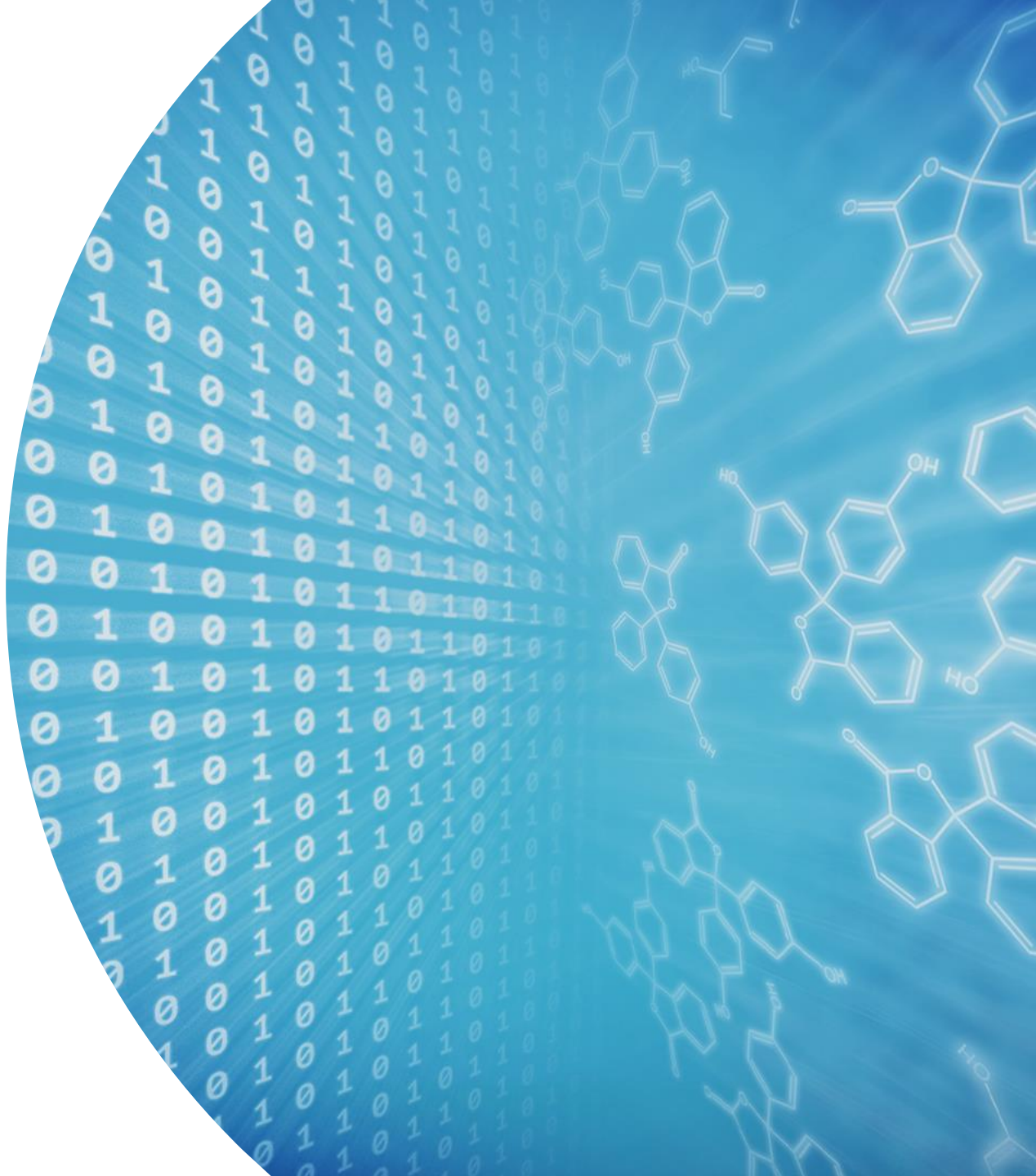
AstraZeneca

Multiobjective Reaction Similarity with the RDKit

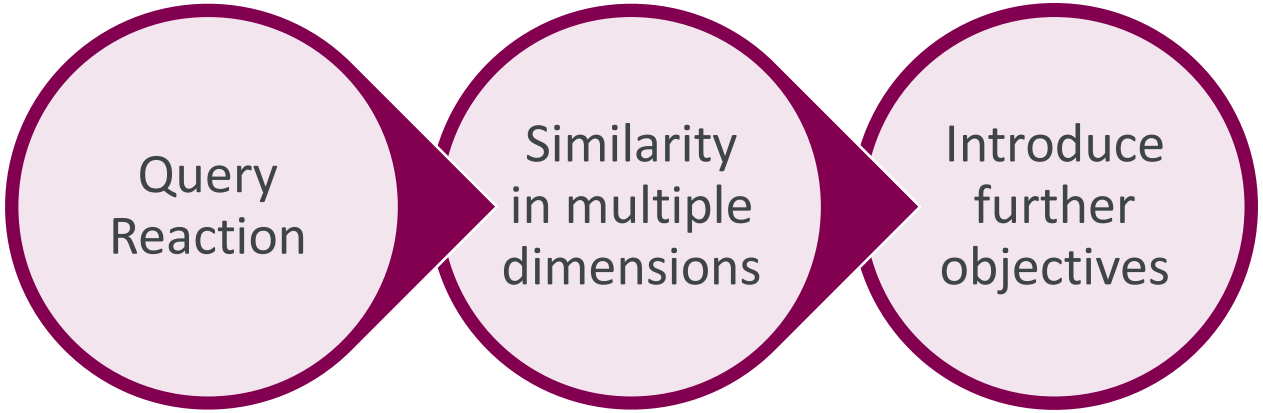
Christoph Bauer,

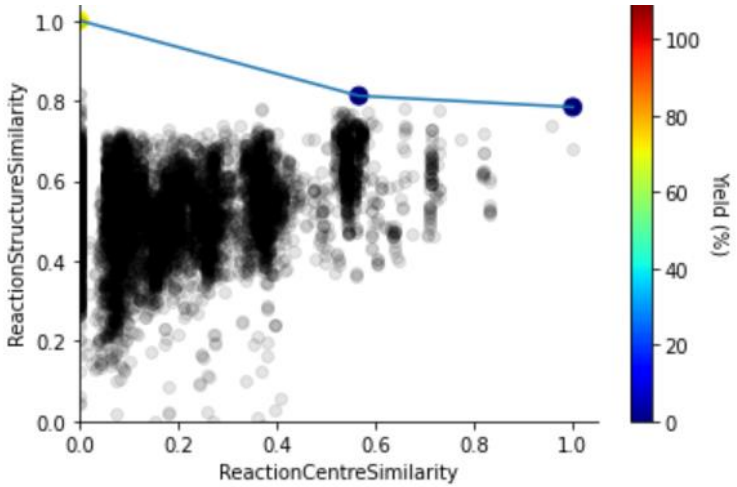
Data Science & Modelling, Pharmaceutical Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

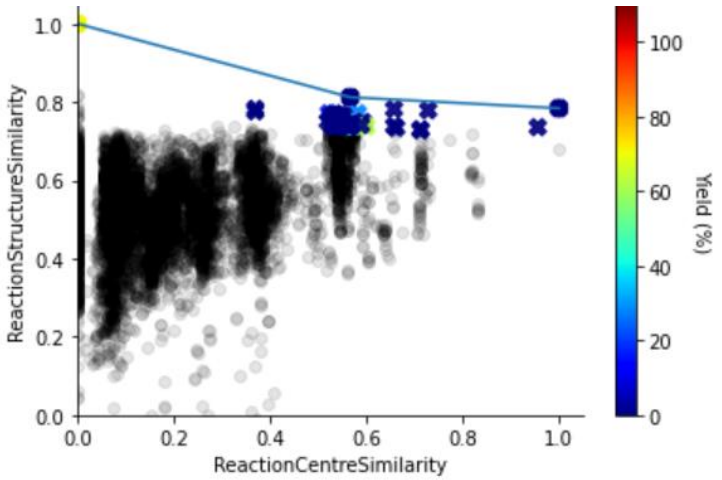
2021-10-14

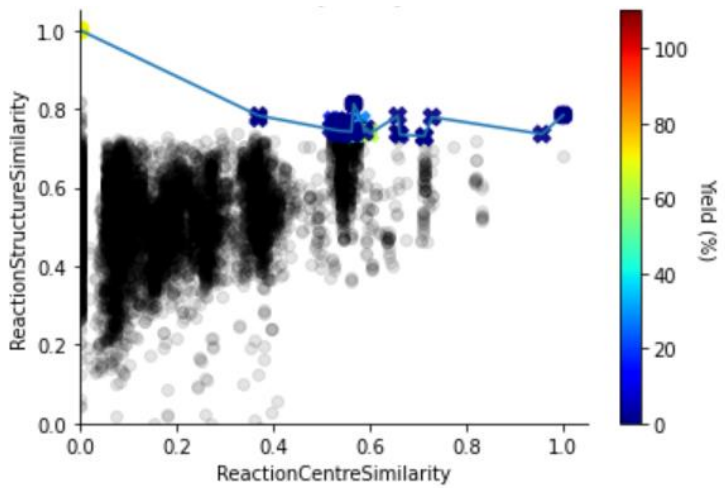


Motivation: Synthetic Planning with Multiple Objectives

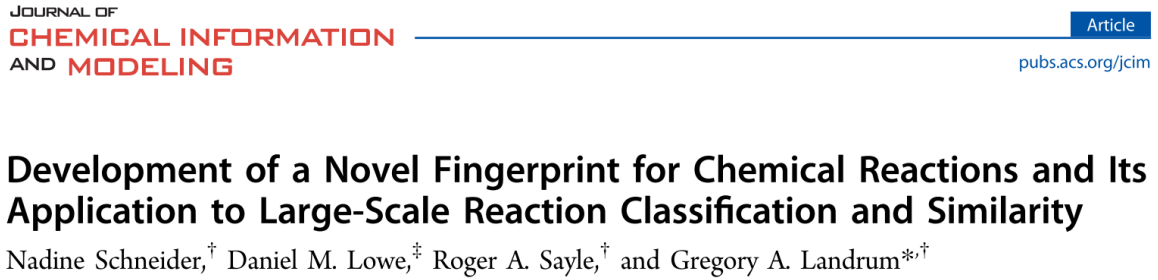








Public Dataset and Subset for Demonstration



Source:
J. Chem. Inf. Model. **2015**, 55, *1*, 39−53.
erratum: *J. Chem. Inf. Model.* **2015**, 55, *2*, 474.

- Public dataset on which reaction difference usage was demonstrated
- Using 50,000 reactions derived from USPTO data from the supporting information
- 1,000 Chloro-Suzuki type reaction subset (reaction class = 3.1.6)

Axes of Reaction Similarity

Reaction Structure Similarity

- Representation: Reaction difference fingerprint (2048 bits) + Morgan fingerprint of the product (radius = 3, 2048 bits)
- Computed with rdkit
- Similarity measure: Dot product

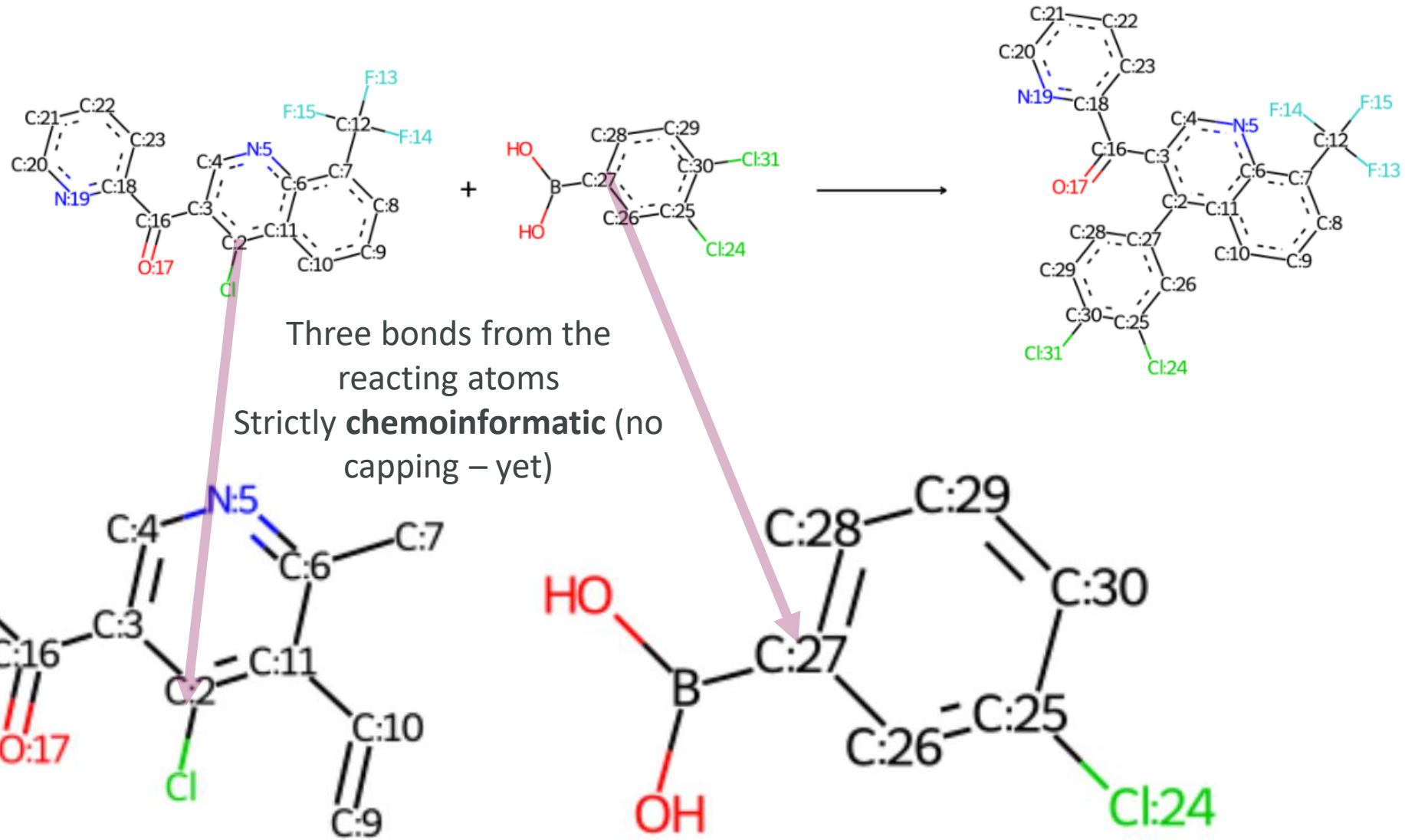
→ Represents the reaction as a **whole** and weights product.

Reaction Centre Similarity

- Representation: Reaction centres (submol, radius=3) as fingerprints: Morgan fingerprint (radius = 3, 1024 bits per reactant) or rdkit fingerprint
- `rxn.GetReactingAtoms(mappedAtomsOnly=True)`
- Similarity measure: Tanimoto similarity

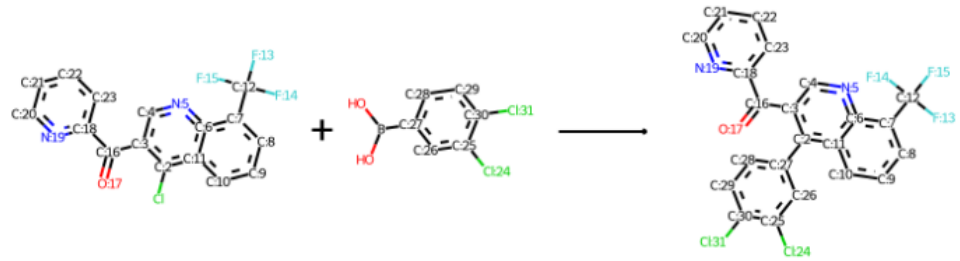
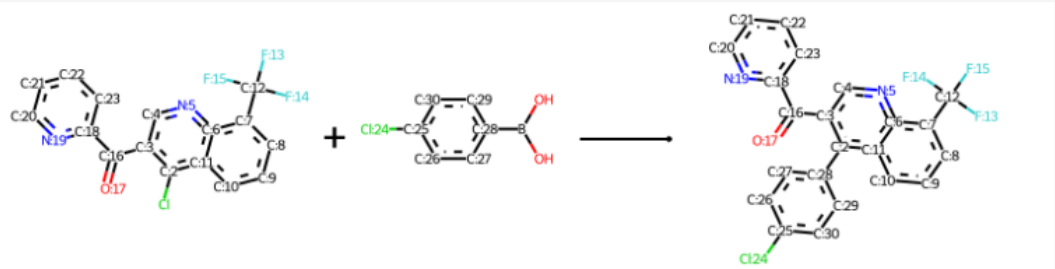
→ Represents the reaction in its **centre**.

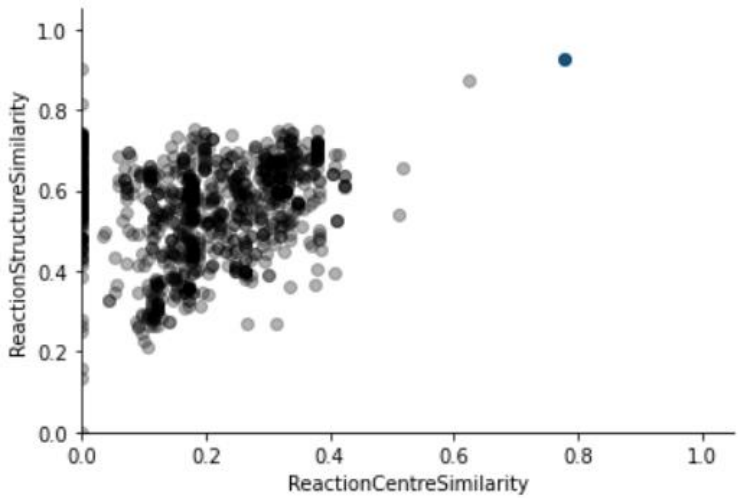
Reaction Centre Fingerprint Computation Details



US Patent ID:US08383620

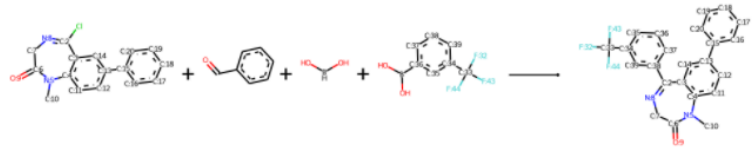
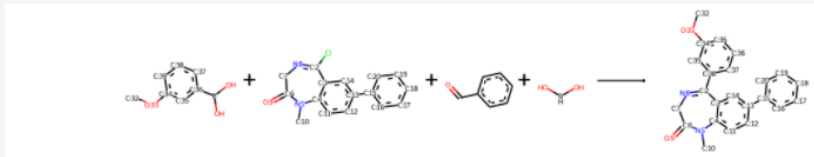
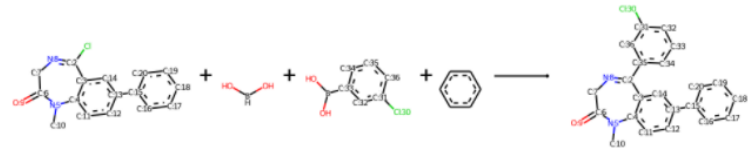
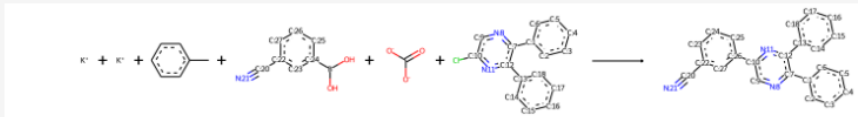
Results: Case 1 – One Dominant Point

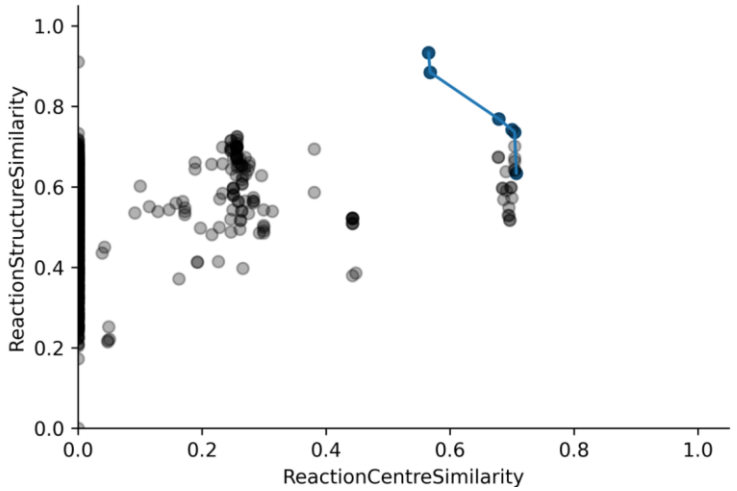
	ID	rxn Image	ReactionCentreSimilarity	ReactionStructureSimilarity	USPatentID
0	Query		1.000000	1.000000	US08383620
1	512		0.681818	0.923848	US07576215



- Retrieval of a clear ”favourite”
- Minor pattern difference leads to minor subtraction in reaction structure similarity
- Penalty in reaction centre similarity for ”missing” Cl atom is high
  - → Should incorporate electronic effects

Results: Case 2 – Pareto Front

	ID	rxn Image	ReactionCentreSimilarity	ReactionStructureSimilarity	USPatentID
0	Query		1.000000	1.000000	US08383620
1	894		0.565355	0.933571	US07410963
2	236		0.567843	0.884806	US07410963
3	343		0.678533	0.768790	US08329903

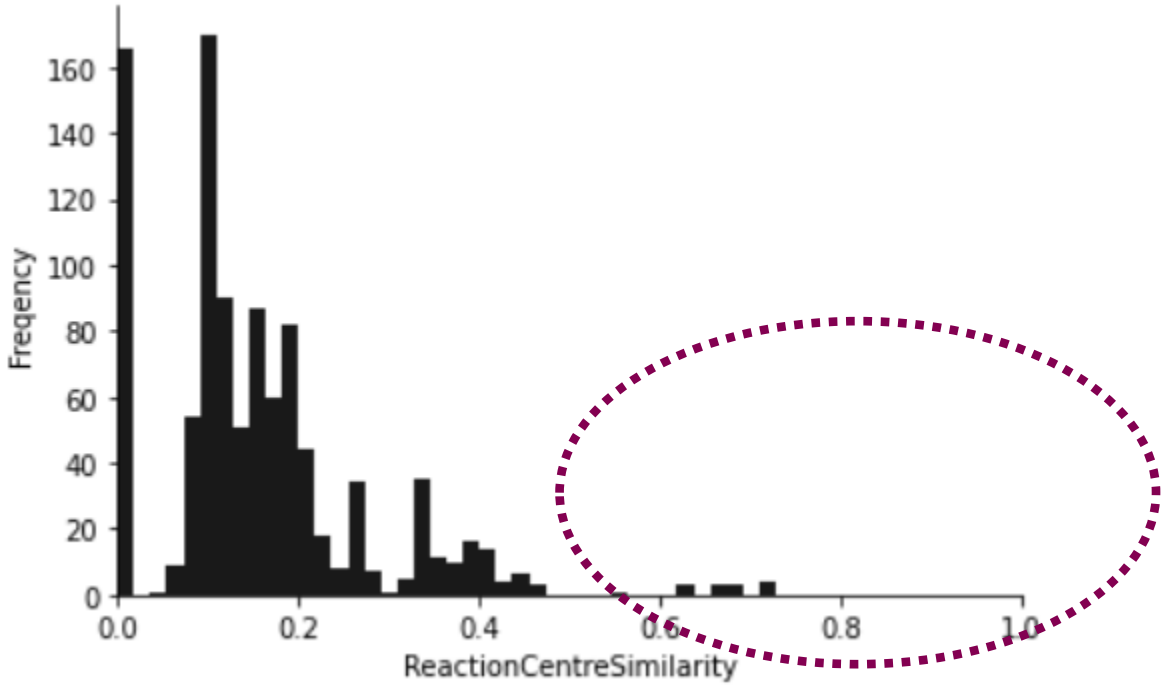


- Retrieval of Pareto Front in two dimensions
- Seven-rings give expected neighbors in reaction structure similarity
- Few close neighbours in reaction centre similarity

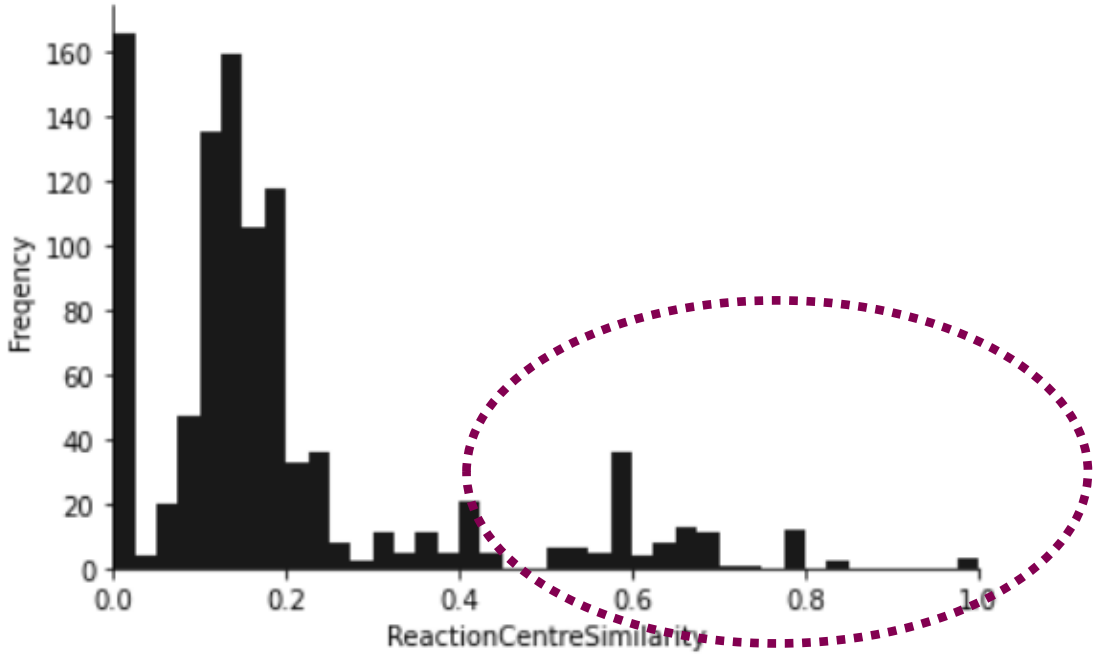
Reaction Centre Similarity: Effect of Representation

US Patent ID: US08383620

Morgan Fingerprint (radius =3)



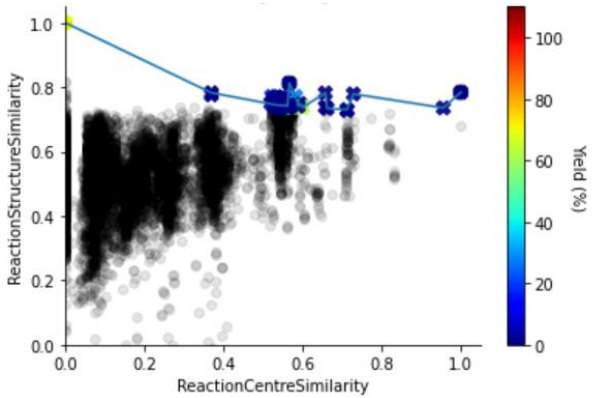
Rdkit Fingerprint



→ RDKit fingerprint-based reaction centre comparisons vs. 1000 Cl-Suzuki type reactions are more spread out.

Summary / Acknowledgements

1. We are using the rdkit to retrieve reactions from a data set with multiple objectives:
   i. Most similar overall
   ii. Most similar in the reaction centre
   → further objectives: yield, date, …



2. Preliminary investigations have shown that for reaction centres, rdkit fingerprints give a more widely distributed similarity histogram than Morgan fingerprints.

Acknowledgements:

❖Per-Ola Norrby, Data Science & Modelling, Pharmaceutical Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

❖Thierry Kogej, Molecular AI, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden