Best practices for developing predictive QSAR models

Alexander Tropsha
Laboratory for Molecular Modeling and
Carolina Center for Exploratory Cheminformatics Research
School of Pharmacy
UNC-Chapel Hill
OUTLINE

• Introduction: Brief outline of the QSAR approach

• Why models fail (bad practices)

• Good practices.
  – Predictive QSAR Modeling Workflow
  – Examples of the Workflow applications
  – Emerging applications of QSAR: chemocentric informatics

• Conclusions: QSAR modeling is a decision support
The rumors of QSAR demise have been greatly exaggerated

Graphs are courtesy of Prof. A. Cherkasov
Principles of QSAR modeling

Quantitative Structure Activity Relationships

COMPOUNDS  DESCRIPTORS  ACTIVITY

0.613
0.380
-0.222
0.708
1.146
0.491
0.301
0.141
0.956
0.256
0.799
1.195
1.005
Principles of QSAR/QSPR modeling

\[ f(\text{Pattern matrix}) = \text{PROPERTY}(i) \]

With \( m \) molecules and \( n \) descriptors
The utility of QSAR models

CHEMICAL STRUCTURES → CHEMICAL DESCRIPTORS → PREDICTIVE QSAR MODELS → PROPERTY/ACTIVITY

~10⁶ – 10⁹ molecules

VIRTUAL SCREENING → HITS

INACTIVES
### QSAR Modeling

**Goal:** Establish correlations between descriptors and the target property capable of predicting activities of novel compounds

<table>
<thead>
<tr>
<th>Chemistry (IC50, Kd...)</th>
<th>Biology</th>
<th>Cheminformatics (Molecular Descriptors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp.1</td>
<td>Value1</td>
<td>D₁</td>
</tr>
<tr>
<td>Comp.2</td>
<td>Value2</td>
<td>&quot;</td>
</tr>
<tr>
<td>Comp.3</td>
<td>Value3</td>
<td>&quot;</td>
</tr>
<tr>
<td>Comp.N</td>
<td>ValueN</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

\[ q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (\bar{y} - y_i)^2} \]

\[ BA = F(D) \{ \text{e.g., \ldots} \} \]

(e.g., \(-\text{LogIC50} = k_1 D_1 + k_2 D_2 + \ldots + k_n D_n\))
But … the unbearable lightness of model building for training sets…

...leads to unacceptable prediction accuracy.

EXTERNAL TEST SET PREDICTIONS

Training Linear (Training)

$y = 0.5958x + 2.3074$

$R^2 = 0.2135$

$y = 0.4694x + 2.9313$

$R^2 = 0.1181$
Only a small fraction of “predictive” training set models with LOO $q^2 > 0.6$ is capable of making accurate predictions ($r^2 > 0.6$) for the test sets.
Major components of QSAR modeling

• **Target properties (dependent variable)**
  – Continuous (e.g., IC50)
  – Categorical unrelated (e.g., different pharmacological classes)
  – Categorical related (e.g., subranges described as classes)

• **Descriptors (or independent variables)**
  – Continuous (allows distance based similarity)
  – Categorical related (allows distance based similarity)
  – Categorical unrelated (require special similarity metrics)

• **Correlation methods (with and w/o variable selection)**
  – Linear (e.g., LR, MLR, PCR, PLS)
  – Non-linear (e.g., kNN, RP, ANN, SVM)

• **Validation and prediction**
  – Internal (training set) vs. external (test set) vs. independent evaluation set

• **Examples of applications and pitfalls**
Complexity of QSAR modeling: Choices and Practices

- Descriptors (thousands and counting)
- Data-analytical methods (dozens and counting)
- Validation approaches (unfortunately (!) only a handful but counting)
- Experimental validation as part of model building (very rare)

BUT

- We typically use one (or at best very few) modeling techniques
- Publish successes only
- Compete but (mostly) indirectly
Why models may fail

- Incorrect data (structures and activities) in the dataset
- Modeling set is too small
- No external validation
- Incorrect selection of an external test set
- Incorrect division of a dataset into training and test sets
- Incorrect measure of prediction accuracy
- Insufficient statistical criteria to estimate predictive power of models
- Lack or incorrect definition of applicability domain
- No Y-randomization test (overfitness)
- Presence of leverage (structure) and activity outliers

Also, see Dearden JC, Cronin MT, Kaiser KL. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR QSAR Environ Res. 2009;20(3-4):241-66
Some reasons why QSAR models may fail: using incorrect target function in classification QSAR for biased datasets:

- **A typical target function (Classification Rate):**
  \[ CR = \frac{N(\text{classified correctly})}{N(\text{total})} \]

  A dataset:
  - **Class 1:** 80 compounds
  - **Class 2:** 20 compounds

  **Model:** assign all compounds to Class 1.
  **Target function:** CR=0.8

  The model appears to have high classification accuracy

- **Better target function:**
  \[ CCR = 0.5(\text{Sensitivity} + \text{Specificity}) \]

  In the above example, CCR = 0.5

- **General formula:**
  \[
  CCR = \frac{1}{K} \sum_{k=1}^{K} \frac{N_{k}^{corr}}{N_{k}^{total}}
  \]

  - \( K \) – the number of classes
  - \( N_{k}^{corr} \) – the number of compounds of class \( k \) assigned to class \( k \)
  - \( N_{k}^{total} \) – total number of compounds of class \( k \)

- For categorical response variable, target functions can depend also on the absolute errors (differences between predicted and observed classes).
HOW TO DEFINE A PREDICTIVE QSAR MODEL

Regression
\[ \tilde{y}^r = a'y + b' \]
\[ a' = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum (y_i - \bar{y})^2} \]
\[ b' = \bar{\tilde{y}} - a'\bar{y} \]
\[ b = \bar{y} - a'\bar{y} \]
\[ y^0 = k\bar{y} \]
\[ k = \frac{\sum y_i\tilde{y}_i}{\sum \tilde{y}_i^2} \]

Correlation coefficient
\[ R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \]

Regression through the origin
\[ \tilde{y}^0 = k'y \]
\[ k' = \frac{\sum y_i\tilde{y}_i}{\sum y_i^2} \]

Coefficients of determination
\[ R^2 = 1 - \frac{\sum (\tilde{y}_i - y^0_i)^2}{\sum (y_i - \bar{y})^2} \]
\[ R^0_2 = 1 - \frac{\sum (y_i - \tilde{y}^0_i)^2}{\sum (y_i - \bar{y})^2} \]

CRITERIA
\[ q^2 > 0.5; R^2 > 0.6; \]
\[ k \text{ or } k' \approx 1.0; R^2 \text{ or } R^0_2 \approx R^2 \]
Some reasons why QSAR models may fail: No Applicability Domain is defined for the Model

• Compounds which are highly dissimilar from all compounds of the training set (according to the set of descriptors selected) cannot be predicted reliably

Lack of the AD:
  - unjustified extrapolation
  - wrong prediction

Typical situation:
  a compound of the test set for which error of prediction is high is considered an outlier

HOWEVER: a compound of the test set dissimilar from all compounds of the training set can be predicted accurately
Applicability domain of QSAR models

For a given model, two parameters are calculated:
- $\langle D_k \rangle$ : average Euclidian distance between each compound of the training set and its $k$ nearest neighbors in the descriptors space;
- $s_k$ : standard deviation of the distances between each compound of the training set and its $k$ nearest neighbors in the descriptors space.
**Applicability domain of QSAR models**

For a given model, two parameters are calculated:
- \( <D_k> \): average euclidian distance between each compound of the training set and its \( k \) nearest neighbors in the descriptors space;
- \( s_k \): standard deviation of the distances between each compound of the training set and its \( k \) nearest neighbors in the descriptors space.

For each test compound \( i \), the distance \( D_i \) is calculated as the average of the distances between \( i \) and its \( k \) nearest neighbors in the descriptors space.

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).

*Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:…*  
Applicability domain vs. prediction accuracy (Ames Genotoxicity dataset)
Some reasons why QSAR models may fail: Y-randomization test is not carried out

- **Y-randomization test:**
  - Scramble activities of the training set
  - Build models and get model statistics.
  - If statistics are comparable to those obtained for models built with real activities of the training set, the last are unreliable and should be discarded.

**Frequently, Y-randomization test is not carried out.**

**Y-randomization test is of particular importance, if there is:**
- a small number of compounds in the training or test set
- response variable is categorical
Activity randomization: model robustness

The lowest $q^2 = 0.51$ in the top 10 models
The highest $q^2 = 0.14$ for randomized datasets

Training set with real property values is expected to produce much higher $q^2$ values than the same set with randomized property values.
Some reasons why QSAR models may fail: outliers

- Many potential outliers can be detected in the dataset prior to QSAR studies, but typically this is not done.

- Two types of outliers
  - **Leverage outliers**: compounds dissimilar from all other compounds in a dataset in the chemistry space.
  - **Activity outliers**: compounds similar to some other compounds in the dataset, but with activities quite different from those of their nearest neighbors (activity cliffs).
Why QSAR models may fail: insensitive descriptors.

[Example: Optimal (left panel) and traditional (right panel) orientations of androgen (DHT shown in gold) and estrogen (estradiol shown in green) within human SHBG steroid-binding site].

Identical $q^2$ (CoMFA*) of 0.53

*CoMFA – Completely Misleading Famous Aberration

A. Cherkasov, *JMC*, 2008
Recently, D. Young et al. pointed out the importance of cleaning data, especially, in the context of QSAR modeling.

They investigated several public and commercial databases to calculate their error rates: the latter were ranging from 0.1 to 3.4% depending on the database.

Their main conclusions were that small structural errors within dataset could lead to significant loss of predictive abilities for the QSAR models which have been built using those erroneous input data.
Why can’t we get it Right? Have not we tried enough?

- Descriptors? No, we have plenty (e.g., 1000’s in Dragon)
- Datamining methods? No, we also have plenty (e.g., SAS)
- Training set statistics? NO, it does not work
- Test set statistics? Maybe, but it is still insufficient

So...what else can we do??????

- Change the success criteria! Leave behind the phase of “narcissistic” modeling and focus on external predictivity and experimental validation.
- Recognize QSAR as an empirical data modeling approach: just do it any (all) way you like but VALIDATE on independent datasets!
Revising QSAR Modeling Process: Predictive QSAR Modeling Workflow*

• Model Building: Combination of various descriptor sets and variable selection data modeling methods (Combi-QSAR)

• Model Validation
  – Y-randomization
  – Training, test, AND evaluation set selection
  – Model sampling and selection criteria
  – Applicability domain

• Consensus prediction using multiple models

Predictive QSAR Modeling Workflow*

Original Dataset

Structure Curation/Harmonization

Split into Training, Test and External Validation sets

Multiple Training Sets

Y-randomization

Combi-QSAR Modeling

Multiple Test Sets

Activity Prediction

Only accept models that passed both internal and external accuracy filters

Database Screening Using Applicability Domain

External validation Using Applicability Domain

Validated Predictive Models with High Internal & External Accuracy

Experimental Validation

* Tropsha, A., Gramatica, P., Gombar, V. The importance of being earnest:…
The importance of data curation: What do these two men have in common?

But in order to satisfy the skeptics on the U.S. side that any agreement could be carried out, obviously, one settled for even more stringent verification measures than some people thought were necessary at the time in order to win order skeptics in agreement. We all remember, you know, trust but verify - that Russian proverb that Ronald Reagan and Felix Dzerzhinsky liked so much. But it wasn’t trust but verify. It was we don’t trust you, and therefore, we have to verify. And we have to verify very rigorously. That was the atmosphere at the time. And, you know, there are visages of that today.
INITIAL LIST OF SMILES

1. Removal of mixtures, inorganics (and eventually organometallics)

2. Structural conversion
   Cleaning/removal of salts
   Normalization of specific chemotypes
   Treatment of tautomeric forms

3. Analysis/removal of duplicates

4. Manual inspection

CURATED DATASET

SOFTWARE

ChemAxon - Standardizer
OpenEye - Filter

ChemAxon - Standardizer
OpenBabel
Molecular Networks - CHECK, TAUTOMER

ISIDA - Duplicates
HiT QSAR
CCG - MOE

ISIDA - EdiSDF
Hyleos - ChemFileBrowser
OpenBabel
ChemAxon - MarwinView

Division of the Dataset into Three Subsets and External Validation

Random division

- Dataset
  - Modeling Subset
    - Training Set
    - Test Set
  - External Validation Set
    - External Validation
      - Predictive Models

Rational division

- "Predictive" Models
  - Model Validation


The OECD Principles of model validation

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm;
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation, if possible;

Application of Predictive QSAR Workflow to GGTase-I Inhibitors*

48 GGTase-I Inhibitors

Multiple Training Sets

Y-Randomization

Multiple Test Sets

$k$ Nearest Neighbor ($k$NN) QSAR

Automatic Lazy Learning (ALL) QSAR

Partial Least Square (PLS) QSAR

Activity Prediction

104 models that have a $R^2 > 0.60$ and $q^2 > 0.60$

Experimental Validation

79 Hits predicted as GGTase-I Inhibitors

Screen $9.5 \times 10^6$ Compound Database

Validated Predictive Models with High Internal & External Accuracy

*Collaboration with P. Casey and Y. Peterson, Duke
Database Mining Reveals Unique Chemical Entities

2 Training Set Scaffolds

GGTI x Series Peptidomimetics
Hamilton & Sebti

GGTI DUX Series Pyrazoles
Peterson & Casey

Novel Scaffolds Discovered

Amide
Amide
Furan

Pyrazole Amide

Tetrazole
**Database Mining: Similarity Search vs. QSAR Search**

A Large Commercial Database of 515,000 Compounds

- **Similarity Search**
  - Similarity Metric: Tanimoto Coefficient; of every single compound in the training set
  - Fingerprint: MACCS Structural Keys
  - 425 hits obtained for TC=0.80; 2 hits obtained for TC=0.90

- **QSAR Database Search**
  - Global search based on the whole chemical space (MZ 4.09 des.) of training set
  - 12 hits obtained after global search (Z = 0.5) and subjected to consensus predictions
  - 2 selected for experimental validation based on high predicted activity, uniqueness of structure & availability

There was NO overlap between the hits from two protocols; All 12 QSAR hits were below TC=0.80 of training set.
Recent examples of experimentally validated QSAR-based predictions

- **HDAC inhibitors**: Wang, S. *et al.*, (JCIM, 2009, 49, 461-76)
- **GGT-I inhibitors**: Wang, Peterson, *et al* (JMC, 2009, 52(14):4210-20; provisional patent)
- **5HT7 binders**: Zhao *et al* (in preparation)
Characteristic AmpC Ligands and Decoys and Their Ranks by Different Scoring Functions. Blue = DOCK, magenta = ScreenScore, yellow = FlexX, cyan = PLP, purple = PMF, and red = SMoG (SMoG ranks are based on a ranking, which does not include halogenated compounds).

QSAR vs. Docking: Application of QSAR Approaches to the Analysis of Binding Decoys

Decoys are frequently indistinguishable from binders using typical SB scoring functions.*

Study Design (AmpC β-lactamase dataset)

Class 1

- 21 inhibitors
- 10 compounds
- 64 HTS ‘hits’ (non-binders)

Class 2

- 80 nonbinders

Modeling Set

- 51 compounds
- Binary kNN QSAR Model Building (MolConnZ descr)

Predictive Models

- 342

Database Mining

Correct Classification Rate (CCR) = \(0.5 \times \frac{TP}{N_1} + \frac{TN}{N_0}\)

\(N_1\): Total number of inhibitors
\(N_0\): Total number of nonbinders

Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A
J Comput Aided Mol Des. 2008;22(9):593-609
The QSAR models do not predict the majority of the 64 HTS ‘Hits’ as binders in agreement with experimental study by Shoichet group.

Z = 0.5; **Accuracy = 20/25 = 0.8**

Z = 3.0; **Accuracy = 47/55 = 0.85**

---

## Descriptor Interpretation

<table>
<thead>
<tr>
<th>Rank</th>
<th>Descriptor ID</th>
<th>Frequency</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nHCsatu</td>
<td>32.2</td>
<td>CH\textsubscript{n} (unsaturated)</td>
</tr>
<tr>
<td>2</td>
<td>Hsulfonamide</td>
<td>28.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>nnitrile</td>
<td>27.5</td>
<td>C≡N</td>
</tr>
<tr>
<td>4</td>
<td>Hmin</td>
<td>27.2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>naaO</td>
<td>26.3</td>
<td>:O: (aromatic)</td>
</tr>
<tr>
<td>6</td>
<td>naaS</td>
<td>26.3</td>
<td>:S: (aromatic)</td>
</tr>
<tr>
<td>7</td>
<td>SaaCH</td>
<td>26.0</td>
<td>:CH:</td>
</tr>
<tr>
<td>8</td>
<td>n3Pad24</td>
<td>26.0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>SssCH2</td>
<td>26.0</td>
<td>-CH\textsubscript{2}</td>
</tr>
<tr>
<td>10</td>
<td>SHBint5</td>
<td>25.4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Xvch5</td>
<td>24.3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>n2Pag23</td>
<td>24.3</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>IDW</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>htets2</td>
<td>23.7</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>nimine</td>
<td>23.7</td>
<td>N≡C</td>
</tr>
</tbody>
</table>

**inhibitor**

![Inhibitor structure](image1)

**nonbinder**

![Nonbinder structure](image2)
Virtual Screening of a PubChem AMPc HTS dataset of 69,653 Compounds

Hit selection criteria:
- Within AD of at least 50% of models
- 80% of those predict a compound as an inhibitor

This leads to 15 Hits
One “inactive” compound (CID 69951) Shows Micro-molar Inhibitory Activity

Kd = 18µM, Ki = 135µM

Experiments done by Dr. D. Teotico at UCSF
Pairwise potential (PPL) descriptors are applied to characterize protein-ligand interactions

1. Delaunay tessellation of protein-ligand interface
   
   Each tetrahedron is categorized by:
   a) receptor/ligand atoms
   b) Chemical atom type
   In total, there are 554 theoretical descriptor types (m) [2].

2. For each atom at the protein-ligand interface, assign the maximal charge transfer (MCT) value calculated by Conceptual Density Functional theory DFT [1].

3. Each descriptor’s value is the SUM of protein (p)-ligand (l) pairwise potential for the same tetrahedral type at the interface (n)

\[
\text{PPL}_m = \sum_{k=1}^{n} \sum_{p=1}^{3} \sum_{l=1}^{3} (\text{MCT}_p \times \text{MCT}_l / d_{pl})_k
\]

[2]: Zhang S et al. J. Med. Chem. 2006; 49(9); 2713-2724
Target-specific filter construction

1. X-ray structure of the protein + its bound native ligand
2. Docking (e.g. Fred) to generate ~1000 geometric poses
3. Generate pairwise potential descriptors for each pose
4. Classify the poses into native-like/decoy based on RMSD cutoff (4Å)
5. Training/test set
6. Build binary classification models using k-NN or LIBSVM
7. Select acceptable models (Accuracy > 95%) for consensus prediction
8. External validation set

Accuracy > 95%
A proposed screening protocol for structure-based virtual screening

1. **Protein-ligand complex(es)**
   - Generate geometric poses
   - Generate pairwise potential descriptors
   - Build QSAR classification model
   - Select acceptable models

2. **A mix of ligands and decoys from DUD**
   - Docking (e.g. Fred) and save multiple poses for each compound

3. **Target-specific filter/multi-target filter**
   - Exclude poses predicted as decoys

4. **Re-ranking (e.g. MedusaScore*)**

5. **Performance evaluation**

*MedusaScore was developed in Dr. N. Dokholyan’s lab in UNC. J. Chem. Inf. Model. 2008, 48, 1656–1662*
Docking enrichment plot for DHFR using DUD

DHFR (201 ligands + 7145 decoys)

- Ideal
- Chemgauss3
- Fred_consensus
- MedusaScore
- Consensus: | + | (after filter)
- 2D similarity
- MedusaScore (after filter)
- Chemgauss3 (after filter)
Chemocentric Informatics: Integration of QSAR modeling with other approaches to drug discovery: structural hypothesis fusion.

Application to 5-HT$_6$ receptor linked to Alzheimer’s disease
Disease-related genes or proteins

Network mining

Disease-related proteins

Disease gene signatures

Text/database mining

PubMed

CTD

HMDB

ChemoText

cmap

New hypothesis about connectivity between chemicals and diseases

Accept common hits only

Disease-Target Association

Target related ligands

Functional data

Binding data

QSAR

Predictive models

Database mining

Structural hypothesis “putative drug candidates”

New testable hypothesis with higher confidence

New hypothesis about connectivity between chemicals and diseases
QSAR Modeling of 5-HT$_6$ Ligands

5-HT$_6$ Dataset:
- 79 Binders ($Ki < 10$ µM),
- 99 Non-binders ($Ki \geq 10$µM)

Source: PDSP Ki Database
Comparison of the QSAR Approaches to Classify 5-HT₆ Receptor Ligands
The Connectivity Map

Step 1: upload signature

Step 2: query the cmap

Step 3: list of correlated compounds

Querying the cmap with Alzheimer’s Disease Gene Signatures

Alzheimer’s disease gene signatures “Two different signatures” from hippocampus (S1) and cerebral cortex (S2) from two independent reports

Upload signature

Query the cmap

List of compounds

Positive Connectivity “possible causes for disease state”

Negative Connectivity “possible treatments for disease state”

Identification of possible treatments (A,B,C) and causes (F)
Virtual Screening Results

Integrated Informatics

300 5-HT₆ Binders Hits

COMMON HITS

Negative Connection HITS

34 Potential Anti-AD agents
## Selected Common Hits from QSAR and the cmap

<table>
<thead>
<tr>
<th>WDI Name</th>
<th>Pred. Value</th>
<th>cmap Score S1</th>
<th>cmap Score S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clomiphene</td>
<td>0.768</td>
<td>-0.425</td>
<td>-0.611</td>
</tr>
<tr>
<td>Tamoxifene</td>
<td>-0.414</td>
<td>-0.741</td>
<td>-0.619</td>
</tr>
<tr>
<td>Toremifene</td>
<td>-0.581</td>
<td>-0.491</td>
<td>-0.408</td>
</tr>
<tr>
<td>Raloxifene</td>
<td>-0.514</td>
<td>-0.541</td>
<td>-0.489</td>
</tr>
<tr>
<td>Clomipramine</td>
<td>-0.541</td>
<td>-0.388</td>
<td>-0.683</td>
</tr>
<tr>
<td>Second Generation</td>
<td>-0.790</td>
<td>-0.591</td>
<td>-0.591</td>
</tr>
<tr>
<td>Mebeverine</td>
<td>-0.698</td>
<td>-0.742</td>
<td>-0.611</td>
</tr>
<tr>
<td>Sodium Channels</td>
<td>-0.514</td>
<td>-0.514</td>
<td>-0.750</td>
</tr>
</tbody>
</table>
| Selective Estrogen Receptor Modulators (SERMs) predicted as 5-HT6 receptor ligands and potential therapeutics for AD: A power of the integrated chemogenomic approach
Raloxifene is a 5-HT\textsubscript{6} Binder and Potential Anti-Alzheimer’s

**A Power of the Integrated Chemogenomic Approach**

- Raloxifene binds to 5-HT\textsubscript{6} receptor with a $K_i = 750$ nM.
- Raloxifene given at a dose of 120 mg/day led to reduced risk of cognitive impairment in post-menopausal women.

[Graph showing 5-HT\textsubscript{6} Receptor binding data]

Raloxifene (blue triangle) and Chlorpromazine (square) versus [3H] LSD competition binding at 5-HT\textsubscript{6} receptors. Tested by our collaborators at PDSP.
Raloxifene is predicted to bind several receptor families using QSAR-based VS

Classification models used prospectively to predict raloxifene’s promiscuity

<table>
<thead>
<tr>
<th>Receptor family</th>
<th>Type of model</th>
<th>Number of models</th>
<th>Average score</th>
<th>Total number of models</th>
<th>CCR_models (tr, ts &amp; ex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-HT</td>
<td>Classification</td>
<td>512</td>
<td>0.57</td>
<td>650</td>
<td>≥0.7</td>
</tr>
<tr>
<td>Alpha 2</td>
<td>Classification</td>
<td>1686</td>
<td>0.76</td>
<td>2045</td>
<td>≥0.9</td>
</tr>
<tr>
<td>Dopamine</td>
<td>Classification</td>
<td>350</td>
<td>0.70</td>
<td>482</td>
<td>≥0.7</td>
</tr>
<tr>
<td>Muscarinic</td>
<td>Classification</td>
<td>444</td>
<td>0.66</td>
<td>500</td>
<td>≥0.7</td>
</tr>
<tr>
<td>Sigma</td>
<td>Classification</td>
<td>730</td>
<td>0.69</td>
<td>898</td>
<td>≥0.7</td>
</tr>
</tbody>
</table>

Regression models used retrospectively to predict raloxifene’s binding affinity

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Type of models</th>
<th>Number of models</th>
<th>plogKi</th>
<th>SD</th>
<th>Ki</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha 2A</td>
<td>Regression</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Alpha 2B</td>
<td>Regression</td>
<td>25</td>
<td>6.2</td>
<td>0.00</td>
<td>631</td>
</tr>
<tr>
<td>Alpha 2C</td>
<td>Regression</td>
<td>1</td>
<td>6.8</td>
<td>0.42</td>
<td>158</td>
</tr>
</tbody>
</table>
Comprehensive screening results

<table>
<thead>
<tr>
<th>CMPD</th>
<th>PI</th>
<th>Tier</th>
<th>5ht1a</th>
<th>5ht1b</th>
<th>5ht1d</th>
<th>5ht1e</th>
<th>5ht2a</th>
<th>5ht2b</th>
<th>5ht2c</th>
<th>5ht3</th>
<th>5ht4</th>
<th>5ht5a</th>
<th>5ht6</th>
<th>5ht7</th>
</tr>
</thead>
<tbody>
<tr>
<td>13505</td>
<td>Hajjo</td>
<td>Raloxifene</td>
<td>2,330.00</td>
<td>624</td>
<td>1,222.00</td>
<td>1,868.00</td>
<td>1,049.00</td>
<td>69</td>
<td>1,642.00</td>
<td>5,050.00</td>
<td>1,219.00</td>
<td>750</td>
<td>1,220.00</td>
<td></td>
</tr>
<tr>
<td>14821</td>
<td>Hajjo</td>
<td>Fendiline</td>
<td>3,550.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>894</td>
</tr>
<tr>
<td>678</td>
<td>PDSP (MUL)</td>
<td>Tamoxifen</td>
<td>3,477.00</td>
<td>1,618.00</td>
<td>2,596.00</td>
<td>4,282.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,123.00</td>
<td>931.1</td>
<td>1,077.00</td>
</tr>
<tr>
<td>10572</td>
<td>PDSP</td>
<td>Tamoxifen</td>
<td>&gt;10,000</td>
<td>7,857.00</td>
<td>2,720.00</td>
<td>1,952.00</td>
<td>5,787.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7,821.00</td>
<td>1,698.00</td>
</tr>
</tbody>
</table>

**Ca-channel blocker** (L-type “as neuronal”) And has high permeability through BBB  
Could be a positive **potentiator of GABA_B** (to be tested: because derivatives are)  
And has high permeability through BBB
## Importance of hypothesis fusion

<table>
<thead>
<tr>
<th>WDI Name</th>
<th>cmap Name</th>
<th>No. of Models</th>
<th>Av. Pred. Value</th>
<th>cmap Score S1</th>
<th>cmap Score S2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLOZAPINE</td>
<td>clozapine</td>
<td>900</td>
<td>1.00</td>
<td>-0.398</td>
<td>-0.366</td>
</tr>
<tr>
<td>TAMOXIFEN</td>
<td>tamoxifen</td>
<td>910</td>
<td><strong>0.99</strong></td>
<td>0.358</td>
<td>-0.507</td>
</tr>
<tr>
<td>FLUSPIRILENE</td>
<td>fluspirilene</td>
<td>854</td>
<td>0.99</td>
<td>-0.493</td>
<td>-0.551</td>
</tr>
<tr>
<td>ZUCLOPENTHIXOL</td>
<td>zuclopenthixol</td>
<td>883</td>
<td>0.98</td>
<td>-0.609</td>
<td>-0.746</td>
</tr>
<tr>
<td>BI-2</td>
<td>imipramine</td>
<td>898</td>
<td>0.98</td>
<td>-0.503</td>
<td>-0.415</td>
</tr>
<tr>
<td>CIDOXEPIN</td>
<td>doxepin</td>
<td>908</td>
<td>0.97</td>
<td>-0.463</td>
<td>-0.777</td>
</tr>
<tr>
<td>NORTRIPTYLINE</td>
<td>nortriptyline</td>
<td>883</td>
<td>0.96</td>
<td>-0.555</td>
<td>-0.410</td>
</tr>
<tr>
<td>BI-3</td>
<td>clomipramine</td>
<td>893</td>
<td>0.95</td>
<td>-0.768</td>
<td>-0.425</td>
</tr>
<tr>
<td>ENCLOMIFENE</td>
<td>clomifene</td>
<td>899</td>
<td><strong>0.91</strong></td>
<td>-0.414</td>
<td>-0.611</td>
</tr>
<tr>
<td>DO-897</td>
<td>Prestwick-559</td>
<td>858</td>
<td>0.79</td>
<td>-0.741</td>
<td>-0.619</td>
</tr>
<tr>
<td>LY-294002</td>
<td>LY-294002</td>
<td>866</td>
<td><strong>0.70</strong></td>
<td>-0.351</td>
<td>-0.303</td>
</tr>
<tr>
<td>ACEFYLLINE-PRENYLAMINE</td>
<td>prenylamine</td>
<td>679</td>
<td>0.69</td>
<td>-0.589</td>
<td>-0.457</td>
</tr>
<tr>
<td>NISOXETINE</td>
<td>nisoxetine</td>
<td>899</td>
<td>0.68</td>
<td>-0.491</td>
<td>-0.408</td>
</tr>
<tr>
<td>IFENPRODIL</td>
<td>ifenprodil</td>
<td>900</td>
<td>0.66</td>
<td>-0.541</td>
<td>-0.489</td>
</tr>
<tr>
<td>FENDILINE</td>
<td>fendiline</td>
<td>765</td>
<td>0.66</td>
<td>-0.388</td>
<td>-0.683</td>
</tr>
<tr>
<td>NAFTIFINE</td>
<td>naftifine</td>
<td>724</td>
<td>0.66</td>
<td>-0.790</td>
<td>-0.591</td>
</tr>
<tr>
<td>RALOXIFEN</td>
<td>raloxifene</td>
<td>809</td>
<td><strong>0.64</strong></td>
<td>-0.378</td>
<td>-0.482</td>
</tr>
<tr>
<td>MEBEVERINE</td>
<td>mebeverine</td>
<td>820</td>
<td>0.57</td>
<td>-0.543</td>
<td>-0.798</td>
</tr>
<tr>
<td>LOBELANIDINE</td>
<td>lobelanidine</td>
<td>826</td>
<td>0.56</td>
<td>-0.508</td>
<td>-0.488</td>
</tr>
<tr>
<td>LOBELINE</td>
<td>lobeline</td>
<td>882</td>
<td>0.55</td>
<td>-0.514</td>
<td>-0.750</td>
</tr>
<tr>
<td>AZACYCLONOL</td>
<td>azacyclonol</td>
<td>840</td>
<td>0.54</td>
<td>-0.448</td>
<td>-0.556</td>
</tr>
</tbody>
</table>
ToxRefDB: >$1 Billion Million Dollars Worth of In Vivo Chronic/Cancer Bioassay Effects and Endpoints

ToxCast Phase I Chemicals

- Chemical/Study-centric
- Detailed toxicity data
- Toxicity standards/Data model
- Exportable
- Compatible with multiple platforms (ACCESS, xml, MySQL)

http://www.epa.gov/ncct/toxrefdb/
Chemocentric view of biological data

Toxicity Risk Assessment

NO₂

increasing uncertainty

increasing relevance to RA

increasing complexity

SAR
structure-activity relationships

Slide courtesy of Dr. Ann Richard (EPA)
Chemical Structure & Properties

Toxicity Signature Development

Genomic Signatures

Cellular Assays

Biochemical Assays

Toxicology Endpoints

In silico Predictions
Poor global relationships between in vivo and in vitro assays in ToxCast™ (based on Matthew’s Correlation Coefficient, MCC*)

75 (49+26) in vivo and 409 in vitro endpoints

\[ *MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]
Chemical Structure – *in vitro* – *in vivo* Toxicity Data Continuum.

Slide is courtesy of Dr. Ivan Rusyn (UNC)
Use of hybrid descriptors for structure - in vitro – in vivo modeling

Use of HTS based biological descriptors improves predictive power of QSAR Models of chemical carcinogenicity*

A new “hierarchical QSAR” approach* relying on the relationship between in vitro and in vivo ToxCast assays results affords highly predictive models.

Introducing QNTR modeling

Nanoparticle fingerprints

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

Experimental measurements (size, relaxivities, zeta potential etc.)

Quantitative Nanostructure Toxicity Relationships

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

High-throughput cellular-based assays
Case Study 1

Recently¹, 51 diverse NPs were tested in-vitro against 4 cell lines in 4 different assays at 4 different concentrations (→ 51x64 data matrix).

**NANOPARTICLES**

→ cross-linked iron oxide (CLIO)-based (23 NPs)
→ pseudocaged nanoparticle (PNP)-based (19 NPs)
→ monocrystalline iron oxide nanoparticle (MION)-based (4 NPs)
→ quantum dot-based with a CdSe core, a ZnS shell, and a polymer coating (3 NPs)
→ two other iron-based MNPs: Feridex IV (approved for in vivo imaging) and Ferrum Hausmann (approved for iron supplementation)

¹ Shaw et al. Perturbational profiling of nanomaterial biologic activity. PNAS, 2008, 105, 7387-7392
Is it possible to predict whether a given particle will induce low or high biological effects using QNTR models?

<table>
<thead>
<tr>
<th>Effect</th>
<th>Size</th>
<th>Zeta pot.</th>
<th>Relaxitivities</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP-01</td>
<td>High</td>
<td>0.4865</td>
<td>0.5278</td>
</tr>
<tr>
<td>NP-02</td>
<td>Low</td>
<td>0.4054</td>
<td>0.7222</td>
</tr>
<tr>
<td>NP-03</td>
<td>High</td>
<td>0.4324</td>
<td>0.5833</td>
</tr>
<tr>
<td>NP-04</td>
<td>Low</td>
<td>1.0000</td>
<td>0.5833</td>
</tr>
<tr>
<td>NP-05</td>
<td>High</td>
<td>0.3649</td>
<td>0.4722</td>
</tr>
<tr>
<td>NP-06</td>
<td>High</td>
<td>0.3919</td>
<td>0.6111</td>
</tr>
<tr>
<td>NP-07</td>
<td>High</td>
<td>0.5135</td>
<td>0.5833</td>
</tr>
</tbody>
</table>

For 44 NPs, size, zeta potential and relaxitivities were available, and then normalized between 0 and 1, to form the QNTR matrix.
CS1. QNTR modeling results of 44 diverse NPs using MML-WinSVM and a 5 fold external cross-validation

<table>
<thead>
<tr>
<th>Fold</th>
<th>n</th>
<th># models</th>
<th>% accuracy internal 5-fold CV</th>
<th>% accuracy</th>
<th>n</th>
<th>% accuracy</th>
<th>% CCR&lt;sup&gt;a&lt;/sup&gt;</th>
<th>% Sensitivity</th>
<th>% Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>11</td>
<td>51.4 – 60.0</td>
<td>71.4 – 82.9</td>
<td>9</td>
<td>78</td>
<td>83</td>
<td>67</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>13</td>
<td>51.4 – 60.0</td>
<td>71.4 – 77.1</td>
<td>9</td>
<td>78</td>
<td>75</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>16</td>
<td>57.1 – 62.9</td>
<td>74.3 – 82.9</td>
<td>9</td>
<td>78</td>
<td>78</td>
<td>80</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>11</td>
<td>60.0 – 62.9</td>
<td>77.1 – 88.6</td>
<td>9</td>
<td>56</td>
<td>55</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>4</td>
<td>66.7</td>
<td>83.3 – 86.1</td>
<td>8</td>
<td>75</td>
<td>67</td>
<td>33</td>
<td>100</td>
</tr>
</tbody>
</table>

<sup>a</sup>CCR – Correct Classification Rate.

Prediction performances are surprisingly good: the overall prediction accuracy for those 44 NPs is equal to 73%
QSAR and toxicity prediction: QSAR Modeling* of the TETRATOX aquatic toxicity

- A short-term, static protocol using the common freshwater ciliate Tetrahymena pyriformis (strain GL-C) to test aquatic toxicity.
- The 50% impairment growth concentration (IGC50) is the recorded endpoint.
- Website: http://www.vet.utk.edu/TETRATOX/

*Zhu et al, JCIM, J Chem Inf Model 2008; (48): 766-784
International Virtual Collaboratory* of Computational Chemical Toxicology

- **USA**: UNC-Chapel Hill (UNC) - H. Zhu and A. Tropsha
- **France**: University of Louis Pasteur (ULP) – D. FOURCHES and A. VARNEK
- **Italy**: University of Insubria (UI) – E. PAPA and P. GRAMATICA
- **Sweden**: University of Kalmar (UK) – T. ÖBERG
- **Germany**: Munich Information Center for Protein Sequences/Virtual Computational Chemistry Laboratory (VCCLAB)– I. TETKO
- **Canada**: University of British Columbia (UBC) – A. CHERKASOV

*a new networked organizational form that also includes social processes; collaboration techniques; formal and informal communication; and agreement on norms, principles, values, and rules
Different countries, different groups, different tools – shared basic principles

- Explore and combine various QSAR approaches
- Use extensive model validation and applicability domains
- Consider external prediction accuracy as the ultimate criteria of model quality

\[
Q_{abs}^2 = 1 - \frac{\sum_Y (Y_{\text{exp}} - Y_{\text{LOO}})^2}{\sum_Y (Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2} \quad (1)
\]

\[
R_{abs}^2 = 1 - \frac{\sum_Y (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum_Y (Y_{\text{exp}} - \langle Y \rangle_{\text{exp}})^2} \quad (2)
\]

\[
\text{MAE} = \sum_Y \left| Y - Y_{\text{pred}} \right| / n \quad (3)
\]
## Overview of the Approaches (15 methodologies total)

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Modeling Techniques</th>
<th>Descriptor Type</th>
<th>Applicability Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNC</td>
<td>$k$NN, SVM</td>
<td>MolConnZ, Dragon</td>
<td>Euclidean distance threshold between a test compound and compounds in the modeling set</td>
</tr>
<tr>
<td>ULP</td>
<td>MLR, $k$NN, SVM</td>
<td>Fragments</td>
<td>Euclidean distance threshold between a compound and compounds in the modeling set; bounding box</td>
</tr>
<tr>
<td>UI</td>
<td>OLS</td>
<td>Dragon</td>
<td>Leverage approach</td>
</tr>
<tr>
<td>UK</td>
<td>PLS</td>
<td>Dragon</td>
<td>Residual standard deviation and leverage within the PLSR model</td>
</tr>
<tr>
<td>MIPS</td>
<td>ASNN</td>
<td>E-state</td>
<td>Maximal correlation coefficient of the test molecule to the training set molecules in the space of models</td>
</tr>
<tr>
<td>UBC</td>
<td>MLR, ANN, SVM, PLS</td>
<td>IND_I</td>
<td>Descriptor variability</td>
</tr>
</tbody>
</table>
# Individual vs. Consensus Models for the Modeling Set

<table>
<thead>
<tr>
<th>Model</th>
<th>Group ID</th>
<th>$q^2$</th>
<th>SE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$NN-Dragon</td>
<td>UNC</td>
<td>0.93</td>
<td>0.23</td>
<td>100%</td>
</tr>
<tr>
<td>$k$NN-MolconnZ</td>
<td>UNC</td>
<td>0.92</td>
<td>0.26</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM-Dragon</td>
<td>UNC</td>
<td>0.93</td>
<td>0.26</td>
<td>100%</td>
</tr>
<tr>
<td>SVM-MolconnZ</td>
<td>UNC</td>
<td>0.89</td>
<td>0.33</td>
<td>100%</td>
</tr>
<tr>
<td>$k$NN-Fragmental</td>
<td>ULP</td>
<td>0.77</td>
<td>0.44</td>
<td>100%</td>
</tr>
<tr>
<td>SVM-Fragmental</td>
<td>ULP</td>
<td>0.95</td>
<td>0.23</td>
<td>100%</td>
</tr>
<tr>
<td>MLR</td>
<td>ULP</td>
<td>0.94</td>
<td>0.25</td>
<td>100%</td>
</tr>
<tr>
<td>MLR-CODESSA</td>
<td>ULP</td>
<td>0.72</td>
<td>0.47</td>
<td>100%</td>
</tr>
<tr>
<td>OLS</td>
<td>UI</td>
<td>0.86</td>
<td>0.35</td>
<td>92.1%</td>
</tr>
<tr>
<td>PLS</td>
<td>UK</td>
<td>0.88</td>
<td>0.34</td>
<td>97.7%</td>
</tr>
<tr>
<td>ASNN</td>
<td>MISP</td>
<td>0.92</td>
<td>0.27</td>
<td>83.9%</td>
</tr>
<tr>
<td>PLS-IND_I</td>
<td>UBC</td>
<td>0.76</td>
<td>0.39</td>
<td>100%</td>
</tr>
<tr>
<td>MLR-IND_I</td>
<td>UBC</td>
<td>0.77</td>
<td>0.39</td>
<td>100%</td>
</tr>
<tr>
<td>ANN-IND_I</td>
<td>UBC</td>
<td>0.77</td>
<td>0.39</td>
<td>100%</td>
</tr>
<tr>
<td>SVM-IND_I</td>
<td>UBC</td>
<td>0.79</td>
<td>0.31</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Consensus Model</strong></td>
<td>-</td>
<td>0.92</td>
<td>0.22</td>
<td>100%</td>
</tr>
</tbody>
</table>
Which model is best?

- Observation: Models that afford most accurate predictions for the validation sets are not necessarily ranked as top models for the modeling set.

- Back to choices and practices: So how do we choose “the best” models?

  Should we choose!?!?

- Consensus Prediction
  - Only predict compounds within the applicability domain of most models
  - For each compound, exclude predictions that have high deviations from the mean value
  - Final predicted value is the average over all predictions.
Consensus Model gives the lowest MAE of prediction (Validation Set)
Principles of “Safe” QSAR modeling

• Establish an SAR database (target property, descriptor set).
• Rationally divide the dataset into training and test sets
• Develop training set models and characterize them with internal validation parameters.
• Validate training set models using external test set and calculate the external validation parameters.
• Finally, explore and exploit validated QSPR models for possible mechanistic interpretation and prediction.*

Important vs. Less Important Directions in QSAR modeling (it is about PREDICTIONS)

• Less important (model development)
  – Descriptor development and/or integration (some exceptions)
  – “novel” data analytical techniques
  – Training set statistics
  – (Harmonizing) definitions (SAR, QSAR, etc.)
  – Mechanistic interpretation (except for validated models)

• More important (model validation)
  – Quality and representation of biological data
  – Analysis of common descriptors and most successful combinations (of descriptors and data modeling techniques) that increase the experimental hit rate
  – Training vs. test vs. evaluation set selection (three-way)
  – Outlier analysis (experimental accuracy or descriptor incapability)
  – Applicability domain (in the context of modeling technique AND TEST SET STATISTICS)
  – The real power of QSAR models is in their ability to design novel active compounds or identify such compounds in databases or virtual libraries

• Independent model evaluation in competitive fashion: CoErPA (similar to CASP) and benchmark dataset depository
Final Word

Nothing that worth knowing can be taught.  

Oscar Wilde

• **Best time ever to be a cheminformatics scholar**
  – Growth of databases
  – Tool development
  – Collaborations with computational and experimental scientists

• **Extending** cheminformatics approaches to new areas
  – Structure based virtual screening
  – “-omics” data analysis
  – Structure – in vitro – in vivo correlations
  – Toxico-cheminformatics

• **Focus on Knowledge Discovery** (accurate testable predictions!) in Chemical Databases