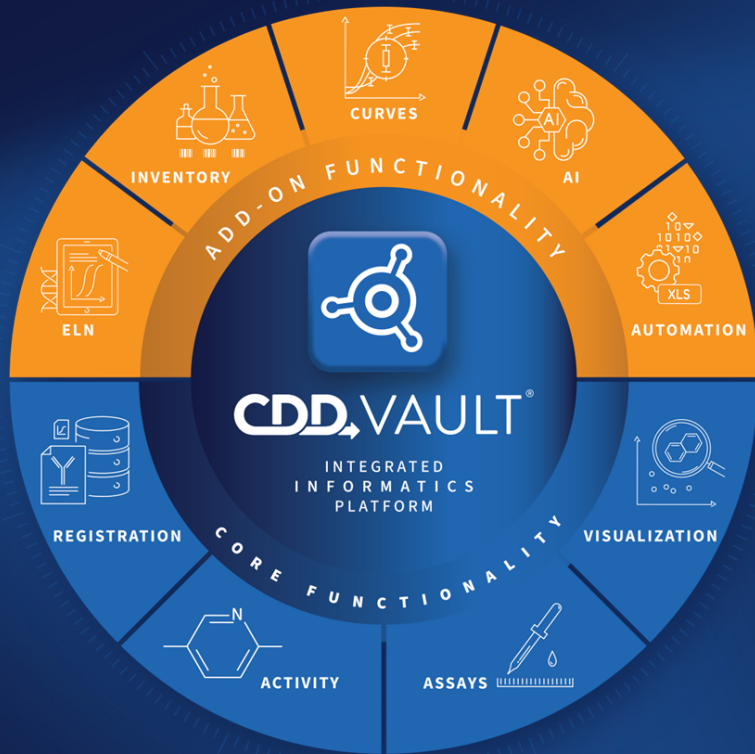




Cheminformatics for all shapes & sizes



Alex M. Clark

- Molecular data
 - ◆ **good**
 - ◆ **big**
 - ◆ **free**
- Pick one



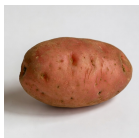
It is a mistake to think you can solve any major problems just with potatoes

- Douglas Adams

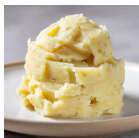
DIRECT



Drugs



Inorganics



Macromolecules

WRAPPED



Polymers



Mixtures

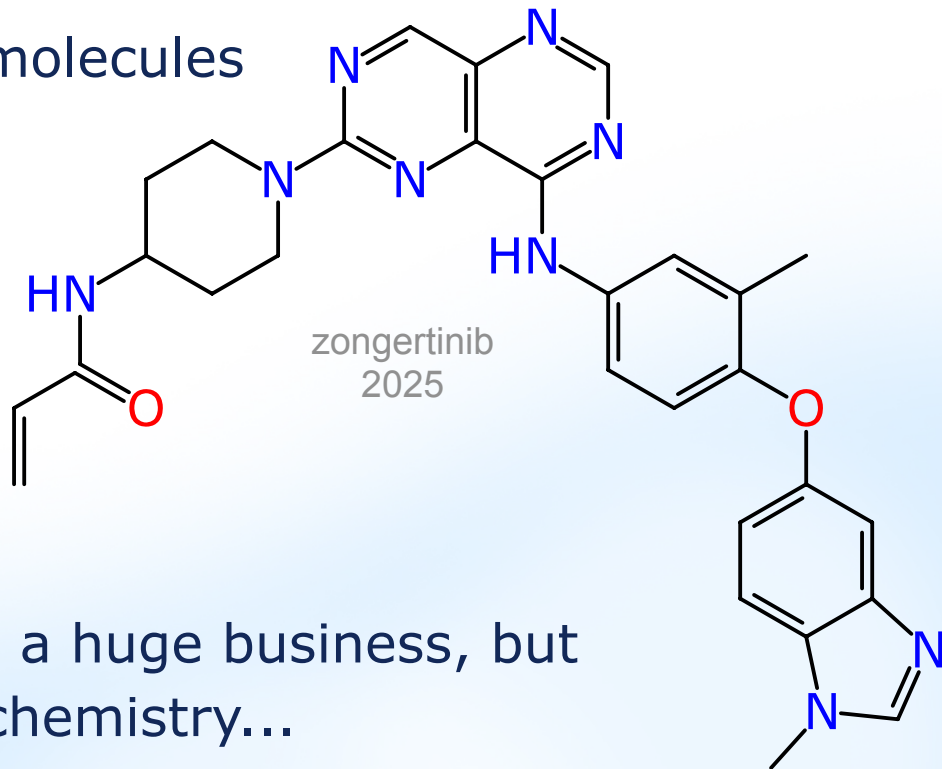


Reactions



Drugs

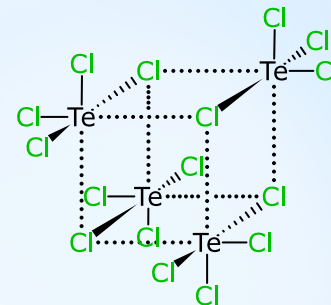
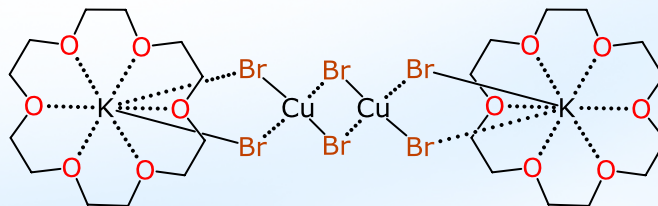
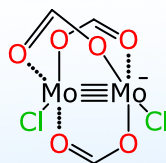
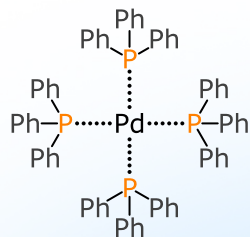
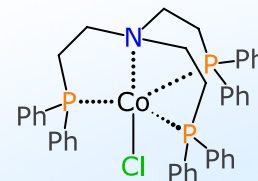
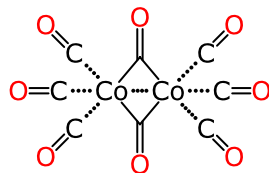
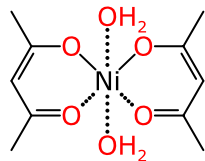
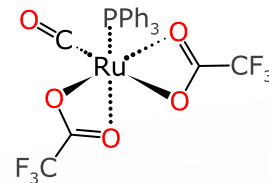
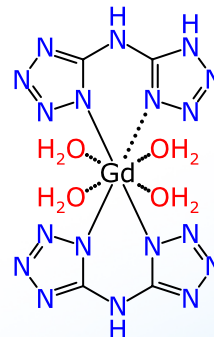
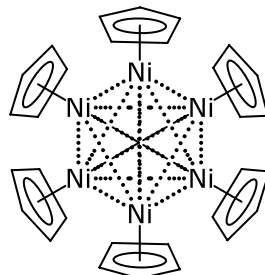
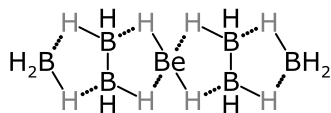
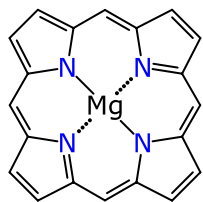
- Designed for organic small molecules
- Limited valence options
- Good enough stereochemistry
- 2D layout: preserves drawing nuances
- Organic small molecules are a huge business, but there is **so much more** to chemistry...



Inorganics



CDD VAOULT
Complexity Simplified

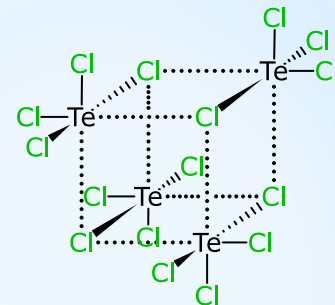
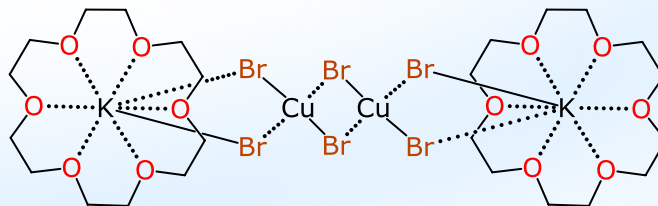
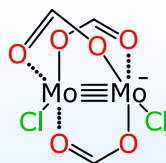
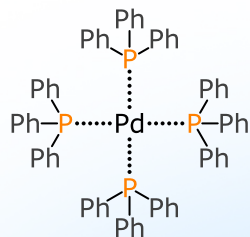
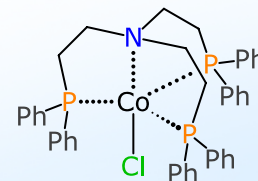
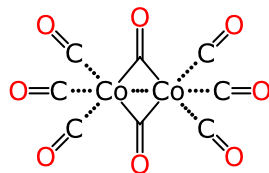
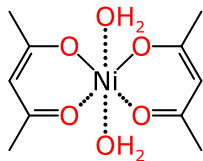
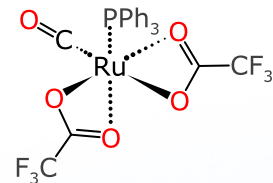
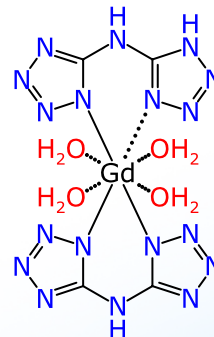
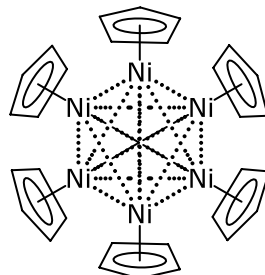
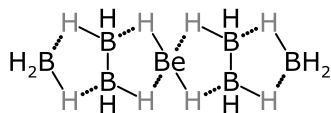
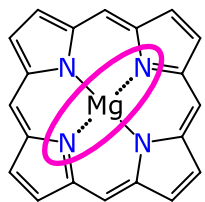


Inorganics



CDD VAOULT
Complexity Simplified

zero order/dative bonds



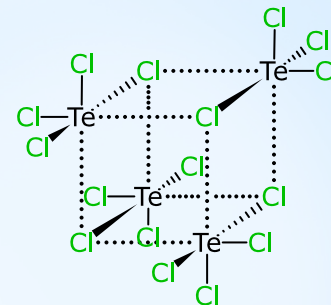
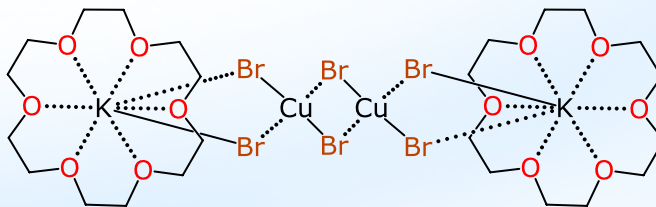
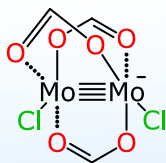
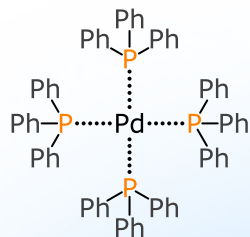
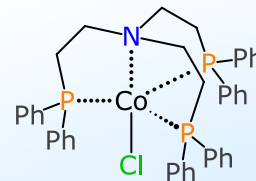
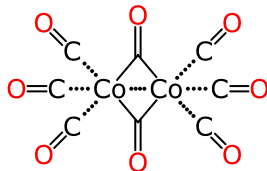
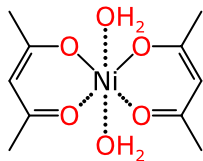
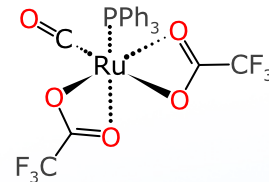
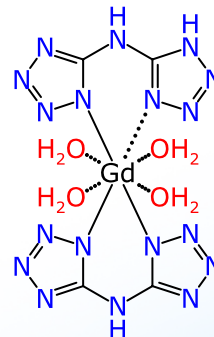
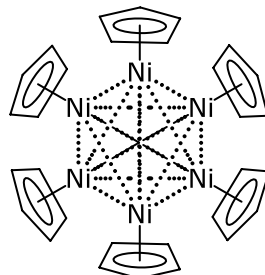
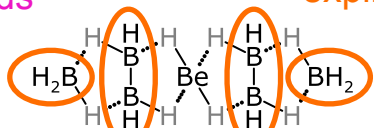
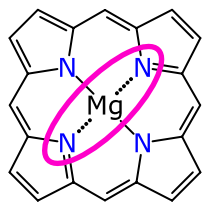
Inorganics



CDD VAOULT
Complexity Simplified

zero order/dative bonds

explicit valence/H-count



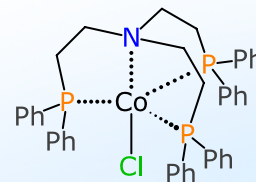
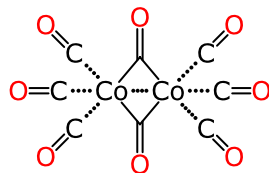
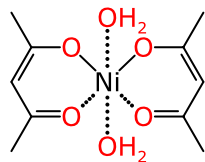
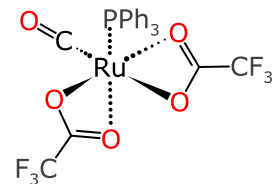
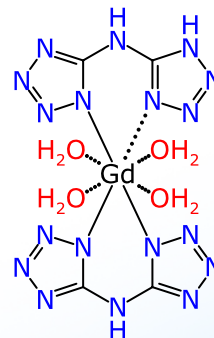
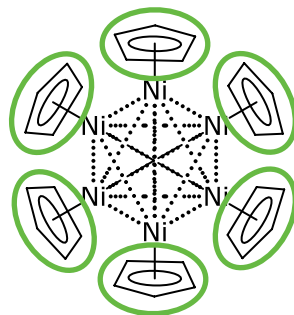
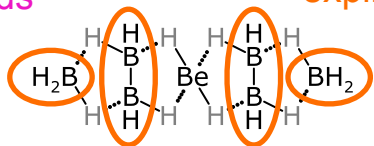
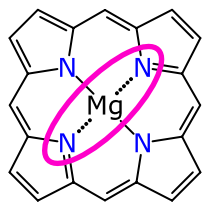
Inorganics



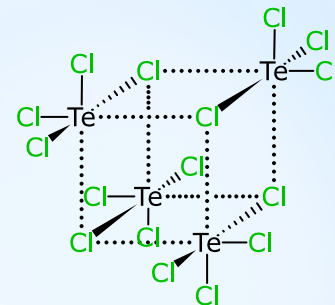
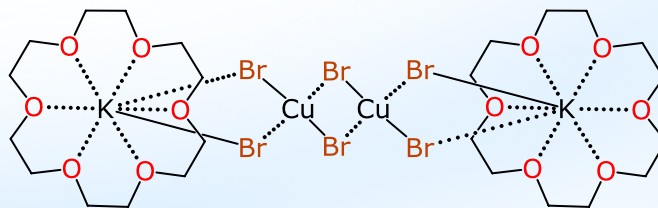
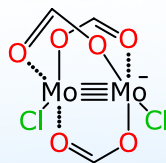
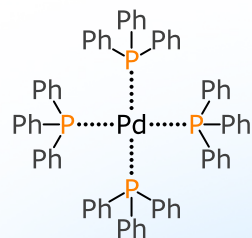
CDD VAOULT
Complexity Simplified

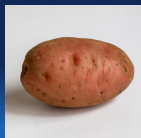
zero order/dative bonds

explicit valence/H-count



haptic/multicoordinate





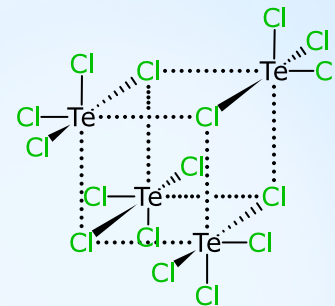
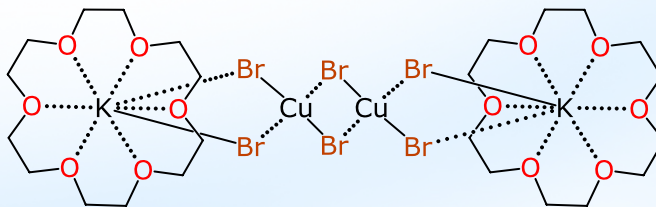
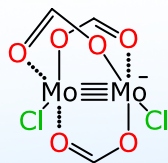
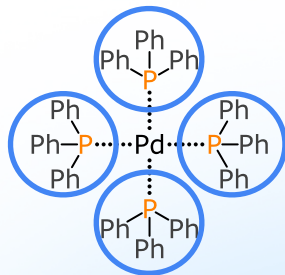
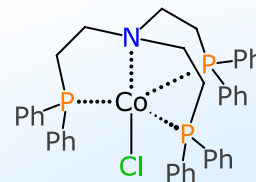
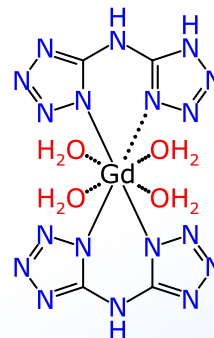
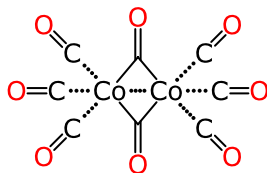
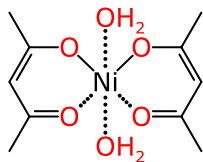
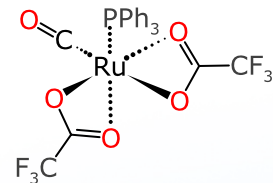
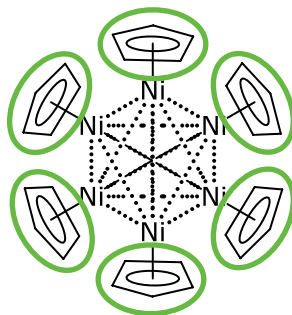
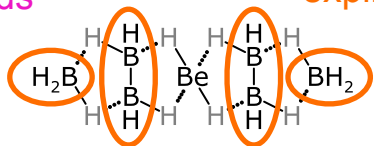
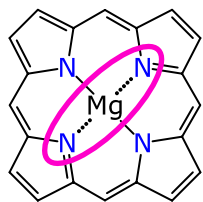
Inorganics



CDD VAOULT
Complexity Simplified

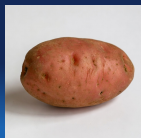
zero order/dative bonds

explicit valence/H-count



haptic/multicoordinate

inline superatom abbreviations

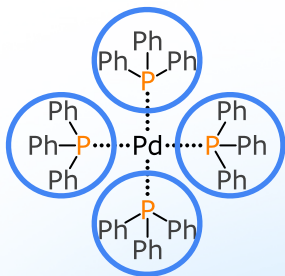
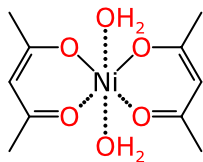
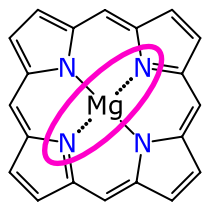


Inorganics

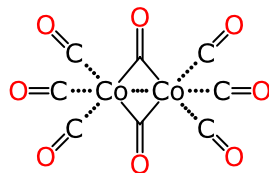
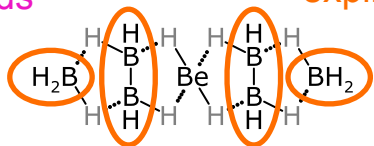


CDD VAOULT
Complexity Simplified

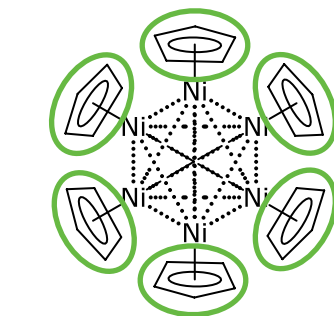
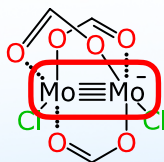
zero order/dative bonds



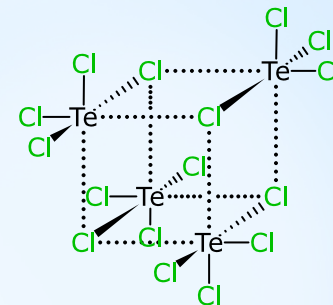
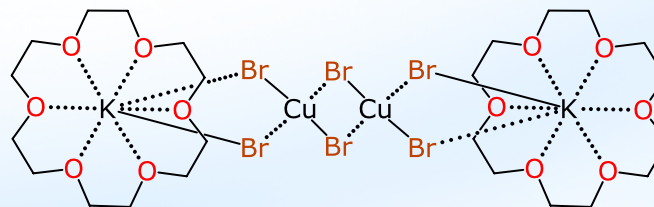
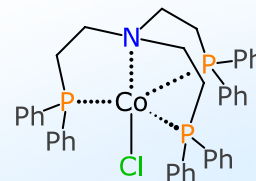
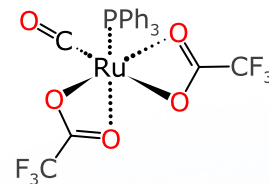
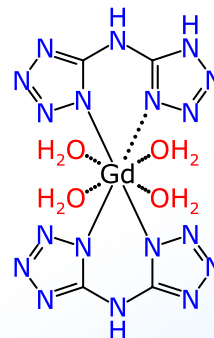
explicit valence/H-count



higher order bonds



haptic/multicoordinate



inline supereatom abbreviations

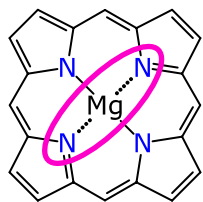


Inorganics

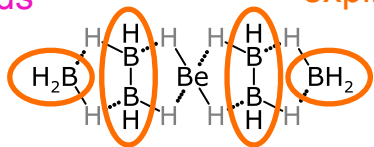


CDD VAOULT
Complexity Simplified

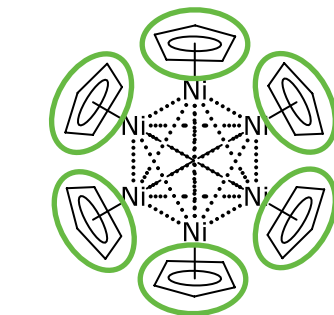
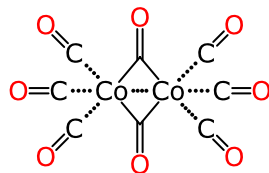
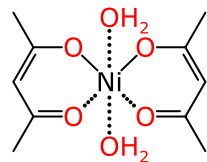
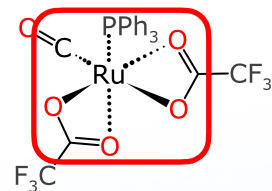
zero order/dative bonds



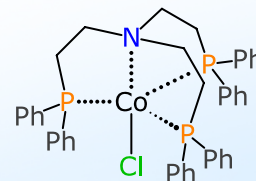
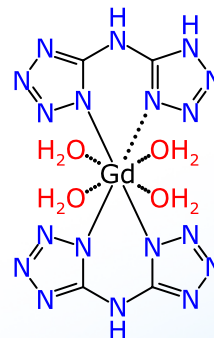
explicit valence/H-count



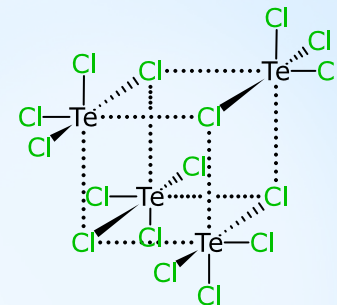
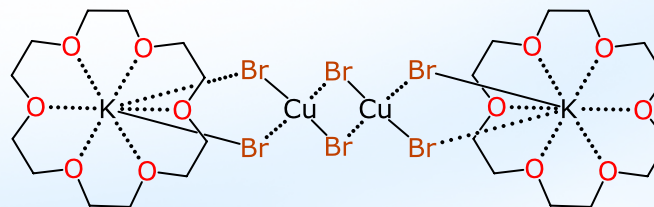
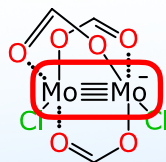
coordination
stereochemistry



haptic/multicoordinate



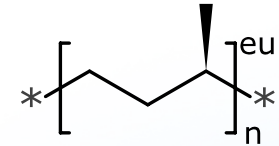
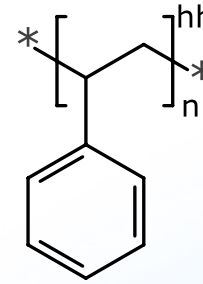
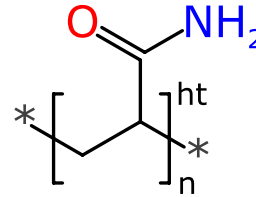
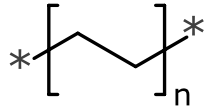
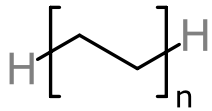
higher order bonds



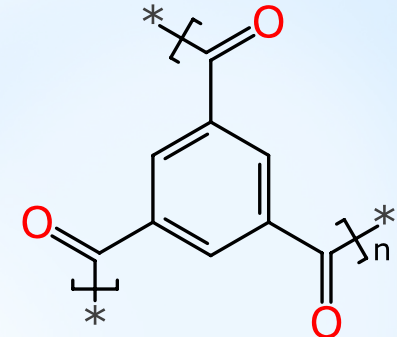
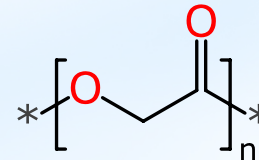
inline supereatom abbreviations



Polymers

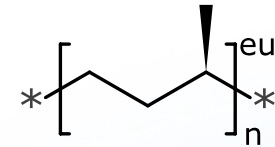
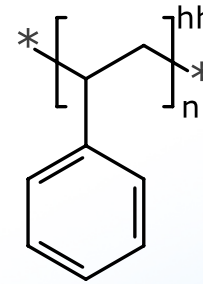
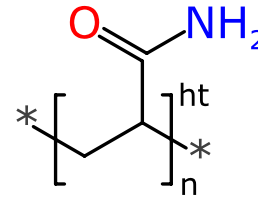
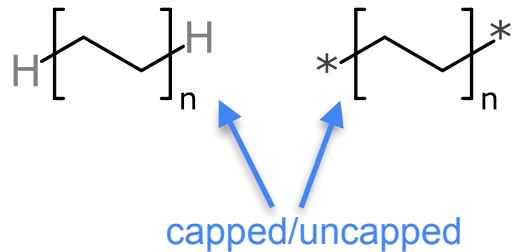


- Repeat units quite well specified
- Statistical distribution: external

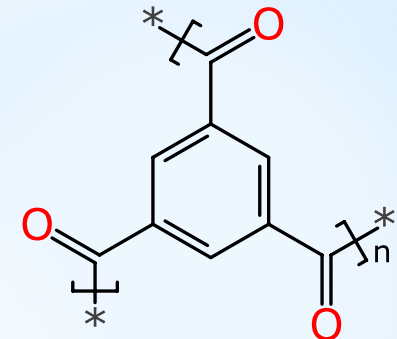
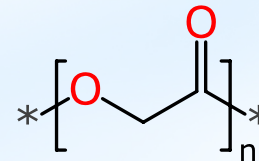




Polymers

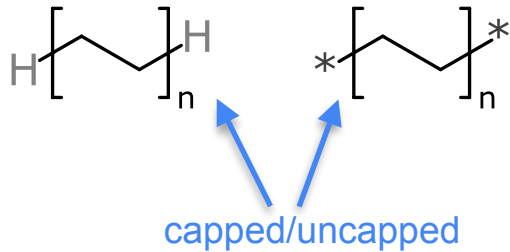


- Repeat units quite well specified
- Statistical distribution: external

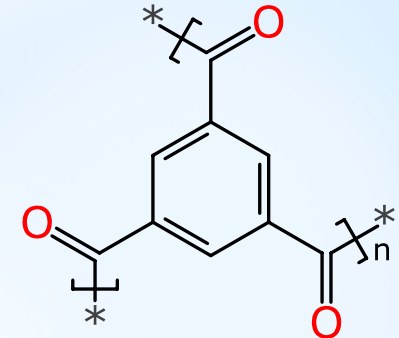
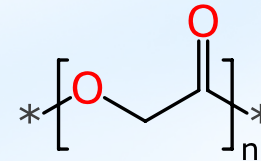
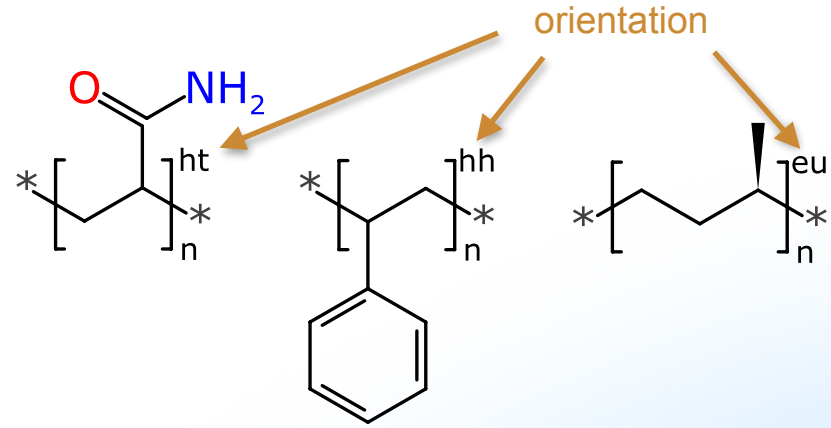




Polymers

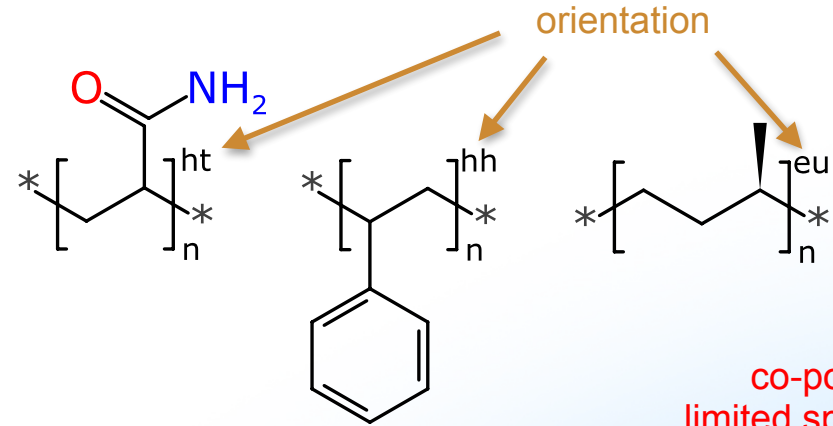
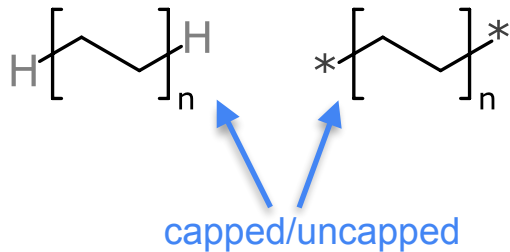


- Repeat units quite well specified
- Statistical distribution: external



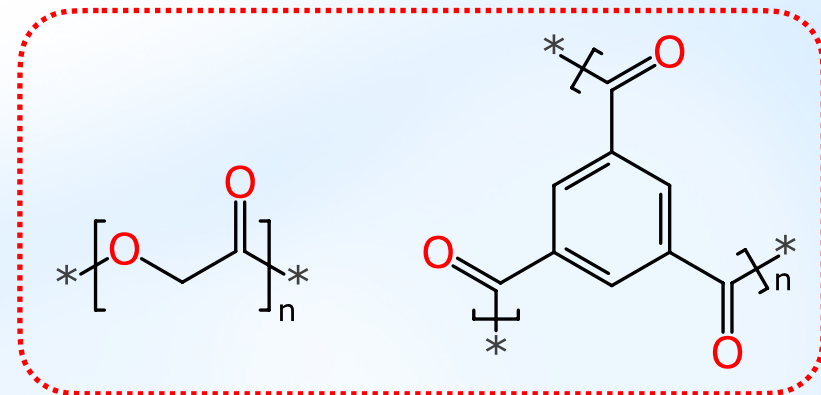


Polymers



co-polymers:
limited specificity

- Repeat units quite well specified
- Statistical distribution: external

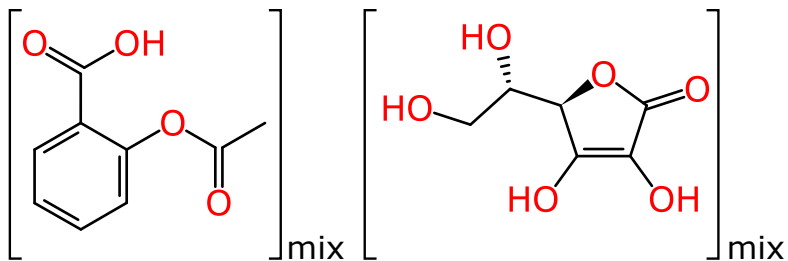




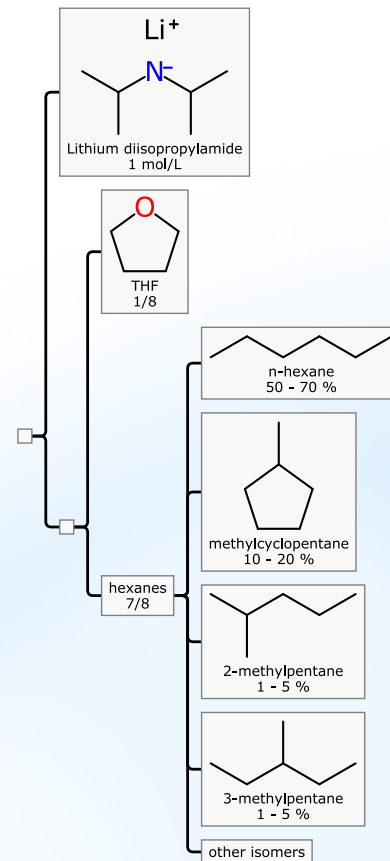
Mixtures



CDD VAULT[®]
Complexity Simplified



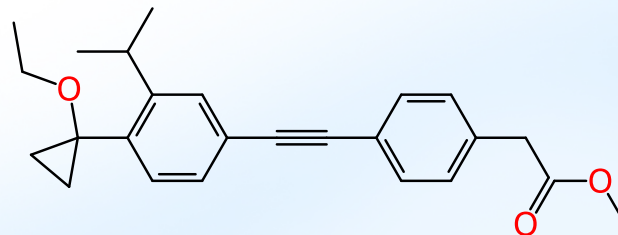
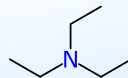
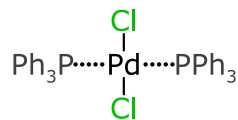
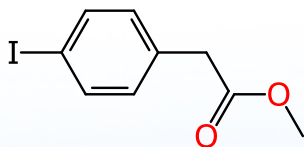
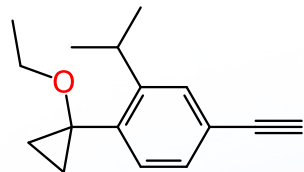
- Inside molfile: limited ability to mark mixture components
- Preferable: specify a *hierarchy* (mixfile)
- Each node can have its own molfile
- Captures precision & uncertainty





Reactions

Cu-I

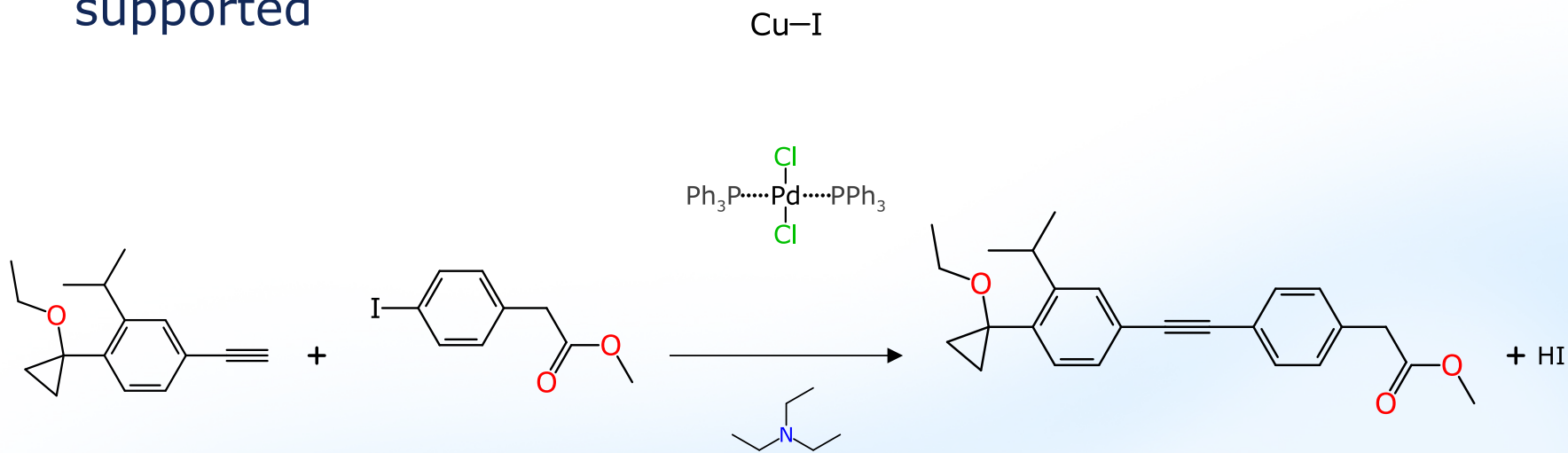


HI



Reactions

- Canvas style not supported

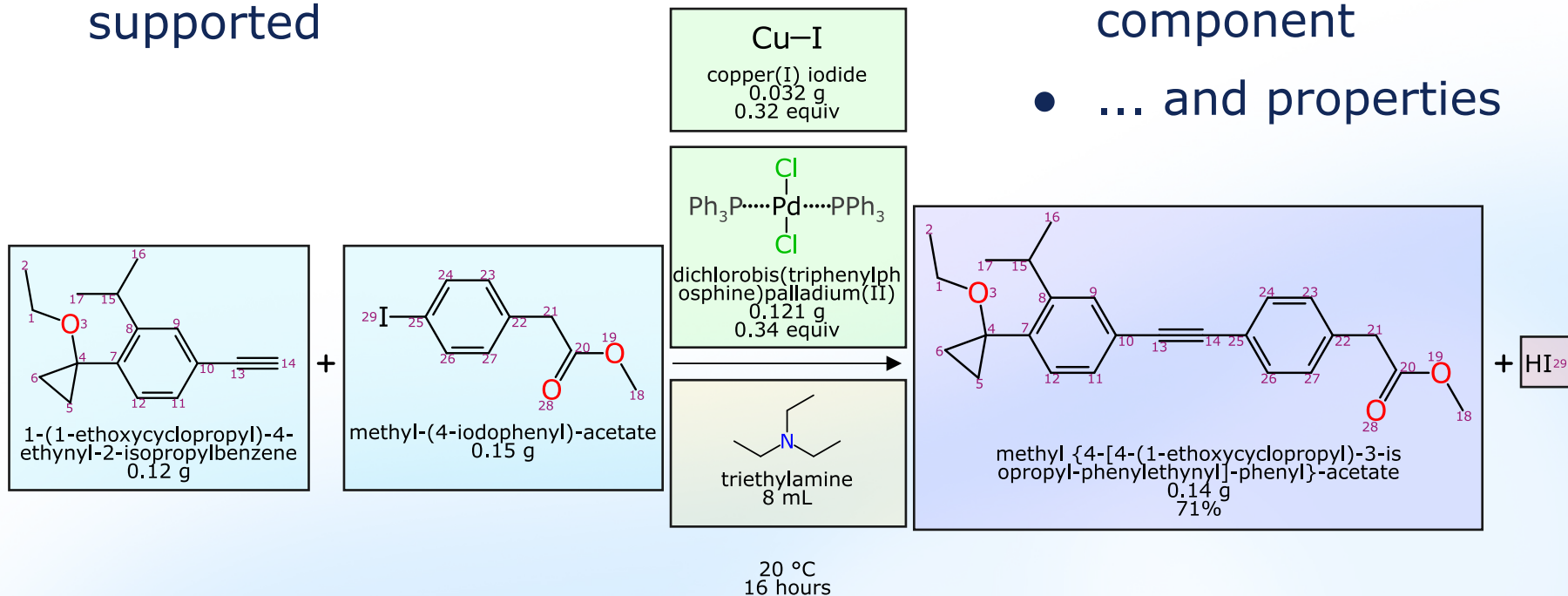




Reactions

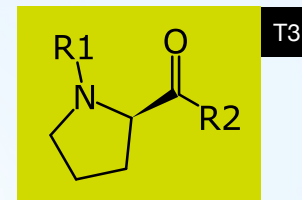
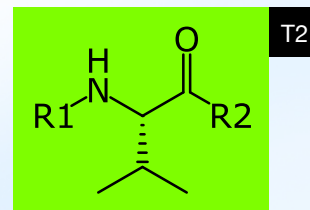
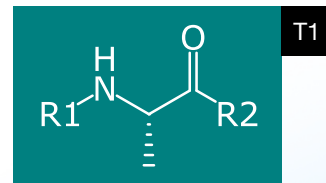
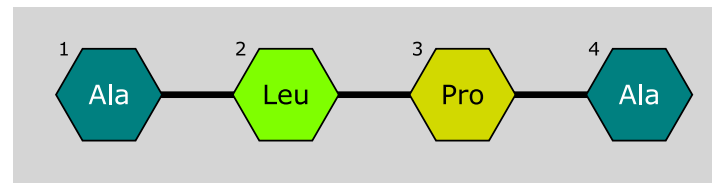
- Canvas style not supported

- One molfile per component
- ... and properties

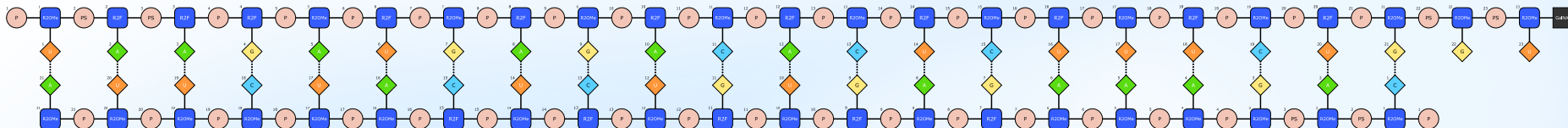
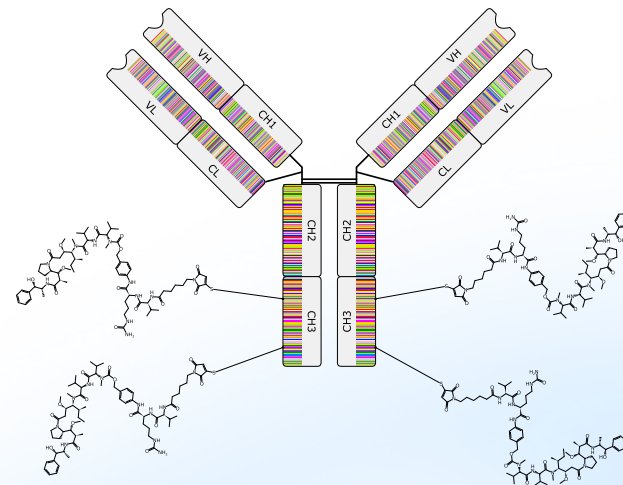
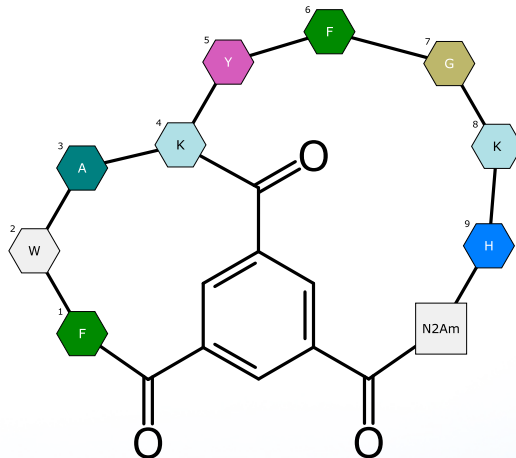
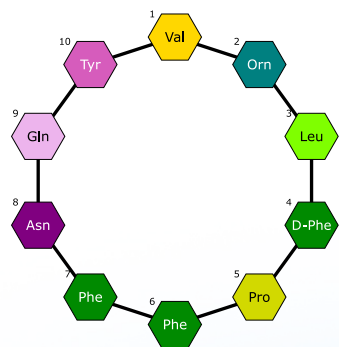
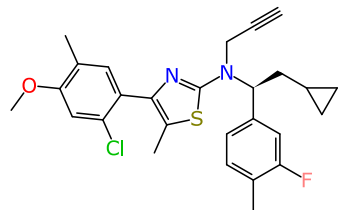


Macromolecules (hiding in plain sight)

- V3000 molfile feature called **SCSR**
- **Layout** information as 2D sketch
- Monomers as **templates**: scalable
- Self contained, no external **dictionary**
- Full **atomic** structure is implied
- Contains the **sequence** data for bioinformatics



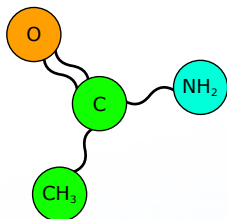
Creeping scope of molecules



Layers of metadata

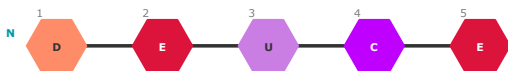
CHEMISTRY

atoms & bonds

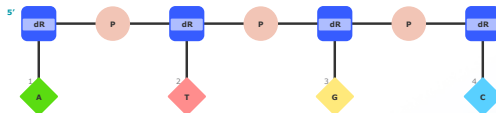


BIOINFORMATICS

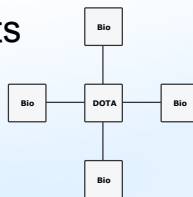
peptides



nucleotides

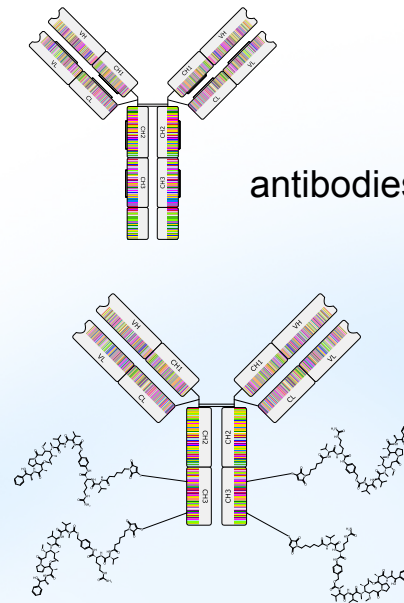


fragments



BIOLOGY

antibodies



- Everything built on chemistry: no formula left behind

Peptides

- Sometimes just sequence is enough

SAFYRVRVCRS
PDPQFSYCVHVF
TFLKLLWSIGIF

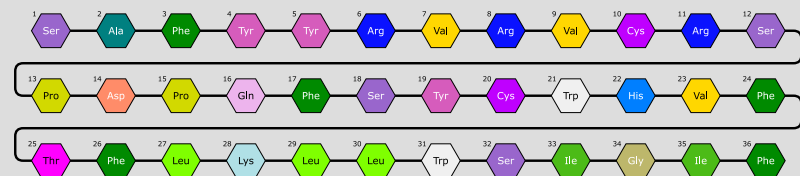


- Exotic oligomers: the sweet spot for **macromolecule** feature

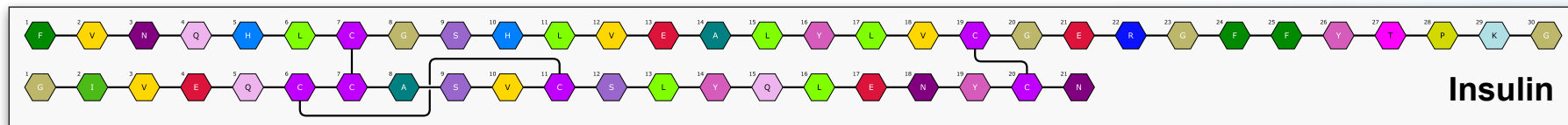
Peptides

- Sometimes just sequence is enough

SAFYRVRVCRS
PDPQFSYCWHF
TFLKLLWSIGIF



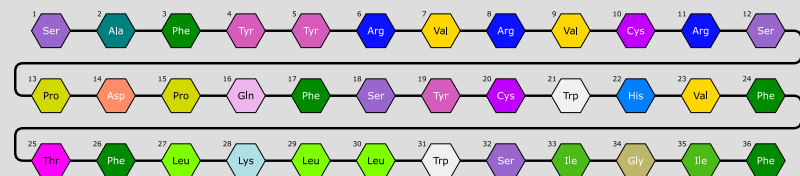
- Exotic oligomers: the sweet spot for **macromolecule** feature



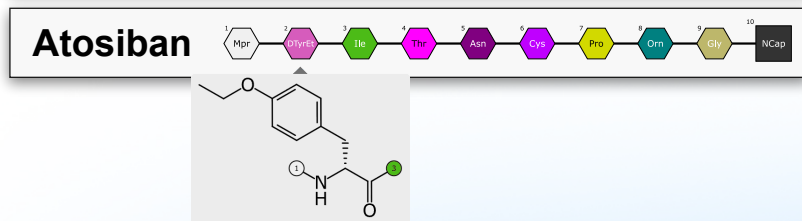
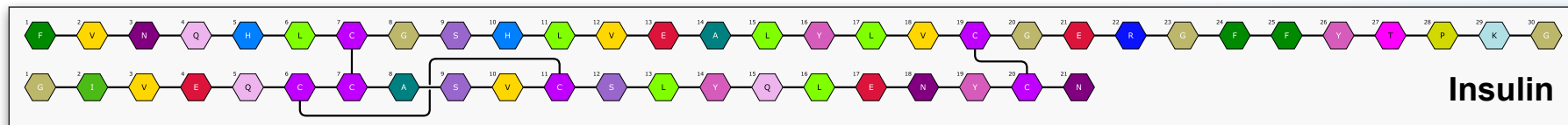
Peptides

- Sometimes just sequence is enough

SAFYRVRVCRS
PDPQFSYCWHF
TFLKLLWSIGIF



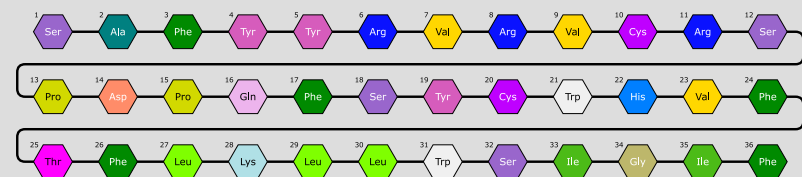
- Exotic oligomers: the sweet spot for **macromolecule** feature



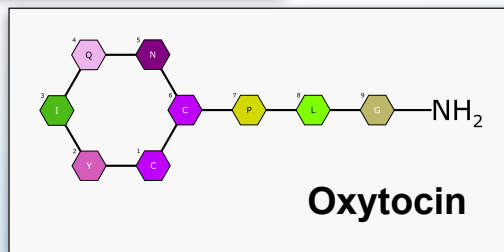
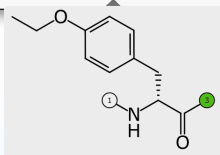
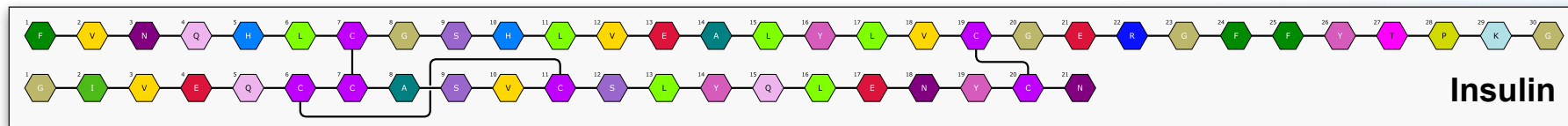
Peptides

- Sometimes just sequence is enough

SAFYRVRVCRS
PDPQFSYCWHF
TFLKLLWSIGIF



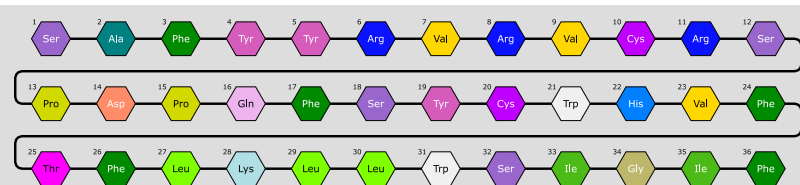
- Exotic oligomers: the sweet spot for **macromolecule** feature



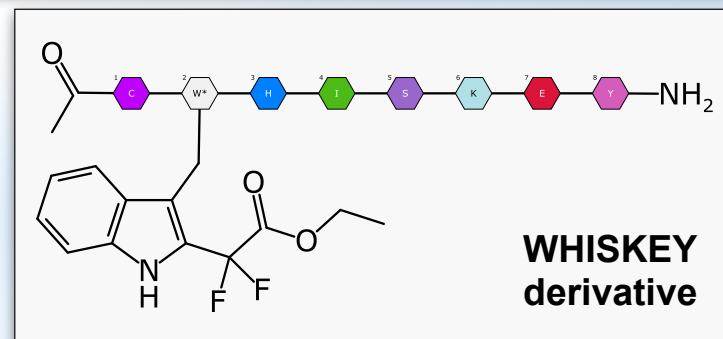
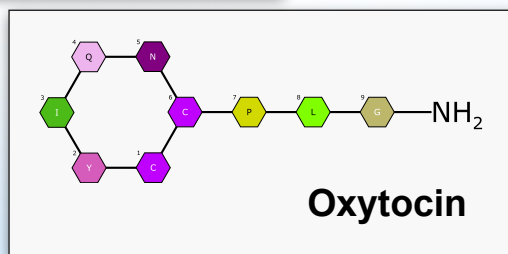
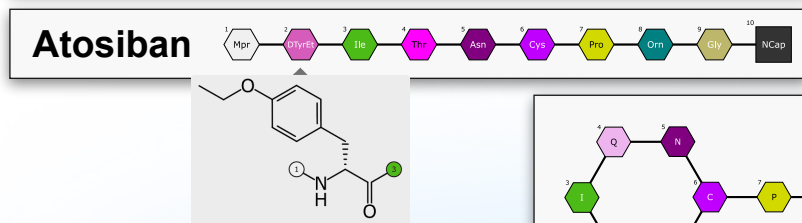
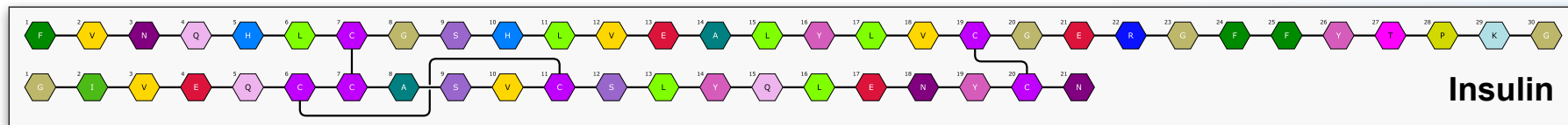
Peptides

- Sometimes just sequence is enough

SAFYRVRVCRS
PDPQFSYCWHF
TFLKLLWSIGIF



- Exotic oligomers: the sweet spot for **macromolecule** feature



Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT - McKerrow Vault

AI Chat Help Log out

Full-Access User
full_access

Explore Data ELN Inventory **Import Data** Reports Settings 4

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

File: syntholigo.csv Project: Macromolecule Owner: Full-Access User

Compose macromolecules from columns [Edit](#)

☐ Not part of macromolecule ☒ Linear peptide sequence ☐ Cyclic peptide sequence ☐ Nucleotide sequence ☐ Antibody/ADC ☐ siRNA strands

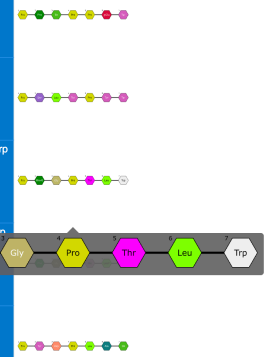
Wrapping: ☒ Off ☐ On

Width: 12 units

☐ Translate DNA/RNA codons into amino acid codes

280 custom peptides,
29 custom nucleotides,
11 custom fragments

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



Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT - McKerrow Vault

AI Chat Help Log out

Full-Access User
full_access

Explore Data ELN Inventory Import Data Reports Settings 4

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

File: syntholigo.csv Project: Macromolecule Owner: Full-Access User

Compose macromolecules from columns [Edit](#)

☐ Not part of macromolecule ☒ Linear peptide sequence ☐ Cyclic peptide sequence ☐ Nucleotide sequence ☐ Antibody/ADC ☐ siRNA strands

Orientation: ☒ North ☐ East ☐ South ☐ West

Direction: ☒ Clockwise ☐ Anti-clockwise

280 custom peptides,
29 custom nucleotides,
11 custom fragments

	A	B
	Name	Sequence
1	Synth1	Pro,Phe,Ile,Pro,Pro,Glu,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,Leu,Tyr
5	Synth5	Pro,Tyr,Asp,Pro,Leu,Ala,

Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT - McKerrow Vault

Explore Data ELN Inventory **Import Data** Reports Settings 4

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

File: syntholigo.csv Project: Macromolecule Owner: Full-Access User

Preview Macromolecule

SVG PNG Molfile Original Identifiers

Size 172 x 166 ☐ Full atom layout

☐ Not part of macromolecule
☐ Linear peptide sequence
☒ Cyclic peptide sequence
☐ Nucleotide sequence
☐ Antibody/ADC
☐ siRNA strands

	A	B
	Name	Sequence
1	Synth1	Pro,Phe,Ile,Pro,Pro,g
2	Synth2	Pro,Ser,Leu,Tyr,Pro,T
3	Synth3	Pro,Phe4,Gly,Pro,Th
4	Synth4	Pro,Phe,Gly,Pro,Gln,U
5	Synth5	Pro,Tyr,Asp,Pro,Leu,U

280 custom peptides,
29 custom nucleotides,
11 custom fragments

Download Copy

Close

Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT - McKerrow Vault

Explore Data ELN Inventory Import Data Reports Settings 4

AI Chat Help Log out Full-Access User - full_access

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

File: syntholigo.csv Project: Macromolecule Owner: Full-Access User

Preview Macromolecule

SVG PNG Molfile Original Identifiers

Size 615 x 518 ☒ Full atom layout

☐ Not part of macromolecule
☐ Linear peptide sequence
☒ Cyclic peptide sequence
☐ Nucleotide sequence
☐ Antibody/ADC
☐ siRNA strands

A	B
Name	Sequence
1 Synth1	Pro,Phe,Ile,Pro,Pro,g
2 Synth2	Pro,Ser,Leu,Tyr,Pro,T
3 Synth3	Pro,Phe4I,Gly,Pro,Th
4 Synth4	Pro,Phe,Gly,Pro,Gln,U
5 Synth5	Pro,Tyr,Asp,Pro,Leu,A

280 custom peptides,
29 custom nucleotides,
11 custom fragments

scale 81%

Download Copy

Close

Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT · McKerrow Vault

AI Chat Help Log out

Full-Access User
full_access

Explore Data ELN Inventory **Import Data** Reports Settings 4

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

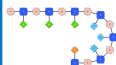
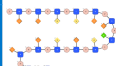
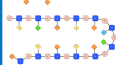
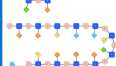
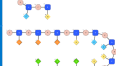
File: ma.csv Project: Macromolecule Owner: Full-Access User

Compose macromolecules from columns [Edit](#)

☐ Not part of macromolecule
☐ Linear peptide sequence
☐ Cyclic peptide sequence
☒ Nucleotide sequence
☐ Antibody/ADC
☐ siRNA strands

Backbone: ☒ RNA ☐ DNA
Strands: ☒ Single ☐ Double
☐ Cleave terminal 5' phosphate

276 custom peptides,
29 custom nucleotides,
11 custom fragments

	A	B	
	Name	Sequence	
1	NonCode1	AAACCGU	
2	NonCode2	UUGGUUAGUGUUUU	
3	NonCode3	GAUGUCAUGUGUAU	
4	NonCode4	UUGCCGAUUUGCCG	
5	NonCode5	UUUGCCGGAAUUU	



Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT · McKerrow Vault

AI Chat Help Log out

Full-Access User

Full_access

Explore Data ELN Inventory Import Data Reports Settings 4

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

File: ma.csv Project: Macromolecule Owner: Full-Access User

Preview Macromolecule

SVG PNG Molfile Original Identifiers

Size 697 × 461 ☐ Full atom layout

☐ Not part of macromolecule
☐ Linear peptide sequence
☐ Cyclic peptide sequence
☒ Nucleotide sequence
☐ Antibody/ADC
☐ siRNA strands

	A	B
	Name	Sequence
1	NonCode1	AAACCGU
2	NonCode2	UUGGUUAGUGU
3	NonCode3	GAUGUCAUGUG
4	NonCode4	UUGCCGAUUUG
5	NonCode5	UUUGCCGGAAM

276 custom peptides,
29 custom nucleotides,
11 custom fragments

Download Copy

Close



Bulk import with Vault

- Incoming data:
 - sequences
 - metadata
- Markup to macromolecules
- Custom types
- Custom monomers

CDD.VAULT · McKerrow Vault

Explore Data · ELN · Inventory · **Import Data** · Reports · Settings · 4

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

File: dna.csv

☐ Not part of macromolecule
☐ Linear peptide sequence
☐ Cyclic peptide sequence
☒ Nucleotide sequence
☐ Antibody/ADC
☐ siRNA strands

A	B
Name	Sequence
1 NonCode1	AAACCGT
2 NonCode2	TTGGTTAGTGT
3 NonCode3	GATGTCATGTGT
4 NonCode4	TTGCCGATTGG
5 NonCode5	TTTGCCGGAAT

Preview Macromolecule

SVG PNG Molfile Original Identifiers

Size 940 x 820 ☐ Full atom layout

276 custom peptides, 29 custom nucleotides, 11 custom fragments

scale 81%

Download Copy

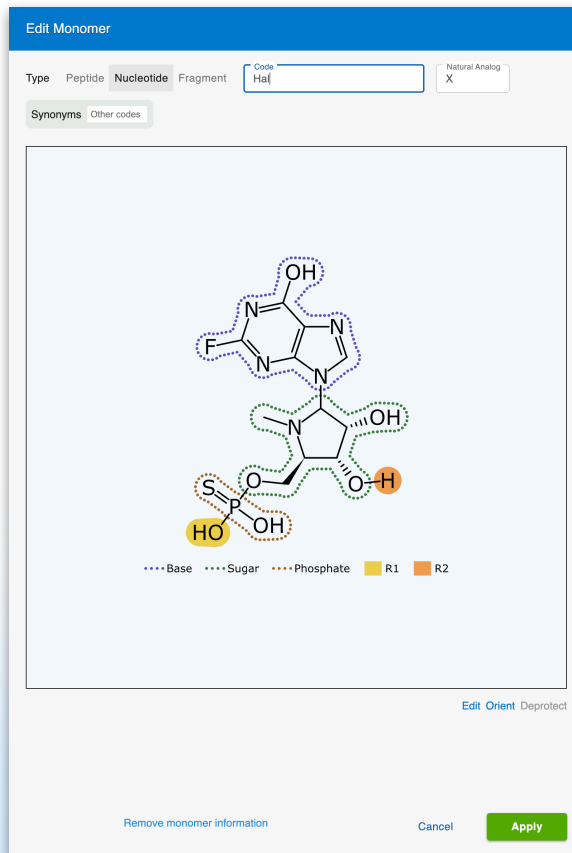
Close

Custom monomers



CDD VAULT
Complexity Simplified

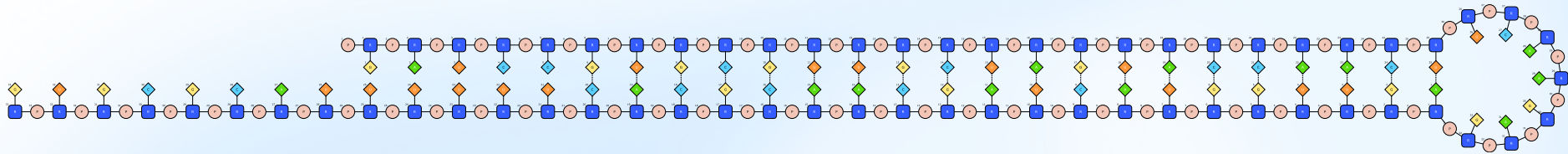
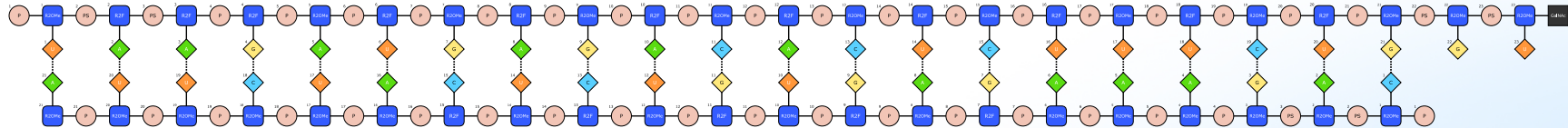
- Monomer management is essential
- Available at time of building
- Then internal to molfile



A → Structure			C → Natural	
SDF Molfile			Name	Natural
1		Aad	Aad	K
2		Abu	A	
3		Aca	X	
4		Alb	X	
5		Aprm	X	
6		App	X	
7		Asu	X	



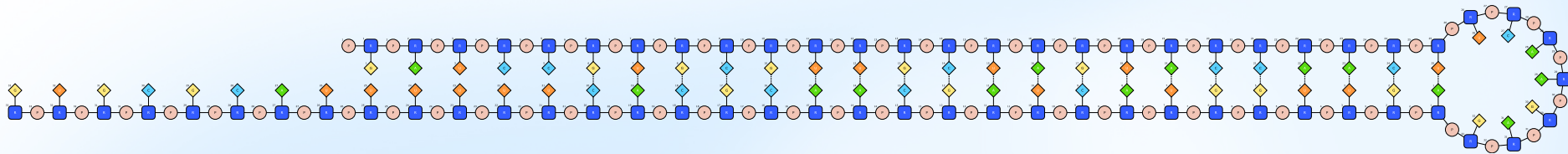
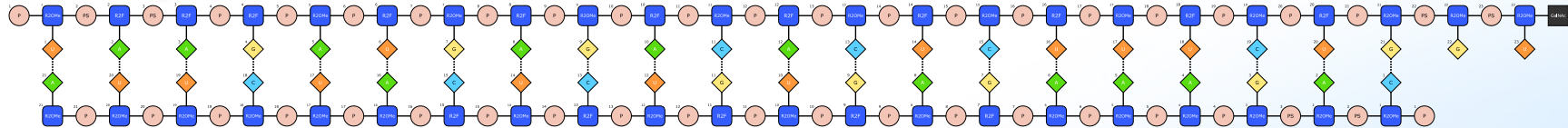
- Two-stranded RNA/DNA: more control





- Two-stranded RNA/DNA: more control

custom phosphates & sugars

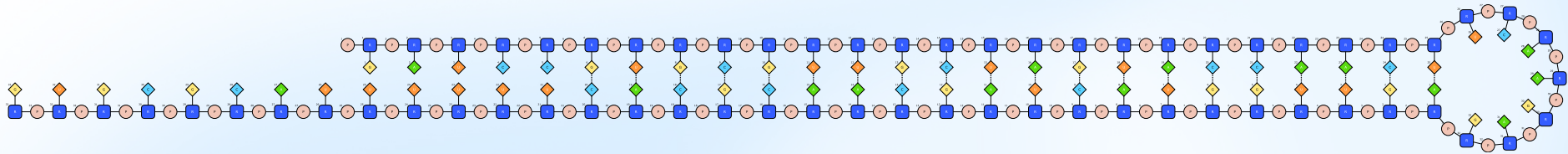
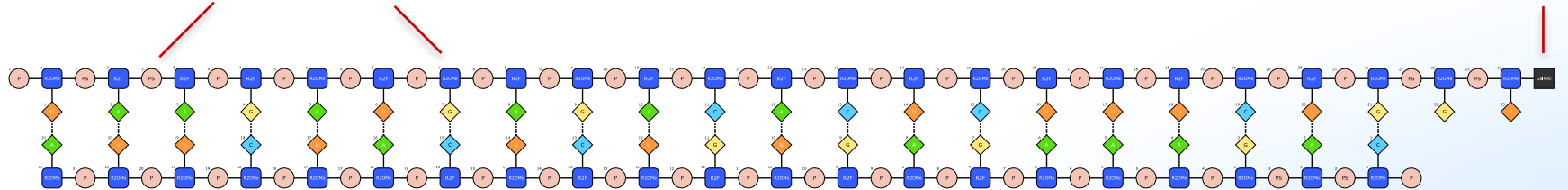




- Two-stranded RNA/DNA: more control

custom phosphates & sugars

end-cap





end-cap

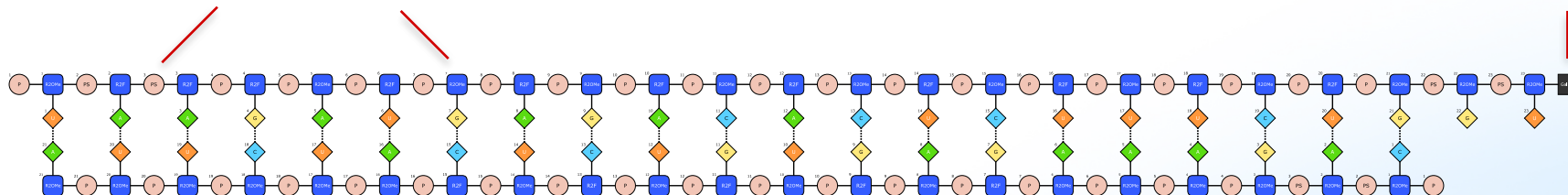




- Two-stranded RNA/DNA: more control

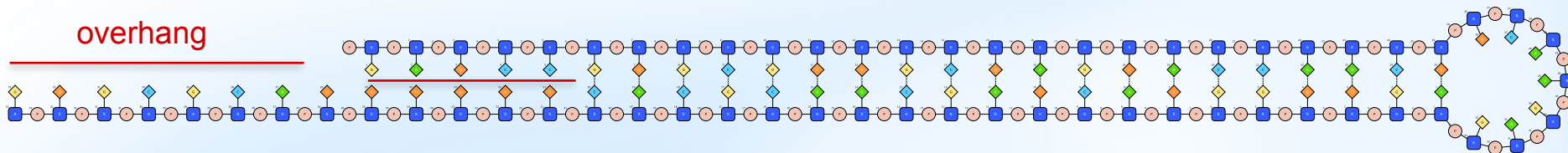
custom phosphates & sugars

end-cap



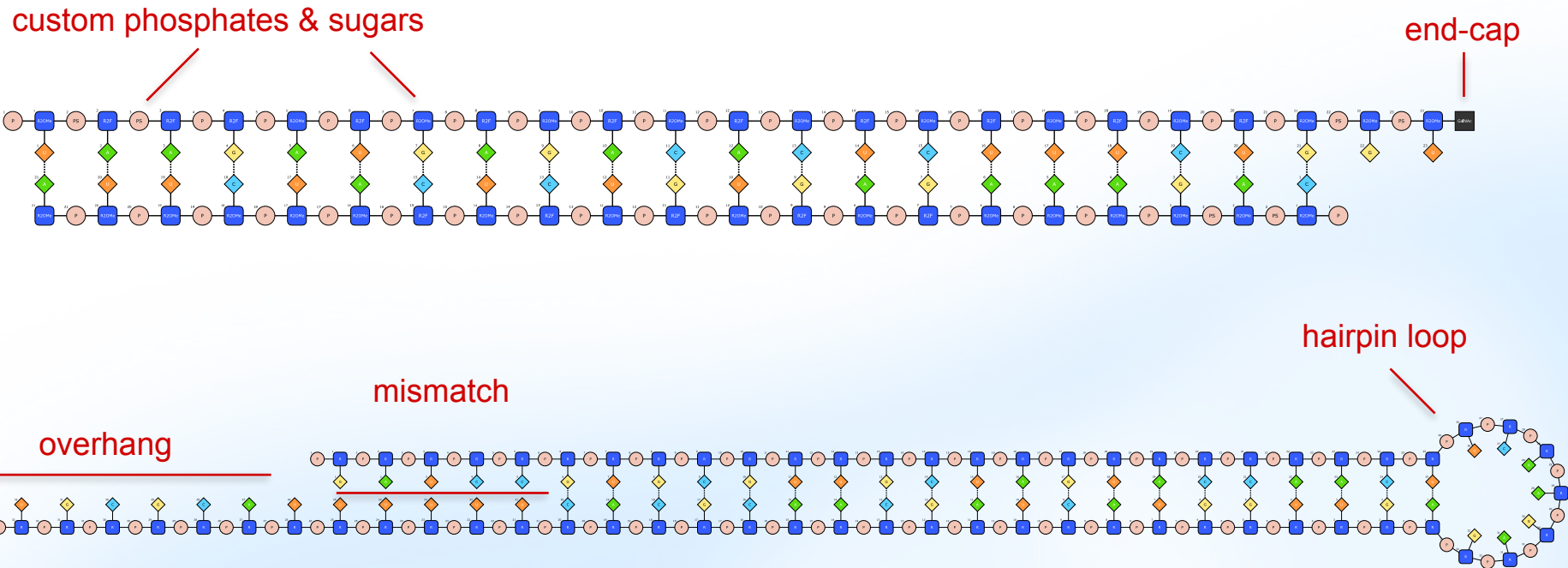
mismatch

overhang





- Two-stranded RNA/DNA: more control



Antibodies

Light

DIQMTQSPSSLSASVGDRVTITCRASQDVNTAVAWYQQKPGKAPKLLIYSASFLYSGVPSRFSGSR
SGTDFTLTISLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVV
CLLNNFYPPREAKVQWKVDNALQSGNSQESVTEQDSKDYSLSSSTLTLSKADYEKHKVYACEVTH
QGLSSPVTKSFNRGEC{VL:1-107}{CL:108-214}{@214:1}

Heavy

EVQLVESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVK
GRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLLTVSSASTKGPSVF
PLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVTVTPSSS
LGTQTYICNVNHKPSNTKVDKKVEPKSCDKHTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRTPE
VTCVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKEYKC
KVSNAKALPAIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESNGQP
ENNYKTTTPVLDSDGSFFLYSKLTVDKSRWQQGNVFCFSVMHEALHNHYTQKSLSLSPGK
{VH:1-120}{CH1:121-218}{CH2:234-343}{CH3:344-450}{H:219-233}{@223:1}{@229:2}
{@232:3}{S264:1}{S324:1}{S370:1}{S428:1}

Domain Region

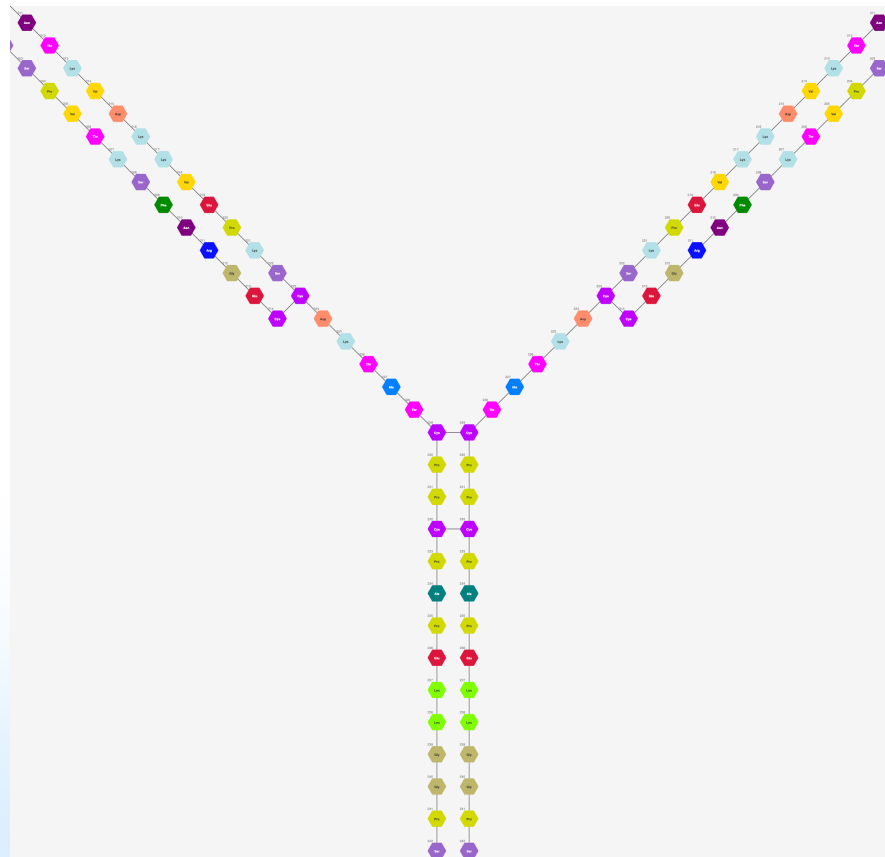
Disulfide Bridge

Hinge Region

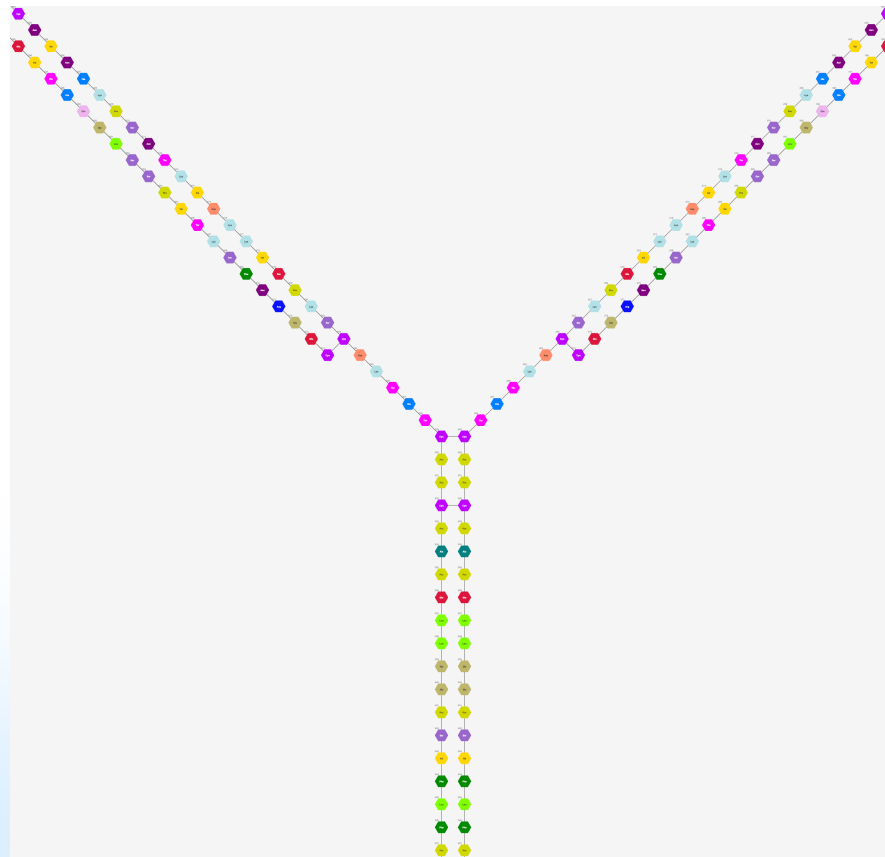
Drug Conjugate Site



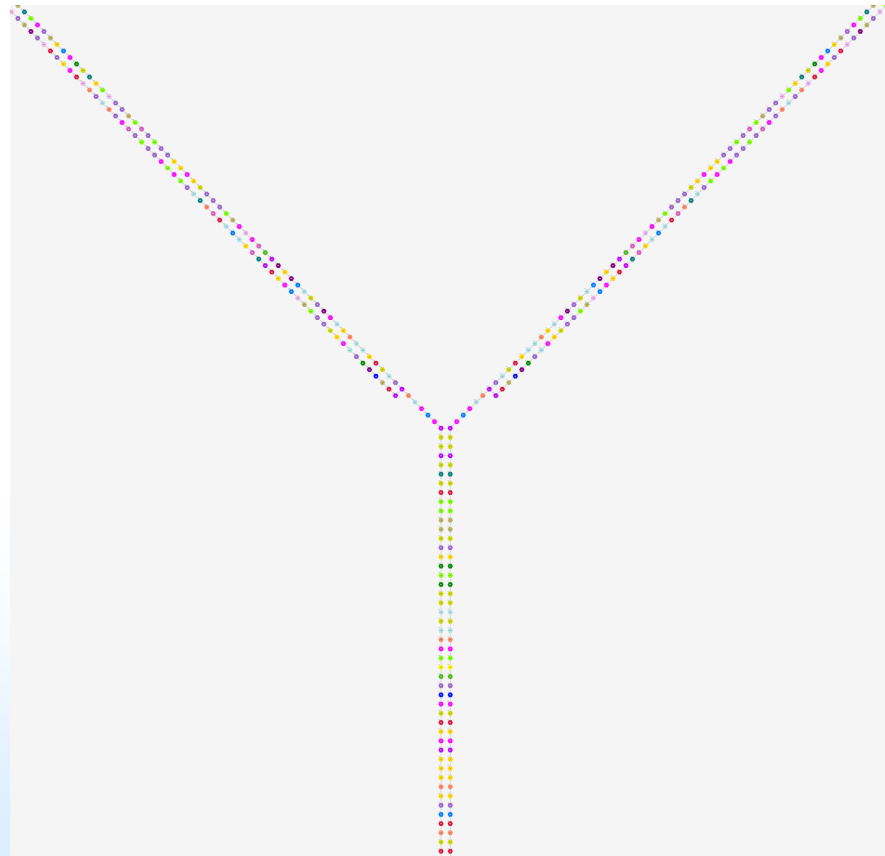
Antibodies



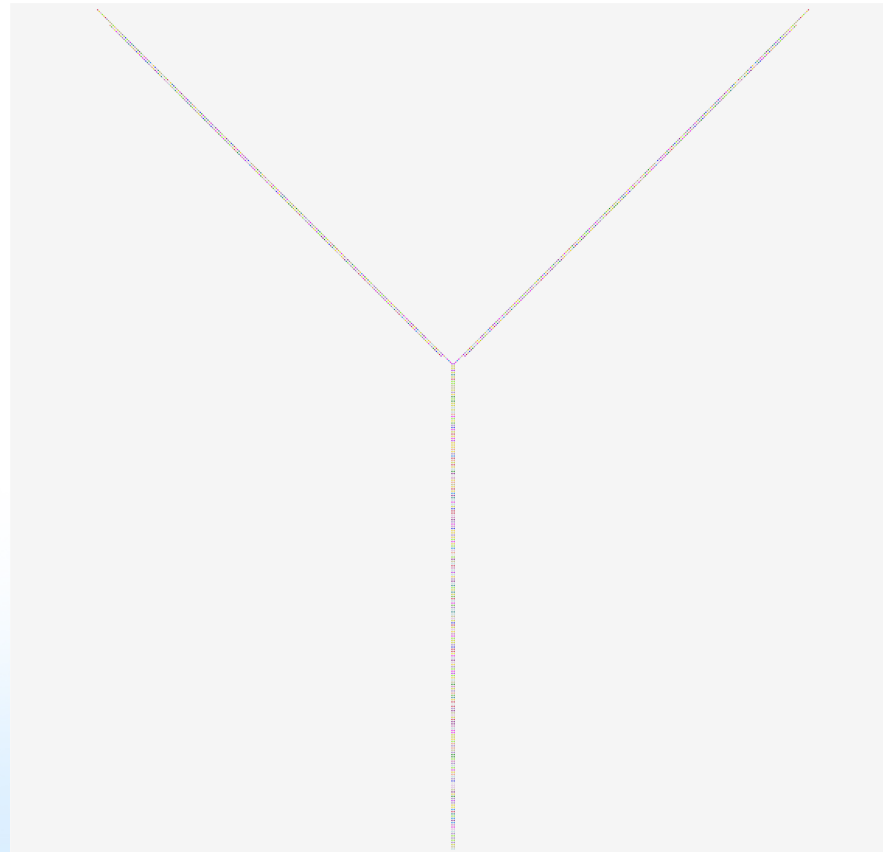
Antibodies



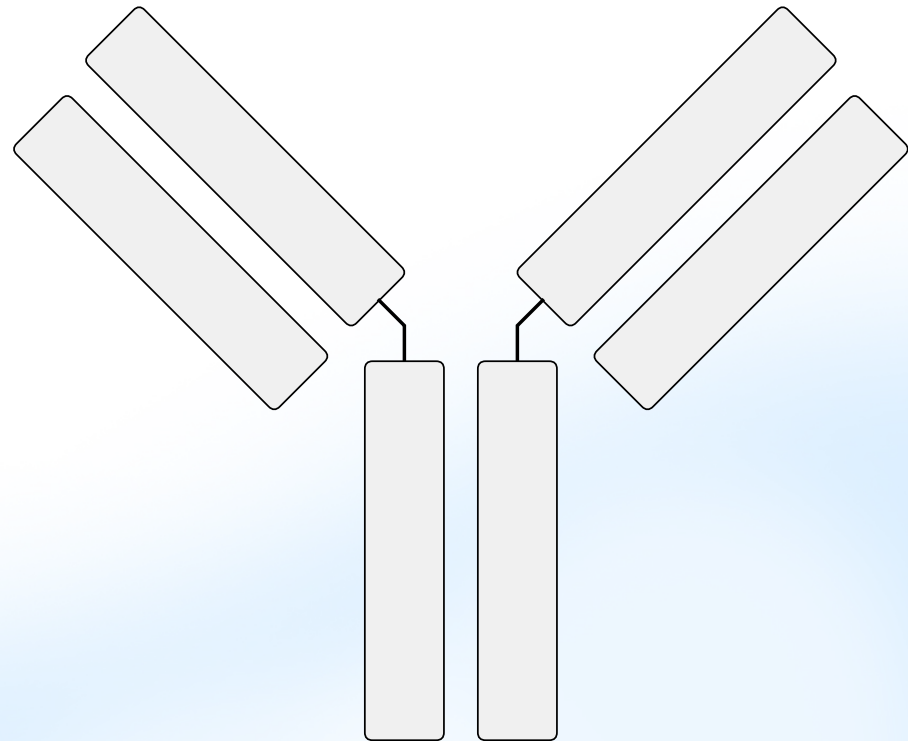
Antibodies



Antibodies



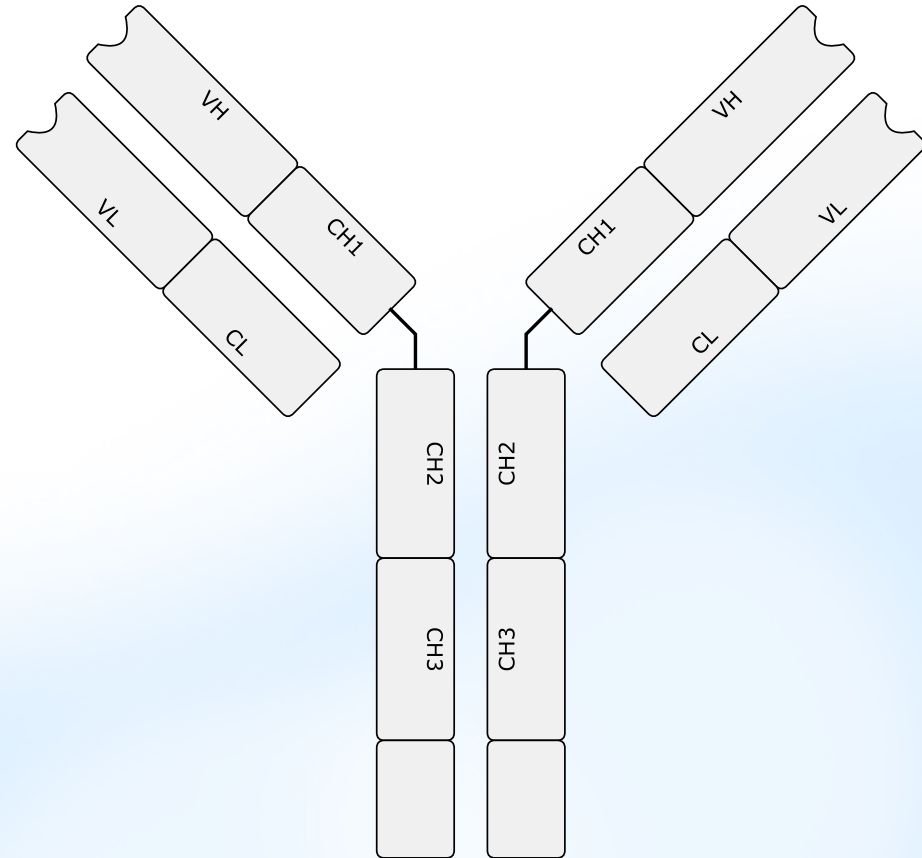
- Draw basic Y-shape outline





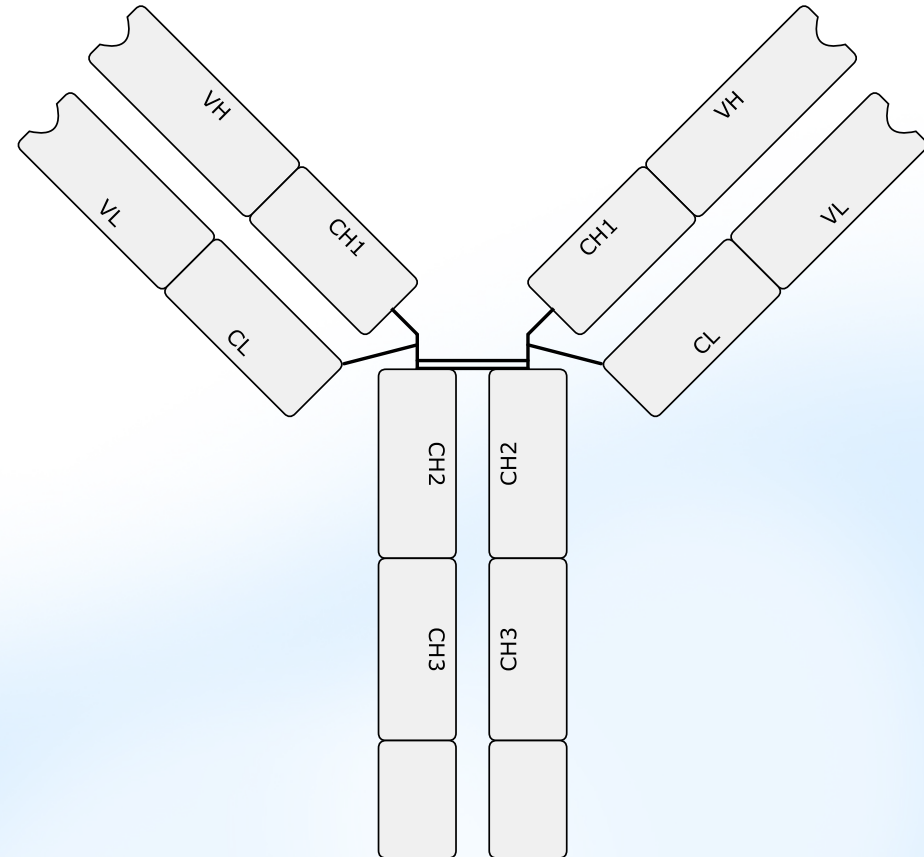
Biology layer

- Draw basic Y-shape outline
- Segment domain regions



Biology layer

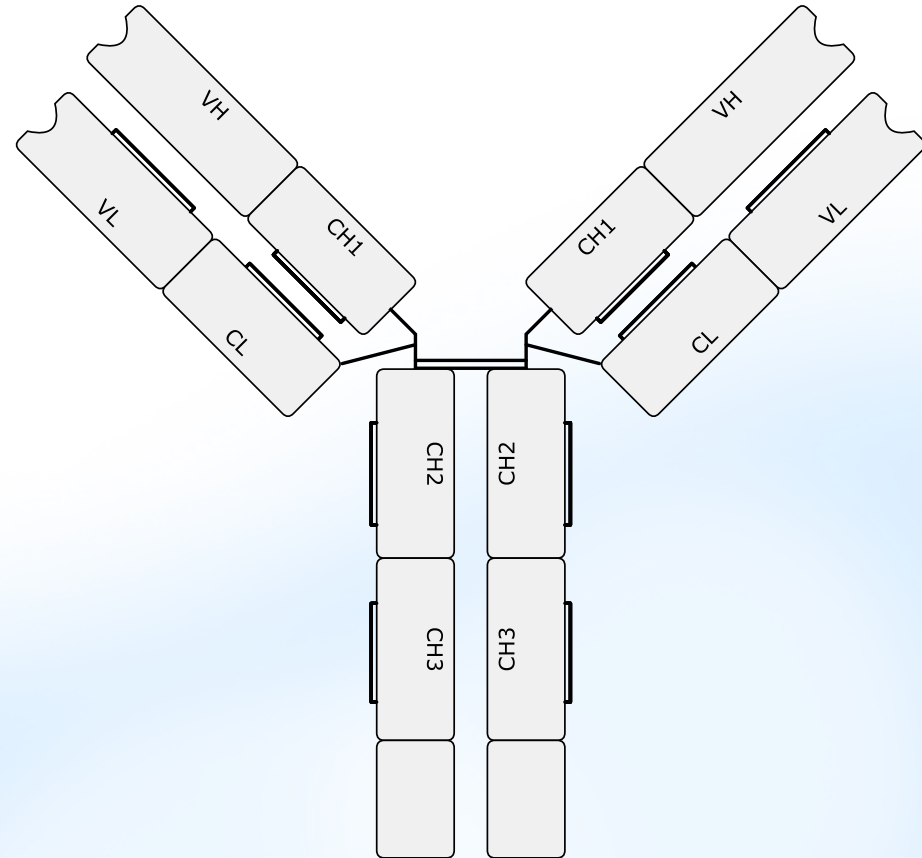
- Draw basic Y-shape outline
- Segment domain regions
- Inter-chain disulfide links





Biology layer

- Draw basic Y-shape outline
- Segment domain regions
- Inter-chain disulfide links
- Intra-chain disulfide links

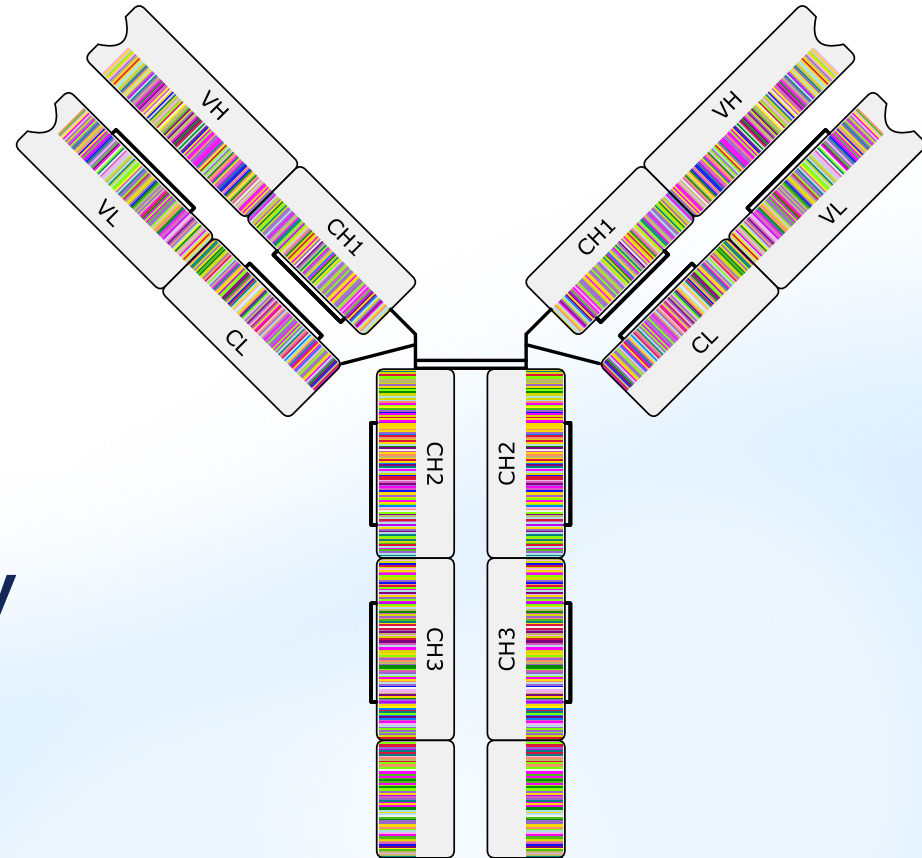


Biology layer



CDD.VAULT
Complexity Simplified

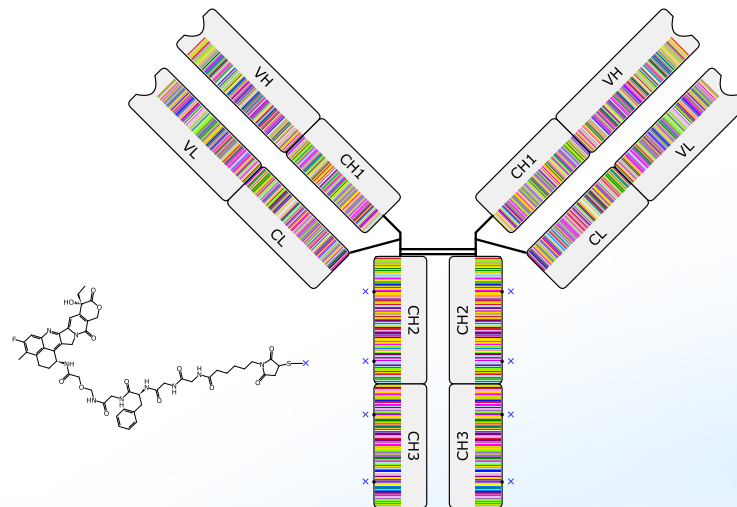
- Draw basic Y-shape outline
- Segment domain regions
- Inter-chain disulfide links
- Intra-chain disulfide links
- Sequence barcoding
- ... **chemistry, biochemistry**
and **biology**





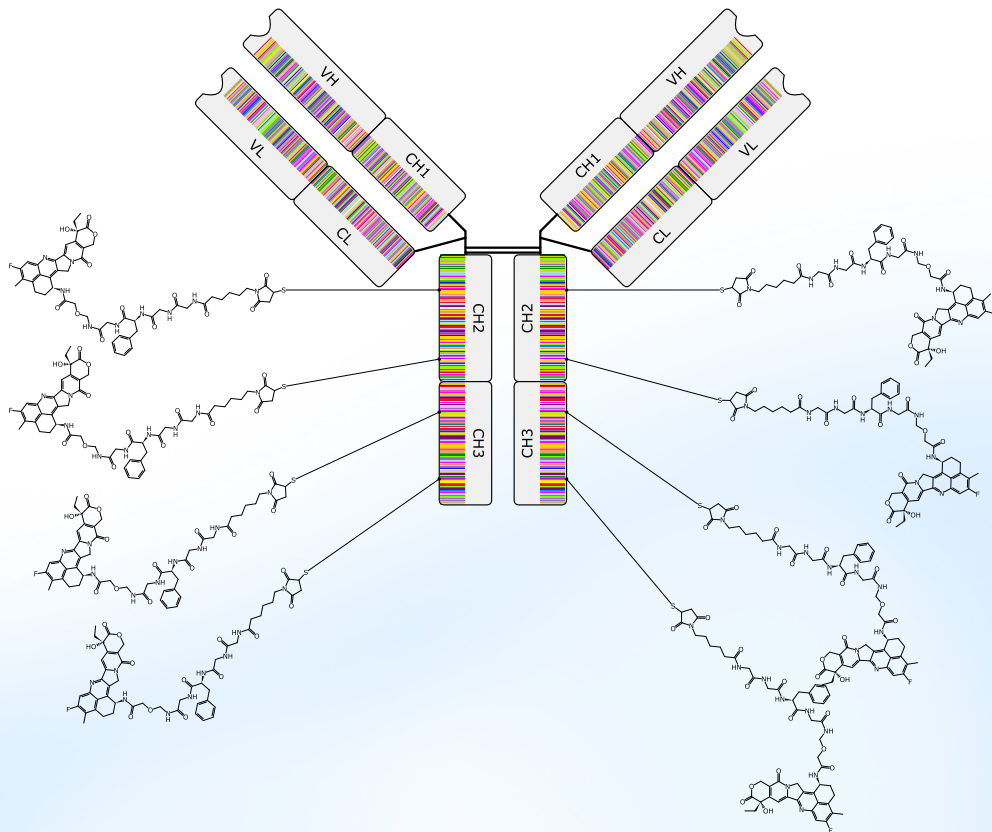
Drug conjugates

- Anchor point is indicated



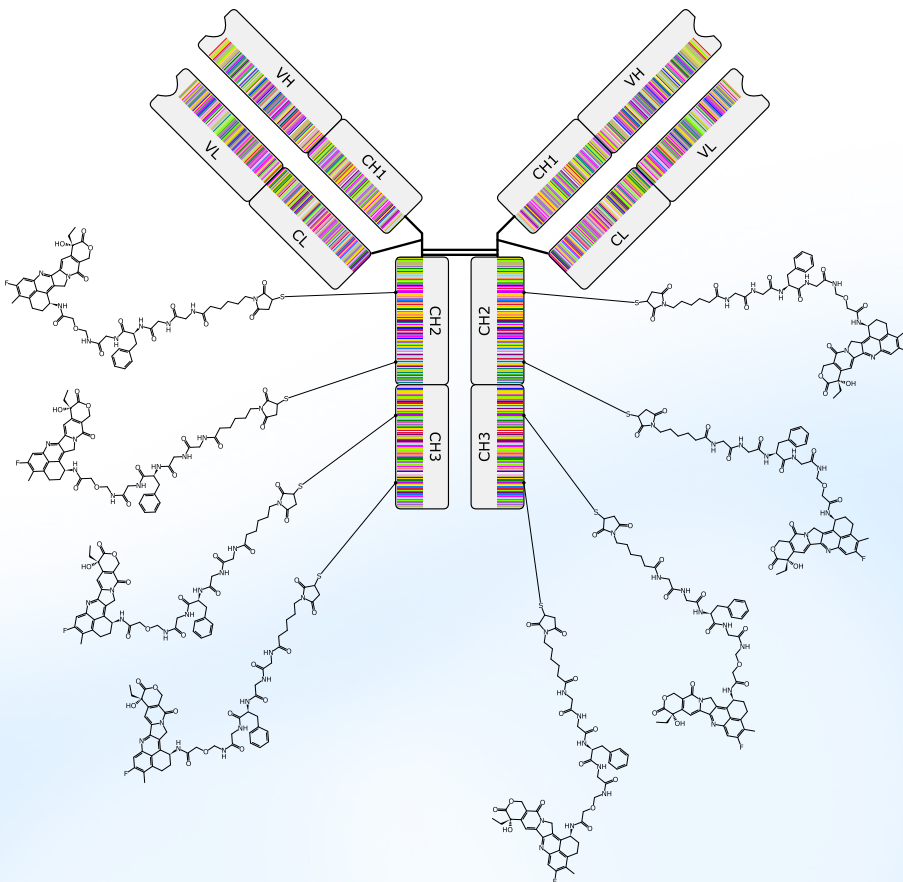
Drug conjugates

- Anchor point is indicated
- Tethered initial placements



Drug conjugates

- Anchor point is indicated
- Tethered initial placements
- Rigid body position optimization
- Underlying chemistry is fully preserved



- Molfiles can represent a **lot of chemistry and biology**
- Sometimes need to **encapsulate** molfiles
- Everything can be reduced to the **correct set** of atoms & bonds with sensible valences
- Preserving **layout**: helps for perception
- Self-contained **context**: no global dictionaries
- Sometimes need **conventions** (e.g. antibodies)
- A lot of work to be done with **using** these capabilities