PROPERTY AND POTENCY PREDICTION BY GENERALISED SIMILARITY

David Buttar
Oncology iMED
AstraZeneca

Geoff Skillman
Openeye Scientific Software

Terry Stouch
Science for Solutions

UK-QSAR Autumn Meeting, 2013
Outline

• Background to Project
• Datasets Studied
• Prediction by Generalised Similarity (PGS) approach
• Results
• Conclusions
• Next Steps
Introduction

• Drug Design often based on the similarity principle

"Structurally similar compounds are more likely to exhibit similar properties”

• Nearest-neighbour Models (NN) are the simplest QSAR Models

• No parameters other than the choice of similarity measure
  • NN approaches avoid the N-parameters Structural Risk Issue

• Despite their simplicity can they be improved?
  • Prediction by Generalised Similarity applicable with large data sets of molecular pairs with measured data

[Diagram of No. Parameters vs. No. Data Points with Structural Risk gradient]
Effective Nearest Neighbour QSAR models

- Work Reported by Abbott labs with Openeye indicated benefits of effective NN

Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping

- The average $R^2$ is over fourteen standard deviations higher than NN - less than 0.5% chance of NN being better than PGS.

Anthony Nicholls - OpenEye Scientific Software (EuroQSAR 2010)

- Collaboration with Openeye Scientific to Explore this approach with AstraZeneca datasets

- How much better than a best Nearest neighbor model are our QSAR methods?
PGS: Prediction by General Similarity

Use the individual distributions for each molecule to build up a composite prediction of the likely property.

*Anthony Nichols - OpenEye Scientific Software (EuroQSAR)*
The Experiment

- Can simple NN Models be Improved though the PGS approach?
- Utilise AstraZeneca Data sets as test case.
- Comparison of PGS models to simple NN and AZ inhouse QSAR models.

Hypothesis: PGS models perform better than simple NN models.
### Data Sets

<table>
<thead>
<tr>
<th>Model</th>
<th>Area</th>
<th>Results</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Solubility</em></td>
<td>Phys props</td>
<td>32784</td>
<td>Thermodynamic solubility@pH 7.4 from solid.</td>
</tr>
<tr>
<td><em>LogD</em></td>
<td>Phys props</td>
<td>25165</td>
<td>LogD@pH7.4 – shake flask approach</td>
</tr>
<tr>
<td>†Herg (pIC50)</td>
<td>Safety</td>
<td>57468</td>
<td>Herg Channel inhibition electrophysiology IonWorks™ device</td>
</tr>
<tr>
<td>Kinase (pIC50)</td>
<td>Project</td>
<td>30203</td>
<td>Kinase enzyme Inhibition Caliper off-chip incubation mobility shift assay</td>
</tr>
</tbody>
</table>

- All data supplied to Openeye with meta data on assay protocols, assay standards, compound purity, compound properties.
- Selection of Training sets for PGS models was based on data only.
- No compound structural information passed to Openeye.


Datasets
Data/Training Split

Solubility
- N = 10987 (32784)
- μ = 4.5 (4.5)*
- σ = 0.9 (1.2)
- Exp error = ± 0.3

Herg pIC50
- N = 11224 (57468)
- μ = 5.1 (5.6)
- σ = 0.5 (0.71)
- Exp error = ± 0.3

logD
- N = 8667 (25165)
- μ = 2.6 (2.6)
- σ = 0.9 (1.0)
- Exp error = ± 0.2

FGFR1 pIC50
- N = 7189 (30203)
- μ = 5.6 (5.3)
- σ = 1.1 (1.3)
- Exp error = ± 0.3

* Complete Population
Parameters in parenthesis

- Training data sets used to construct AZ- AutoQSAR models and NN models.
Data Sets – Diversity Analysis

- Cluster Analysis of the Training Data sets
- AZ – path based fingerprints approach

![Graph showing similarity distributions for Solubility, FGFR1, LogD, and Herg with probabilities and cluster sizes.]
PGS Approach
Fitting Property Similarity Profiles

Property (Y)

Non-parametric regression via Nadaraya-Watson (NW) Kernel Regression

• Provides a similarity based weighting scheme based on training set data. (local averaging)

Multivariate NW Estimator

\[ \hat{m}_h(x) = \frac{1}{n} \sum_{i=1}^{n} W_{hi}(x) Y_i \]

\[ W_{hi}(x) = \frac{n K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)} \]

• Key is optimizing the \( \alpha \) parameter (bandwidth)
• Only optimized parameter in the PGS model

\[ K(x/\alpha) = (2\pi)^{-1/2} e^{-x^2/2\alpha^2} \]

Nonparametric Regression Applied to Quantitative Structure–Activity Relationships
Nadaraya-Watson Kernel Regression

• For each training data set
  • 10% Excluded for an external Test Set Validation.
  • Remaining 90% of Training Set
  • Loop over N (default =40)
    • 90% Train / 10% test
    • Loops over kernel α values (min, max, count – user controlled)
    • Use lowest upper 95% CI of mean residuals to select optimum α
  • Applies “best” α to external validation set and evaluates key criteria for model performance
PGS - Fitting Training Set Data
(Solubility/Gaussian/Circular FP)

- Optimization of \( \alpha \) parameter for each Kernel/FP
- Each point is data from 40 runs (random 90/10 train:test split)

Function defining Contribution to prediction for each compound

- Mean 95% Conf. Limit
- "best \( \alpha \)"

Gaussian Kernel \( (\alpha=4.36) \)

- Optimization of \( \alpha \) parameter for each Kernel/FP
- Each point is data from 40 runs (random 90/10 train:test split)

Looking for majority of points to be below the 95% CI

Prediction error 0.6
Kernel selection

Choice of kernel has little influence on prediction accuracy

Model locality has large effect on prediction accuracy
PGS Solubility Model

- Gaussian weight/MACCS166
- Mean residual on 10% external validation set = 0.6
- Correlation = 0.61 ($R^2 = 0.36$)

**Fingerprints**

**Root Mean Square Error**

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Path</th>
<th>Lingo</th>
<th>MACCS166</th>
<th>Tree</th>
<th>Circular</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiQuadratic</td>
<td>0.61</td>
<td>0.60</td>
<td>0.59</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>BoxCar</td>
<td>0.64</td>
<td>0.63</td>
<td>0.62</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td>Exponential</td>
<td>0.61</td>
<td>0.58</td>
<td>0.59</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>Gaussian</td>
<td>0.61</td>
<td>0.58</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Linear</td>
<td>0.62</td>
<td>0.61</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.63</td>
<td>0.60</td>
<td>0.60</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>SmoothBoxCar</td>
<td>0.62</td>
<td>0.60</td>
<td>0.57</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>TriCubic</td>
<td>0.61</td>
<td>0.60</td>
<td>0.59</td>
<td>0.61</td>
<td>0.60</td>
</tr>
</tbody>
</table>

- GraphSimToolKit utilised for similarity calculations.
- Choice of Kernel does not influence results significantly.
- $\alpha$ is the Key parameter
- Differences between approaches and therefore the degree of locality in the regression.

Regression to the mean for AZ_LogSol RMSE = 0.78
PGS - New Prediction

- Size of the weights applied during smoothing determined by $\alpha$

Local Averaging for prediction

Training Data

-pSol

Weights

Similarity

Width determined by $\alpha$

New Compound
PGS - New Prediction

- Local Averaging (weights > 0.2, N=151)
Benchmarking - AZ AutoQSAR Platform

## External Validation – PGS vs AutoQSAR

<table>
<thead>
<tr>
<th>Target</th>
<th>Modelling Method</th>
<th>Approach</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herg</td>
<td>AQ</td>
<td>SVM-Signature</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>PGS</td>
<td>Gaussian/Circular</td>
<td>0.27</td>
</tr>
<tr>
<td>Solubility</td>
<td>AQ</td>
<td>SVM-Signature</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>PGS</td>
<td>Gaussian/Circular</td>
<td>0.60</td>
</tr>
<tr>
<td>Kinase(FGFR1)</td>
<td>AQ</td>
<td>SVM-Signature</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>PGS</td>
<td>Gaussian/Circular</td>
<td>0.42</td>
</tr>
<tr>
<td>logD</td>
<td>AQ</td>
<td>SVM-Signature</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>PGS</td>
<td>Exponential/Tree</td>
<td>0.48*</td>
</tr>
</tbody>
</table>

- 10% random external test set utilised for Both methods
- Best performing AutoQSAR/PGS model reported for test set data.
- PGS best model in 3 End-points
- *In all cases PGS performed better than PLS/RF AQ approaches.
Results

• Additional External test set selected from Complete Datasets (excluding training set compounds)

• Compounds with at least 3 Near Neighbours as defined by AZ similarity metric. (Path based fingerprints ($t_c = 0.7$))

• Modelling within the domain of applicability

• Benchmarked against
  
  • Best AutoQSAR model built using same training set data (90/10 random train:test split)
  
  • Two simple K-NN models based on AZ path-based fingerprints
    
    • One-nearest neighbour Model (1-NN)
    • Mean five nearest neighbours (5-NN)
### Kinase (FGFR1) Results

External Validation Set, N = 4875

Model Accuracies  Summary Statistics

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>% ± 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.46</td>
<td>0.95</td>
<td>1.25</td>
<td>39.1</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.62</td>
<td>0.71</td>
<td>0.98</td>
<td>49.3</td>
</tr>
<tr>
<td>PGS Gauss/Circular</td>
<td>0.73</td>
<td>0.63</td>
<td>0.83</td>
<td>50.2</td>
</tr>
<tr>
<td>AutoQSAR SVM-Signature*</td>
<td>0.72</td>
<td>0.69</td>
<td>0.87</td>
<td>42.9</td>
</tr>
<tr>
<td>Regression to Mean</td>
<td>1.30</td>
<td>1.5</td>
<td>20.7</td>
<td></td>
</tr>
</tbody>
</table>

*MAE (Mean absolute error), RMSE (Root mean Squared error), % of compounds predicted with less than 0.5 log unit error, $R^2$ – correlation coefficient for line of fit.

*Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms
Hongming Chen; Lars Carlsson; Mats Eriksson; Peter Varkonyi; Ulf Norinder; Ingemar Nilsson;
Results – Kinase (FGFR1)

Number of Compounds in external Validation Set, N = 4875

- Improving Model Performance
- AutoQSAR/PGS models considered to be suitable for design.
Virtual Screening with PGS Models

- FGFR1 PGS Model used to Screen AstraZeneca Compound Collection.
- Requires similarity calculations for all compounds against PGS training set.
  - ~50 Billion similarity calculations
  - Intel Xenon X5650@2.67GHZ (Linux Workstation)
  - Each of 5 similarity measures runs on Thread (C++ OpenMP)
  - ~12 hours to complete the screen.
- PGS approach applicable for fast knowledge based similarity searching.
- Identified key Project Chemotypes, and untested series from other projects.
- Screened top 100 compounds in FGFR1 enzyme assay.
  - 61% active (sub 100nM)
  - Random back ground rate of 12%


†Protein–Ligand Crystal Structures Can Guide the Design of Selective Inhibitors of the FGFR Tyrosine Kinase. Richard A. Norman; Anne-Kathrin Schott; David M. Andrews; Jason Breed; Kevin M. Foote; Andrew P. Garner; Derek Ogg; Jonathon P. Orme; Jennifer H. Pink; Karen Roberts; David A. Rudge; Andrew P. Thomas; Andrew G. Leach; J. Med. Chem. 2012, 55, 5009-5012.
## Results – Herg

External Validation Set, N = 6414

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>% ± 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.13</td>
<td>0.41</td>
<td>0.53</td>
<td>68.0</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.22</td>
<td>0.33</td>
<td>0.42</td>
<td>78.5</td>
</tr>
<tr>
<td>PGS Gauss/Circular</td>
<td>0.27</td>
<td>0.29</td>
<td>0.37</td>
<td>83.5</td>
</tr>
<tr>
<td>AutoQSAR SVM-Signature</td>
<td>0.44</td>
<td>0.26</td>
<td>0.33</td>
<td>87.8</td>
</tr>
<tr>
<td>Regression to Mean</td>
<td>0.34</td>
<td>0.43</td>
<td>0.53</td>
<td>74.3</td>
</tr>
</tbody>
</table>

*MAE (Mean absolute error), RMSE (Root mean Squared error),
% of compounds predicted with less than 0.5 log unit error,
$R^2$ – correlation coefficient for line of fit.
Results – Herg

Number of Compounds in external Validation Set, N = 6414

- Improving Model Performance
- Not as marked as Kinase2 dataset
## Results – Solubility

External Validation Set, N = 8491

### Model Accuracies Summary Statistics

<table>
<thead>
<tr>
<th>Method</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>% ± 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.08</td>
<td>1.06</td>
<td>1.33</td>
<td>28.2</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.20</td>
<td>0.91</td>
<td>1.11</td>
<td>31.0</td>
</tr>
<tr>
<td>PGS Gauss/Circular</td>
<td>0.38</td>
<td>0.76</td>
<td>0.98</td>
<td>41.4</td>
</tr>
<tr>
<td>AutoQSAR SVM-Signature</td>
<td>0.67</td>
<td>0.53</td>
<td>0.71</td>
<td>55.6</td>
</tr>
<tr>
<td>Regression to Mean</td>
<td>1.05</td>
<td>1.21</td>
<td></td>
<td>24.5</td>
</tr>
</tbody>
</table>

*MAE (Mean absolute error), RMSE (Root mean Squared error), % of compounds predicted with less than 0.5 log unit error, $R^2$ – correlation coefficient for line of fit.*
Results – Solubility

(-logSol (Mol/litre))

- Solubility always difficult to predict
- NN-models no signal
- PGS outliers often ionization state related.

Number of Compounds in external Validation Set, N = 8491
## Results – logD

External Validation Set, N = 2297

<table>
<thead>
<tr>
<th>Method</th>
<th>R²</th>
<th>MAE</th>
<th>RMSE</th>
<th>% ± 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.05</td>
<td>1.02</td>
<td>1.34</td>
<td>33.6</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.10</td>
<td>0.88</td>
<td>1.13</td>
<td>35.6</td>
</tr>
<tr>
<td>PGS Exp/Tree</td>
<td>0.26</td>
<td>0.73</td>
<td>0.93</td>
<td>41.8</td>
</tr>
<tr>
<td>AutoQSAR SVM/Signatures</td>
<td>0.47</td>
<td>0.53</td>
<td>0.85</td>
<td>69.5</td>
</tr>
<tr>
<td>AutoQSAR RF/Mol. Descriptors</td>
<td>0.75</td>
<td>0.41</td>
<td>0.58</td>
<td>67.9</td>
</tr>
<tr>
<td>Regression to Mean</td>
<td></td>
<td>0.90</td>
<td>1.11</td>
<td>31.9</td>
</tr>
</tbody>
</table>

*MAE (Mean absolute error), RMSE (Root mean squared error), % of compounds predicted with less than 0.5 log unit error, R² – correlation coefficient for line of fit.*
Results – LogD

Number of Compounds in external Validation Set N = 2297

- Improving Model Performance
- PGS outliers often ionization state related.
Results

- External test set selected from Recent AstraZeneca Data
  - No filtering on molecular similarity,

- Prediction moving out of the domain of applicability

  - Projects Exploring New Areas of Chemistry
  - No FGFR1 project compounds
Results – FGFR1  (New Chemistry)

Number of Compounds in external Validation Set, N = 1753

Model Accuracies  Summary Statistics

<table>
<thead>
<tr>
<th>Method</th>
<th>R²</th>
<th>MAE</th>
<th>RMSE</th>
<th>% ± 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>0.06</td>
<td>1.05</td>
<td>1.32</td>
<td>29.2</td>
</tr>
<tr>
<td>5-NN</td>
<td>0.00</td>
<td>1.17</td>
<td>1.43</td>
<td>26.0</td>
</tr>
<tr>
<td>PGS Gauss/Circular</td>
<td>0.12</td>
<td>0.89</td>
<td>1.07</td>
<td>32.3</td>
</tr>
<tr>
<td>AutoQSAR SVM-Signature</td>
<td>0.25</td>
<td>0.81</td>
<td>1.03</td>
<td>40.7</td>
</tr>
<tr>
<td>Regression to Mean</td>
<td>0.94</td>
<td>1.13</td>
<td>28.4</td>
<td></td>
</tr>
</tbody>
</table>

*MAE (Mean absolute error), RMSE (Root mean Squared error),
% of compounds predicted with less than 0.5 log unit error,
R² – correlation coefficient for line of fit.

- Equivalent Observation seen with all Datasets
- Performance of all modelling approaches significantly deteriorates.
Results – Kinase2 (New Chemistry)

- As expected, model performance deteriorates as we move to new chemical space.
- Keeping models up-to-date improves the accuracy of our predictions.

Quantitative Structure–Activity Relationship Models That Stand the Test of Time
Andrew M. Davis; David J. Wood; Mol. Pharmaceutics 2013, 10, 1183-1190
Conclusions

Can simple NN Models be Improved through the PGS approach?

- For the AstraZeneca datasets studied results indicate PGS approach consistently improves simple Near-Neighbour models.

- For a simple 3-parameter model it can perform as well as classical multi-parameter QSAR approaches.

- Provide Intuitive similarity-based QSAR models

- PGS models are suitable Base-line QSAR Models for benchmarking against.
Next Steps

• Effect of Data Set Size/Selection on PGS predictions.
• Introduction of additional parameters to model
  • Improved treatment of Ionization
  • Expand to 3D similarity
• Explore transferability of PGS weights to alternative data sets
• Faster estimation of kernel alphas
• Automating PGS models
Acknowledgements

• AstraZeneca
  • Sorel Muresan
  • Thierry Kogej

• Openeye Scientific Software
  • Anthony Nicholls
  • Tom Darden