

Leveling up chemical information for the era of big data

Alex M. Clark



CDD VAULT[®]
Complexity Simplified

Background

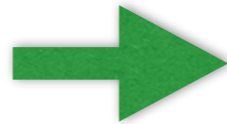
- ◆ **Big Data**: more like **Large Data** (+ inconvenience)
- ◆ We can have aspirations
- ◆ For any established subset of chemistry: vast amounts of data exists...
- ◆ ... but not in a way computers can use it

History

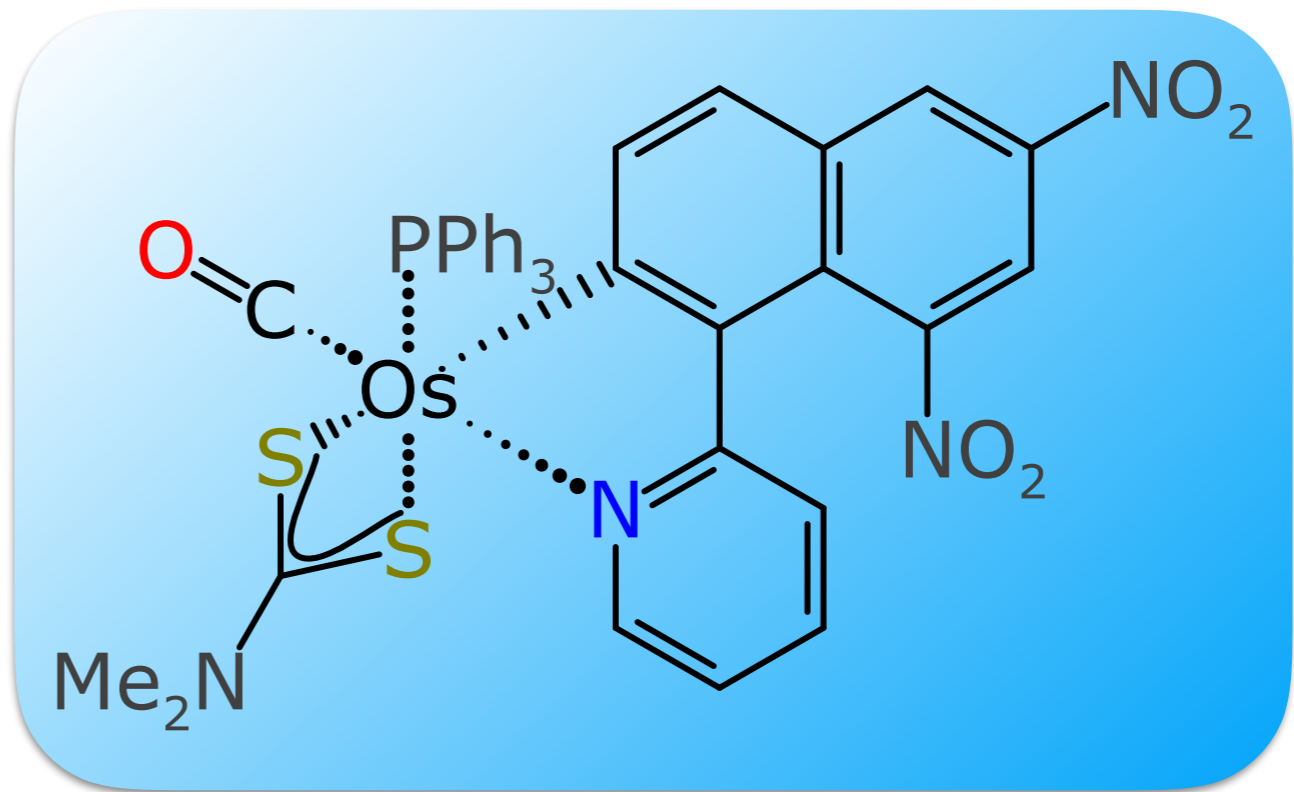
- ◆ Early 20th century: science was a small town, everyone knew each other
- ◆ The dead tree model of publishing scaled nicely
- ◆ 1990s: my grad advisor (W.R. Roper) knew all the players in organometallics, read everything
- ◆ 2000s: my postdoc advisor (C.A. Reed) opined we were *drowning in peer reviewed literature*

Grains of sand on a beach

◆ From my graduate research...



◆ ... and a hundred more much like it.



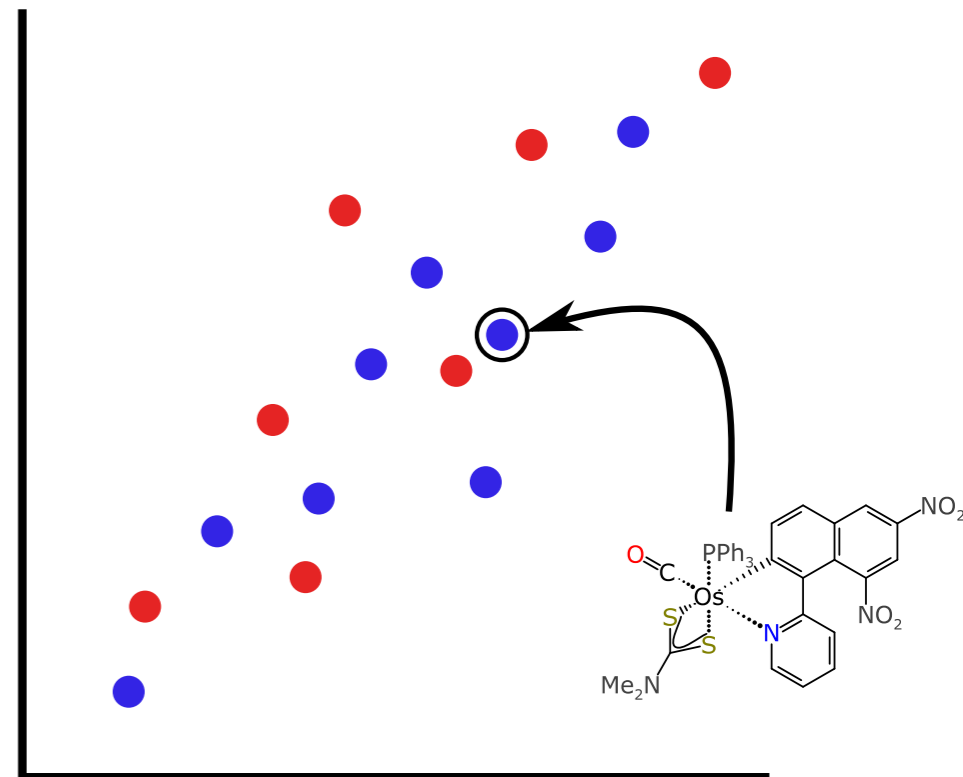
◆ Pure academic research: to see if we can

◆ Published in respectable journals: a few citations, mostly forgotten in dusty tomes & PDF files

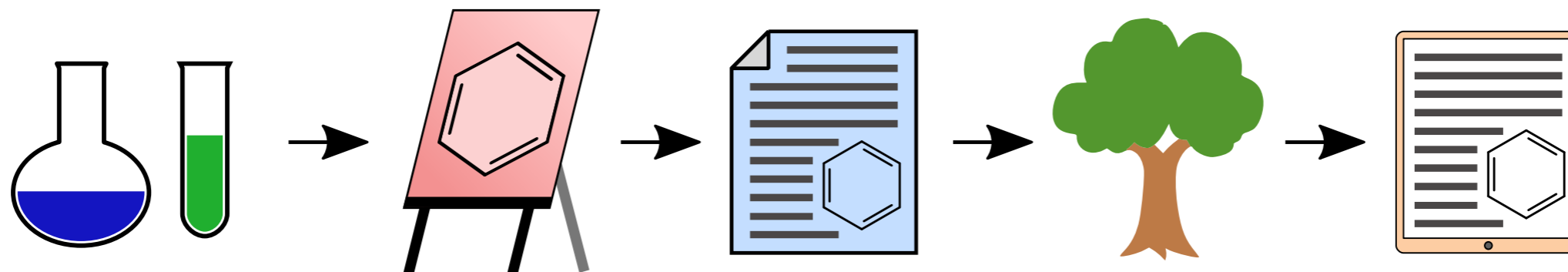
◆ All that time & effort & funding... for what?

To be a datapoint

- ◆ Every unit of science has *potential* value, but no way to be sure when or to whom
- ◆ Nobody can read it all, but the solution is not to publish less
- ◆ We need to rethink the target audience: it's not humans... it's **machines**
- ◆ They can handle the scale...
- ◆ ... but they cannot interpret the input data.



Digital trees



- ◆ Computers for writing & reading since 1990s
- ◆ Yet we use tech to pretend we're with Newton & friends in early days of the Royal Society
- ◆ Each step is digital, but the process is modelled on typewriters and wood carvings
- ◆ What little data existed in the intermediate documents is destroyed during the final production

Progress?

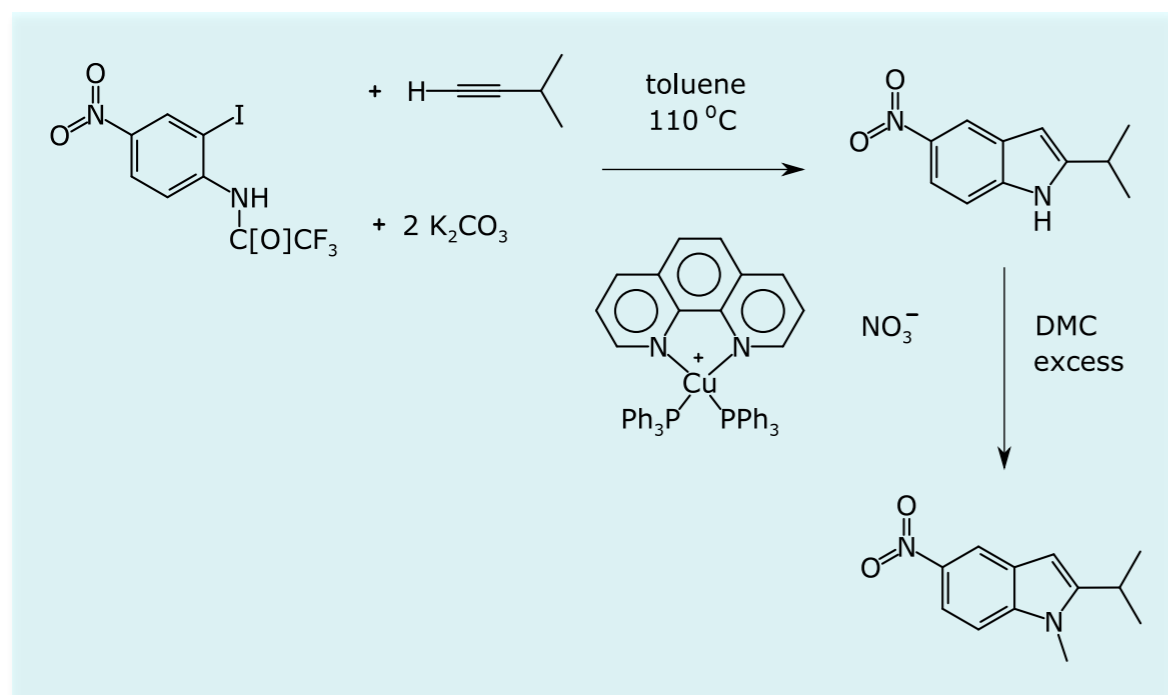
- ◆ Most chemical data is silo'd and opaque
- ◆ Drug discovery is very motivated: there are structure-activity datasets that are **free, large, useful** *and accurate* (previously: pick zero)
- ◆ Now: **ChEMBL, PubChem** & others (SAR data)
- ◆ Improved accessibility:
 - ▷ open access is growing (still not data though)
 - ▷ APIs to the patent literature
 - ▷ FAIR data movement: building steam

Dark data

- ◆ Almost all published chemical research is visible only to humans, and effectively dark to machines
- ◆ This is almost as true now as it was in 1700
- ◆ Solutions have distinct categories:
 - ▶ **fix the past**: curate old publications using descriptive data formats - machine learning and text mining can help, but still need expensive expert humans
 - ▶ **fix the present** (+ future): publish new data using file formats that target machines *and* humans - need better tools and different attitudes

Case 1: Figures

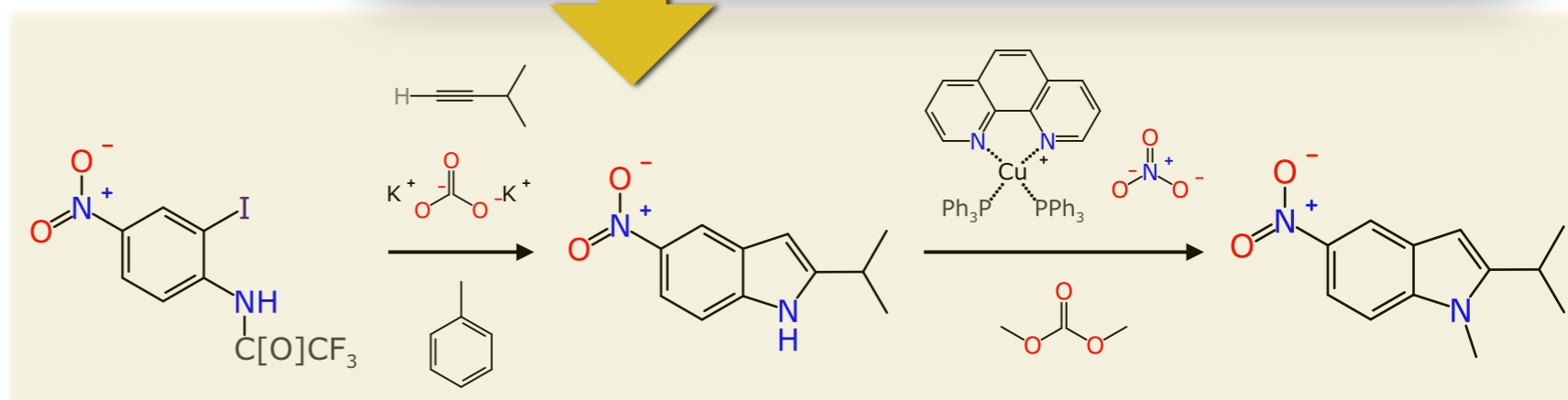
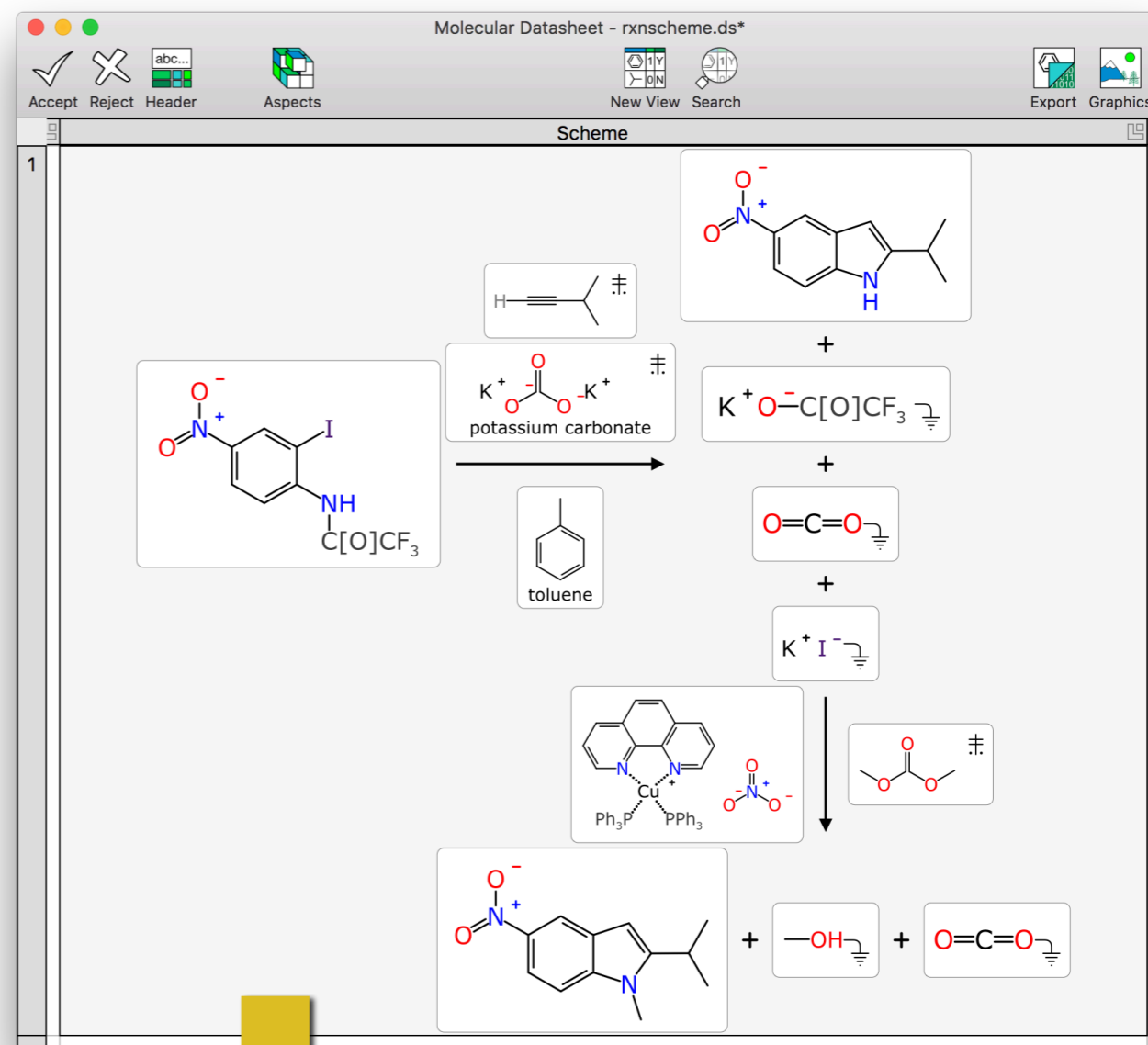
- ◆ Chemists painstakingly draw molecules & reactions



- ◆ Interpretability is high for other chemists...
- ◆ ... and surprisingly low for machines.

Cheminformatics first

- ◆ Similar amount of effort to draw fully defined content
- ◆ e.g. **Molecular Notebook**
- ◆ Draw structures & reaction components for computers
- ◆ Let algorithms do the rendering...



Internet-era manuscripts

- ◆ Using standard wordprocessor is not good for data
- ◆ WordPress plugin **MolPress**:
 - ▷ insert & retain raw data
 - ▷ plugin handles rendering
 - ▷ open source
- ◆ Use blogging software as an electronic lab notebook

The screenshot shows the MolPress website interface. At the top, the browser address bar displays "molpress.com". The page title is "MOLPRESS Open source chemistry content plugin for WordPress".

Individual Molecules

Four molecular structures are displayed: isopropanol, benzene, a mercury complex, and a complex organic molecule with a sulfur atom.

Collection of Solvents

| MOLECULE | NAME | CASRN |
|----------|------------------------|----------|
| | acetic acid | 64-19-7 |
| | acetic anhydride | 108-24-7 |
| | formic acid | 64-18-6 |
| | methane sulphonic acid | 75-75-2 |

Scaffold Breakdown

| MOLECULE | SCAFFOLD | R1 | R2 | R3 | R4 | R5 | ID |
|----------|----------|----|----|----|----|----|----|
| | | | | | | | 1a |
| | | | | | | | 1c |
| | | | | | | | 1d |
| | | | | | | | 1e |

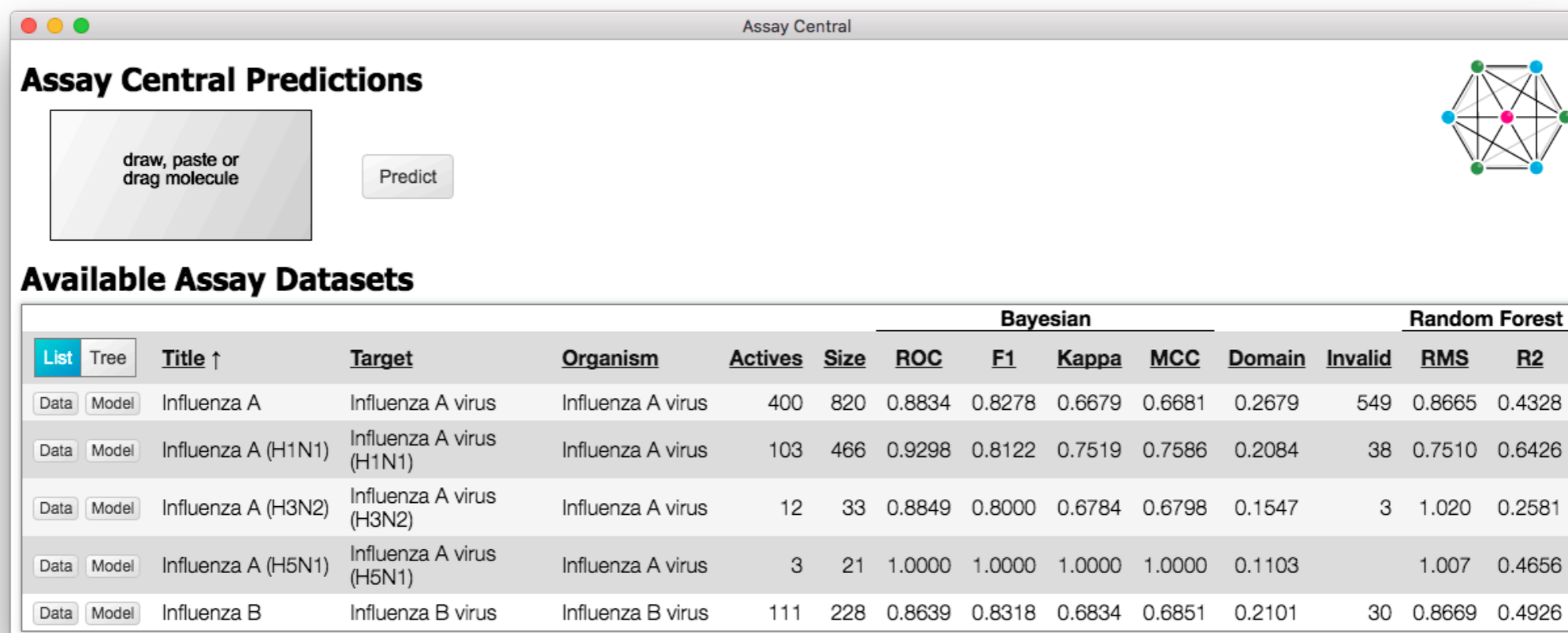
Reaction Scheme

Case 2: Assay Central

- ◆ Rare & neglected diseases: long tail, few resources
- ◆ Premise: the data is out there
- ◆ Pragmatism: grab & fix, by any means necessary
- ◆ Sources:
 - ▷ **ChEMBL, PubChem** (parts thereof)
 - ▷ curation, collaboration, importing, scraping
- ◆ Goal: seed *all* targets with only good data

Provenance

- ◆ Some data is good: just have to label (e.g. target), others must sample, validate by flagging
- ◆ Merging: standardise, group by, build models
- ◆ Project → **GitHub** repository: a molecular build system



Assay Central Predictions

draw, paste or drag molecule

Available Assay Datasets

| | | Title ↑ | Target | Organism | Actives | Size | Bayesian | | | | Random Forest | | | |
|------|-------|--------------------|--------------------------|-------------------|---------|------|----------|--------|--------|--------|---------------|---------|--------|--------|
| List | Tree | | | | | | ROC | F1 | Kappa | MCC | Domain | Invalid | RMS | R2 |
| Data | Model | Influenza A | Influenza A virus | Influenza A virus | 400 | 820 | 0.8834 | 0.8278 | 0.6679 | 0.6681 | 0.2679 | 549 | 0.8665 | 0.4328 |
| Data | Model | Influenza A (H1N1) | Influenza A virus (H1N1) | Influenza A virus | 103 | 466 | 0.9298 | 0.8122 | 0.7519 | 0.7586 | 0.2084 | 38 | 0.7510 | 0.6426 |
| Data | Model | Influenza A (H3N2) | Influenza A virus (H3N2) | Influenza A virus | 12 | 33 | 0.8849 | 0.8000 | 0.6784 | 0.6798 | 0.1547 | 3 | 1.020 | 0.2581 |
| Data | Model | Influenza A (H5N1) | Influenza A virus (H5N1) | Influenza A virus | 3 | 21 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.1103 | | 1.007 | 0.4656 |
| Data | Model | Influenza B | Influenza B virus | Influenza B virus | 111 | 228 | 0.8639 | 0.8318 | 0.6834 | 0.6851 | 0.2101 | 30 | 0.8669 | 0.4926 |

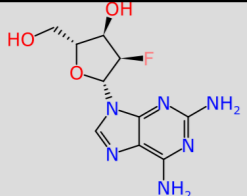
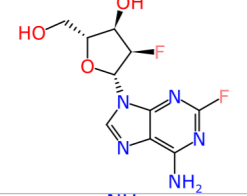
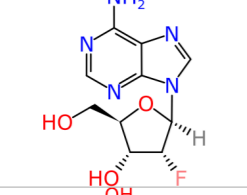
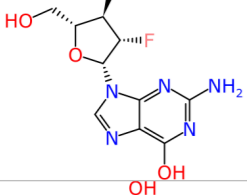
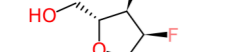
Virtuous cycle

- ◆ Initial data integration is labour intensive
- ◆ Build models & other decision support
- ◆ Can generate *prescribed* molecules
 - ▷ high predictions for desired target(s)
 - ▷ low predictions for undesirable off-targets
- ◆ Collaborators can order & test these candidates, and return the data in **model-ready format**

Case 3: CDD Vault

◆ Entry point is an *import* process: effort is up front...

Molecular Datasheet - assay.ds

| | Molecule | Name | Activity | SourceURI | Or |
|---|---|------|-----------|-------------------|-------|
| 1 |  | | = 4.82391 | mol:CHEMBL3143553 | 1:799 |
| 2 |  | | < 4 | mol:CHEMBL3143526 | 1:786 |
| 3 |  | | = 4.284 | mol:CHEMBL309579 | 1:790 |
| 4 |  | | = 4.72125 | mol:CHEMBL3143549 | 1:785 |
| 5 |  | | | | |

Add readouts (protocol data) only Add or update molecules

| | A | B | C | D | E | F |
|---|-------------------|---------|----------|-------------------|--------|-------|
| 1 | Molfile | Value | Relation | SourceURI | Origin | Error |
| 2 | Extended... M END | 4.82391 | = | mol:CHEMBL3143553 | 1:799 | |
| 3 | Generate... M END | 4 | < | mol:CHEMBL3143526 | 1:786 | |
| 4 | Generate... M END | 4.284 | = | mol:CHEMBL309579 | 1:790 | |

Structure

pConc
→ pConc
→ 2018-08-11 (Alex ...)

Molecule Name or Synor

▼ Molecule Fields

- Molecule Name or Synonym
- Structure
- User-defined Field
- ▶ Batch Fields
- ▶ Plate and Well
- Readouts (Protocol Data)
- Do not Import

Molfile is mapped to Structure

Structure will be modified according to the following rules.

Next

Save this mapping as a template... ?

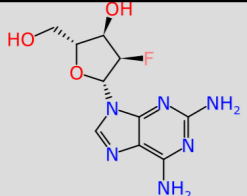
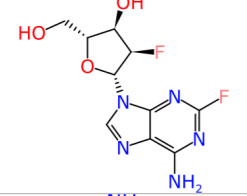
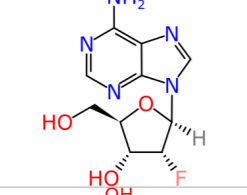
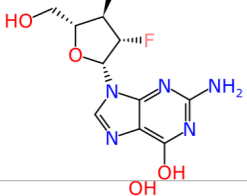
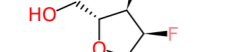
Process File

◆ ... after that, everything else just works.

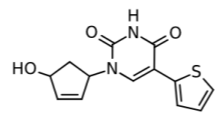
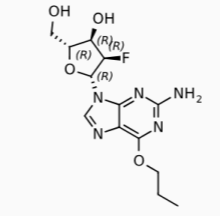
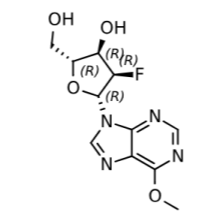
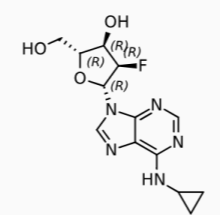
Case 3: CDD Vault

◆ Entry point is an *import* process: effort is up front...

Molecular Datasheet - assay.ds

| | Molecule | Name | Activity | SourceURI | Or |
|---|---|------|-----------|-------------------|-------|
| 1 |  | | = 4.82391 | mol:ChEMBL3143553 | 1:799 |
| 2 |  | | < 4 | mol:ChEMBL3143526 | 1:786 |
| 3 |  | | = 4.284 | mol:ChEMBL309579 | 1:790 |
| 4 |  | | = 4.72125 | mol:ChEMBL3143549 | 1:785 |
| 5 |  | | | | |

789 Selected: Launch Vision Plot Export Add to collection Build model Flag outliers Customize your report Save this search

| Select... | Molecule | pConc |
|-------------------------------------|--|---------|
| all · none | Alex Clark Sandbox | pConc |
| <input checked="" type="checkbox"/> |  mol:ChEMBL3217926 Alex Clark Sandbox | 6.09691 |
| <input checked="" type="checkbox"/> |  mol:ChEMBL3143559 Alex Clark Sandbox | 4.4437 |
| <input checked="" type="checkbox"/> |  mol:ChEMBL3143558 Alex Clark Sandbox | 4 |
| <input checked="" type="checkbox"/> |  mol:ChEMBL3143557 Alex Clark Sandbox | 4 |

◆ ... after that, everything else just works.

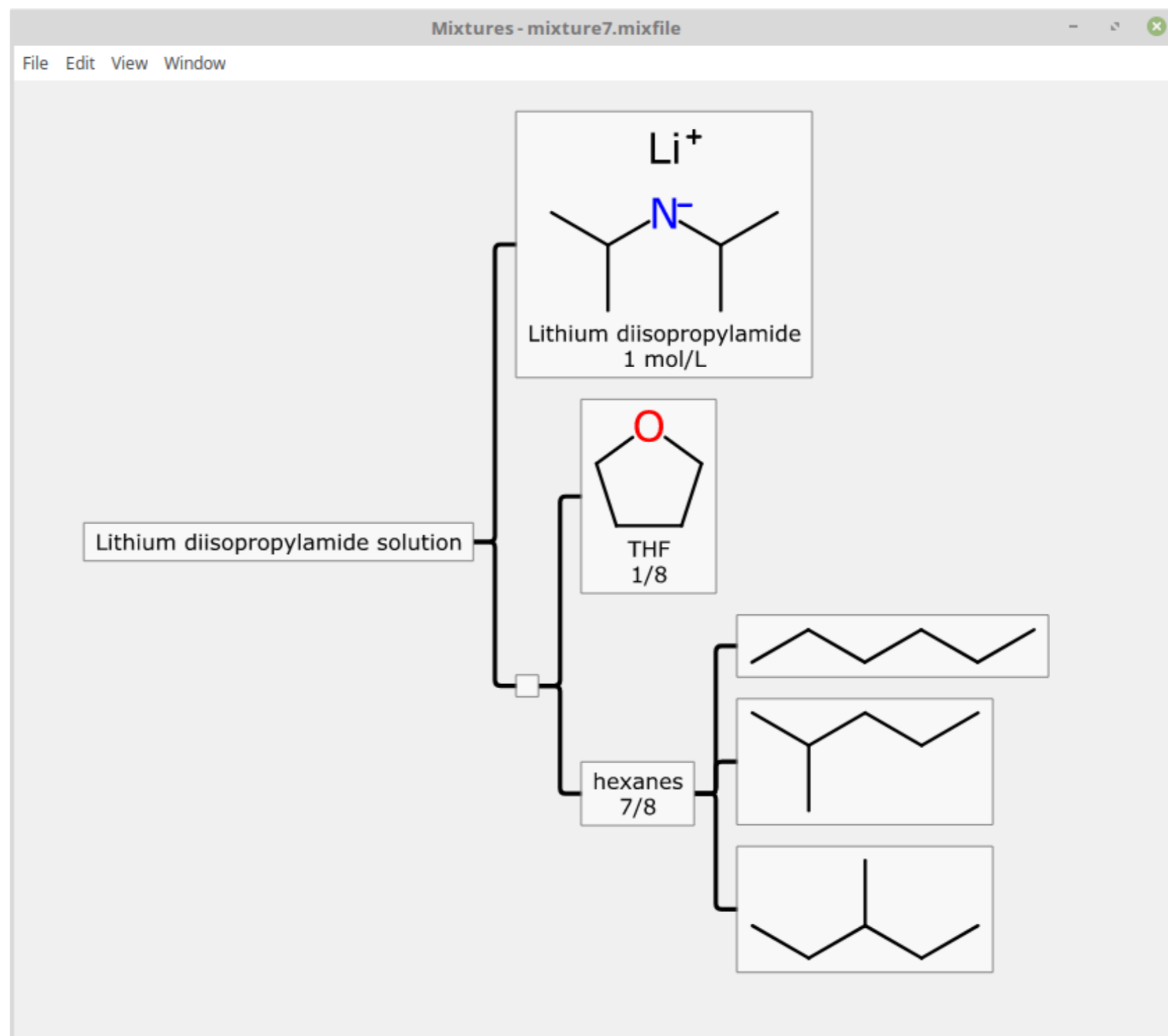
Case 4: Mixtures

- ◆ Mixtures of compounds: ad hoc is the way, e.g.
 - ▷ `osmium tetroxide 2.5 wt. % in tert-butanol`
 - ▷ `n-butyllithium 2.5M in hexanes`
- ◆ Long overdue for a well defined open format
 - ▷ **Mixfile**
- ◆ High priority: text-extraction importer
- ◆ Working with IUPAC: data feedstock

| | | |
|----------------|---|---------------|
| Molfile | → | InChI |
| Mixfile | → | MInChI |

Mixfile editor

- ◆ Verbose
- ◆ Hierarchical
- ◆ Cross-platform
- ◆ Open source
- ◆ JSON-based



Case 5: Bioassay protocols

◆ Assay protocols usually just text: enrich by describing them with *semantic web terms* (e.g. BAO, DTO, CLO...)

◆ Curated assays are machine readable

◆ Legacy: ML-support

◆ Current: ELN-style

◆ Began 2014

◆ See Wednesday

The screenshot shows the BioAssay Express web interface for editing a PubChem assay (ID 743084). The interface is divided into several sections:

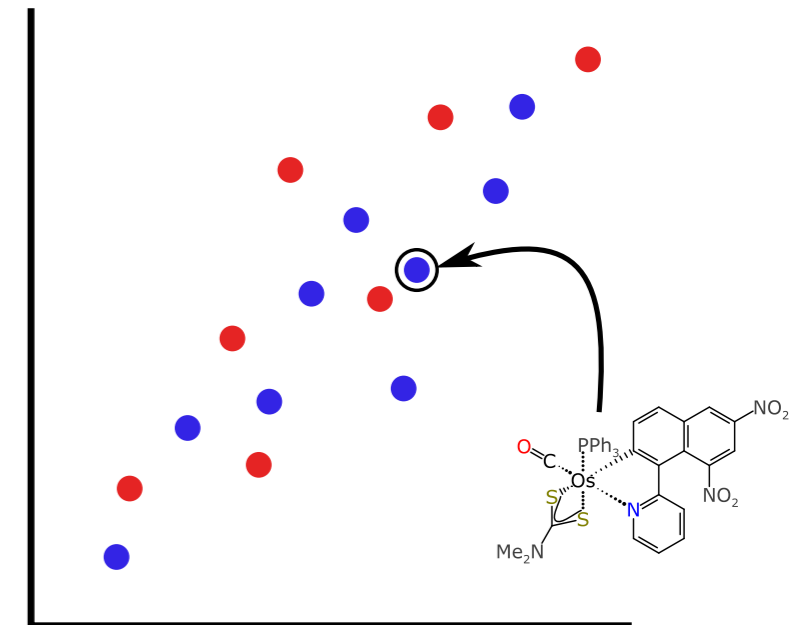
- Header:** "Assign PubChem Assay ID 743084" and "Assign Assay" button. Navigation buttons: Edit, New, Clone, Random. Search bar: "find by ID".
- Search:** "find term to annotate with" input field.
- Assay Information:** PubChem Assay ID: 743084, common assay template, Template, Sections.
- Protocol Full Text:** U.S. Tox21 Program, National Center for Advancing Translational Sciences [NCATS], NIH Chemical Genomics Center [NCGC], U.S. Environmental Protection Agency [EPA], National Institutes of Environmental Health Sciences [NIEHS], National Toxicology Program [NTP], U.S. Food and Drug Administration [FDA]. Tox21 Assay Overview: Endocrine disrupting chemicals (EDCs) interfere with the biosynthesis and normal functions of steroid hormones including estrogen and androgen in the body. Aromatase catalyzes the conversion of androgen to estrogen and plays a key role in maintaining the androgen and estrogen balance in many of the EDC-sensitive organs. MCF-7 aro ERE cell line (provided by Dr. Shiu-an Chen at Beckman Research Institute of the City of Hope) was used to screen the Tox21 10K compound library collection for identification of aromatase inhibitors. MCF-7 aro ERE is human breast carcinoma cell line that was stably transfected with a promoter plasmid, pGL3-Luc, containing three repeats of estrogen responsive element (ERE). The cytotoxicity of the Tox21 compound library against the MCF-7 aro ERE cell line was tested in parallel by measuring the cell viability using CellTiter-Fluor assay (Promega, Madison, WI) in the same wells.
- Annotation Fields:** assay title, bioassay type, bioassay, assay format, assay design method, assay supporting method, assay cell line, organism, biological process, target, applies to disease, assay mode of action, result, screening campaign stage. Each field has a dropdown menu with selected terms and a search icon.
- Similar Assays:** A list of similar assays with their IDs: AID 651634, AID 743041, AID 743081, AID 720634, AID 743074, AID 1224875, AID 1224876, AID 720693, AID 743033, AID 426. A search bar is also present.

Case 6: Data are valuable

- ◆ The only way forward: machine readability at source
- ◆ Scientists must be “**persuaded**” to get on board
- ◆ Better tools can help, but mostly social engineering
- ◆ Who controls much of the culture of science?

Journal publishers

- ◆ Only the human-readable parts of a manuscript are peer reviewed (text & figures)
- ◆ What if the **data** were part of the formal process...
- ◆ ... if an algorithm can't understand it, send it back.
- ◆ Supplementary information becomes the most important part of the article
- ◆ Open it to everyone's model



Example

◆ Can start small, e.g. with organic/drug discovery

Journal of Medicinal Chemistry

Article

protein kinase inhibitor, but was found in a zebrafish embryo dorsalization assay to selectively inhibit BMP signaling through SMAD1/5/8.⁶ Further development led to LDN-193189,⁷ LDN-212854,⁸ and ML347,⁹ which had improved microsomal stability, potency, and selectivity. In addition, these molecules demonstrated efficacy in mouse models of FOP.⁸ However, these compounds have a number of kinase off-targets and display dose-limiting toxicity with a 10% loss in body weight in animal models.¹⁰ A second series of inhibitor based on a pyridine core (e.g., K02288¹¹ and LDN-214117¹²), with equivalent biochemical potency and improved kinome selectivity, has also been reported (Figure 1).

RESULTS

We identified 6-pyrazole quinazolinone, **1**, as a ligand efficient inhibitor of ALK2 ($IC_{50} = 8.2 \mu M$; $LE = 0.45$) through cross-screening of a focused kinase fragment library, using Invitrogen's LanthaScreen binding assay. **1** shares features with reported ALK5 inhibitors such as PF-03671148¹³ and compound **19**.¹⁴ These are known to bind to the hinge of ALK5 through a single polar contact at the N-1 position of the quinazolinone moiety, with the 2-methylpyridine directing toward the ALK5 Ser280 gatekeeper residue, forming a key water mediated hydrogen bond to Lys232 (Figure 2).¹⁴

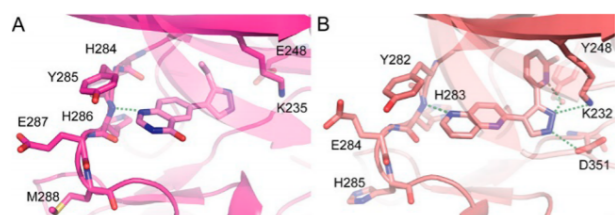


Figure 2. (A) Docking pose of **4** in ALK2 (structure used: 3Q4U). (B) X-ray structure of analogous naphthyridine-based inhibitor (compound **19**) of ALK5¹⁴ (PDB 1VJY).

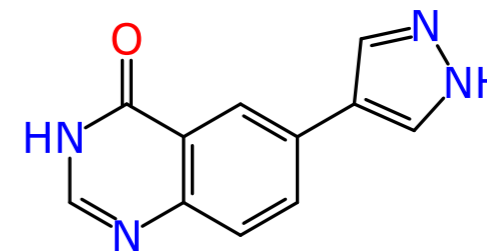
Interestingly, PF-03671148 has been shown to be selective against ALK1 likely due a larger gate keeper (Thr) causing a clash with the pyridine substituent.¹³ Given that ALK2 also

Table 1. SAR at Quinazolinone 6-Position^a

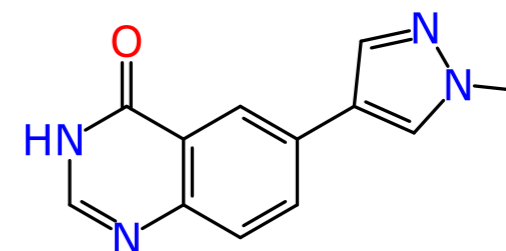
| No ^o | | ALK2 IC_{50} (μM) | LE | LiPE | TPSA |
|-----------------|--|-------------------------------|------|------|-------|
| 1 | | 8.20 | 0.45 | 4.79 | 70.14 |
| 2 | | 18.2 | 0.39 | 4.43 | 59.28 |
| 3 | | 2.30 | 0.46 | 5.38 | 70.14 |
| 4 | | 1.00 | 0.47 | 5.21 | 70.14 |
| 5 | | 0.386 | 0.47 | 5.71 | 70.14 |
| 6 | | 0.293 | 0.51 | 6.30 | 70.14 |
| 7 | | 0.167 | 0.47 | 4.90 | 54.35 |
| 8 | | 0.344 | 0.45 | 4.60 | 54.35 |
| 9 | | 0.653 | 0.43 | 4.65 | 58.76 |

^a IC_{50} data is an average of 2–4 measurements by LanthaScreen Eu kinase binding assay.

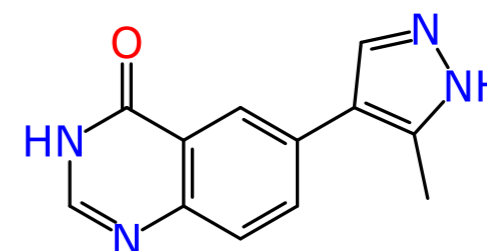
1.mol



2.mol



3.mol



data.csv

```
ID,ALK2IC50,LE,LiPE,TPSA  
-,uM,uM,uM,uM
```

```
1,8.20,0.45,4.79,70.14
```

```
2,18.2,0.39,4.43,59.28
```

```
3,2.30,0.46,5.38,70.14
```

◆ Author provides the files, middleware does basic validation, reviewer opens and verified

What are you waiting for?

- ◆ A whole industry & community exists for support
- ◆ Best of breed tools have open source options
- ◆ One publisher has to go first...
- ◆ ... **who?**



ACS
Chemistry for Life®



Springer



ELSEVIER



WILEY

PeerJ

Acknowledgments

- ◆ Royal Society of Chemistry & ACS CINF
- ◆ Barry Bunin (**CDD**)
- ◆ Peter Gedeck, Hande Küçük-McGinty (**BAE**)
- ◆ Sean Ekins, Kim Zorn (**CPI**)
- ◆ Leah McEwen (**IUPAC**)

alex@collaborativedrug.com
@aclarkxyz
cheminf20.org