Characterizing, Navigating, and Modeling the Chemical Space Using Next-Generation Cheminformatics Methods

Denis Fourches, PhD

Department of Chemistry, Bioinformatics Research Center, North Carolina State University, USA

www.fourches-laboratory.com
$C_{33}H_{35}F_{2}N_{2}O_{5}$

$0D/1D$  $2D$  $3D$
LIPITOR (atorvastatin)  
Competitive inhibitor of HMG-CoA reductase in liver  
Lower blood cholesterol  

World's best-selling drug of all time ($125 billion over 15 years)  

PDB code = 1HWK
Cheminformatics is becoming an essential element in the chemist’s toolbox

- Characterizing molecules’ structural properties
- Navigating the chemical space
- Building predictive models
- Screening virtual libraries
- Prioritizing new compounds to be tested experimentally
Thousands of molecular descriptors are available for organic compounds. They include constitutional, topological, structural, quantum mechanics based, fragmental, steric, pharmacophoric, geometrical, thermodynamical, conformational, etc.

- Building of models using machine learning methods (NN, SVM, RF)
- Validation of models according to numerous statistical procedures, and their applicability domains.

Thousands of molecular descriptors are available for organic compounds, constitutional, topological, structural, quantum mechanics based, fragmental, steric, pharmacophoric, geometrical.

- Validation of models according to numerous statistical procedures, and their applicability domains.

Correct Classification Rate (CCR) for QSAR models discriminating sensitizers from non-sensitizers was 71–88% when evaluated on several external validation sets, within a broad AD, with positive (for sensitizers) and negative (for non-sensitizers) predicted rates of 85% and 79% respectively.

Can any QSAR model, even if well validated, be applied to any molecule?

Robustness of QSPR models:
- Descriptors type;
- Descriptors selection;
- Machine-learning methods;
- Validation of models.

Applicability domain of models:
Is a test compound similar to the training set compounds?
The new compound will be predicted by the model, only if:

\[ D_i \leq <D_k> + Z \times s_k \]

with \( Z \), an empirical parameter (0.5 by default)

The applicability domain of a given QSAR model is given by:

\[ AD_i = \frac{D_i}{(<D_k> + Z \times s_k)} \times 100 \]

- \( AD_i \leq 100\% \) will be predicted by the model
- \( AD_i > 100\% \) will not be predicted by the model

Development of New Computational Tools Adapted to Hyper-Dimensional HTS data to Analyze Drugs’ Polypharmacology

Visualizing and comparing chemical datasets using the ADDAGRA approach

ADDAGRA for Dataset Fingerprints
(based on distribution of vertex degrees)

ADDAGRA for the analysis of prediction outliers

Many prediction outliers also correspond to outliers in the descriptor space.
ADDAGRA for the analysis of prediction outliers

But still, some large outliers cannot be identified in the chemical space only.

**ACTIVITY CLIFF**

Mol 60

**EXPERIMENT**

Mol 441

**MISANNOTATED**
Hybrid modeling using both chemical and biological descriptors

Environ Health Perspect. 2011, 119(3):364-70

Chemical descriptors

Biological descriptors

High Throughput Screening

Molecular weight, compositions and geometrical parameters, physico-chemical properties (acidic, basic, neutral, amphi- or lipophilic etc.)

In Silico models

Toxicity testing

Human health risk
Compound 19

Chemical Neighbors

Structural Descriptors

Integrative Chemical Biological Read Across CBRA

CBRA Radial Plots

Low et al. CRT. 2013, 26(8):1199

Fourches et al. JCIM, 2016, In Preparation
Cheminformatics Approaches To Analyze the Similarity and the Effects of Environmental Chemical Mixtures
Cheminformatics Approaches To Analyze the Similarity and the Effects of Environmental Chemical Mixtures

CBRA Radial Plots Based on Both Chemical and Children's Exposures Similarity
Concept of Quantitative Structure-Exposure-Toxicity Relationships (QSETR)

Need to develop new modeling workflow …
ERK2 – Bad results with Glide 2015-1

$\rho = 0.74$

$\rho = 0.65$

$\rho = 0.43$

$\rho = 0.51$
GPU-accelerated molecular dynamics simulations

Example: 1 ns simulation of ERK2 using Desmond
65Å*90Å*70Å, 42k atoms, explicit solvent (TIP3P water), step= 1 femtosecond

Up to 1 µs per day on high-end GPU workstations!
New MD-QSAR modeling approach

~10^3 Compounds

~10^3 3D Descriptors

MD-based Descriptor Matrix

~10^6 Time Steps

Deep Learning

Hyper-Predictive QSAR models
Computation of MD Descriptors

- 3D descriptors computed for all conformations sampled along ligand MD simulations

- MD descriptors were constructed by taking the mean and standard deviation of each 3D descriptor distribution for each ligand:

\[
\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}
\]

\[
s_i = \sqrt{\frac{\sum_{j=1}^{n} (x_{ij} - \bar{x})^2}{n - 1}}
\]
Chemical Descriptors Associated with Ligand Activity

Dataset: 87 ERK2 kinase inhibitors; pKi ranging from 4.6 to 9

Atomic Masses Weighted WV
- Cohen's D = 1.1
- Mean of 3D descriptor distribution

Chemical Descriptors Associated with Ligand Activity
Ash and Fourches. JCIIM, 2017. Under Review
For the 3D and MD descriptors, a clear difference in the profiles of active and inactive compounds is observed. Compounds were classified as:
- Active: pKi $\geq 7.5$
- Inactive: pKi $< 7.5$

For MACCS and 2D descriptors, the difference in active and inactive profiles is less apparent.
Hierarchical Clustering of ERK2 Ligands Using 2D and MD fingerprints

**MACCS Fingerprints**

Cophenetic correlation coefficient: **0.89**

**MD Descriptors**

Cophenetic correlation coefficient: **0.74**

Ash and Fourches. JCIM 2017. Under Review
Characterizing the MD Chemical Space of ERK2 Inhibitor Conformations

Ash and Fourches. JCIM 2017. Under Review

Fourches et al. 2015, In Press.

Fourches D, Pu D, Tropsha A. Comb Chem High Throughput Screen. 2011
Functionalized Carbon Nanotubes

Fourches et al.
Nanotoxicology. In Preparation
QNAR Modeling of Carbon Nanotubes

In 2008, Zhou et al* published *in vitro* protein binding, acute toxicity and immune toxicity assays for 84 Carbon NanoTubes (CNTs) decorated with different surface modifiers.

*Zhou et al. Nano Lett., Vol. 8, No. 3, 2008*
Computer-aided design of novel carbon nanotubes with desired biological properties
(in collaboration with Dr. Bing Yan, St. Jude Children's Research Hospital)

Fourches et al. Nanotoxicology 2016 In Press.
Computer-aided design of novel carbon nanotubes with desired biological properties (in collaboration with Dr. Bing Yan, St. Jude Children's Research Hospital)

All rationally prioritized, synthesized, and tested CNTs predicted as non-toxic were confirmed experimentally.

6 out of 10 rationally prioritized, synthesized, and tested CNTs predicted as toxic were confirmed experimentally.

Fourches et al. Nanotoxicology 2016 In Press.
QNAR models for the enhanced set of f-CNTs

**Classification Models (LOO)**

<table>
<thead>
<tr>
<th>Cytotoxicity</th>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.45</td>
<td>79.25%</td>
<td>62.00%</td>
<td>70.87%</td>
</tr>
<tr>
<td>SVM-C Model</td>
<td>0.45</td>
<td>74.40%</td>
<td>56.00%</td>
<td>66.99%</td>
</tr>
</tbody>
</table>

**Continuous Models (LOO)**

[Graph showing cytotoxicity regression tree consensus model predictions]
HLA-induced drug adverse effects

Small molecule drugs bind specifically to certain type of HLA proteins.

Binding pocket of HLA proteins is slightly modified

Range of “self” peptides able to bind the given HLA type is modified

Immune reaction

HLA proteins as important off-targets

Need predictive models to assess HLA-induced ADR
Clozapine-induced agranulocytosis is associated with rare HLA-DQB1 and HLA-B alleles.
Molecular Docking Study at HLA-B*57:01

Develop a virtual screening model using molecular docking at the HLA-B*57:01 variant using three X-ray crystals of abacavir bound to HLA-B*57:01.


3VRI 3VRJ 3UPR
Summary

- Cheminformatics is becoming mandatory in all projects involving the curation, integration, characterization, analysis, testing, modeling, visualization, screening of chemicals.

- The skyrocketing amount of freely-available data in the public domain is boosting the development of new cheminformatics approaches to fully exploit that data, especially when it comes to chemical risk assessment and *in silico* toxicity predictions.

- New methods such as MD-QSAR or QSETR are poised to boost the prediction performances of cheminformatics predictors.

- Structure-based docking and pan-target screening have never been so relevant for chemical risk assessment, especially for key targets such as ER, AhR, or HLA.
Acknowledgements

Lab members
George Van Den Driessche
Jeremy Ash
Melaine Kuenemann, PhD
Ryan Lougee
Phyo Phyo Kyaw Zin
Bethany Cook

Funding
NCSU CFEP Program
RISF
CMI