RELAY
THERAPEUTICS

Eight Factors to Consider When Validating a Model

Pat Walters | CADD GRC | July 16, 2019

Deju Vu?

Computer Aided Drug Design
Gordon Research Conference

July 21 - 26, 2013

Chair
Martin Stahl

Vice Chair
Anthony Nicholls

Mount Snow
89 Grand Summit Way
West Dover, VT, US
Venue and Travel Information

Venue
Office Manager: Katie Lamb
Office Manager Email: MountSnow_SM@grc.org
Office Phone: 802-464-7069

MONDAY	
7:30 am - 8:30 am	Breakfast
9:00 am - 10:30 am	**Practical Statistics, Part I**
	Discussion Leader: **Haihong Ni** (BioDuro, Beijing, China)
9:00 am - 9:15 am	**Peter W. Kenny** (Universidade de São Paulo, Brazil) "Tales of Correlation Inflation: Eu prefiro a minha comida cozida e meus dados brutos"
9:15 am - 9:20 am	Discussion
9:20 am - 9:35 am	**Maria A. Gallardo** (Universidad de Zaragoza, Zaragoza, Spain) "Outliers in Theory and Experiment: Lessons from Small Molecule Dipole Moments"
9:35 am - 9:40 am	Discussion
9:40 am - 9:55 am	**Pat Walters** (Vertex Pharmaceuticals, Boston, USA) "Just Because You Published It Doesn't Make it Right"
9:55 am - 10:00 am	Discussion
10:00 am - 10:30 am	Group Photo / Coffee Break
10:30 am - 12:30 pm	**Deeper Issues for Statistics in CADD: Non-ideality**
	Discussion Leader: **Martin Stahl** (Roche, Basel, Switzerland)
10:30 am - 11:15 am	**Tom Darden** (OpenEye Scientific Software, Santa Fe, NM, USA) "Ideality is Overrated: Making the Most of an Imperfect World"
11:15 am - 11:30 am	Discussion
11:30 am - 12:15 pm	**Anna Linusson** (Umeå University, Umeå, Sweden) "Cost and Correlation: Designing Better Data From the Start"
12:15 pm - 12:30 pm	Discussion
12:30 pm	Lunch

TUESDAY	
7:30 am - 8:30 am	Breakfast
9:00 am - 10:30 am	**Practical Statistics, Part II**
	Discussion Leader: **Haihong Ni** (BioDuro, Beijing, China)
9:00 am - 9:15 am	**Martha S. Head** (GSK, Upper Merion, PA, USA) "One of these things is not like the others?"
9:15 am - 9:20 am	Discussion
9:20 am - 9:35 am	**Paul Czodrowski** (Merck KG, Germany) "The Kappa Statistic: Taking Care of Background Rates"
9:35 am - 9:40 am	Discussion
9:40 am - 10:00 am	**Short Poster Presentations**
10:00 am - 10:30 am	Coffee Break
10:30 am - 12:30 pm	**Deeper Issues for Statistics in CADD: NULL Models**
	Discussion Leader: **Anna Linusson** (Umeå University, Umeå, Sweden)
10:30 am - 11:15 am	**Woody Sherman** (Schrödinger, New York, NY, USA) "Null Models to Improve Assessment of Virtual Screening Enrichments and Lead Optimization Scoring"
11:15 am - 11:30 am	Discussion
11:30 am - 12:15 pm	**Marcel Verdonk** (Astex Pharmaceuticals, UK) "Statistics don't lie. Or do they?"
12:15 pm - 12:30 pm	Discussion
12:30 pm	Lunch
1:30 pm - 6:00 pm	Free Time

WEDNESDAY	
7:30 am - 8:30 am	Breakfast
9:00 am - 10:30 am	**Practical Statistics, Part III**
	Discussion Leader: **Haihong Ni** (BioDuro, Beijing, China)
9:00 am - 9:15 am	**Ullrika Sahlin** (Lund University, Lund, Sweden) "Uncertainty in QSAR Predictions - Bayesian Inference and the Magic of Bootstrap"
9:15 am - 9:20 am	Discussion
9:20 am - 9:35 am	**Stephen Johnson** (BMS, Princeton, NJ, USA) "How to Measure to Anything"
9:35 am - 9:40 am	Discussion
9:40 am - 9:55 am	**Eric Martin** (Novartis, San Francisco, CA, USA) "Hamming Distance and Hemple's Ravens: How to Build a Euclidean Property Space from Sparse Binary Fingerprints"
9:55 am - 10:00 am	Discussion
10:00 am - 10:30 am	Coffee Break
10:30 am - 12:30 pm	**Deeper Issues for Statistics in CADD: Parameters**
	Discussion Leader: **Ajay Jain** (UCSF, San Francisco, CA, USA)
10:30 am - 11:15 am	**Kim Branson** (Hessian Informatics, Emerald Hills, CA, USA) "Curb Your Enthusiasm with Bayes Factors"
11:15 am - 11:30 am	Discussion
11:30 am - 12:15 pm	**Vijay Pande** (Stanford, Palo Alto, CA, USA) "Less Art, More Science: A Systematic, Reproducible, Statistically Significant Scheme for Parameterizing Force Fields"
12:15 pm - 12:30 pm	Discussion

Critical Issues

Datasets

Visualization

Statistics

Reproducibility

Critical Issues

Datasets

Visualization

Statistics

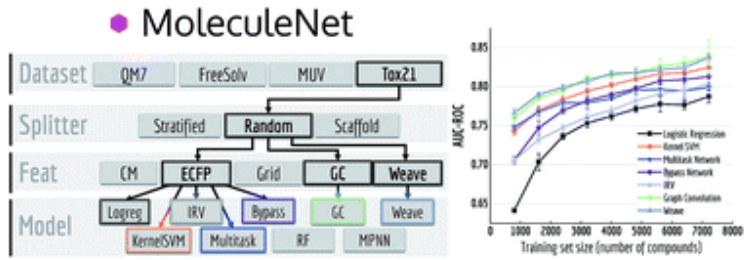
Reproducibility

The “Big 3” Validation Sets

The “BIG 3” Validation Sets

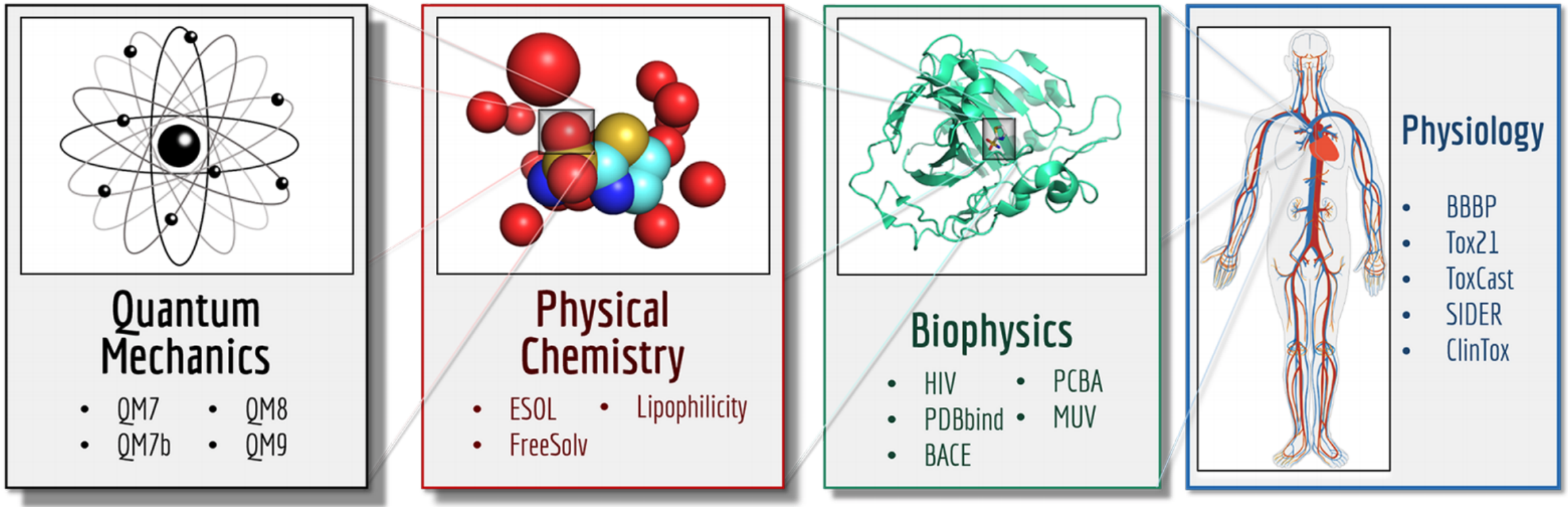
http://dude.docking.org/

http://www.pdbbind.org.cn/



http://moleculenet.ai

MoleculeNet – A Set of Machine Learning Benchmark

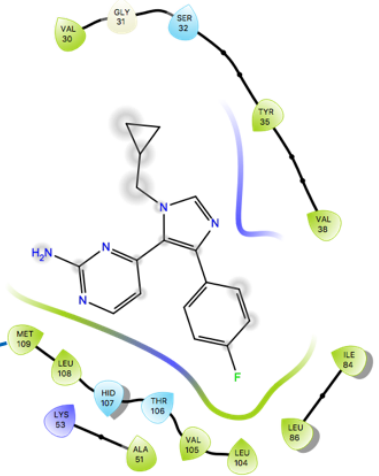


Benchmarks are great – ***when applied properly***

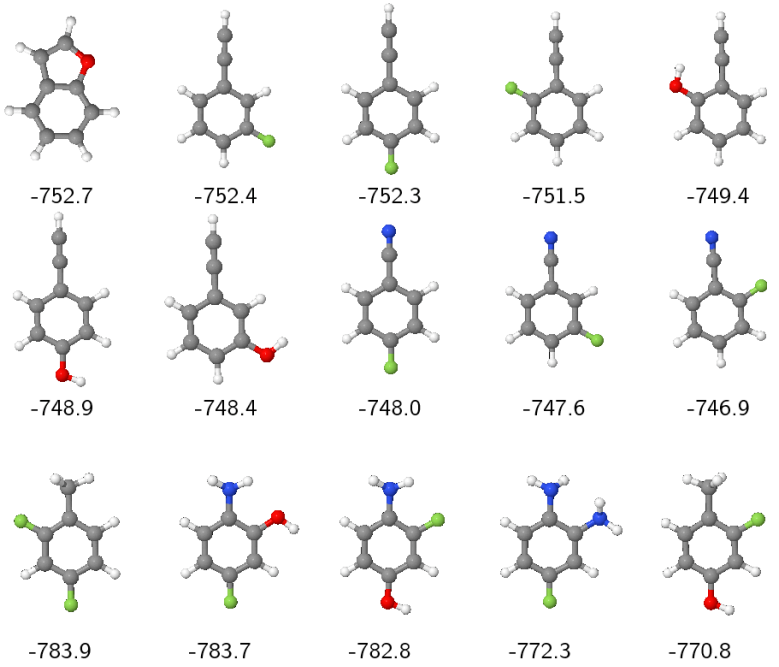
What Should We Be Able to Predict?

PDBBind – Predict Activity from Protein-Ligand Complex

Name	Number	Polar	Hydrophobic	Acceptor	Donor	Aromatic	Charged
TYR	35	0	1	0	0	1	0
VAL	38	0	1	0	0	0	0
ALA	51	0	1	0	0	0	0
LYS	53	1	0	0	1	0	1
ILE	84	0	1	0	0	0	0
LEU	104	0	1	0	0	0	0
THR	106	1	0	0	0	0	0
LEU	108	0	1	0	0	0	0
MET	109	0	1	1	1	0	0

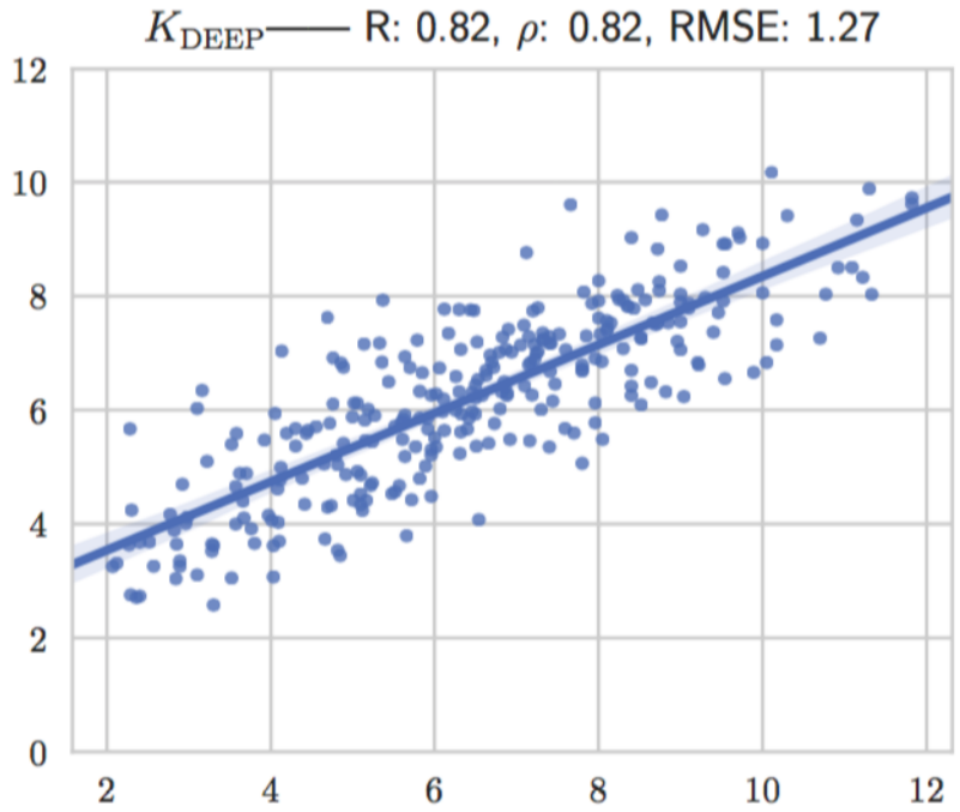


QM7, QM8, QM9 – Predict Quantum Chemical Properties

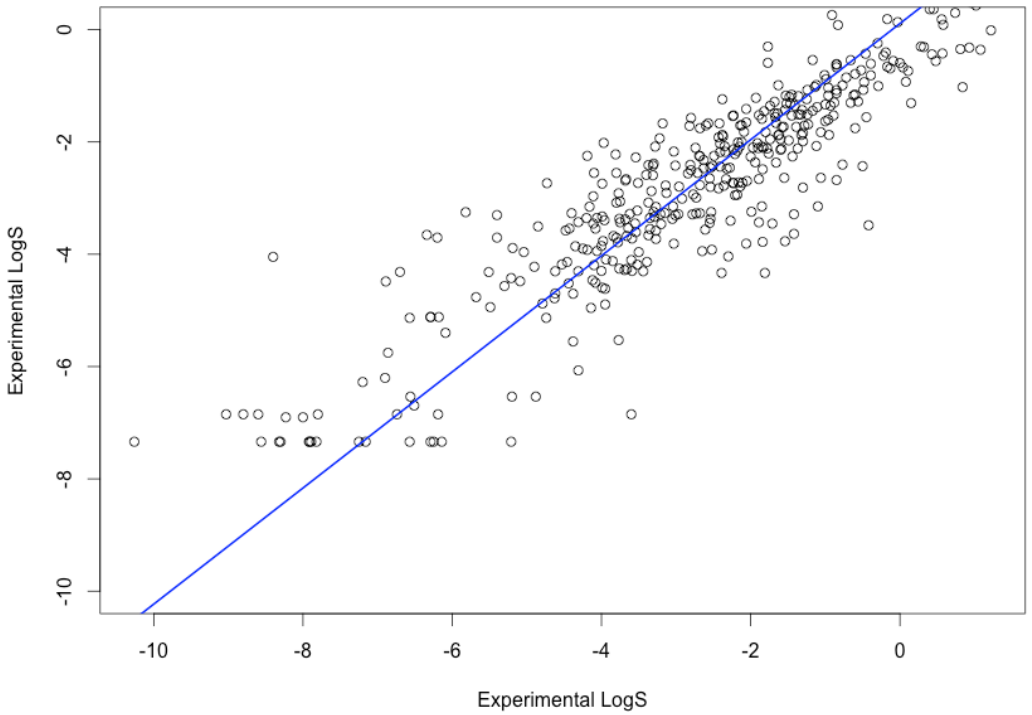


Should we be able to predict 3D properties from SMILES?

Is the Dynamic Range Appropriate?

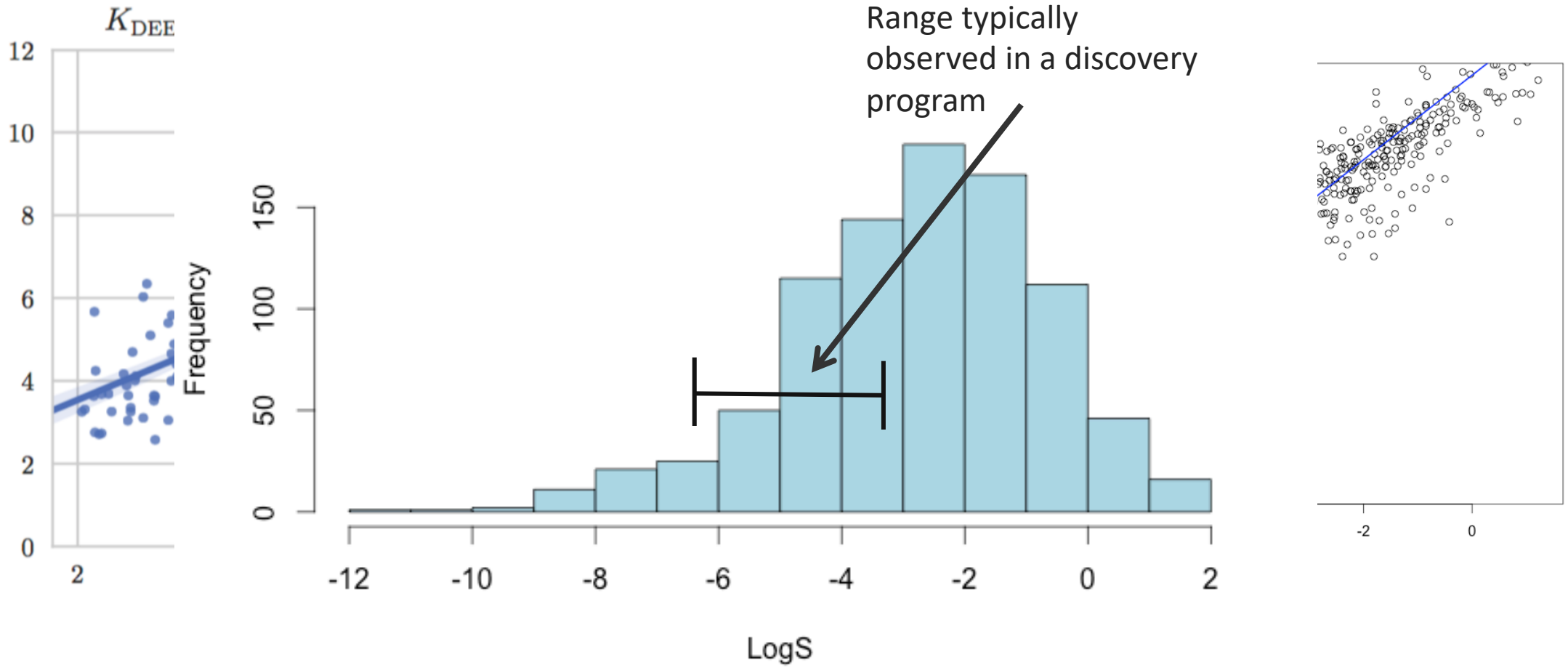


PDBBind 10 logs



ESol 12 logs

Is the Dynamic Range Appropriate?



Include Correlation and Error for Regression Models

You won't make everyone happy, so why not include everything?

- **Pearson R^2**
- **Spearman ρ**
- **Kendall τ**
- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**

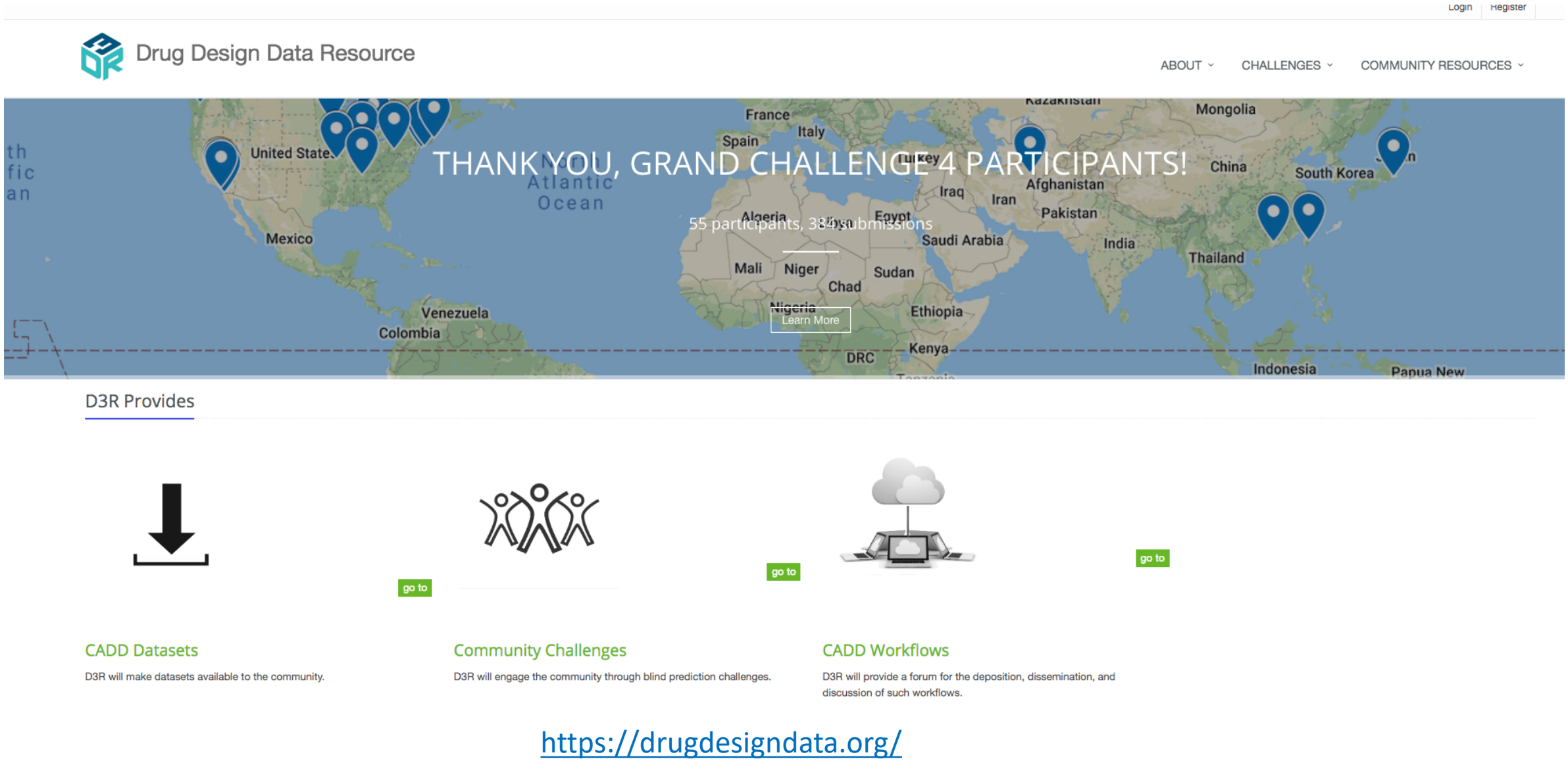
Table 1 Dataset details: number of compounds and tasks, recommended splits and metrics

Category	Dataset	Data type	Tasks		Compounds	Rec – split	Rec – metric
Quantum mechanics	QM7	SMILES, 3D coordinates	1	Regression	7165	Stratified	MAE
	QM7b	3D coordinates	14	Regression	7211	Random	MAE
	QM8	SMILES, 3D coordinates	12	Regression	21 786	Random	MAE
	QM9	SMILES, 3D coordinates	12	Regression	133 885	Random	MAE
Physical chemistry	ESOL	SMILES	1	Regression	1128	Random	RMSE
	FreeSolv	SMILES	1	Regression	643	Random	RMSE
	Lipophilicity	SMILES	1	Regression	4200	Random	RMSE

From the MoleculeNet paper, no correlation statistics

https://github.com/PatWalters/metk

Consider Prospective Validation



https://drugdesigndata.org/

Critical Issues

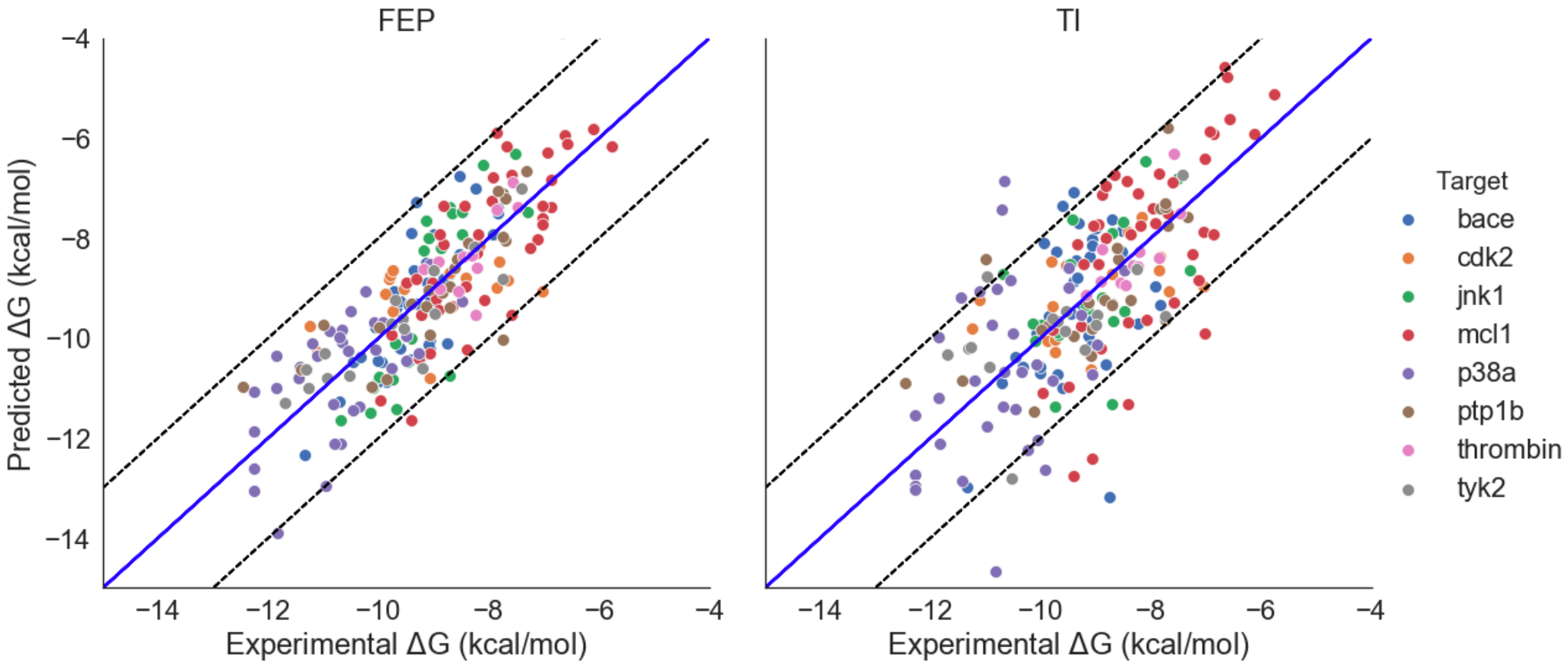
Datasets

Visualization

Statistics

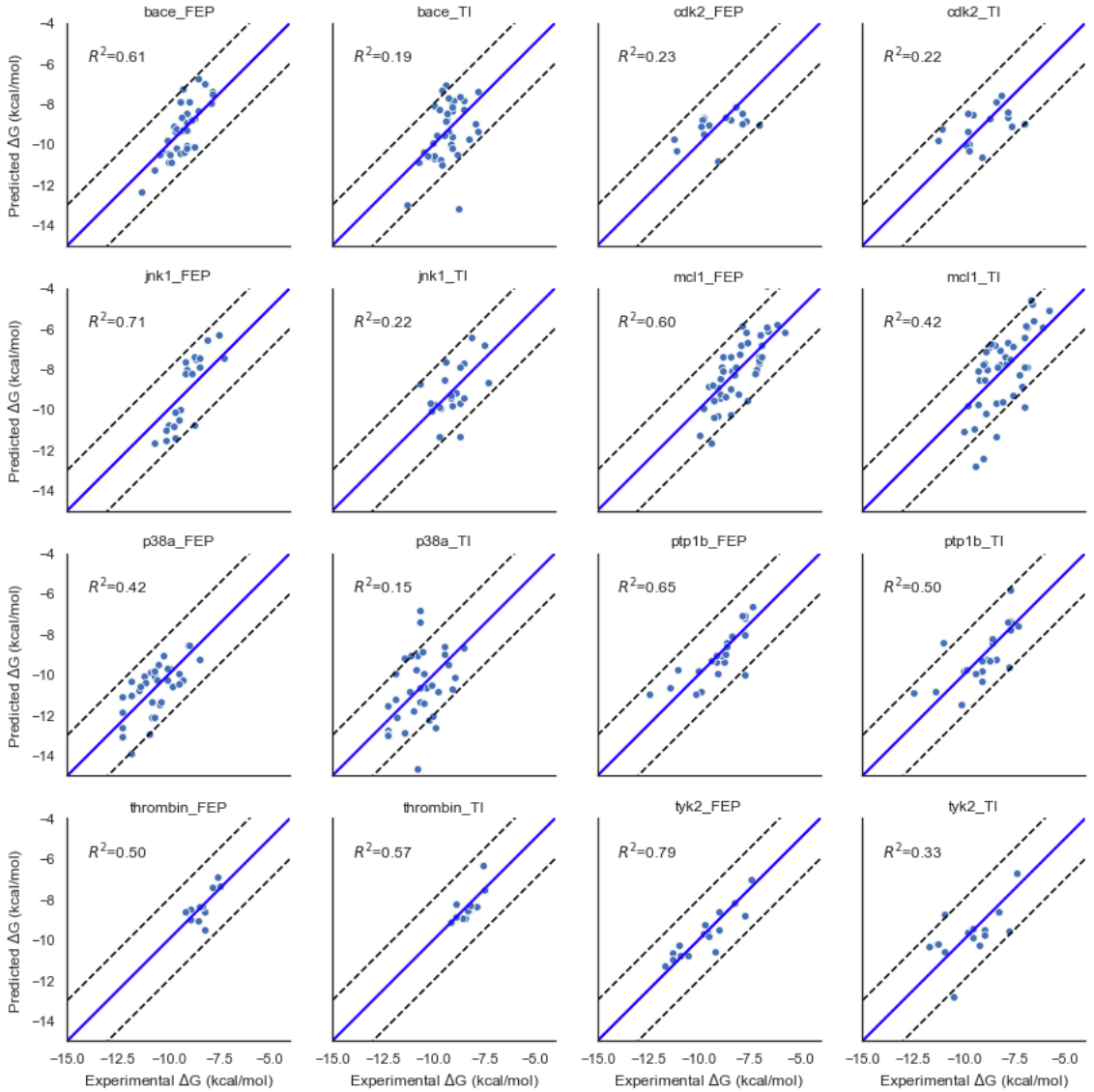
Reproducibility

Are Your Plots Meaningful?



https://practicalcheminformatics.blogspot.com/2019/02/some-thoughts-on-evaluating-predictive.html

Consider Trellising Your Plots as an Alternative



Critical Issues

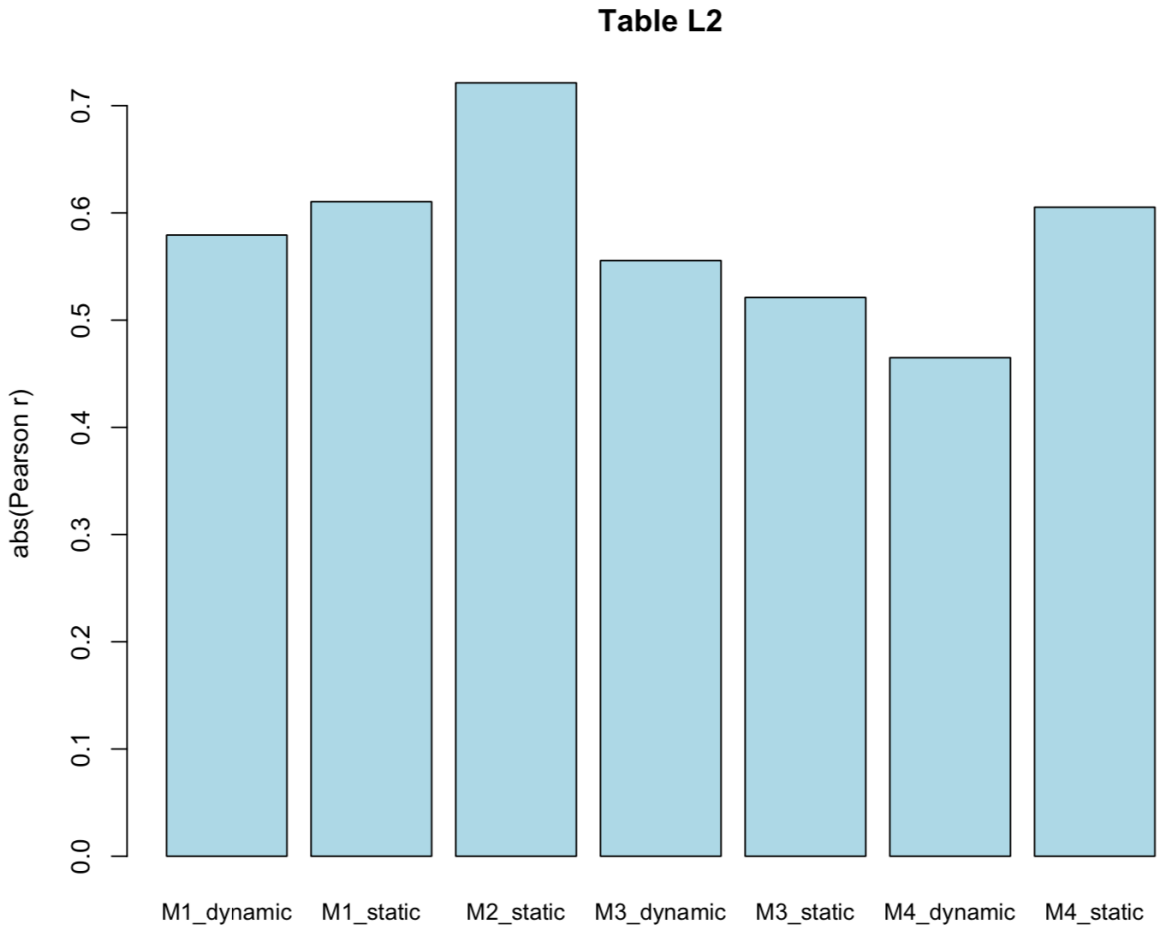
Datasets

Visualization

Statistics

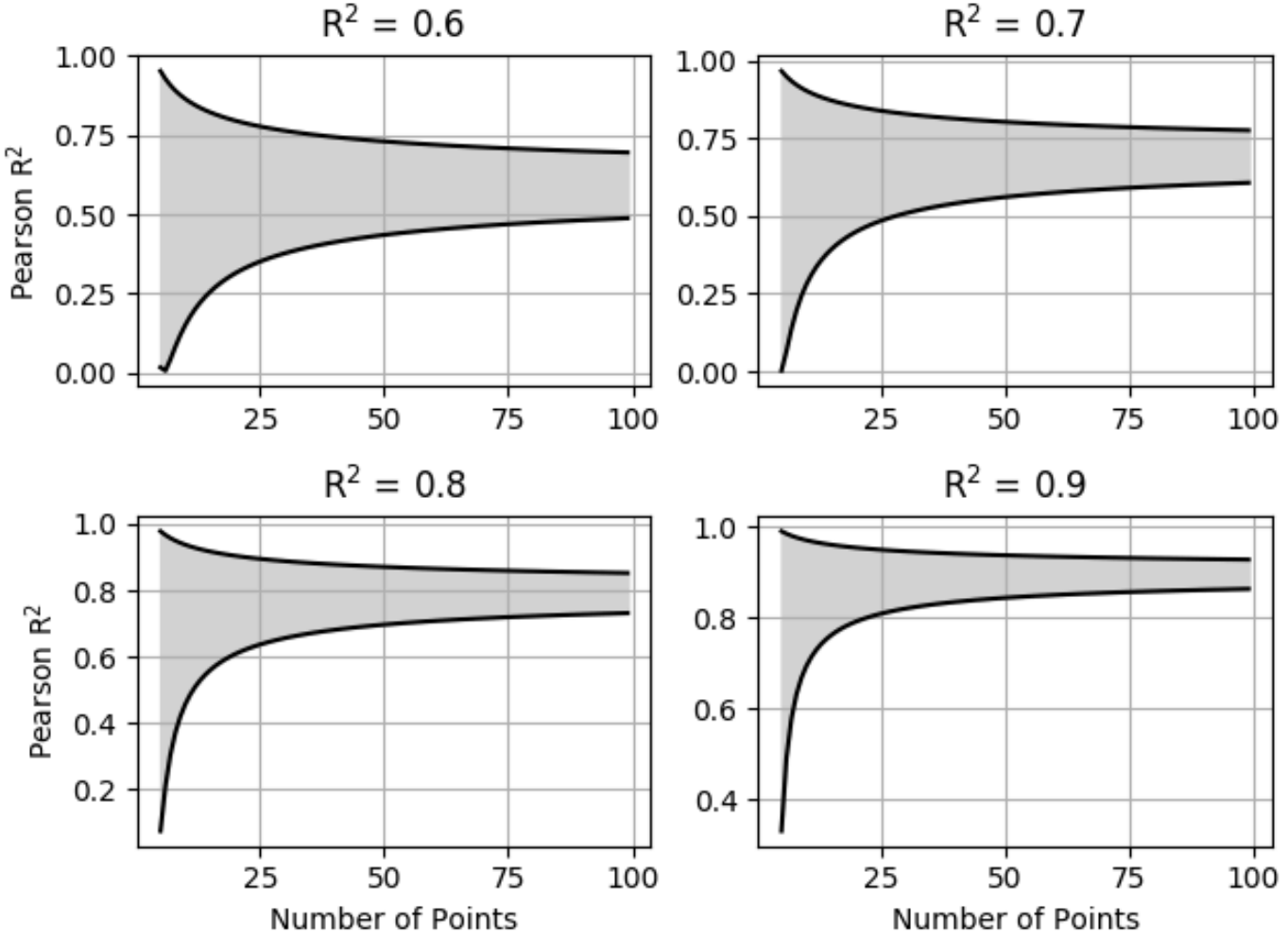
Reproducibility

Did You Perform the Appropriate Statistical Comparisons?



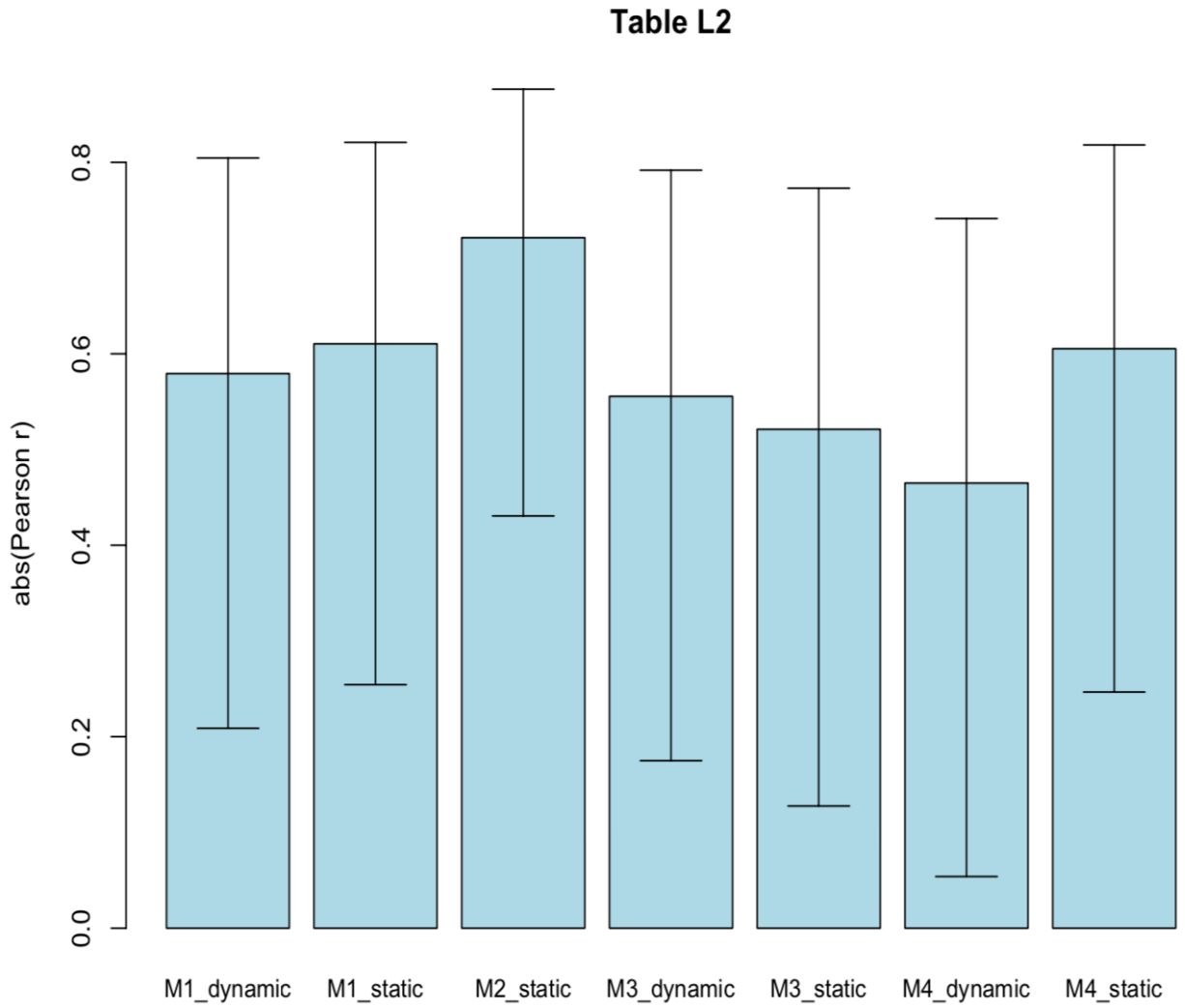
A literature comparison of 7 methods for scoring protein-ligand interactions

Correlations Have an Associated Error



https://practicalcheminformatics.blogspot.com/2018/09/predicting-aqueous-solubility-its.html

Adding Error Bars Dramatically Alters the Conclusion



Report Effect Size

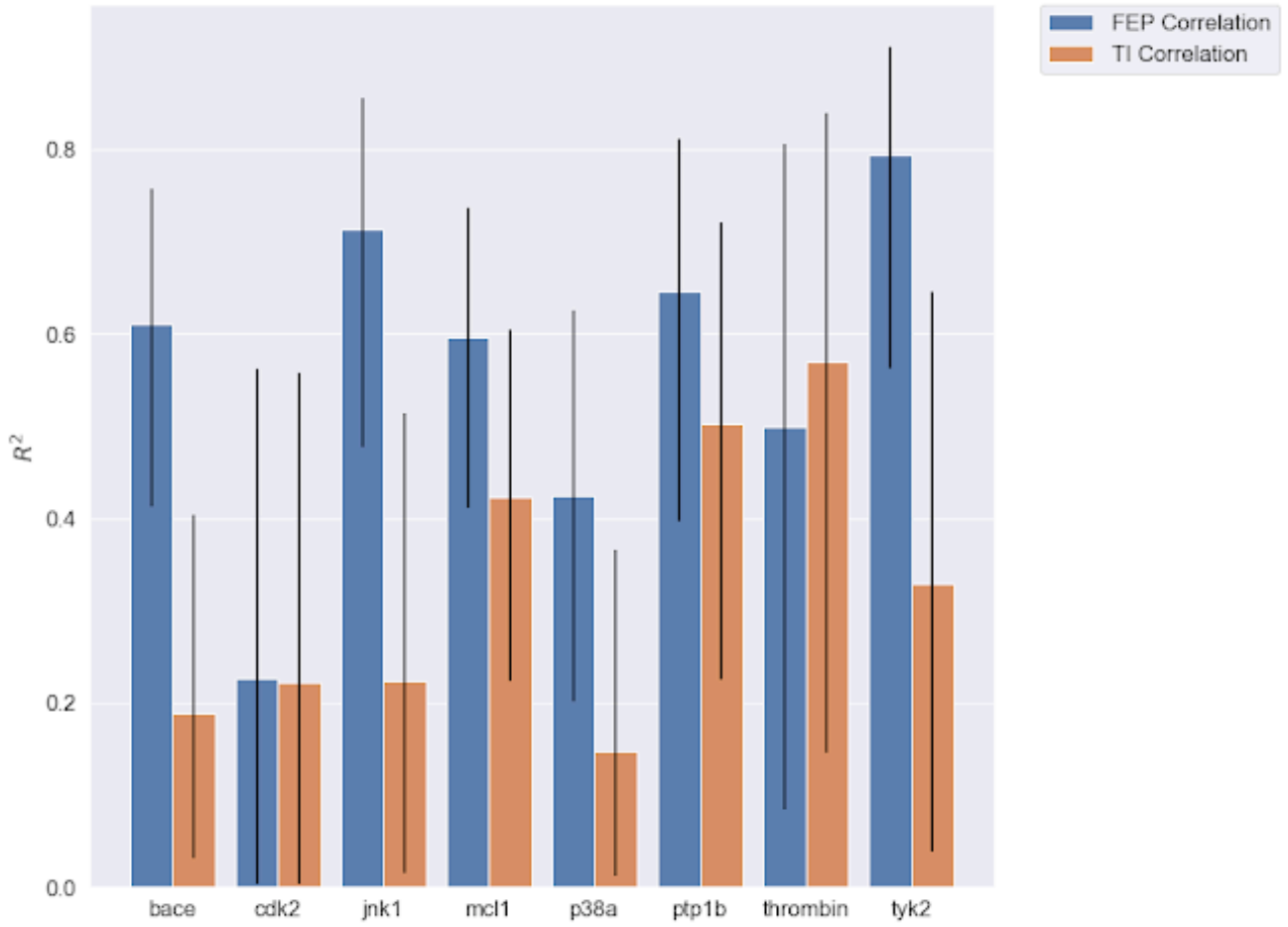
Samples are not independent

- **Can't just look at error bars**

Cohen's d = $\frac{mean(u1-u2)}{std(u1-u2)}$

Effect	Cohen's d
Small	0.2
Medium	0.5
Large	0.8

p(B>A) ≈ 0.25d + 0.5



d = 1.2
p = 0.8

How Good Can/Should Your Model Be?

Start with experimental data

Add Gaussian error

- **Mean = 0.0**
- **Standard deviation = experimental error**

Calculation correlation Repeat 1000 times

Drug Discovery Today
Volume 14, Issues 7–8, April 2009, Pages 420-427

Review

Post Screen

Healthy skepticism: assessing realistic model performance

Scott P. Brown, Steven W. Muchmore, Philip J. Hajduk

Show more

https://doi.org/10.1016/j.drudis.2009.01.012

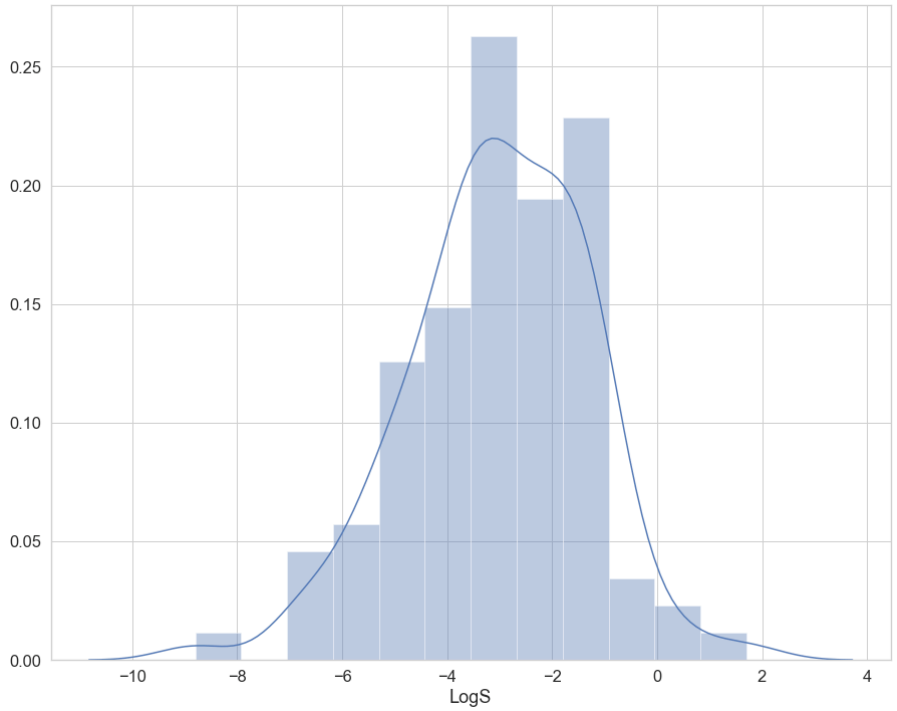
Get rights and content

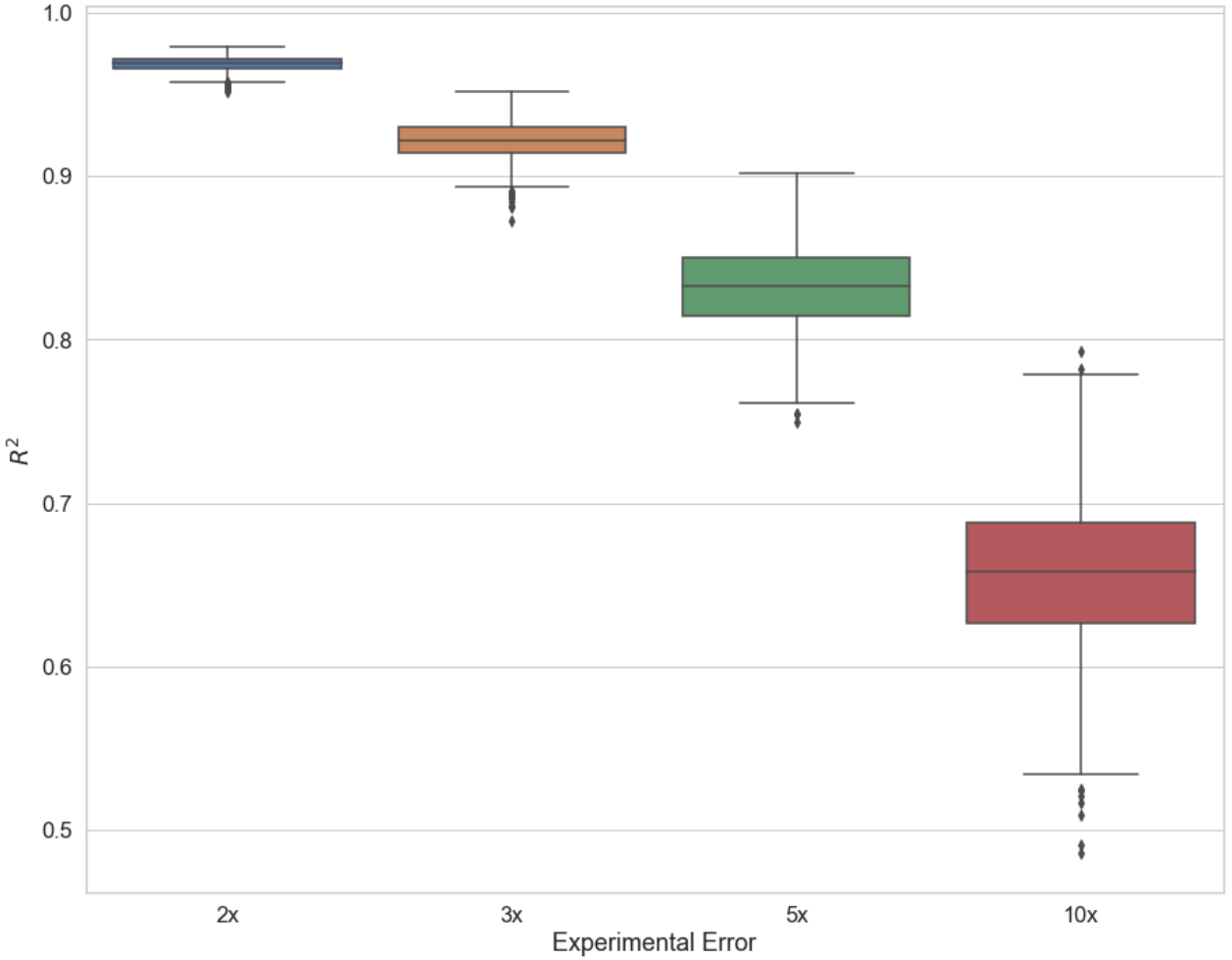
https://www.sciencedirect.com/science/article/pii/S1359644609000403

The Impact of Experimental Error on Correlation

DLS 100 Dataset

- 100 compounds
- Measured by Dynamic Light Scattering





https://risweb.st-andrews.ac.uk/portal/en/datasets/dls100-solubility-dataset(3a3a5abc-8458-4924-8e6c-b804347605e8).html

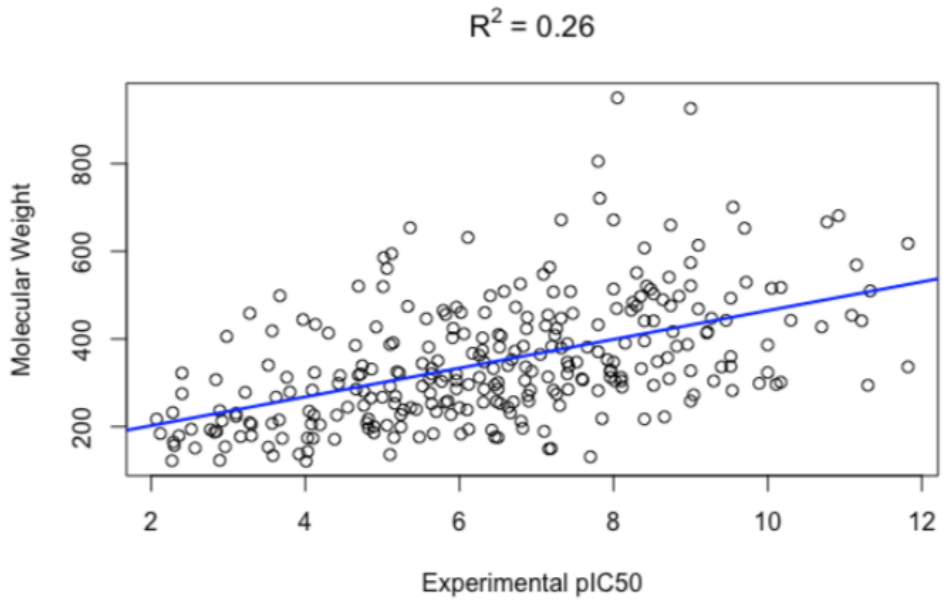
What Is Your Null Model?

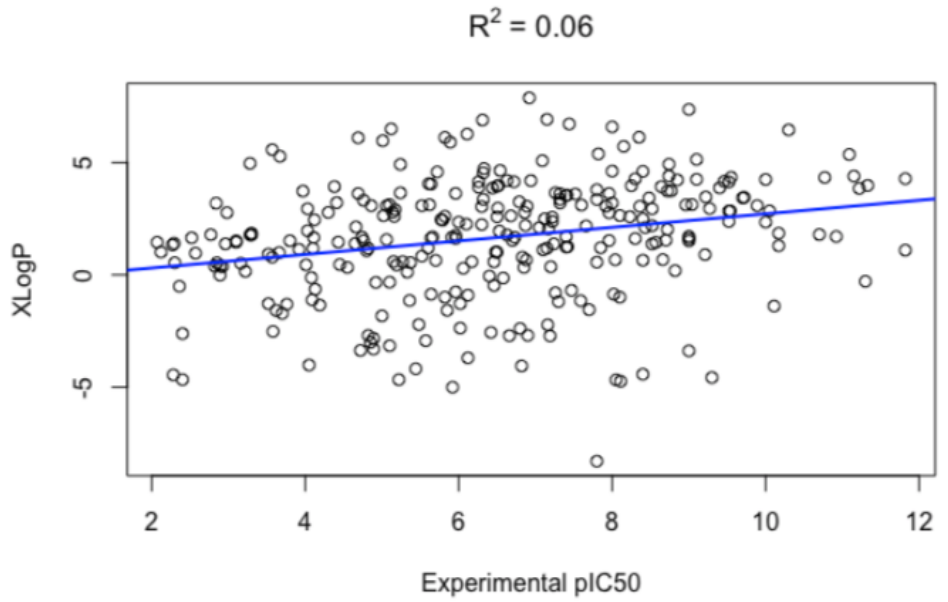
Compare with molecular weight and LogP

Compare with generally accepted models (e.g. Random Forest)

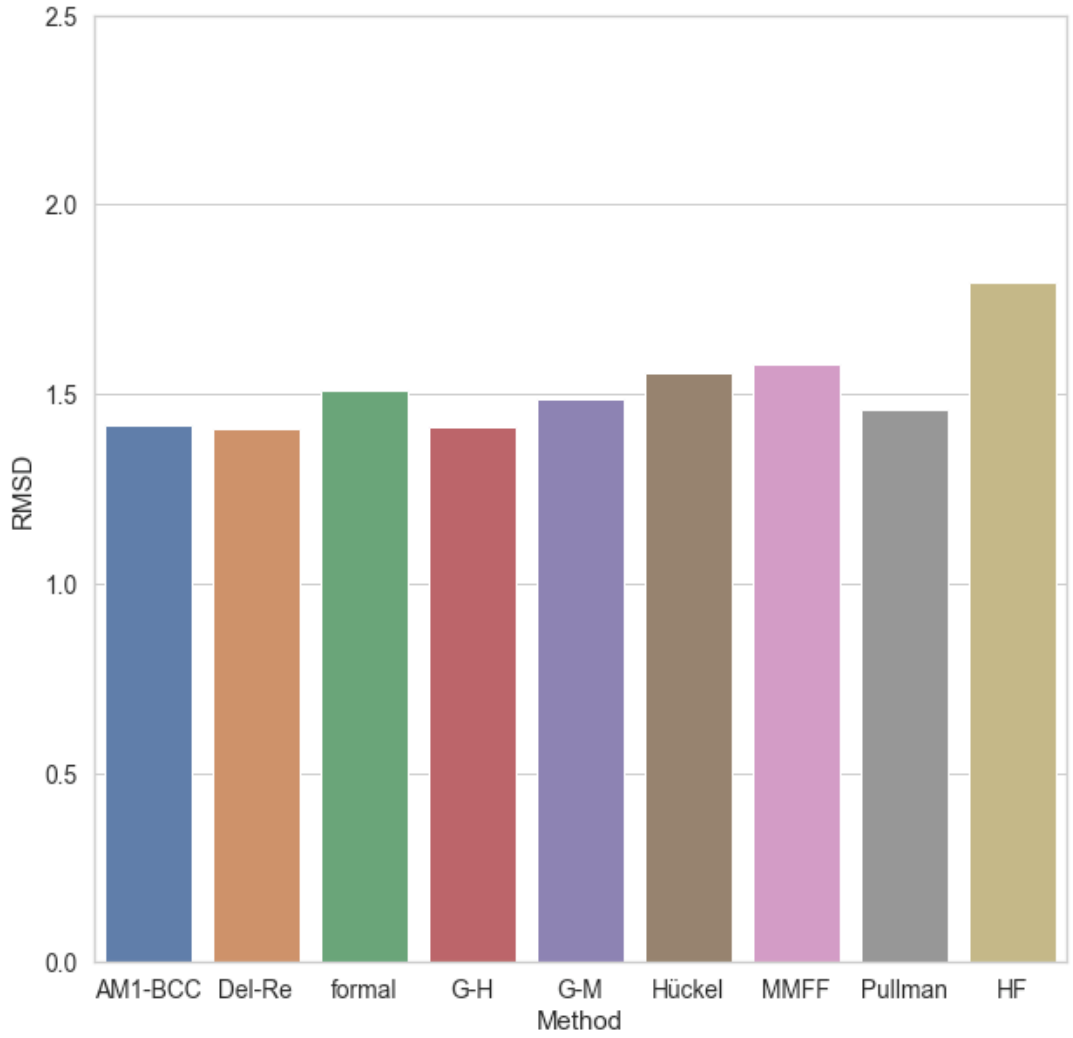
Docking scores or MM[G,P]BSA vs free energy calculations

Evaluate random distribution when calculating RMSE or MAE

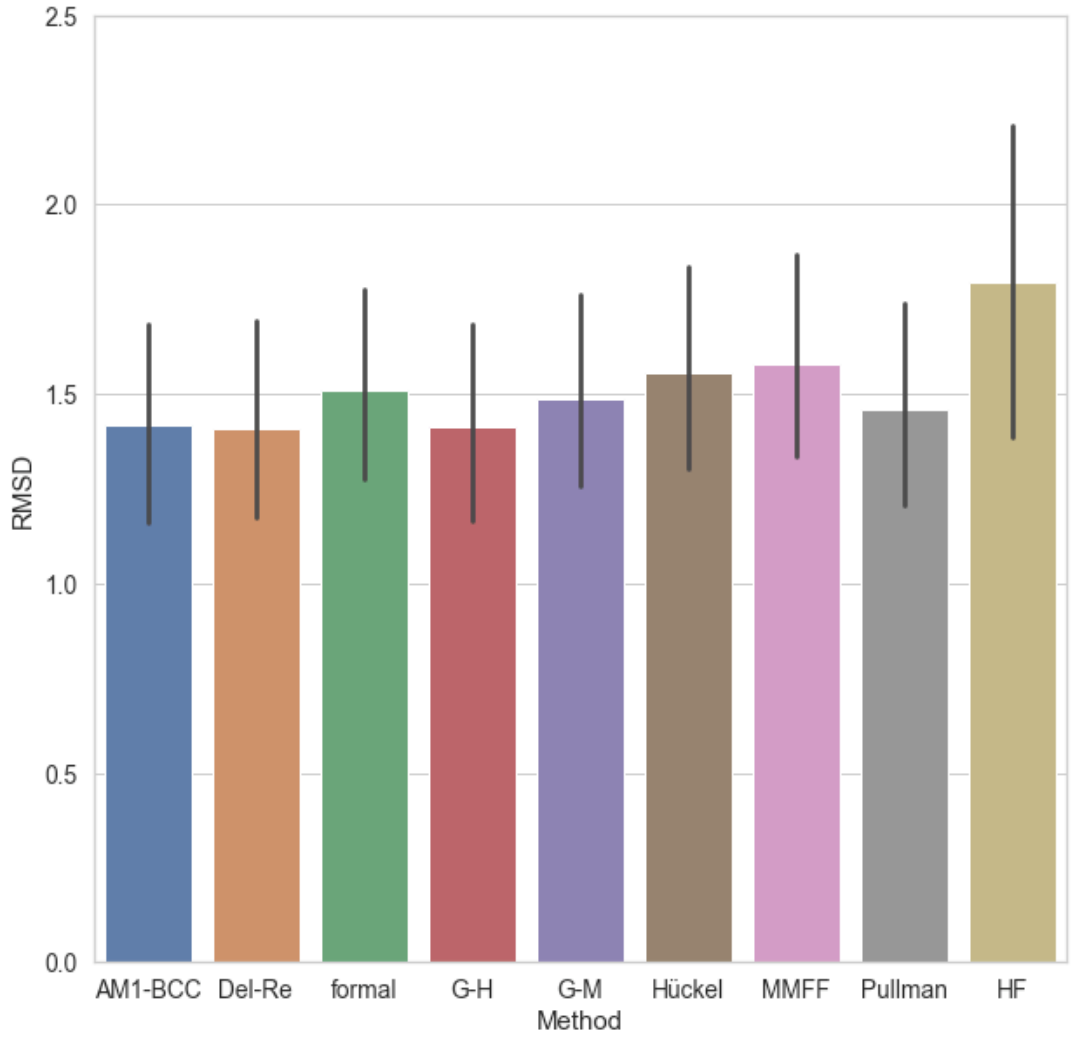




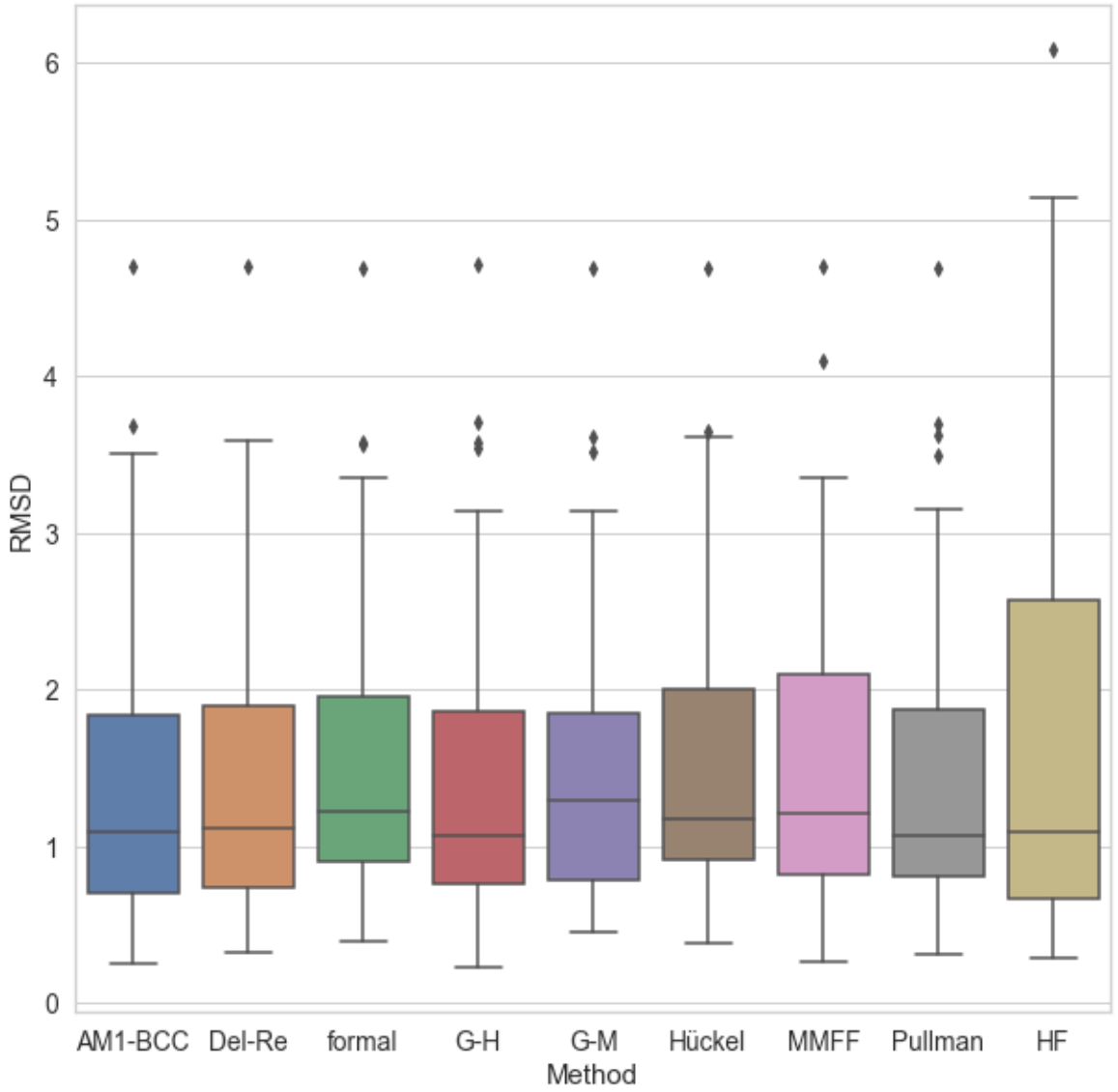
Did You Consider the Impact of Multiple Comparisons?



Error Bars Already Make This Appear Problematic



Box Plots Provide a Better Visualization of Distributions



Considering Multiple Comparisons

The more hypotheses we check, the higher the probability of a Type I error (false positive)

Bonferroni-corrected p value = $\frac{\alpha}{n}$

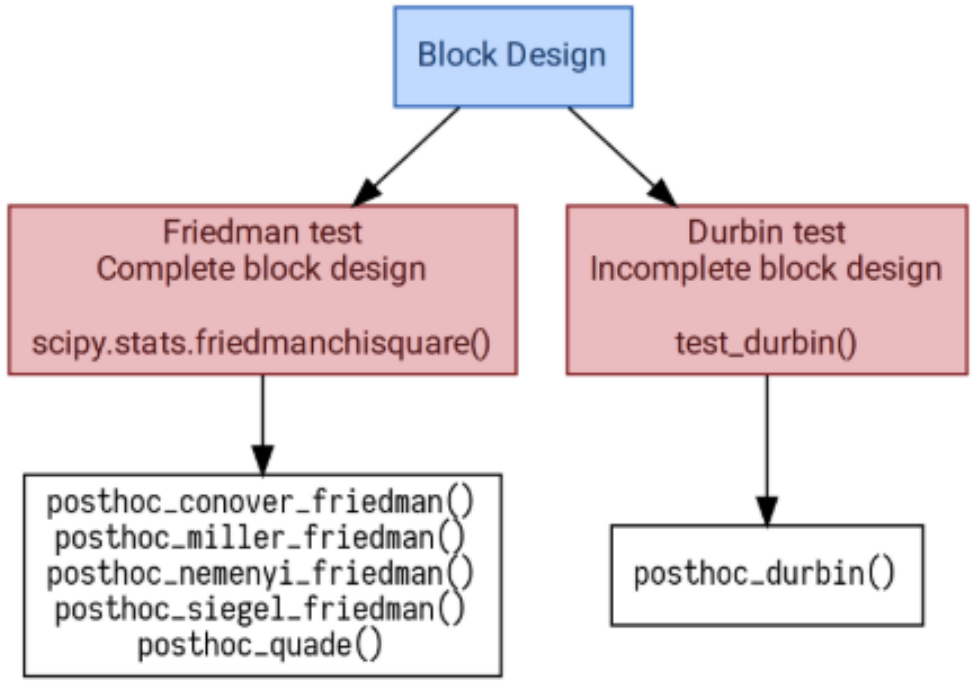
original p value (points to α)

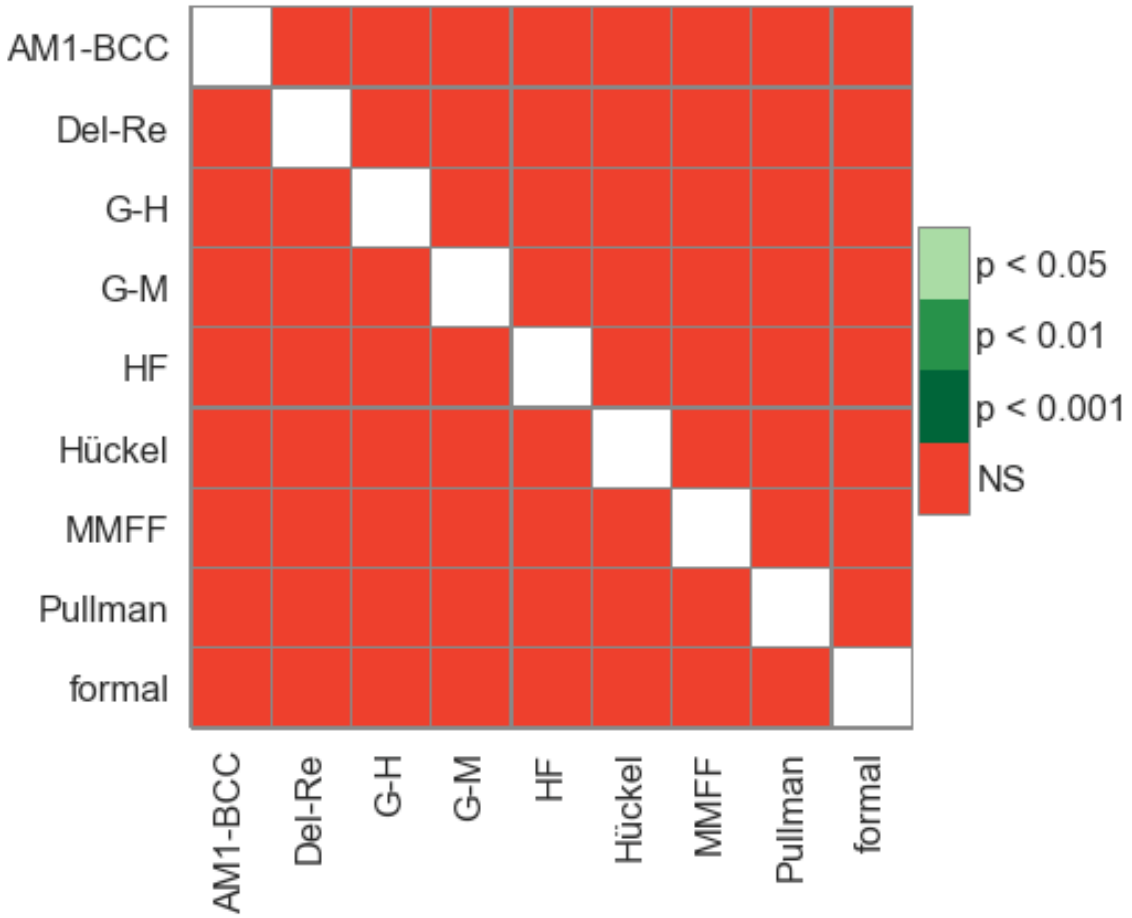
number of tests performed (points to n)

Holm-Bonferroni slightly less conservative – avoids Type II error (false negative)

- **Order p-values from smallest to largest**
- **HB = Target α / (n – rank + 1)**

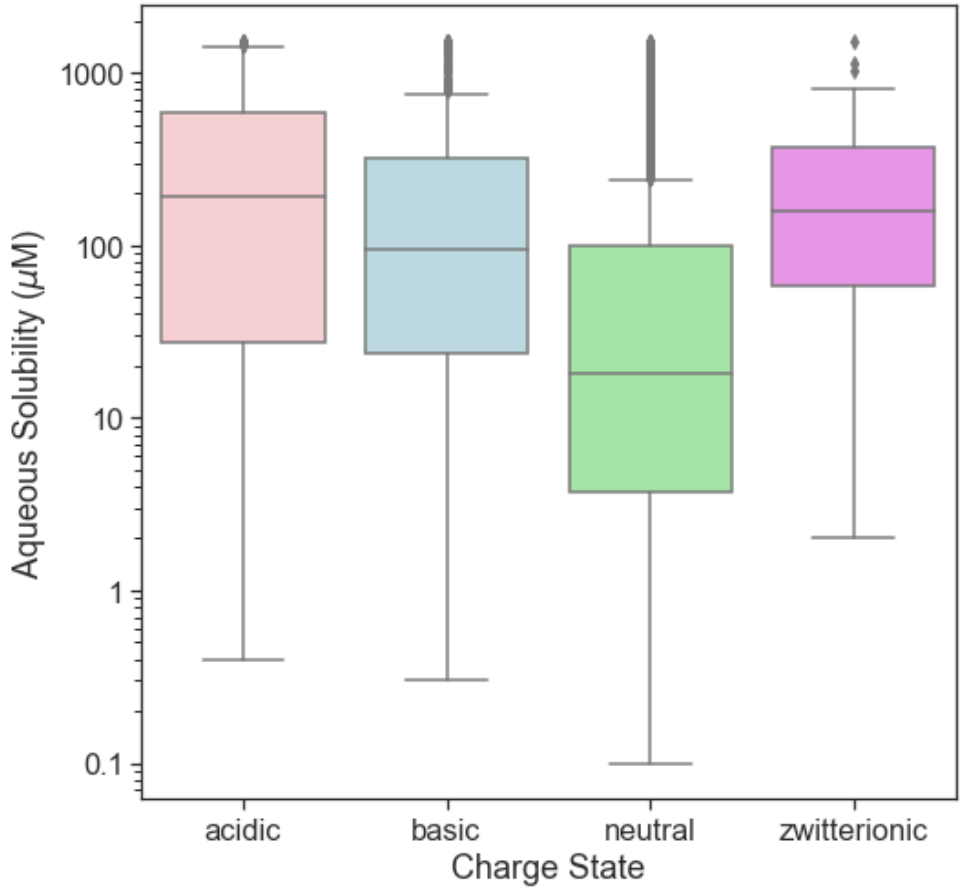
scikit-posthocs Tests for Pairwise Multiple Comparisons

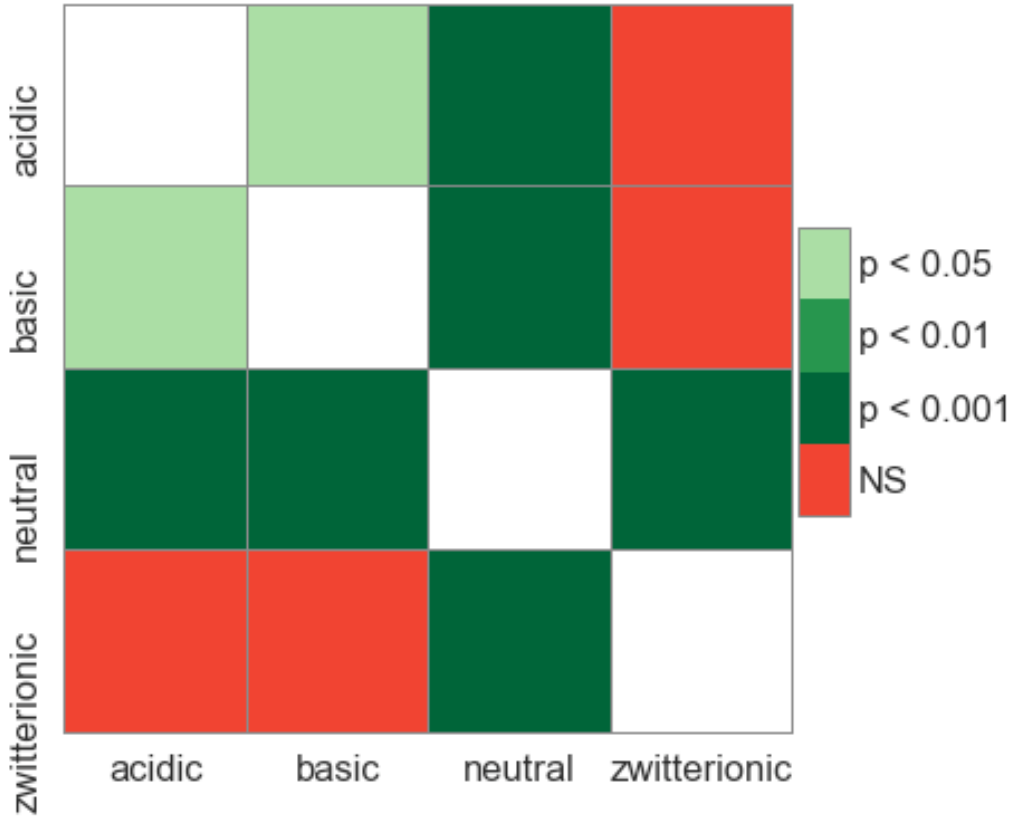




https://pypi.org/project/scikit-posthocs/

A Somewhat More Compelling Example





Charifson, P. S., & Walters, W. P. (2014) *Journal of Medicinal Chemistry*, *57*(23), 9701-9717.

Critical Issues

Datasets

Visualization

Statistics

Reproducibility

Include Appropriate Supporting Information

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

- Yes, I know about Tabula
- No, I shouldn't have to use it

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

Need to enable readers to evaluate correlations and errors

Create a Git Repo With Your Supporting Material

README.md

D3R-Grand Challenge 4

D3R Grand Challenge 4: Comparison of ligand affinity ranking using Autodock-GPU and MM-GBSA scores

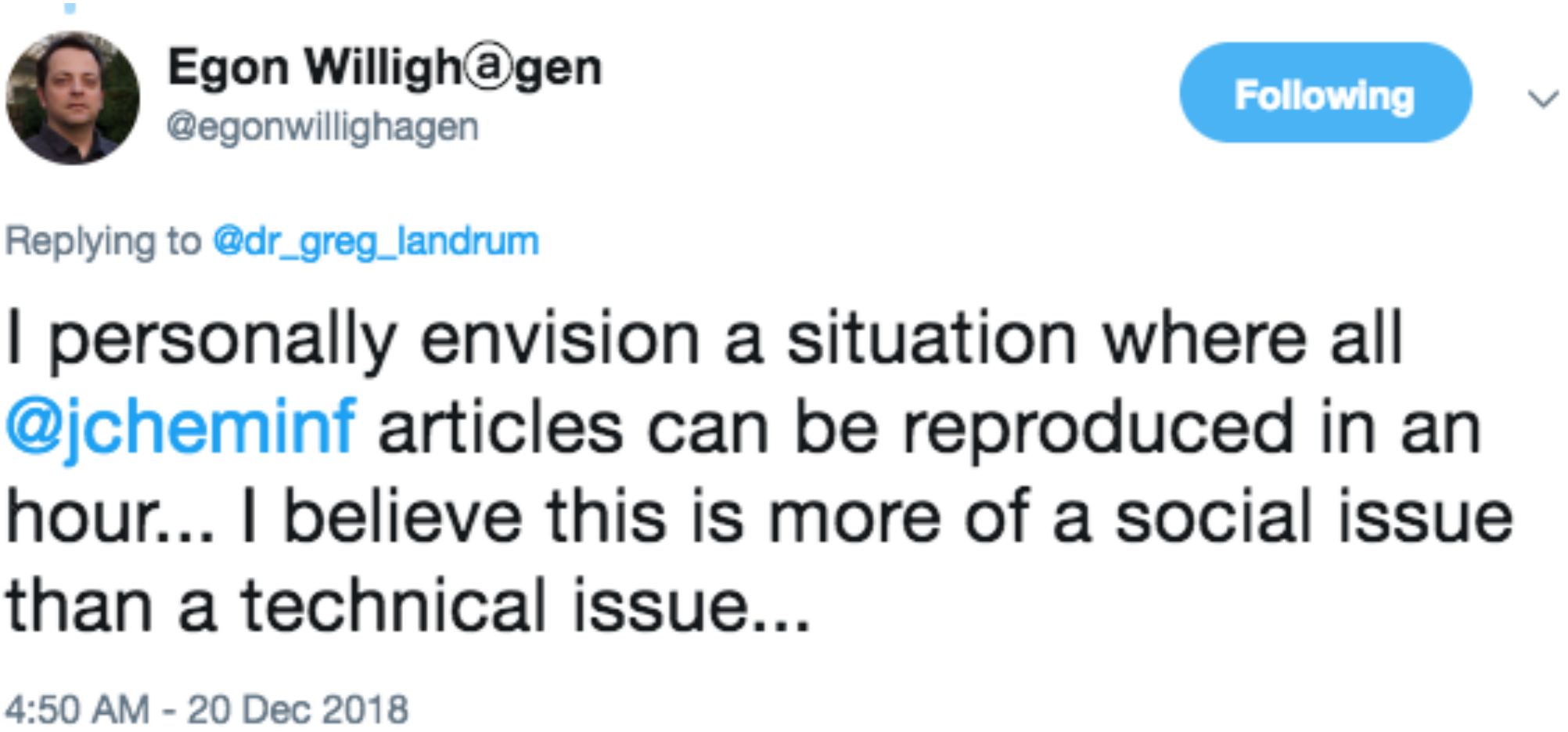
This repository houses all the files that we used for our MM-GBSA calculations for our D3R Grand Challenge 4 participation, as detailed in our collaborative manuscript entitled "Comparison of ligand affinity ranking using AutoDock-GPU and MM-GBSA scores in the D3R Grand Challenge 4". Here you will find all the files needed to perform MM-GBSA end-point free energy estimates on the BACE-1 subsets provided by the D3R Grand Challenge 4 organizers for the affinity ranking subchallenge. Here, we provide the input files of the receptors (pdb files) and the 154 ligands (mol2 files) that we prepared for Stage 2 affinity ranking subchallenge. The scripts used to perform MD simulations and MM-GBSA calculations are also available.

Manifest

- `ligands` : directory containing the mol2 files of the 154 ligands and the respective MM-GBSA results.
- `protein_oe` : directory containing the pdb files of the structures that we used to dock and prepare the 154 receptor-ligand complexes.
- `scripts` : directory containing the scripts that we used to conduct the parametrization of the receptors and the ligands using amber and tleap respectively, the MD simulations and the MM-GBSA calculations.
- `ligand_4EWO_singleProt` : directory containing the MM-GBSA results of the protein-ligand complexes using the BACE-1 structure (PDB code: 4EWO) with a single protonated aspartyl dyad (Asp32H Asp228-).
- `ligand_2wf3_redo` : directory containing the MM-GBSA results of the protein-ligand complexes using the BACE-1 structure (PDB code: 2WF3); the ligands in these complexes were originally docked in 4EWO and during our retrospective analysis, we used 2WF3 to dock and model them.
- `fetch_score.sh` : script used to fetch all the MM-GBSA scores of the receptor-ligand sets that we modeled.
- `charge.txt` : file containing the charges that we used to model each ligand in the BACE-1 dataset.

https://github.com/MobleyLab/D3R-2018-AutoDock-MMGBSA/

Is Your Method Reproducible?



Egon Willigh@gen
@egonwillighagen

Following

Replying to @dr_greg_landrum

I personally envision a situation where all @jcheminf articles can be reproduced in an hour... I believe this is more of a social issue than a technical issue...

4:50 AM - 20 Dec 2018

Tilting at Windmills

JOURNAL OF CHEMICAL INFORMATION AND MODELING

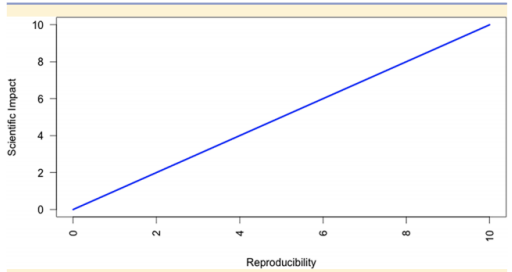
Letter

pubs.acs.org/jcim

Modeling, Informatics, and the Quest for Reproducibility

W. Patrick Walters*

Vertex Pharmaceuticals, Inc., 130 Waverly St., Cambridge, Massachusetts 02139, United States



ABSTRACT: There is no doubt that papers published in the Journal of Chemical Information and Modeling, and related journals, provide valuable scientific information. However, it is often difficult to reproduce the work described in molecular modeling and cheminformatics papers. In many cases the software described in the paper is not readily available, in other cases the supporting information is not provided in an accessible format. To date, the major journals in the fields of molecular modeling and cheminformatics have not established guidelines for reproducible research. This letter provides an overview of the reproducibility challenges facing our field and suggests some guidelines for improving the reproducibility of published work.

in our field to not release source code. Some may have been motivated by a desire to secure intellectual property that would lead to future financial gains. Others, who are not professional programmers, may not consider their "research code" worthy of publication. Academic groups may feel that proprietary code provides an advantage when applying for grants. This argument could, of course, be removed if funding agencies required grant recipients to release their source code, as suggested in a recent editorial in *Science*.[1]

It is possible that the wide array of computer hardware platforms available 20 years ago would have made supporting a particular code base more difficult. However, over the last 10 years, the world seems to have settled on a small number of hardware platforms, dramatically reducing any sort of support burden. In fact, modern virtual machine technologies have made it almost trivial to install and run software developed in a different computing environment.

Other factors may have also affected the situation. Technology transfer offices at some universities have become more aggressive in pushing groups to monetize their research. Actually, cost is a small component of the difficulty created by tech transfer offices. Many industrial groups will not consider licensing academic code due to the time-consuming and often tedious process of negotiating a licensing agreement. Industrial groups are typically no better than academics when it comes to releasing code. Companies may believe that software provides a competitive advantage and prevent employees from publishing

Walters, W. P. (2013). Modeling, informatics, and the quest for reproducibility, *JCIM*, *53*(7), 1529-1530.

The Tide Has Turned (For Some Journals)

Science AAAS

We require that all computer code used for modeling and/or data analysis that is not commercially available be deposited in a publicly accessible repository upon publication.

nature
International weekly journal of science

If published, the software application/tool should be readily available to any scientist wishing to use it for non-commercial purposes, without restrictions (such as the need for a material transfer agreement).

Policy Changes Are Necessary for Some Other Journals

	Editorial Policy for Reproducibility	Git Repository	Frequent Inclusion of Source Code
Journal of Computer-Aided Molecular Design	✗	✗	✗
molecular informatics (models – molecules – systems)	✗	✗	✗
Journal of Cheminformatics	✅	✅	✅
JCIM Journal of Chemical Information and Modeling	✗	✗	✗
JCTC Journal of Chemical Theory and Computation	✗	✗	✗

Guidelines Can Have a Profound Impact on Publication Quality

Journal of Medicinal Chemistry Editorial

Martin Stahl
Editorial Advisory Board Member

Jürgen Bajorath
Senior Editor

J. Med. Chem. **2011,** *54,* 1–2 **1**
DOI: 10.1021/jm1013055

Computational Medicinal Chemistry

Approximately one-third of current submissions to the *Journal of Medicinal Chemistry* (Journal) report computational work of varying weight and complexity. To ensure a high level of consistency in evaluating computational studies, the Journal extends and further refines the current requirements and acceptance criteria for computational manuscripts (specified in the January 2010 revision of the Guidelines for Authors, sections 2.3.5 and 2.3.6.). These revisions especially focus on combined experimental and computational studies, given the large number of submissions that fall into this category.

1. Predictive Use of Computational Methods

The submission of manuscripts that report the prospective computational design and experimental evaluation of new chemical entities is highly encouraged. Applications of model-

- Models and hypothetical statements must be clearly distinguished from experimental observations both in the text and in figure captions.
- Computational methods must be described in sufficient detail for the reader to reproduce the results.
- Computational methods must be thoughtfully selected and not used uncritically. It should be explained why the applied method is an appropriate choice and why it was chosen over similar methods, if they exist. Calculation results, in particular those of automated modeling software, must be critically examined.
- Conclusions from modeling must be drawn with an appropriate amount of caution, under consideration of all assumptions made, and within the accuracy limitations of the applied computational methods.

Maybe We Need a Checklist?

Are your datasets realistic?

Did you report both correlation and regression statistics?

Did do your statistics have confidence intervals?

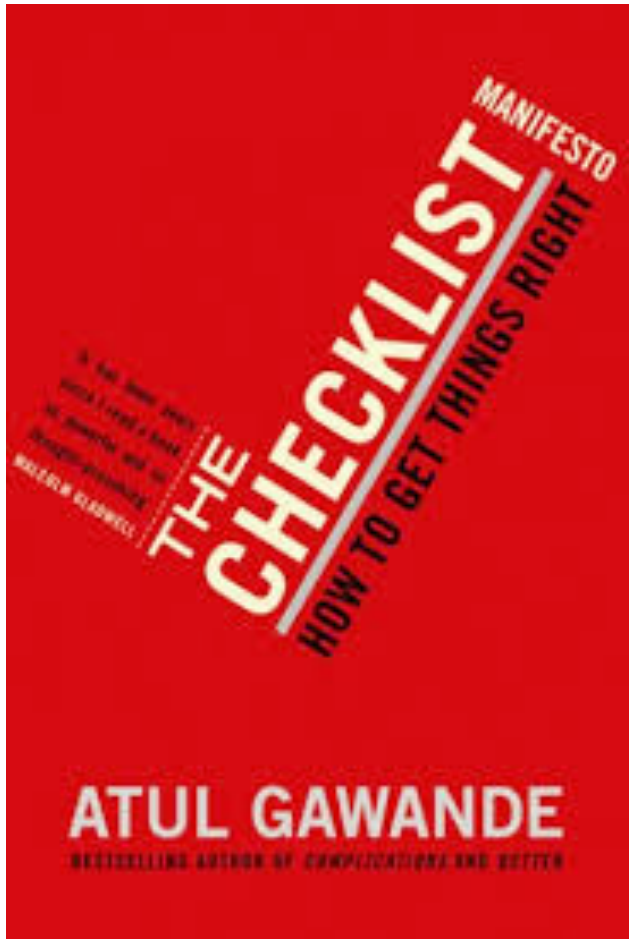
Are your plots meaningful?

Did you perform appropriate statistical comparisons?

Did you consider the impact of multiple comparisons?

Did you estimate the maximum possible correlation?

Can your method and results be easily reproduced?



Maybe We Need Something Else?

Are your datasets realistic?

Did you report both correlation and regression statistics?

Did do your statistics have confidence intervals?

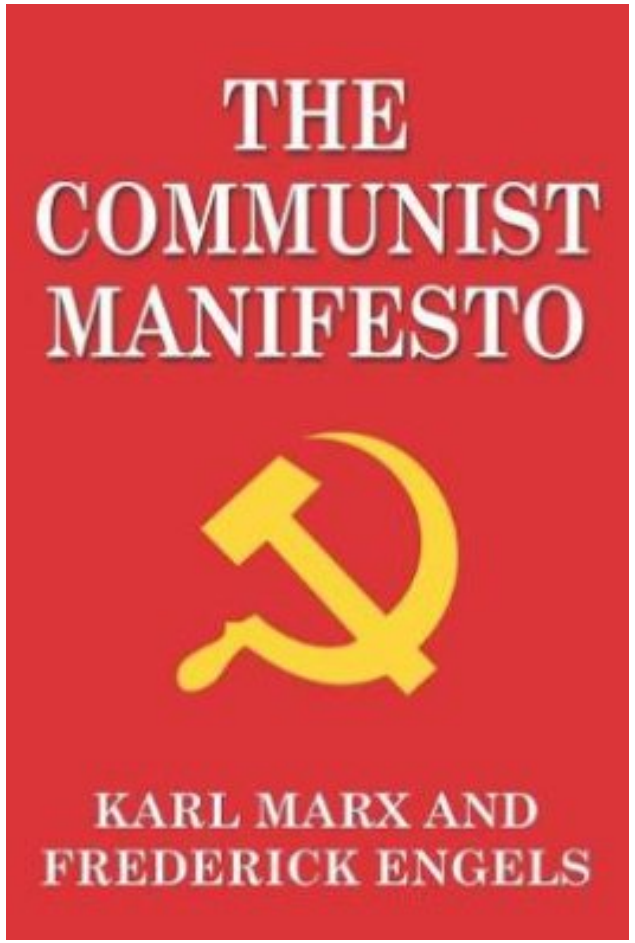
Are your plots meaningful?

Did you perform appropriate statistical comparisons?

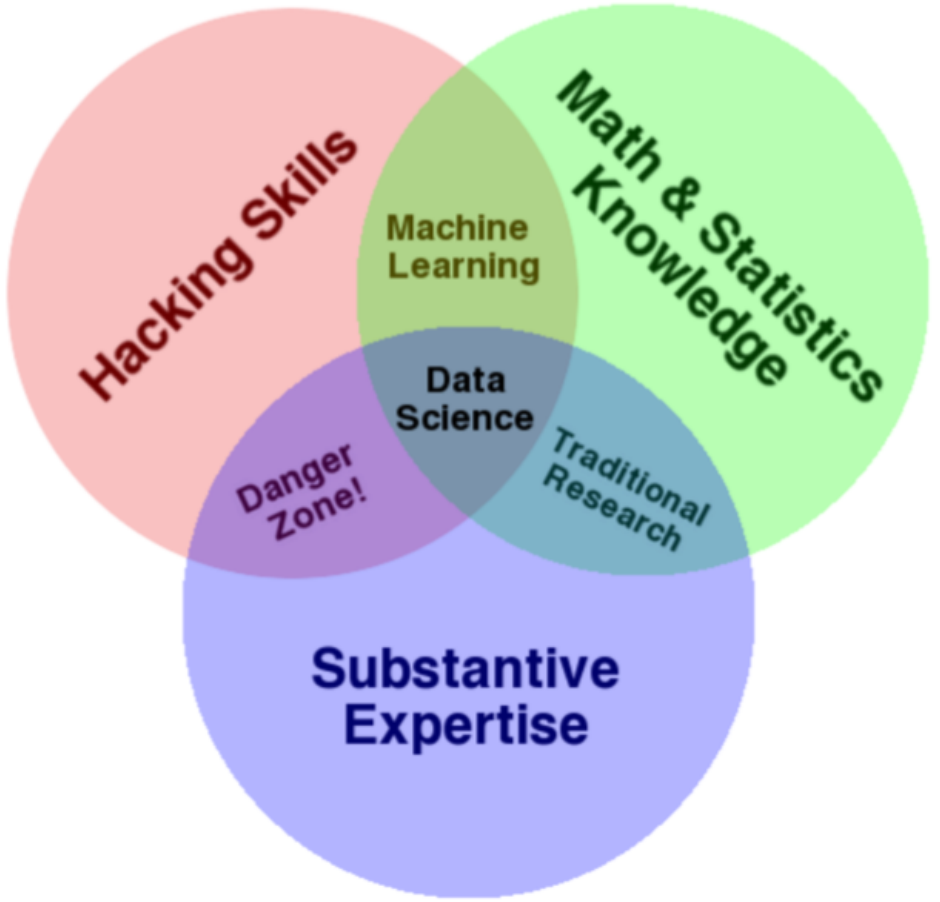
Did you consider the impact of multiple comparisons?

Did you estimate the maximum possible correlation?

Can your method and results be easily reproduced?



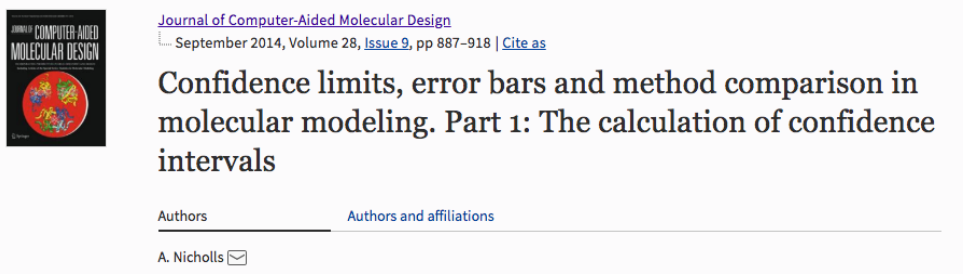
What Do We Need To Succeed?



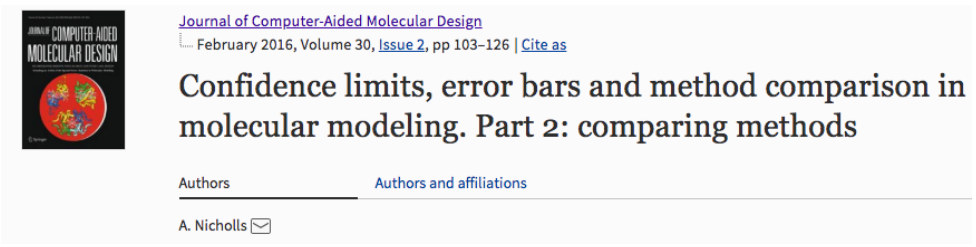
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Essential Reading

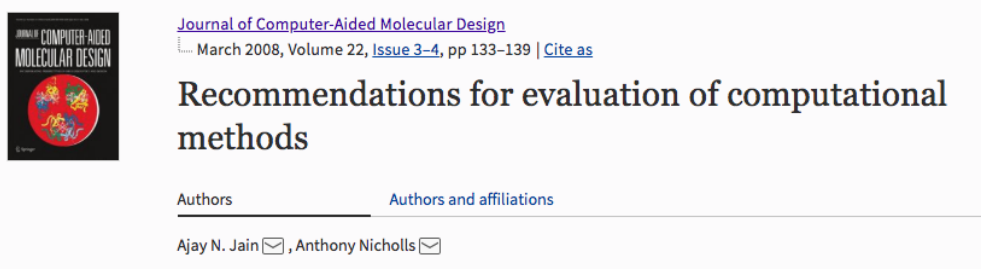
Nicholls A. What do we know and when do we know it?. *J Comput Aided Mol Des*. 2008;22(3-4):239–255. doi:10.1007/s10822-008-9170-2



Nicholls A. Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. *J Comput Aided Mol Des*. 2014;28(9):887–918. doi:10.1007/s10822-014-9753-z



Nicholls A. Confidence limits, error bars and method comparison in molecular modeling. Part 2: comparing methods. *J Comput Aided Mol Des*. 2016;30(2):103–126. doi:10.1007/s10822-016-9904-5



Jain AN, Nicholls A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des*. 2008;22(3-4):133–139. doi:10.1007/s10822-008-9196-5

https://github.com/PatWalters/metk

metk

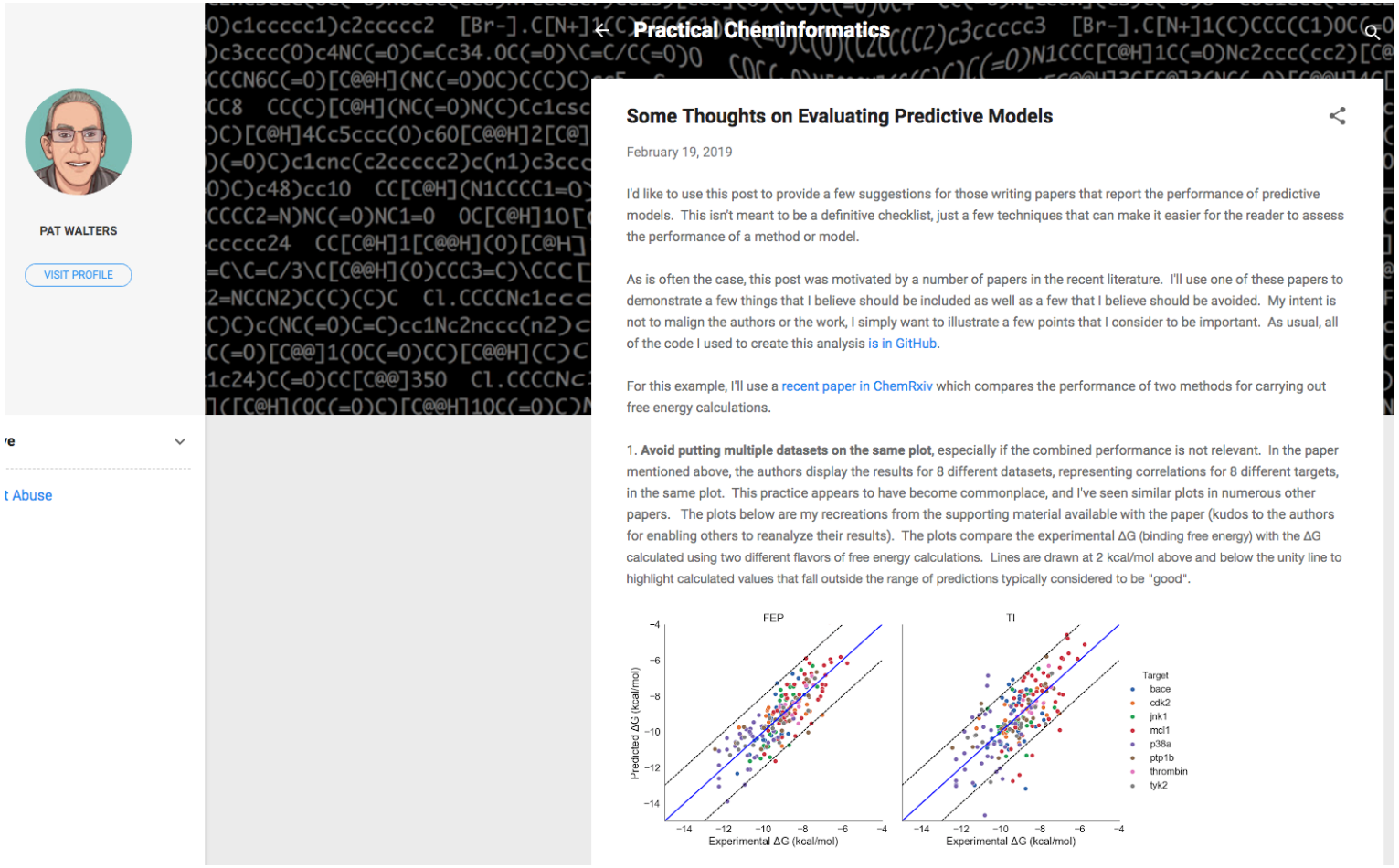
Model Evaluation Toolkit

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 confience limits)
- Maximum possible correlation given a specific experimental error. This is based on on a 2009 paper by Brown, Muchmore and Hajduk

Acknowledgements

https://practicalcheminformatics.blogspot.com




SPECIAL ISSUE
New Trends in Virtual Screening
Guest Editors: Renxiao Wang and Patrick Walters
JCIM
JOURNAL OF CHEMICAL INFORMATION AND MODELING
Deadline: December 31, 2019

RELAY

THERAPEUTICS