# GSRS Development Update

15 May 2019

TYLER PERYEA
DANNY KATZEL
NIH/NCATS

NCATS

# Agenda

## High Level GSRS Core Status

- 2.3.3 Release
- Ongoing Software Updates
- Data Export Update
- New Image to Structure Feature

# GSRS Release Status

- GSRS FDA Production Releases
  - 2.3.1 - October 2018 - previous public release
  - 2.3.2 - December 2018
  - 2.3.3 - February 2019 - most recent public release including public data set
  - 2.3.4 - March 2019
  - 2.3.5 - April 2019
- Upcoming GSRS FDA Releases
  - 2.3.6 - scheduled for June 2019
  - 2.4 development - 3rd quarter 2019
  - Overall goals:  Improved data exchange / communication

# GSRS 2.3.3 Release Update

## Updates available in the 2.3.3 public release

- Structure rehash / reprocessing task to regenerate hashes when structure hash algorithm changes
- Allow scheduled task for remaking backups
- Minor UI Improvements
- Improved speed of facet filtering, sequence searching
- Integrated Image to Structure in registration pages
- Bug fixes, additional validation rules

# Ongoing Software Updates

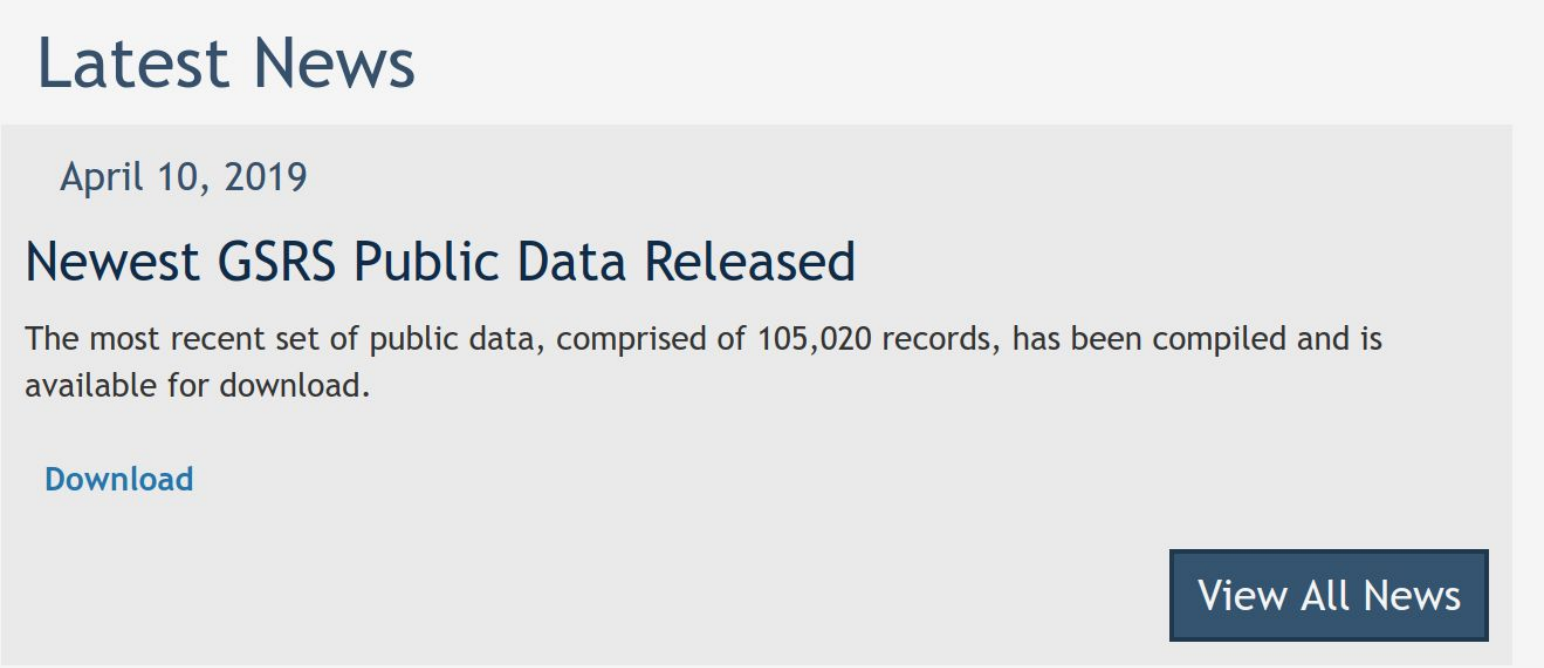
## Completed Updates to be Included in Next Public Release

- Initial Beta UI Redesign - View Forms, Global Search (basic), Structure Search, Sequence Search
- Improved Name Display - added Additional Listing name to view of names, added name details to Names card
- Add a button on browse for privileged users to copy the structure/sequence directly to a registration page without bringing the references

## Upcoming Updates

- Beta UI Improvements
- Migrating Bookmarklets / Powertools to UI
- Improved Molvec

# New Dataset Released

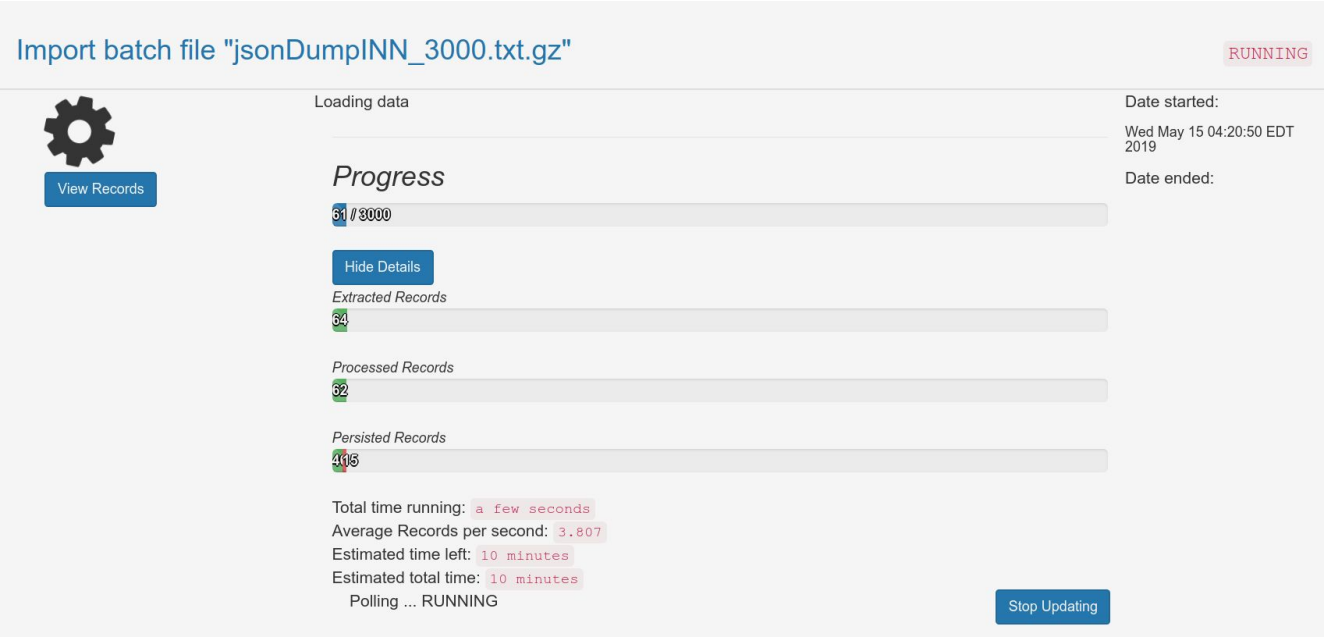



| Data Type | Count |
|---|---|
| Chemical | 71,098 |
| Structurally Diverse | 25,038 |
| Protein | 3,547 |
| Mixture | 2,479 |
| Polymer | 2,080 |
| Nucleic Acid | 172 |
| Group 1 Specified Substance | 15 |
| Names | 735,102 |
| Codes | 426,409 |
| Relationships | 55,323 |

# New Dataset Released

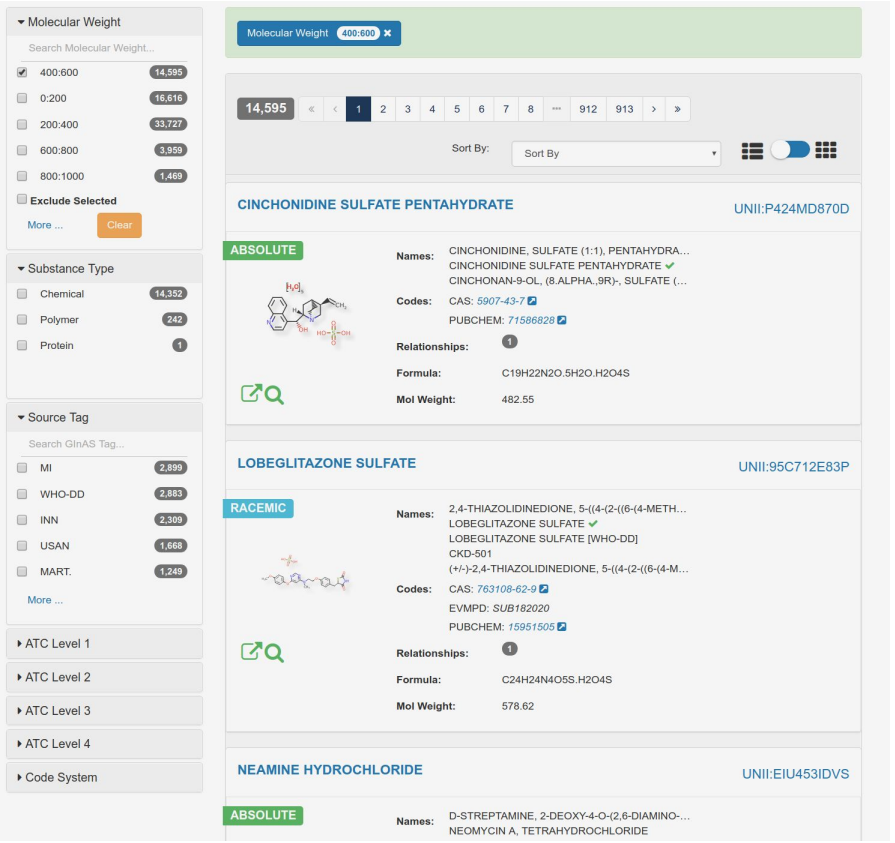
## A few words about the Data Dump format

- Native file format for G-SRS (.gsrs)
  - gzipped, tab-delimited file with 1 G-SRS substance JSON per row
  - loadable directly by the G-SRS application
  - convertible (lossy) to several other formats (txt, sdf)



# New Dataset Released

- Browsable on main website, or via API




https://ginas.ncats.nih.gov/ginas/app/substances

https://tripod.nih.gov/ginas/#/api

# Links

Full dataset json

https://tripod.nih.gov/ginas/downloads/dump-public-2019-04-03.gsrs

**Collaborator Slack channel (ask us to add you)**

https://gsrscollaborator.slack.com

# GSRS New User Interface (UI)

## Main Goals

- Decouple UI/ client side logic from monolithic application
- Continue to support current application until decouping finished

## Why

- Use newer web technology typescript, angular 6 instead of legacy Play scala templates that are harder to write, slower, and require backend processing
- Better performance
  - Use REST API only  remove back end processing
  - SPA loads most assets at once
- Leads to more efficient way of adding new features
- Easier for future development of custom UIs for different organizations

# Beta UI

Record Status

- Validated (UNII) 494
- FAILED 6
- pending 1

Substance Type

- chemical 480
- polymer 15
- protein 6

Source Tag

Code System

ATC Level 1

ATC Level 2

ATC Level 3

ATC Level 4

Moiety Type

Stereochemistry

Relationships

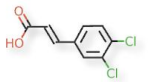
Sort By

Items per page: 10 1 - 10 of 501

## 3,4-DICHLOROCINNAMIC ACID

480625A7SY

ACHIRAL



**Names:** 3,4-DICHLOROCINNAMIC ACID ✓
2-PROPENOIC ACID, 3-(3,4-DICHLOROPHENYL)-
3',4'-DICHLOROCINNAMIC ACID
CINNAMIC ACID, 3,4-DICHLORO-
NSC-518800

**Codes:** **CAS**: 1202-39-7
**EC (EINECS)**: 214-866-4
**BDNUM**: 0000113AB

**Mol. Weight:** 217.049

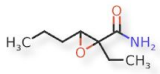
**Formula:** C9H6Cl2O2

**Created:** 5/14/19
**Status:** Validated (UNII)
**Last Modified:** 5/14/19
**Version:** 1

## OXANAMIDE

050271194T

MIXED

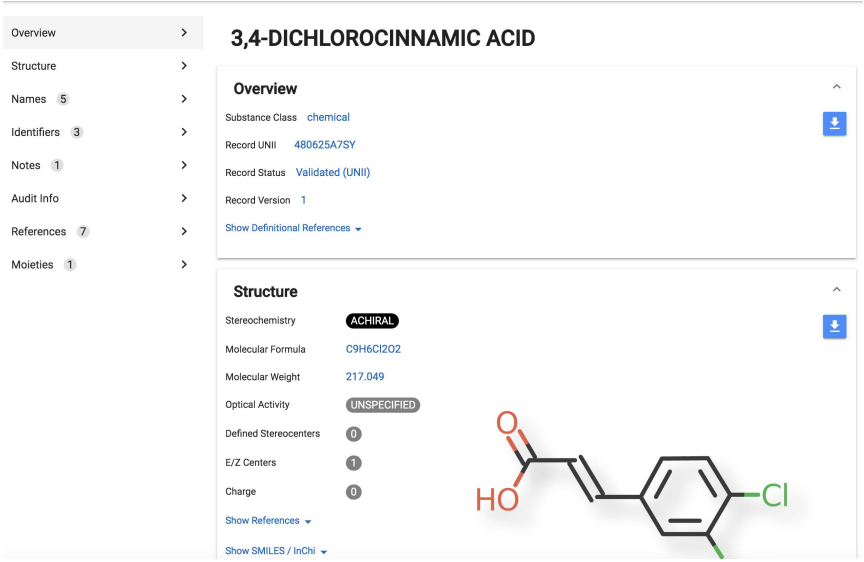


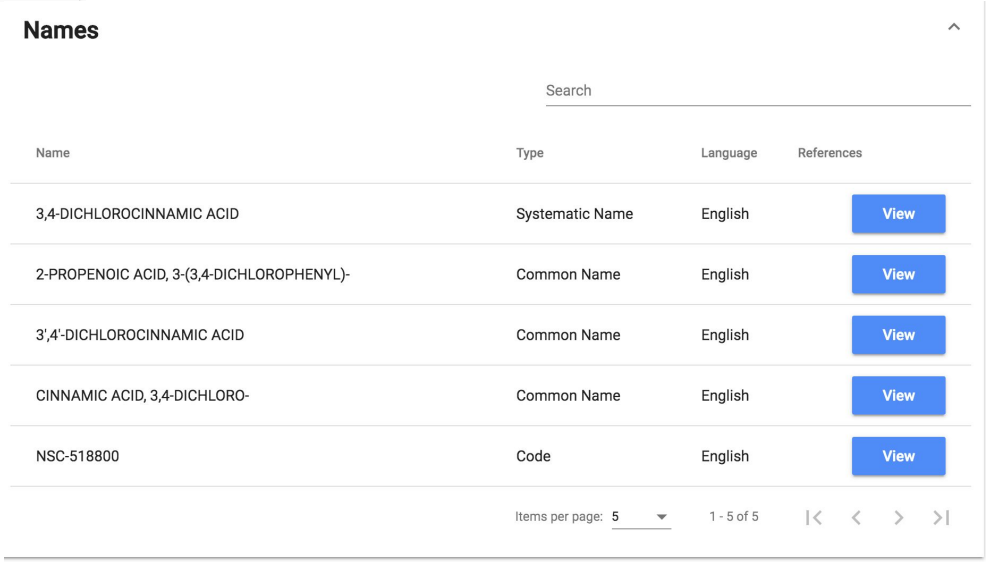
**Names:** OXANAMIDE ✓
oxanamida
oxanamidum
оксанамид
أوكسانااميد

**Codes:** **CAS**: 126-93-2
**WIKIPEDIA**: OXANAMIDE
**EVMPD**: SUB09495MIG
**INN**: 812
**BDNUM**: 0000114AB

**Created:** 5/14/19
**Status:** Validated (UNII)
**Last Modified:** 5/14/19
**Version:** 1

# Beta UI - Substance View

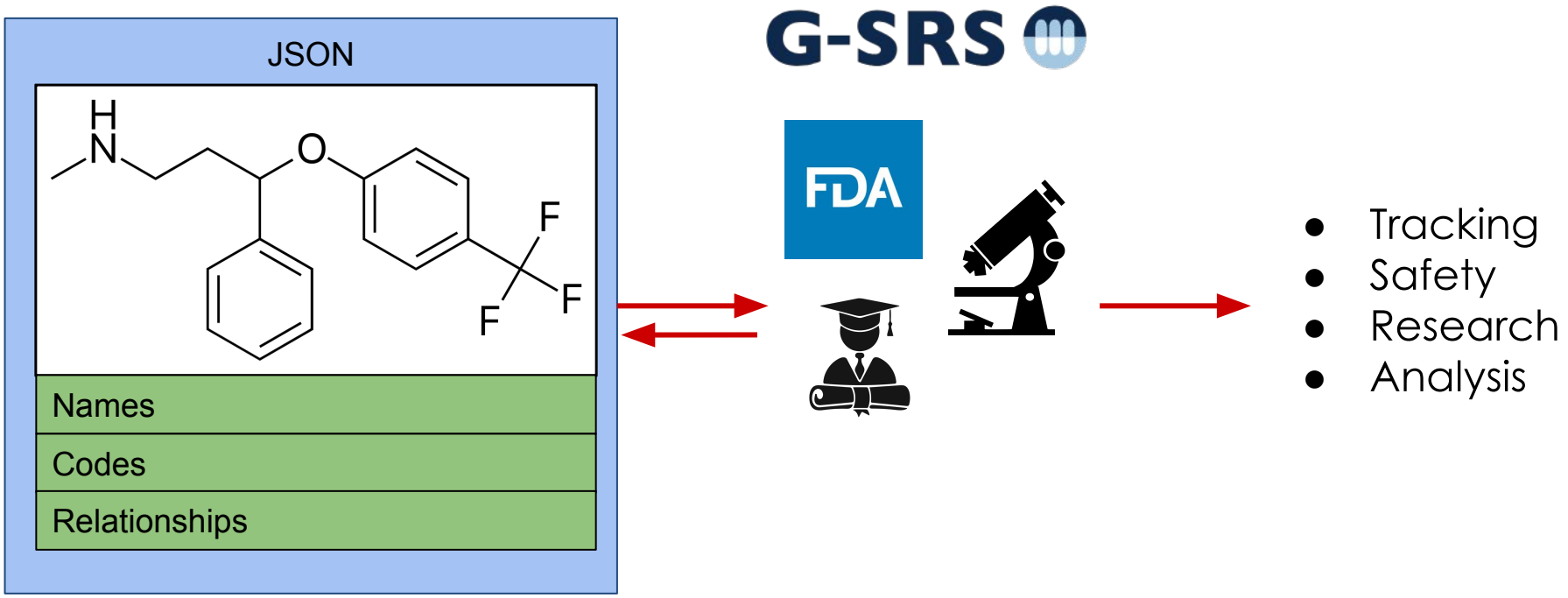


# Questions

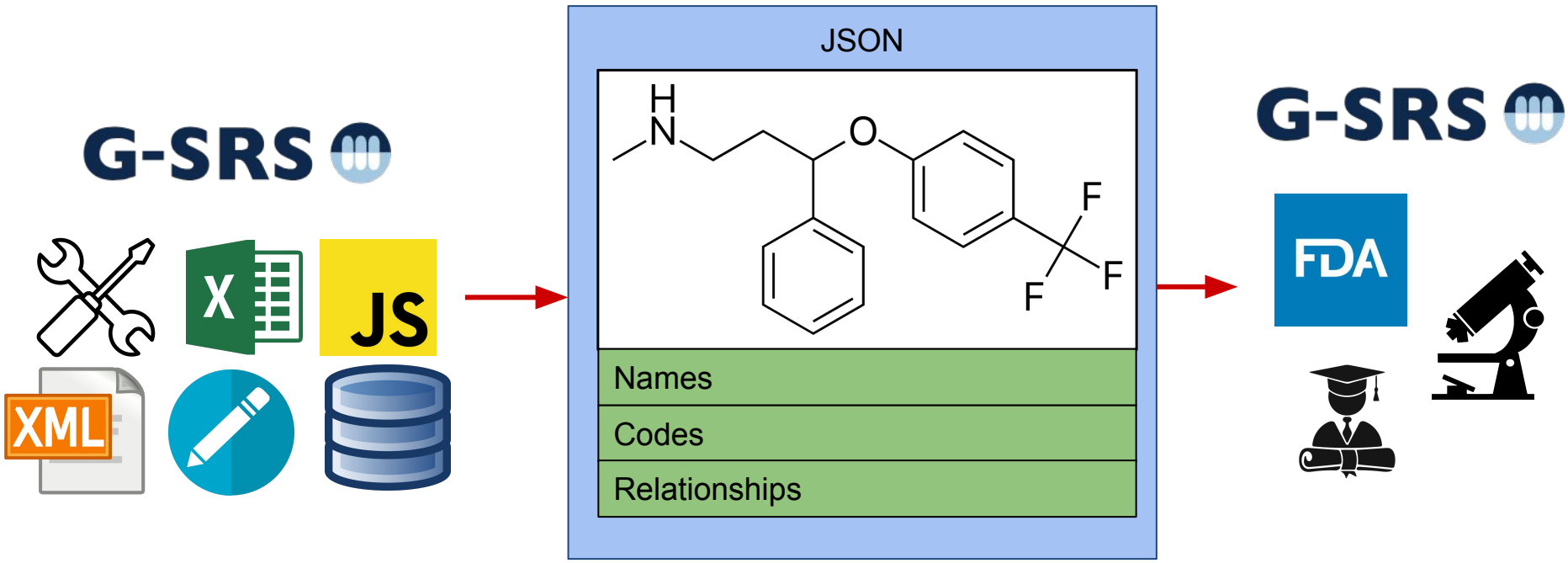
# Molecular Structure Image Recognition

- G-SRS is intended to produce, consume, curate and exchange highly-structured substance data



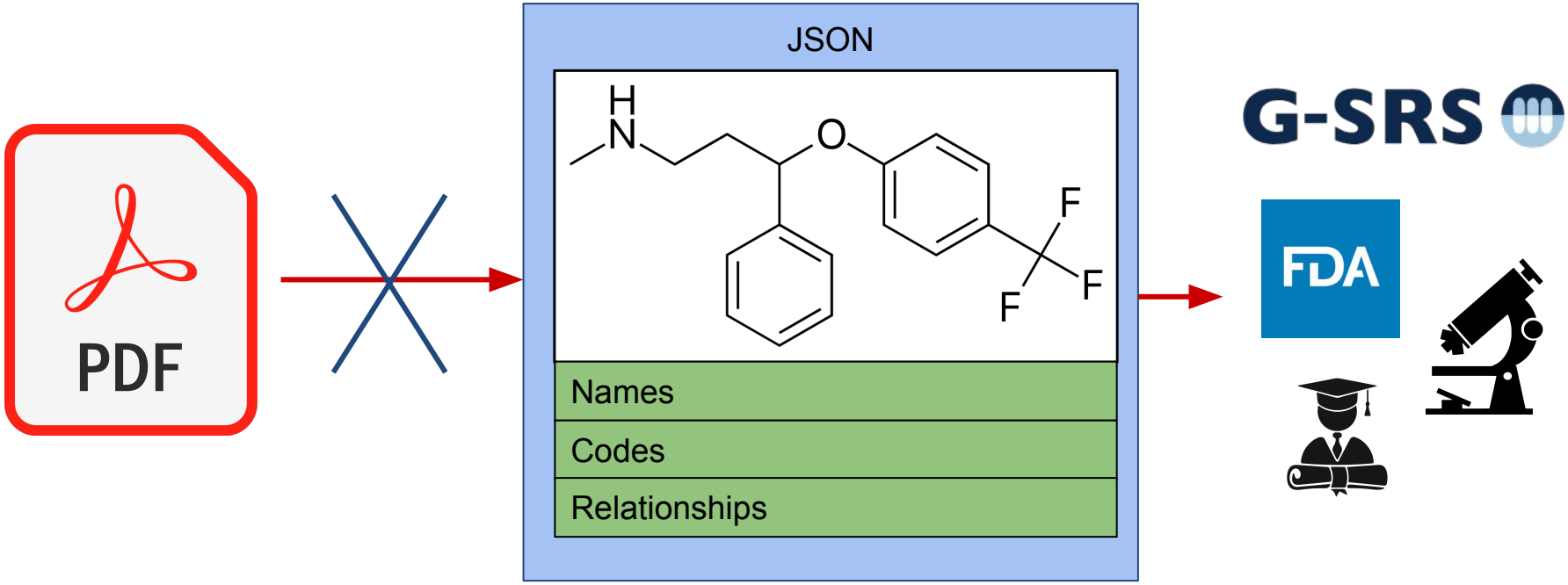
# Molecular Structure Image Recognition

- For structured formats and systems transformation, exchange of data is pretty easy



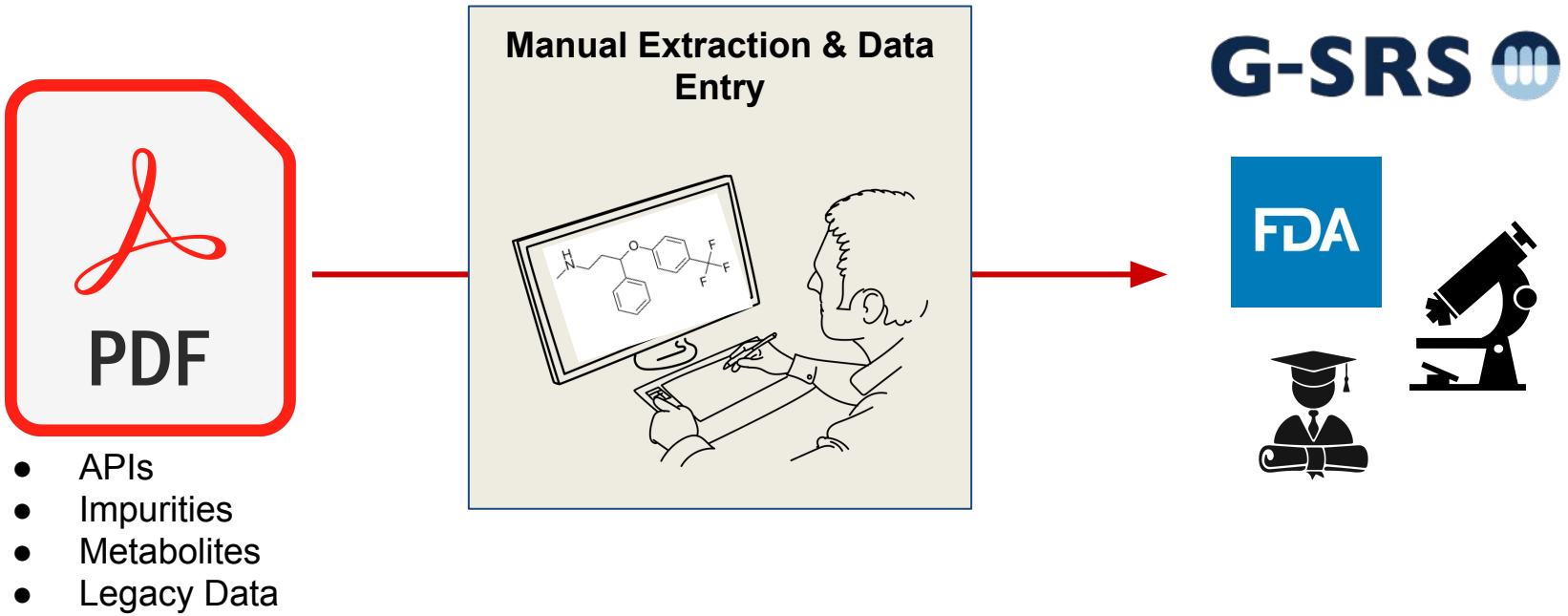
# Molecular Structure Image Recognition

- For structured formats and systems transformation, exchange of data is pretty easy
- Unstructured data is harder to process



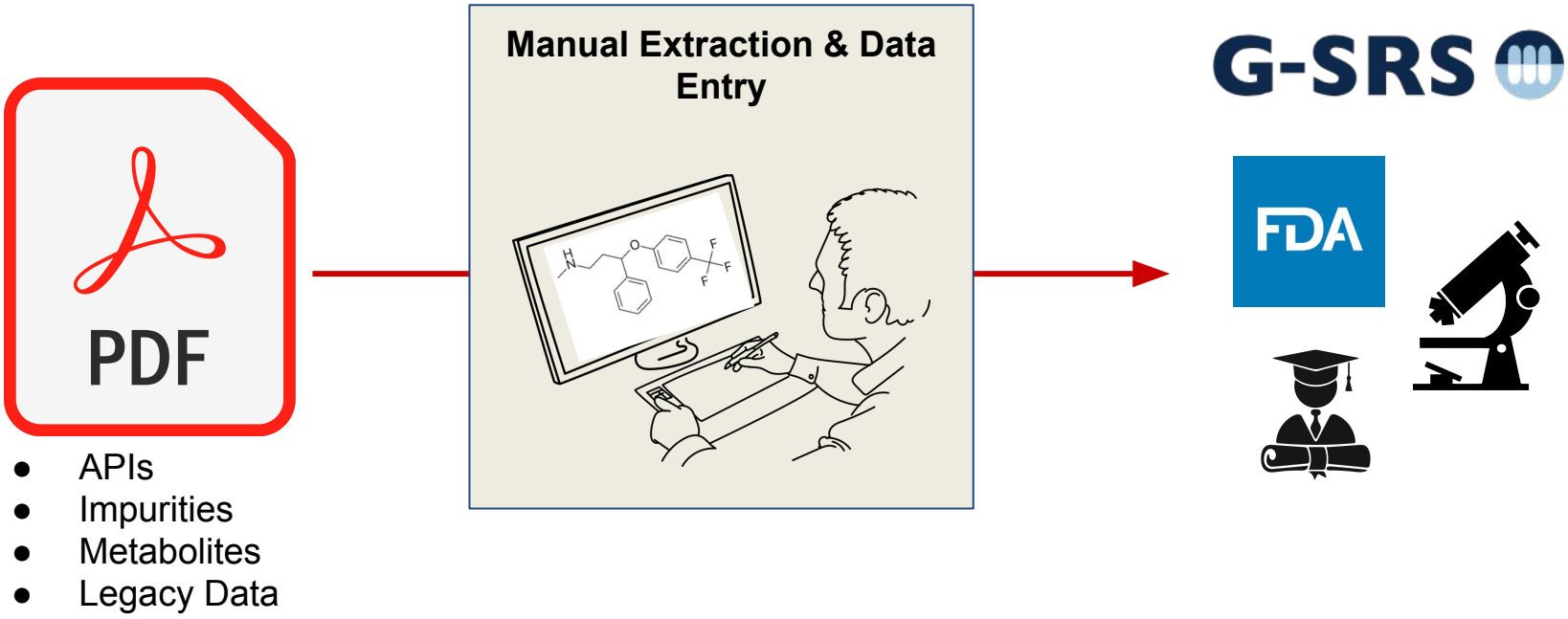
# Molecular Structure Image Recognition

- For structured formats and systems transformation, exchange of data is pretty easy
- Unstructured data is harder to process
- A large portion of legacy data ***is*** unstructured




- APIs
- Impurities
- Metabolites
- Legacy Data

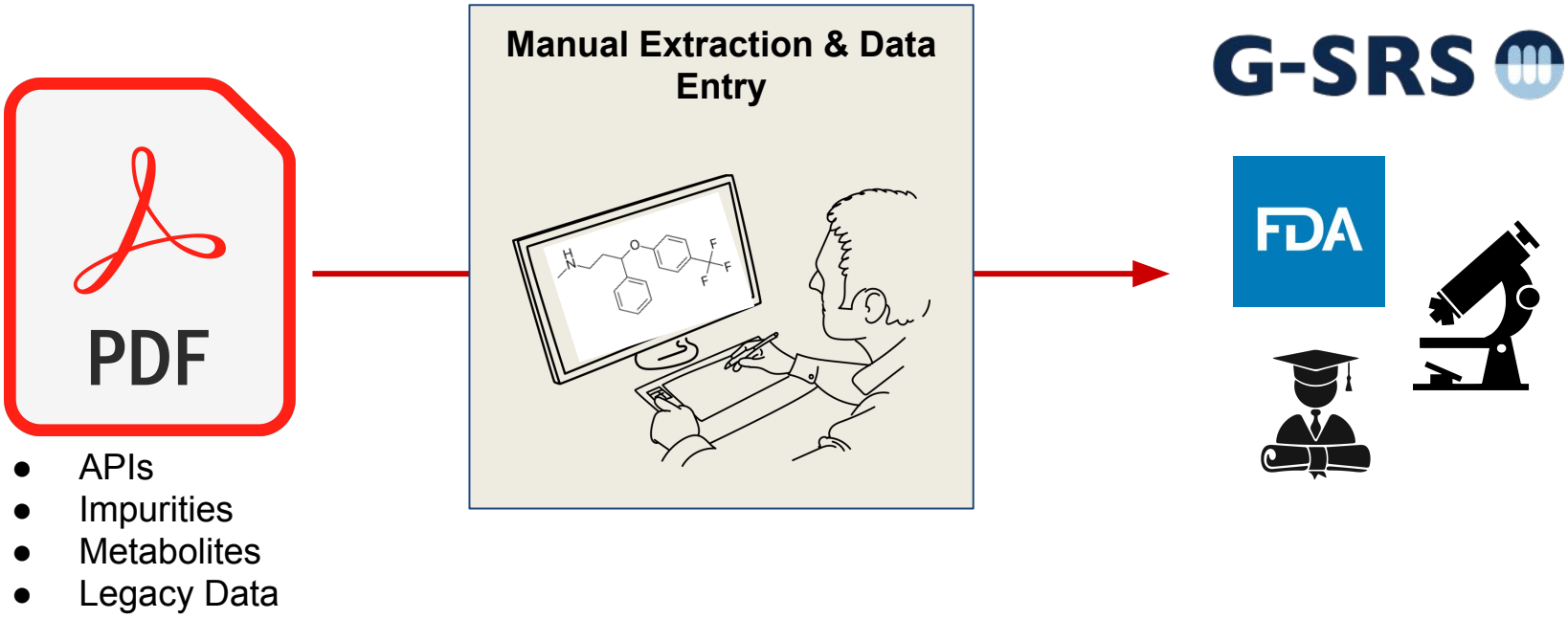
# Molecular Structure Image Recognition




- APIs
- Impurities
- Metabolites
- Legacy Data

- Tedious and slow
- Error-prone

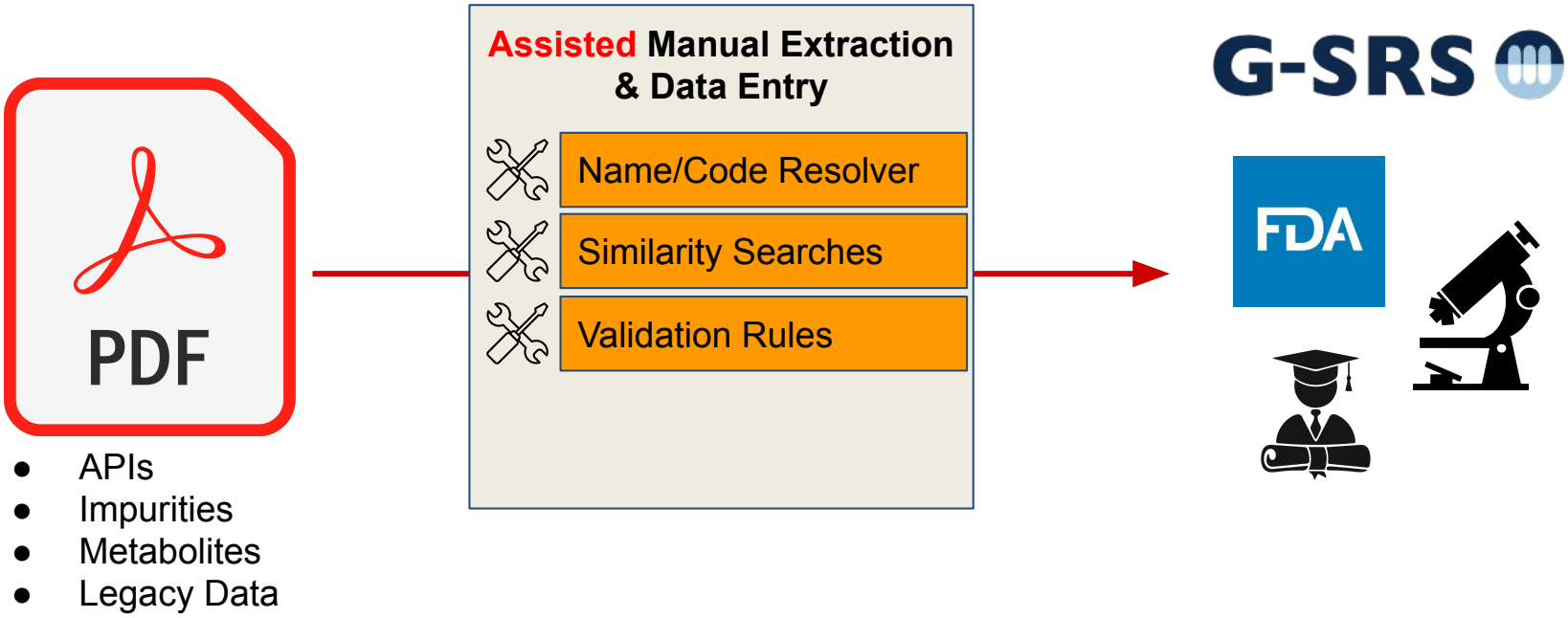
# Molecular Structure Image Recognition




- APIs
- Impurities
- Metabolites
- Legacy Data

- Tedious and slow
- Error-prone
- Carpal Tunnel Syndrome

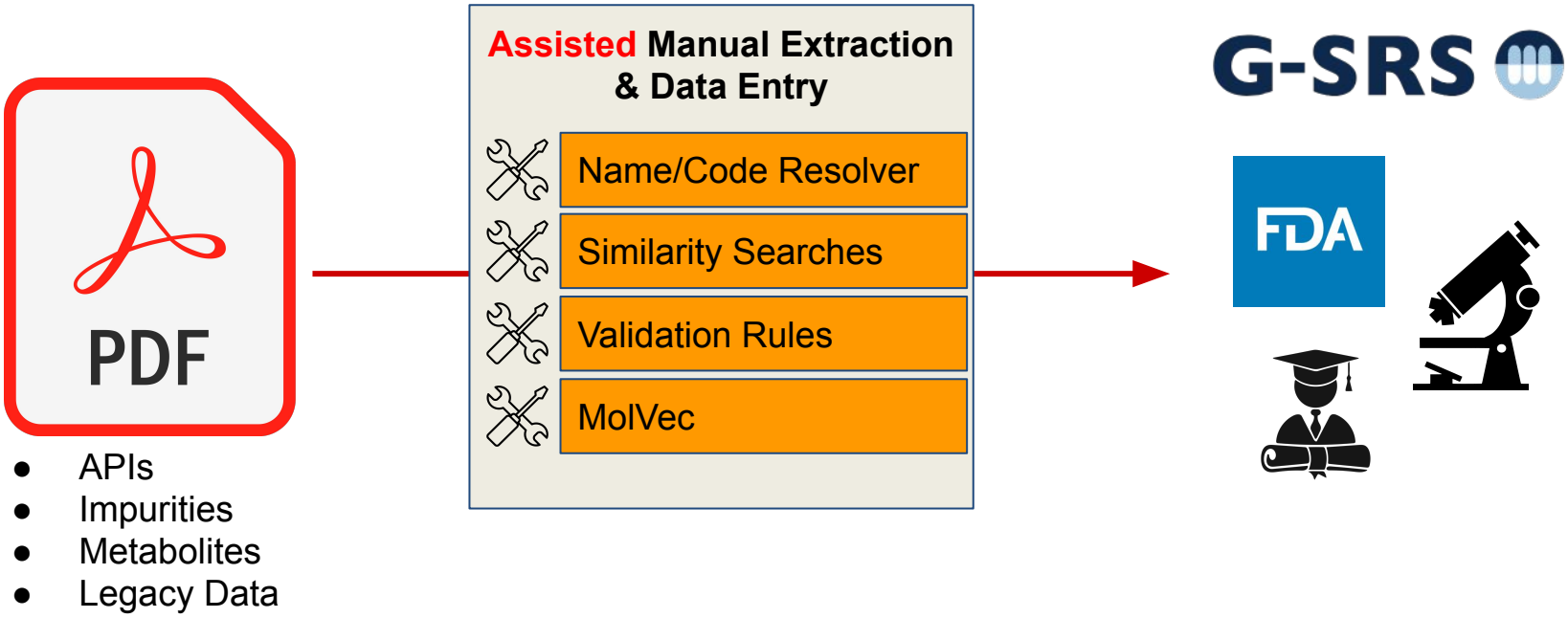
# Molecular Structure Image Recognition




- APIs
- Impurities
- Metabolites
- Legacy Data

- ***Less*** Tedious and slow
- ***Less*** Error-prone
- Carpal Tunnel Syndrome

# Molecular Structure Image Recognition




- APIs
- Impurities
- Metabolites
- Legacy Data

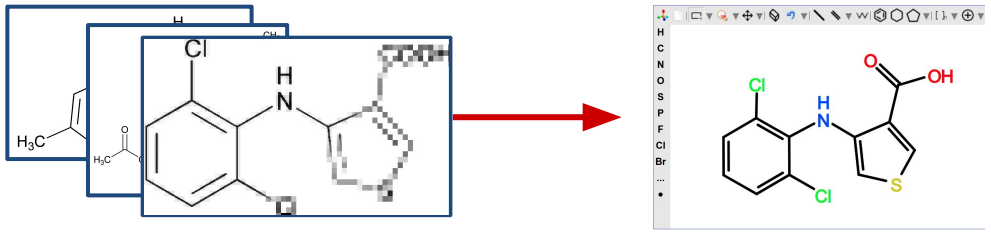
- ***Less*** Tedious and slow
- ***Less*** Error-prone
- ***Less*** Carpal Tunnel Syndrome*



*claim not evaluated or endorsed by FDA

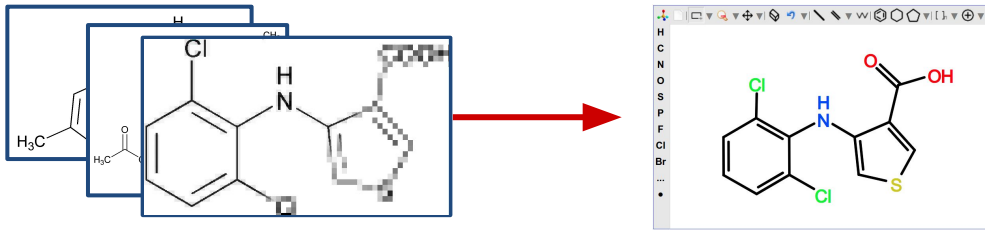
# G-SRS: MolVec

- **MolVec** takes chemical structure ***images*** and turns them into chemical structure ***data***



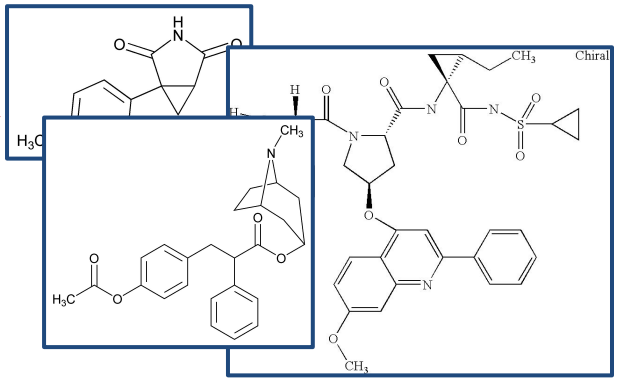
# G-SRS: MolVec

- **MolVec** takes chemical structure ***images*** and turns them into chemical structure ***data***



PDF

- APIs
- Impurities
- Metabolites
- Legacy Data

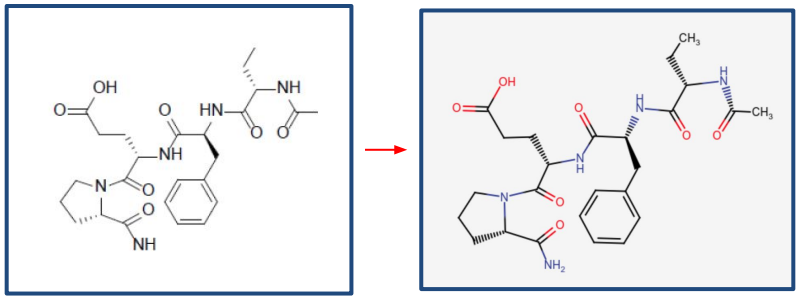


\+ manual extraction / curation

G-SRS

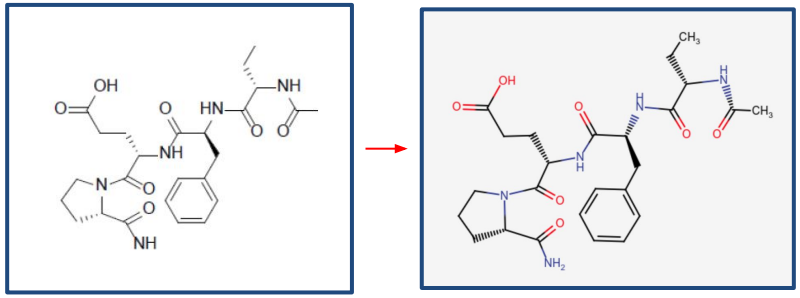
# G-SRS: MolVec

## How it Works



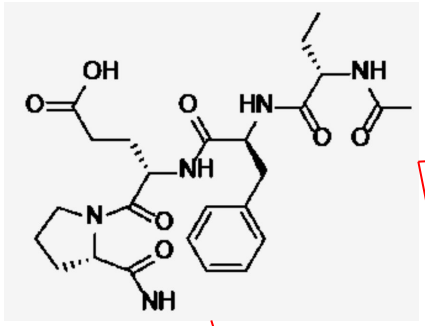
# G-SRS: MolVec

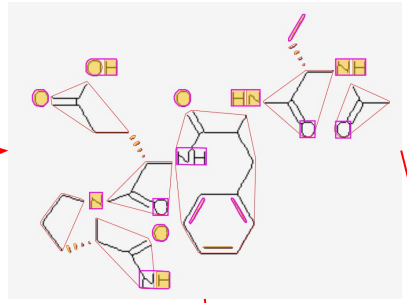
## How it Works

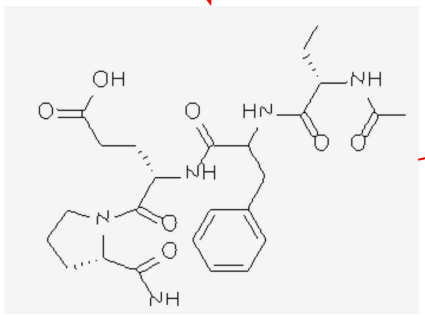


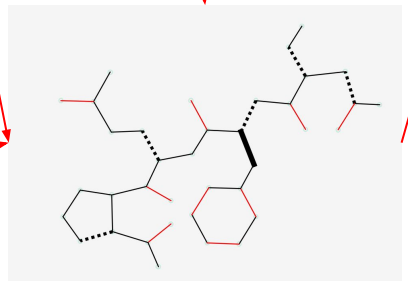
Shape and Feature Detection

Thresholding



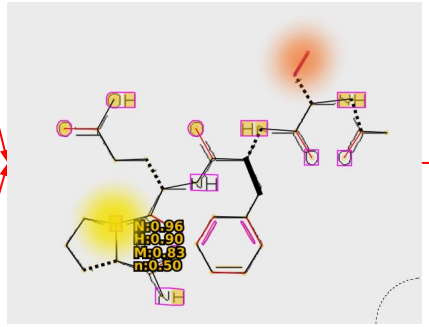




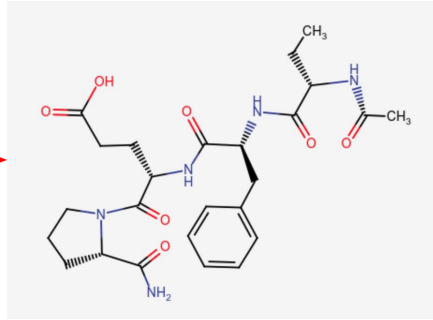


Thinning

Node and Edge Detection




Heuristics and Adjustments

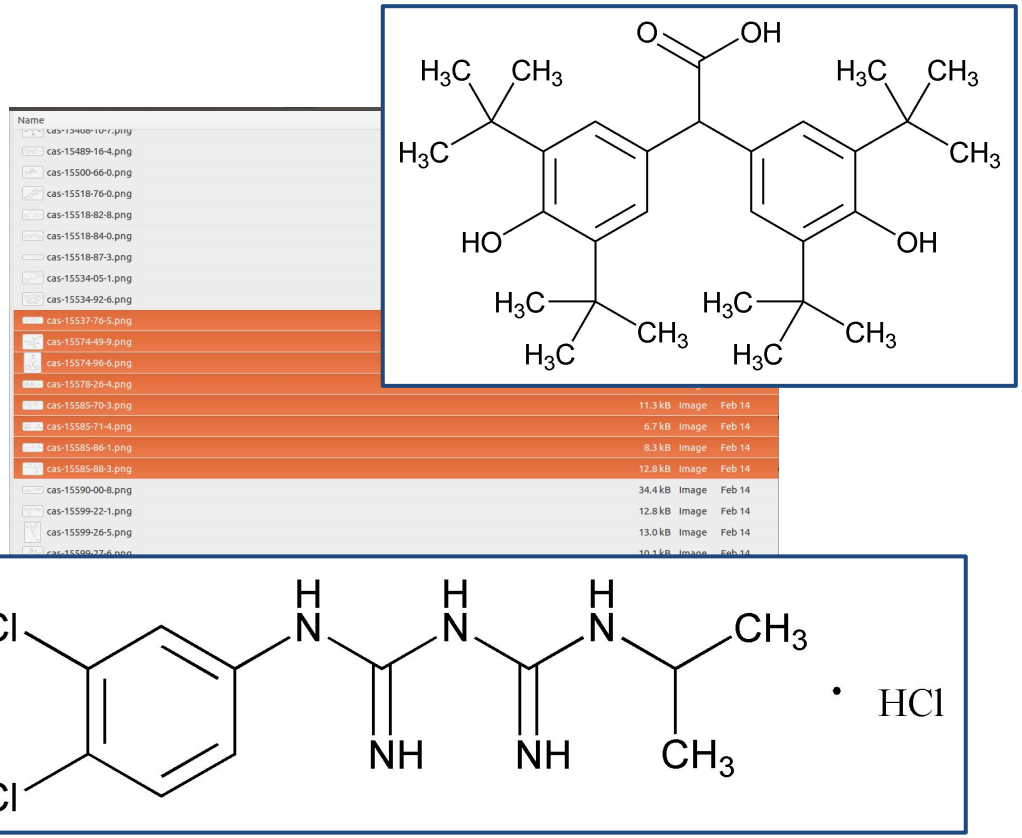


Assembly

# G-SRS: MolVec

## Evaluation Use Case

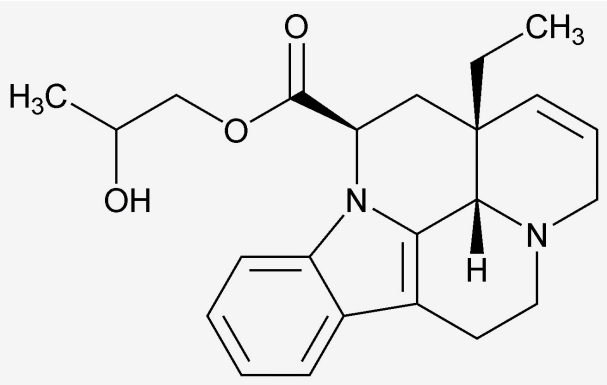
- **USAN** chemical structure image files
  - ~7500 png files
  - Well-drawn
  - Highly-curated
  - Real and meaningful
  - Definitions already known to FDA G-SRS



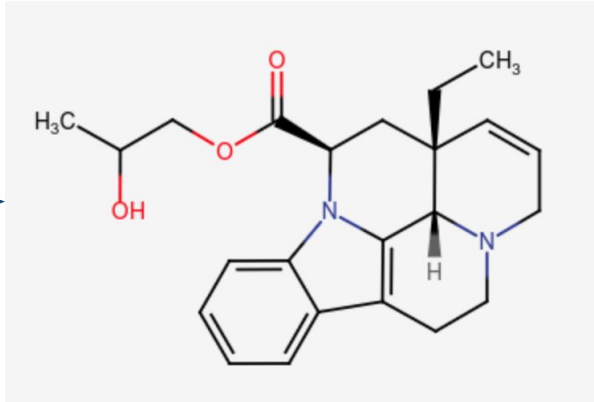
# G-SRS: MolVec

## Evaluation Use Case

**USAN Image**



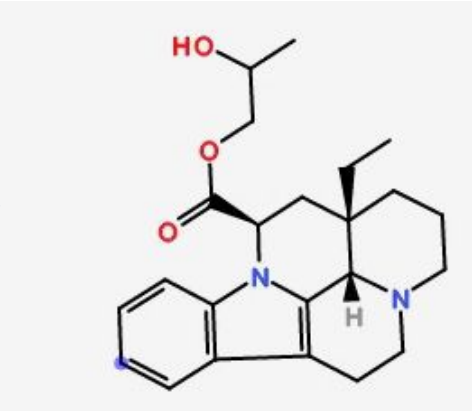

MolVec




File Name

**Name or Code**

G-SRS REST API

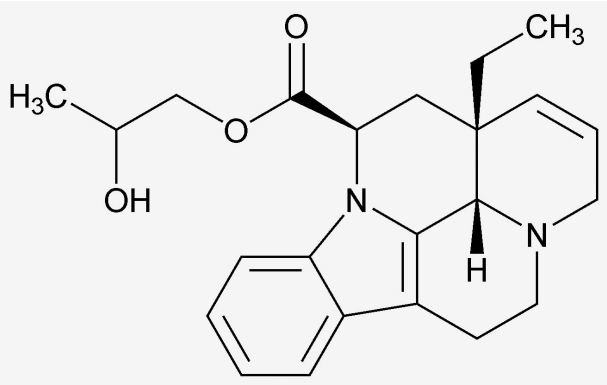



**G-SRS** Structure

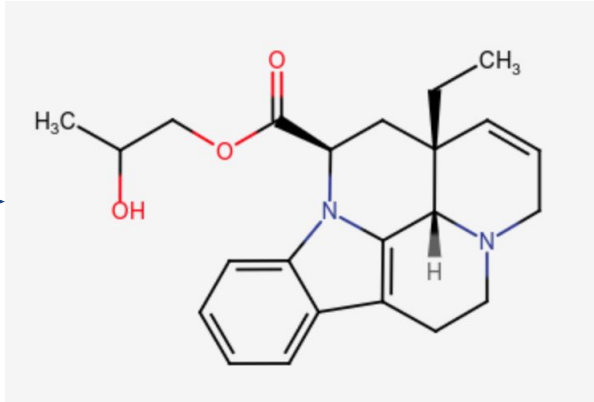
# G-SRS: MolVec

## Evaluation Use Case

**USAN Image**




MolVec

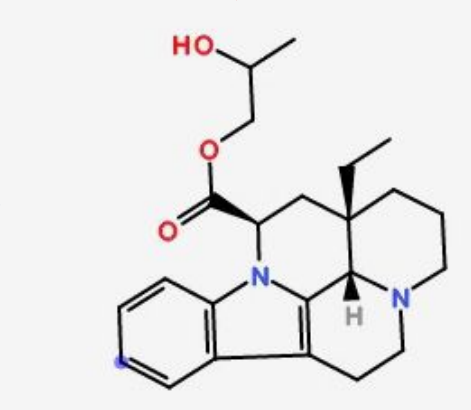



?

File Name

Name or Code

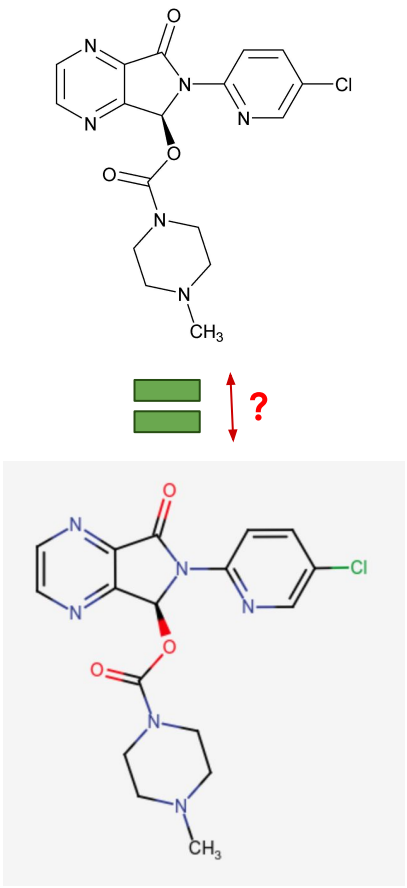
G-SRS REST API

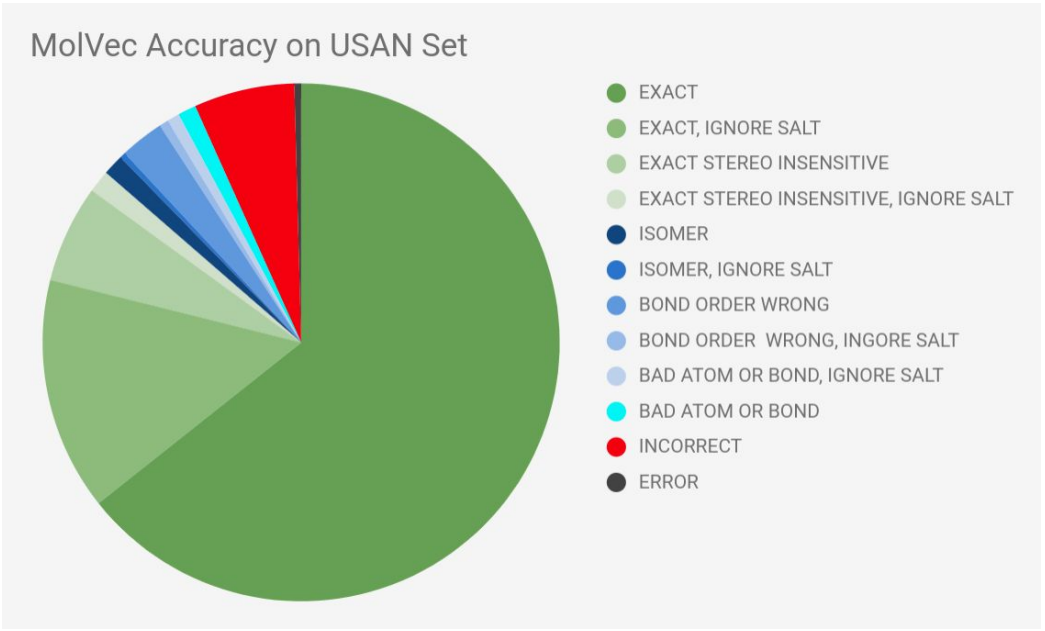



**G-SRS** Structure

# G-SRS: MolVec

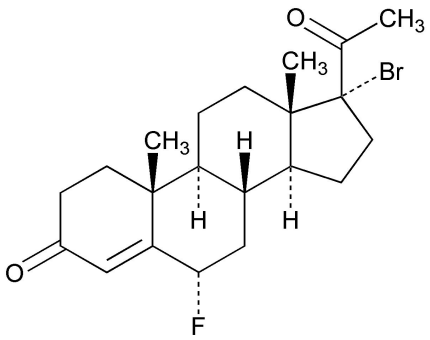
## Evaluation Use Case

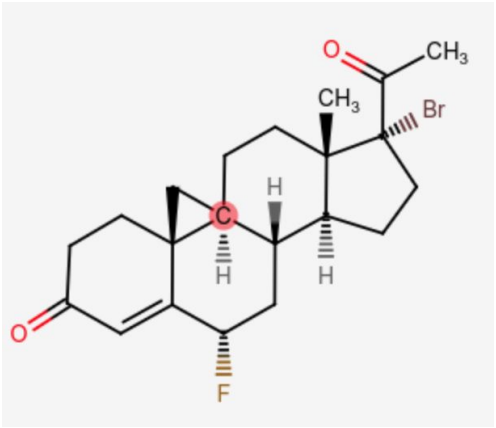


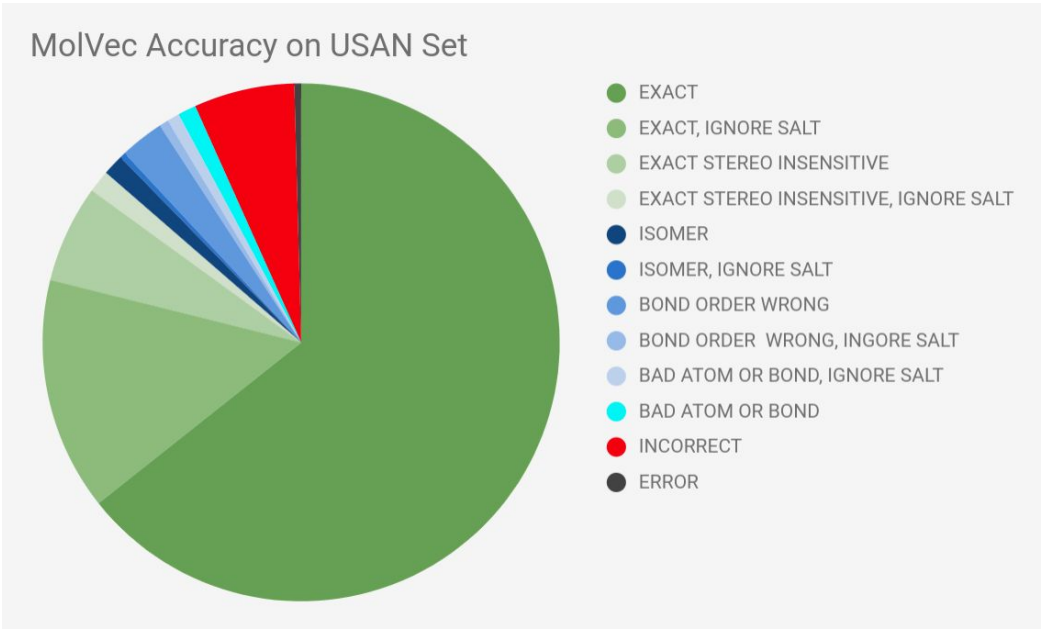



# G-SRS: MolVec

## Evaluation Use Case

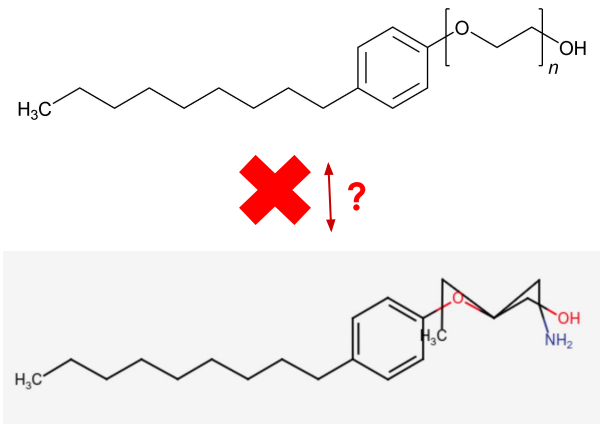


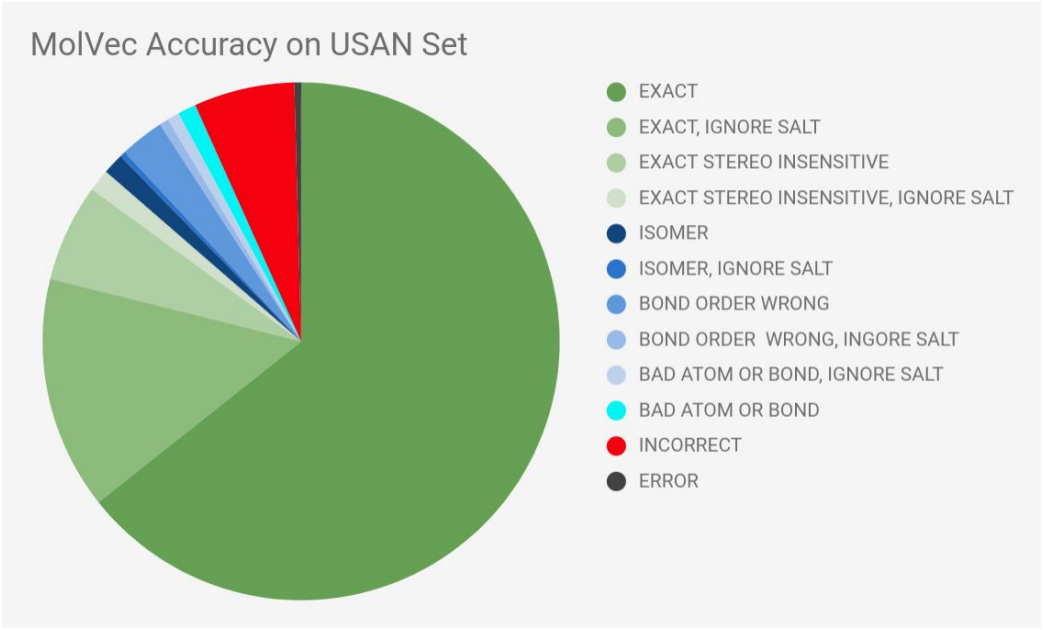



# G-SRS: MolVec

## Evaluation Use Case

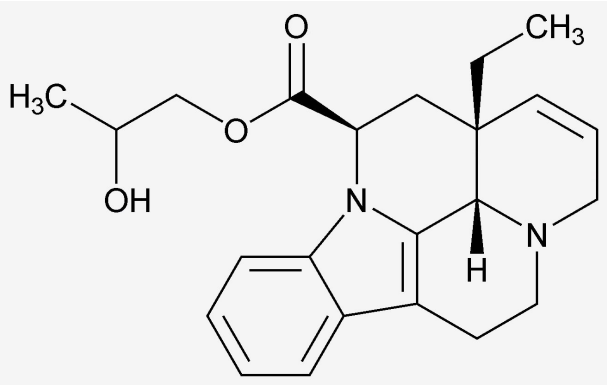


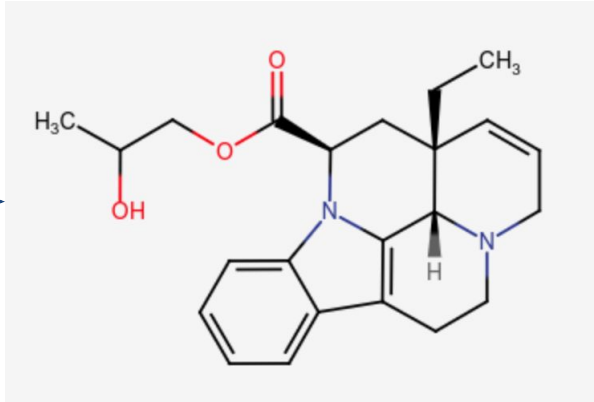
# G-SRS: MolVec

## Evaluation Use Case

**USAN Image**




MolVec

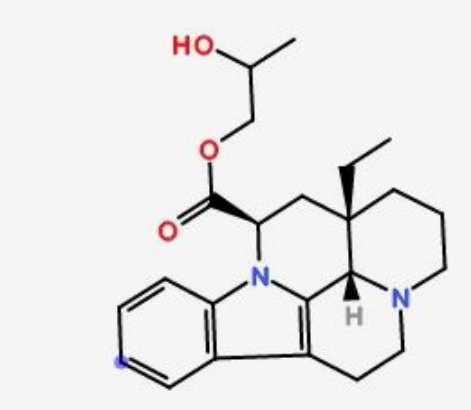



?

File Name

Name or Code

G-SRS REST API

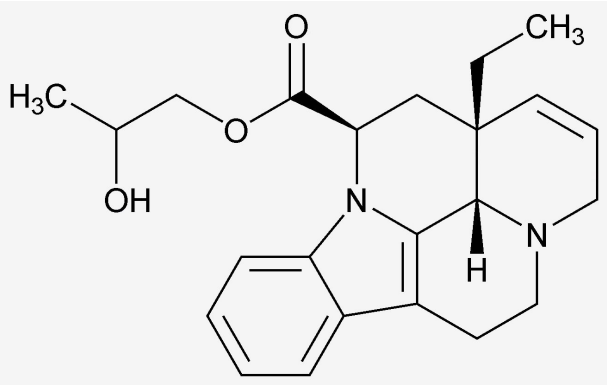



**G-SRS** Structure

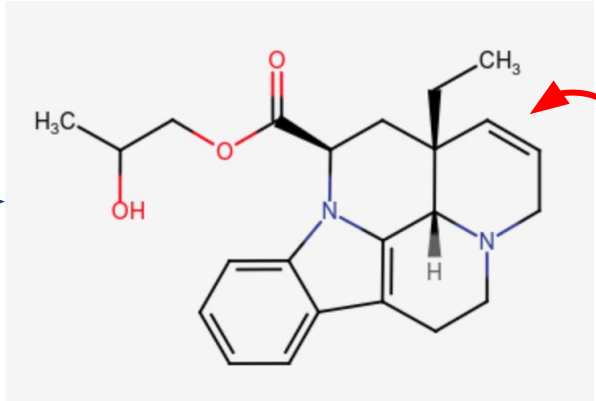
# G-SRS: MolVec

## Evaluation Use Case

**USAN Image**




MolVec

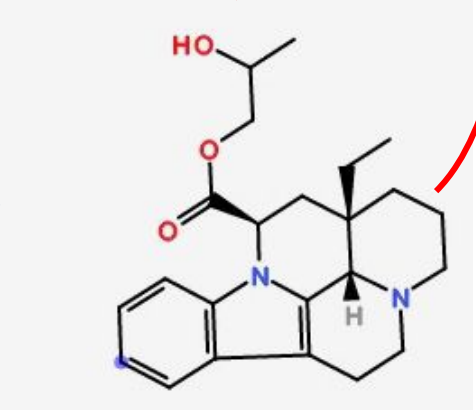



?

File Name

Name or Code

G-SRS REST API

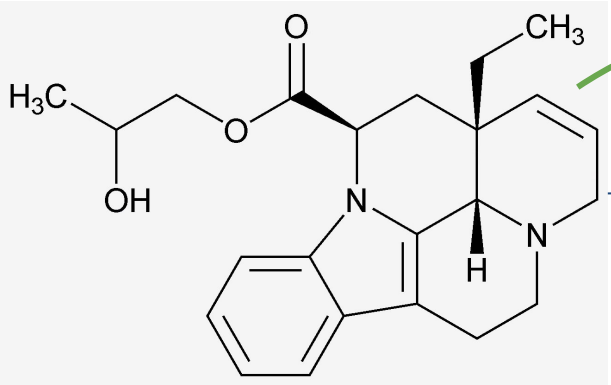



**G-SRS** Structure

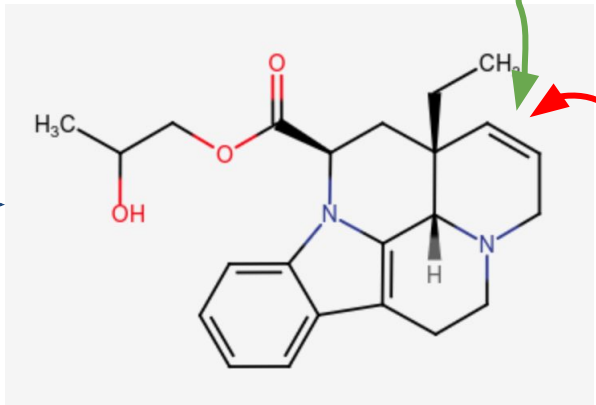
# G-SRS: MolVec

## Evaluation Use Case

**USAN Image**




MolVec

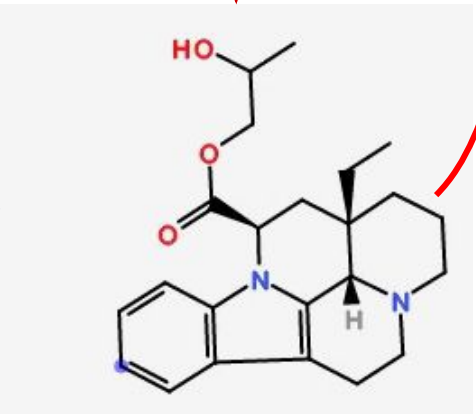



?

File Name

Name or Code

G-SRS REST API



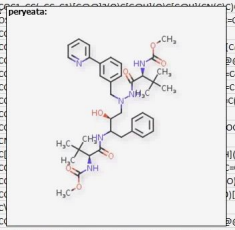

**G-SRS** Structure

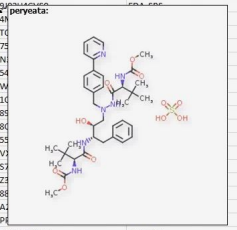
# G-SRS: MolVec

## Evaluation Use Case

- **20%** of "wrong" **MolVec** images were actually *accurate* to the image
- Either the **USAN Image** or **FDA G-SRS** must be incorrect

| | A | B | C | D | E |
|---|---|---|---|---|---|
| | CAS | Smiles MOLVEC | Smiles GSRS | UNII | Source |
| | 103238-56-8 | C(c1c2ccccc2[nH]c1)C(NCC(O)COc3cccc4c3[nH]cc4)C | CC(c)(CC1=CNC2=C1C=CC=C2)NCC(O)COC3=CC=CC4=C3C=CN4 | 0Q1O6G98Q8 | FDA-SRS |
| | 106083-71-0 | O1[C@]([C@H](C)NC(C1)(C)C)(c2ccc(cc2)Cl)O | Cl.C[C@@H]1NC(C)(C)CO[C@@]1(O)C2=CC(Cl)=CC=C2 | SYD411H23S | FDA-SRS |
| | 121288-39-9 | [C@@H]1(O)[C@@H](CO)O[C@H]([C@@H]1O)N2C3 | NC1=NC2=C(N(CC=C)C(=O)N2[C@@H]3O[C@H](CO)[C@@H](O)[C@H]3O)C(=O)N1 | 9CAS0V66OI | FDA-SRS |
| | 1336-29-4 | c1ccc(C(c2ccc(OC(C)=O)cc2)c3ccc(cc3)OC(C)=O)nc1 | CC(=O)OC1=CC=C(C=C1)C(C2=CC=C(OC(C)=O)C=C2)C3=CC=CC=N3.OC4=CC(=CC(O)=C4O)C( | T0I8L1B7J6 | FDA-SRS |
| | 153242-02-5 | C(=O)(CNC(=O)Nc1cc([C@@](C)([H])C(O)=O)ccc1)N2 | C[C@H](C(O)=O)C1=CC(NC(=O)NCC(=O)N2[C@@H](CS[C@@H]2C3=C(F)C=CC=C3)C(=O)O | G2M24F1I2A | FDA-SRS |
| | 187219-99-4 | O[C@]1([C@H]([CC[C@H](C1)O)CN(C)C)c2ccccc(c2)OC | C | | |
| | 229975-97-7 | c1cnc(cc1)-c2ccccc(c2)CN[NC([C@H](C)C)(C)C)NC(=O)O | =CC=C1)[C@@H](O)CN(CC2=CC=1 | | |
| | 23580-33-8 | C=1(C(=O)c2c(cc3c(cc(o3)C(=O)O)e2)C1)CC | T | | |
| | 300832-84-2 | C(C)(Nc1nc(cs1)-c2nc3cc(OC)ccc3c(c2)C[C@@H]4C[C | C@H](CCCCCC=C[C@@H]5C[C@] | 75 | |
| | 35457-80-8 | C(=O)(O[C@H]1[C@@H](O[C@H](C[C@@]1(C)O)O[C | @@H]2[C@@H](C)O[C@@H](O[C@H](N | | |
| | 42971-09-5 | C(OCC)(=O)C=1n2c3ccccc3c4CCN5[C@]([C@](CC)(C1) | =C4C=CC=C5 | 54 | |
| | 5061-22-3 | c1(cc2ccccc2cc1)C(C(=O)OCCN3CCN(CC3)CCOC(=O)C | =C3)CC1)C4=CC=CC5=C4C=CC=C5 | W | |
| | 50-76-0 | C1C[C@@H]2N(C(=O)[C@H](NC([C@@]([C@H](OC( | =C(C)C=C2)=C(C)C(=O)C(N)=C4C(= | 1( | |
| | 54340-62-4 | c1cc(CC)c2occ(C(CNC(C)(C)C)O)c2c1 | C | 85 | |
| | 57694-27-6 | C(C(O)C)OC([C@@H]1n2c3c(cccc3)c4c2[C@]5(N(CC4 | OCC(C)O)C5=C4C=CC=C5 | 8( | |
| | 60-56-0 | S=C1N(CC=N1)C | CH | 55 | |
| | 66112-59-2 | C(=O)(N[C@@H]1[C@H]([C@@H]([C@@H](CO)OC1 | H](O)[C@H](O)CO)[C@@H](NC(C | V1 | |
| | 66644-81-3 | C(c1c(c(OC)c(cc1C)S(N)(=O)=O)OC)(NCC2CCCN2CC= | =C | S1 | |
| | 69739-16-8 | c1(c(sc(n1)SCC2=C(N3[C@@]([C@](C3=O)([H])NC(=O | O)=O)S3)=C(N2C1=O)C(O)=O)C4= | Z3 | |
| | 74817-61-1 | [C@@H]1([C@H](O[C@@H](C(=O)N[C@@H](C)C(=O | )][C@@H](C)O[C@@H]([C@H](O | 8( | |
| | 75078-91-0 | c1ccccc1\C(\C)=C\c2cc3c(cc2)C(C)(C)CCC3(C)C | A1 | | |
| | 84-36-6 | c1(OC)cc(cc(c1OC)OC)C(=O)O[C@@H]2C[C@]3([C@@ | B@H]3CN4CCC5=C(NC6=C5C=CC(P | | |
| | 85118-44-1 | OC(c1nc2c(c(c1)NC)cc3c(c2CCC)OC(C(O)=O)CC3=O)=O | CCCC1=C2OC(=CC(=O)C2=CC3=C(NC)C=C(N=C13)C(O)=O)C(O)=O | F37A9VKY5Q | FDA-SRS |
| | 86227-47-6 | C(CCC/C=C\C/C=C\C/C=C\C/C=C\C/C=C\CC)(OCC)=O | CCOC(=O)CCC\C=C/C\C=C/C\C=C/C\C=C/C\C=C/CC | 6GC8A4PAYH | FDA-SRS |
| | 93-54-9 | c1c(cccc1)CCCO | CCC(O)C1=CC=CC=C1 | 0F8970O4M | FDA-SRS |
| | 99248-33-6 | [C@H]1N(C([C@H](c2ccccc2)NC([C@H]([C@@H](C)C | CC(O)=O.CC(C)[C@@H]1NC(=O)[C@H](CCCCN)NC(=O)[C@@H](CC2=CNC3=C2C=CC=C3)N | URM20C99DG | FDA-SRS |

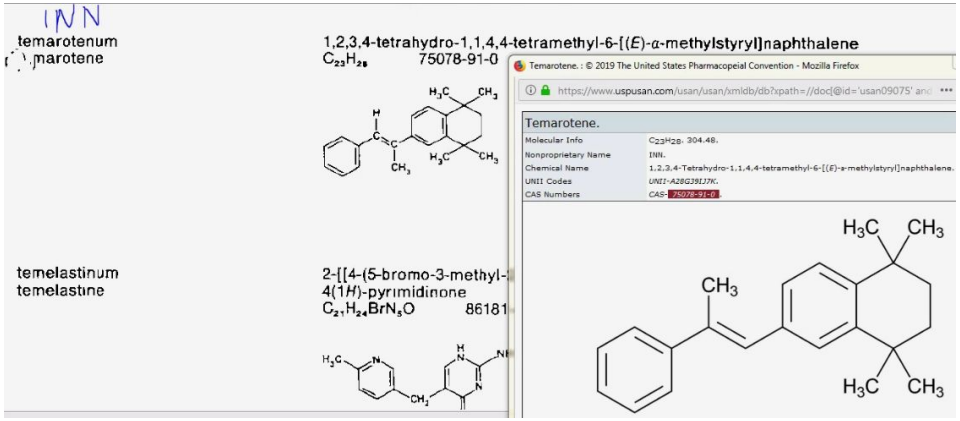


# G-SRS: MolVec

## Evaluation Use Case

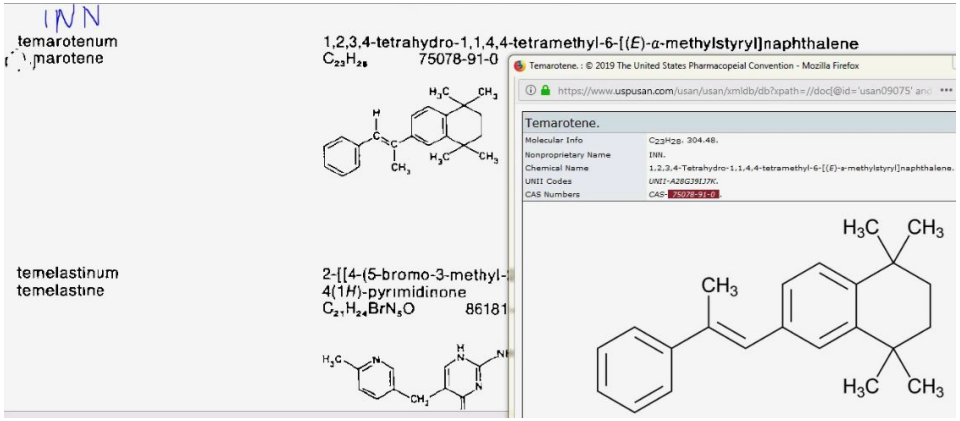
- **1** structure was shown to be wrong in G-SRS, and corrected
- **25** structures were shown to be wrong in the **USAN Images**
  - All but **9** had already been fixed (all changes minor)



# G-SRS: MolVec

## Evaluation Use Case

- **1** structure was shown to be wrong in G-SRS, and corrected
- **25** structures were shown to be wrong in the **USAN Images**
  - All but **9** had already been fixed (all changes minor)

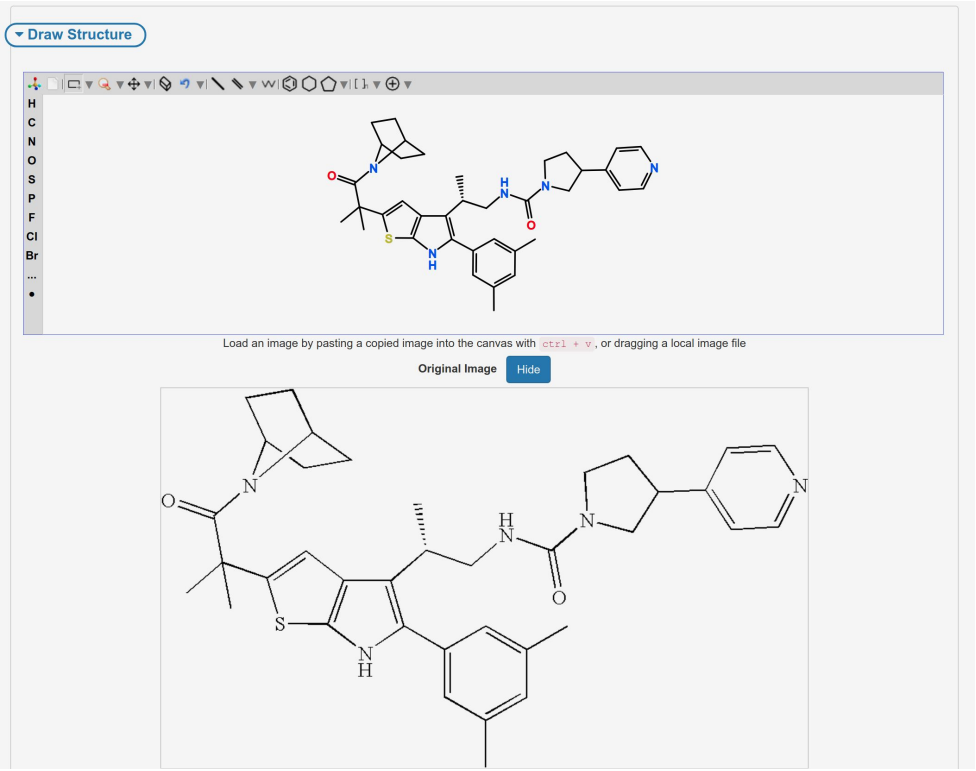



**MolVec**, while imperfect, can help in bulk curation

# G-SRS: MolVec

## Evaluation Use Case

- Useful to bootstrap structure searches from *inside* G-SRS
- Useful to bootstrap chemical registrations

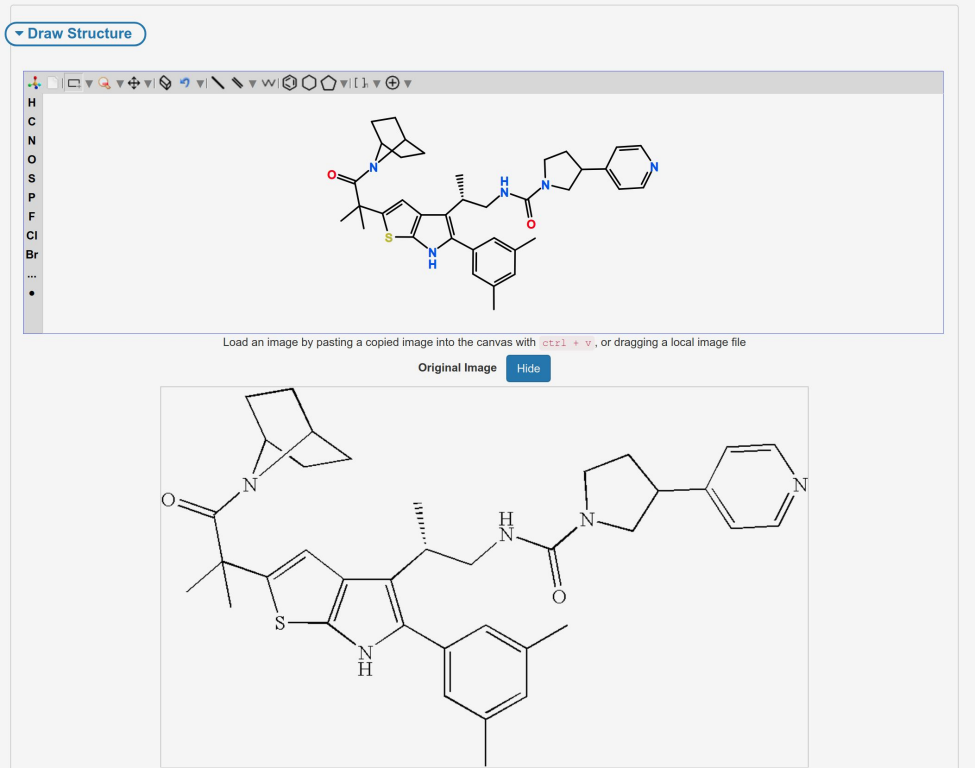


# G-SRS: MolVec

**DEMO**

# G-SRS: MolVec

- Currently used within FDA both via API and via G-SRS registration forms / search
- Rapidly improving due to feedback



# G-SRS: MolVec

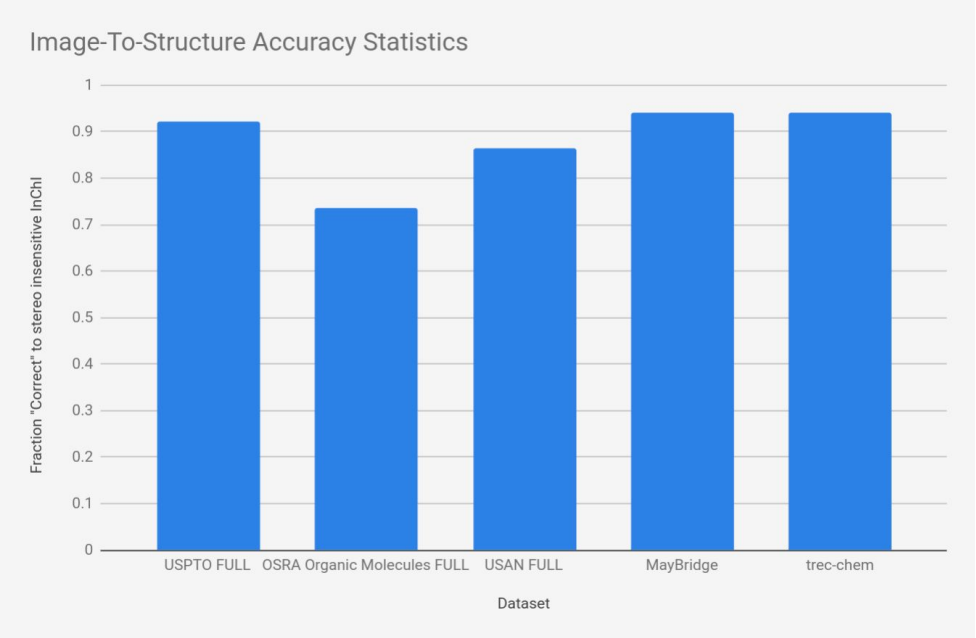
## Features

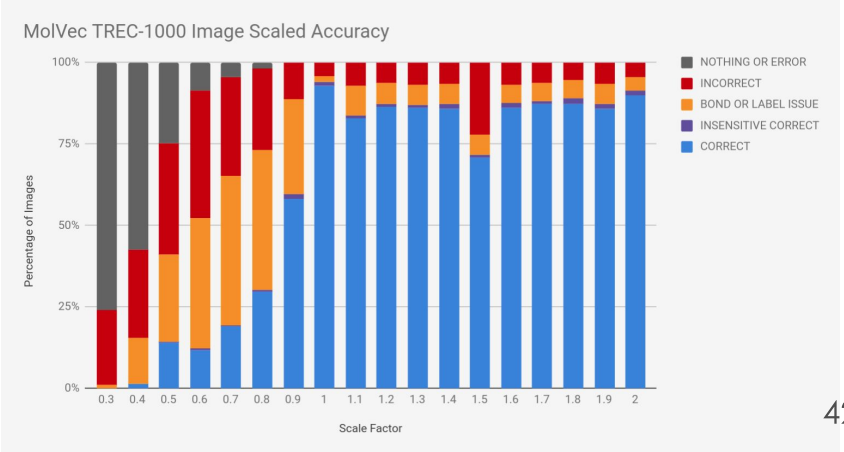
- **Free, self-contained Java code, distributable, no server configuration**
- Pretty accurate
- Accurately pretty
- Fairly Fast

# G-SRS: MolVec

## Features

- Free, self-contained Java code, distributable, no server configuration
- **Pretty accurate**
- Accurately pretty
- Fairly Fast

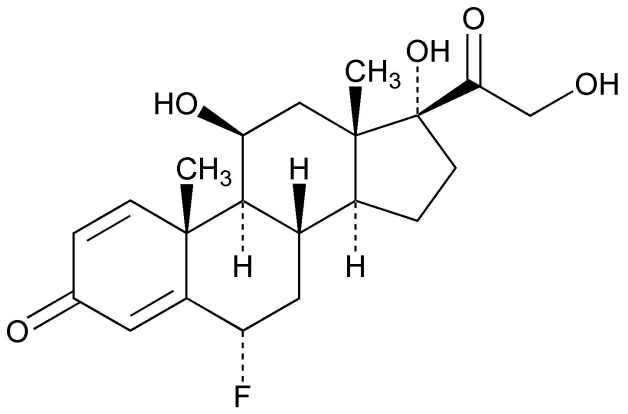


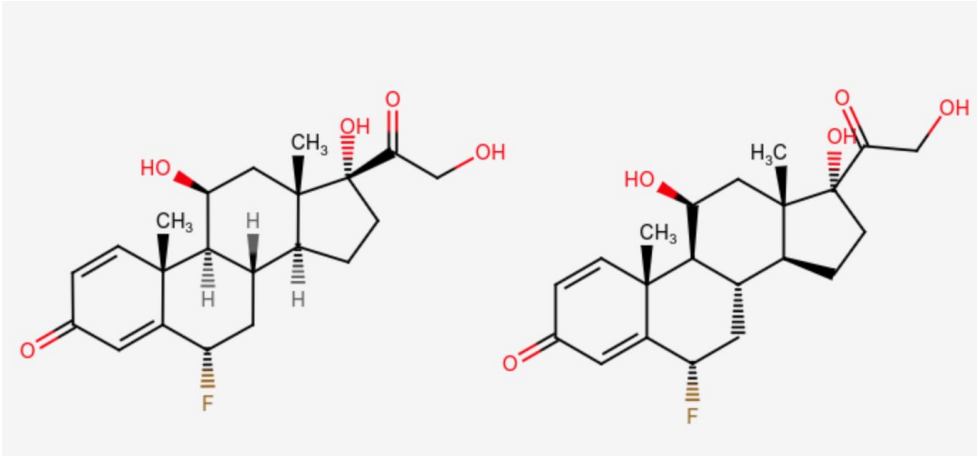



# G-SRS: MolVec

## Features

- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- **Accurately pretty**
- Fairly Fast



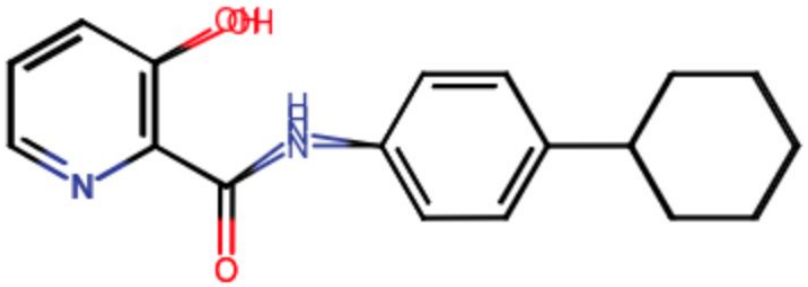

From MolVec From IUPAC Name

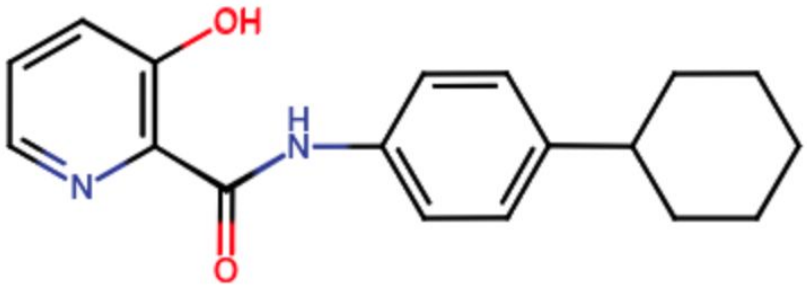
# G-SRS: MolVec

## Features

- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- **Accurately pretty**
- Fairly Fast

Pretty Good Alignment To Image



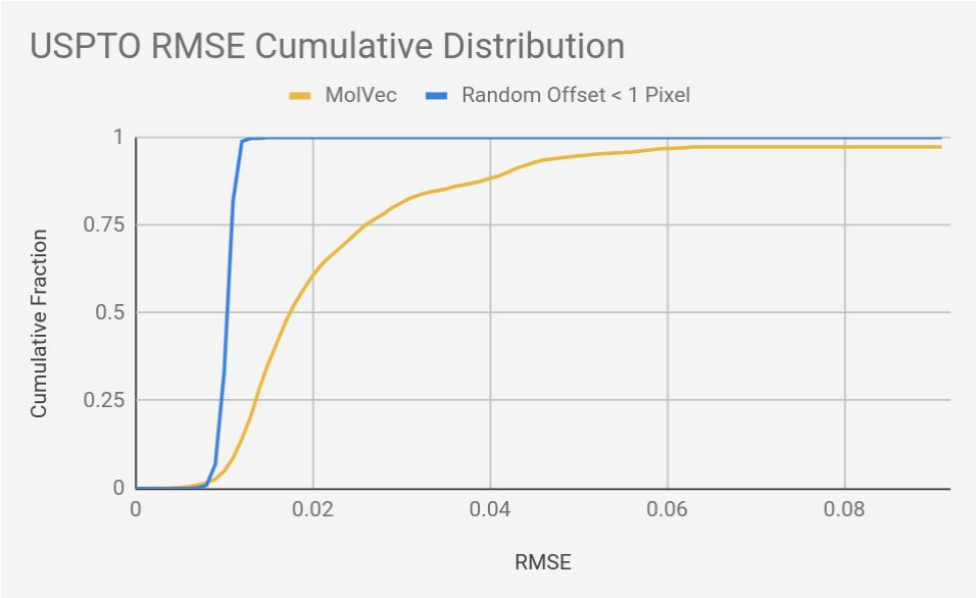

Better Alignment To Image

# G-SRS: MolVec

Evaluation Use Case

## Features

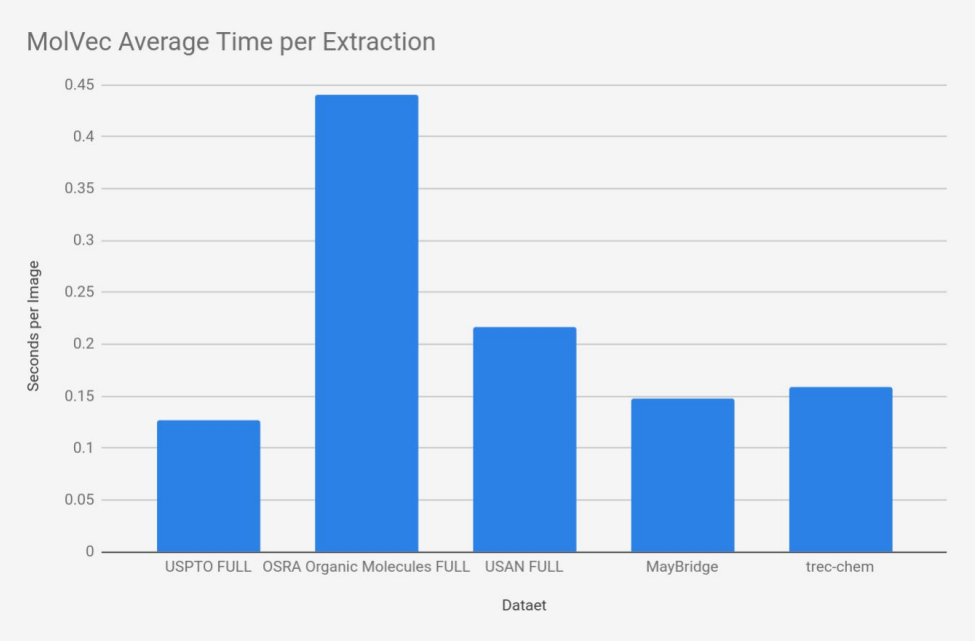
- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- **Accurately pretty**
- Fairly Fast



# G-SRS: MolVec

## Features

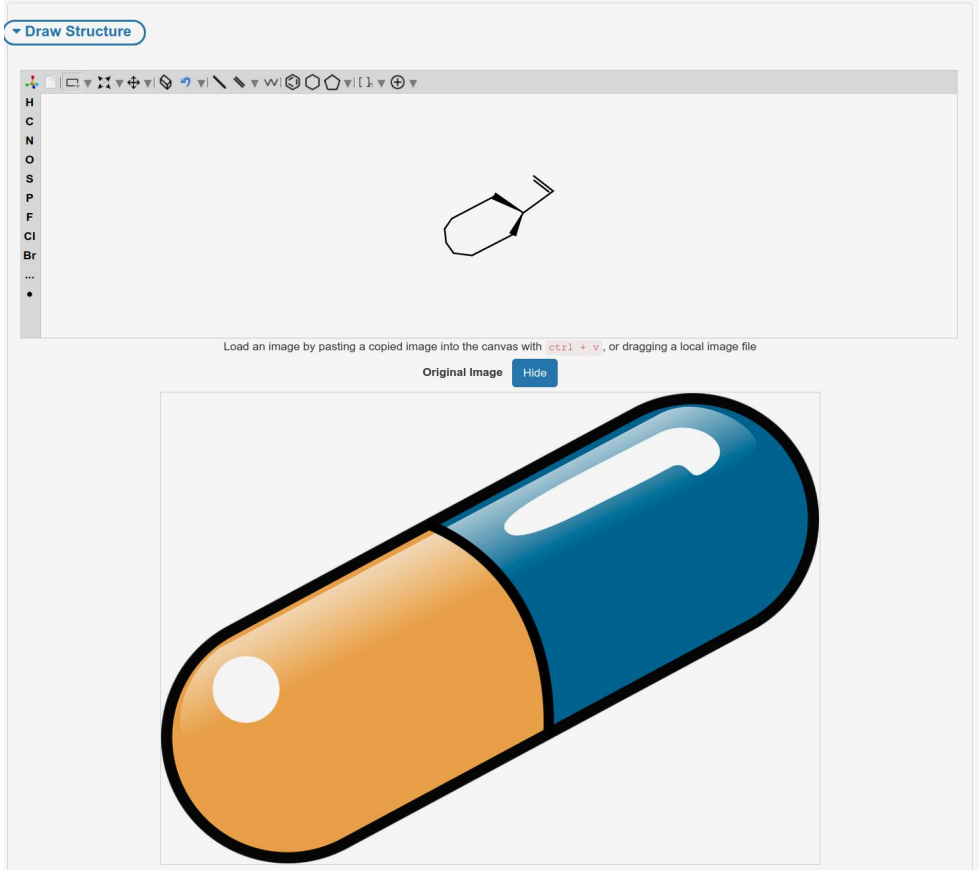
- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- Accurately pretty
- **Fairly Fast**



# G-SRS: MolVec

## Features

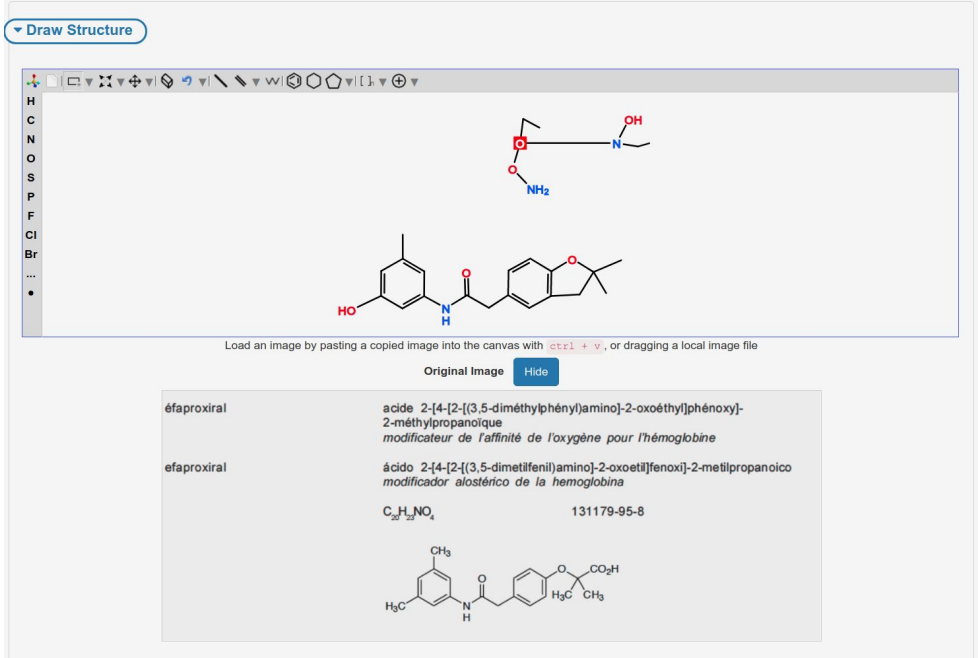
- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- Accurately pretty
- Fairly Fast
- **NOT MAGIC**



# G-SRS: MolVec

## Features

- Free, self-contained Java code, distributable, no server configuration
- Pretty accurate
- Accurately pretty
- Fairly Fast
- **NOT MAGIC**



# G-SRS: MolVec

**Try it out Online:**
https://ginas.ncats.nih.gov/ginas/app/structure

or

https://predictor.ncats.io

(structure prediction -> use image)

**As part of GSRS 2.3.3:**

https://tripod.nih.gov/ginas/

**MolVec Git Repository:**

https://spotlite.nih.gov/ncats/molvec

# Acknowledgements


## NCATS

Tyler Peryea
Danny Katzel
Jorge Neyra
Niko Anderson
Mitch Miller
Sarah Stemann
Noel Southall
Dac-Trung Nguyen
Ivan Grishagin
Dammika Amugoda
Chris LeClair
Paul Shinn

## FDA

Larry Callahan
Frank Switzer
Archana Newatia
Yulia Borodina
Ramez Ghazzaoui
Elaine Johansen
Ta-Jen Chen
Mary-Ann Slack
Alex Welsch
Yoshiyuki Tokiwa
George Washburn
Sabrina Mosley

## Medicines Evaluation Board

Herman Diederik
Marcel Hoefnagel
Joris Kampmeijer
Ciska Matai
Burt Kroes

## US Pharmacopeial Convention

Fouad Atouf
Andrej Wilk
Tina Morris
Ahmed Abdulhadi
Steven Emrick

**Health Canada**

Craig Anderson
Daniel Buijs
Ted Kim

# Questions

# Thank You