Just.®

biotherapeutics for all

Standardization and Quality Considerations for Machine Learning From Physical Protein Samples

Jeremy Shaver, Ph.D.

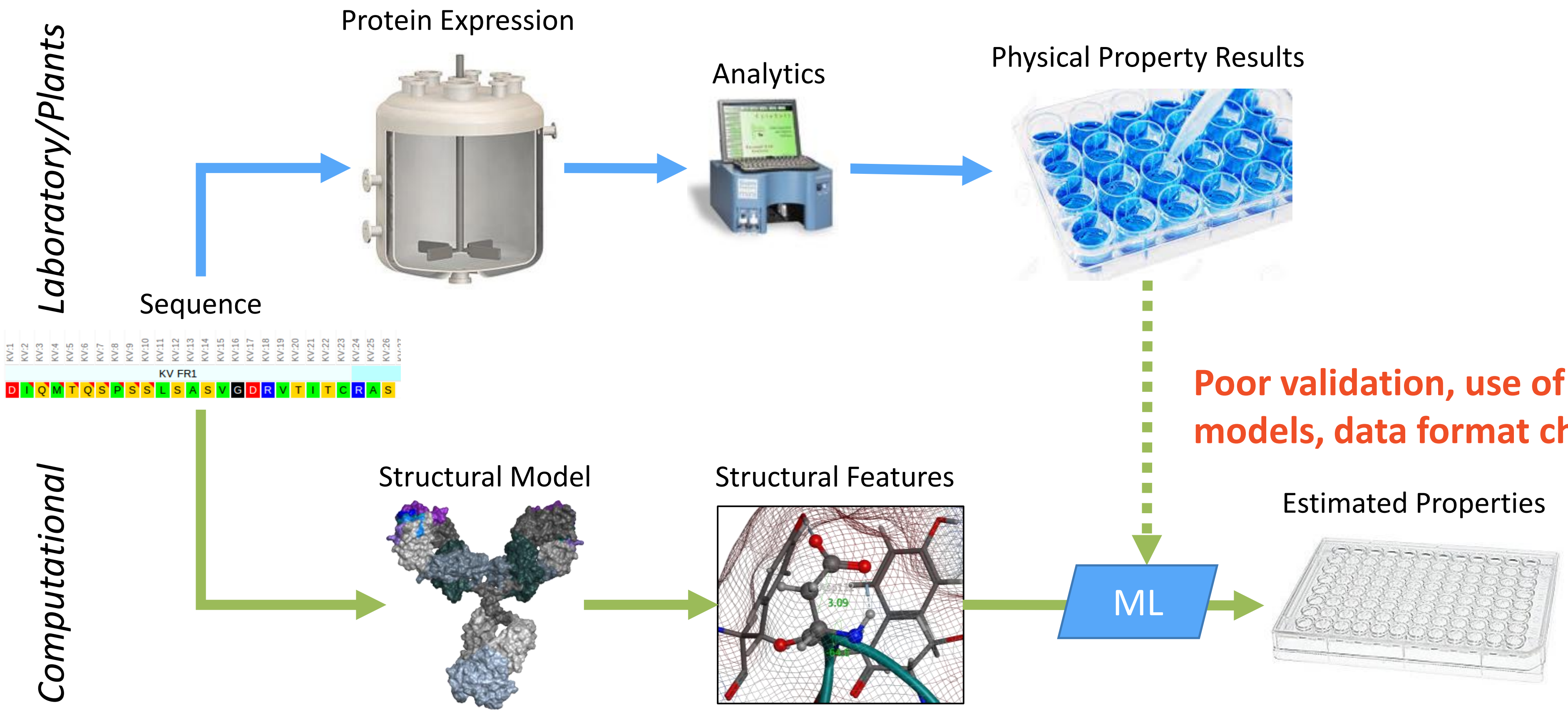
November 14, 2018

Computational Drug Development for Biologics

Mission: Design and apply innovative technologies to dramatically expand global access to biotherapeutics

Challenges are present at various points in the QSAR pipeline

Variations in: glycosylation, formulation, instruments, procedures + EXPENSE $$$$



Poor validation, use of greedy models, data format challenges

Inaccuracies in: structural estimates, feature abstractions, static vs. dynamic models

Challenges of machine learning demonstrated in some recent biologics QSAR publications

Two scenarios published in the last 18 months, but typical of MANY publications

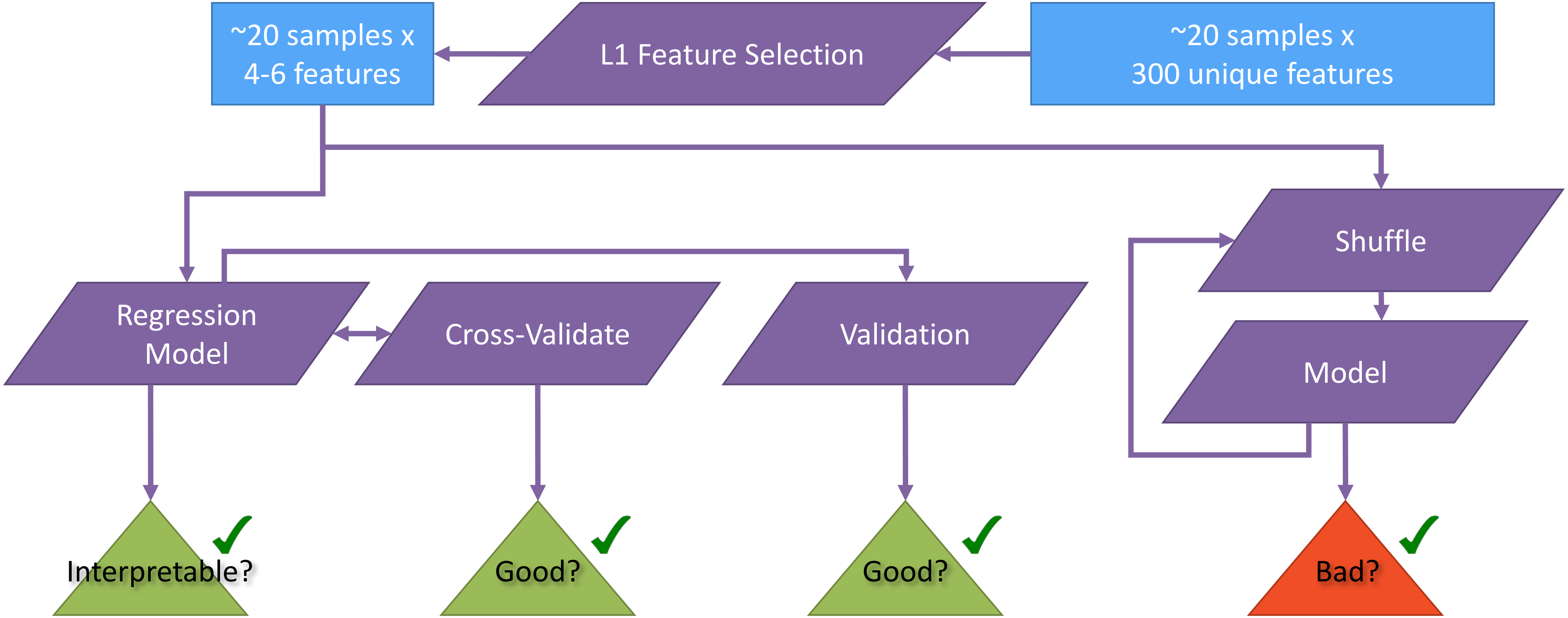
A. Prediction of downstream purification behavior

   Feature selection from set of highly-refined structure features followed by SVM and PLS

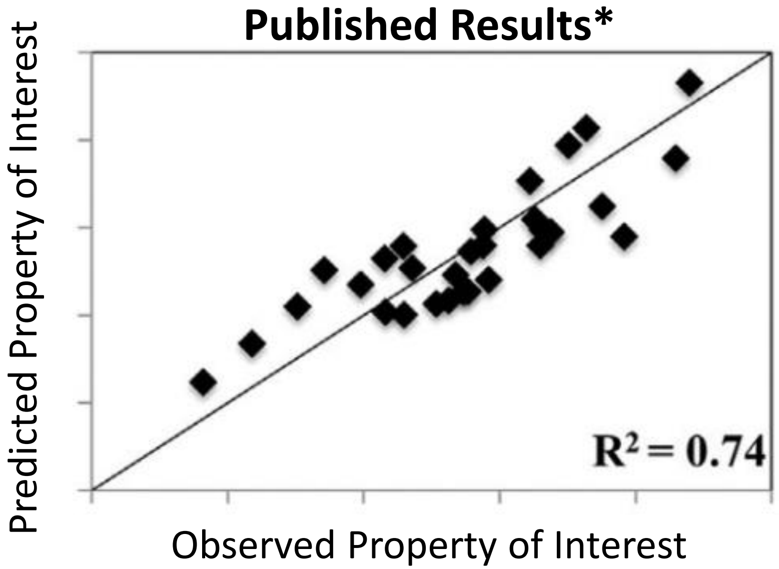
B. Prediction of molecular properties

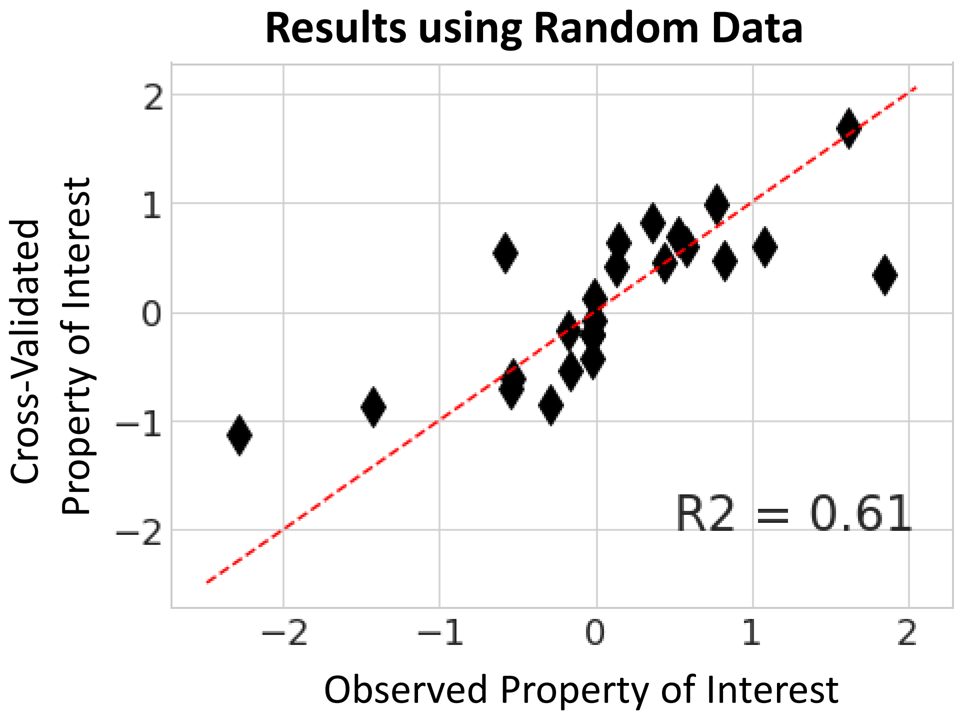
   Data augmentation (for non-linear behaviors) followed by selection and linear modeling

Example A: Downstream properties from structure properties



Results are indistinguishable from random chance

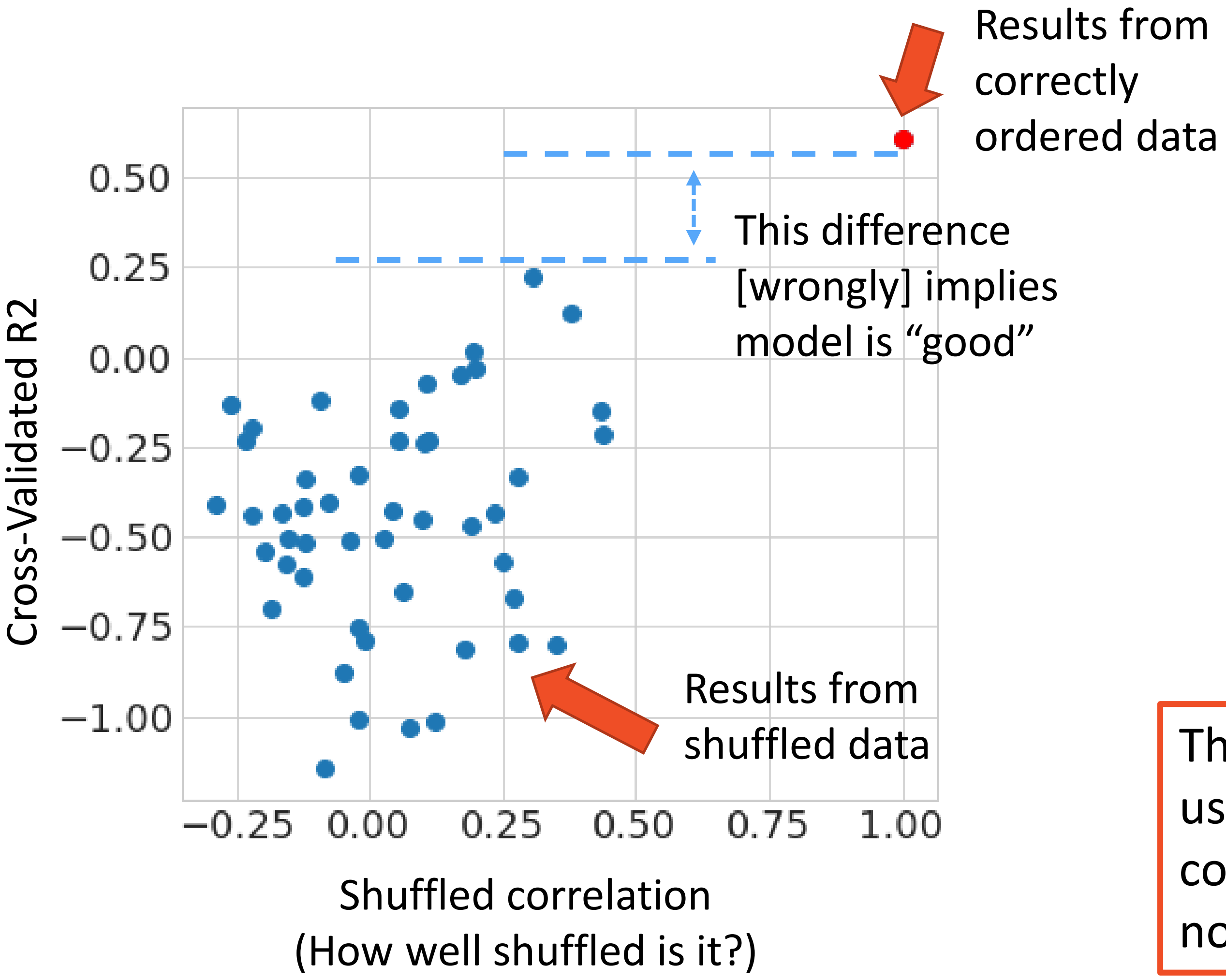




* It's not clear from the publication if this was self-prediction or cross-validation

From random data created in the same size as in paper

Shuffled-data test shows equally “good” results with random data



Results from correctly ordered data

This difference [wrongly] implies model is “good”

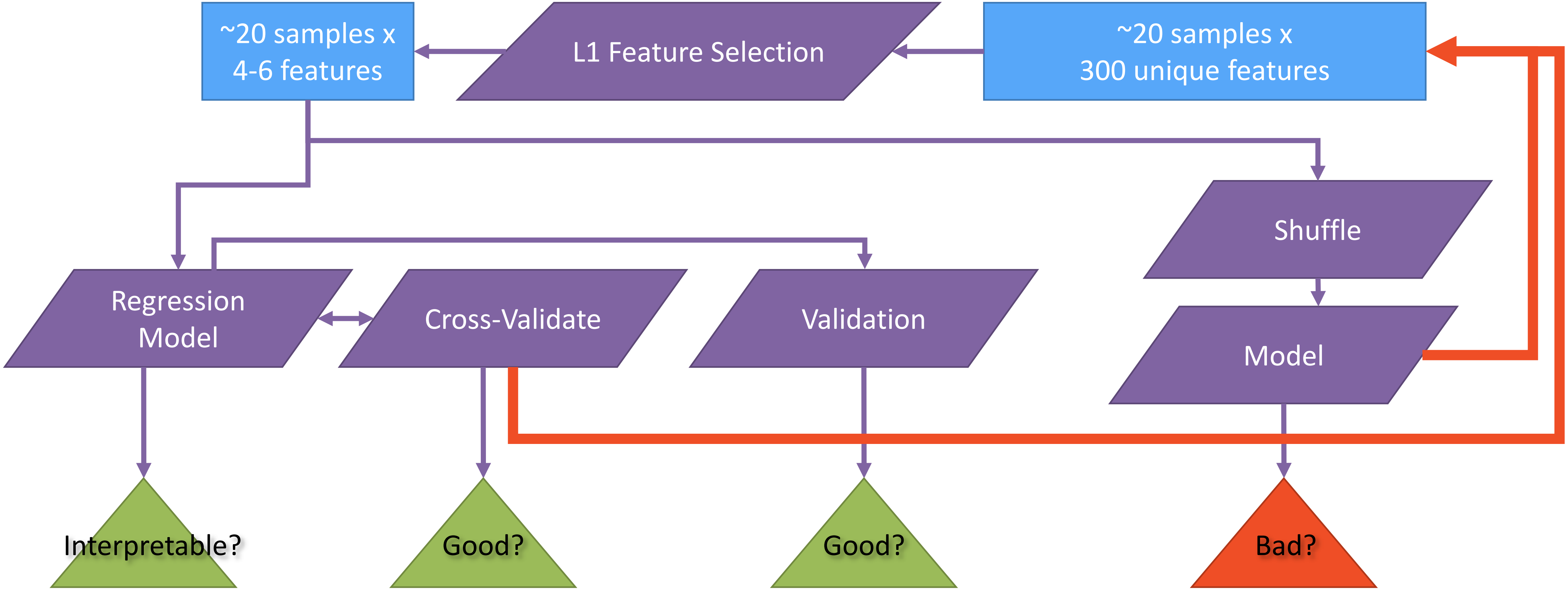
Results from shuffled data

If a model is fitting random data, shuffling the samples should give the same results as when the samples are in the “correct” order.

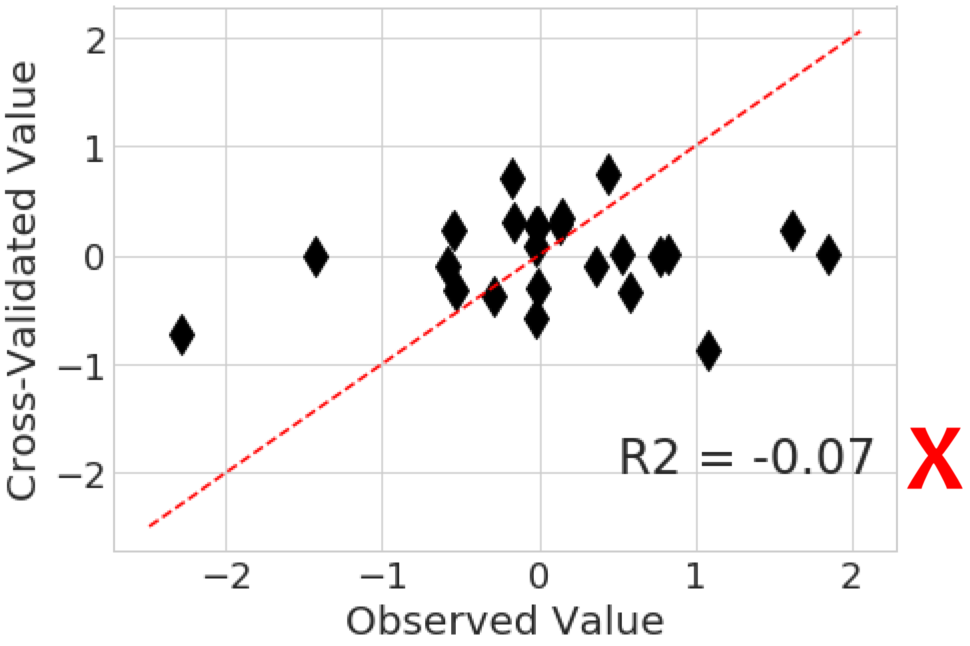
This test appears to pass because we have pre-selected the subset of features!

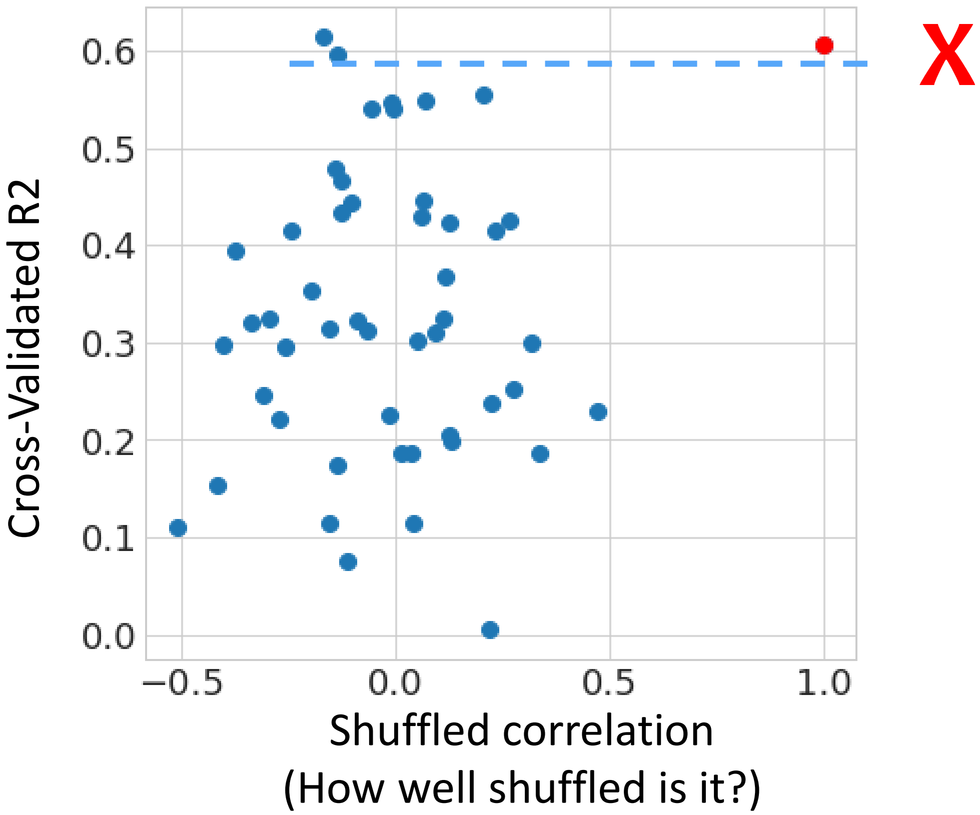
This paper referenced a website as the tool used to calculate their models. That website contains the modeling flaw, but the website is no longer supported or operational.

Correct testing requires looping over entire process

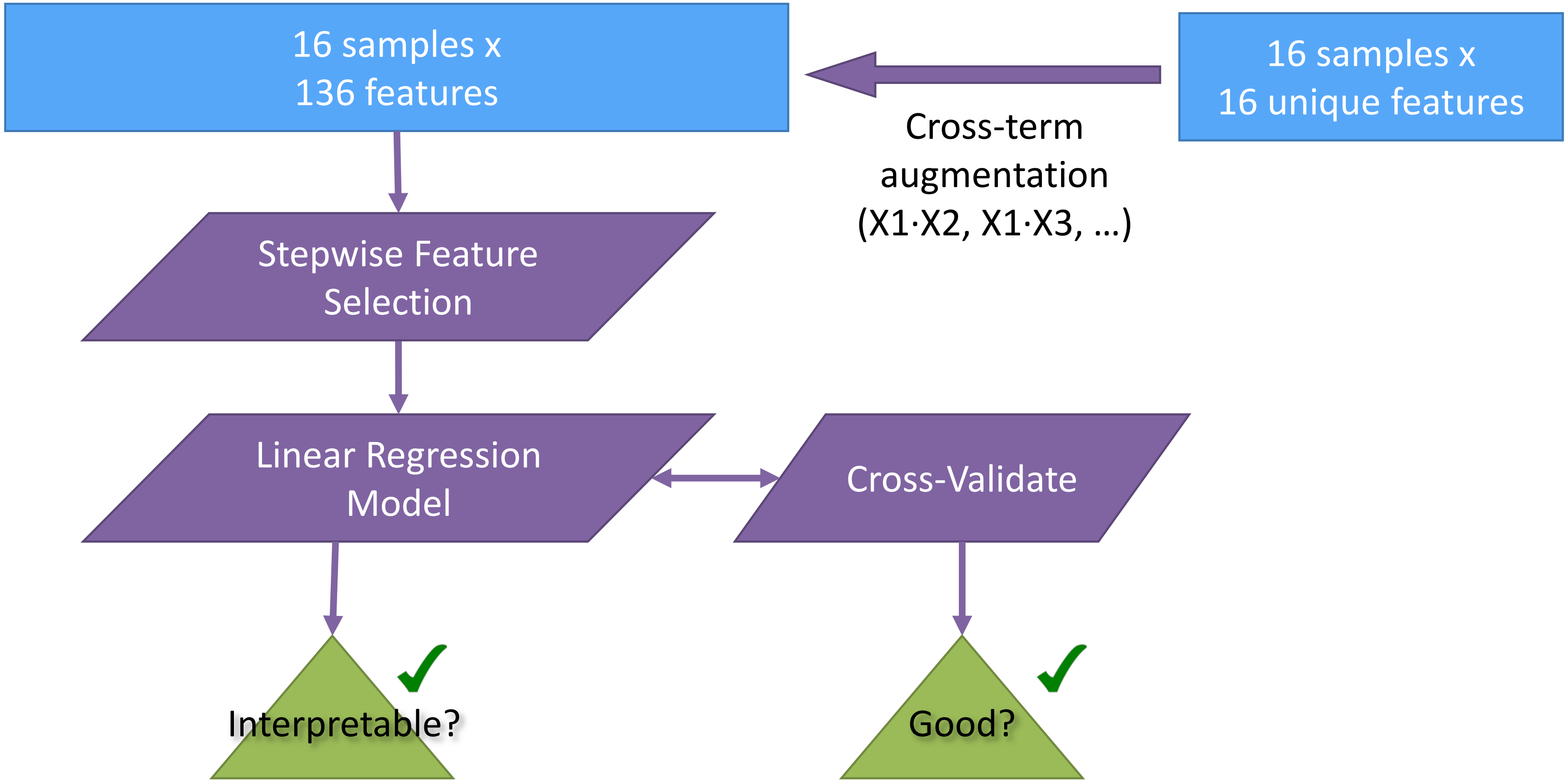


Including the feature selection step in cross-validation and shuffle testing correctly identifies the failure

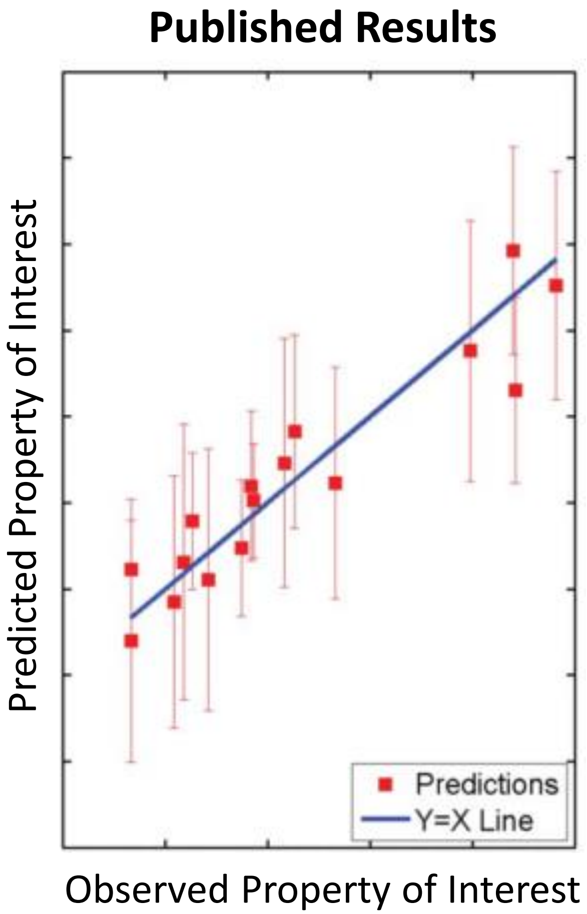


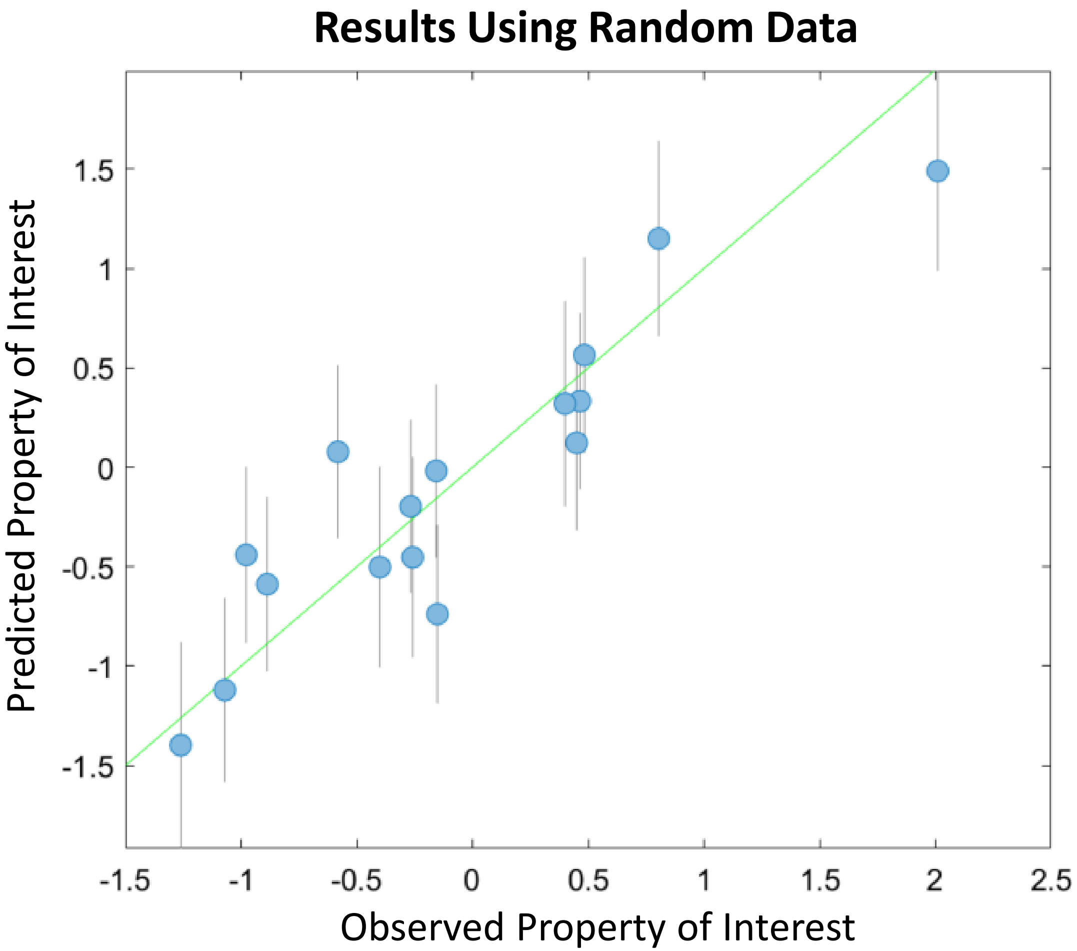


Example B: Molecular properties from structure features



Example B: Cannot distinguish real “interesting results” from overfit results





16 samples, 16 features augmented with with 120 interaction terms (X1*X2, etc)

In spite of the fewer features and higher relatedness of them, the model is still indistinguishable from random chance.

How to actually address issue of troublesome machine learning?

You **cannot** avoid this with more machine learning. It can only be avoided by using more data.

More data from where?

- “Meta analyses” combining data from public data sources?

  Challenging without metadata, consistent analytical methodology and/or bridging studies

- Joining data from multiple actively-collaborating labs?

  Better chance for merging, but still challenging

Case study: selected molecules from

Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci.* 201616408 (2017). doi:10.1073/pnas.1616408114

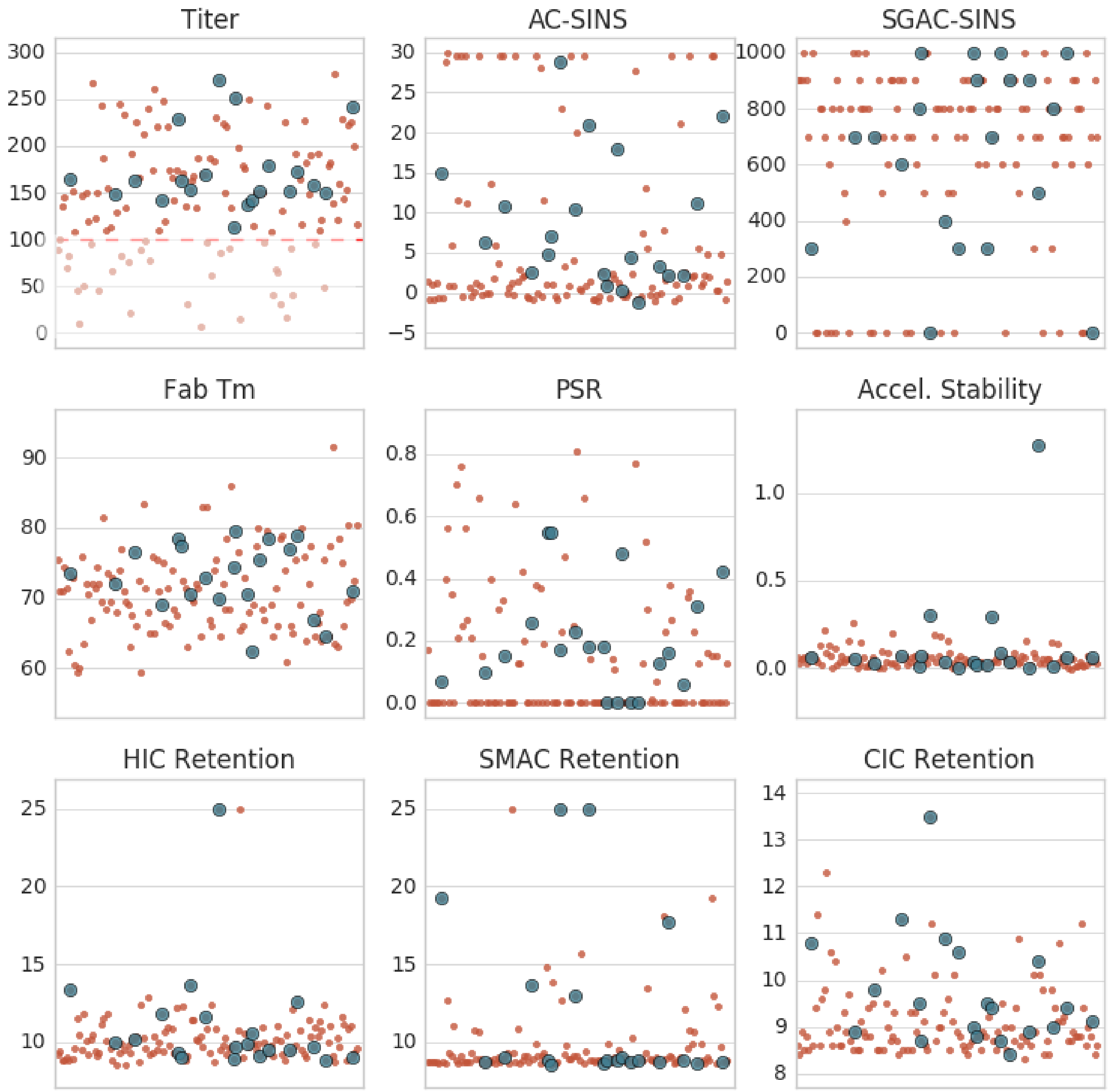
Selecting an “Interesting” set of Molecules to produce in-house

Using published data:

- Hard threshold on Titer (>100 required)
- Choose 20 molecules maximizing distribution across AC-SINS/SGAC-SINS space (Kennard Stone selection)
- Manually substitute 2 molecules to cover Accelerated Stability and Retention data (out of interest)

● Available Jain molecules

● Selected molecules



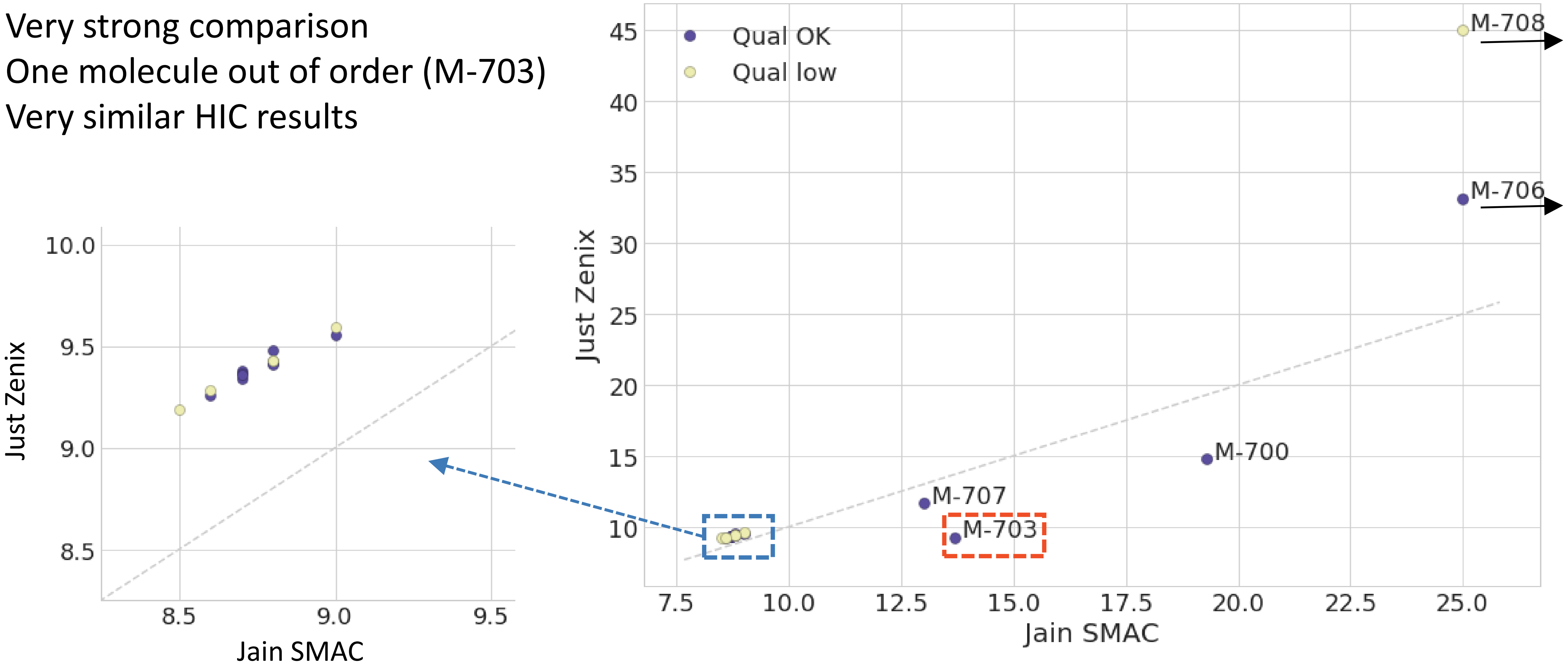
137 Commercial Antibodies from Jain 2017

Comparison of SMAC to Zenix retention times

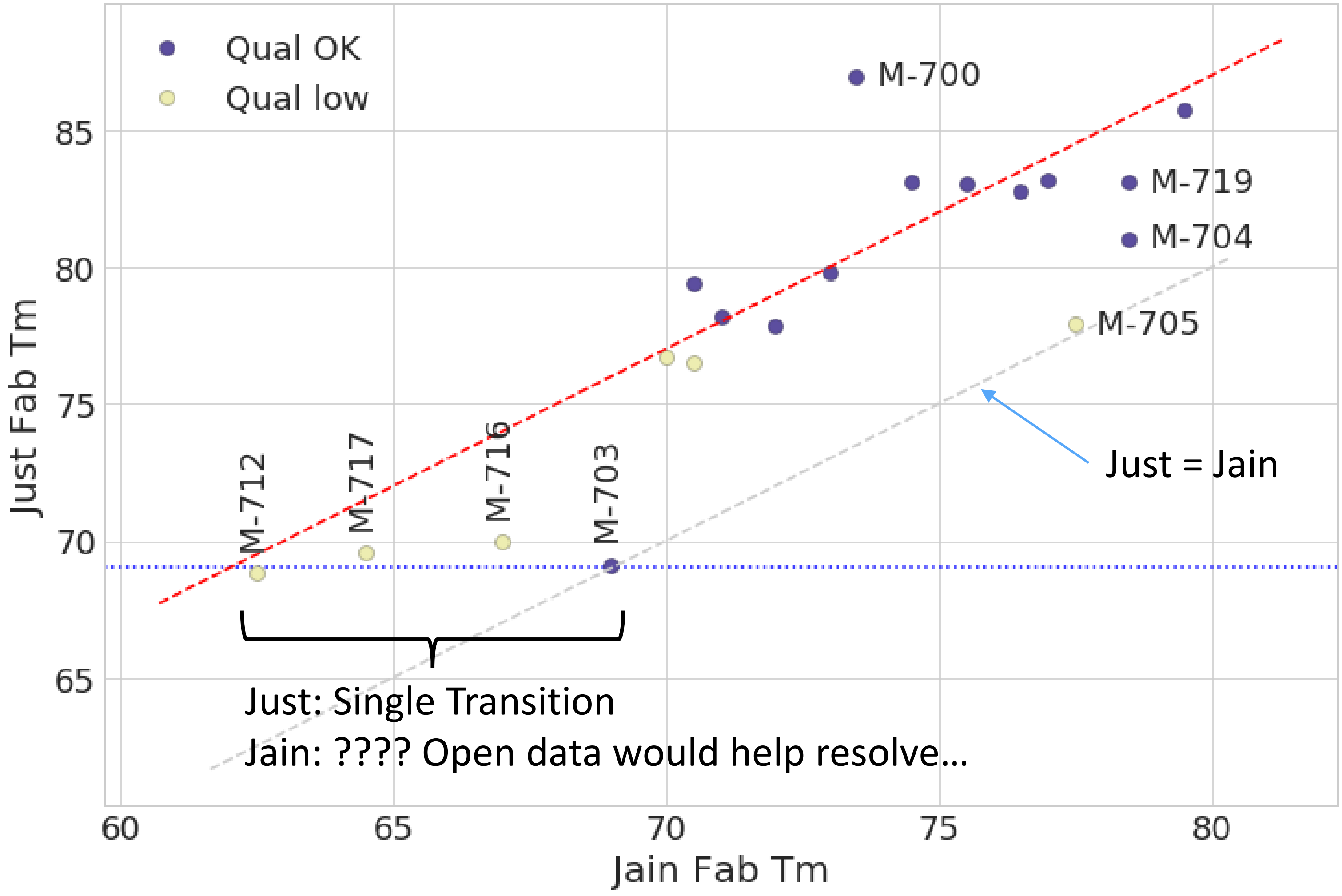
Very strong comparison

One molecule out of order (M-703)

Very similar HIC results



Comparison of Thermal Stability: Similar linear trend



Just: Single Transition

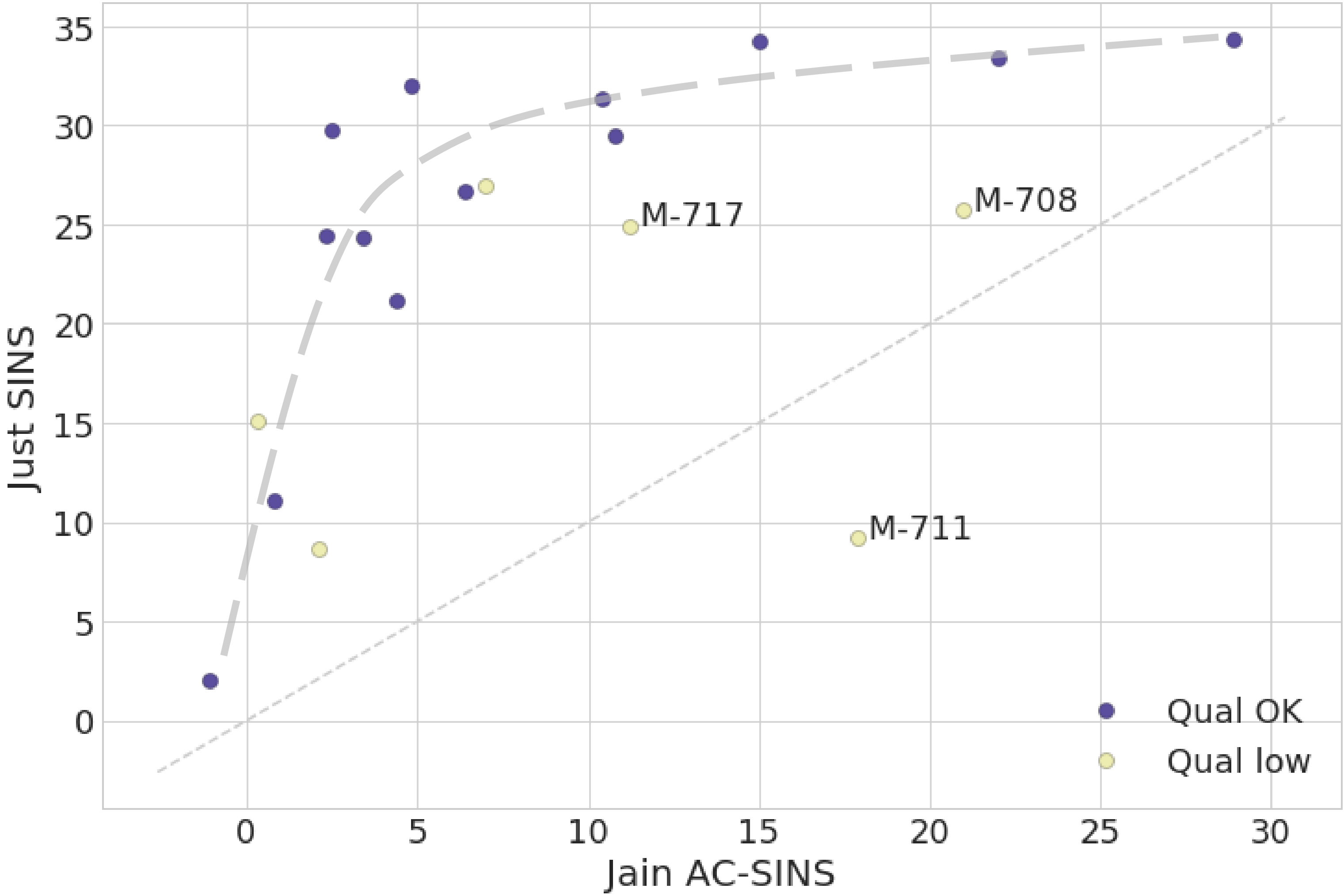
Jain: ???? Open data would help resolve…

Results for good quality molecules show similar trend, but offset (shows as bias in regression model)

Primary experimental difference: higher concentration of dye in Jain work. Possibly inducing chemical denaturation (thus lower temperature unfolding)

Red line: Just = Jain + ~7 deg

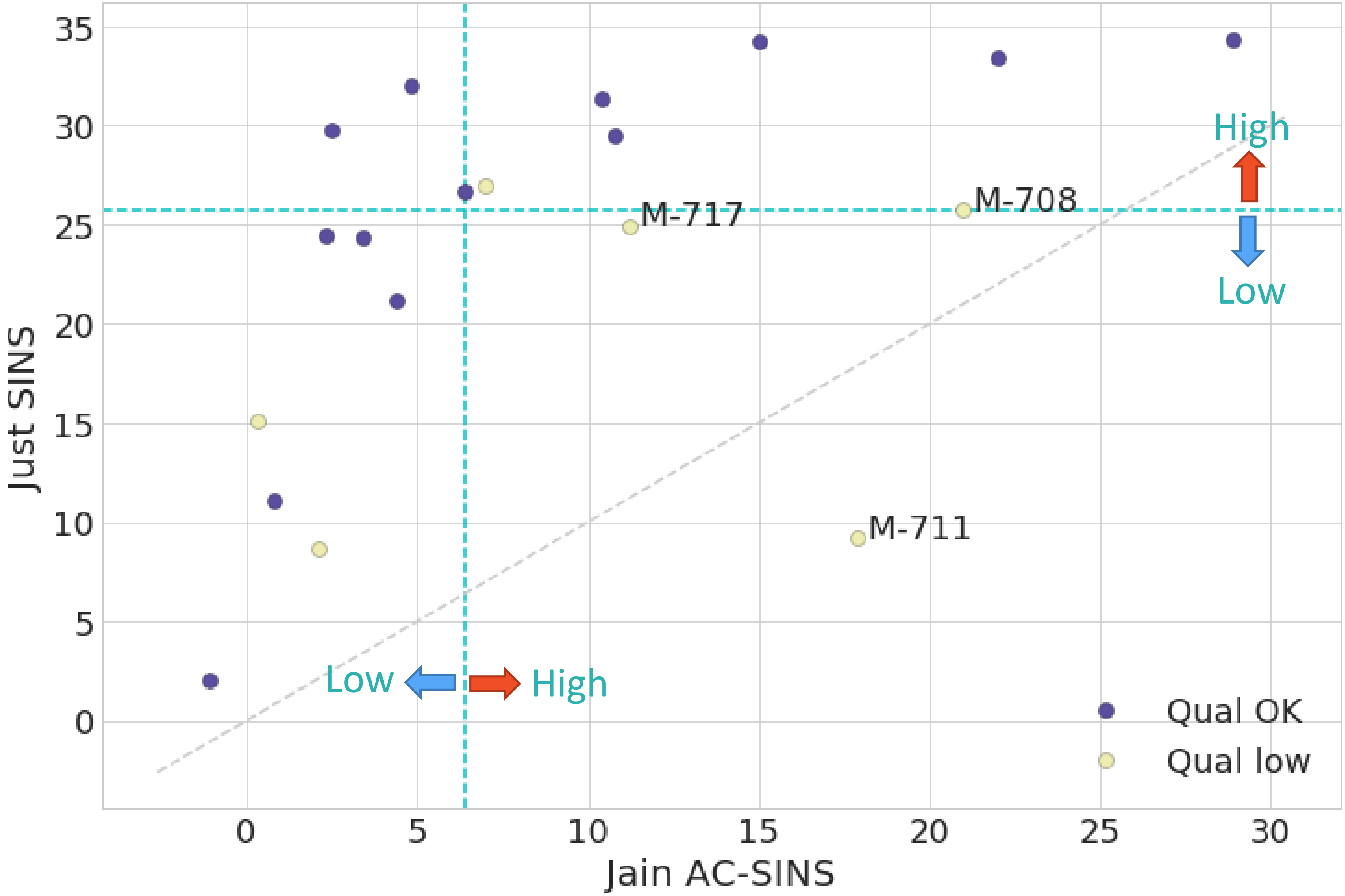
Comparison of AC-SINS results (suspected relation to viscosity)



Non-linear relationship (adjusting conditions gave similar results)

Regression would be seriously challenged by these results

Comparison of AC-SINS results (suspected relation to viscosity)



If treated like a "High SINS" classification model for top 50%, samples in top left and bottom right quadrants would be differently classified (4 out of 19) = 20% error rate

Conclusions (1)

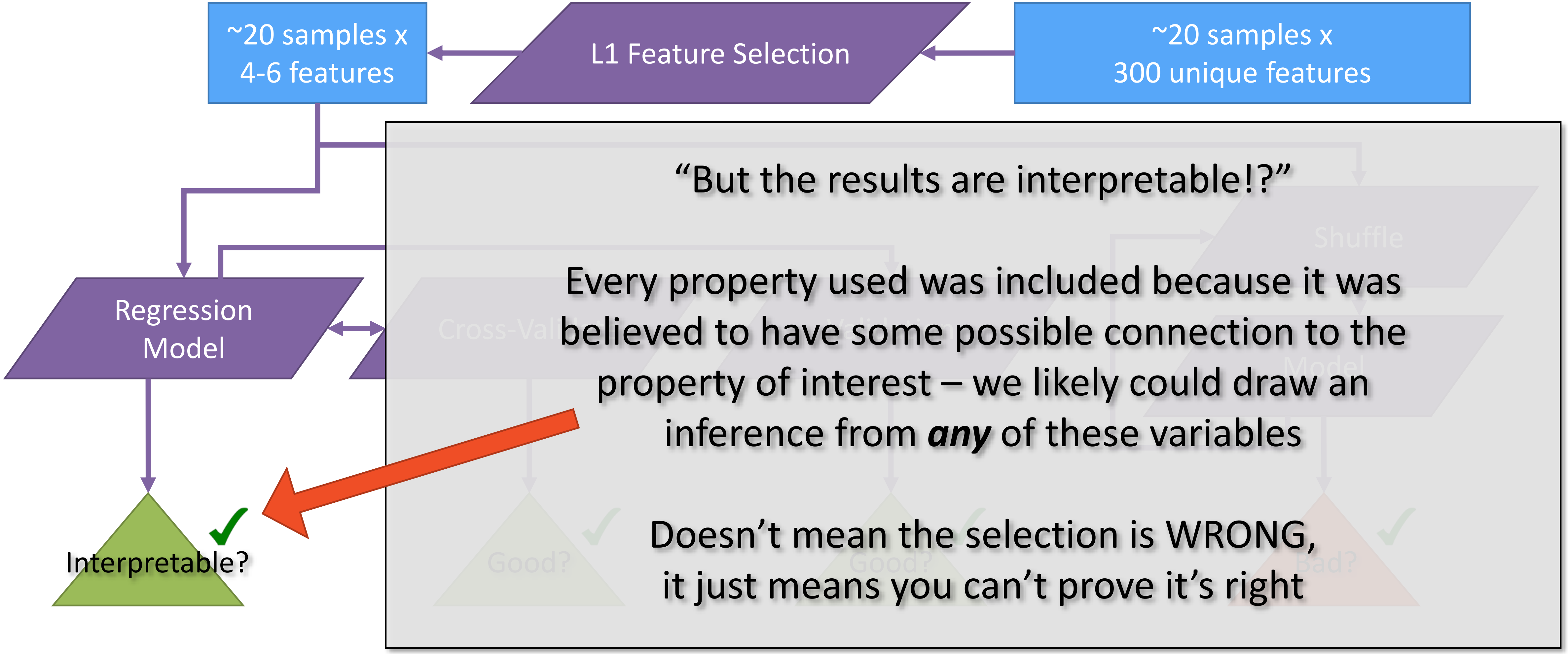
- Detect machine learning faults with:
  - Use of *full* cross-validation, scrambling, and strong validation tests. Test using random data.
  - More careful use of methods – researchers need to watch for traps and look for these errors when reviewing publications
  - Vendors must add machine learning CAREFULLY

- Joining data from well-documented publications is challenging
  - Expression differences, assay differences, instrumental differences
  - Meta-analyses of data from different publications is likely almost impossible
  - Unless…

Conclusions (2)

- Open code? Access to modeling code would allow validation of methods
  - Other scientific communities do this
- Open data? Raw data access would allow comparison of computational methods
  - What format?
  - Do we need to prove reproducibility?
  - Sequence??
- Open *materials?* Access to physical material and sequences? NIST?!

But what about the “Interpretability” test?



“But the results are interpretable!?”

Every property used was included because it was believed to have some possible connection to the property of interest – we likely could draw an inference from ***any*** of these variables

Doesn’t mean the selection is WRONG,
it just means you can’t prove it’s right