



**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



Reaxys®

Unlocking the power of data from disparate sources:

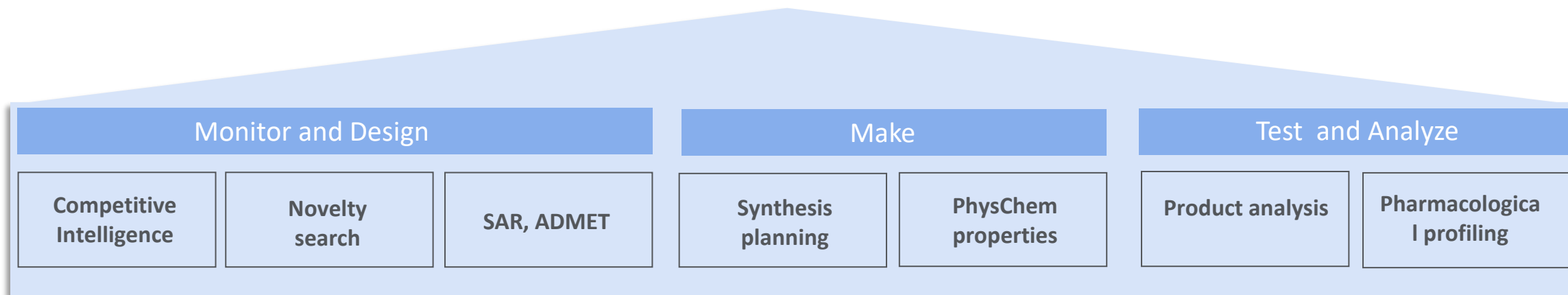
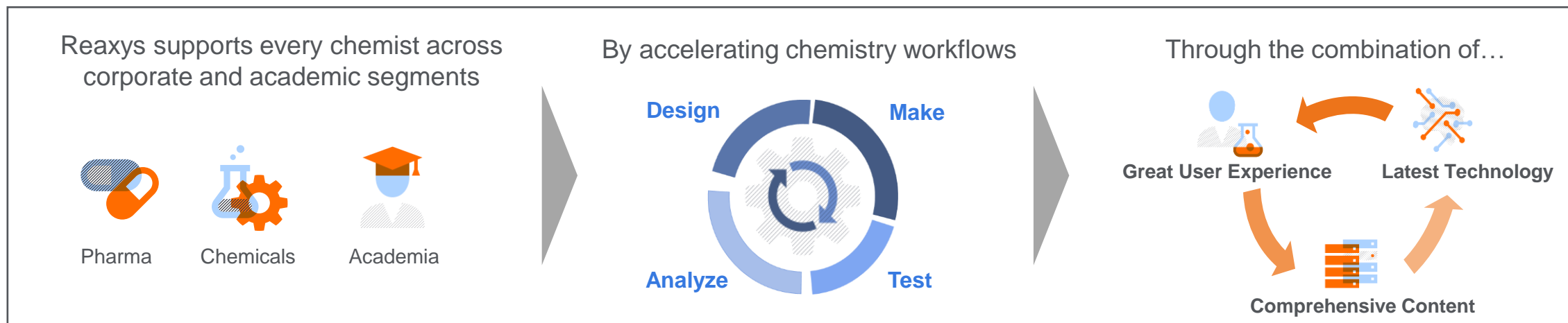
Elsevier's journey toward accurate reaction
outcome predictions

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

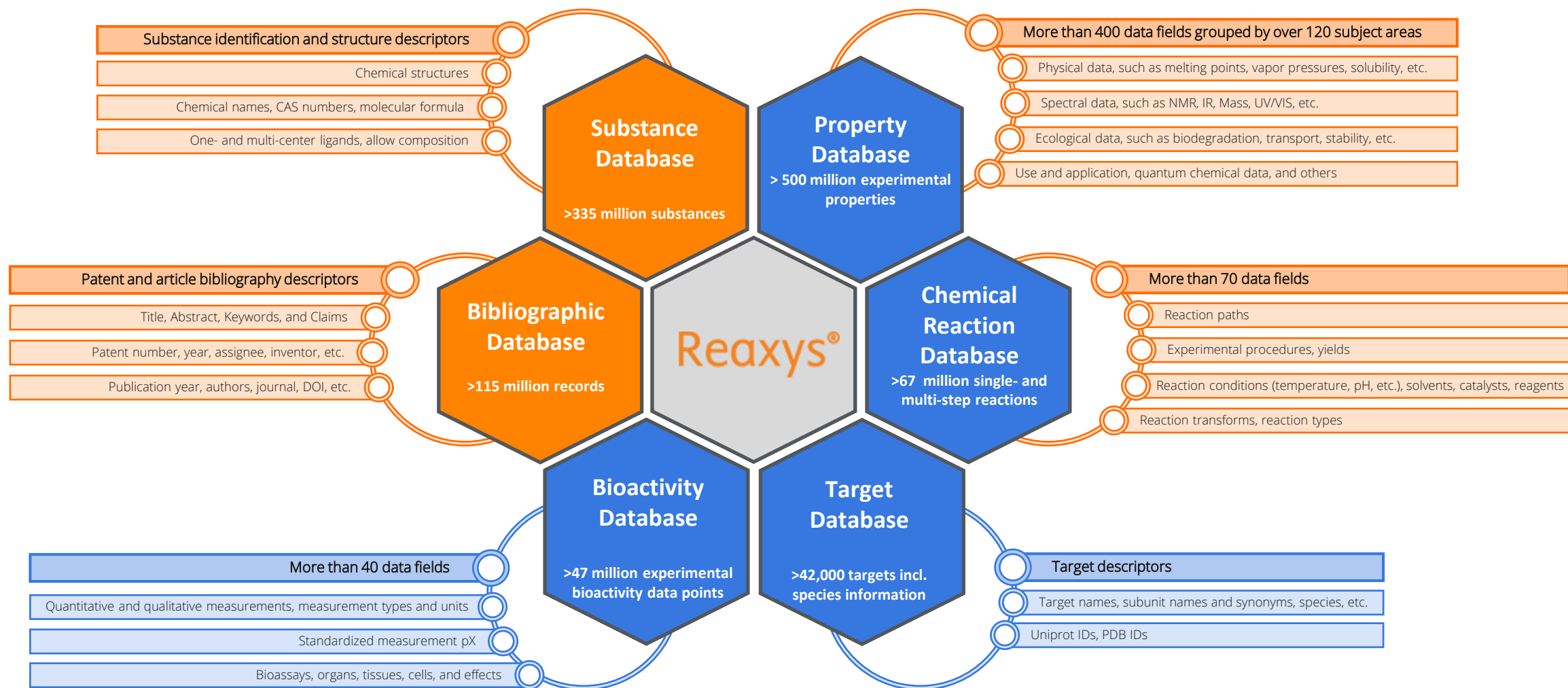
September 25, 2024



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



Reaxys today*



Manually annotated

Both automatically and manually annotated

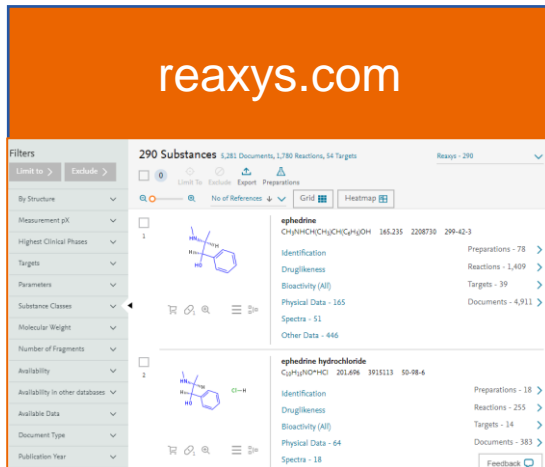
*data valid for September 16, 2024

Data is a new gold



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


reaxys.com



The screenshot shows the Reaxys web interface with a search results page. The top header displays 'reaxys.com'. Below it, there are search filters and a list of results. Two results are visible: 'ephedrine' and 'ephedrine hydrochloride'. Each result includes a chemical structure, identification information (name, SMILES, CAS, EINECS), and various data points like 'Preparations', 'Reactions', 'Targets', 'Documents', 'Spectra', and 'Other Data'.

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

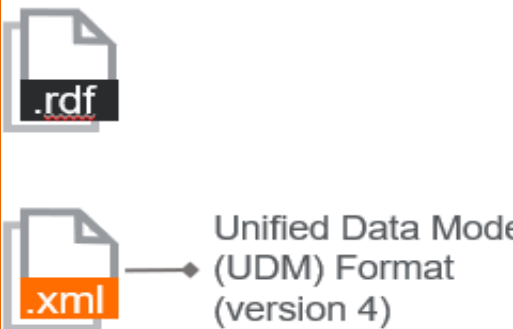
API



The logo for the Reaxys API, featuring the word 'Reaxys' in orange above three interlocking gears (two blue, one orange) and the word 'API' in black below them, all enclosed in a circular orange border.

- **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



The diagram illustrates data export options. It shows a document icon with a '.rdf' label and another document icon with a '.xml' label. An arrow points from the '.xml' icon to the text 'Unified Data Model (UDM) Format (version 4)'.

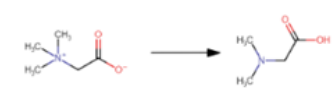
- **Machine** readable format for ML/AI applications or internal data lakes
- **Case-specific information** accessible (bibliography, reactions, molecules/properties, bioactivities)
- **Dataset** delivered to user

Intuitive and integrated exploration of reactions using content integration


Search Reaxys



Substance CAS Registry Number, e.g. 102625-70-7 Find >

AND



Reactions

Reaction Query :  as drawn; included: only absolute stereo, additional ring closures allowed, salts, mixtures, isotopes, charges, radicals

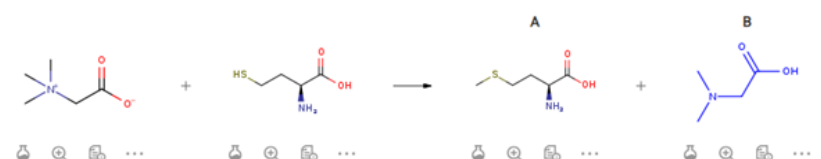
Edit in Query Builder  Create Alert 

17,438	in Reaxys	Preview Results	View Results
0	A ELN		
0	B ELN		
1,675	C ELN	Preview Results	View Results
1	D ELN	Preview Results	View Results

Reaxys - 17,438 A ELN - 0 B ELN - 0 **C ELN - 1,675** D ELN - 1

1,675 Reactions out of 0 Documents, containing 17 Substances

0 [Limit To](#) [Exclude](#) [Export](#) [Hide Conditions](#) [Search](#) [Ranking](#) [Down Arrow](#)



1 **Conditions** [Find Similar](#) **Reaction ID: 981903** [Open in database](#)

Conditions	Yield	Reference/Experiment
platinum in dmsa at 591.24896°C; under 829.5416 - 5005.5254 Torr; for 37.3183h; pH=2.295665 - 7.316745; argon;	A 8.7% B 74.8%	Oncology-FJ0 page 97 section 3
Experiment Procedure		Experiment
NextMove Reaction Type: reduction MW Largest Product: 149.2134 g/mol		Scientist: Karen, Howell; Experiment ID: Chem-79511-Notebook-39287a/r6 Project: Oncology Created: 2010-04-01-04:00 Modified: 2010-04-03-04:00 Experimental Data Sheet
Experiment Type: synthesis type 3/w Identifiers: 22336(masterId); approved(status); [MA, FO, CB](qualifiers) Source: Central lab sourcing		
Preparative: N Conclusion Phrase: completed Analytical Data Exists: N		
betaine + L-homocysteine => L-methionine + N,N-dimethylglycine (equation)		

Available through [Close](#)

original record

Additional Information on Reactants/Products

Reactants [Show/Hide columns](#)

Name	Amount	Volume	Color	Equivalents
CHEBI:17750	0.0693 mol	0.948 l	blue	0.62
Details				
CHEBI:58199	0.0538 mol	0.518 l	muddy	1.71
Details				

[Products](#)

Additional Information on Reagents/Solvents/Catalysts

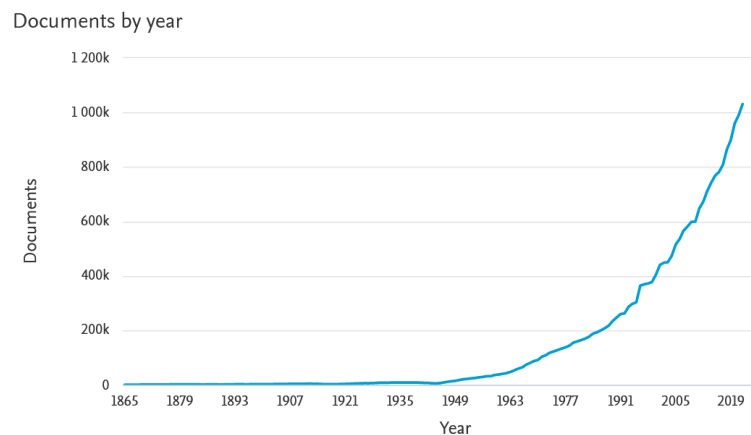
[Reagents/Solvents/Catalysts](#)

Data is a new **gold** **curse**

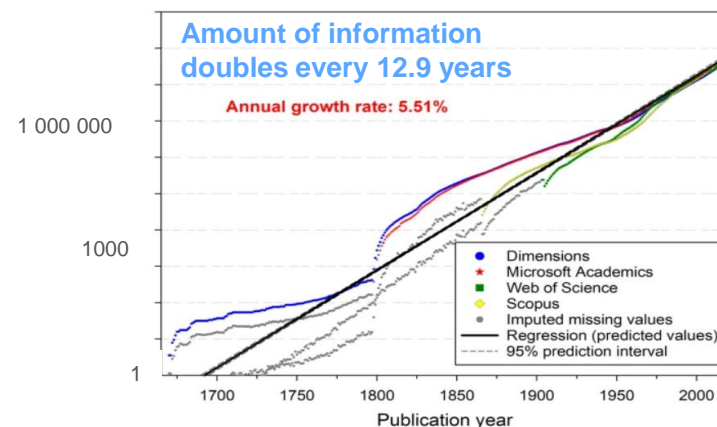


Data challenges

- 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

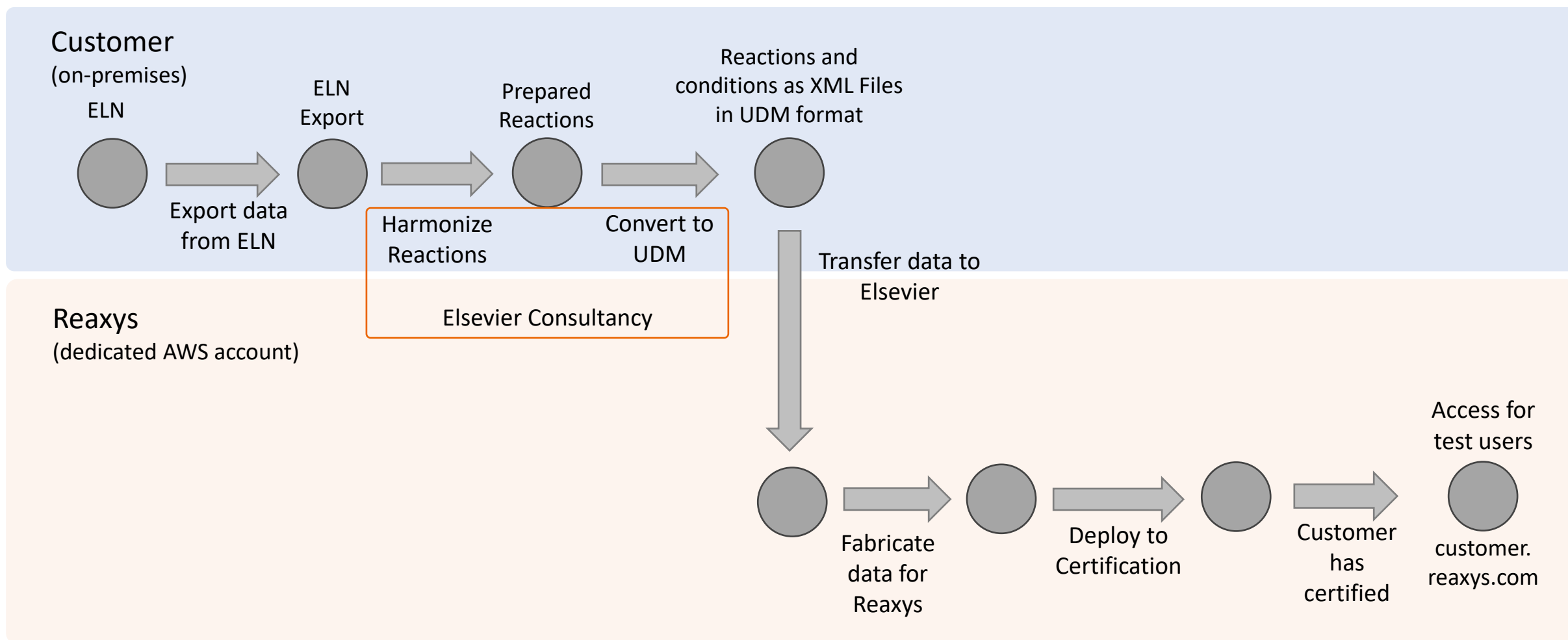
Data preparation and curation is essential

- 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

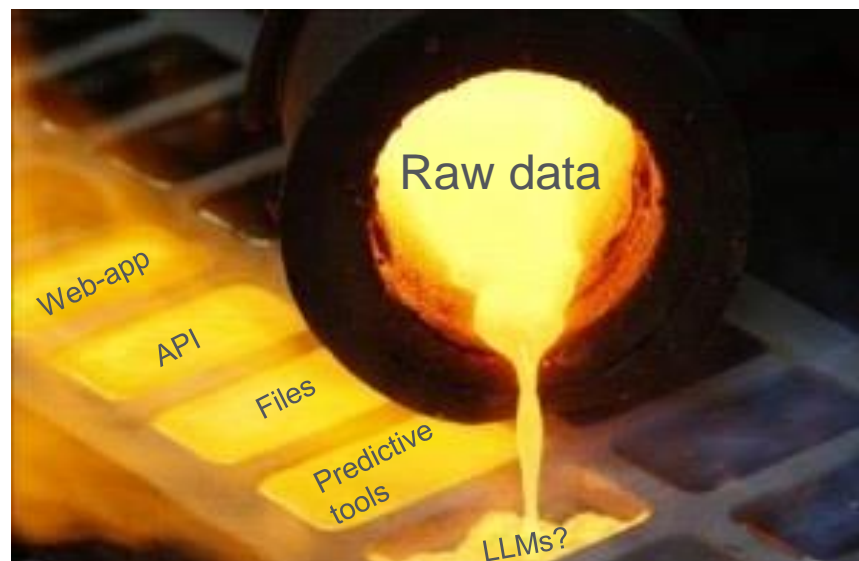


Elsevier stock photo

Content integration onboarding process



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



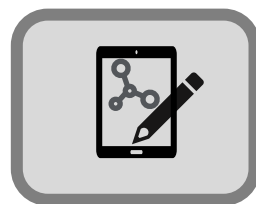
How should reaction data look like to enable easy modelling?



Can be processed by regular cheminformatics software



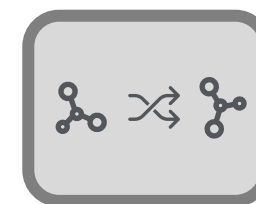
Homogeneous representation



Free of obvious errors



SMILES compatible



Atom mapped

Challenging cases:

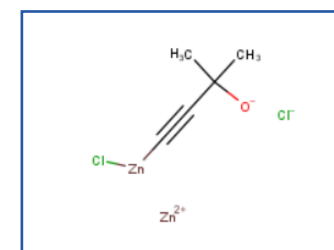
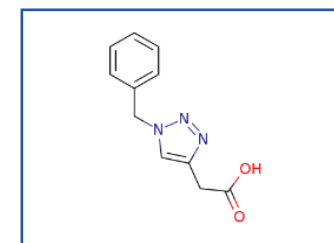
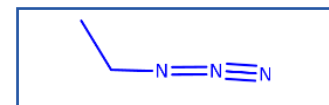
- Organometallics
- Inorganics
- Complexes
- Stereochemistry
- S-group fields

Challenging cases:

- Radicals
- Multiplicity
- Salts/mixtures
- Cycles
- Non-covalent bonds

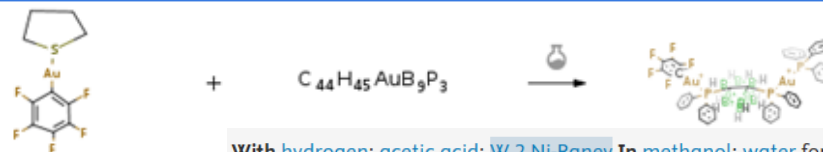
Reaxys reaction data:

- High-quality – only manually annotated
- Comprehensive
- Well-curated
- Standardized
 - Functional group representation is unique
 - Kekule structures are given
 - Salts, mixtures, cocrystals 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;

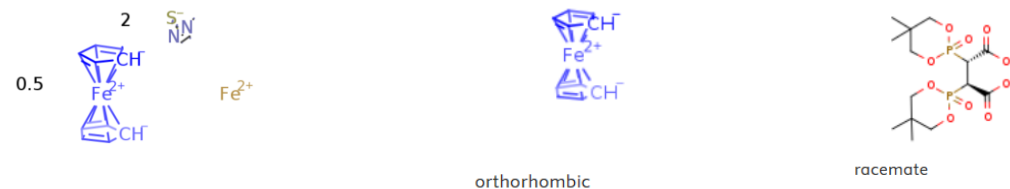
Simplicity of visual representation



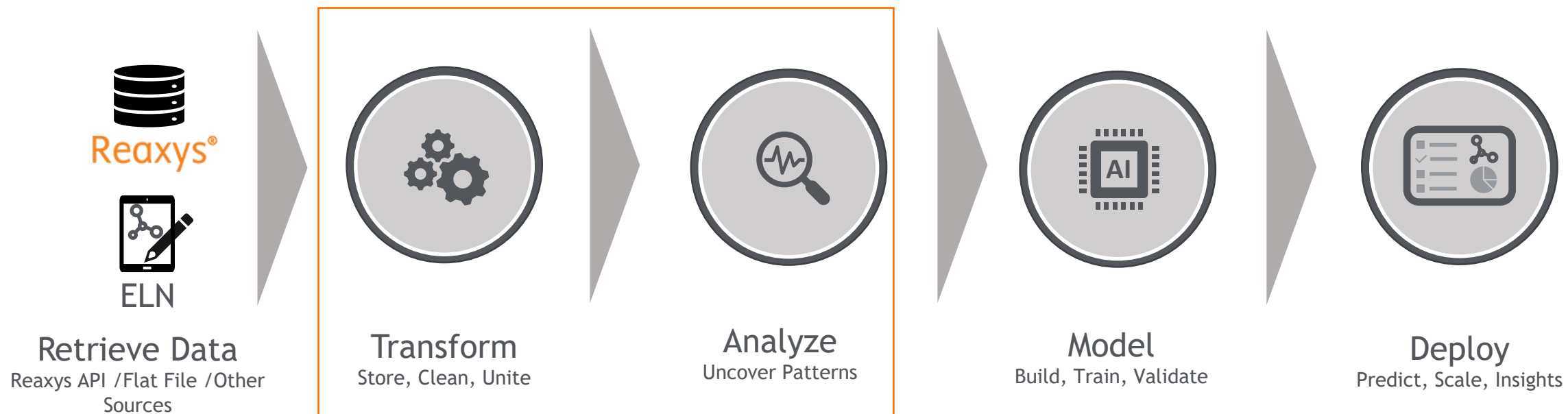
Discoverability



Coverage of whole chemistry



Typical Data Science Workflow



Role assignment

Deduplication

Atom mapping

Format conversion

Class creation

Fingerprints

Rule extraction

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



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

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



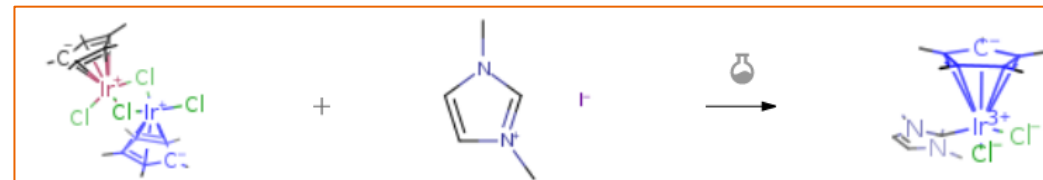
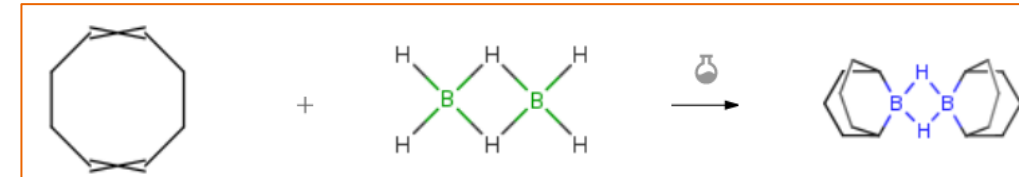
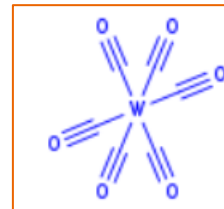
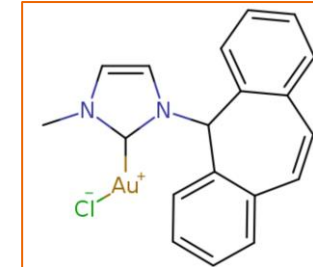
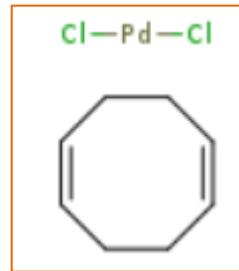
Pending.AI feedback:

- **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%)**



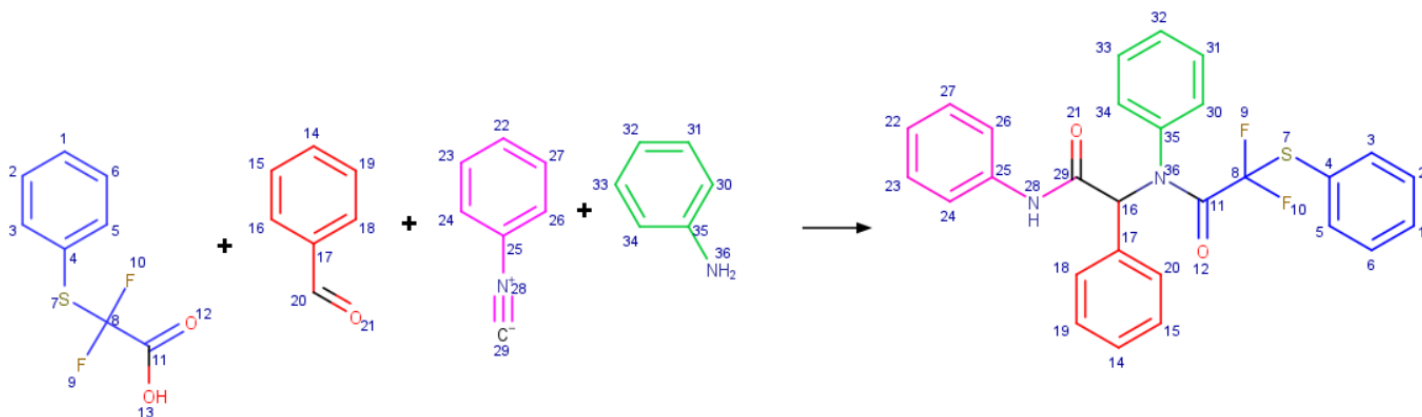
Pending.AI

Questions to community: how to represent these compounds? Chemical conventions contradict cheminformatics 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

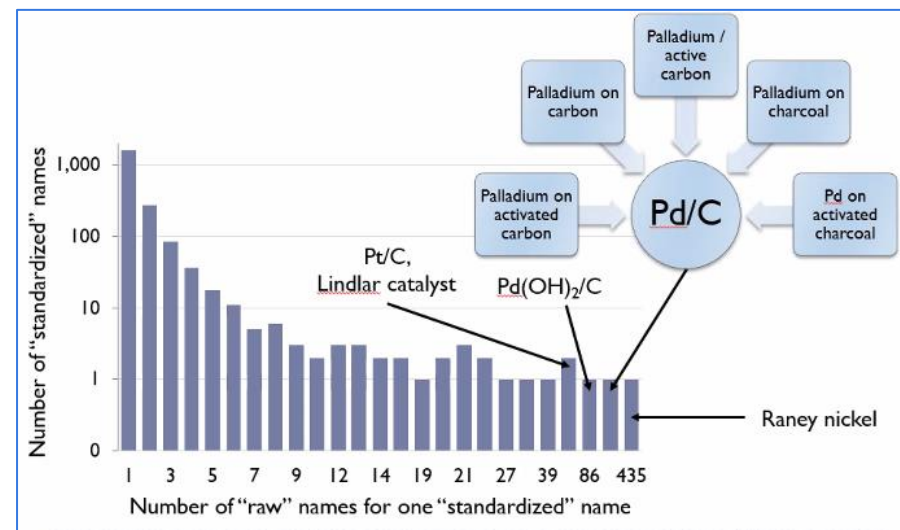
Advertisement

	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:



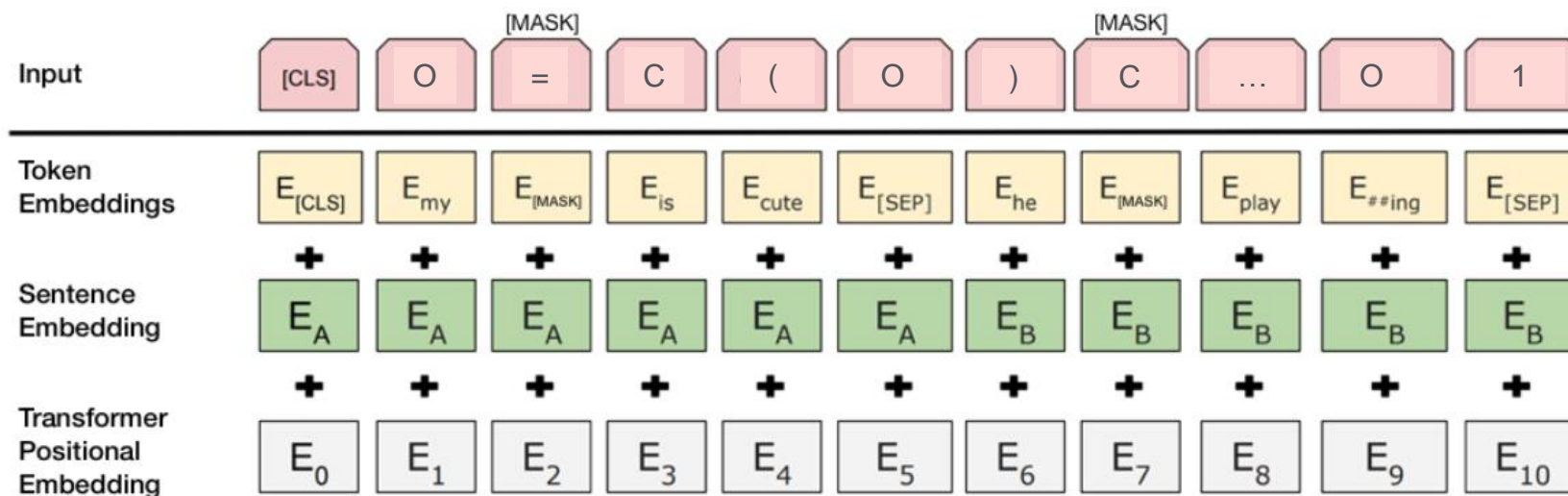
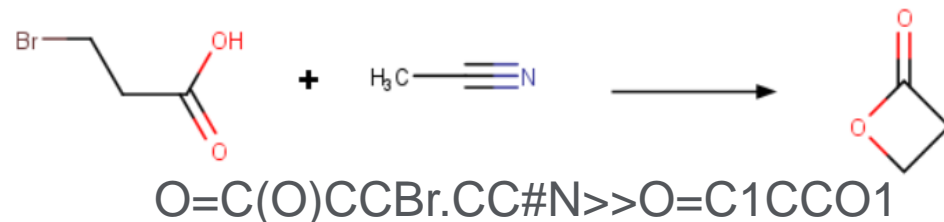
Objective

- **Goal:** Create a model to predict reaction success (classification: yield \geq 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?

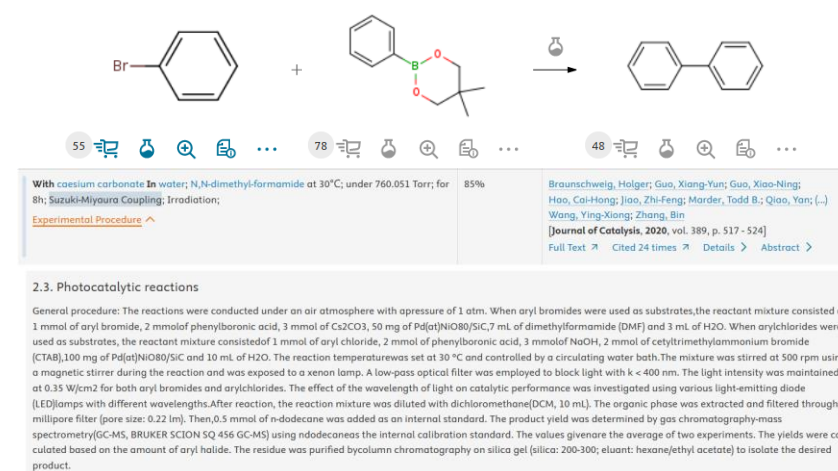
- literature yields biased, not always reliable.
- ELN data contains more lower-yielding rxns.

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



How can we improve the model?

- Add chemical knowledge!
 - Add context from text reaction description
 - Add context by augmenting pretraining task



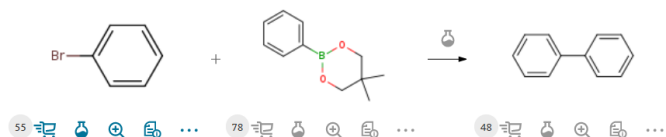
Adding chemical context from reaction description: encoders that describe reaction procedure and SMILES are trained together to minimize the distance between actual text and SMILES embedding

With cesium carbonate in water; N,N-dimethyl-formamide at 30°C, under 760.051 Torr; for 85%
8h; Suzuki-Miyaura Coupling; Irradiation;
[Experimental Procedure](#)

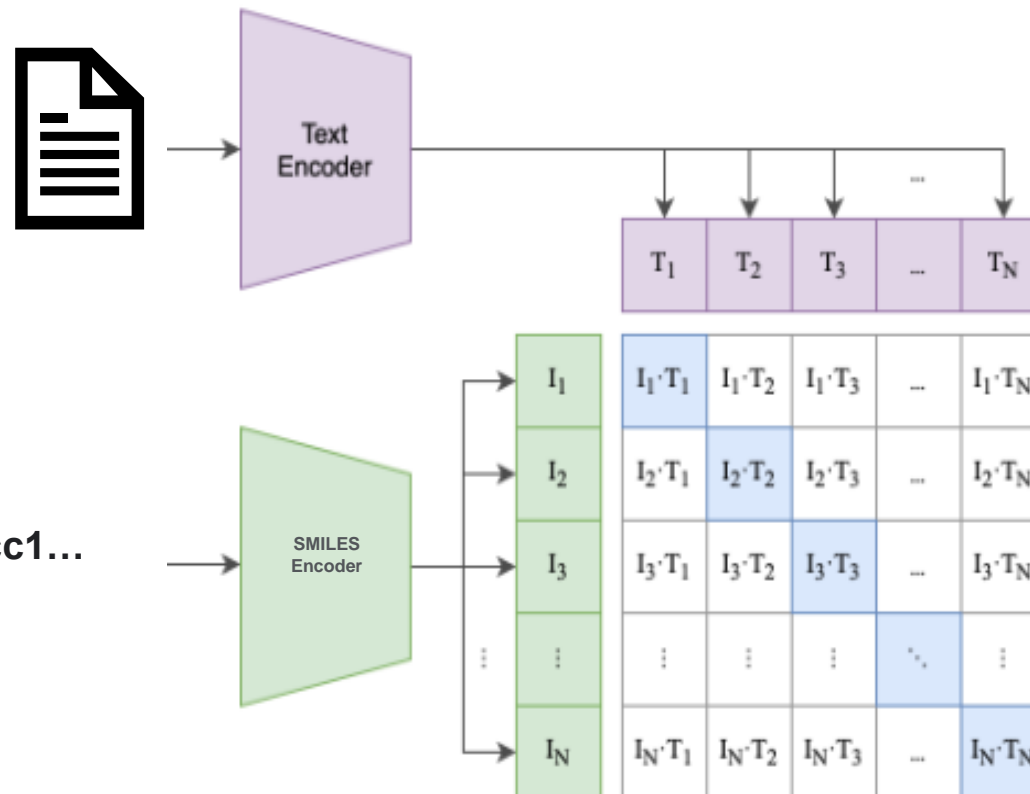
Braunschweig, Holger; Guo, Xiang-Yun; Guo, Xiao-Ning; Hao, Cai-Hong; Jiao, Zhi-Feng; Marder, Todd B.; Qian, Yan; (...) Wang, Ying-Xiang; Zhong, Bin
Journal of Catalysis, 2020, vol. 389, p. 517-524
[Full Text](#) [Cited 24 times](#) [Details](#) [Abstract](#)

2.3. Photocatalytic reactions

General procedure: The reactions were conducted under an air atmosphere with a pressure of 1 atm. When aryl bromides were used as substrates, the reactant mixture consisted of 1 mmol of aryl bromide, 2 mmol of phenylboronic acid, 3 mmol of Cs2CO3, 50 mg of Pd(dppf)Cl2·CH2Cl2, 7 mL of dimethylformamide (DMF) and 3 mL of H2O. When aryl chlorides were used as substrates, the reactant mixture consisted of 1 mmol of aryl chloride, 2 mmol of phenylboronic acid, 3 mmol of NaOH, 2 mmol of cetyltrimethylammonium bromide (CTAB), 100 mg of Pd(dppf)Cl2·CH2Cl2 and 10 mL of H2O. The reaction temperature was set at 30 °C and controlled by a circulating water bath. The mixture was stirred at 500 rpm using a magnetic stirrer during the reaction and was exposed to a xenon lamp. A low-pass optical filter was employed to block light with $\lambda < 400$ nm. The light intensity was maintained at 0.35 W/cm² for both aryl bromides and aryl chlorides. The effect of the wavelength of light on catalytic performance was investigated using various light-emitting diode (LED) lamps with different wavelengths. After reaction, the reaction mixture was diluted with dichloromethane (DCM, 10 mL). The organic phase was extracted and filtered through a millipore filter (pore size: 0.22 μ m). Then, 0.5 mmol of n-dodecane was added as an internal standard. The product yield was determined by gas chromatography-mass spectrometry (GC-MS; BRUKER SCIION SQ 456 GC-MS) using n-dodecane as the internal calibration standard. The values given are the average of two experiments. The yields were calculated based on the amount of aryl halide. The residue was purified by column chromatography on silica gel (silica: 200-300; eluent: hexane/ethyl acetate) to isolate the desired product.



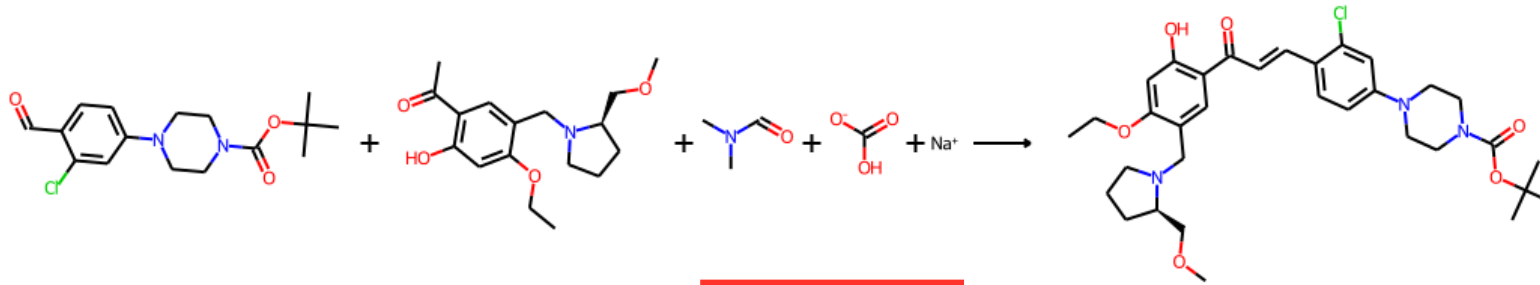
...c1ccccc1...



The model can be used to spot inconsistencies in data

- Cosine similarity between SMILES and text embeddings: -0.227

SMILES:

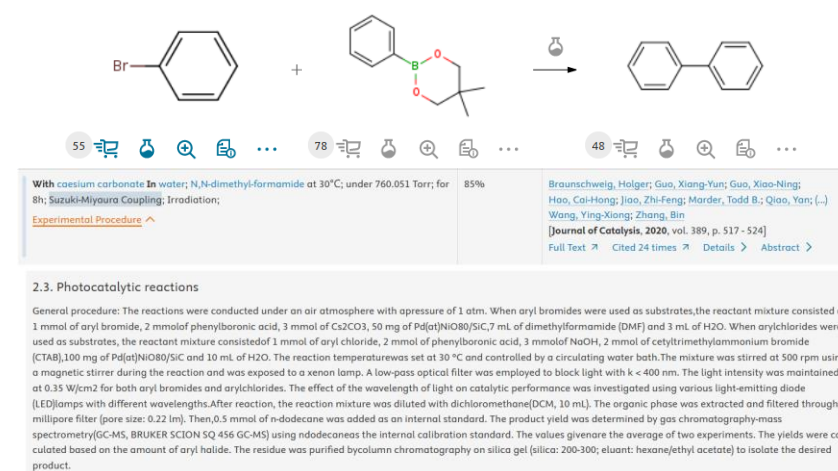


Text:

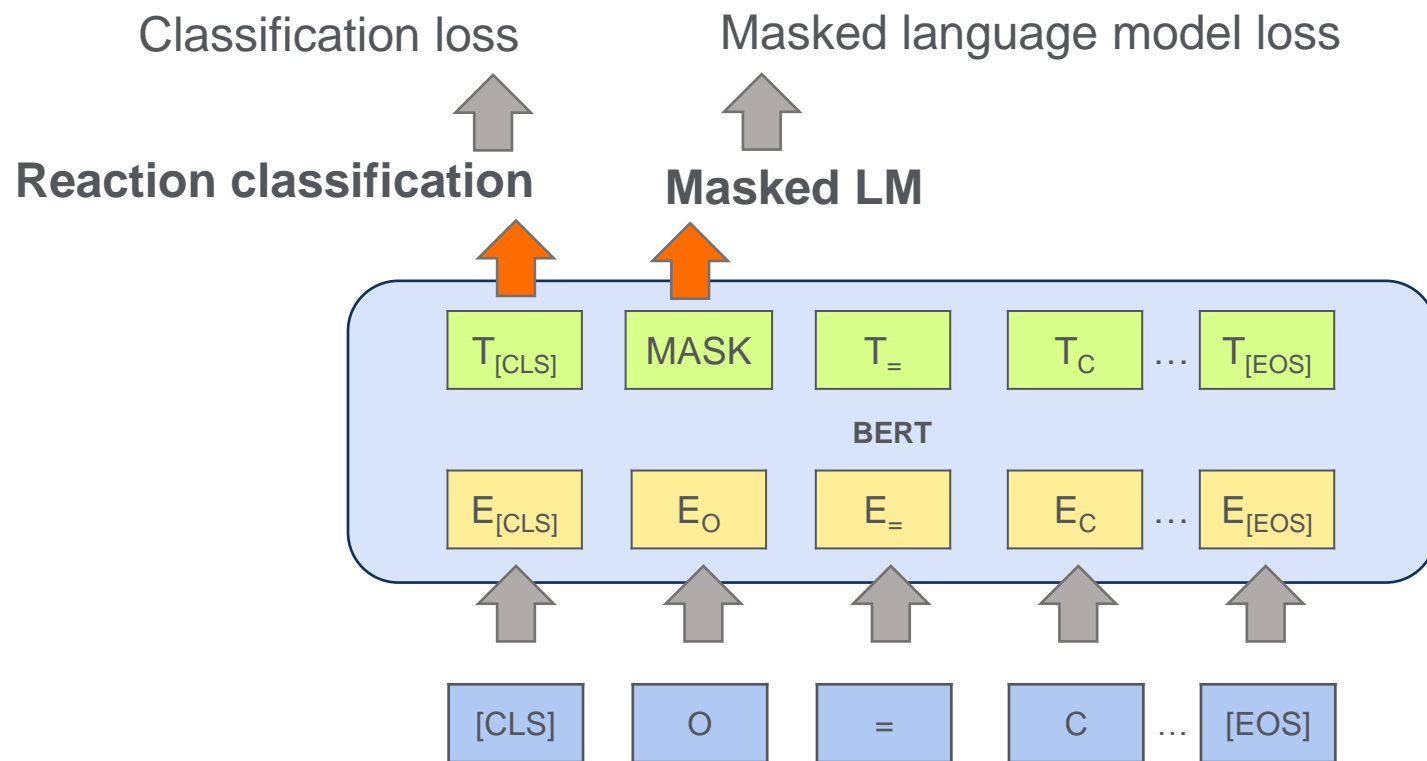
Add 303mg B-122 to 4ml EtOH, then add 221mg KOH, 640mg B-10, stir at RT, and monitor LC-MS until there is no B-122 left.

How can we improve the model?

- Add chemical knowledge!
 - Add context from text reaction description
 - Add context by augmenting pretraining task

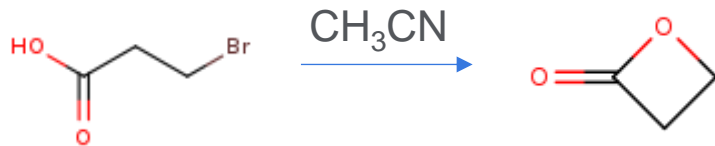


Augmented Pretraining of BERT Model: train model not only to predict masked token in reaction SMILES but also predict whether reaction looks reasonable



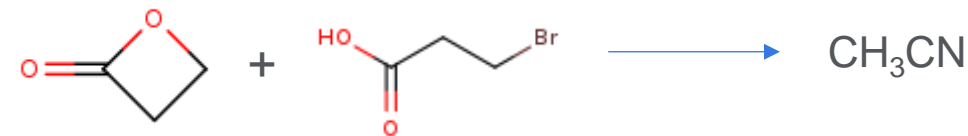
MLM Task

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



rxn classification task

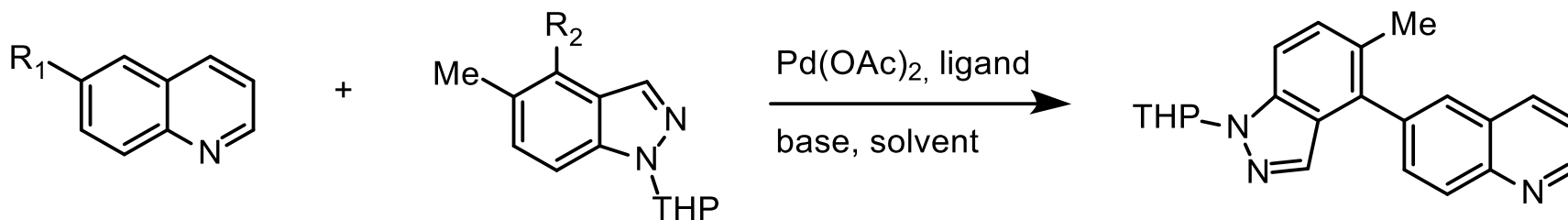
O=C(O)CCBr.O=C1CCO1>>CC#N



Model shows the best quality of prediction on Suzuki Reaction Benchmark

5760 reactions¹

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