What can your library do for you?

Rajarshi Guha, Dac-Trung Nguyen,
Alexey Zhakarov, Ajit Jadhav

NIH NCATS

ACS Fall Meeting 2016, Philadelphia

August 21, 2016

Library Design

- Historical collections and assay data provide information on how a set of compounds has faired
- Use (dis)similarity and machine learning to construct new collections that show similar behavior
  - Plus various constraints
- Libraries can be designed for certain target families or specific screening paradigms

If sufficiently annotated, compound behavior may be correlated to assay and biology characteristics

Library Design

- Historical collections and assay data provide information on how a set of compounds has faired
- Use (dis)similarity and machine learning to construct new collections that show similar behavior
  - Plus various constraints
- Libraries can be designed for certain target families or specific screening paradigms

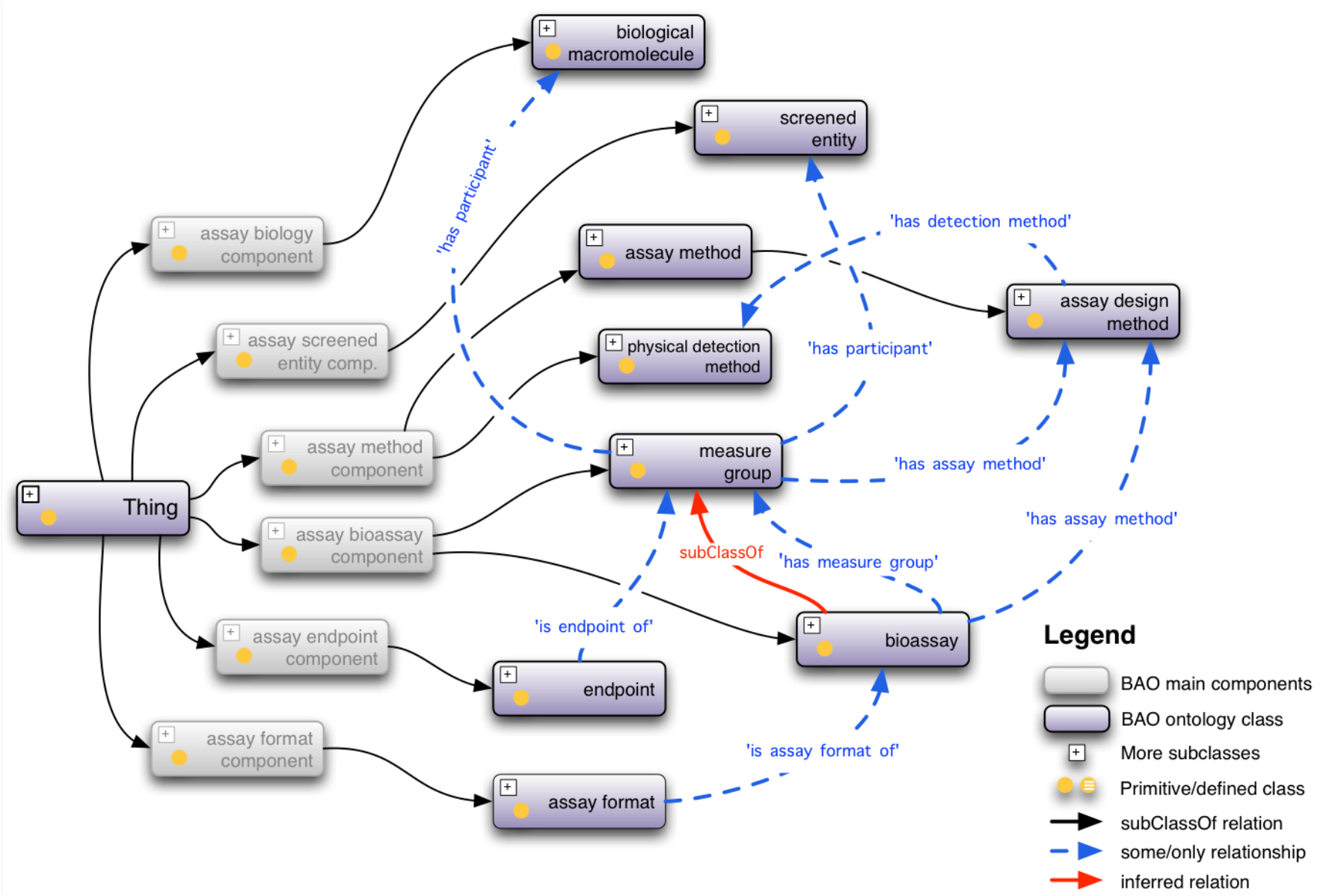
If sufficiently annotated, compound behavior may be correlated to assay and biology characteristics

Two Questions

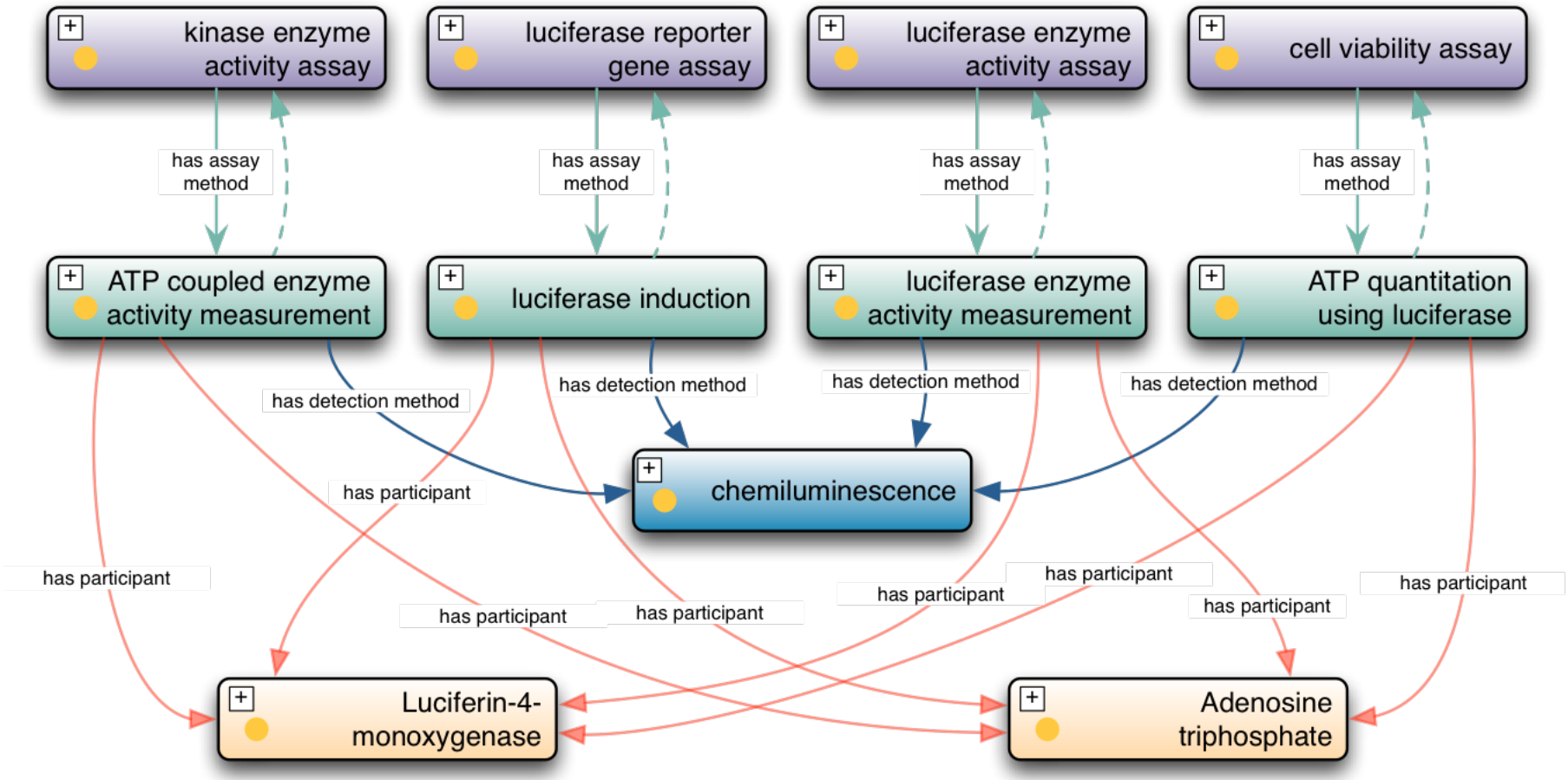
How likely are compounds, associated with a given annotation, identified as active?

Given a new set of compounds, what sets of assay conditions (as implied by the annotations) will they be active in?

BAO 2.0


biological macromolecule
screened entity
assay biology component
assay method
assay design method
assay screened entity comp.
physical detection method
assay method component
measure group
Thing
assay bioassay component
assay endpoint component
bioassay
endpoint
assay format component
assay format
'has participant'
'has detection method'
'has participant'
'has assay method'
'has assay method'
subClassOf
'has measure group'
'is endpoint of'
'is assay format of'
Legend
BAO main components
BAO ontology class
More subclasses
Primitive/defined class
subClassOf relation
some/only relationship
inferred relation

Assay Modeling



Prior Work

- BAO annotated datasets
  - de Souza et al, 2014; Vempati et al, 2012
- Analyzing HTS datasets using BAO
  - Zander-Balderud et al, 2015; Schürer et al, 2011
- Semi-automated annotation of assay descriptions using the BAO
  - Clark et al, 2014

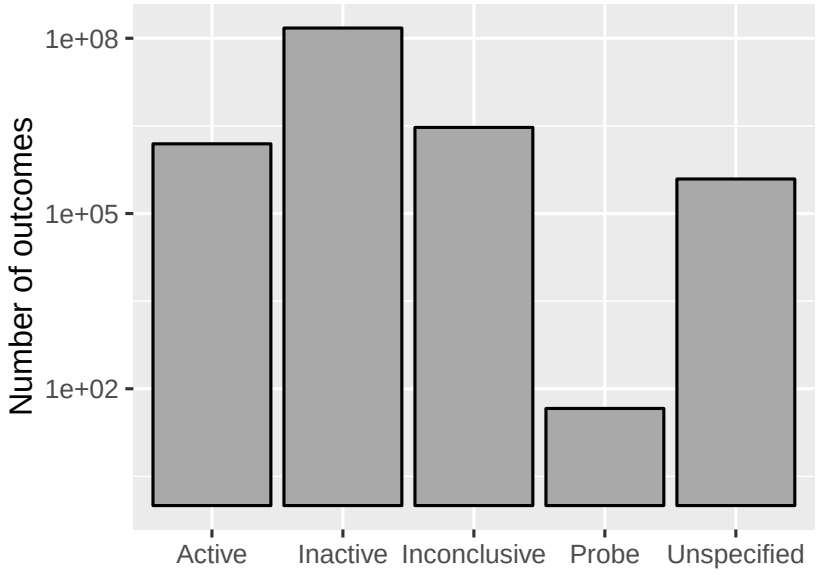
Workflow

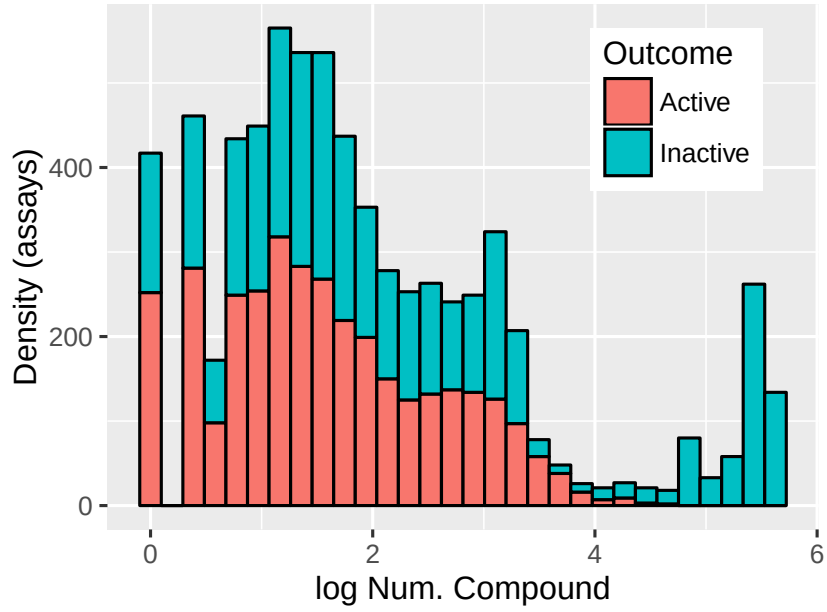
- Extract unique BAO terms and for each term identify annotated assays
- Extract active compounds from this set of assays
- Compute fingerprint bit distribution
- Use these conditional bit distributions to identify the BAO terms that describe the assay that they are likely to be active in

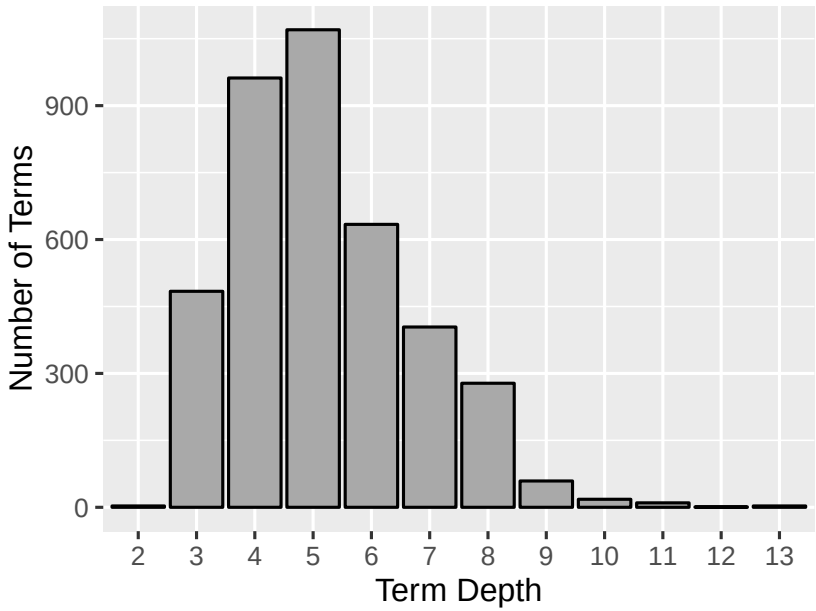
Dataset Overview

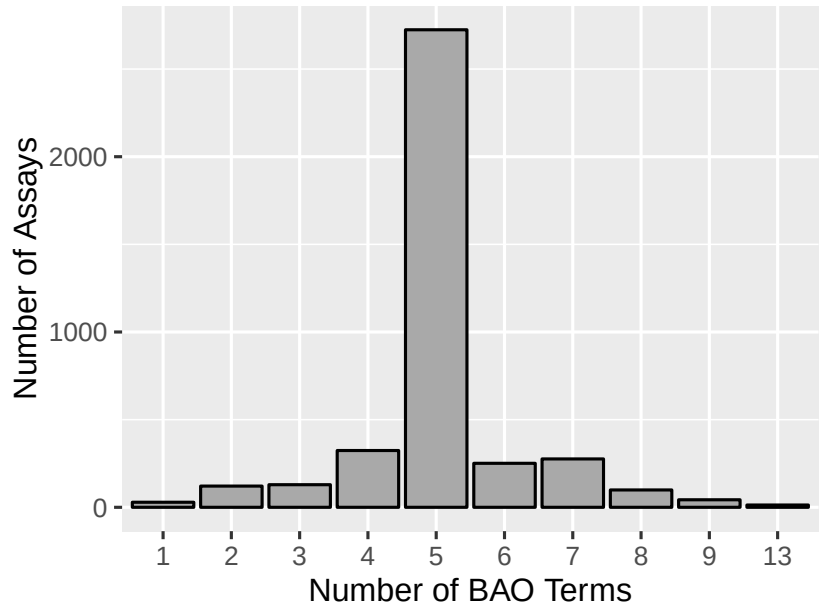
- Extracted 4010 Pubchem AIDs from BARD
- Primary, confirmation, counterscreening assays
- 154M outcomes
- 740K compounds
- Pubchem 881-bit keys using CDK and NCGC implementations
- 192 unique BAO terms

Dataset Overview



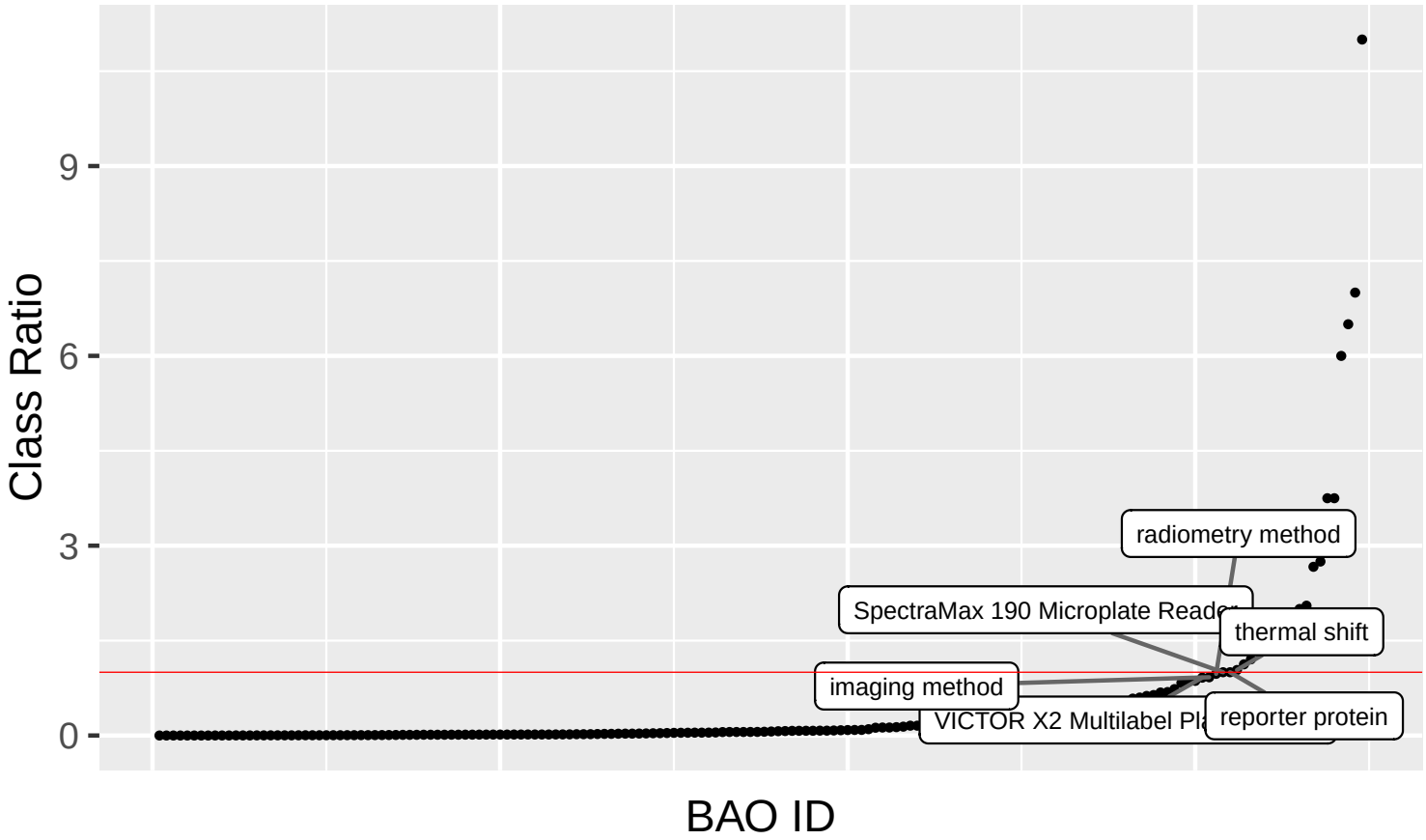






Class Imbalances

Imbalanced classes are problematic, and some of the terms with near-balanced classes are not very specific (e.g., `imaging method`)



Problem formulation

For a given library of compounds X, we would like to calculate a ranked list *relevant* T of BAO terms that are most likely associated with X. Let $\mathbf{x} \in X$ and t is a BAO term. The list T is an ordered list based on the following:

$$\underset{i}{\text{argmax}} \left\{ \sum_j p(t_i|\mathbf{x}_j) \right\}, \tag{1}$$

where $p(t_i|\mathbf{x}_j)$ is the probability that BAO term t_i is associated with compound $\mathbf{x}_j$. From Bayes' rule, we have

$$p(t_i|\mathbf{x}_j) = \frac{p(\mathbf{x}_j|t_i)p(t_i)}{p(\mathbf{x}_j)} \quad \text{or} \quad p(t_i|\mathbf{x}_j) \propto p(\mathbf{x}_j|t_i)p(t_i).$$

Given that BAO terms are annotated at the assay level, we instead have

$$p(t_i|\mathbf{x}_j) \propto p(t_i) \sum_k p(\mathbf{x}_j|a_k)p(a_k|t_i), \tag{2}$$

where a_k is a BAO annotated assay.

A Bayesian Approach for Ranking

Note that $p(\mathbf{x}_j|a_k)$ is the sampling function specified over only *active* compounds in assay a_k. In our model, $\mathbf{x}_j$ is defined as independent Bernoulli distribution with parameter θ, i.e.,

$$p(\mathbf{x}_j|a_k) = \prod_i \theta_i^{x_{ji}} (1-\theta_i)^{1-x_{ji}},$$

where $x_{ji} \in \{0,1\}$ is the i-th bit of the PubChem substructural fingerprint.

Learning BAO terms for a library of compounds amounts to estimating θ, $p(t_i)$, and $p(a_k|t_i)$.

Per-Term Activity Classifier

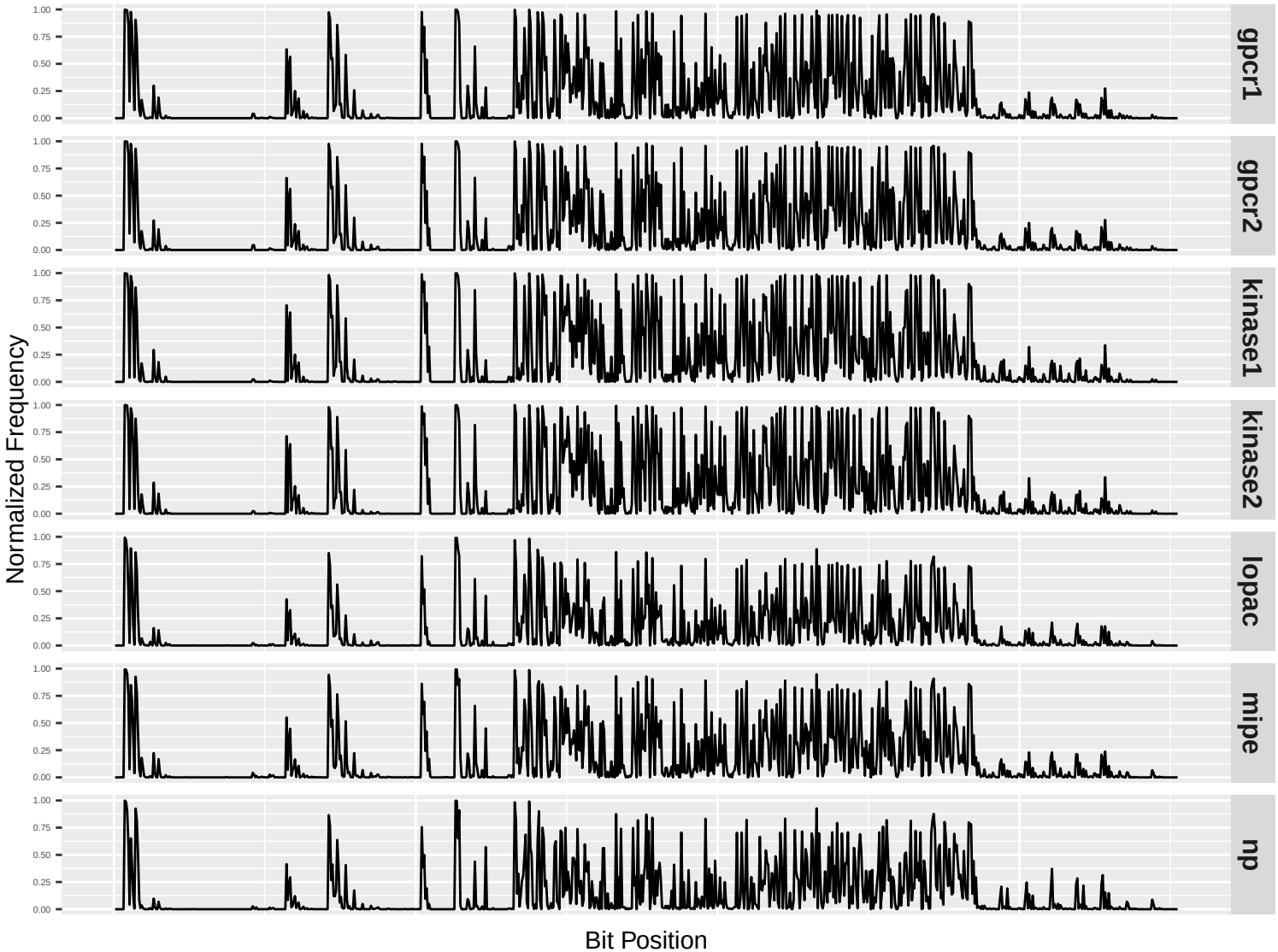
- For a given ontology term T_i, predict whether a compound will be active or not
- Model this using Naïve Bayes, where we extract set of actives and inactives from assays annotated with T_i
- Results in a set of models $\{M_1, M_2, \cdots, M_N\}$,
- For a new compound in library, obtain probability of being active for term T_i for all i and take top k terms
- Aggregate top k terms from all compounds in library
- **Represents the set of ontology terms defining an assay in which these compounds would likely be active**

Test Libraries

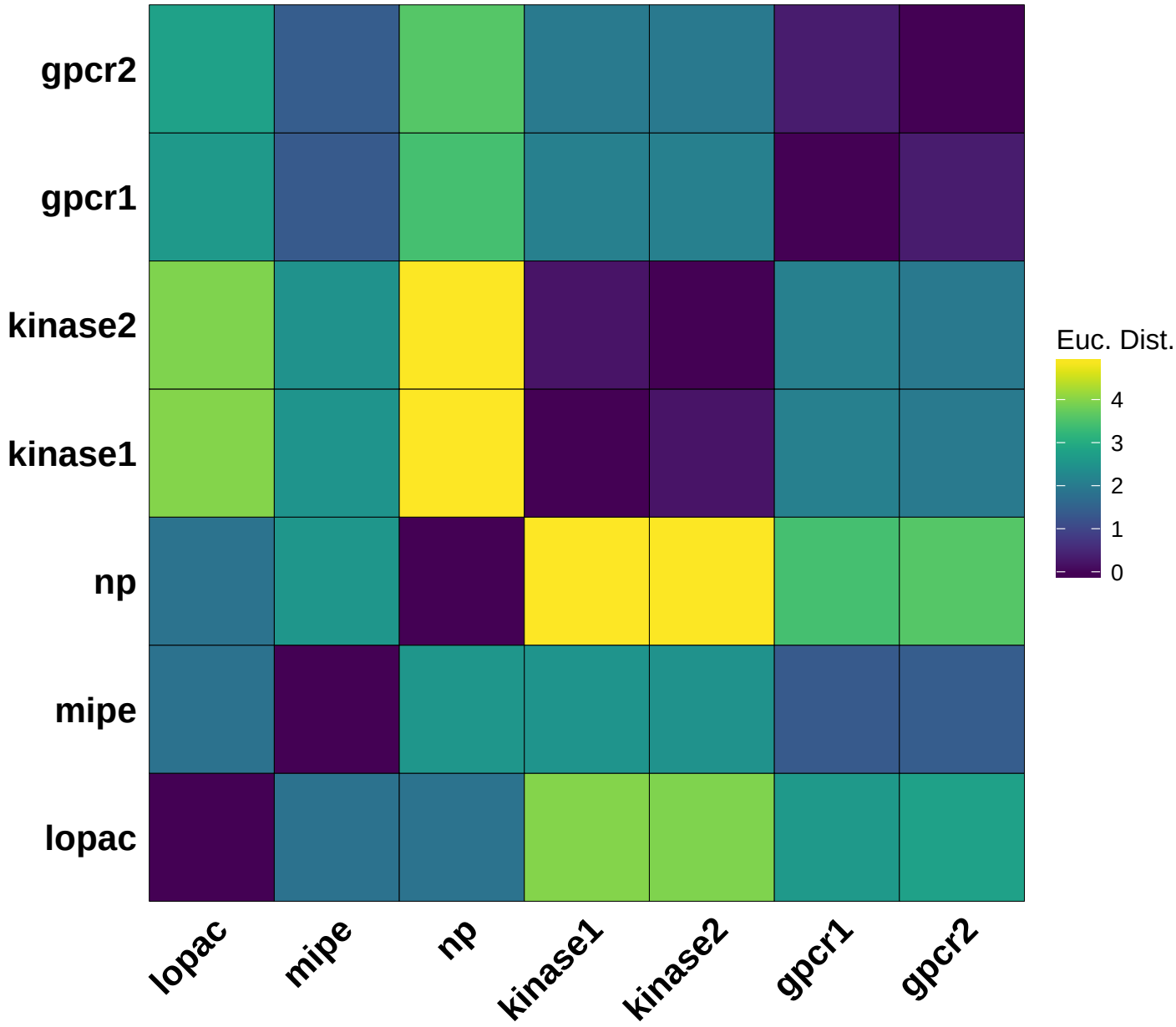
- Considered several libraries to test out the approach
- MIPE (1912 compounds) - Approved, investigational drugs, constructed for functional diversity
- LOPAC (1280 compounds) - Diverse library, designed for enrichment of bioactivity
- Natural Products (5000 compounds)
- 1000 member subset of ChEMBL GPCR collection
- 1000 member subset of ChEMBL Kinase collection

Test Libraries

In the Pubchem fingerprint space, the libraries are not very different



Test Libraries - Distance Matrix



Prediction Workflow

Bayesian Ranking

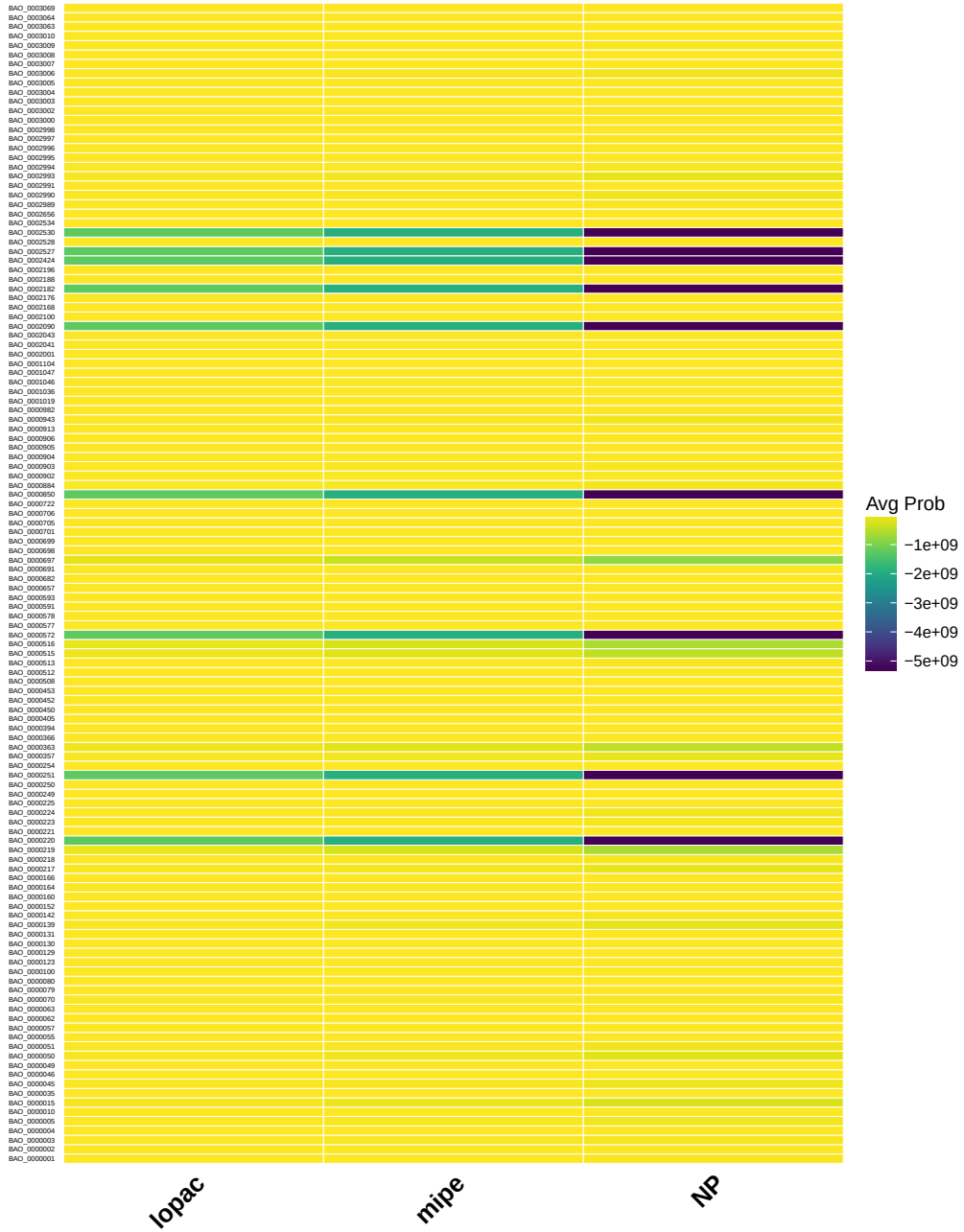
- Compute liklihood of all terms for each compound
- Aggregate across library (mean liklihood) and take top k

Activity Models

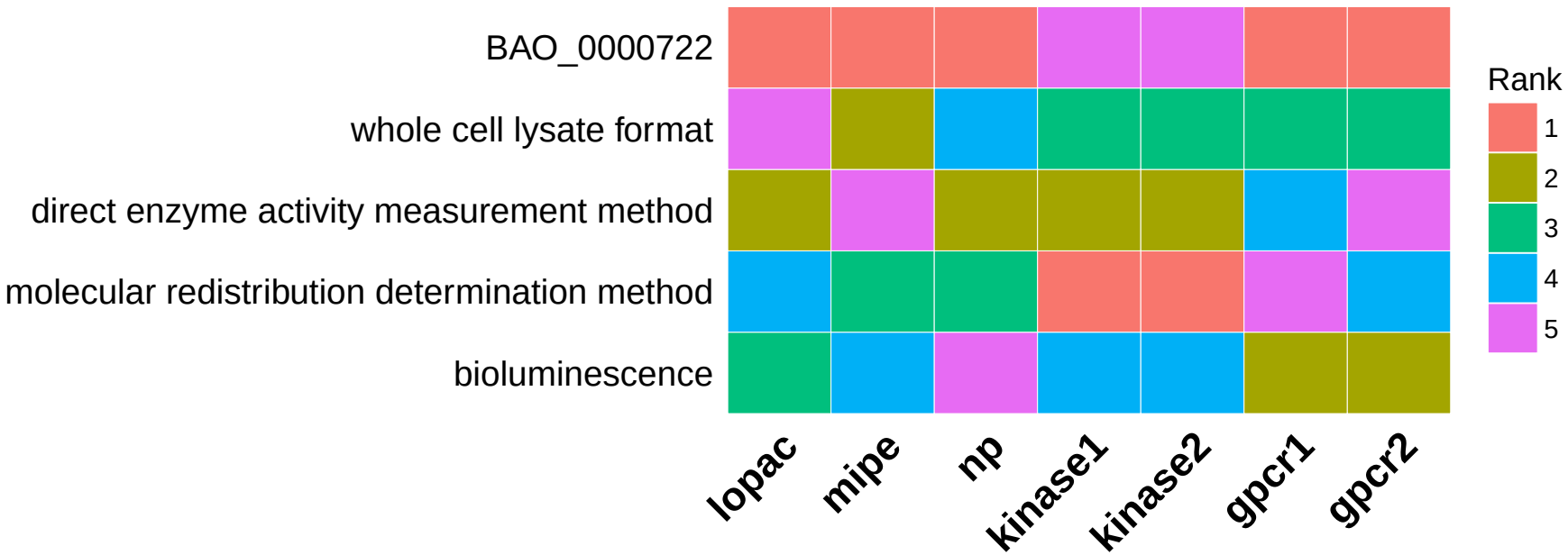
- For molecules predicted active, collect corresponding terms
- Retain the top k most frequent terms across the library

We take the top k terms as the set of annotations describing an assay in which the library will show activity in

Result - Bayesian Ranking



Result - Per Term Activity Classifier



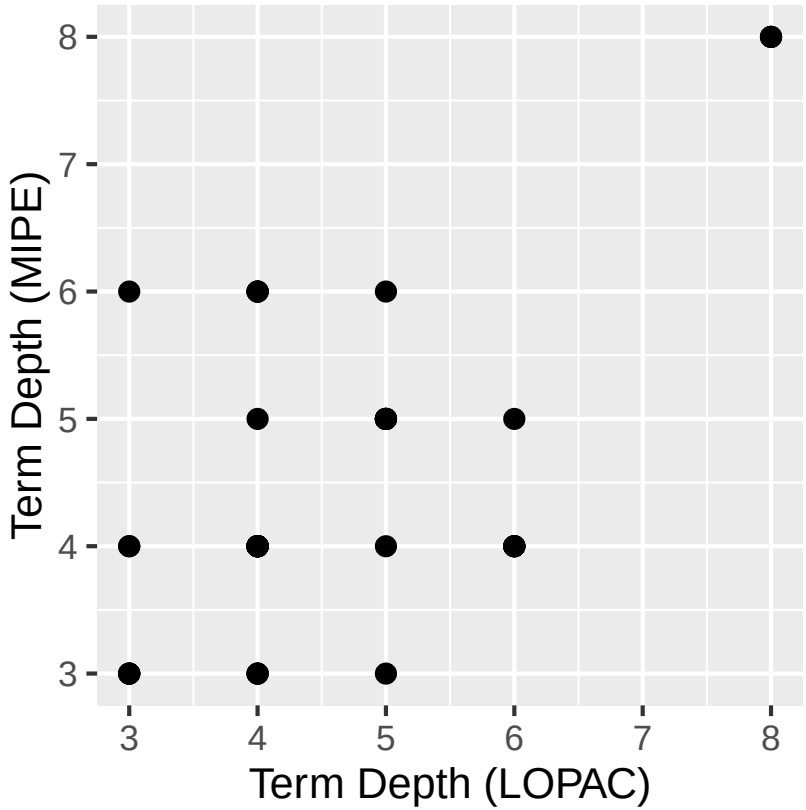
What's Different Between Libraries?

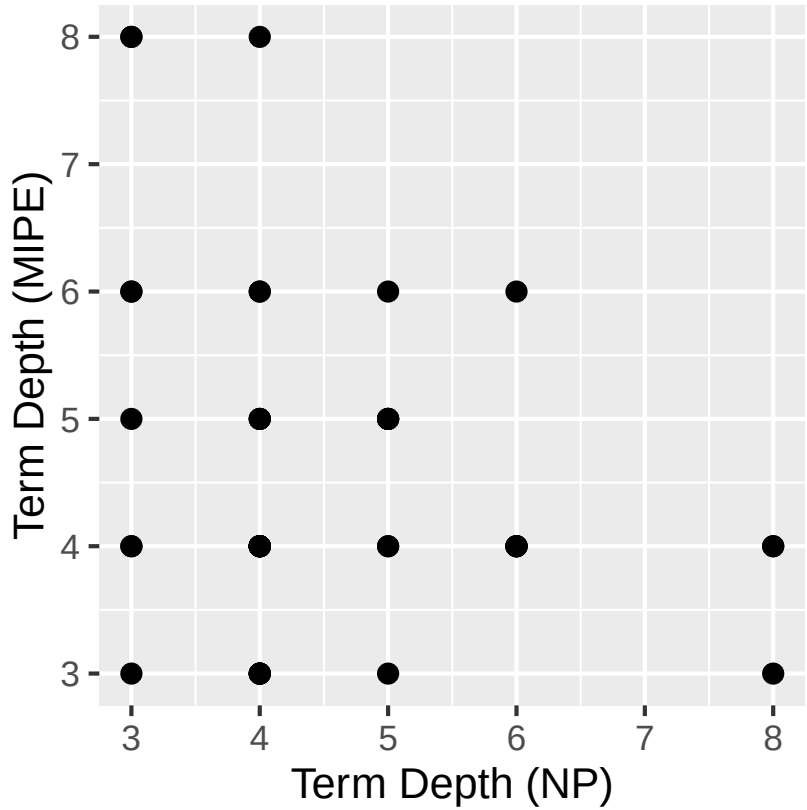
lopac	mipe
cell morphology assay	protein-turnover assay
protein-turnover assay	cell morphology assay
TopCount NXT Microplate Scintillation Luminescence	standard deviation
confocal microscopy	TopCount NXT Microplate Scintillation Luminescence C
8453 UV-Visible Spectrophotometer	confocal microscopy
standard deviation	8453 UV-Visible Spectrophotometer
measured entity	drug abuse assay
drug abuse assay	drug interaction assay
drug interaction assay	measured entity
FlexStation II Microplate Reader	optical based
optical based	FlexStation II Microplate Reader
patch clamp	cell membrane format
cell membrane format	assay footprint
assay footprint	SpectraMax 190 Microplate Reader
tissue-based format	patch clamp
SpectraMax 190 Microplate Reader	tissue-based format
protein folding assay	functional
functional	cytokine secretion assay
cytokine secretion assay	protein folding assay
dehydrogenase activity determination	time resolved fluorescence resonance energy transfer
time resolved fluorescence resonance energy transfer	phosphorylation assay
phosphorylation assay	dehydrogenase activity determination
positive control	negative control
negative control	positive control
Opera QEHS	luciferase induction
luciferase induction	Opera QEHS
enzyme activity assay	cell viability assay
cell viability assay	enzyme activity assay
chemiluminescence	reporter gene assay
reporter gene assay	chemiluminescence

What's Different Between Libraries?

NP	mipe
nuclear magnetic resonance	cell-free format
cell-free format	nuclear magnetic resonance
cell morphology assay	protein-turnover assay
8453 UV-Visible Spectrophotometer	cell morphology assay
TopCount NXT Microplate Scintillation Luminescence Counter	target
target	standard deviation
confocal microscopy	TopCount NXT Microplate Scintillation Luminescence Counter
protein-turnover assay	confocal microscopy
cuvette	8453 UV-Visible Spectrophotometer
standard deviation	cuvette
measured entity	imaging method
imaging method	drug interaction assay
drug interaction assay	measured entity
immunoassay	membrane potential assay
membrane potential assay	immunoassay
reporter gene	purified
purified	reporter gene
patch clamp	FDSS7000
cell membrane format	optical based
FDSS7000	FlexStation II Microplate Reader
FlexStation II Microplate Reader	cell membrane format
optical based	assay footprint
tissue-based format	SpectraMax 190 Microplate Reader
assay footprint	patch clamp
SpectraMax 190 Microplate Reader	tissue-based format
protein folding assay	functional
functional	protein folding assay
Fluorometer	scintillation counting
scintillation counting	Fluorometer
protein-nucleotide interaction assay	EnVision Multilabel Reader
EnVision Multilabel Reader	protein-nucleotide interaction assay
dehydrogenase activity determination	time resolved fluorescence resonance energy transfer
time resolved fluorescence resonance energy transfer	phosphorylation assay
phosphorylation assay	dehydrogenase activity determination
positive control	negative control
negative control	positive control
binding assessment method	positively correlated
positively correlated	localization assay
localization assay	binding assessment method

Term Depth for the 'Differential' Terms





Pitfalls

If sufficiently annotated, compound behavior may be correlated to assay and biology characteristics

- A very abstract, possibly lossy, view of the effect of compounds on biology
- Depends on correct and meaningful annotations
- Annotations terms are context dependent, but this may not be considered when annotating a dataset
- BAO terms exhibit hierarchical relationships and ignoring them is simplistic

Acknowledgements

- Qiong Cheng (U. Miami)
- Stephan Schürer (U. Miami)

Source code and slides