation procedure

Remove inorganics and mixtures
Clean salts and counterions
Normalize of tautomers
Remove of duplicates
Final inspection

Aim of the workflow:
• Combine different procedures and ideas
• Minimize the differences between the structures used for prediction
• Produce a flexible free and open source workflow to be shared

QSAR-ready structures

KNIME workflow

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267
Molecular structures in the computer

Fragmental keys & fingerprints
- substructural search
- read-across
- similarity search
Classification methods

- **kNN: k Nearest Neighbors**
  - classification according to the majority class of the $k$ neighbors

- **SVM: Support Vector Machines**
  - Kernel function maximizing the margin between the classes

Other methods: Self organized maps (SOM), Kohonen maps, PLSDA, LDA
Regression methods

- **MLR**: Multiple Linear Regression
  \[ \hat{y} = bX \]
  \[ b = (X'X)^{-1}X'y \]

- **PLS**: Partial Least Squares
  \[ X = TP' + E \]
  \[ Y = UQ' + F \]

PLS is the vector on the PCR ellipse upon which MLR has the longest projection.

Other methods: Artificial Neural Networks (ANN), Random Forest, LASSO, PCR…
Variable selection procedure

- Many more descriptors than chemicals
- Many irrelevant descriptors

Only the most important descriptors are selected

Create initial descriptor population

Evaluate fitness of the populations

Select and reproduce (Crossover, Mutation)

MLR (Multiple Linear Regression)
PLS (Partial Least squares)
SVM (Support Vector Machines)

Replace the descriptors of old populations with new descriptors

Stopping criteria

Final models

The Genetic Algorithms diagram
Cross-validation and test-set to avoid the “by chance” correlation problem

“There is a concern in West Germany over the falling birth rate. The accompanying graph might suggest a solution that every child knows makes sense”.
Defining the Applicability Domain (AD)

Sahigara, Mansouri et al. Molecules 17 (5), 4791-4810
An overview of Different AD Approaches

**AD Approaches**

- Geometric
  - Convex Hull
- Range Based
  - Bounding Box
  - Bounding Box with PCA
- Distance Based
  - Centroid Approach
  - Fixed Knn
  - Variable Knn

Sahigara, Mansouri et al. Molecules 17 (5), 4791-4810
Structure-Activity landscape

Smooth landscape:
Congenericity principle fulfilled

Rugged landscape:
Activity cliffs & structural cliffs

Maggiora (2006): The difference between “the gently rolling hills found on the Kansas prairie” and “the rugged landscapes of Utah’s Bryce Canyon”
Activity cliffs/Structural cliffs

**Activity Cliffs:**
Two structurally similar compounds with diverse values of the activity

**Structural Cliffs:**
Two structurally diverse compounds with similar values of the activity
Discontinuous SARs

Any similarity method must recognize these compounds as being “similar”...

Structure-Activity Landscape index (SALI)

\[ SALI_{st} = \frac{|A_s - A_t|}{1.01 - \text{sim}(s, t)} \]

A: the activity of a given molecule
Sim: the similarity coefficient
ER & AR modeling projects: Background and Goals

• U.S. Congress mandated that the EPA screen chemicals for their potential to be endocrine disruptors

• This led to the development of the Endocrine Disruptor Screening Program (EDSP)

• The initial focus was on environmental estrogens, but the program was expanded to include androgens and thyroid pathway disruptors
CERRAP: Collaborative Estrogen Receptor Activity Prediction Project

40 scientists, 17 research groups

- EPA/NCCT: U.S. Environmental Protection Agency / National Center for Computational Toxicology. USA
- DTU/food: Technical University of Denmark/ National Food Institute. Denmark
- FDA/NCTR/DBB: U.S. Food and Drug Administration. USA
- FDA/NCTR/DSB: U.S. Food and Drug Administration. USA
- Helmholtz/ISB: Helmholtz Zentrum Muenchen/Institute of Structural Biology. Germany
- ILS&EPA/NCCT: ILS Inc & EPA/NCCT. USA
- IRCSS: Istituto di Ricerche Farmacologiche “Mario Negri”. Italy
- JRC_Ispra: Joint Research Centre of the European Commission, Ispra. Italy
- LockheedMartin&EPA: Lockheed Martin IS&GS/ High Performance Computing. USA
- NIH/NCATS: National Institutes of Health/ National Center for Advancing Translational Sciences. USA
- NIH/NCI: National Institutes of Health/ National Cancer Institute. USA
- RIFM: Research Institute for Fragrance Materials, Inc. USA
- UMEA/Chemistry: University of UMEA/ Chemistry department. Sweden
- UNC/MML: University of North Carolina/ Laboratory for Molecular Modeling. USA
- UniBA/Pharma: University of Bari/ Department of Pharmacy. Italy
- UNIMIB/Michem: University of Milano-Bicocca/ Milano Chemometrics and QSAR Research Group. Italy
- UNISTRA/Infochim: University of Strasbourg/ ChemoInformatique. France

Mansouiri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267
CERAPP data and results

Datasets of the project
- Training set: 1,677 chemicals (EPA ToxCast data)
- Prediction set: 32,464 chemicals (The Human Exposure Universe)
- Evaluation set: 7,000 chemicals (Literature: Tox21, FDA, METI…)

40 Models received:

- Classification / Qualitative:
  - Binding: 22 models
  - Agonists: 11 models
  - Antagonists: 9 models

- Regression / Quantitative:
  - Binding: 3 models
  - Agonists: 3 models
  - Antagonists: 2 models

Consensus modeling:
Weighted vote based on rankings of the predictions accuracy scores

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267
Consensus Qualitative Accuracy

Prediction Accuracy Strongly Depends on Data Quality

Total binders: 3961
Agonists: 2494
Antagonists: 2793

<table>
<thead>
<tr>
<th></th>
<th>ToxCast data (training set)</th>
<th>Literature data (test set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed\Predicted</td>
<td>Actives</td>
<td>Inactives</td>
</tr>
<tr>
<td>Actives</td>
<td>83</td>
<td>6</td>
</tr>
<tr>
<td>Inactives</td>
<td>40</td>
<td>1400</td>
</tr>
</tbody>
</table>

Sensitivity: 0.93 0.30 0.87
Specificity: 0.97 0.91 0.94
Balanced accuracy: 0.95 0.61 0.91

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267

ROC curve of the external validation set (literature)
Consensus Quantitative Accuracy

Box plot of the active classes of the consensus model.

- positive concordance < 0.6 => Potency class = Very weak
- 0.6 <= positive concordance < 0.75 => Potency class = Weak
- 0.75 <= positive concordance < 0.9 => Potency class = Moderate
- positive concordance >= 0.9 => Potency class = Strong

Variation of the balanced accuracy with positive concordance thresholds

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267
Concordance of the qualitative models

Most models predict most chemicals as inactive

Only 757 chemicals have >75% positive concordance

Prioritization

Only a small fraction of chemicals require further testing!

Mansouri et al. (2016) EHP 124:1023–1033 DOI:10.1289/ehp.1510267
A renaissance of neural networks in drug discovery

ToxCast chemical landscape: Paving the road to:

Phytoestrogens and Mycoestrogens Induce Changes on Estrogen Receptor α

Identifying known unknowns using the US EPA’s Dashboard

EDSP Prioritization: Collaborative Estrogen Receptor Prediction Project (CERAPP) (SOT)

Humans are potentially exposed to tens of thousands of man-made chemicals in the environment. It is well known that some environmental chemicals mimick natural hormones. Estrogen receptor (ER) ligands are known to disrupt normal hormonal balance. Therefore, the identification and prioritization of chemicals that interact with ER is critical to prevent adverse health effects.
From CERAPP to CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity

- Follow the CERAPP framework
- Use larger size prioritization set
- Use data from the combined EPA ToxCast AR assays
- Collect and curate data from the literature for validation
- Use agonists, antagonists, and binding data
- Build continuous and classification models
- Similar approach for consensus modeling and validation
CoMPARA participants: 34 international groups

New research groups

- NCSU. Department of Chemistry, Bioinformatics Research Center. USA
- EPA/NRMRL. National Risk Management Research Laboratory. USA
- INSUBRIA. University of Insubria. Environmental Chemistry. Italy
- Tartu. University of Tartu. Institute of Chemistry. Estonia
- NIH/NTP/NICEATM. USA
- Chemistry Institute. Lab of Chemometrics. Slovenia
- SWETOX. Swedish toxicology research center. Sweden
- Lanzhou University. China
- BDS. Biodetection Systems. Netherlands
- MTI. Molecules Theurapetiques in silico. France
- IBMC. Institute of Biomedical Chemistry. Russia
- UNIMORE. University of Modena Reggio-Emilia. Italy
- UFG. Federal University of Golas. Brazil
- MSU. Moscow State University. Russia
- ZJU. Zhejiang University. China
- JKU. Johannes Kepler University. Austria
- CTIS. Centre de Traitement de l'Information Scientifique. France
- IdeaConsult. Bulgaria
- ECUST. East China University of Science and Technology. China

From CERAPP

- EPA/NCCT. USA
- DTU/food. Denmark
- FDA/NCTR/DBB. USA
- Helmholtz. Germany
- ILS&EPA/NCCT. USA
- IRCSS. Italy
- LockheedMartin&EPA. USA
- NIH/NCATS. USA
- NIH/NCI. USA
- UMEA/Chemistry. Sweden
- UNC/MML. USA
- UniBA/Pharma. Italy
- UNIMIB/Michem. Italy
- UNISTRA/Infochim. France
- VCCLab. Germany
Developing “OPERA Models”

• Interest in physicochemical properties to include in exposure modeling, augmented with ToxCast HTS *in vitro* data etc.

• Our approach to modeling:
  – Obtain high quality training sets
  – Apply appropriate modeling approaches
  – Validate performance of models
  – Define the applicability domain and limitations of the models
  – Use models to predict properties across our full datasets
PHYSPROP Data: Available from:
http://esc.syrres.com/interkow/EpiSuiteData.htm

- Water solubility
- Melting Point
- Boiling Point
- LogP (KOWWIN: Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Ready biodegradability
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOAs (Octanol/Air Partition Coefficient)
- LogKOCs (Soil Adsorption Coefficient)
- Vapor Pressure
KNIME Workflow to Evaluate the Dataset

LogP dataset: 15,809 chemicals (structures)

- CAS Checksum: 12163 valid, 3646 invalid (>23%)
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES (>24%)
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry (~8%)
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds (~3%)

### Examples of Errors

#### Valence Errors

<table>
<thead>
<tr>
<th>Mol Block</th>
<th>CAS</th>
<th>NAME</th>
<th>SMILES</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Mol Block" /></td>
<td>000274-87-3</td>
<td>TETRAZOCOLE, 1,5-APRYLONE</td>
<td><img src="image2.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Mol Block" /></td>
<td>000542-65-0</td>
<td>ETHYL SODIUM CYANIDE</td>
<td><img src="image4.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Mol Block" /></td>
<td>000707-58-2</td>
<td>5-PROPYL ADENINE</td>
<td><img src="image6.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Mol Block" /></td>
<td>000718-40-0</td>
<td>1-METHYL-HYDRAZINE-1-OXIDE</td>
<td><img src="image8.png" alt="SMILES" /></td>
</tr>
</tbody>
</table>

#### Different Compounds

<table>
<thead>
<tr>
<th>Mol Block</th>
<th>CAS</th>
<th>NAME</th>
<th>SMILES</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image9.png" alt="Mol Block" /></td>
<td>000016-37-7</td>
<td>FLUVANARINE</td>
<td><img src="image10.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Mol Block" /></td>
<td>000017-91-5</td>
<td>1,1-DIHYDROXYMETHYLPROPAINE</td>
<td><img src="image12.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Mol Block" /></td>
<td>000037-91-1</td>
<td>CORTISONE-3A-METHYLOXY</td>
<td><img src="image14.png" alt="SMILES" /></td>
</tr>
<tr>
<td><img src="image15.png" alt="Mol Block" /></td>
<td>000032-38-2</td>
<td>DISPERSE RED 9</td>
<td><img src="image16.png" alt="SMILES" /></td>
</tr>
</tbody>
</table>

---

Examples of Errors

Duplicate Structures

Covalent Halogens

<table>
<thead>
<tr>
<th>Property</th>
<th>Initial file flagged</th>
<th>Updated 3-4 STAR</th>
<th>Curated QSAR ready</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOP</td>
<td>818</td>
<td>818</td>
<td>745</td>
</tr>
<tr>
<td>BCF</td>
<td>685</td>
<td>618</td>
<td>608</td>
</tr>
<tr>
<td>BioHC</td>
<td>175</td>
<td>151</td>
<td>150</td>
</tr>
<tr>
<td>Biowin</td>
<td>1265</td>
<td>1196</td>
<td>1171</td>
</tr>
<tr>
<td>BP</td>
<td>5890</td>
<td>5591</td>
<td>5436</td>
</tr>
<tr>
<td>HL</td>
<td>1829</td>
<td>1758</td>
<td>1711</td>
</tr>
<tr>
<td>KM</td>
<td>631</td>
<td>548</td>
<td>541</td>
</tr>
<tr>
<td>KOA</td>
<td>308</td>
<td>277</td>
<td>270</td>
</tr>
<tr>
<td>LogP</td>
<td>15809</td>
<td>14544</td>
<td>14041</td>
</tr>
<tr>
<td>MP</td>
<td>10051</td>
<td>9120</td>
<td>8656</td>
</tr>
<tr>
<td>PC</td>
<td>788</td>
<td>750</td>
<td>735</td>
</tr>
<tr>
<td>VP</td>
<td>3037</td>
<td>2840</td>
<td>2716</td>
</tr>
<tr>
<td>WF</td>
<td>5764</td>
<td>5076</td>
<td>4836</td>
</tr>
<tr>
<td>WS</td>
<td>2348</td>
<td>2046</td>
<td>2010</td>
</tr>
</tbody>
</table>

### OPERA models

<table>
<thead>
<tr>
<th>Prop</th>
<th>Vars</th>
<th>5-fold CV (75%)</th>
<th>Training (75%)</th>
<th>Test (25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q2</td>
<td>RMSE</td>
<td>N</td>
</tr>
<tr>
<td><strong>BCF</strong></td>
<td>10</td>
<td>0.84</td>
<td>0.55</td>
<td>465</td>
</tr>
<tr>
<td><strong>BP</strong></td>
<td>13</td>
<td>0.93</td>
<td>22.46</td>
<td>4077</td>
</tr>
<tr>
<td><strong>LogP</strong></td>
<td>9</td>
<td>0.85</td>
<td>0.69</td>
<td>10531</td>
</tr>
<tr>
<td><strong>MP</strong></td>
<td>15</td>
<td>0.72</td>
<td>51.8</td>
<td>6486</td>
</tr>
<tr>
<td><strong>VP</strong></td>
<td>12</td>
<td>0.91</td>
<td>1.08</td>
<td>2034</td>
</tr>
<tr>
<td><strong>WS</strong></td>
<td>11</td>
<td>0.87</td>
<td>0.81</td>
<td>3158</td>
</tr>
<tr>
<td><strong>HL</strong></td>
<td>9</td>
<td>0.84</td>
<td>1.96</td>
<td>441</td>
</tr>
</tbody>
</table>
## OPERA models

<table>
<thead>
<tr>
<th>Prop</th>
<th>Vars</th>
<th>5-fold CV (75%)</th>
<th>Training (75%)</th>
<th>Test (25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q2</td>
<td>RMSE</td>
<td>N</td>
</tr>
<tr>
<td>AOH</td>
<td>13</td>
<td>0.85</td>
<td>1.14</td>
<td>516</td>
</tr>
<tr>
<td>BioHL</td>
<td>6</td>
<td>0.89</td>
<td>0.25</td>
<td>112</td>
</tr>
<tr>
<td>KM</td>
<td>12</td>
<td>0.83</td>
<td>0.49</td>
<td>405</td>
</tr>
<tr>
<td>KOC</td>
<td>12</td>
<td>0.81</td>
<td>0.55</td>
<td>545</td>
</tr>
<tr>
<td>KOA</td>
<td>2</td>
<td>0.95</td>
<td>0.69</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td>BA</td>
<td>Sn-Sp</td>
<td></td>
<td>BA</td>
</tr>
<tr>
<td>R-Bio</td>
<td>10</td>
<td>0.8</td>
<td>0.82-0.78</td>
<td>1198</td>
</tr>
</tbody>
</table>
LogP Model: Weighted kNN Model, 9 descriptors

Weighted 5-nearest neighbors
9 Descriptors
Training set: 10531 chemicals
Test set: 3510 chemicals

5 fold Cross-validation:
Q2=0.85  RMSE=0.69
Fitting:
R2=0.86   RMSE=0.67
Test:
R2=0.86   RMSE=0.78
The iCSS Chemistry Dashboard at https://comptox.epa.gov
OPERA Models: Melting Point

4-Acetylaminobiphenyl
4075-P90 | UTXSD062639043

Model Results

Predicted value: 143 °C
Global applicability domain: moderate
Local applicability domain index: 0.89
Confidence level: 0.7

Model Performance

Nearest Neighbors from the Training Set

- 3-Methylbenzanilide
  Measured: 145
  Predicted: 140

- 2-Methylbenzanilide
  Measured: 145
  Predicted: 141
Acknowledgements

National Center for Computational Toxicology
Thank you for your attention