

Integrated Computational Platform for Chemistry Automation

Gergely Zahoranszky-Kohalmi

Informatics Lead, ASPIRE

Samuel G. Michael

Branch Chief, ITRB

G. Sitta Sittampalam

Senior Advisor to the Director

Alexander G. Godfrey

Lead, ASPIRE

Ultra-Large Chemistry Databases NIH Workshop

Wednesday, December 2nd, 2020

Outline

- ASPIRE
- Integrated Computational Platform
- ELN Data Curation
- HCASE chemical space embedding method
- What We Need...



ASPIRE



[Site Map](#) | [Contact](#)

[Research](#) [Funding & Notices](#) [News & Media](#) [About Translation](#) [About NCATS](#)

[Home](#) > [About NCATS](#) > [NCATS Programs & Initiatives](#) > A Specialized Platform for Innovative Research Exploration (ASPIRE)

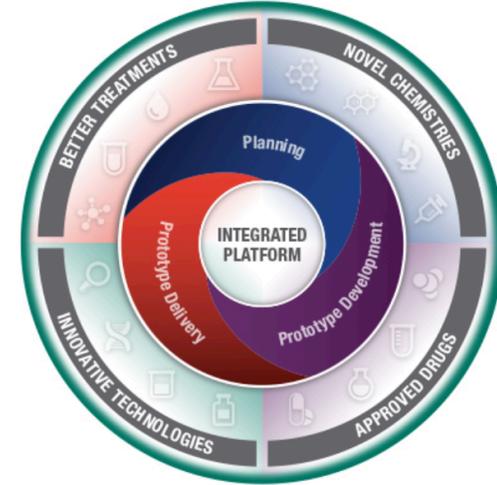
A Specialized Platform for Innovative Research Exploration (ASPIRE)

By addressing long-standing challenges in the field of chemistry, including lack of standardization, low reproducibility and inability to predict how new chemicals will behave, ASPIRE is designed to bring novel, safe and effective treatments to more patients more quickly at lower cost. [Learn More.](#)



Apply: ASPIRE Reduction-to-Practice Challenge Accepting Submissions Starting Nov. 30

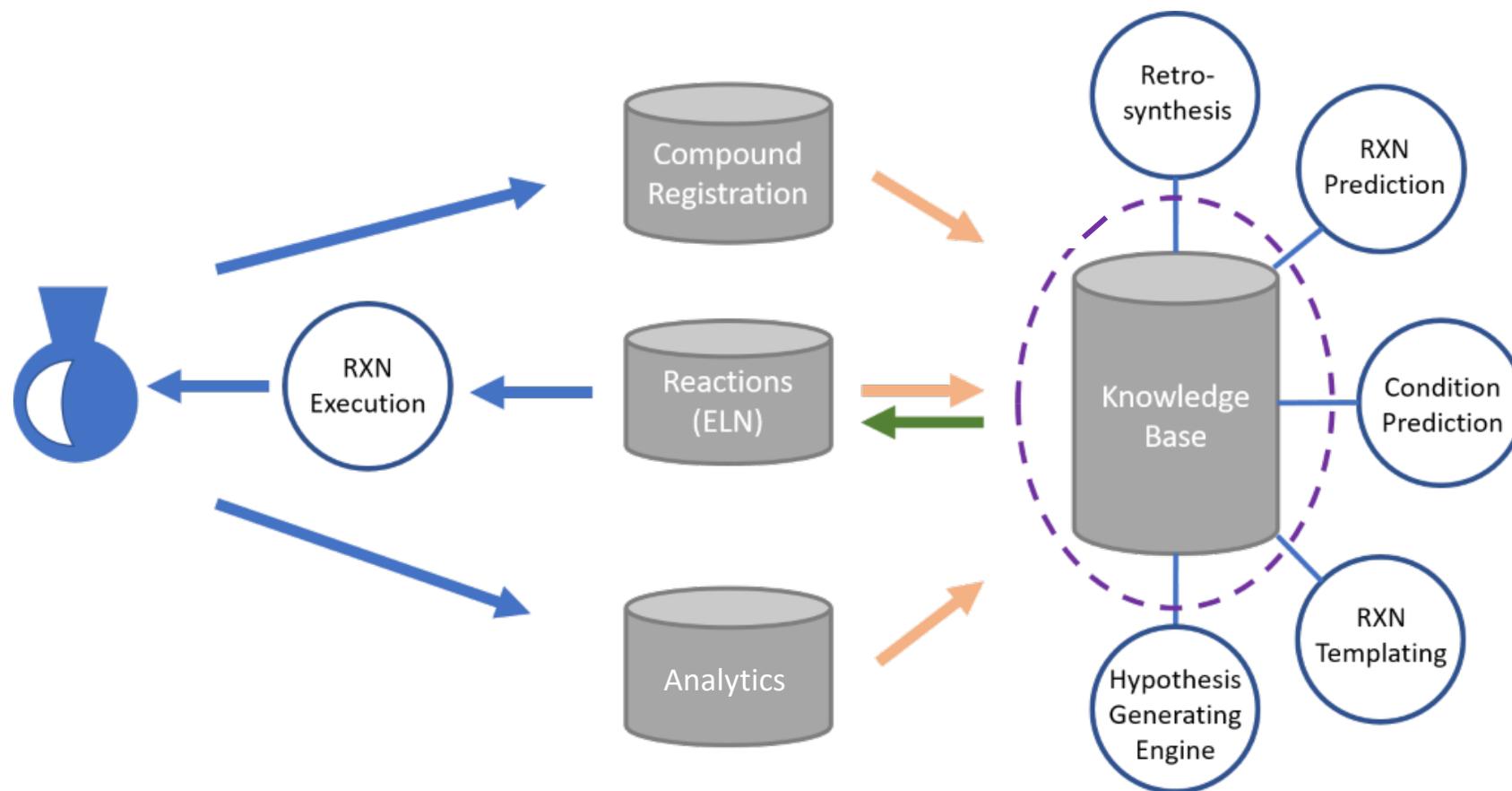
The NCATS ASPIRE Reduction-to-Practice Challenge is open for submissions. ASPIRE is supported through the Helping to End Addiction Long-term InitiativeSM, or NIH Heal InitiativeSM. [▶](#)



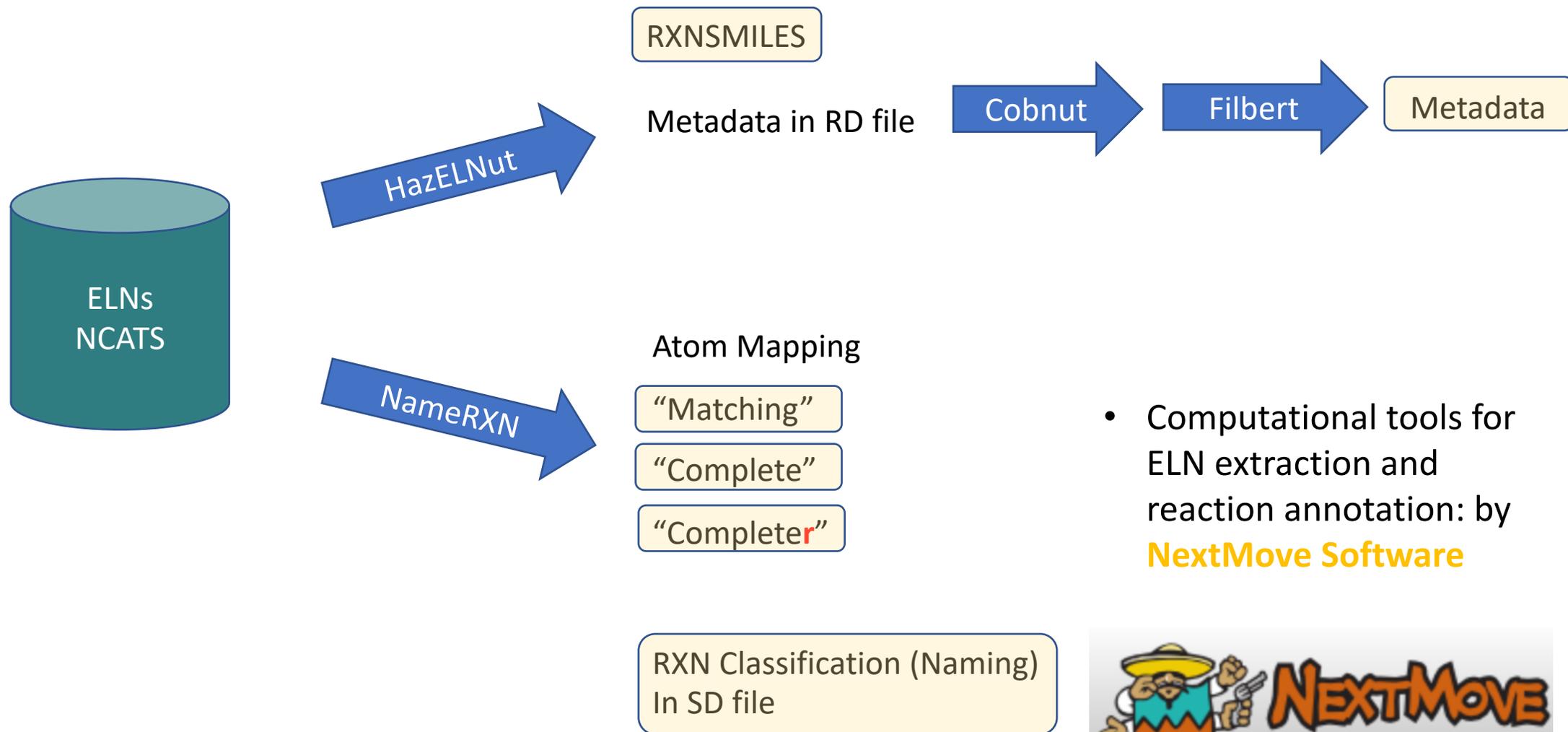
National Center for Advancing Translational Sciences

<https://ncats.nih.gov/aspire>

ASPIRE Integrated Computational Platform (AICP)



ELN Extraction - Raw Output



- Computational tools for ELN extraction and reaction annotation: by **NextMove Software**



Image credit: <https://www.nextmovesoftware.com/>



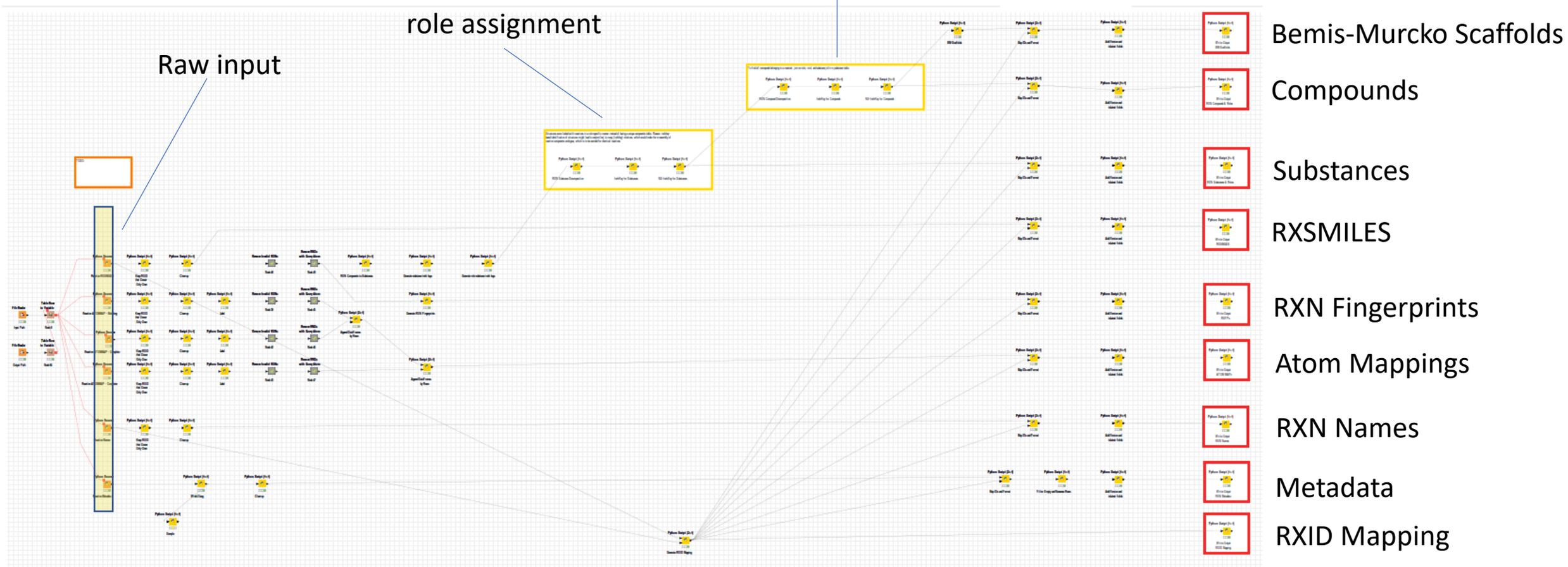
National Center
for Advancing
Translational Sciences

Data Post-Processing

Substance decomposition
&
role assignment

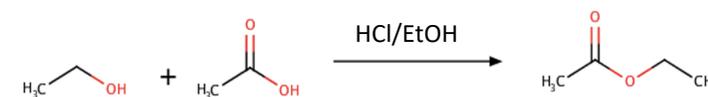
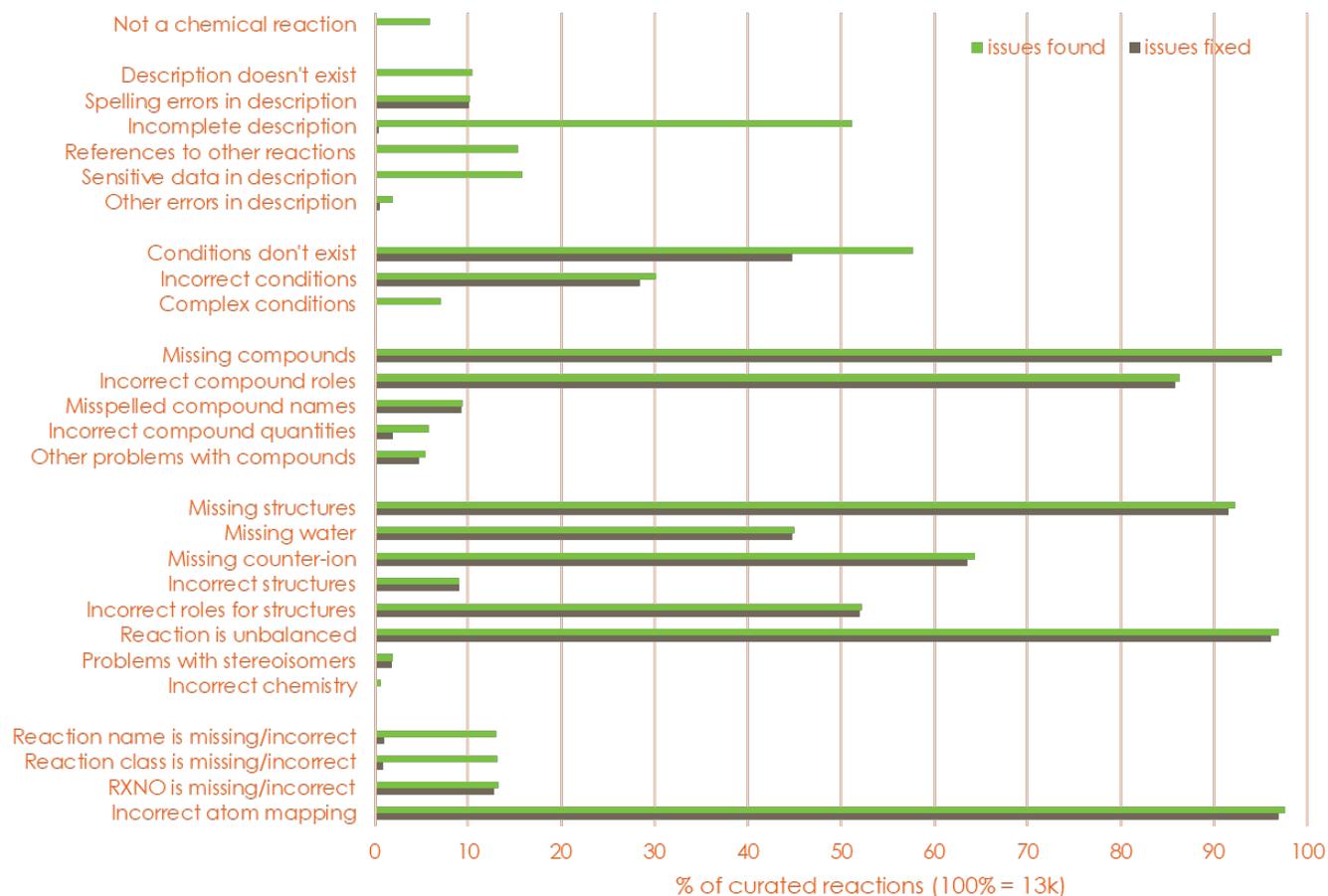
Compound decomposition

Raw input



Pipeline prototype in KNIME

Lessons Learned from Curating 16K Reactions



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

- Unbalanced reactions
- Inconsistent use of roles (reactant vs. reagent)
- Multistep reactions
- Lack of systematic capture of failed/successful reactions
- Lack of standardized process description
- Data curation: by **Rancho BioSciences**

Rancho
bio**sciences**

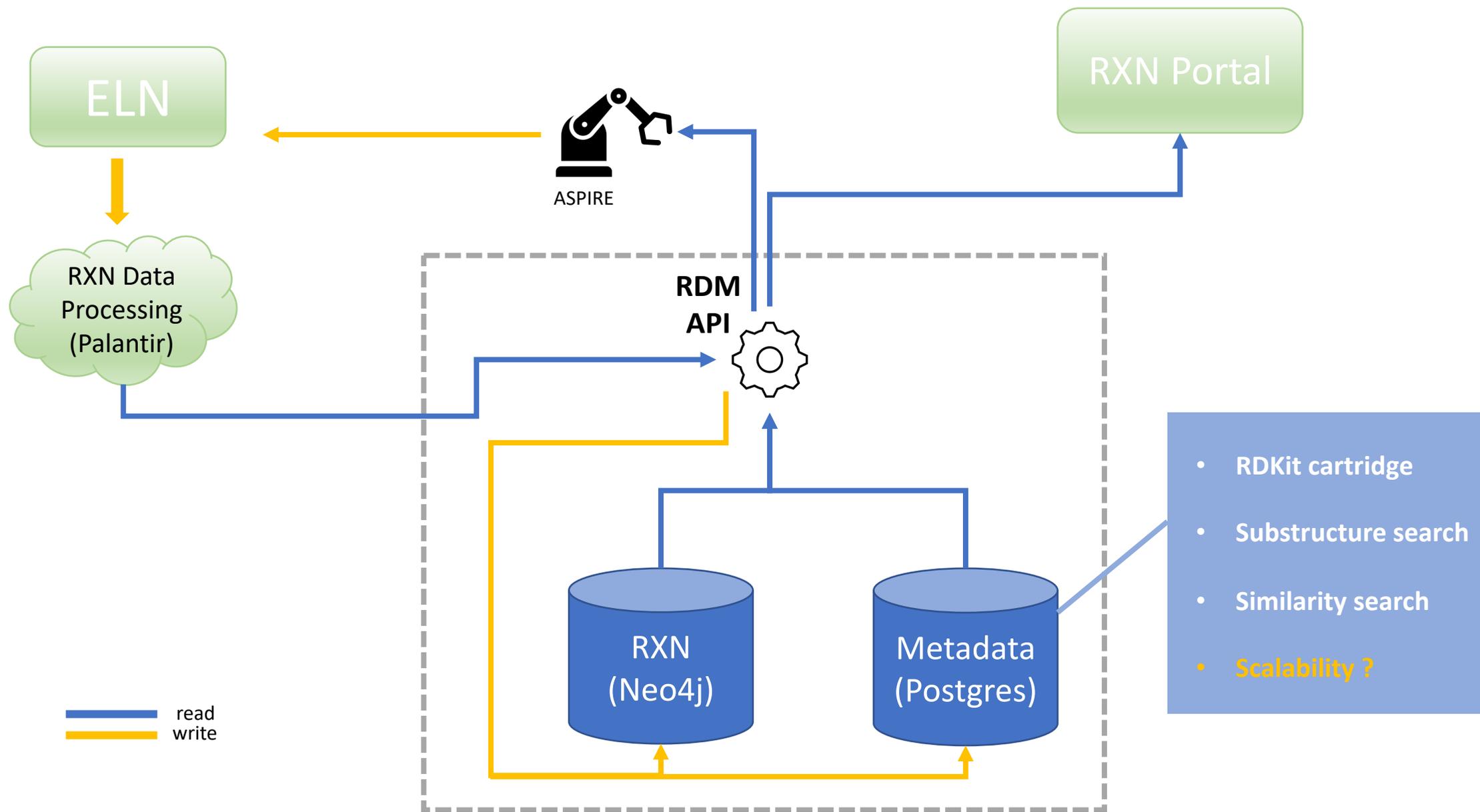
Image credit: <https://ranchobiosciences.com/>

Slide from Oleg Stroganov, PhD (Rancho BioSciences)



National Center
for Advancing
Translational Sciences

ASPIRE Knowledge Base Architecture

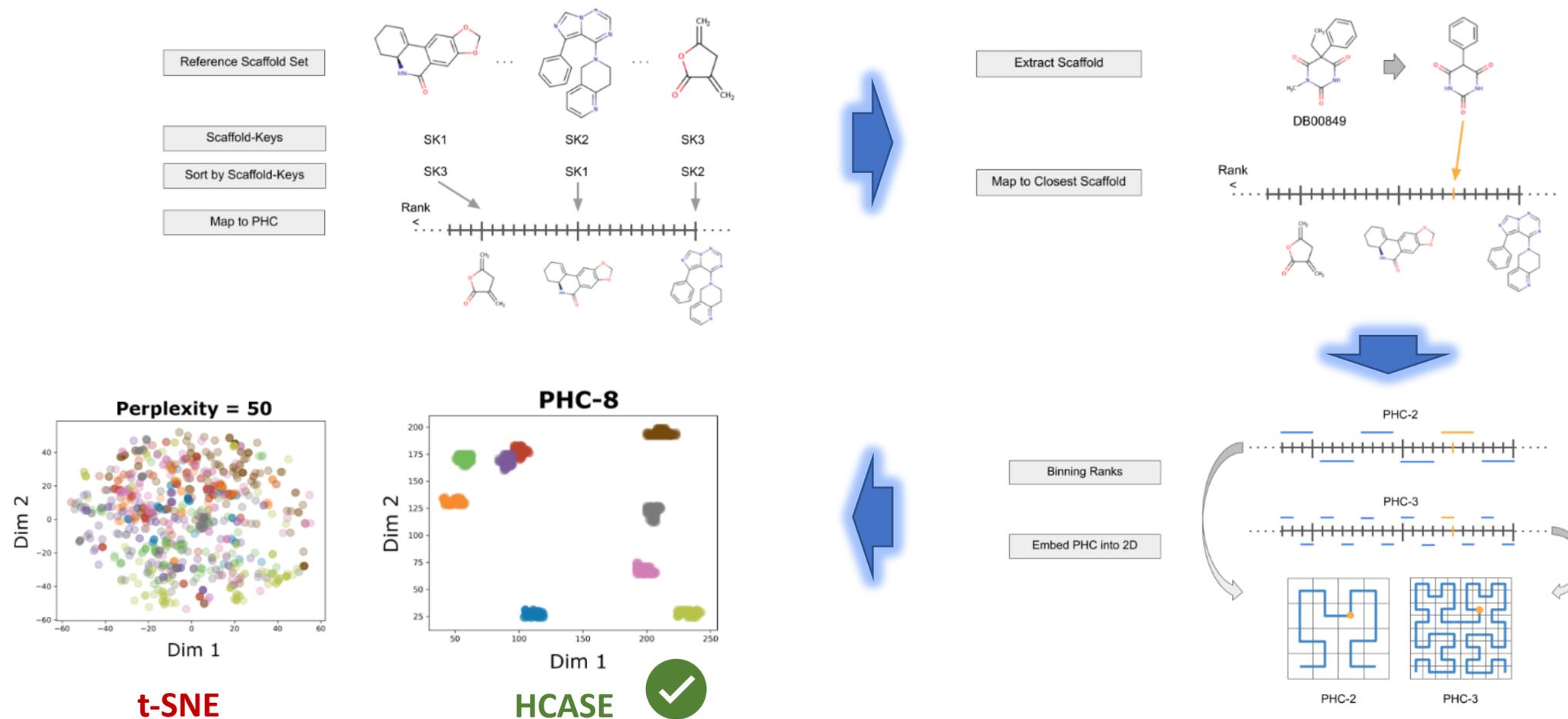


Treats of an Ideal Chemical Space Embedding Method

- Intuitive interpretation by medicinal chemist
 - Placement of scaffolds reflects medicinal chemist's thought process.
 - Molecules of similar structure and complexity are clustered in the generated map.
- Generated map is robust with regards to embedding/overlaying new datasets.
- Facilitate the comparison of chemical space coverage across multiple datasets.
- Scales to large datasets.



Hilbert-Curve Assisted Structure Embedding (HCASE) Method



Zahoranszky-Kohalmi *et al.* (2020), ChemRxiv. Preprint. <https://doi.org/10.26434/chemrxiv.11911296.v1>

Code available at: <https://github.com/ncats/hcase>

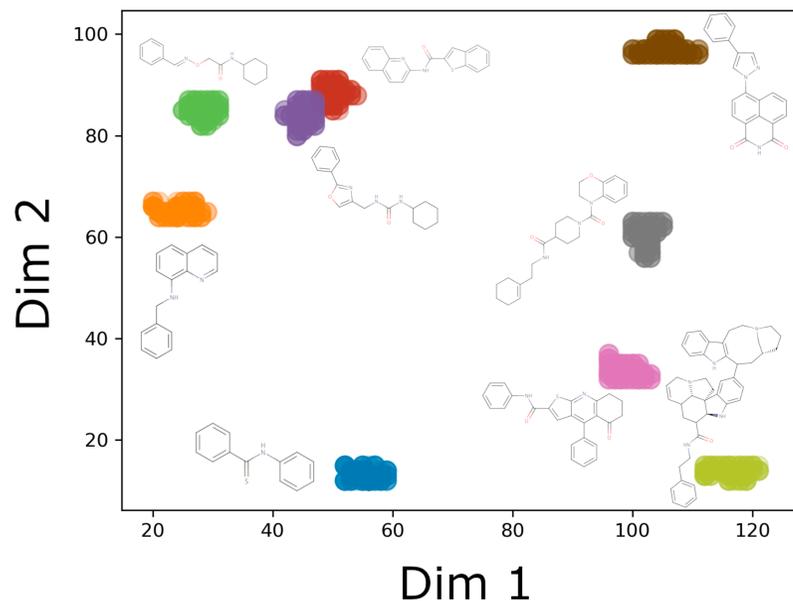


National Center
for Advancing
Translational Sciences

Comparison of the Clustering of Reference Scaffolds

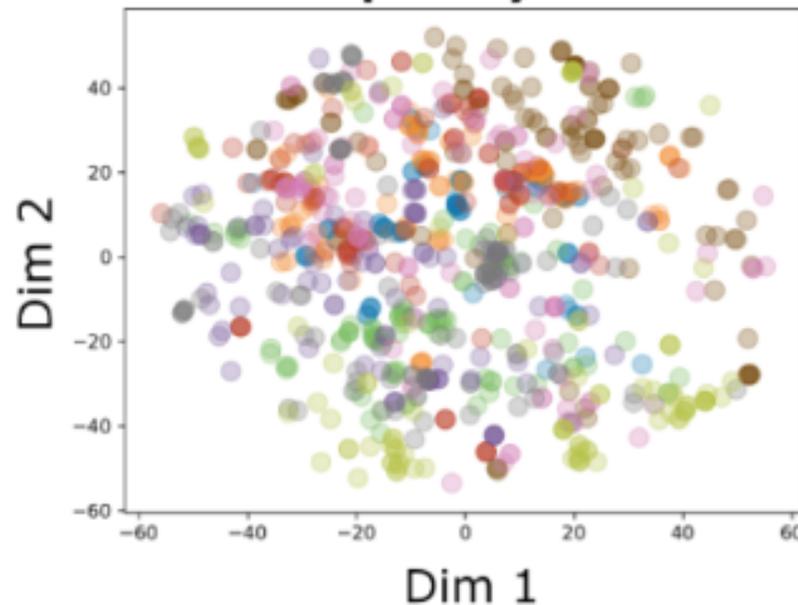
HCASE

PHC-7

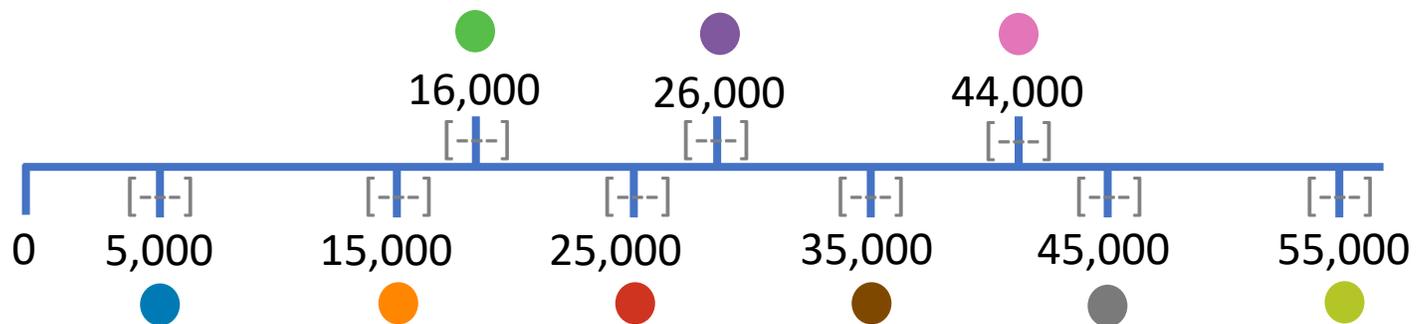


t-SNE

Perplexity = 50



55,961 unique Bemis-Murcko scaffolds were extracted from ChEMBL.

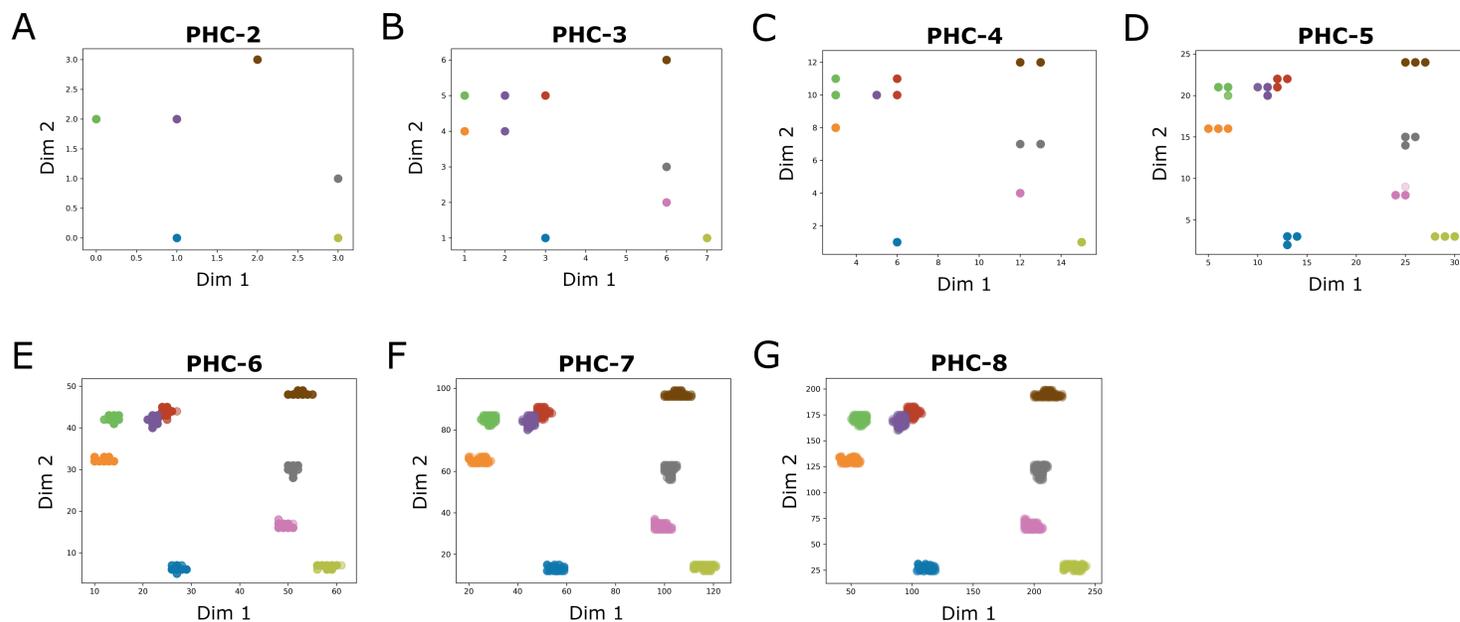


[---] : +/- 50 neighbors

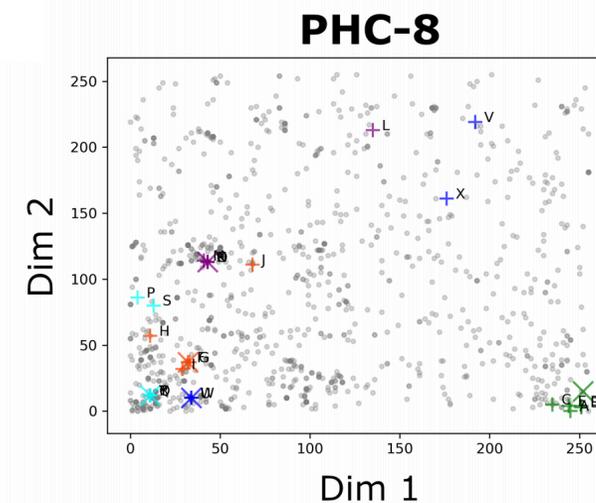
Scaffold-Key rank

Convergent Property of the Pseudo-Hilbert Curve

Mapping to PHC translates to the clustering of reference scaffolds



HCASE-embedding of DrugBank compounds



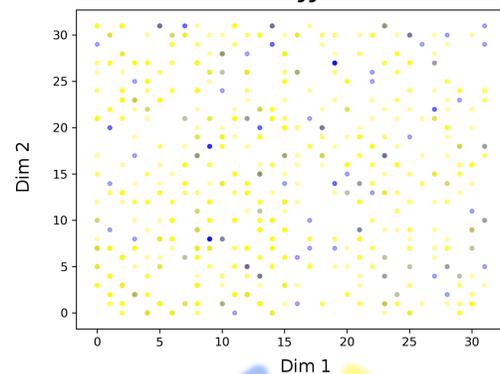
- Highlighted are: 9 scaffolds and their +/- 50 immediate neighbors on the PHC.
- Resolution (order of PHC) can be varied.
- Increasing the resolution leads to convergence in space.

- Five randomly selected compounds and their 5-NNs are highlighted.
- Space defined by 55,961 unique scaffolds from ChEMBL.

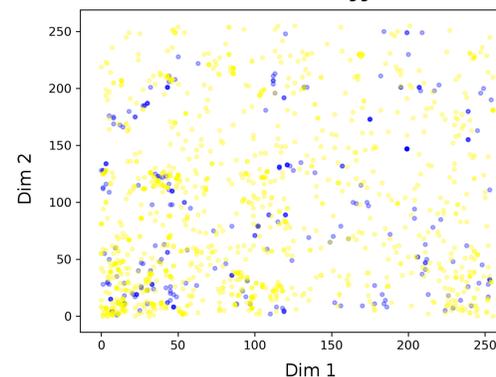


Analysis of Chemical Space Coverage

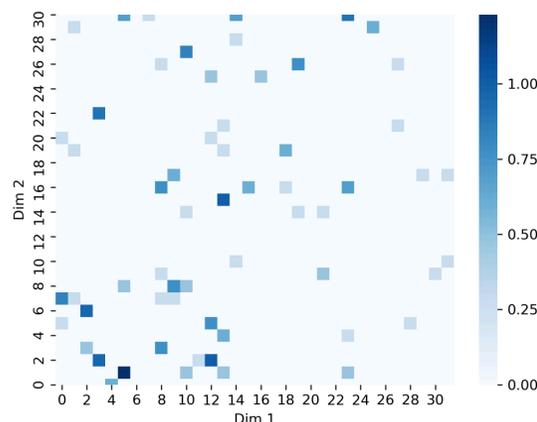
ChEMBL NatProd space
546 scaffolds



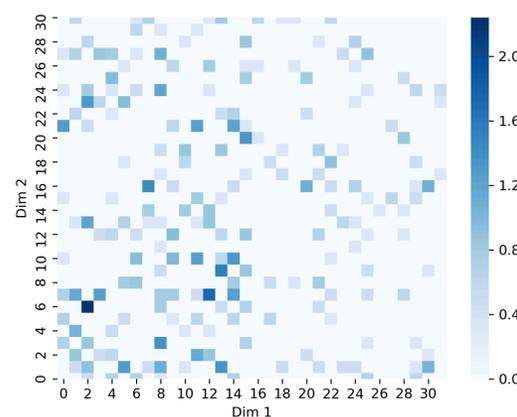
Entire ChEMBL space
55,961 scaffolds



- CANVASS (natural products) compounds (344)
- DrugBank compounds (2,073)



CANVASS compounds



DrugBank compounds

- **Chemical space:** ChEMBL NatProd
- **Color:** number of compounds associated with a specific coordinate in the chemical space expressed in a log scale.

What We Need ...

- Molecular properties
 - State of matter
 - Solubility
- Novel tools & integration of existing ones to aid data exploration in an ULCD
- Standardized synthesis protocols
- Large dataset of annotated reactions
 - Standardized and machine interpretable reaction mechanism representation
 - Reaction outcomes
 - Analytical data: LC-MS, NMR
- API access



ELN of Future

- Scales to support High-Throughput Chemical Synthesis
- Individual vs. blocks of reactions
- API access
- Focus on information integration & visualization
- Storage agnostic



Conclusions

- Graph-relational hybrid database for high-performance reaction informatics operations and search.
- A small scale (16K reactions) data curation project to guide the development of the ELN of the future.
- HCASE: a novel chemical space embedding method that produces intuitively interpretable chemical maps.
- Identification of areas related to ULCDs the ASPIRE project would benefit from.



Acknowledgements

Ewy Mathé
Informatics Director

Nikita Lysov
Developer

NCATS / ASPIRE - IT

Dimitrios Metaxatos
Lead

Biju Mathew
Project Manager

Busola Grillo

Mark Backus

Manideep Gurumurthy

Rafat Sarosh

Reid Simon

Tim Mierzwa

NCATS

Qian Zhu

Ivan Grishagin

Vishal Siramshetty

Dac-Trung Nguyen

Nathan Hotaling

Nick Schaub

Yuhong Wang

Radha Krishnakumar

Dave Calabrese

Cullen Klein

& All Chemists

Rancho BioSciences

Laura Brovold

Oleg Stroganov

Rob Gill

NextMove Software

Roger Sayle

Palantir

Nabeel Qureshi

Amin Manna

Mark Bissel

“This research was supported in part by the Intramural research program of the NCATS, NIH.”

