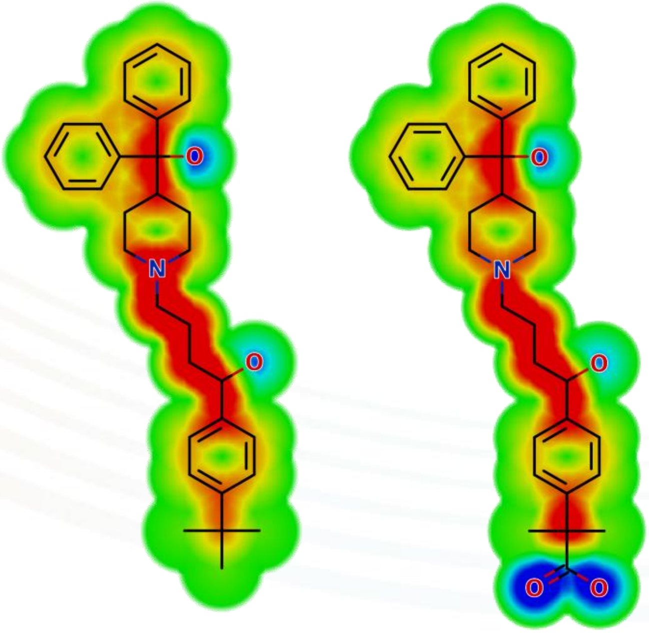
Quality Data to Quality Models

19th March 2018

Travis P. Hesketh – travis@optibrium.com

Modelling PK-ADME Targets

- QSAR models are a well established methodology
  - often used in industry
  - widely utilised in drug discovery
- The information they can provide is useful for prioritising synthesis
  - i.e. flagging up potential toxicity ensures that less time is wasted
- Where interpretable descriptors are used, this information can be used in design
  - if we know what makes a molecule have poor activity, we can change it



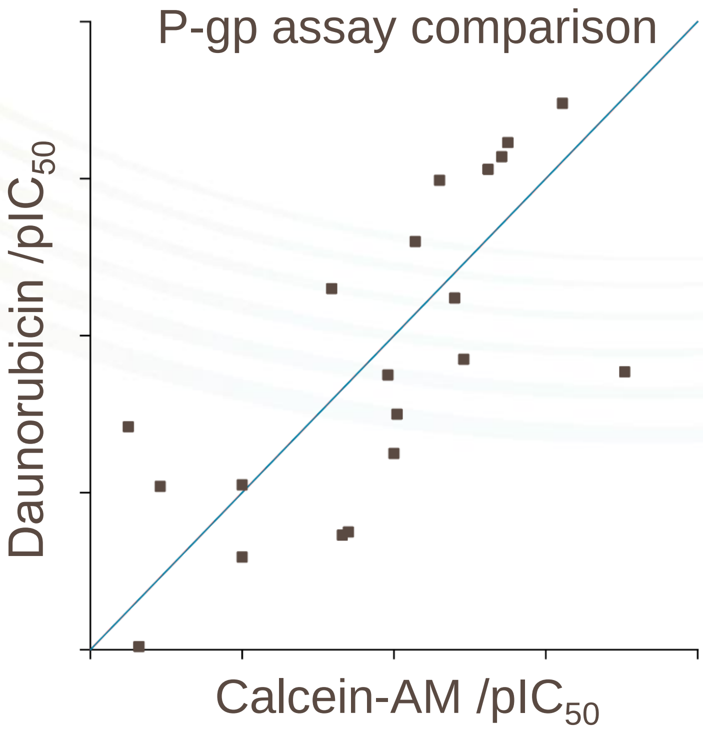
StarDrop's Glowing Molecule Visualisation of hERG inhibition for terfenadine (L) and fexofenadine (R)

Modelling Public Data

- Our biggest problem lies not in modelling the data, but in deciding *what data to model*.

- Public data sources are an incredibly useful resource, **but** suffer from:
  - inter-lab and inter-assay variability
  - misreported values
  - mis-abstracted values
  - structural variations

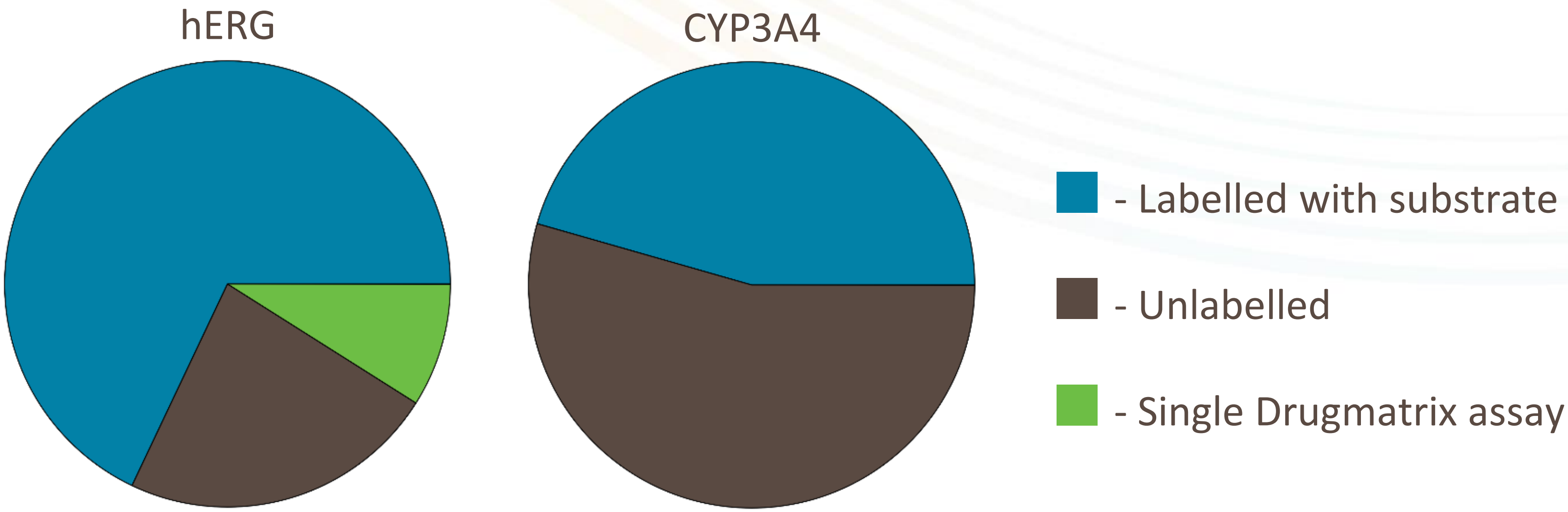
- Knowledge about measurement conditions *(metadata)* is critical



Modelling Public Data (Continued)

- In many cases this metadata is completely missing,
  - Data simply labelled 'Inhibition of X’
  - Problem is worse for some targets than for others (see below)

- This is often due to unreported conditions (or long chains of ‘see citation from paper Y’) in the primary literature.

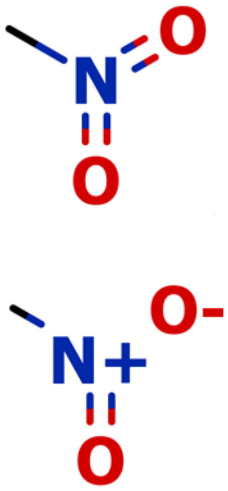


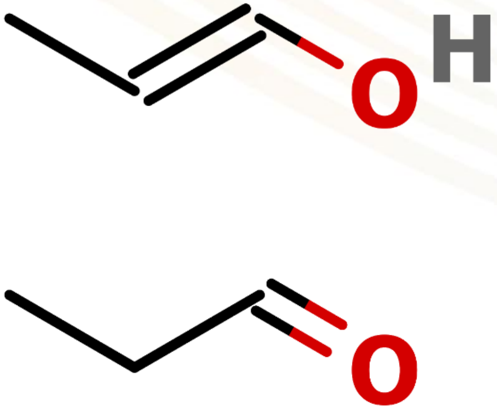
Data from **ChEMBL23** – http://ebi.ac.uk/chembl

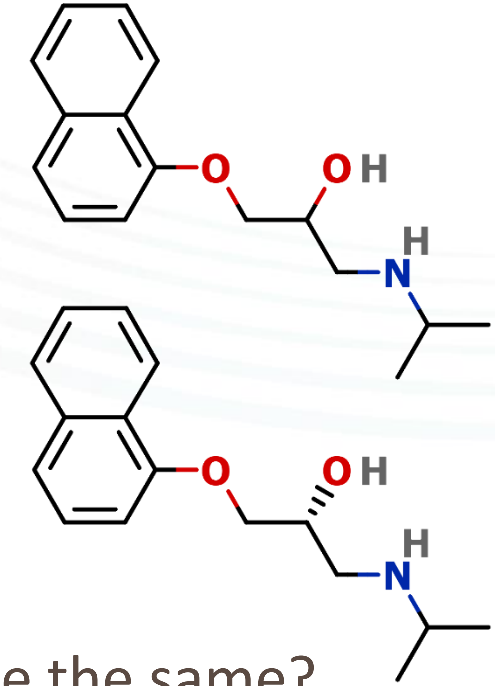
Chemical Structure Problems

- Other important considerations include treatment of group representations, tautomers and stereochemistry (pictured left to right below).

- These issues can occur in all databases.







- Can these pairs of molecules be said to be the same?

What Makes a Good Model?

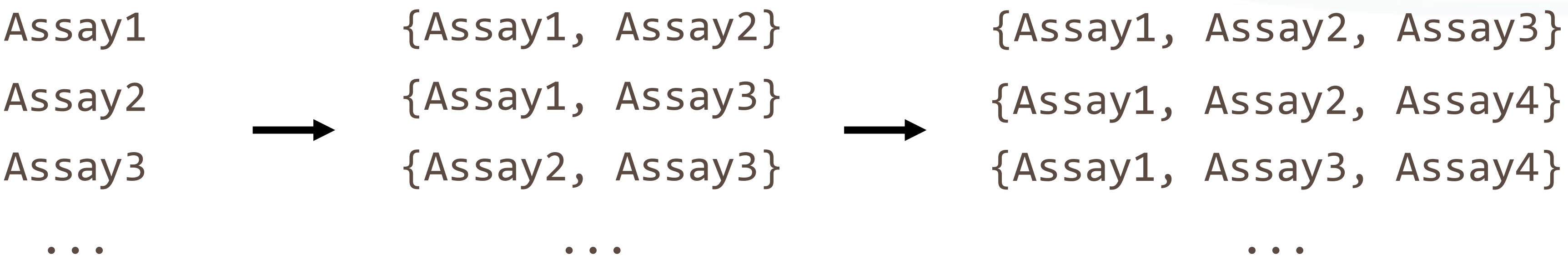
- A good model is highly subjective but some desirable qualities include:
  - large domain of applicability
  - high accuracy
  - a regression model

- To get a more accurate model, the training data should be as consistent as possible
  - Unchecked public data too variable to produce accurate models
  - Checking for consistency takes a very long time
  - Modelling only well labelled data can greatly decrease available data

- ‘QSARSetBuilder’ (QSB) helps with this process

The Rationale

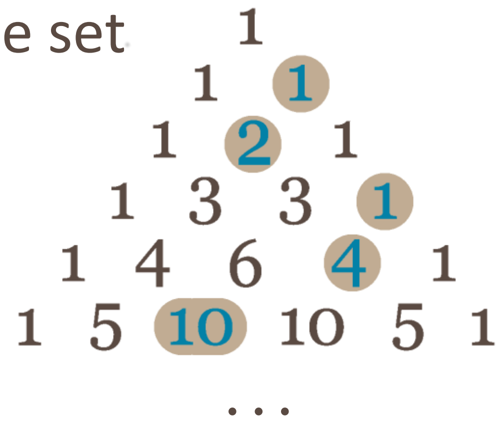
- Inconsistent data should produce poorer models, can we use potentially consistent data and then add to it whilst monitoring performance?
- Consider each ChEMBL 'assay' as a non-separable block of data and test models built from *every* combination of these blocks
- We could use the information about which assays commonly produce these good models to pick out better data



```
Assay1          {Assay1, Assay2}          {Assay1, Assay2, Assay3}
Assay2    →     {Assay1, Assay3}    →     {Assay1, Assay2, Assay4}
Assay3          {Assay2, Assay3}          {Assay1, Assay3, Assay4}
  ...                 ...                           ...
```

The Problem with Testing 'All the Sets'

- Too many combinations to test every possible set
  - Sample sets of assays of varying size instead



1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

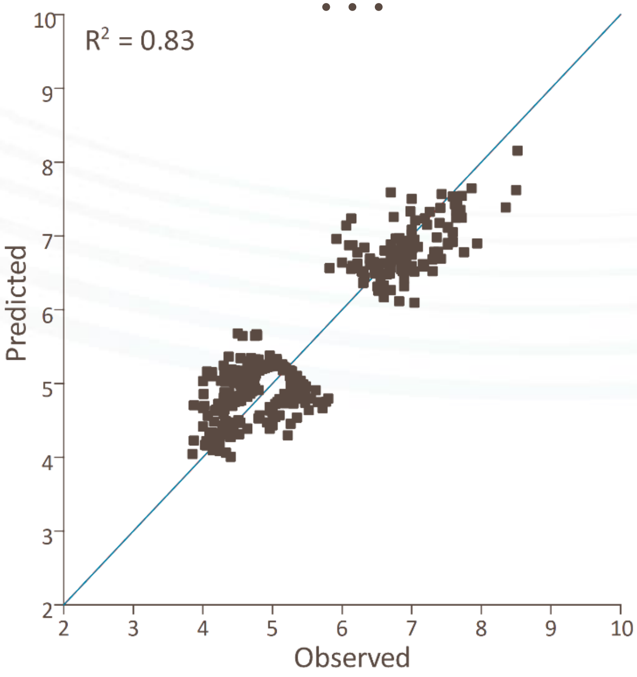
1 5 10 10 5 1

. . .

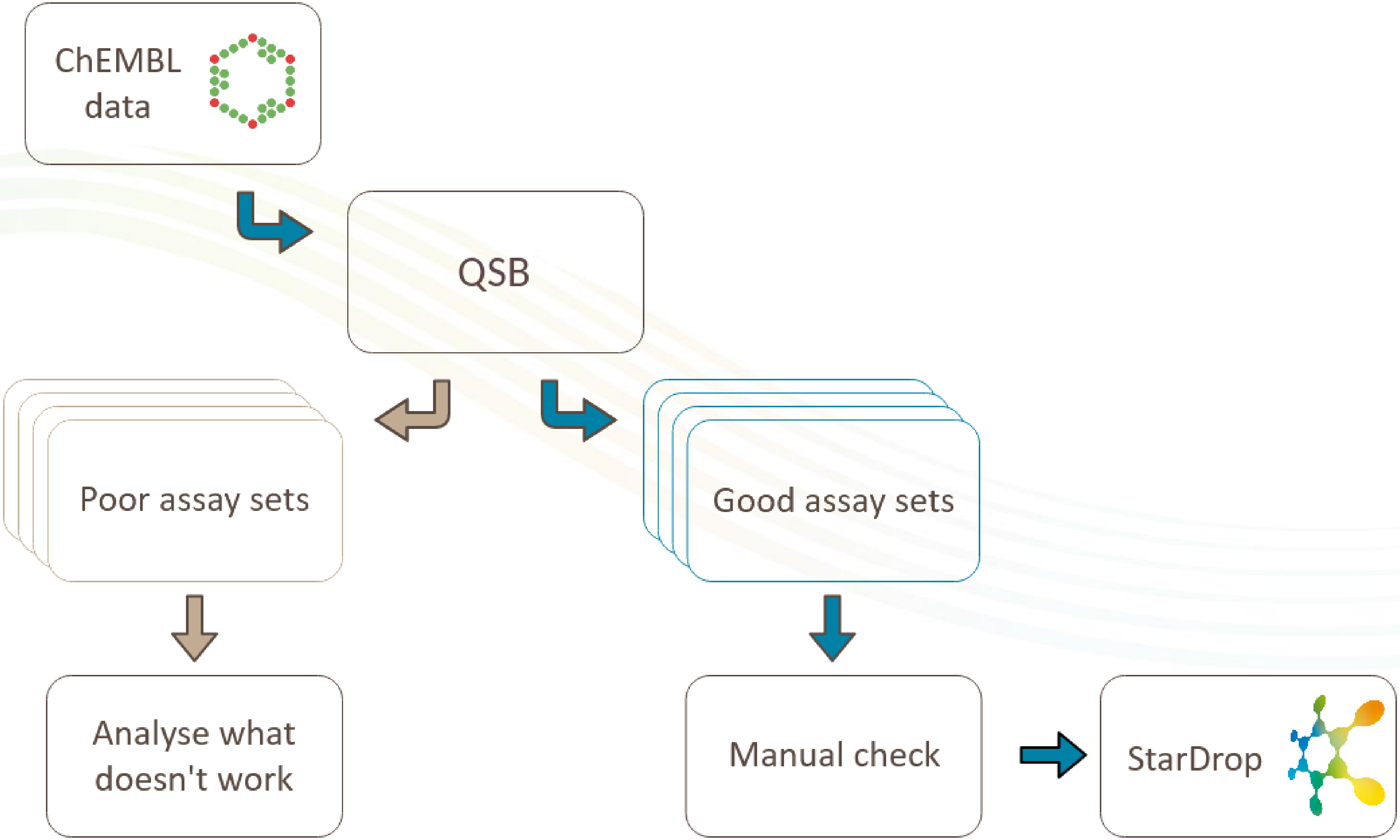
- Testing many sets reduces the influence of poor set choices which report good statistics (e.g. bottom right)

- The data we collect can be used to produce a finalised dataset for modelling: an *assisted* QSAR modelling approach

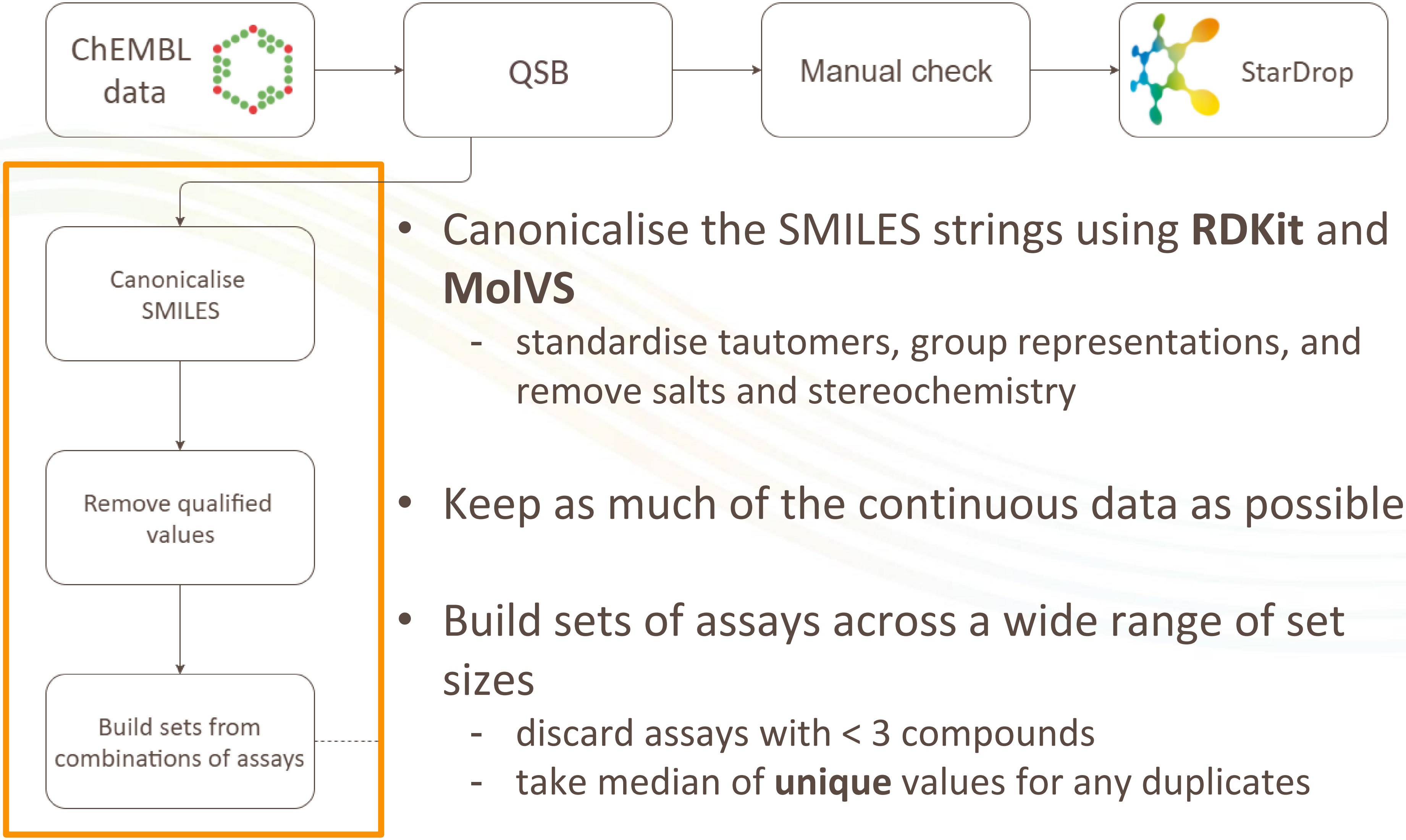
- R^2 is coefficient of determination: how well the points fit the identity line



The Current QSB Workflow



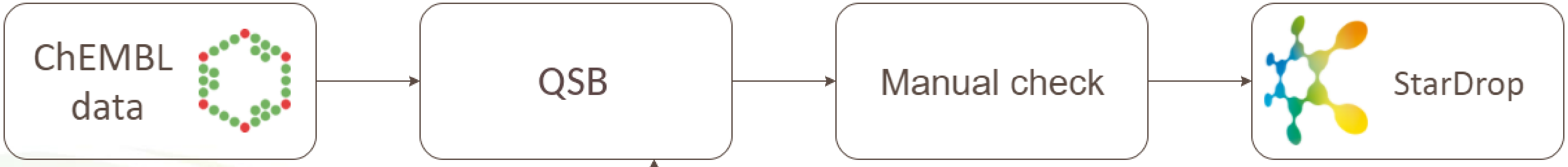
An Expanded Workflow



- Canonicalise the SMILES strings using **RDKit** and **MolVS**
  - standardise tautomers, group representations, and remove salts and stereochemistry
- Keep as much of the continuous data as possible
- Build sets of assays across a wide range of set sizes
  - discard assays with < 3 compounds
  - take median of **unique** values for any duplicates

MolVS – https://github.com/mcs07/MolVS RDKit – https://www.rdkit.org

An Expanded Workflow



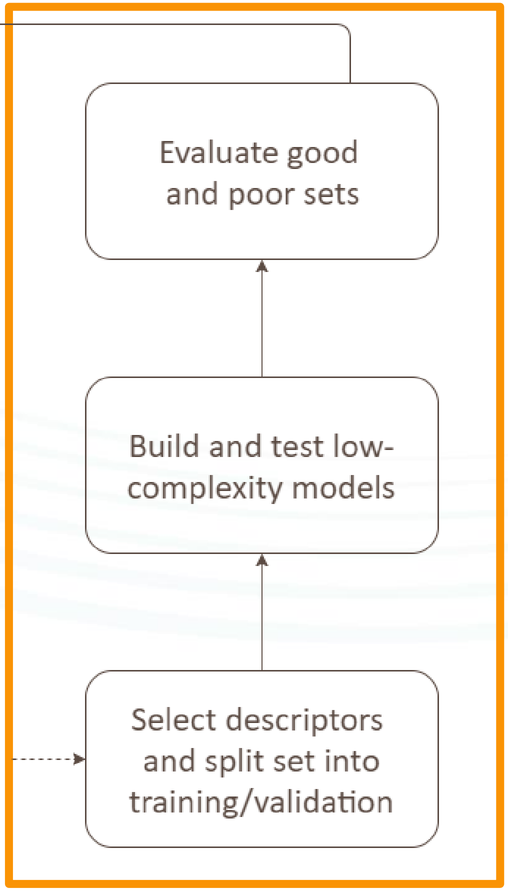
ChEMBL data

QSB

Manual check

StarDrop

- Build low complexity models for each set using **scikit-learn**
  - 70:30 split for training/validation set
  - Random Forests models with 30 descriptors
- Test model further using 5-fold cross validation over training set
  - Can also use an external test set from file
  - R^2 or Matthews Correlation Coefficient > 0.6 for both tests are ‘good’
- Produce report based on the good sets



Evaluate good and poor sets

Build and test low-complexity models

Select descriptors and split set into training/validation

Additional Detail

- Sets are built at first from favourably overlapping assays
  - having compounds in common whose PCHEMBL activities differ by < 0.5
  - additional sets are randomly assembled to target sizes (in compounds)

- The initial 97 descriptors include RDKit’s fragment library and some whole molecule descriptors (Log P, VABC, MWt, HBD sites, etc.)
  - Selection from these is done using scikit-learn’s Recursive Feature Elimination

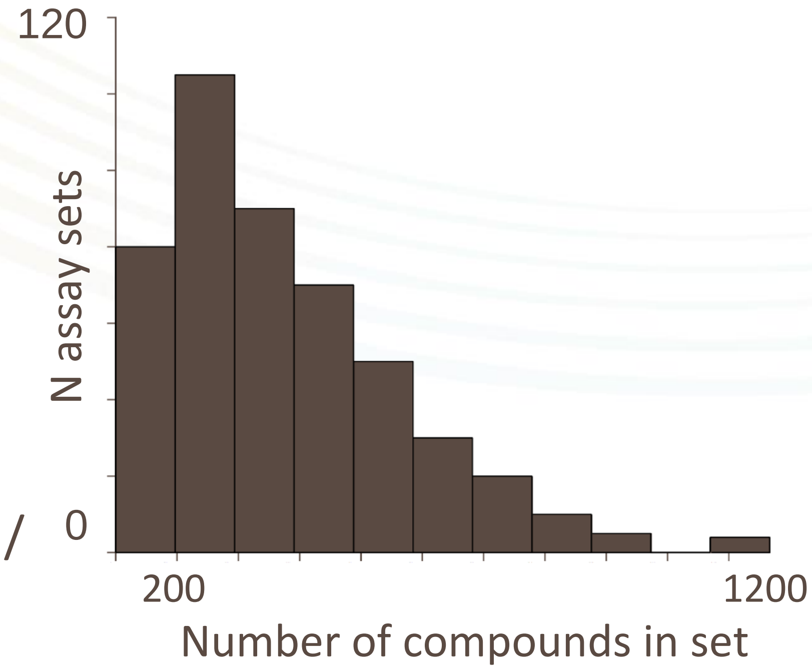
- Sets are split into training/validation sets using the RDKit MinMax picker and Tanimoto similarity of Morgan circular fingerprints.

What Information Is Obtained?

- *Relative appearance* of assays
  - **(number of good set appearances / total number of appearances)**
  - Look for features which could determine what makes an assay ‘good’

```
Assays:
CHEMBL3096731   0.083
CHEMBL2051179   0.057
CHEMBL1039568   0.043
Regression Descriptors:
logP            1.00
fr_NH1          1.00
fr_NH0          1.00
```

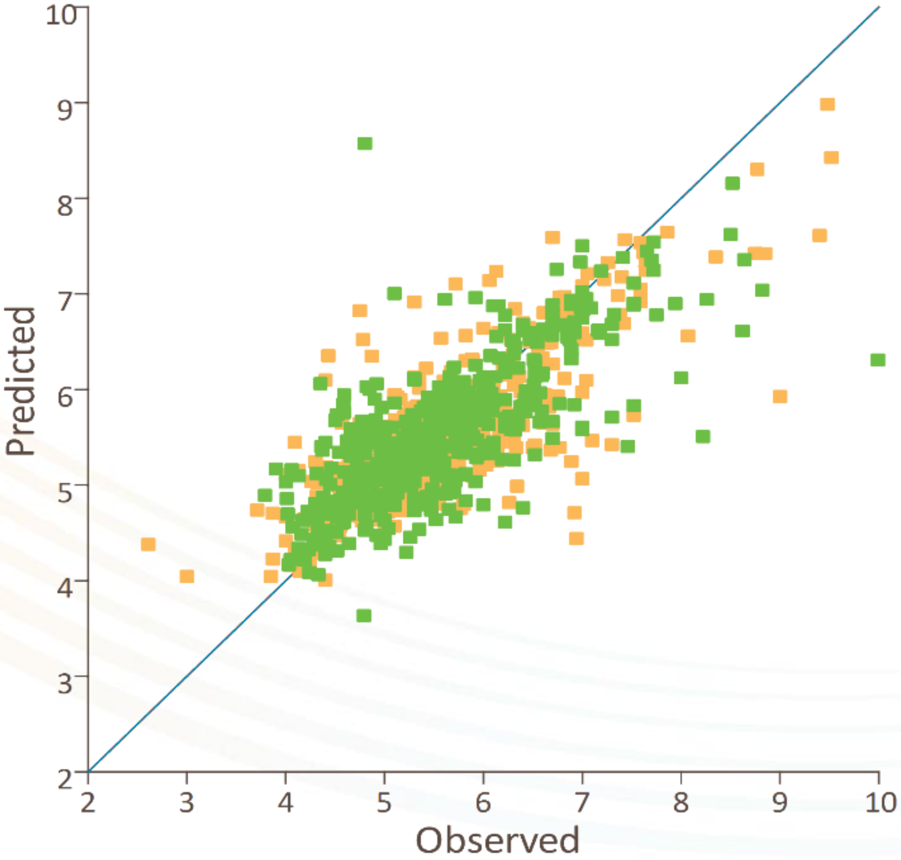
- The distribution of set sizes
  - can help to guide expectations about domain of applicability



- How often descriptors are selected in good sets
  - (number of good sets using descriptor / number of good sets)

Example – CYP3A4 (All IC_{50} Data)

- Building a model from all ChEMBL IC_{50} data leads to a poor model
  - Used the same data cleaning process as in QSARSetBuilder, 3921 compounds remaining
- Running 10,000 set combinations using QSB, we get 294 good sets
  - Take all sets with **relative appearance** >= 0.02 (48)
  - Early termination flag can end run if no good sets produced in last *N*

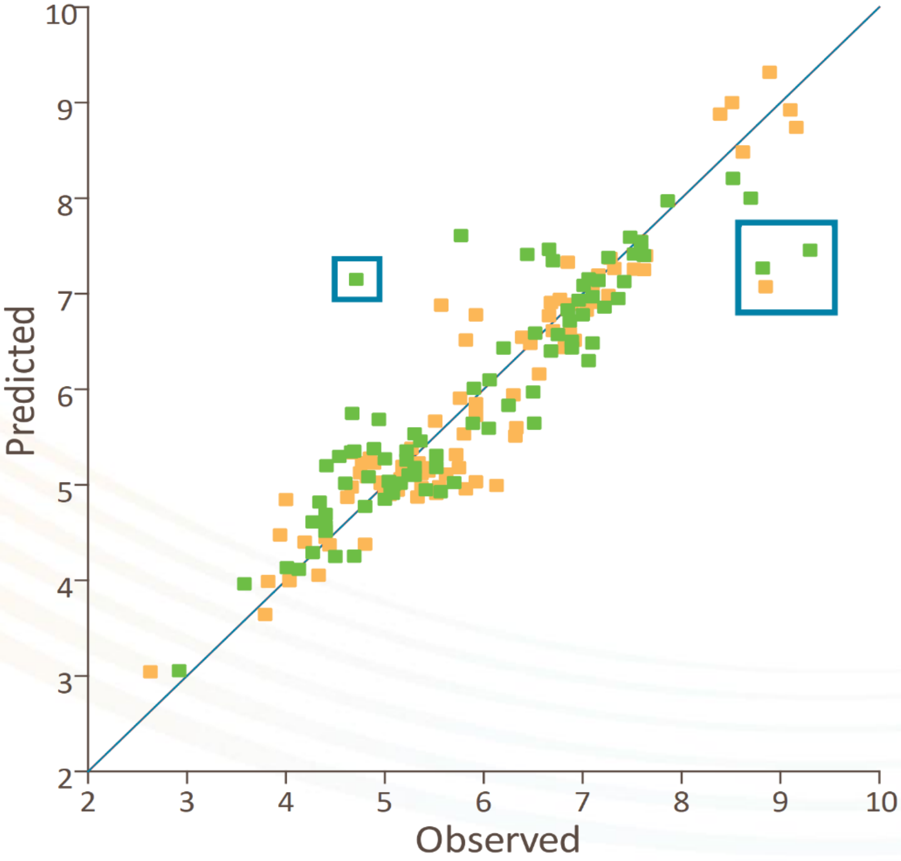


Validation (N = 493)		Test (N = 493)	
R^2	RMSE	R^2	RMSE
0.625	0.574	0.544	0.632

RBF Models generated using StarDrop Auto-Modeller™, figures generated in StarDrop™

Example – CYP3A4 (Initial Post CBsort)

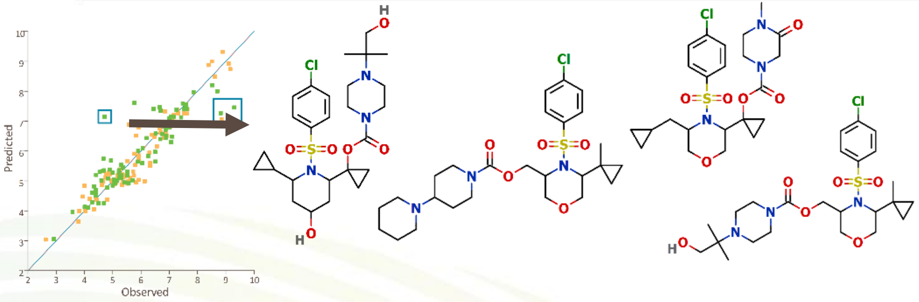
- Highlighted outliers are a series from a single assay
  - The paper has an activity cliff in SAR and doesn’t sample much of the space around it
- Should we ignore these outliers? Two ideas for improvement
  - Add another assay which samples more of this space
  - Alternatively, as we consider assays as ‘blocks’, we should remove the whole assay



Validation (N = 84)		Test (N = 84)	
R^2	RMSE	R^2	RMSE
0.882	0.458	0.804	0.578

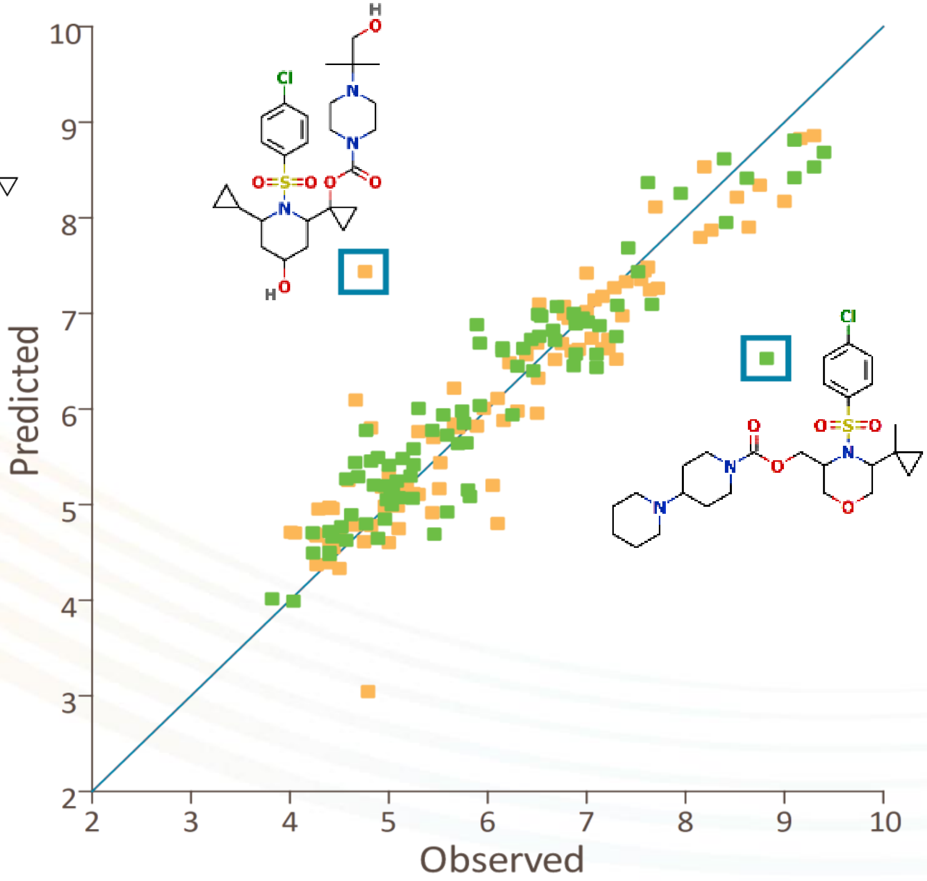
RBF Models generated using StarDrop Auto-Modeller™, figures generated in StarDrop

Example – CYP3A4 (Plus Additional Assay)



- Outliers appear in another, larger assay
  - More space sampled around problem molecules
  - Inclusion could improve model

- When this assay is included and a new model generated, some outliers still poor



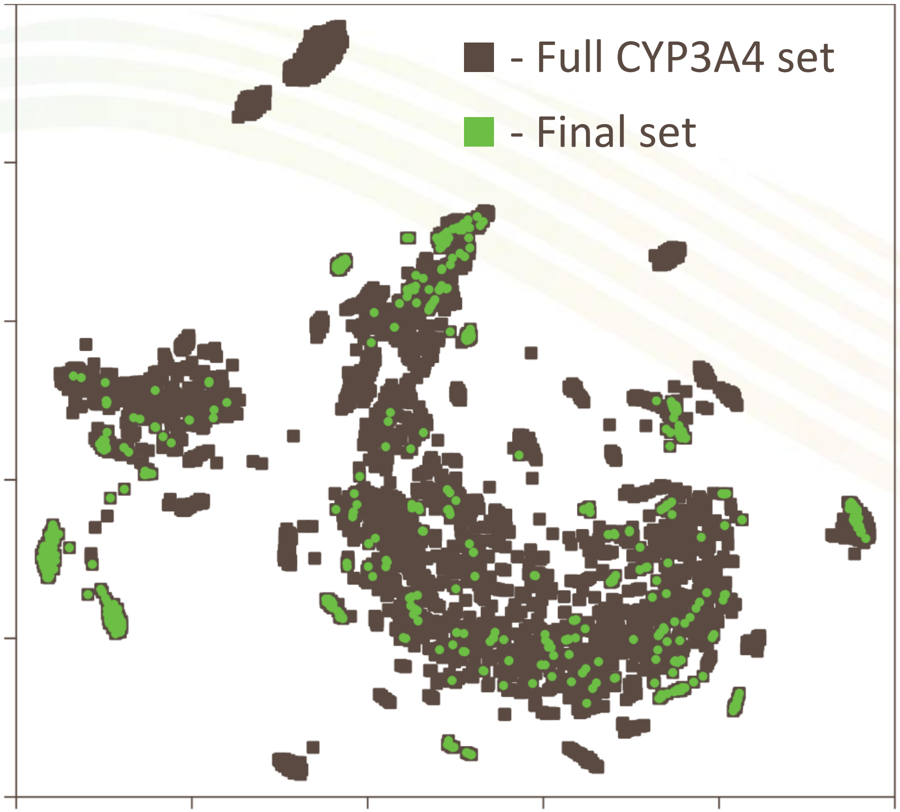
Validation (N = 87)		Test (N = 87)	
R^2	RMSE	R^2	RMSE
0.842	0.548	0.872	0.481

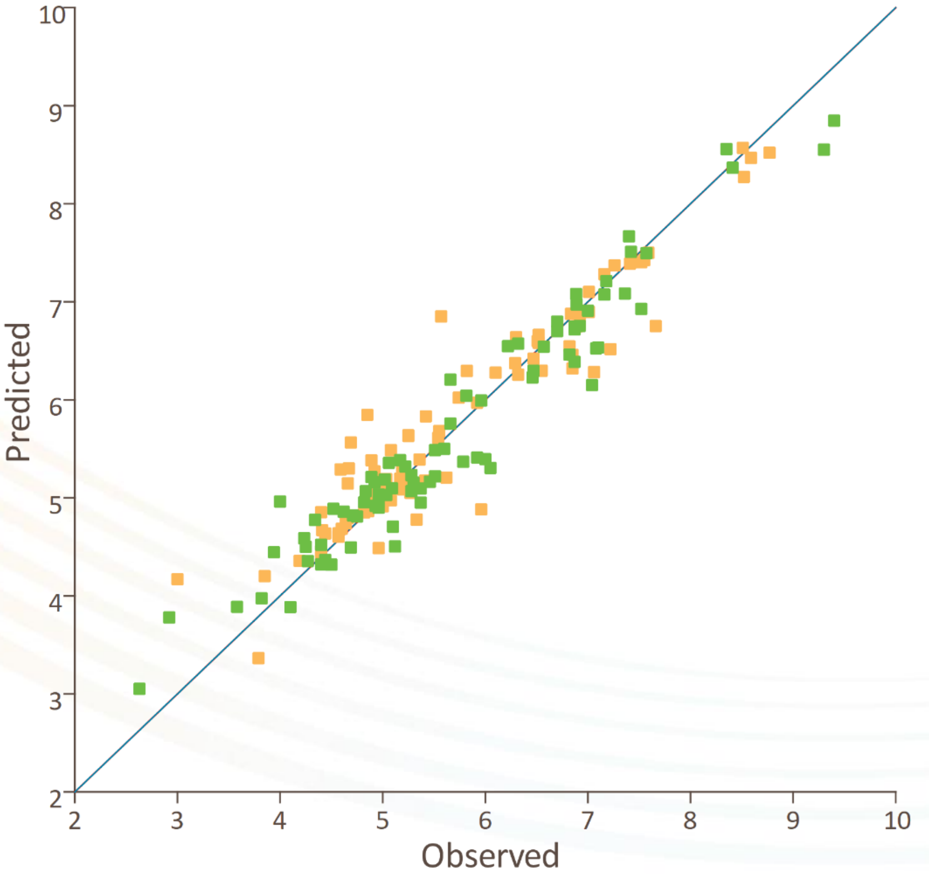
RBF Models generated using StarDrop Auto-Modeller™, figures generated in StarDrop™

Example – CYP3A4 (Final with Removed Assay)

- Removal leads to better stats

CYP3A4 Chemical Space t-SNE





Validation (N = 82)		Test (N = 82)	
R^2	RMSE	R^2	RMSE
0.894	0.397	0.933	0.344

- Data primarily uses midazolam as probe

RBF Models generated using StarDrop Auto-Modeller™, figures generated in StarDrop™

Implementation and Availability

- Implemented in Python 3 and tested with version 3.6
- Cross platform – tested on macOS®, Windows® and Linux® operating systems.
- The code is freely available (GNU GPL) and can be downloaded from the Optibrium website
- Makes use of multiprocessing to run on more than one core
  - Can set process priority to avoid system slowdown
- Isomeric -> canonical smiles changes, descriptors and fingerprints are cached

macOS® is a trademark of Apple Inc., registered in the U.S. and other countries.
Windows® is a trademark of Microsoft Inc., registered in the U.S. and other countries.
Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.

Thoughts for the Future

- Is there a better method to use for building sets than randomly combining assays?
  - Initial building is done using overlap - we don't totally discard that information

- Can we use information about assays which commonly appear in a good set together?

- Natural language processing
  - Analyse potential probe substrates in assays using **chemlistem**
  - Can text analysis be expanded to consider whole articles?

Acknowledgements

- With thanks to:

Academic supervisor: Dr. David Palmer
Industrial supervisor: Dr. Peter Hunt
The team at Optibrium

University of Strathclyde

ROYAL SOCIETY OF CHEMISTRY

optibrium

- Software and libraries used: **MolVS**, **chemlistem, NumPy,**

scikit learn

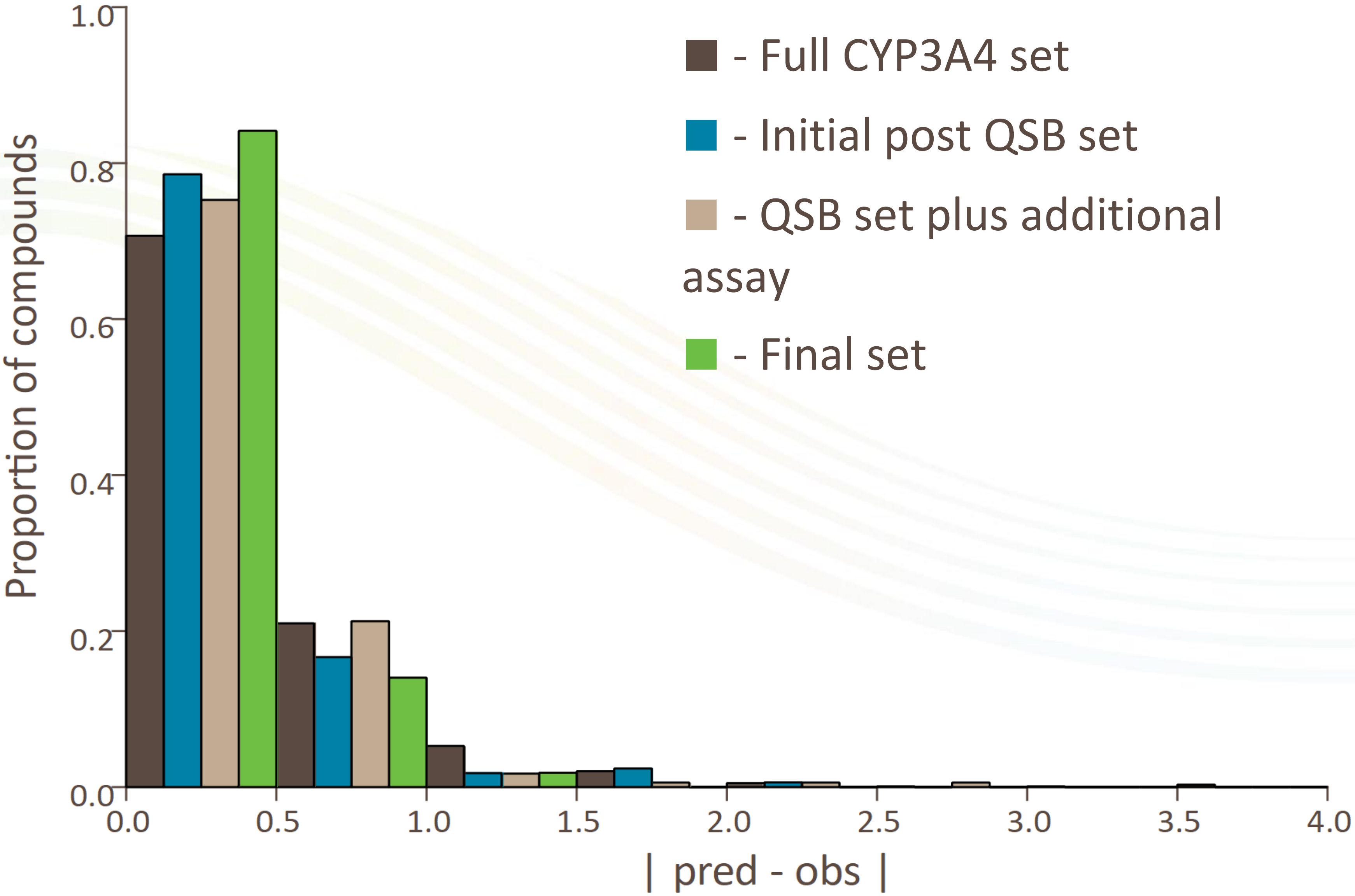
RDKit

python

matplotlib

StarDrop™

Comparison of Error Distributions



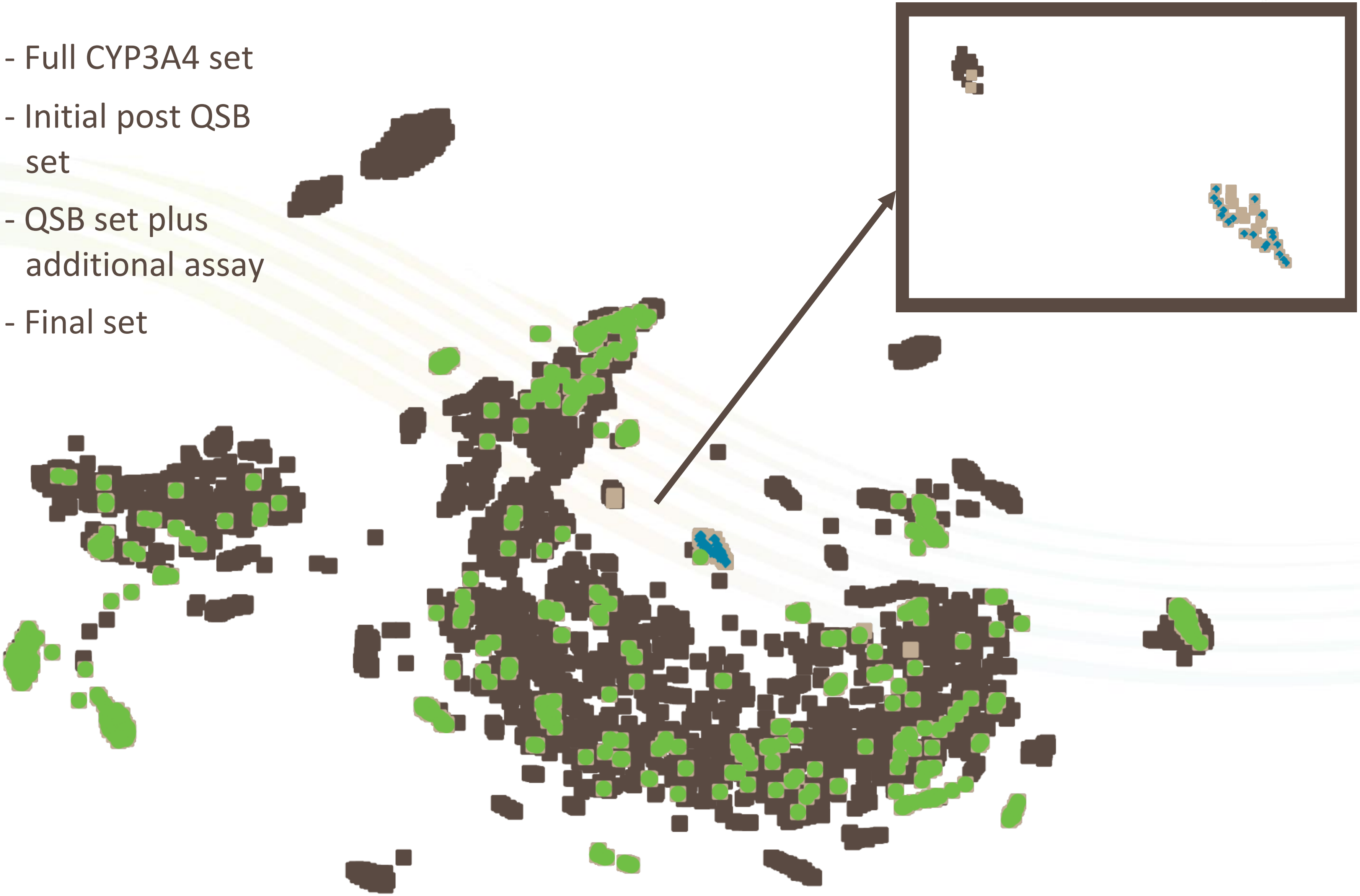
Chemical Space of Plus/Initial vs Final

🟫 - Full CYP3A4 set

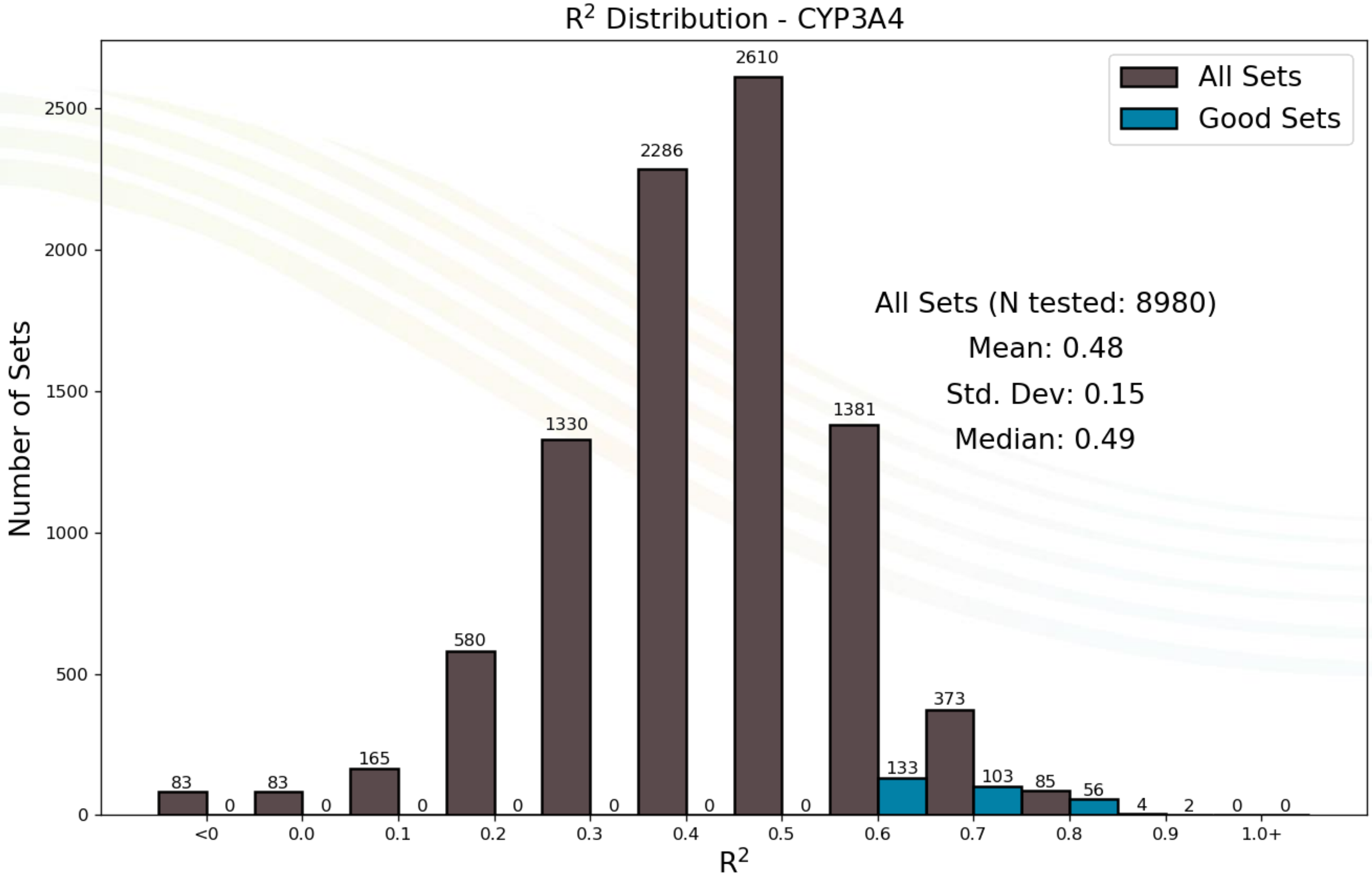
🟦 - Initial post QSB set

■ - QSB set plus additional assay

🟩 - Final set



Distribution of Rsq Values


R^2 Distribution - CYP3A4
Number of Sets
R^2
<0
All Sets
Good Sets
All Sets (N tested: 8980)
Mean: 0.48
Std. Dev: 0.15
Median: 0.49