Are Self-driving Labs for Chemistry Next?

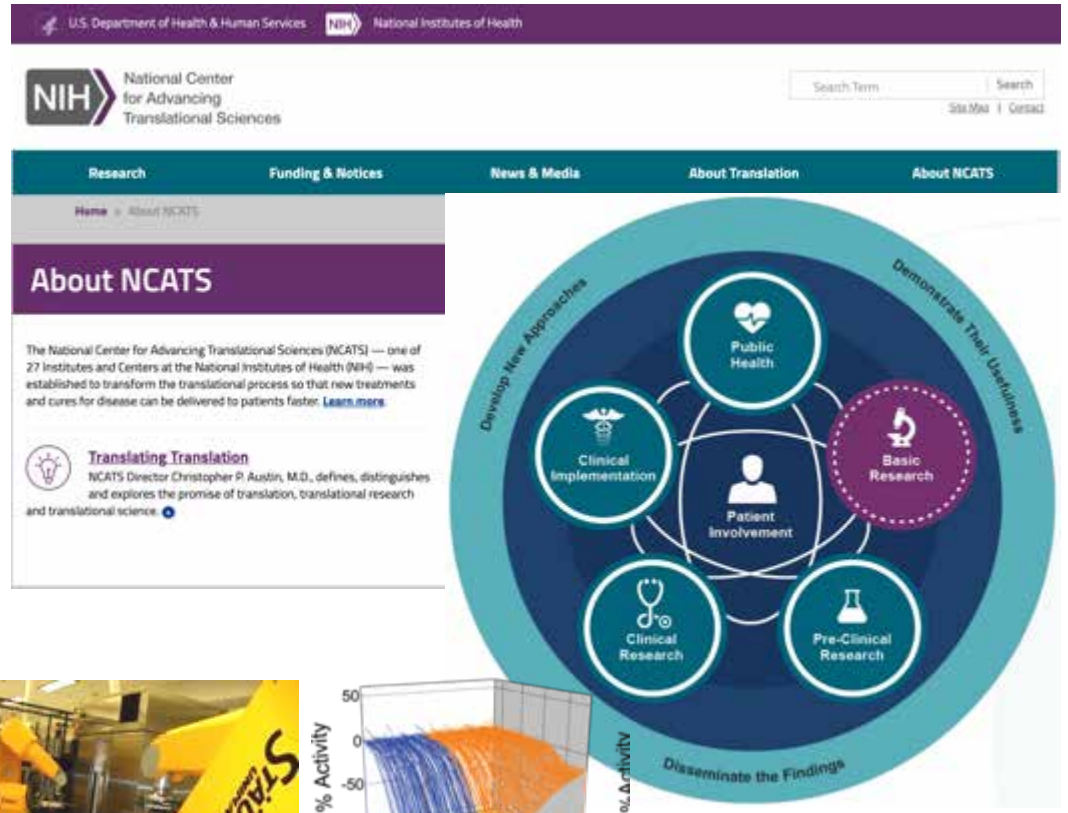
Alex Godfrey

NAS Space Science Week
CBPSS Meeting
March 31, 2020

National Center for Advancing Translational Sciences

National Center for Advancing Translational Sciences

- Founded 2004 as part of NIH Molecular Libraries Roadmap Initiative
- NIH Chemical Genomics Center (NCGC)
- MLPCN (screening & chemical synthesis; compound repository; PubChem database; funding for assay, library and technology development)
  - Develop new chemical probes for basic research and leads for therapeutic development, particularly for rare/neglected diseases
  - New paradigms & applications of HTS for chemical biology / chemical genomics
- Incorporated into NCATS in January 2012



ncats.nih.gov

NIH National Center for Advancing Translational Sciences

NCATS Mission

To catalyze the generation of innovative methods and technologies that will enhance the development, testing and implementation of diagnostics and therapeutics across a wide range of human diseases and conditions.

2

ASPIRE

A Specialized Platform for Innovative Research Exploration

Modern
Lab Bench

Re-thinking and Re-designing the Research Laboratory Ecosystem

Integrated
Automated
Solutions

“Self-Driving Lab”

Implies a system that can autonomously learn and improve its function by analyzing and generating its own data.

Schneider, G. (2019). Mind and machine in drug design. *Nature Machine Intelligence* 1(3)**,** 128-130. doi: 10.1038/s42256-019-0030-7.

ASPIRE

Modern Lab Bench

Integrated Automation

Space Chemistry??

Many Challenges in Automating Chemistry

Methodology	Robust, efficient, safe, mild conditions and high yielding
Physical Operations	Reagent delivery Product Isolation (workup) Evaporation (solvent removal or exchange)
Analysis	Invasive (sampling, chemical derivation, e.g. TLC, HPLC) Non-invasive (spectral analysis, UV-vis, FTIR, Raman, NMR)
Informatics	Data mining, curation, and feature engineering Contextual representation

ASPIRE's Technology Ecosystem

ASPIRE

Translational

Next Gen Medchem

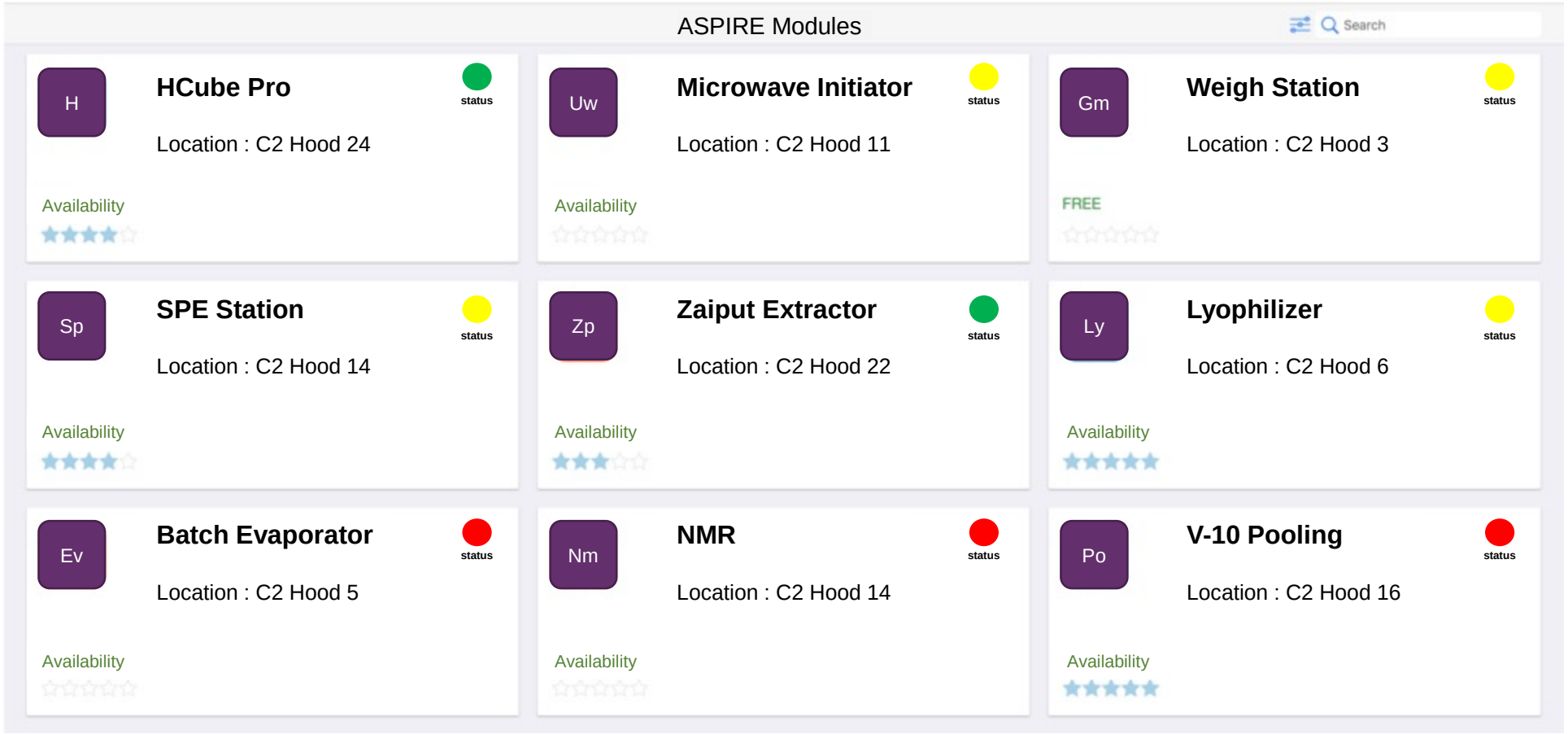
Collaboration-Oriented

Advanced Technology Development Lab

Technology-Driven

ASPIRE Next Gen MedChem Dashboard

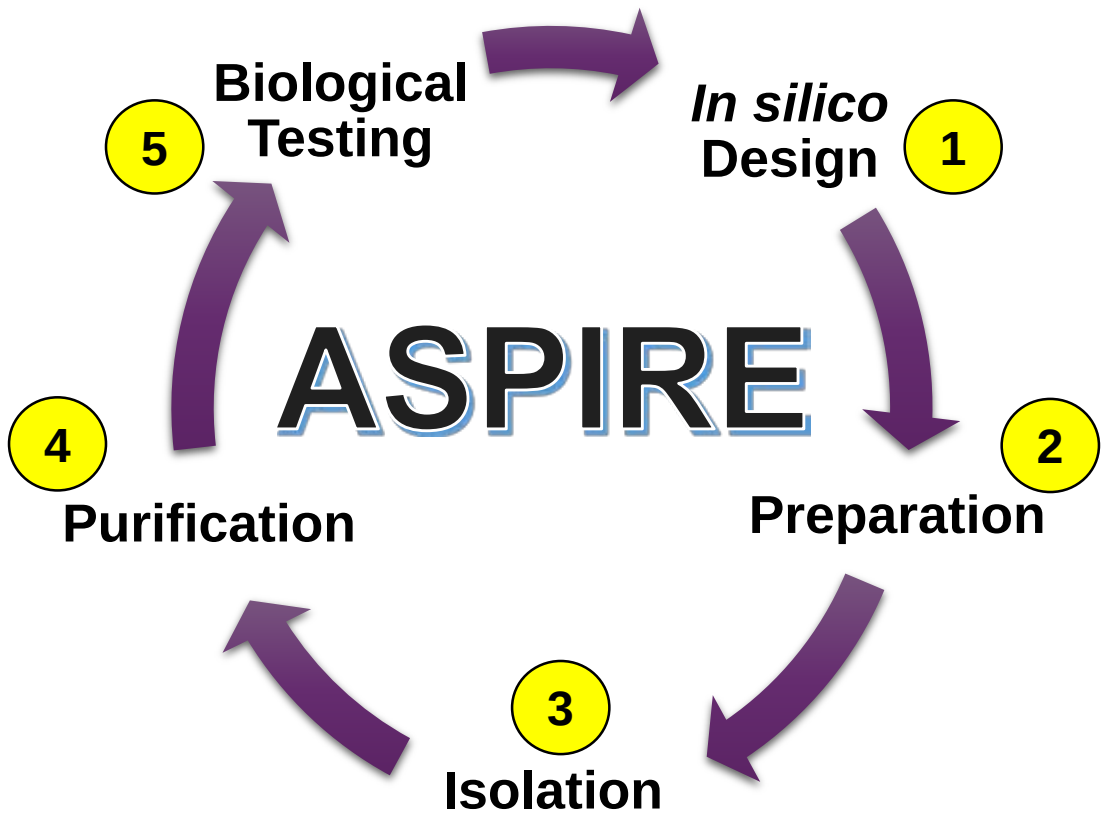
“Complete Situational Awareness”



Facilitated Development & Deployment Through Azure IOT Platform

An Evolution from Research in Silos to Research in Real-time!

ASPIRE – Automated, Closed Loop Processing



Process Detail

1 In silico Design
- Target Design
- Scored Synthetic Target(s)
- Retrosynthetic Analysis
- Route Selection
- Reaction Templating

2 Preparation
- Reaction Setup
- Reaction Control
- Reaction Monitoring
- Reaction Completion

3 Isolation
- Workup Selection
- Workup Platform(s)
- Purification Submission

4 Purification
- Method Selection
- Purification Platform(s)
- Fraction Handling
- Characterization
- Delivery (Storage Logistics)
- Reaction Outcome Reporting

5 Biological Testing
- Assay Selection
- Assay Platform(s)
- Data Collection / Analysis
- Knowledge Assessment
- New Hypothesis (next iteration)

A New Renaissance in Chemical Synthesis Automation
(In a Research Setting)

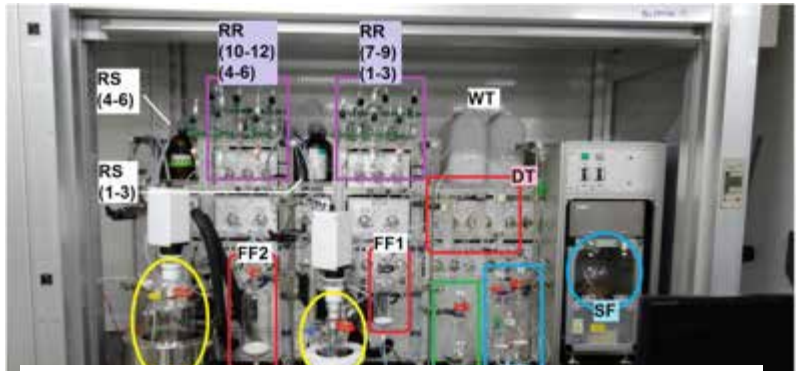
ASL
2009

Eli Lilly, Indianapolis, IN

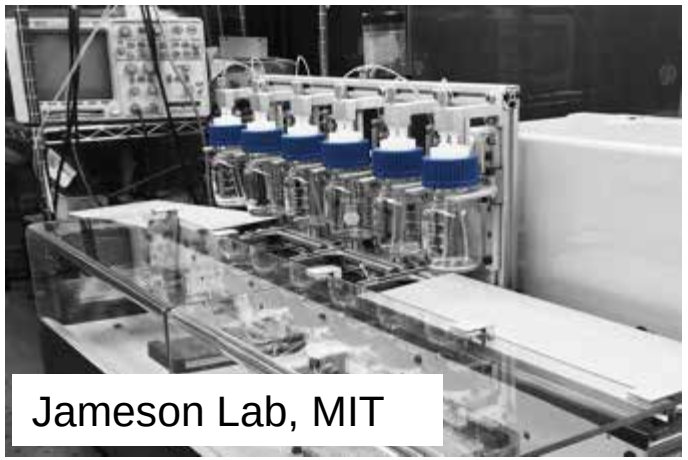
Lilly Life Sciences Studio (L2S2)
2020

Eli Lilly / Strateos, San Diego, CA

Additional Contemporary Examples

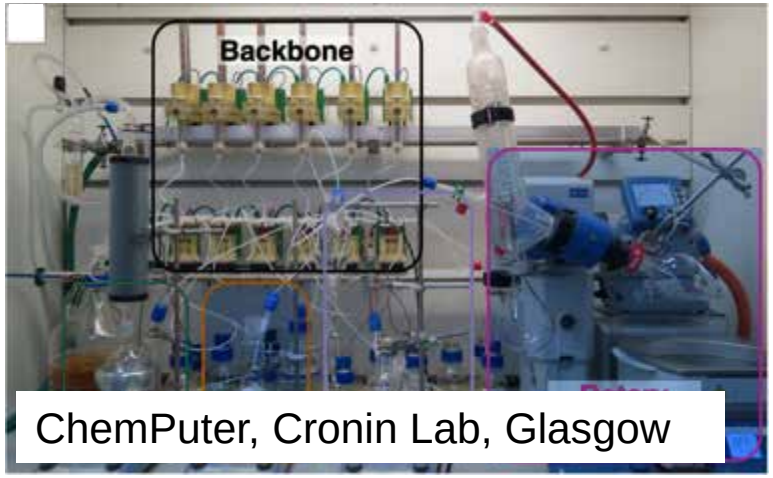


ChemKonzert, Takahashi Lab, Tokyo



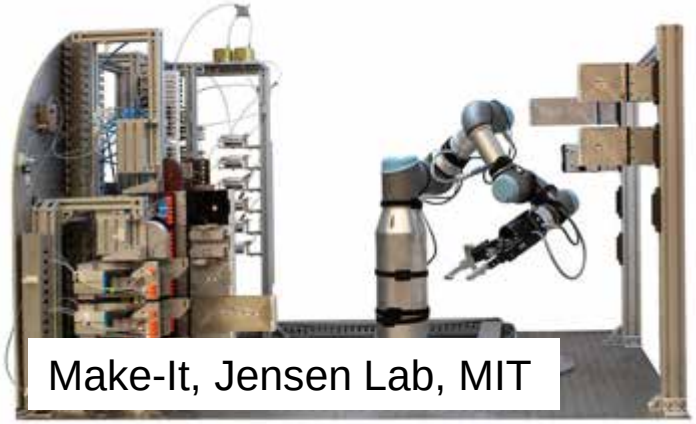
Jameson Lab, MIT

POD, Jensen Lab, MIT



ChemPuter, Cronin Lab, Glasgow

Radial Synthesis, Gilmore Lab, Max Planck Inst.



Make-It, Jensen Lab, MIT

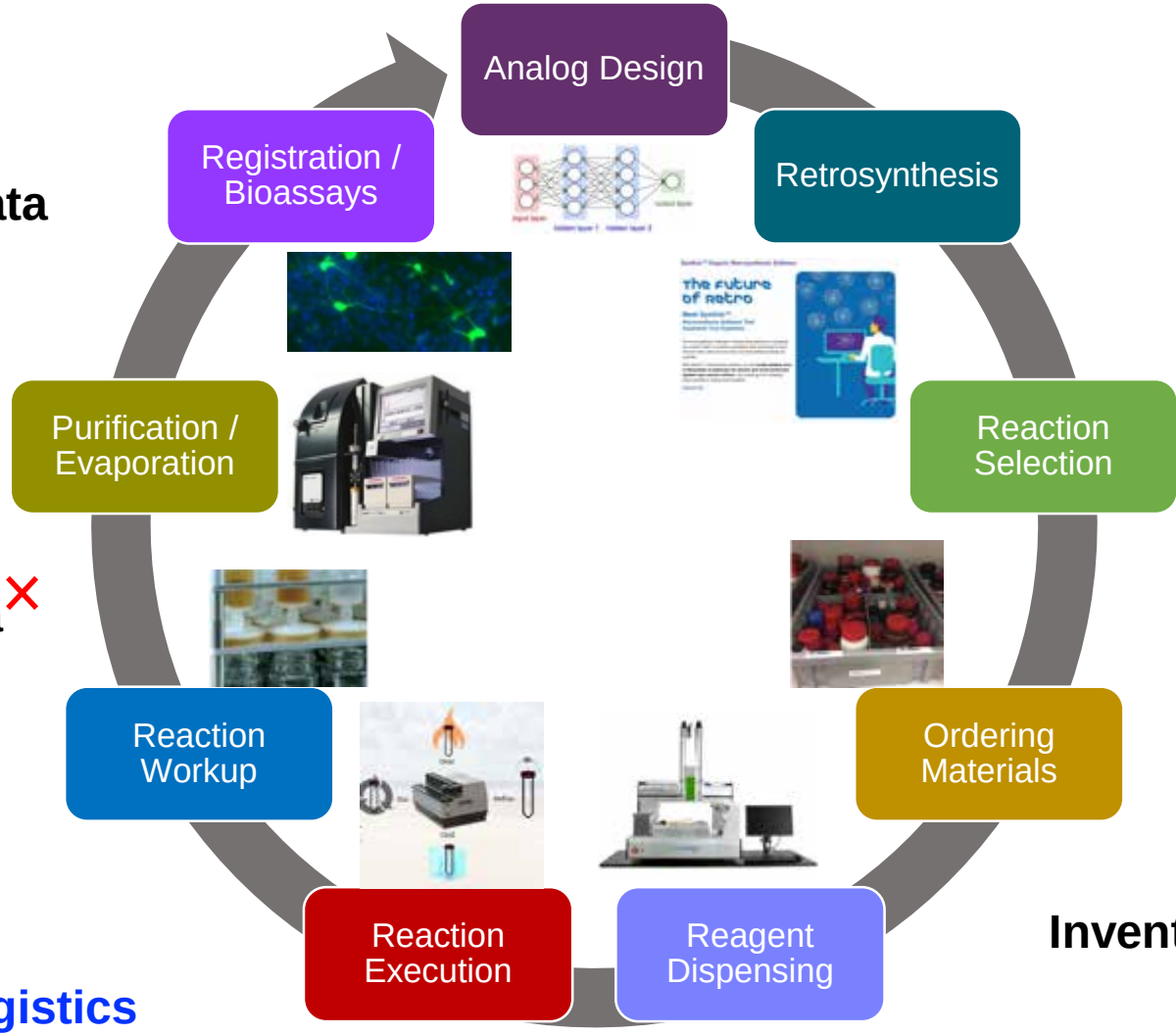
The Value of Integration

✓ **Biological Data**

ELN Data ✗

Analytical Data ✗

Inventory Data ✗

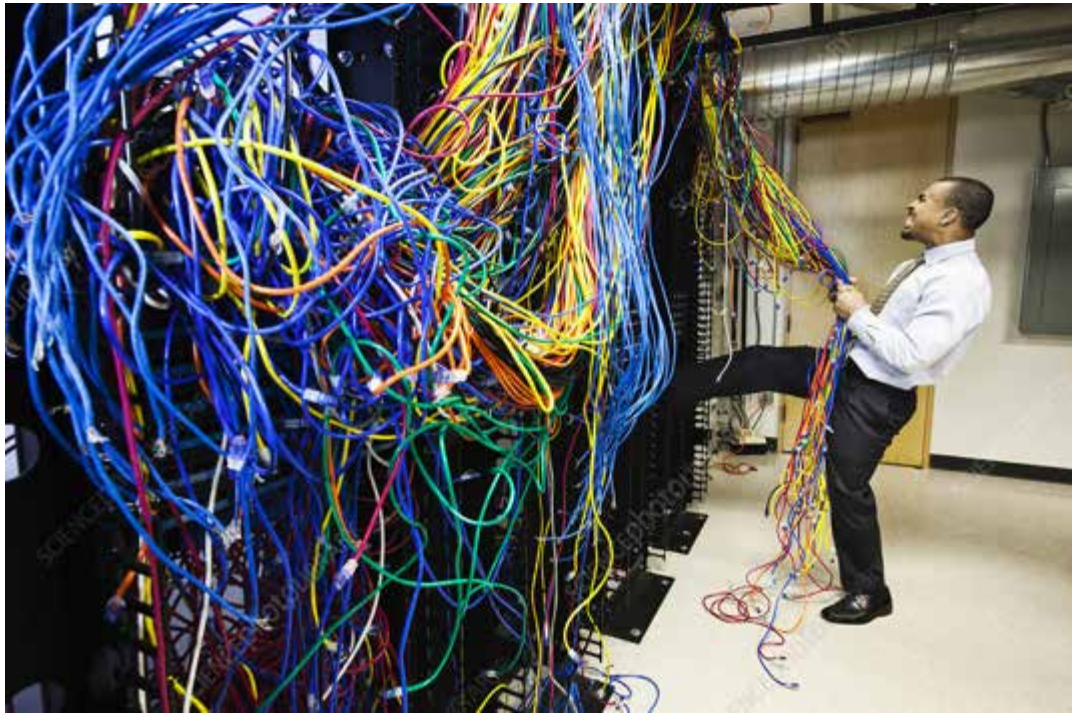


✓ **Value in improving efficiency in logistics**

✓ **Value in providing greater scientific connectivity & context**

Our Love-Hate Relationship with the Electronic Notebook

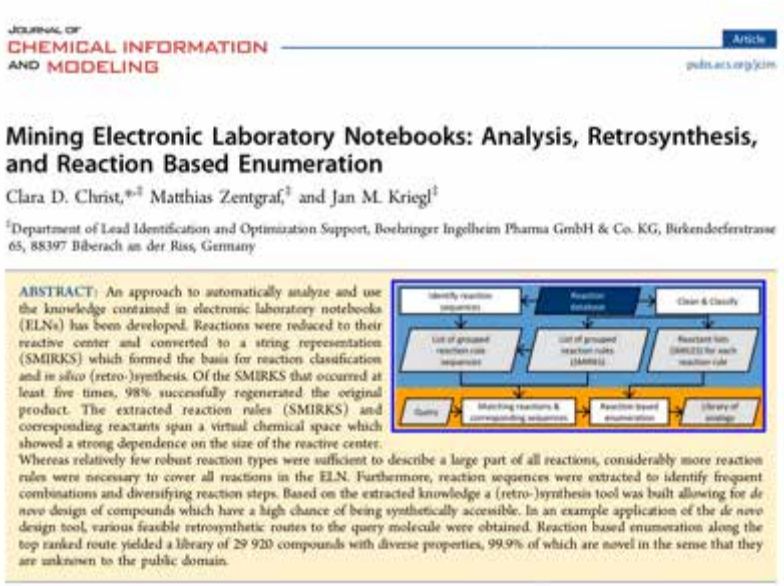
What we wish we had



What we really have

Lots of other folks have been looking at this problem as well

eLNs


JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration

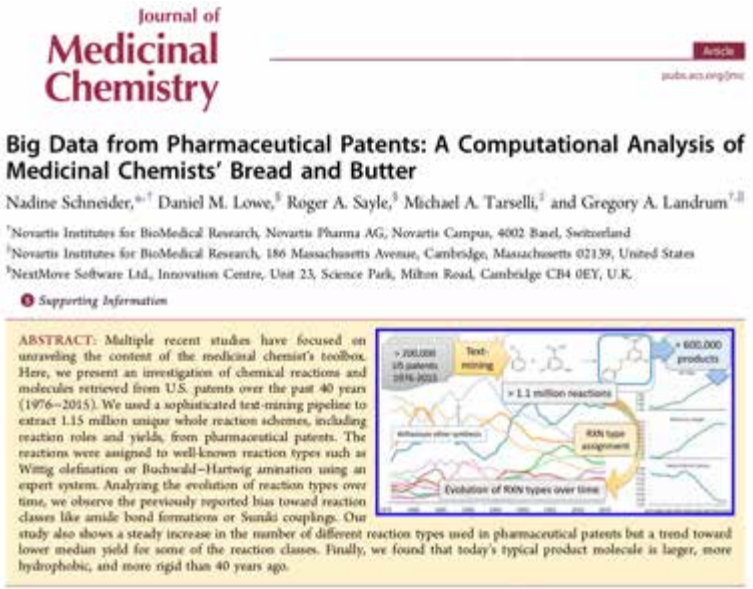
Clara D. Christ,*[†] Matthias Zentgraf,[‡] and Jan M. Kriegl[‡]

[‡]Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany

ABSTRACT: An approach to automatically analyze and use the knowledge contained in electronic laboratory notebooks (ELNs) has been developed. Reactions were reduced to their reactive center and converted to a string representation (SMIRKS) which formed the basis for reaction classification and *in silico* (retro-)synthesis. Of the SMIRKS that occurred at least five times, 98% successfully regenerated the original product. The extracted reaction rules (SMIRKS) and corresponding reactants span a virtual chemical space which showed a strong dependence on the size of the reactive center. Whereas relatively few robust reaction types were sufficient to describe a large part of all reactions, considerably more reaction rules were necessary to cover all reactions in the ELN. Furthermore, reaction sequences were extracted to identify frequent combinations and diversifying reaction steps. Based on the extracted knowledge a (retro-)synthesis tool was built allowing for *de novo* design of compounds which have a high chance of being synthetically accessible. In an example application of the *de novo* design tool, various feasible retrosynthetic routes to the query molecule were obtained. Reaction based enumeration along the top ranked route yielded a library of 29 920 compounds with diverse properties, 99.9% of which are novel in the sense that they are unknown to the public domain.

Christ, C. D.; Zentgraf, M.; Kriegl, J. M.,
J. Chem. Inf. Model. **2012,** *52* (7), 1745-1756.

Patents


Journal of Medicinal Chemistry

Article

pubs.acs.org/jmc

Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter

Nadine Schneider,*[†] Daniel M. Lowe,[‡] Roger A. Sayle,[‡] Michael A. Tarselli,[†] and Gregory A. Landrum*[†]

[†]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland

[‡]Novartis Institutes for BioMedical Research, 186 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

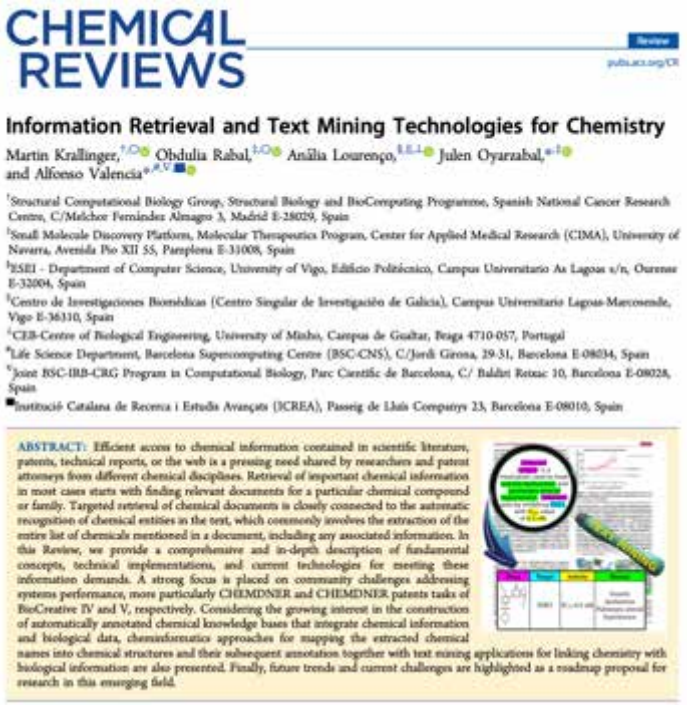
[§]NextMove Software Ltd., Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, U.K.

Supporting Information

ABSTRACT: Multiple recent studies have focused on unraveling the content of the medicinal chemist's toolbox. Here, we present an investigation of chemical reactions and molecules retrieved from U.S. patents over the past 40 years (1976−2015). We used a sophisticated text-mining pipeline to extract 1.15 million unique whole reaction schemes, including reaction roles and yields, from pharmaceutical patents. The reactions were assigned to well-known reaction types such as Wittig olefination or Buchwald−Hartwig amination using an expert system. Analyzing the evolution of reaction types over time, we observe the previously reported bias toward reaction classes like amide bond formations or Suzuki couplings. Our study also shows a steady increase in the number of different reaction types used in pharmaceutical patents but a trend toward lower median yield for some of the reaction classes. Finally, we found that today's typical product molecule is larger, more hydrophobic, and more rigid than 40 years ago.

Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A.,
J. Med. Chem. **2016,** *59* (9), 4385-4402.

Literature


CHEMICAL REVIEWS

Review

pubs.acs.org/CR

Information Retrieval and Text Mining Technologies for Chemistry

Martin Krallinger,[†,⊥] Obdulia Rabal,[‡,⊥] Anália Lourenço,[§,∥,⊥] Julen Oyarzabal,*[‡] and Alfonso Valencia*[#,▽]

[†]Structural Computational Biology Group, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, C/Melchor Fernández Almagro 3, Madrid E-28029, Spain

[‡]Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avenida Pío XII 55, Pamplona E-31008, Spain

[§]ESEI - Department of Computer Science, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, Ourense E-32004, Spain

[∥]Centro de Investigaciones Biomédicas (Centro Singular de Investigación de Galicia), Campus Universitario Lagoas-Marcosende, Vigo E-36310, Spain

[⊥]CEB-Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga 4710-057, Portugal

[#]Life Science Department, Barcelona Supercomputing Centre (BSC-CNS), C/Jordi Girona, 29-31, Barcelona E-08034, Spain

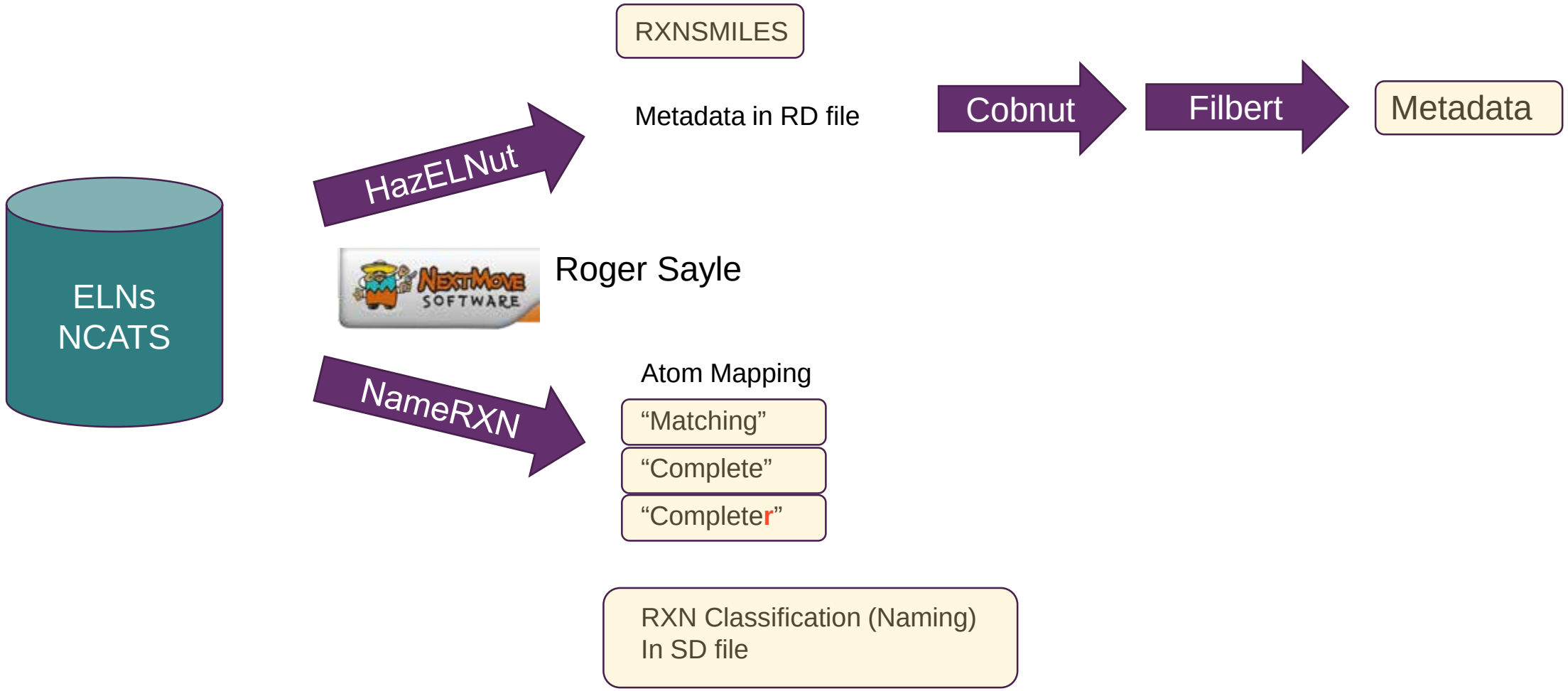
[▽]Joint BSC-IRB-CRG Program in Computational Biology, Parc Científic de Barcelona, C/ Baldiri Reixac 10, Barcelona E-08028, Spain

[■]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, Barcelona E-08010, Spain

ABSTRACT: Efficient access to chemical information contained in scientific literature, patents, technical reports, or the web is a pressing need shared by researchers and patent attorneys from different chemical disciplines. Retrieval of important chemical information in most cases starts with finding relevant documents for a particular chemical compound or family. Targeted retrieval of chemical documents is closely connected to the automatic recognition of chemical entities in the text, which commonly involves the extraction of the entire list of chemicals mentioned in a document, including any associated information. In this Review, we provide a comprehensive and in-depth description of fundamental concepts, technical implementations, and current technologies for meeting these information demands. A strong focus is placed on community challenges addressing systems performance, more particularly CHEMDNER and CHEMDNER patents tasks of BioCreative IV and V, respectively. Considering the growing interest in the construction of automatically annotated chemical knowledge bases that integrate chemical information and biological data, cheminformatics approaches for mapping the extracted chemical names into chemical structures and their subsequent annotation together with text mining applications for linking chemistry with biological information are also presented. Finally, future trends and current challenges are highlighted as a roadmap proposal for research in this emerging field.

Krallinger, M.; Rabal, O.; Lourenco, A.; Oyarzabal, J.; Valencia, A.,
Chem. Rev. **2017,** *117* (12), 7673-7761.

ELN Extraction - Raw Output



Reaction Representation

- Reaction SMILES

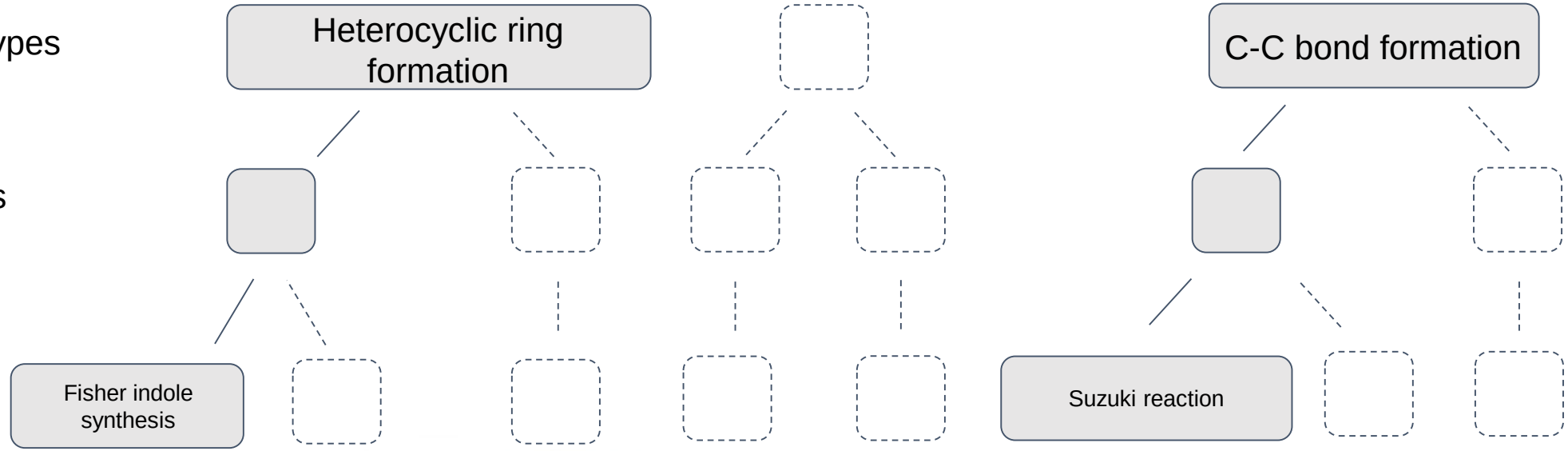
OCC.CC(=O)O>[H+].[Cl-].OCC>CC(=O)OCC |f:2.3|

Ref: https://www.daylight.com/meetings/summerschool01/course/basics/smirks.html

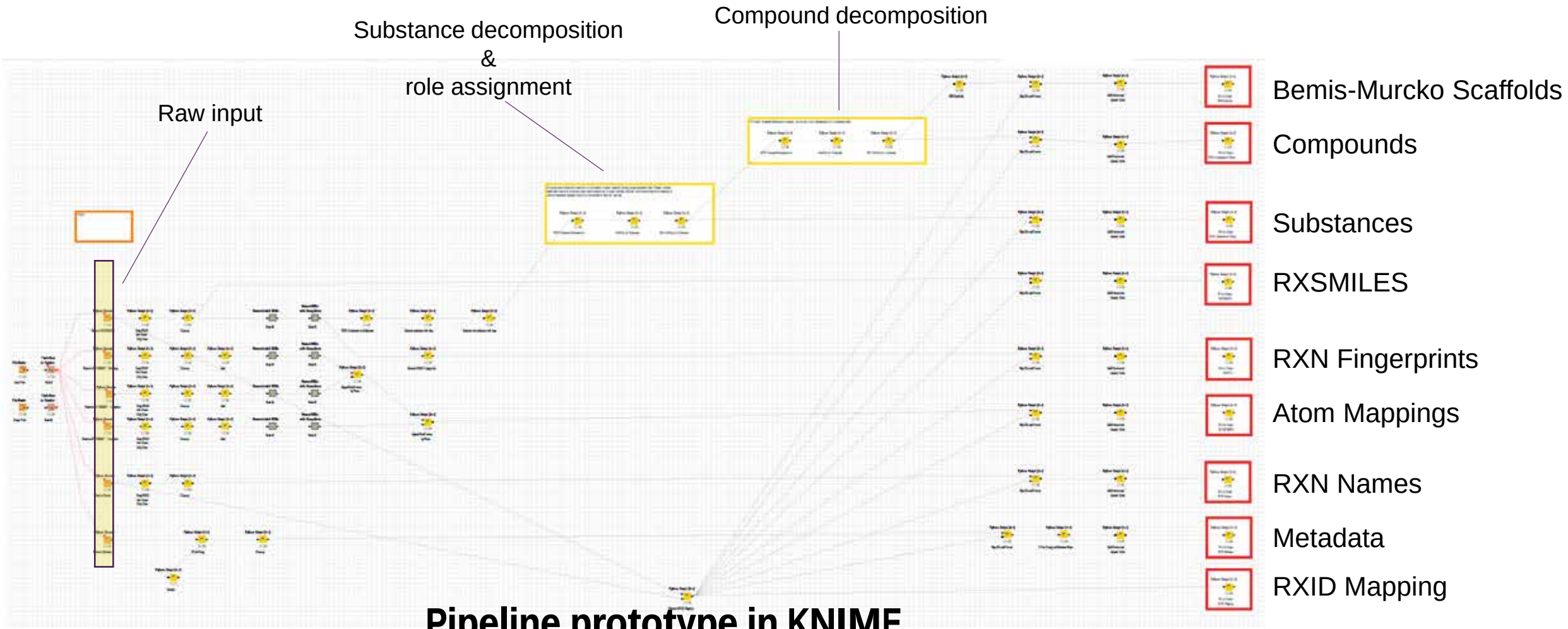
Reaction Ontology

- RXNO reaction ontology invented by Royal Society of Chemistry
- Ontology available at: https://github.com/rsc-ontologies/rxno
- Implemented by NextMove
- Can be utilized in curation and data mining

- 12 basic types
- 84 classes
- 500+ named reactions



Data Post-processing



Pipeline prototype in KNIME

Overall Process

ELN Extraction (~90K expts)

Data Curation (~16K expts)

Data Processing

Palantir

Database Upload

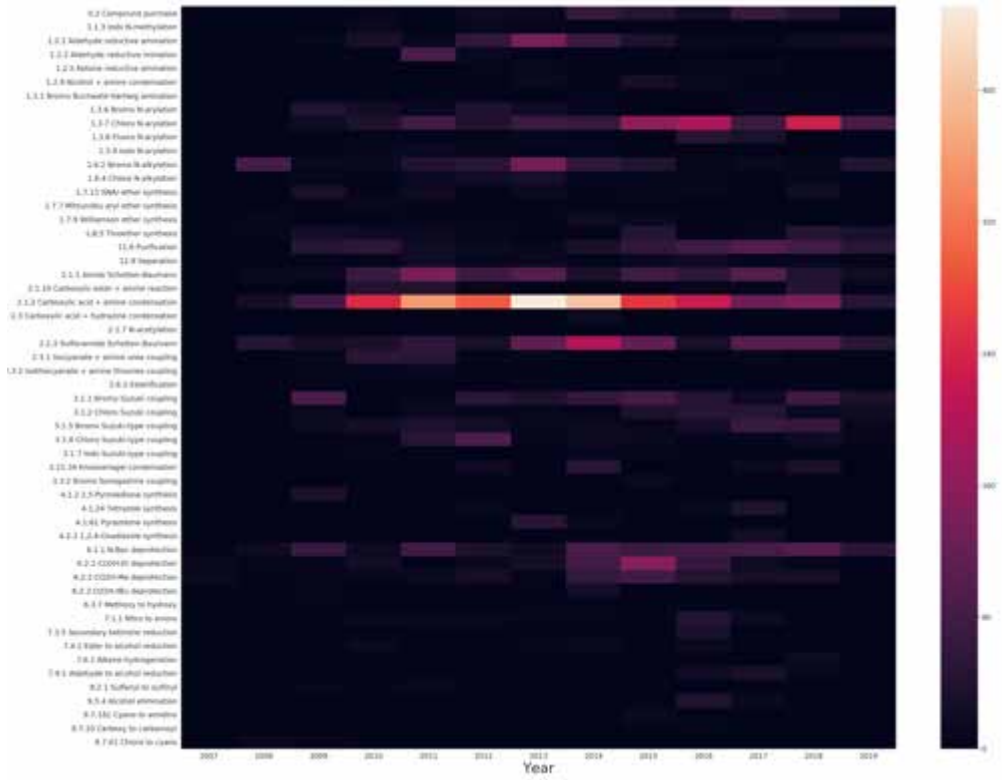
NCATS RXN

Reaction Analytics Dashboard

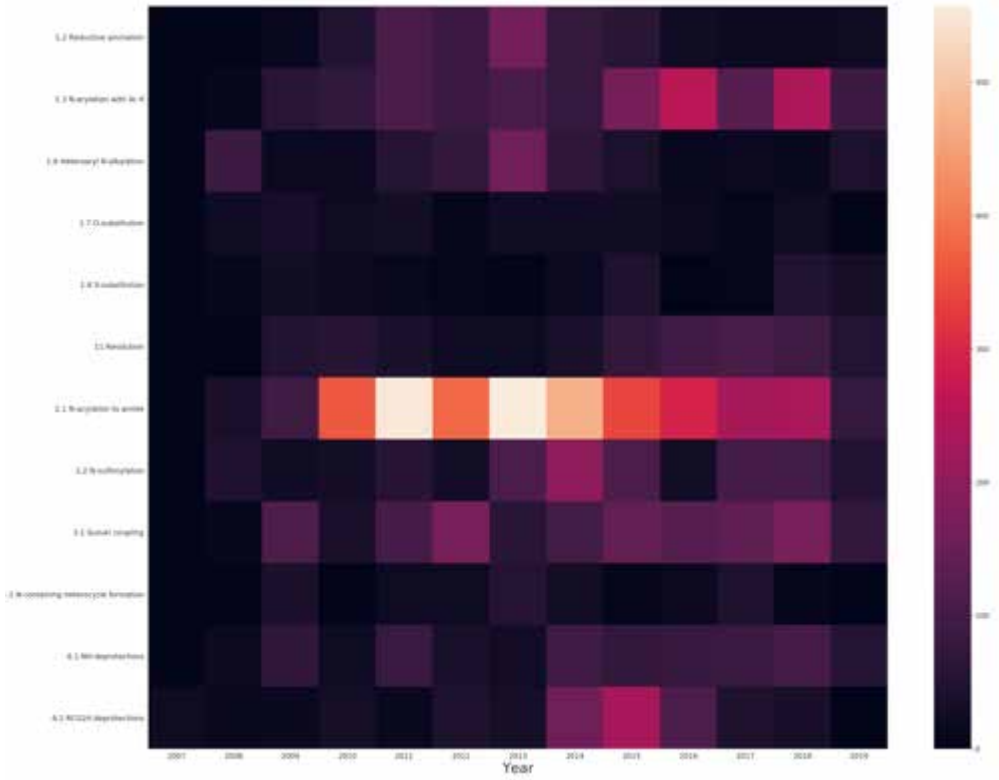
Some Preliminary Observations

Most Frequently Utilized Reactions per Year

Reaction Name (Top 25% of 216)

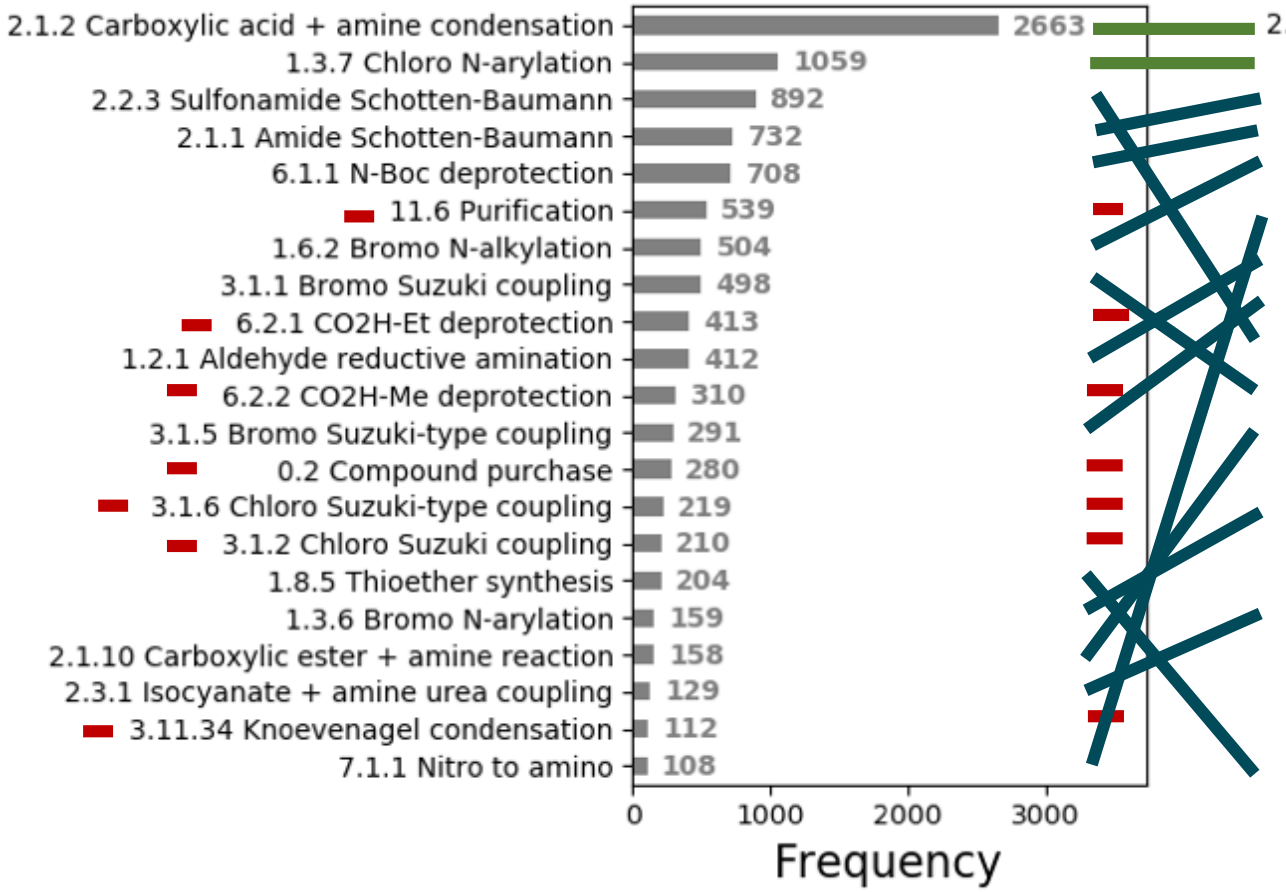


Reaction Class (Top 25% of 50)

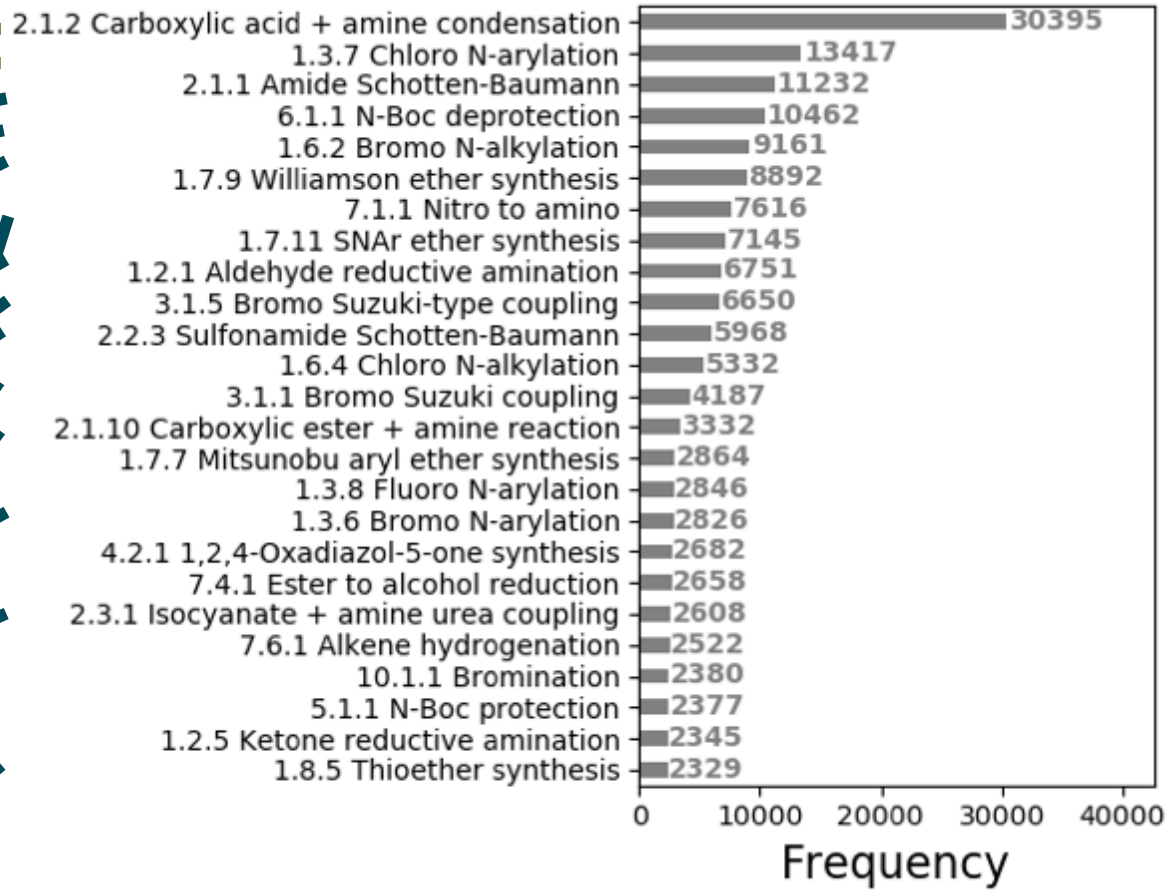


Comparison of Name Reaction Frequencies

Top 10% of 210 Reactions - ASPIRE



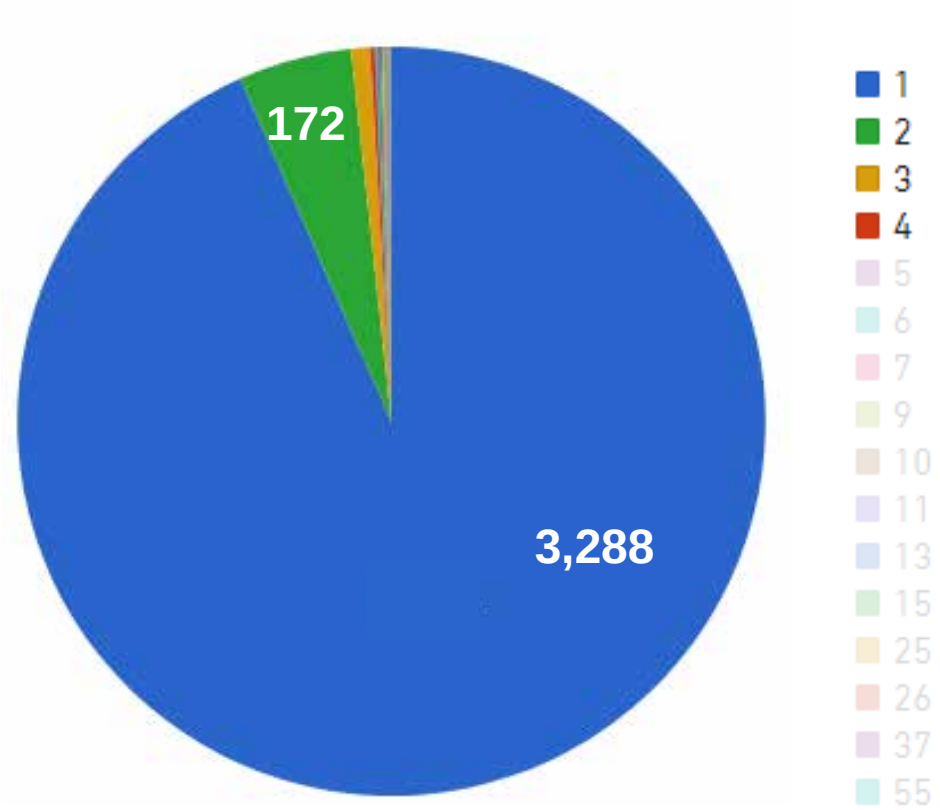
Top 5% of 505 Reactions - USPTO



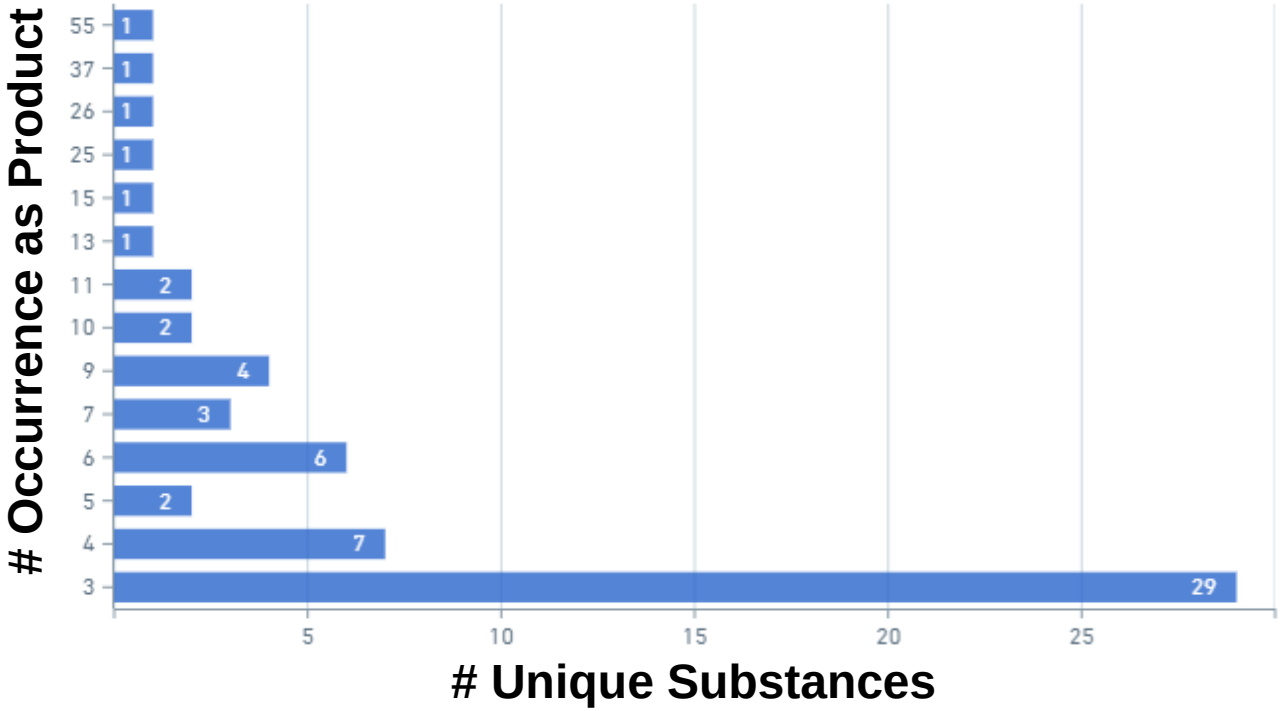
Substance Synthesis Frequency Distribution

Closed Reactions Only (~4K expts)

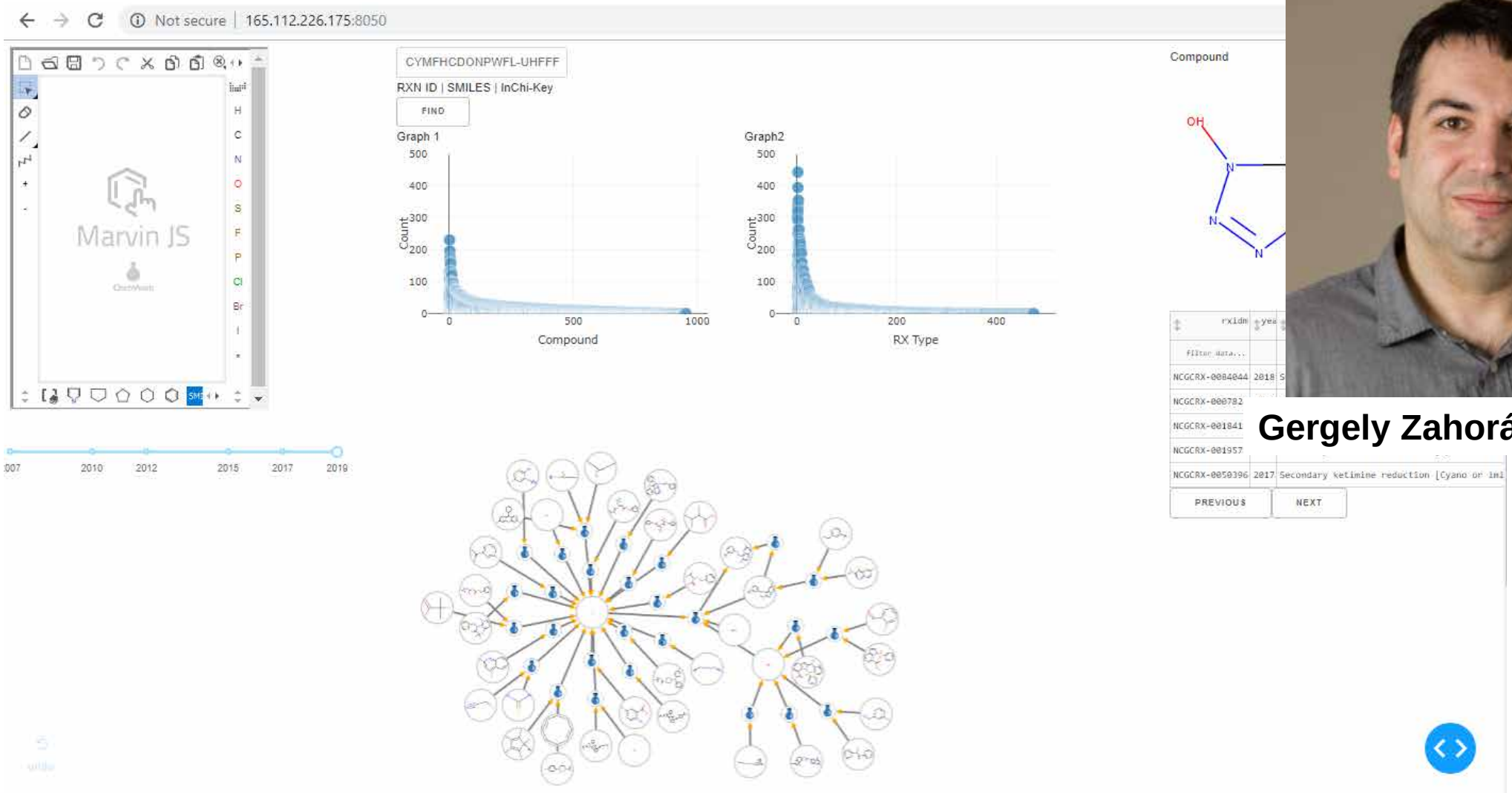
Occurrence as product >= 1



Occurrence as product >= 3

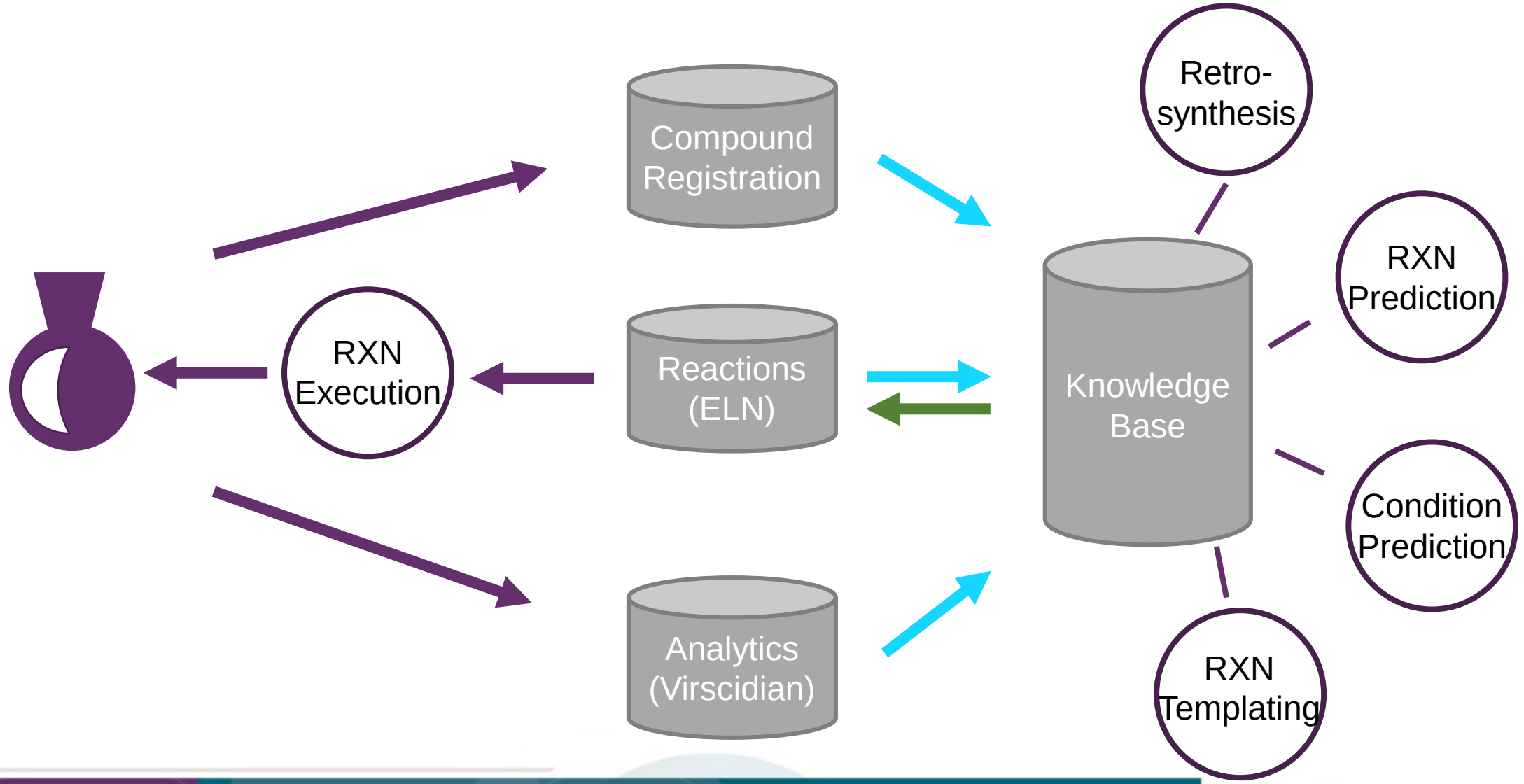


Reaction Analytics Dashboard Web App Prototype



Gergely Zahoránszky-Köha

Computational Platform to Support Chemistry Automation



25

Reaction Analytics – From Information to Actionable Knowledge

Frequency Distributions:

- Reaction Type (overall or by year)
- Reagent use by name or type
- Reagent use by quantity
- Catalyst use (most common vs least common)
- Catalyst use (by yield)
- Reactant use

Reaction Networks:

- Average or median length of synthesis
- Longest synthetic sequence
- Reaction types represented / NOT represented
- Reactant diversity analysis
  - Total count & frequency
  - Performance by functionality

Cross Comparisons with Inventory:

- Reactants infrequently used or never used
- Use of rare (long lead time) reactants

Productivity Metrics:

- Product production cycle times
- Cycle time by reaction type
- Yield Distributions

26

Reaction Analytics by Platform (i.e. Target Space)

Reaction Networks:

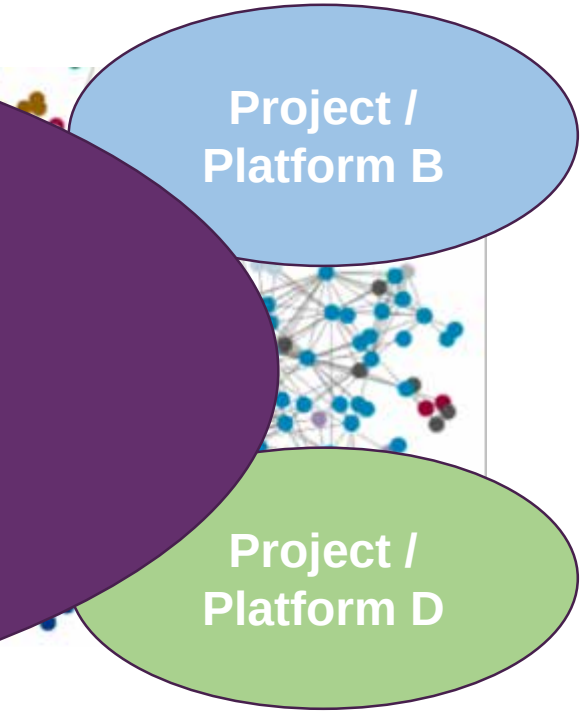
- Average o
- Longest
- Reactio
- Reactan
  - Tota
  - Perform

How is our reaction space evolving over time?

How would we like to see it evolve?

Where can we intervene most effectively?

What opportunities might we be overlooking?



27

Reaction Templating

“How do we get reactions to automatically set themselves up?”

Think: Object + [Role] + Properties ➡ Robotic Process

Features:

1) Reaction Ontology
2) Complete Reaction Annotation (“No atom left behind”)
3) Standardized Reaction Equations (scale-independent)
4) Machine-Interpretable Temporal Instructions

The Work Ahead

- Clearly defined annotation strategy for benchmark reactions
- Test how discretized scale factors will work in practice
- Complete the reaction template lexicon / graph
- Build and test the hardware specification map
- Detail module design and construction
- Integrate-test, integrate-test, integrate-test…

Thank you!

Work presented here was supported in part by the Intramural/Extramural research program of the NCATS, NIH