ICANN24

33rd International Conference on Artificial Neural Networks

Curating reagents in chemical reaction data with an interactive reagent space map

Mikhail Andronov[1], Natalia Andronova, Michael Wand[1], Jürgen Schmidhuber[1,3], Djork-Arné Clevert[2]

USI/SUPSI IDSIA — 1

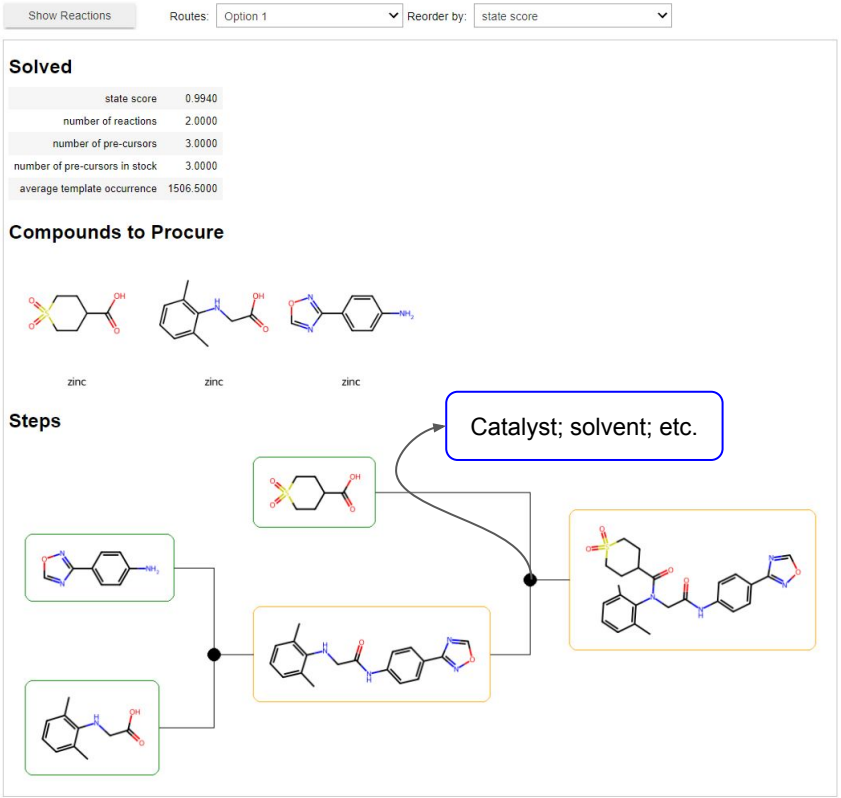
Pfizer — 2

KAUST — 3

Chemical reaction data curation

Currently, CASP (computer-aided synthesis planning) systems are often powered by machine learning.

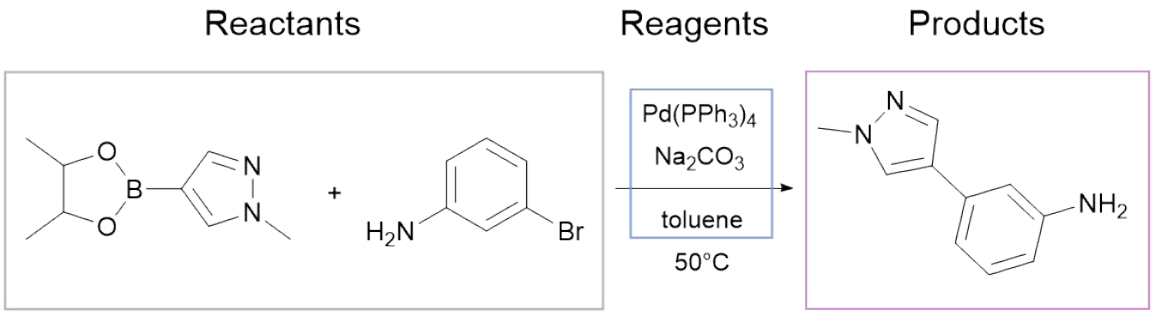
There should be no errors in training data for product, reagent/condition, or single-step retrosynthesis prediction.



Reagents

Reactants contribute atoms to products.
Reagents are catalysts, solvents, and other auxiliary substances that make the reaction possible.

All molecules can be written as strings in the SMILES format. SMILES are not unique for a molecule.



Problems with reagents in reaction data

Besides erroneous reactions, missing molecules and wrong atom mapping, there are problems with reagent records.

1. Missing reagent roles: only “catalyst”, “solvent”, and “other”, and all assigned by a machine.
2. Inconsistent SMILES:
   - `[OH-].[Na+]` or `O[Na]` for sodium hydroxide
   - `[CH2-]CCC.[Li+]` or `[Li]CCCC` for n-Butyllithium
3. Reactants recorded as reagents (above the arrow):
   - Building blocks for Suzuki coupling
   - Grignard reagents
   - Building blocks for amide coupling
   - etc…

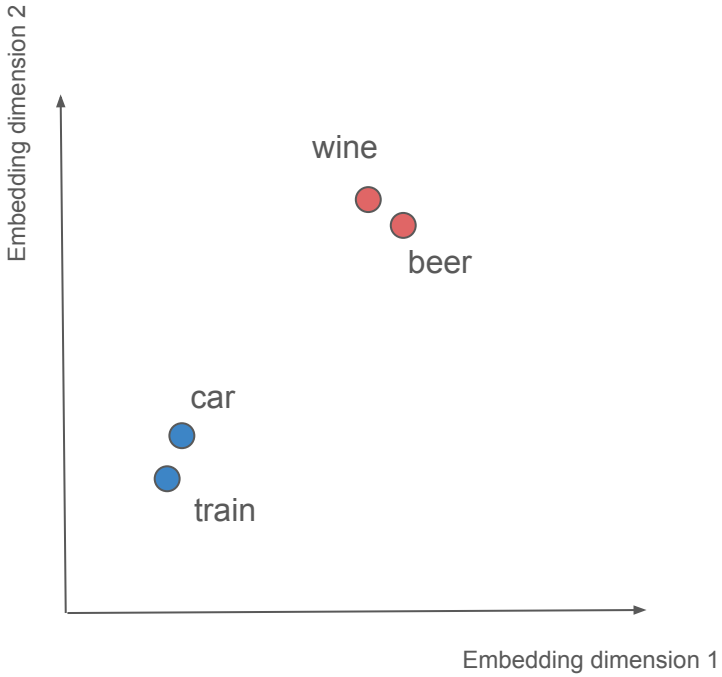
Distributional hypothesis

Comes from natural language processing.

"*You shall know a word by the company it keeps*"

"*Similar words occur in similar contexts*"

You drank a bottle of wine and got drunk.
You drank a bottle of beer and got drunk.
…
I usually travel by car.
I usually travel by train.

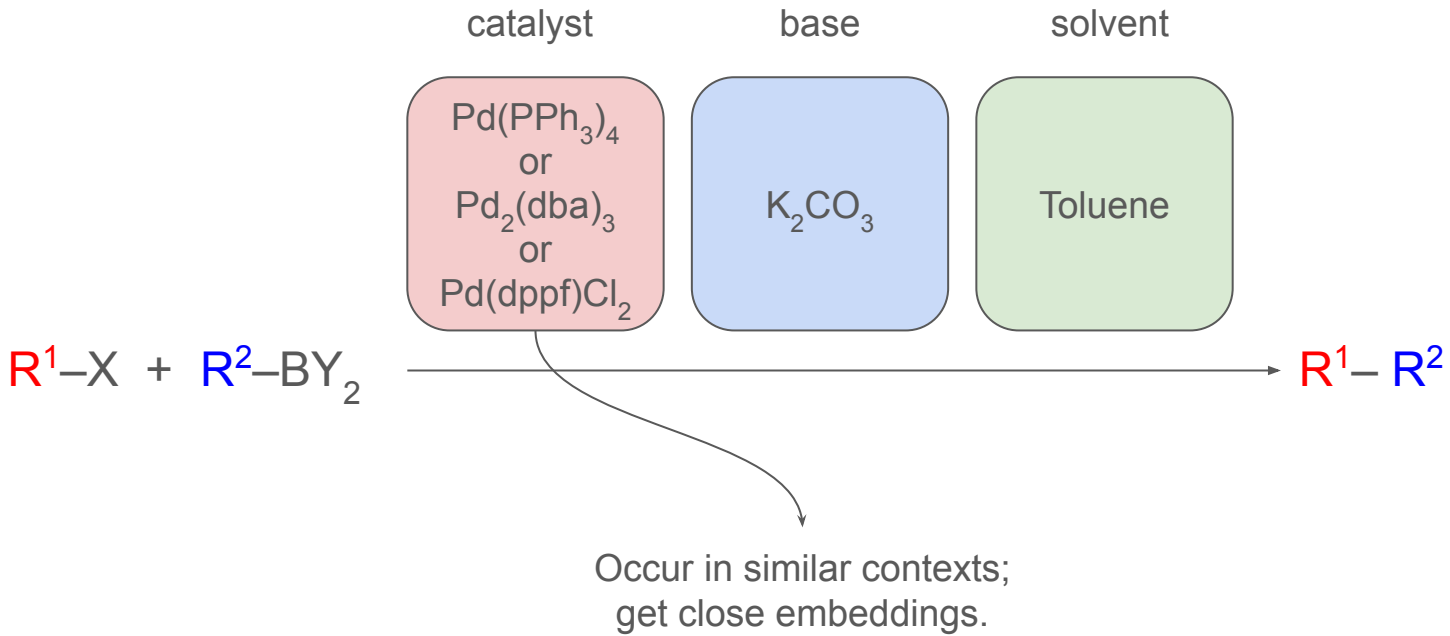


Vector representations of words cluster by meaning [1]

[1]. Mikolov et al. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Distributed reagent representations

Suzuki coupling example:



Distributed reagent representations

Count reagents in tuples

```
...
CCO;c1ccccc1
[Na+].[H-];C1CCOC1;
CO;c1cccc1
...
```

and get the PMI matrix.

Obtain reagent embeddings by factorizing the PMI matrix with SVD (singular value decomposition).

Equivalent to word2vec [2]

pointwise mutual information score

$$\underset{N\times N}{\mathrm{PMI}} \approx \underset{N\times d}{U} \cdot \underset{d\times d}{\Sigma} \cdot \underset{d\times N}{V}$$

$$\mathrm{PMI}(r_1, r_2) = \log_2 \frac{P(r_1,r_2)}{P(r_1)P(r_2)}$$

reagents

relative occurrence frequency

reagent embeddings

reagent number

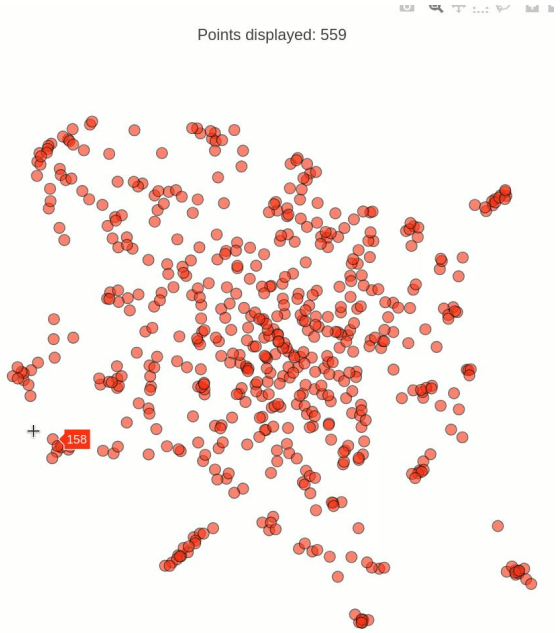
embedding size

[2]. Levy et al. 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics, 3, pp.211-225.

Embeddings in the web application

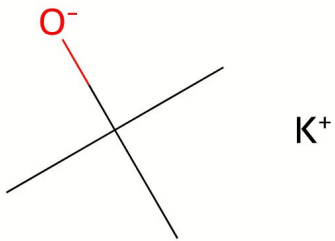
Rh·

[Rh]



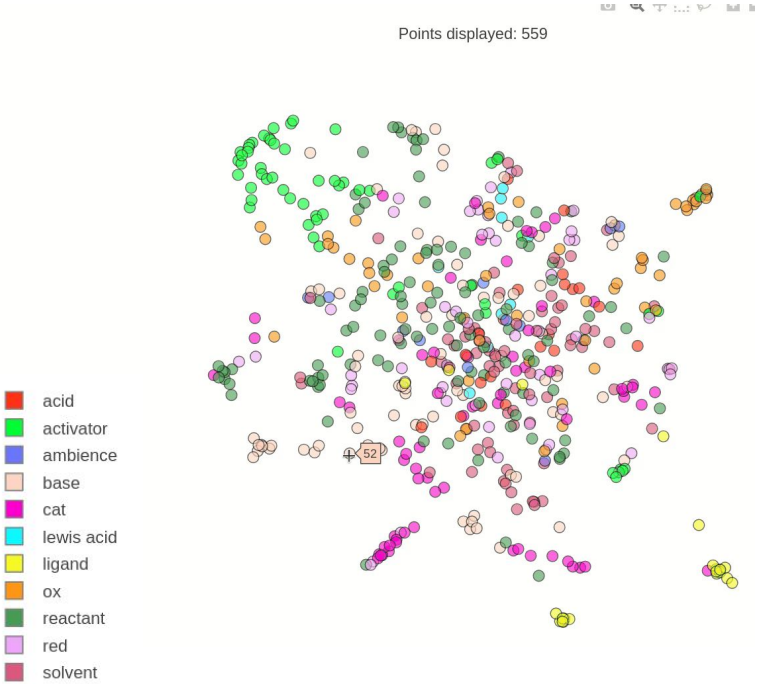
We project embeddings on the plane with UMAP and explore it in a custom web application

Labeling reagents with roles (manually)



CC(C)(C)[O-].[K+]

Potassium t-butoxide

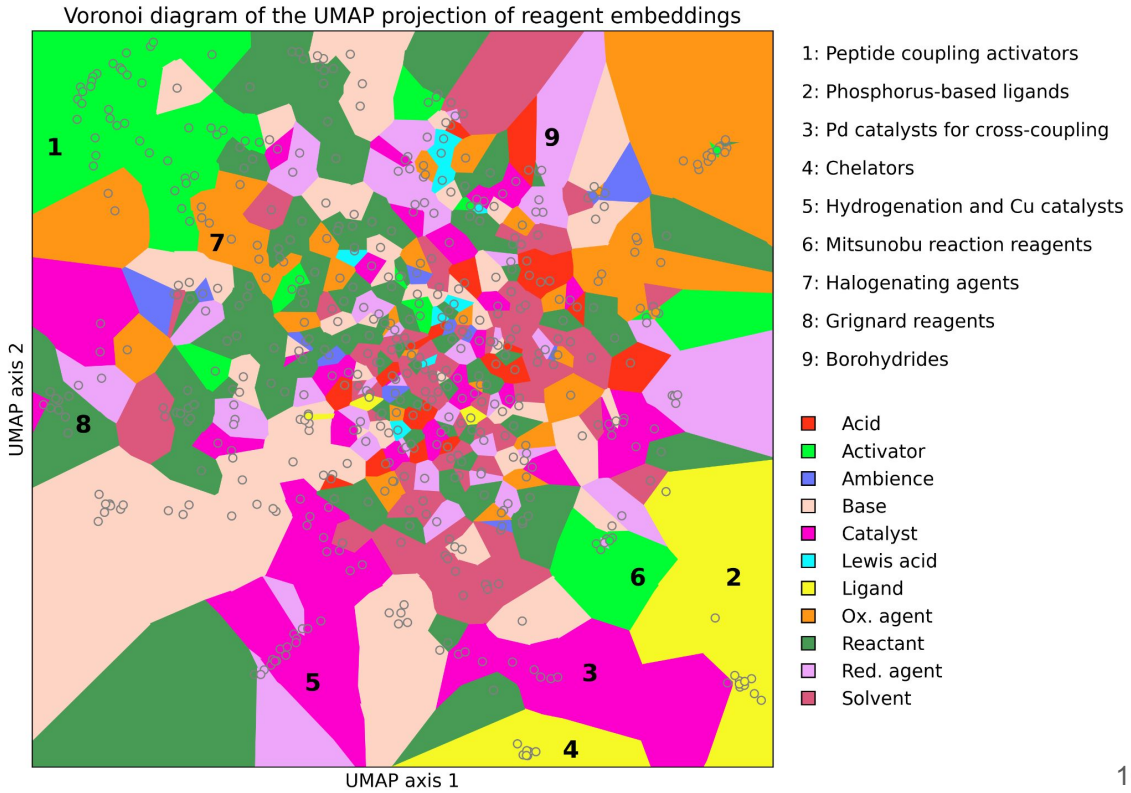


We categorize the USPTO reagents into ten detailed roles. Emergent role clusters help with that.

Regions of similar reagents

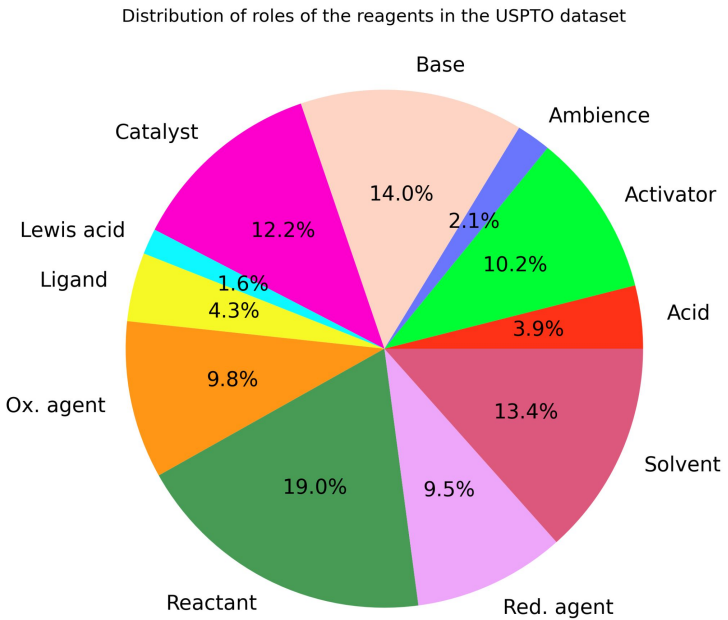
Reagents tend to cluster together according to their action in reactions. There are clusters of bases, catalysts, etc.

We highlight it with a Voronoi diagram.



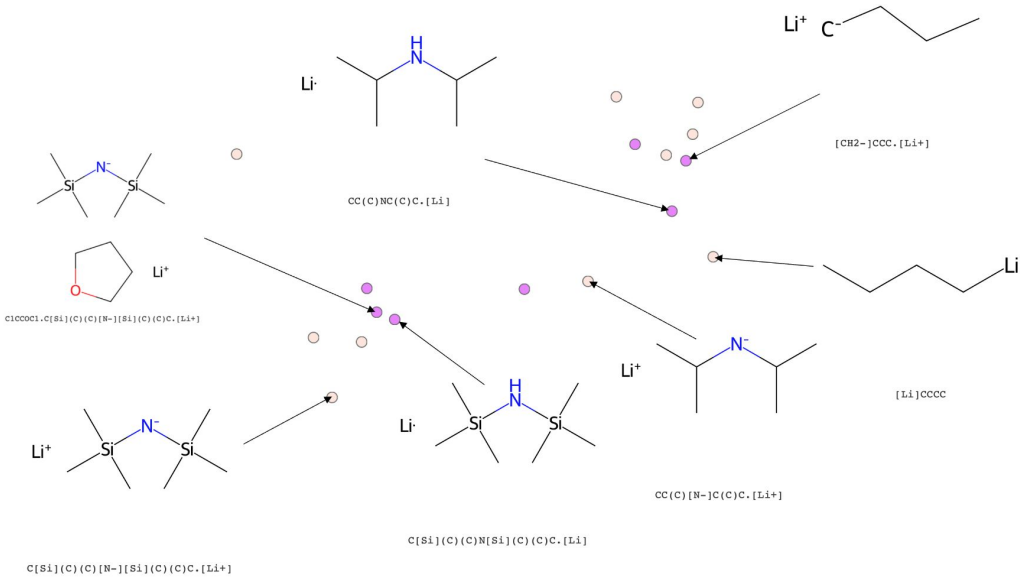
Reagent roles in USPTO

After role assignment, 19% of reagents turn out to be reactants. Placing reactants correctly will improve the quality of the data.



Detecting redundant SMILES

We easily detect redundant reagent SMILES and standardize them, improving the data quality.



Seven SMILES, three unique reagents

Conclusion

- We propose a new approach to visual exploration of large reaction datasets based solely on reagents.
- We label about 600 reagents present in USPTO into their detailed roles. Role information is often instrumental for tasks such as reagent prediction.We build an interactive web application suitable for the exploration of any reaction datasets.
- We build an interactive web application suitable for the exploration of any reaction datasets.

https://github.com/Academich/reagent_emb_vis

Thank you for your attention!