How are RDKit and MolVS used to prepare Elsevier's reaction data for synthetic route prediction?

Michael Collingsworth[2], Gerd Blanke[4], Marcus Stamm[1], Hinnerk Rey[1], Benoit Pasquereau[2], Markus Fischer[1], Jürgen Swienty-Busch[1], Frederik van den Broek[3], Ani Marrs-Riggs[2], Tim Miller[2], Elena Herzog[1] and Ivan Krstic[1]

[1]Elsevier Information Systems GmbH, Frankfurt, Germany; [2]Elsevier RELX Group, UK, [3]Elsevier, Amsterdam, The Netherlands, [4]StructurePendium Technologies GmbH, Essen, Germany

Introduction

- There are three prerequisites for AI/ML predictive modelling: availability of predictive models, high quality and considerable quantity of data, and the IT infrastructure to train and to run predictive models.
- Reaxys with its 19 million single step reactions provides a good base for AI/ML predictive modelling and the data has been used by various academic and commercial partners.
- Entellect's Reaction Workbench allows you to combine the Reaxys data with your own data for further AI/ML augmented model development.
- Data preparation for synthesis route predictions using AI/ML models is tedious and time-consuming due to variations in how the reactions data is extracted from scientific literature, how reactions are annotated, stored and transformed to be usable for the modelling work.

Data challenges in chemistry

- Agreed standards for storing chemistry data are missing
- Inconsistent and incomplete data indexing within datasets due to errors and/or different concepts in indexing
- Finding correct method to properly load and filter datasets from different sources for AI/ML methods
- Normalizing chemical structures and reactions to provide a consistent chemical representation for the AI/ML model building
- Training AI/ML models on duplicated data - to recognize at the late stage the chemically obvious identity of different depictions of the same compound .
- Setting up an infrastructure for deploying the models in production is time consuming
- Data processing is time intensive

Entellect's Reaction Workbench

Data Preparation

(using semantic standards for identification and provenance)

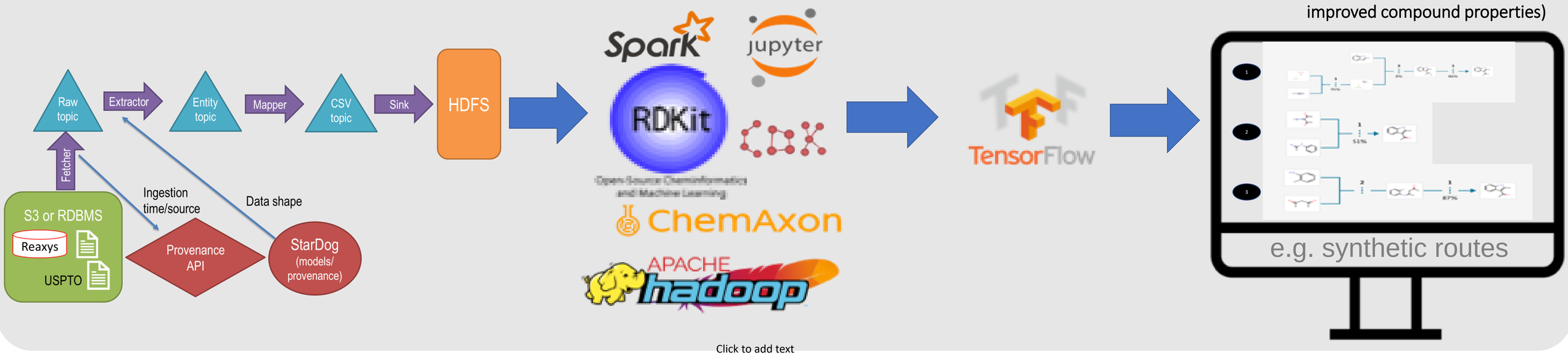
Data Normalization

(using tools & applications to normalize and de-duplicate data)

Algorithm/ Model Development

AI Predictions

(chemistry and pharma predictions of novel routes, optimization of routes and improved compound properties)



Click to add text

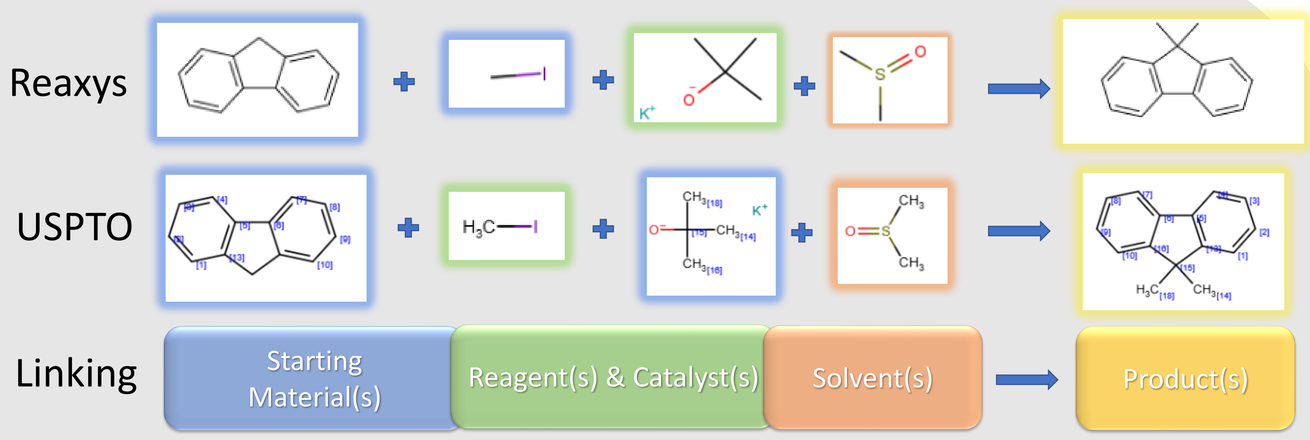
Method for linking and deduplicating data

Challenges for normalization, deduplication & linking of reactions

- Missing gold-standard guidelines for drawing chemical structures (e.g., salts)
- Inconsistent assignment of molecules to reaction roles (e.g., reactant vs. catalyst vs. solvent)
- Differences in substance assignments to reactions (main reaction vs. clean-up steps)

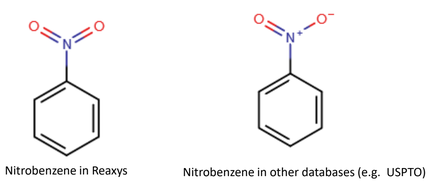
The Reaction Workbench provides a clean & structured data set for AI/ML applications:

- By normalization of chemical structures (e.g., with RDKit)
- By removing assignments of reaction roles on starting material & product level
- By applying mappings to identify relevant molecules



Develop a common structure normalization for Reaxys and upload "in-house" structures

- Because there are no common rules available, chemical structures and reactions are normalized to satisfy customary requirements of individual organizations working with these data
  - A common example: Nitro groups



Nitrobenzene in Reaxys

Nitrobenzene in other databases (e.g. USPTO)

  - Chemical representations in Reaxys are not always aligned with the "sanitized" concept of RDKit

Proof of concept for the upload of external "in-house" data and their normalization

- Python 3.7 with Jupyter Notebook, RDKit 2019.2 and MolVS 1.0.
- Because MolVS already provides a major set of applicable rules out of the box that represents most of the Reaxys rules, the first structure normalization step is done by "standardizing" the structures and reactions by MolVS
  - E.g. the nitro groups are sanitized to the charged form by MolVS
- To adapt to any additional chemical representation of Reaxys that is not corresponding with MolVS, RDKit scripts are added to normalize towards the Reaxys drawing rules.
  - E.g. the Nitro group is transferred into the Reaxys typical double bonded form in this second step
- RDKit "adapted" functions are used to create displayable pictures of structures and reactions

How does cheminformatician become productive on RWB?

- Develop a consistent rule set to normalize and display structures and reactions of Reaxys and external datasets.
  - To work on chemical and pharmacological predictors using "in-house" data, normalize structures and reactions in a way that adhere to the "in-house" drawing rules.
  - Normalizing the structures involves choices such as
    - Ignoring or including aromaticity
    - Handling counter-ions for salts
    - Ignoring or including stereochemistry
    - …
- Practical programming considerations
  - Adapt MolVS to your own rule set.
  - Use RDKit as alternative to build the structure normalization
  - Use Python examples in the Jupyter notebooks to guide and support you in the programming.

Conclusions

- RWB provides 19 million single step reactions indexed from scientific literature and patents with reaction conditions available for 88%. The reaction set represents a diverse collection of organic chemistry covering more than 16,000 reaction classes and including common reactions such as Suzuki coupling, Michael addition as well as rare reaction types (e.g. Julia olefination, Atherton-Todd synthesis)
- The transformation of structures and reactions by RDKit and MolVS allows easy normalization and integration of "in-house" Reaxys data for AI/ML model development.
- Benefits of working with Entellect's Reaction Workbench besides the Reaxys reactions dataset include:
  - ✓ Data scientists - solid data foundations (data sourcing and normalization)
  - ✓ Bench chemists - new ideas for synthetic routes and easier way of incorporating data scientist outputs into their workflow
  - ✓ Organization - FAIR reaction data, scalable process, time & cost savings

References

1. Planning chemical synthesis with deep neural networks and symbolic AI, Marwin H. S. Segler, Mike Preuss, Mark P. Waller, Nature, 2018, 555, 604-610
2. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain, Thakkar, Amol et al., Chemical Science, 2019, vol. 11, # 1, p. 154 - 168]
3. D. Lowe, Chemical reactions from US patents, 1976-Sep 2016, https://figshare.com/articles
4. Reaxys is a trademark of Elsevier Limited and PAI is Pending AI Pvt. Ltd.
5. Entellect is a registered trademark of Elsevier Inc.