CDD BioAssay Express

Bringing bioassay protocols to the world of informatics, using semantic annotations

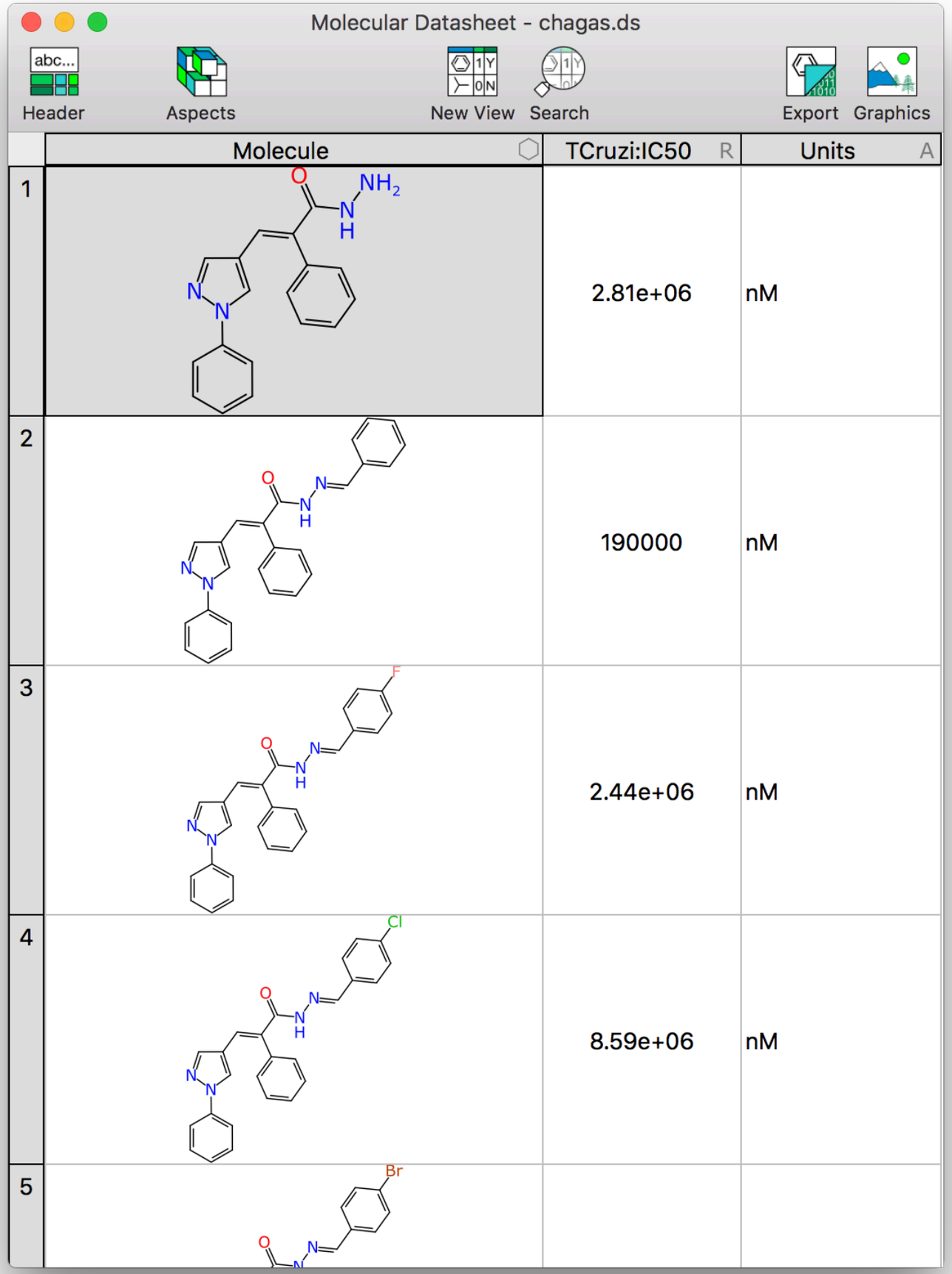
Alex M. Clark

CDD VAULT®
Complexity Simplified

SAR

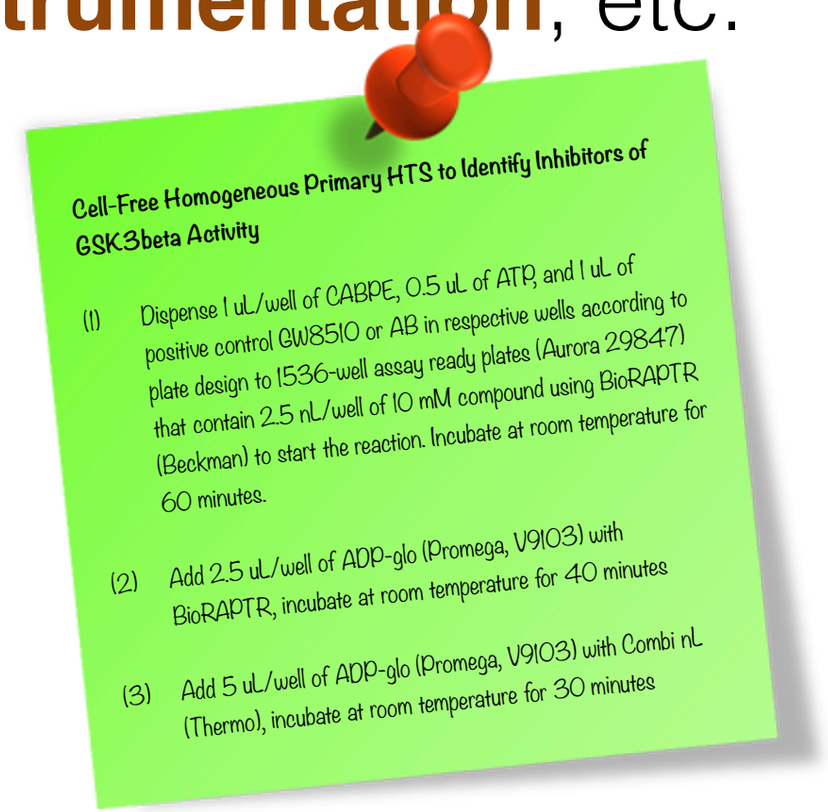
- The **activity column**:
  - `TCruzi` (chagas)
  - `IC50`
  - `nM`

- Underdefined in a typical SDfile or table
- Only alternative is a verbose publication



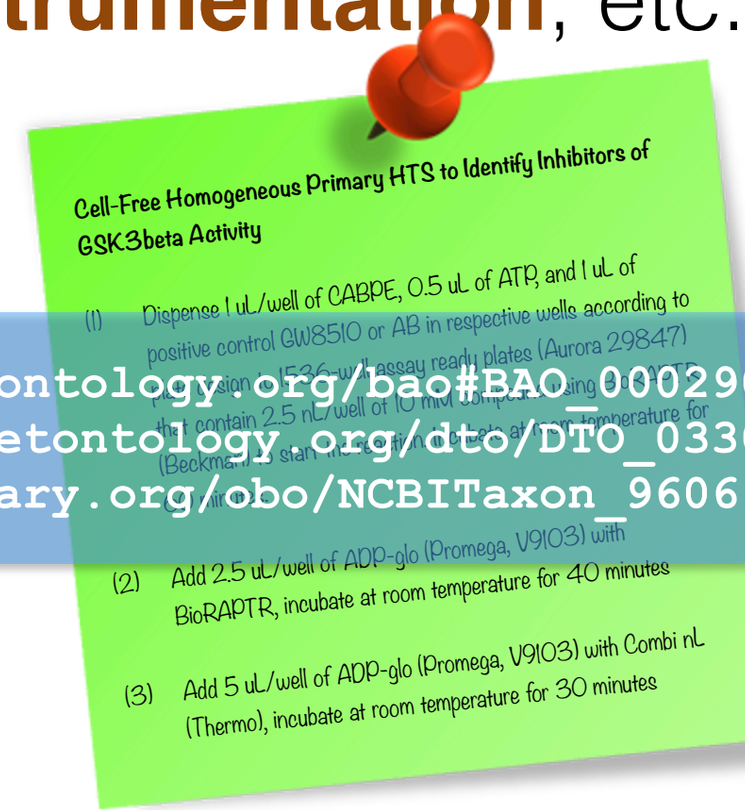
Assay Informatics

- Measurement of an activity has many details:
  - **target**, **organism**, **cell line**, **measurement type**, **units**, **experimental design**, **instrumentation**, etc.
- Usually only available as **text**...
- ... needs to be **informatics**.
- Searching for experiments, comparing datasets, spotting trends, evaluating completeness



Assay Informatics

- Measurement of an activity has many details:
  - **target**, **organism**, **cell line**, **measurement type**, **units**, **experimental design**, **instrumentation**, etc.
- Usually only available as **text**...




```
http://www.bioassayontology.org/bao#BAO_0002909
http://www.drugtargetontology.org/dto/DTO_03300101
http://purl.obolibrary.org/obo/NCBITaxon_9606
```

- ... needs to be **informatics**.
- Searching for experiments, comparing datasets, spotting trends, evaluating completeness

The Mission

- To make the world's bioassay protocol data **machine readable**
- Much progress toward making this goal an actual reality...
- With the help of ontology markup:
  - BAO, DTO, CLO, GO, UO, UniProt...

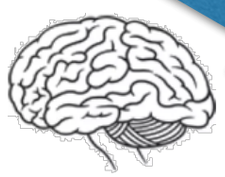
Ontologies

BAO, DTO, CLO, UO, UniProt

Legacy Data

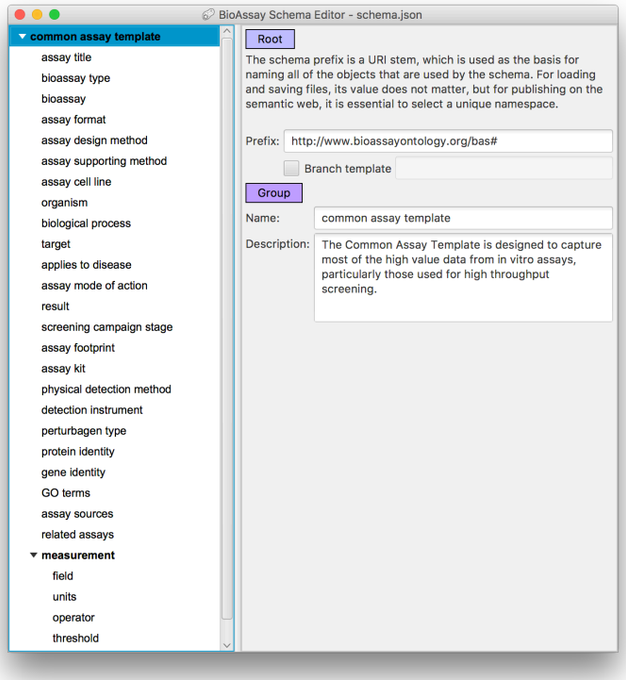
PubMed
PubChem
MLPCN
Tox21
ChEMBL

ML

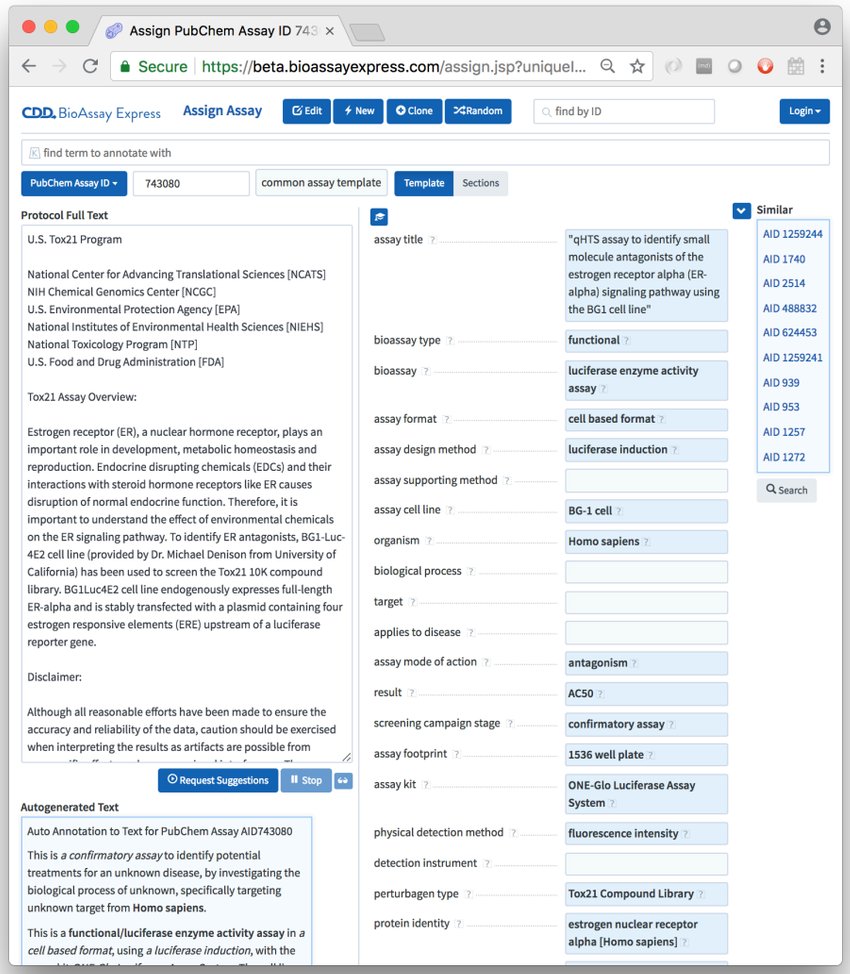


New Data

ELN



Common Assay Template



Fully Annotated Assays

Common Assay Template

- A recipe for how to use ontologies
- Capture most important information for most assays

assay title	"GSK3beta Surface Plasmon Resonance (SPR) Assay Measured in Biochemical System Using Microfluidics - 2133-08_Inhibitor_Dose_DryPowder_Activity"	assay footprint	Sensor chip CM5 384 well plate
bioassay type	binding type	assay kit	
bioassay	binding assay	physical detection method	
assay format	biochemical format	detection instrument	Biacore 3000 SPR System (GE Healthcare)
assay design method	binding assessment method	perturbagen type	compound library
assay supporting method	single perturbagen	protein identity	glycogen synthase kinase-3 beta isoform 1 [Homo sapiens]
assay cell line	SF-21 cell	gene identity	GSK3B: glycogen synthase kinase 3 beta
organism	Homo sapiens	GO terms	
biological process		assay sources	Broad Institute (Harvard-MIT)
target	Kinase GSK3B	related assays	AID 2119
applies to disease		measurement	
assay mode of action	modulation inhibition	field	KD_(nM)
result		units	nanomolar
screening campaign stage		operator	
		threshold	

Legacy Curation

- Text-centric data: millions of assays...
  - private collections (**ELNs**)
  - published literature (**PubMed**)
  - assay databases (**PubChem**)
- Most accessible sources: **MLPCN**, **Tox21**
- Data comes in as plain text + several fields + SAR

Hybrid machine learning

1. Text extraction
2. Probabilistic:
   - ▷ natural language models
   - ▷ correlation models
3. Axiomatic rules

◈ Each of these *accelerates* the human curator

Text extraction

- Optimised for low false positives

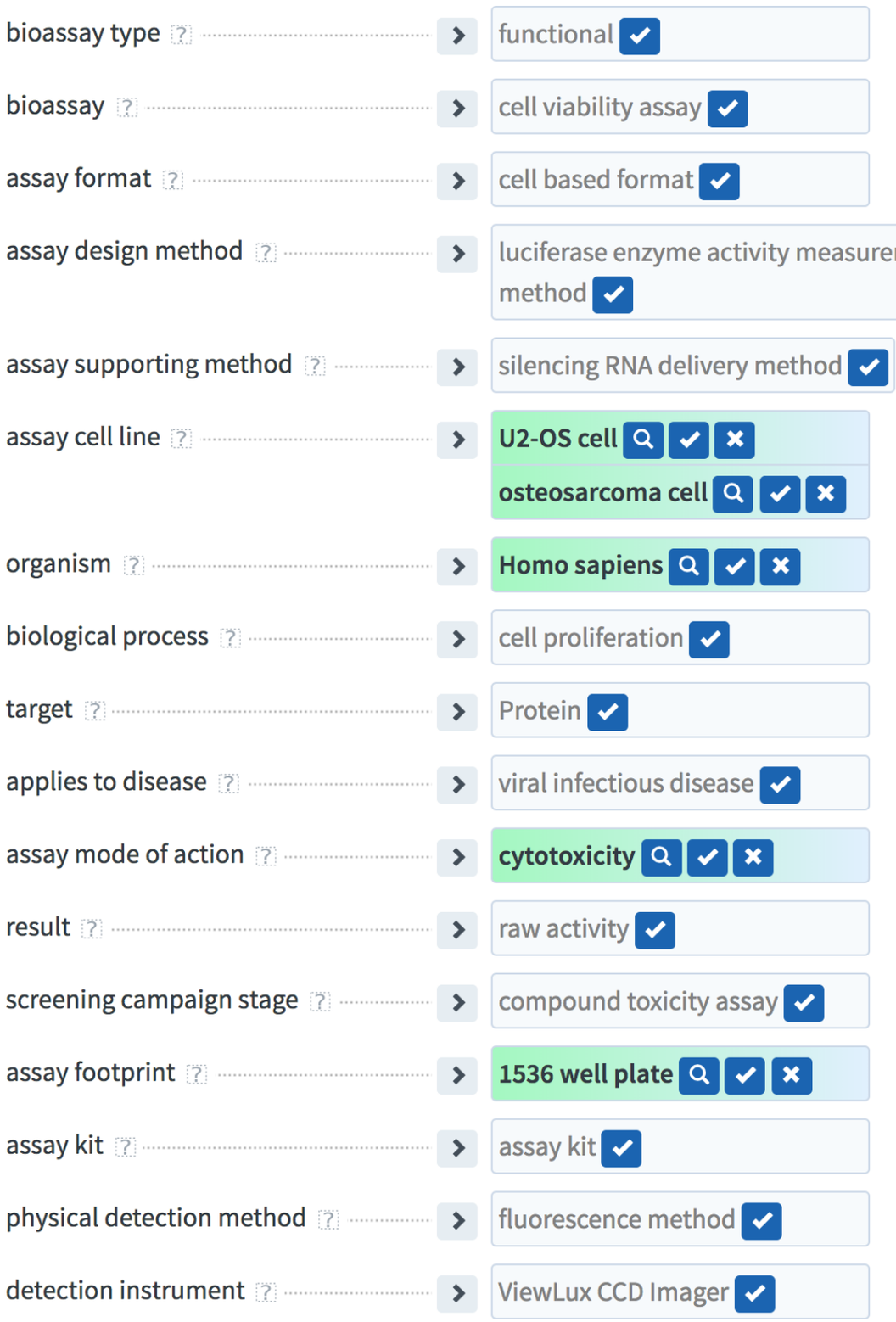
qHTS Assay for Small Molecule Inhibitors of the **Human** hERG Channel Activity

NIH Chemical Genomics Center [NCGC]
National Institutes of Environmental Health Sciences [NIEHS]
National Toxicology Program [NTP]

NCGC Assay Overview:

We have developed a **1536-well** cell-based assay for quantitative high throughput screening (qHTS) to determine in vitro hERG channel blockage as a measure of cardio **toxicity** with small molecules. This particular assay uses the **U2OS** cell line which is derived from human **osteosarcoma**.

...



Text extraction

- Optimised for low false positives

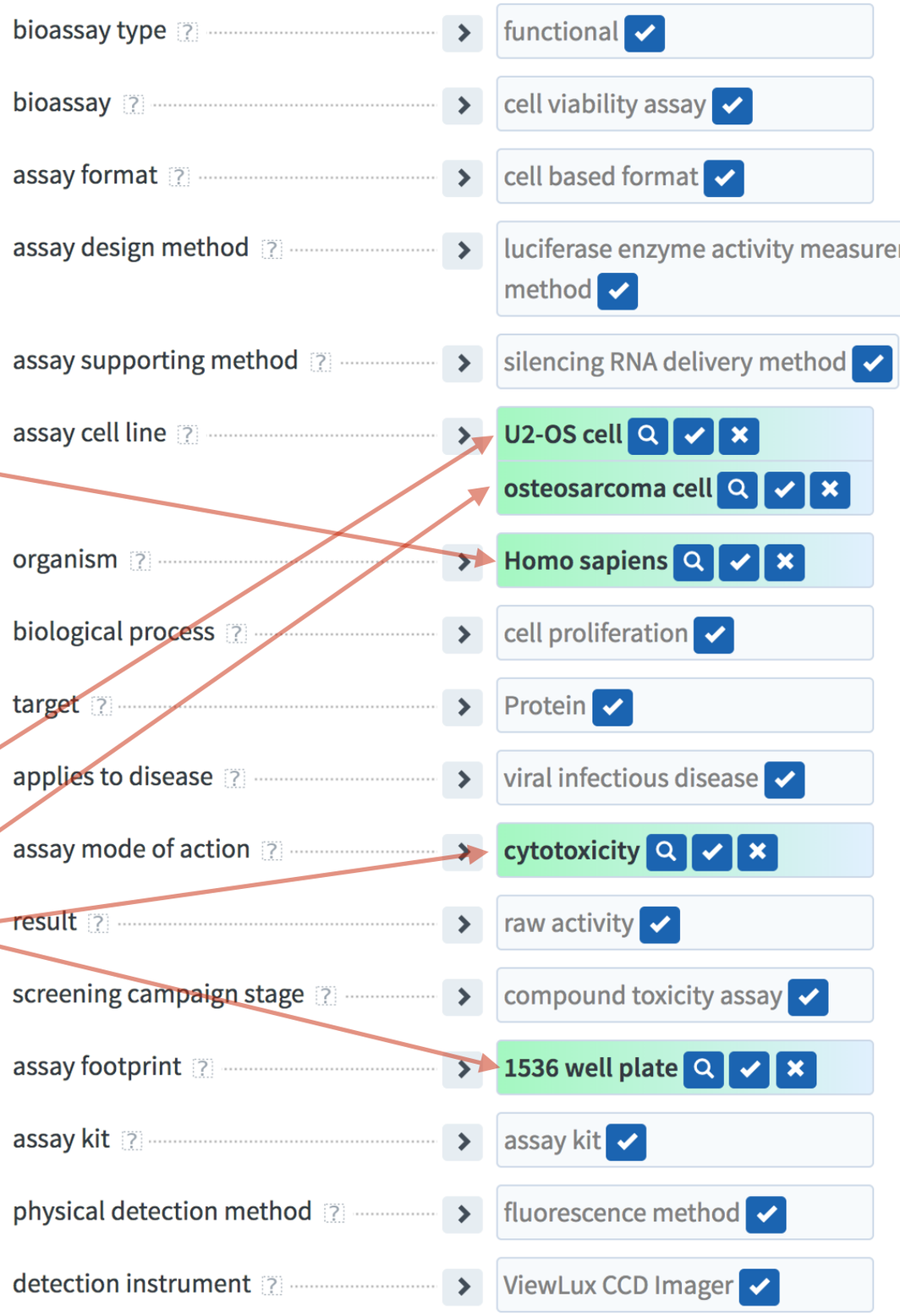
qHTS Assay for Small Molecule Inhibitors of the **Human** hERG Channel Activity

NIH Chemical Genomics Center [NCGC]
National Institutes of Environmental Health Sciences [NIEHS]
National Toxicology Program [NTP]

NCGC Assay Overview:

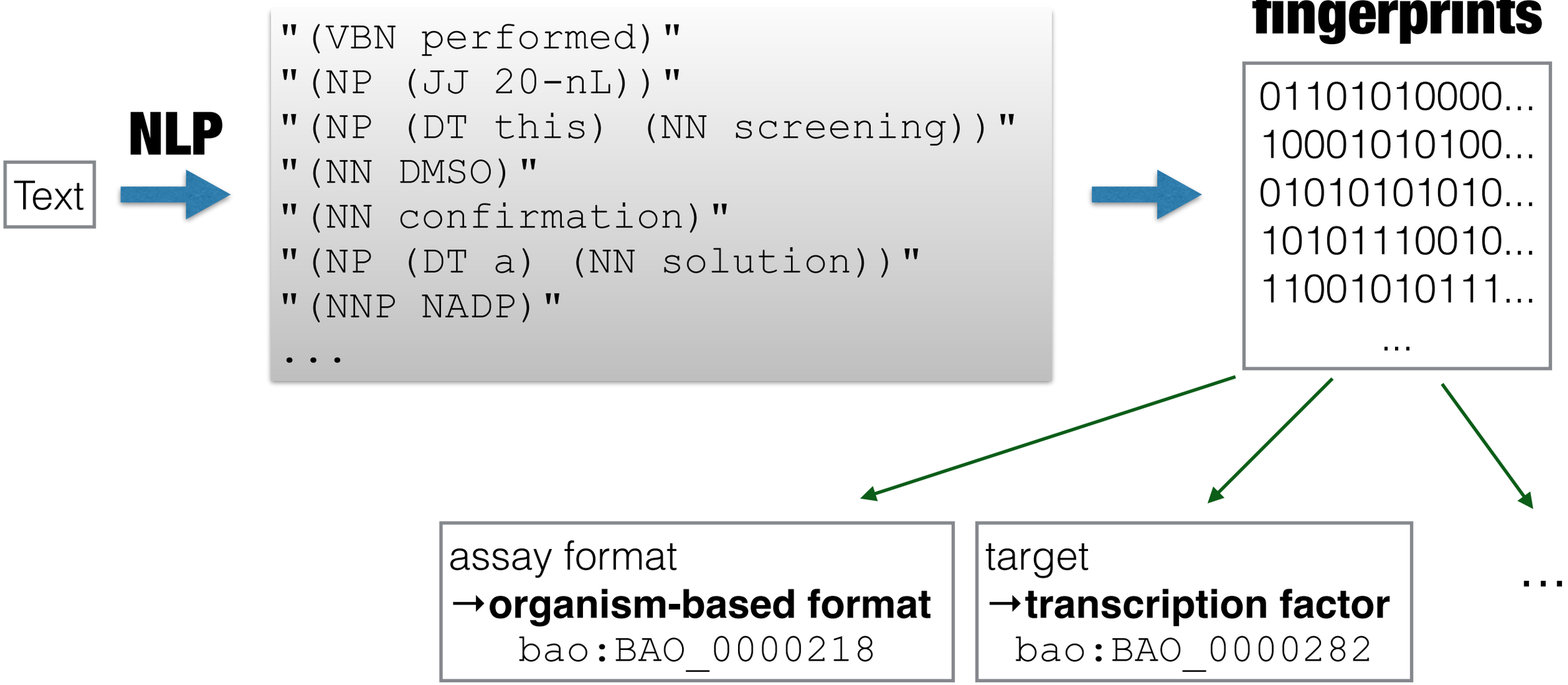
We have developed a **1536-well** cell-based assay for quantitative high throughput screening (qHTS) to determine in vitro hERG channel blockage as a measure of cardio **toxicity** with small molecules. This particular assay uses the **U2OS** cell line which is derived from human **osteosarcoma**.

...



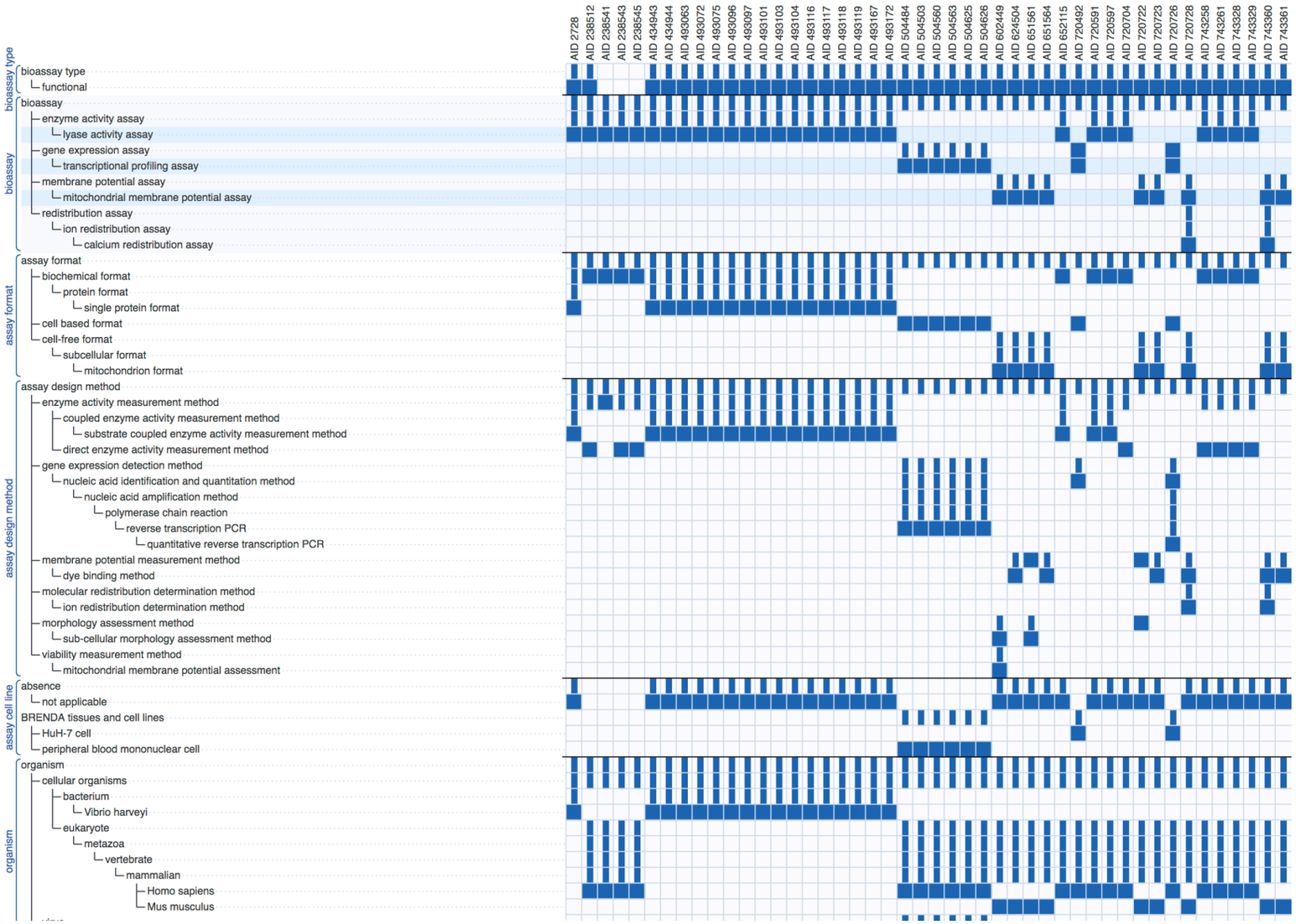
Natural language models

- Use **NLP** to create **fingerprints** for each text document
- One Bayesian model for each possible annotation term



Correlation models

- Similar principle: existing annotations are fingerprints



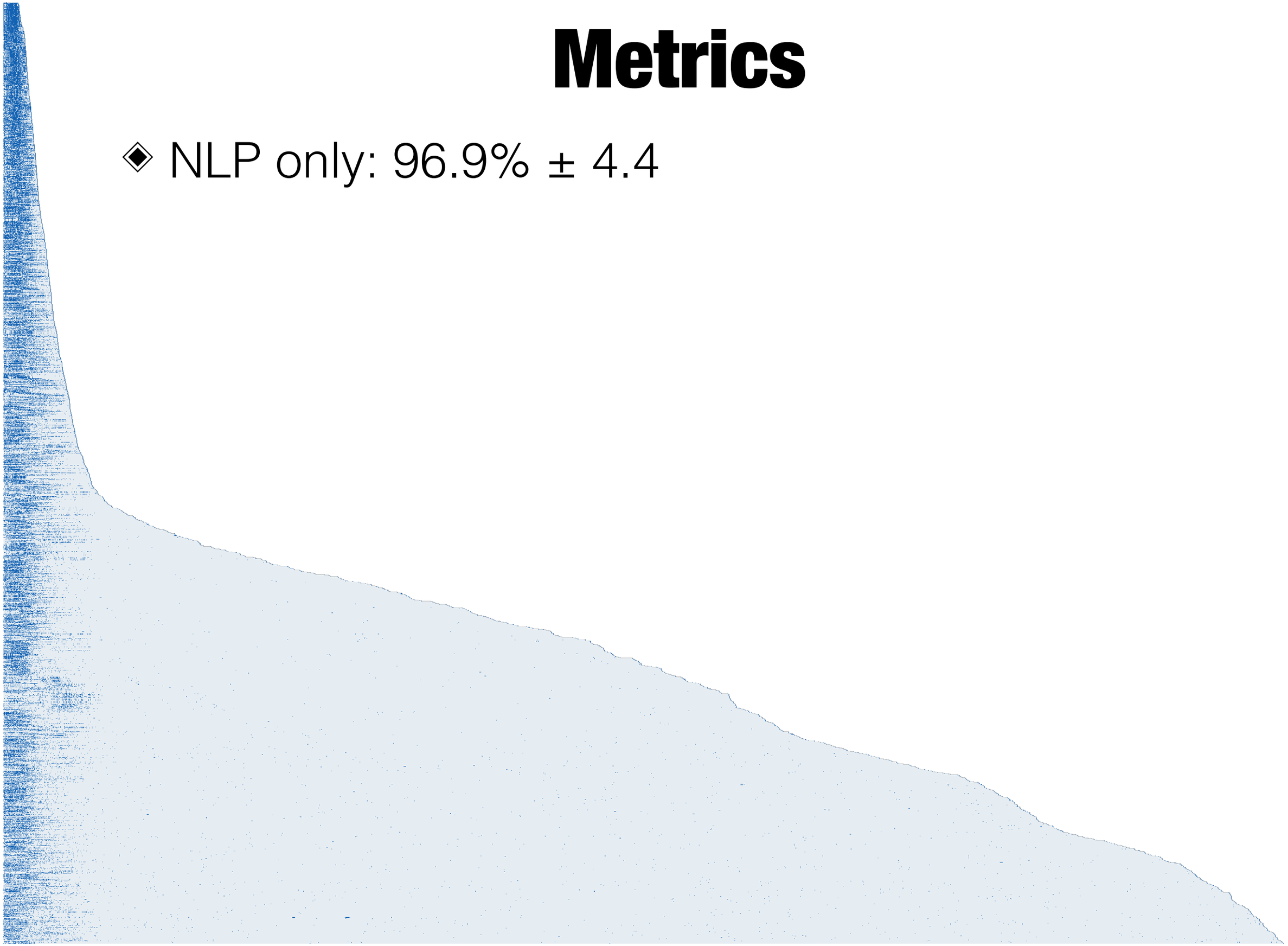
one model per term

- Combine *natural language* + *correlation* = **ranking**

Metrics

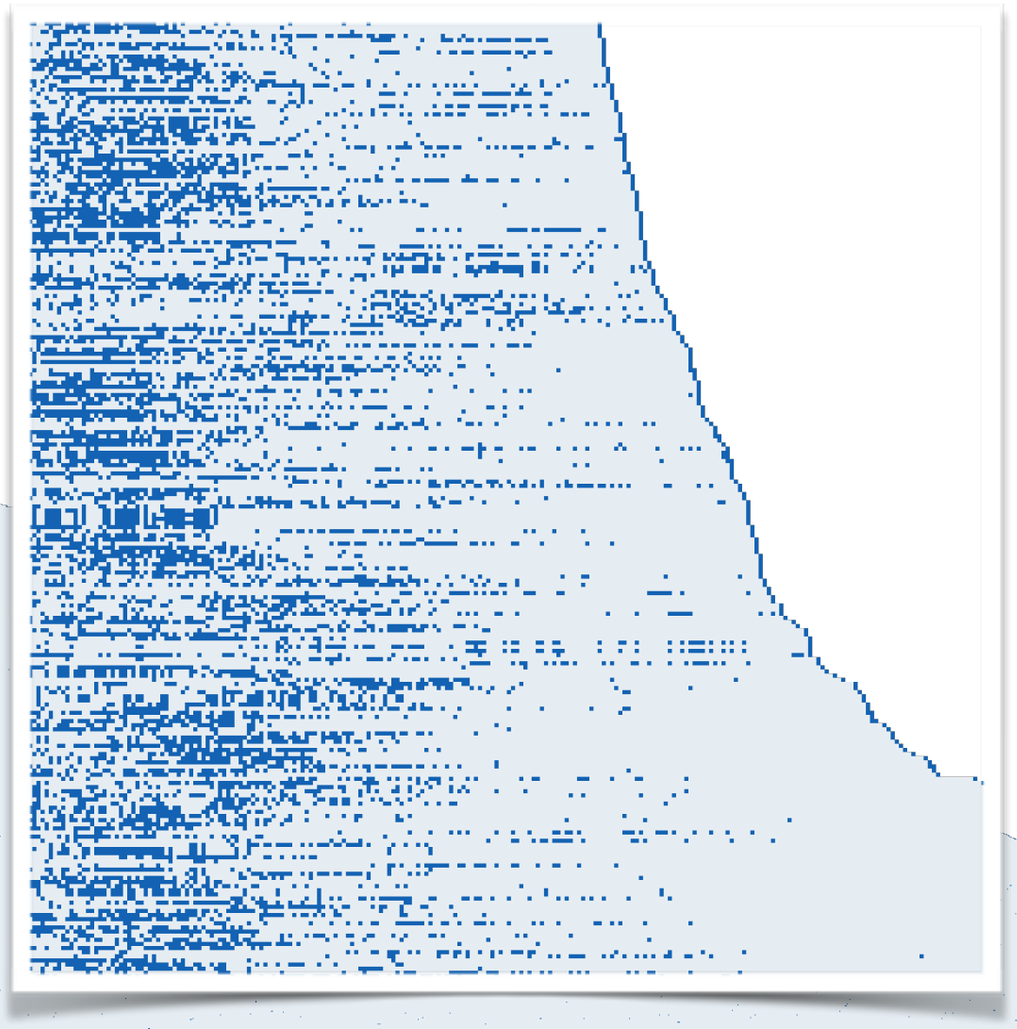
Metrics

- NLP only: 96.9% ± 4.4



Metrics

- NLP only: 96.9% ± 4.4

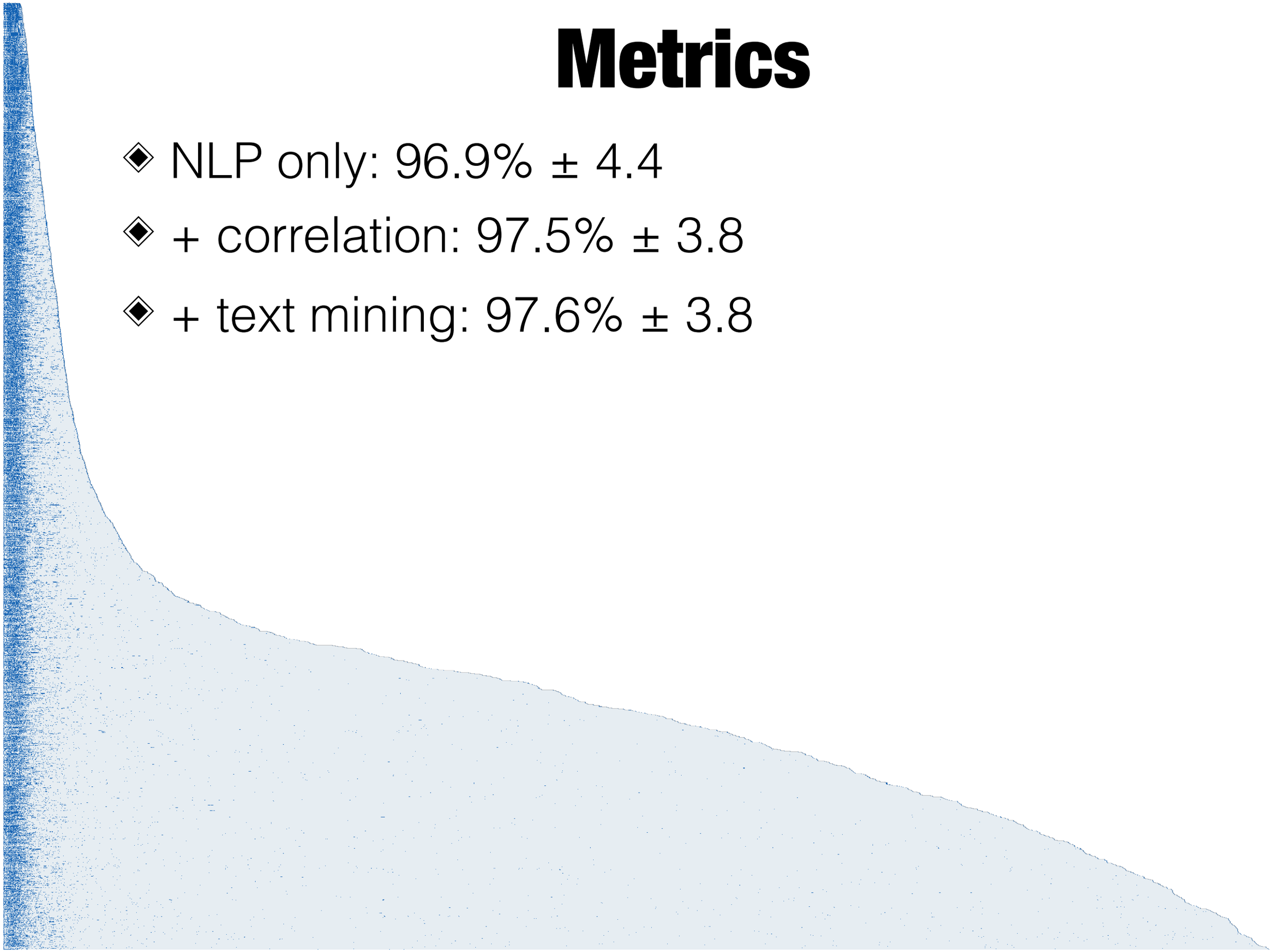


Metrics

- NLP only: 96.9% ± 4.4
- + correlation: 97.5% ± 3.8

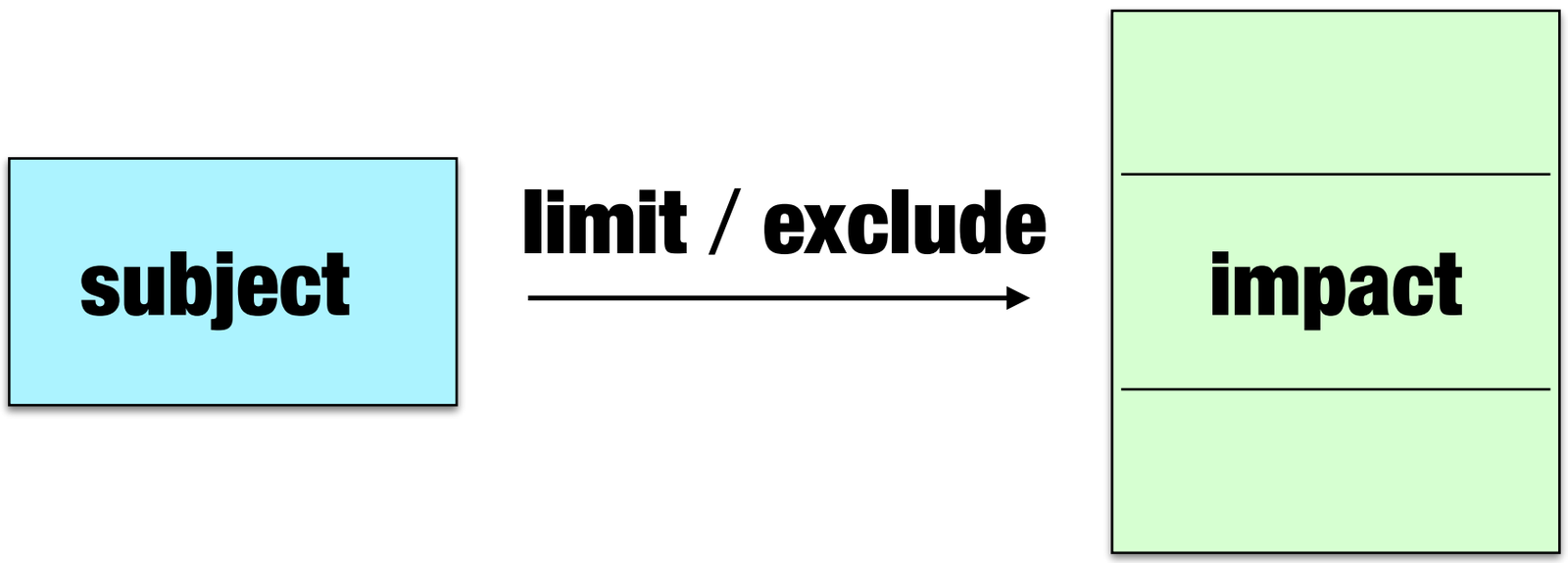
Metrics

- NLP only: 96.9% ± 4.4
- + correlation: 97.5% ± 3.8
- + text mining: 97.6% ± 3.8



Axioms

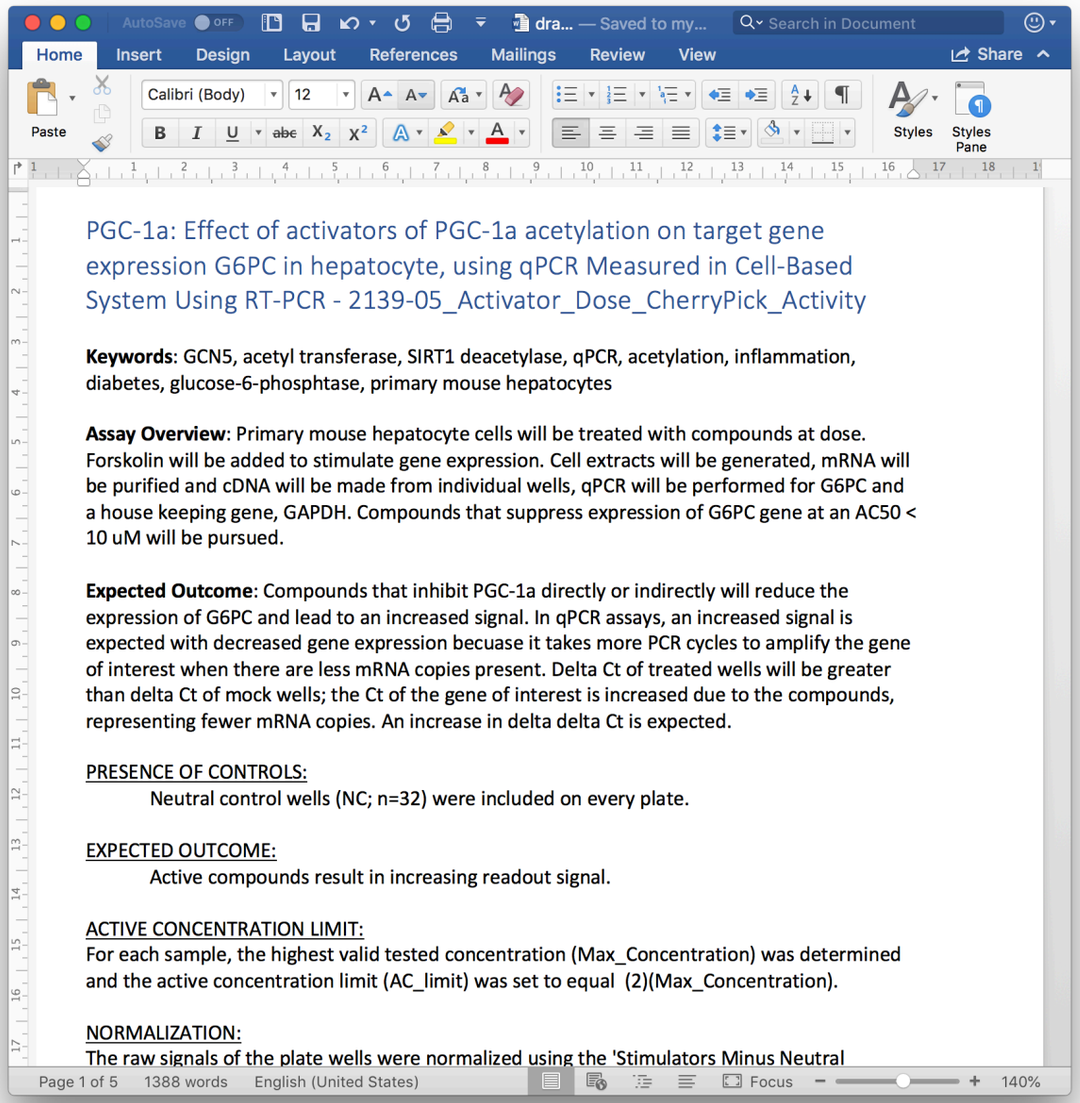
- Ontologies like BAO & DTO have *axioms*
- Convert these into **rules**:



- Use to constrain predictions at each step
- Like correlation models, except absolute

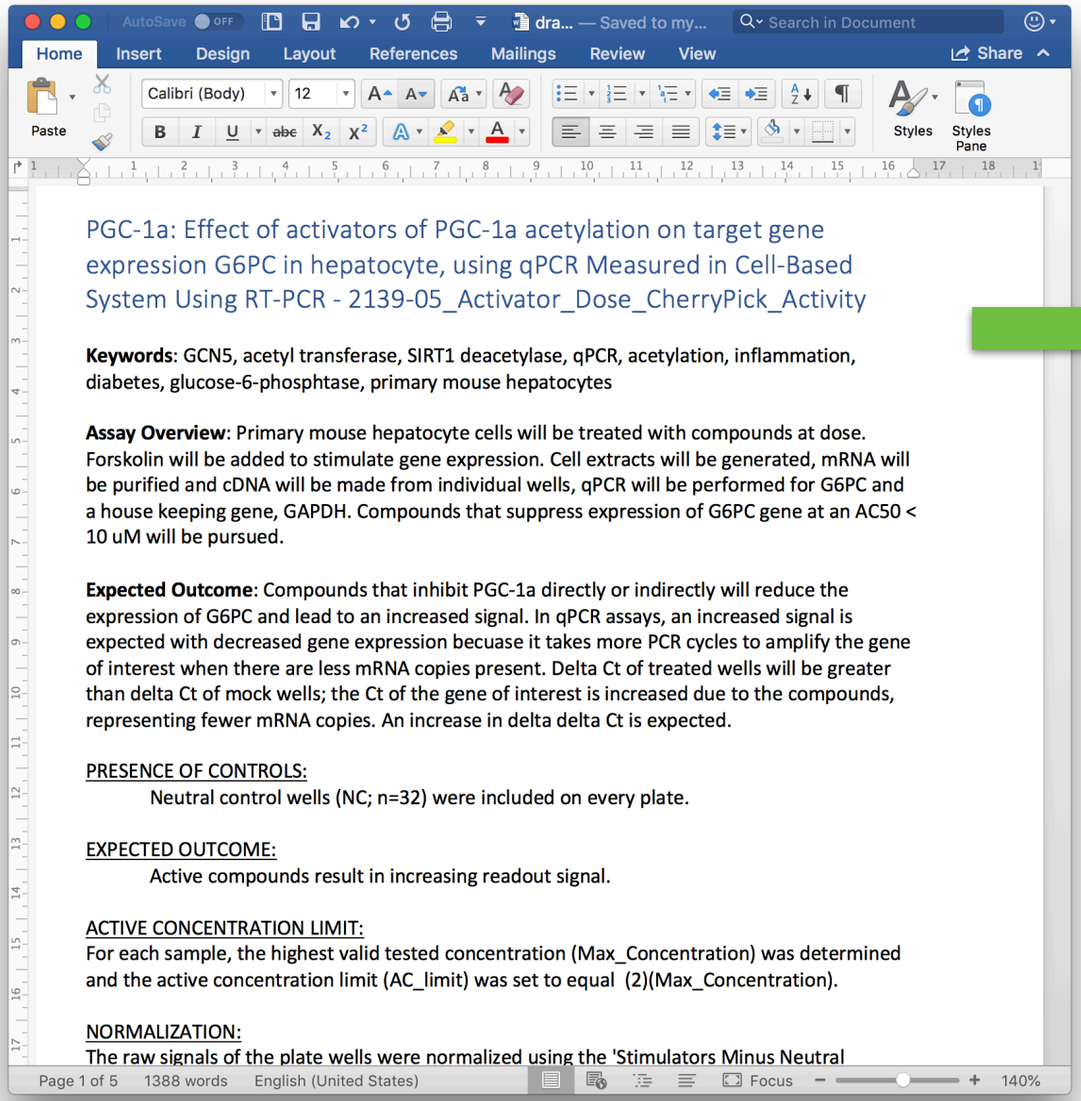
New data: ELN

- Paste text or drag a document...



New data: ELN

- Paste text or drag a document...



PGC-1a: Effect of activators of PGC-1a acetylation on target gene expression G6PC in hepatocyte, using qPCR Measured in Cell-Based System Using RT-PCR - 2139-05_Activator_Dose_CherryPick_Activity

Keywords: GCN5, acetyl transferase, SIRT1 deacetylase, qPCR, acetylation, inflammation, diabetes, glucose-6-phosphtase, primary mouse hepatocytes

Assay Overview: Primary mouse hepatocyte cells will be treated with compounds at dose. Forskolin will be added to stimulate gene expression. Cell extracts will be generated, mRNA will be purified and cDNA will be made from individual wells, qPCR will be performed for G6PC and a house keeping gene, GAPDH. Compounds that suppress expression of G6PC gene at an AC50 < 10 uM will be pursued.

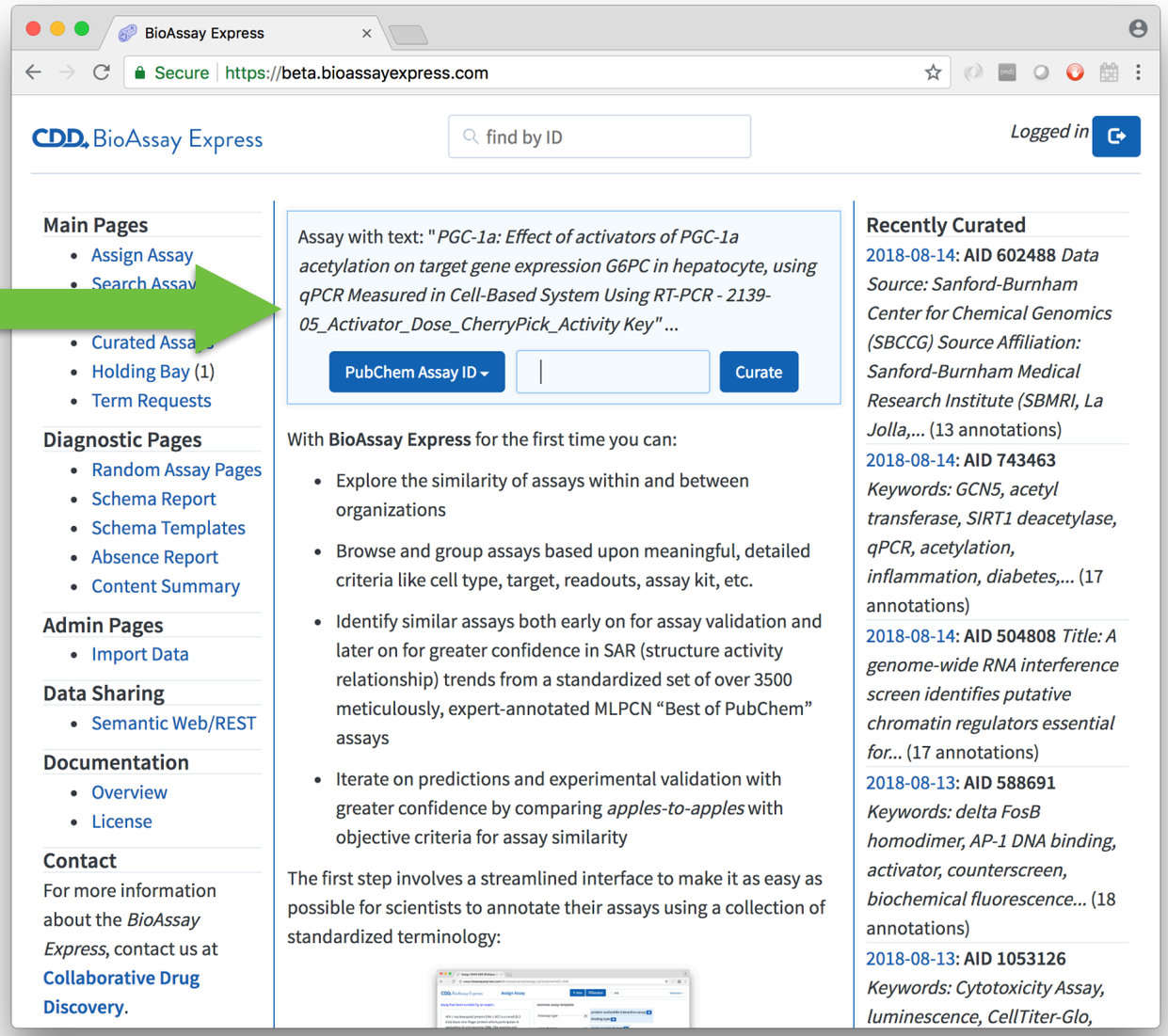
Expected Outcome: Compounds that inhibit PGC-1a directly or indirectly will reduce the expression of G6PC and lead to an increased signal. In qPCR assays, an increased signal is expected with decreased gene expression becuase it takes more PCR cycles to amplify the gene of interest when there are less mRNA copies present. Delta Ct of treated wells will be greater than delta Ct of mock wells; the Ct of the gene of interest is increased due to the compounds, representing fewer mRNA copies. An increase in delta delta Ct is expected.

PRESENCE OF CONTROLS:
Neutral control wells (NC; n=32) were included on every plate.

EXPECTED OUTCOME:
Active compounds result in increasing readout signal.

ACTIVE CONCENTRATION LIMIT:
For each sample, the highest valid tested concentration (Max_Concentration) was determined and the active concentration limit (AC_limit) was set to equal (2)(Max_Concentration).

NORMALIZATION:
The raw signals of the plate wells were normalized using the 'Stimulators Minus Neutral



BioAssay Express

Secure https://beta.bioassayexpress.com

CDD BioAssay Express

find by ID

Logged in

Main Pages
- Assign Assay
- Search Assay
- Curated Assays
- Holding Bay (1)
- Term Requests

Diagnostic Pages
- Random Assay Pages
- Schema Report
- Schema Templates
- Absence Report
- Content Summary

Admin Pages
- Import Data

Data Sharing
- Semantic Web/REST

Documentation
- Overview
- License

Contact
For more information about the *BioAssay Express*, contact us at **Collaborative Drug Discovery**.

Assay with text: "*PGC-1a: Effect of activators of PGC-1a acetylation on target gene expression G6PC in hepatocyte, using qPCR Measured in Cell-Based System Using RT-PCR - 2139-05_Activator_Dose_CherryPick_Activity Key*"...

PubChem Assay ID | Curate

With **BioAssay Express** for the first time you can:

- Explore the similarity of assays within and between organizations
- Browse and group assays based upon meaningful, detailed criteria like cell type, target, readouts, assay kit, etc.
- Identify similar assays both early on for assay validation and later on for greater confidence in SAR (structure activity relationship) trends from a standardized set of over 3500 meticulously, expert-annotated MLPCN "Best of PubChem" assays
- Iterate on predictions and experimental validation with greater confidence by comparing *apples-to-apples* with objective criteria for assay similarity

The first step involves a streamlined interface to make it as easy as possible for scientists to annotate their assays using a collection of standardized terminology:

Recently Curated

2018-08-14: **AID 602488** *Data Source: Sanford-Burnham Center for Chemical Genomics (SBCCG) Source Affiliation: Sanford-Burnham Medical Research Institute (SBMRI, La Jolla,...* (13 annotations)

2018-08-14: **AID 743463** *Keywords: GCN5, acetyl transferase, SIRT1 deacetylase, qPCR, acetylation, inflammation, diabetes,...* (17 annotations)

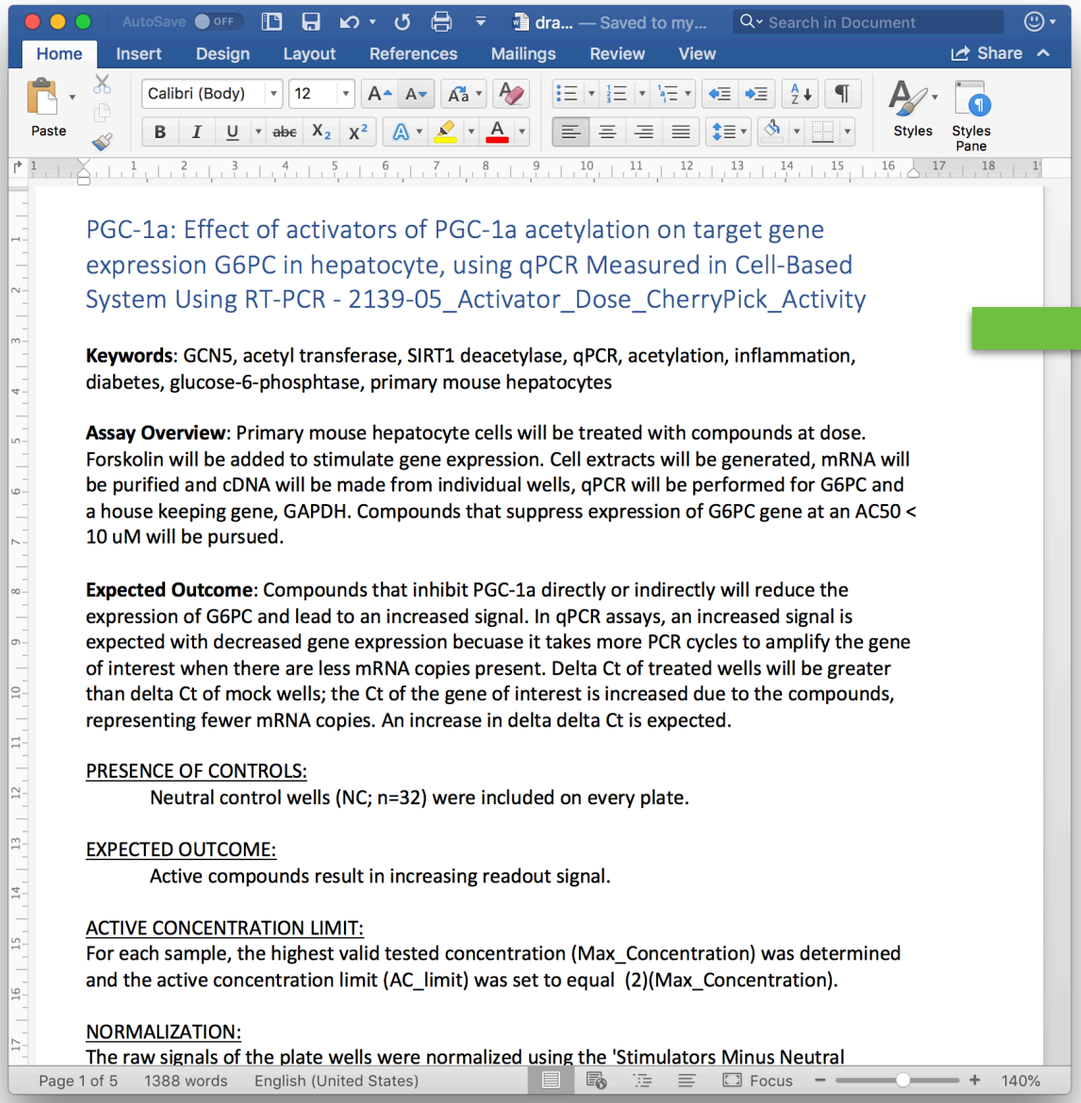
2018-08-14: **AID 504808** *Title: A genome-wide RNA interference screen identifies putative chromatin regulators essential for...* (17 annotations)

2018-08-13: **AID 588691** *Keywords: delta FosB homodimer, AP-1 DNA binding, activator, counterscreen, biochemical fluorescence...* (18 annotations)

2018-08-13: **AID 1053126** *Keywords: Cytotoxicity Assay, luminescence, CellTiter-Glo,*

New data: ELN

- Paste text or drag a document...



PGC-1a: Effect of activators of PGC-1a acetylation on target gene expression G6PC in hepatocyte, using qPCR Measured in Cell-Based System Using RT-PCR - 2139-05_Activator_Dose_CherryPick_Activity

Keywords: GCN5, acetyl transferase, SIRT1 deacetylase, qPCR, acetylation, inflammation, diabetes, glucose-6-phosphtase, primary mouse hepatocytes

Assay Overview: Primary mouse hepatocyte cells will be treated with compounds at dose. Forskolin will be added to stimulate gene expression. Cell extracts will be generated, mRNA will be purified and cDNA will be made from individual wells, qPCR will be performed for G6PC and a house keeping gene, GAPDH. Compounds that suppress expression of G6PC gene at an AC50 < 10 uM will be pursued.

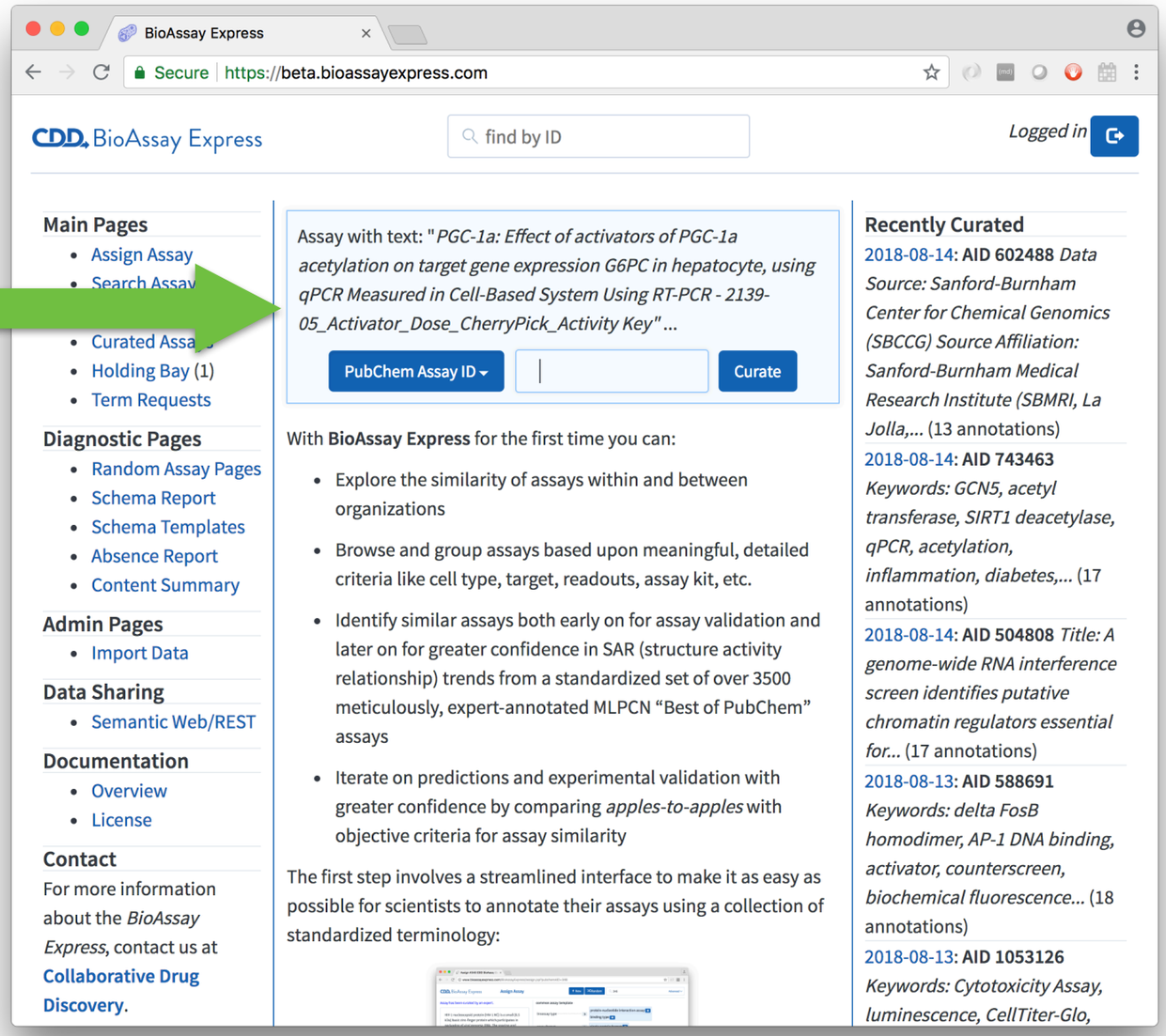
Expected Outcome: Compounds that inhibit PGC-1a directly or indirectly will reduce the expression of G6PC and lead to an increased signal. In qPCR assays, an increased signal is expected with decreased gene expression becuase it takes more PCR cycles to amplify the gene of interest when there are less mRNA copies present. Delta Ct of treated wells will be greater than delta Ct of mock wells; the Ct of the gene of interest is increased due to the compounds, representing fewer mRNA copies. An increase in delta delta Ct is expected.

PRESENCE OF CONTROLS:
Neutral control wells (NC; n=32) were included on every plate.

EXPECTED OUTCOME:
Active compounds result in increasing readout signal.

ACTIVE CONCENTRATION LIMIT:
For each sample, the highest valid tested concentration (Max_Concentration) was determined and the active concentration limit (AC_limit) was set to equal (2)(Max_Concentration).

NORMALIZATION:
The raw signals of the plate wells were normalized using the 'Stimulators Minus Neutral



BioAssay Express

Secure https://beta.bioassayexpress.com

CDD BioAssay Express

find by ID

Logged in

Main Pages
- Assign Assay
- Search Assay
- Curated Assays
- Holding Bay (1)
- Term Requests

Diagnostic Pages
- Random Assay Pages
- Schema Report
- Schema Templates
- Absence Report
- Content Summary

Admin Pages
- Import Data

Data Sharing
- Semantic Web/REST

Documentation
- Overview
- License

Contact
For more information about the *BioAssay Express*, contact us at **Collaborative Drug Discovery**.

Assay with text: "*PGC-1a: Effect of activators of PGC-1a acetylation on target gene expression G6PC in hepatocyte, using qPCR Measured in Cell-Based System Using RT-PCR - 2139-05_Activator_Dose_CherryPick_Activity Key*"...

PubChem Assay ID | Curate

With **BioAssay Express** for the first time you can:

- Explore the similarity of assays within and between organizations
- Browse and group assays based upon meaningful, detailed criteria like cell type, target, readouts, assay kit, etc.
- Identify similar assays both early on for assay validation and later on for greater confidence in SAR (structure activity relationship) trends from a standardized set of over 3500 meticulously, expert-annotated MLPCN "Best of PubChem" assays
- Iterate on predictions and experimental validation with greater confidence by comparing *apples-to-apples* with objective criteria for assay similarity

The first step involves a streamlined interface to make it as easy as possible for scientists to annotate their assays using a collection of standardized terminology:

Recently Curated

2018-08-14: **AID 602488** *Data Source: Sanford-Burnham Center for Chemical Genomics (SBCCG) Source Affiliation: Sanford-Burnham Medical Research Institute (SBMRI, La Jolla,...* (13 annotations)

2018-08-14: **AID 743463** *Keywords: GCN5, acetyl transferase, SIRT1 deacetylase, qPCR, acetylation, inflammation, diabetes,...* (17 annotations)

2018-08-14: **AID 504808** *Title: A genome-wide RNA interference screen identifies putative chromatin regulators essential for...* (17 annotations)

2018-08-13: **AID 588691** *Keywords: delta FosB homodimer, AP-1 DNA binding, activator, counterscreen, biochemical fluorescence...* (18 annotations)

2018-08-13: **AID 1053126** *Keywords: Cytotoxicity Assay, luminescence, CellTiter-Glo,*

- Or re-use...

Autogenerated text

Auto Annotation to Text for PubChem Assay AID743463

This is *a confirmatory assay* to identify potential treatments for an unknown disease, by investigating the biological process of **regulation of protein acetylation**, specifically targeting unknown target from **Mus musculus**.

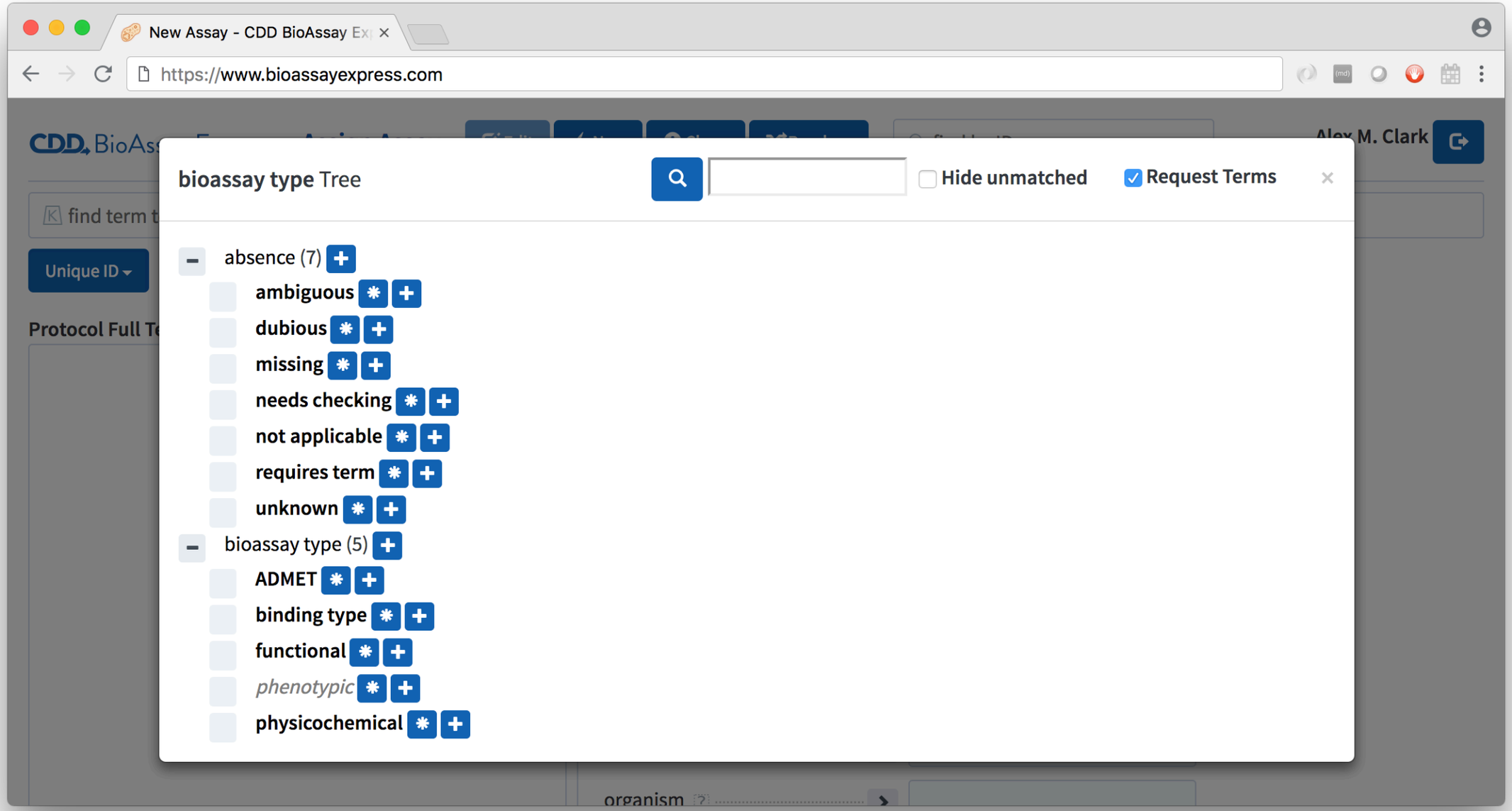
This is a **functional/gene expression assay** in *a cell based format*, using *a gene expression detection method*, with the assay kit *Cell to CT RT Kit (Ambion)*. The cell line *hepatocyte* was used. It was conducted in *384 well plates*, with the detection method of *fluorescence method*, using *a LightCycler 480 Real-Time PCR detection instrument*. Results are reported as *Ct*, in units of unknown. This assay tested the mode of action of *inhibition* by unknown perturbagens.

- Annotations-to-text is much easier
- Transliteration template:

```
[
    {
        "schemaURI": ["http://www.bioassayontology.org/bas#"],
        "boilerplate":
        [
            "<p>",
            "Auto Annotation to Text for PubChem Assay ",
            {"field": "uniqueID"},
            "</p>",

            "<p>",
            "This is ",
            {
                "ifany": "bao:BAO_0000210",
                "then": {"term": "bao:BAO_0000210", "article": "indefinite", "style": "italic"},
                "else": "an <i>unknown stage assay</i>"
            },
            " to identify potential treatments for ",
            {"term": "bao:BAO_0002848", "style": "bold", "empty": "an unknown disease"},
            ", by investigating the biological process of ",
            {"term": "bao:BAO_0002009", "style": "bold", "empty": "unknown"},
            ", specifically targeting ",
            {"term": "bao:BAO_0000211", "style": "bold", "empty": "unknown target"},
            " from ",
            {"term": "bao:BAO_0002921", "style": "bold", "empty": "unknown organism"},
            ".</p>",

            "<p>",

            "This is a ",
            {"terms": [["bao:BAO_0002854"], ["bao:BAO_0002855"]], "sep": "/", "style": "bold", "empty": "?"},
            " in ",
            {"term": "bao:BAO_0000205", "style": "italic", "article": "indefinite"},
            ", using ",
            {"term": "bao:BAO_0095009", "style": "italic", "article": "indefinite"},

            {
                "ifany": "bao:BAO_0002663",
                "then":
                {
                    "ifbranch": ["bat:Absence", "bao:BAO_0002663"],
                    "else":
                    [
                        ", with the assay kit ",
```

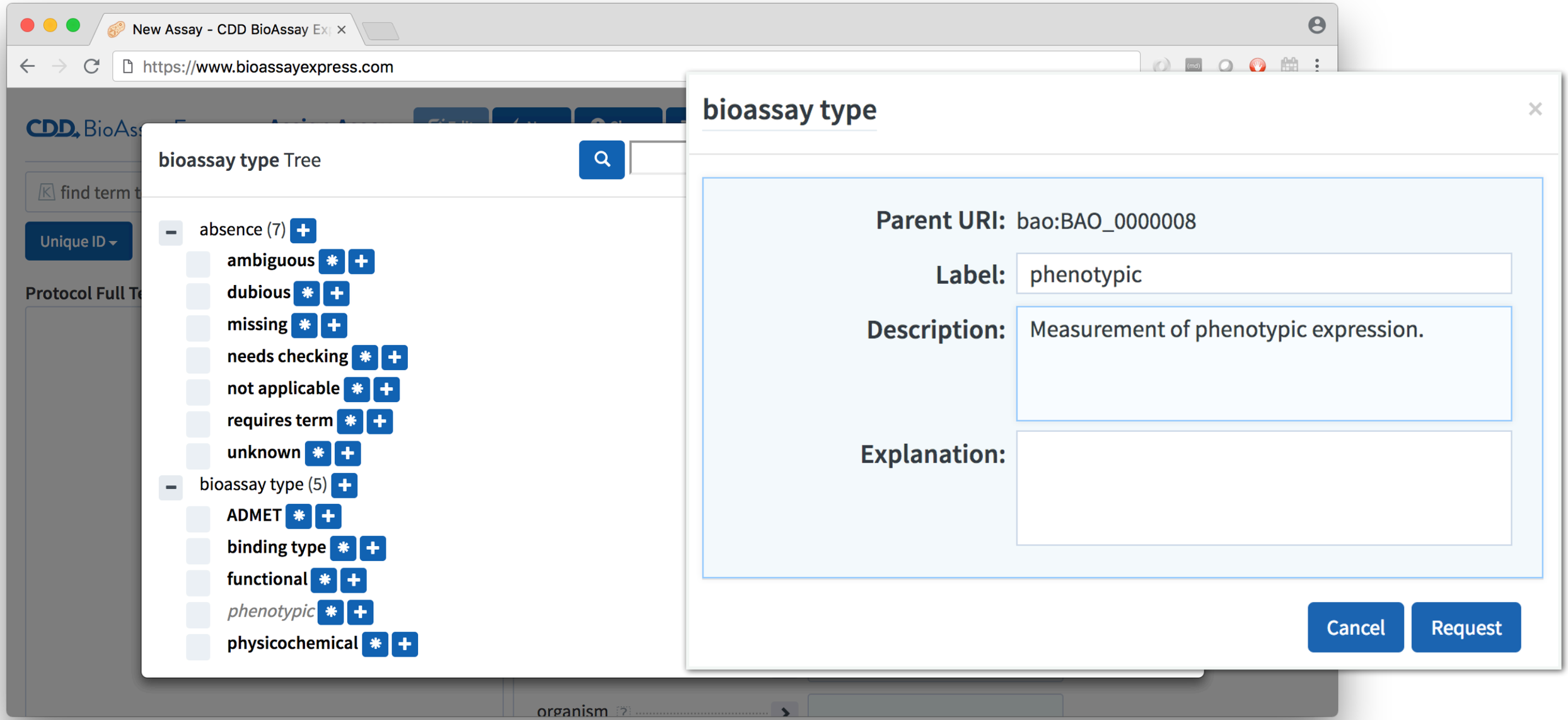
Extending ontologies

- Science is never complete: diction evolves
- Curators can request a *provisional term*



Extending ontologies

- Science is never complete: diction evolves
- Curators can request a *provisional term*



Upgrading ontologies

- Next step: submit *provisional terms* to **OntoloBridge**
- Joint collaboration between CDD and:
  - Stanford (BioPortal, CEDAR)
  - University of Miami (BAO, DTO)
- Ontology owner receives new additions via API
- Experts vet all new terms, outcome made available
- Users of BAE are unimpeded...

Detailed templates

- The **Common Assay Template** is a summary
- Can insert extra branch groups
- Ultimate goal:
  - capture whole experiment
  - replace text entirely
- Syncing with ontology maintainers...

★**measurement**
- measurement details

★**assay components**
- buffer concentration
- carrier protein concentration
- chelator concentration
- detergent concentration
- metal salt concentration
- reducing agent concentration
- solvent concentration

★**control details**

★**organism details**
- cell line details
- microbe/virus details

★**protocol details**
- antibody details
- assay molecule
- biochemical assay
- coupled substrate incubation
- enzyme reaction
- ligand incubation
- perturbagen incubation
- substrate incubation

★**screened entity details**

Content expansion

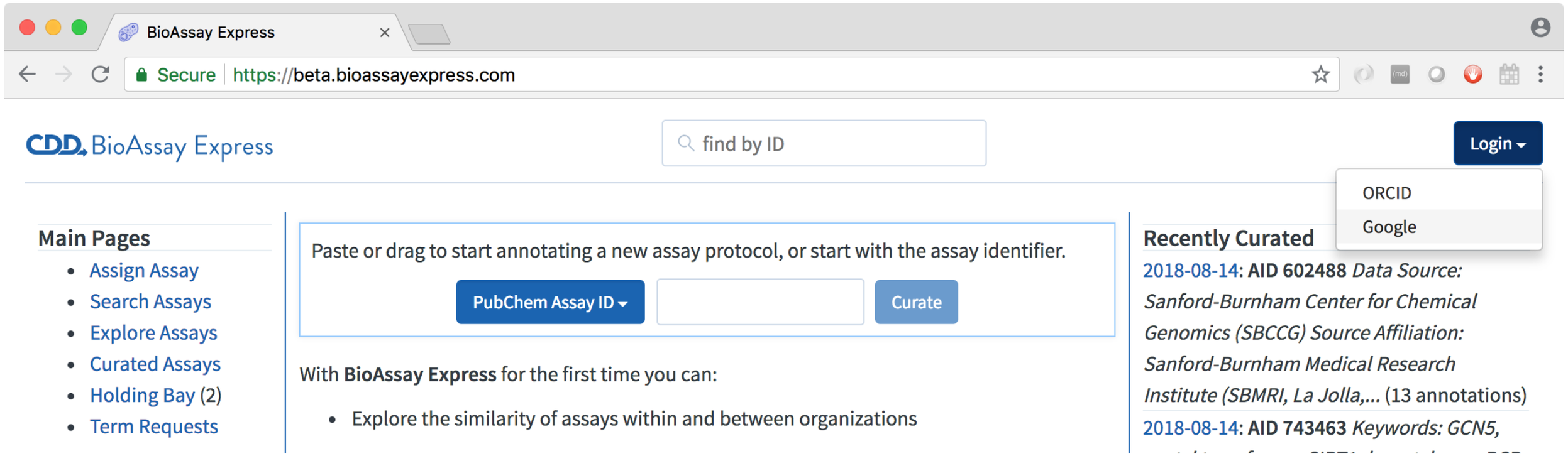
- Public data originally from **PubChem** subset:
  - **MLPCN**: thousands of screening assays
  - **Tox21**: hundreds of assays from EPA
- Detailed text, and compounds + measurements (SAR)
- Done most of them: where's the next goldmine?

PubMed

- PubMed is massive (1.7M open access), but:
  - only some have relevant assays
  - no compounds or other useful metadata
- Current project:
  - rank assays based on likelihood of assay description
  - pick out relevant text
- Hundreds of thousands of candidate assays for BAE
- Overlap with ChEMBL: ~600 (1.7M ∩ 70K) - metadata, SAR

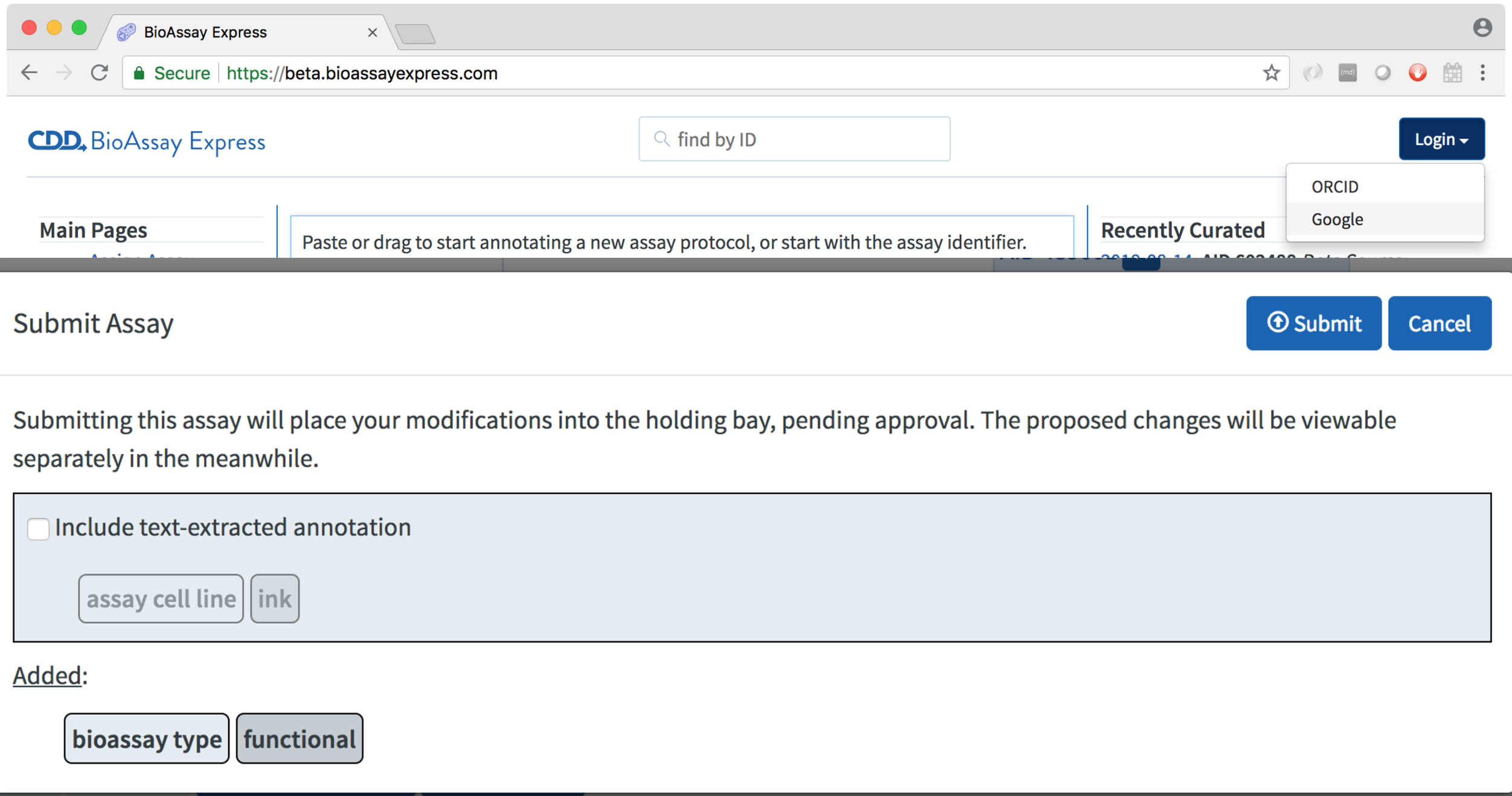
Crowdsourcing

- Anyone can authenticate on **beta.bioassayexpress.com**



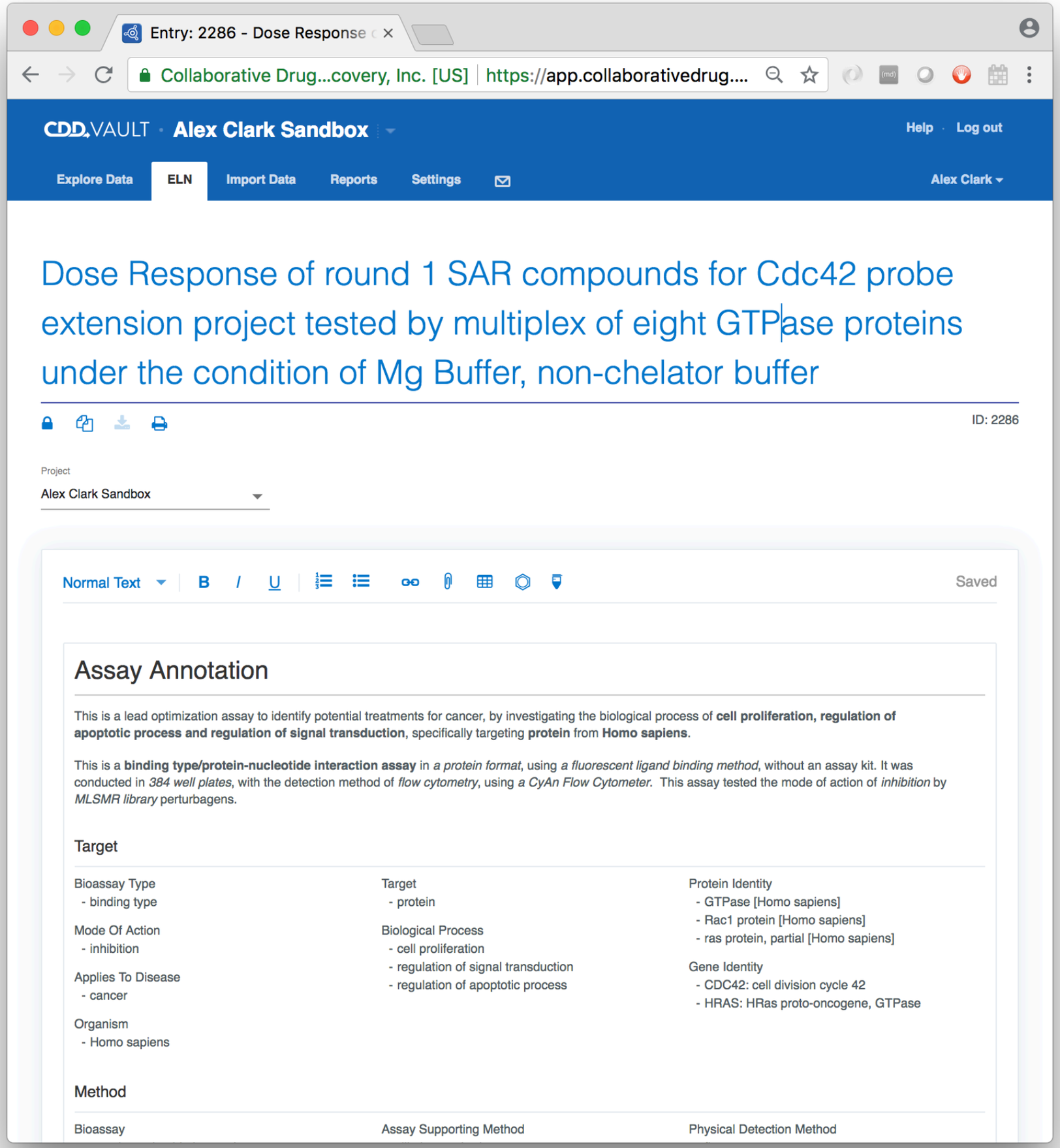
Crowdsourcing

- Anyone can authenticate on **beta.bioassayexpress.com**



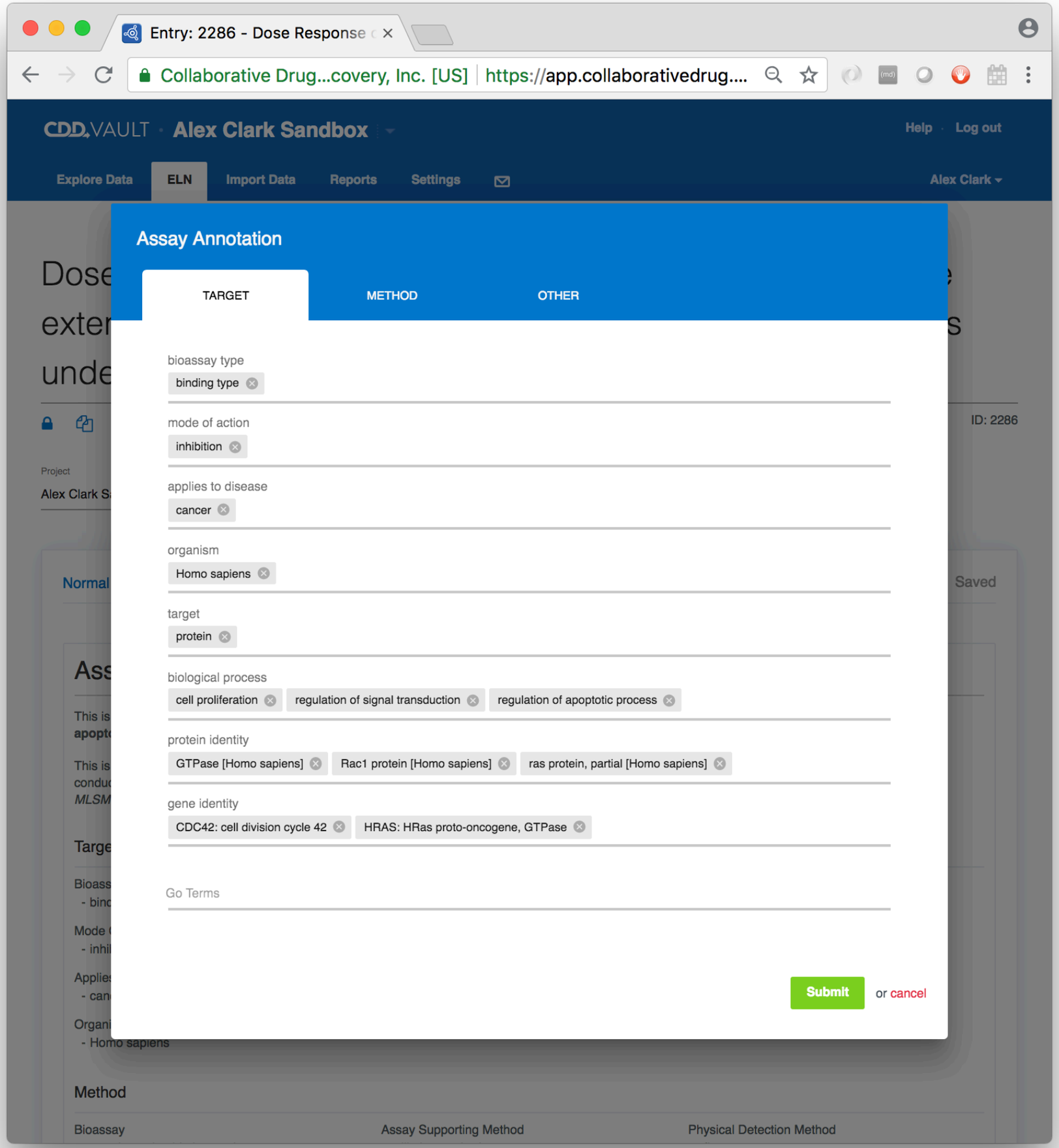
Integration

- Basic functionality integrated into **CDD Vault**
- ELN users can insert an **assay**
- Annotations are semantic, like standalone BAE



Integration

- Basic functionality integrated into **CDD Vault**
- ELN users can insert an **assay**
- Annotations are semantic, like standalone BAE



Acknowledgments

BioAssay Express Team

- Peter Gedeck
- Hande Küçük
- Barry Bunin
- Jason Harris
- Charlie Read
- Samantha Jeschonek

Collaborators

- René Witte & Bahar Sateli (Knowlet)
- Julia Davies (UCSF)
- Mark Musen & John Graybeal (Stanford)
- Stephan Schürer (U. of Miami)

More information

http://www.bioassayexpress.com

http://collaborativedrug.com

Funding
NIH SBIR