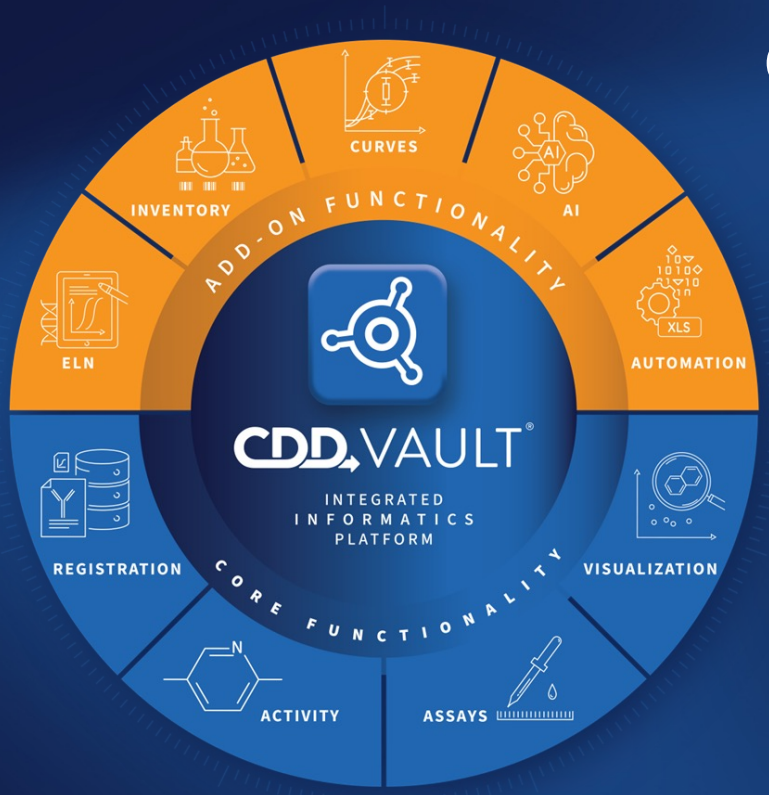
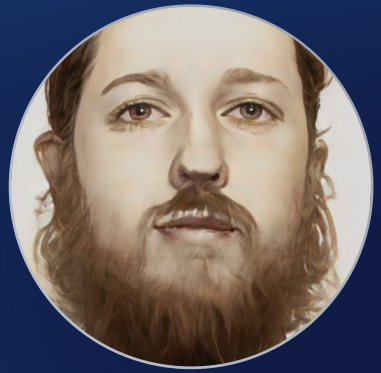
CDD VAULT®
Complexity Simplified



Cheminformatics Meets Biology: Hybrid Molecular Representations in Drug Discovery



Jonathan Bisson
Alex Clark – Tad Hurst

IF YOU HAVE TO REMEMBER ONE THING

CDD VAULT®
Complexity Simplified

*Don't **ever** throw away the original file!*

Dr. Alex M. Clark

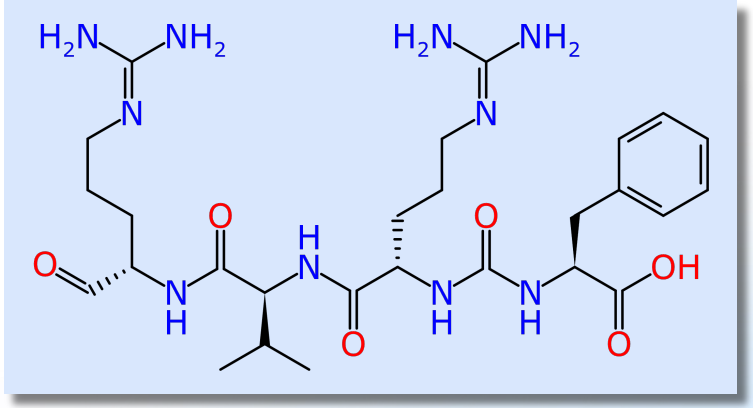
https://cheminf20.org/2014/05/05/on-the-myth-of-chemical-structure-format-conversion/

Macromolecules

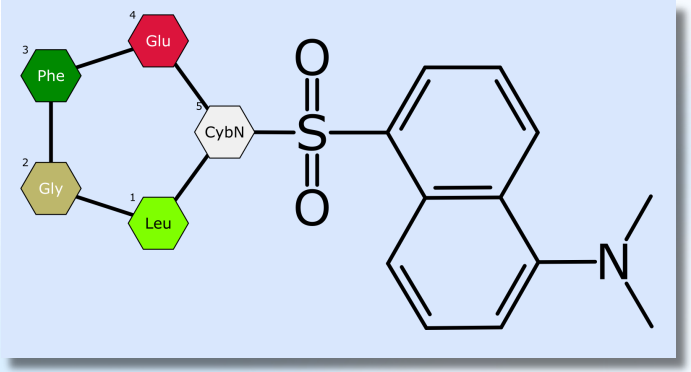
H	R	R	T	S
Q	R	C	E	E
E	R	Y	V	P
V	N	Y	C	C
Q	W	G	M	

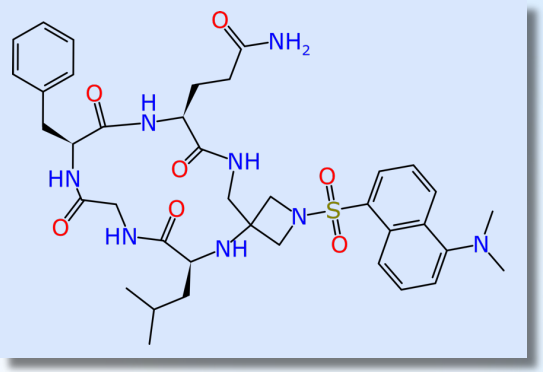
Bioinformatics or Cheminformatics

We need both...



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





Datastructure

V3000 molfile feature: **SCSR**

Self-**C**ontained **S**tructure **R**epresentations

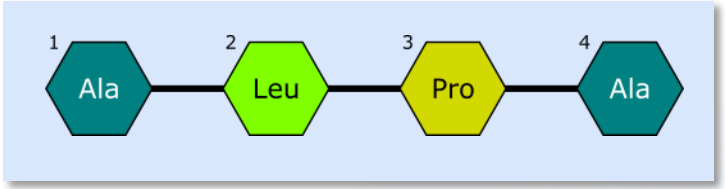
Layout information as 2D sketch

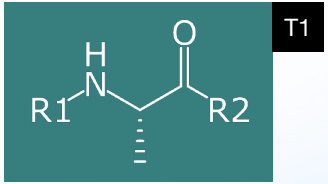
Monomers as **templates**: scalable

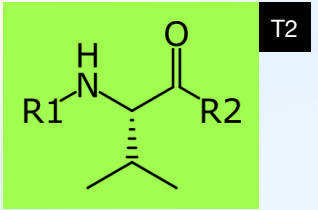
Self contained

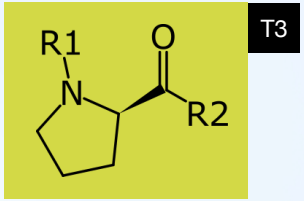
Full **atomic** structure is implied

Contains the **sequence** data



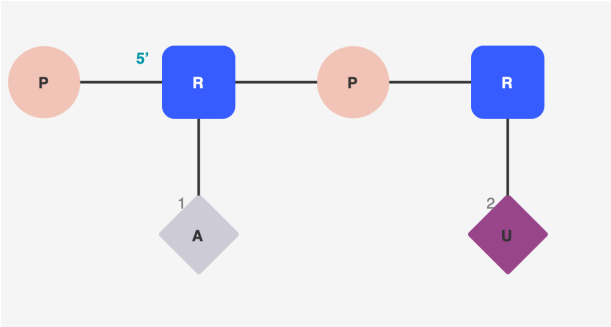






Datastructure

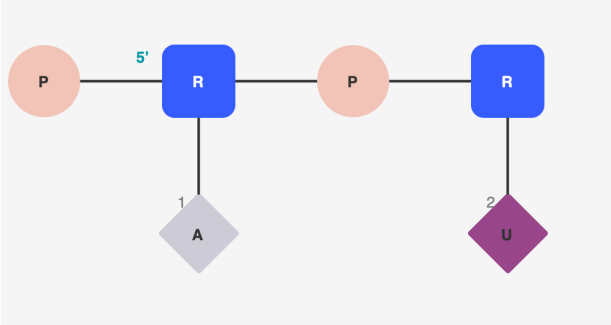
https://www.rdkit.org/docs/RDKit_Book.html#id32




```
Generated by WebMolKit

  0  0  0     0  0            999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 6 5 0 0 1
M  V30 BEGIN ATOM
M  V30 1 P 0.0000 0.0000 0.0000 0 CLASS=PHOSPHATE SEQID=1 ATTCHORD=(2 3 Br)
M  V30 2 P 3.0000 0.0000 0.0000 0 CLASS=PHOSPHATE SEQID=2 ATTCHORD=(4 3 Al 4 Br)
M  V30 3 R 1.5000 0.0000 0.0000 0 CLASS=SUGAR SEQID=1 ATTCHORD=(6 1 Al 2 Br 5 Cx)
M  V30 4 R 4.5000 0.0000 0.0000 0 CLASS=SUGAR SEQID=2 ATTCHORD=(4 2 Al 6 Cx)
M  V30 5 A 1.5000 -1.5000 0.0000 0 CLASS=BASE SEQID=1 ATTCHORD=(2 3 Al)
M  V30 6 U 4.5000 -1.5000 0.0000 0 CLASS=BASE SEQID=2 ATTCHORD=(2 4 Al)
M  V30 END ATOM
```

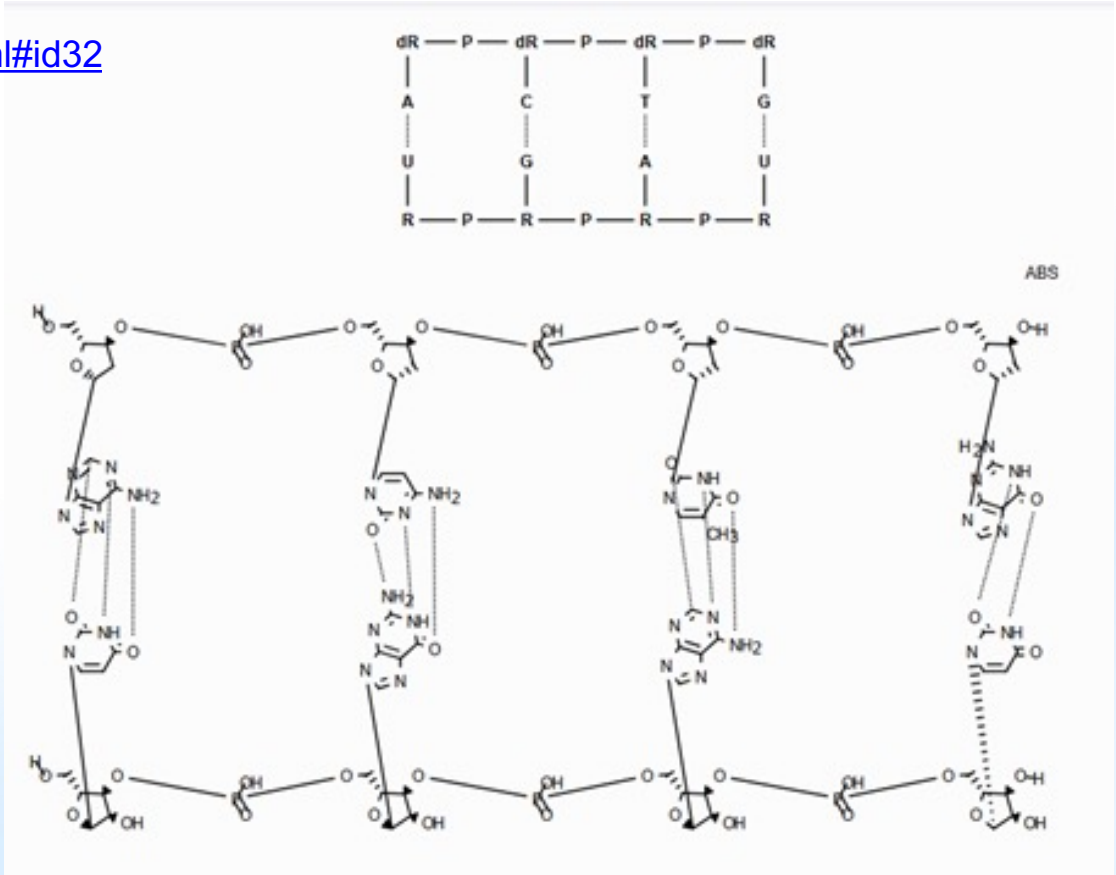
SCSR – Template and leaving groups




```
M  V30 BEGIN TEMPLATE
M  V30 TEMPLATE 1 PHOSPHATE/P/P/ NATREPLACE=PHOSPHATE/P
M  V30 BEGIN CTAB
M  V30 COUNTS 5 4 3 0 1
M  V30 BEGIN ATOM
M  V30 1 P 0.0000 0.0000 0.0000 0
M  V30 2 O 1.5000 0.0000 0.0000 0
M  V30 3 O -1.5000 0.0000 0.0000 0
M  V30 4 O 0.0000 -1.5000 0.0000 0
M  V30 5 O 0.0000 1.5000 0.0000 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 1 4
M  V30 4 2 1 5
M  V30 END BOND
M  V30 BEGIN SGROUP
M  V30 1 SUP 0 LABEL=P ATOMS=(3 1 4 5) XBONDS=(2 2 1) CLASS=PHOSPHATE … SAP=(3 1 3 Al) SAP=(3 1 2 Br)
M  V30 2 SUP 0 LABEL=O ATOMS=(1 3) XBONDS=(1 2) CLASS=LGRP
M  V30 3 SUP 0 LABEL=O ATOMS=(1 2) XBONDS=(1 1) CLASS=LGRP
M  V30 END SGROUP
M  V30 END CTAB
```

SCSR – Wobble bonds

https://www.rdkit.org/docs/RDKit_Book.html#id32



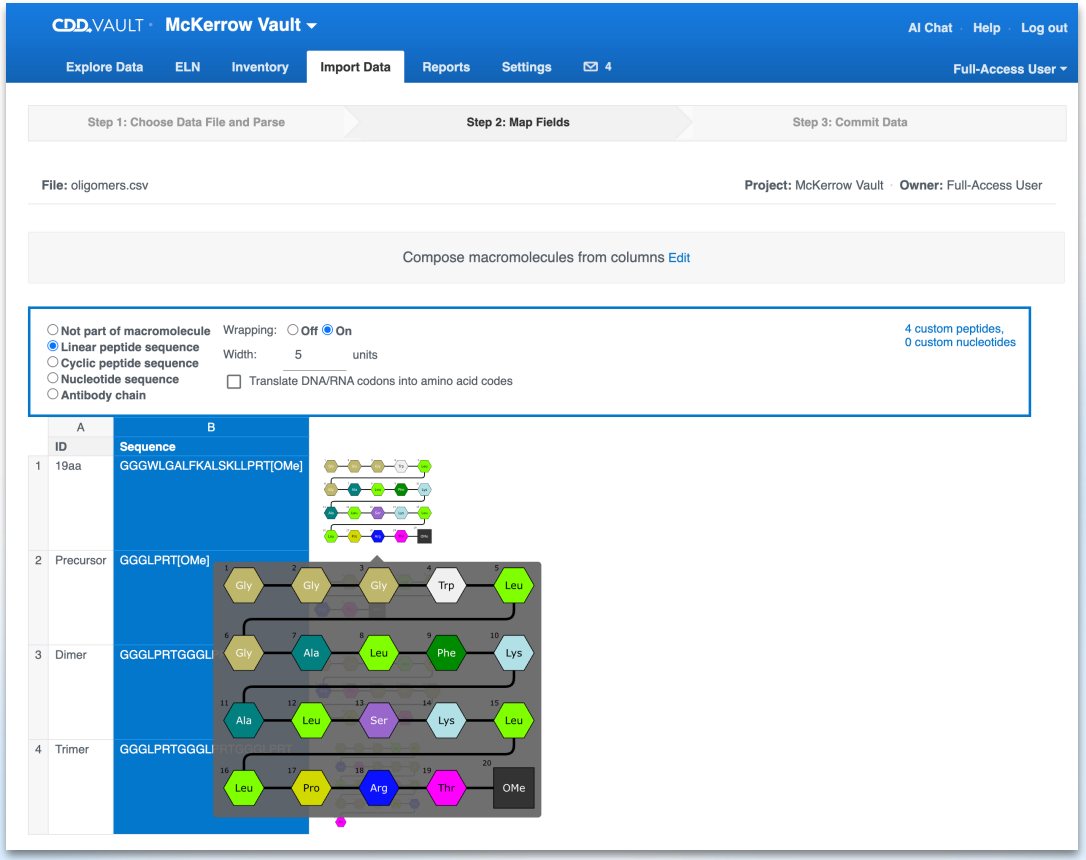
Content creation

CDD VAULT®
Complexity Simplified

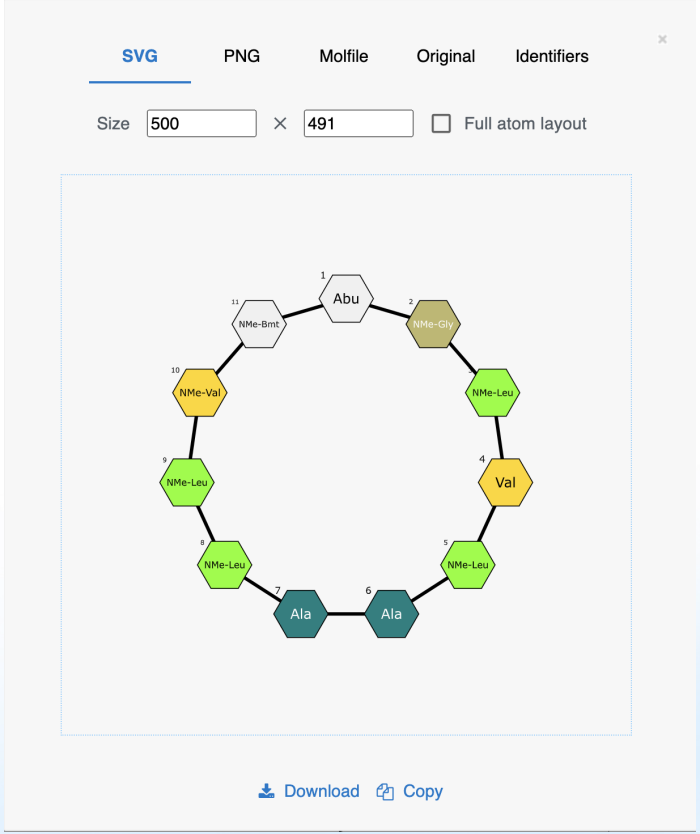
Draw anything with EPAM'Ketcher

Bulk import: markup sequences

Custom monomers



Cyclic peptides

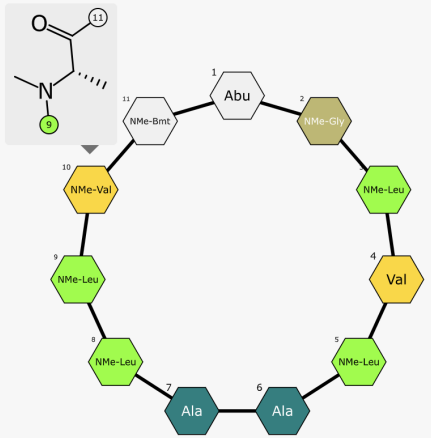


Cyclic peptides

CDD VAULT®
Complexity Simplified

SVG PNG Molfile Original Identifiers

Size 500 × 491 Full atom layout



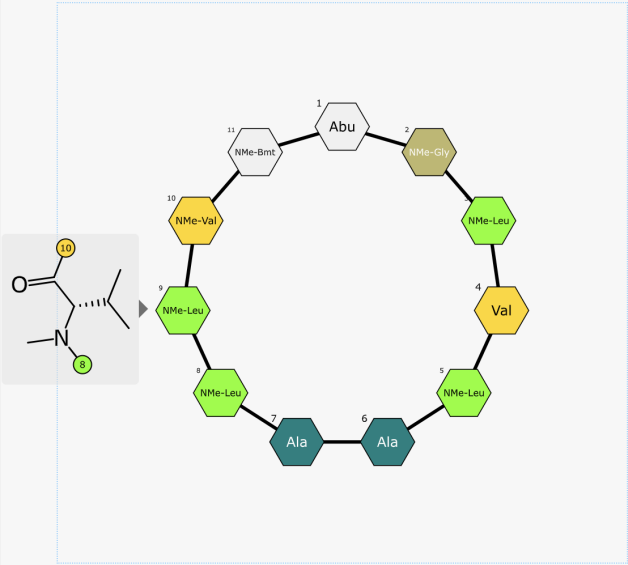
Download Copy

Cyclic peptides

CDD VAULT®
Complexity Simplified

SVG PNG Molfile Original Identifiers

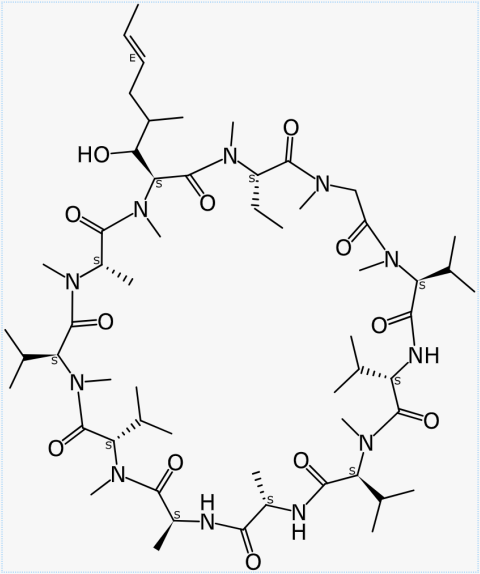
Size 500 × 491 Full atom layout



Download Copy

SVG PNG Molfile Original Identifiers

Size 631 × 758 Full atom layout

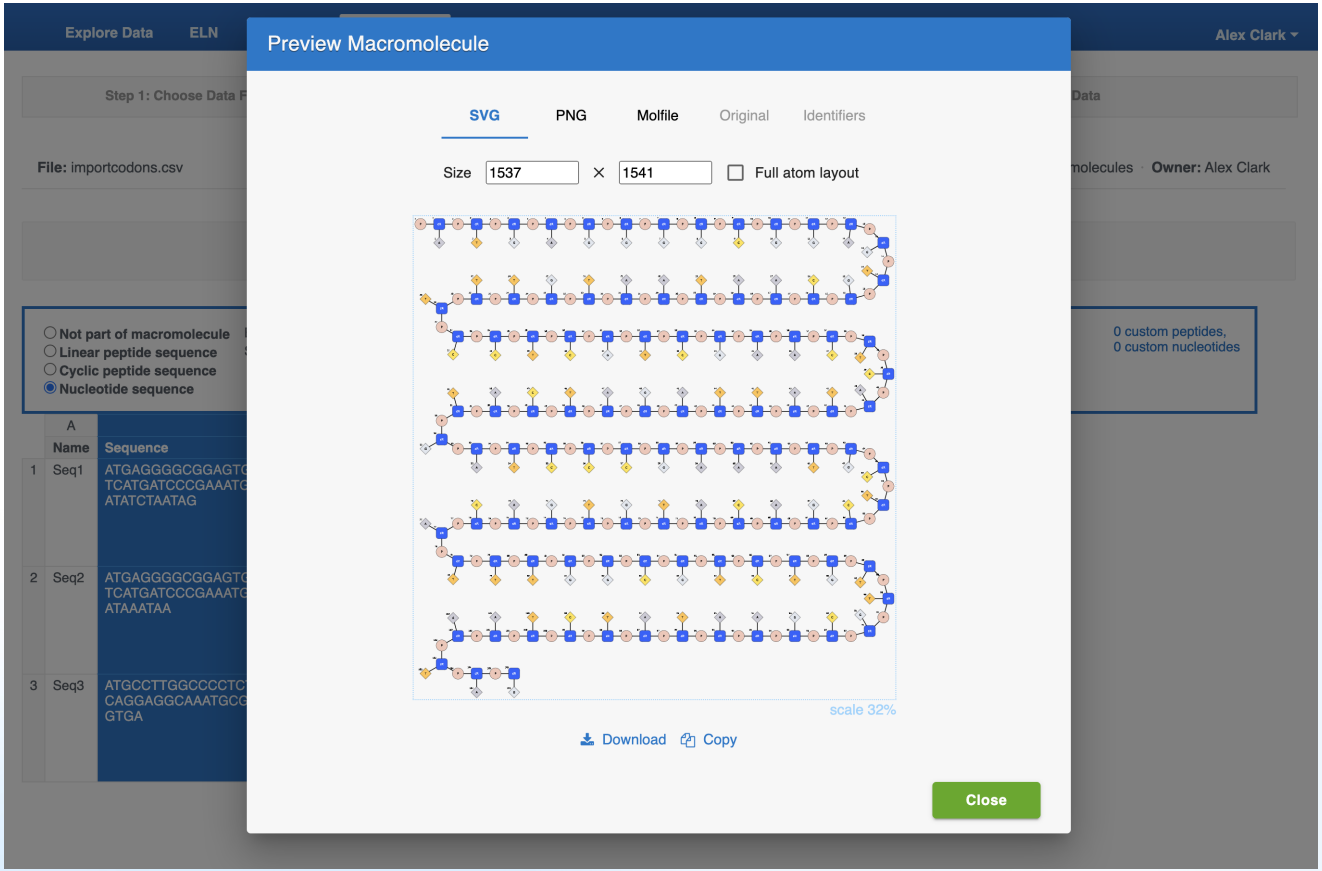


scale 65%

Download Copy

Nucleotides

CDD VAULT®
Complexity Simplified



Nucleotides

CDD VAULT®
Complexity Simplified

Explore Data | ELN | Inventory | Import Data | Reports | Settings | Alex Clark

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

File: importcodons.csv

Project: Macromolecules · **Owner:** Alex Clark

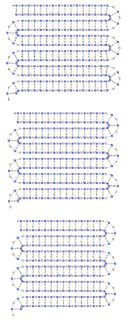
Compose macromolecules from columns Edit

- ○ **Not part of macromolecule**
- ○ **Linear peptide sequence**
- ○ **Cyclic peptide sequence**
- ◉ **Nucleotide sequence**

Backbone: ○ **RNA** ◉ **DNA**
Strands: ○ **Single** ◉ **Double**

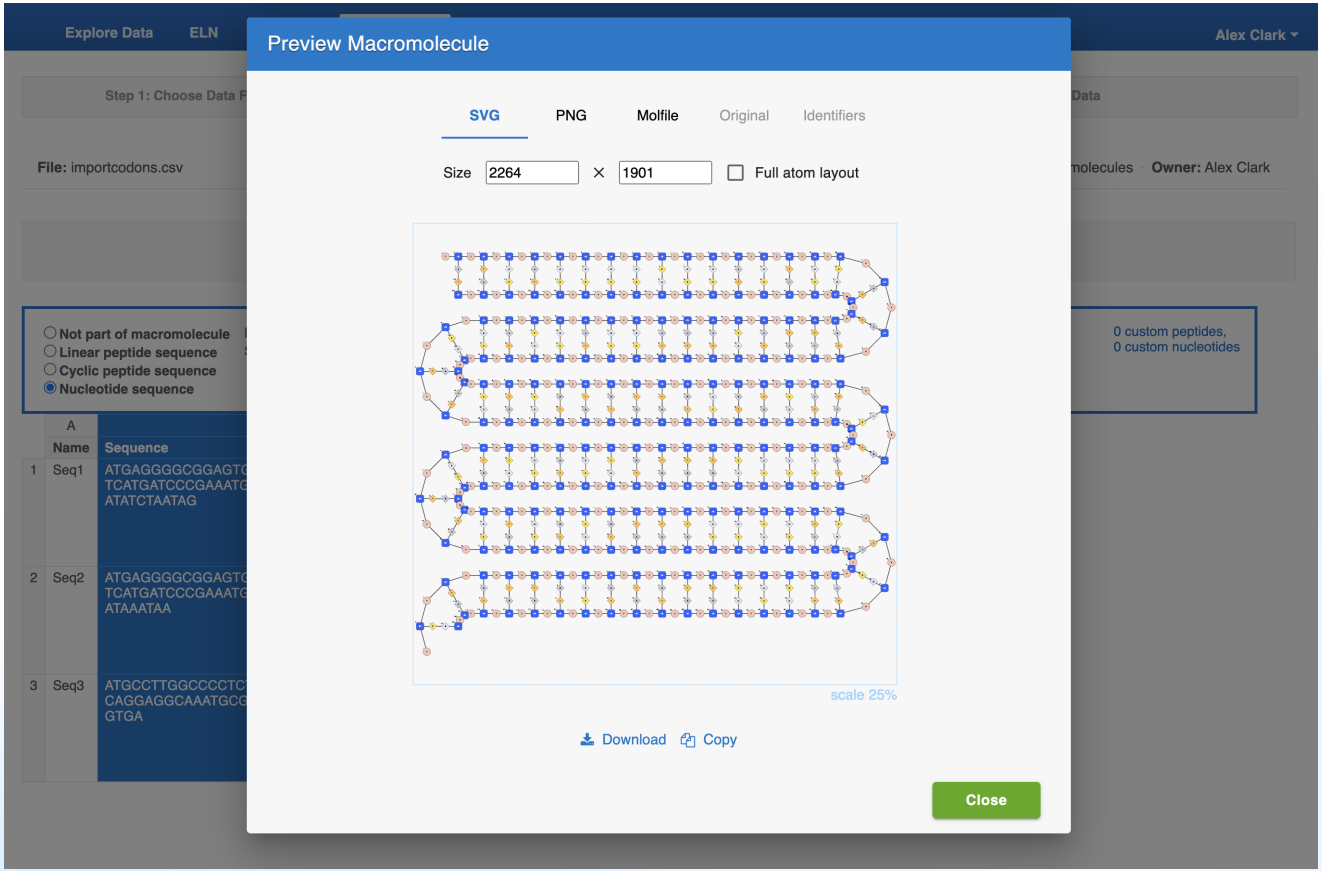
0 custom peptides,
0 custom nucleotides

	A	B
	Name	**Sequence**
1	Seq1	ATGAGGGGCGGAGTGCAATAATGTTTCCTCGTCGAACTCATATTAGATCATGATCCCGAAATGCTCGACATGTGACATTTGGCGTCTGTTGCGAATATCTAATAG
2	Seq2	ATGAGGGGCGGAGTGCAATAATGTTTCCTCGTCGAACTCATATTAGATCATGATCCCGAAATGCTCGACATGTGACATTTGGCGTCTGTTGCGAATAAATAA
3	Seq3	ATGCCTTGGCCCCTCTAAGCCAGCAAAAGCCTTTGTGGCCTGCTGACAGGAGGCAAATGCGGAAGAATTTATGCGAACCATTTGGATGACACGTGA



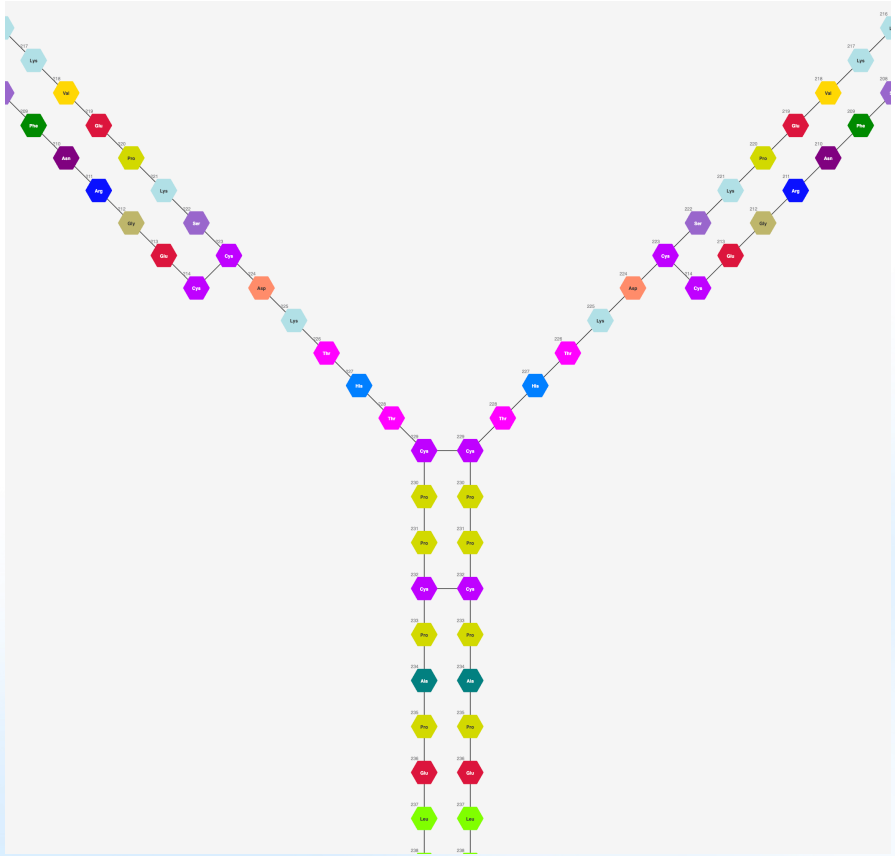
Nucleotides

CDD VAULT®
Complexity Simplified



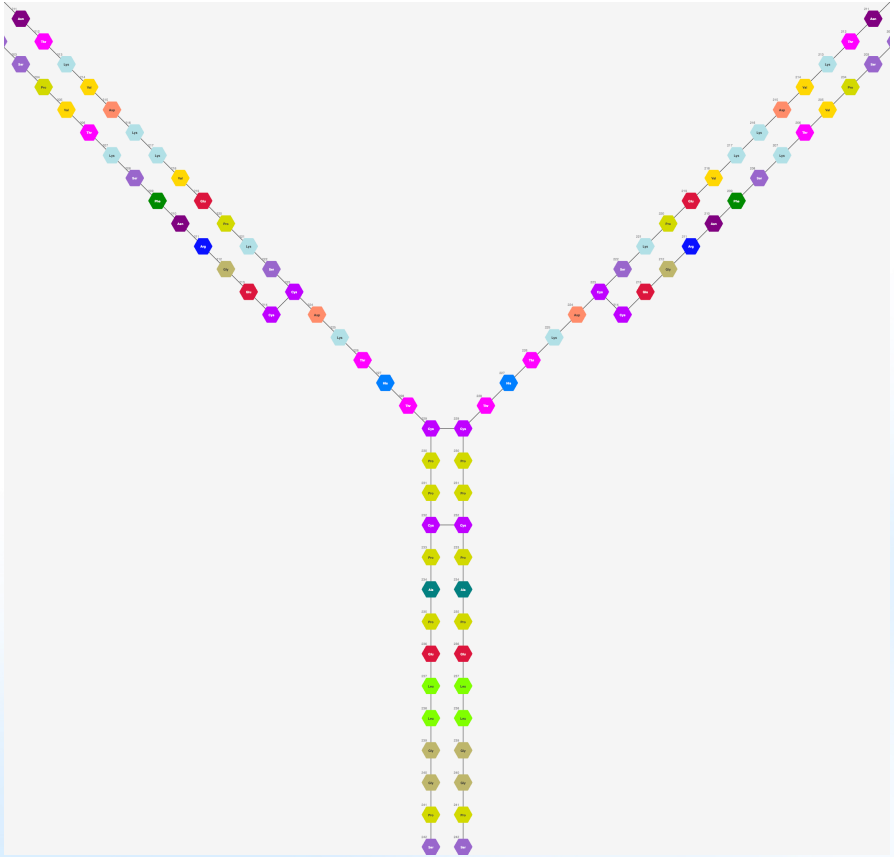
The problem of scale

CDD VAULT®
Complexity Simplified



The problem of scale

CDD VAULT®
Complexity Simplified

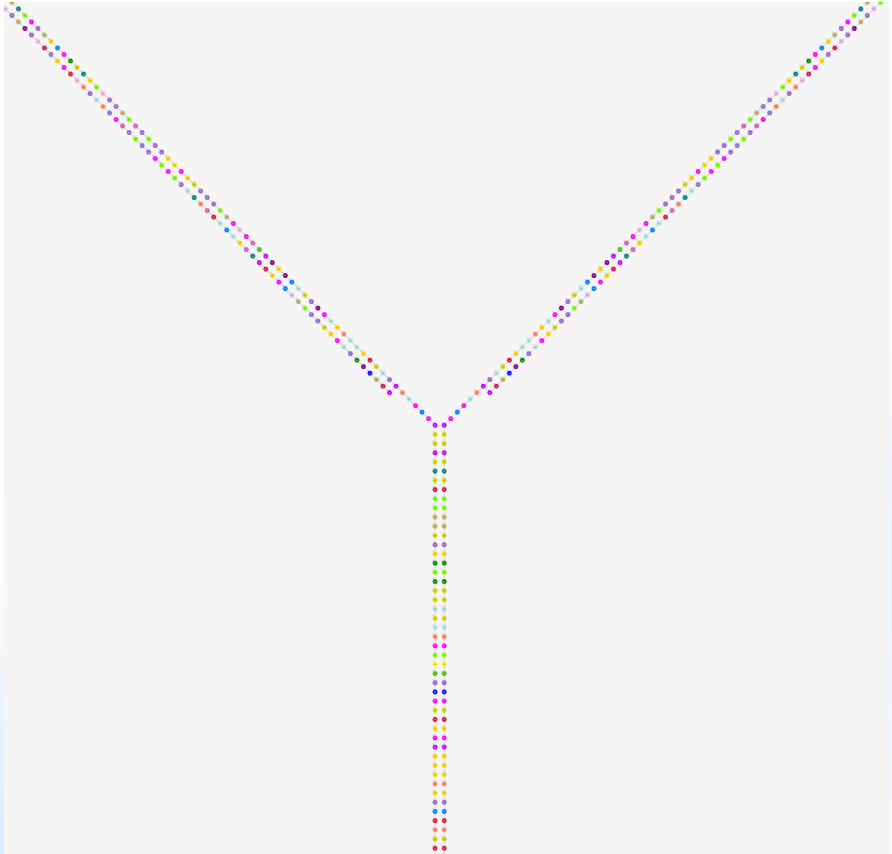


The problem of scale

CDD VAULT®
Complexity Simplified

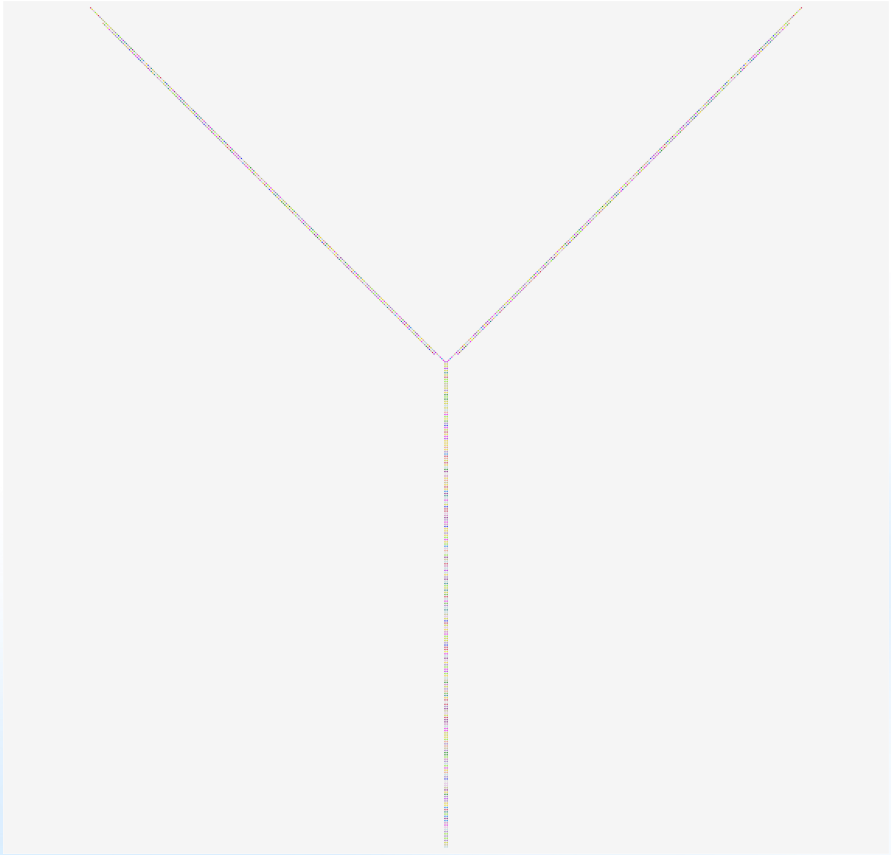
The problem of scale

CDD VAULT®
Complexity Simplified



The problem of scale

CDD VAULT®
Complexity Simplified

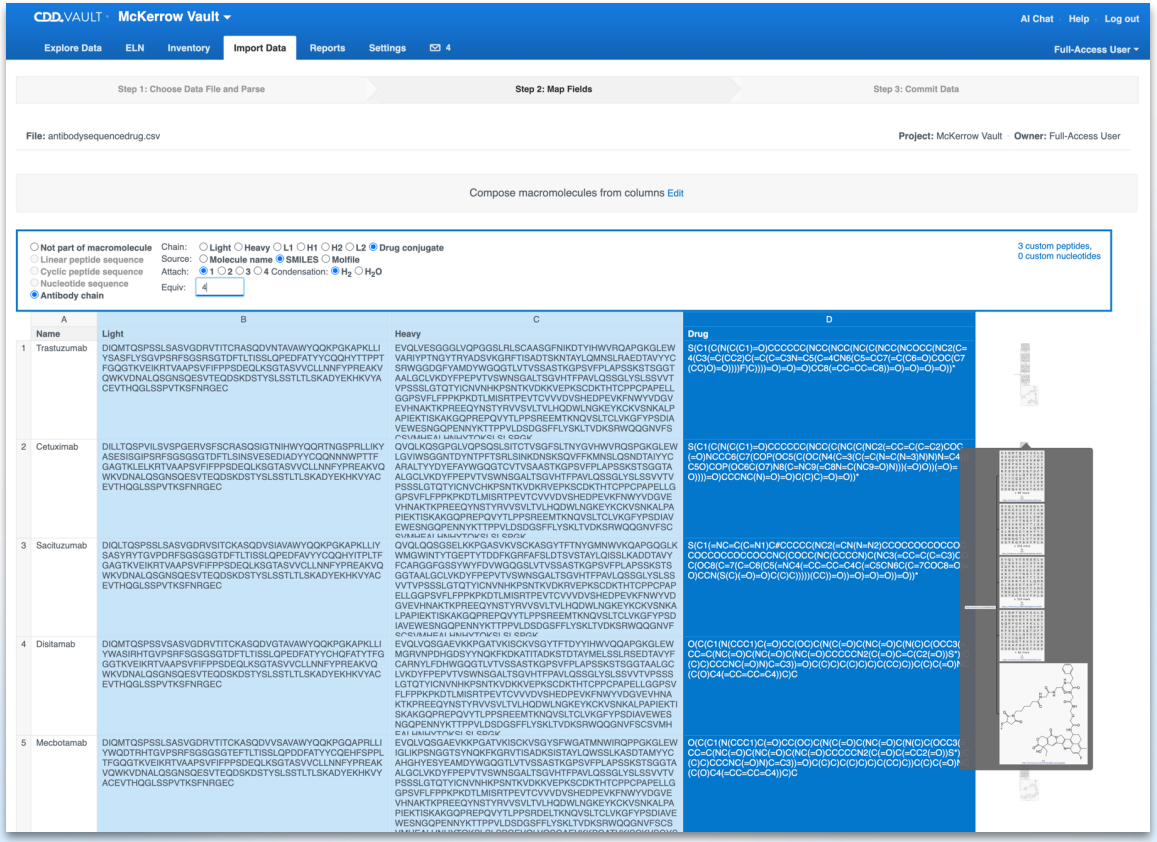


Antibodies: Basic

CDD VAULT®
Complexity Simplified

Just have sequences and drug conjugate structures?

Not a problem: collection of objects



Antibodies: Metadata

CDD VAULT®
Complexity Simplified

Light

DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSR
SGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVV
CLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTH
QGLSSPVTKSFNRGEC**{VL:1-107}{CL:108-214}{@214:1}**

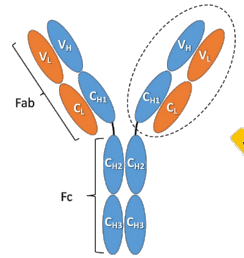
Heavy

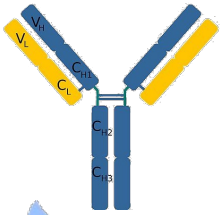
EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVK
GRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSSASTKGPSVF
PLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSS
LGTQTYICNVNHKPSNTKVDKKVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPE
VTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKC
KVSNKALPAPIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQP
ENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK
{VH:1-120}{CH1:121-218}{CH2:234-343}{CH3:344-450}{H:219-233}{@223:1}{@229:2}
{@232:3}{$264:1}{$324:1}{$370:1}{$428:1}

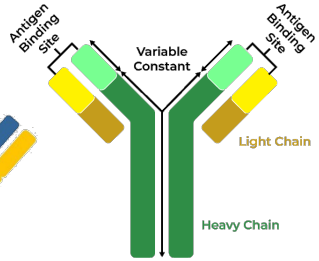
Domain Region
Disulfide Bridge
Hinge Region
Drug Conjugate Site

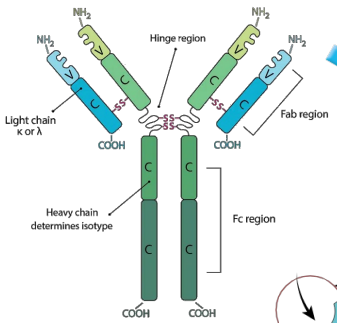
Diversity of displays

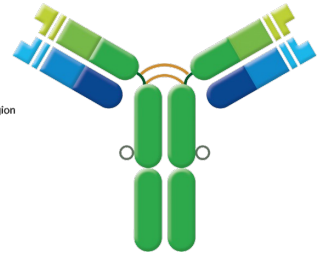
CDD VAULT®
Complexity Simplified

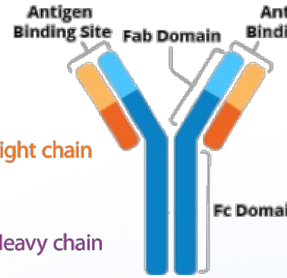


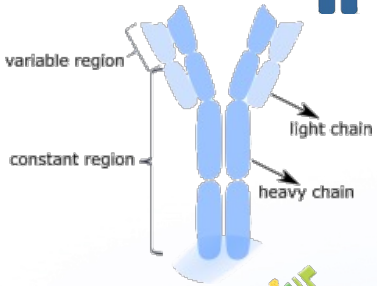


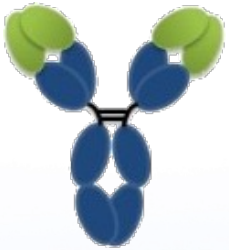


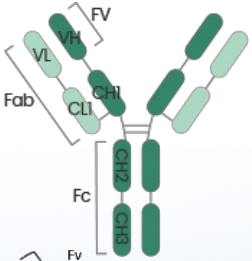


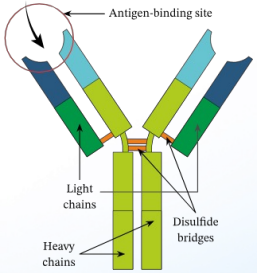


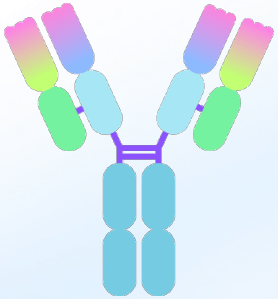


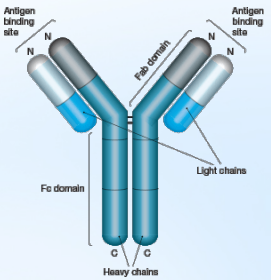


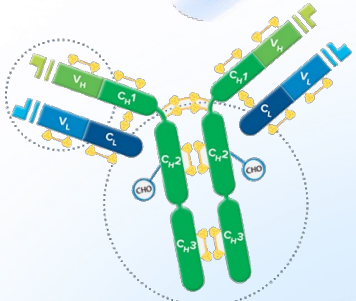


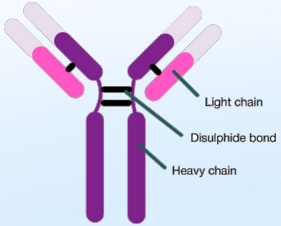


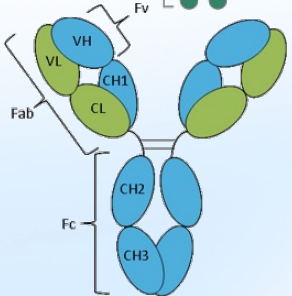


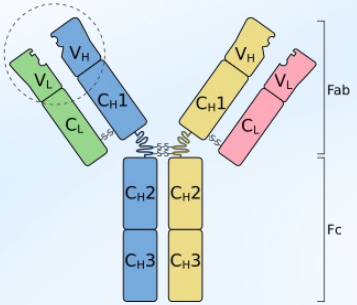


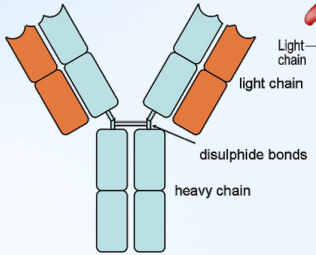


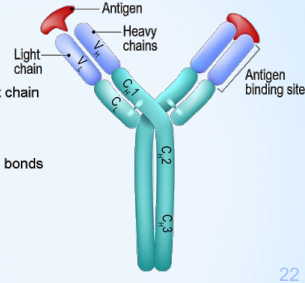






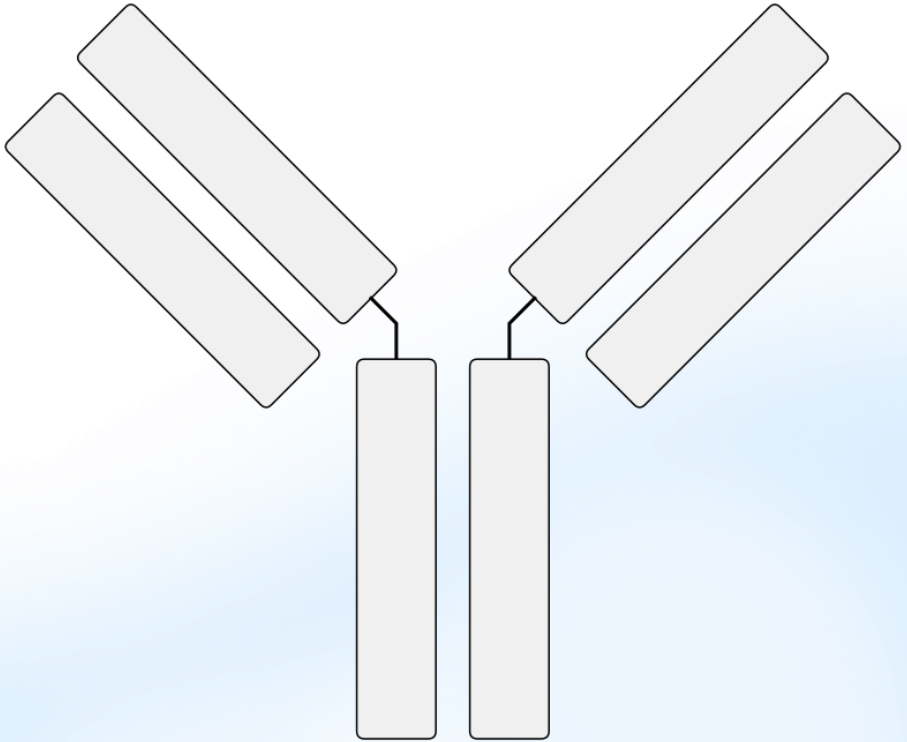






Cartoon and content

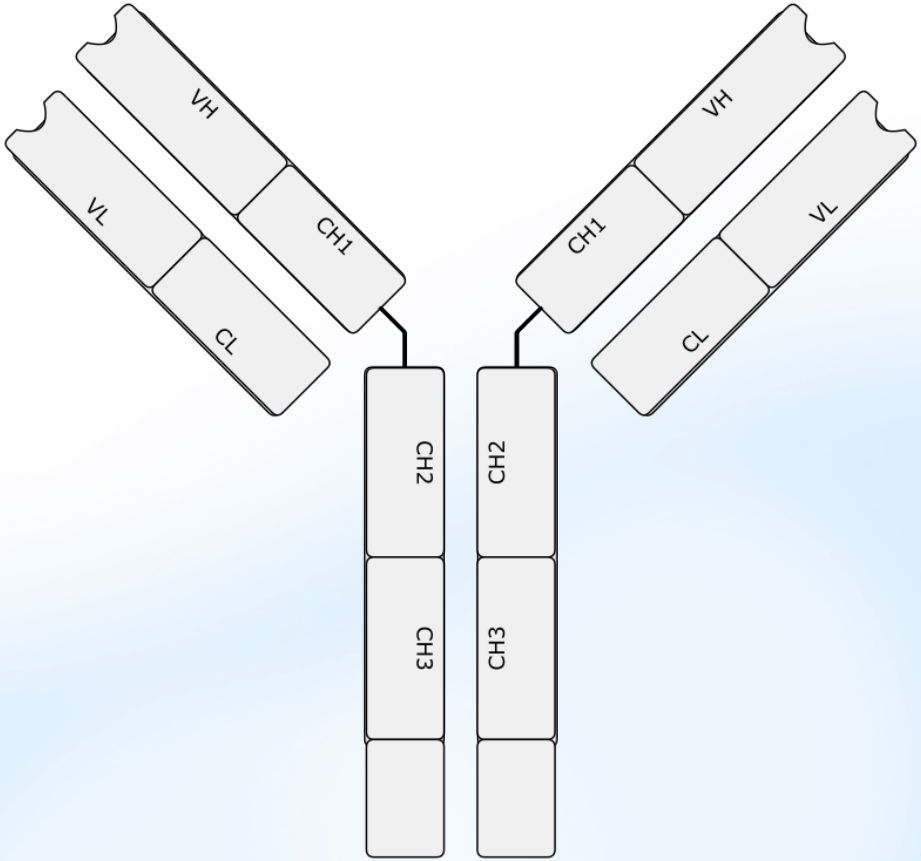
Draw basic Y-shape outline



Cartoon and content

Draw basic Y-shape outline

Segment domain regions

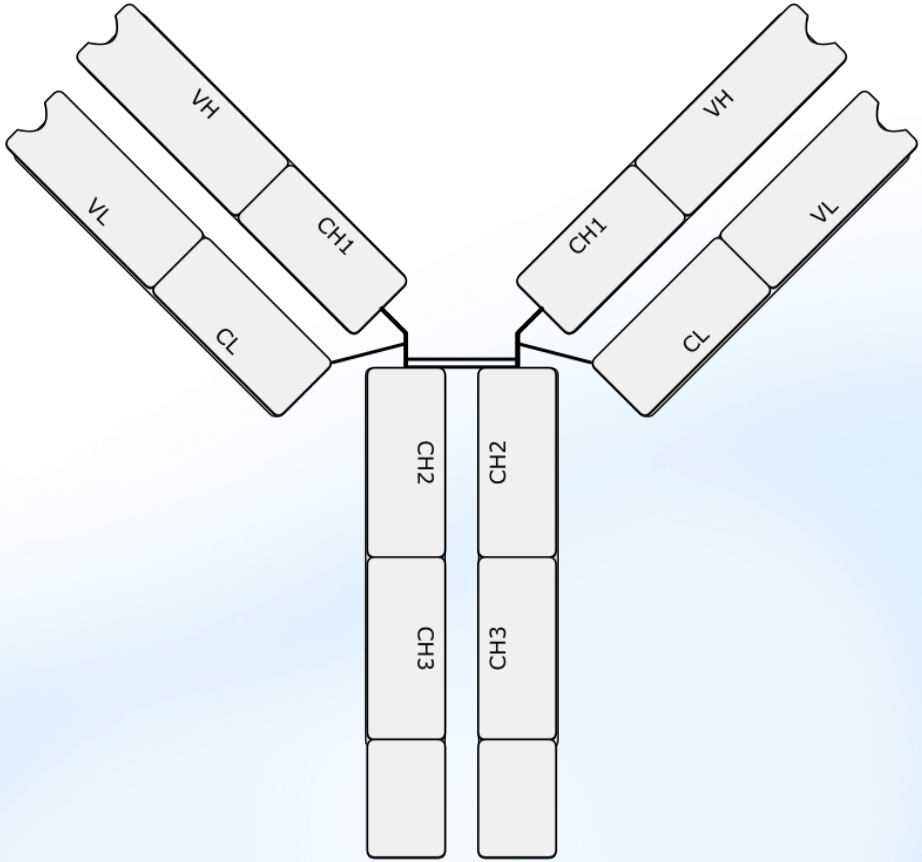


Cartoon and content

Draw basic Y-shape outline

Segment domain regions

Inter-chain disulfide links



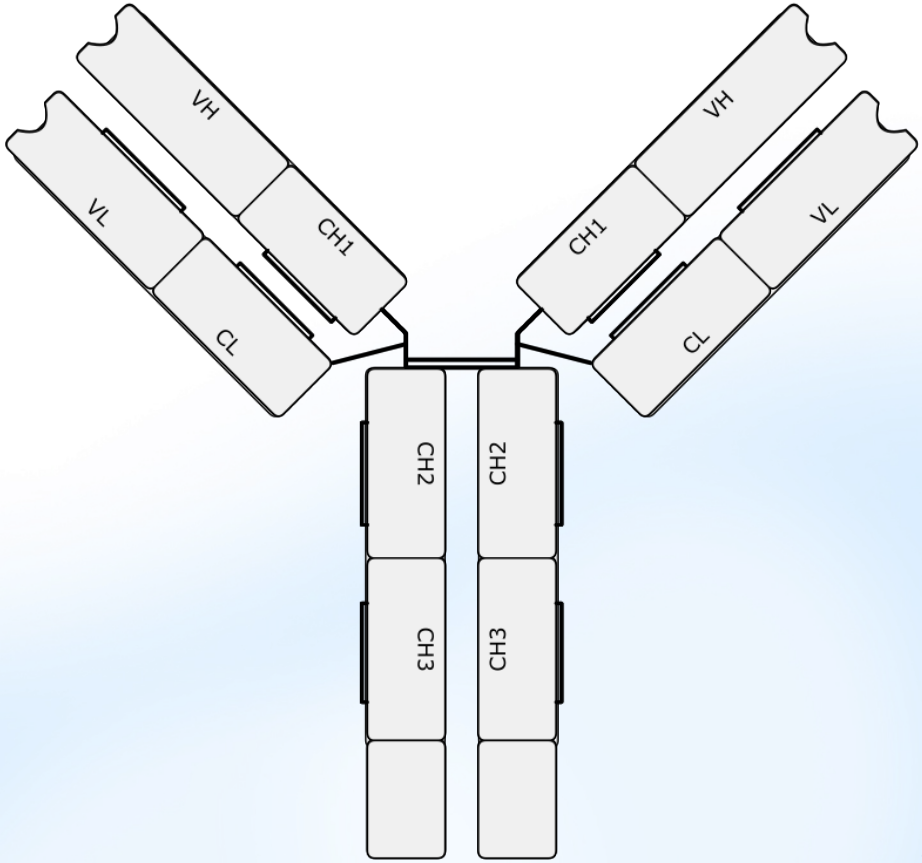
Cartoon and content

Draw basic Y-shape outline

Segment domain regions

Inter-chain disulfide links

Intra-chain disulfide links



Cartoon and content

CDD VAULT®
Complexity Simplified

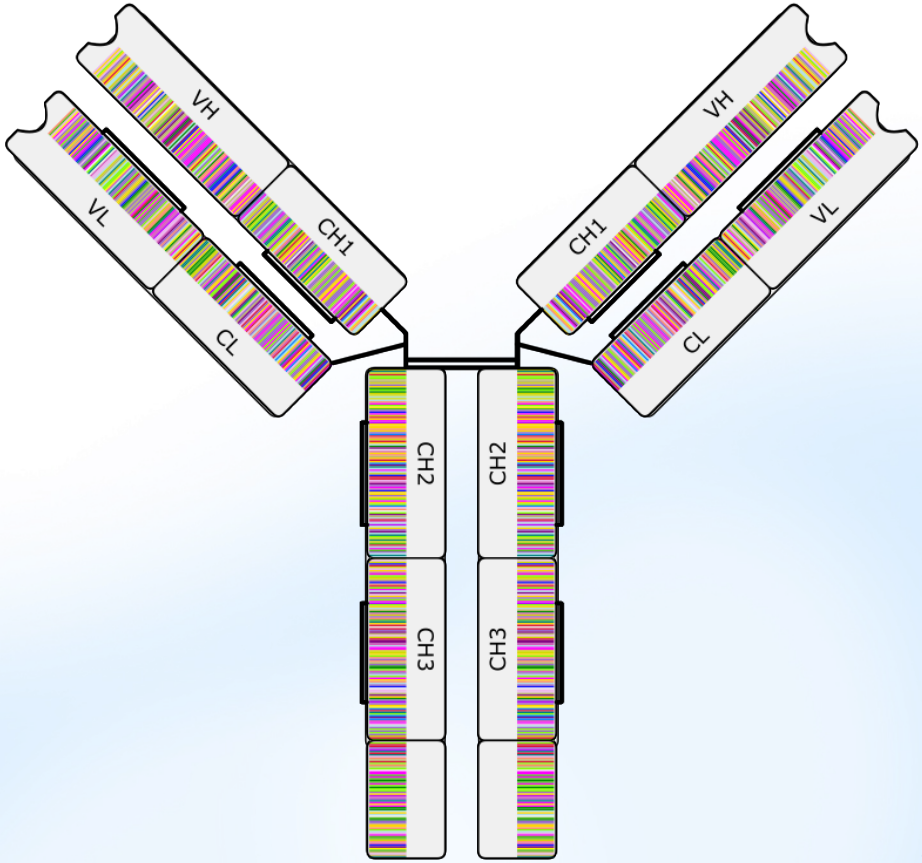
Draw basic Y-shape outline

Segment domain regions

Inter-chain disulfide links

Intra-chain disulfide links

Sequence barcoding

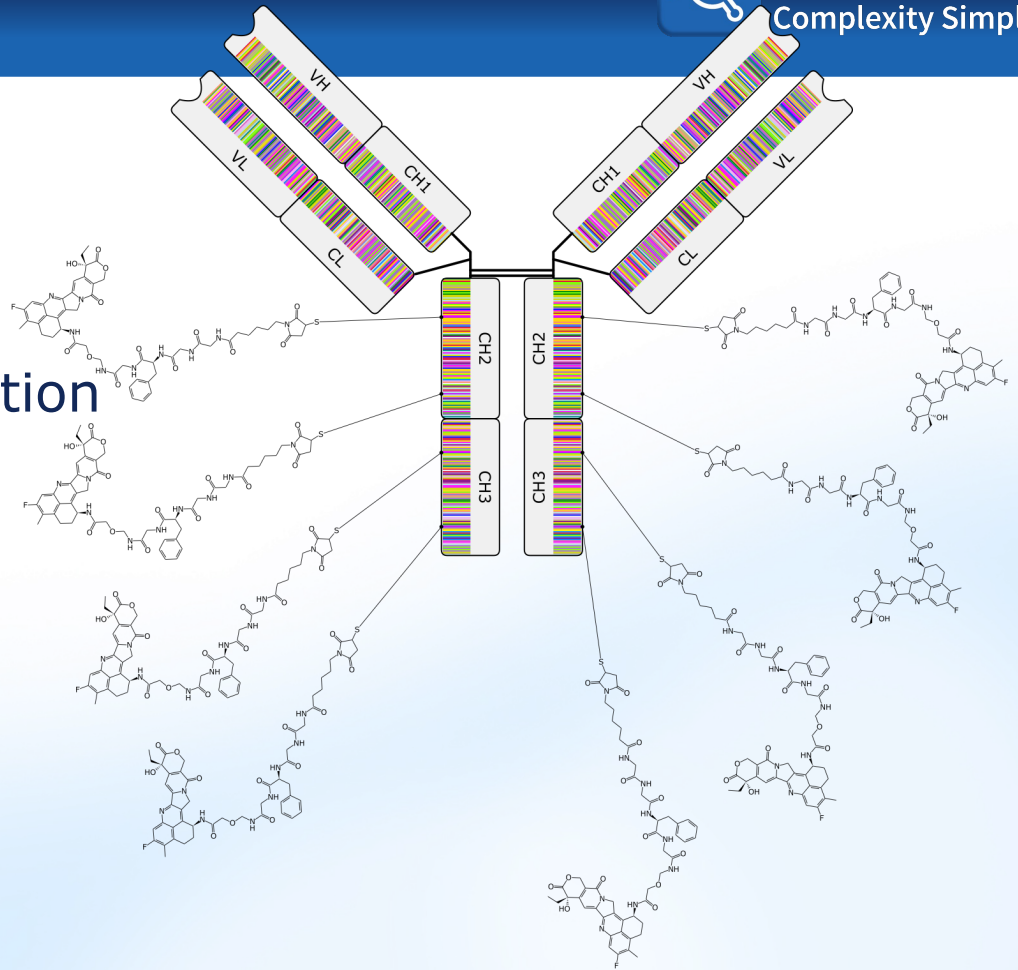


Drug conjugates

Tethered initial placements

Rigid body position optimization

Anchor point is indicated



The RDKit at CDD: It is everywhere

CDD VAULT®
Complexity Simplified

One dedicated employee for new features/fixes: **Tad Hurst**

SCSR, MRV, stereochemistry, atropisomers…

PostgreSQL cartridge

Python microservice for i/o & layered **fingerprint**

lambdas for **standardization**

WASM for client-side

Conclusion

CDD VAULT®
Complexity Simplified

Diagrams for biologists

Sequences for bioinformaticians

Atoms & **bonds** for chemists

Standard formats and **open-source** tools where possible

Provide **all** information: get all the functionality

Excessive flexibility without constraints creates
unusable standards