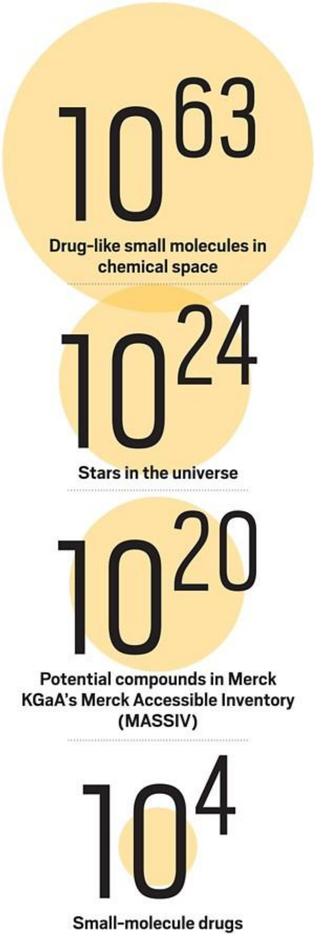
ChemClusterPL

Efficient searching of large chemical databases

Zuzanna Milczarska, Julia Szkóp, Ignacy Makowski, Ryszard Kobiera

Introduction



The need for rapid search for analogues of existing drugs

Project Idea ??

Google for Molecules: Fast, Smart, Scalable

Goal: Enable fast search and grouping of similar molecules in large databases (e.g., ZINC22)

Our proposal:

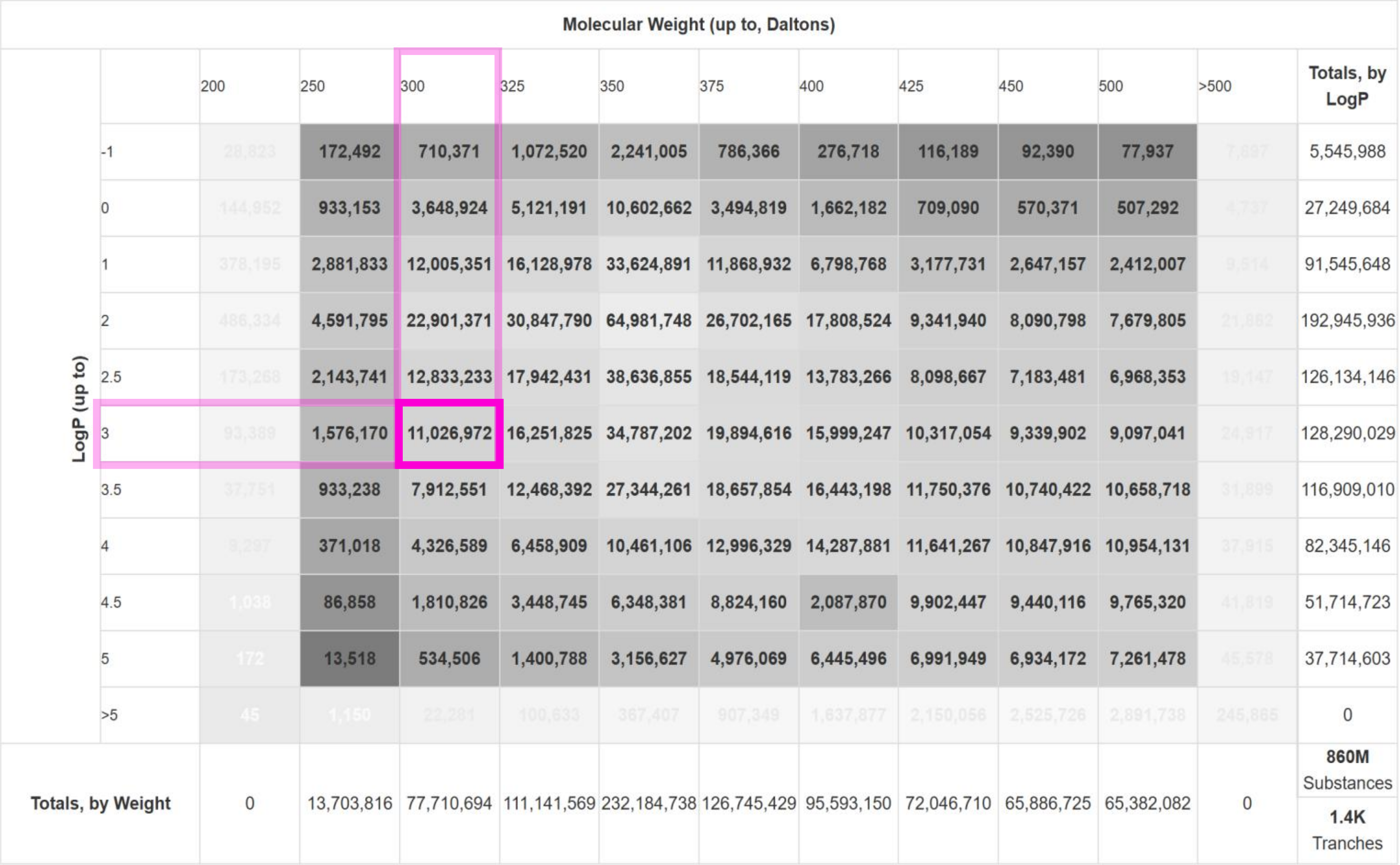
- Scalable similarity search using molecular embeddings
- Fast approximate search
- Clustering at scale

ZINC freely accessible repository of commercially available compounds

ZINC
Data Base

Compounds Selected by:

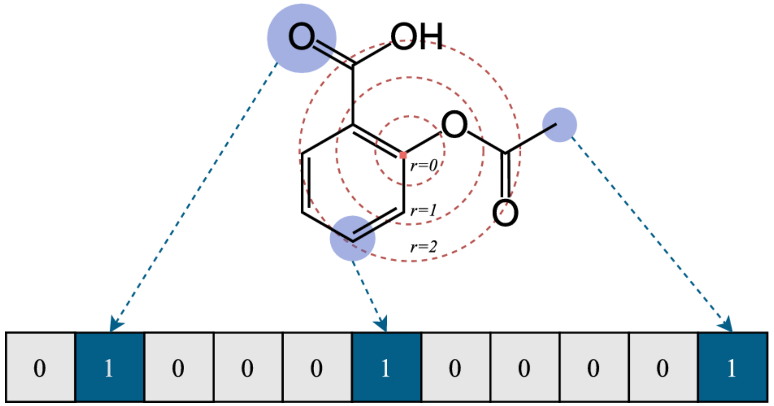
- Availability Type
- Chemical and Pharmacophore Properties
- Structure Format
- Physicochemical Filters



	Molecular Weight (up to, Daltons)											
LogP (up to)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	28,823	172,492	710,371	1,072,520	2,241,005	786,366	276,718	116,189	92,390	77,937	7,697	5,545,988
0	144,952	933,153	3,648,924	5,121,191	10,602,662	3,494,819	1,662,182	709,090	570,371	507,292	4,737	27,249,684
1	378,195	2,881,833	12,005,351	16,128,978	33,624,891	11,868,932	6,798,768	3,177,731	2,647,157	2,412,007	9,514	91,545,648
2	486,334	4,591,795	22,901,371	30,847,790	64,981,748	26,702,165	17,808,524	9,341,940	8,090,798	7,679,805	21,862	192,945,936
2.5	173,268	2,143,741	12,833,233	17,942,431	38,636,855	18,544,119	13,783,266	8,098,667	7,183,481	6,968,353	19,147	126,134,146
3	93,389	1,576,170	11,026,972	16,251,825	34,787,202	19,894,616	15,999,247	10,317,054	9,339,902	9,097,041	24,917	128,290,029
3.5	37,751	933,238	7,912,551	12,468,392	27,344,261	18,657,854	16,443,198	11,750,376	10,740,422	10,658,718	31,899	116,909,010
4	9,297	371,018	4,326,589	6,458,909	10,461,106	12,996,329	14,287,881	11,641,267	10,847,916	10,954,131	37,915	82,345,146
4.5	1,038	86,858	1,810,826	3,448,745	6,348,381	8,824,160	2,087,870	9,902,447	9,440,116	9,765,320	41,819	51,714,723
5	172	13,518	534,506	1,400,788	3,156,627	4,976,069	6,445,496	6,991,949	6,934,172	7,261,478	45,578	37,714,603
>5	45	1,150	22,281	100,633	367,407	907,349	1,637,877	2,150,056	2,525,726	2,891,738	245,865	0
Totals, by Weight	0	13,703,816	77,710,694	111,141,569	232,184,738	126,745,429	95,593,150	72,046,710	65,886,725	65,382,082	0	**860M** Substances **1.4K** Tranches

Measures of similarity between chemical compounds

- Fingerprints (ECFP - 1024)

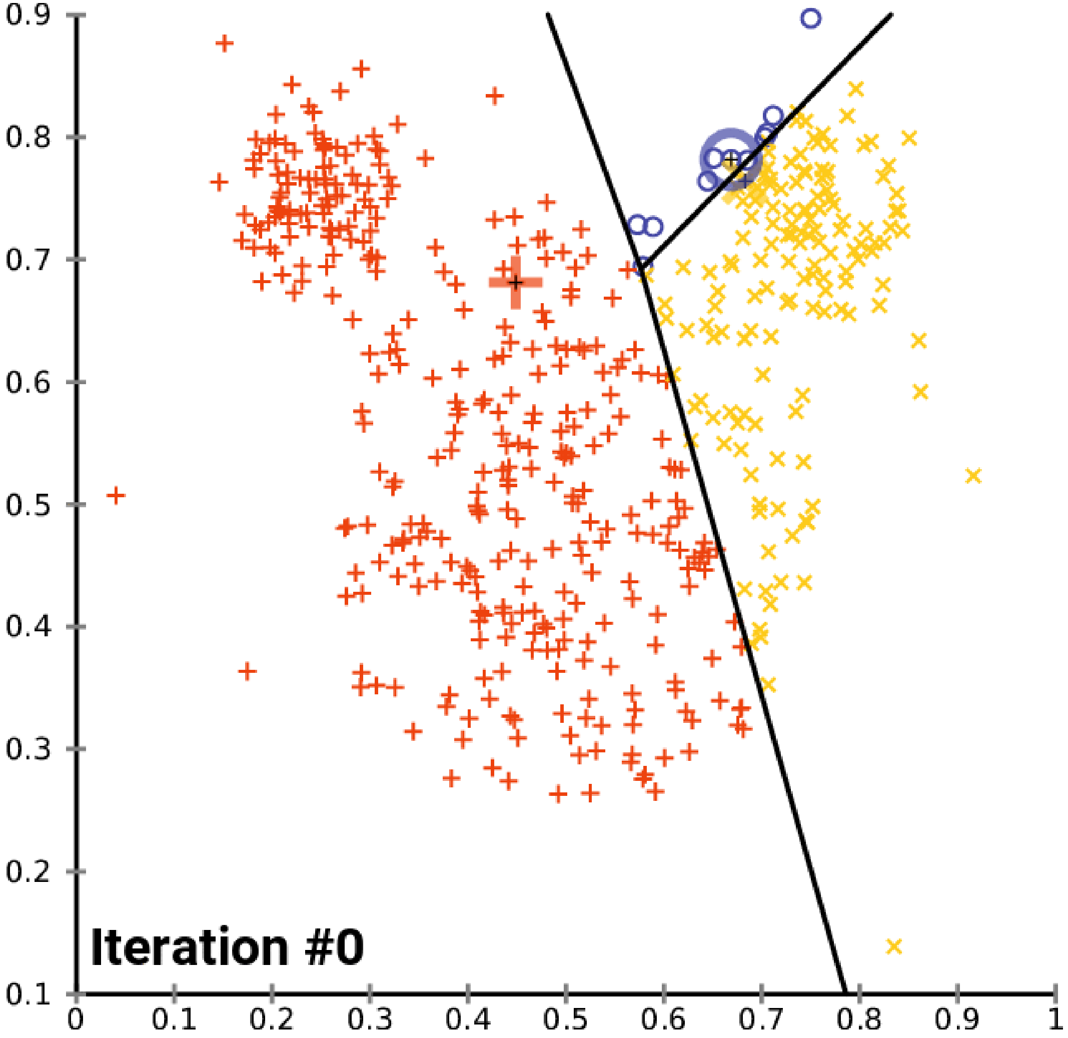


- Tanimoto / Cosine distance

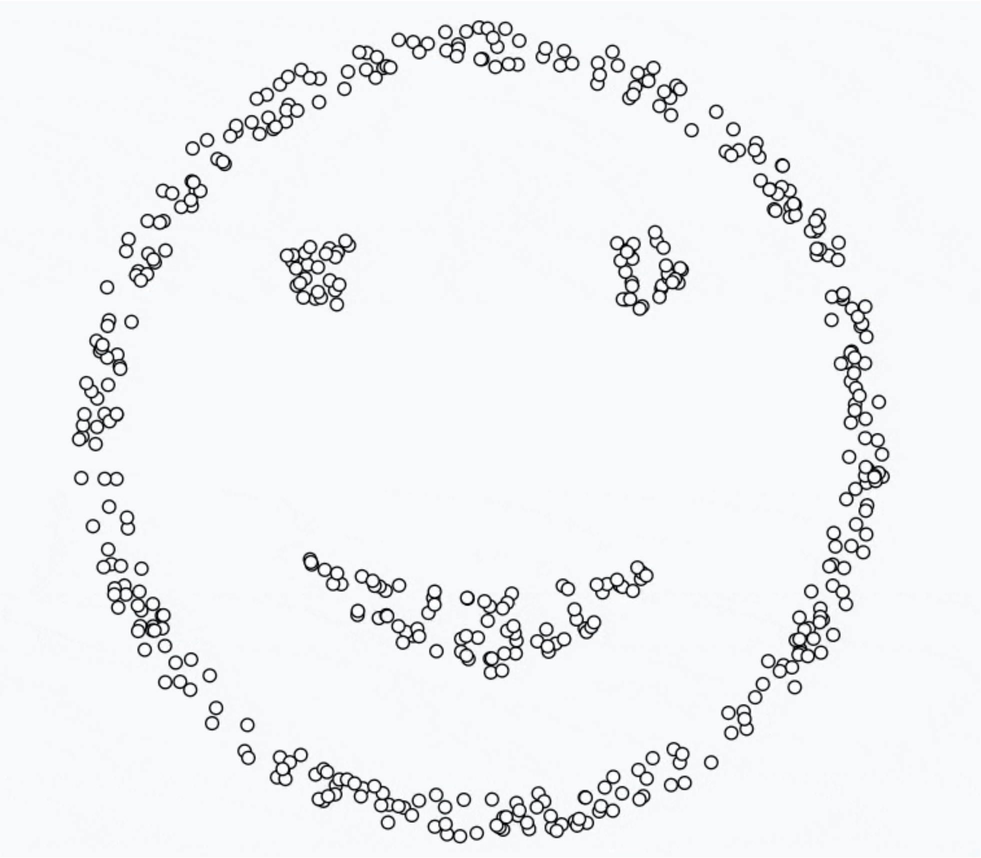
$$T(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

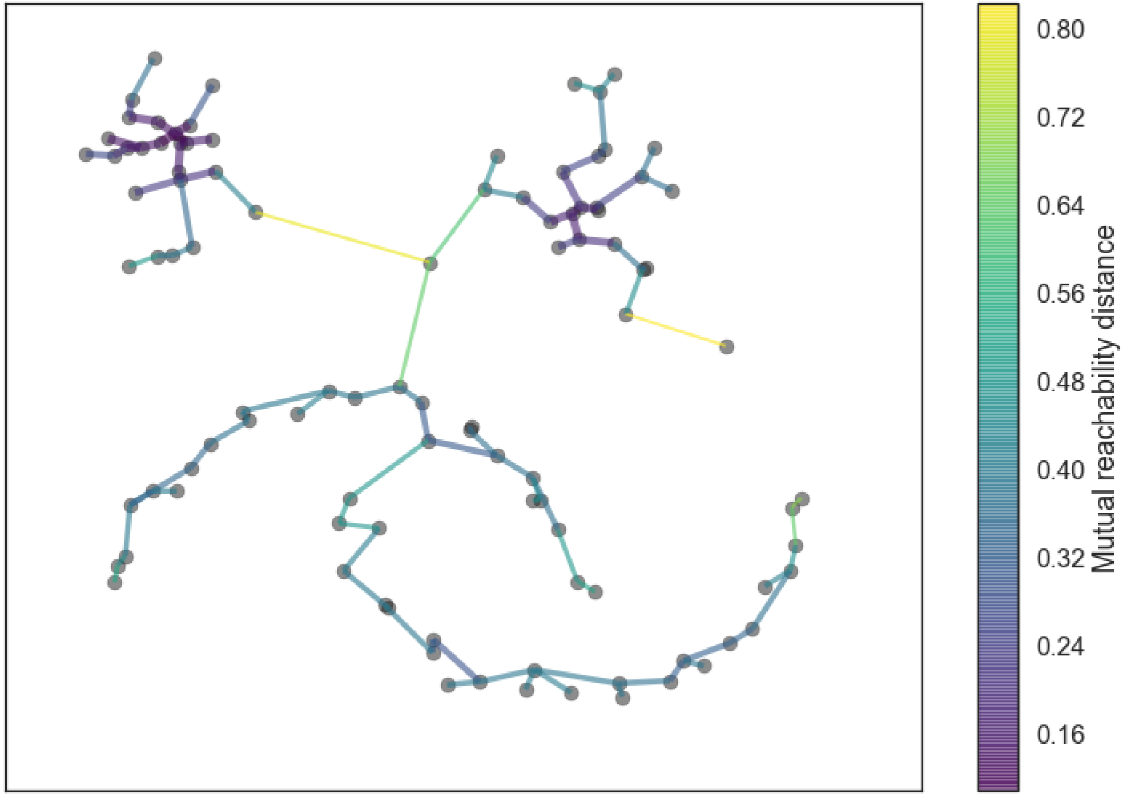
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Clustering methods – K-medoids

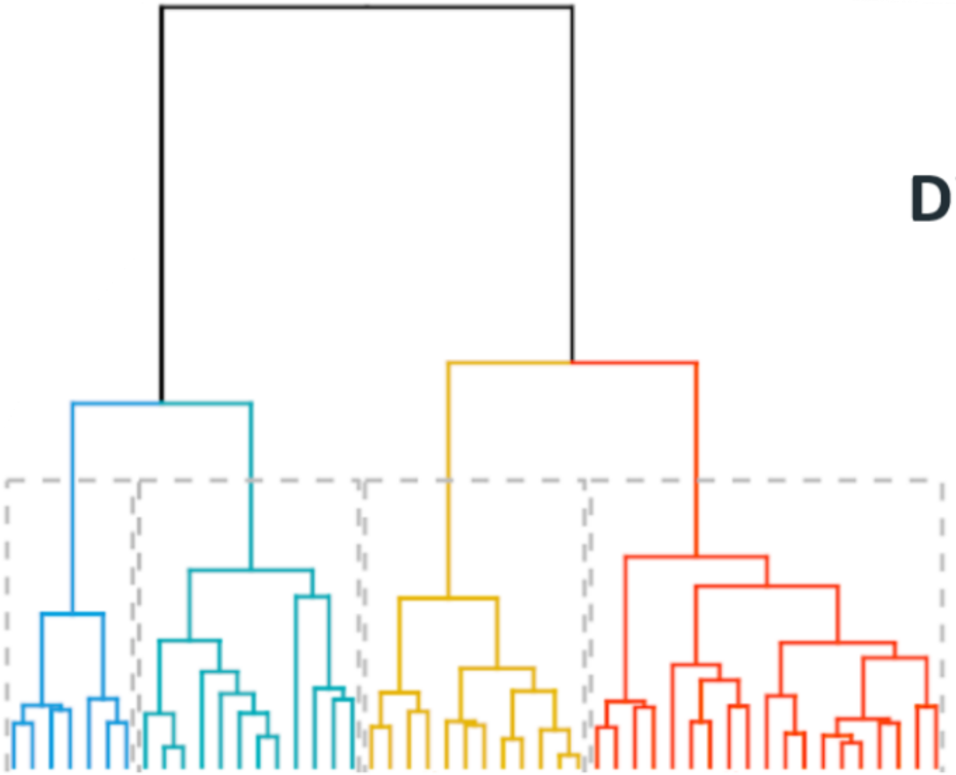


Clustering methods - HDBSCAN



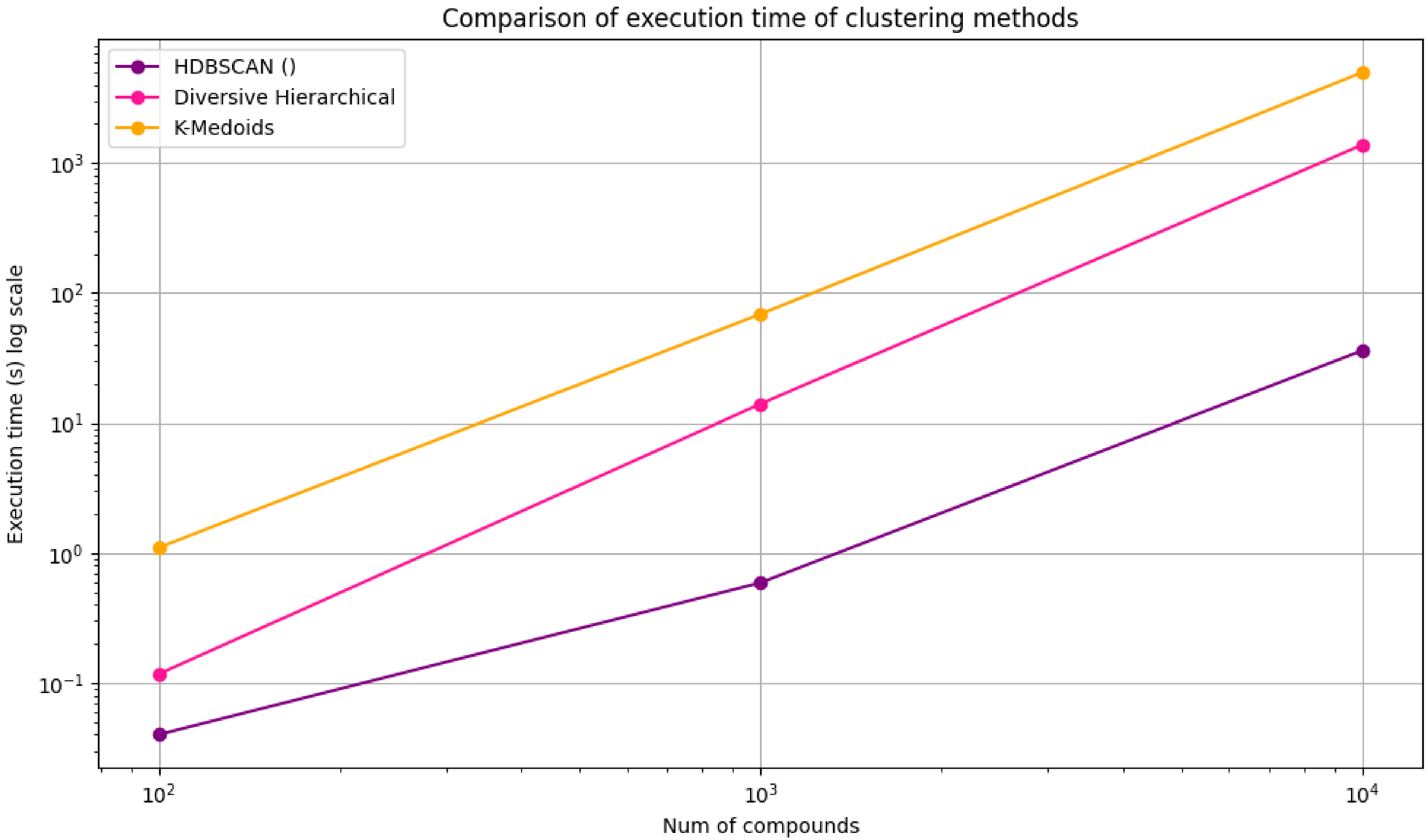


Clustering methods – Divisine hierarchical

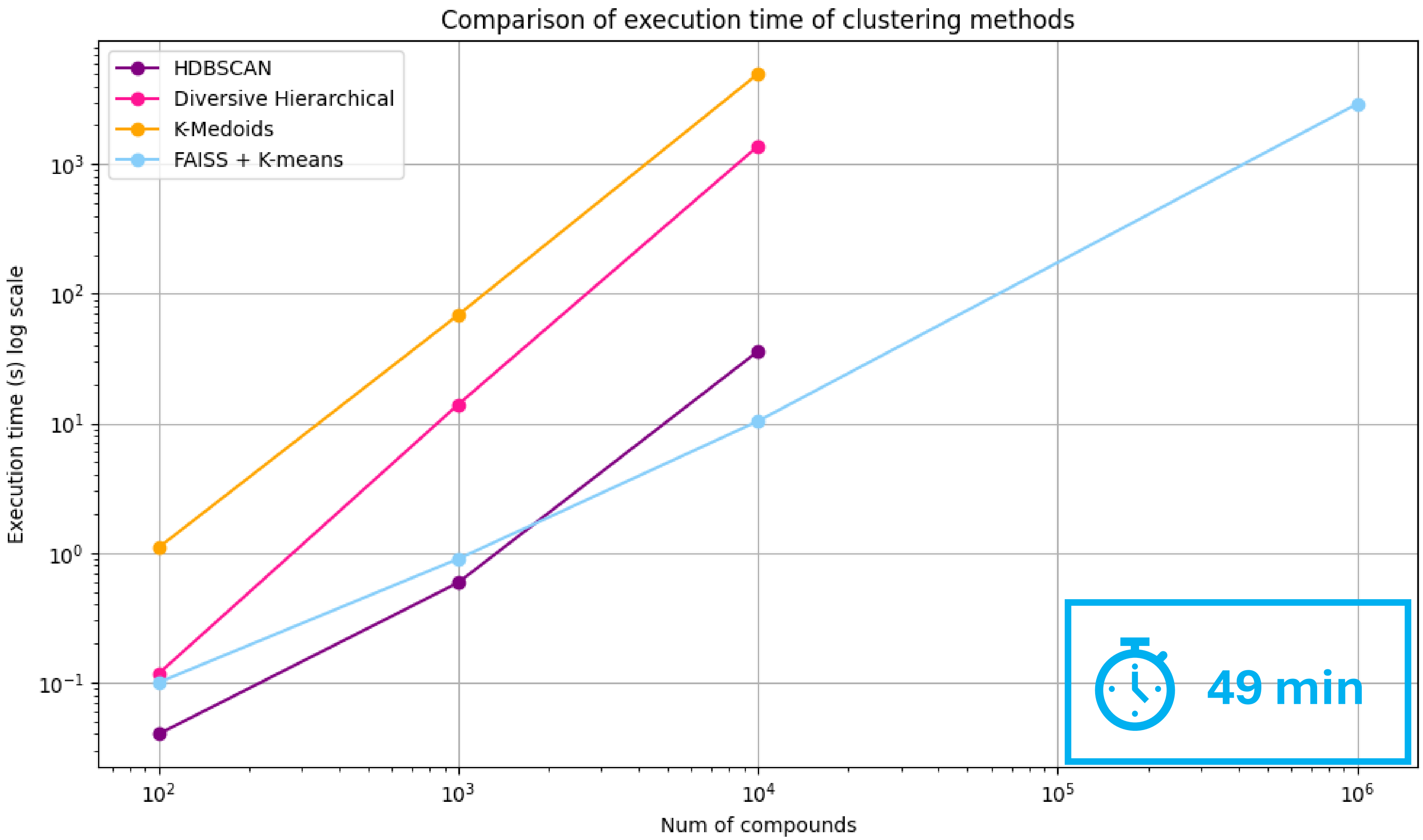


Divisive

Complexity and performance analysis

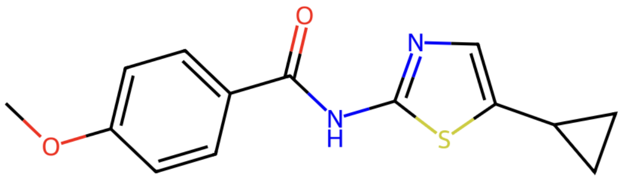


Implementation with FAISS

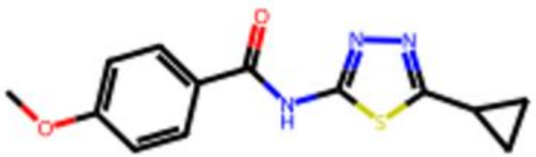


49 min

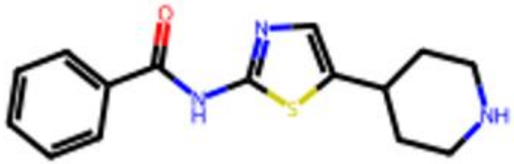
Searching in entire database



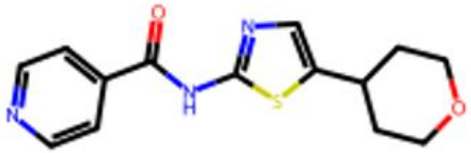
COc1ccc(cc1)C(=O)Nc2ncc(s2)C3CC3



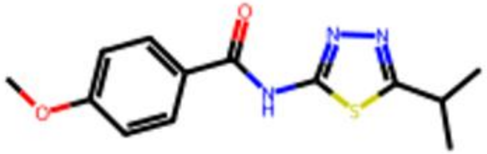
T:0.66 C:0.79



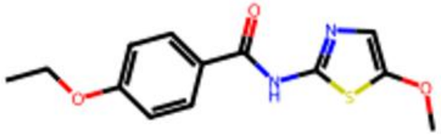
T:0.59 C:0.74



T:0.58 C:0.73



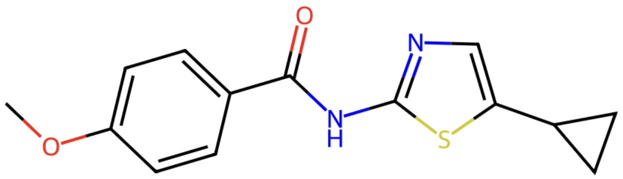
T:0.56 C:0.72



T:0.56 C:0.72

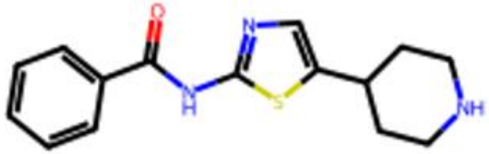
7m 23s

Searching in ChemClusterPL

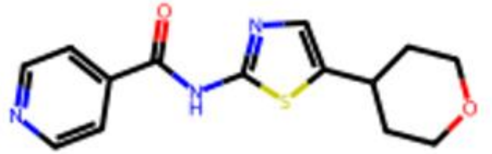


COc1ccc(cc1)C(=O)Nc2ncc(s2)C3CC3

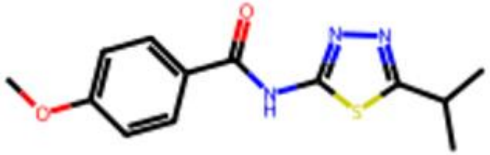
5,424 s



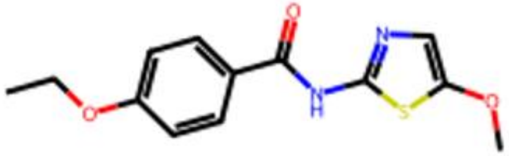
Tanimoto: 0.59 Cosine: 0.74



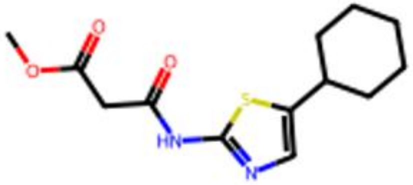
Tanimoto: 0.58 Cosine: 0.73



Tanimoto: 0.56 Cosine: 0.72



Tanimoto: 0.56 Cosine: 0.72



Tanimoto: 0.54 Cosine: 0.70

Balancing cluster sizes

Method	B2 Index
TanimotoNN	0,0010
HDBSCAN	0,0024
FAISS_KMeans	0,1522
SKLearn_KMeans	0,1226
FAISS_Hierarchical	0,0536

Conclusions

- Achieved **high accuracy** in similarity grouping
- Reached **very fast runtimes**, even on large datasets
- Successfully **scaled clustering to databases** that are infeasible using traditional methods (e.g., ZINC22-scale)

Further work

- Deploy solution on **GPUs for even faster performance**
- Extend to **larger datasets** (tens of billions of molecules)
- Integrate with existing **drug discovery pipelines**

Bibliography

https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

https://en.wikipedia.org/wiki/K-means_clustering#:~:text=k%2Dmeans%20clustering%20is%20a,a%20prototype%20of%20the%20cluster.

https://arxiv.org/abs/1702.08734