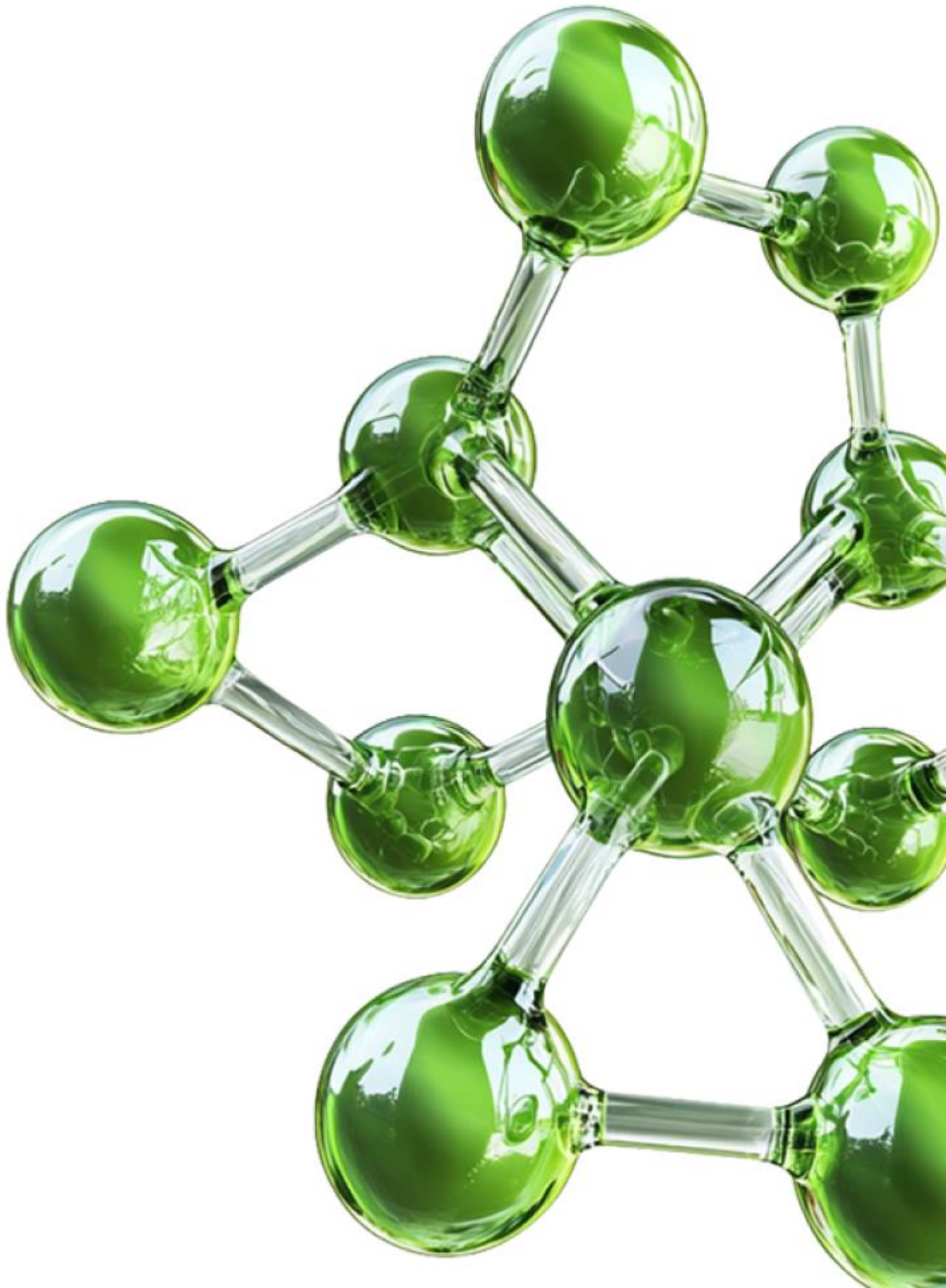
FREEDOM SPACE 4.0 - CREATION OF ULTRA-LARGE CHEMICAL SPACES USING ML-BASED REAGENT FILTERING

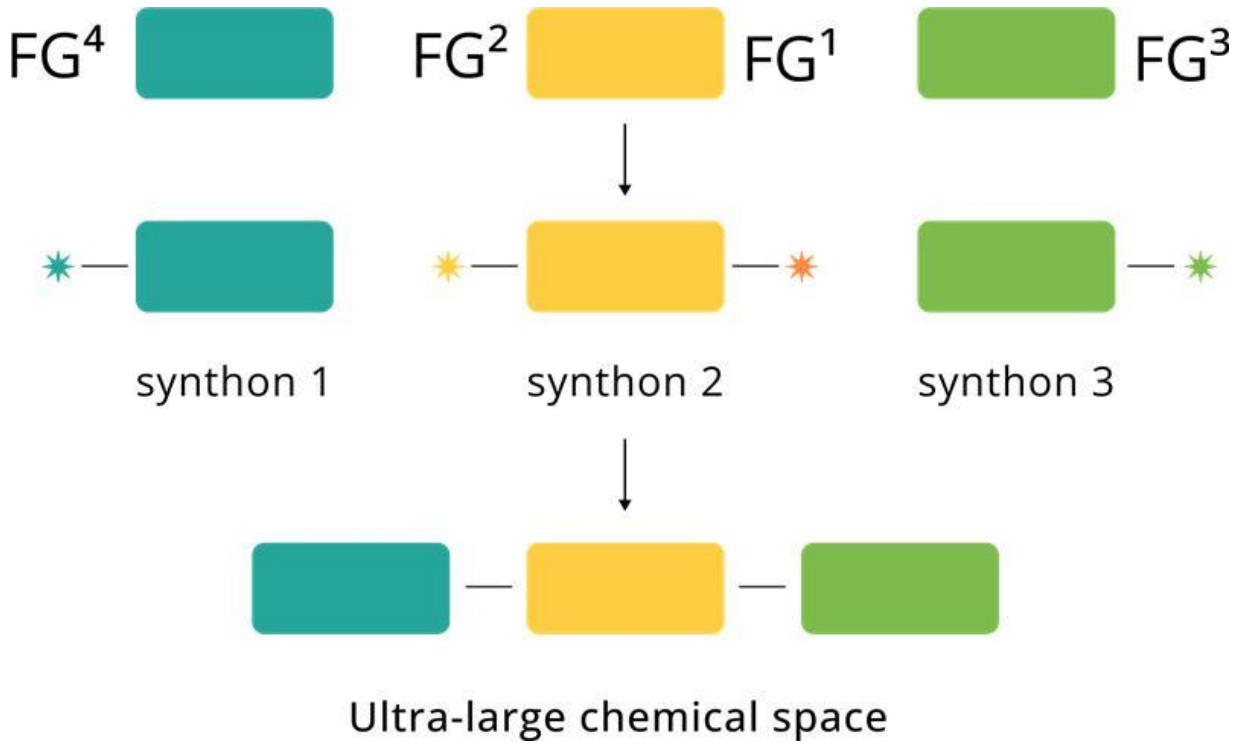
Anna Kapeliukha, Chemspace

a.kapeliukha@chem-space.com



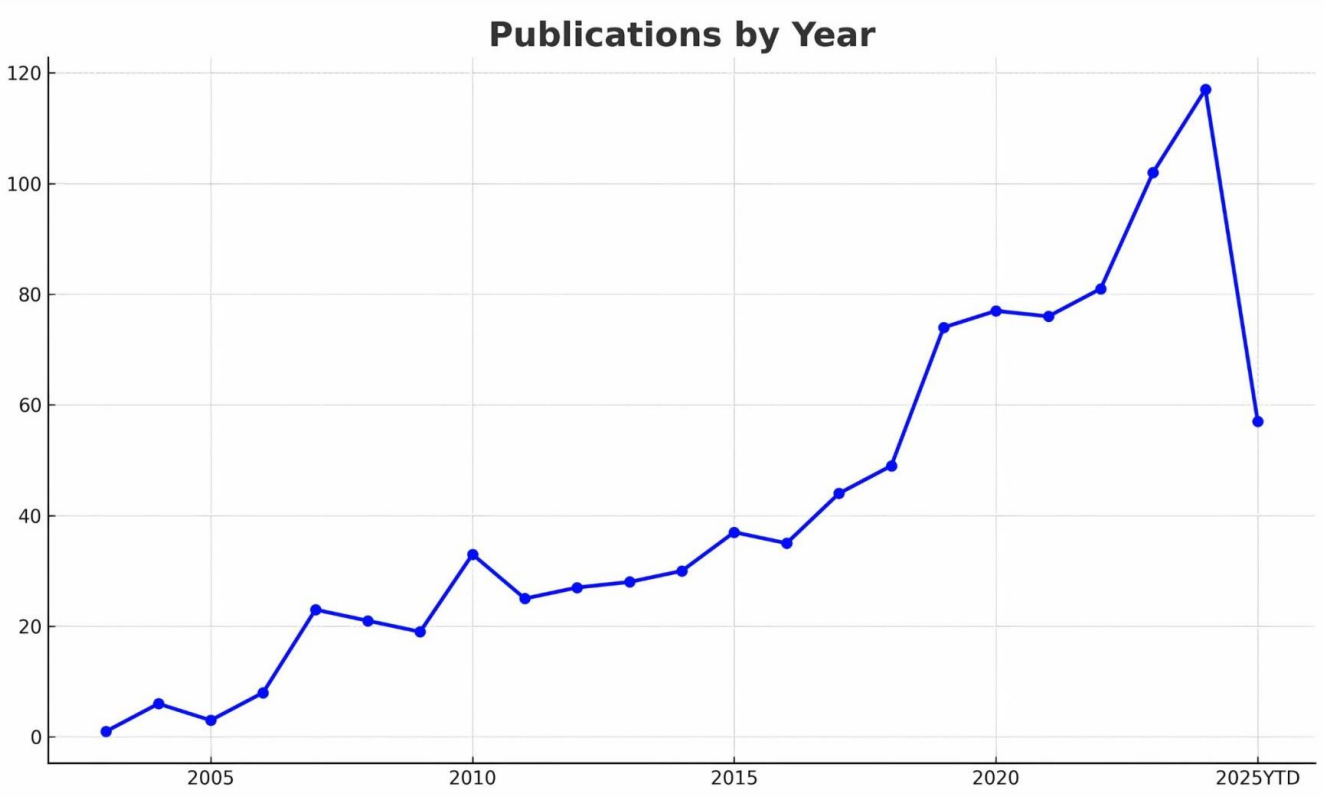

eMolecules eXplore
10^{12}
Enamine REAL Space
10^{10}
Chemspace Freedom
Space 4.0
10^{9}
WiXi Galaxi
10^{9}
Otava CHEMriya
10^{9}

Synthon-based nature of combinatorial chemical spaces



Enamine REAL: The Most Explored Commercially Available Chemical Space

Enamine REAL Compounds - **5 Trillion molecules**



nature

Article | Published: 09 March 2020

This is an unedited manuscript that has been accepted for publication. Nature Research are providing this early version of the manuscript as a service to our customers. The manuscript will undergo copyediting, typesetting and a proof review before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

An open-source drug discovery platform enables ultra-large virtual screens

Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D. Fischer, Paul Coote, Krishna M. Padmanabha Das, Yehor S. Malets, Dmytro S. Radchenko, Yurii S. Moroz, David A. Scott, Konstantin Fackeldey, Moritz Hoffmann, Iryna Iavniuk, Gerhard Wagner & Haribabu Arthanari

Nature (2020) | Cite this article

nature

Article | Published: 10 February 2020

This is an unedited manuscript that has been accepted for publication. Nature Research are providing this early version of the manuscript as a service to our customers. The manuscript will undergo copyediting, typesetting and a proof review before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

Virtual discovery of melatonin receptor ligands to modulate circadian rhythms

Reed M. Stein, Hye Jin Kang, John D. McCorvy, Grant C. Glatfelter, Anthony J. Jones, Tao Che, Samuel Slocum, Xi-Ping Huang, Olena Savych, Yurii S. Moroz, Benjamin Stauch, Linda C. Johansson, Vadim Cherezov, Terry Kenakin, John J. Irwin, Brian K. Shoichet, Bryan L. Roth & Margarita L. Dubocovich

Nature (2020) | Cite this article

nature
International journal of science

Article | Published: 06 February 2019

Ultra-large library docking for discovering new chemotypes

Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmache Shoichet, Bryan L. Roth & John J. Irwin

Nature **566**, 224–229 (2019) | Download Citation

Synthesizability Aspects of Combinatorial Spaces

POTENTIAL PROBLEMS

- Low yields
- Side products
- Failed purification
- No product at all

POTENTIAL SOLUTIONS

✅ Experimental validation of reagents

x long and expensive, requires investments to buy reagents
x not applicable to external chemical collections

✅ Computer-assisted synthesis prediction (CASP):

ML-assisted feasibility, yield, and optimal conditions for a chemical reaction prediction.

x limited training data – not optimal performance
x require full enumeration
x large inference time and compute

Synthesizability Aspects of Combinatorial Spaces

POTENTIAL PROBLEMS

- Low yields
- Side products
- Failed purification
- No product at all

OUR SOLUTION – ML-BASED FILTERING OF BBS FOR SPACE GENERATION

- does not require full enumeration
- cheap predictions
- synthon-based space creation
- applicable to commercial compound collections of choice

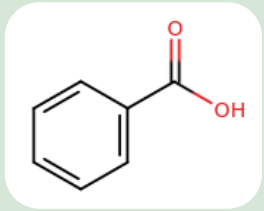
ML-Assisted BBs Filtering For Space Generation

Training data sets:

We use the data obtained from years of operation of Enamine REAL, where they recorded experimental statistics about the building block success rate.

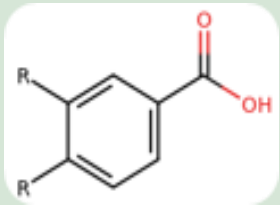
Datasets contain **1,000 – 20,0000** datapoints for each reagent.

EXAMPLE*:



20 reactions total
18 successful
2 failed

GOOD



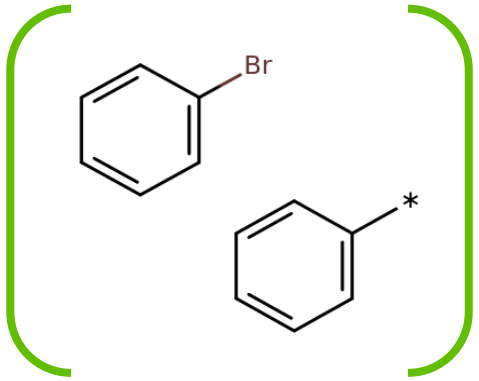
22 reactions total
1 successful
21 failed

BAD

* The presented data is made up for understanding and is not part of the actual training set

ML-Assisted BBs Filtering For Space Generation

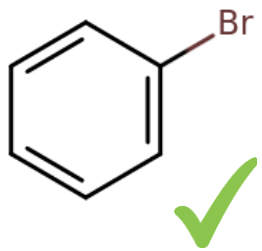
Synthon-building block molecular representation - Accounts for BBs that share the synthon.

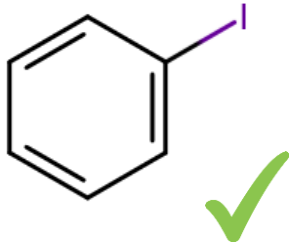


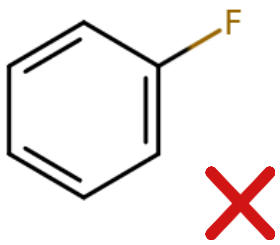
chemprop

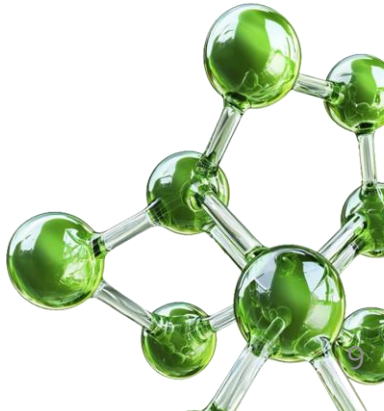
Graph Neural Network

Probability of being suitable for the reaction



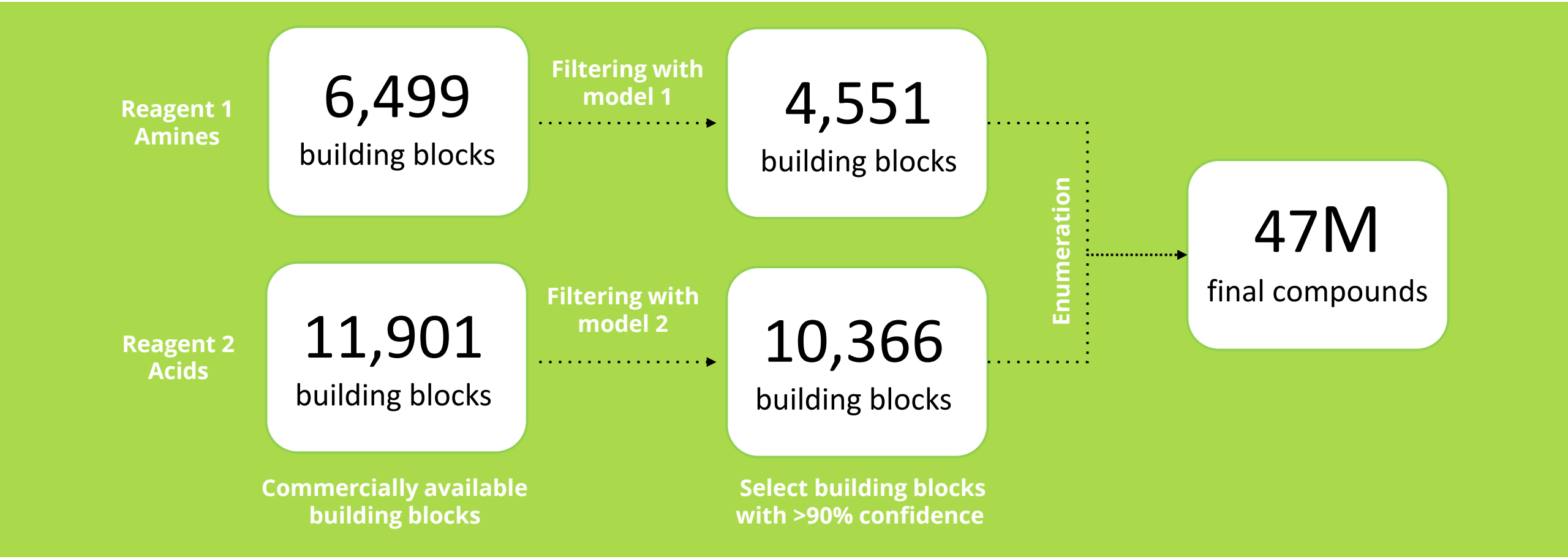






Example of Reaction Processing

Example: Amide Coupling (reaction id: 22)



ML-Assisted Space Generation: General Workflow

Forming the BBs collection

BBs from trusted suppliers with price and lead time cutoffs

Create BB pools for reaction types

Building blocks are passed through a set of SMARTS filters for specific reaction types to ensure presence of functional groups

Filtering BBs with ML models

Each building block set is filtered using in-house machine learning models trained on Enamine REAL reaction success statistics

Enumeration

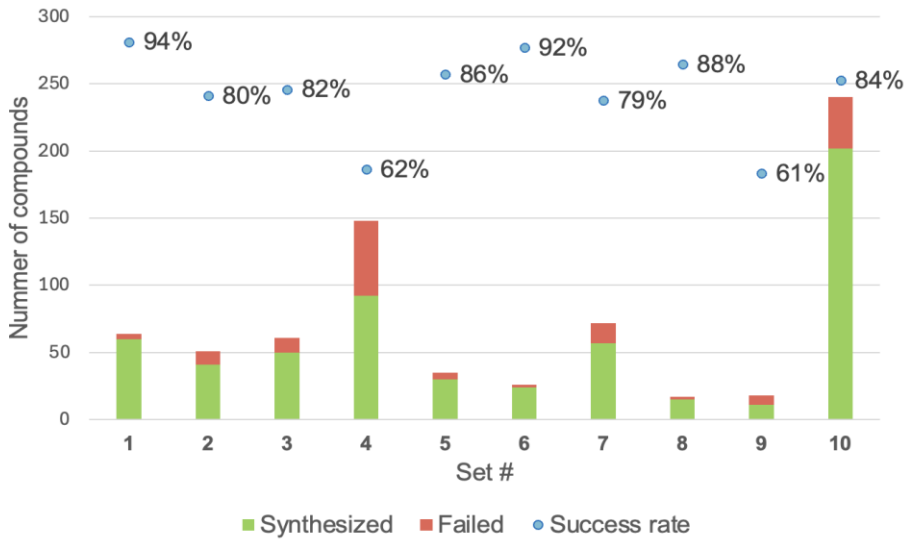
40
reaction protocols

2 and 3 component reactions

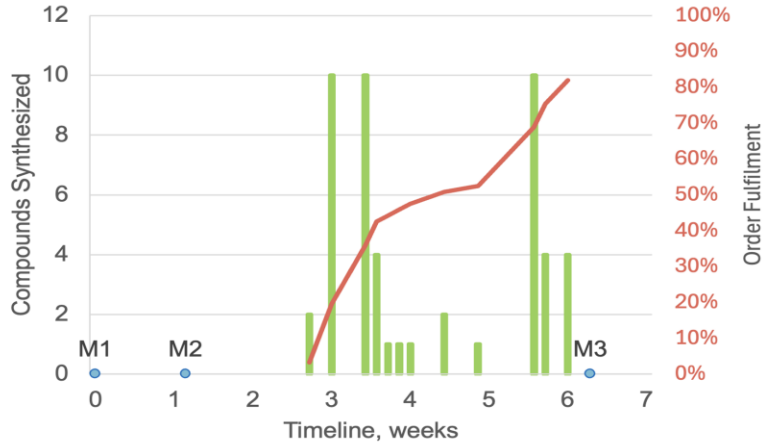
Examples:
- Amide formation
- Condensation
- Arylation
- Reductive amination
- Urea formation
- Suzuki coupling

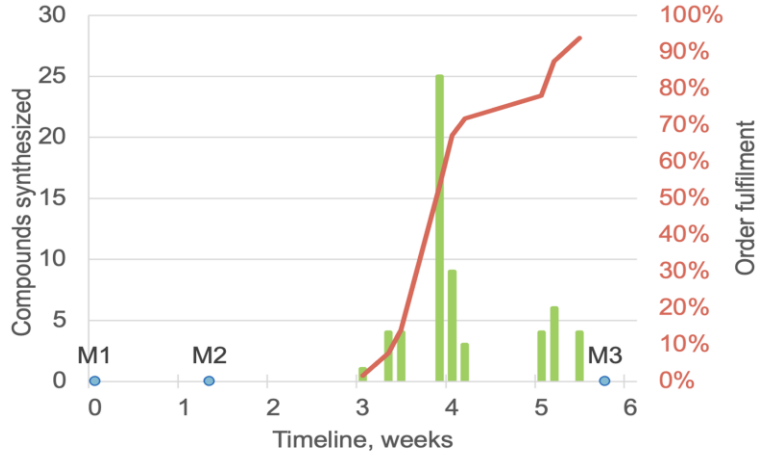
Full list with reaction schemes available here

ML-Assisted Space Generation Success Rate – Freedom Space 3.0



Examples of orders timelines





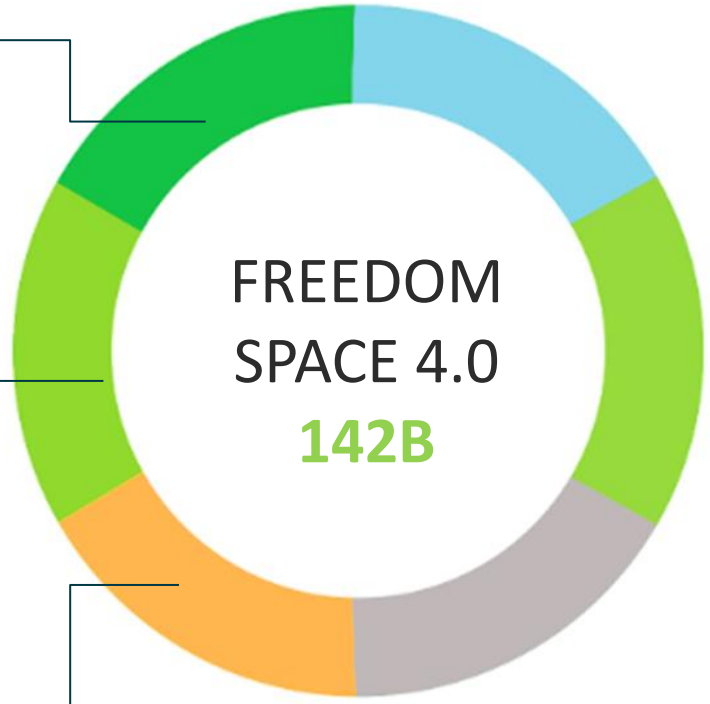
- 5B molecules, little overlap with Enamine REAL
- 10 chemical transformations (7 ML-, 3 human-filtered)
- Constructed from BBs of trusted suppliers
- Competitive pricing, 5-6 weeks lead time
- **Experimentally confirmed synthesis success rate 80%+**

Freedom Space 4.0 - Overview

Created using ML-based reagent filters
40 reaction protocols

Expected synthesis
success rate 80%

4-6 weeks lead time,
competitive fixed price



Ro5 Enumerated Subset

20 billion molecules

- Molecular Weight: 0-500
- Hydrogen Bond Donors: 0-5
- Hydrogen Bond Acceptors: 0-10
- LogP: ≤ 5

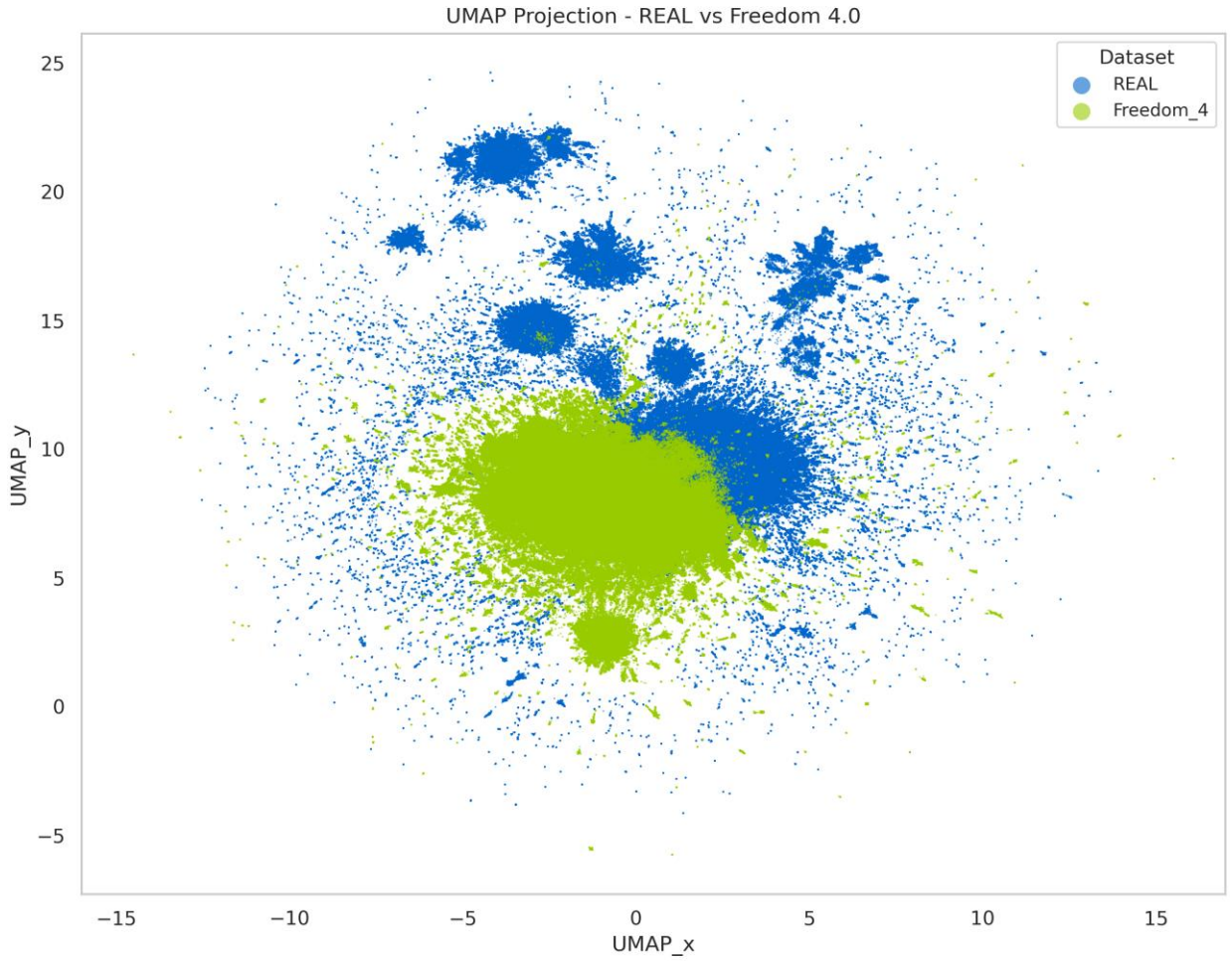
Beyond Ro5 Enumerated Subset

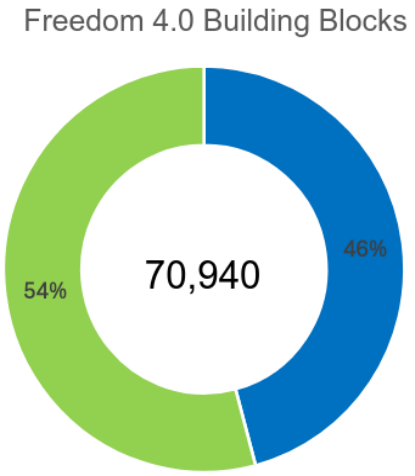
68 billion molecules

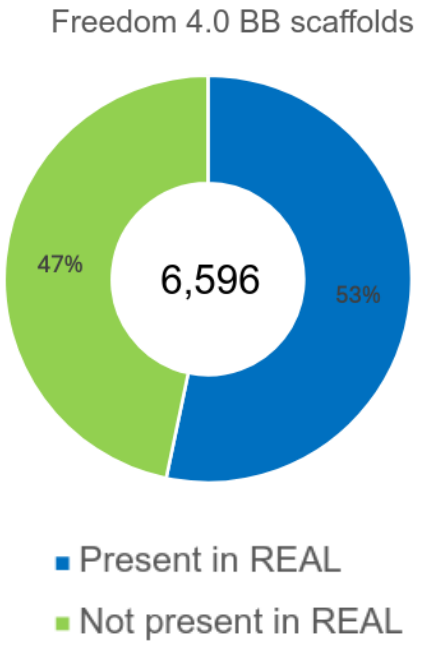
- Molecular Weight: 0-800 (excluding Ro5 subset)
- Hydrogen Bond Donors: 0-8
- Hydrogen Bond Acceptors: 0-15
- LogP: ≤ 6
- Rotatable Bonds: 0-15

Freedom Space 4.0 vs Enamine REAL

The space is designed to be complementary to Enamine REAL







The UMAP is based on 1M random selections with retained reaction proportions

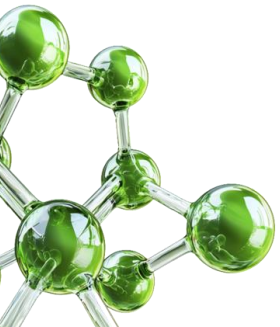
Access Options

The **enumerated subsets** can be downloaded using FTP.

The full **142B synthon-based space** is available through our partners BiosolveIT (InfiniSee), Alipheron (Hyperspace, Pharos3D), and RDKit (RDKit Library).

We provide **raw synthons and reactions file** upon request to support synthon-based developments.

Free **Freedom Search Platform** can be used to perform searches without registration.

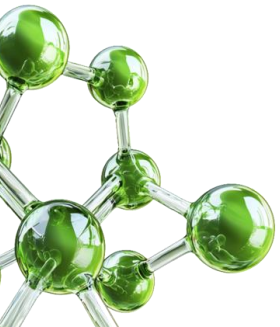


cs_sales@chem-space.com

Completely free, need to sign a licensing agreement

Just email “I want Freedom synthons”

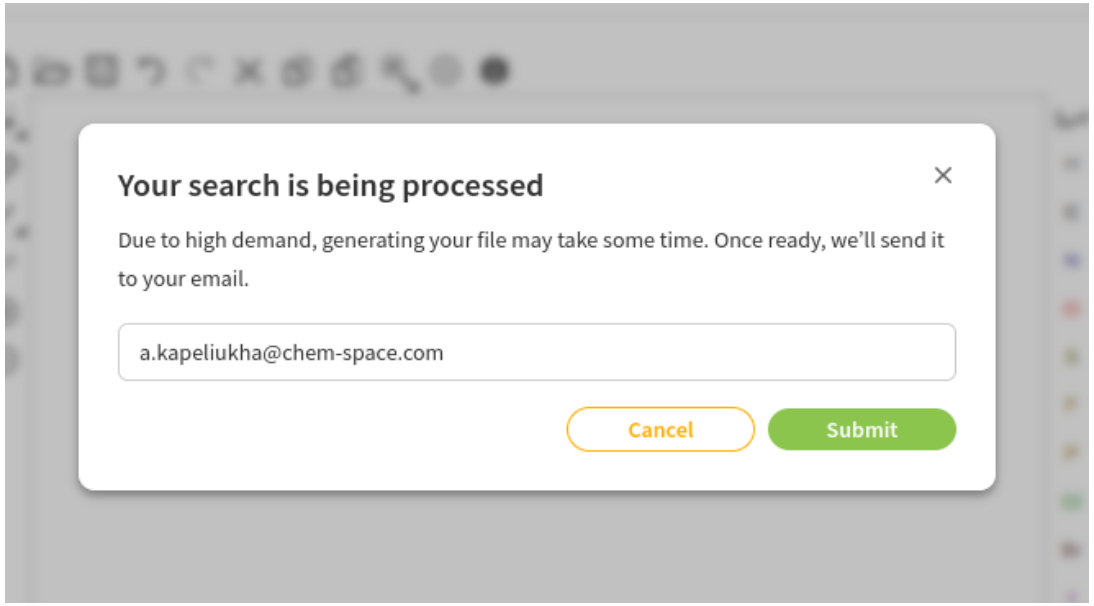
Freedom Search Platform – Powered by RDKit Synthon-Based Search



Freedom.chem-space.com

Freedom Search Platform - Results

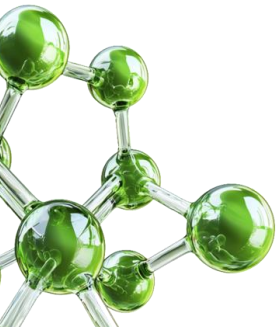
The results are sent to the provide email when the search is complete



Output up to 10k molecules per query

File columns:

- Query SMILES
- Freedom_SMILES
- Freedom_ID



RDKit Synthon-Based Search: Benchmarking

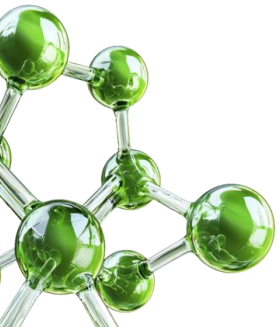
Participants of the study

Freedom Space 4.0 – 142B molecules

Enamine REAL Space – 77B molecules

Freedom Space 3.0 – 9B molecules

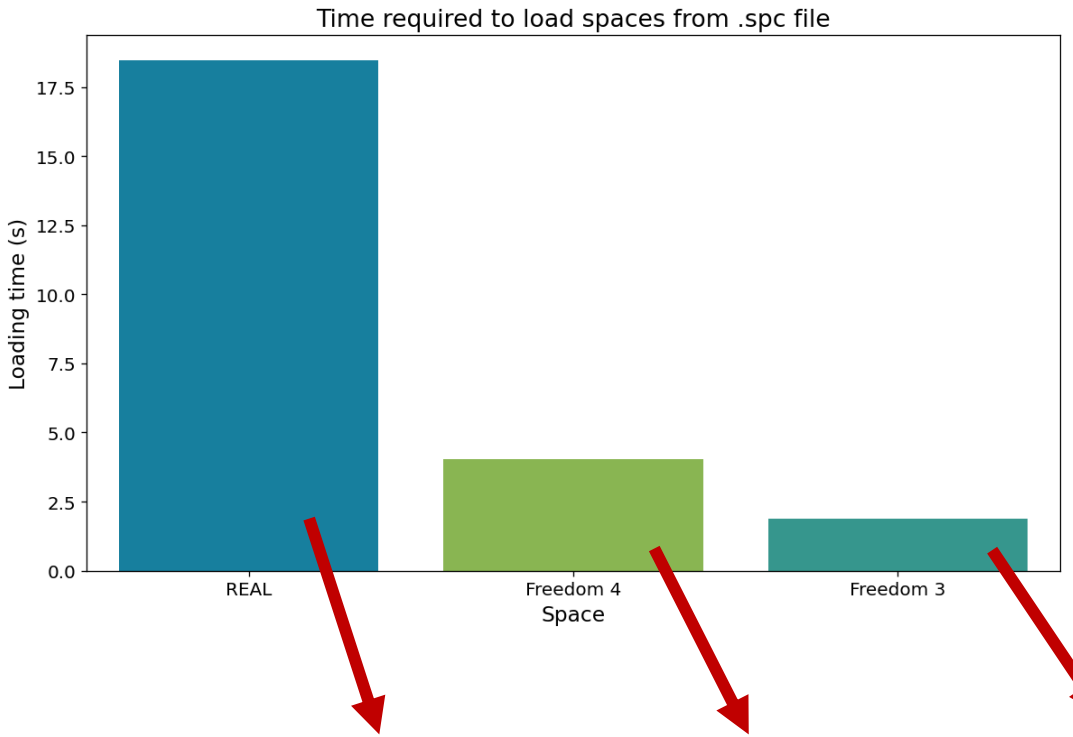
RDKit Synthon-based search

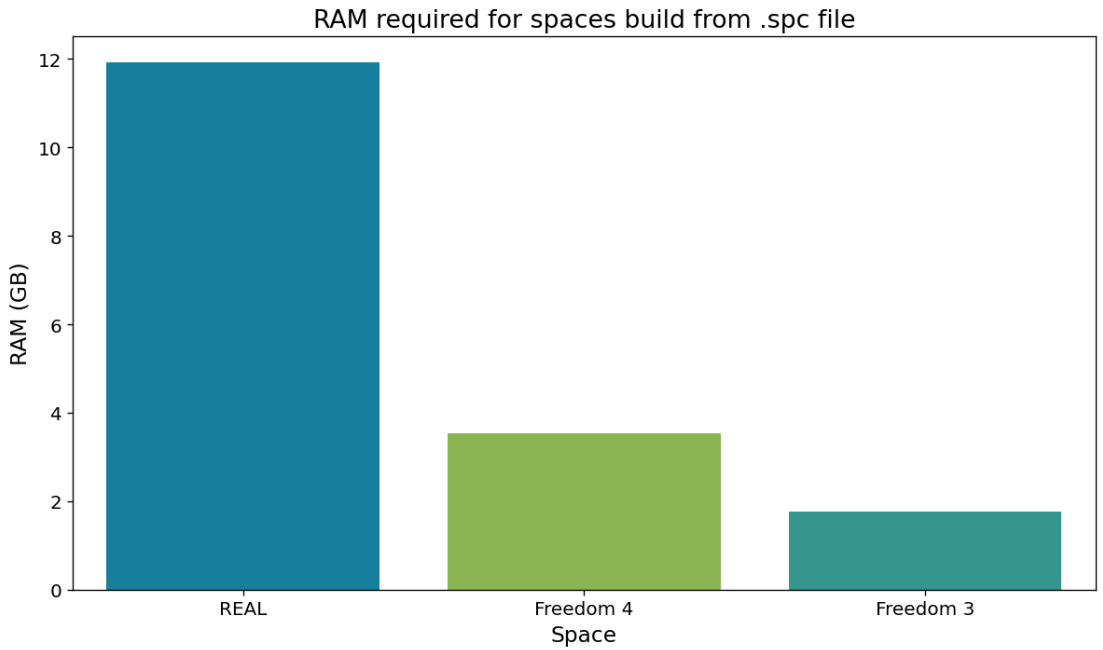


All calculations were performed on Intel(R) Core(TM) i7-14700F with 64GB RAM, RDKit version 2025.03.5

RDKit Synthon-Based Search: Benchmarking

Step 1: Loading the space using the synthons

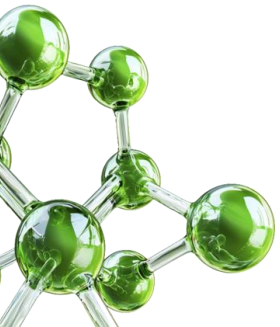




916 sub reactions

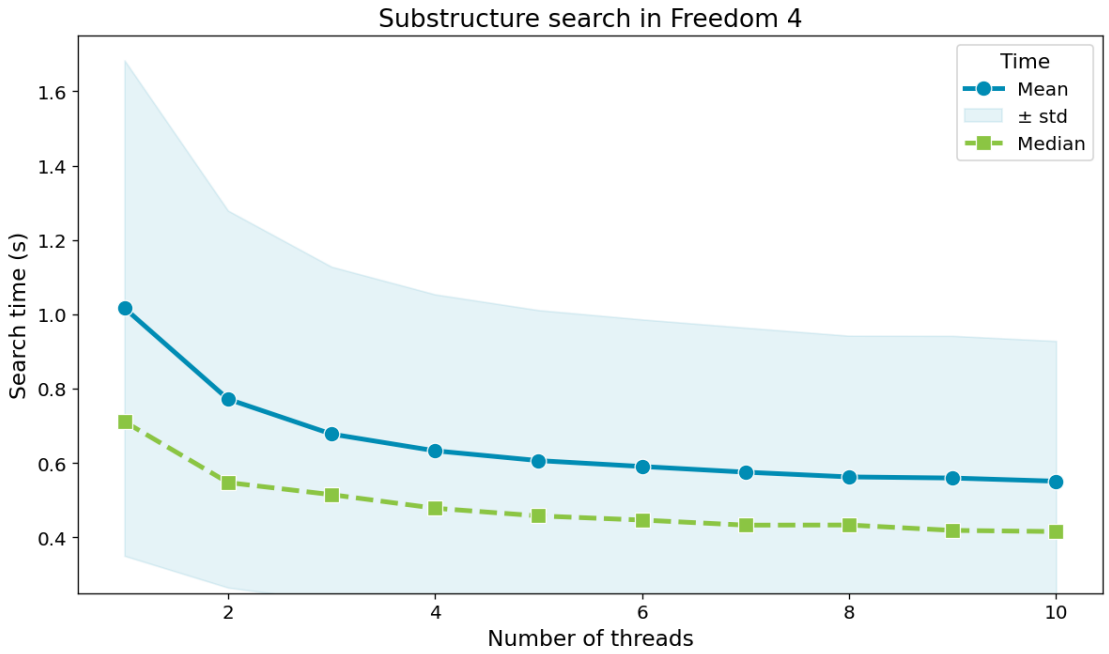
42 sub reactions

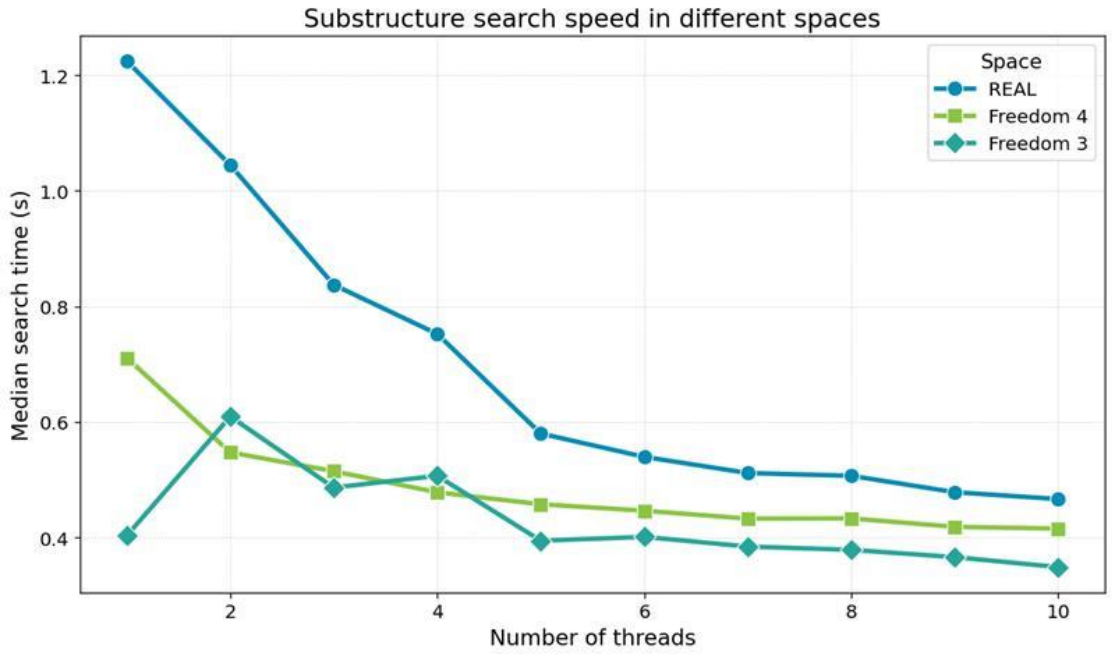
10 sub reactions

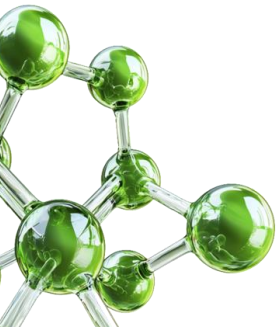


RDKit Synthon-Based Search: Benchmarking

Step 2: Substructure search – searching 33 Murcko scaffolds of approved drugs from ChEMBL

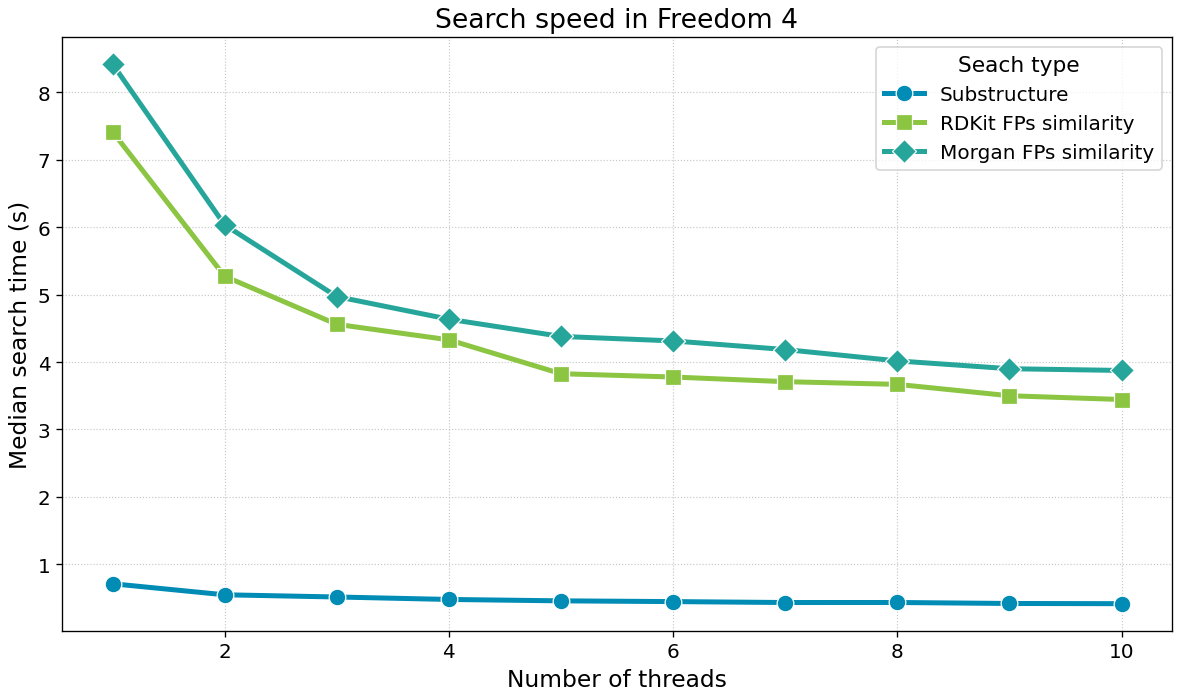


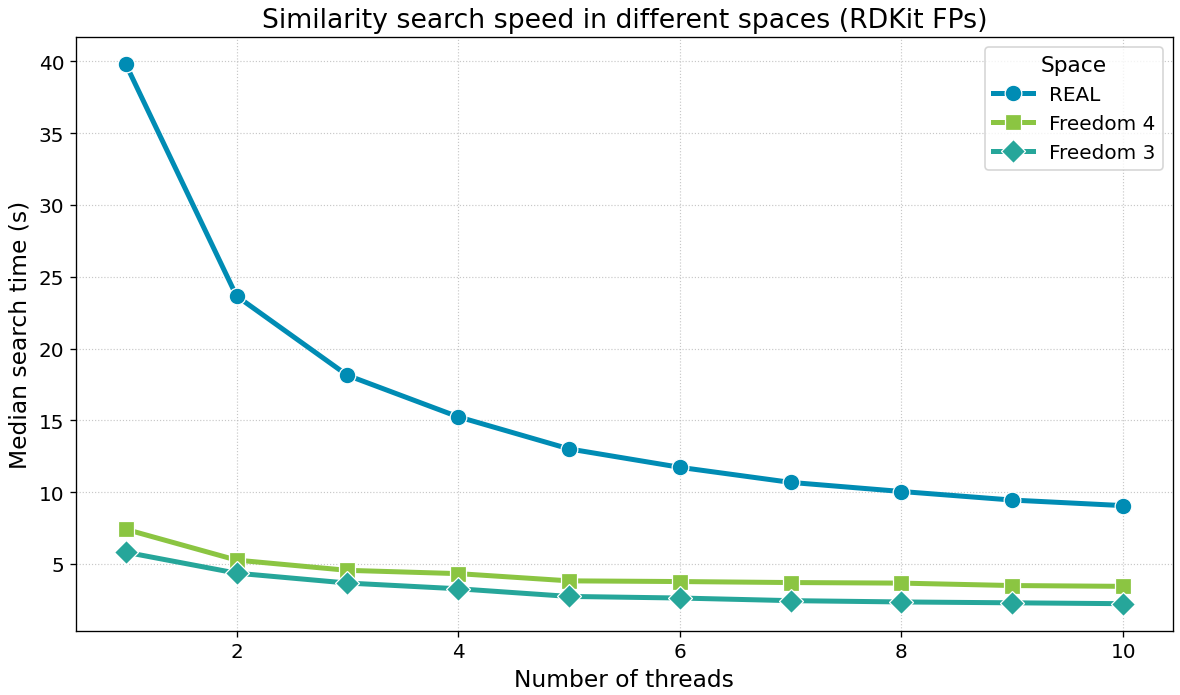


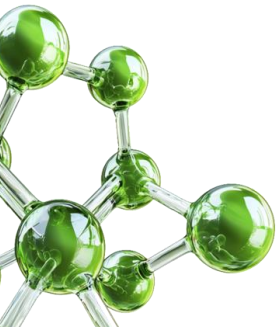


RDKit Synthon-Based Search: Benchmarking

Step 3: Similarity search – 38 diverse compounds from different sources (ChEMBL, in-stock)

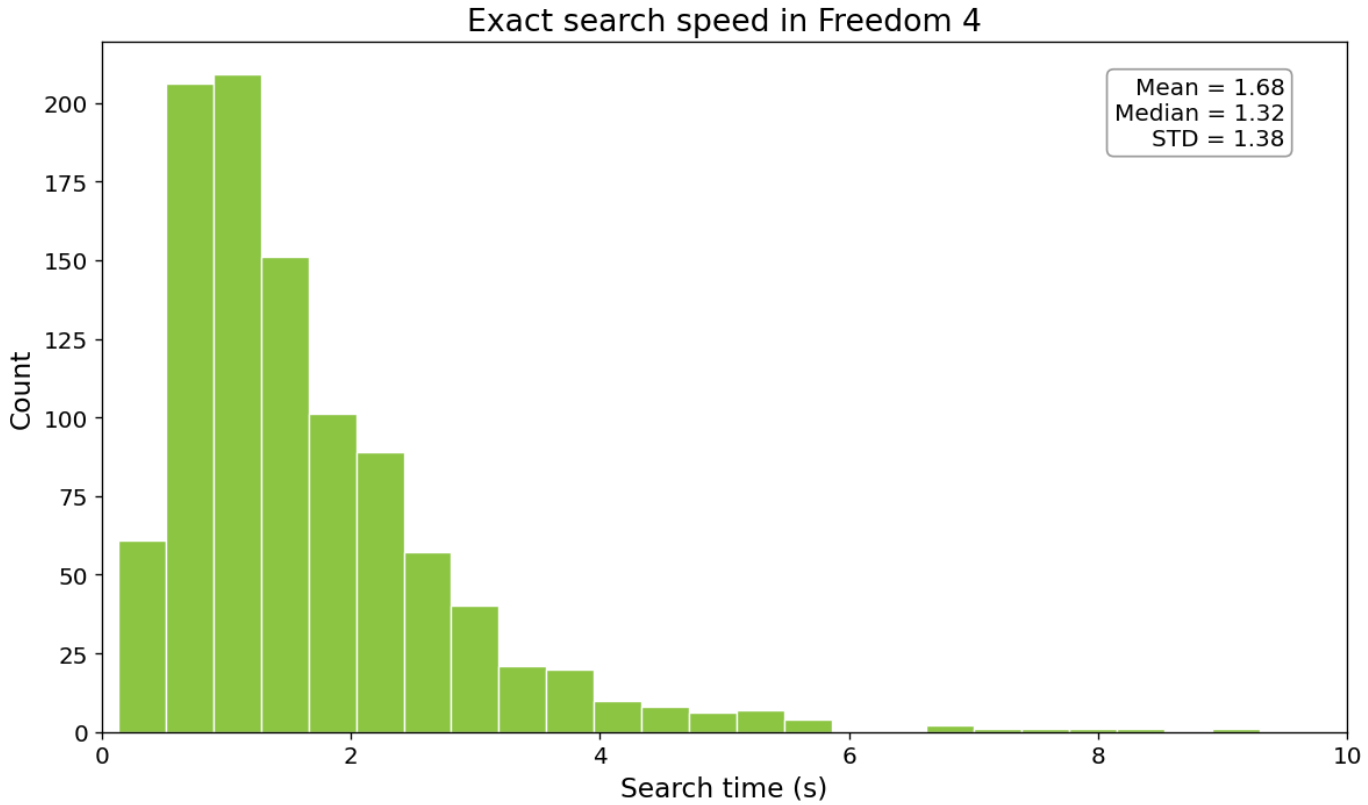


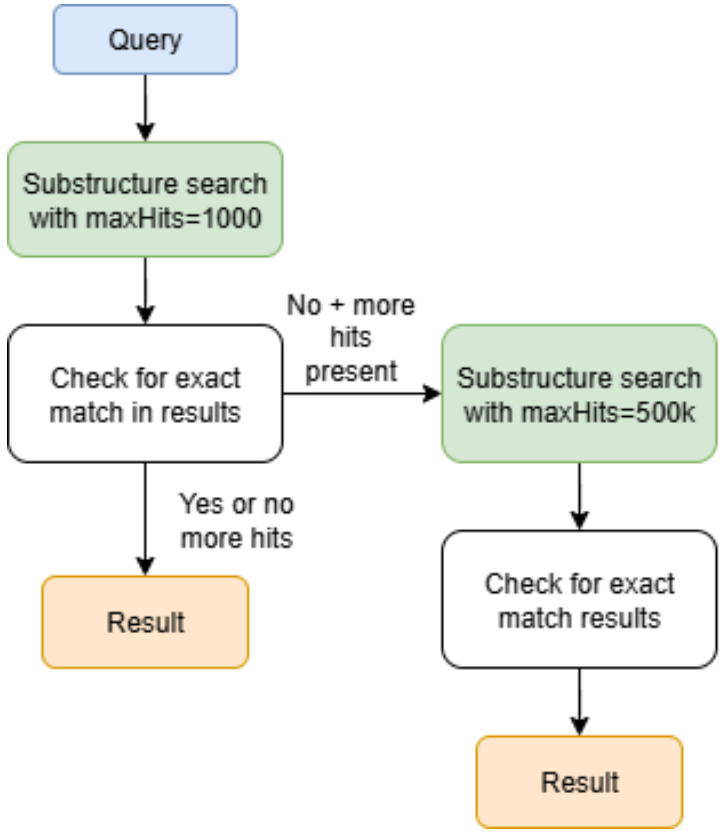


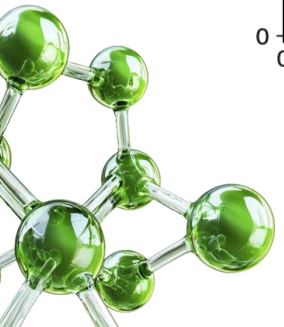


RDKit Synthon-Based Search: Benchmarking

Step 4: Using substructure search for exact matching – 1000 compounds from Freedom 4







Acknowledgements

Olga Tarkhanova, PhD
CEO

Serhii Hlotov
Data Scientist

Maryna Vasylchyk
Data Scientist

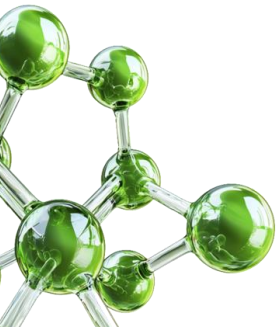
Oleksandr Mosia
Data Scientist

Mykola Protopopov, PhD
Head of Computational Chemistry

Enamine

Yurii Moroz, PhD
VP of Sales and Marketing, Enamine

David Cosgrove, PhD
Freelance Cheminformatics Developer



Thank you!

sales@chem-space.com