MOLVEC: Open source library for chemical structure recognition

Dac-Trung Nguyen

National Center for Advancing Translational Sciences
National Institutes of Health
ACS Meeting, San Diego, 2019

Acknowledgements

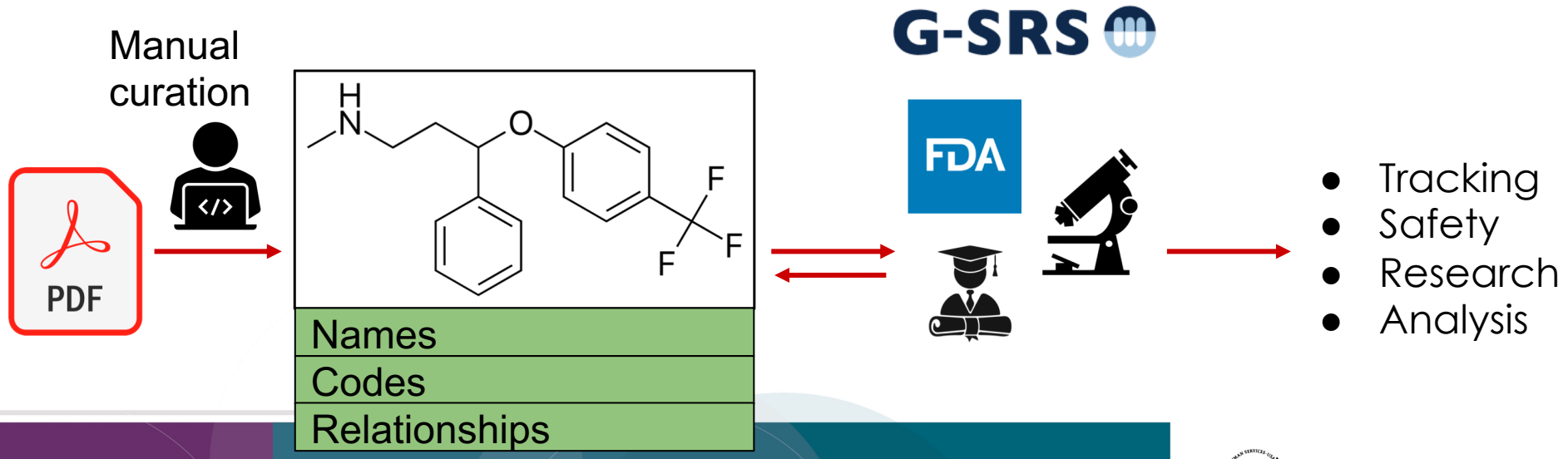
- **Tyler Peryea**
- **Daniel Katzel**
- Tongan Zhao
- Noel Southall

Outline

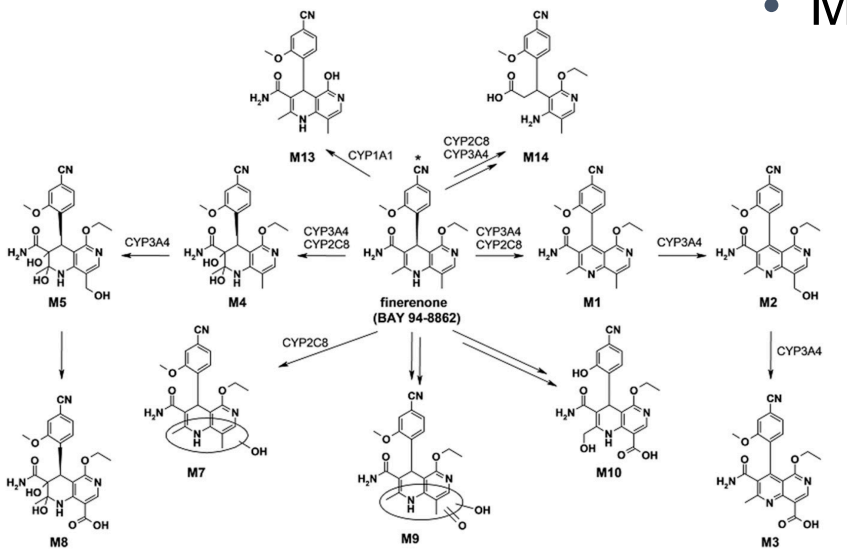
- Background & motivation
- MolVec capabilities and performance
- Examples
- Conclusion and future direction

Background

- G-SRS (Global Substance Registration System)
  - ISO 11238 reference implementation
  - Produce, consume, curate, and exchange highly-structured substance data
  - A large portion of legacy data ***is*** unstructured (e.g., APIs, impurities, metabolites, etc.) in PDF files



Background (cont'd)



- Metabolite database construction
  - In-house DMPK scientists manually extract metabolite data from literature
  - Slow, tedious, and error prone

Background (cont'd)

- A number of available options: OSRA, IMAGO, CLiDE, ChemOCR, Kekule, ChemReader, etc.
- Only OSRA and IMAGO are readily accessible
  - Complex library dependencies
  - Require significant efforts to build
  - Not easily extendable and distributable
  - Does not fit well into our Java-based stack
- MolVec is our feeble attempt at making chemical extraction from images *less* painful

MolVec *is*…

- Small, portable, and self-contained
- Pretty accurate
- Accurately pretty
- Fairly fast

Molvec *isn't*…

- A general purpose page reader engine (yet)

MolVec *is*…
Small, portable, and self-contained

- Only require Java 8 or higher
- Support standard image formats: tiff, png, gif, jpeg
- Available on Maven

  https://search.maven.org/search?q=a:molvec
- Self-contained jar file is less than 2Mb
- Source code available at

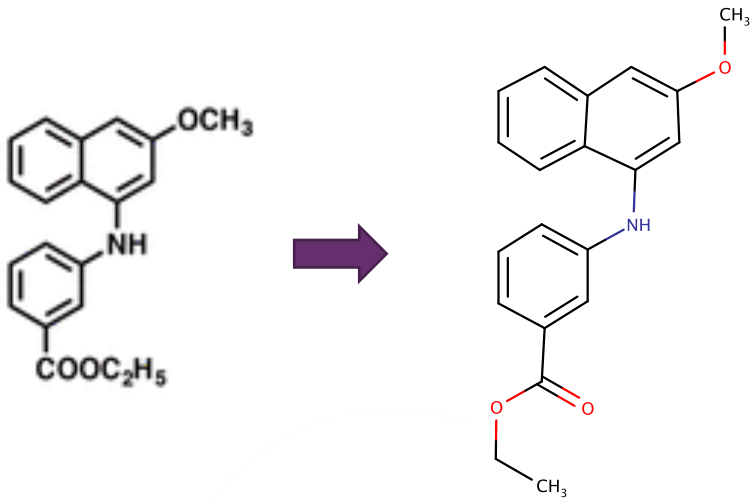
  https://spotlite.nih.gov/ncats/molvec
- Easily extendable with simple command-line interface

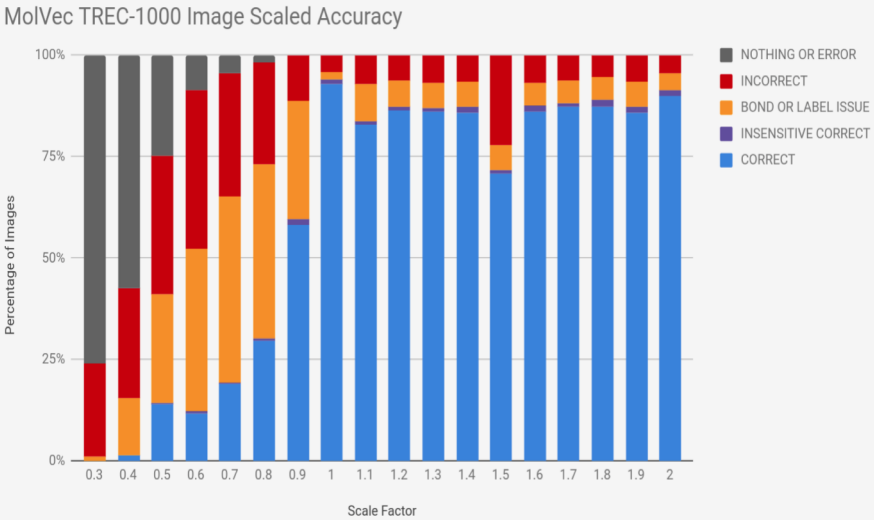
```
java -jar molvec-0.9.4-jar-with-dependencies.jar -f structures.sdf
```

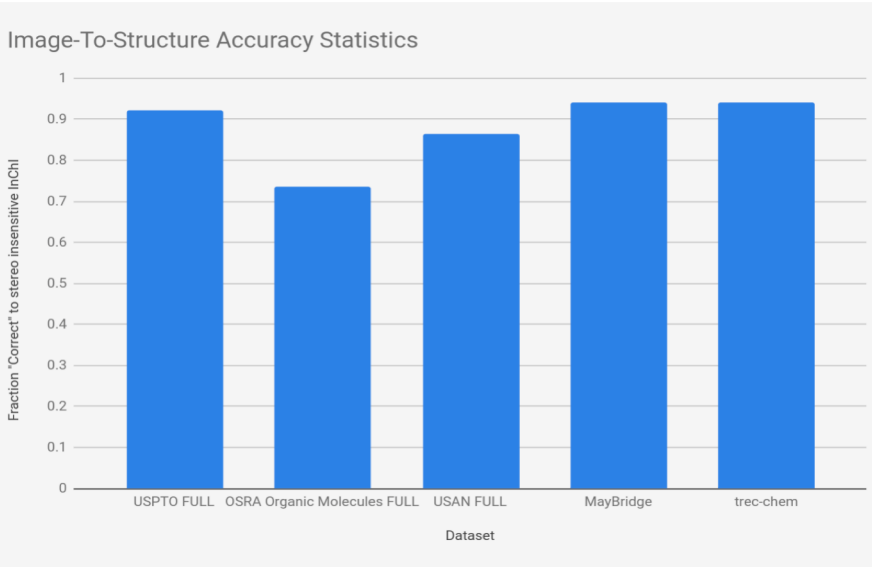
- REST API and demonstration available at https://molvec.ncats.io
- An integral part of G-SRS and is used daily within the FDA for registration of legacy substances
- Public deployments
  - https://predictor.ncats.io
  - https://ginas.ncats.nih.gov

MolVec *is*…
Pretty accurate

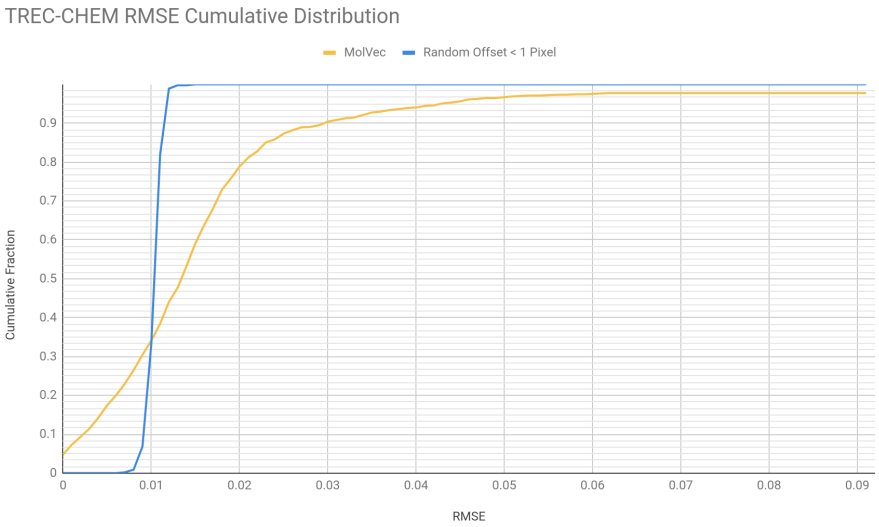
- Robust across different image resolutions
- Handle complex superatoms (S-groups)

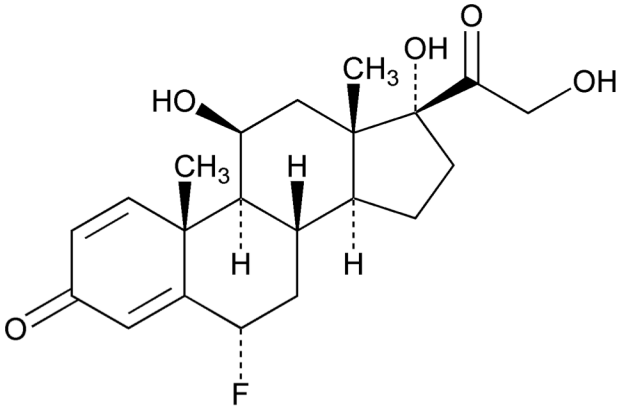


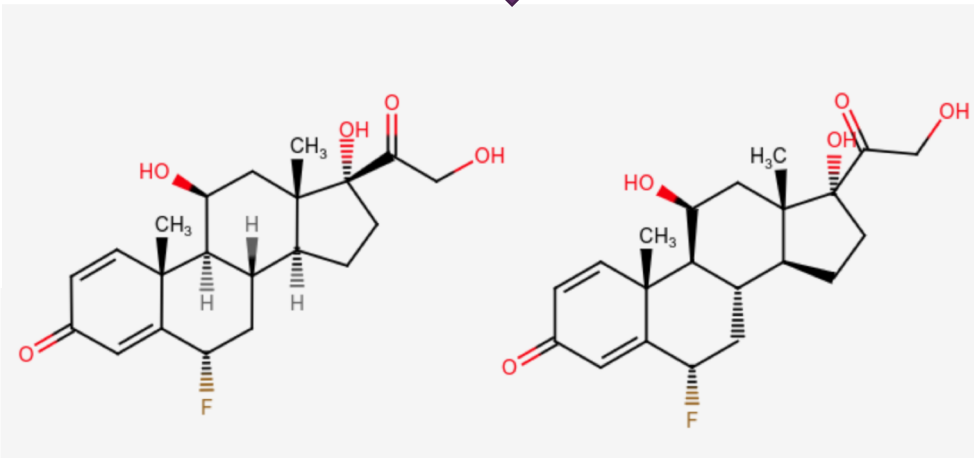




MolVec *is*…
Accurately pretty

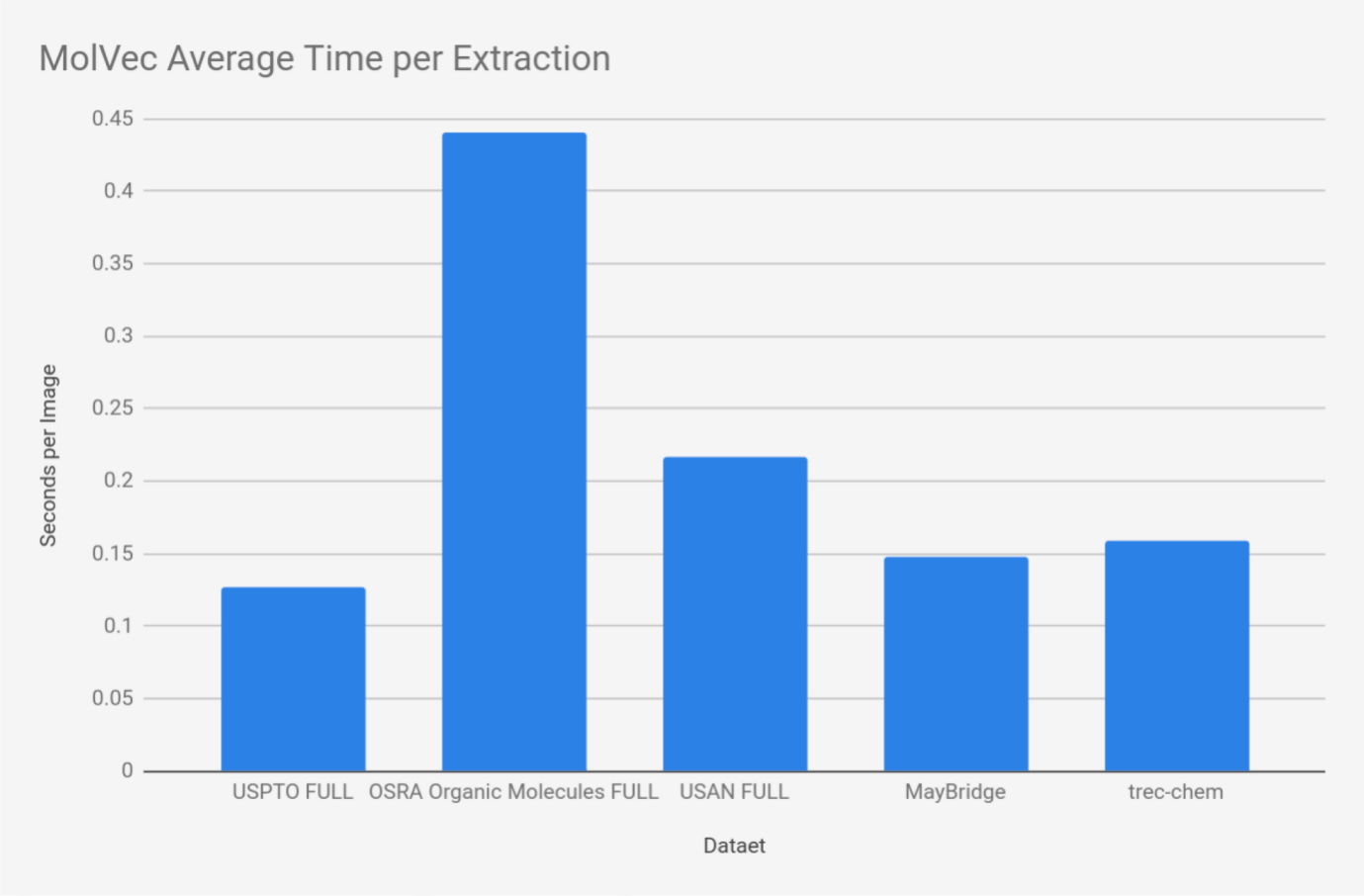




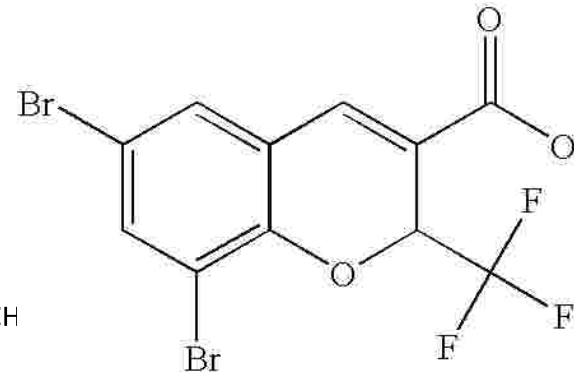


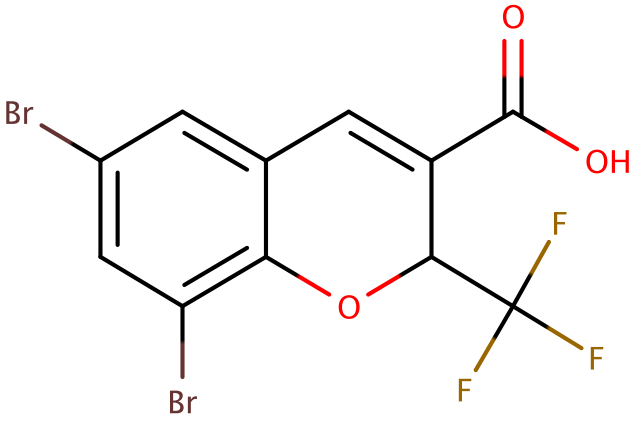
MolVec IUPAC

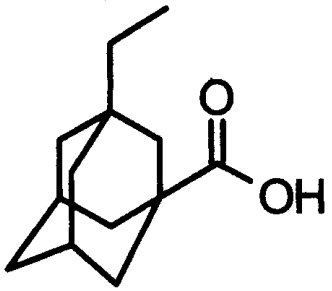
MolVec *is*…
Fairly fast

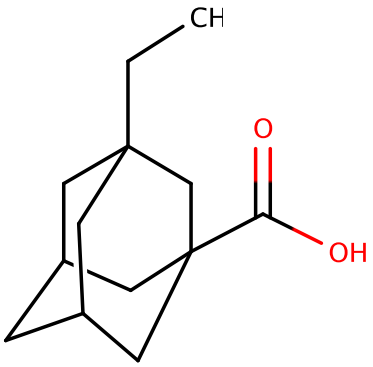


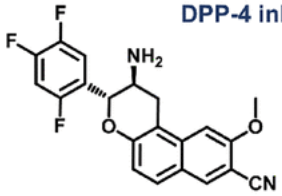
Examples



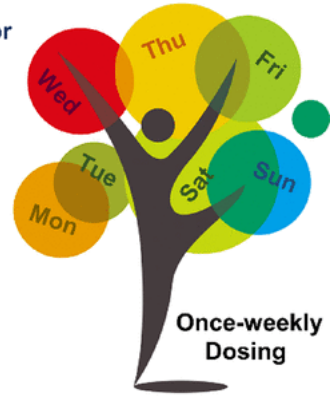






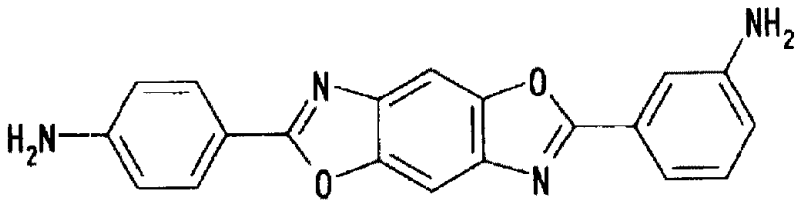


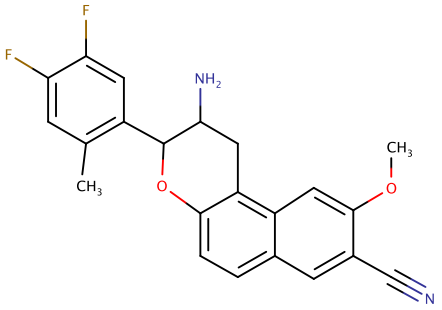
Long-acting
DPP-4 inhibitor

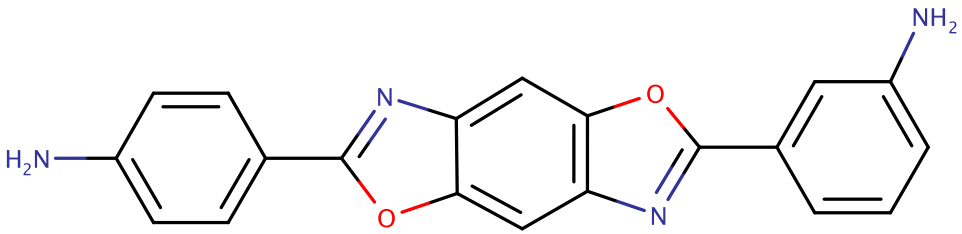


2，6-(3，4'-ジアミノジフェニル)ベンゾ[1，2-d：4，5-d']ビスオキサゾ-

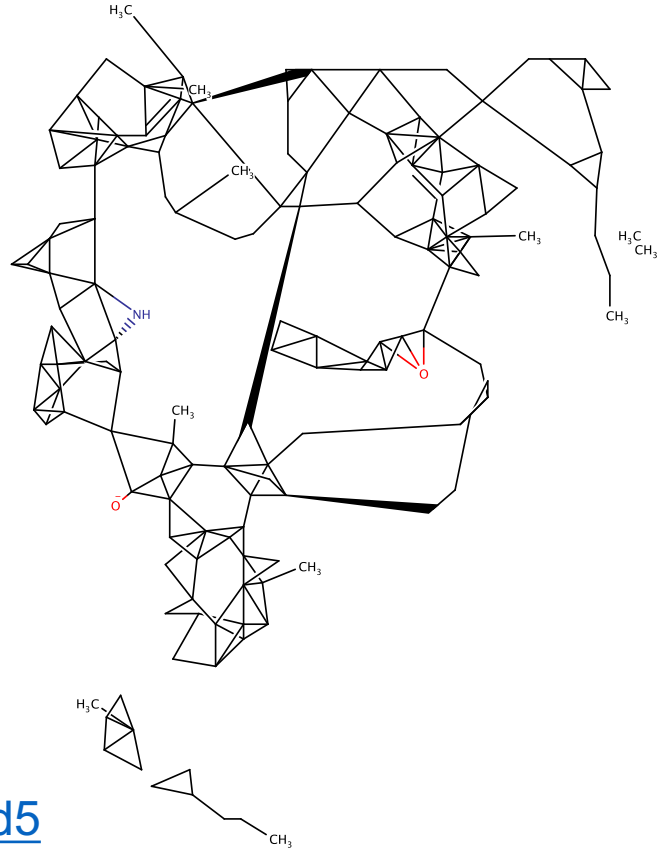
$K_D = 1.77 \times 10^{-10}$ nM
$k_{on} = 1.29 \times 10^{7}\ M^{-1}s^{-1}$
$k_{off} = 2.29 \times 10^{-3}\ s^{-1}$
$T_{1/2} > 24$ h
$F(\%) = 98.54$







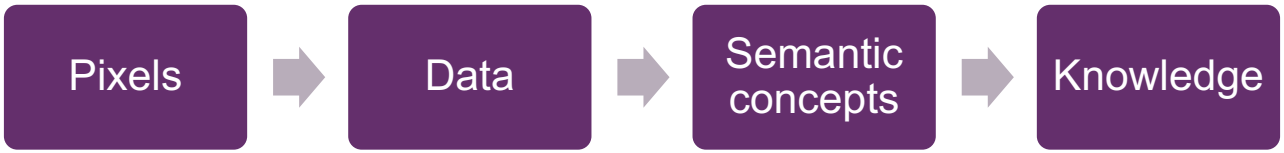
Molecular rendition of Roger Sayle



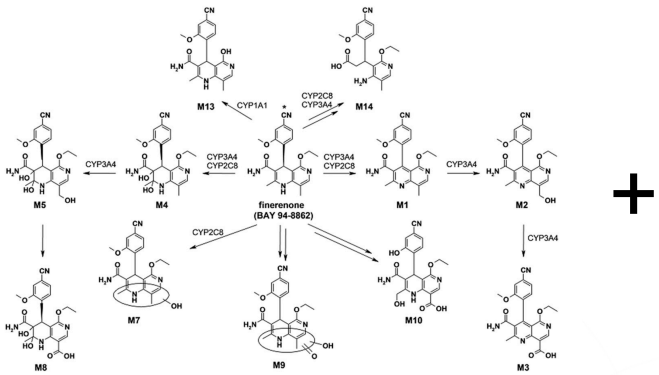
https://molvec.ncats.io/d3d24314d5

Conclusion and future direction

- MolVec is a portable, lightweight, and accurate image to structure recognition engine
- Only the first step toward our larger goal



- Next step: how to combine MolVec and NLP to generate data?



+

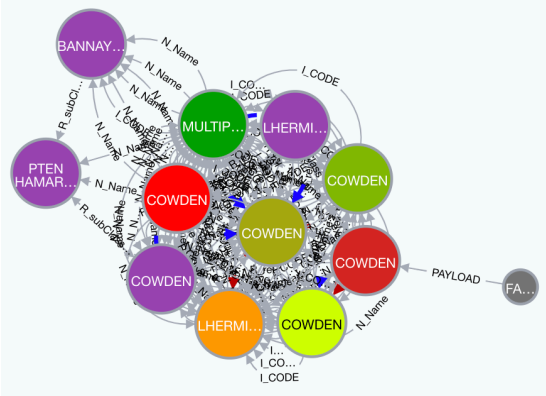
TABLE 4

Atropisomer ratios in plasma and excretion (total balance) in humans, dogs, and rats following single-dose administration of [^{14}C]finerenone, and from incubates in liver microsomes and recombinant CYP3A4 and CYP2C8

	M1a %	M1b %	M2a %	M2b %	M3a %	M3b %
Plasma in vivo						
Humans	79.3	20.7	94.6	5.4	99.1	0.9
Dogs	91.2	8.8	89.1	10.9	90.7	9.3
Rats	na	na	na	na	na	na
Total mass balance in vivo						
Humans	86.0	14.0	93.6	6.4	97.3	2.7
Dogs	87.5	12.5	83.3	16.7	83.8	16.2
Rats	87.7	12.3	93.9	6.1	87.5	12.5
Liver microsomes						
Humans	94.6	5.4	96.2	3.8	—	—
Dogs	92.5	7.5	91.9	8.1	—	—
Rats	95.5	4.5	96.8	3.2	93.7	6.3
Recombinant enzymes						
CYP2C8	70.7	29.3	—	—	—	—
CYP3A4	95.7	4.3	97.8	2.2	—	—

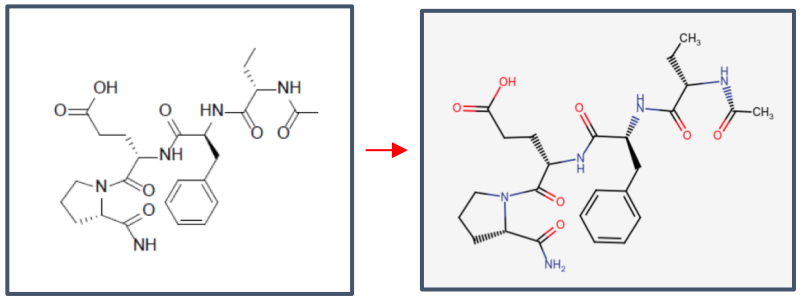
na, Not applicable.

=



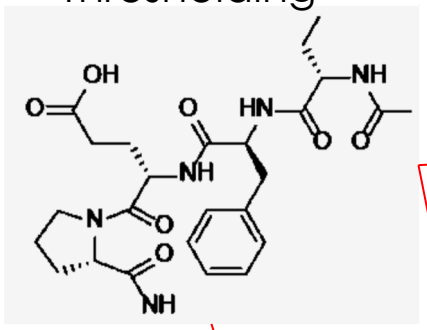
Thank you

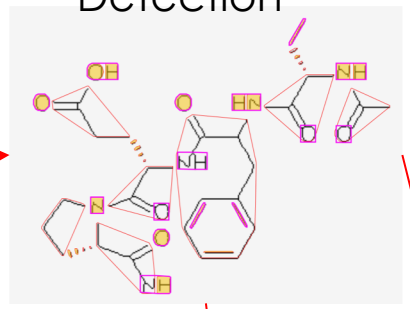
Implementation

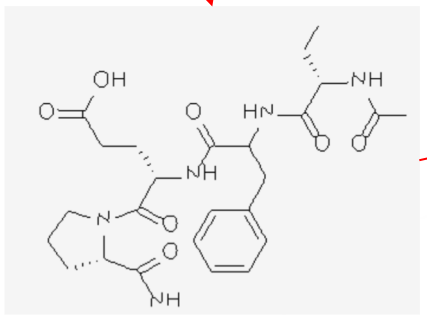


Shape and Feature Detection

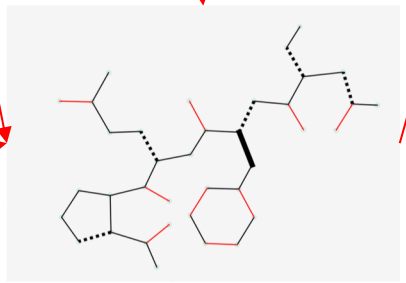
Thresholding



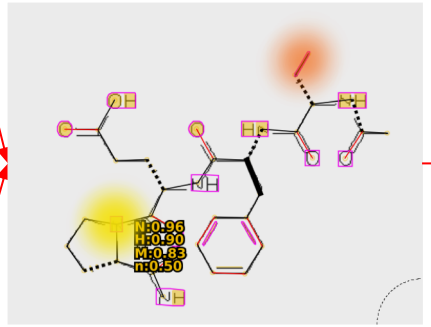




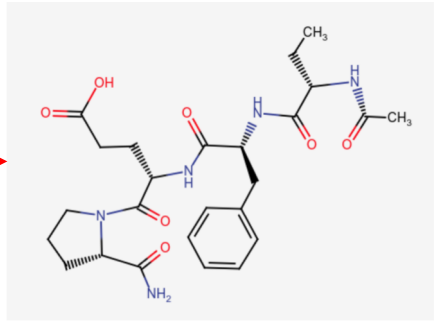
Thinning



Node and Edge Detection



Heuristics and Adjustments



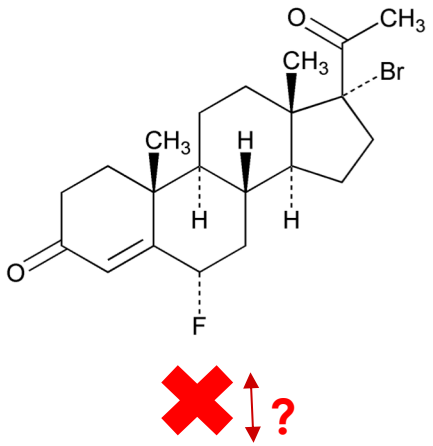
Assembly

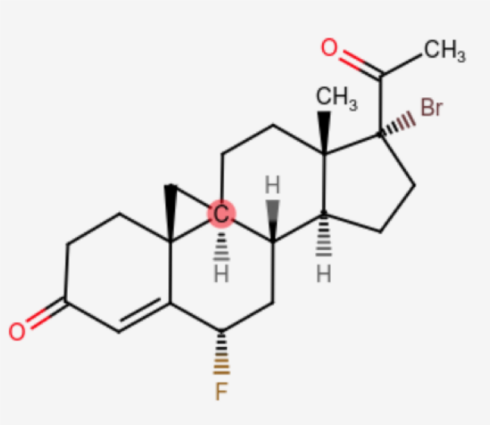
- No AI magic
- Simple image processing and pattern recognition techniques

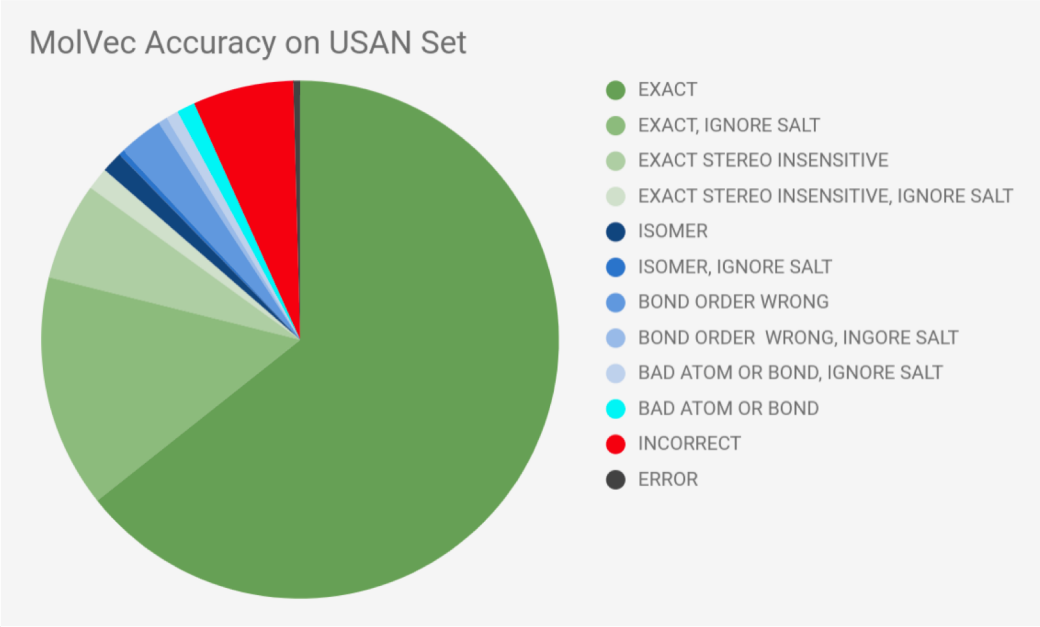
Benchmark datasets

- Available at https://spotlite.nih.gov/ncats/molvec
- TREC-CHEM (1000 binary images + molfiles)
- USPTO subset (5,710 binary images + molfiles)
- MayBridge subset (2,665 binary images + molfiles)
- OSRA organic molecule set (441 binary images + molfiles)
- USAN image set (7,546 binary images + CAS)

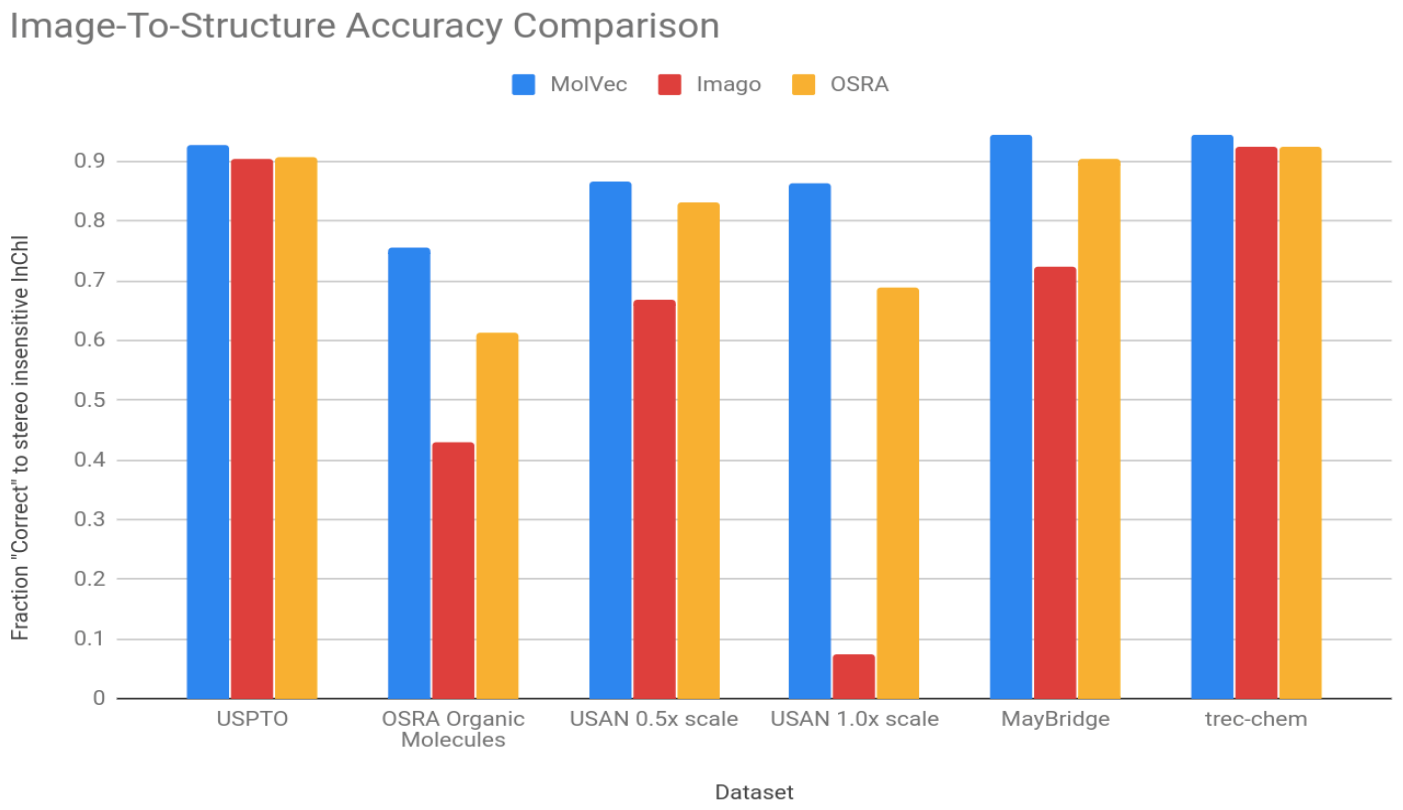
Performance benchmark
Accuracy evaluation



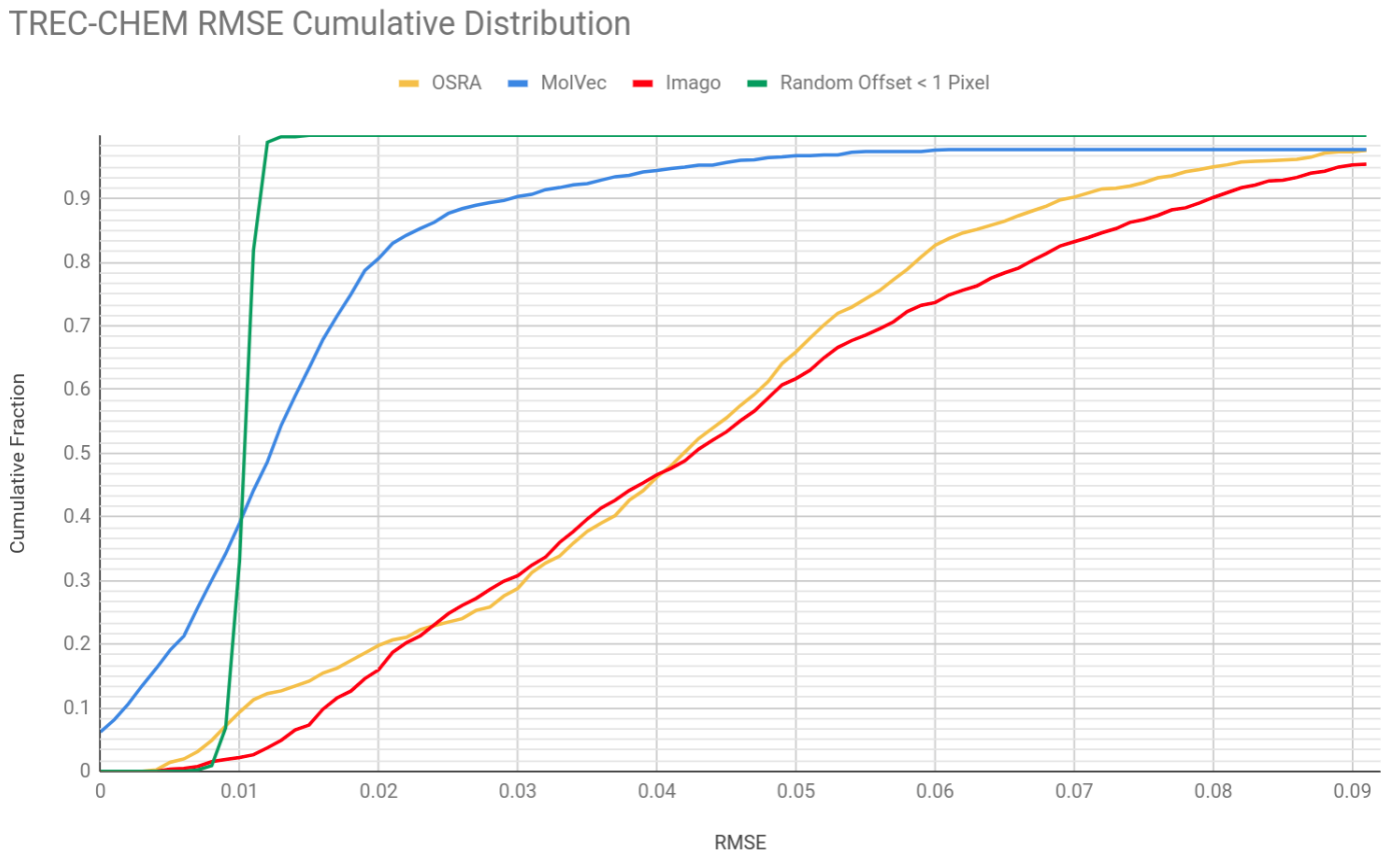




Performance benchmark
Accuracy



Performance benchmark
Layout preservation



Performance benchmark
Speed