How to compare ultra-large chemical spaces?

Uta Lessel
Medicinal Chemistry

Boehringer Ingelheim

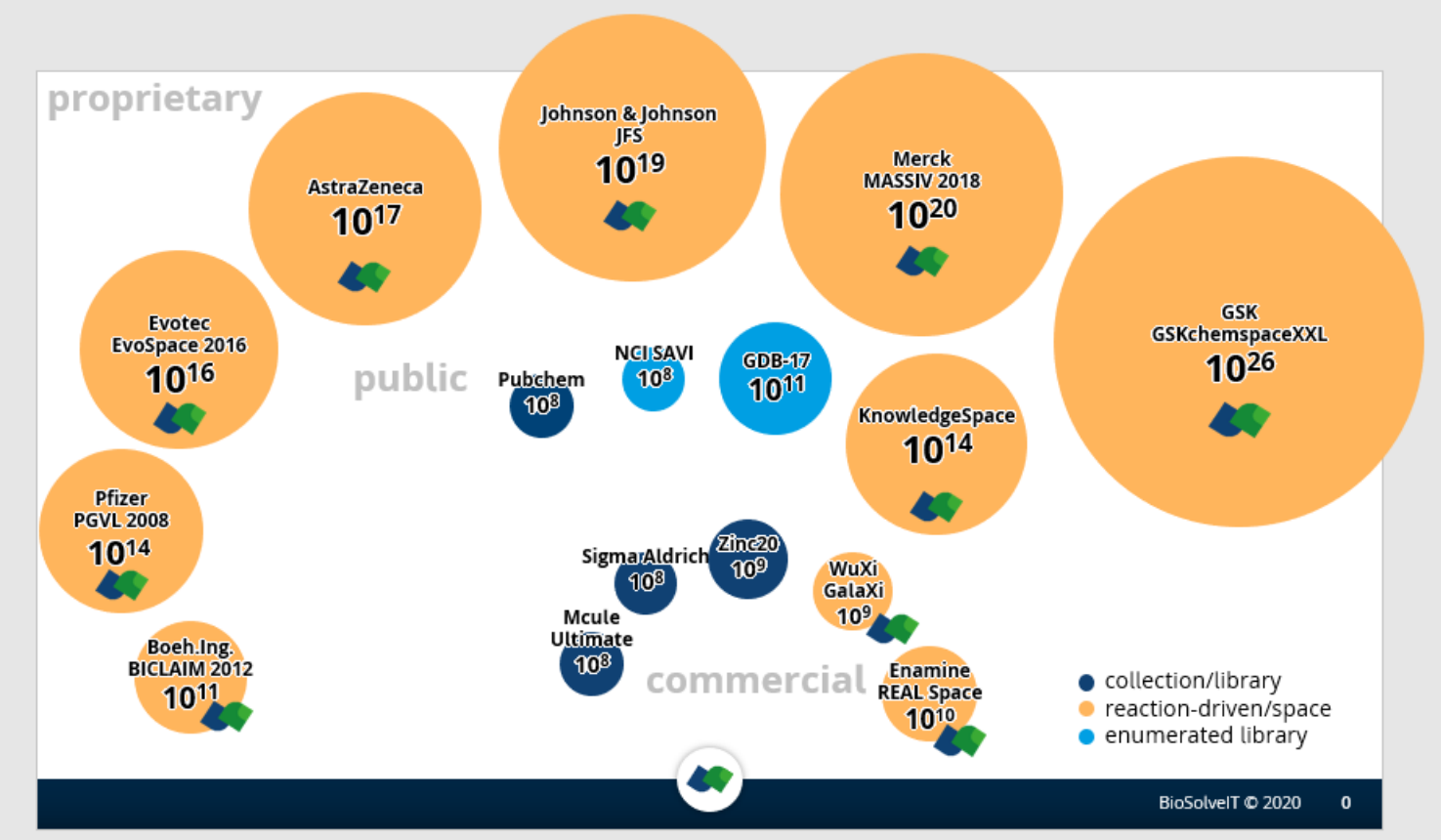
Acknowledgements

Christian Lemmen
Marcus Gastreich

Lessel, Lemmen, *ACS Med. Chem Lett.*, **2019**, 10, 1504-1510

Introduction

More and more ultra-large chemical spaces erase and they are growing exponentially



Motivation

Two main questions:

1. Is it reasonable to expand existing ultra-large chemical spaces further and further?
2. If you have access to one ultra-large chemical space, is it useful/ necessary to search in additional spaces?

How to compare ultra-large chemical spaces?

Method

Comparison of chemical spaces is commonplace
(exact matches, range of physicochemical properties)
Only suitable for enumerated structures

Obvious solution:
compare enumerated random subsets
gives at best rough estimate of any overlap due to the vast size

Our solution:
compare search results in the spaces for a panel of query compounds
Idea behind this:
Any overlap of vast chemical spaces – if present - should be detected when comparing molecules similar to particular queries.

Fragment spaces

Three chemistry spaces used in this study:

- **BICLAIM (BI)** with a size comparable to that of the Knowledge Space: thousands of combinatorial libraries with different scaffolds and variable R-groups
- **Real Space (Enamine)** with ~10^{10} compounds: reliable reactions and validated in-stock building blocks
- **Knowledge Space (BioSolveIT)** with ~10^{12} compounds: literature reactions and commercially available fragment-size reagents

❖ **ZINC15 collection** with ~10^{7} compounds: vast collection of commercially available compounds from various sources

(Sterling, T.; Irwin, J. J. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337)

Procedure

Random selection of 100 query molecules from known drugs with the following properties:

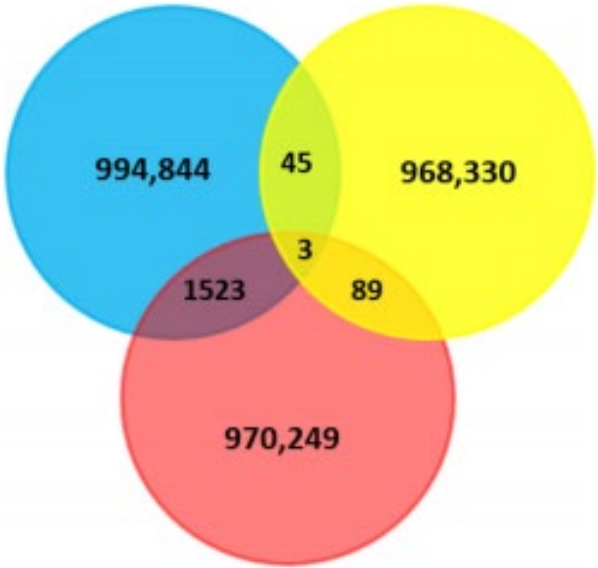
- number of violations of Lipinski's rules < 2
- molecular weight < 600 Da
- clogP < 6
- total polar surface area < 150 $Å^2$
- number of rotatable bonds < 12
- number of H-bond donators and acceptors > 0

Determine 10,000 nearest neighbours for each of the query molecules in each of the three fragment spaces using **FTrees FS**

Overlap

10,000 hits for 100 queries => 1,000,000 compounds from each space
(numbers slightly reduced as some of the hits showed up for more than one query)

Exact match:



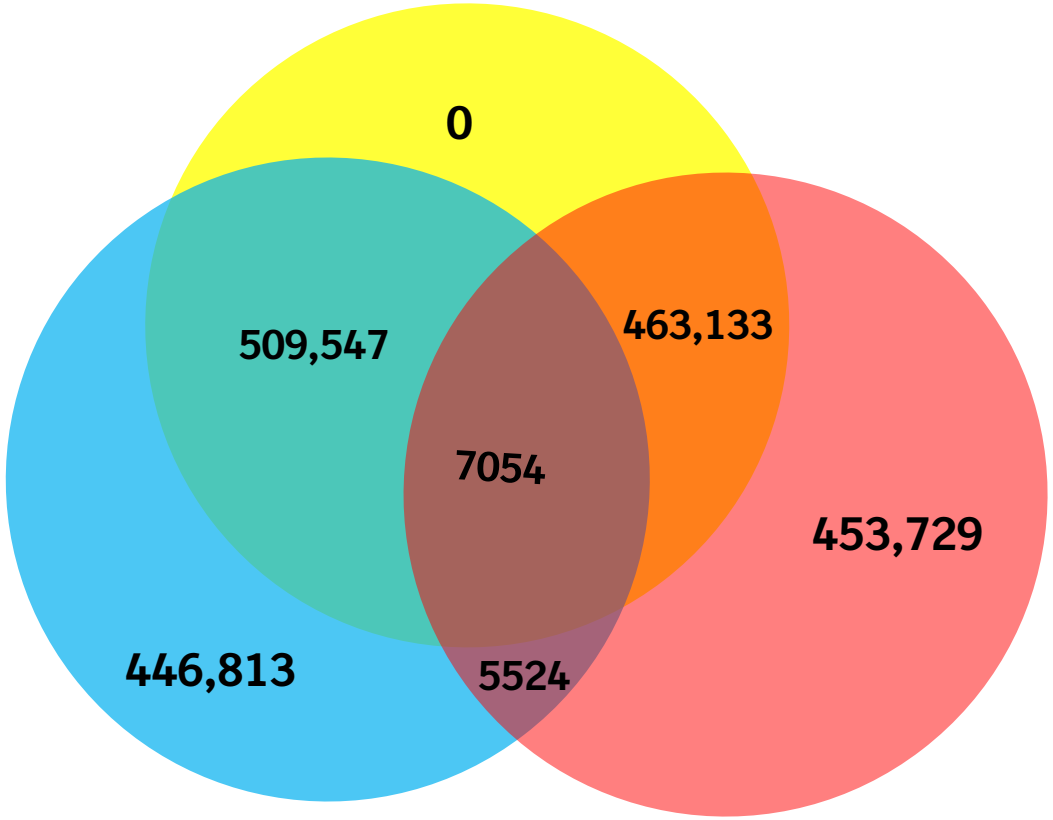
- Overlap surprisingly low
  (< 0.2% in at least 2 spaces)
- Only 3 compounds retrieved from all three spaces (1 query)
- 49 out of 100 queries do not show any overlap among their hits.
- For the other 51 queries on average 32 of 10,000 hits were retrieved from two different spaces.

Validation: Apply method to fragment spaces with known overlap

Knowledge Space built from 118 reactions encoding 10^{12} compounds

Split 1: 58 reactions
Split 2: 60 reactions
Encoding ~550 * 10^9 compounds each

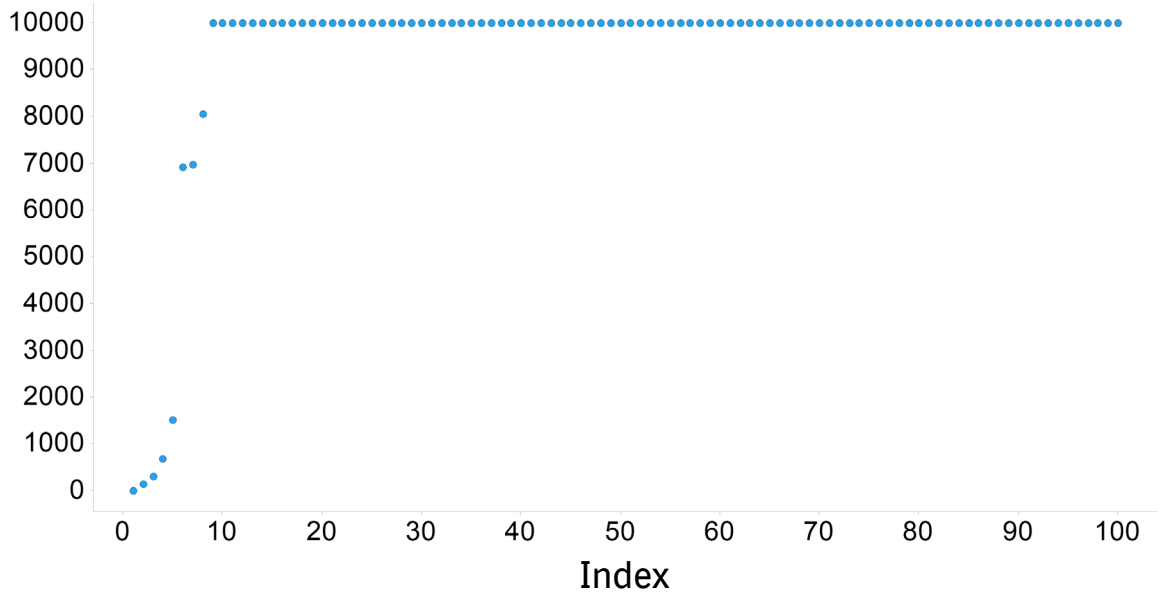
*As expected hits from Split 1 as well as those from Split 2 show an **overlap of ~50%** with the hits of the Knowledge Space*



Coverage of relevant pharmacophore space

Goal from our experience in Virtual Screening: FTrees similarity > 0.9

Determine number of hits with FTrees similarity > 0.9 for each of the queries in the hit sets of the three chemical spaces

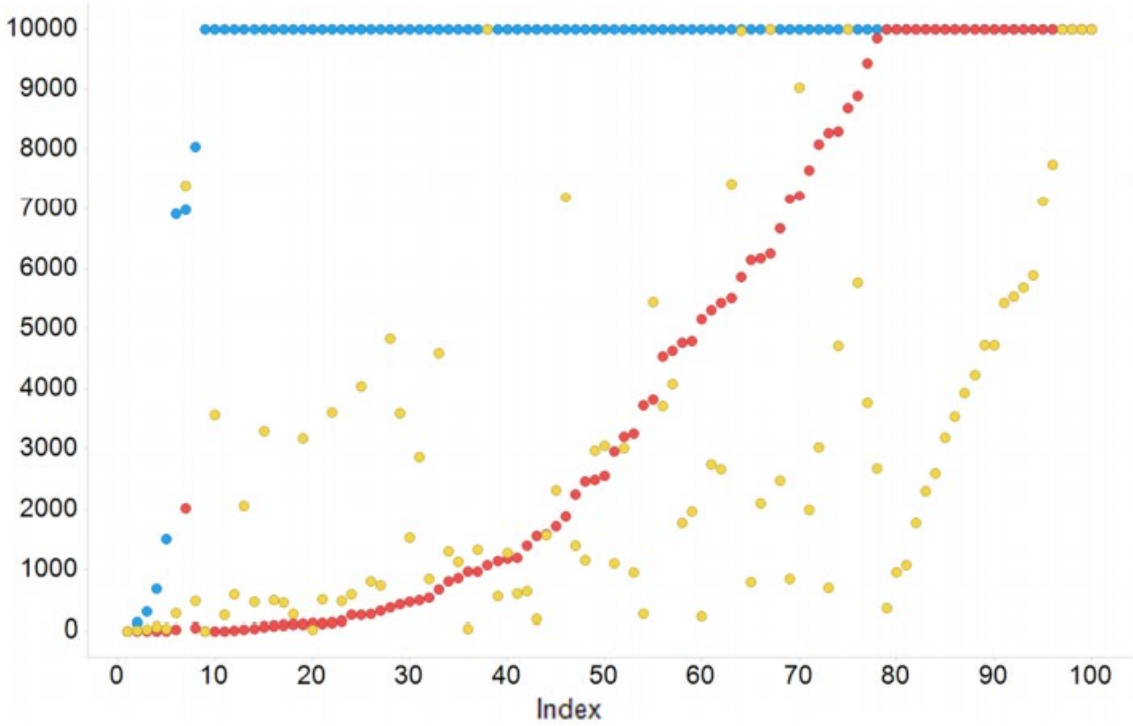


Complementarity

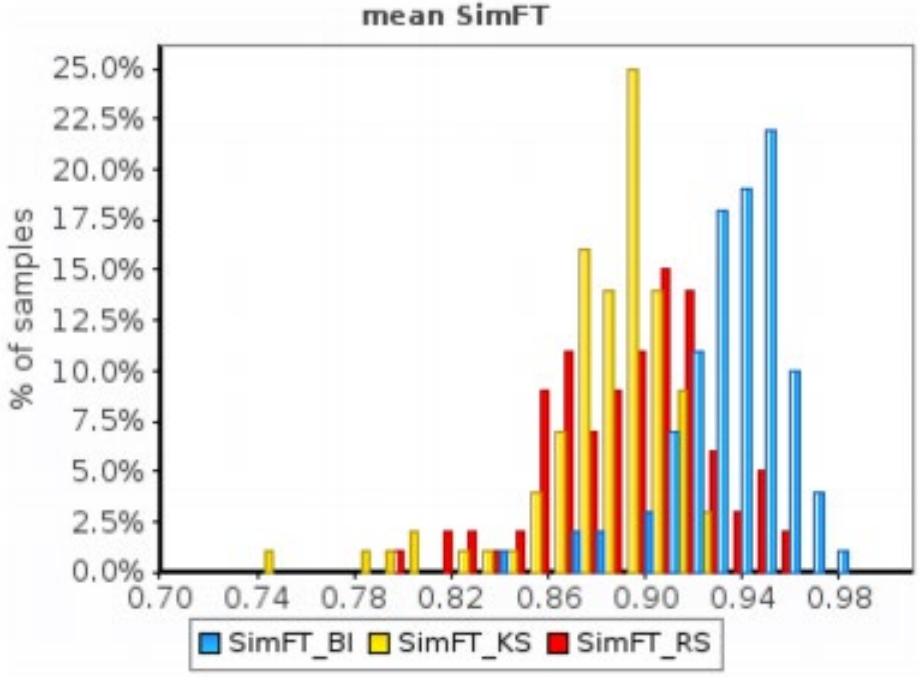
The number of hits with FTrees similarity > 0.9 for each of the random queries varies significantly from query to query and from space to space.

Valuable hits can usually be found from all three spaces.

Number of hits with FTrees similarity > 0.9:

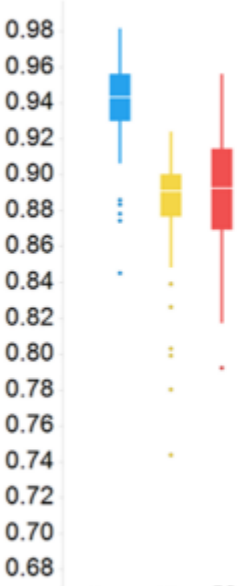


Coverage: average FTrees similarity of the hits to each of the queries



- Average FTrees similarities clearly higher for the BICLAIM hits
- KnowledgeSpace and REAL Space show comparable results

Coverage: broadening the statement



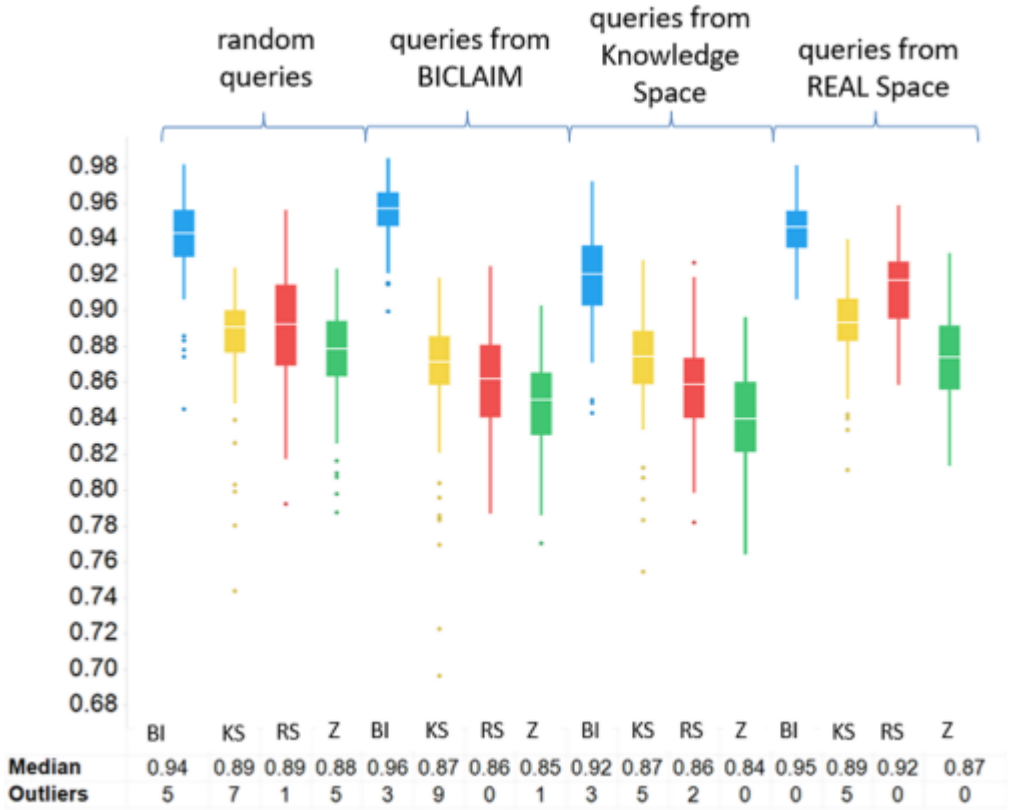
	BI	KS	RS
Median	0.94	0.89	0.89
Outliers	5	7	1

1. Added FTrees similarity searches in ZINC15 collection

2. Repeated study three times with 100 random queries from each of the spaces

Boehringer Ingelheim

Coverage: broadening the statement

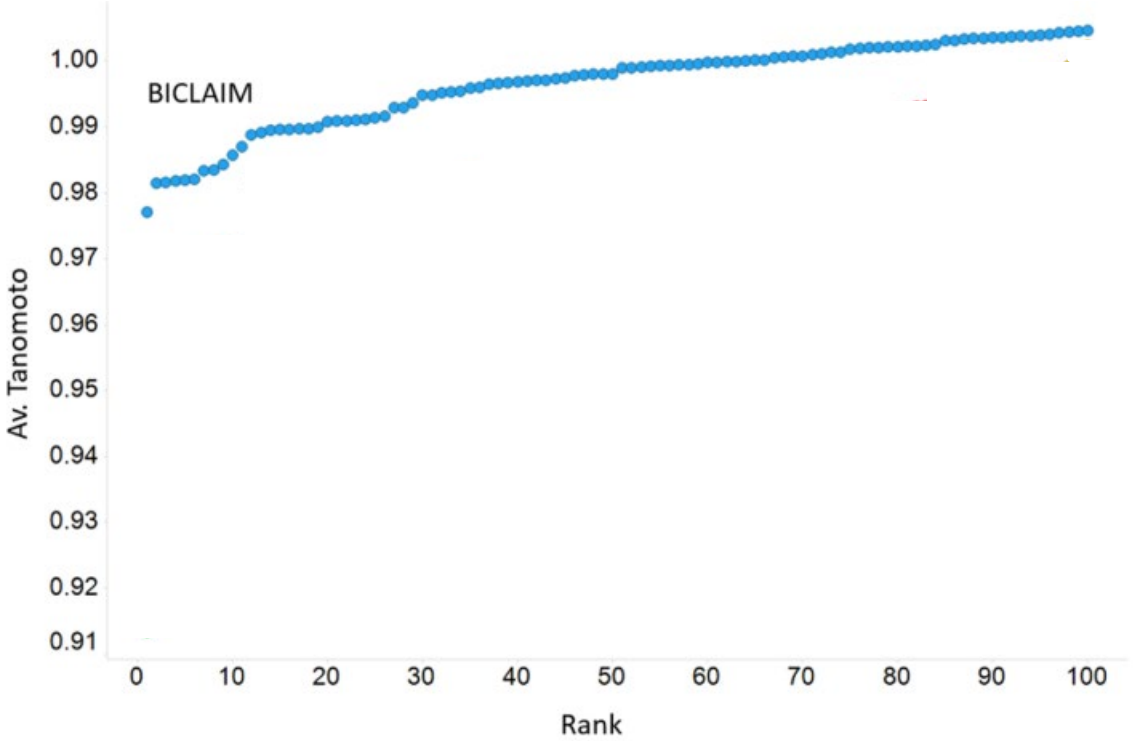


	BI	KS	RS	Z	BI	KS	RS	Z	BI	KS	RS	Z	BI	KS	RS	Z
Median	0.94	0.89	0.89	0.88	0.96	0.87	0.86	0.85	0.92	0.87	0.86	0.84	0.95	0.89	0.92	0.87
Outliers	5	7	1	5	3	9	0	1	3	5	2	0	0	5	0	0

- Hits detected in BICLAIM show higher average FTrees similarities
- Hit sets from ZINC database consistently have a lower average FTrees similarity

- Trend to slightly higher average FTrees similarities of hit sets for queries coming from the search space

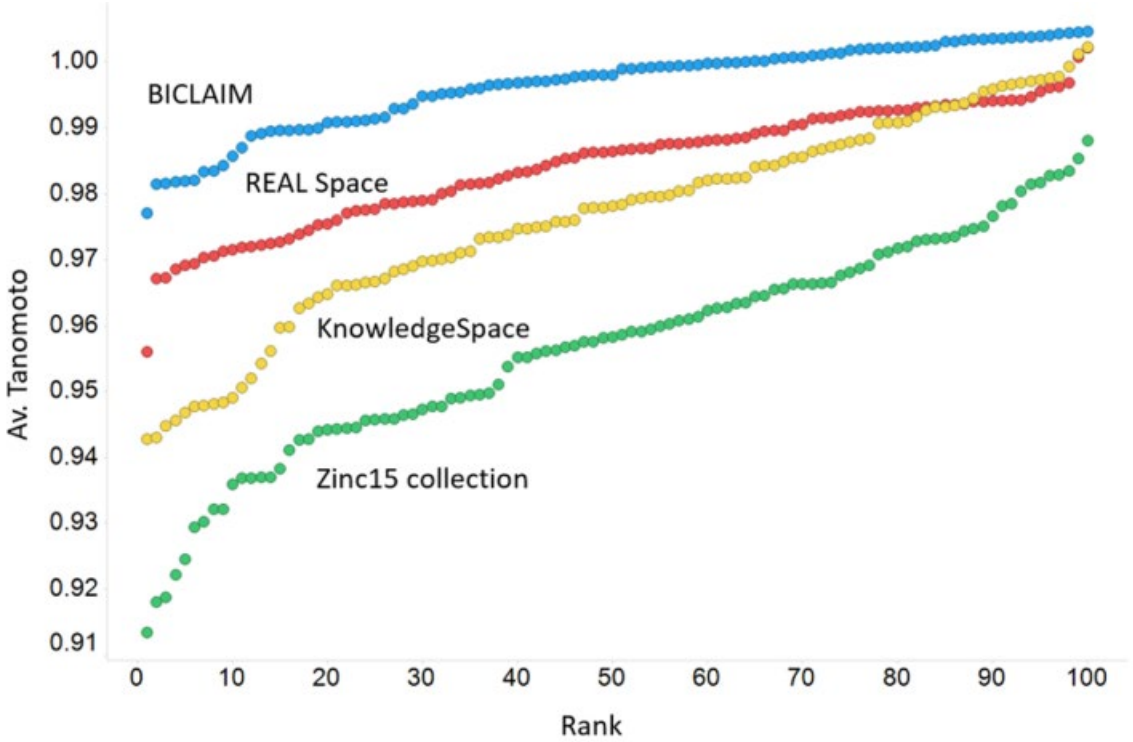
Density: structural similarity of compounds within a hit set (MDL keys*)



- For each member of a hit set determine its nearest neighbor within this hit set using MDL keys
- Determine the average Tanimoto similarity to the nearest neighbors for each hit set
- Sort the queries according to ascending average Tanimoto similarity and plot them

*: Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *J. Chem. Inf. Comput. Sci.* **2002**, *42 (6)*, 1273–1280

Density: results

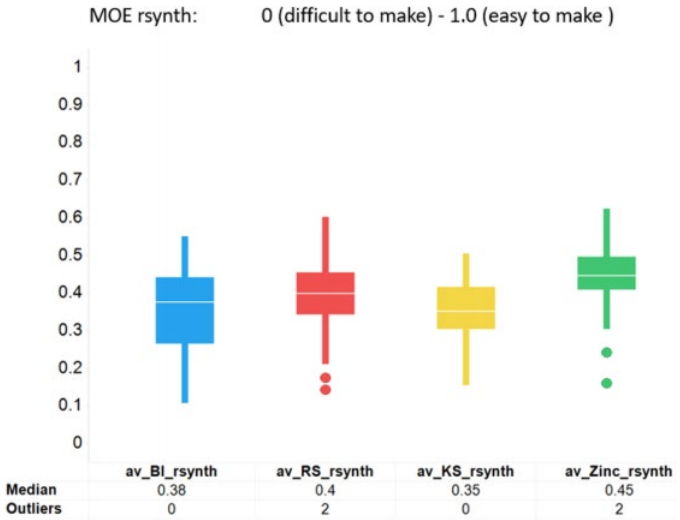


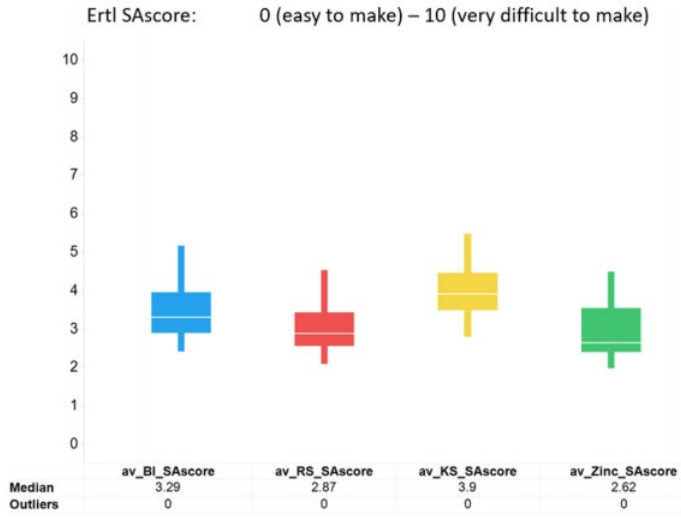
BICLAIM is the most densely occupied space
probably due to the special scaffold based set up of the fragment space

Number of unique scaffolds in BICLAIM is significantly higher than the number of reactions in the other spaces.

Chemical Feasibility

- Chemical Computing Group ULC. Molecular Operating Environment (MOE). Version 2018.01. Chemical Computing Group ULC: Montreal, QC, Canada 2018.
- Ertl, P.; Schuffenhauer,A., *J. Cheminform.* **2009**, *1 (1)*, 8–18.





- Hits from the chemistry spaces show only slightly worse predicted chemical feasibility compared to hits from ZINC15
- Narrow range and only small differences
- Similar trend REAL space > BICLAIM > Knowledge Space according to set up (optimized for chemical feasibility – different levels of chemical feasibility up to ideas – generalized literature reactions)

Summary

- Comparison of ultra-large chemical spaces can be achieved by analyzing the results of similarity searches with a panel of query molecules
- This way the comparison is focused on parts of the spaces, where they might overlap.
- In this study the overlap of the three analyzed spaces is extremely low.
- The coverage of the drug-like portion of the chemical universe – illustrated by the similarity distribution of the most similar hits – is generally quite high for the three spaces.
- The predicted chemical feasibility is reasonably high compared to values obtained for existing compounds.
- The slight differences between the spaces can be explained by their design and setup.

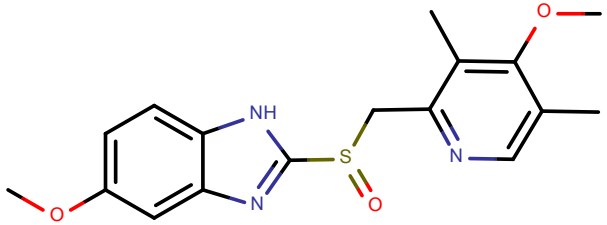
Conclusions

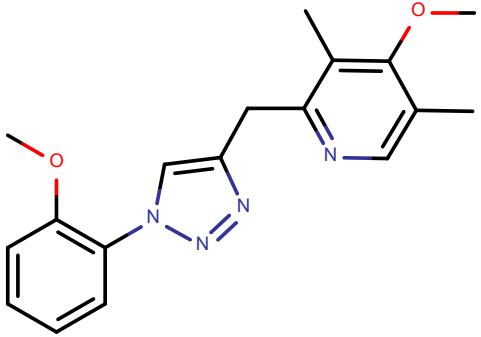
- Ultra-large virtual chemistry spaces contain a lot of valuable starting points for drug discovery
- Searches in these spaces have a very high impact in practice
- For the detection of alternative hits and potential leads it is worthwhile not only to extend existing spaces but also to explore different spaces

BACKUP

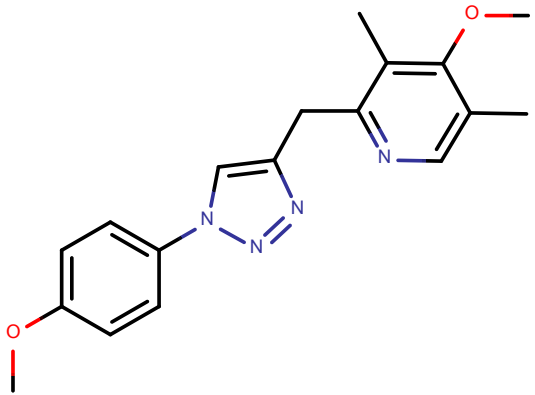
All 3 hits detected in all three spaces

Query: Omeprazole

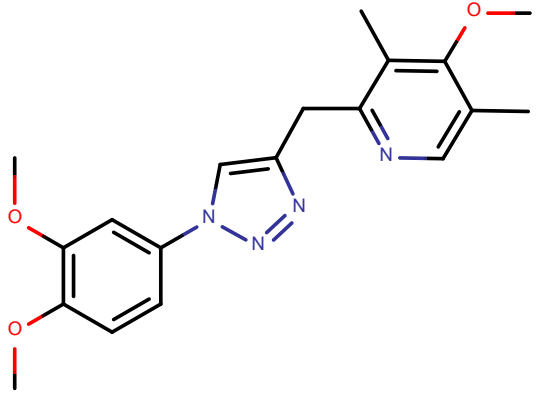




Sim^{FT} = 0.93



Sim^{FT} = 0.93



Sim^{FT} = 0.93