The BADAPPLE promiscuity plugin for BARD Evidence-based promiscuity scores

Jeremy Yang, UNM
Oleg Ursu, UNM
Cristian Bologa, UNM
Anna Waller, UNM
Larry Sklar, UNM
Tudor Oprea, UNM

BioAssay Research Database

Translational Informatics Division

UNM | Center for Molecular Discovery

ACS National Meeting – Indianapolis, IN -- Sep. 8-12, 2013

CINF: Integrative Chemogenomics Knowledge Mining Using NIH Open Access Resources

What is BADAPPLE?

- **B**io**A**ctivity **D**ata **A**ssociative **P**romiscuity **P**attern **L**earning **E**ngine
- Bioassay data analysis algorithm
- Scaffold association patterns
- Evidence-based
- Robust to noise and errors

What is promiscuity?

- Un-selective bioactivity
- Normally undesirable
- Evolving conceptions:
  - Polypharmacology
  - Systems biology
  - Systems chemical biology

Promiscuity & bioassay data analysis: Selected references

- Frequent hitters (2002, Schneider &al.)
- Aggregators (2003, Shoichet &al.)
- ALARM NMR (2004, Hajduk &al.)
- Promiscuous scaffolds *[talk]* (2007, Bologa)
- Pan-Assay Interference (2010, Baell &al.)
- PubChem Promiscuity (2011, Canny &al.)

Prerequisites for Promiscuity Data Analysis

- Definition of unique chemical entity?
  - Yes

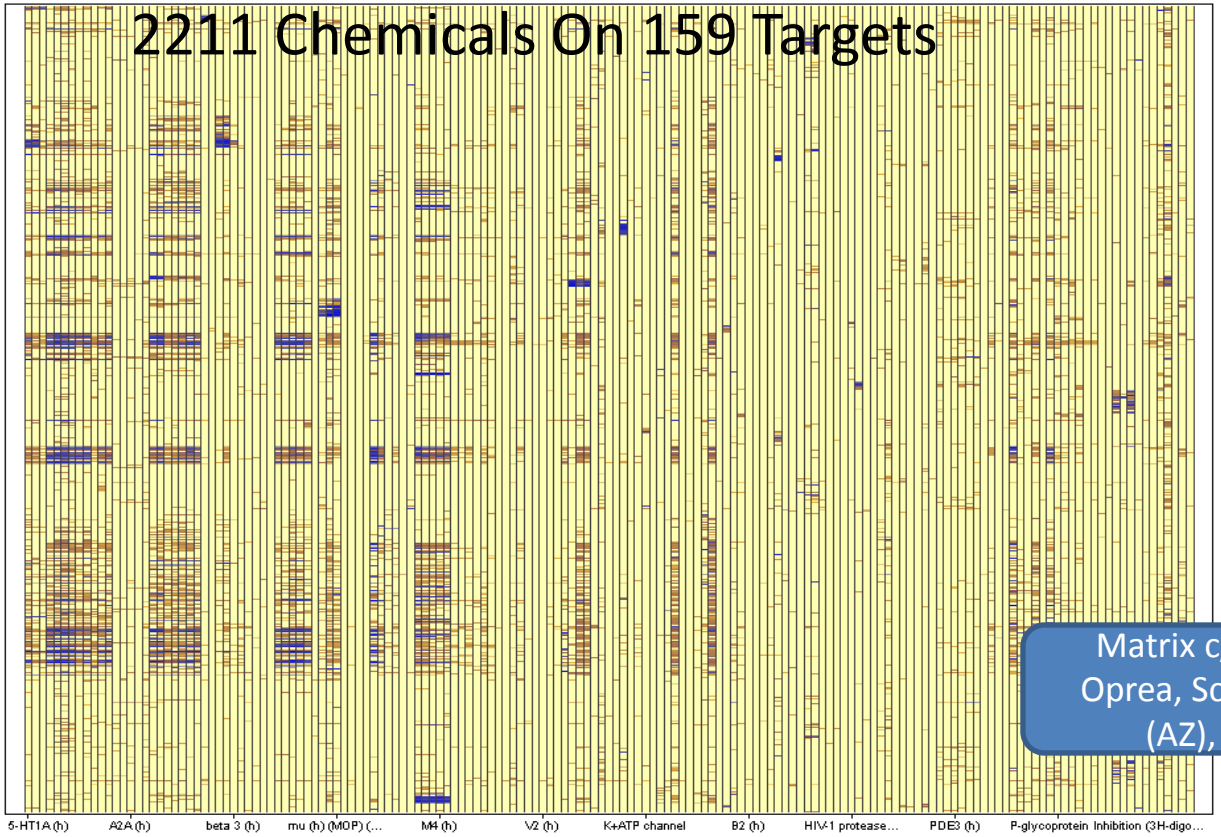
CN/C(=C\[N+](=O)[O-])/NCCSCC1=CC=C(O1)CN(C)C.Cl

- Definition of unique biological entity?
  - Challenging

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

- I.e., Rigorously calibrated bioactivity matrix

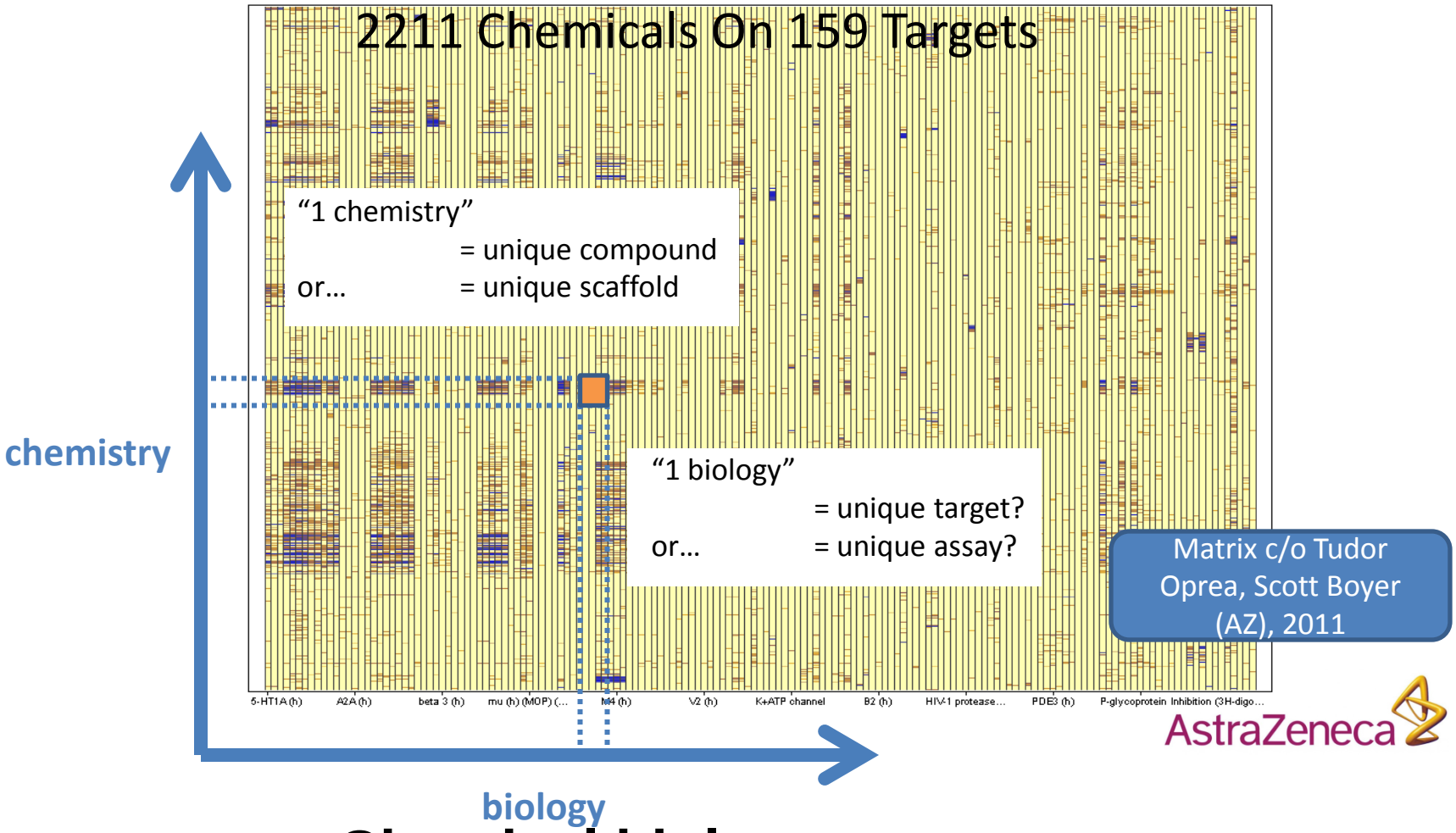
Bioactivity matrices depend on rigorous informatics



Matrix c/o Tudor Oprea, Scott Boyer (AZ), 2011

AstraZeneca

Bioactivity matrices depend on rigorous informatics



Chemical biology space:
To define a space must define a point

Promiscuity: related concepts

- Assay-interferers
- Experimental artifacts
- Frequent hitters
- False positives
- True positives
- True non-selective actives
- Aggregators
- Reactives
- Cytotoxic

Badapple promiscuity score: a practical definition for bioassay data analysis

- **Purpose:** Streamline molecular discovery project workflow.
- **Hence:** Score designed to detect "false trails" (true-promiscuous OR false positives) unlikely to be productive leads.

What is evidence-based?

- Empirical data analysis
- Mistakes can be un-learned.
- Data driven
- Not: Expert systems

Drawbacks of evidence-based

- Theories proven wrong. Ouch.
- Reality is messy.
- GIGO.

Benefits of evidence-based

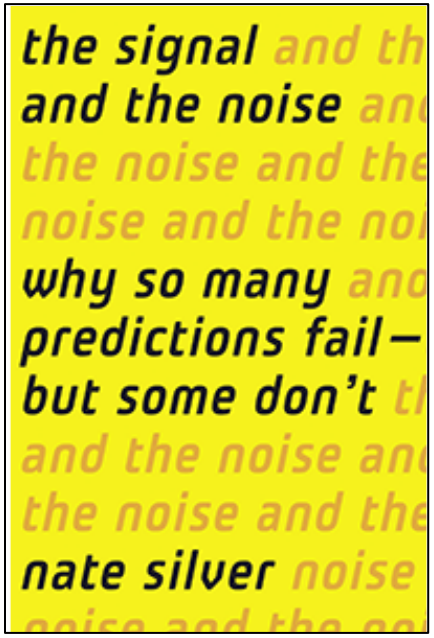
When the Nerds Go Marching In

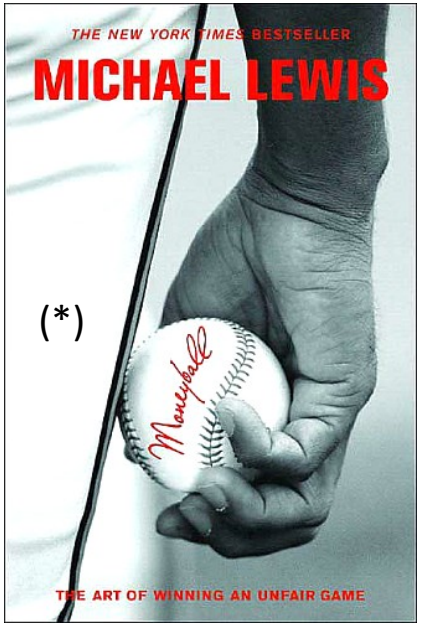
ALEXIS C. MADRIGAL | NOV 16 2012, 7:00 AM ET

How a dream team of engineers from Facebook, Twitter, and Google built the software that drove Barack Obama's reelection

Three members of Obama's tech team, from left to right: Harper Reed, Dylan Richard, and Mark Trammell (Photo by Daniel X. O'Neil).

The Obama campaign's technologists were tense and tired. It was game day and everything was going wrong.





More wins. More knowledge. Lower costs. Progress.

(*) N.B. key role of Dick Cramer, ACS CINF 2013 Skolnik award winner.

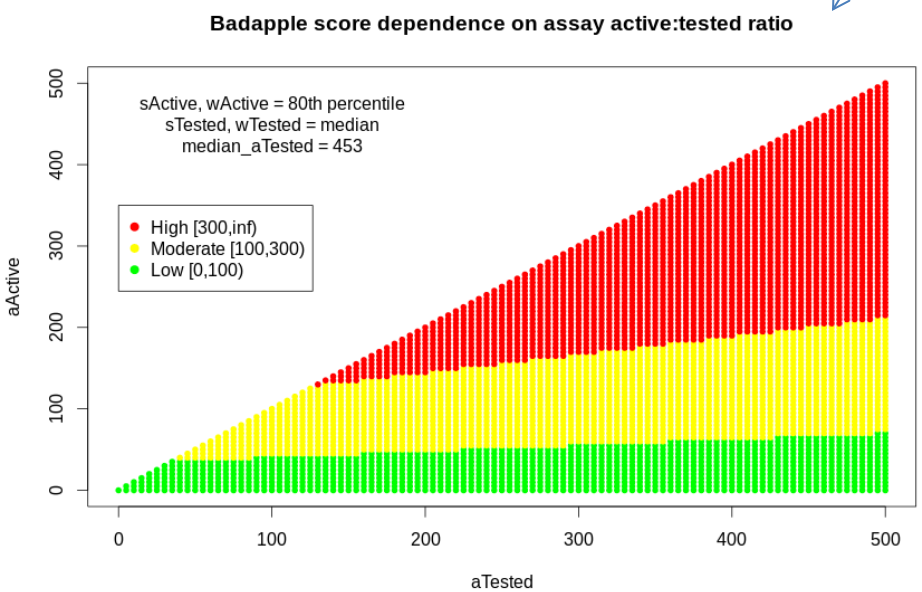
BADAPPLE scoring function

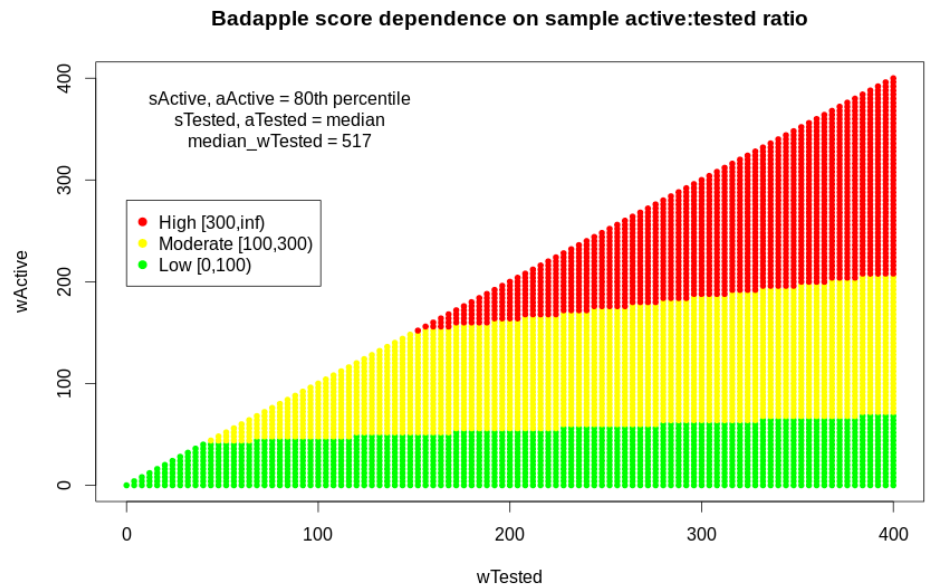
$$score = \frac{subActive}{subTested + median(subTested)} * \frac{asyActive}{asyTested + median(asyTested)} * \frac{samActive}{samTested + median(samTested)} * 10^5$$

- Scaffold score
- By scoring scaffolds, more relevant evidence
- Substances, assays and samples considered
- Penalize under-sampling
- Score is a **statistic**, not a "model"!
- Inherently "validated"

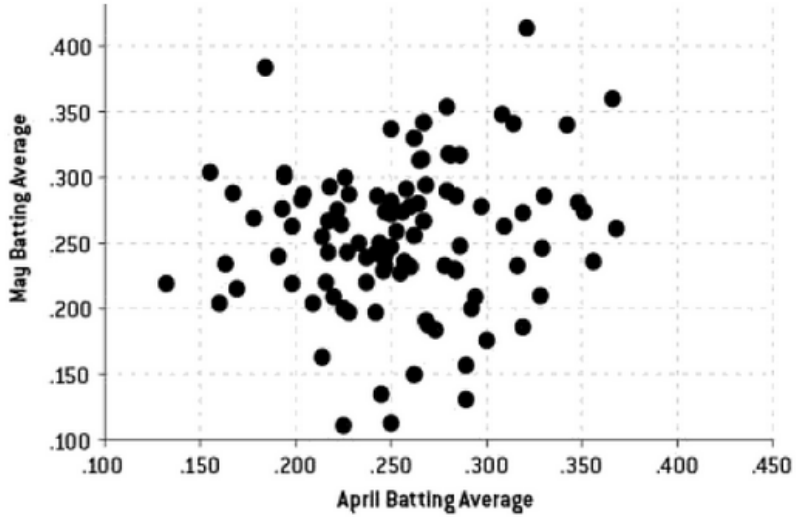
Avoiding overfitting; being skeptical of scanty evidence

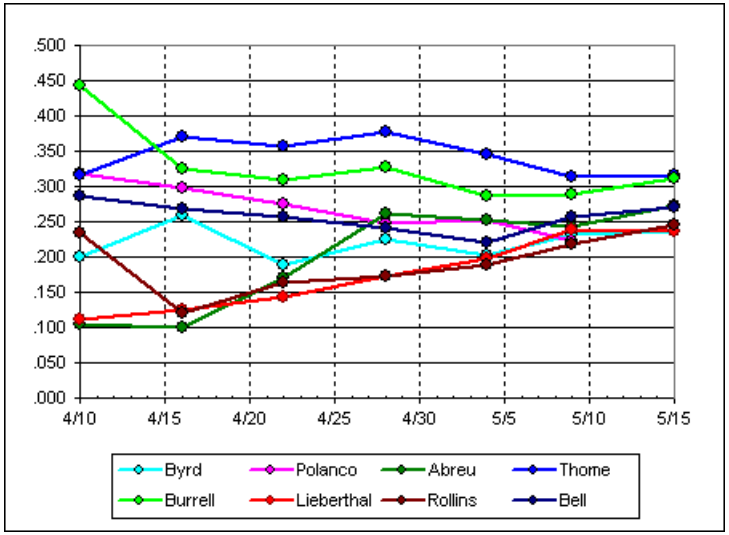
$$score = \frac{subActive}{subTested + median(subTested)} * \frac{asyActive}{asyTested + median(asyTested)} * \frac{samActive}{samTested + median(samTested)} * 10^5$$





Avoiding overfitting: Moneyball style





The Signal and the Noise: Why So Many Predictions Fail—But Some Don't, Nate Silver (2012).

http://mark.stubbornlights.org/phils/archives/2004_05.html

Sabermetrics leads the way.

BADAPPLE public web app

badapple - BioActivity Data Associative Promiscuity Pattern Learning Engine

help

mode: single (1st) multi

input mol[s]: format:

automatic

upload: Choose File No file chosen

...or paste: file2txt

```
clccc2c(cl)c(=O)n(s2)c3ccccc3C(=O)N4CCCC4
```

or draw query...

output:

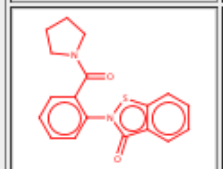
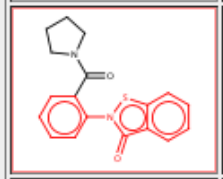
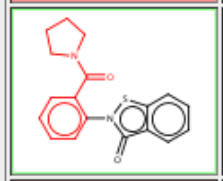
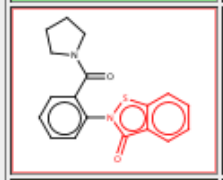
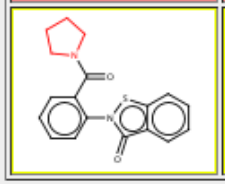
depict: mol scaf none S

sort scaffolds by: size id pScore

verbose

submit molecule[s]

rows returned: 5

scaffold	pScore	advisory	sTested	sActive	aTested	aActive	wTested	wActive	InDrug	ID
	0	none	0	0	0	0	0	0	FALSE	~
	835	High pScore.	40	37	526	127	7898	615	FALSE	22888
	18	Low pS								.4
	815	High pS								89
	148	Modera pScore								

```
Apache Derby 10.5
Apache Derby Embedded JDBC Driver 10.5.3.0 - (802917)
BADAPPLE database (all MLSMR compounds, selected MLP HTS assays)
total compounds: 373802
total scaffolds: 146024
```

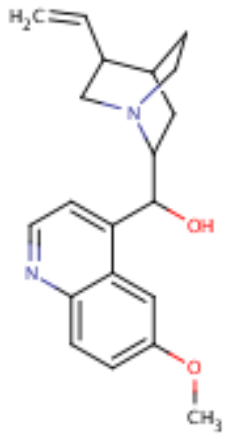
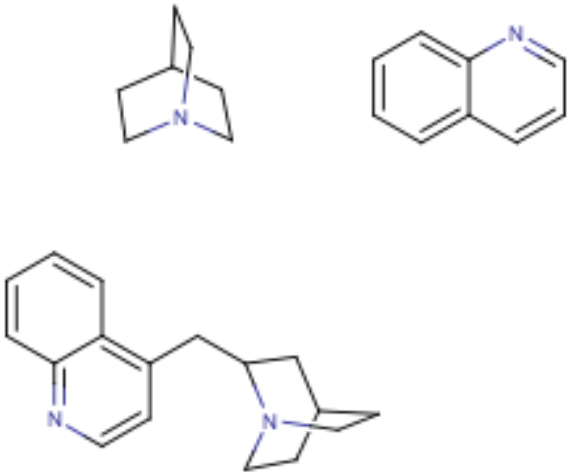
pScore range	advisory
~	unknown; no data
0.0-100	low pScore; no indication
100-300	moderate pScore; weak indication of promiscuity
>300	high pScore; strong indication of promiscuity

ChemAxon

*http://pasilla.health.unm.edu/tomcat/biocomp/badapple

Why Scaffolds?

biology	-	birds, bees, nature
chemistry	-	chemists
SAR	-	drugs
IP	-	lawyers

	mol	scaffolds
1.	 quinine	

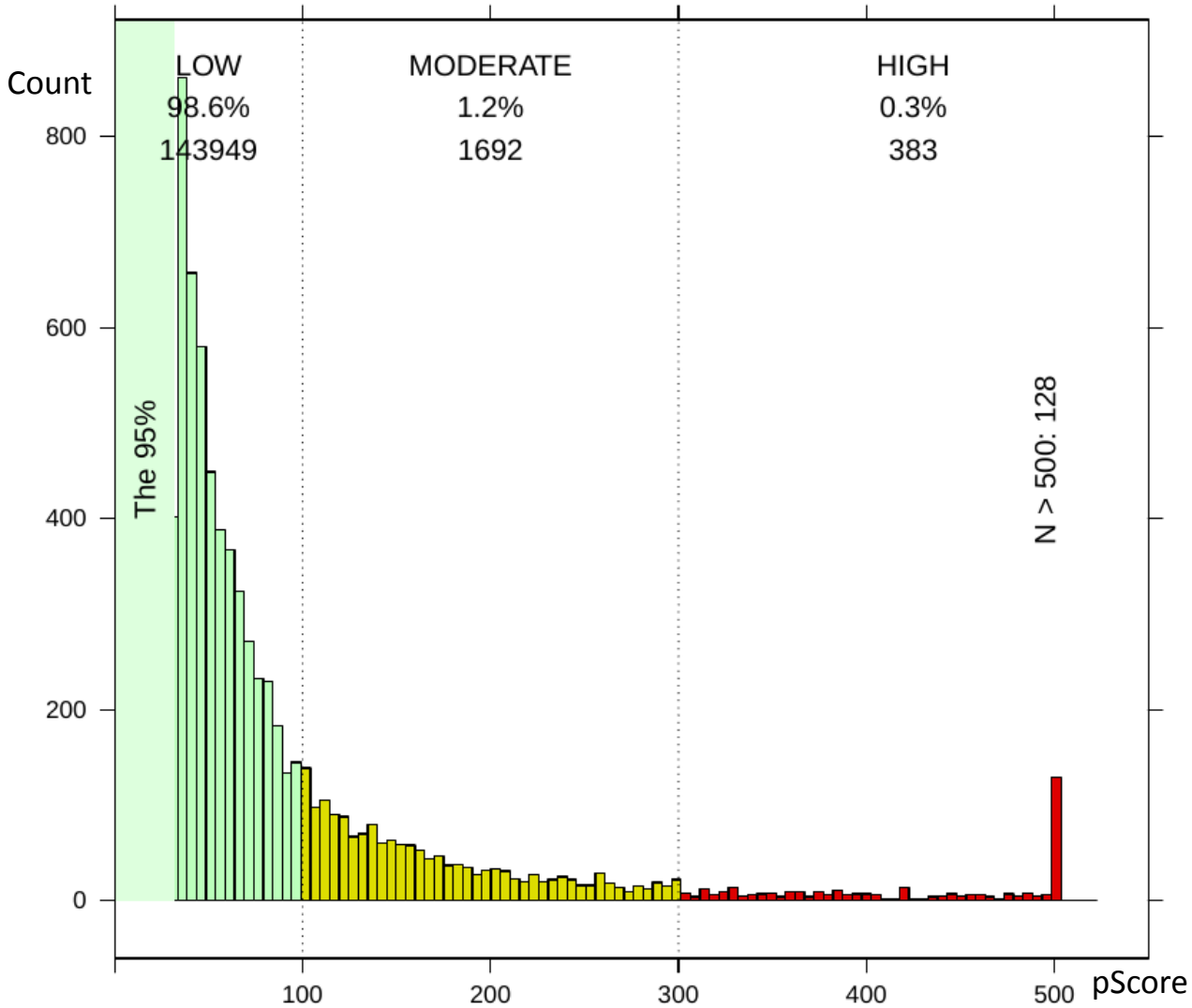
CINF grad student symposium 9/8/13 & ACS SciMix 9/9/13 8pm

Molecular scaffolds are special and useful guides for discovery

Jeremy Yang, UNM & IU
Cristian Bologa, UNM
David Wild, IU
Tudor Oprea, UNM

ACS National Meeting - Sept. 8-12, 2013 - Indianapolis, IN

There's something about scaffolds... Especially the "privileged few"...

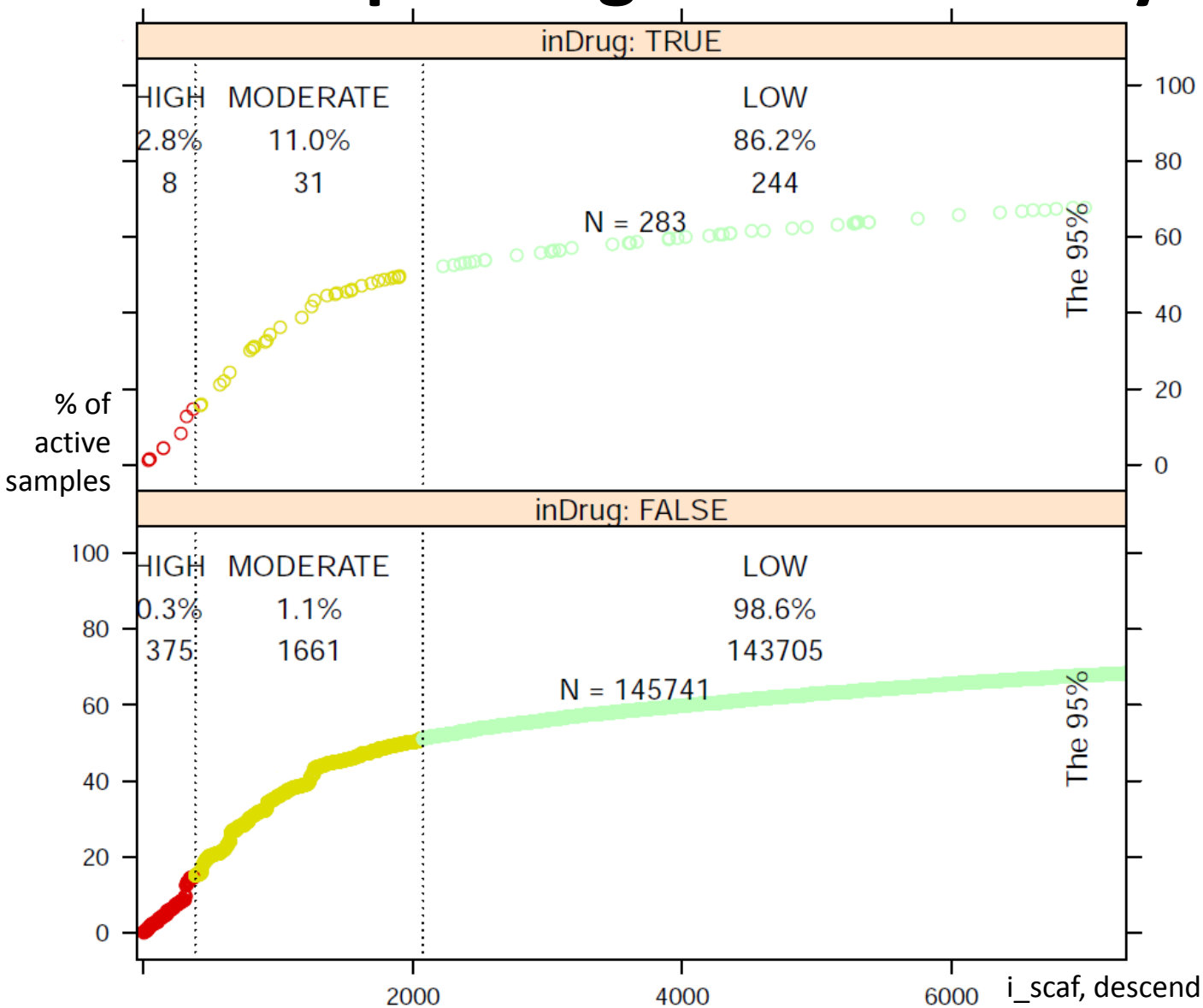


Promiscuity score distribution

Dataset:
BARD,
MLSMR,
MLP HTS

Totals:
compounds:
373,802 ;
scaffolds:
146,024 ;
assays: 528 ;
wells/results:
30,612,714

Scaffolds & drug-scaffolds, the privileged few explaining a lot of activity...



Dataset: BARD, MLSMR, MLP HTS

Totals: compounds: 373,802 ; scaffolds: 146,024 ; assays: 528 ; wells/results: 30,612,714; drugs: 283; drugscafs: 1958

Scaffolds & drug-scaffolds, the privileged few explaining a lot of activity...

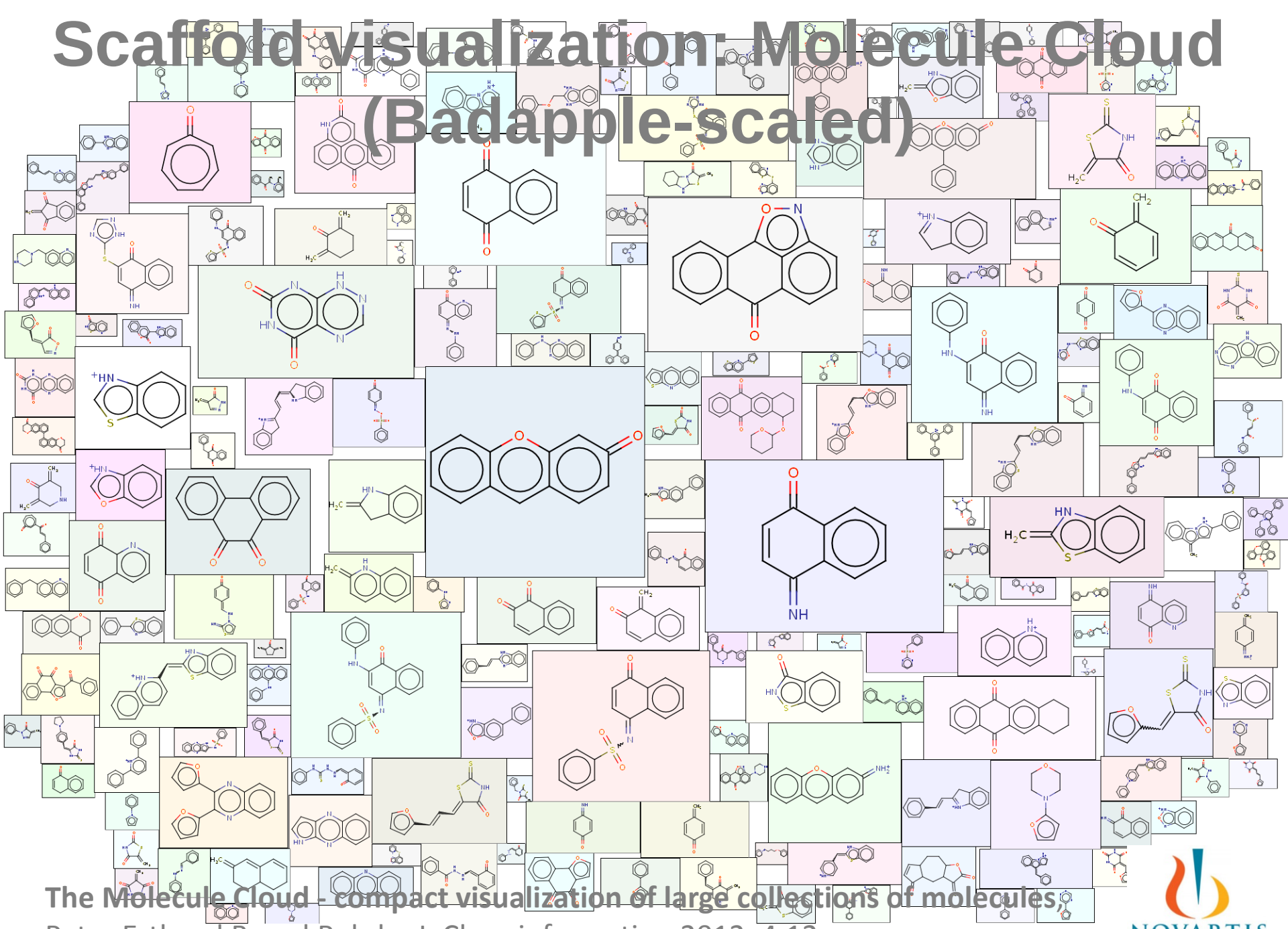
	% total activity	# scaffolds	% scaffolds
All	50%	1979	1.4%
All	75%	11,645	8%
Drugs	50%	54	2.8%
Drugs	90%	327	16.7%

"total activity" = active scaffold-instances

Dataset: BARD, MLSMR, MLP HTS

Totals: compounds: 373,802 ; scaffolds: 146,024 ; assays: 528 ; wells/results: 30,612,714; drugs: 283; drugscafs: 1958

Scaffold visualization: Molecule Cloud (Badapple-scaled)



The Molecule Cloud - compact visualization of large collections of molecules,
Peter Ertl and Bernd Rohde, J. Cheminformatics, 2012, 4:12.

NOVARTIS

What is BARD?

- BioAssay Research Database, http://bard.nih.gov
- MLP: 1000+ assays, 400k+ cpds, 200M data
- Manual assay annotations + QA
- BARD Assay Ontology
- Assay Data Standard
- Community platform for bioassay data analysis & computation

BioAssay Research Database

BARD + Badapple Synergy

- BARD ontology, based on BAO
- BARD raising "semantic IQ" of public bioassay data.
- Evidence = information = data + metadata

BioAssay Research Database

http://bard.nih.gov/

BARD Plugin Platform: Community development Enterprise deployment

- BARD Plugin spec (IPlugin interface)
- BARD PluginValidator class
- Java, JAX-RS, Jersey, REST
- BARD REST API
- WAR deployment, discoverable

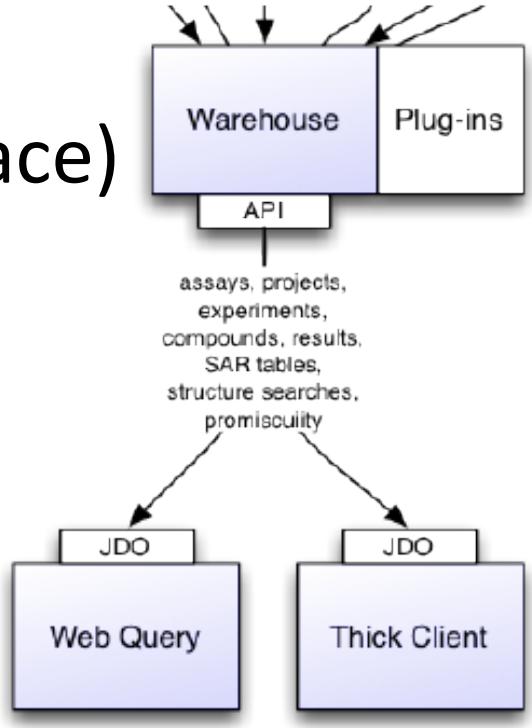
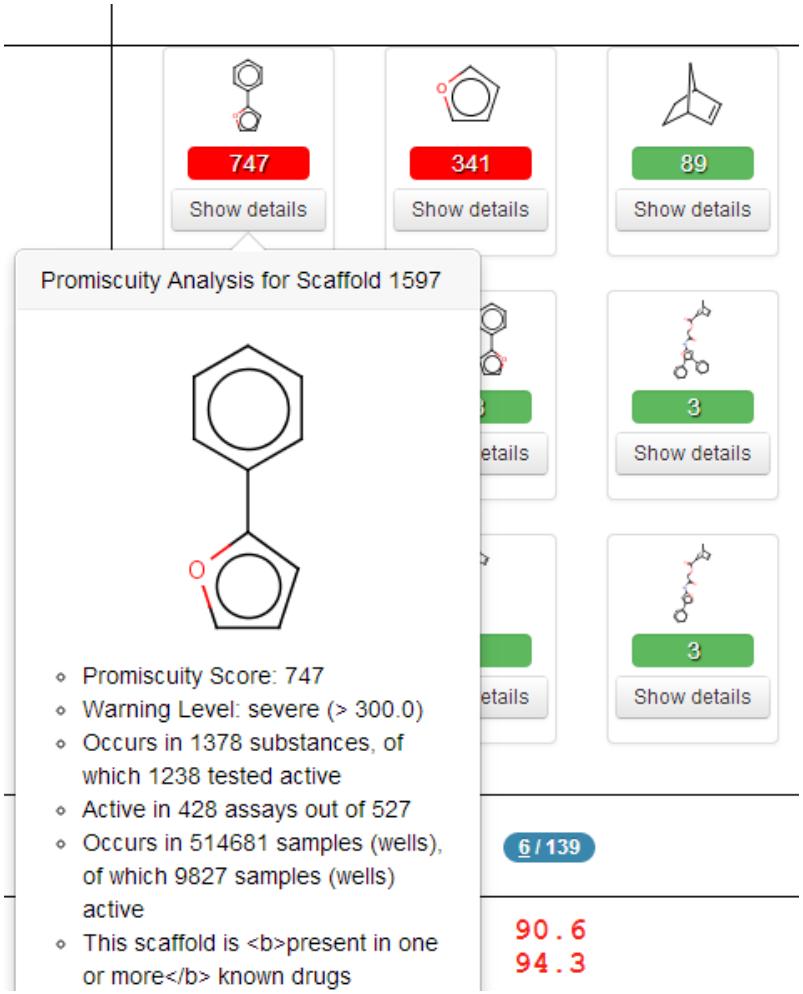
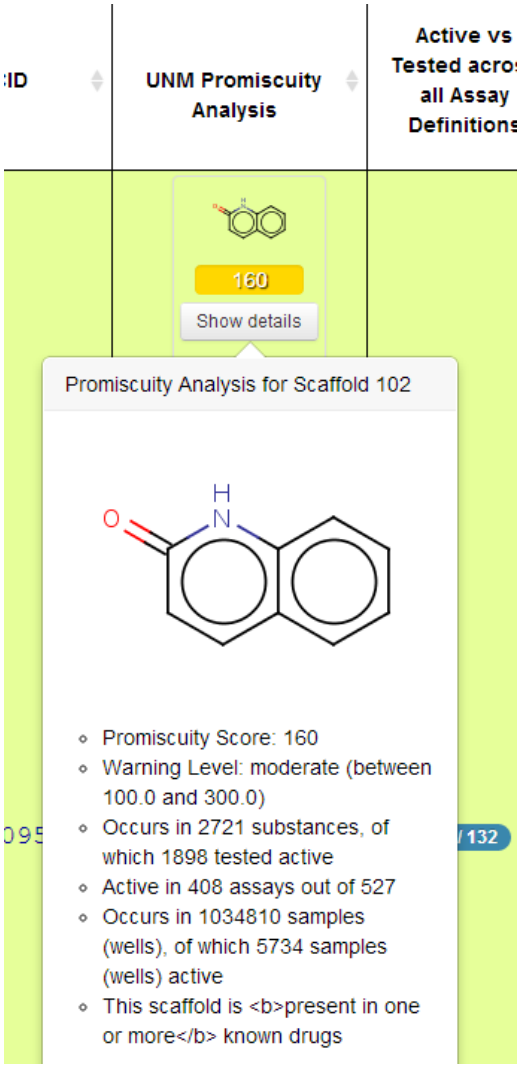


fig c/o Steve Brudz, Broad

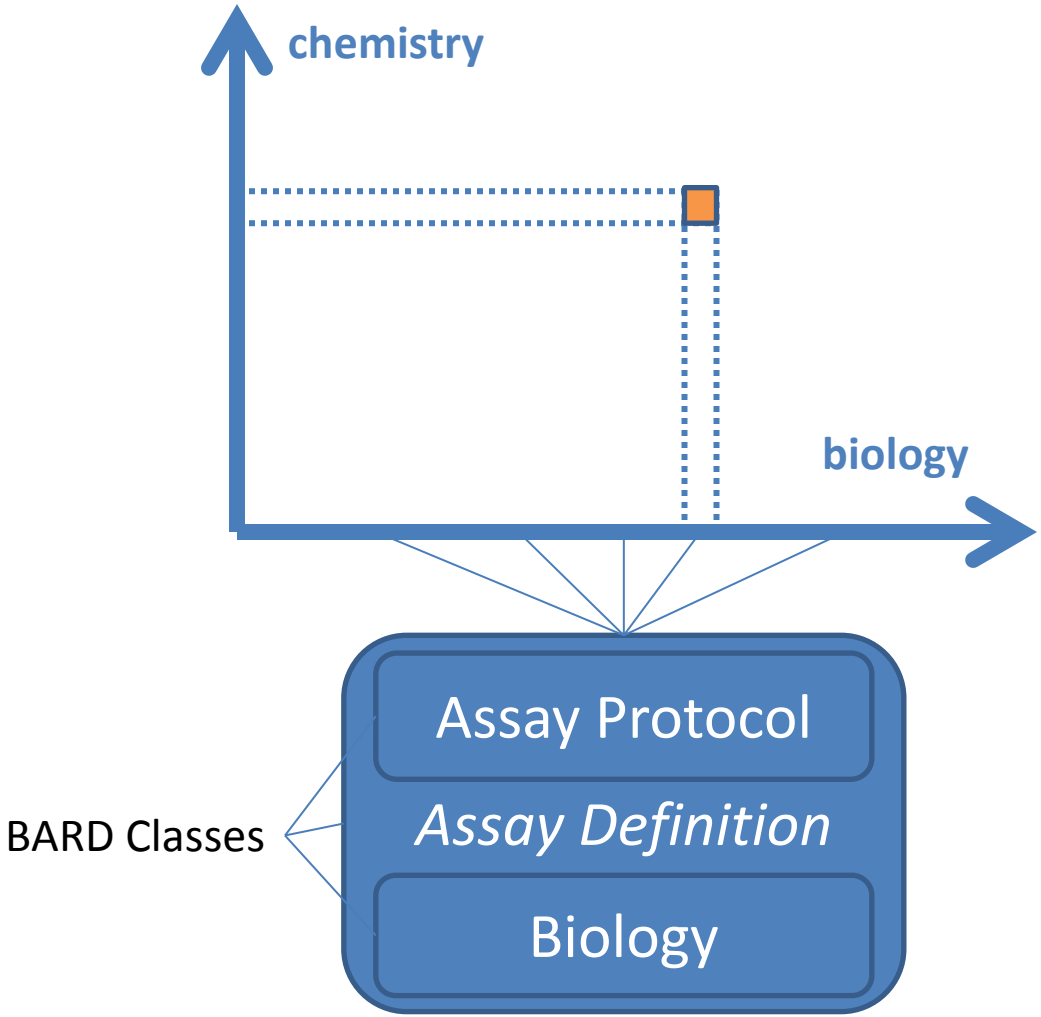
Badapple Plugin via BARD web client



Discovery scenario: Rapidly flags potentially problematic (*notorious*) scaffolds.



BARD Synergy, next steps: Raising Badapple to the next level



Re-calibrating the bioactivity matrix for improved rigor, accuracy & sensitivity, using the BARD ontology.

BARD Synergy, next steps: Raising Badapple to the next level

Re-calibrating the bioactivity matrix for improved rigor, accuracy & sensitivity, using the BARD ontology.

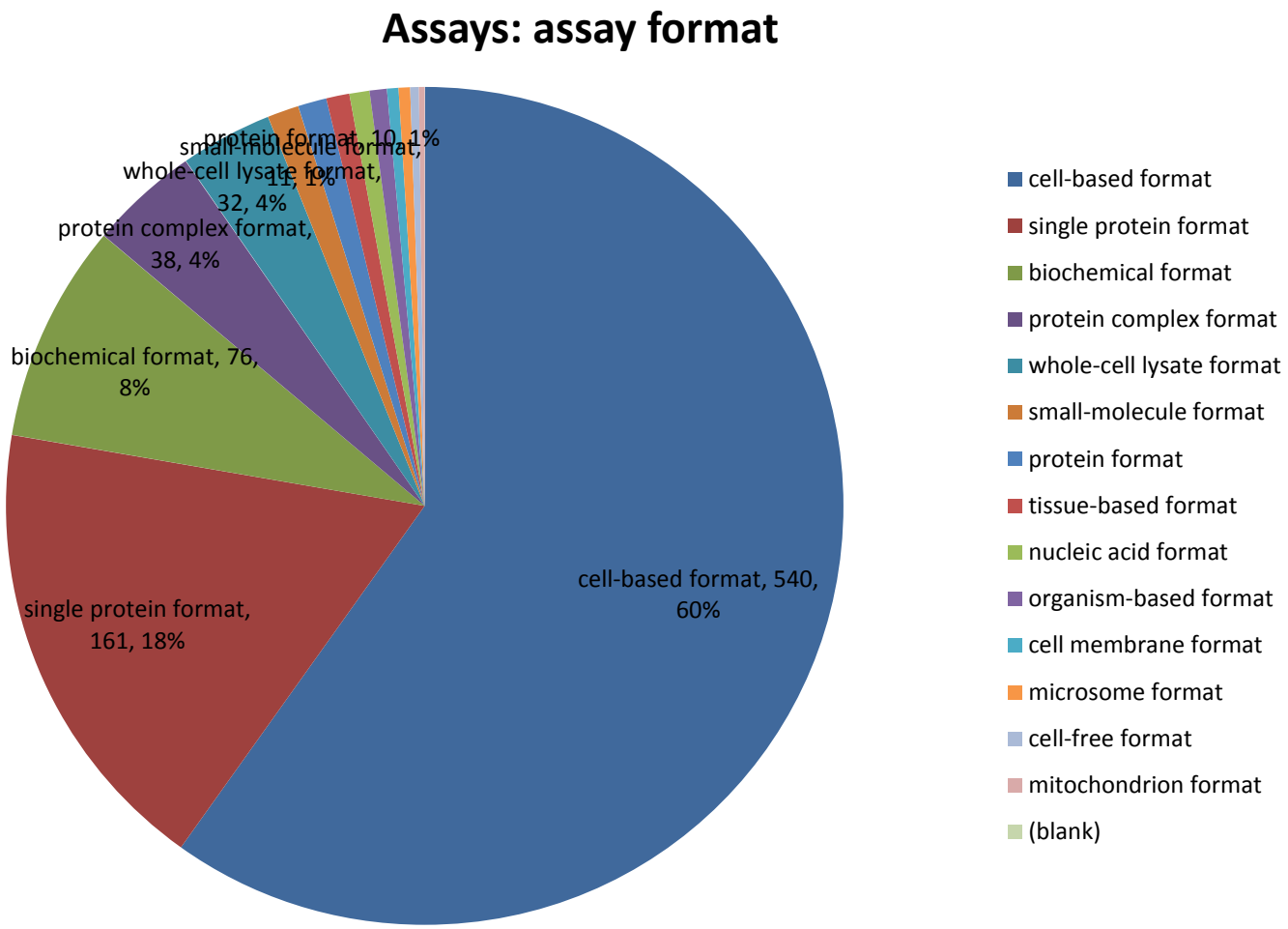
Some re-calibrations of interest:

Assay Formats

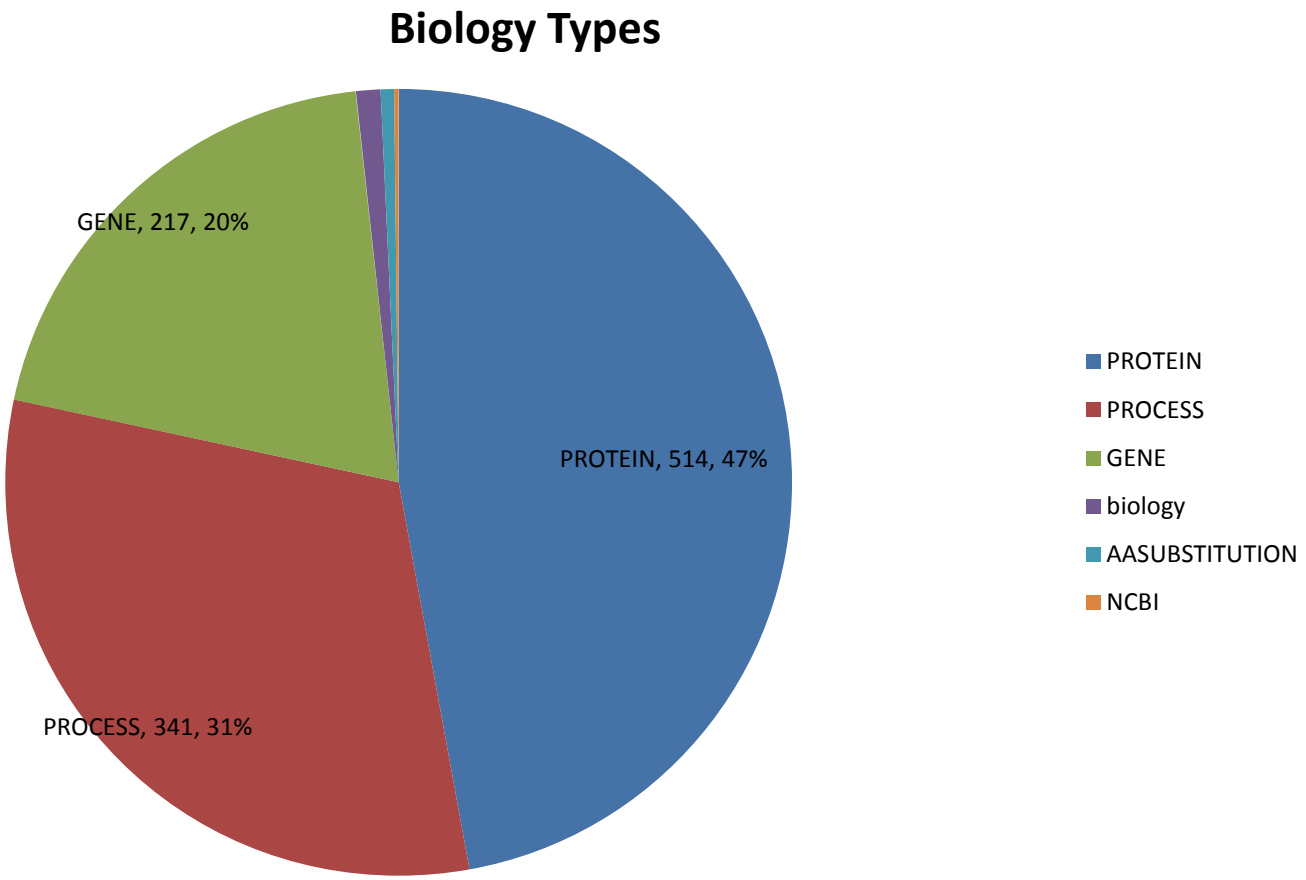
Biology Types

Protein Classes

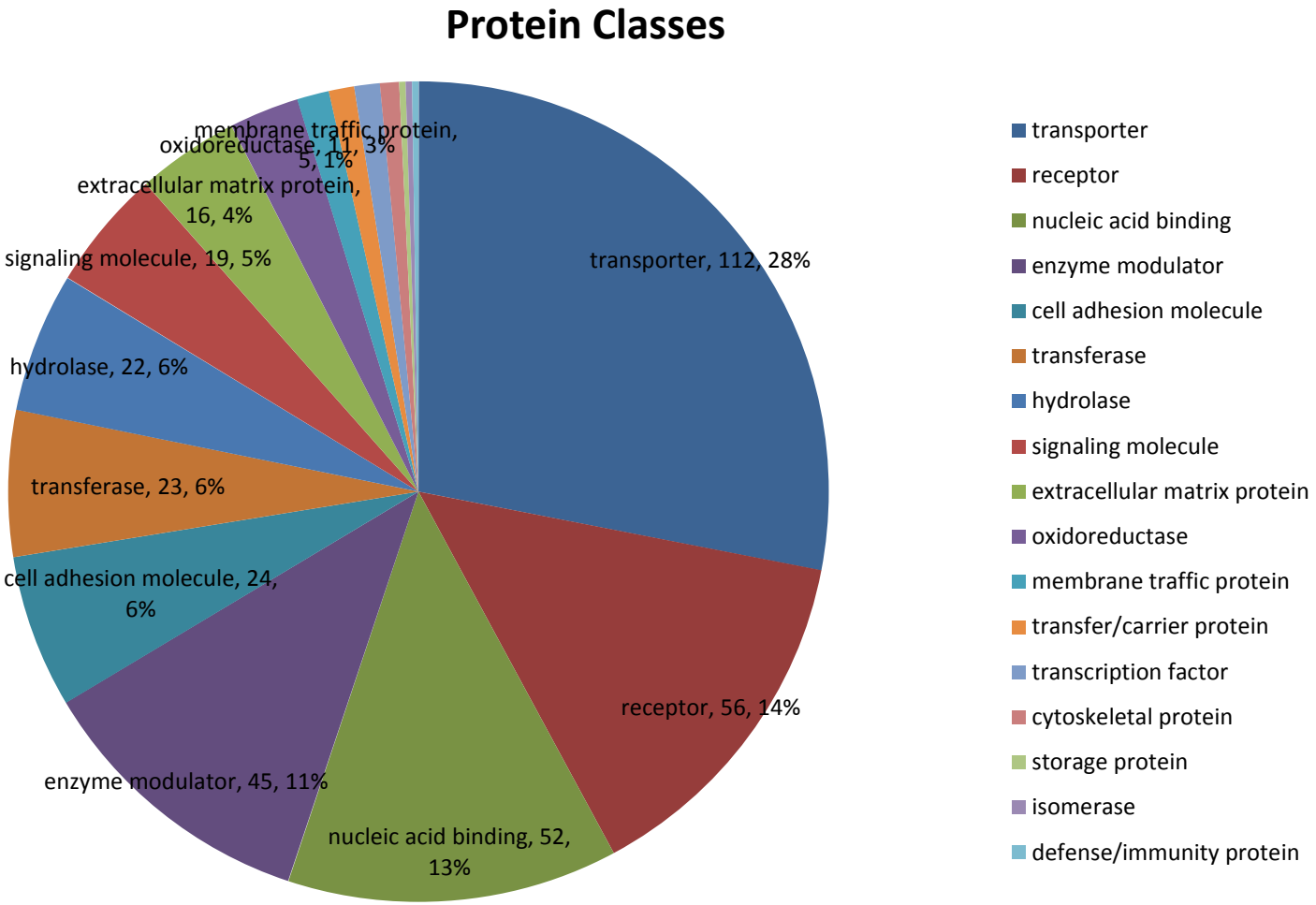
BARD Synergy, next steps: Raising Badapple to the next level



BARD Synergy, next steps: Raising Badapple to the next level



BARD Synergy, next steps: Raising Badapple to the next level



Conclusions

- Badapple exemplar as BARD plugin
- Badapple exemplar as evidence-based algorithm
- New BARD semantic capabilities will elevate Badapple to next level.
- Promiscuity a complex issue.

Acknowledgements

- Steve Mathias, UNM
- Chris Lipinski, Melior
- Rajarshi Guha, NCATS
- Stephan Schurer, UMiami
- Uma Vempati, UMiami
- Mark Southern, Scripps
- BARD Engineering Working Group

BioAssay Research Database