Center for Computer Assisted Synthesis

FAIRies, Ghosts and Trolls: Data Challenges in the Age of AI

Open – Access & FAIR Data Workshop

NASEM, 02/21-22/2024

owiest@nd.edu

FAIRies in the age of AI

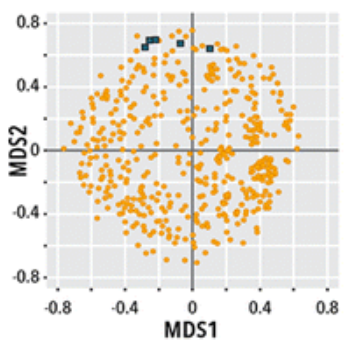
Findability
Accessibility
Interoperability
Reuse

Sustainable → funding models that demonstrate value

Qualitity
→ Trustworthy &Transparent
→ Curated
→ Known uncertainties
→ Complete & consistent

Precompetitive → incentivize deposition/usage of data by many stakeholders

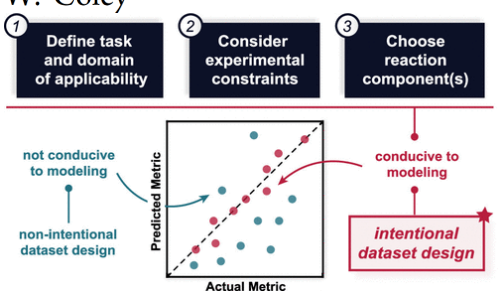
Design



Dataset Design for Building Models of Chemical Reactivity

Priyanka Raghavan, Brittany C. Haas,[⊥] Madeline E. Ruos,[⊥] Jules Schleinitz,[⊥] Abigail G. Doyle, Sarah E. Reisman, Matthew S. Sigman, and Connor W. Coley*

Cite This: *ACS Cent. Sci.* 2023, 9, 2196–2204



Brown Lady of Raynham Hall
H.C. Provand, *Country Life, 1936*

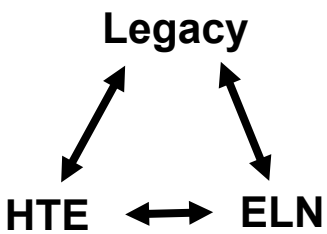
Ghosts

descriptions of ghosts vary widely, from an invisible presence to translucent or barely visible wispy shapes to realistic, lifelike forms.

https://en.wikipedia.org/wiki/Ghost, accessed 2/15/2024

When is data not data ?

- Inferred data
- Implicit data
- **Explicit data**



JOC The Journal of Organic Chemistry

pubs.acs.org/joc

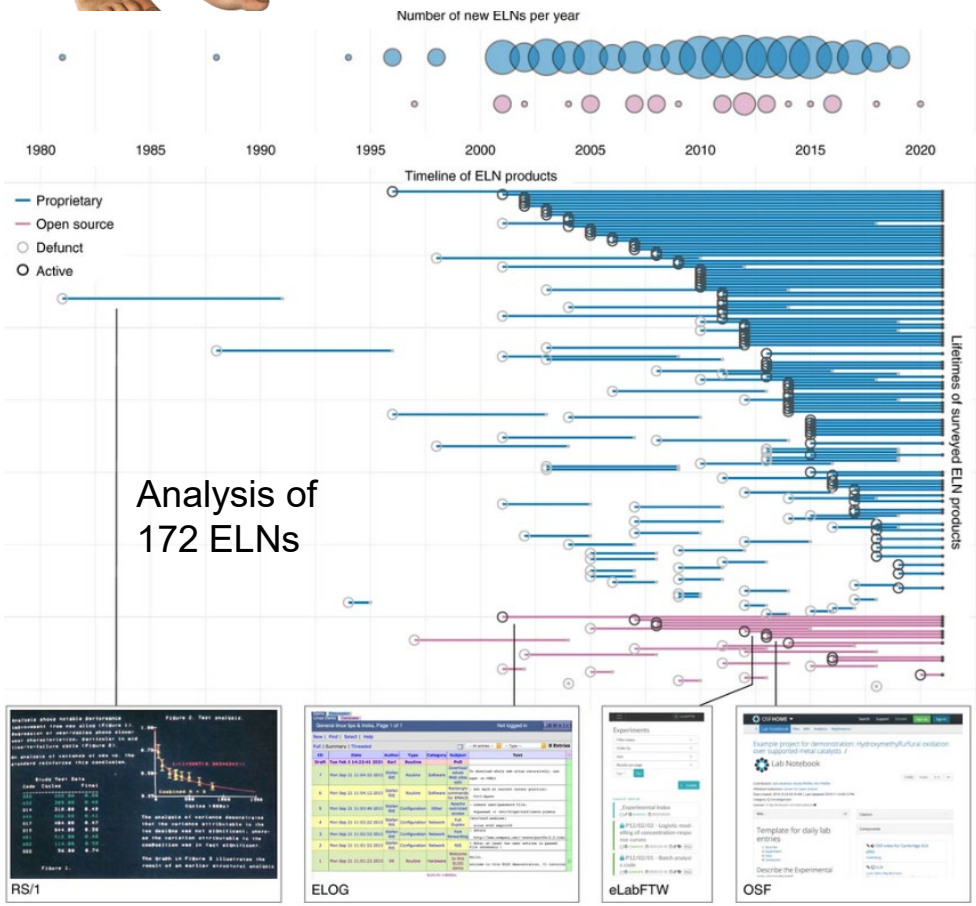
Negative Data in Data Sets for Machine Learning Training

Cite This: *J. Org. Chem.* 2023, 88, 5239–5241 Read Online

- Isolated vs. crude reaction yields
- Other measures used as proxies of yield, especially in HTE.
- A mandatory conclusion before closing the experiment, which could take the form of a drop-down menu in an ELN, with the following options:
  A. Significant amount of product was detected (success)
  B. No significant product was detected, but starting material remains
  C. Neither starting material nor intended product was detected
  D. The reaction was not run as intended (incorrect setup, physical error, reaction cancelled, other). In this case, a free-text comment describing the observation would be beneficial, but not essential.

Trolls

In Old Norse sources, beings described as trolls dwell in isolated areas […] may be ugly and slow-witted, and are rarely helpful or even dangerous to human beings.
https://en.wikipedia.org/wiki/Troll, accessed 2/10/24



- Extensive use in industry
- Not widely adopted in academia

- Most are proprietary
- Formats not interoperable
- Limited lifetime

- Data entries inconsistent/incomplete
- Contradictory data
- Missing data
- Significant data in unstructured text
- Different languages/units/formats

Higgins et al. *Nature Protocols* **2022**, *17*, 179

Center for Computer Assisted Synthesis

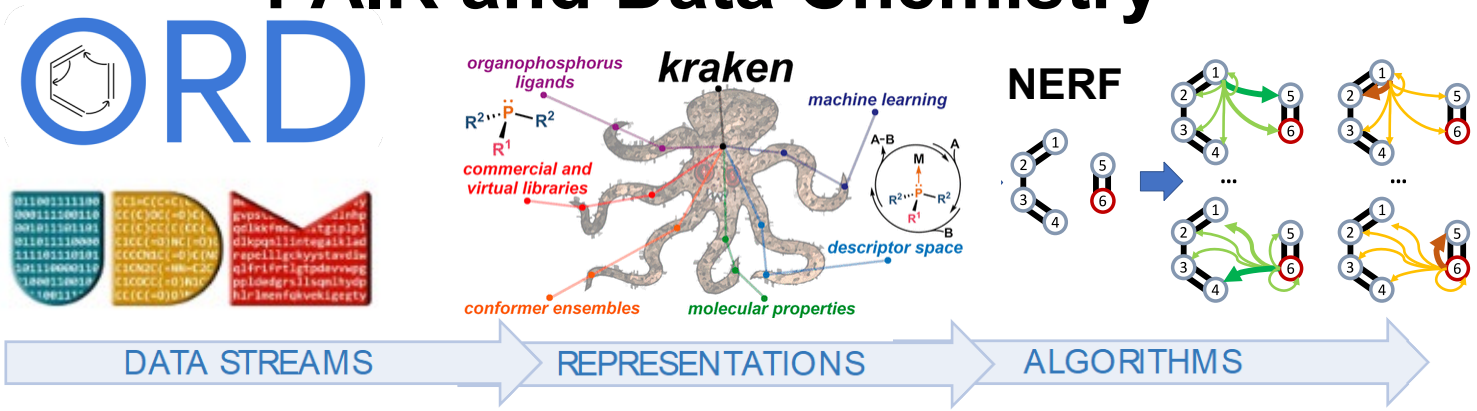
Vision

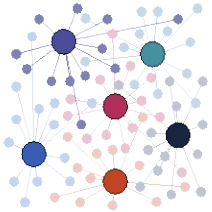
To transform how chemists discover, optimize, interrogate, and apply new reactions to the synthesis of functional molecules through “data chemistry”

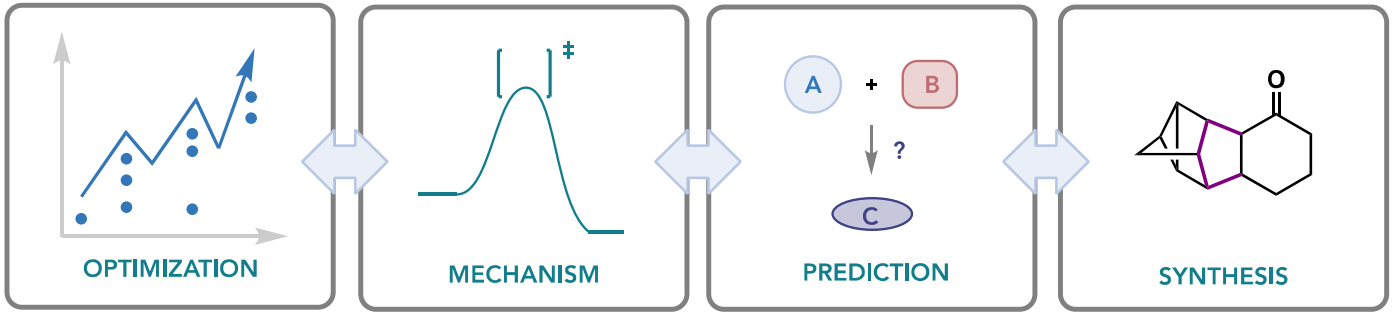
- **Phase II NSF Center for Chemical Innovation**
- **Flagship program of NSF Chemistry division**
- **Seven Phase II CCIs in US**
- **Building on Phase I, 9/2019-8/2022**
- **Phase II started Sept 1, 2022**
- **$20M over 5 years, renewable once**
- **Additional funding from industry**

ccas.nd.edu **@NSF_ccas**

FAIR and Data Chemistry







EDBO+

Bayesian reaction optimization as a tool for chemical synthesis

Build

Upload

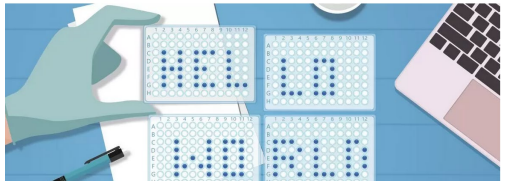
Delete

Download

Optimize Reaction

ASKCOS/**ASKCOS**

Software package for computer aided synthesis planning



CoScientist

The Open Reaction Database

1. Provide a structured data format for chemical reaction data
2. Provide an interface for easy browsing and downloading of data
3. Make reaction data freely and publicly available for anyone to use
4. Encourage sharing of precompetitive proprietary data

SCHMIDT FUTURES

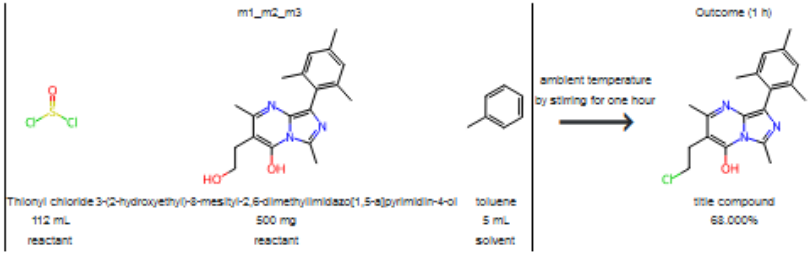
Table 1. Example Data Sets Currently Available in the ORD

category	description	ref	size
single-step batch	deoxyfluorination reaction screening as a function of substrate, base, and fluoride source (entries in Figure 1)	3	80
single-step batch	microwave synthesis of a small library using the Biginelli multicomponent condensation reaction	4	48
kinetic profiling	online monitoring of a Suzuki coupling reaction by HPLC	5	7
high throughput	subset of "chemistry informer" screen of copper-catalyzed Buchwald–Hartwig aminations (entries 11–15 in Figure 4)	6	90
high throughput	C–N cross-coupling reaction yields varying aryl halide, additive, Pd catalyst, and base identities	7	4312
high throughput	Suzuki coupling reaction performance as a function of aryl halide, boronic acid, ligand, base, and solvent performed under pseudoflow conditions	8	5760
high-throughput	C–N cross-coupling reaction performance of 3-bromopyridine with various nucleophiles, varying precatalysts and bases (entries in Experiment 2)	9	1536
high throughput	combinatorial nanochemistry screen of a complex aryl halide library using dual-metal photoredox C–N coupling (entries in Figure 6)	10	1728
photochemistry	substrate scope tables regarding coupling of α-carboxyl sp^3 carbons with aryl halides	11	24
photochemistry	Ir-catalyzed debromination conversions as a function of photocatalyst ligands	12	1152
electrochemistry	electroreductive coupling of alkenyl and benzyl halides via nickel catalysis (entries in Figures 2 and 3)	13	27
flow chemistry	sulfonamide library synthesis in flow	14	39
enzymatic	multistep biocatalytic cascade for the manufacture of islatravir	15	3
multistep	copper-catalyzed enantioselective hydroamination of alkenes	16	3
literature extracted	reactions extracted by text-mining United States published patents; imported from CML documents	1	1771017

https://open-reaction-database.org/

Kearnes et al. *JACS* **2021**, *143,* 18820

Summary



Identifiers

REACTION_CXSMILES S(Cl)([Cl:3])=O.O[CH2:6][CH2:7][C:8]1[C:9]([CH3:28])=[N:10][C:11]2[N:12]([C:15]([CH3:27])=[N:16][C:17]=2[C:18]2[C:23]([CH3:24])=[CH:22][C:21]([CH3:25])=[CH:20][C:19]=2[CH3:26])[C:13]=1[OH:14]>C1(C)C=CC=CC=1>[Cl:3][CH2:6][CH2:7][C:8]1[C:9]([CH3:28])=[N:10][C:11]2[N:12]([C:15]([CH3:27])=[N:16][C:17]=2[C:18]2[C:23]([CH3:24])=[CH:22][C:21]([CH3:25])=[CH:20][C:19]=2[CH3:26])[C:13]=1[OH:14]

Inputs

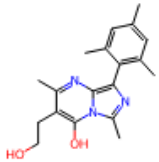
m1_m2_m3

Details

Addition Order 0

Components

Compound	Amount	Role	Raw
	112 milliliter	reactant	⇔

Compound	Amount	Role	Raw
	500 milligram	reactant	⇔

Compound	Amount	Role	Raw
	5 milliliter	solvent	⇔

Conditions

Temperature | Stirring | Other

Control Type AMBIENT

Setpoint None

Notes

Procedure details Thionyl chloride (112 mL, 1.54 mmol) was added to a solution of 3-(2-hydroxyethyl)-8-mesityl-2,6-dimethylimidazo[1,5-a]pyrimidin-4-ol (500 mg, 1.54 mmol) in toluene (5 mL) at 80° C., followed by stirring for one hour. After cooling to room temperature, the resulting crystals were washed with diethyl ether, to give the title compound (360 mg) as pale brown crystals.

- **Facilitate data intake**
  - Interfaces to ELNs etc.
  - Automated curation
- **Prioritize data over format**
  - Flexible, structured fields
- **Interface to other format/data**
  - UDM
  - Free to use

ORD

Summary

Identifiers

Inputs

Conditions

Notes

Workups

Outcomes

Provenance

Full Record

FAIRify Ghosts and Trolls

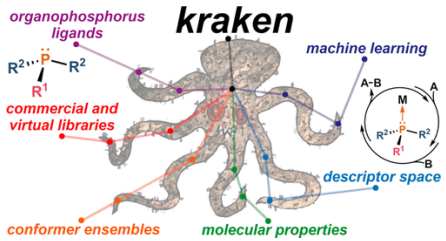
UNITED STATES
PATENT AND TRADEMARK OFFICE

uspto

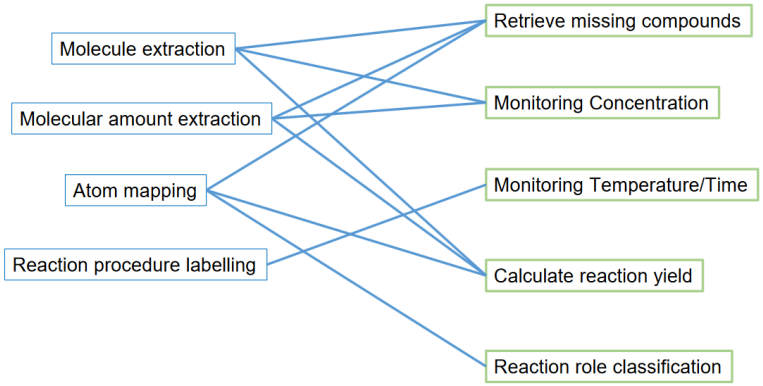
SI Supporting Info (1) »

CAS SciFinder^n

Reaxys®



ORD



Model	GPT 3.5	GPT4
Correctness	89.34%	100.00%
Missing rate	12.25%	3.57%