How can we get better at this?

Pat Walters – D3R Workshop
February 23, 2018
I’ve Done This a Few Times

How I Spend My Time On Challenges

- Dealing with poorly formatted submissions: 100%
- Validating evaluations: 0%
- Making Slides: 0%
The Evaluation Process

Pat Evaluate

Connor and Zied Evaluate

Final Comparisons
The Literature Makes It Look Like Activity Prediction is a Solved Problem

Pearson r

0.82

0.80

0.66

0.65
Scoring Performance From GC2 and GC3

Kendall's Tau

Spearman Rho

FXR 2017  CatS 2018

FXR 2017  CatS 2018
Guidelines For Reviewing "Scoring Function" Papers

We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
We need to agree on

- What constitutes a reasonable dataset
- How data should be reported
- Evaluation metrics
- Statistics for comparison
- What constitutes a null model
- Format of supporting material
- Criteria for reproducibility
When evaluating a regression model, the dataset should have a dynamic range similar to those observed in drug discovery projects (typically 4-6 logs).

This dataset (PDBind v.2016 core set) spans 10 logs and doesn’t provide an appropriate representation of correlation.
Correlations Can Change Dramatically With Dynamic Range

This is the same dataset. On the left we consider the entire set, which has an unrealistically large (~10 log) dynamic range. On the right we consider a more realistic subset with a 3 log dynamic range. Note the change in correlation.

\[ R^2 = 0.76 \]
\[ MAE = 0.55 \]

\[ R^2 = 0.22 \]
\[ MAE = 0.69 \]
GC3 CatS Dataset Spans a Realistic Dynamic Range
We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
Don’t Cram Multiple Datasets on to the Same Plot

http://pubs.acs.org/doi/abs/10.1021/ja512751q
Even My Friends Are Guilty

Docking vs Experiment


Mill and Neysa (Yesterday)
Trellising provides a much more effective means of comparing datasets
Guidelines For Reviewing "Scoring Function" Papers

We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
Always report correlations appropriately

Report Pearson, Spearman and Kendall correlations
Favor $R^2$ over $R$ when reporting a Pearson correlation coefficient
Report MAE and/or RMSE

Figure 3. QM/MM LIE calculated binding energy (kcal/mol) vs experimental binding energy (kcal/mol) for BACE1 (red squares), HSP90 (blue diamonds), PERK (orange triangles), and TYK2 (green triangles). The best-fitted line (dotted line), which has a correlation between measured and calculated values of 0.69, has slope = 0.82 and intercept = -1.79.

http://pubs.acs.org/doi/abs/10.1021/acs.jpcb.7b07224
Start with experimental data

Add Gaussian error
- Mean = 0.0
- Standard deviation = 0.3 log

Calculation correlation
Repeat 1000 times

Maximum Achievable Correlation - HPS90 D3R1

Maximum Achievable Correlation

Spearman Rank Correlation

0.940
0.950
0.960
0.970

Confidential | © 2017 Relay Therapeutics
Open Source Evaluation Code (More to Come)

https://github.com/PatWalters/metak

metk

Model Evaluation Toolkit

In metk, I've collected a set of routines for evaluating predictive models. I put a lot of this code together when I was doing the evaluation for the TDT and D3r projects, as well as a book chapter I wrote in 2013. I'm releasing this project as a way for the community to collaborate and (hopefully) agree on best practices for model evaluation. Most of the initial release is oriented toward the evaluation of free energy calculations.

This is just a start and I plan to add a lot more. Currently, there are routines to calculate

- Root mean squared (RMS) error
- Mean absolute error (MAE)
- Pearson correlation coefficient (with confidence limits)
- Spearman rank correlation (rho) (still need to add confidence limits)
- Kendall tau (still need to add confidence limits)
- Maximum possible correlation given a specific experimental error. This is based on a 2009 paper by Brown, Muchmore and Hajduk

Most of the statistics is done with routines from scikitlearn and scipy.

The toolkit also includes code to generate a few diagnostic plots that I find helpful when looking at model performance. Examples of these plots can be found here

- A scatter plot of experimental vs predicted ΔG. Lines are drawn at 1 and 2 kcal error
- A histogram of the error distribution.
- The two plots above with ΔG converted to a binding affinity (in uM or nM). On the scatter plot, lines are drawn at 5-fold and 10-fold error. I find that I mentally relate to a fold error in binding affinity better than I do to error expressed in kcal/mol. However, if you like looking at error in kcal/mol, use that plot.

Ultimately, the plan is to implement a number of other methods for model evaluation including those described in papers by Anthony Nicholls.
Guidelines For Reviewing "Scoring Function" Papers

We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• **Statistics for comparison**
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
In particular, both MM-PB/SA and MM-GB/SA produced better results by using a representative structure (R = 0.72-0.79) rather than averaging over the conformational ensemble of each given complex (R = 0.61-0.74).
A literature comparison of 7 methods for scoring protein-ligand interactions

Table L2

<table>
<thead>
<tr>
<th>Method</th>
<th>abs(Pearson r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1_dynamic</td>
<td>0.4</td>
</tr>
<tr>
<td>M1_static</td>
<td>0.2</td>
</tr>
<tr>
<td>M2_static</td>
<td>0.7</td>
</tr>
<tr>
<td>M3_dynamic</td>
<td>0.6</td>
</tr>
<tr>
<td>M3_static</td>
<td>0.4</td>
</tr>
<tr>
<td>M4_dynamic</td>
<td>0.5</td>
</tr>
<tr>
<td>M4_static</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Remember that correlations have confidence intervals and report these intervals.
Table L2

<table>
<thead>
<tr>
<th></th>
<th>abs(Pearson r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1_dynamic</td>
<td>0.0</td>
</tr>
<tr>
<td>M1_static</td>
<td>0.2</td>
</tr>
<tr>
<td>M2_static</td>
<td>0.4</td>
</tr>
<tr>
<td>M3_dynamic</td>
<td>0.6</td>
</tr>
<tr>
<td>M3_static</td>
<td>0.8</td>
</tr>
<tr>
<td>M4_dynamic</td>
<td></td>
</tr>
<tr>
<td>M4_static</td>
<td></td>
</tr>
</tbody>
</table>
Guidelines For Reviewing "Scoring Function" Papers

We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
Molecular weight and calculated LogP are poor null models
Simple QSAR as a Null Model

Generate RDKit fingerprints for ligands
Train on PDB bind refined set (n=4057)
Test on PDB bind core set (n=290)
Wall clock time < 5 min

\[ R^2 = 0.37 \]
What Constitutes an Appropriate Null Model

Molecular Weight

XLogP

Simple QSAR

$R^2 = 0.26$

$R^2 = 0.06$

$R^2 = 0.37$
A Null Model for RMSE

1. Sample N observed values
2. Calculate RMS
3. Repeat 1 and 2 * 1000
Null Model for GC1 HSP90 Free Energy Challenge

RMSE (kcal/mol)

FreeEnergiesSet1  FreeEnergiesSet2  FreeEnergiesSet3
Comparing RMS vs Null for GC1 HSP90 Challenge

Dashed line indicates the null model
We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• **Format of supporting material**
• Criteria for reproducibility
Always provide a machine readable table (e.g. csv) of predicted and experimental values

A table in a paper is not sufficient, it is often very difficult to extract tables from pdf files

Chemical structures should be included as SDF or, where appropriate, SMILES to facilitate comparison with other methods

Need to enable readers to evaluate correlations and errors
Guidelines For Reviewing “Scoring Function” Papers

We need to agree on

• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
Can I Reproduce Your Method?

Letter

Modeling, Informatics, and the Quest for Reproducibility

W. Patrick Walters*
Vertex Pharmaceuticals, Inc., 130 Waverly St., Cambridge, Massachusetts 02139, United States

DOI: 10.1021/ci400197w
Publication Date (Web): June 12, 2013
Copyright © 2013 American Chemical Society

*Cite this: J. Chem. Inf. Model. 53, 7, 1529-1530

E-mail: pat_walters@vrx.com Phone: (617) 341-6242.

ACS AuthorChoice - Terms of Use
What Constitutes Reproducibility?

Code !!!
A thorough description of your method
A web implementation
None of the above
Guidelines For Reviewing "Scoring Function" Papers

We need to agree on
• What constitutes a reasonable dataset
• How data should be reported
• Evaluation metrics
• Statistics for comparison
• What constitutes a null model
• Format of supporting material
• Criteria for reproducibility
How Can You Help?
Docking Challenges Have Become More Challenging

HSP90 2015

FXR 2017

CaS 2018
Questions on Docking Challenges

Are we spending enough time understand compounds that docked poorly?

• Insufficient conformational sampling
• Insufficient pose sampling
• Inadequate scoring
• Ligand poses with limited density

Is everyone missing the same compounds?

Can groups work together to improve their methods?
Acknowledgements

D3R Participants
CSAR Participants
TDT Participants
SAMPL Participants

Rommie Amaro
Mike Gilson

Mill Lambert
Neysa Nevins

Connor Parks
Zied Gaieb

Shuai Liu
Open Source Evaluation Code (More to Come)

https://github.com/PatWalters/metk

metk

Model Evaluation Toolkit

In metk, I’ve collected a set of routines for evaluating predictive models. I put a lot of this code together when I was doing the evaluation for the TDT and D3R projects, as well as a book chapter I wrote in 2013. I’m releasing this project as a way for the community to collaborate and (hopefully) agree on best practices for model evaluation. Most of the initial release is oriented toward the evaluation of free energy calculations.

This is just a start and I plan to add a lot more. Currently, there are routines to calculate

- Root mean squared (RMS) error
- Mean absolute error (MAE)
- Pearson correlation coefficient (with confidence limits)
- Spearman rank correlation (rho) (still need to add confidence limits)
- Kendall tau (still need to add confidence limits)
- Maximum possible correlation given a specific experimental error. This is based on a 2009 paper by Brown, Muchmore and Hajduk

Most of the statistics is done with routines from scikitlearn and scipy.

The toolkit also includes code to generate a few diagnostic plots that I find helpful when looking at model performance. Examples of these plots can be found here

- A scatter plot of experimental vs predicted ΔG. Lines are drawn at 1 and 2 kcal error
- A histogram of the error distribution.
- The two plots above with ΔG converted to a binding affinity (in nM or mM). On the scatter plot, lines are drawn at 5-fold and 10-fold error. I find that I mentally relate to a fold error in binding affinity better than I do to error expressed in kcal/mol. However, if you like looking at error in kcal/mol, use that plot.

Ultimately, the plan is to implement a number of other methods for model evaluation including those described in papers by Anthony Nicholls.
BACKUP
Looks Like Activity Prediction is a Solved Problem

Pearson r

- $K_{DEEP}$: $R = 0.82$, $\rho: 0.82$, RMSE: 1.27
- RF-Score: $R = 0.80$, $\rho = 0.80$, RMSE: 1.39
- X-Score: $R = 0.66$, $\rho: 0.66$, RMSE: 1.71
- cyScore: $R = 0.65$, $\rho: 0.65$, RMSE: 4.13
What Constitutes an Appropriate Null Model

Molecular Weight

$R^2 = 0.26$

XLogP

$R^2 = 0.06$

Simple QSAR

$R^2 = 0.37$
What Constitutes an Appropriate Null Model

Molecular Weight

\[ R^2 = 0.26 \]

XLogP

\[ R^2 = 0.06 \]

Simple QSAR

\[ R^2 = 0.37 \]
Evaluate maximum possible correlation for a dataset given experimental error

FIGURE 2
Plots illustrating the procedure for introducing sampling error into simulated data. (a) Two independently sampled snapshots (open and filled squares) are shown for three different numbers of points initially distributed over 2 log units. (b) Distribution of possible R-values generated by the 'snapshots' of the data in (a). 50,000 iterations were used to generate the R-value distributions.

Maximum Achievable Correlation

Start with experimental data
Add Gaussian error
  • Mean = 0.0
  • Standard deviation = 0.3 log
Calculation correlation
Repeat 1000 times