Clustering large datasets in KNIME for model building and hitlist triaging

Valery Polyakov*, Eric Martin, and Li Tian

Senior Investigator I

Novartis Institute for Biomedical Research, Emeryville, CA

ICM UGM 2016

Friday, March, 18, 2016 from 8:55 AM – 9:20 AM

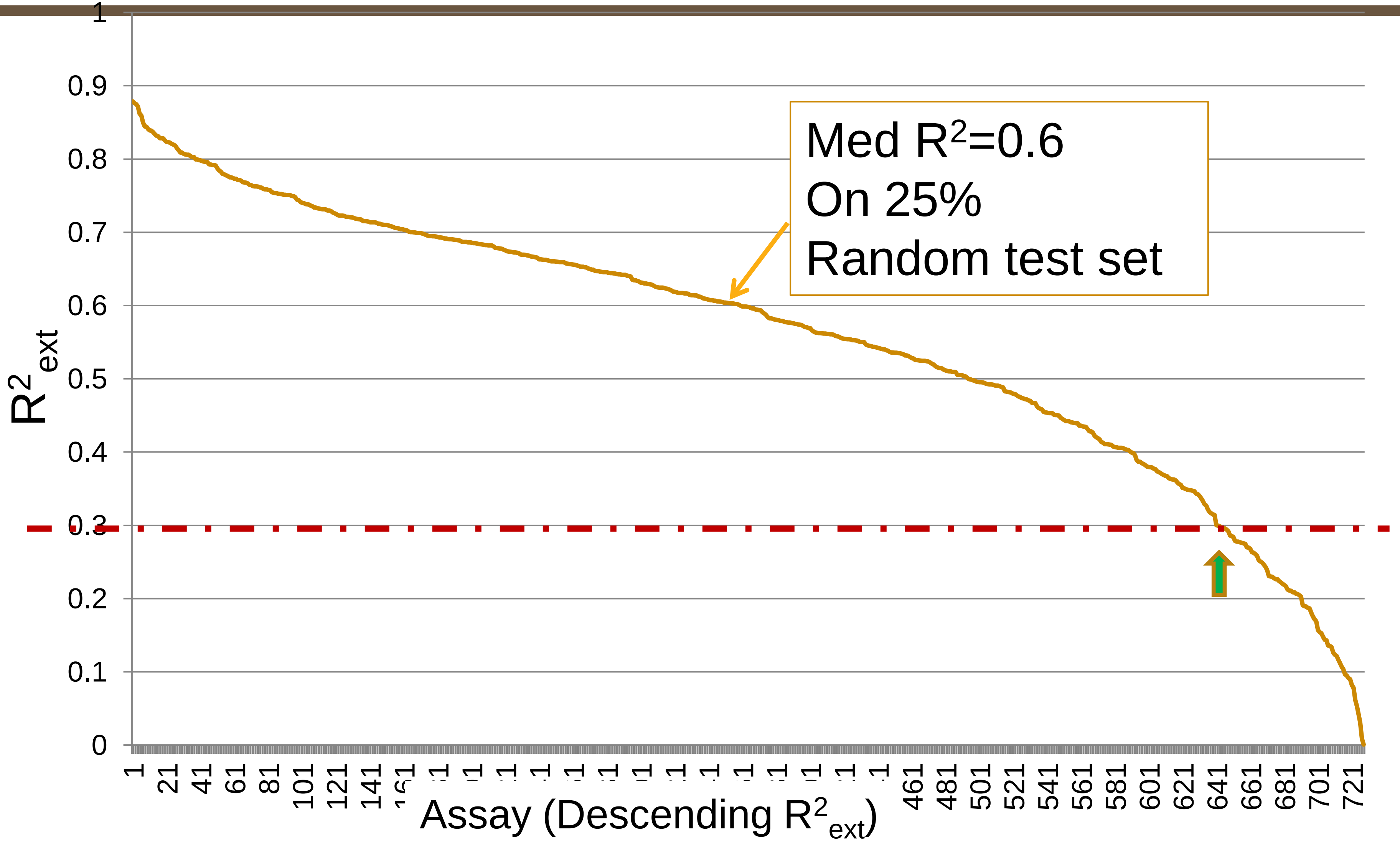
NOVARTIS

Outline

- Using clustering in model building (training v. test sets)
- Clustering for hit list triaging

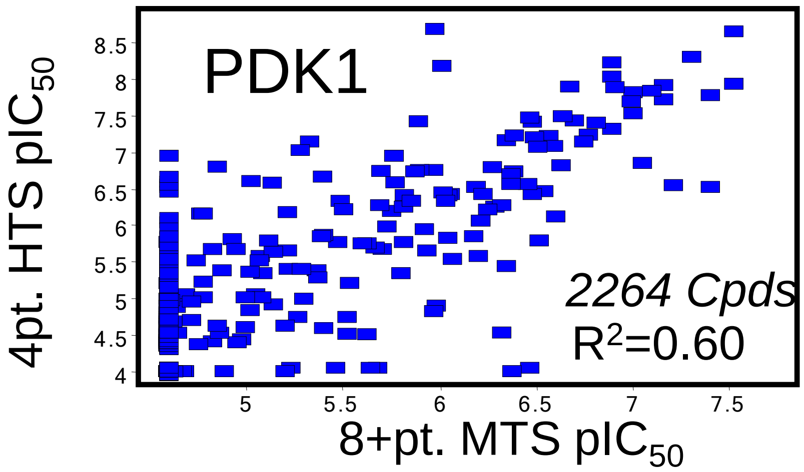
“Small-data” Random Forest Kinase Models for Virtual Screening

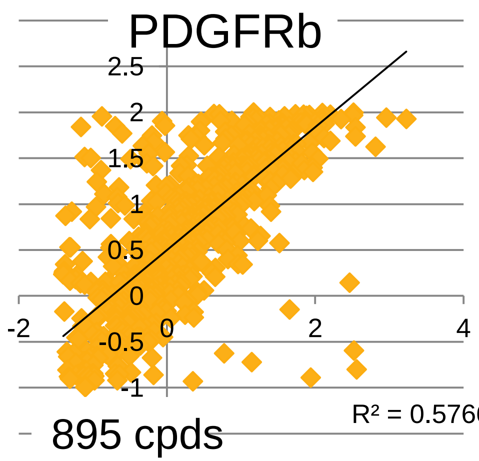
25% RANDOM held-out test set looks good for 728 kinase assays.

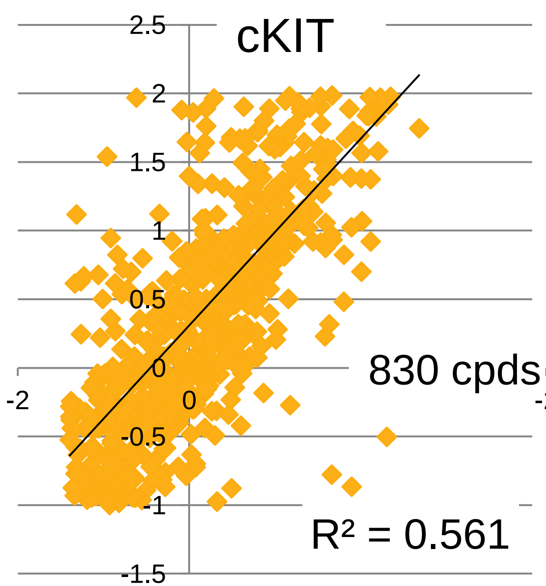


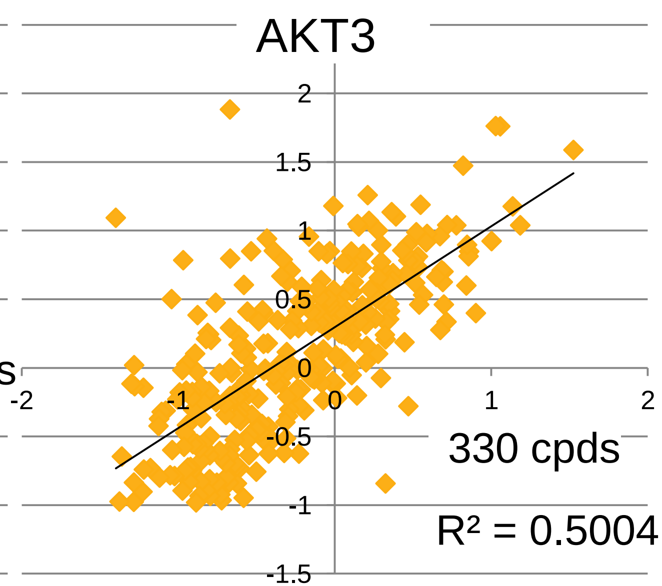
NOVARTIS

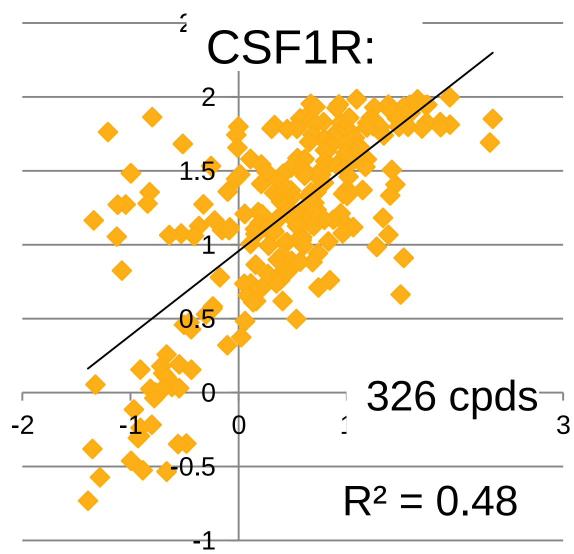
4 Pt. MTS IC50 vs 8-12 pt. project IC50: **average R^2=0.54**







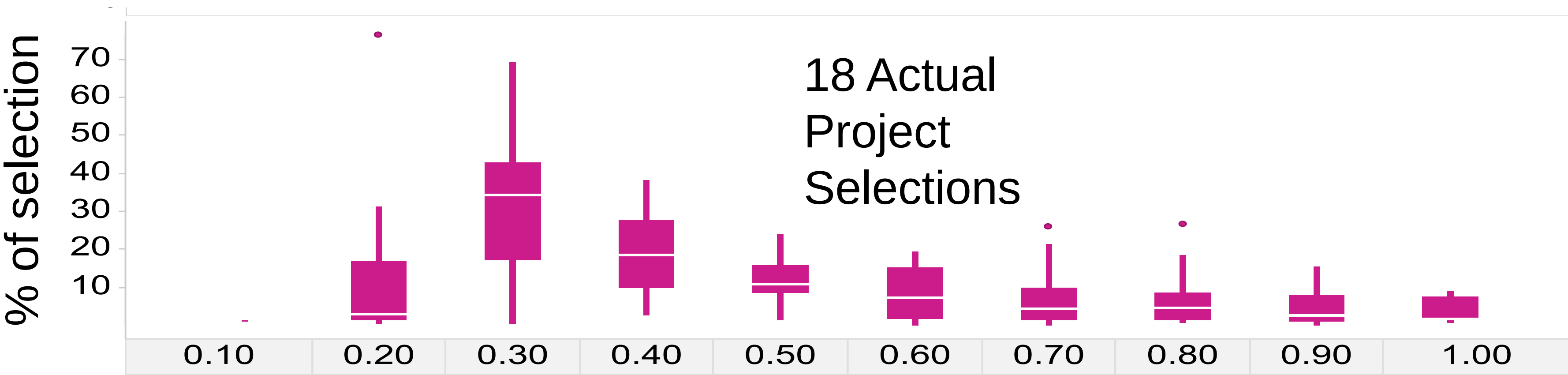




NOVARTIS

Compounds Teams Actually Order are Unlike Training Sets

Box plot summary of test/train similarity histograms from 16 virtual screens

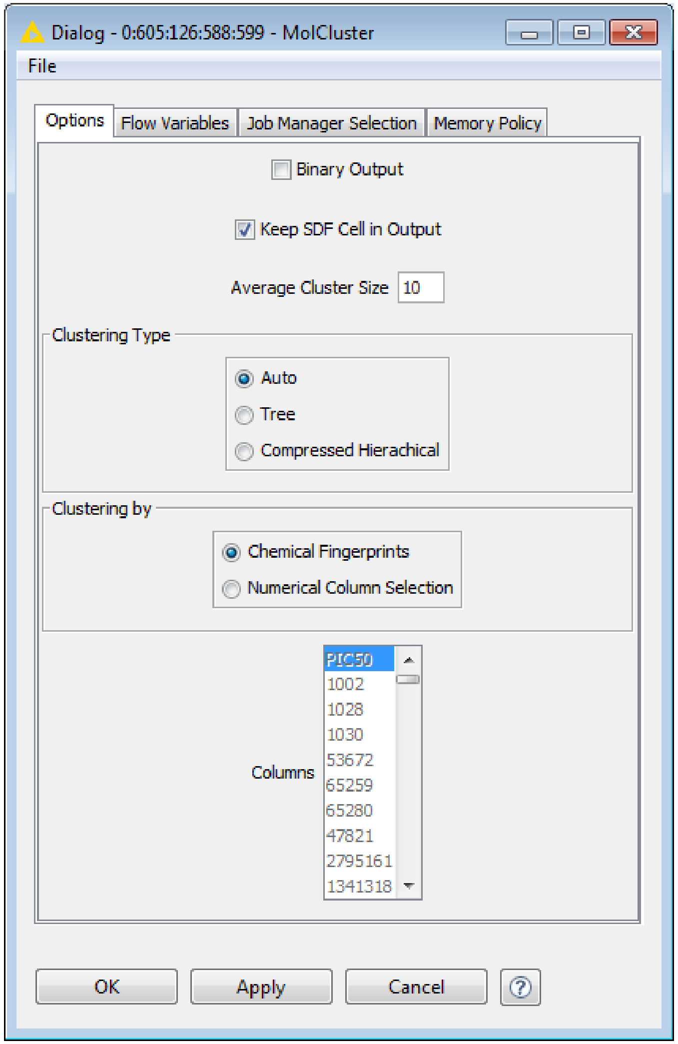


Similarity of newly selected compounds to the closest "old" compound

Can we develop a realistic test set that resembles experimental distribution?

MolCluster node was the fastest available in KNIME

Ideally suited for the large datasets



- We checked the following nodes for clustering:
  - k-Means
  - Cluster MCS
  - MolCluster
  - ...
- MolCluster scaled almost linearly with the number of molecules
- 64-bit version is available > 100 K datasets

Mimicking compound selection

Large clusters were tested before during the previous attempts

Cluster
- Compound clustered
- Using MolCluster

Sort
- Sorted
- Largest clusters first

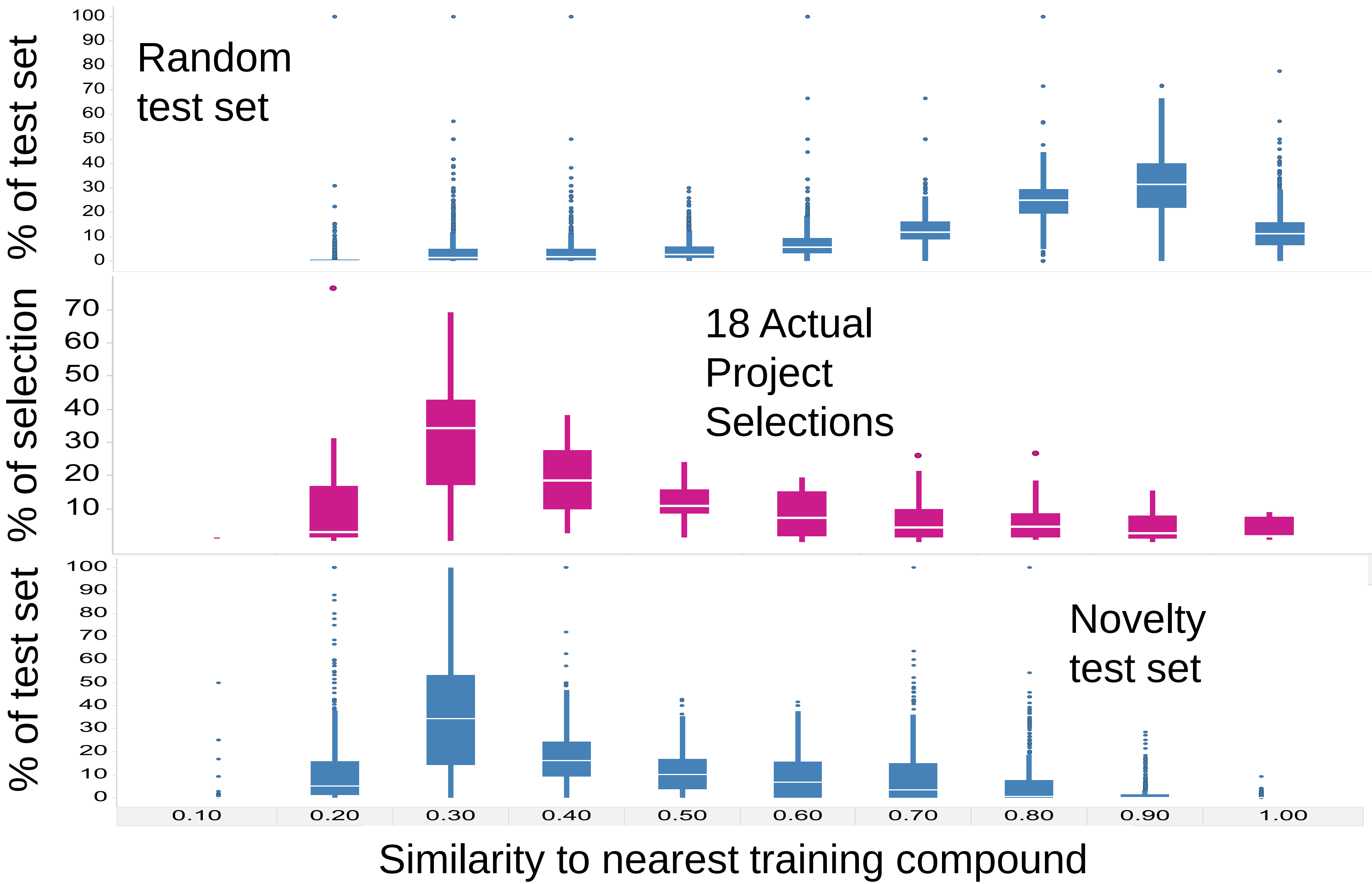
Select
- ~ 75% molecules -> training from largest clusters
- ~ 25% -> test set from small clusters and singletons

- If some data for the projects were known before, it could be from the previous optimization runs
- Limited number of large series was extensively tested before
- Other non-optimized chemical matter in the database is of interest
- Is likely to be in small clusters an singletons

NOVARTIS

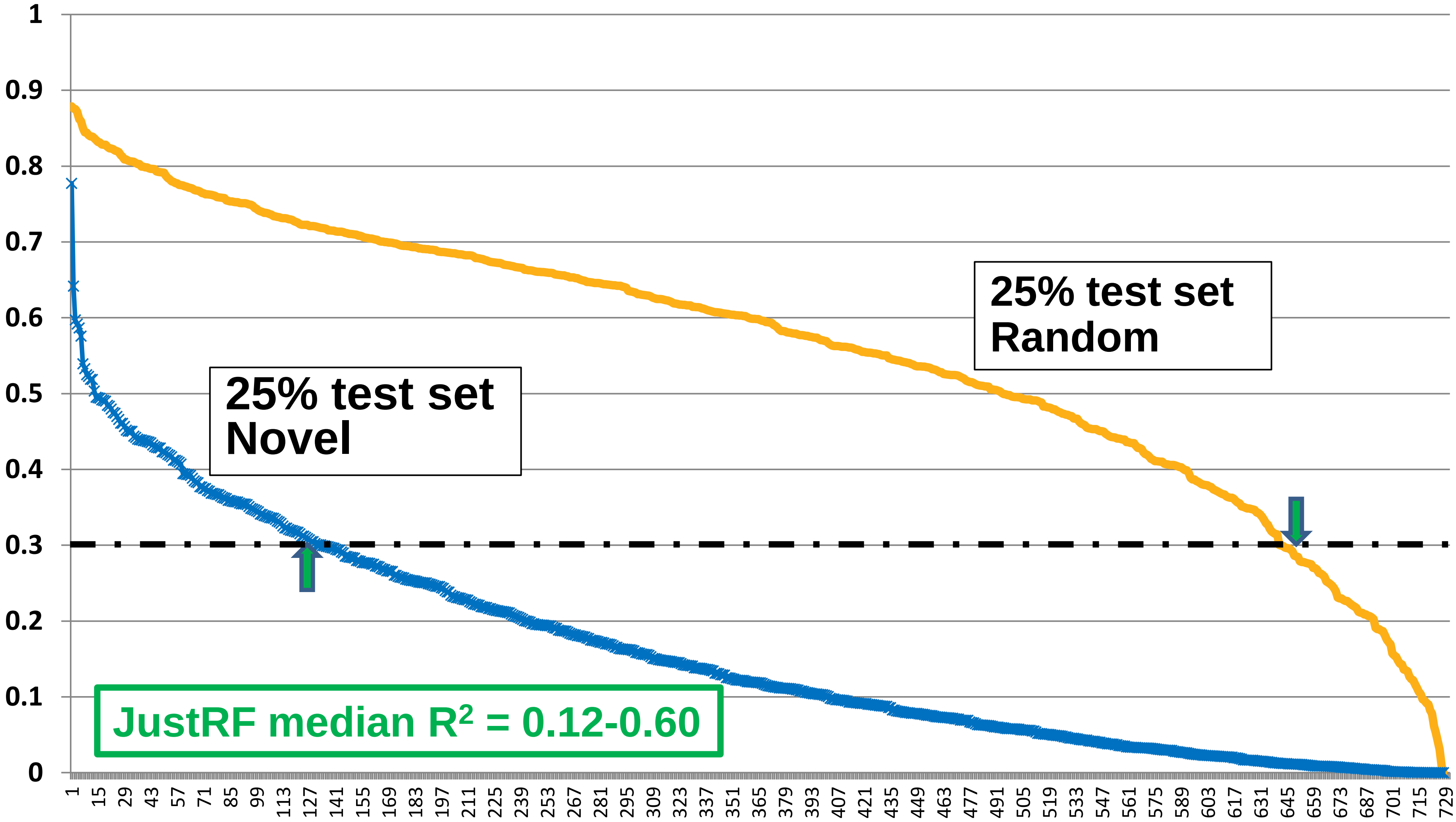
Realistic Domain of Application: Test Set Unlike Training Set

“Outlier” test set distribution mirrors actual selections by project teams



80% of RF Models Fail on Realistic Test Sets

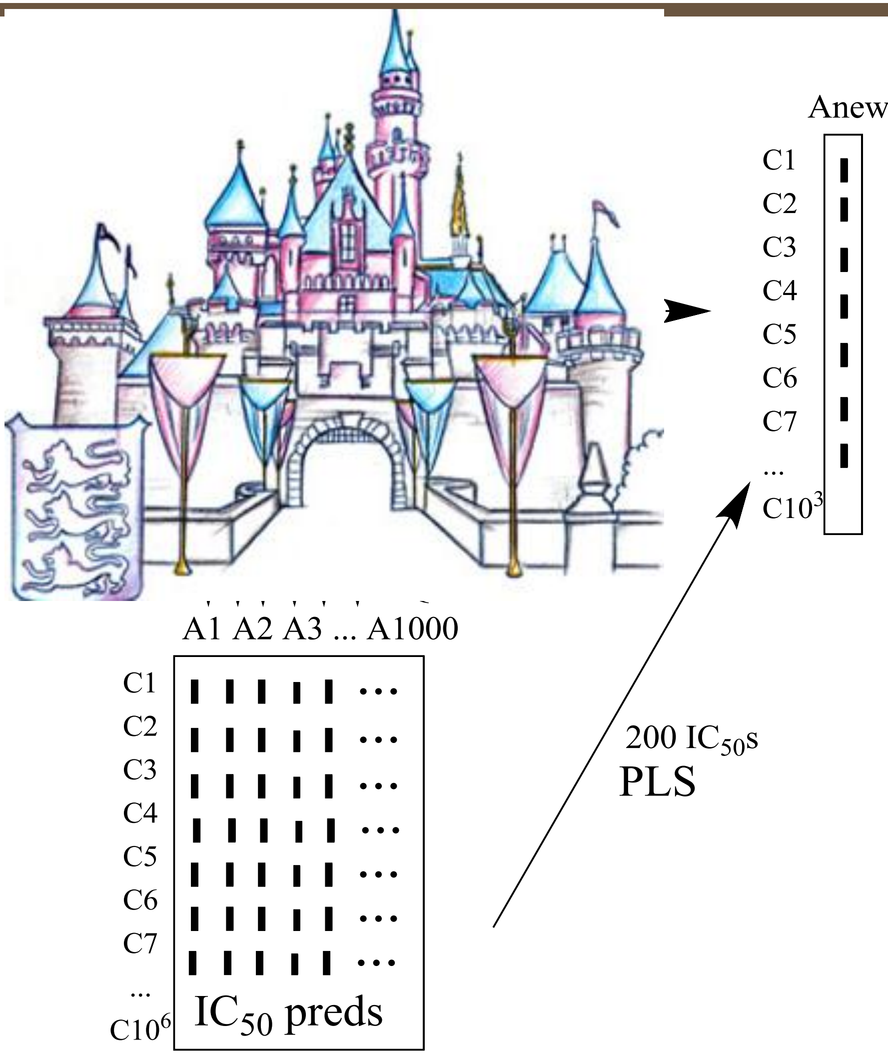
Train on large clusters, predict on outliers



NOVARTIS

pQSAR is a two step process

Correcting for the non-ideal world: Filling matrix with predicted values



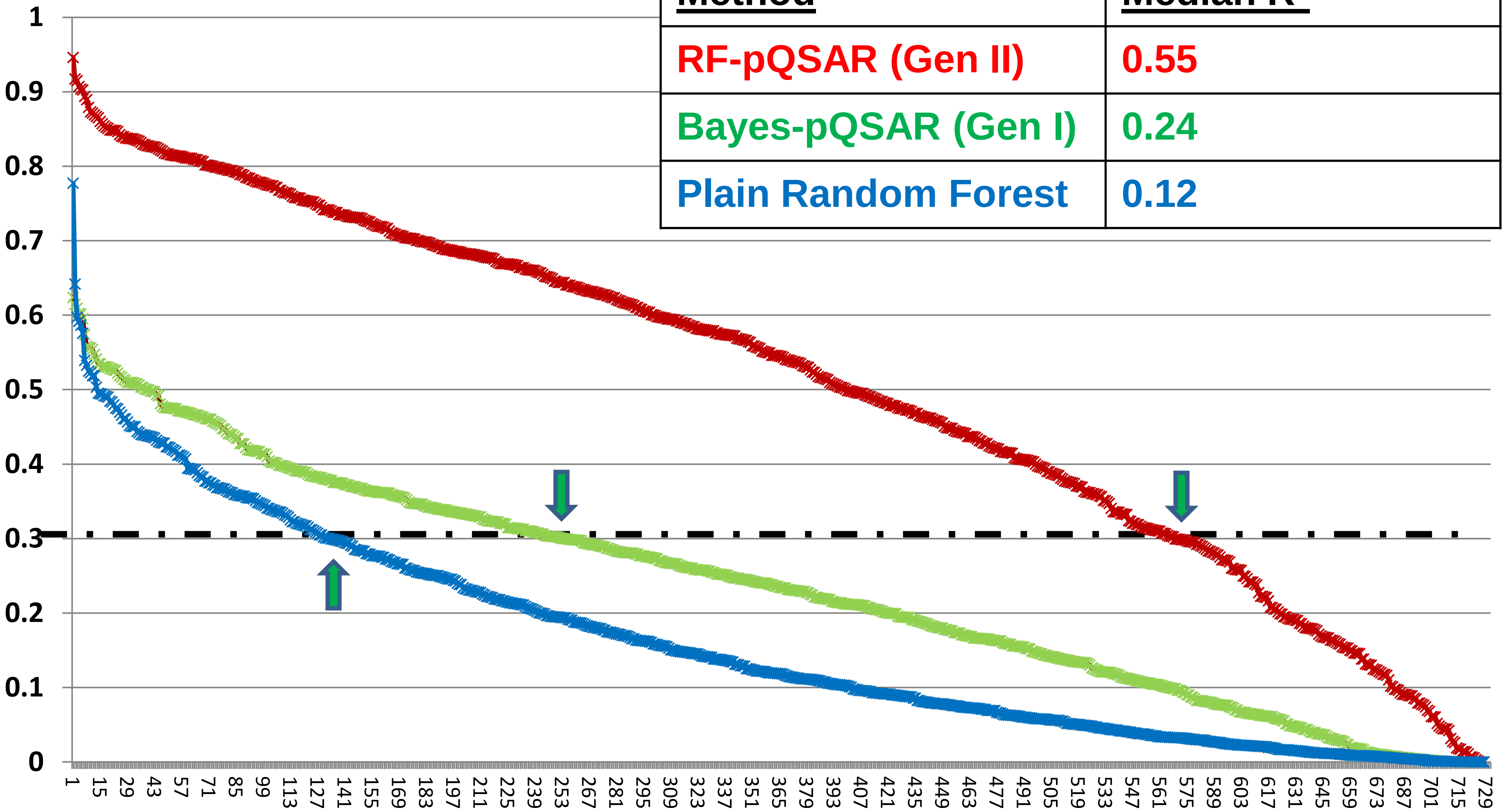
- Before we use PLS, we need to fill in the matrix of missing IC_{50}s for all assay in the superfamily
- We created a set of RF models for each assay in the set
- Models are superfamily specific – Kinases, GPCRs, proteases, etc.

Naive Bayes -> Random Forest of Profile-QSAR

Now ~ 4-pt IC50 for realistic novelty test set.

25% Novelty test set

Method	Median R^2
RF-pQSAR (Gen II)	0.55
Bayes-pQSAR (Gen I)	0.24
Plain Random Forest	0.12



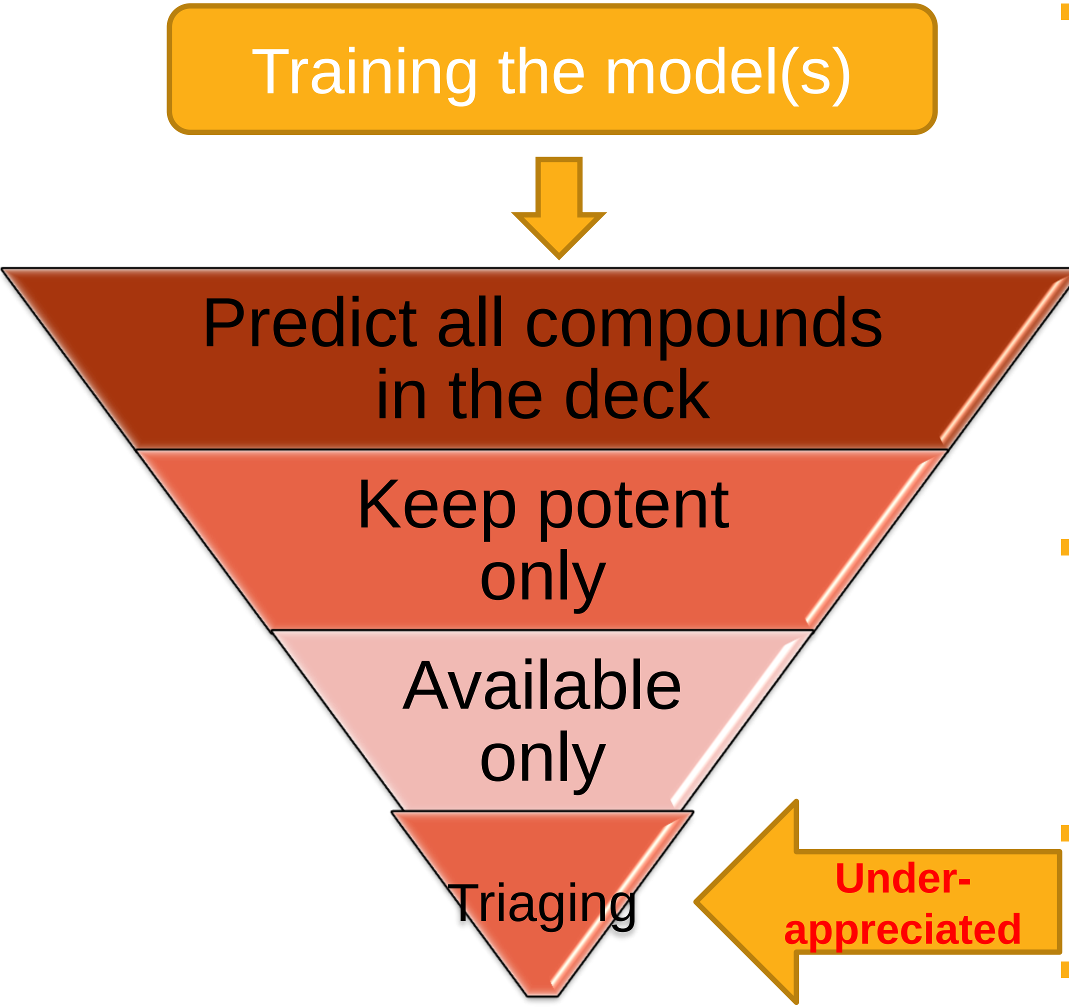
NOVARTIS

Outline

- Using clustering in model building (training v. test sets)
- Clustering for hit list triaging

Using the models for lead generation

Unselected compounds may be lost forever



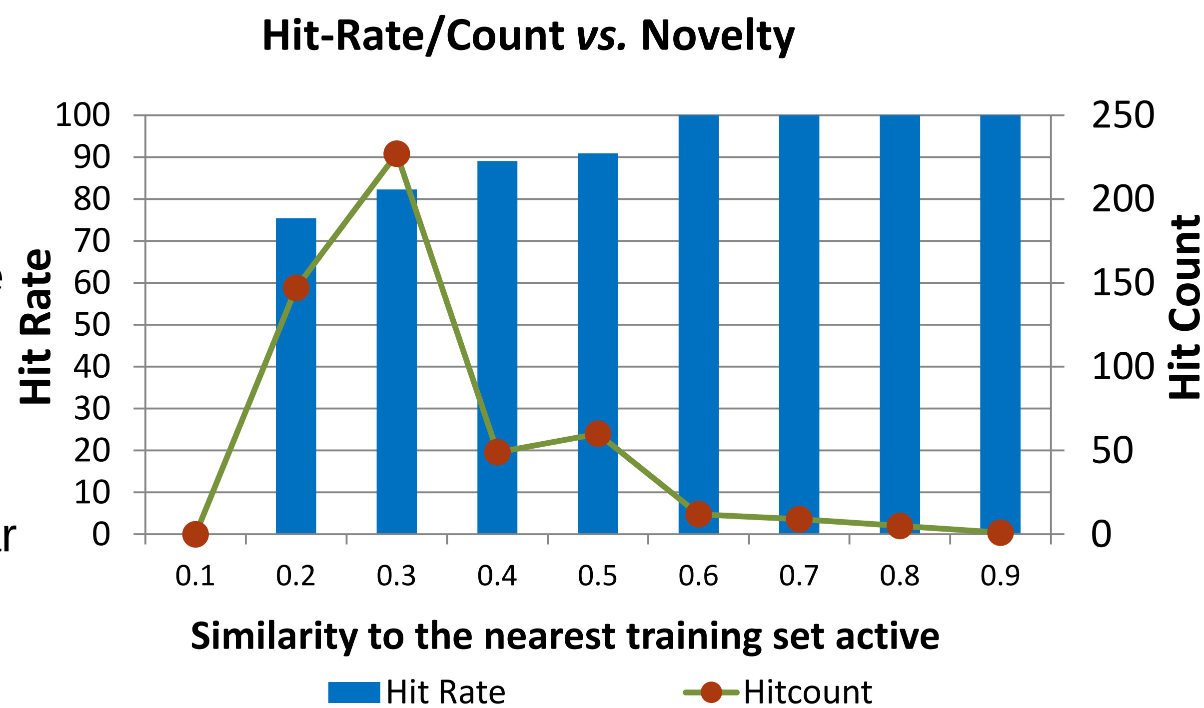
- Triaging process involves a dialog with the project team to understand their objectives:
  - Diversity
  - Novelty
  - Potency
  - Selectivity
  - Ligand efficiency
  - Phys-chem properties
- Sophisticated clustering:
  - Selects representatives by properties: LE, LiPE, Potency, or Selectivity
  - Cluster and sample bins at different chemical densities
- Selecting top compounds from each cluster by desired properties
- Combine the selections to ensure the best possible diversity

Triaging could be simple

If needed

- **300k** — Predicted kinofiles
- **34k** — pIC50 > 5
- **33k** — Not similar
- **32k** — Both AlogP < 6 and MW < 600
- **30k** — Less than 2 lipinski violations and no I and P atoms
- **5k** — At least 10 fold selective (predicted by our models)
- **2,5k** — Available from HTCP Basel A

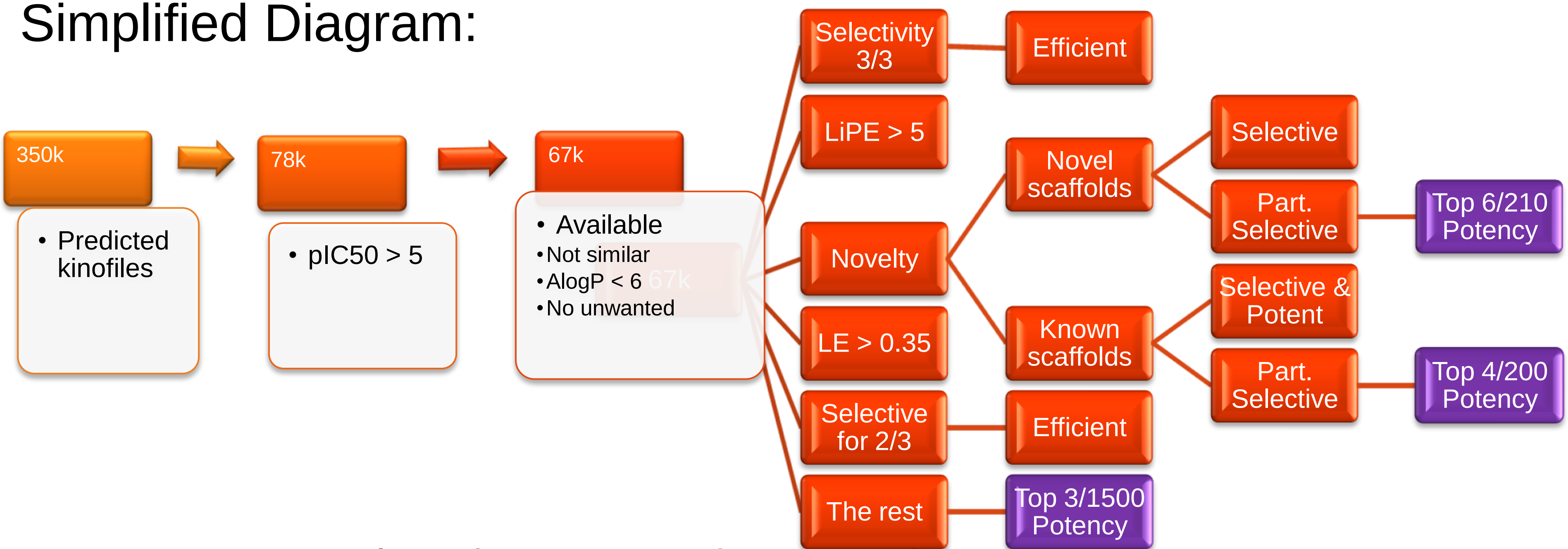
- The process of triaging depends on the needs of the project and the number of compounds they can test
- Different criteria could be used for: tool cpd, medchem lead, catchup screen, archive enhancement, priming the pump (initial training set for new target), etc.
- The method also finds compounds that are dissimilar to the training set!



Clustering is used for variable density sampling

A kinase project – combining chemical and property space diversities

Simplified Diagram:



- Important to sample various groups of compounds
- Some groups were large, while some were small
- Large groups were clustered to ensure maximal chemical diversity
- Clusters were sampled with different densities and parameters
- 15k selected for testing

Conclusions

- Ability to cluster large datasets is vital for
  - Model building (getting realistic quality estimates)
  - Triaging -> trimming large sets & retaining diversity in compound & property distributions
- Not too many tools are available for doing so...

NOVARTIS

Acknowledgements

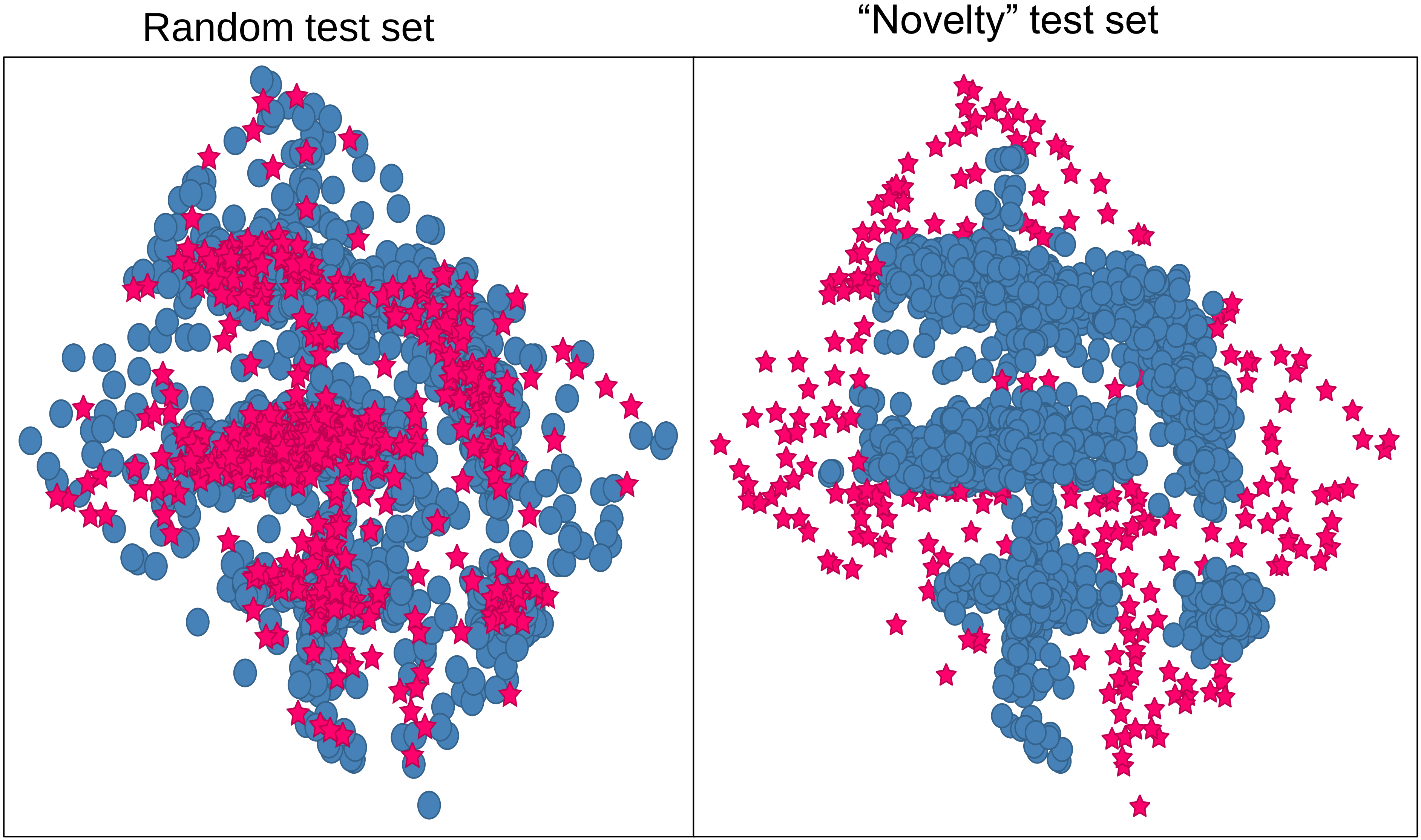
- Hanneke Jansen
- Ansgar Schuffenhauer
- Brian Kelley
- Eric Vangrevelinghe

Backup slides

- Questions?

“Novelty” Test Set to Mimic Real Team Selections:

Train on 75% from big clusters, test 25% singletons and small clusters

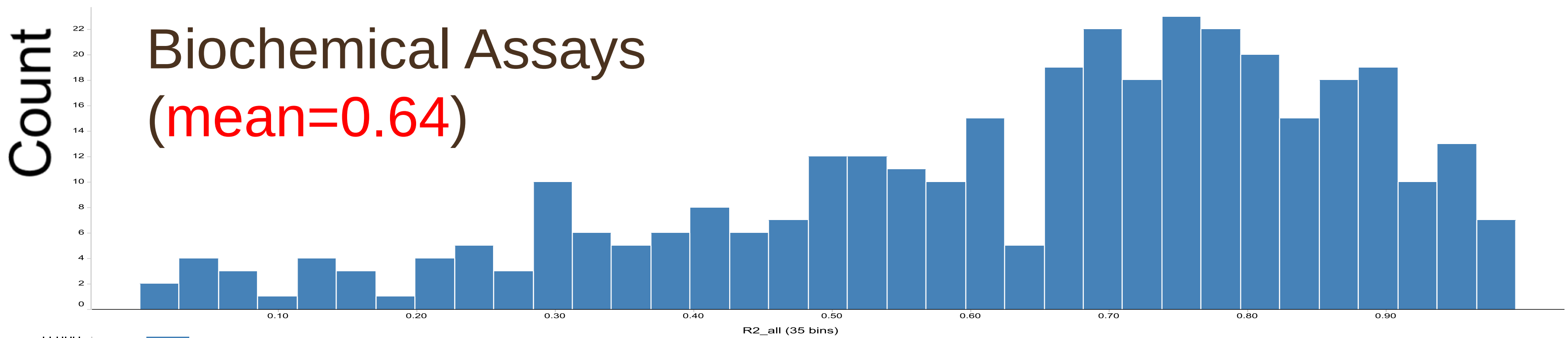


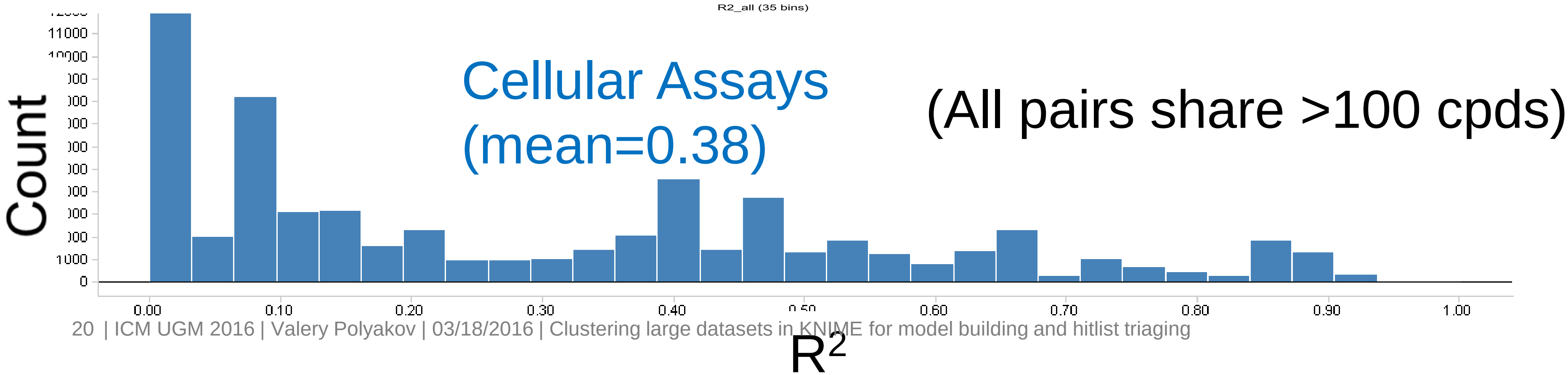
Kinase pQSAR median R^2=0.55-0.73 *vs* within Assay Variation

Correlations (R^2) between Assays for the Same Kinase

Assay Factors Varied by Biochemist:

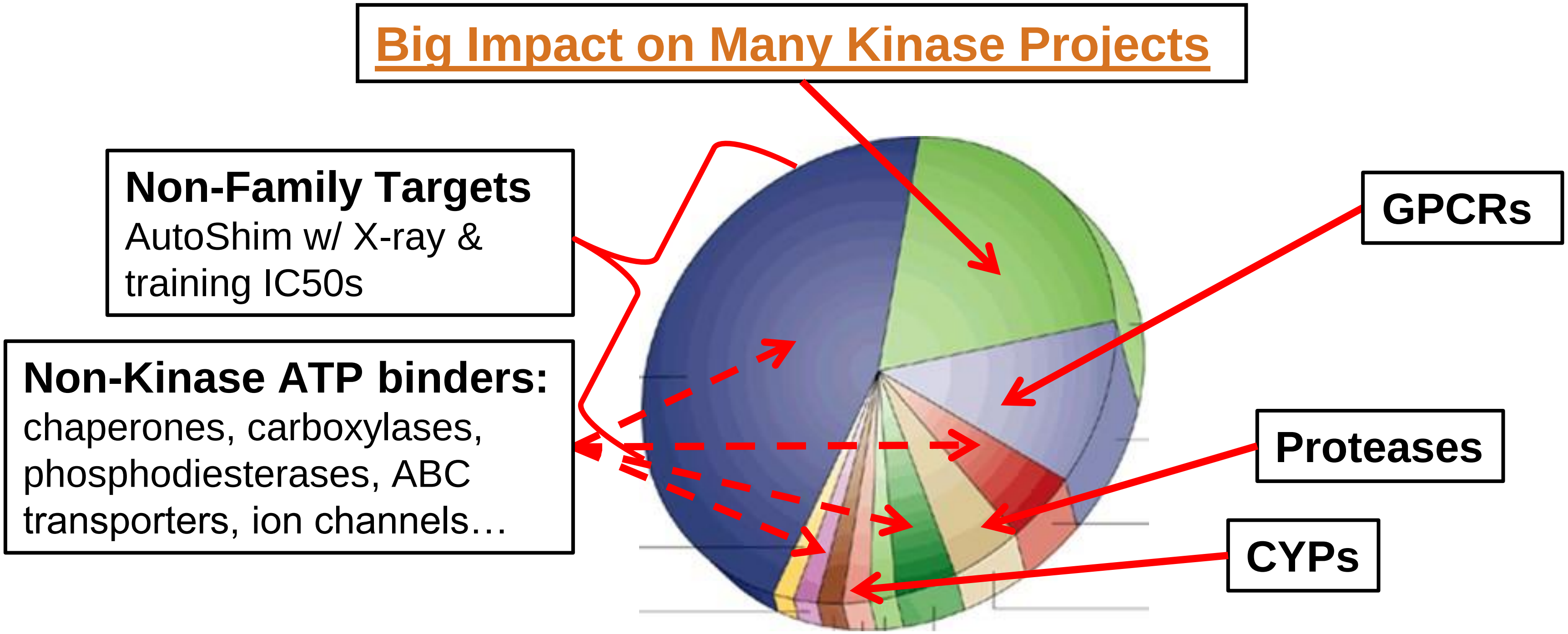
1) protein construct,
2) protein concentration,
3) post-translational mod's (phos or gluc),
4) temperature,
5) pH,
6) buffer salts, concentrations, detergents,
7) cofactor concentrations (ATP),
8) substrate,
9) incubation time,
10) order of adding reagents,
11) assay format,
12) label type and location





Extend profile-QSAR to More Families

GPCR, proteases, CYP, non-kinase adenosine binding proteins

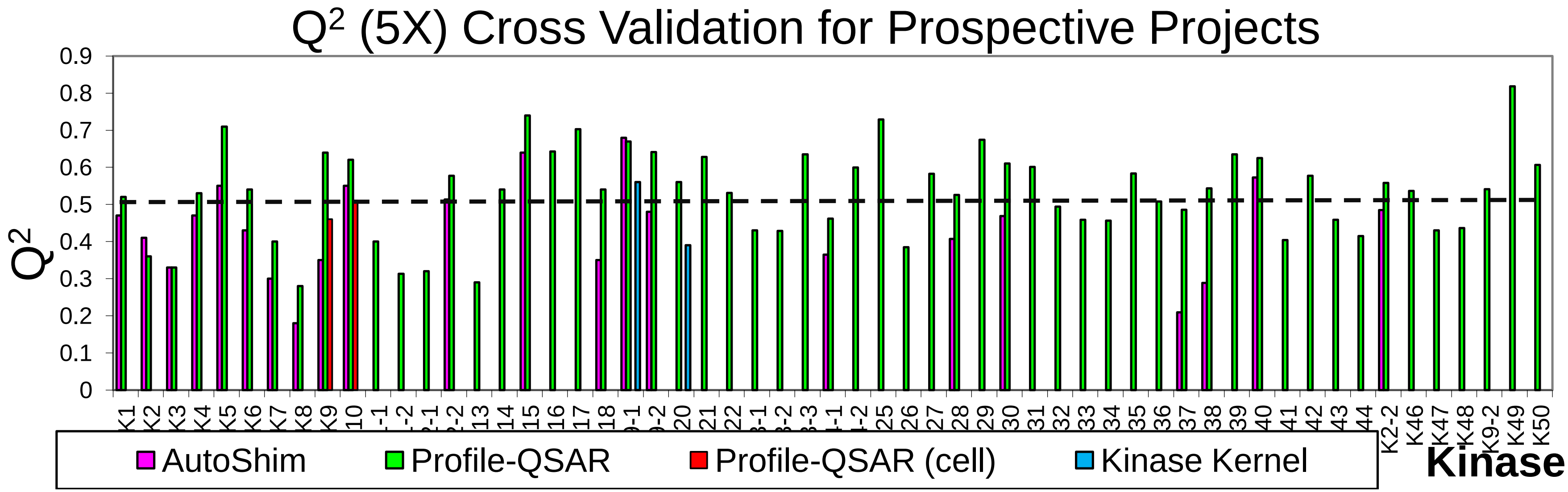


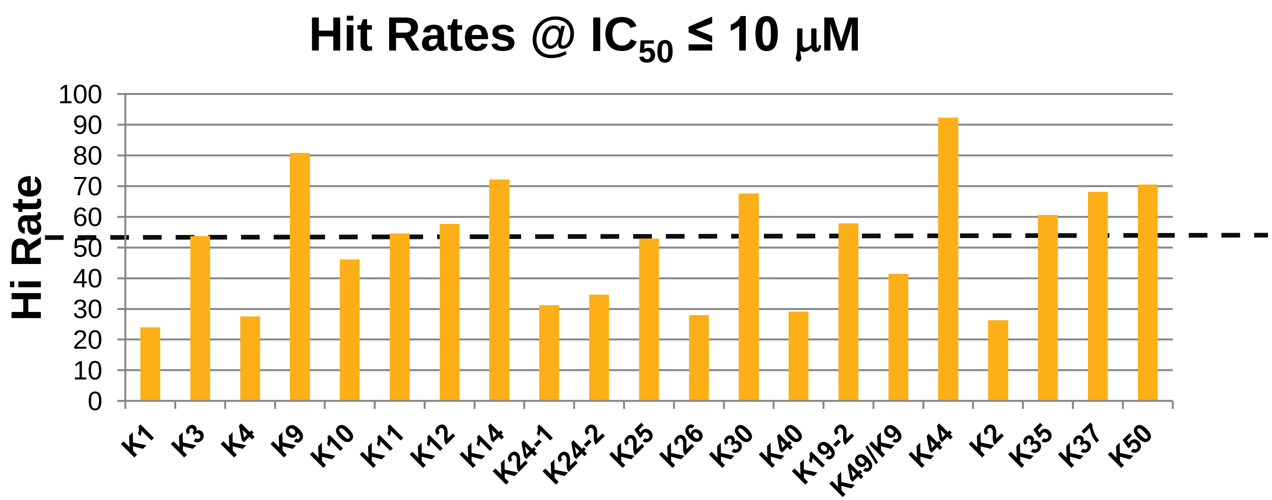
- > 2,000 assays: form virtual panels for safety and selectivity profiling
- Pre-calculated predictions of 3 million cpds stored for MoA analysis

Druggable genome from Hopkins et al., *Nature Reviews Drug Discovery* 1, 727-730 (2002)

Experience with (Gen1) Profile-QSAR

Median Q^2 > 0.5 for 50+ "Live" Kinase Projects





Hit Rates:

- Typical eHTS ~1%
- Median PFVS>50%

* Special application to find type II inhibitors

NOVARTIS

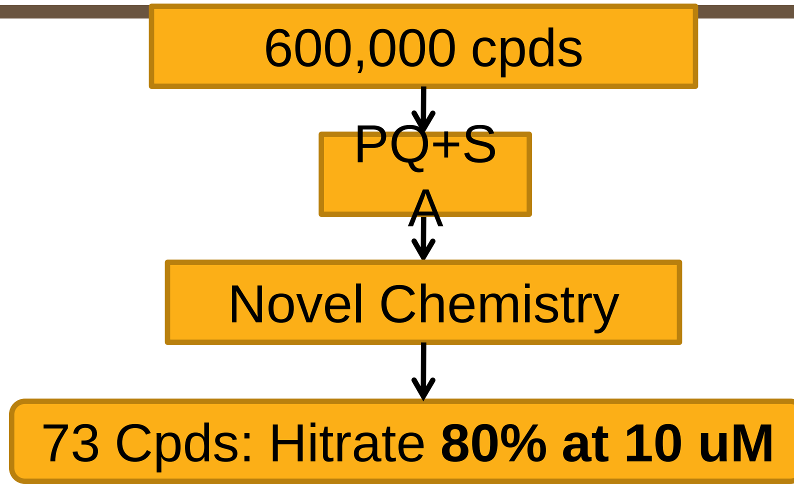
Similarity Distribution from Real Projects Mimic Novelty Test Set

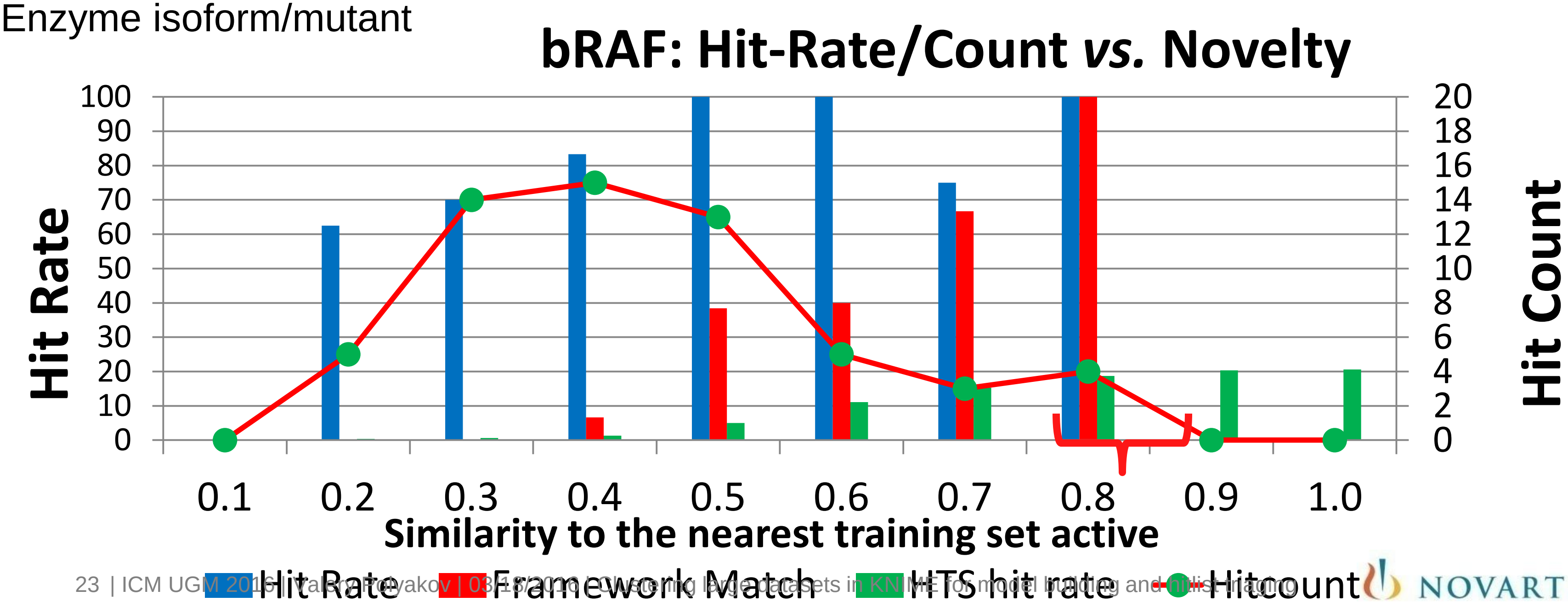
bRaf: 60% to 80% hit rate for completely novel compounds

Model Summary

Assay	R^2_{ext}	N
Biochem*	0.56	769
Biochem*	0.7	34839
Biochem*	0.8	8745
TM	0.7	3543
CP	0.51	2211

* Enzyme isoform/mutant

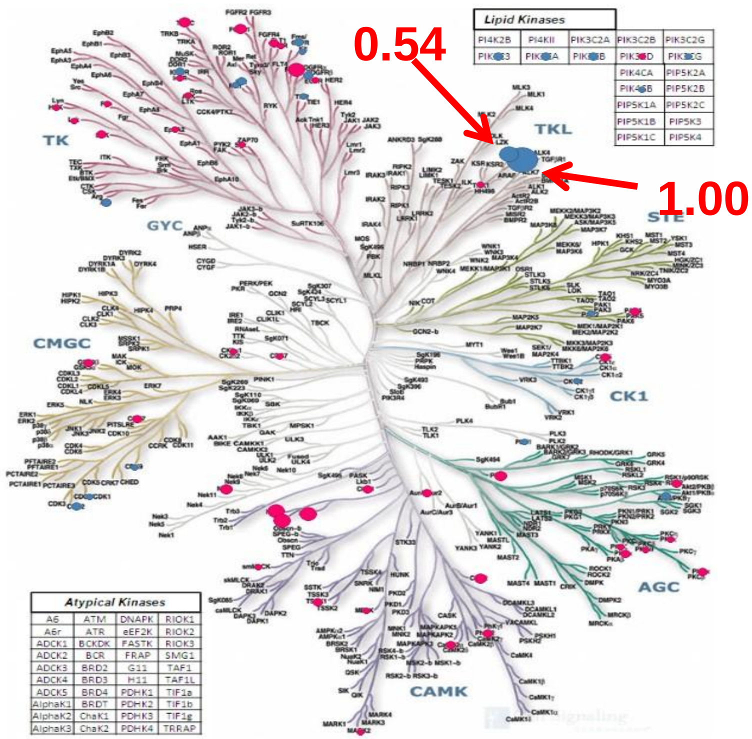




Components of “Hybrid” Model Carry Real Biochemical Information

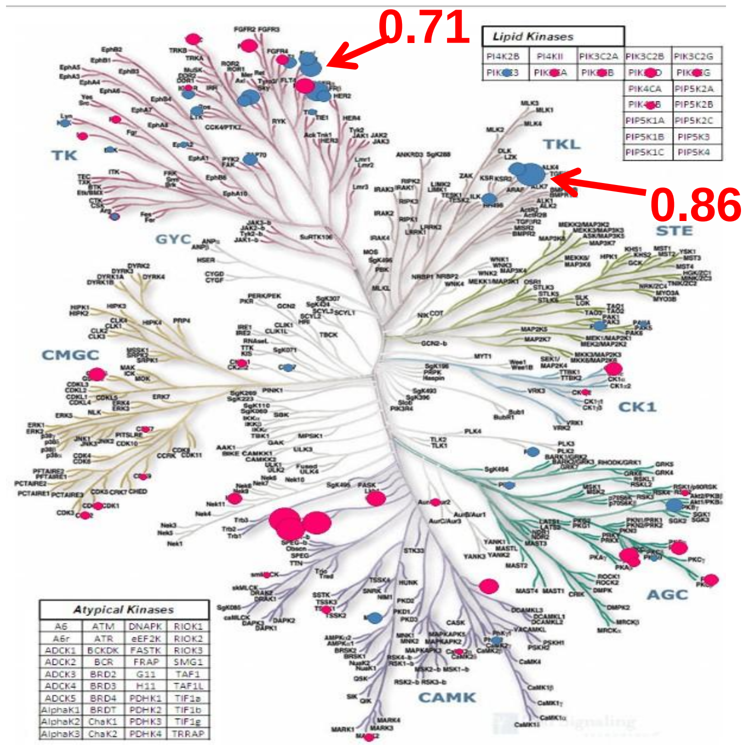
bRaf: Cellular Model Coefficients Anticipate Clinical Escape MoA

a) **TM:** RAF is Dominant Coefficient



PNAS, 2010, 4299-4304
Selective dual PDGFRb/bRaf inhibitor shows in-vivo efficacy in murine cancer model

b) **CP**: Raf is Largest but also TK’s, esp. PDGFRb



Nature, 2010, 973-977; NEJM, 2010, 772-774
Clinical Raf escapes from PDGFRb upregulation fixed by co-treatment with PDGFRb inhibitor.

NOVARTIS

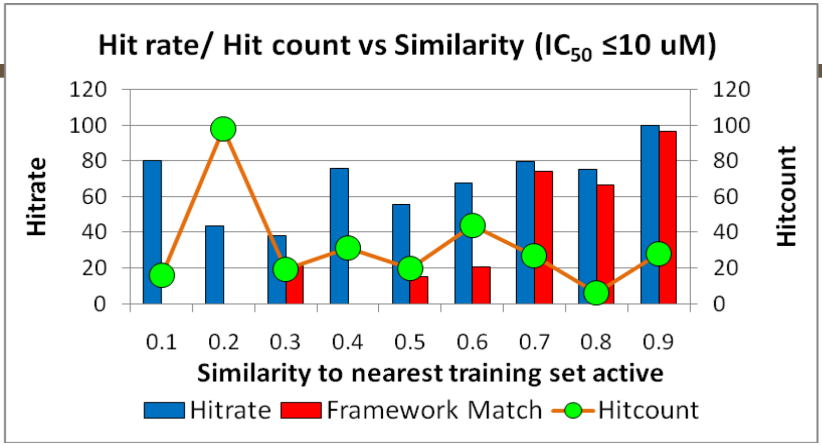
SGK3: Selective Tool Compound for Chemical Validation:
Hit-list Triaging: Virtual counter-screen against all TK, AGC, or PI3Ks

~100,000 focused expt'l 1-pt. MTS
4,000 %I > 30%

PQ for selectivity against ***TK, AGC, or PI3Ks*** and **TM/Biochem** =>cellular exposure

IC_{50}s for 900 Cpds

20 cpds selected for 70-kinase Panel



PQ+SA (R^2_{ext}=0.53): ~2500 cpds

Hundreds of novel cpds,
so HTS to be complex cascade assay

Best from PFVS	
SGK3	**150 nM**
Kinases < 1 uM	**0/69**
CaCO-2	**AB=3, BA=2**
Solubility	**97 µM**
cLogD (pH 7.4)	**0.44**

1.5 FTE
4 mo.

Improved by SBDD	
SGK3	**50 pM**
Next Kinase	10 nM
Panel < 1 uM	3/59
Target Mod.	**15 nM**
Cell Prolif.	**4 uM**

10^5X

300X

Project parked pending further biology

NOVARTIS

MELK: Need for Speed in patent race

Extensive novel chemistry identified by PFVS

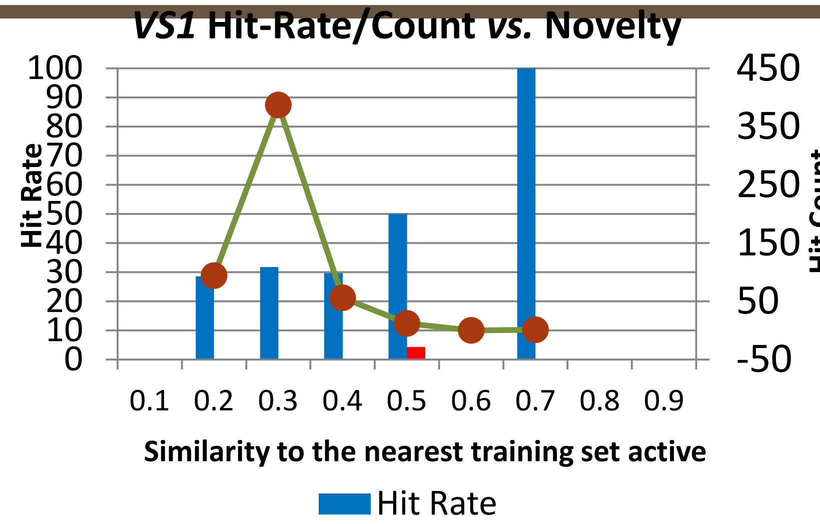
1123 IC_{50}s from Emeryville MTS in 2001
Model R^2_{ext}=0.52 biased training cpds.

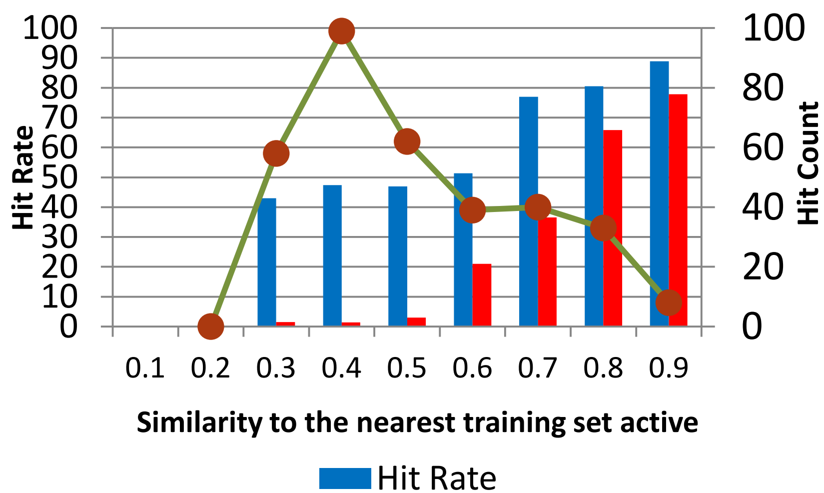
↓

VS1*: 1.5 million cpds** screened,1762 cpds tested (novel and lead-like***); **550<10 uM (31%), 188<1 uM**

↓

VS2*:** Model includes new cpds: R^2_{ext}=0.62. ***Novel, lead-like, and ***SAR-series***. ~1700 cpds tested; **585<10 uM (35%), 281<1uM**



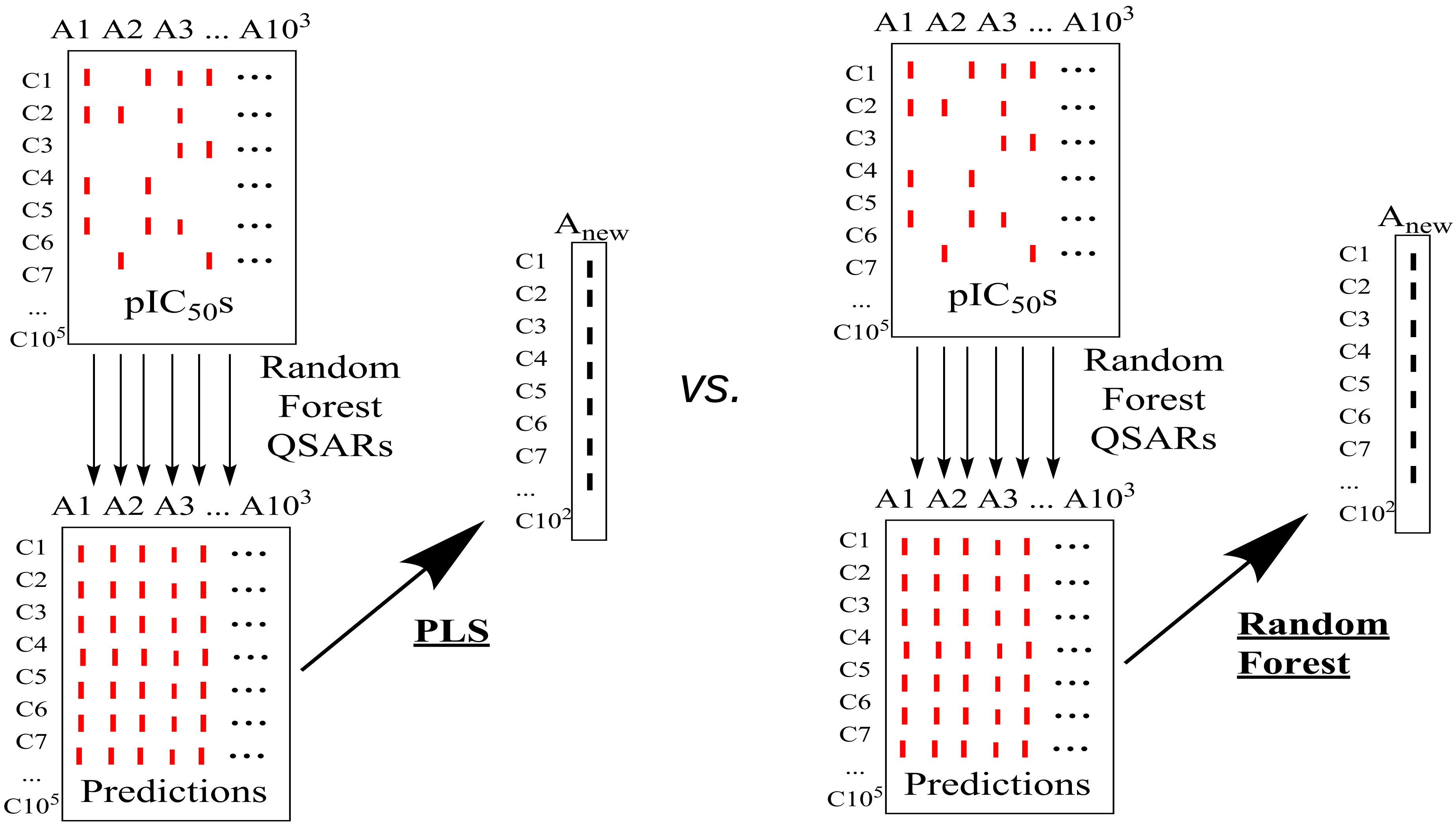


- Discovered 3 cpd series with good selectivity, phys props & *in vitro* PK.
  - One series has no IP.
- 1 ligand-efficient fragment size series, not yet selective.
- Cancelled HTS.
- Optimized to sub-nM cpds

NOVARTIS

Is PLS Obsolete?

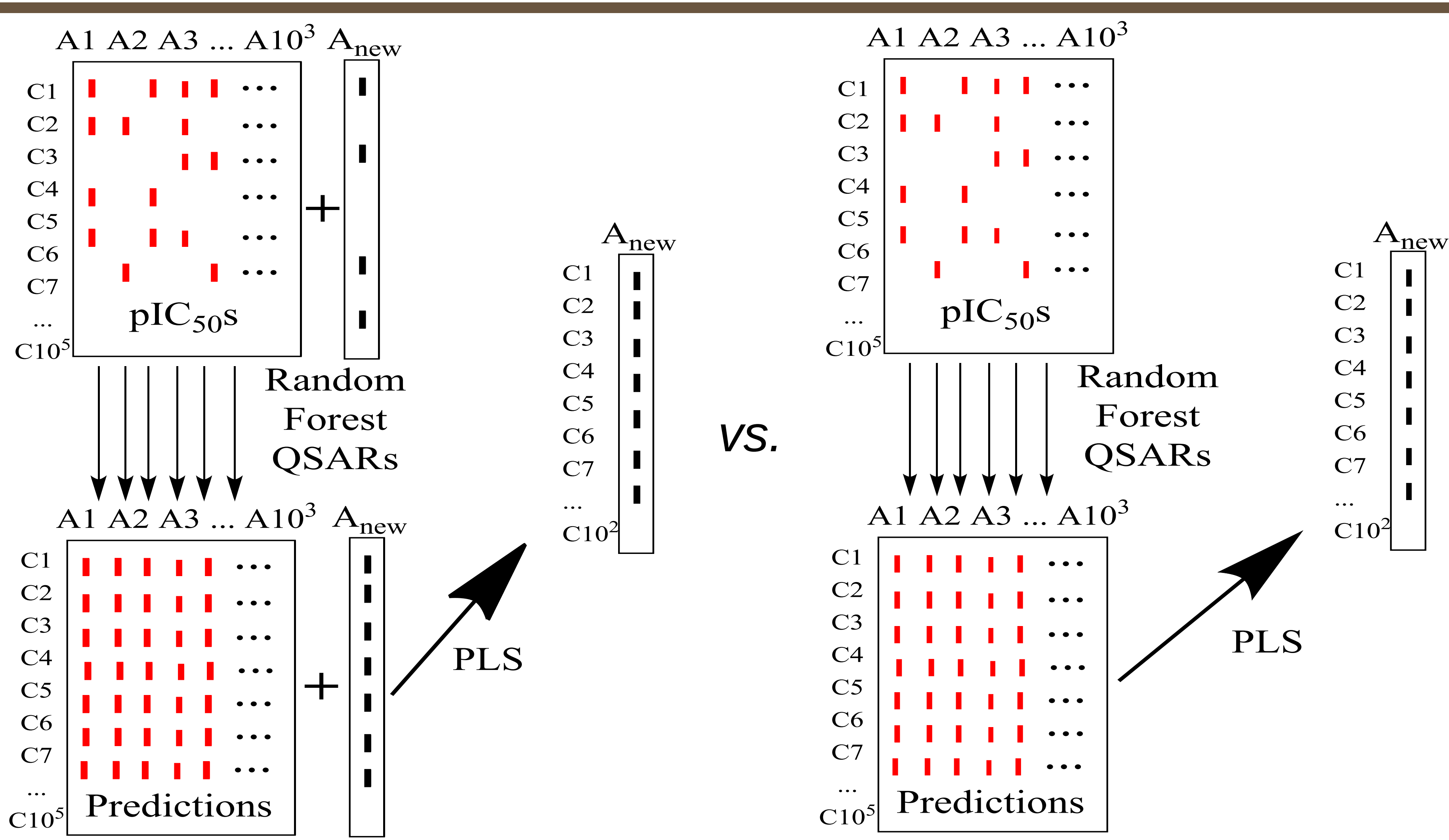
Use modern, non-linear random forest in second stage?



NOVARTIS

Is RF perfect?–Yes, but it fits data too well & “hogs” the signal

Rather remove from basis set with R^2>0.95!



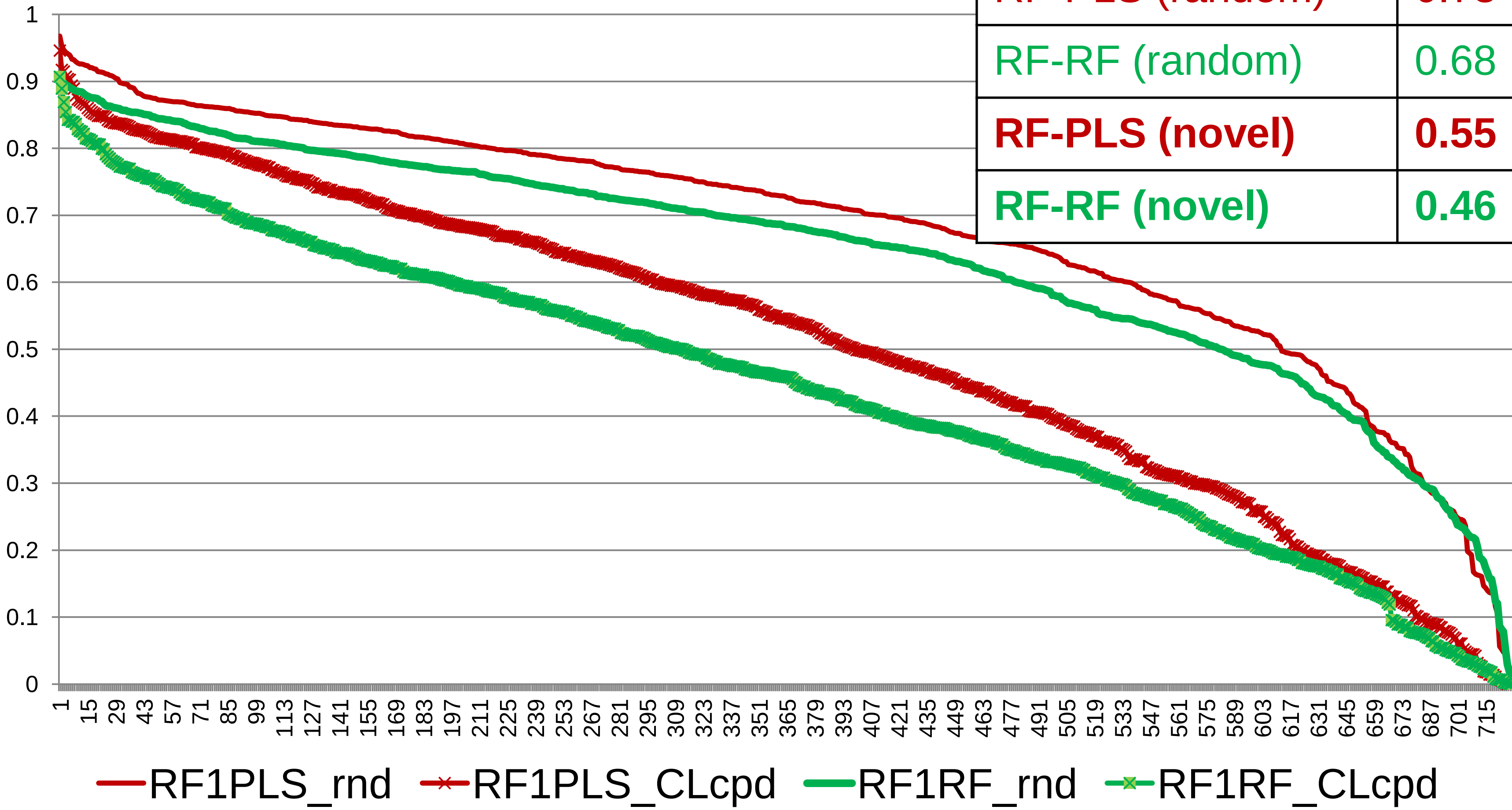
Perfection, the enemy of the good!

NOVARTIS

PLS is not Obsolete!

PLS extrapolates better than random forest in the 2nd layer

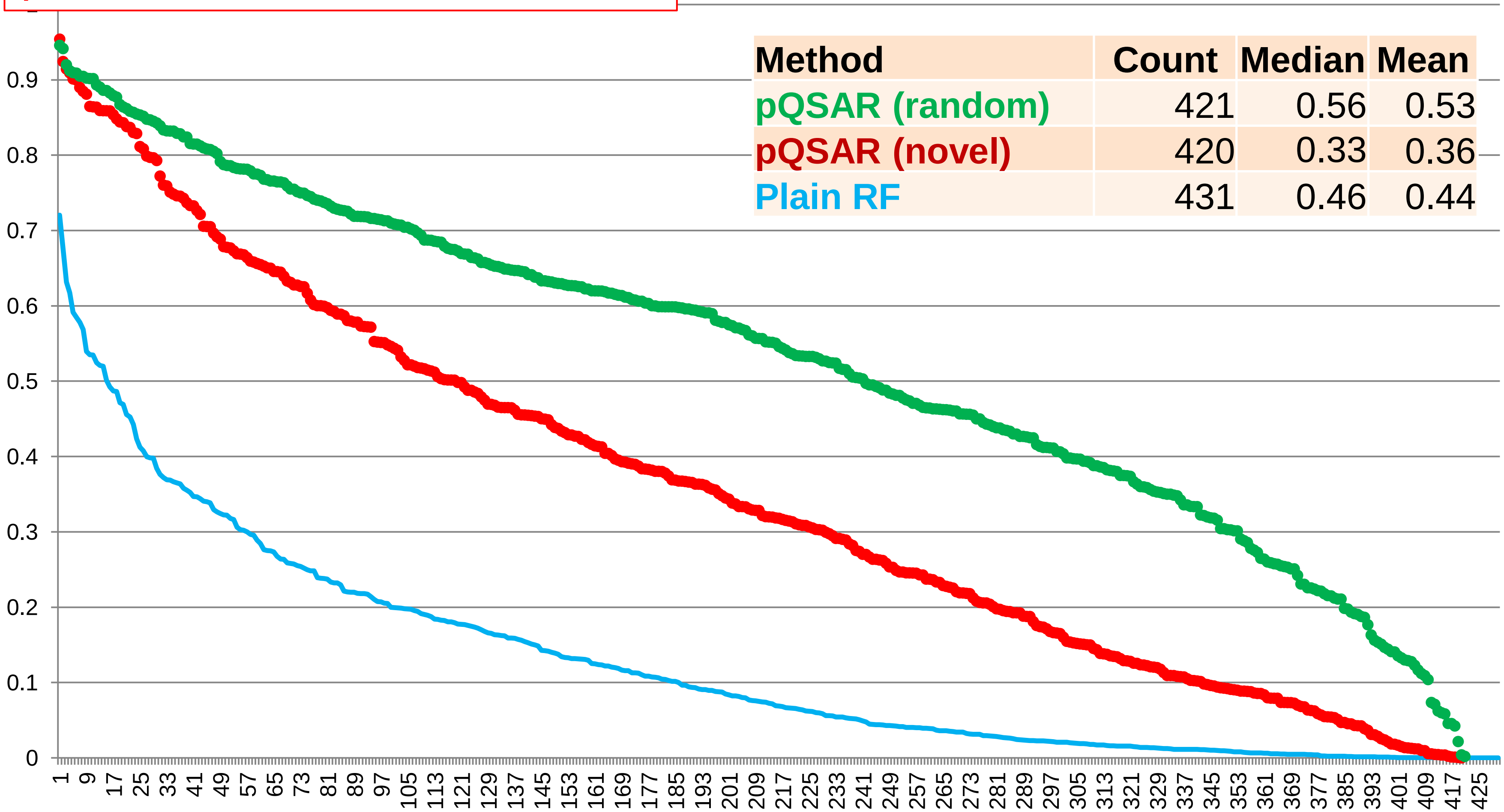
Method	**Median R^2**
RF-PLS (random)	0.73
RF-RF (random)	0.68
RF-PLS (novel)	**0.55**
RF-RF (novel)	**0.46**



GPCR: Profile-QSAR Models

Less accurate than kinases, but still very good

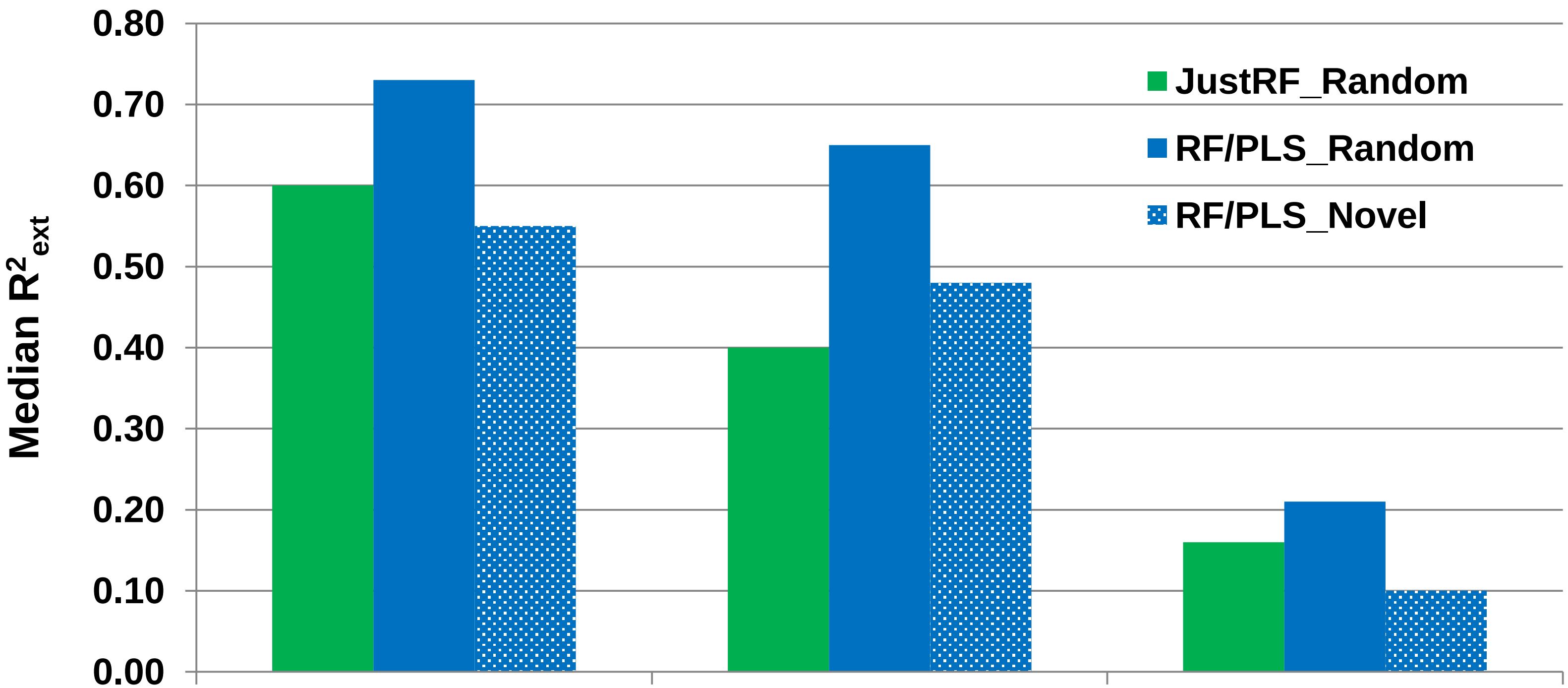
pQSAR median $R^2 = 0.33 - 0.56$



Method	Count	Median	Mean
pQSAR (random)	421	0.56	0.53
pQSAR (novel)	420	0.33	0.36
Plain RF	431	0.46	0.44

Public Data Sets: Strong Models for Kinases

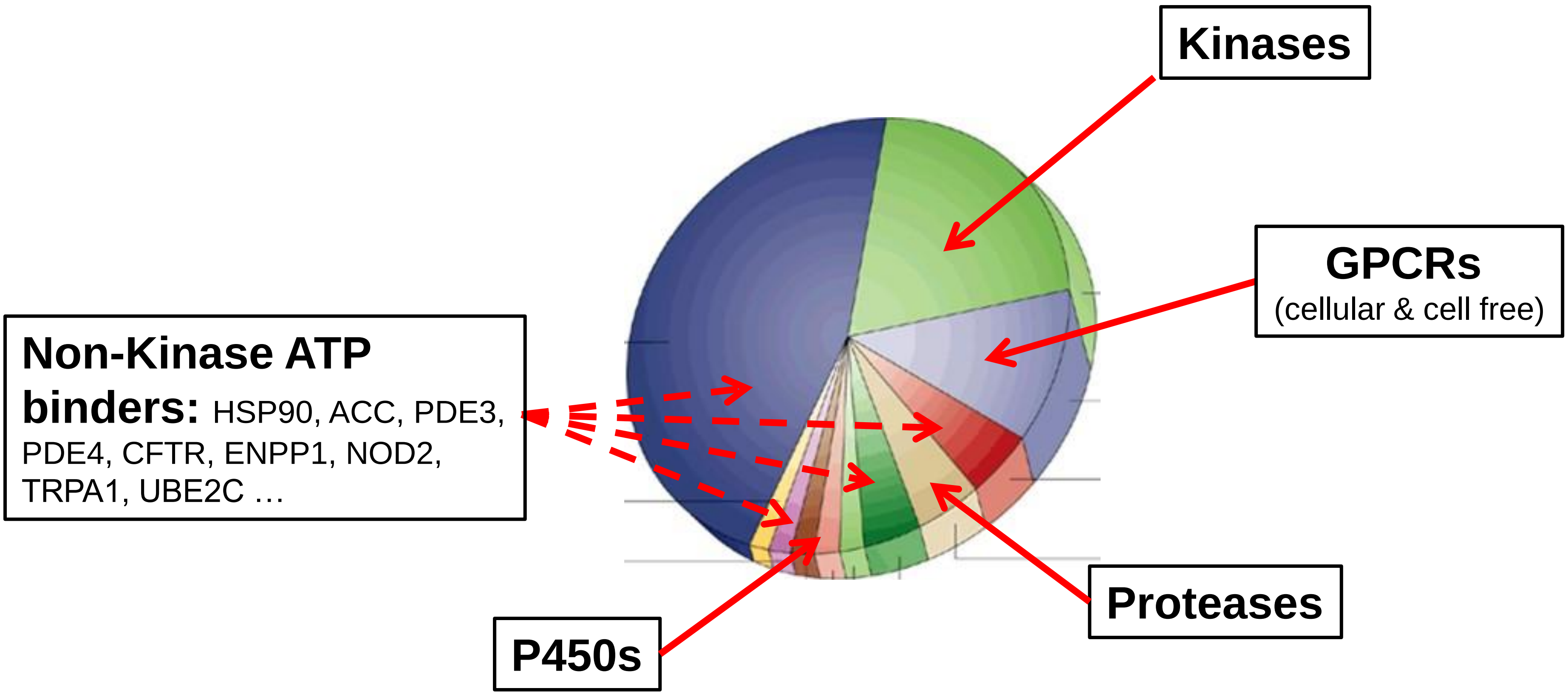
Prediction power transferable w/ high quality data



	Novartis	ChEMBL	PubChem
assay_cnt	728	159	171
cpd_cnt	377k	13k	233k
data_cnt	3.2m	0.11m	0.45m
median_cpd/assay	1621	683	484
data/cpd	8.4	8.7	1.9

NOVARTIS

Profile-QSAR now covers 50% of Estimated Druggable Genome



Druggable genome from Hopkins et al., *Nature Reviews Drug Discovery* 1, 727-730 (2002)