The Validation & Interpretation of QSAR Models

Rajarshi Guha

Department of Chemistry
Pennsylvania State University

December 13, 2004

Outline

1. An Introduction to QSAR
2. Validating QSAR Models
3. Interpreting QSAR Models

Outline

1. An Introduction to QSAR
   - The Goals of QSAR
   - QSAR Methodology
   - An Application of the Methodology

2. Validating QSAR Models

3. Interpreting QSAR Models

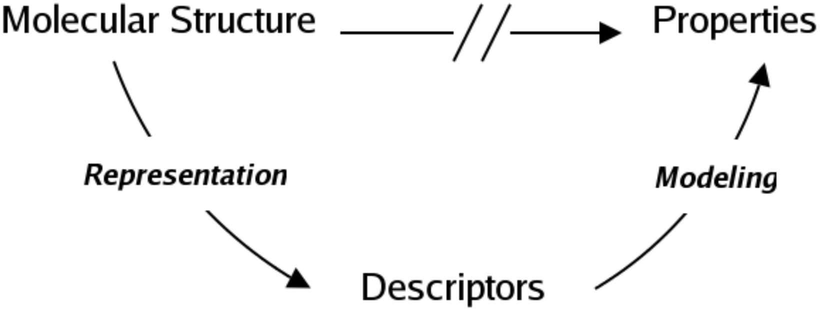
What is the Aim of a QSAR / QSPR Model?

- Predict properties of molecules or classifiy molecules based on structural features
- Properties can include
  - Physical properties like *boiling point* or *aqueous solubility*
  - Biological activities like *carcinogenicity* or LD_{50}
- QSAR modeling can be considered to be an application of data mining

Why do We Need QSAR Models?

- Compound screening, especially for virtual libraries
- ADME/Tox modeling - *fail early, fail cheap* principle
- Can be used to focus on specific compounds
- A model can provide insight into mechanism or mode of action

The QSAR Pipeline



Molecular Structure

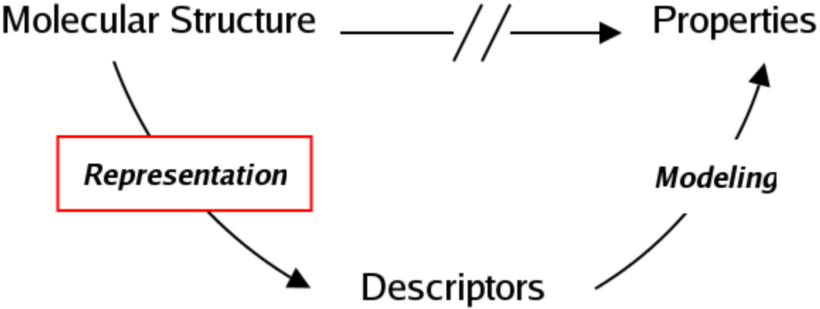
Properties

Representation

Modeling

Descriptors

The QSAR Pipeline



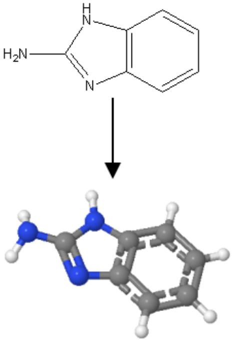
Structure Representation

Data Entry

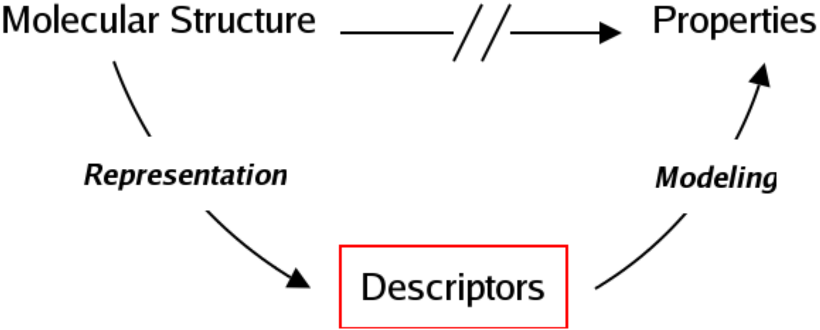
- Directly draw 3D structures in Hyperchem
- Convert 2D structures (e.g., SMILES) to 3D using Corina or Concord

Structure Optimization

- Geometry optimization is carried out using MOPAC with the PM3 Hamiltonian
- Electronic optimization uses the AM1 Hamiltonian



The QSAR Pipeline



Molecular Descriptors

- Molecular descriptors can be broadly divided into 3 groups
  - Topological
  - Geometric
  - Electronic
- The above types can be combined to generate hybrid descriptors

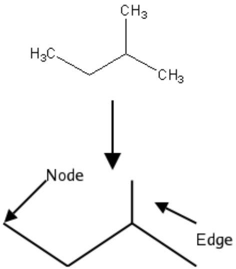
Molecular Descriptors - Topological

Characteristics

- Considers a molecule as a graph
- The descriptors are various graph invariants

Examples

- Connectivity indices
- Substructure counts
- Path length descriptors



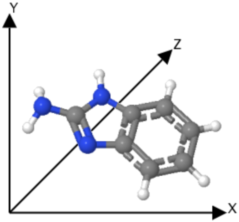
Molecular Descriptors - Geometric

Characteristics

- Characterizes the geometry of the molecule
- Dependent on accurate 3D conformations

Examples

- Moments of inertia
- Molecular surface area and volume
- Length to breadth ratio



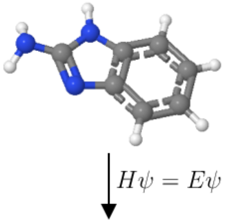
Molecular Descriptors - Electronic

Characteristics

- Derived from *ab initio* or semi-empirical calculations
- Characterizes the electronic environment of a molecule

Examples

- HOMO energies
- Dipole moments
- Partial charges



$H\psi = E\psi$

Charges
Dipole moments
HOMO / LUMO Energies
Electronegativity

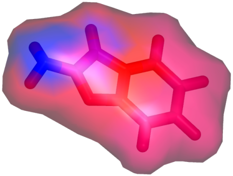
Molecular Descriptors - Hybrid

Characteristics

- These descriptors usually combine electronic features and geometric or topological features
- These descriptors are usually information rich

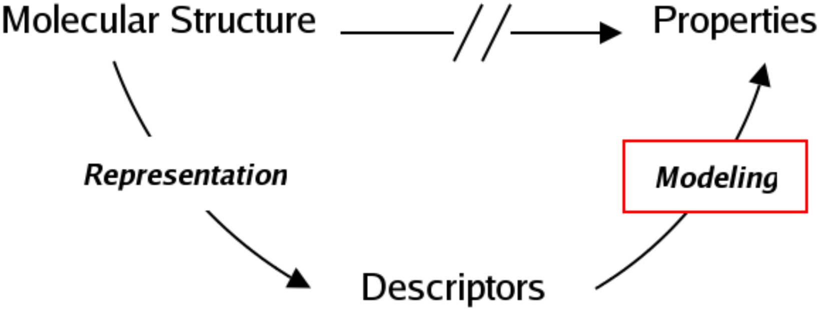
Examples

- Charged Polar Surface Areas
- Hydrophobic Polar Surface Areas
- H-bond descriptors



Hydrophobic Surface Area

The QSAR Pipeline



Building Predictive Models

- At this stage we have a large pool of descriptors for each molecule
- Before we build a predictive model we need to reduce this pool to work with *relevent* and *information rich* descriptors
- Thus modeling can be broken into two steps:
  - Feature selection
  - Model development

Building Predictive Models - Feature Selection

Objective

- Uses only independent variables
- Correlation test
- Identical test
- Vector space analysis

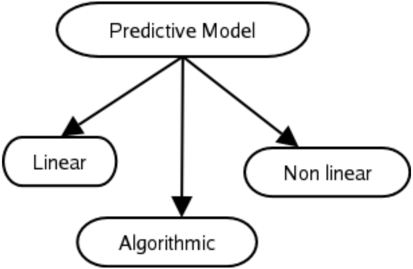
Subjective

- Uses the dependent variable
- Searches for good descriptor subsets
- Genetic algorithms
- Simulated annealing

Model Development

Model Characteristics

- Complexity
- Computational needs
- Flexibility
- Accuracy



Model Development

Linear Models

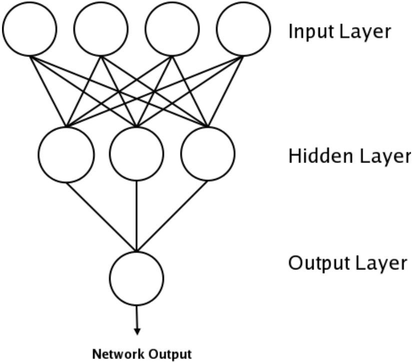
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

- Multiple linear regression, PLS, . . .
- Simple and fast to compute
- Not very flexible
- Amenable to interpretation

Model Development

Non-linear Models

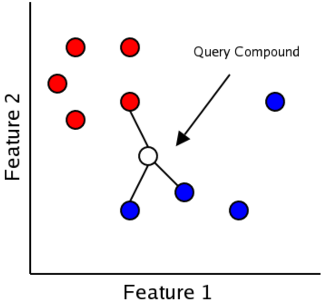
- Neural networks
- Models are complex and computationally intensive to train
- Very flexible
- Black box methodology



Model Development

Algorithmic

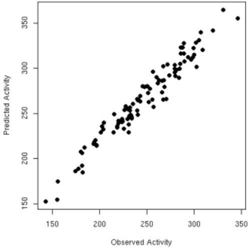
- *k*NN, random forests, . . .
- Models are of low complexity and rapid to compute
- Very flexible
- Can be interpreted in some cases

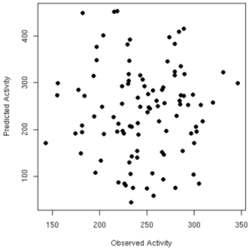


Model Validation

Y Scrambling

- This procedure ensures that the model is not due to chance
- Scramble the dependent variable (Y) and make predictions
- A random scatter plot indicates that the model was probably not due to chance

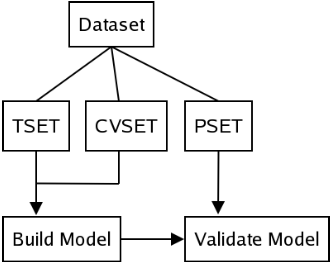




Model Validation

External Prediction Test

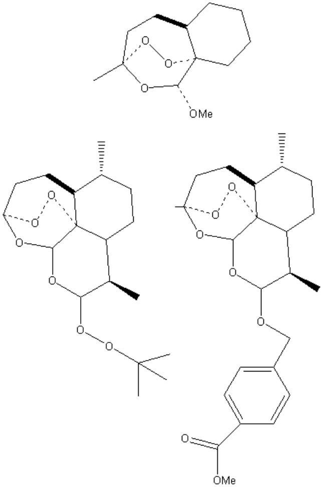
- This procedure tests the model's generalizability
- The PSET is used *only* during this stage
- Characterizes the behavior of the model when faced with new data



Artemisinin Dataset

- Potent analogs of artemisinin (anti-malarial) exist but are neurotoxic
- The original dataset was studied using CoMFA to try and design less toxic analogs

- 179 analogs of artemisinin
- Measured property was the logarithm of the relative activity
- A number of molecules had the same value of log RA but diverse structures



Avery, M.A. et al., *J. Med. Chem.*, **2002**, *45*, 292-303

QSAR Preliminaries

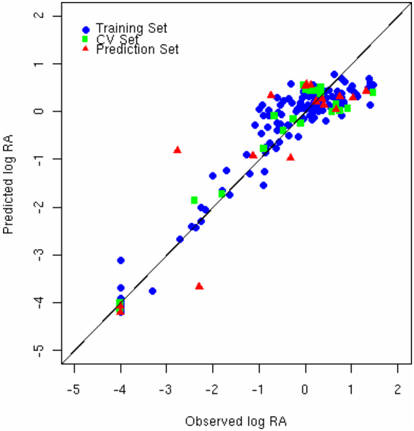
- 179 molecules were divided into:
    - TSET - 144
    - CVSET - 17
    - PSET - 18
- The sets were generated using an activity binning method
- 299 ADAPT descriptors calculated, reduced to 65 descriptors
- Linear and non-linear models were built

Guha, R.; Jurs, P.C; *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1440-1449

Summary of the Best CNN Model

- The model architecture was 10-5-1
- Relatively complex model
- Good statistics

	R^2	RMSE
TSET	0.96	0.47
PSET	0.88	0.74



Guha, R.; Jurs, P.C; *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1440-1449

Outline

1. An Introduction to QSAR

2. Validating QSAR Models
   - Extending Model Validation
   - Approaches to Model Applicability
   - A Classification Approach

3. Interpreting QSAR Models

Types of Validation

Model Validation

- Goal is to test the reliability of the model
- Ensures that the model is not due to chance factors
- Based on dataset used to develop the model

Types of Validation

Model Validation

- Goal is to test the reliability of the model
- Ensures that the model is not due to chance factors
- Based on dataset used to develop the model

Model Applicability

- Goal is to test the applicability of the model to new data
- Tells us: The model will predict the activity well (or not)
- Similar to confidence measures

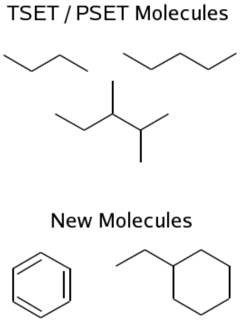
Why Isn’t Model Validation Enough?

Training

- Aim is to capture molecular features related to activity
- Features not captured by the model will not be recognized

Prediction

- The PSET is used to see how well the model captured molecular features
- PSET is taken from the same dataset as the TSET
- It will have features in common with the TSET



Extrapolation Is Not A Good Idea!

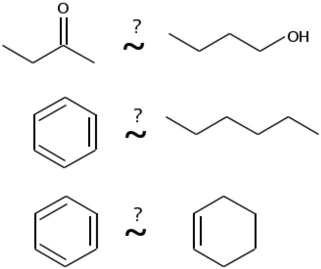
What Is Model Applicability?

Question?

How will a model perform when faced with molecules that it has not been trained on or validated with?

Aspects

- Similarity to the TSET?
- Structural or statistical similarity?
- Quantitative or qualitative?



How To Assess Model Applicability

Define Model Performance

Performance is measured by prediction residuals. The model performs well on a new molecule if it predicts its activity with low residual error.

Correlate *'X'* With Performance

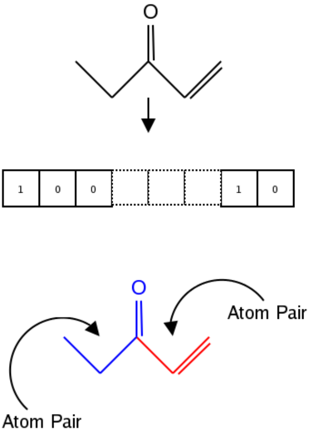
- 'X' could be similarity between an unseen molecule and the original training set
- 'X' could be derived from a cluster membership approach
- Alternatively, *predict performance* itself

Guha, R.; Jurs, P.C; *J. Chem. Inf. Comput. Sci.*, in press

Structural Similarity As a Measure of Applicability

Features

- Intuitive
- Evaluate fingerprints or atom pairs
- Use variety of similarity measures
- Correlate similarity to prediction residuals



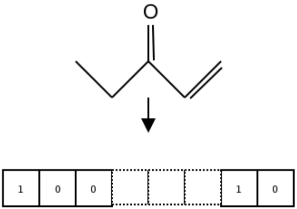
Structural Similarity As a Measure of Applicability

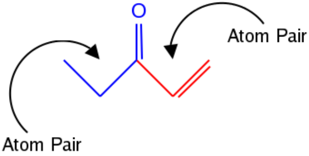
Features

- Intuitive
- Evaluate fingerprints or atom pairs
- Use variety of similarity measures
- Correlate similarity to prediction residuals

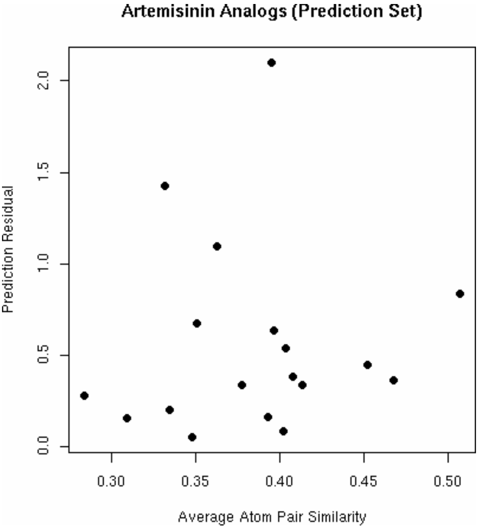
Problem!

Does not work (well)





Structural Similarity As a Measure of Applicability



Classifying Performance

Why?

- Our interest is in the model itself
- We can quantify applicability

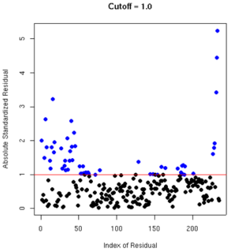
How?

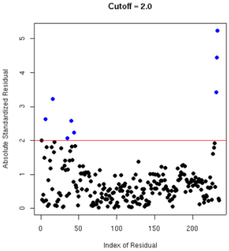
1. Consider residuals for TSET
2. Choose a cutoff - residuals above the cutoff are bad and below are good
3. Build a classifier with these class assignments
4. Predict class of residual for new molecules

Classifying Performance

Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?





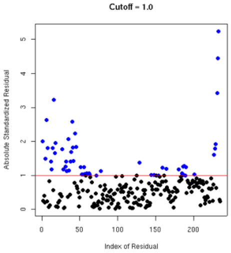
Classifying Performance

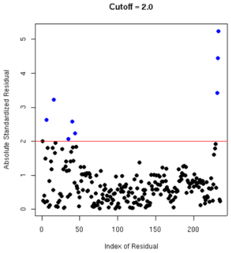
Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?

Possibilities

- Visual inspection
- Regression diagnostics





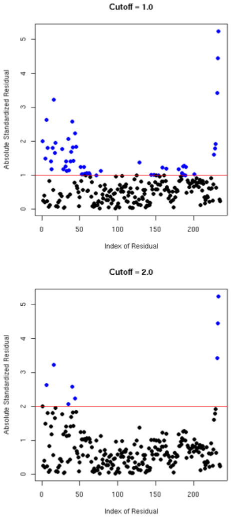
Classifying Performance

Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?

Possibilities

- Depends on the size of the dataset
- More classes allow for finer analysis



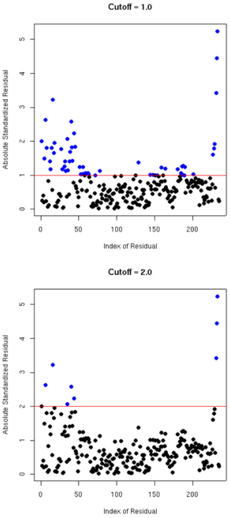
Classifying Performance

Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?

Possibilities

- Linear: LDA and PLS
- Non-linear: CNN



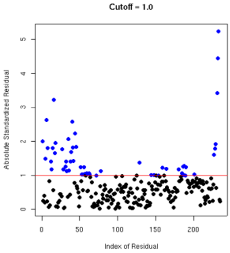
Classifying Performance

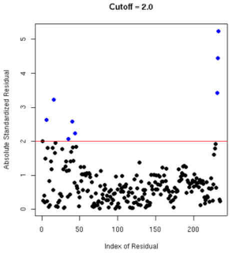
Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?

Possibilities

- Oversampling or undersampling
- Use pseudo convex data





Breiman, L., *Technical Report 513*, **1998**, Dept. of Statistics, UC Berkely

Classifying Performance

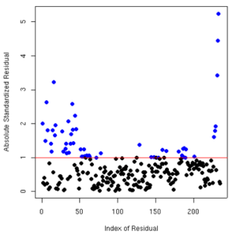
Choices

- How do we choose a cutoff?
- How many classes do we take?
- What classifier do we use?
- How do we handle unbalanced classes?
- Which descriptors do we use for the classifier?

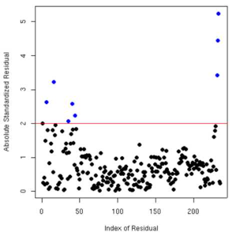
Possibilities

- The descriptors used in the original model
- Global descriptors

Cutoff = 1.0



Cutoff = 2.0



Methodology

Choices Made

- Cutoffs obtained via visual inspection giving 2 classes
- PLS, LDA, CNN
- Pseudo convex data
- Descriptors from the original models
- Original models were linear regression

Datasets

- Boiling point (TSET = 235, PSET = 42)
- Activity of artemisinin analogs (TSET = 161, PSET = 18)

Goll, E.S. et al., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 974-983

Avery, M.A. et al., *J. Med. Chem.*, **2002**, *45*, 292-303

Results

Class Breakup

		Class Size	
Dataset	**Cutoff**	**Good**	**Bad**
Artemisinin	1.0	133	46
Boiling Point	1.0	213	64

Results

Weighted Success Rates

Method	Dataset	TSET	PSET
LDA	Artemisinin	0.51	0.50
	Boiling Point	0.52	0.53
PLS	Artemisnin	0.51	0.46
	Boiling Point	0.36	0.53
CNN	Artemisinin	0.79	0.80
	Boiling Point	0.98	0.93

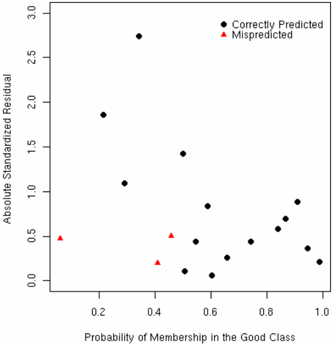
Weston, J. et al., *Bioinformatics*, **2003**, *19*, 764-771

Results

Artemisinin / CNN Classifier (4-3-1)

TSET	**Predicted**	
Actual	bad	good
bad	38	4
good	27	92

PSET	**Predicted**	
Actual	bad	good
bad	4	0
good	3	11

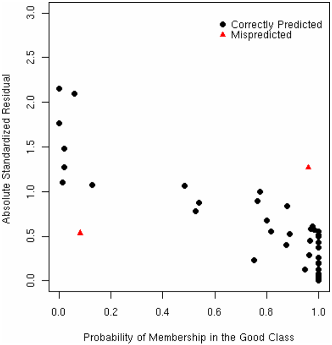


Results

Boiling Point / CNN Classifier (4-3-1)

TSET	**Predicted**	
Actual	bad	good
bad	54	0
good	5	176

PSET	**Predicted**	
Actual	bad	good
bad	9	1
good	1	31



Summary

- Model validation is required to ensure model reliability
- Model applicability allows us to decide whether the model will be useful for new data
- The classification approach can be applied to *any* quantitative model
- The role of structural similarity needs further investigation

Outline

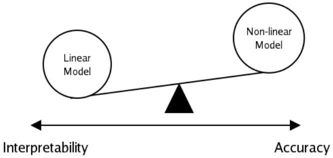
1. An Introduction to QSAR
2. Validating QSAR Models
3. Interpreting QSAR Models
   - The Problem of Interpretation
   - Creating a Model To Interpret
   - Aspects of the Interpretation

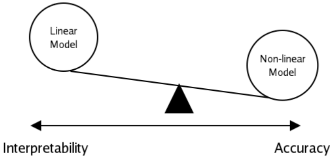
Isn't a Prediction Enough?

- Predictive models are good for screening purposes
- To understand *why* a compound is active we need an interpretation
- Interpretation is one way to approach the inverse QSAR problem
- Interpretability depends on modeling technique & descriptors involved

Interpretability and Accuracy

- Interpretability generally involves a trade off with accuracy
- Linear regression models are amenable to interpretation, but not very accurate
- Neural networks are black boxes, but are more accurate
- Some techniques lie in between (random forests)





Aspects of Interpretability

Broad Interpretation

- Essentially describes which descriptors are important
- Good for understanding which descriptors to focus on
- Based on randomization

Detailed Interpretation

- Describes how the property (activity) relates to the descriptor
- Gives us conclusions like:
  high value of DESC leads to **low** values of activity
- Allows for a detailed understanding of the SAR in QSAR

Partial Least Squares Based Interpretation

PLS Overview

- Creates a model with *latent variables*
- Latent variables (components) are linear combinations of the original variables (X's)
- Each latent variable is used to predict a *pseudo* dependent variable (Y's)

Interpretation

- The linear model is subjected to PLS analysis
- This also *validates* the model
- Choose the number of components to use
- Interpretation uses the X-weights, X-scores & Y-scores

Stanton D.T.; *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1423-1433

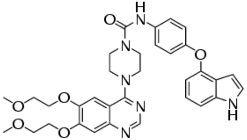
Dataset

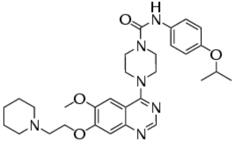
Overview

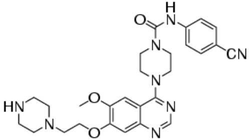
- 79 derivatives of 4-piperazinylquinazoline
- PDGFR phosphorylation inhibitors
- Measured activity was IC_{50}

QSAR Details

- Divided into training set (57), cross validation set (9) and prediction set (13)
- Final reduced pool had 41 descriptors
- Dependent variable was $-\log(IC_{50})$
- Linear regression, CNN and random forest models were built







Pandey, J. et al.; *J. Med. Chem.*, **2002**, *45*, 3772-3793

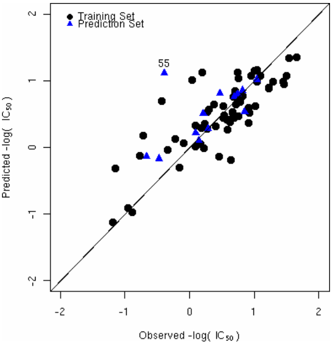
Linear Model

Statistics

- $R^2 = 0.65$, $\mathrm{RMSE} = 0.38$
- F-statistic $= 37.06$ (3,59)

Descriptors

- MDEN-23 - distance edge between N atoms
- RNHS-3 - relative hydrophilic SA
- SURR-5 - ratio of weighted hydrophobic SA to weighted hydrophilic SA



Guha, R., Jurs, P.C.; *J. Chem. Inf. Comput. Sci.*, ASAP

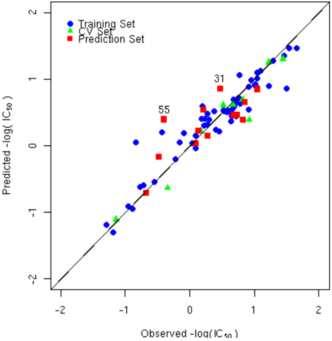
Alternate Models

CNN Model

- 7-3-1 architecture
- $R^2 = 0.94$, $\text{RMSE} = 0.22$
- 2 descriptors (RNHS-3, SURR-5) in common with the linear model

Random Forest Model

- Used to investigate descriptor importance
- Predictive ability not significantly better than other models



PLS Interpretation

Choosing components

- Q^2 allows us to choose how many components
- For a valid model cumulative variance should be 1

	X variance	R^2	Q^2
C1	0.51	0.52	0.45
C2	0.78	0.60	0.56
C3	1.00	0.61	0.56

Descriptor Weights

- Descriptors are ranked by their weights
- Sign of weight indicates how the descriptor correlates to predicted activity

Desc	C1	C2	C3
MDEN-23	-0.16	0.93	0.30
RNHS-3	0.55	-0.17	0.81
SURR-5	-0.82	-0.29	0.48

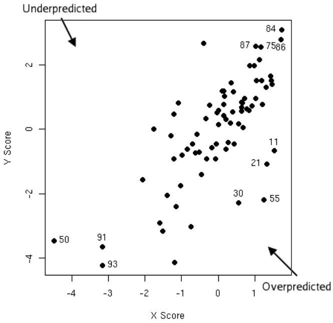
PLS Interpretation

Component 1

- SURR-5 is most weighted
- Low values of SURR-5 ⇒ high values of predicted activity

Interpretation

- Active compounds have high absolute values of SURR-5
- Indicates large hydrophobic surface area
- Consistent with cell based assay which depends on cell membrane transport

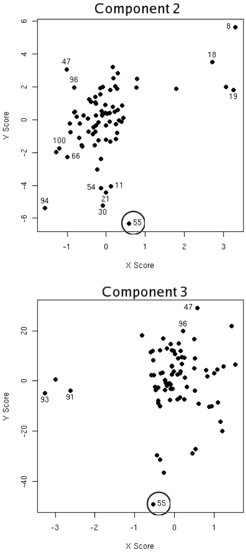


	MDEN-23	RNHS-3	SURR-5
C1	-0.16	0.55	-0.82

PLS Interpretation

Understanding Outliers

- Compound 55 is mispredicted by each component
- It is also an outlier in both linear & CNN models
- Has high absolute value of SURR-5 but low measured activity



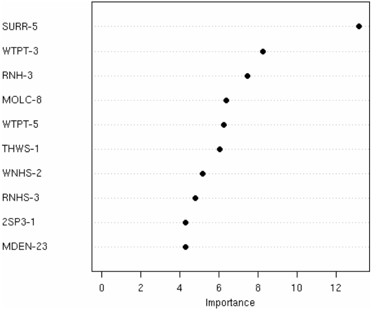
Descriptor Importance

Random Forest Interpretation

- The RF provides a measure of descriptor importance
- Utilizes the whole descriptor pool & ranks the descriptors
- SURR-5 is indeed important

Features of the CNN Model

- The 2 most important descriptors from the RF model are in the CNN model
- Other 5 descriptors are in the top 20 ranked descriptors



Summary

- Interpretability is required to fully utilize 2D QSAR models
- Dependent on model type and descriptors involved
- The model is fundamentally 2D, so we cannot explore 3D features affecting the SAR using this scheme

Conclusions

- Validation & interpretation are two important stops on the QSAR pipeline

- Validation is required to assess reliability & applicability
- A classification approach to validation is quite general in nature and performs well

- Interpretation plays an important role in drug *design*
- Linear regression models have been shown to be easily interpretable
- Work is on to create an interpretation scheme for CNN models

Acknowledgemts

- Dr. Peter Jurs
- Dr. Brian Mattioni
- Dr. Jon Serra
- NSF