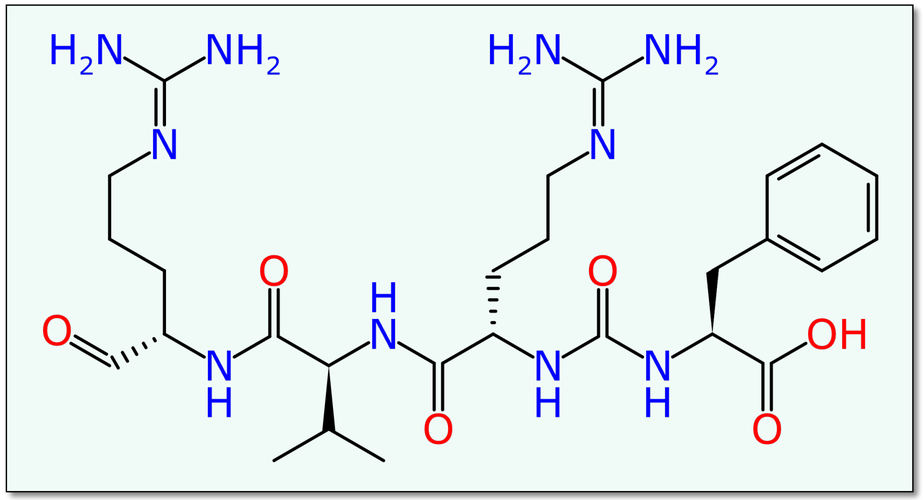
Mixing small molecules and macromolecules in the world of informatics

Alex M. Clark

alex@collaborativedrug.com

Two Worlds

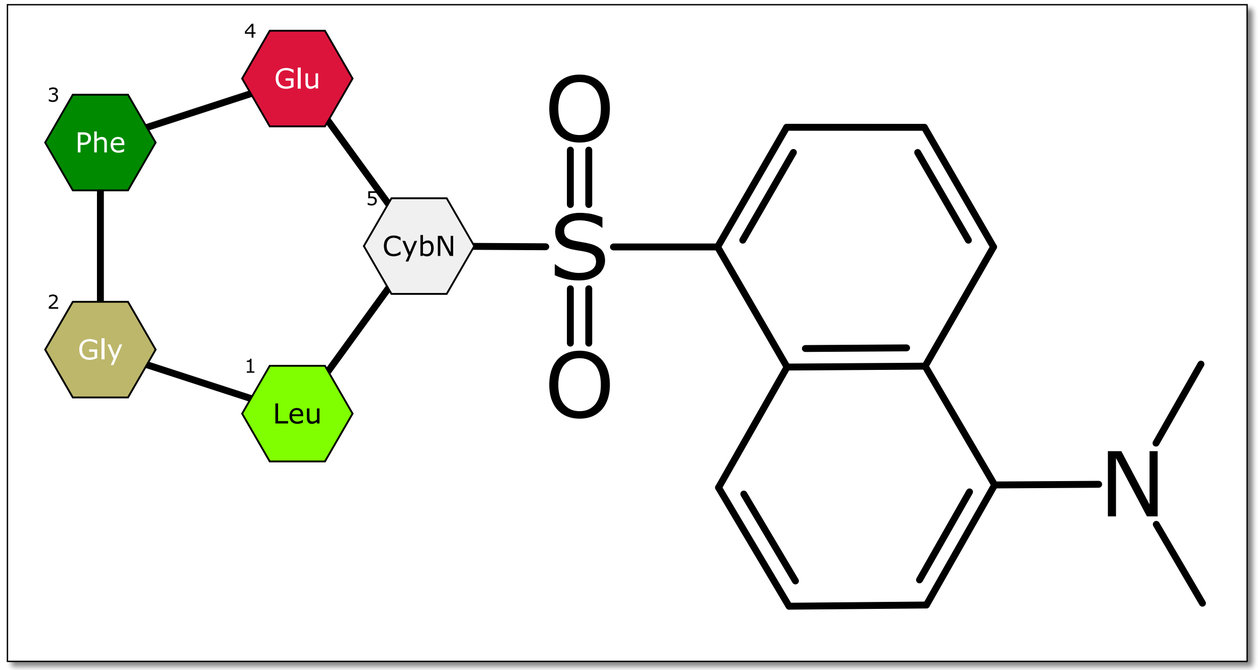


H	R	R	T	S
Q	R	C	E	E
E	R	Y	V	P
V	N	Y	C	C
Q	W	G	M	

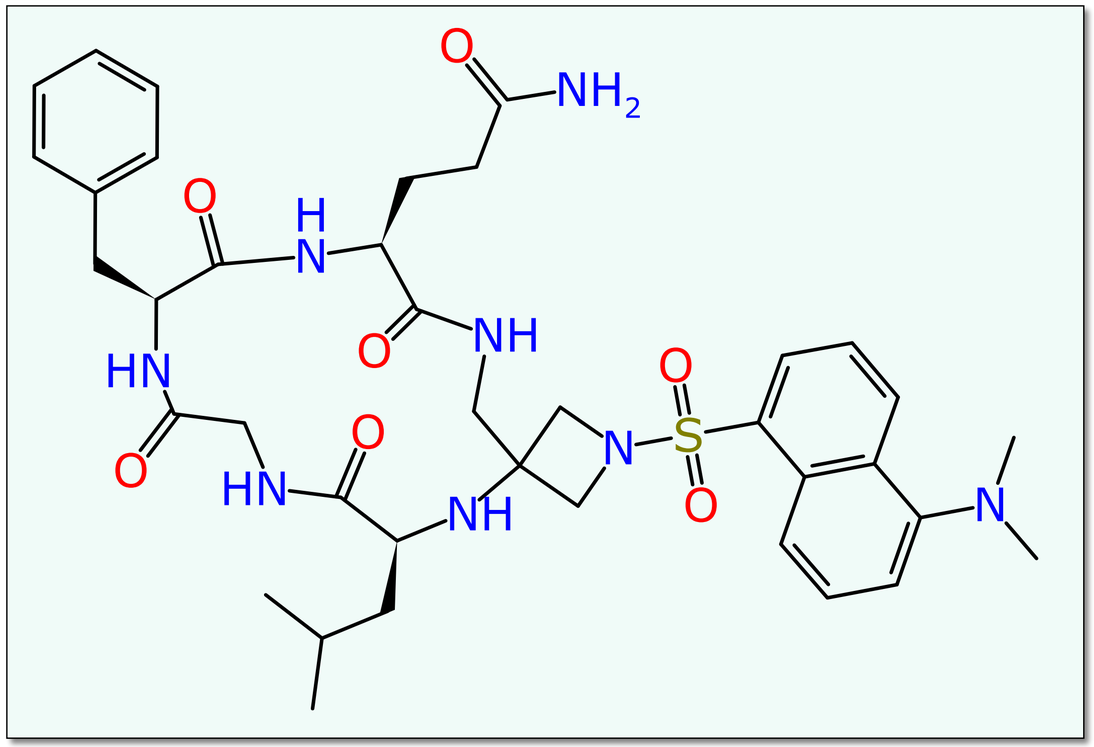
- **Cheminformatics**: graphs of atoms & bonds
- **Bioinformatics**: sequences of monomers
- Different datastructures, different science, different algorithms...
- ... most traditional applications suit one or the other.

Intersection

L,G,F,E,CybN-SO2NphNMe2,*



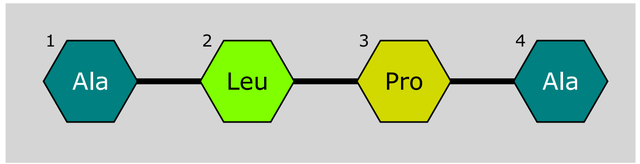
- Needing sequences *and* atom graphs is not really niche anymore
  - custom peptides & nucleotides
  - chemical functionalization
- Scale is making this an informatics problem



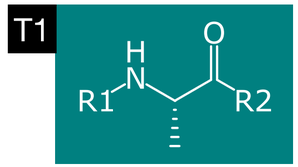
Representation

- 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

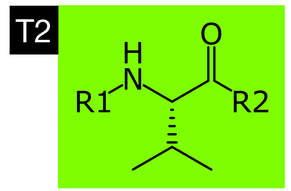
main block



Ala



Leu



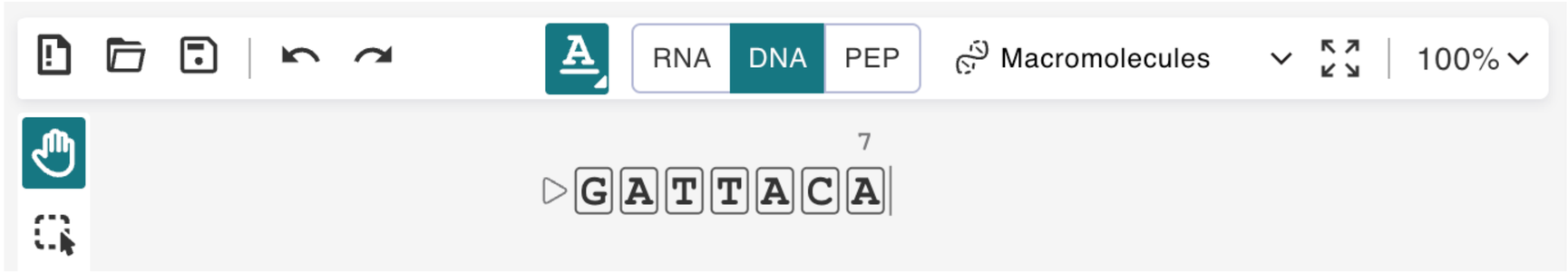
Pro

Software Support

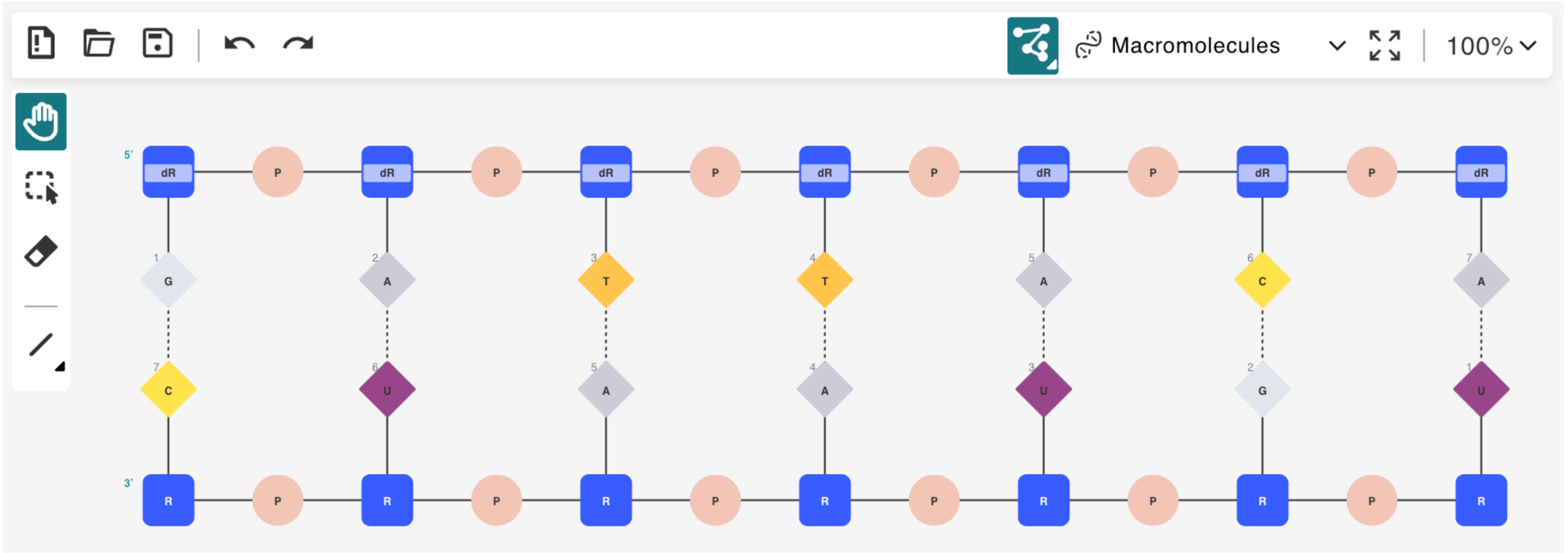
- It is a V3000 standard, but documentation is extremely sparse
- Reference implementation is ***Biovia Draw***
- Open source implementation: ***Ketcher*** from EPAM
  - ... with sponsorship from **Collaborative Drug Discovery**
- Integration, rendering and workflow in ***CDD Vault***
- Contributed code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

Drawing macromolecules - Ketcher

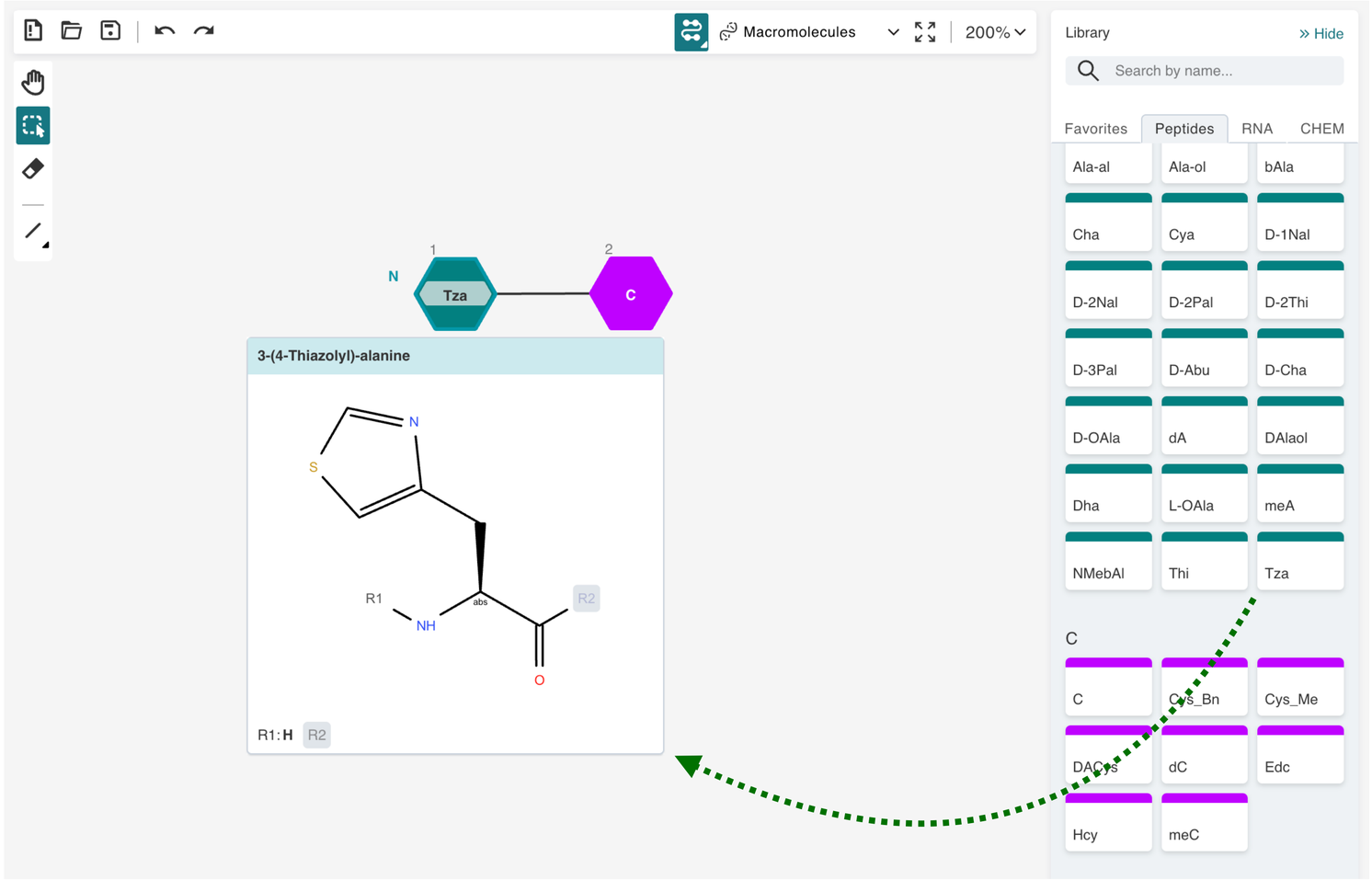
≥3.1



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

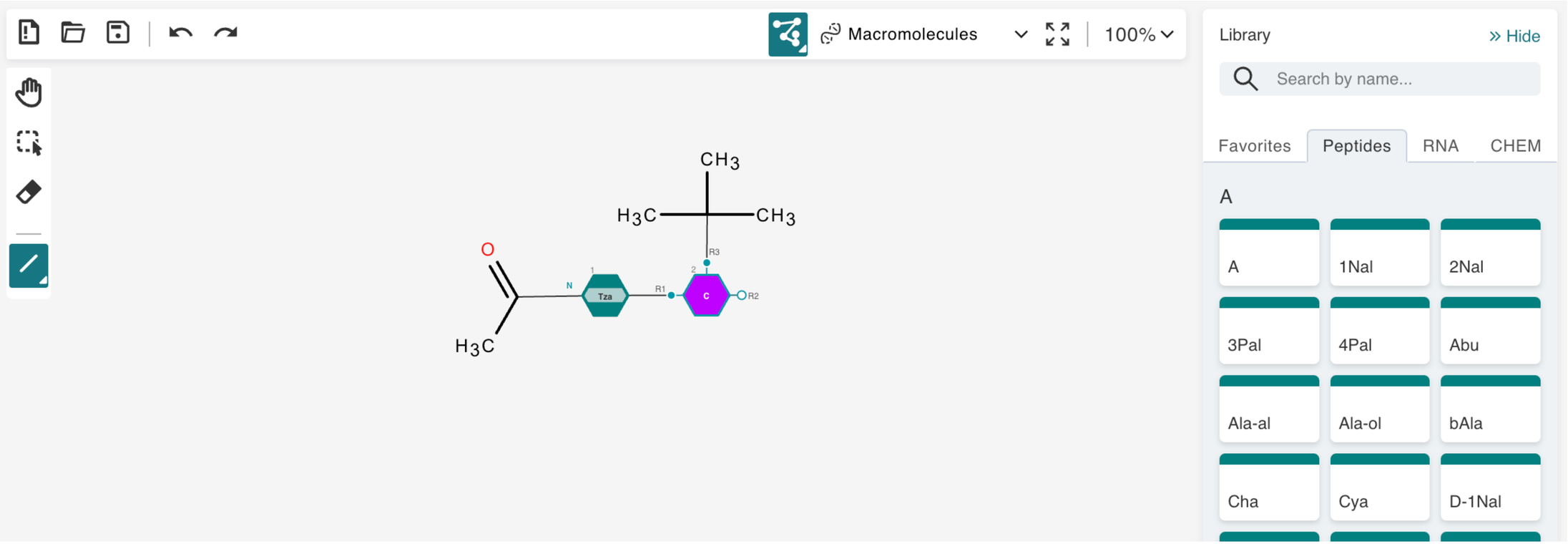


Synthetic monomers



- 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

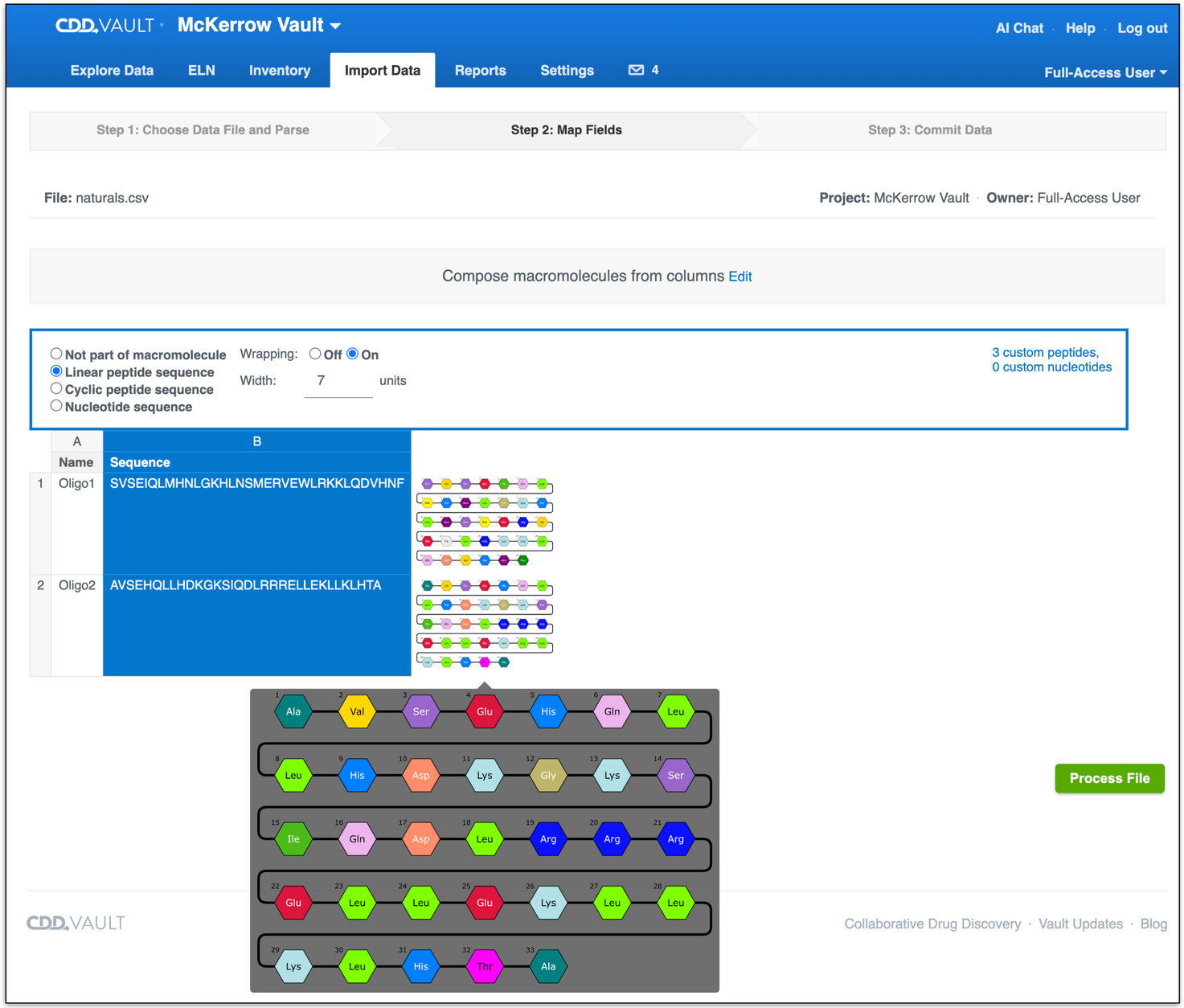
```
  -INDIGO-03112510242D

  0  0  0  0  0  0  0  0  0  0  0 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
```

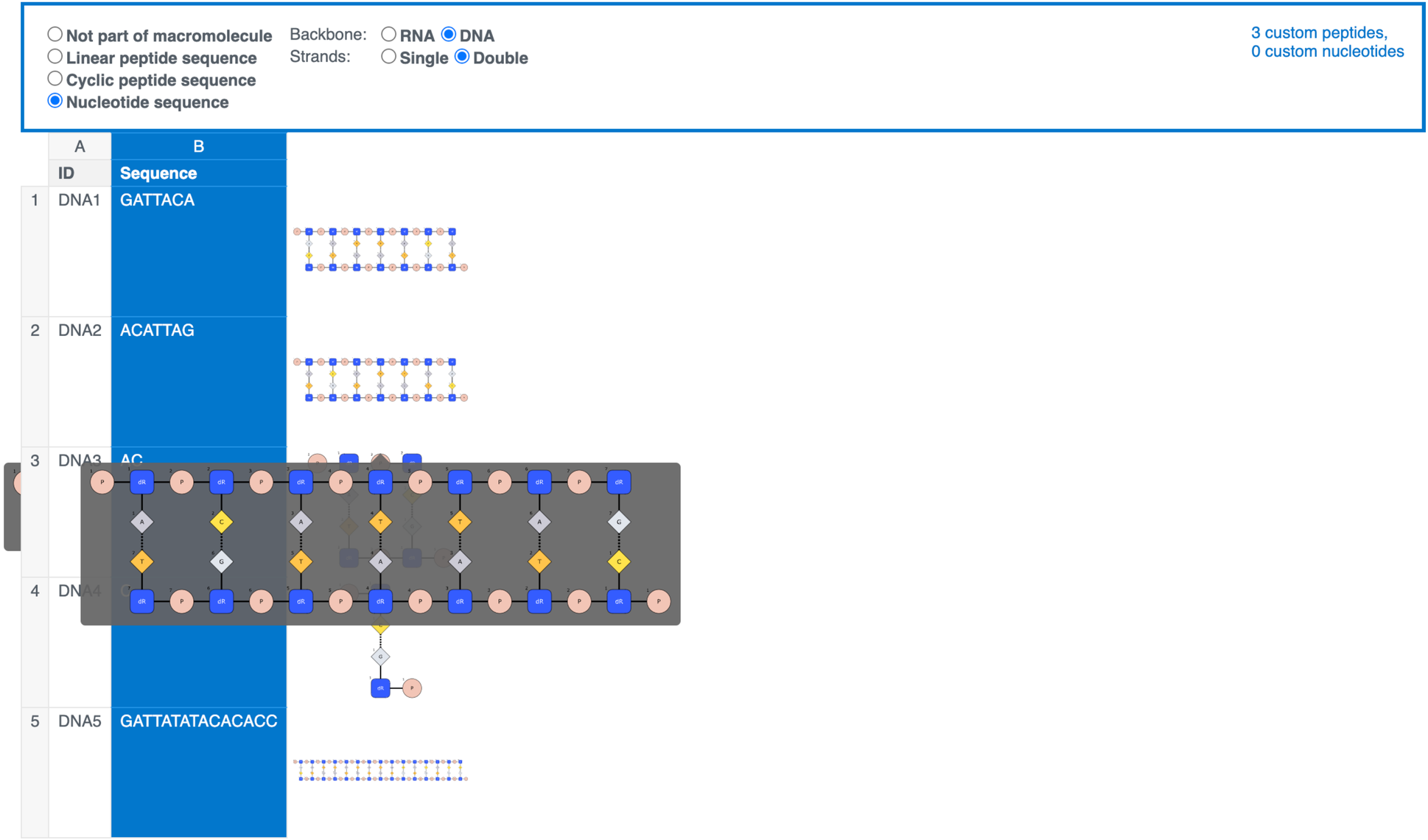
Bulk markup

- 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



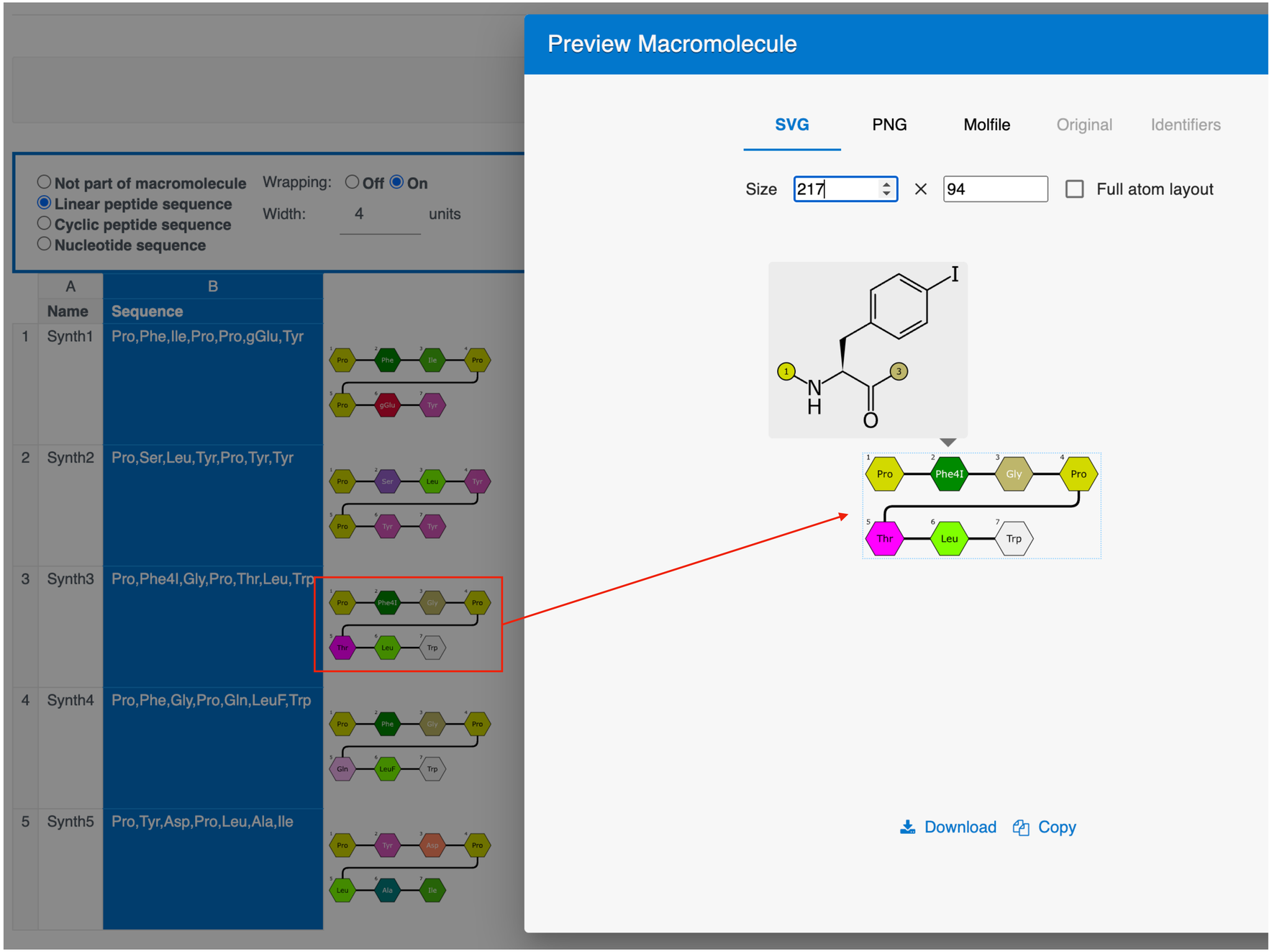
Import modes

- Cyclic peptides
- RNA
- DNA



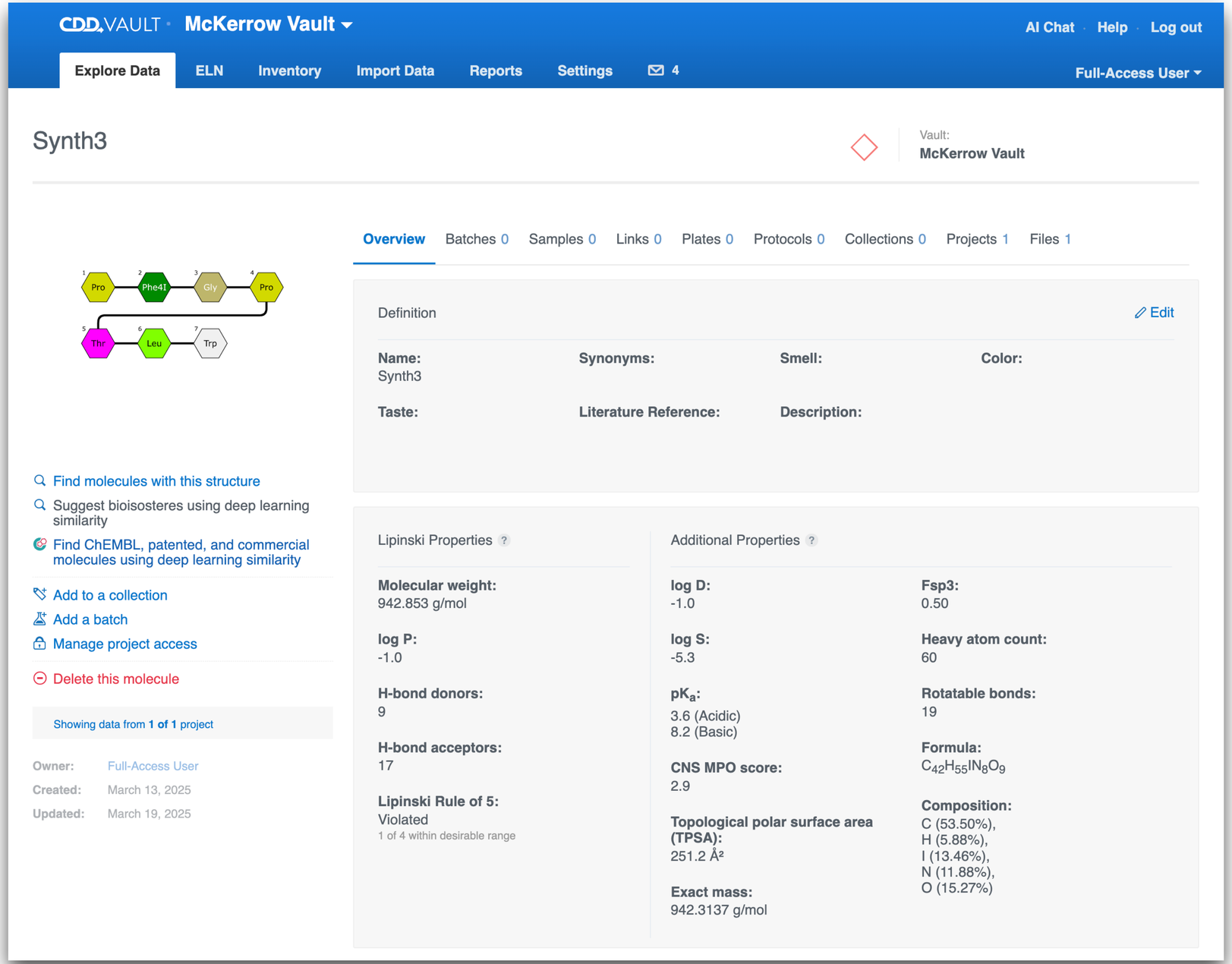
Custom monomers

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



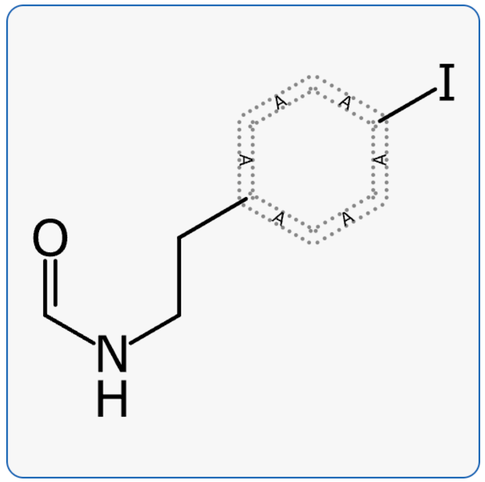
Registration

- Identification and duplication
- Chemical properties

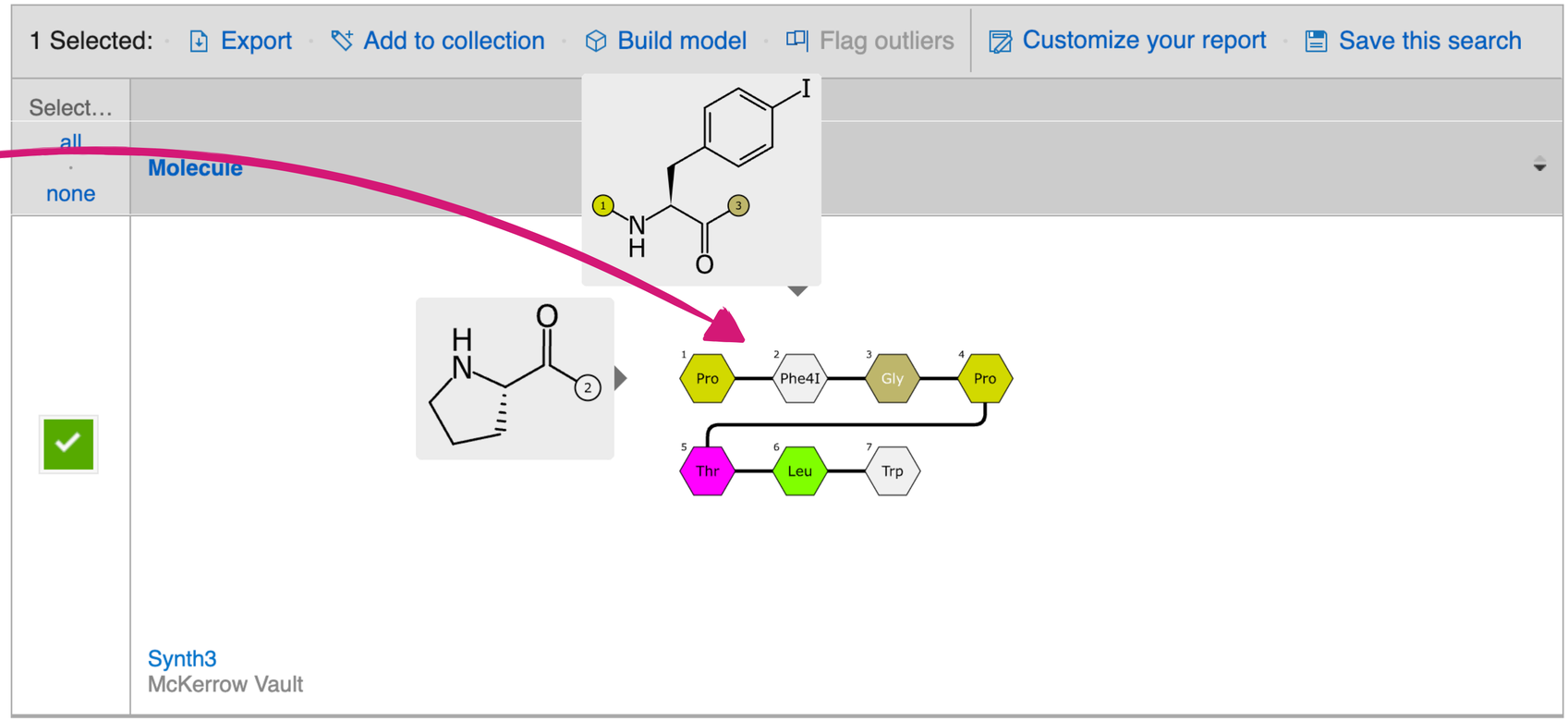


Searching

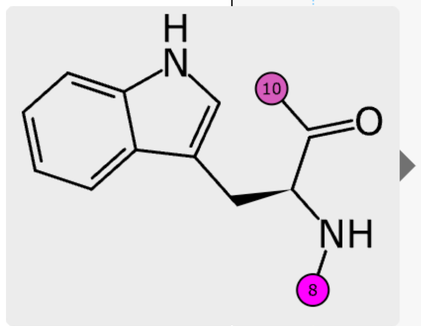
- Structure search queries operate on the full-atom structure



substructure query

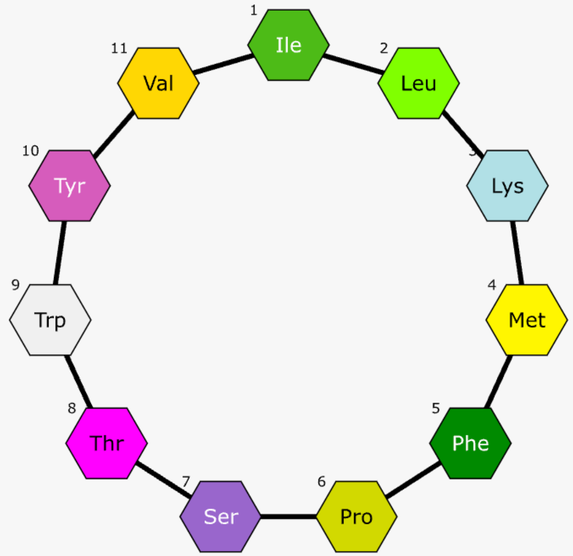


Full atom view



SVG PNG Molfile Original Identifiers

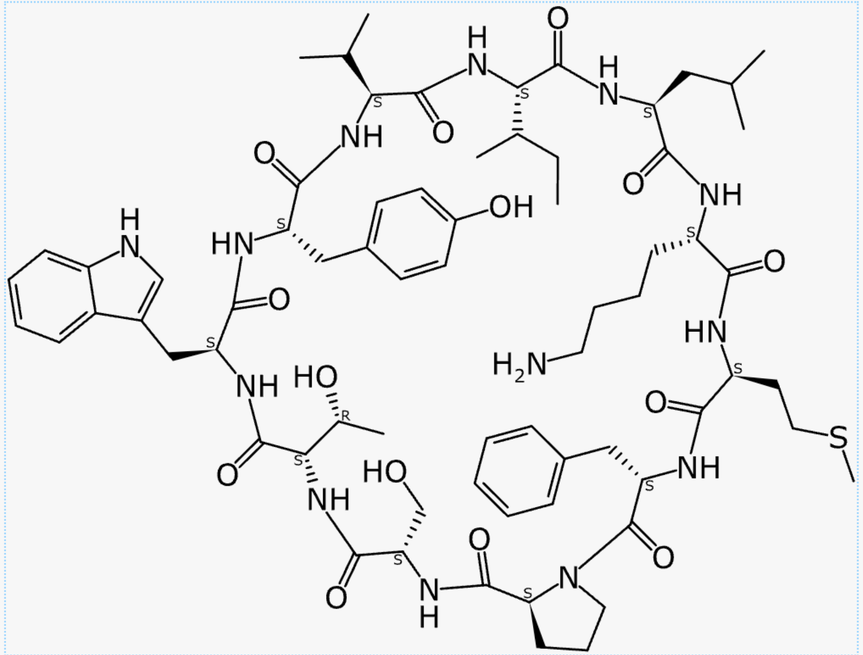
Size 500 × 490 Full atom layout



Download Copy

SVG PNG Molfile Original Identifiers

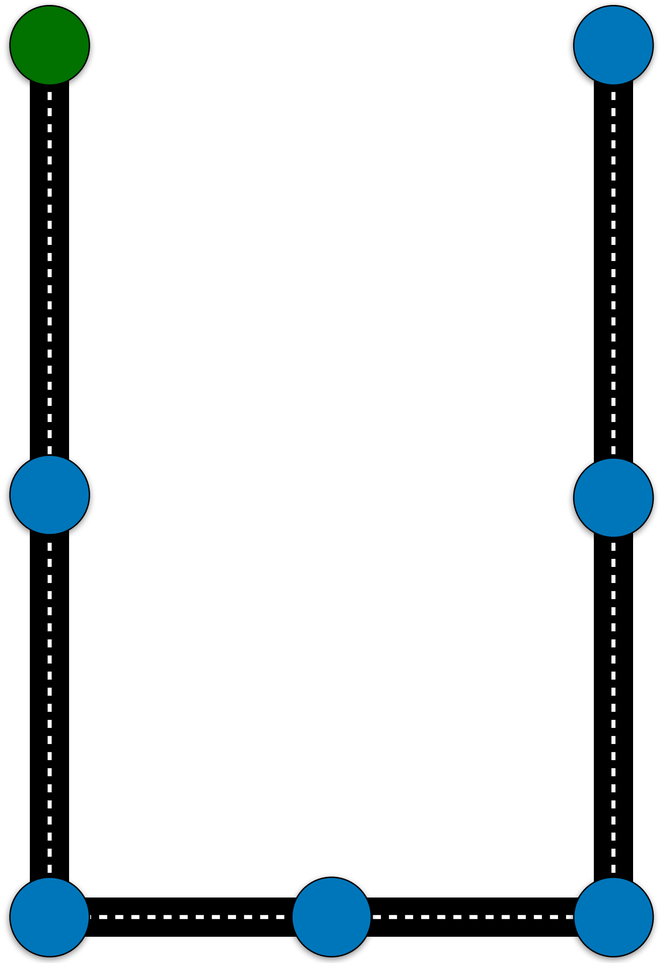
Size 820 × 625 Full atom layout



scale 60%

Download Copy

Roadmap



monomer management

prioritize requests...

import & export

V3000 monomer ecosystem

visualization (e.g. substructure hits)

drawing exotic hybrids with Ketcher

complex entities (e.g. antibody-drug conjugates)

Questions?

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