New Methods for Large Chemical Databases:

Giga-Search: Fast Substructure Search
RIDE: Fast Ligand based screening

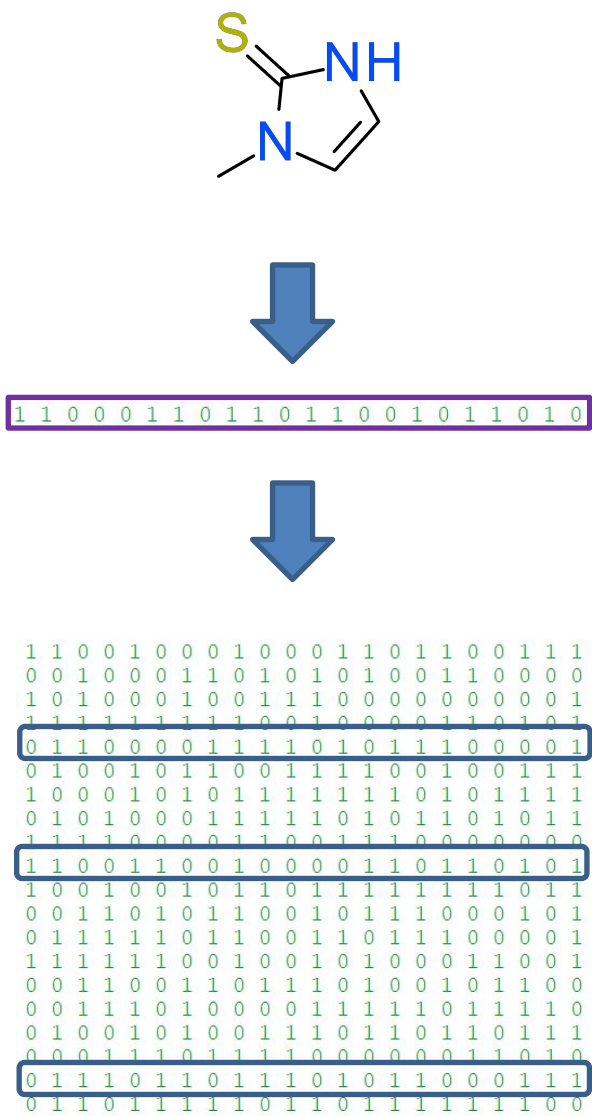
Eugene Raush
Molsoft LLC

Giga-Search

Fast Chemical Substructure Search

“Traditional” approach to substructure searching

- Pre-calculated **column with fingerprints** in the database
  - Each bit in the fingerprint corresponds to some sub-fragment in the molecule.
  - Search goes first through this column and checks if pattern fingerprint is subset of fingerprint from the database
  - For matched rows search performs actual substructure search
- Works reasonably fast on DB size < 10M
  - Most searches complete in in seconds
- Same fingerprint column can be used for similarity search (Tanimoto distance)
- Implemented in **Molcart** as chemical cartridge
  - Set of chemical functions to be used within SQL query: http://molsoft.com/molcart.html



molsoft molecules in silico

New approach for searching in large space

- "Traditional search" on such size may take **minutes to hours** to get the result
  - The bottleneck is the **amount of data** (fingerprints) to scan through (~200G for 1B compounds, for 1536 bits FP)
- The new approach is to:
  - Dynamically find optimal subset of bits for an input chemical pattern/query
  - Design new fingerprint storage to efficiently read only subset of bits

molsoft
molecules in silico

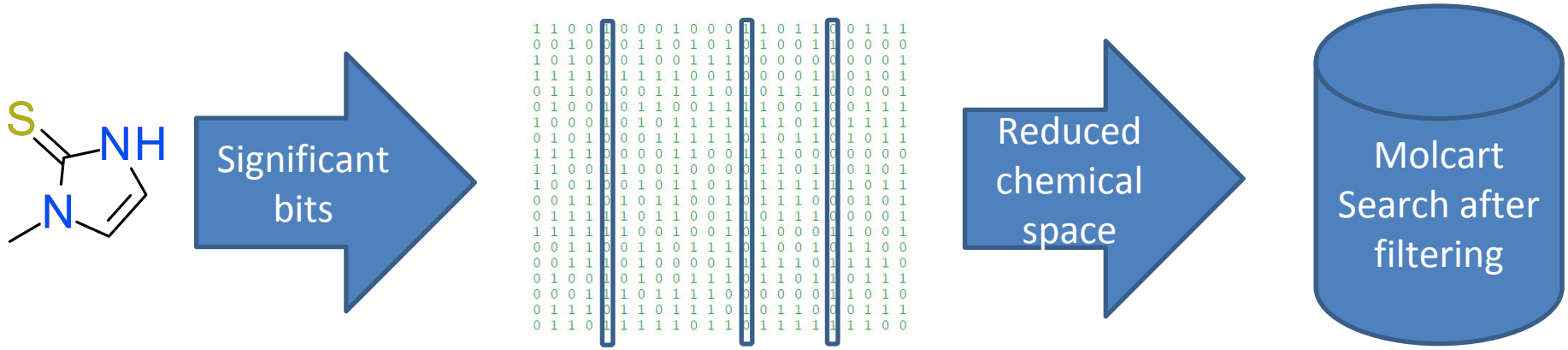
Finding significant bits of a fingerprint

- Significant bits can be found using pre calculated fingerprint statistics for a certain database
- This information can be further used to more efficiently filter the database without scanning the whole table of fingerprints
- We did some experiments and found that for average chemical substructure **set of ~10 significant bits** is enough to significantly reduce chemical search space
  - That means that we reading less than 1% of FP data.

molsoft molecules in silico

Giga-Search method workflow

1. Pre-filtering
   - Find optimal bits subset for a given pattern
   - Reduce chemical space (list of row IDs)
   - This stage takes around 2 sec for 1B DB size for any pattern
   - Reduced space is usually less than 0.5% of the total DB
2. Apply “traditional” search for reduced chemical space to find actual hits

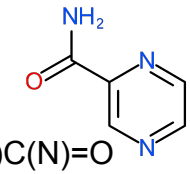
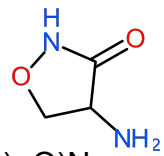
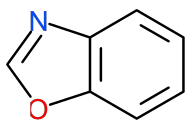
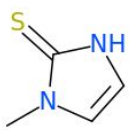
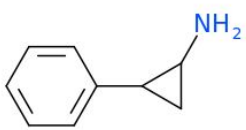


molsoft molecules in silico

Databases used in Giga-search benchmark

- **Enamine REAL** 2020
  - Size: 1.3B of structures
  - Synthesis success rate normally attains 85% and higher
  - More info from Enamine website
    - https://enamine.net/library-synthesis/real-compounds/real-database
- **Synthetically Accessible Virtual Inventory: SAVI**-2020
  - "Plus Class" only, size 1.1B (62.61% of total SAVI products)
  - Proposed products with reactions generated in the first full enumeration of the SAVI project
    - Patel, H., Ihlenfeldt, W. D., Judson, P. N., Moroz, Y., Pevzner, Y., Peach, M., Tarasova, N., & Nicklaus, M. (2020). Synthetically Accessible Virtual Inventory (SAVI) (Version 2020). CADD Group, CBL, CCR, NCI, NIH
    - https://cactus.nci.nih.gov/download/savi_download/

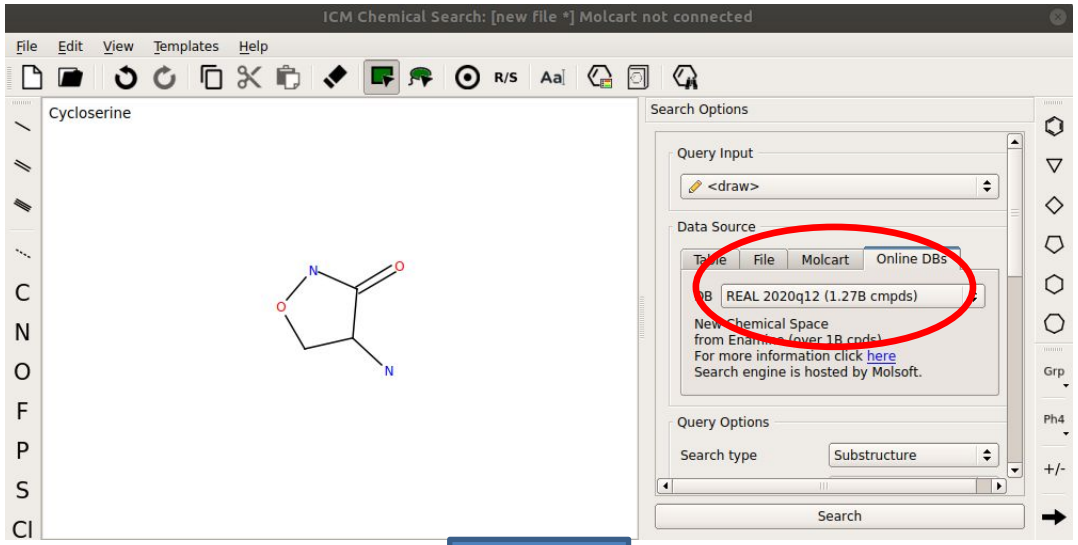
molsoft molecules in silico

Substructure	REAL (1.3B) # hits after pre-filtering # of actual hits	SAVI (1.1B) # hits after pre-filtering # of actual hits	Time to get first 1000 hits
 c1cnc(cn1)C(N)=O	5,934,684 (0.4%) 5,317,178	2,105,092 (0.19%) 1,429,140	3.9 sec
 C1C(C(NO1)=O)N	664,654 (0.05%) 61,841	94,569 (0.008%) 93,052	3.78 sec
 c1ccc2c(c1)nco2	3,976,593 (0.3%) 3,847,767	5,867,937 (0.5%) 5,648,155	4.21 sec
 CN1C=CNC1=S	40,516 (0.002%) 40,004	42,543 (0.003%) 41,592	3.66 sec
 C1C(C1N)c1ccccc1	2,752,329 (0.2%) 2,308,662	2,416,947 (0.2%) 1,979,782	3.94 sec

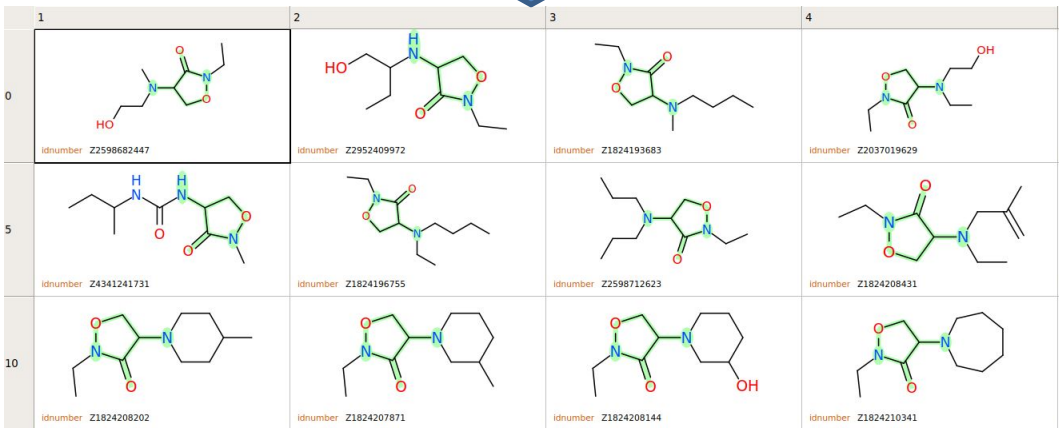
molsoft
molecules in silico

Giga-Search Implementation

- Online service (ICM script) to provide substructure search
  - GUI interface from ICM-Pro or ICM-Chemist
- For Molcart customers the service can be hosted locally on custom database



Giga Search



molsoft molecules in silico

Giga-Search: Hardware Requirements

- Any modern CPU: i7, i9, Xeon
- RAM: >16Gb
- Storage: Fast SCSI HDD or SSD (preferred)

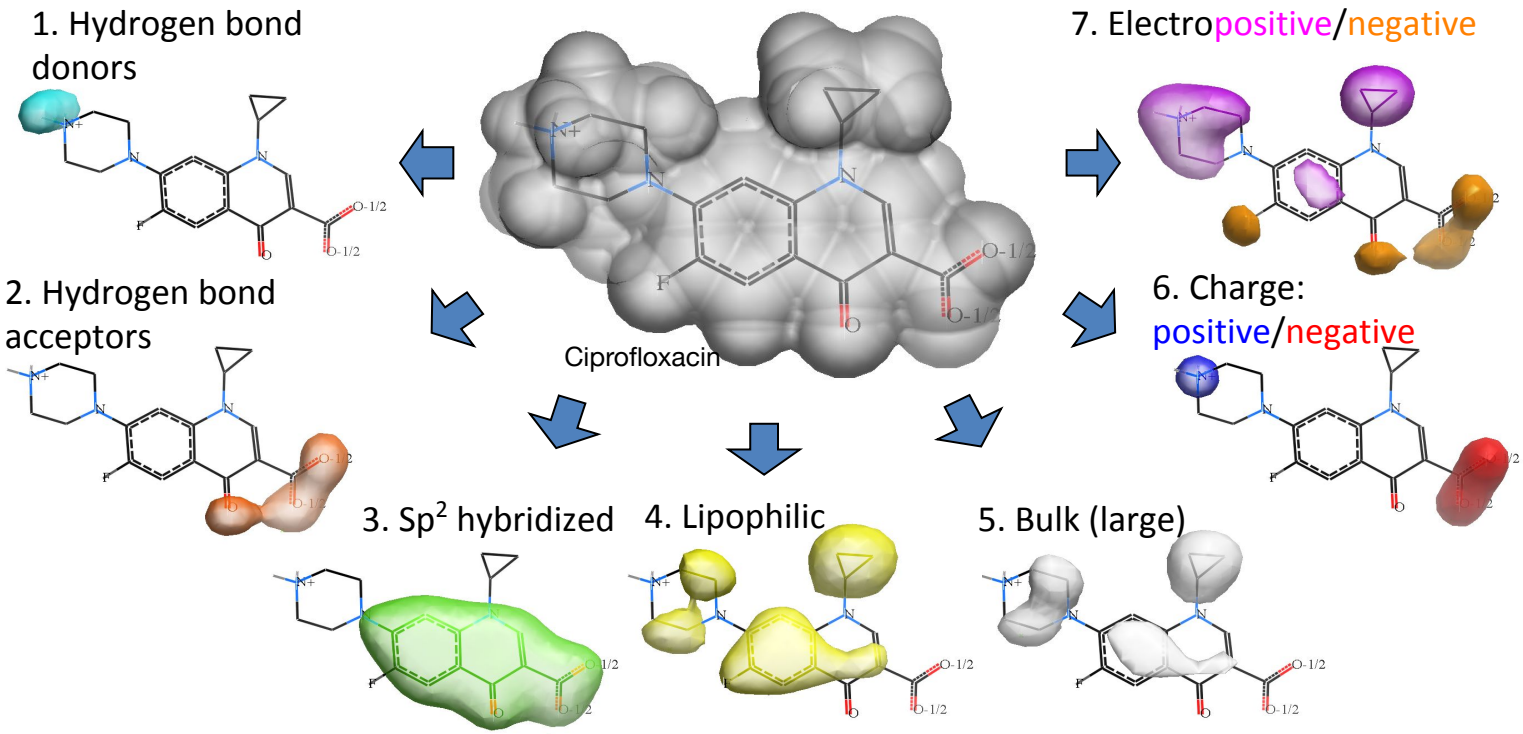
Rapid Isostere Discovery Engine (RIDE)

Fast Ligand Based Screening of Large Databases

molsoft
molecules in silico

Atomic Property Fields (APF)

- APF is a 3D pharmacophoric potential implemented on a continuously distributed grid which can be used for ligand docking and scoring
  - A single ligand atom can contribute to multiple fields
  - Multiple similar ligand atoms in a spatially consistent location result in a strong pharmacophore signal for their features in this location
- APF covariance/score can be considered as a 3D similarity and alignment quality measure



Totrov M. Chem Biol Drug Des. 2008 Jan;71(1):15-27 (https://www.molsoft.com/apf.html)

molsoft molecules in silico

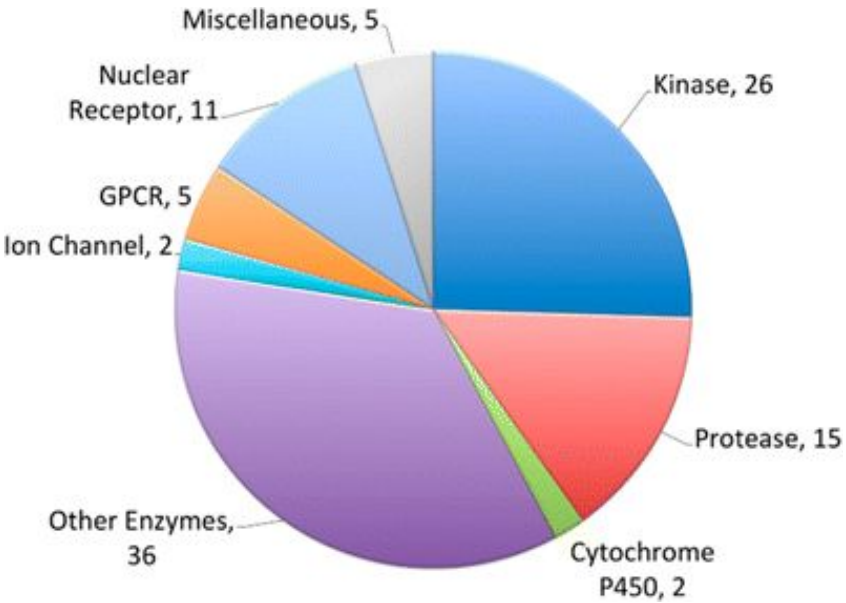
APF applications

- **Chemical superposition**
- **Ligand guided Virtual Screening**
- Multiple chemical alignment
- Scaffold Hopping
- APF-ARM (activity regression model)
- Chemical clustering by pose
- Ligand binding site superposition and comparison

DUDe benchmark

A Database of Useful Decoys: Enhanced

- 102 target proteins
- 22,886 active compounds
  - an average of 224 ligands per target
- 50 decoys for each active
  - similar physico-chemical properties
  - dissimilar 2-D topology
- 1 ligand from X-ray per target as a query for ligand-based screening



Mysinger MM, Carchia M, Irwin JJ, Shoichet BK J. Med. Chem., 2012, Jul 5

molsoft molecules in silico

Accuracy of Flexible Ligand Superposition

- Independent broad benchmark: ligands without X-ray structures but similar chemotype to a solved complex.
- Assessment of superposition quality: 2-good/1-acceptable/0-poor
  - 11 targets from DUDe
- Flexible APF **outperforms** other methods but it is relatively **slow**.

		ADA (39)	CDK2 (72)	DHFR (410)	ER (39)	FXA (146)	HIVRT (43)	NA (49)	P38 (454)	THR (72)	TK (22)	TRP (49)	mean (1100)
	2	12.82	12.5	44.39	56.41	4.11	18.6	18.37	9.69	4.17	68.18	40.82	23.15
	1	35.9	51.39	53.66	43.59	33.56	72.09	75.51	69.6	93.06	31.82	59.18	59.07
	0	51.28	36.11	1.95	0	62.33	9.3	6.12	20.7	2.78	0	0	17.78
	2	12.82	43.06	74.15	41.03	14.38	30.23	79.59	9.47	2.78	86.36	8.16	35.63
	1	20.51	36.11	14.39	56.41	28.77	34.88	14.29	41.19	69.44	9.09	81.63	32.83
	0	66.67	20.83	11.46	2.56	56.85	34.88	6.12	49.34	27.78	4.55	10.2	31.54
	2	15.38	25	56.1	48.72	35.62	16.28	36.73	14.98	30.56	81.82	18.37	33.48
	1	20.51	19.44	11.71	43.59	13.7	46.51	57.14	74.01	5.56	13.64	2.04	35.77
	0	64.1	55.56	32.2	7.69	50.68	37.21	6.12	11.01	63.89	4.55	79.59	30.75
ICM/APF	2	46.15	12.5	86.83	51.28	70.55	18.6	75.51	20.04	88.89	90.91	69.39	54.48
	1	23.08	68.06	11.95	46.15	16.44	46.51	14.29	68.28	9.72	9.09	28.57	36.49
	0	30.77	19.44	1.22	2.56	13.01	34.88	10.2	11.67	1.39	0	2.04	9.03

Giganti et al. J Chem Inf Model 2010, 50, 992-1004

molsoft molecules in silico

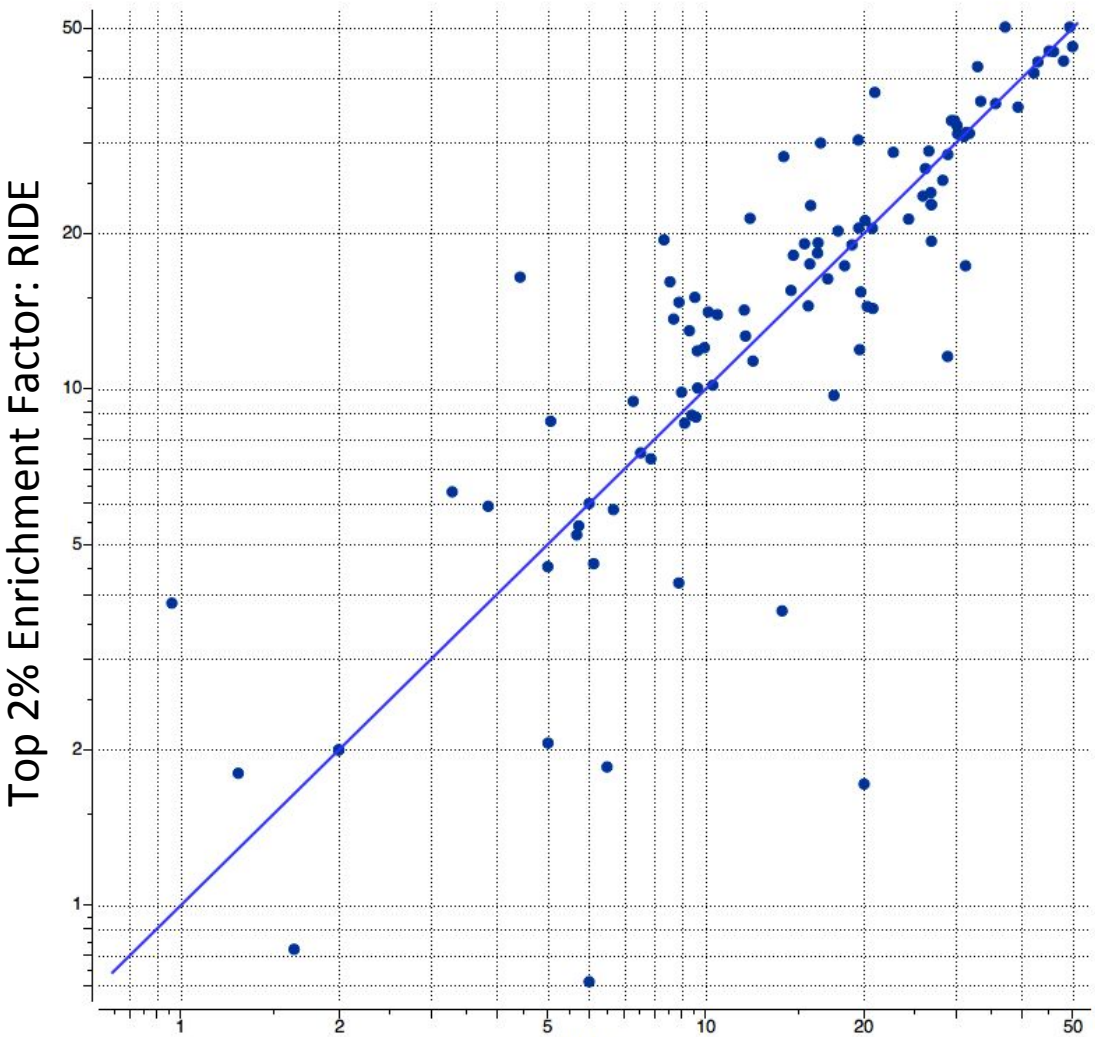
Fast APF screening: Rapid Isostere Discovery Engine (RIDE)

- Use pre-generated conformers, **rigid body APF** superposition only optimizes translations/rotations
- **New systematic search algorithm** designed for speed from the ground up:
  - high level heuristics focusing search on the most relevant space
  - maximized use of pre-computed lookups
  - most compact data storage
  - concurrent evaluation of many poses
- Much faster even on CPU, especially multi-core
- GPU implementation is still orders of magnitude faster
- **How the alignment quality is comparable to flexible APF**?

molsoft
molecules in silico

RIDE versus Flexible APF screen on DUDe

- **EF** = fraction of true positives in top N% divided by fraction of actives in the input database
- **Flexible APF EF2%**
  - 18.1 mean, 15.8 median
- **RIDE EF2%**
  - 18.5 mean, 16.1 median
- RIDE superposition quality is **similar** to slower flexible method



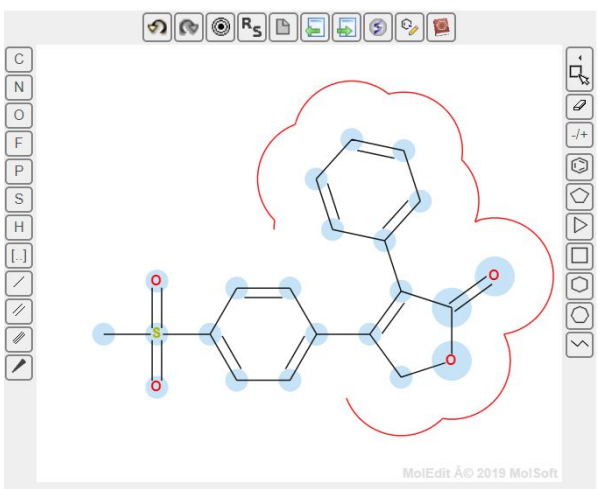
molsoft molecules in silico

RIDE implementation: Input

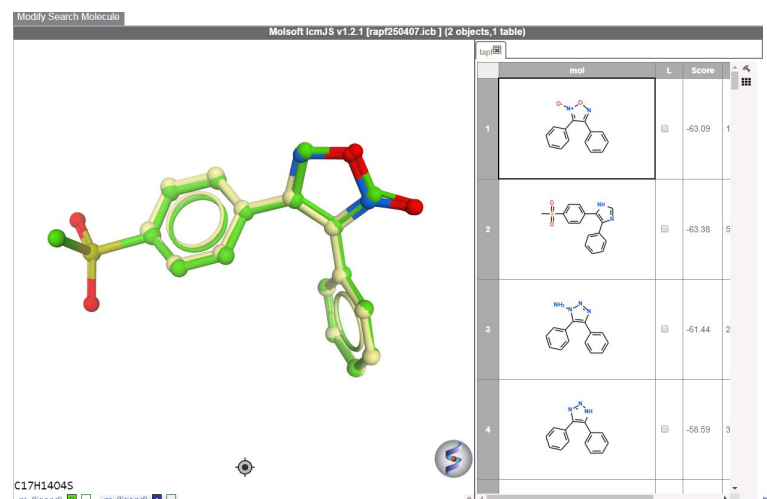
- The method can work with any set of pre-generated conformations
- For most efficient processing we designed a special DB format based on **Sqlite** engine
- Space required ~ **200 bytes** per conformation
  - ~20x time more space efficient than SDF
  - 5Gb per 1M chemicals with ~30 conformations per molecule
- Provides both efficient sequential and random (by id) access
- The database can be either generated by Molsoft conformation generator or by converting multiple external SD files

RIDE features

- **Fine tuning** of the query
  - Change **per-atom contributions** of different portions of the molecule
    - scaffold hopping, shape match, etc.
  - Apply **envelope penalty** to different portions or to the whole molecule
  - Provide receptor object to automatically calculate **excluded volume**
  - **Multiple ligands** can be used as a consensus template



RIDE

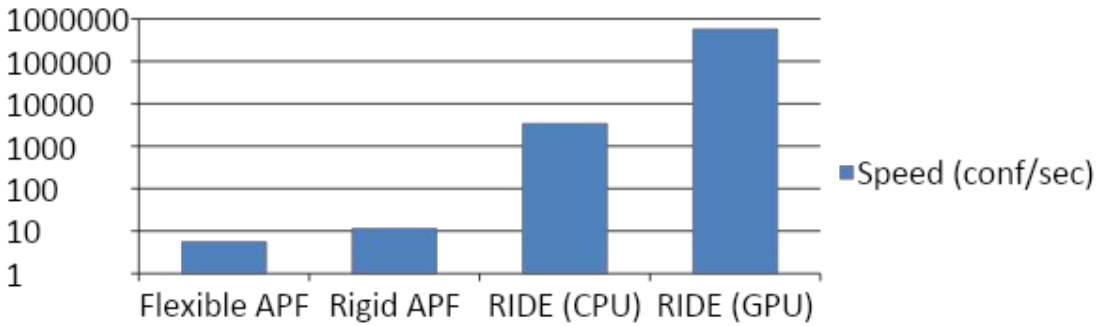


RIDE speed benchmark

Benchmark Database:

Screening collection from (1.5M compounds 30M conf.)

Method	Speed (mol/sec)	Total time
Flexible APF	0.38	44 days
Rigid APF	0.78	22 days
RIDE (CPU)	229	1.78 hours
RIDE (GPU)	38729	**48 sec**



CPU: Intel(R) i7 CPU 870 @ 2.93GHz

GPU: GeForce RTX 2080

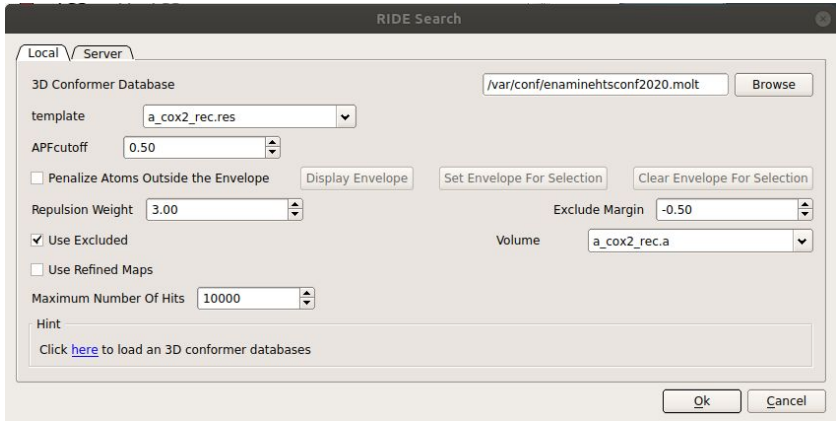
1B compounds can be screened in **8-9 hours** on a single GPU

Speed and quality of superposition make RIDE a **perfect tool** for ligand based virtual screening of large chemical databases.

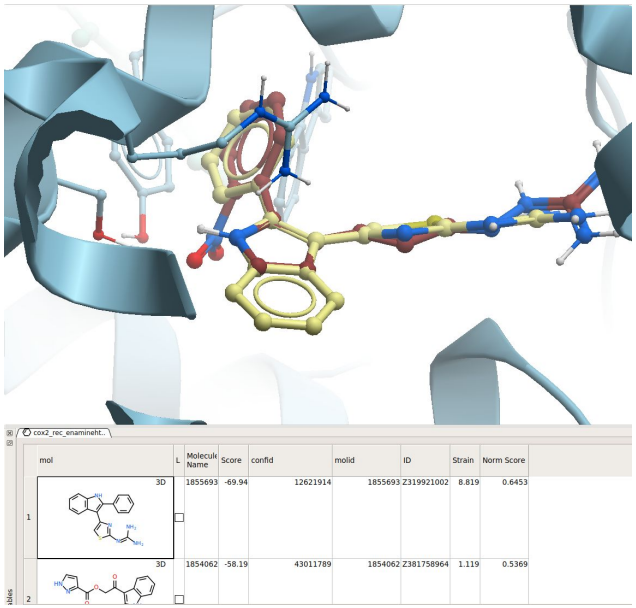
molsoft molecules in silico

RIDE Implementation

- Molsoft provides number of pregenerated conformational databases
  - Screening collections from Chemridge, Chemdiv and Enamine
  - Diverse subset from Enamine REAL (~16M)
  - MolPort and eMolecules aggregation DBs (~7.5-6.5M each)
- GUI interface from ICM-Pro or ICM-Chemist
- Docker container with RIDE service and web front-end



RIDE



molsoft molecules in silico

RIDE: Hardware Requirements

- Any modern CPU: i7, i9, Xeon with 4 or more cores
- RAM: >16Gb
- Storage: Fast SCSI HDD or SSD (preferred)
- For GPU mode:
  - Nvidia GPU with Pascal or higher architecture
  - >8Gb RAM
  - RIDE doesn't use double precision, so cheaper models like GeForce RTX (around $800) work as well as more expensive Quadro and Tesla models

molsoft
molecules in silico

Acknowledgements

Molsoft Team: Max Totrov, Ruben Abagyan, Andy Orry, Polo Lam, Arman Sahakyan

molsoft
molecules in silico