Unlocking the power of data from disparate sources – Elsevier's journey toward accurate reaction outcome predictions

Timur Madzhidov

Elsevier

Senior Product Manager in Chemistry Innovation

Unlocking the power of data from disparate sources:

Elsevier’s journey toward accurate reaction outcome predictions

Timur Madzhidov,
David Wöhlert, Eric Gilbert, Frederik van den Broek

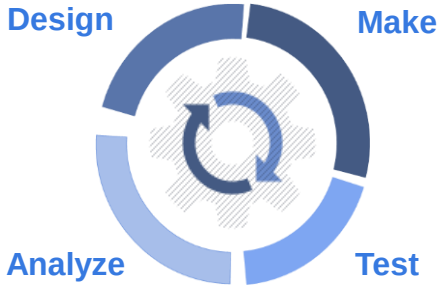
September 25, 2024

Reaxys, the most comprehensive, innovative and intuitive chemistry information system supporting customers' chemistry use cases and digital transformation needs

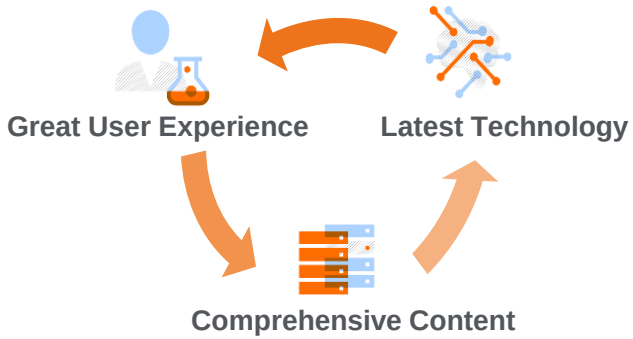
ELSEVIER

Reaxys supports every chemist across corporate and academic segments

By accelerating chemistry workflows

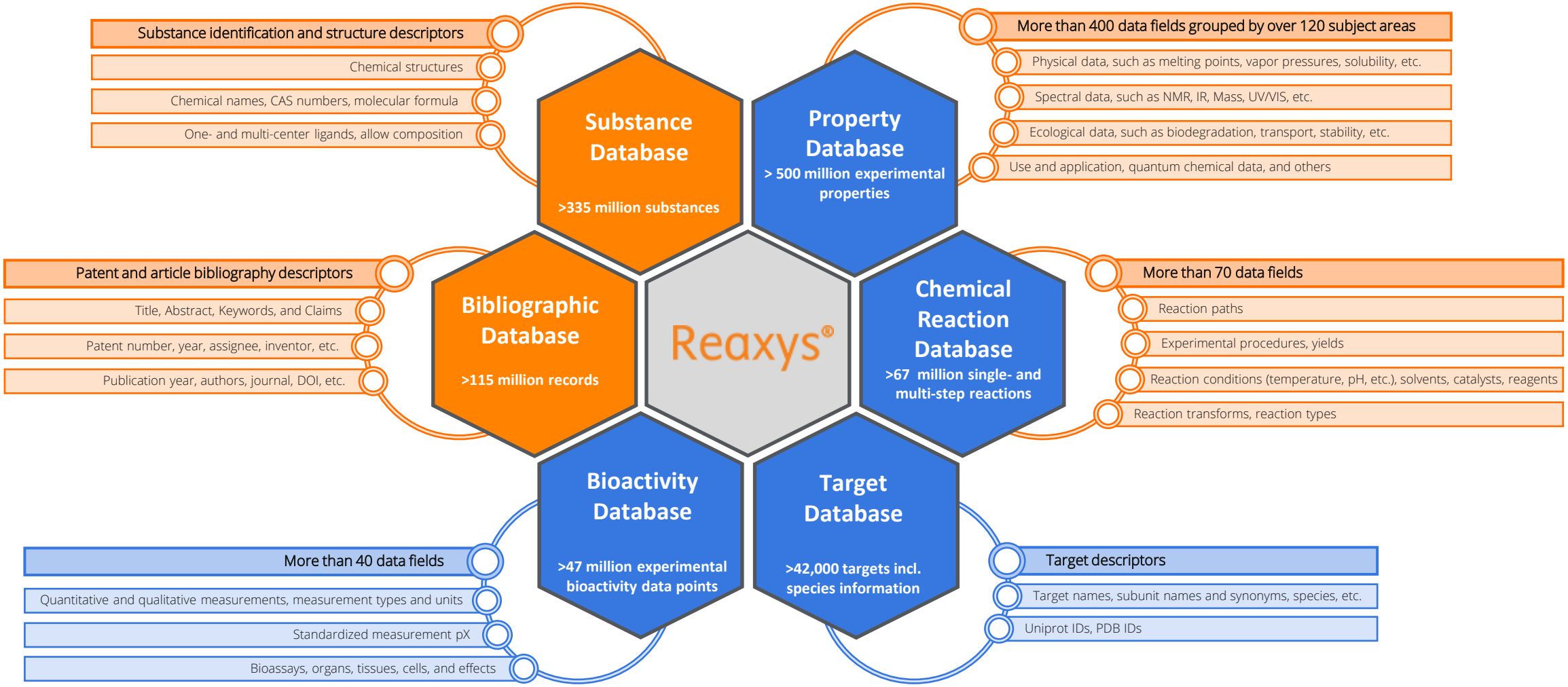


Through the combination of…



Monitor and Design			Make		Test and Analyze	
Competitive Intelligence	Novelty search	SAR, ADMET	Synthesis planning	PhysChem properties	Product analysis	Pharmacological profiling

Reaxys today*



Manually annotated

Both automatically and manually annotated

*data valid for September 16, 2024

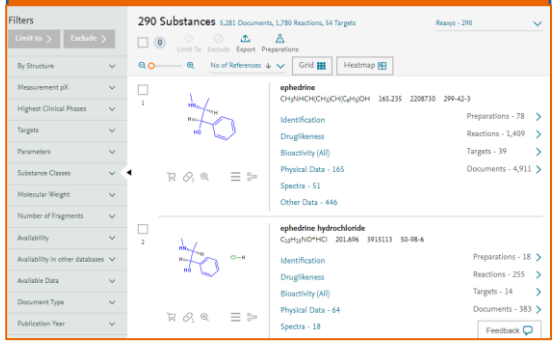
Data is a new gold

Credit: TechCrunch, August 8, 2019
Cultivated data is the next Gold Rush
https://techcrunch.com/2019/08/08/cultivated-data-is-the-next-gold-rush/

ELSEVIER

Reaxys provides access to chemical data tailored to all possible use cases

reaxys.com

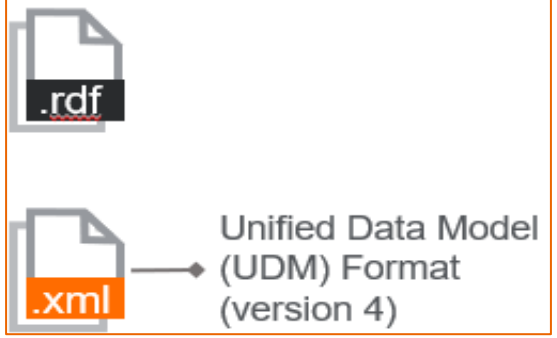


- **Human** friendly for everyday use
- **All information** accessible
- **Search** implemented
- Predictive **retrosynthesis** available
- Limited **export** available

API

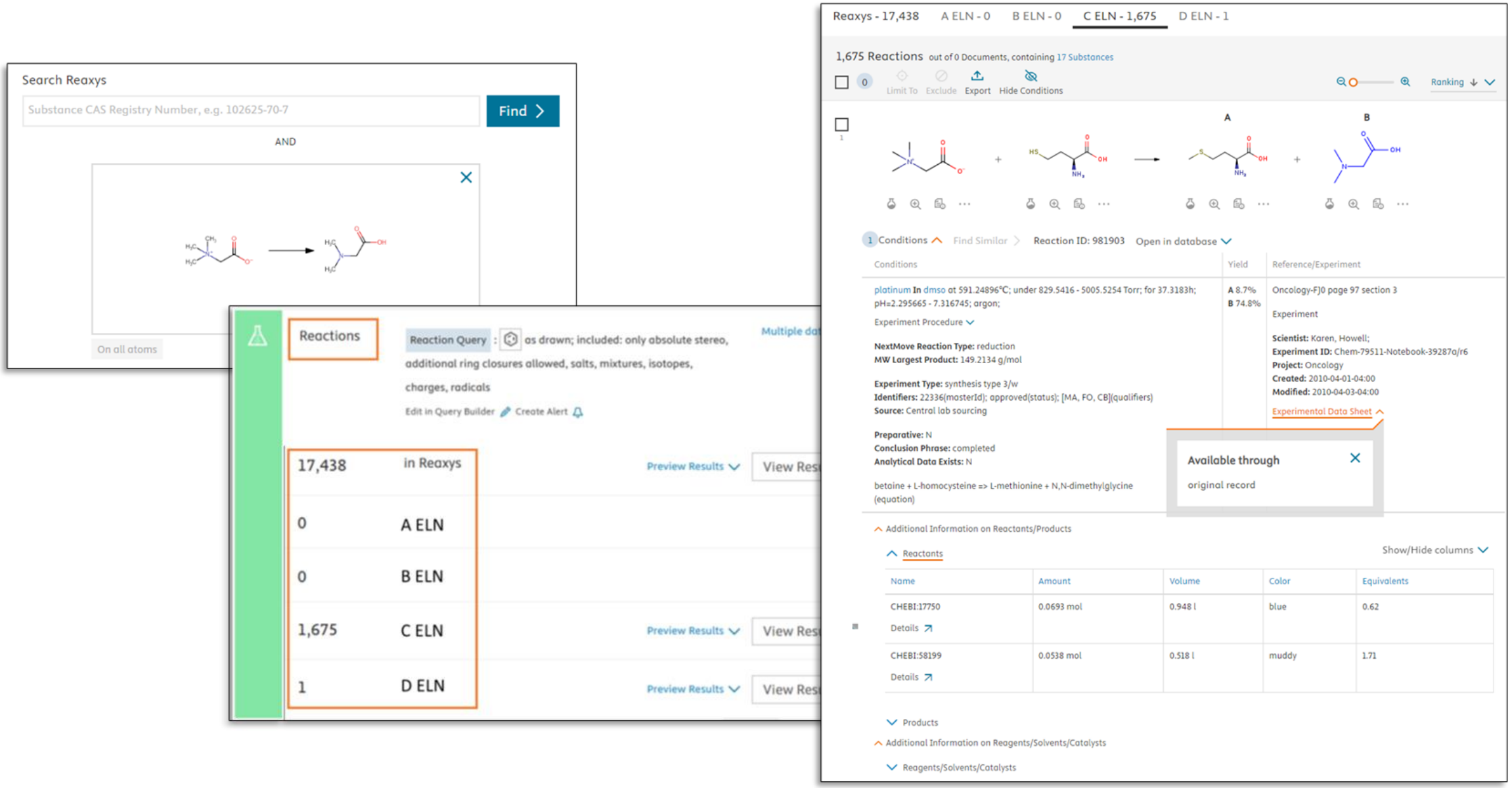
- **Machine-**friendly for embedding into internal pipelines
- **Most of information** accessible
- **Search** implemented
- **Retrosynthesis** and **synthetic accessibility** predictors available (as separate API)
- **Export**-friendly

Flat Files



- **Machine** readable format for ML/AI applications or internal data lakes
- **Case-specific information** accessible (bibliography, reactions, molecules/properties, bioactivities)
- **Dataset** delivered to user

Intuitive and integrated exploration of reactions using content integration

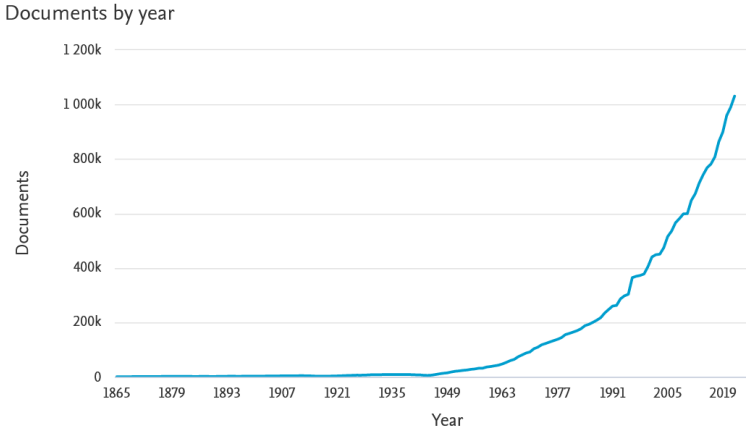


Credit: TechCrunch, August 8, 2019
Cultivated data is the next Gold Rush
https://techcrunch.com/2019/08/08/cultivated-data-is-the-next-gold-rush/

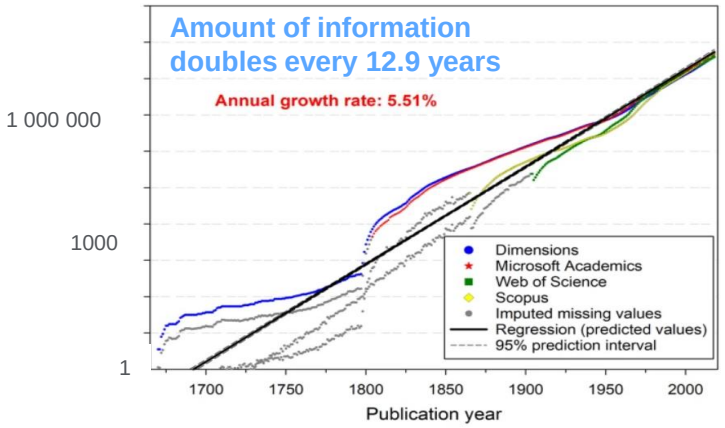
ELSEVIER

Data challenges

- Available data size increases exponentially
  - **More efficient approaches required for annotation**: programmatic extraction or increased costs
  - **Infrastructure challenges**: more efficient search approaches needed, bigger infrastructure
- Data representation
  - New previously **unknown types of chemical compounds**
  - **Rules change with time**: not a problem for visual perception but problem for ML/AI applications
  - **Diversity** in data representation in different sources: how to incorporate new data?



Credit: Scopus.com
subjects: Chem, ChemEnd, MatSci, BioChem



Source: Bornmann, L., et al. Humanit Soc Sci Commun 8, 224 (2021). https://doi.org/10.1057/s41599-021-00903-w

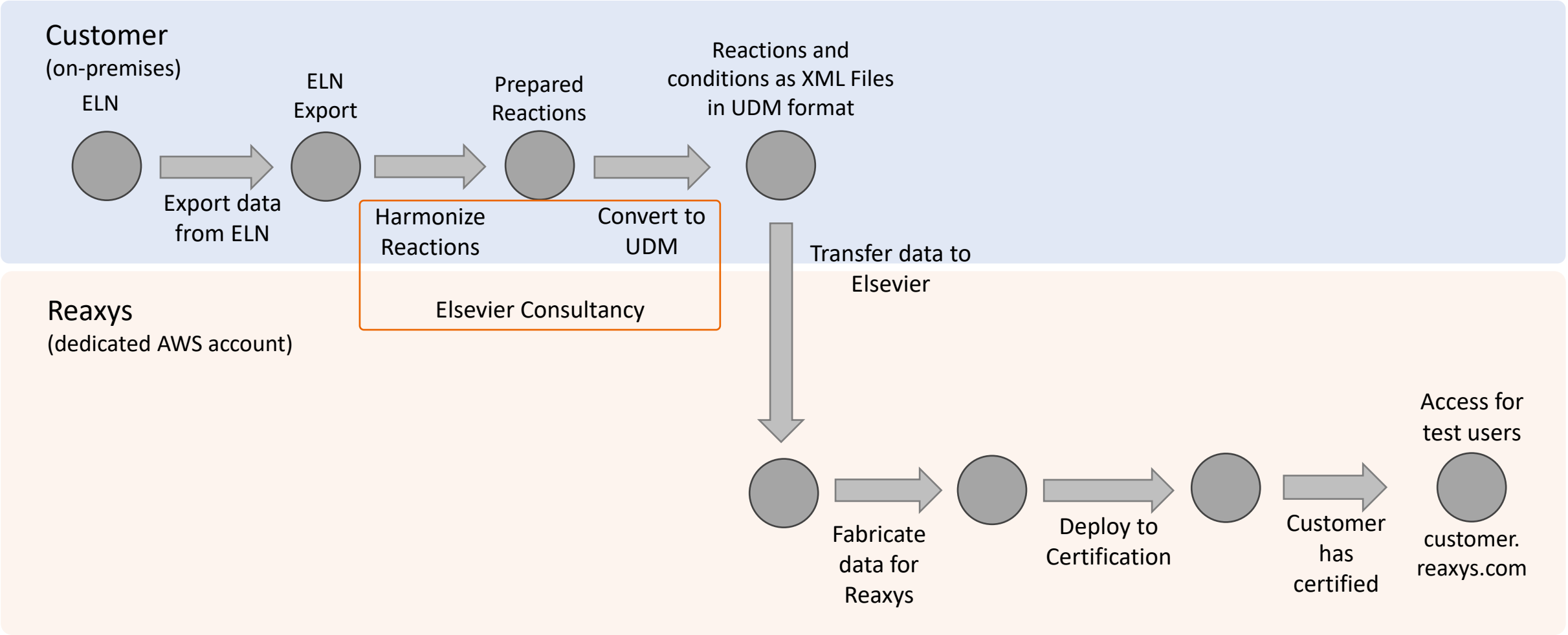
"Reaxys law":
Chemistry information doubles every 140 months (11.6 years)
+7.1% annually

Data preparation and curation is essential

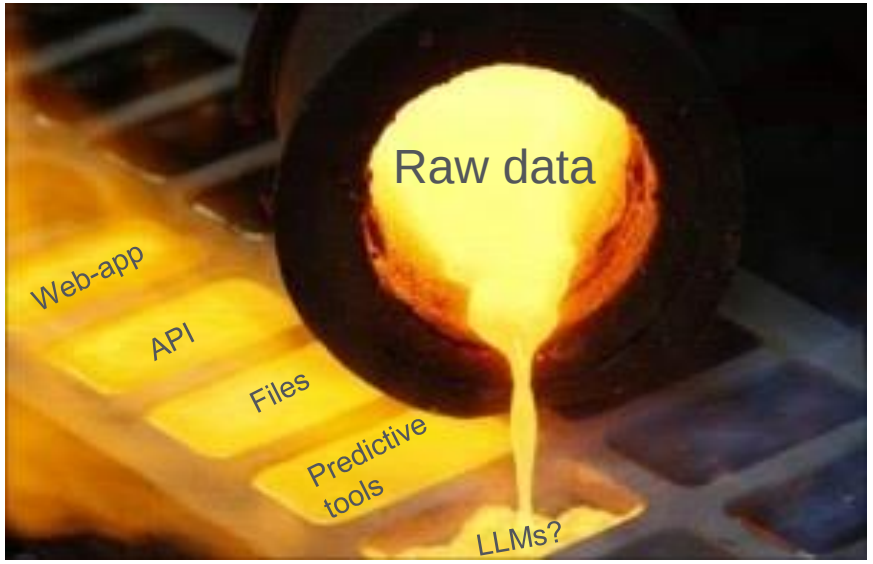
- Data from literature or Electronic Lab Notebooks is often experiment- or document-centric
- This leads to inconsistencies, duplications, etc.
- We need to follow Reaxys style in chemical data representation

Elsevier stock photo

Content integration onboarding process



How can we support chemists applying ML/AI or chemoinformatics?



How should reaction data look like to enable easy modelling?

Can be processed by regular chemoinformatics software

Homogeneous representation

Free of obvious errors

SMILES compatible

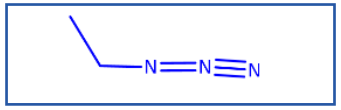
Atom mapped

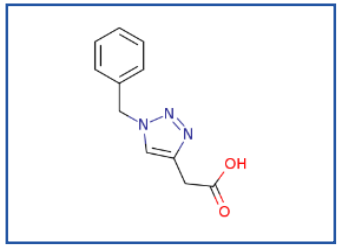
Challenging cases:
- Organometallics
- Inorganics
- Complexes
- Stereochemistry
- S-group fields

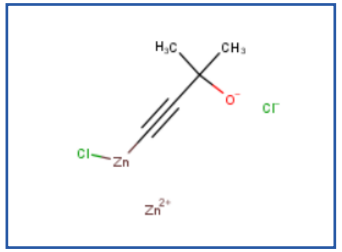
Challenging cases:
- Radicals
- Multiplicity
- Salts/mixtures
- Cycles
- Non-covalent bonds

Reaxys reaction data:

- High-quality – only manually annotated
- Comprehensive
- Well-curated
- Standardized
  - Functional group representation is unique
  - Kekule structures are given
  - Salts, mixtures, cocrustals are represented as one graph

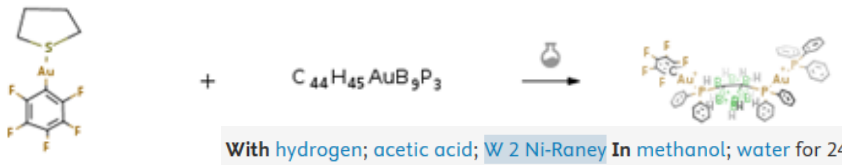






Our drawbacks are consequences of our benefits… what is good for database is not good for modelling

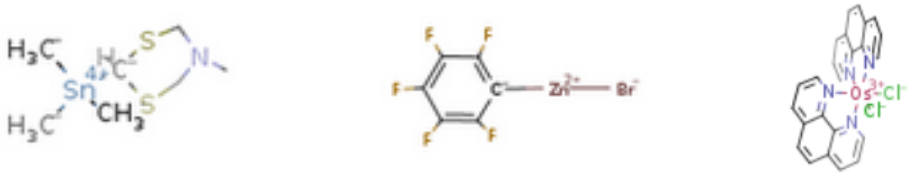
Comprehensiveness



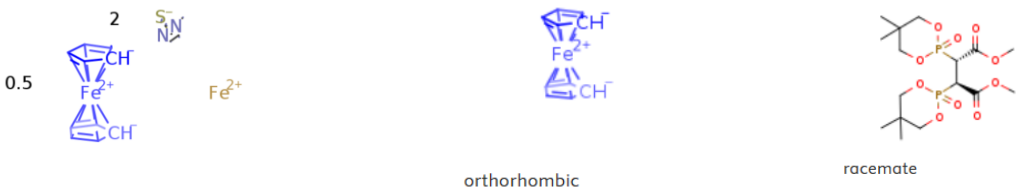
With hydrogen; acetic acid; W 2 Ni-Raney **In** methanol; water for 24h; Ambient temperature;
With eggshell calcinated at 800 grad C at 25°C; under 760.051 Torr; for 2h;

Simplicity of visual representation

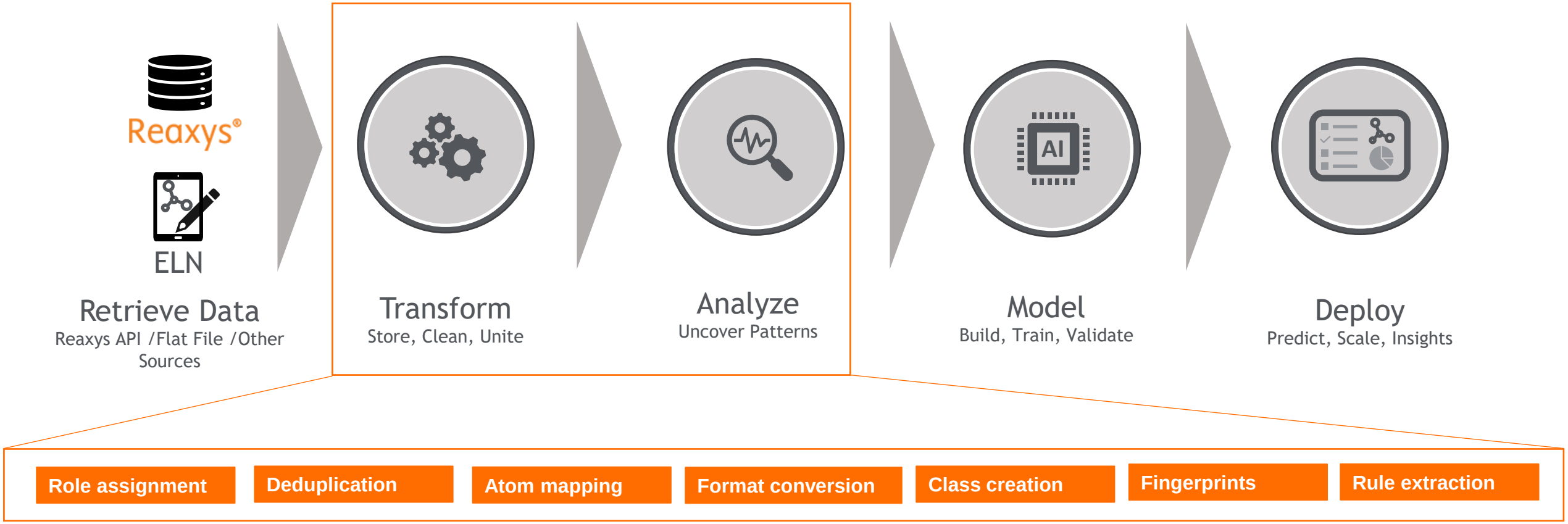
Discoverability



Coverage of whole chemistry



Typical Data Science Workflow



80% of the effort spent doing the tasks above prior to modelling

ELSEVIER

ML-optimized RFF is an extension of Reaxys Flat File to support AI and ML initiatives

Our approach:
We know our data the best, so we should curate them

Phase 1

1. Chemical structures curation

- Structure curation
- Format conversion

Phase 2

2. Transformations curation

- Atom mapping

Phase 3

3-4. Reaction conditions and endpoints curation

- Condition curation
- Unit normalization
- Deduplication
- Role assignment

Expected outcomes

Reduce the time spent on data preparation by 80-90%

More reactions are ready to use with ML/AI models (+15%)

Improved ML/AI models due to larger modelling dataset and greater homogeneity of data

ELSEVIER

Structure curation workflow consists of “transformers” that change structure, and “filters” that delete irrelevant information

Structure transformers

1. Interpret internally used S-group fields
2. Standardization of organometallics
3. Remove explicit hydrogen preserving stereochemistry and valence
4. Split reaction with concurrent products

Filters

1. Delete reactions w/o reactants or products
2. Delete reactions with same reactants and products
3. Delete reactions leading to error by chemoinformatics software

Reaxys®

ML-optimized reaction flat files

Some statistics for structure curated dataset

Initial file	22.8M
Final file	**24.8M**
Remove Data Sgroup	10.9M
Split Multi Product	4.5M
Remove Explicit H	4.4M
Resolve Fragment Multiplicity	1.1M
Interpret molecular charge	0.3M
Remove Unstructured	0.4M
Organometallics	0.3M
SMILES conversion filter	0.2M
No reaction filter	39K
Half-reaction filter	23K

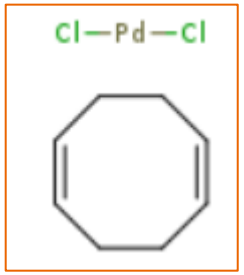
Pending.AI feedback:

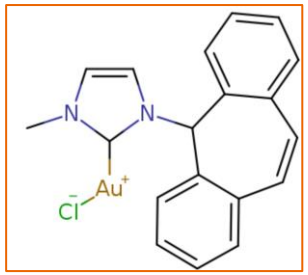
- **Increased number of reactions** available for modelling **(+12%)**
- Increased number of **generated rules** for retrosynthesis planning (+17%)
- Increased **rule coverage** (+1.2%, from 93.2% to 94.4%)
- **Simplified processing** – no script failures
- **More routes** suggested by retrosynthesis solution (+5%)

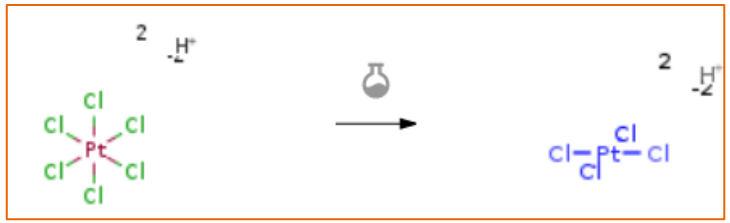
Pending.AI

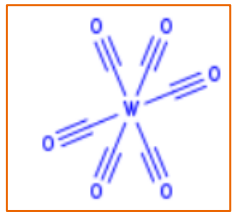
ELSEVIER

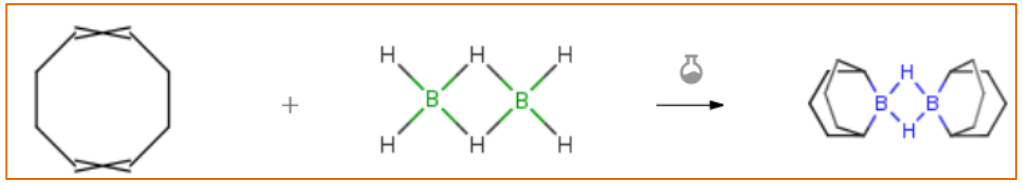
Questions to community: how to represent these compounds? Chemical conventions contradict chemoinformatics software capabilities

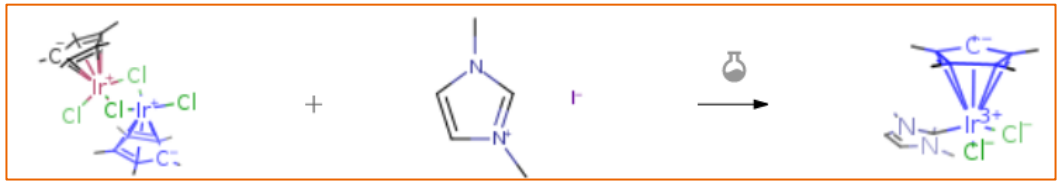






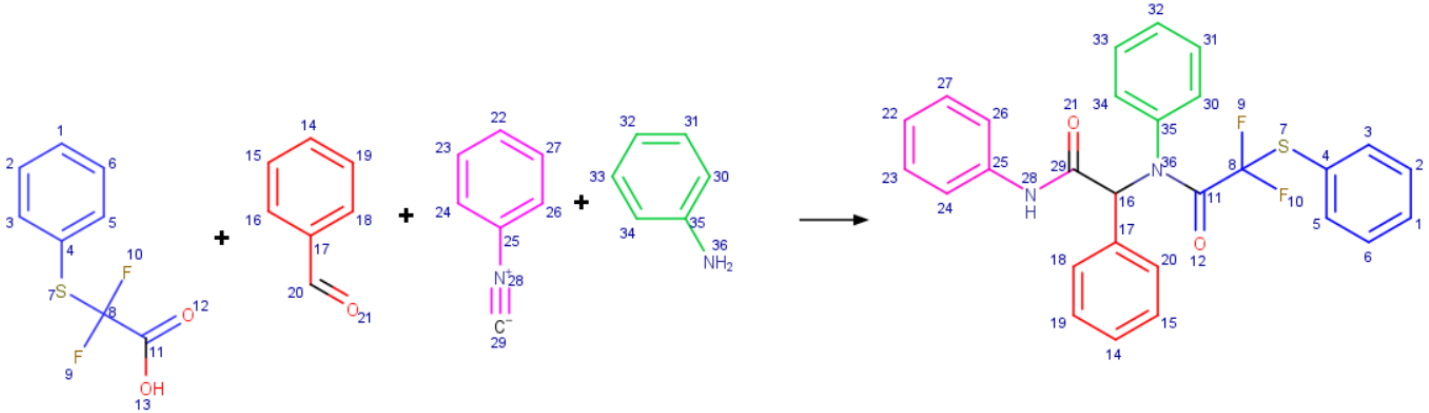






Phase 2: find best atom-to-atom mapping solution and develop atom mapping pipeline

"One-to-one correspondence between atoms of reactants and products that reflects reaction mechanism"



- Homogeneous rather than chemically correct


molecular informatics
models – molecules – systems

Research Article

Atom-to-atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies

Arkadii Lin, Natalia Dyubankova, Timur I. Madzhidov, Ramil I. Nugmanov, Jonas Verhoeven, Timur R. Gimadiev, Valentina A. Afonina, Zarina Ibragimova, Assima Rakhimbekova ... See all authors

Volume 41, Issue 4
April 2022
2100138
This article also appears in:
Chemical Reaction Mining

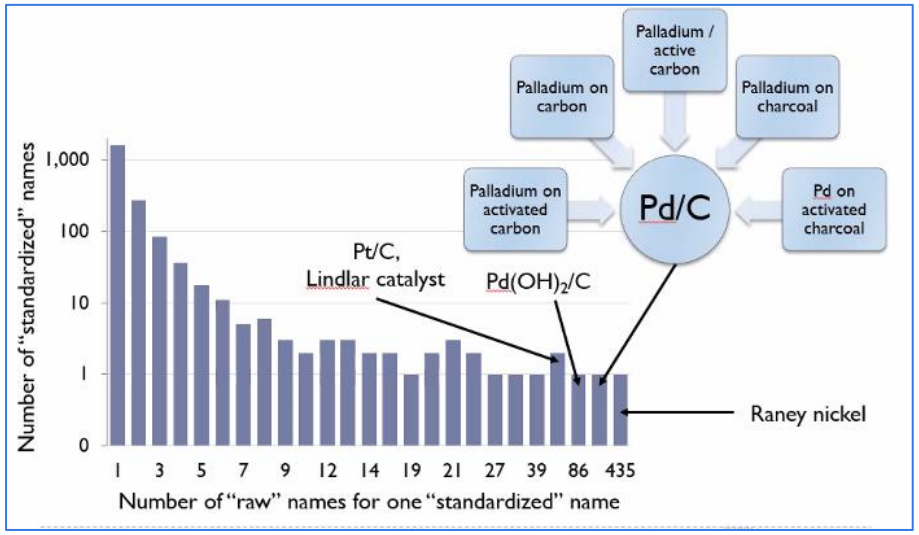
Advertisement

	Processed*	% Correct*
Mapper 1	100.00	83.33
Mapper 2	100.00	76.16
Mapper 3	100.00	74.53
Mapper 4	100.00	74.43
Mapper 5	99.24	73.15
Mapper 6	100.00	72.79
Mapper 7	100.00	72.66
Mapper 8	99.51	53.95
Mapper 9	100.0	83.93
Mapper 10	100.0	**84.2**

* On Reaxy'fied dataset from Lin et al (2022) Mol.Inf., 2100138

Phase 3: Reaction condition curation

Normalization of compound names



Compound role reconsideration (catalyst/reagent/solvent)

Normalization of numerical fields (T, p, yield)

Deletion of conditions corresponding to multi-step reactions

Use case for curated reaction dataset

Eric Gilbert
David Wohlert
Frederik van der Broek

In collaboration with:

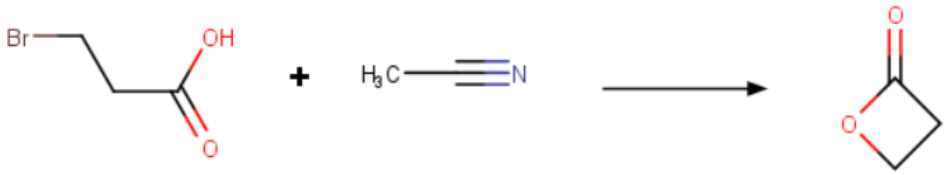
Objective

- **Goal**: Create a model to predict reaction success (classification: yield ≥ 5%).
- **Data**: Reaxys reaction SMILES for pretraining, Suzuki HTE dataset for benchmarking & Janssen Electronic Notebook (ELN) for fine-tuning

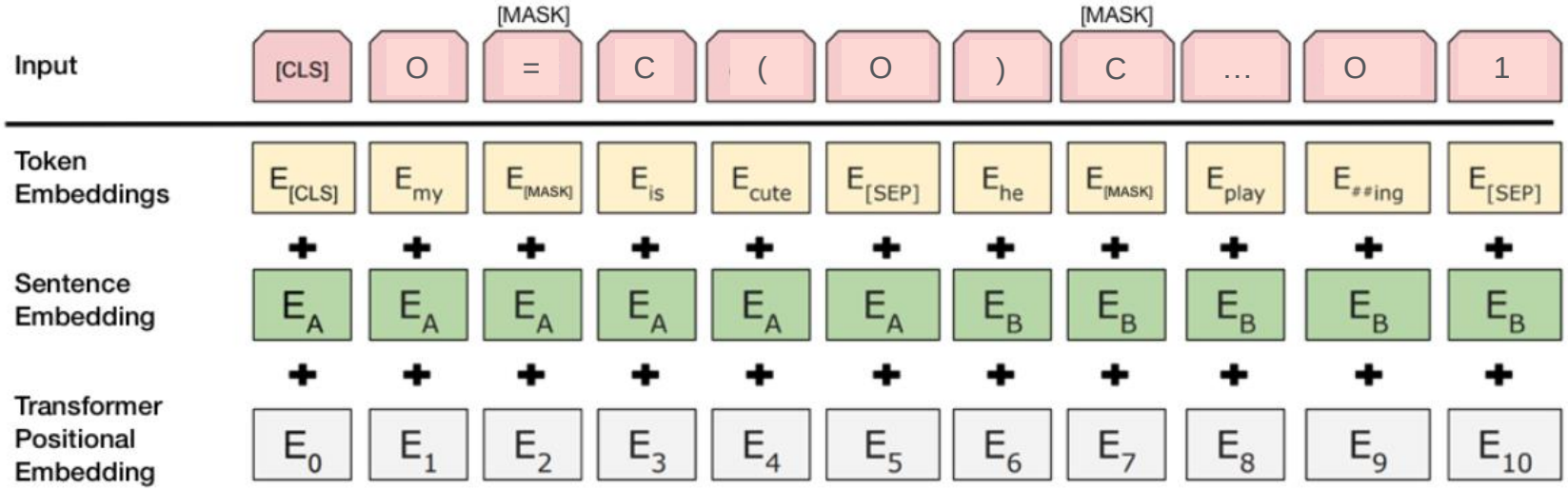
Why use ELN data to fine-tune model?

- literature yields biased, not always reliable.
- ELN data contains more lower-yielding rxns.

By substitution of text to SMILES one can create embedding of reaction



O=C(O)CCBr.CC#N>>O=C1CCO1

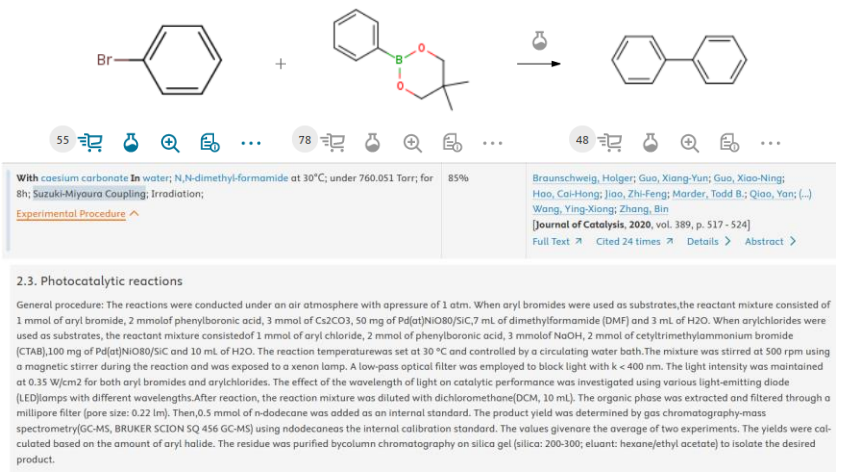


Neves, P., *et al.* (2023). *J Cheminf*. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00685-0
Philippe Schwaller *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 015016

Picture from publication:
https://arxiv.org/pdf/1810.04805

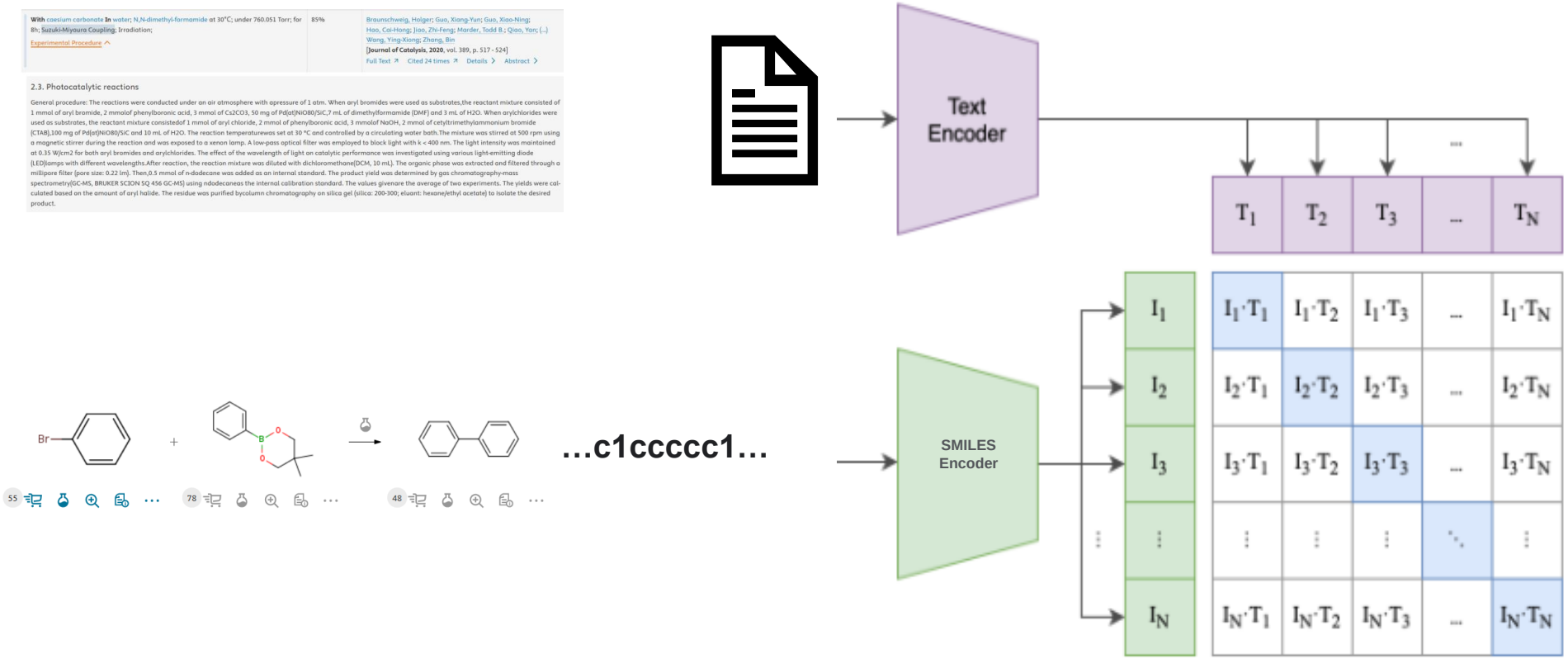
How can we improve the model?

- Add chemical knowledge!
  - Add context from text reaction description



  - Add context by augmenting pretraining task

Adding chemical context from reaction description: encoders that describe reaction procedure and SMILES are trained together to minimize the distance between actual text and SMILES embedding

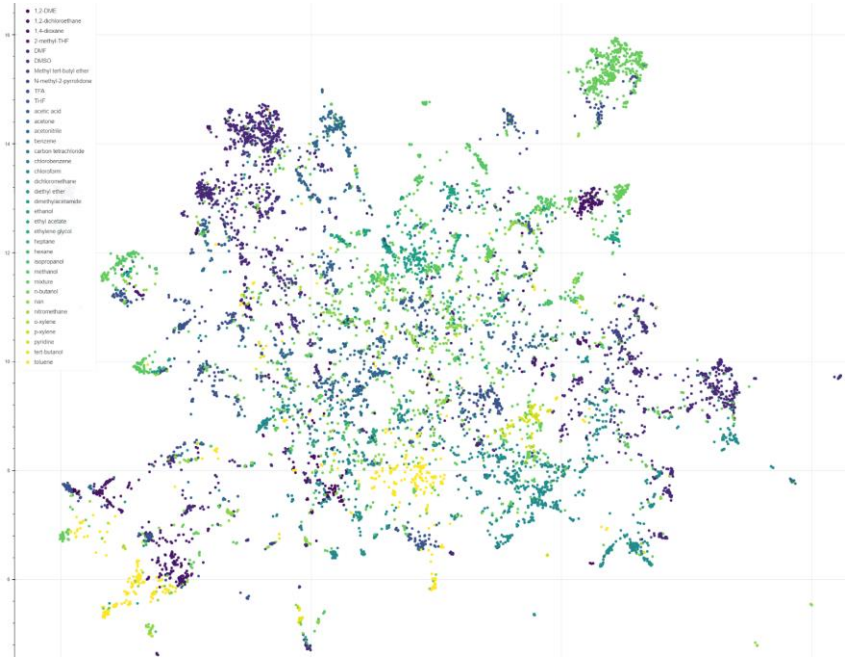


ELSEVIER

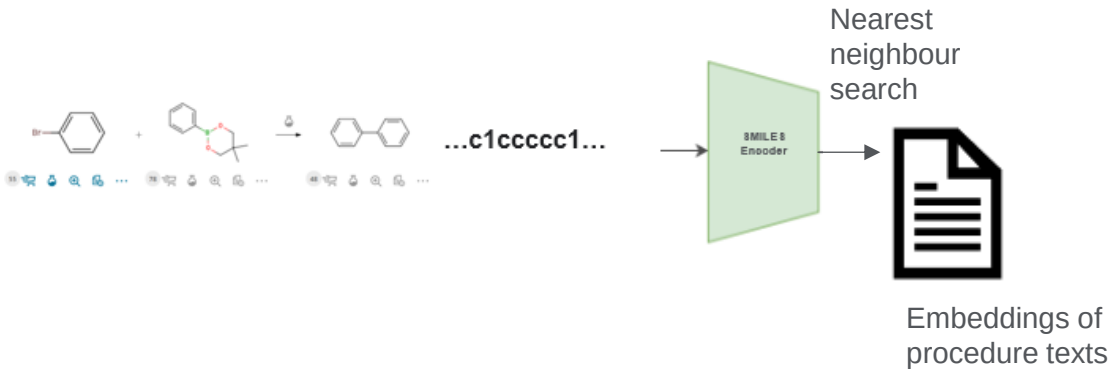
Reaction SMILES space reflects conditions!

Reaction space coloured by solvent used

- Reactions in certain solvent are clustered together



Retrieval of the suitable procedures for reaction

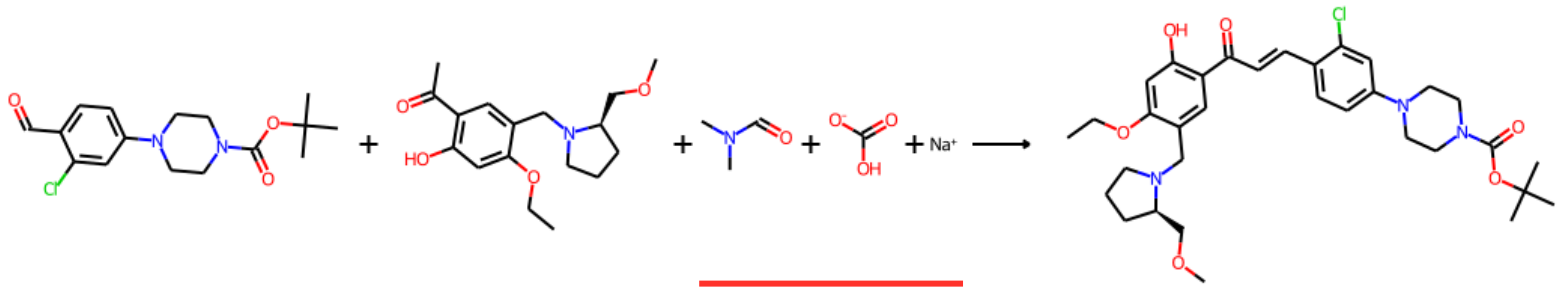


Metric	Precision@K
Text-to-SMILES @1	0.495
SMILES-to-Text @1	0.492
Text-to-SMILES @10	0.859
SMILES-to-Text @10	0.876
Text-to-SMILES @50	0.956
SMILES-to-Text @50	0.968

The model can be used to spot inconsistencies in data

- Cosine similarity between SMILES and text embeddings: -0.227

SMILES:

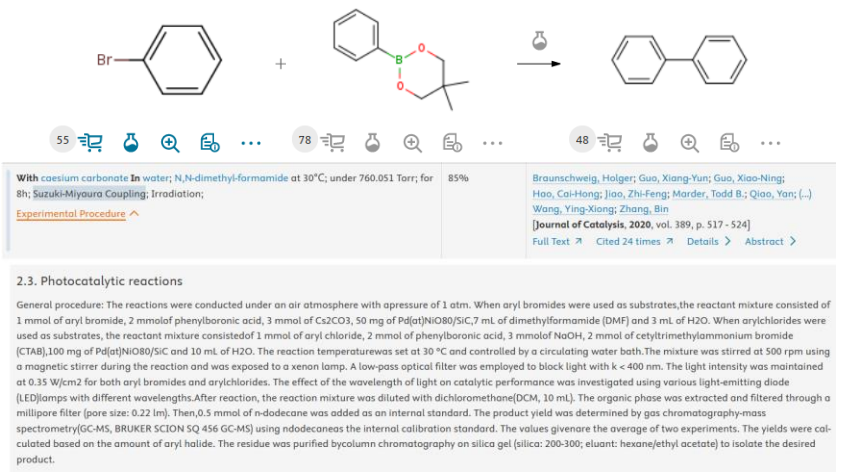


Text:

Add 303mg B-122 to 4ml EtOH, then add 221mg KOH, 640mg B-10, stir at RT, and monitor LC-MS until there is no B-122 left.

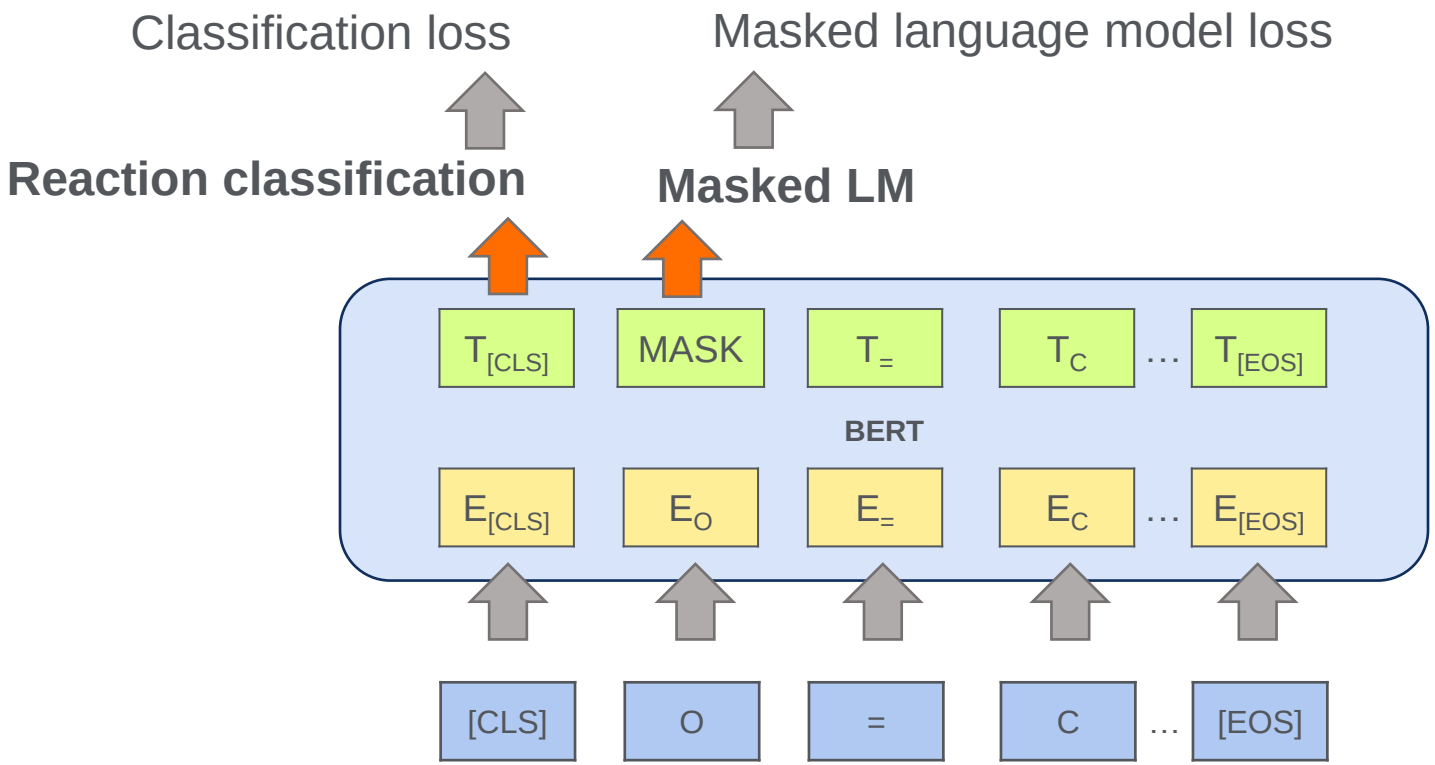
How can we improve the model?

- Add chemical knowledge!
  - Add context from text reaction description



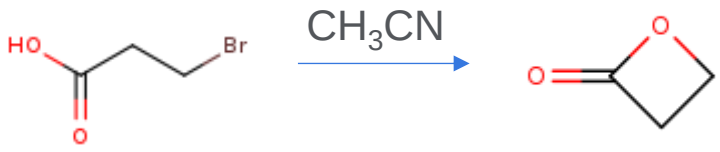
  - Add context by augmenting pretraining task

Augmented Pretraining of BERT Model: train model not only to predict masked token in reaction SMILES but also predict whether reaction looks reasonable



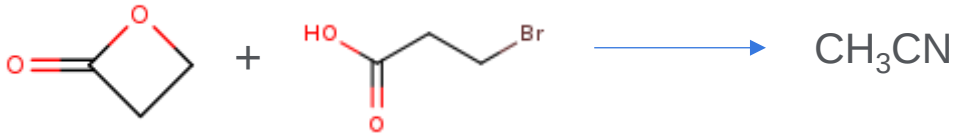
MLM Task

O=C(O)CCBr.CC#N>>O=C1CCO1



rxn classification task

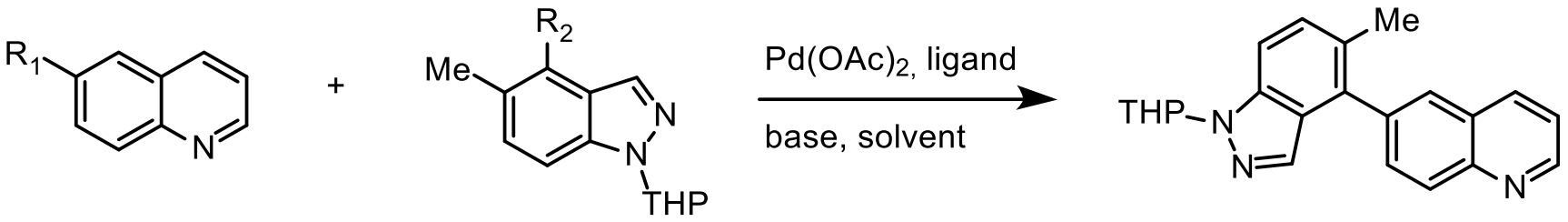
O=C(O)CCBr.O=C1CCO1>>CC#N



Model shows the best quality of prediction on Suzuki Reaction Benchmark

5760 reactions[1]

11 ligands
6 boronic acids
4 aryl halides
7 bases
4 solvents



Base model	MSE loss	R^2
MLM-only pretrain	0.19 ± 0.01	0.80 ± 0.01
Dual pretrain	**0.17 ± 0.01**	**0.82 ± 0.01**
[2]Schwaller rxnfp		0.79 ± 0.01
Schwaller rxnfp fine-tuned on CLS task		0.81 ± 0.01

1) Perera et al., Science 359, 429–434 (2018)
2) Philippe Schwaller *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 015016

Conclusions

- Data should be standardized by data provider
- Let’s discuss what is the best way to represent data for modelling!
  - Standardization for ML applications might be different from the one used for data representation
  - There is fundamental gap between data representation in databases, “chemical beauty”, and chemoinformatics software capabilities
- Model quality can be further improved by incorporation of cross-domain data and knowledge

E-business card:

Thank you

E-mail: t.madzhidov@elsevier.com

Unlocking the power of data from disparate sources – Elsevier's journey toward accurate reaction outcome predictions

Timur Madzhidov

Elsevier

Senior Product Manager in Chemistry Innovation