

Mixing small molecules and macromolecules in the world of informatics

Alex M. Clark

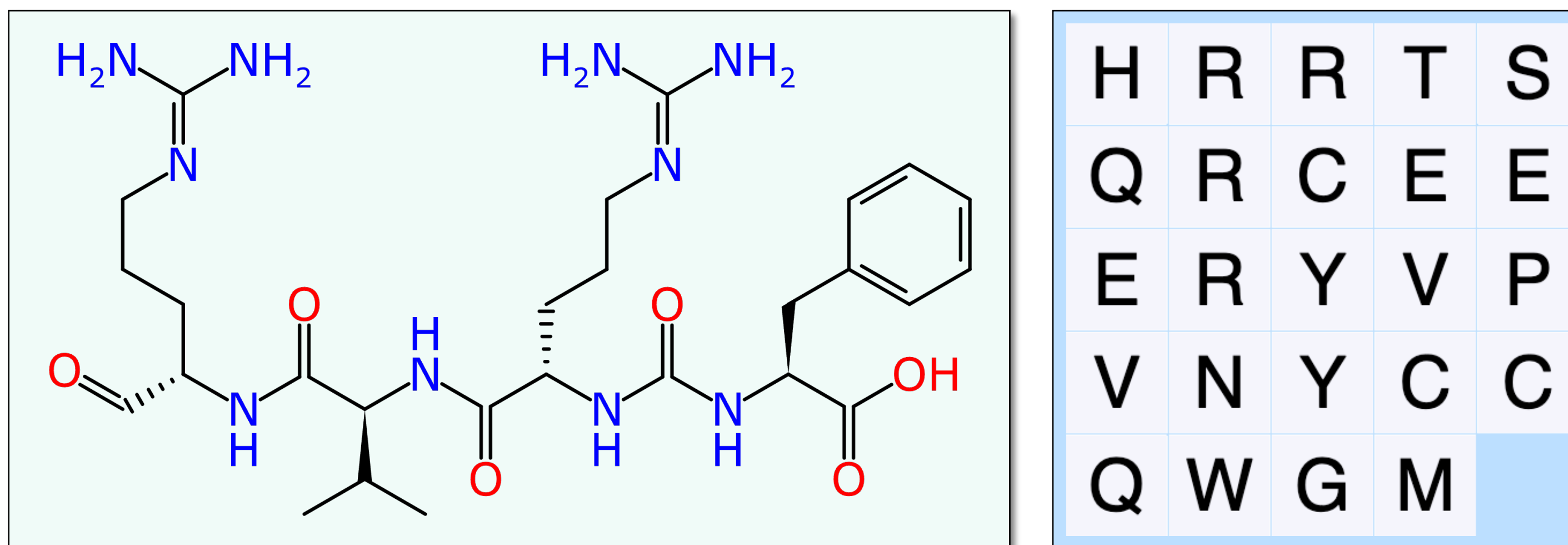
alex@collaborativedrug.com



CDD VAULT[®]
Complexity Simplified



Two Worlds

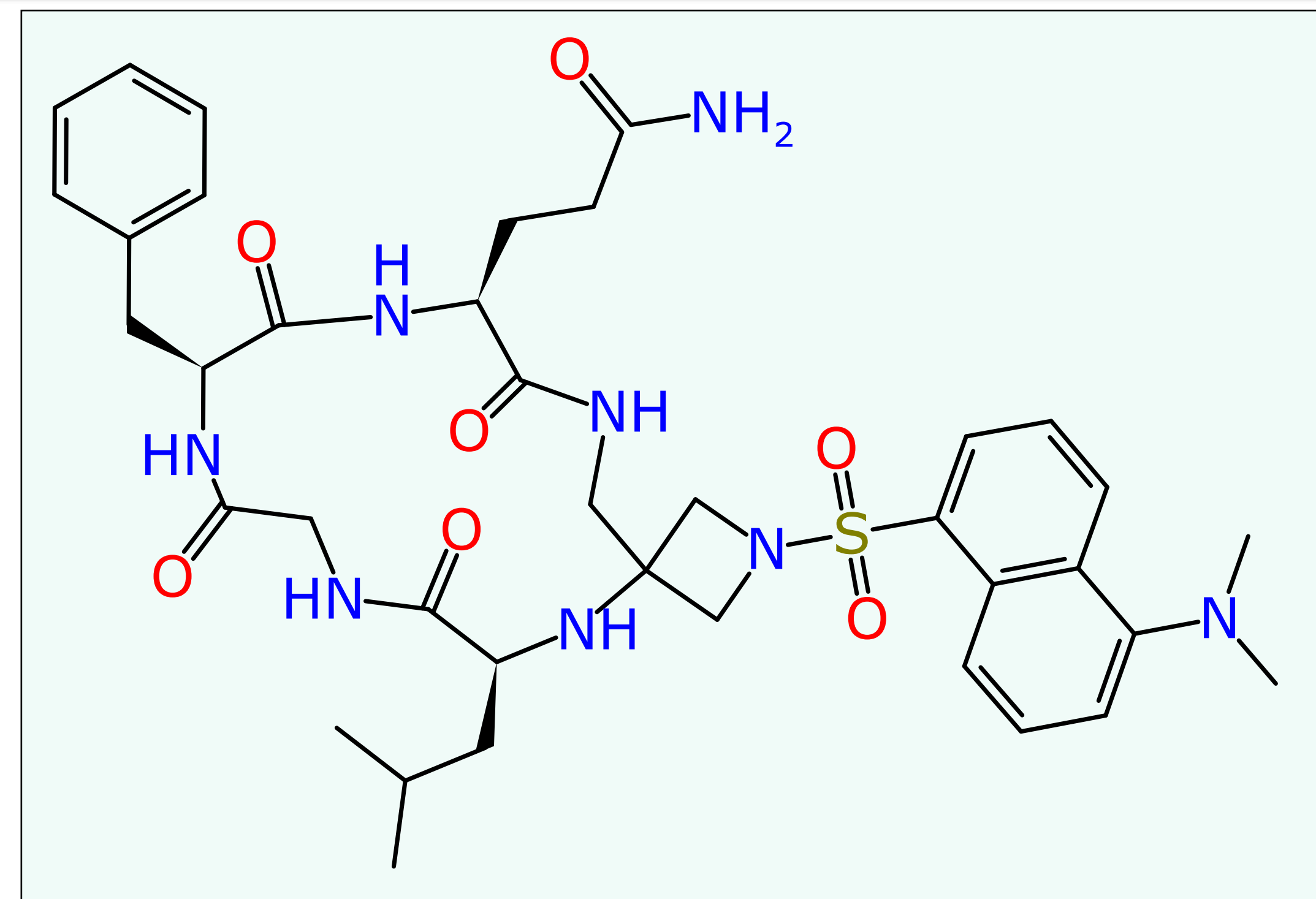
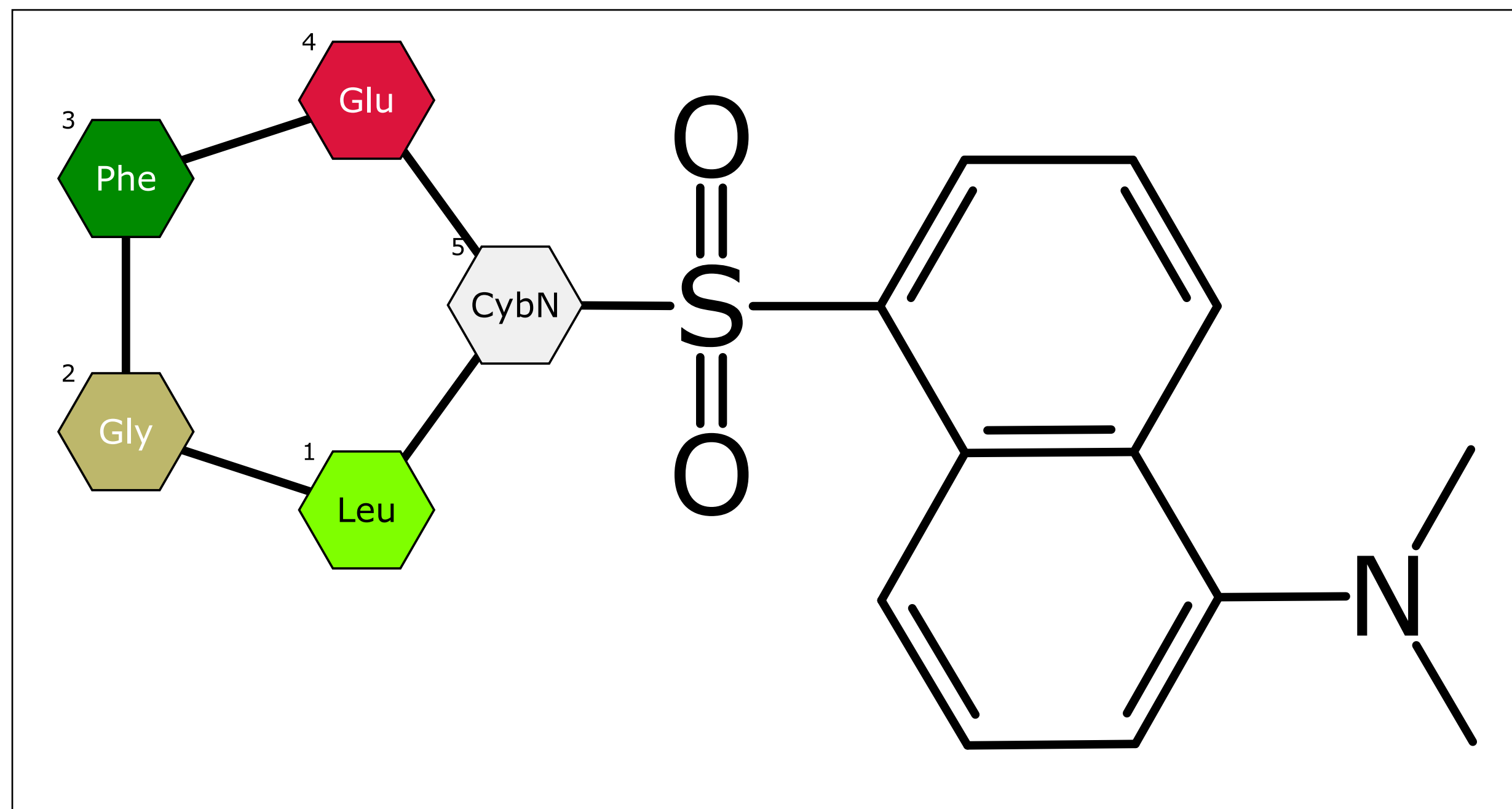


- ❖ **Cheminformatics**: graphs of atoms & bonds
- ❖ **Bioinformatics**: sequences of monomers
- ❖ Different datastructures, different science, different algorithms...
- ❖ ... most traditional applications suit one or the other.

Intersection

L,G,F,E,CybN-SO₂NphNMe₂*

- ❖ Needing sequences *and* atom graphs is not really niche anymore
- ◆ custom peptides & nucleotides
- ◆ chemical functionalization
- ❖ Scale is making this an informatics problem

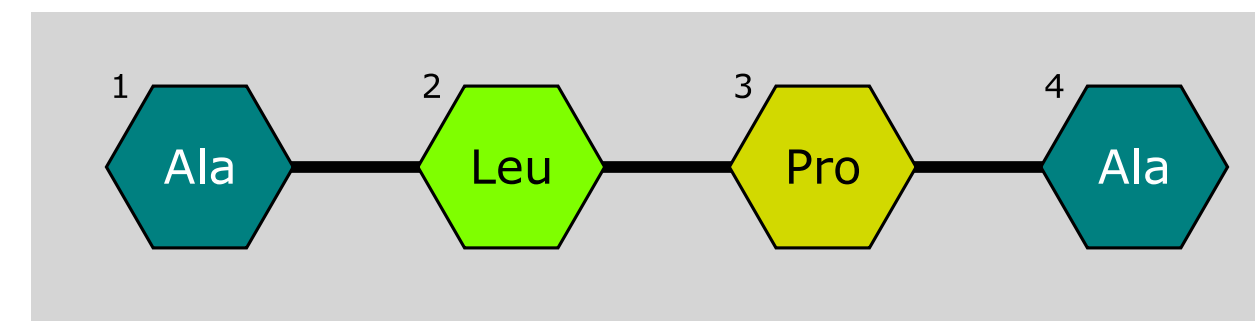




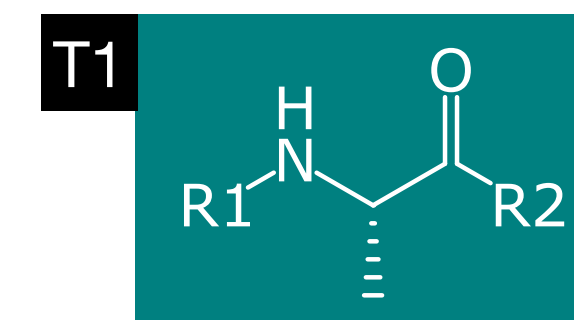
Representation

- ❖ Line notations: HELM, multiple vendor-specific options
- ❖ *or*: there is a V3000 molfile feature called **SCSR** (self-contained sequence representation), since ~2010 but rarely used
 - ◆ stores **layout** information, as a 2D sketch
 - ◆ monomers as **templates**: scalable
 - ◆ self contained, no external **dictionary**
 - ◆ full **atomic** structure is implied
 - ◆ contains the **sequence** data for bioinformatics

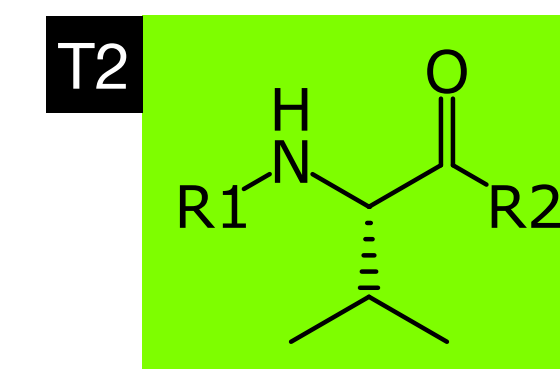
main block



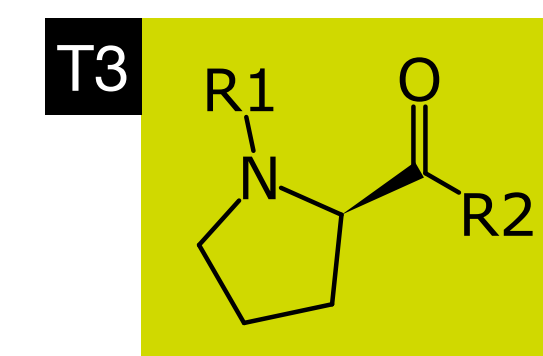
Ala



Leu



Pro



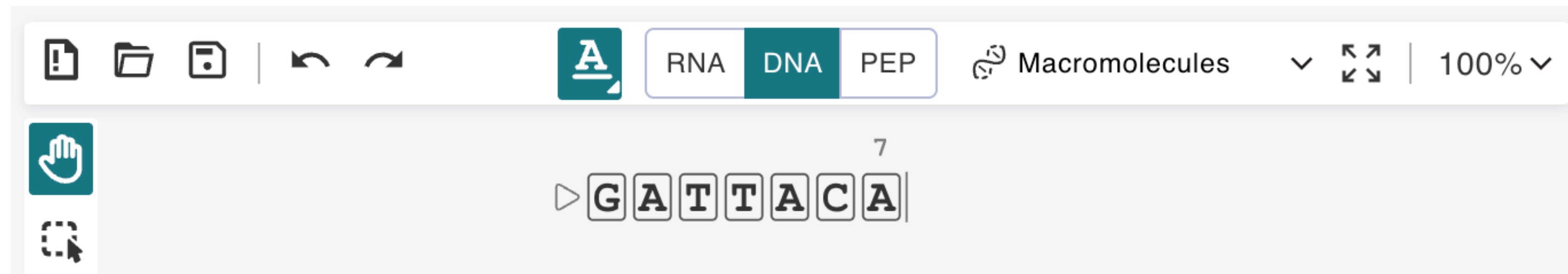


Software Support

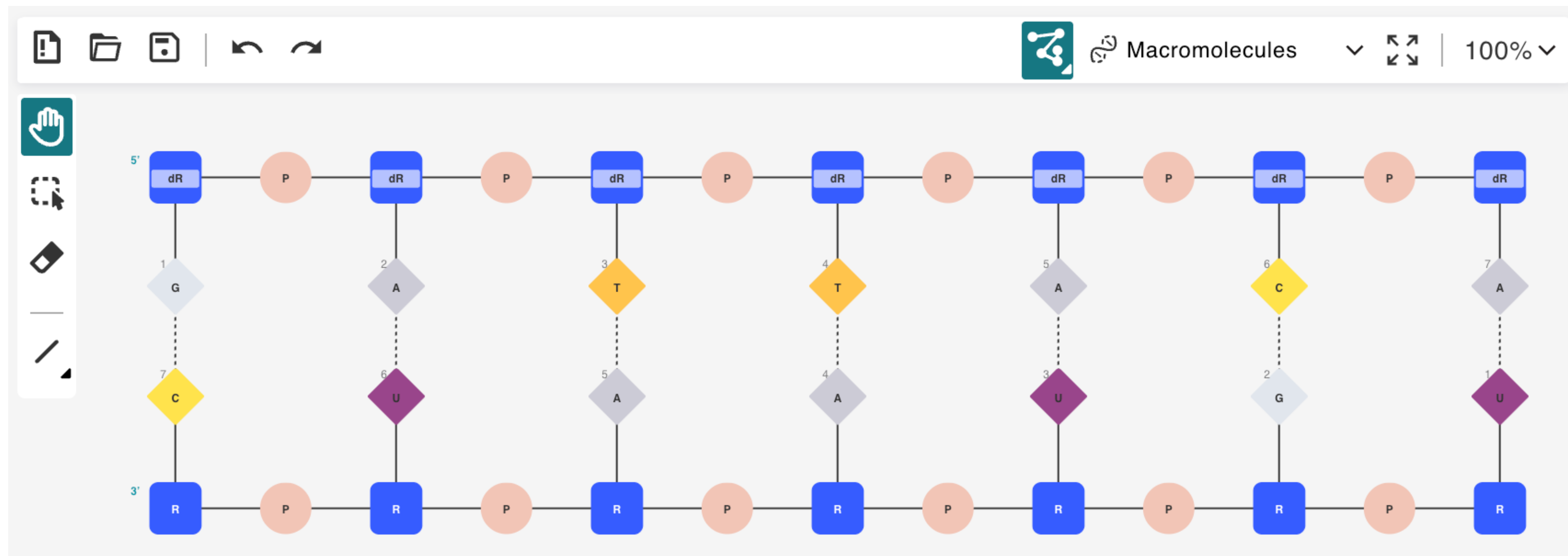
- ❖ It is a V3000 standard, but documentation is extremely sparse
- ❖ Reference implementation is ***Biovia Draw***
- ❖ Open source implementation: ***Ketcher*** from EPAM
 - ✦ ... with sponsorship from **Collaborative Drug Discovery**
- ❖ Integration, rendering and workflow in ***CDD Vault***
- ❖ Contributed code to ***RDKit*** to enable all-atom calculations
- ❖ Recent developments as of late 2024: beginning of an ecosystem?

Drawing macromolecules - Ketcher

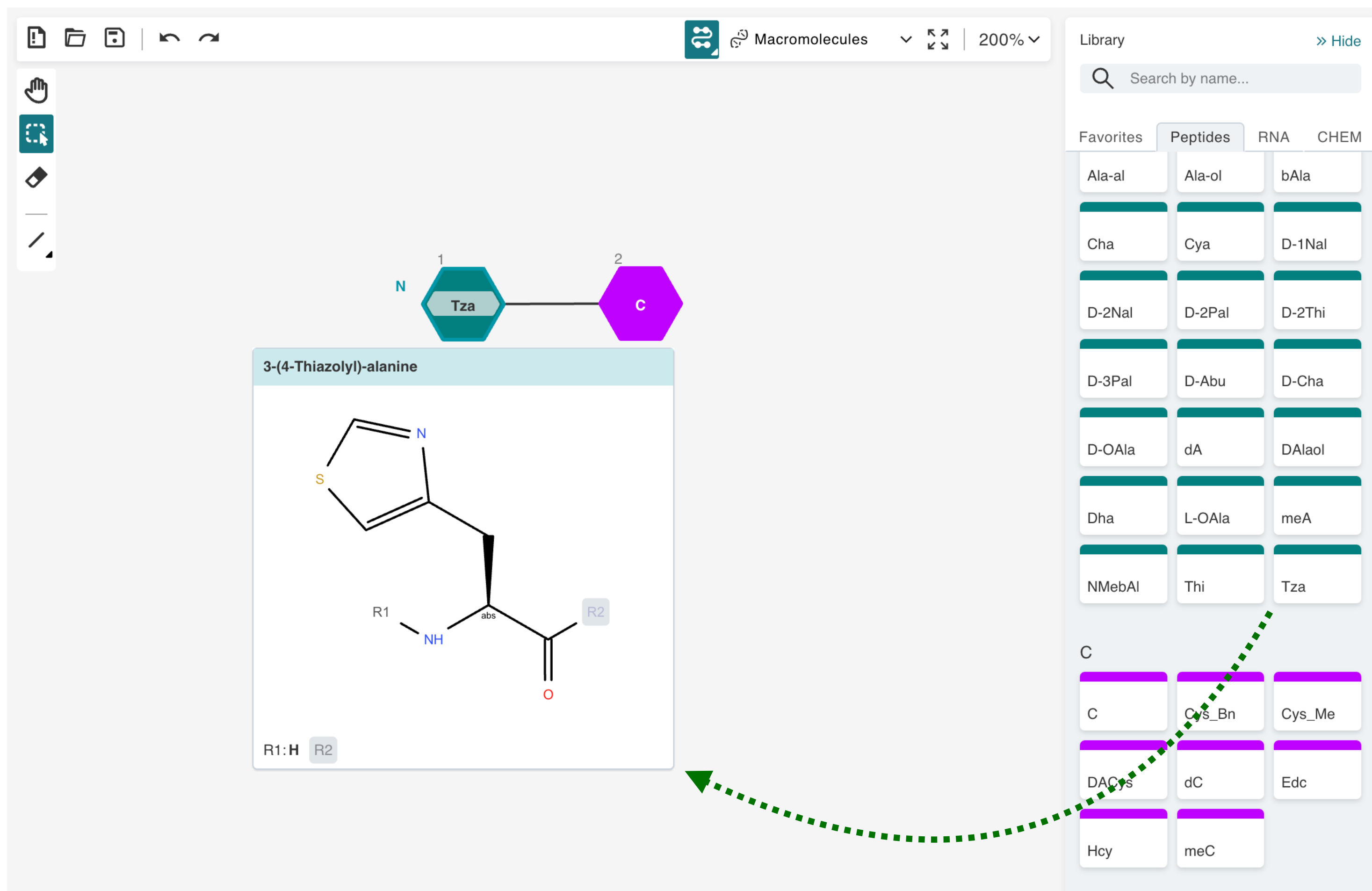
≥3.1



❖ Don't be fooled by the typewriter font... switch to monomer view:

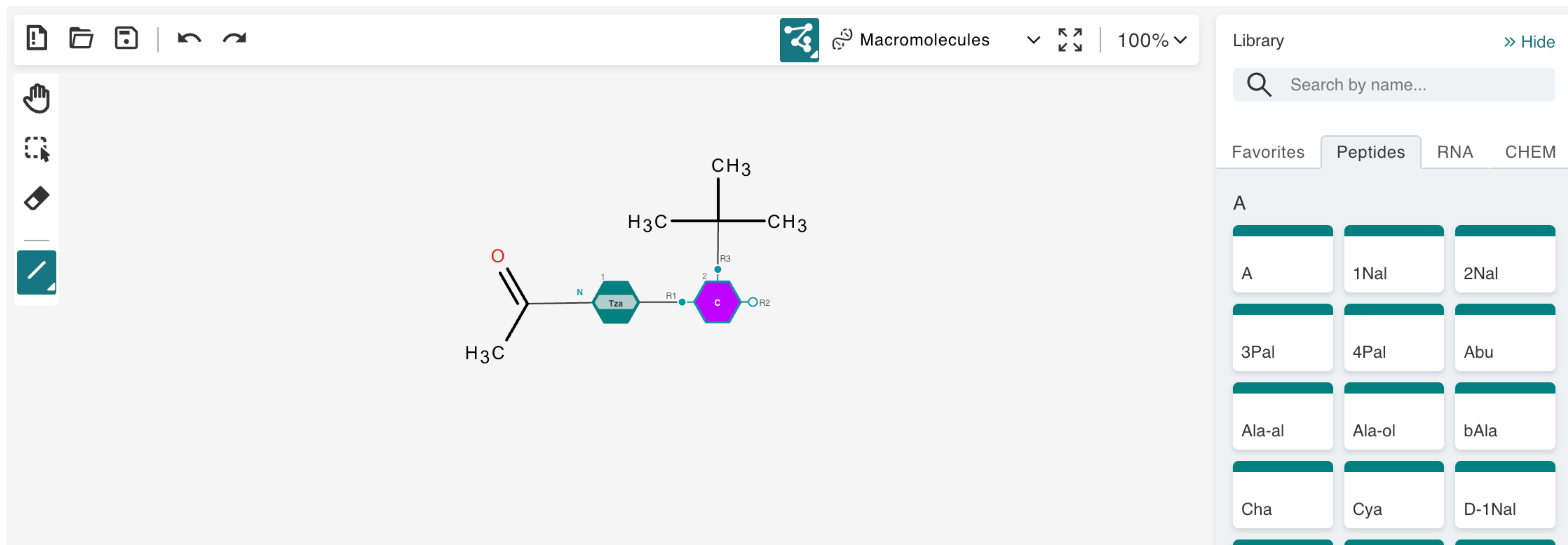


Synthetic monomers



- ❖ Full sketcher capabilities: draw it however you want it
- ❖ Common synthetic peptides and nucleotides
- ❖ Custom monomers coming soon...
 - ◆ Create using Ketcher
 - ◆ Manage and share when using in CDD Vault

Just add chemistry



- ❖ Most monomers have connection points (R1, R2, R3, ...)
- ❖ Draw chemical functionality using the regular molecule sketcher...
- ❖ ... and connect it up.

Under the hood: V3000 all the way down



```
-INDIGO-03112510242D
0 0 0 0 0 0 0 0 0 0 0 0 V3000
M V30 BEGIN CTAB
M V30 COUNTS 9 8 0 0 0
M V30 BEGIN ATOM
M V30 1 C 3.075 0.2 0.0 0
M V30 2 C 4.075 0.2 0.0 0
M V30 3 C 2.075 0.2 0.0 0
M V30 4 C 3.075 1.2 0.0 0
M V30 5 C -0.325 -1.275 0.0 0
M V30 6 O -0.825 -0.408975 0.0 0
M V30 7 C -0.825 -2.14103 0.0 0
M V30 8 Tza 1.25 -1.25 0.0 0
    CLASS=AA SEQID=1
    ATTCHORD=(4 9 Br 5 Al)
M V30 9 C 3.0625 -1.24375 0.0 0
    CLASS=AA SEQID=2
    ATTCHORD=(4 8 Al 1 Cx)
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 1 2
M V30 2 1 1 3
M V30 3 1 1 4
M V30 4 2 5 6
M V30 5 1 5 7
M V30 6 1 8 9
M V30 7 1 9 1
M V30 8 1 8 5
M V30 END BOND
M V30 END CTAB
M V30 BEGIN TEMPLATE
```

```
M V30 TEMPLATE 1 AA/Tza/Tza/ NATREPLACE=AA/A
M V30 BEGIN CTAB
M V30 COUNTS 12 12 3 0 0
M V30 BEGIN ATOM
M V30 1 C 0.8925 0.4863 0.0 0
M V30 2 C 0.8948 -1.0144 0.0 0 CFG=1
M V30 3 C 2.1952 -1.7637 0.0 0
M V30 4 N -0.4056 -1.7637 0.0 0
M V30 5 O 2.197 -2.9637 0.0 0
M V30 6 O 3.2337 -1.1625 0.0 0
M V30 7 H -1.4447 -1.1635 0.0 0
M V30 8 N -0.5669 2.7257 0.0 0
M V30 9 C -0.4057 1.2344 0.0 0
M V30 10 C -1.7742 0.6202 0.0 0
M V30 11 S -2.7811 1.7319 0.0 0
M V30 12 C -2.035 3.0333 0.0 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 2 1 CFG=1
M V30 2 1 4 2
M V30 3 1 2 3
M V30 4 2 3 5
M V30 5 1 3 6
M V30 6 1 4 7
M V30 7 1 8 9
M V30 8 2 9 10
M V30 9 1 10 11
M V30 10 1 11 12
M V30 11 2 8 12
M V30 12 1 9 1
M V30 END BOND
M V30 BEGIN COLLECTION
M V30 MDLV30/STEABS ATOMS=(1 2)
M V30 END COLLECTION
M V30 BEGIN SGROUP
M V30 1 SUP 1 ATOMS=(1 7) XBONDS=(1 6) LABEL=H CLASS=LGRP
M V30 2 SUP 2 ATOMS=(1 6) XBONDS=(1 5) LABEL=OH CLASS=LGRP
M V30 3 SUP 3 ATOMS=(10 1 2 3 4 5 8 9 10 11 12) XBONDS=(2 5 6)
    LABEL=Tza CLASS=AA SAP=(3 4 7 Al) SAP=(3 3 6 Br)
    NATREPLACE=AA/A
M V30 END SGROUP
M V30 END CTAB
```

```
M V30 TEMPLATE 2 AA/Cys/C/ NATREPLACE=AA/C
M V30 BEGIN CTAB
M V30 COUNTS 9 8 4 0 0
M V30 BEGIN ATOM
M V30 1 N -0.866 -1.0 0.0 0
M V30 2 C 0.0 -0.5 0.0 0 CFG=2
M V30 3 H -1.732 -0.5 0.0 0
M V30 4 C 0.866 -1.0 0.0 0
M V30 5 O 1.732 -0.5 0.0 0
M V30 6 O 0.866 -2.0 0.0 0
M V30 7 C 0.0 0.5 0.0 0
M V30 8 S -0.866 1.0 0.0 0
M V30 9 H -0.866 2.0 0.0 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 1 2
M V30 2 1 1 3
M V30 3 1 2 4
M V30 4 1 4 5
M V30 5 2 4 6
M V30 6 1 2 7 CFG=1
M V30 7 1 7 8
M V30 8 1 8 9
M V30 END BOND
M V30 BEGIN COLLECTION
M V30 MDLV30/STEABS ATOMS=(1 2)
M V30 END COLLECTION
M V30 BEGIN SGROUP
M V30 1 SUP 1 ATOMS=(1 3) XBONDS=(1 2) LABEL=H CLASS=LGRP
M V30 2 SUP 2 ATOMS=(1 5) XBONDS=(1 4) LABEL=OH CLASS=LGRP
M V30 3 SUP 3 ATOMS=(1 9) XBONDS=(1 8) LABEL=H CLASS=LGRP
M V30 4 SUP 4 ATOMS=(6 1 2 4 6 7 8) XBONDS=(3 2 4 8) LABEL=C
    CLASS=AA SAP=(3 1 3 Al) SAP=(3 4 5 Br) SAP=(3 8 9 Cx)
    NATREPLACE=AA/C
M V30 END SGROUP
M V30 END CTAB
M V30 END TEMPLATE
M END
```

Bulk markup

- ❖ At informatics scale
- ❖ Real workflows: preparing new oligomers daily
 - ◆ probably have design tools that can emit sequences
 - ◆ need to register them: too many for a spreadsheet
- ❖ **CDD Vault** has a dedicated *import composer* for this

CDD VAULT · McKerrow Vault

AI Chat · Help · Log out

Explore Data · ELN · Inventory · **Import Data** · Reports · Settings · 4 · Full-Access User

Step 1: Choose Data File and Parse → Step 2: Map Fields → Step 3: Commit Data

File: naturals.csv Project: McKerrow Vault · Owner: Full-Access User

Compose macromolecules from columns [Edit](#)

Not part of macromolecule Wrapping: Off On
 Linear peptide sequence Width: 7 units
 Cyclic peptide sequence
 Nucleotide sequence

3 custom peptides, 0 custom nucleotides

	A	B
	Name	Sequence
1	Oligo1	SVSEIQLMHNLGKHLNSMERVEWLRKKLQDVHNF
2	Oligo2	AVSEHQLLDKGGKSIQDLRRRELLEKLLKLHTA

1 Ala 2 Val 3 Ser 4 Glu 5 His 6 Gln 7 Leu
8 Leu 9 His 10 Asp 11 Lys 12 Gly 13 Lys 14 Ser
15 Ile 16 Gln 17 Asp 18 Leu 19 Arg 20 Arg 21 Arg
22 Glu 23 Leu 24 Leu 25 Glu 26 Lys 27 Leu 28 Leu
29 Lys 30 Leu 31 His 32 Thr 33 Ala

[Process File](#)

CDD VAULT Collaborative Drug Discovery · Vault Updates · Blog



Import modes

❖ Cyclic peptides

❖ RNA

❖ DNA

Not part of macromolecule Backbone: RNA DNA
 Linear peptide sequence Strands: Single Double
 Cyclic peptide sequence
 Nucleotide sequence

3 custom peptides,
0 custom nucleotides

	A	B	
	ID	Sequence	
1	DNA1	GATTACA	
2	DNA2	ACATTAG	
3	DNA3	AC	
4	DNA4		
5	DNA5	GATTATATACACACC	

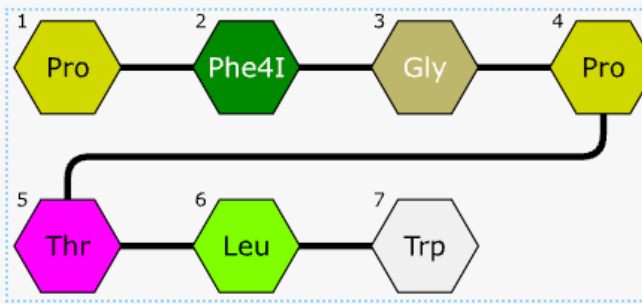
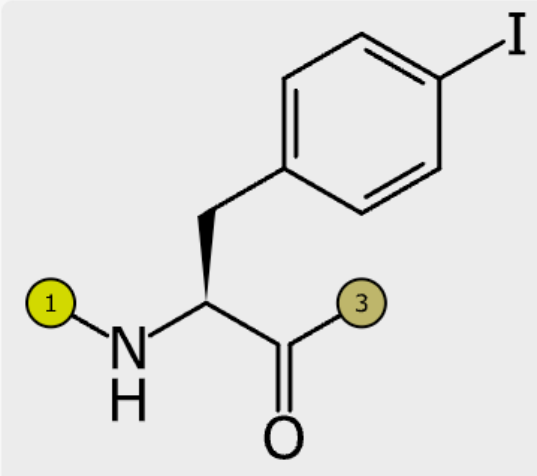
Custom monomers

- ❖ Draw and name monomer fragments with attachment points
- ❖ Store in Vault or use ad hoc

Preview Macromolecule

SVG PNG Molfile Original Identifiers

Size × Full atom layout



○ Not part of macromolecule Wrapping: ○ Off ● On
● Linear peptide sequence Width: 4 units
○ Cyclic peptide sequence
○ Nucleotide sequence

	A	B	
	Name	Sequence	
1	Synth1	Pro,Phe,Ile,Pro,Pro,gGlu,Tyr	
2	Synth2	Pro,Ser,Leu,Tyr,Pro,Tyr,Tyr	
3	Synth3	Pro,Phe4I,Gly,Pro,Thr,Leu,Trp	
4	Synth4	Pro,Phe,Gly,Pro,Gln,LeuF,Trp	
5	Synth5	Pro,Tyr,Asp,Pro,Leu,Ala,Ile	

Download Copy

Registration

- ❖ Identification and duplication
- ❖ Chemical properties

The screenshot displays the CDD.VAULT McKerrow Vault interface. The top navigation bar includes "Explore Data", "ELN", "Inventory", "Import Data", "Reports", "Settings", and a notification icon with "4". The user is logged in as "Full-Access User".

The main content area is titled "Synth3" and shows a chemical structure diagram of a peptide chain with seven residues: Pro (1), Phe4I (2), Gly (3), Pro (4), Thr (5), Leu (6), and Trp (7). The residues are represented by colored hexagons connected by lines.

Below the structure, there are several action items:

- Find molecules with this structure
- Suggest bioisosteres using deep learning similarity
- Find ChEMBL, patented, and commercial molecules using deep learning similarity
- Add to a collection
- Add a batch
- Manage project access
- Delete this molecule

A summary box indicates "Showing data from 1 of 1 project" with the following details:

- Owner: Full-Access User
- Created: March 13, 2025
- Updated: March 19, 2025

The right-hand side of the interface shows the "Overview" tab for the molecule. It includes a table with the following fields:

Name:	Synonyms:	Smell:	Color:
Synth3			
Taste:	Literature Reference:	Description:	

Below the table, there are two sections: "Lipinski Properties" and "Additional Properties".

Lipinski Properties

- Molecular weight: 942.853 g/mol
- log P: -1.0
- H-bond donors: 9
- H-bond acceptors: 17
- Lipinski Rule of 5: Violated (1 of 4 within desirable range)

Additional Properties

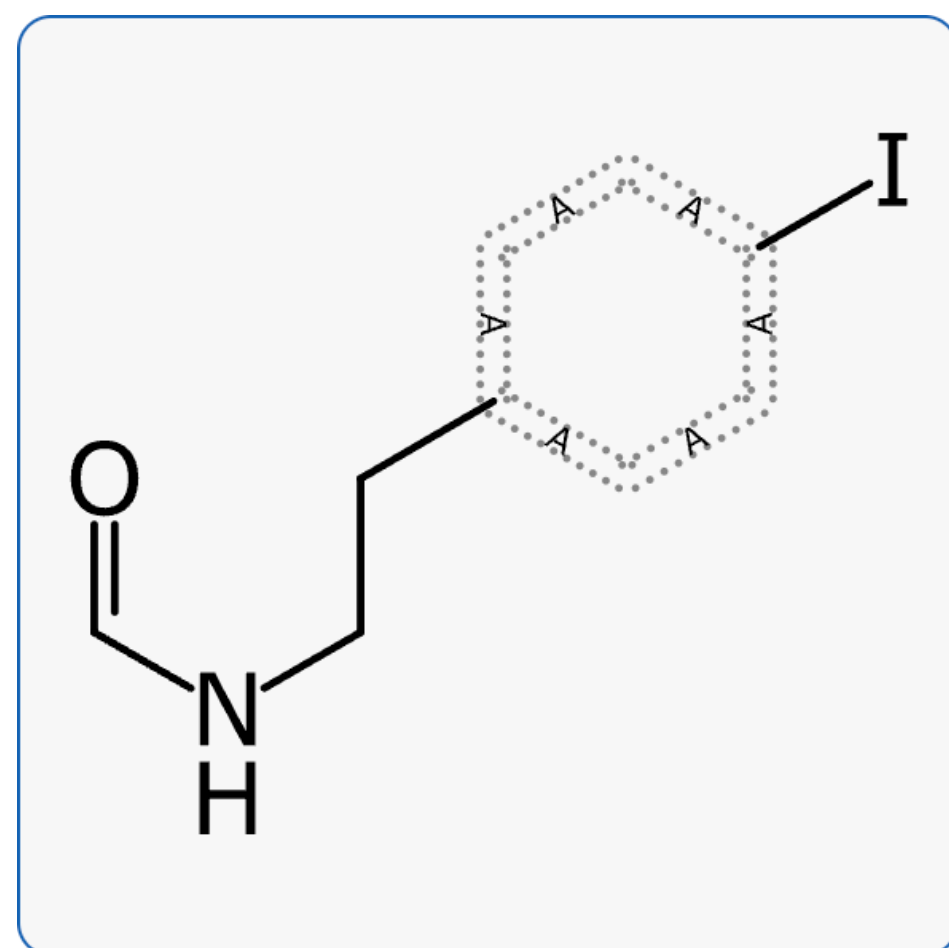
- log D: -1.0
- log S: -5.3
- pK_a: 3.6 (Acidic), 8.2 (Basic)
- CNS MPO score: 2.9
- Topological polar surface area (TPSA): 251.2 Å²
- Exact mass: 942.3137 g/mol
- Fsp3: 0.50
- Heavy atom count: 60
- Rotatable bonds: 19
- Formula: C₄₂H₅₅IN₈O₉
- Composition: C (53.50%), H (5.88%), I (13.46%), N (11.88%), O (15.27%)



Searching



- ❖ Structure search queries operate on the full-atom structure



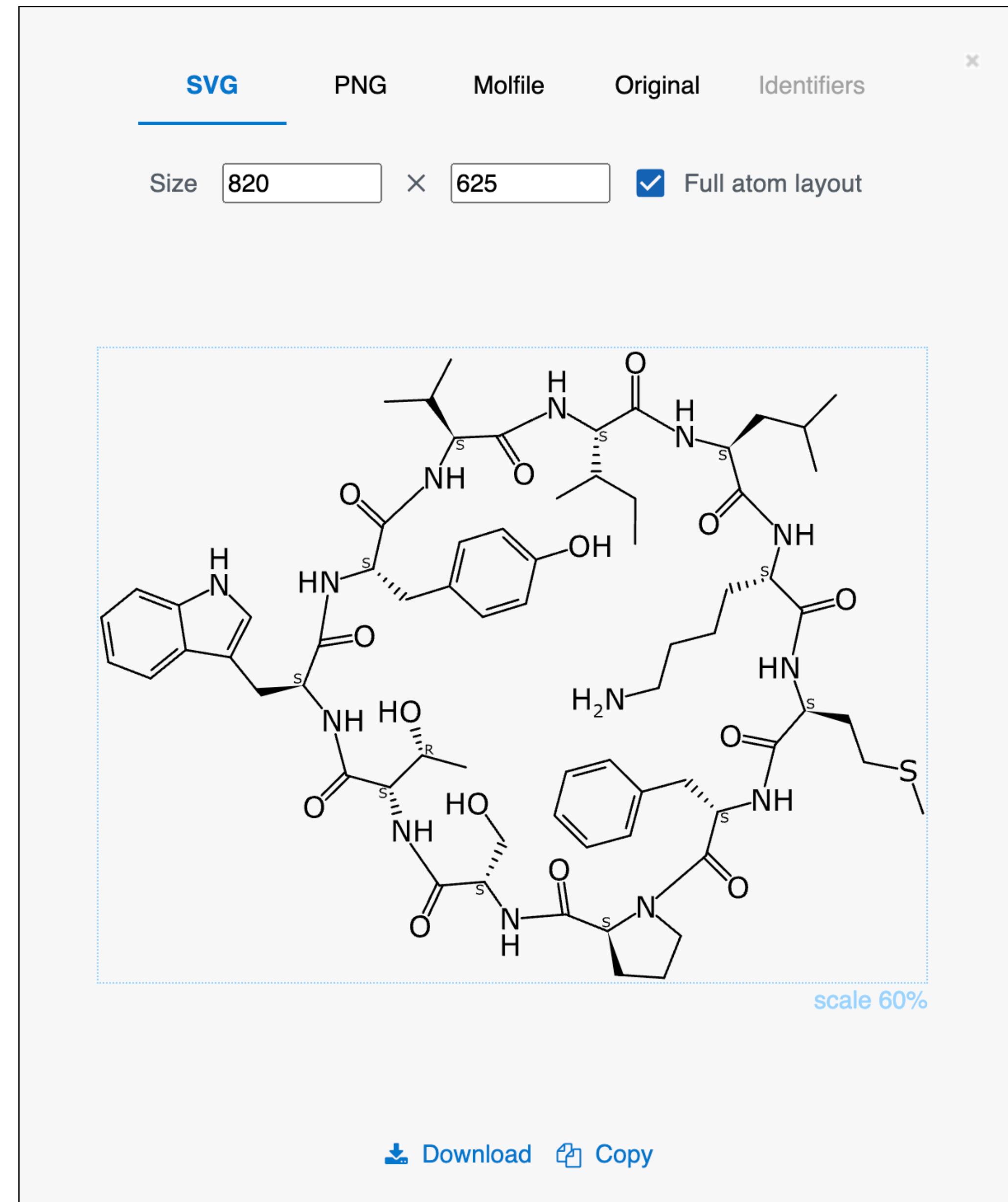
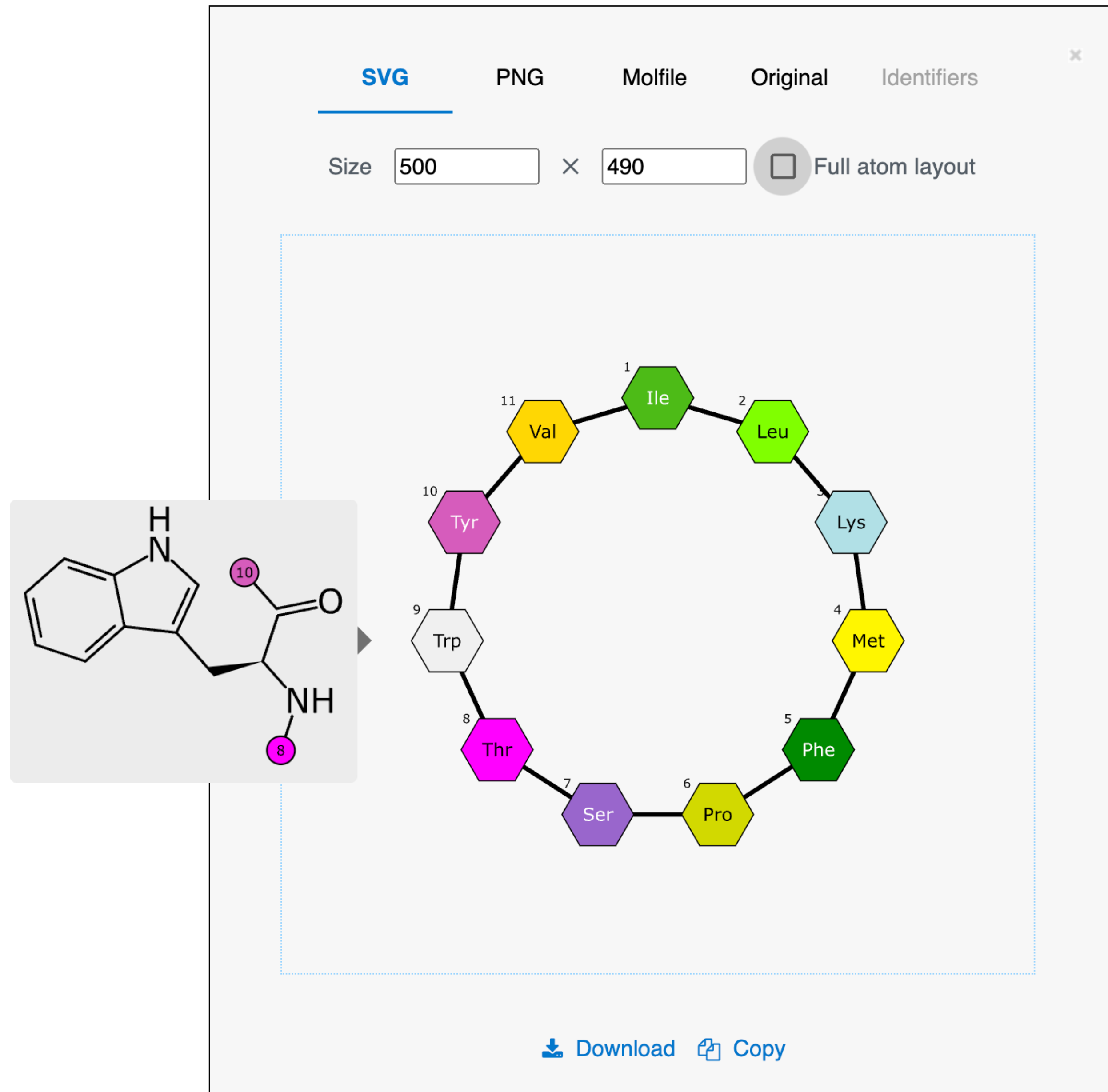
substructure
query

1 Selected: · [Export](#) · [Add to collection](#) · [Build model](#) · [Flag outliers](#) · [Customize your report](#) · [Save this search](#)

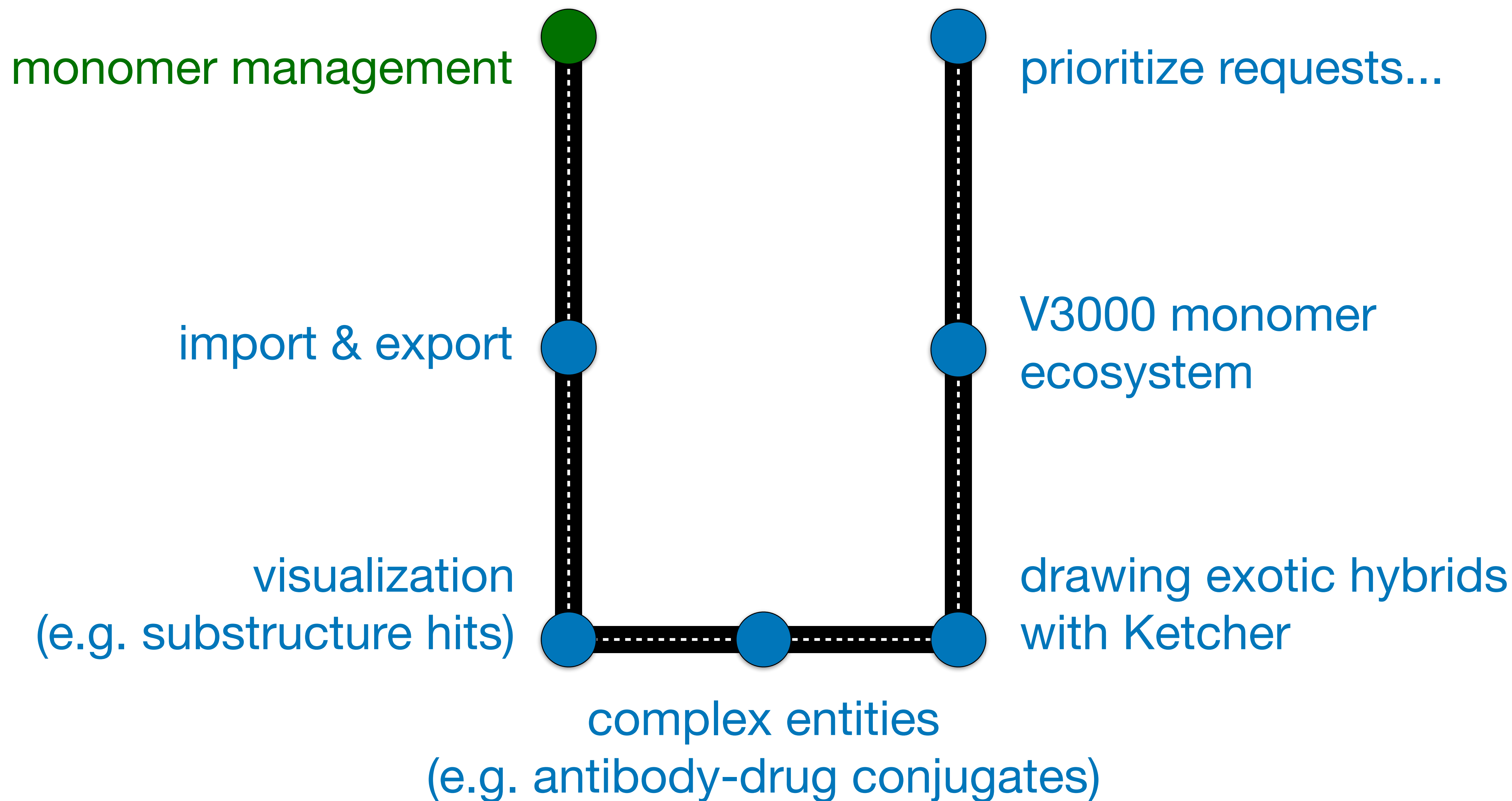
Select...	
all	
·	Molecule
none	

Synth3
McKerrow Vault

Full atom view



Roadmap



Questions?



- ❖ Contact:
 - ◆ Alex M. Clark alex@collaborativedrug.com (Collaborative Drug Discovery)
- ❖ Thanks to the Vault, Ketcher & Research Informatics teams