A Tera-Scale Entropy-regularized Retrieval Augmented Generative Framework for Analog Expansion in Drug Discovery.

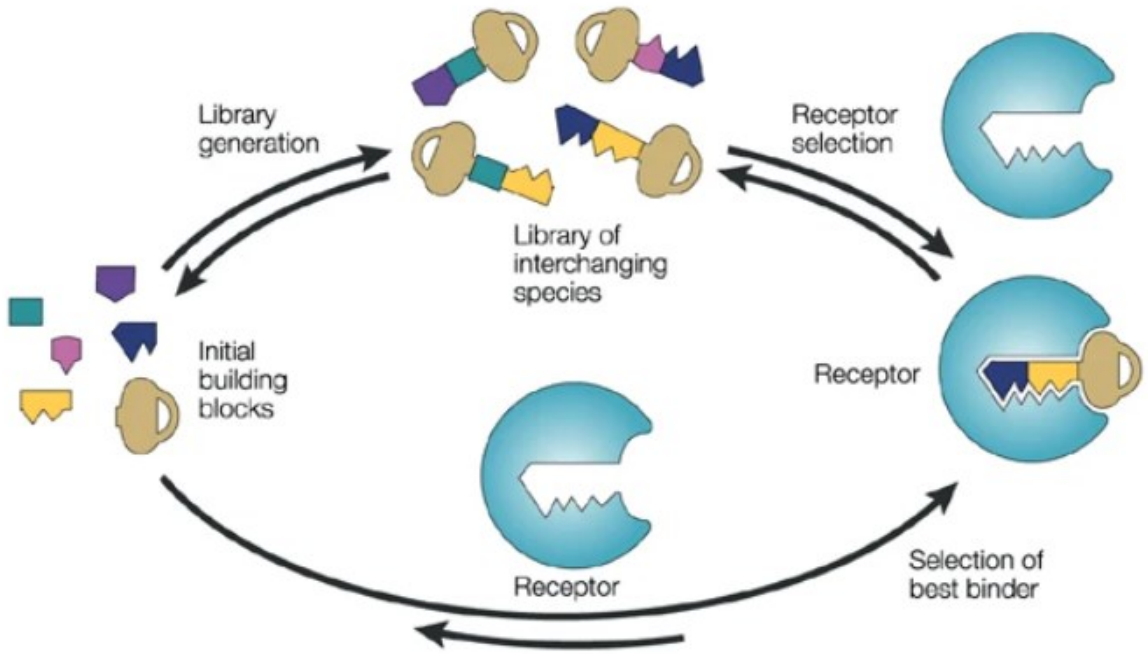
Image Credit: https://cen.acs.org/pharmaceuticals/drug-discovery/Hunting-drugs-chemical-space/100/i23

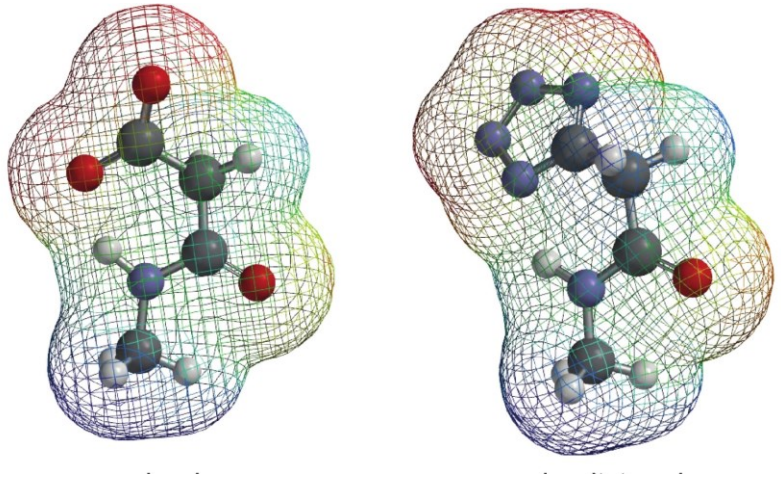
USC Dornsife
Department of Quantitative and Computational Biology

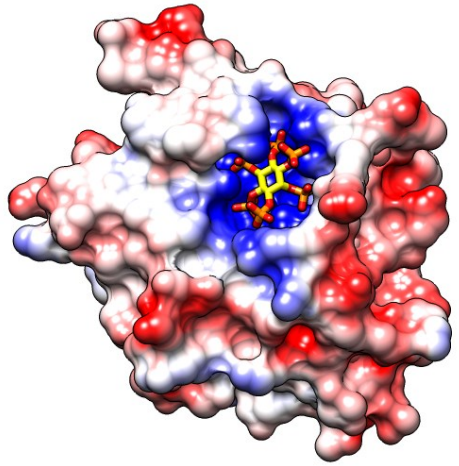
July 2025

Jordy Homing Lam
Katritch Lab
University of Southern California

Drug Discovery is a Combinatorial Problem



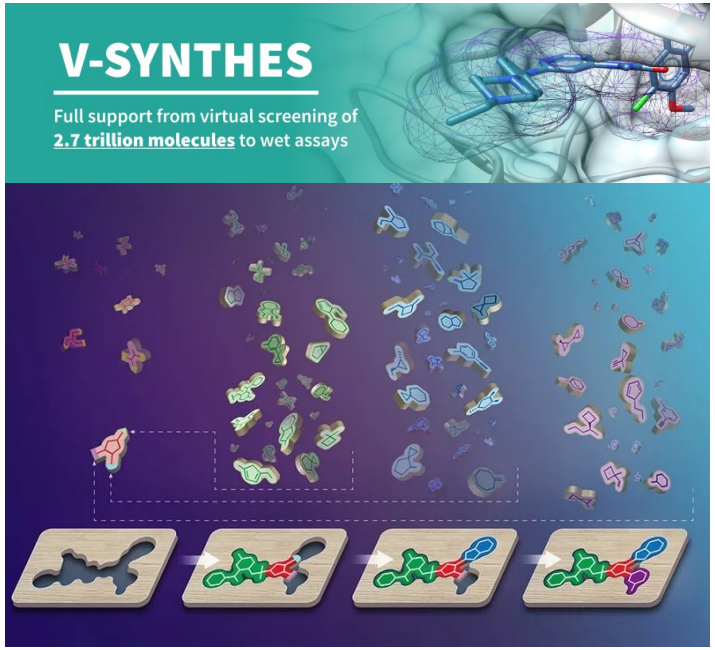


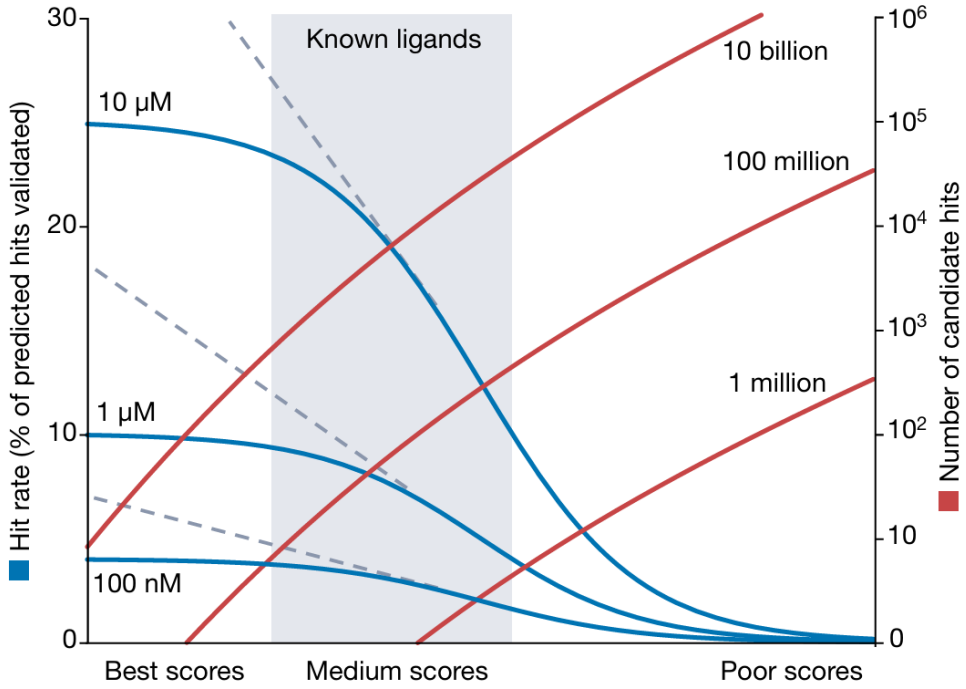


Chemical Physics:
Long/short-ranged electrostatics
Electron Density
Strains, Interactions
Thermodynamics

Ramström, O., Lehn, JM. Drug discovery by dynamic combinatorial libraries. *Nat Rev Drug Discov* **1**, 26–36 (2002). https://doi.org/10.1038/nrd704
Yihang Jing, Sarah E. Bergholtz, Anthony Omole, Rhushi A. Kulkarni, Thomas T. Zengeya, Euna Yoo, Jordan L. Meier Synthesis and Evaluation of a Stable Isostere of Malonyllysine, ChemBioChem 23(1)

Tera-Scale Combinatorial Spaces

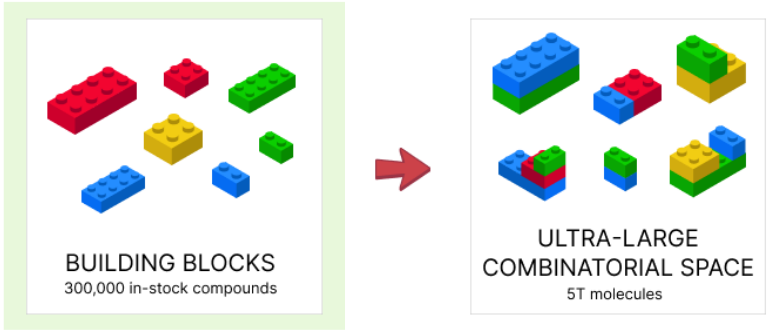




Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Sadybekov, A.A.; Sadybekov, A.V..; Katritch, V. et al Nature 2021, 601 (7893), 452-459.
Anastasiia Sadybekov, Vsevolod Katritch, Computational approaches streamlining drug discovery
Katya Cherezov's Illustration

Three Properties of Combinatorial Space

- 1. Exponential Growth
  - Modular building blocks
  - 10^12 – 10^70



- 2. Baked-in Synthetic Logic
  - > 200,000 synthons for 100+ reactions
  - Separate into 2300 sets.
  - Overlaps but not free to mix-and-match.

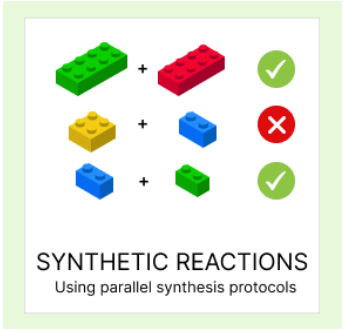
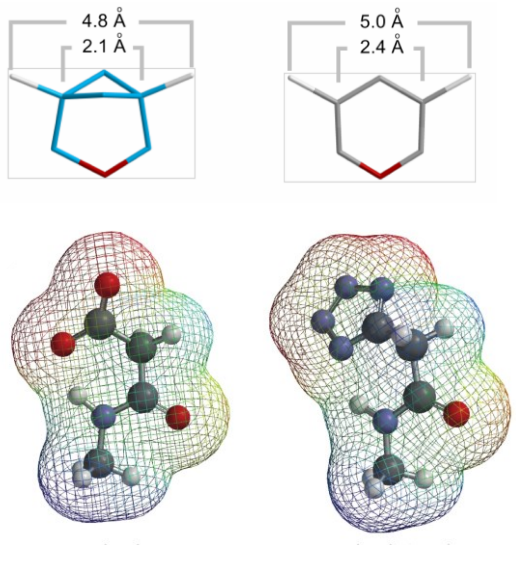
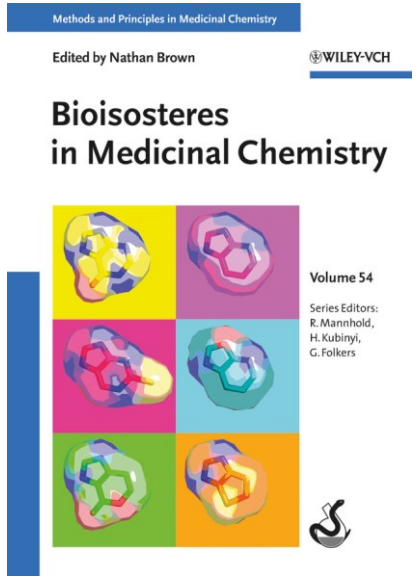


Image credit: ChemSpace

Three Properties of Combinatorial Space

- 3. Property-enriched isosteres
  - Similar geometry + physics
  - ↓ lipophilicity
  - ↑ solubility

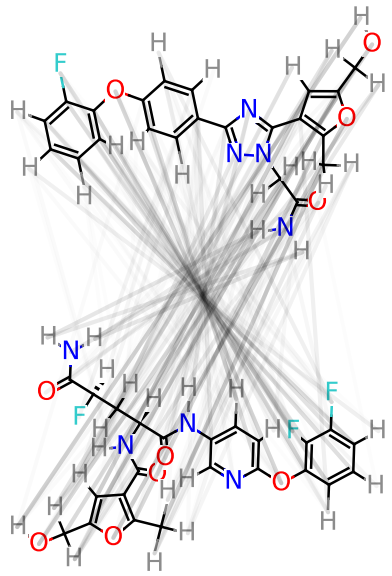




Physical similarity != Graph similarity

VRAGFN – ML-assisted search engine

- V-SYNTHES Retrieval Augmented Generative Flow Network
- A Tera-Scale Entropy-regularized Retrieval Augmented Generative Framework for chemical space exploration



VRAGFN assimilates all 3 properties

1. Exponential Growth + Cross-library Search
2. Baked-in Synthetic Logic
3. Property-enriched isosteres

(Unpublished)

Before we begin…

Do you prefer a search in this space?

“utensils”

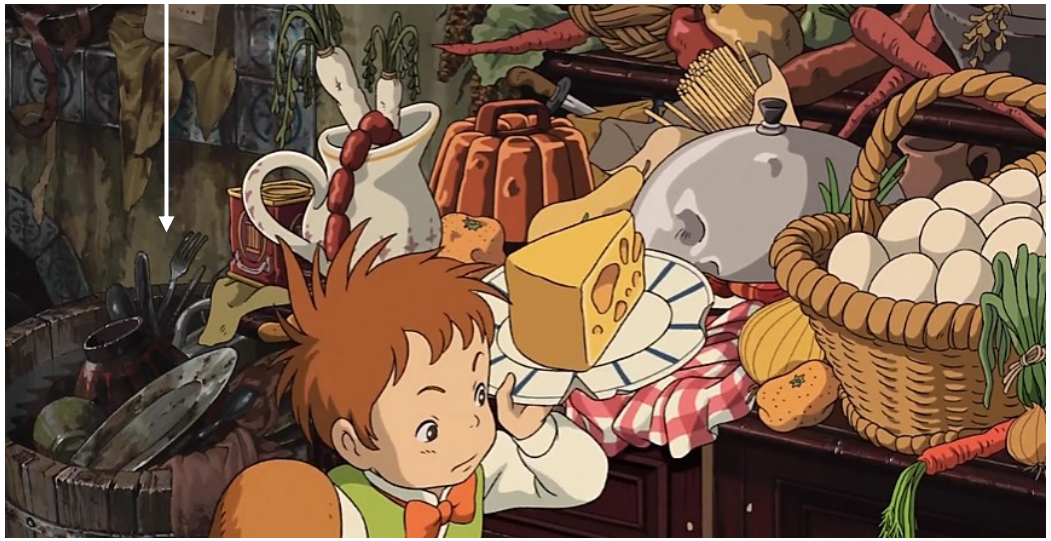


Image Credit. Howl's Moving Castle, Ghibili Studio ‘04

Imagine helping grandma getting her dinner

- Query: A cheese knife for cheese.
- Answer: Somewhere there.

End up searching all over the place but still missed it!

Do you prefer a search in this space?

Image Credit. Howl's Moving Castle, Ghibili Studio ‘04

Imagine helping grandma getting her dinner

- Query: A cheese knife for cheese.
- Answer: Somewhere there.

A Good Search in Supermarket

Q: Grandma’s Cheese Knife at 7:00 PM

- Representation. Aisles clustered with principles.
- Stocked. Unchanging catalog. No hallucination.
- Conditional. The cart size and cuisine characterized
- Flow. Progress of search → Fast inference.

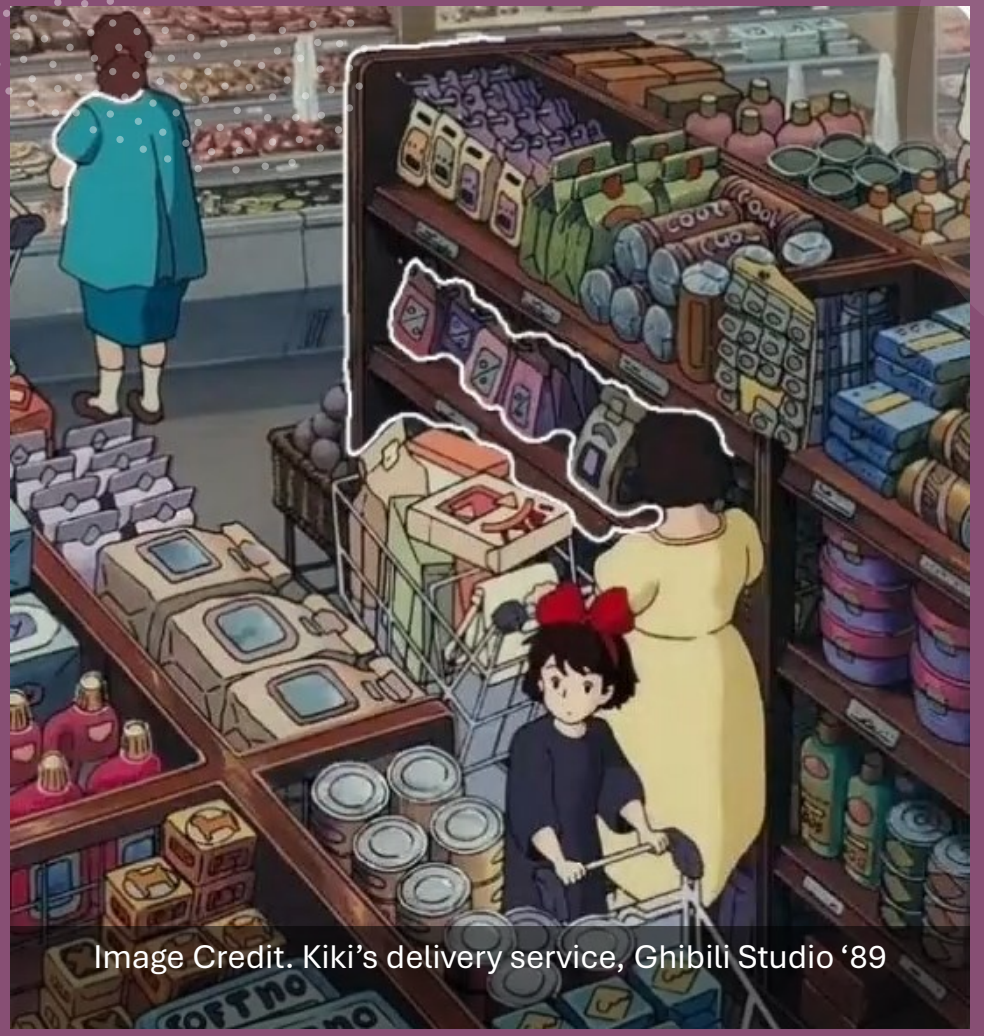
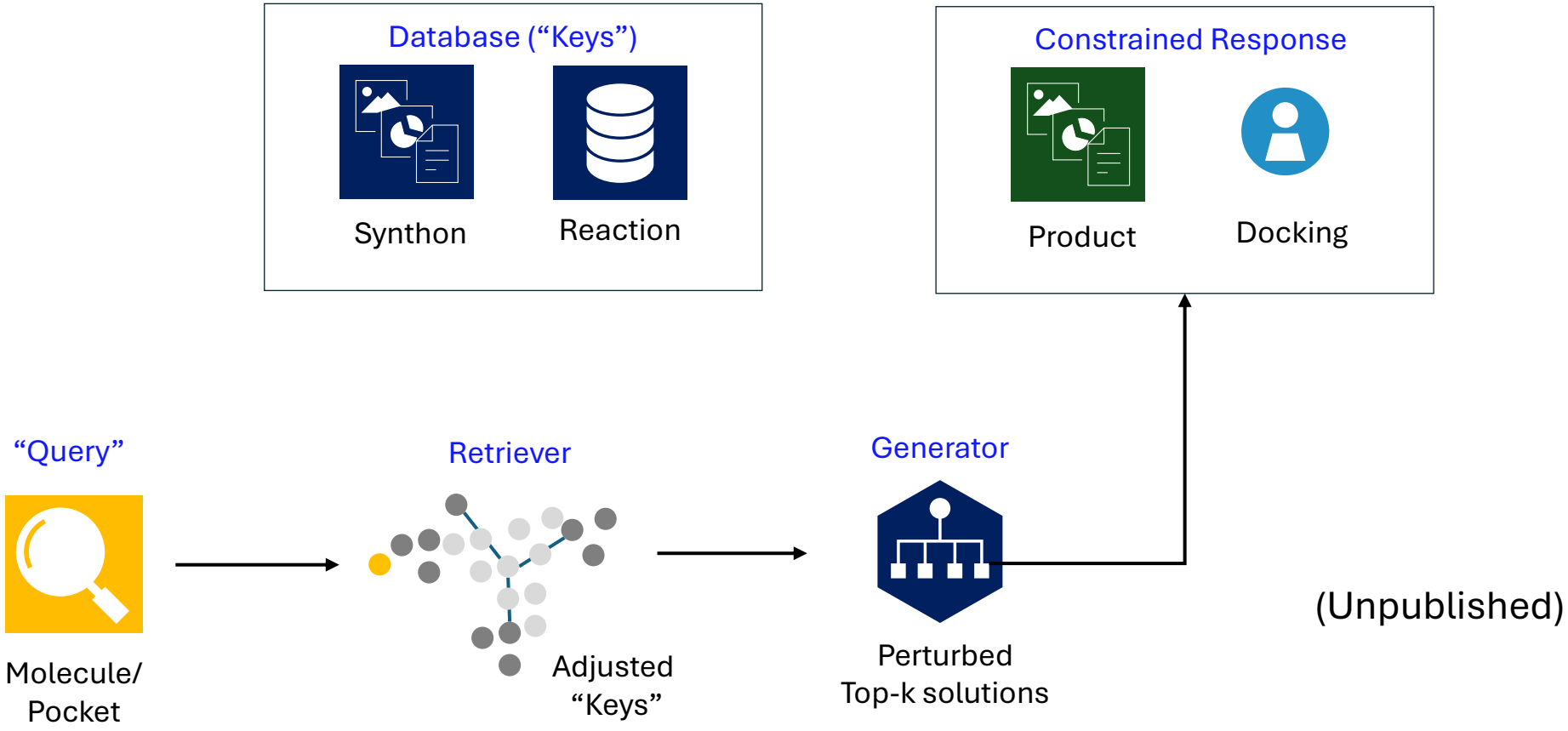


Image Credit. Kiki’s delivery service, Ghibili Studio ‘89

Retrieval Augmented Generation (RAG) is a scheme that learns to make constrained choices



(Unpublished)

Tokens in VRAGFN are physical representations

Database (“Keys”)

Synthon

Reaction

(Unpublished)

Representation

Physics

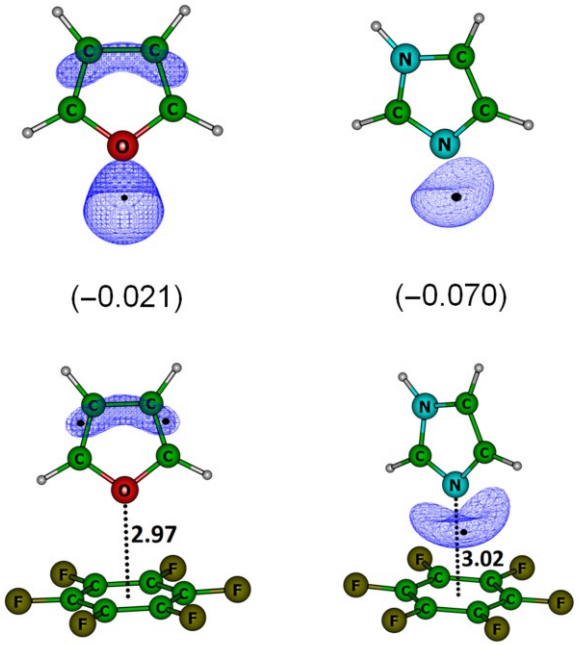
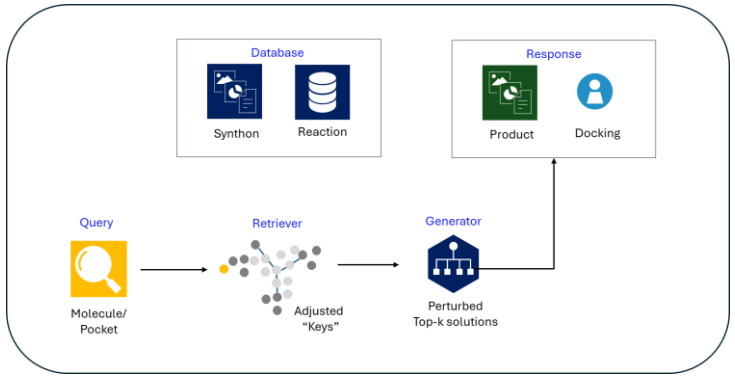
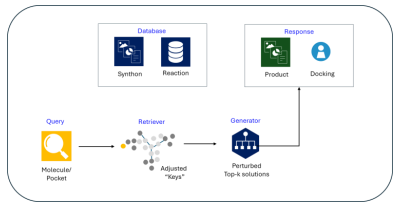


Image Credit. Cherumuttathu H. Suresh, Geetha S. Remya, Puthannur K. Anjalikrishna, https://doi.org/10.1002/wcms.1601

VRAGFN uses the RAG flow to propose chemicals within a constrained chemical space



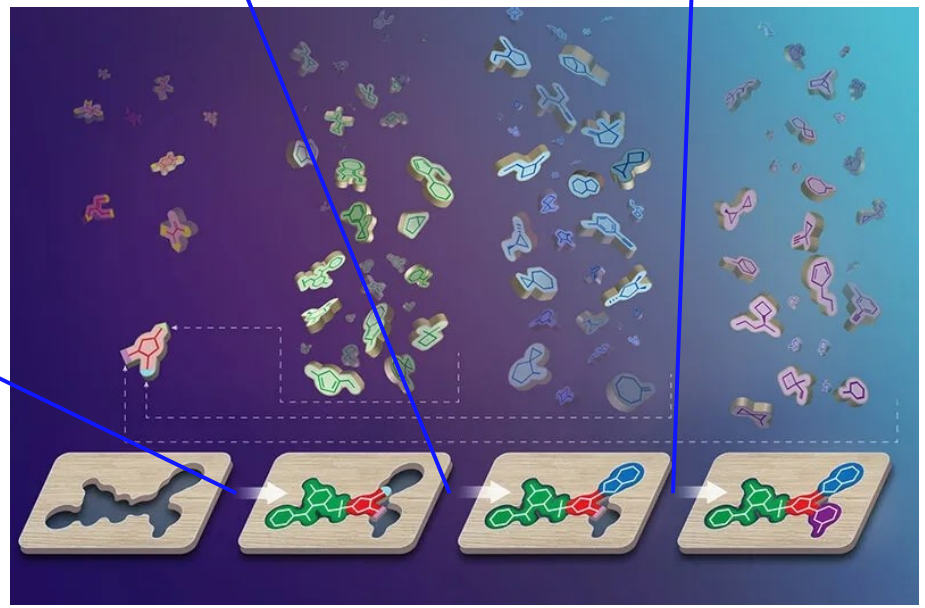


$P_F^{(0)}$

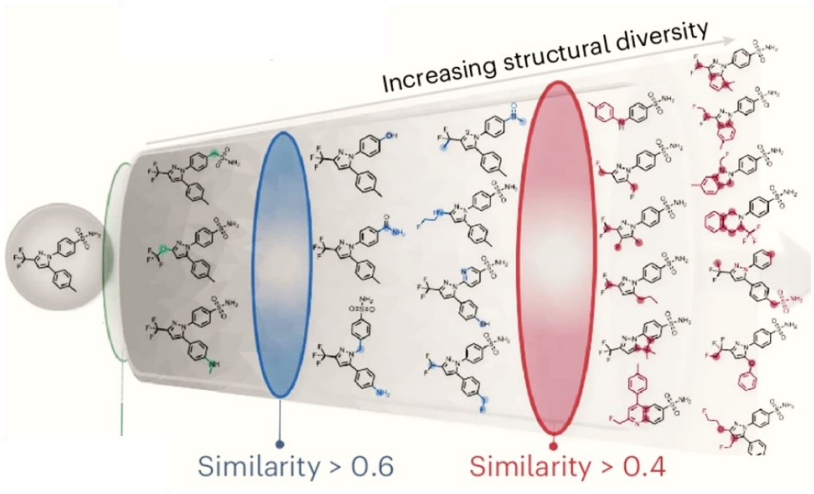
$P_F^{(1)}$

$P_F^{(2)}$

(Unpublished)



Telescoping
Action Space



Hit Expansion with VRAGFN

Application 1.0 on a 2.7 Trillion X-REAL space

Hit Expansion

- Build upon successful hits
  - ↑ Structural diversity
  - ↑ Property

  - "Second opinion" when side effect arises from experiments.

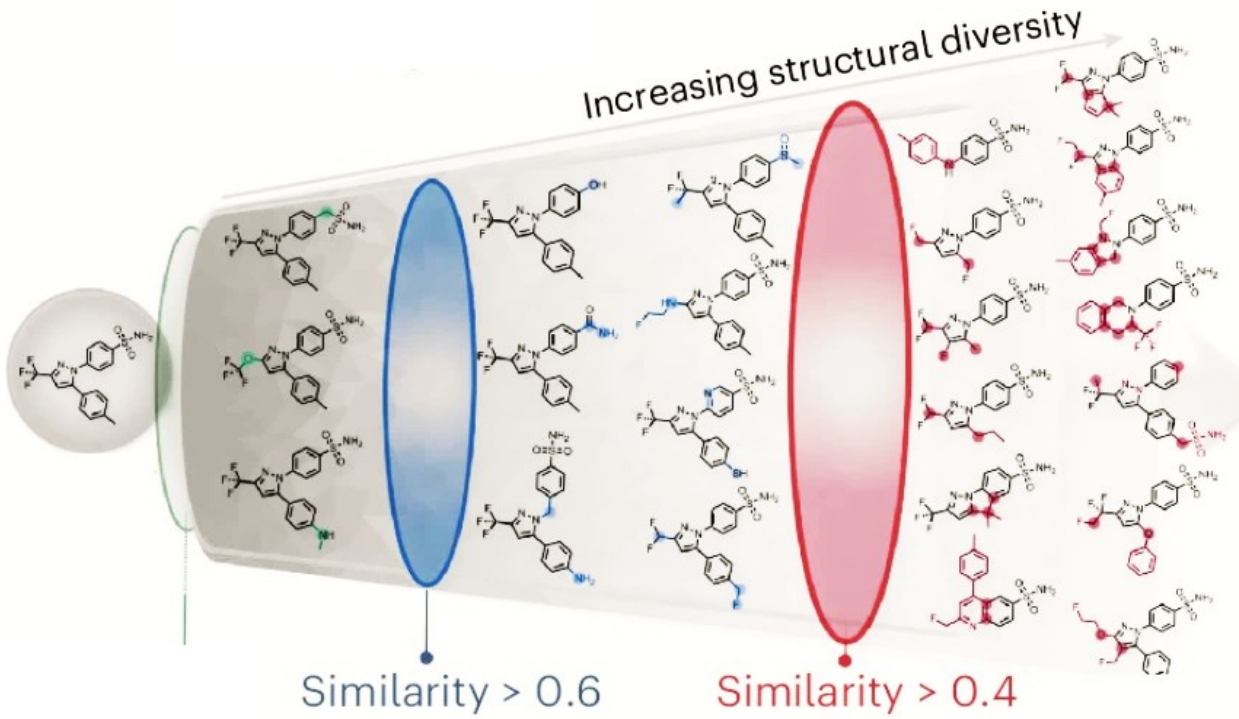
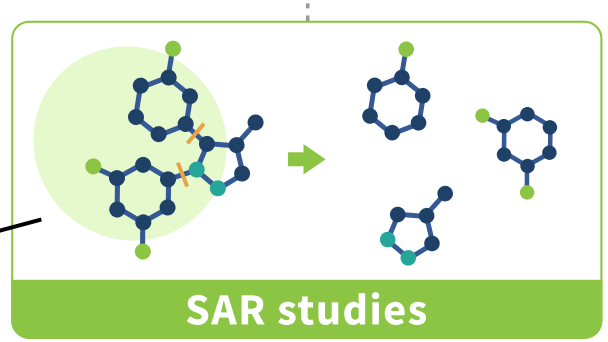
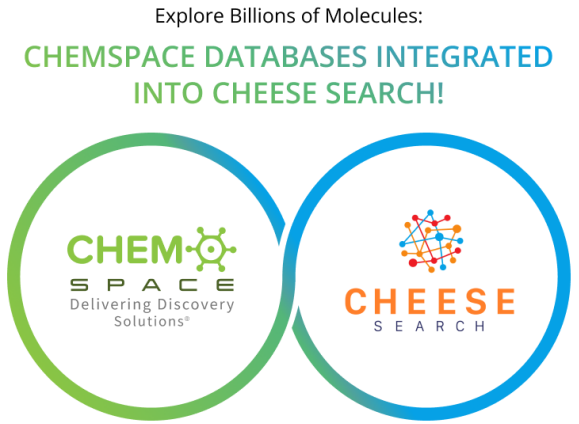


Image Credit: Ghazi Vakili, M., Gorgulla, C., Snider, J. *et al.* Quantum-computing-enhanced algorithm unveils potential KRAS inhibitors. *Nat Biotechnol* (2025). https://doi.org/10.1038/s41587-024-02526-3

SOTA Hit Expansion in Chemical Spaces

1. SAR-type decomposition.
   - RDKit 2025.05.3 rdSynthonSpaceSearch
   - 142B Freedom space
2. Whole-graph, vectorized search
   - Cheese ChemSpace
   - 10B Freedom space, Ro5 filtered.
     - → 1 trillion ~ 116 TB (Unscalable!)



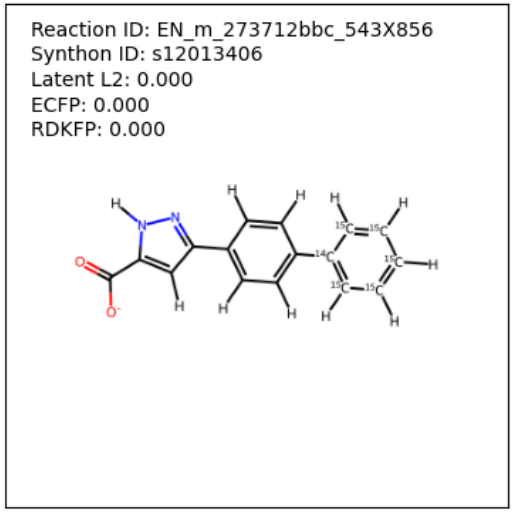


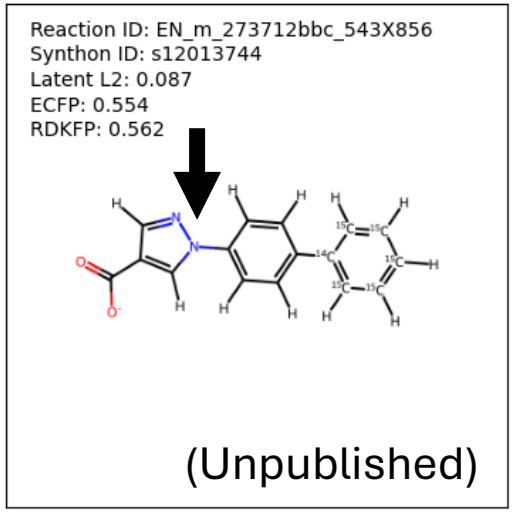
Both are Hash-based search

Hash-based search is ineffective

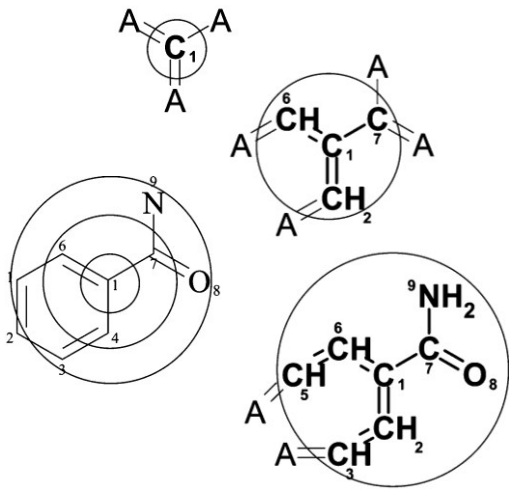
- Graphlet hashes can easily inflate dissimilarity.
  - 6 other types of blunders will be discussed in manuscript.

The cheese knife
not in utensils.
Grandma is confused!



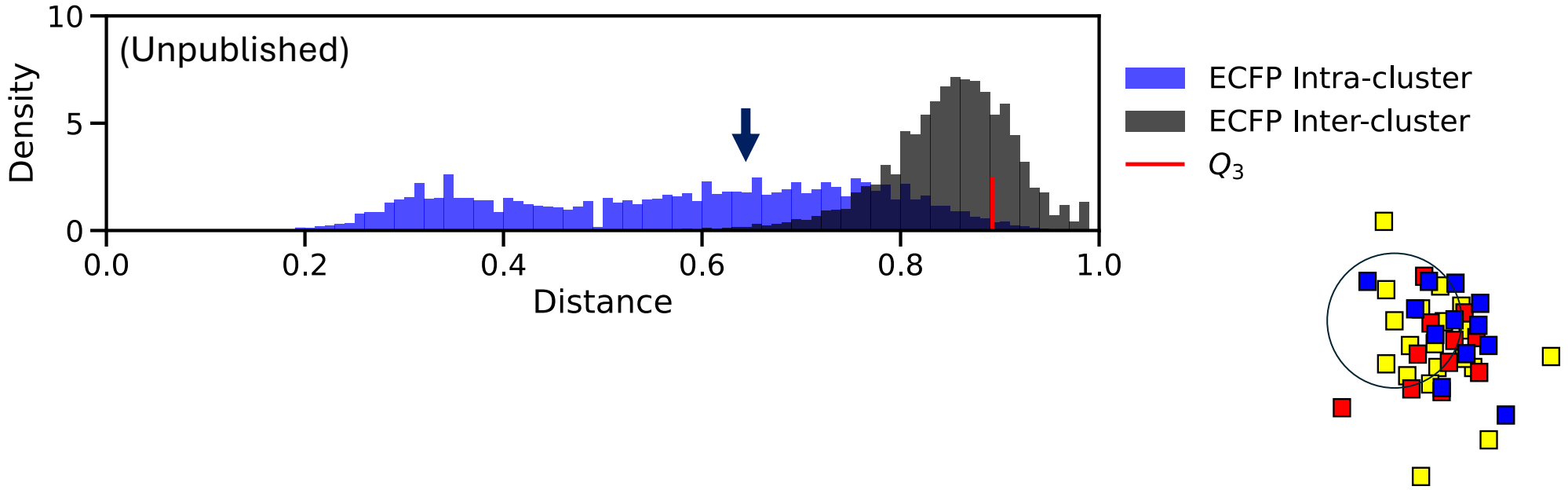


Tanimoto Distance
ECFP4: 0.554
RDKFP: 0.562



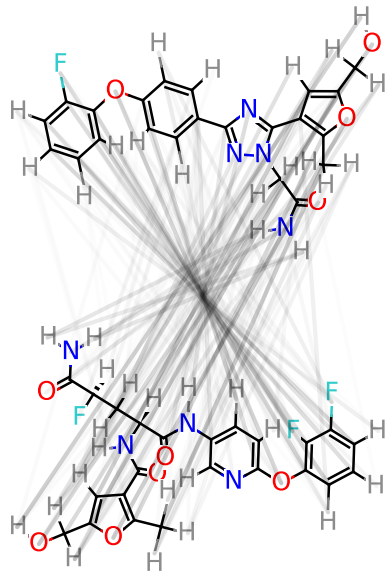
ECFP-Tanimoto inflates dissimilarity among isosteres

- Similarity in geometry and electronic distributions not respected
  - Poor distinguishing power → Ineffective search.



VRAGFN is trained to retrieve isosteres

- V-SYNTHES Retrieval Augmented Generative Flow Network
- A Tera-Scale Entropy-regularized Retrieval Augmented Generative Framework for chemical space exploration



VRAGFN assimilates all 3 properties

1. Exponential Growth + Cross-library Search
2. Baked-in Synthetic Logic
3. Property-enriched isosteres

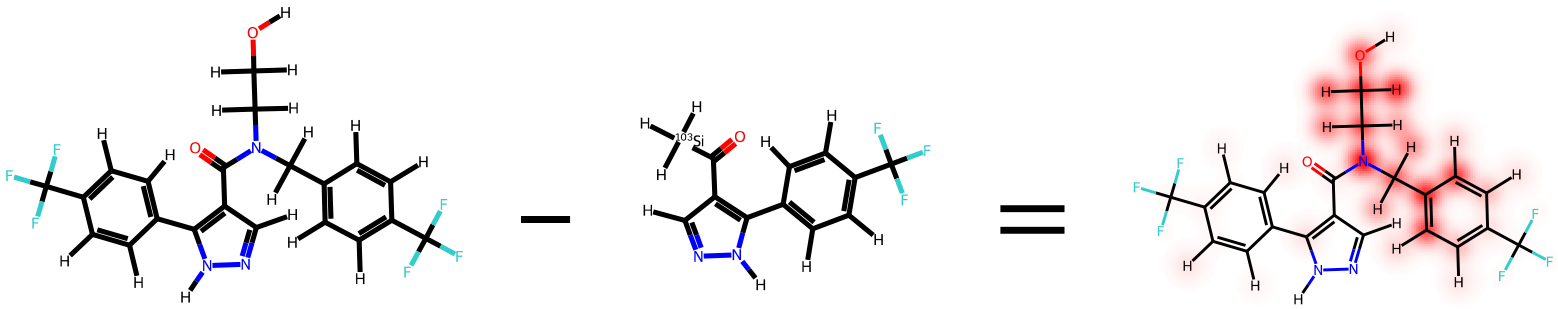
(Unpublished)

VRAGFN is a modular search approach

- The query is matched partially with synthons to reveal remaining parts

➢ We learn to decompose query with isosteric partial matches.

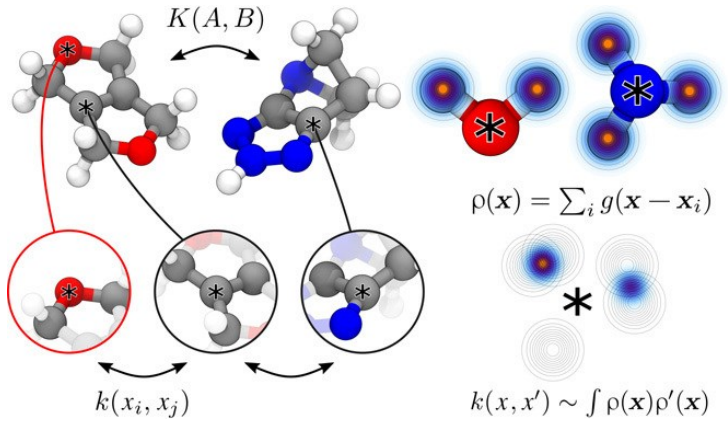
- Synthon similarity is learned from aligning isosteres

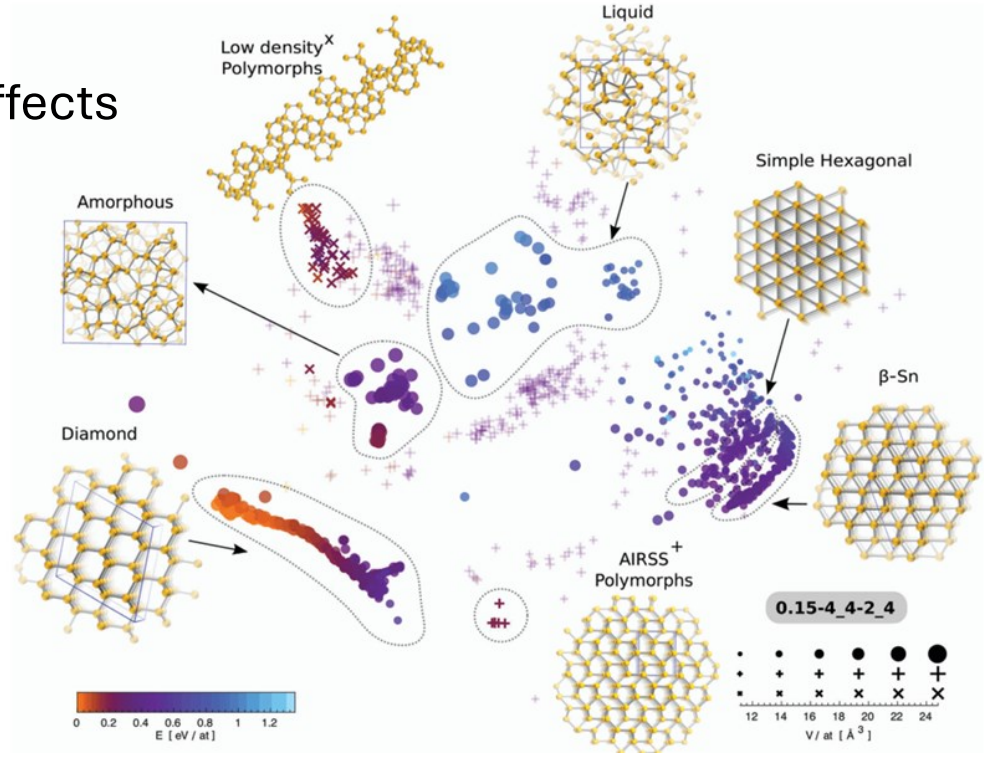


Query Found Remaining

VRAGFN retrieval is built upon works in Optimal Transport

- REMATCH kernel (Csanyi, Ceriotti ‘16)
  - Node-to-node soft-matching → ↓ averaging effects
  - Entropy-regularized → ↓ overconfidence
  - Similarity metric as a probability.



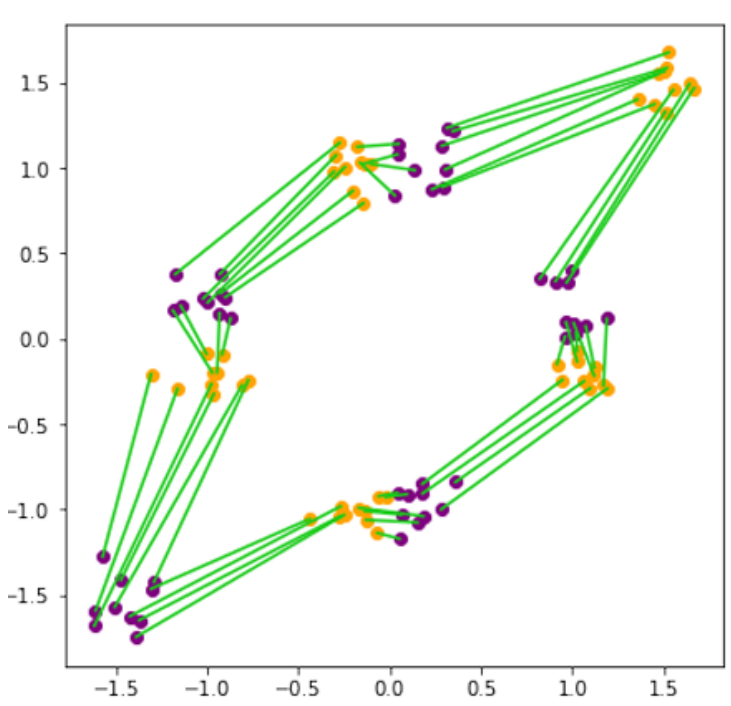


Sandip De, Albert P. Bartók, Gábor Csányi and Michele Ceriotti, Comparing molecules and solids across structural and alchemical space, Phys. Chem. Chem. Phys., 2016,18, 13754-13769

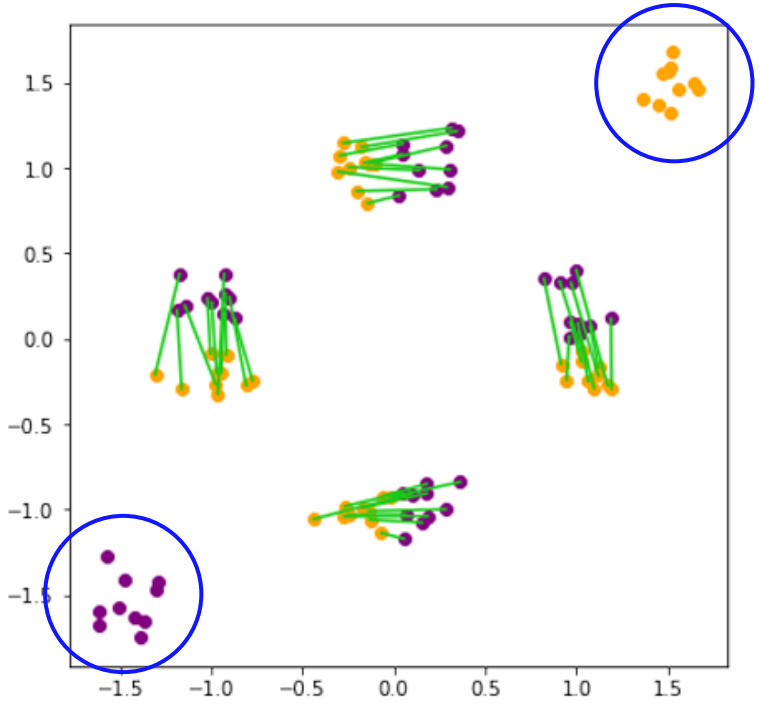
VRAGFN adopts a variant in Unbalanced OT

- Unbalanced OT (Séjourné-Peyré-Vialard ‘22)

It has a “No-Match” kernel.
(c.f. any softmax e.g. transformer)



(a) OT matching



(b) Unbalanced OT matching

Thibault Séjourné, Gabriel Peyré, François-Xavier Vialard, Handbook of Numerical Analysis, Vol 24, p.407-471

Retrieval Mechanism in VRAGFN

- U-REMATCH kernel.
  - Unbalanced Optimal Transport
- Probability mass deflation.
  - Take away the matched

Algorithm 2 Mass deflated transport

Require: Initial source mass $a^{(1)} \in \mathbb{R}_+^n$ (e.g. uniform, $\sum_i a_i^{(1)} = 1$), substructures $\{B_k\}_{k=1}^K$ with target masses $b^{(k)}$, local similarity matrices $C^{(k)} \in \mathbb{R}^{n \times m_k}$, parameters $\alpha > 0$, $\tau_r, \tau_c > 0$

for $k = 1, \ldots, K$ **do**

Solve unbalanced OT:

$$P^{(k)} = \arg\min_{P \geq 0} \sum_{i,j} P_{ij} \left[\frac{1 - C_{ij}^{(k)}}{\alpha} + \ln P_{ij} \right] + \tau_r \, D(P\,\mathbf{1} \| a^{(k)}) + \tau_c \, D(P^T \mathbf{1} \| b^{(k)})$$

Compute u-REMATCH kernel:

$$S_k = \sum_{i,j} P_{ij}^{(k)} \, C_{ij}^{(k)}$$

(Unpublished)

$a^{(k+1)} = a^{(k)} - P^{(k)}\,\mathbf{1}$ ▷ Mass deflation

if $\|a^{(k+1)}\|_1 = 0$ **then**

break ▷ No mass left to match

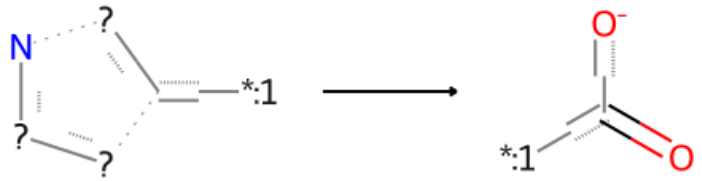
return P, C, S

VRAGFN is augmented with Isosteric Data

6.3 million products
xTB partial charges

Catalog

- Strict Enamine X-REAL Product/Synthon
  - ➢ 423 isosteric transforms in SMARTS
  - ➢ ~16,000 valid synthon clusters (~38%)
  - ➢ Matched molecular pair (Leach '06)



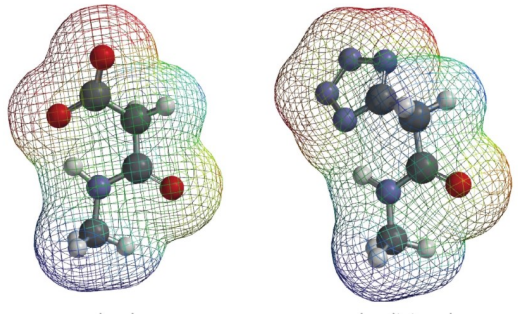

```
[*:1][#6H0]=,:1[#7H0,#6H1]=,:[#7H0][#7H1,#
8H0,#16H0][#7H0,12#6H1]=,:1>>>[*:1]C(=O)[O-]
```

Augmentation ("Playbook")

- Apply transforms randomly on X-REAL product lacking isostere
  - ➢ Just for training. Not in inference
  - ➢ Generalize the "interfacing" isosteres e.g. amide.

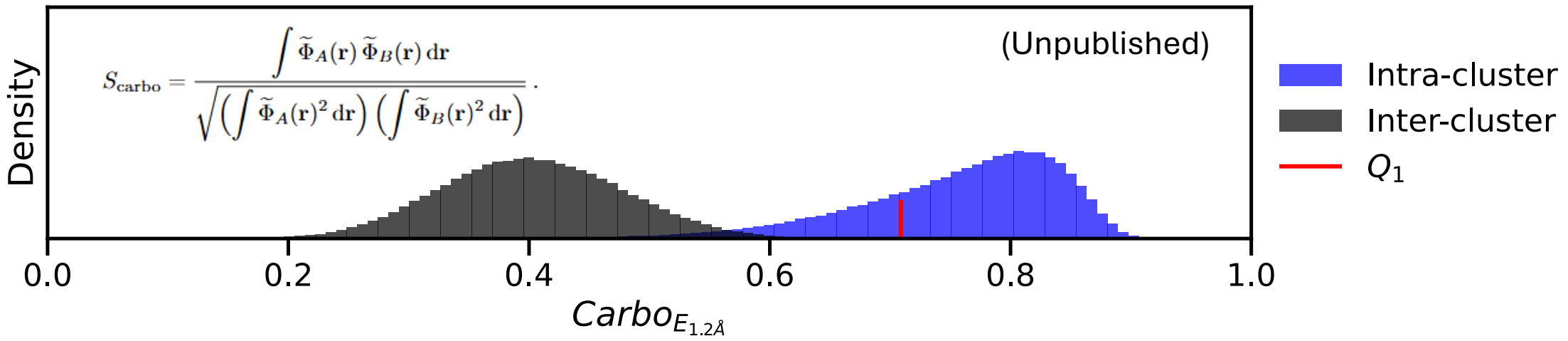
(Unpublished)

The Quality of Data



$$O_{A,B}^{\mathrm{ESP}} = \sum_{a \in \boldsymbol{Q}_A} \sum_{b \in \boldsymbol{Q}_B} \left(\frac{\pi}{2\alpha}\right)^{\frac{3}{2}} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{r}_a - \boldsymbol{r}_b\|^2\right) \exp\left(-\frac{\|\boldsymbol{v}_A[a] - \boldsymbol{v}_B[b]\|^2}{\lambda}\right)$$

- Good-Carbo ‘92 charge overlap
  - Clear separation

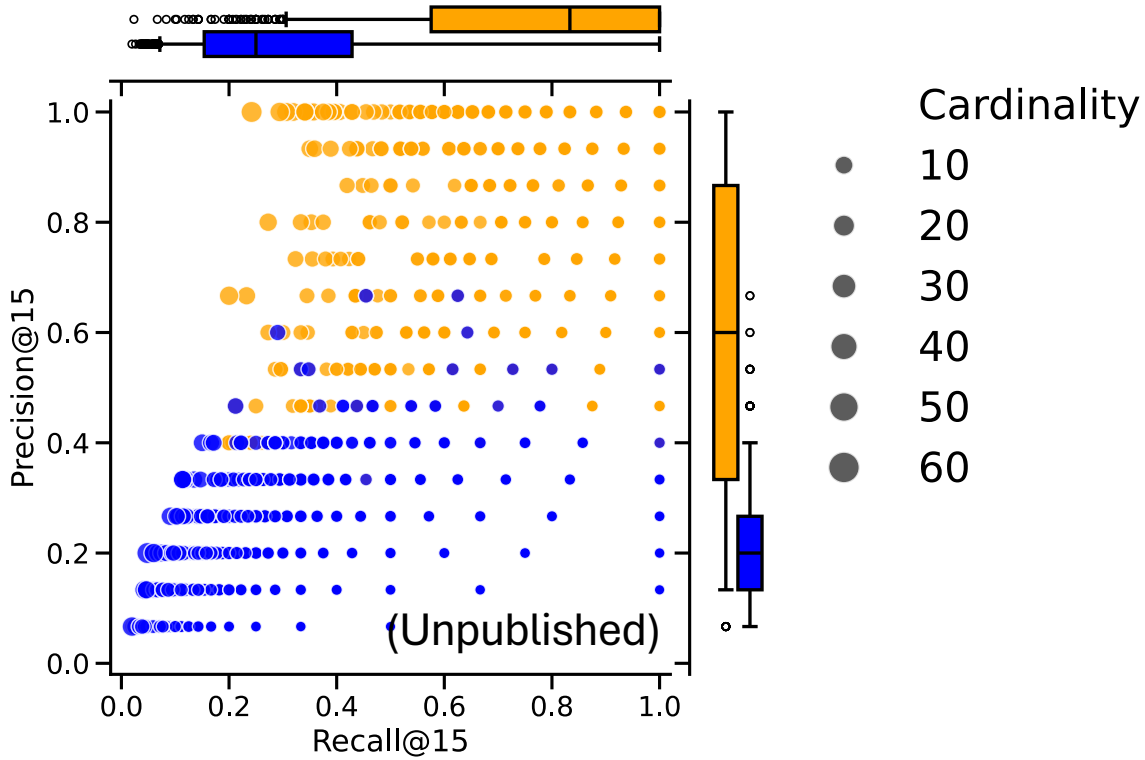


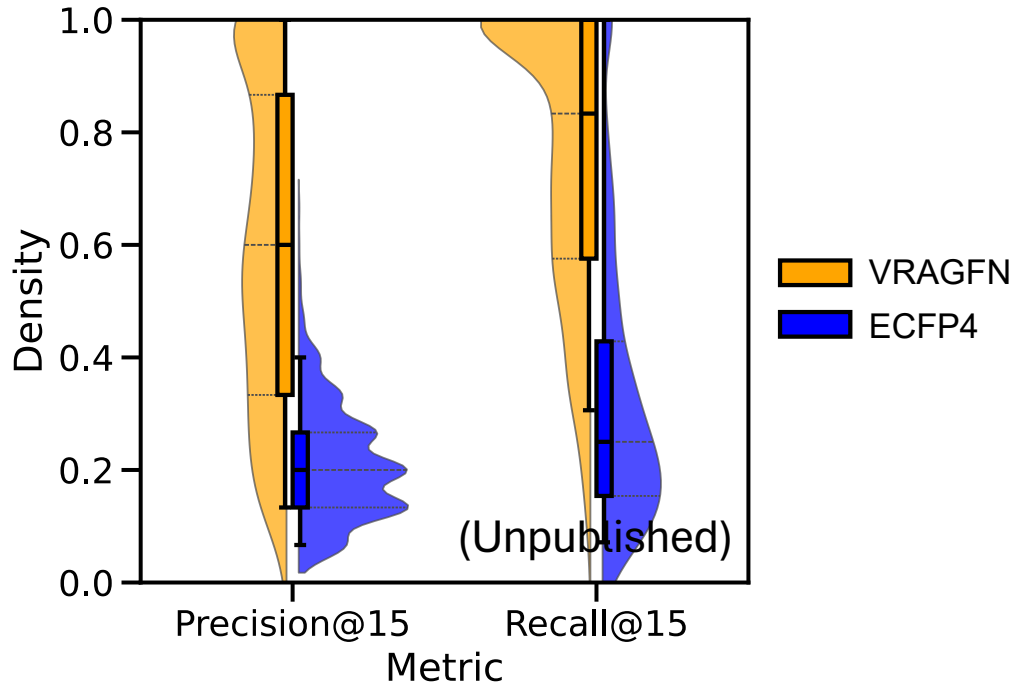
A.C. Good, The calculation of molecular similarity, Journal of Molecular Graphics, Volume 10, Issue 3,1992, Pages 144-151

VRAGFN is effective in 15-NN search

$$Recall@k =: \frac{y^T \hat{y}}{1^T y}$$

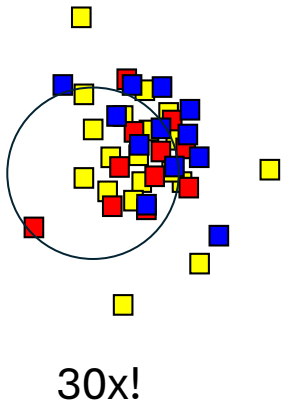
$$Precision@k =: \frac{y^T \hat{y}}{k}$$



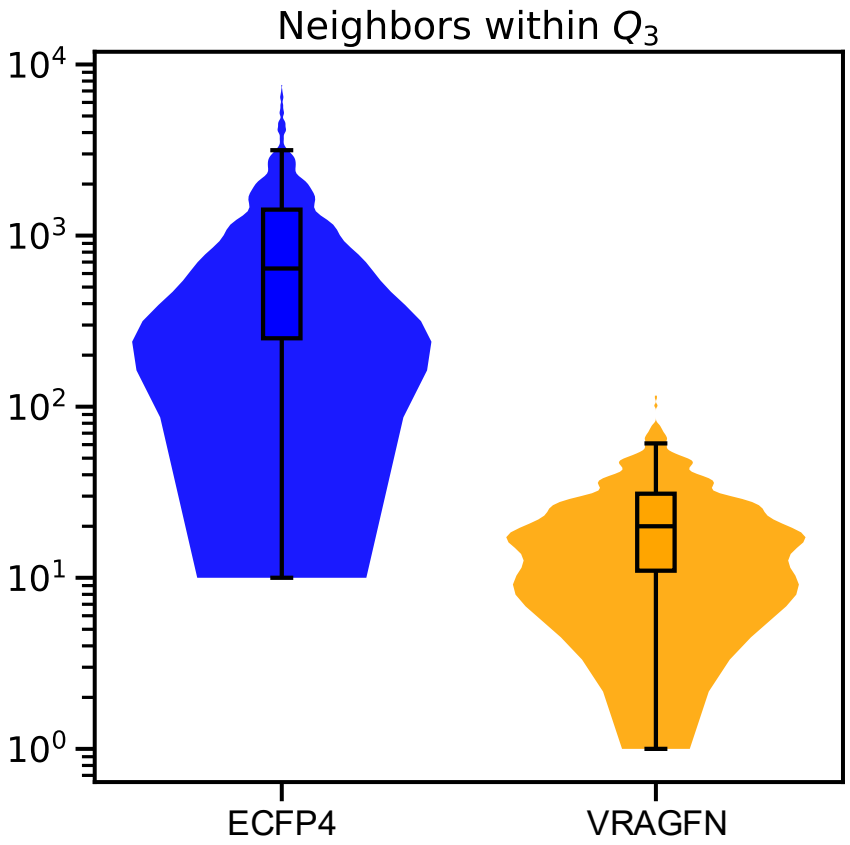


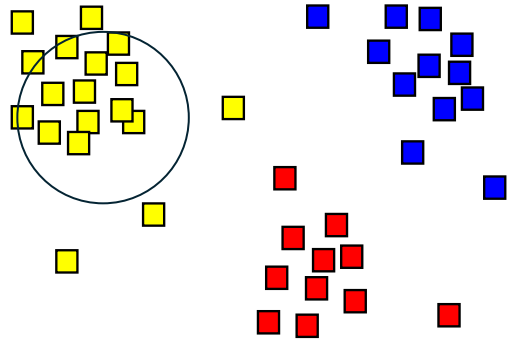
VRAGFN is effective in Radius Search

- While Q_3 guarantees 75% isosteres, ECFP4 gives ‘fat’ search results.



30x!





Technical Milestones

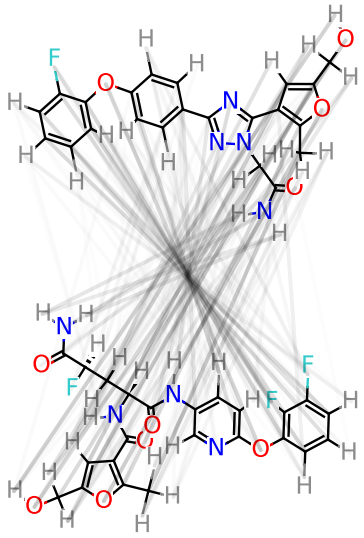
VRAGFN enables

1. Physics-based 3D alignment
2. Matches at decomposition boundaries
3. Cross Library Search

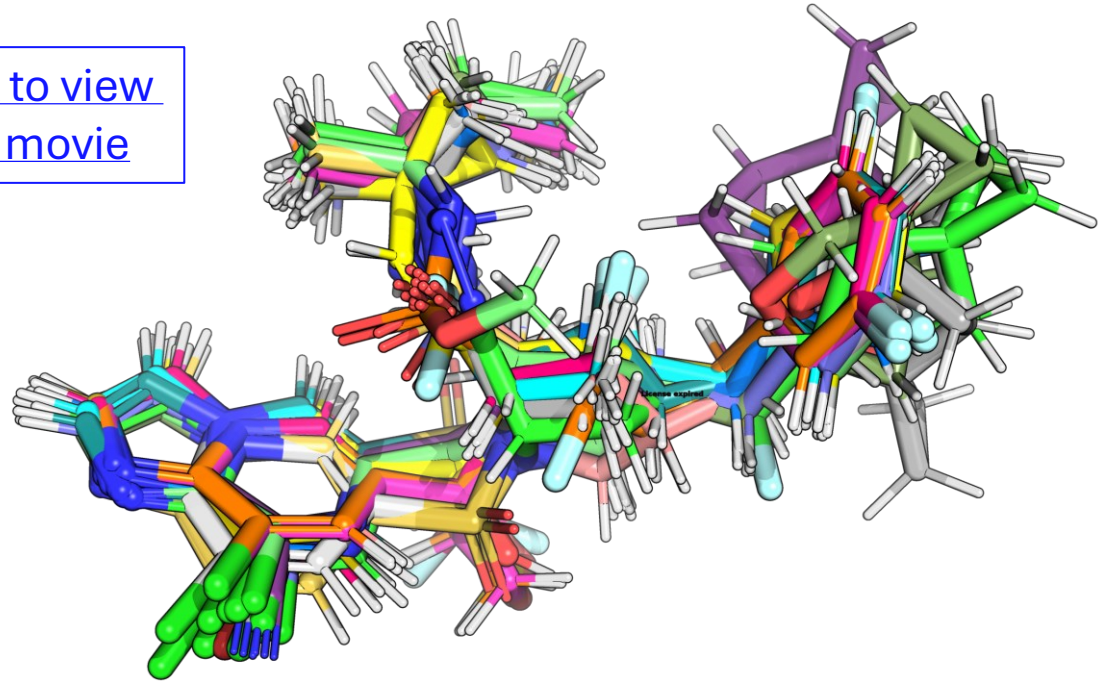
These are new features unencountered by other similarity search engines

VRAGFN enables physics-based 3D alignment

- Matching links (if any!) is induced by the UOT mechanism in 3D

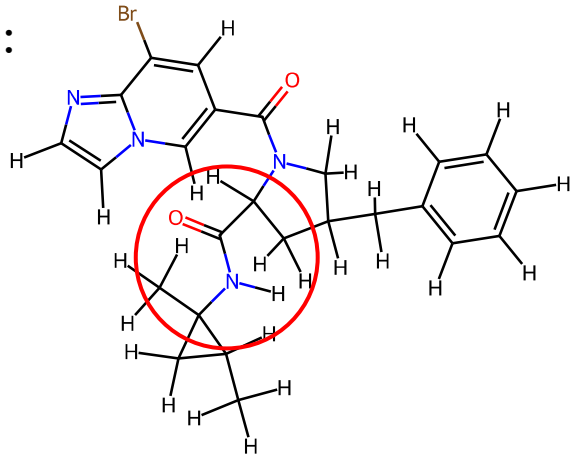


Click to view the movie

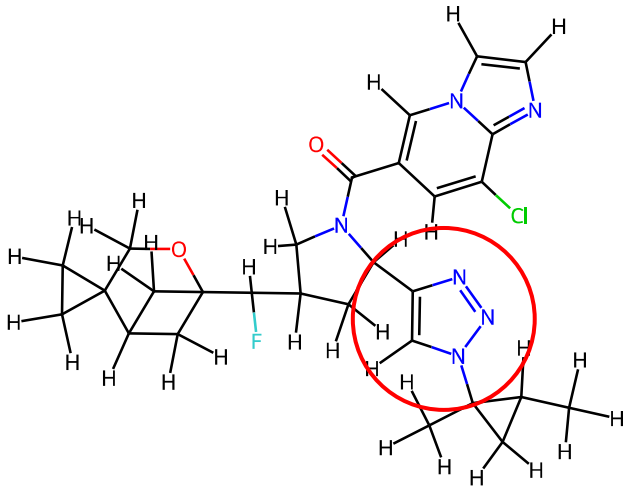


VRAGFN enables matches at decomposition boundary

Q:

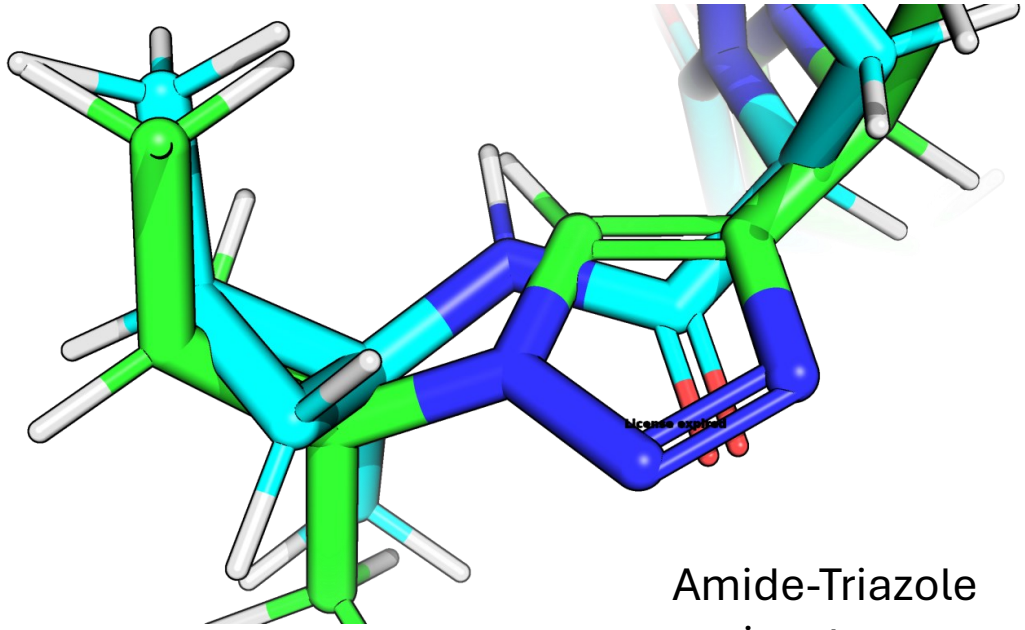


A_5:



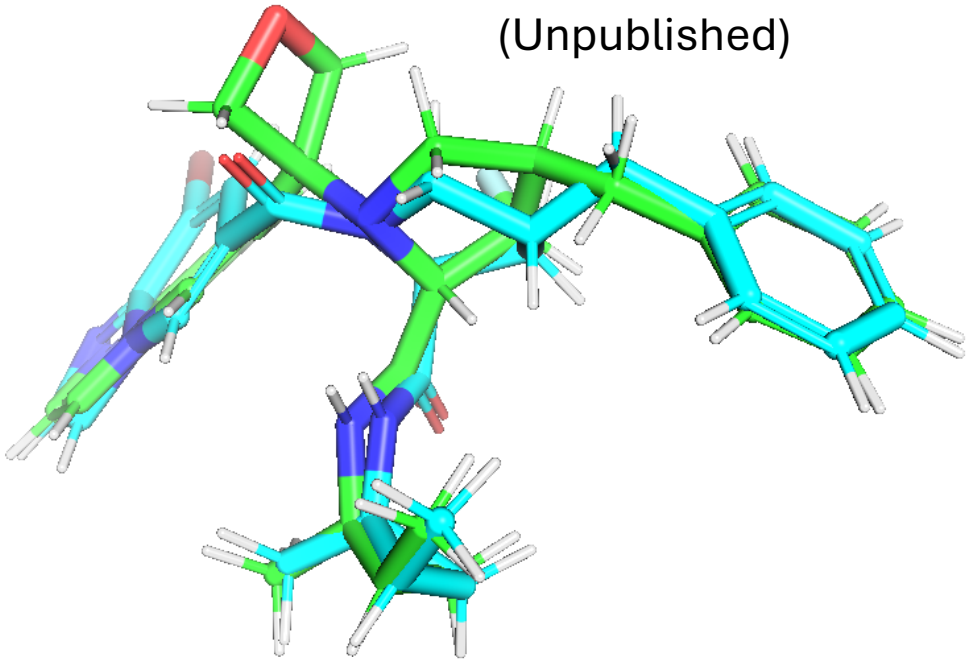
Divide-Conquer-Combine

- Similarity sits right at the bond

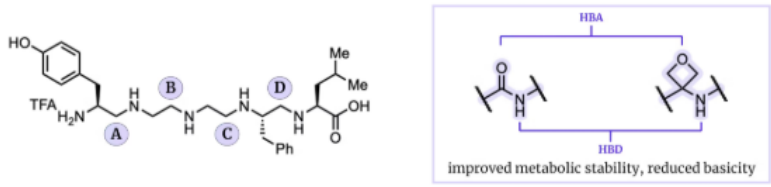


Amide-Triazole isostere

Usage: Scaffold Replacement



(Unpublished)

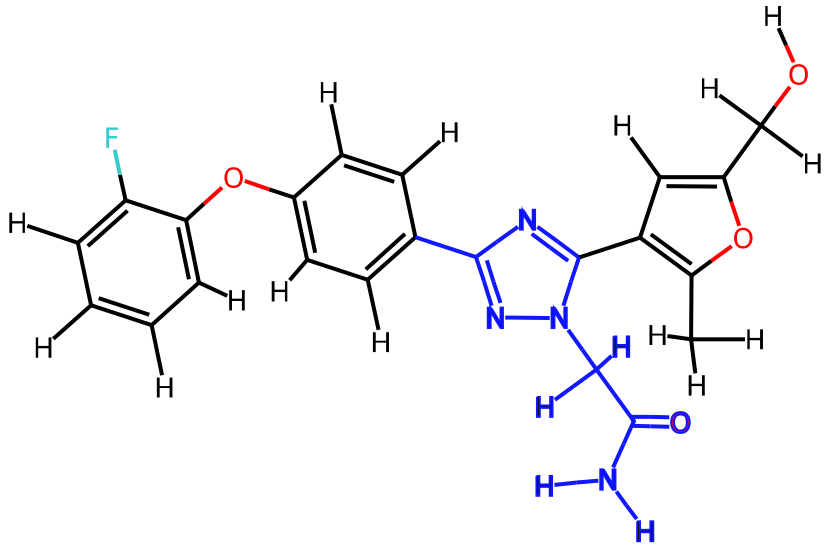


compound	A	B	C	D	Serum ($t_{1/2}$)	δ K_i (nM)
leu-enkephalin	C=O	C=O	C=O	C=O	> 10 min	9.2
compound 20		C=O	C=O	C=O	> 3.2 h	> 1000
compound 21	C=O		C=O	C=O	> 18 h	> 1000
compound 22	C=O	C=O		C=O	> 15 min	157
compound 23	C=O	C=O	C=O		> 26 min	43

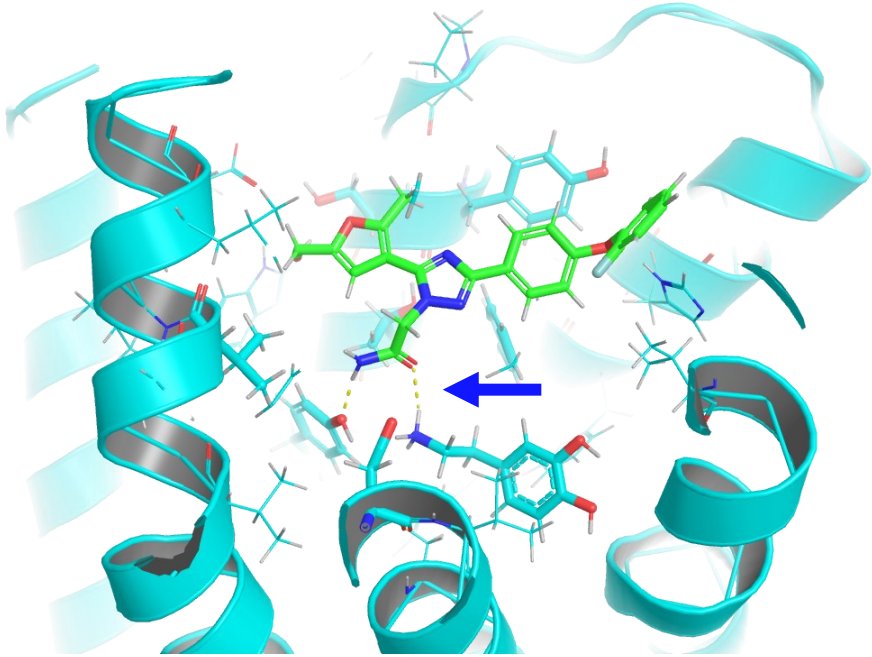
Image Credit (right): https://drughunter.com/resource/bioisosteres-for-drug-hunters-part-1-background-carboxylic-acids-and-amides

VRAGFN enables Cross-Library search

- Input: a 2-comp hit from REAL 2021 containing a retired synthon

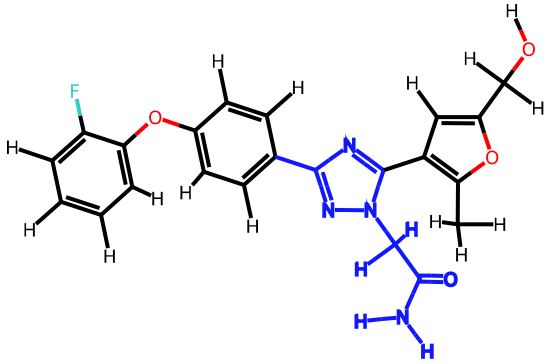


Retired but important

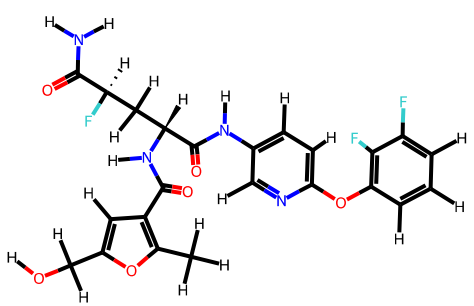


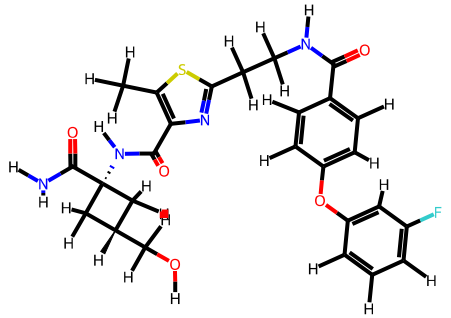
VRAGFN enables Cross-Library search

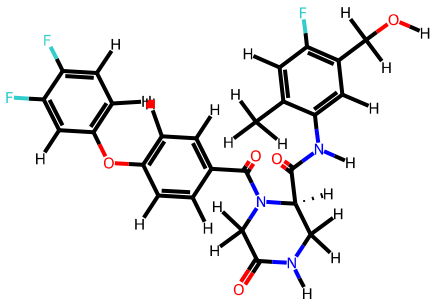
- Input: a 2-comp hit from REAL 2021 containing a retired synthon
- Output: a 3-comp(!) molecule from X-REAL 2024



Emeritus



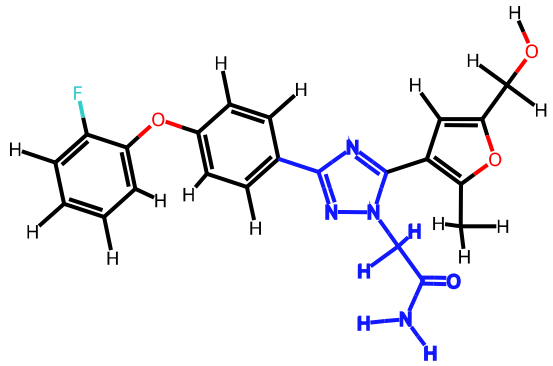




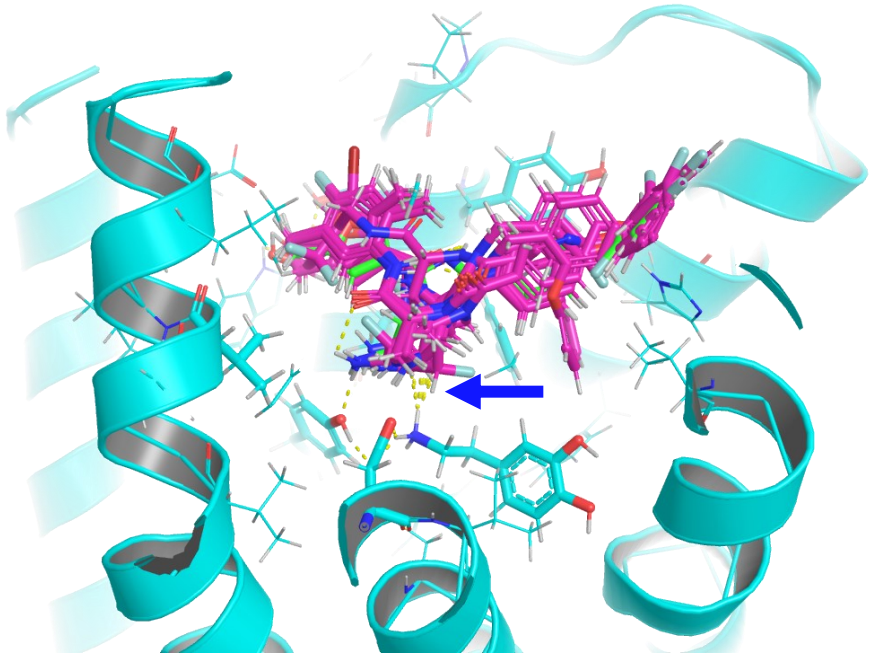
Emeritus part
mimicked
in 3-comp library

VRAGFN enables Cross-Library search

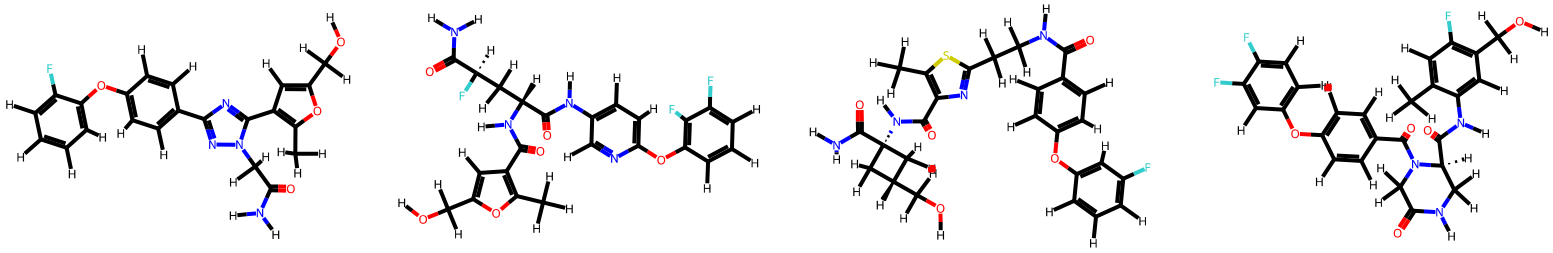
- Input: a 2-comp hit from REAL 2021 containing a retired synthon
- Output: a 3-comp(!) molecule from X-REAL 2024



Emeritus



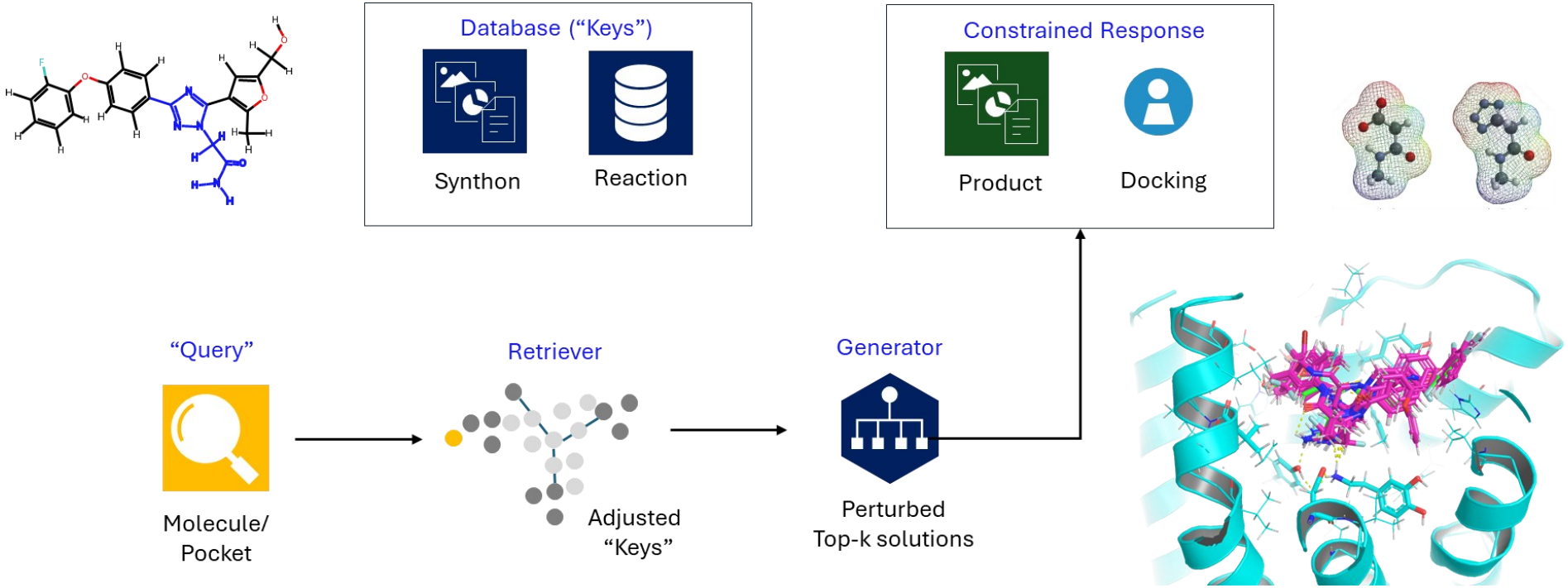
VRAGFN Leads



	Query	Lead 1	Lead 2	Lead 3
Docking Score	-35	-38	-37	-37
MolWeight	422	513	512	439
LogP	2.8	1.7	2.7	2.4
LogS	-3.5	-2.8	-3.5	-3.4

Conclusion

- VRAGFN brings new functionality to similarity search engine



Acknowledgement

Compute

- NVIDIA Academic Grant Program

Collaborators

- Prof. Aiichiro Nakano and Prof. Yan Liu

Fundings and Fellowship

Croucher Foundation
Hong Kong

American Chemistry Society

National Institutes of Health

National Institute of
General Medical Sciences

Nvidia Corp

The Katritch Lab
Department of Quantitative & Computational Biology
Department of Chemistry

USC Dornsife
Department of Quantitative and Computational Biology