In silico generation of heterocycle-containing drug-like small molecules:

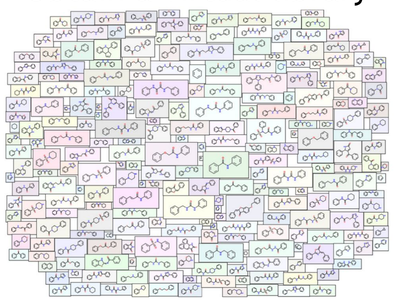
Towards tools for the many different needs of drug discovery projects.

Brian Atwood, Hamza Tajmouati, Vinicius Barros Ribeiro da Silva, Maoussi Lhuillier-Akakpo,
Rohit Arora, Philippe Gendreau, Yann Gaston-Mathé, Nicolas Do Huu, Quentin Perron

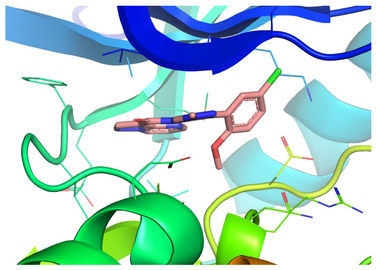
Computer-aided drug design (CADD)

The major paradigm for **hit-finding** in CADD:

Obtain a virtual library



Screen and rank



- How diverse are current virtual libraries?
- How synthesizable are compounds within virtual libraries?
- Screening and searching ultra-large libraries quickly becomes costly or infeasible.
- "Good" compounds need to *already* be a part of your virtual library.

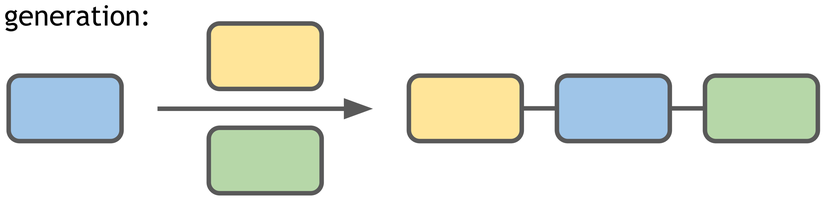
Our largest virtual libraries are massive! On the order of 10^{10} to 10^{20}, but these pale in comparison to the size of chemical space, estimated to be upwards of 10^{60}.

Ultra-large virtual libraries: Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, Yurii S. Moroz, Generating Multibillion Chemical Space of Readily Accessible Screening Compounds, *iScience*, **2020**, 23, *11*, 101681, https://doi.org/10.1016/j.isci.2020.101681.

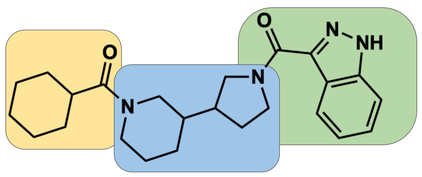
Building virtual libraries

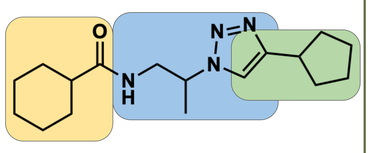
Modern commercial virtual libraries have largely shifted to combinatorial reaction-based composition. These libraries aim to balance synthesizability and novelty, and are often able to deliver make-on-demand compounds with an 80% success rate. We will discuss these libraries, their strengths and weaknesses, and compare with our own approaches.

Generalized approach to combinatorial reaction-based library generation:



Example compounds from Enamine REAL Space:





Some of the drawbacks of these libraries are:

- The need for pre-defined reactions limits virtual library growth along defined vectors.
- The linearity inherent in the stepwise approach leads to linearly constructed molecules.
- They fail to capture many heterocycle syntheses that have non-obvious *a priori* functional group pairings.
- Failure modes for synthesis often occur when failing to take into account the context of the rest of the molecule; i.e. only looking at the reacting functional group.
- As virtual libraries grow larger, screening efforts become prohibitively costly from a computational perspective.

Iktos generative infrastructure

At Iktos, we have developed AI and ML tools for retrosynthesis prediction and reinforcement-learning-based generation of novel molecules.

spaya.ai

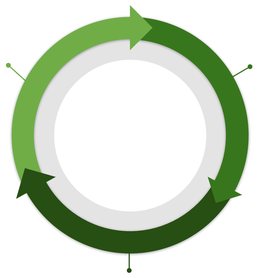
- Reaction templates
- Reaction feasibility models
- Regiochemistry models
- Chemoselectivity models

makya

- Molecule generators
- Reinforcement learning
- AutoML models
- Property optimization

Free-to-react generators and reinforcement learning:

Generate molecules based on chemistry rules, without needing pre-defined reaction vectors.



Score molecules using predictors:

- QSAR models
- 3D Simulations
  - Docking
  - 3D QSAR
- Retrosynthesis
- Generic scores

Adjust the AI for the next set of molecules, in order to optimize predicted scores.

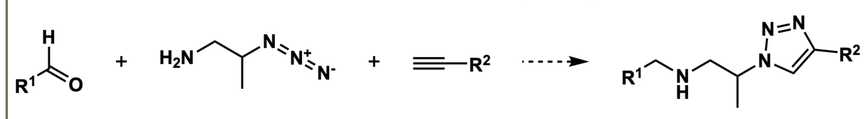
These methods allow us to explore chemical space as we generate new molecules, finding regions with optimal properties and increasing the chances of the "best" virtual molecules being screened.

Generation of heterocycle-focused molecule libraries

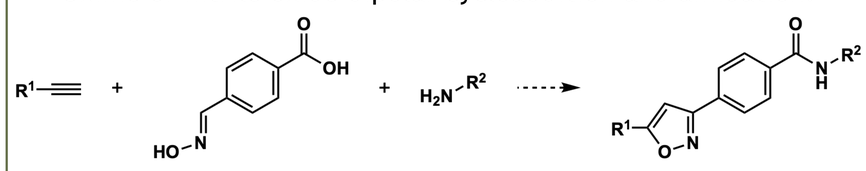
The *de novo* synthesis of heterocycles in most virtual libraries requires explicitly specifying their construction in one of the chemistry steps used to build the library. Recently, the Enamine team demonstrated that a virtual chemical space of more than 29 billion compounds could be generated with the only heterocycle-forming reaction being the azide-alkyne "click" reaction. There is concern that this approach will have the same drawbacks as the combinatorial chemistry efforts of the late 90s: high structure and shape similarity in libraries, low overall diversity, and homogenous library property space.

To investigate the differences between libraries built from pre-defined reactions versus those built from reacting freely *in silico*, we constructed a handful of test cases:

Reductive amination and azide-alkyne cycloaddition:



Oxime or nitrile oxide dipolar cycloaddition and amidation:

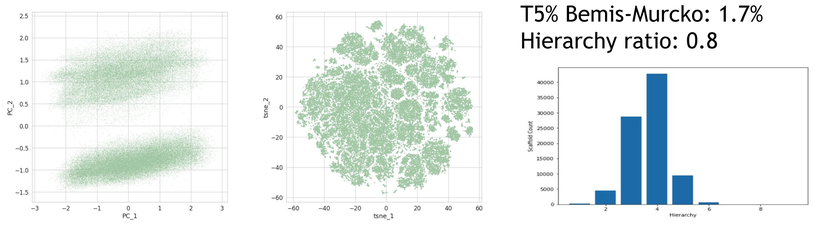


"Freely reacting" seed compounds for library generation were urea and bromoacetone.

Chemical space analysis

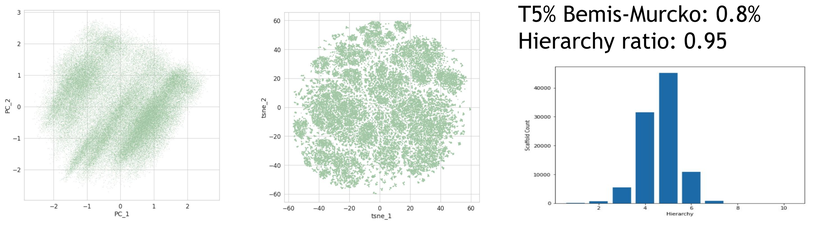
Our preliminary assessment indicates that for this test comparison, the libraries generated using pre-defined reaction patterns were less diverse overall than those generated without such constraints.

Azide-alkyne cycloaddition



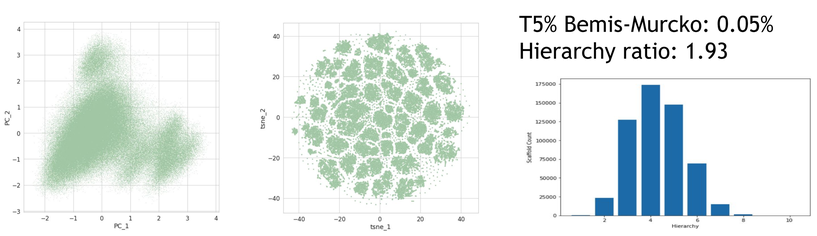
T5% Bemis-Murcko: 1.7%
Hierarchy ratio: 0.8

Oxime or nitrile oxide cycloaddition



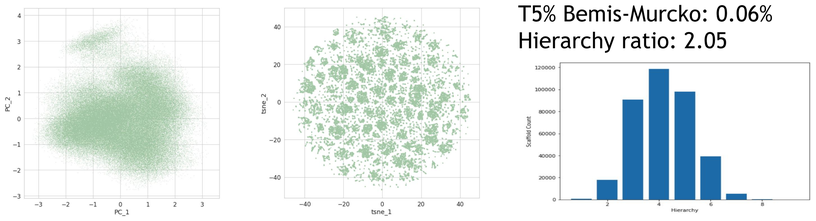
T5% Bemis-Murcko: 0.8%
Hierarchy ratio: 0.95

"Freely reacting" urea



T5% Bemis-Murcko: 0.05%
Hierarchy ratio: 1.93

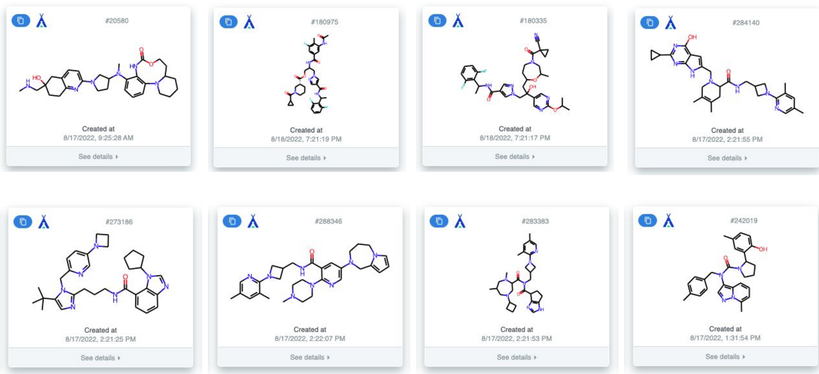
"Freely reacting" bromoacetone



T5% Bemis-Murcko: 0.06%
Hierarchy ratio: 2.05

Chemical space is visualized by projecting molecular fingerprints into the first two principal component analysis (PCA) dimensions, as well as by reducing the first 50 PCA dimensions into two dimensional space with t-distributed stochastic neighbor embedding (tSNE). Each library consists of at least 100,000 molecules. As an additional measure of diversity and chemical space coverage, the percentage of molecules in the top 5 Bemis-Murcko scaffolds and the ratios of hierarchical scaffolds per molecule are reported for each library.

Selected generated compounds:



We have applied our technologies to a number of drug discovery projects, including a publicly disclosed collaboration with Oncodesign wherein we identified novel inhibitors of PIM1 kinase *in silico* and validated with only 11 compounds synthesized.