datagrok

From Sequences to Molecules: An open-source Monomer-Centric Toolkit

Davit Rizhinashvili, Datagrok

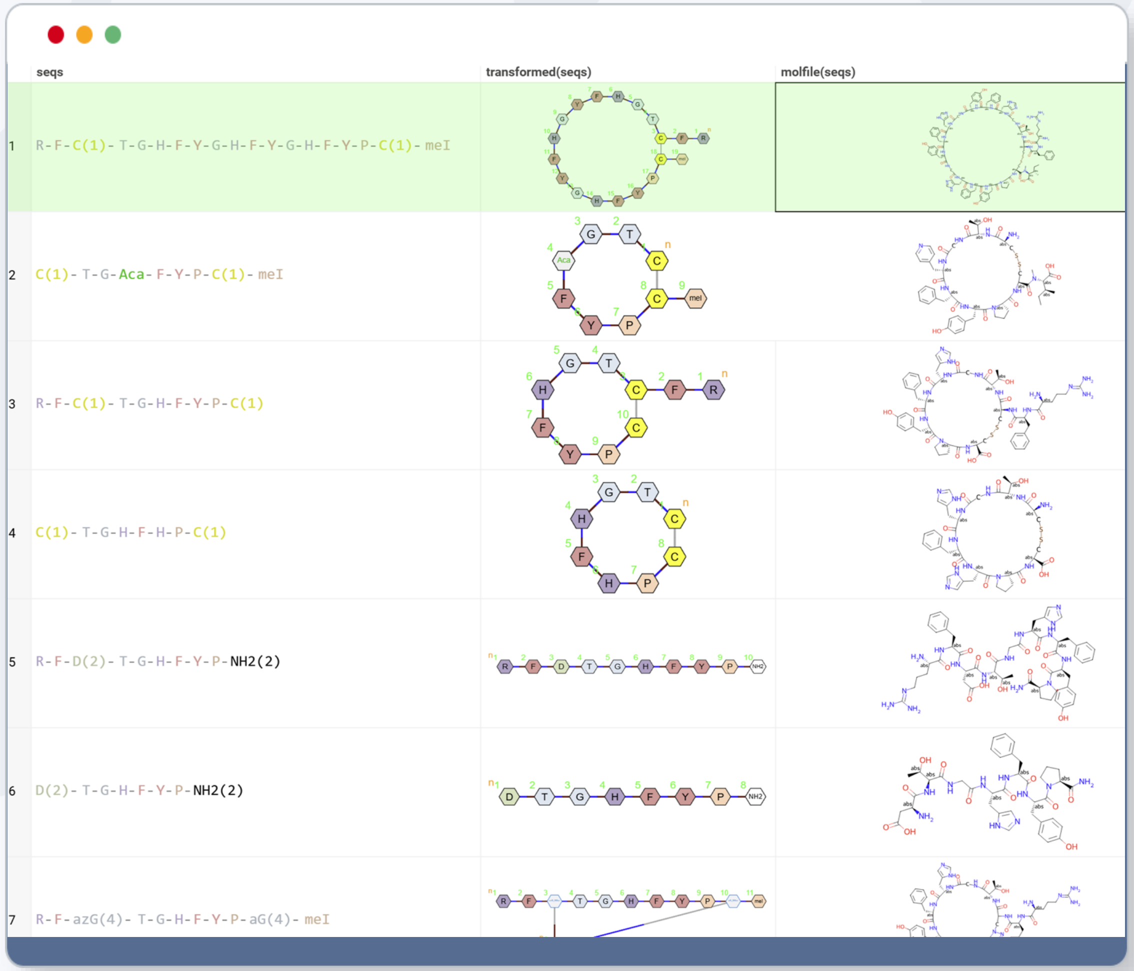
RDKit UGM 2025, Prague

September 11, 2025

Bridging sequences and molecules - Why this is valuable

- Sequence representations are more human readable
- Modular building block approach
- Rapid sequence modification
- Efficient database searches
- Easier computational analysis, access to sequence based tools (MSA, composition analysis, clustering, etc.)
- Error prevention
- Interactive UI

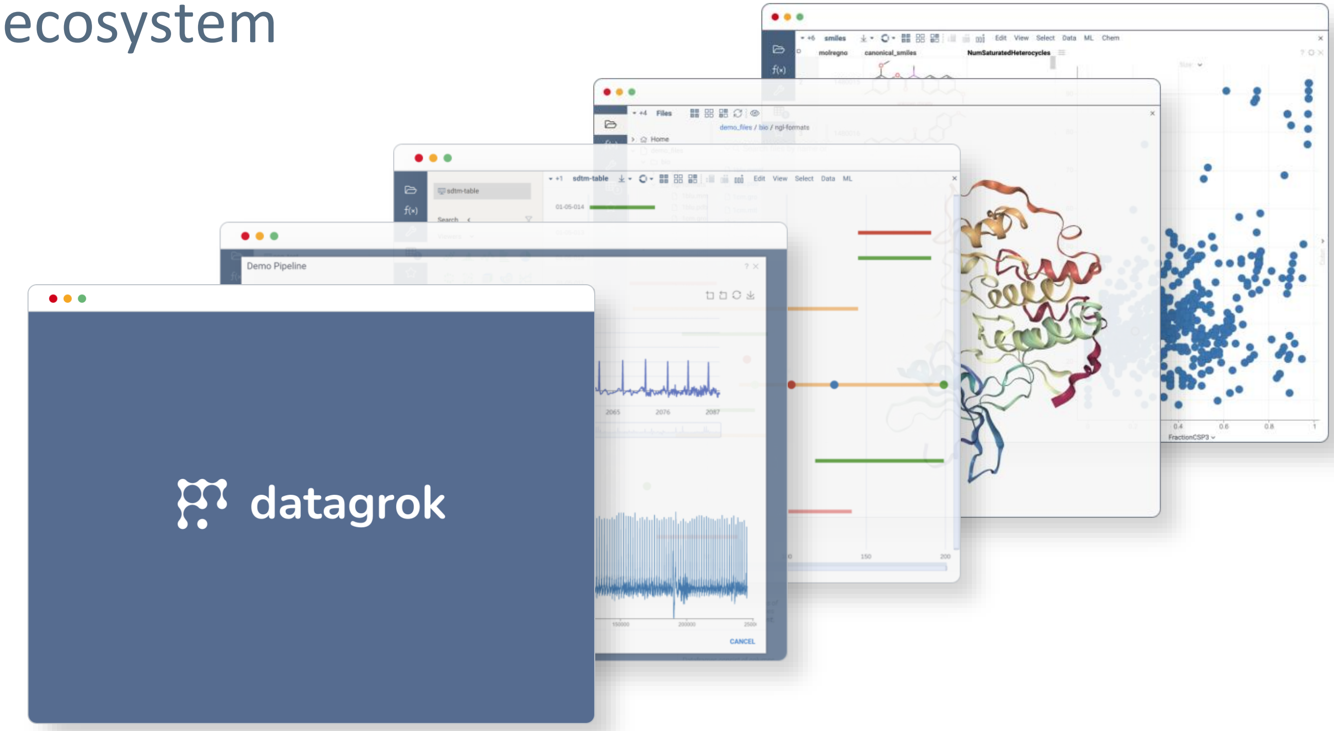
While also being able to calculate properties and conduct analysis on molecular level



	seqs	transformed(seqs)	molfile(seqs)
1	R-F-C(1)-T-G-H-F-Y-G-H-F-Y-G-H-F-Y-P-C(1)-meI		
2	C(1)-T-G-Aca-F-Y-P-C(1)-meI		
3	R-F-C(1)-T-G-H-F-Y-P-C(1)		
4	C(1)-T-G-H-F-H-P-C(1)		
5	R-F-D(2)-T-G-H-F-Y-P-NH2(2)		
6	D(2)-T-G-H-F-Y-P-NH2(2)		
7	R-F-azG(4)-T-G-H-F-Y-P-aG(4)-meI		

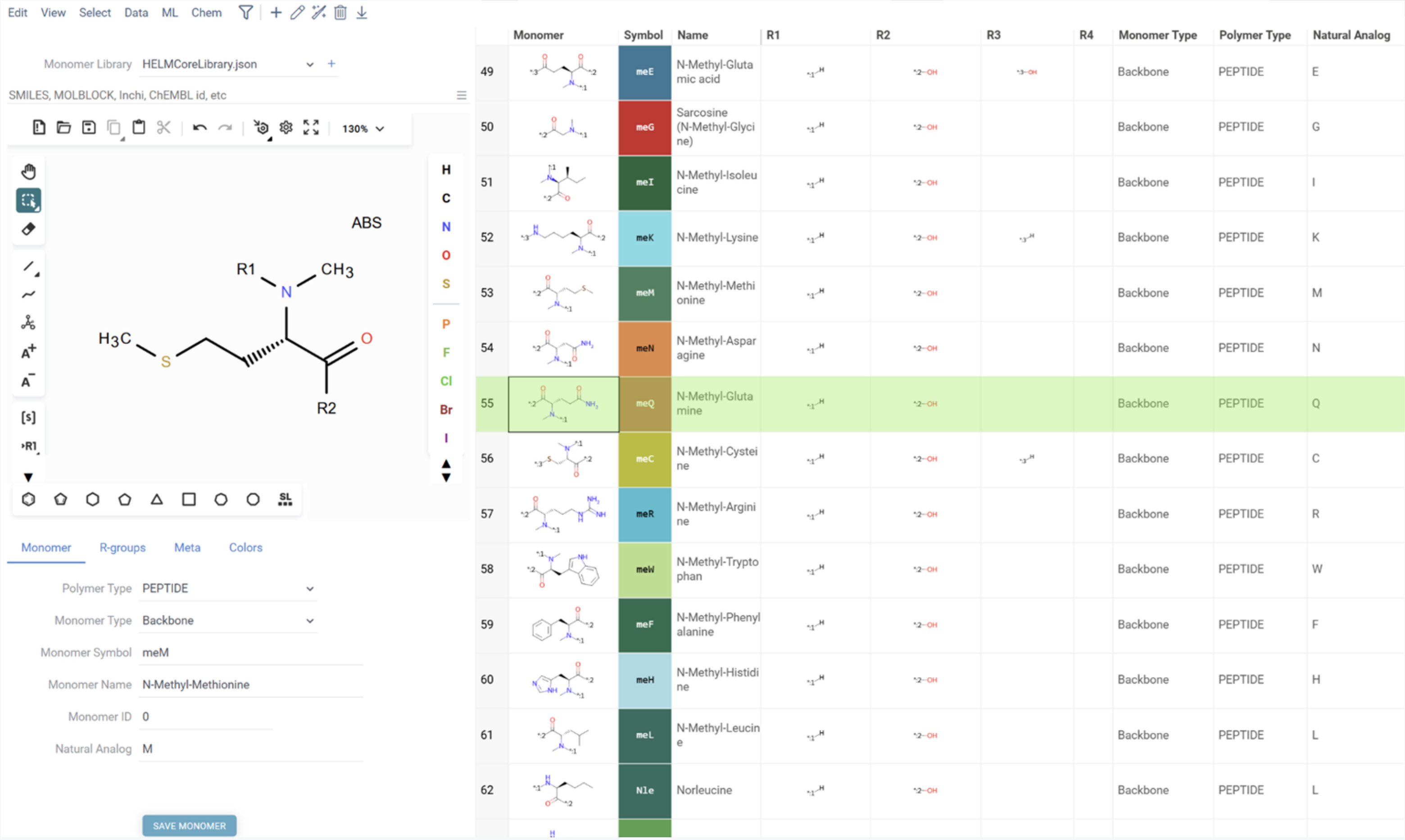
Datagrok: enterprise-ready life sciences platform

- Data access, exploratory data analysis, scientific computing, etc
- Analyzing big datasets completely in the browser
- Proprietary core, open-source plugin ecosystem
- Industry adoption
- Domain-agnostic
- Cheminformatics as a plugin
- RDKit At its cheminformatics core
- Free for personal or academic use



Monomer libraries

- Manage multiple libraries
- Add, edit, delete monomers
- Selectively extract
- Manage duplicates
- Standardization and validation



Monomer Library HELMCoreLibrary.json

SMILES, MOLBLOCK, Inchi, ChEMBL id, etc

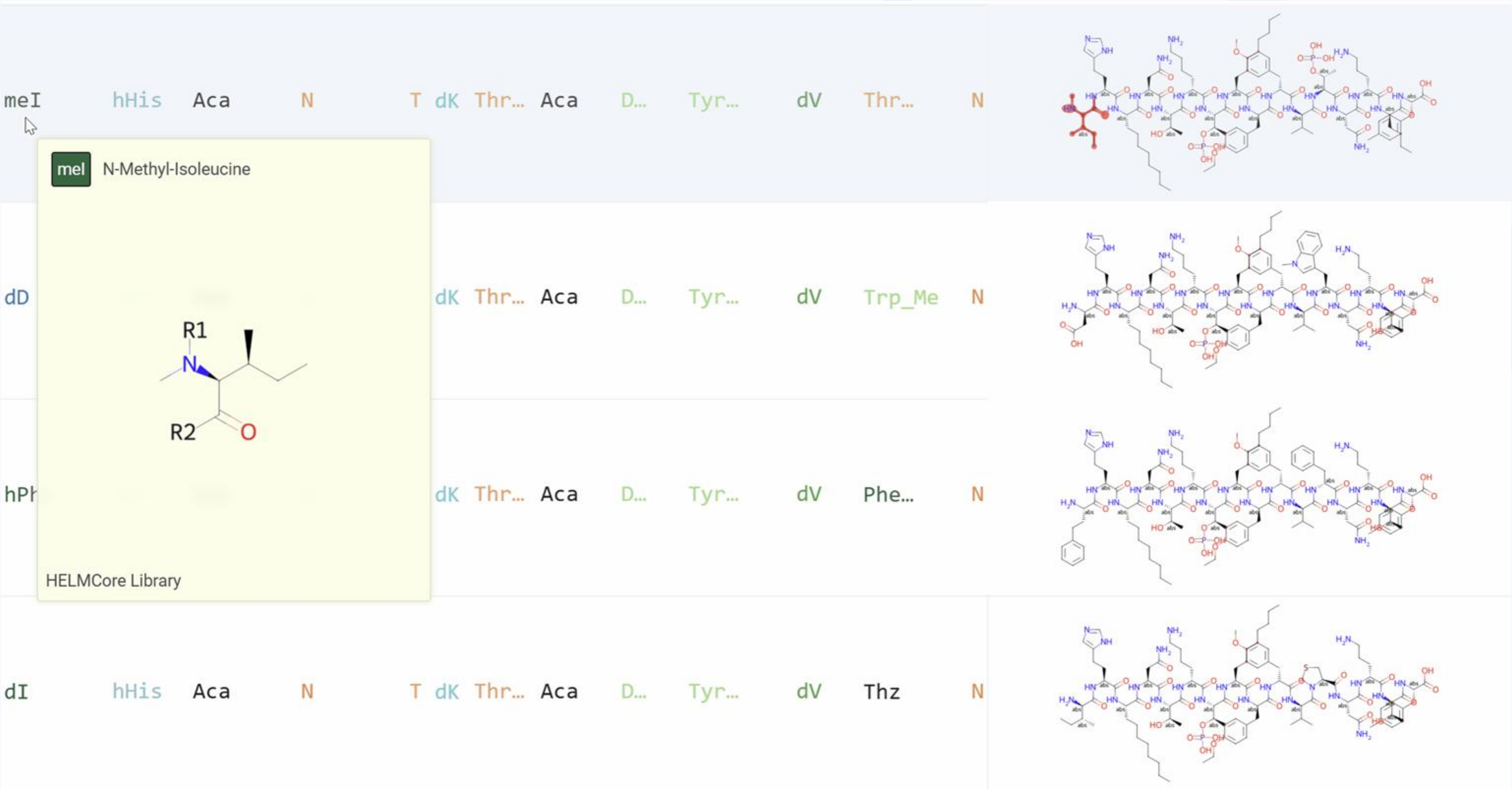
Polymer Type	PEPTIDE
Monomer Type	Backbone
Monomer Symbol	meM
Monomer Name	N-Methyl-Methionine
Monomer ID	0
Natural Analog	M

	Monomer	Symbol	Name	R1	R2	R3	R4	Monomer Type	Polymer Type	Natural Analog
49		meE	N-Methyl-Glutamic acid	*1-H	*2-OH	*3-OH		Backbone	PEPTIDE	E
50		meG	Sarcosine (N-Methyl-Glycine)	*1-H	*2-OH			Backbone	PEPTIDE	G
51		meI	N-Methyl-Isoleucine	*1-H	*2-OH			Backbone	PEPTIDE	I
52		meK	N-Methyl-Lysine	*1-H	*2-OH	*3-H		Backbone	PEPTIDE	K
53		meM	N-Methyl-Methionine	*1-H	*2-OH			Backbone	PEPTIDE	M
54		meN	N-Methyl-Asparagine	*1-H	*2-OH			Backbone	PEPTIDE	N
55		meQ	N-Methyl-Glutamine	*1-H	*2-OH			Backbone	PEPTIDE	Q
56		meC	N-Methyl-Cysteine	*1-H	*2-OH	*3-H		Backbone	PEPTIDE	C
57		meR	N-Methyl-Arginine	*1-H	*2-OH			Backbone	PEPTIDE	R
58		meW	N-Methyl-Tryptophan	*1-H	*2-OH			Backbone	PEPTIDE	W
59		meF	N-Methyl-Phenylalanine	*1-H	*2-OH			Backbone	PEPTIDE	F
60		meH	N-Methyl-Histidine	*1-H	*2-OH			Backbone	PEPTIDE	H
61		meL	N-Methyl-Leucine	*1-H	*2-OH			Backbone	PEPTIDE	L
62		Nle	Norleucine	*1-H	*2-OH			Backbone	PEPTIDE	L

Monomer Manager

The picture can't be displayed.

Monomers



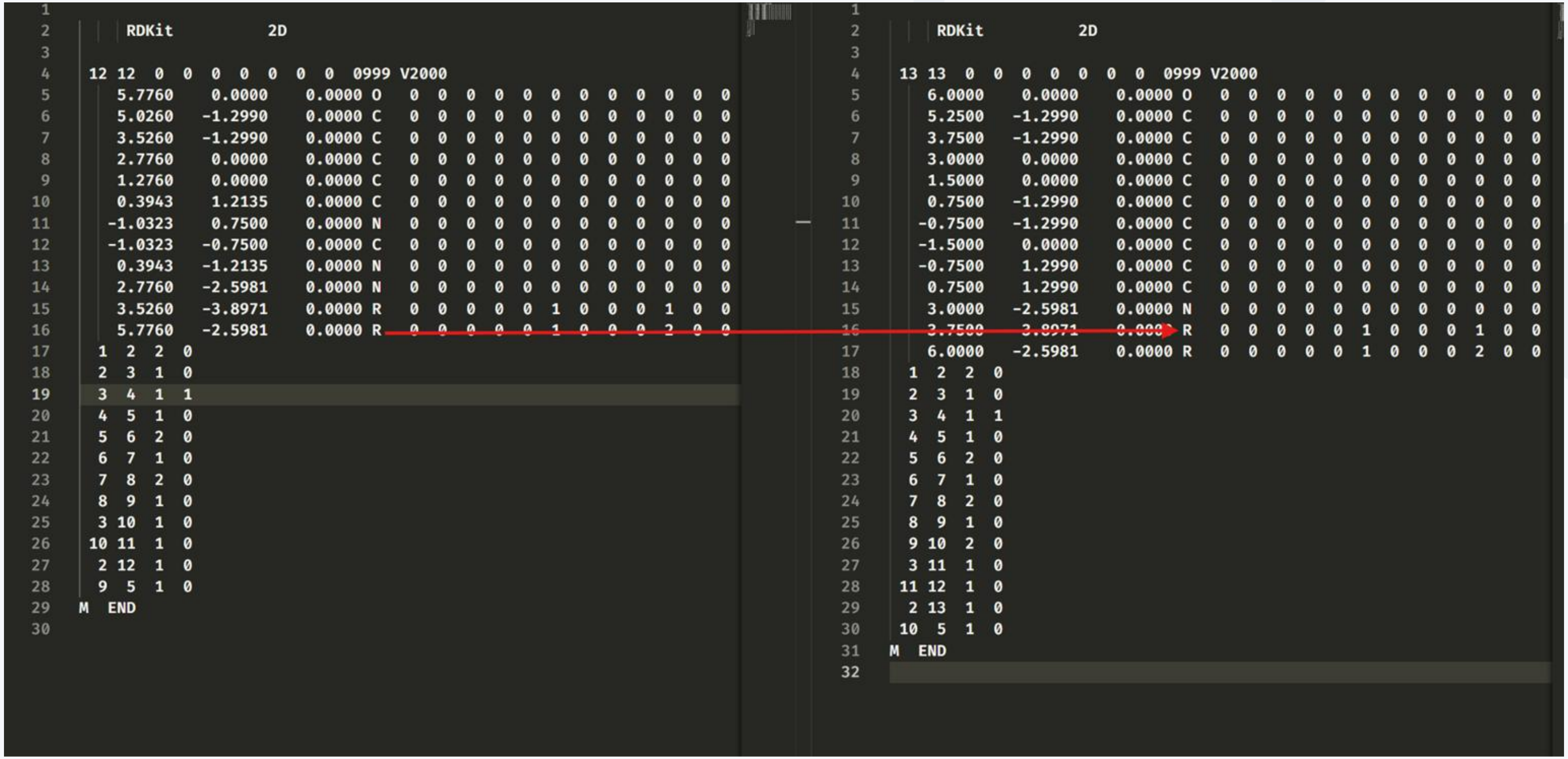
meI	hHis	Aca	N	T	dK	Thr…	Aca	D…	Tyr…	dV	Thr…	N
dD					dK	Thr…	Aca	D…	Tyr…	dV	Trp_Me	N
hPh					dK	Thr…	Aca	D…	Tyr…	dV	Phe…	N
dI	hHis	Aca	N	T	dK	Thr…	Aca	D…	Tyr…	dV	Thz	N

mel N-Methyl-Isoleucine

HELMCore Library

Linear Sequences

- R2 of First monomer connects to R1of second monomer

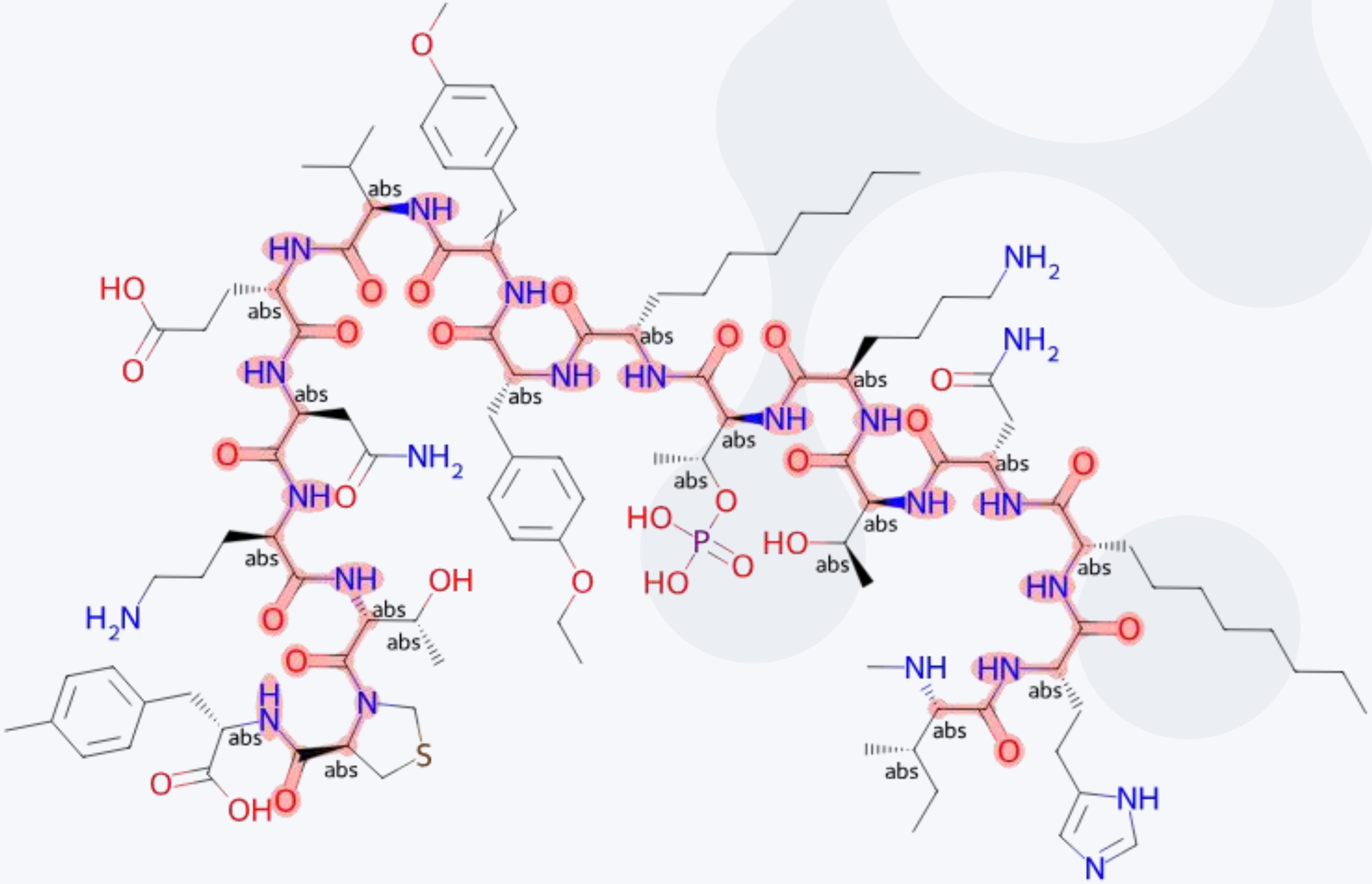


```

     RDKit          2D

 12 12  0  0  0  0  0  0  0  0999 V2000
    5.7760    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.0260   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.5260   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.7760    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2760    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3943    1.2135    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0323    0.7500    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0323   -0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.3943   -1.2135    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.7760   -2.5981    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.5260   -3.8971    0.0000 R   0  0  0  0  0  1  0  0  0  1  0  0
    5.7760   -2.5981    0.0000 R   0  0  0  0  0  1  0  0  0  2  0  0
  1  2  2  0
  2  3  1  0
  3  4  1  1
  4  5  1  0
  5  6  2  0
  6  7  1  0
  7  8  2  0
  8  9  1  0
  3 10  1  0
 10 11  1  0
  2 12  1  0
  9  5  1  0
M  END

```

```

     RDKit          2D

 13 13  0  0  0  0  0  0  0  0999 V2000
    6.0000    0.0000    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    5.2500   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7500   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7500   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7500   -1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.5000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7500    1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7500    1.2990    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0000   -2.5981    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.7500   -3.8971    0.0000 R   0  0  0  0  0  1  0  0  0  1  0  0
    6.0000   -2.5981    0.0000 R   0  0  0  0  0  1  0  0  0  2  0  0
  1  2  2  0
  2  3  1  0
  3  4  1  1
  4  5  1  0
  5  6  2  0
  6  7  1  0
  7  8  2  0
  8  9  1  0
  9 10  2  0
  3 11  1  0
 11 12  1  0
  2 13  1  0
 10  5  1  0
M  END

```

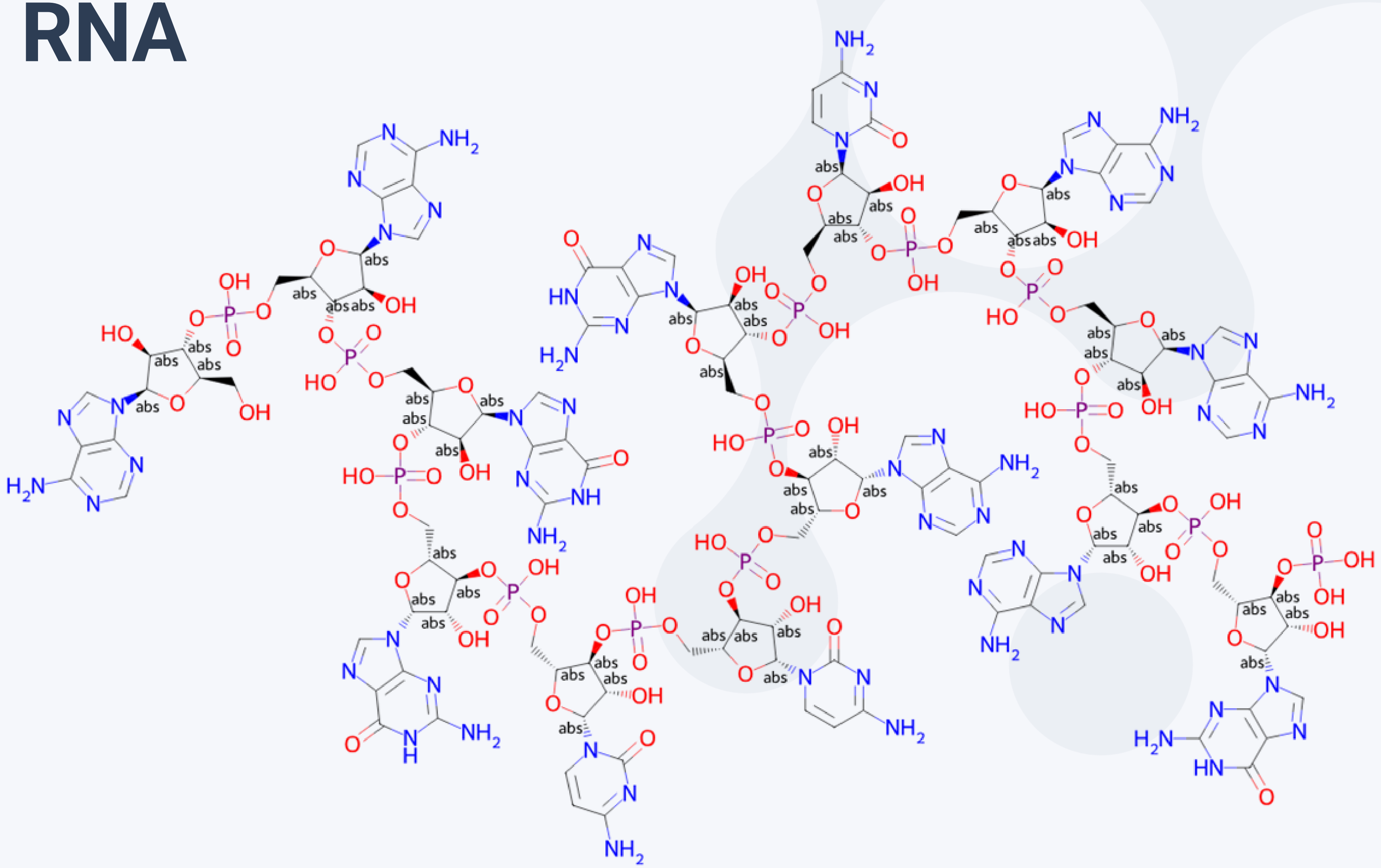
Linear Sequences

Peptides



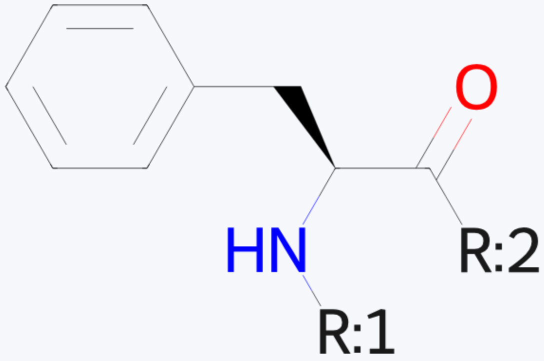
Linear Sequences

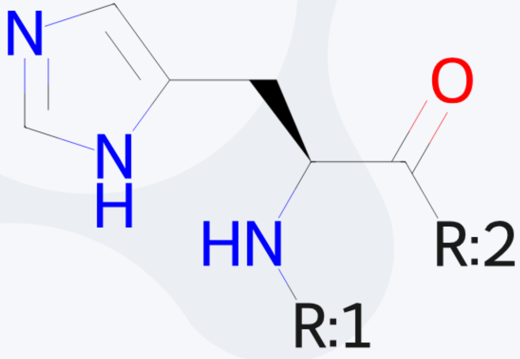
DNA / RNA

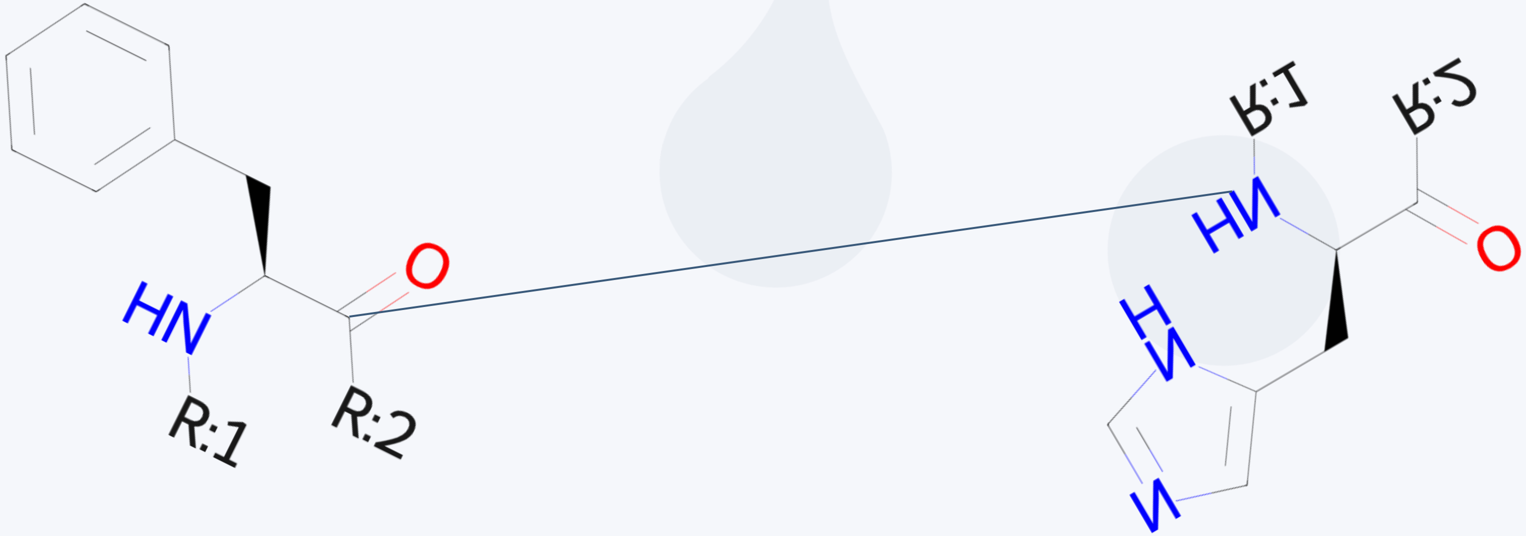


Linear Sequences

More human-friendly depiction

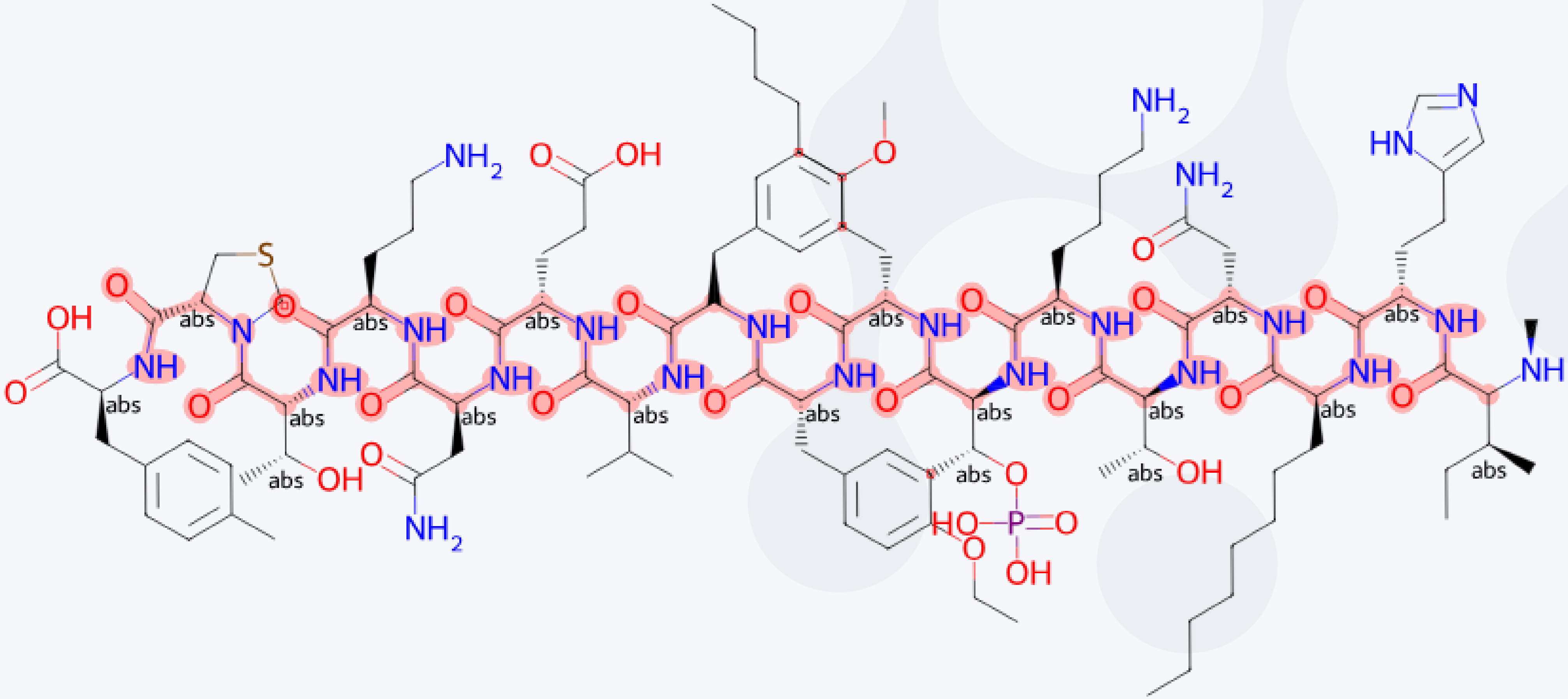






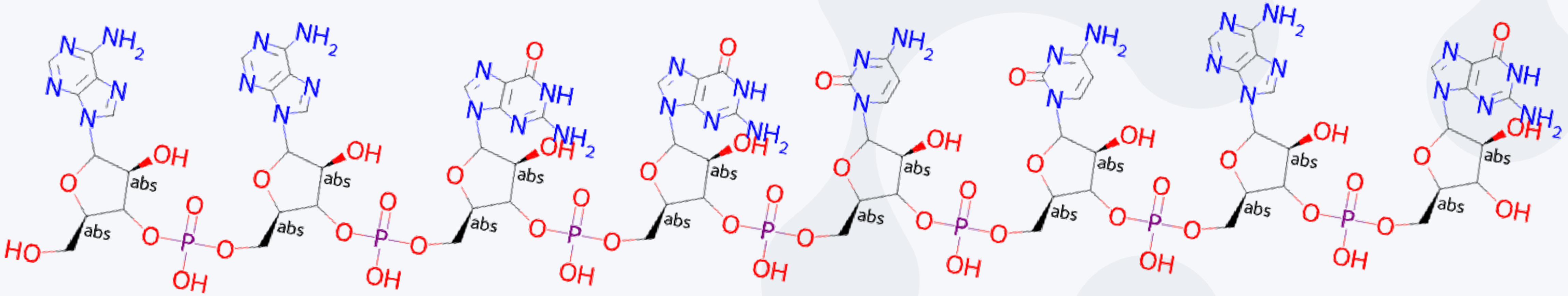
Linear Sequences

Peptides

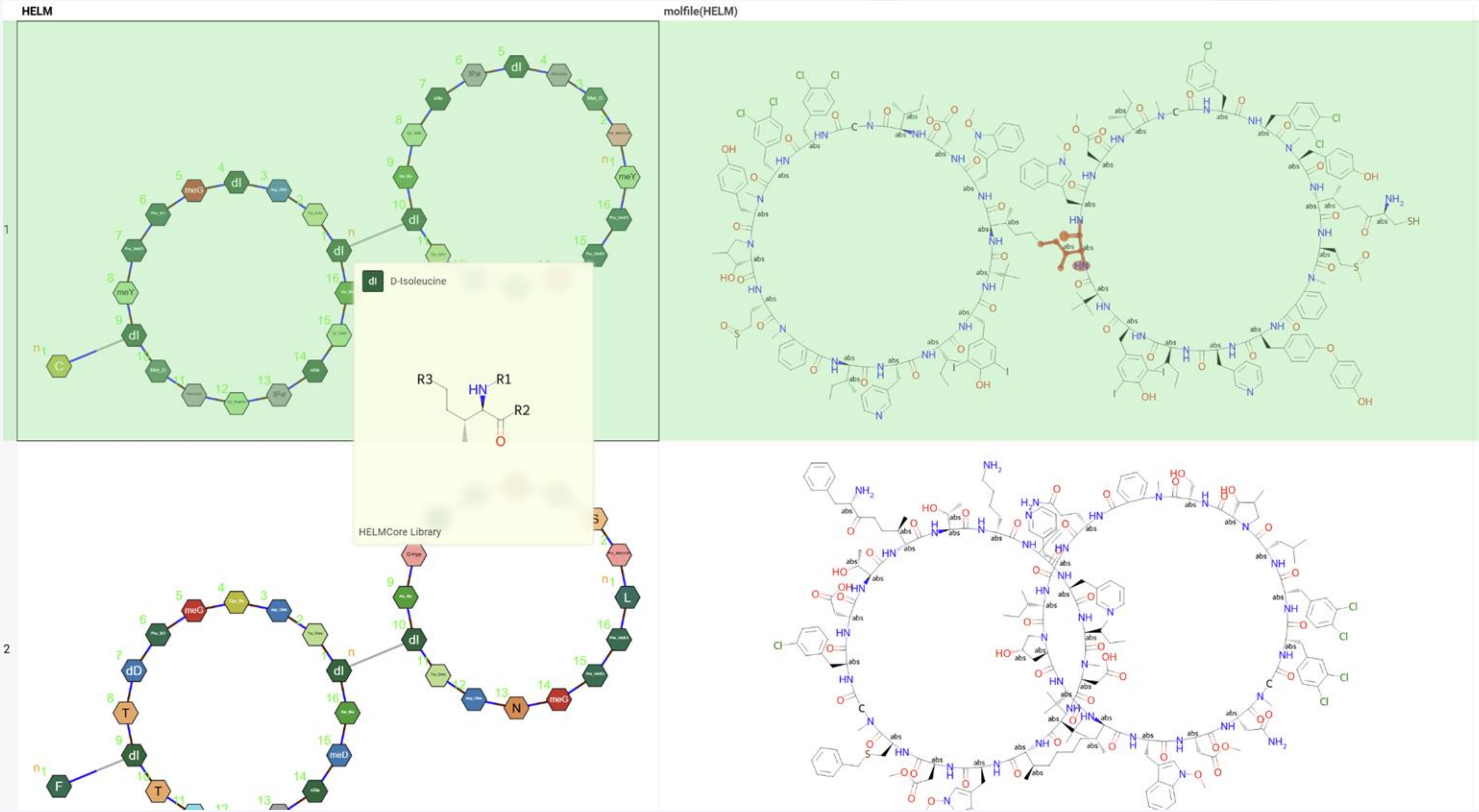


Linear Sequences

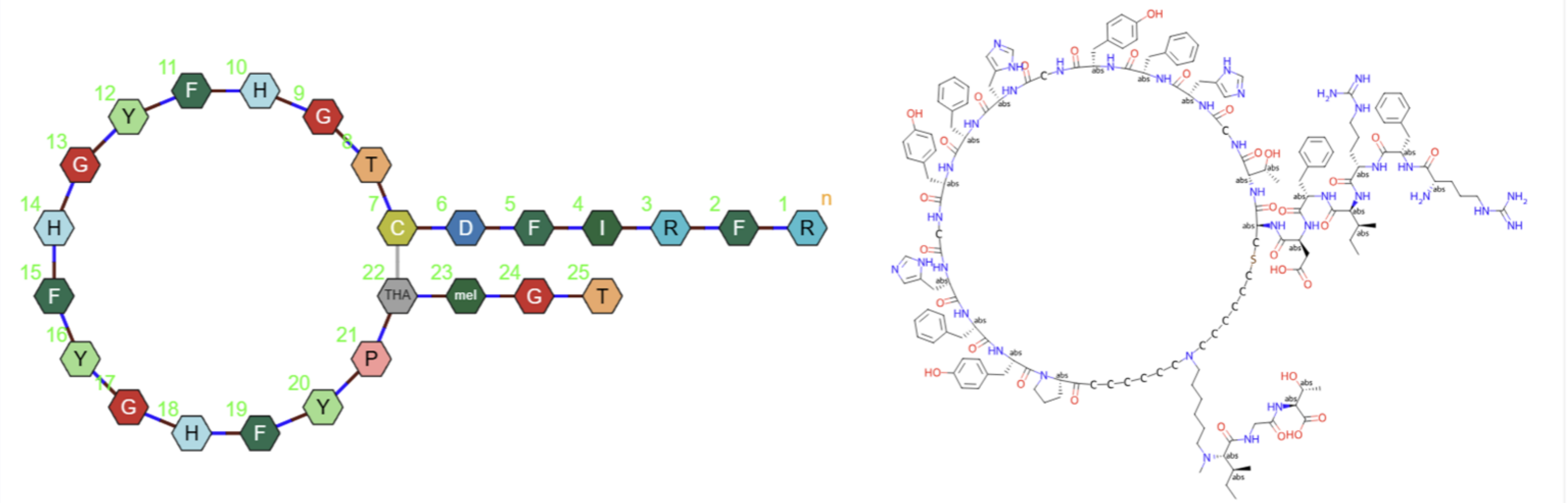
DNA / RNA



Cyclic structures: HELM

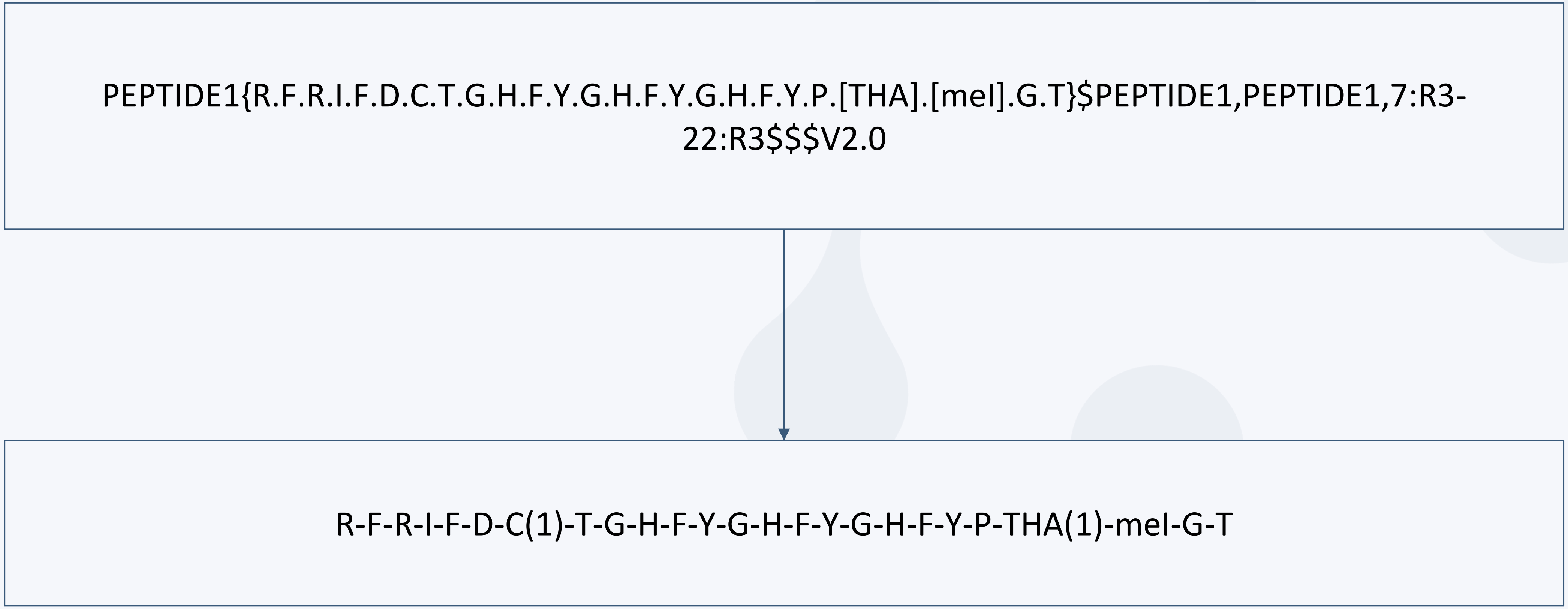


HELM



PEPTIDE1{R.F.R.I.F.D.C.T.G.H.F.Y.G.H.F.Y.G.H.F.Y.P.[THA].[meI].G.T}$PEPTIDE1,PEPTIDE1,7:R3-22:R3$$$V2.0

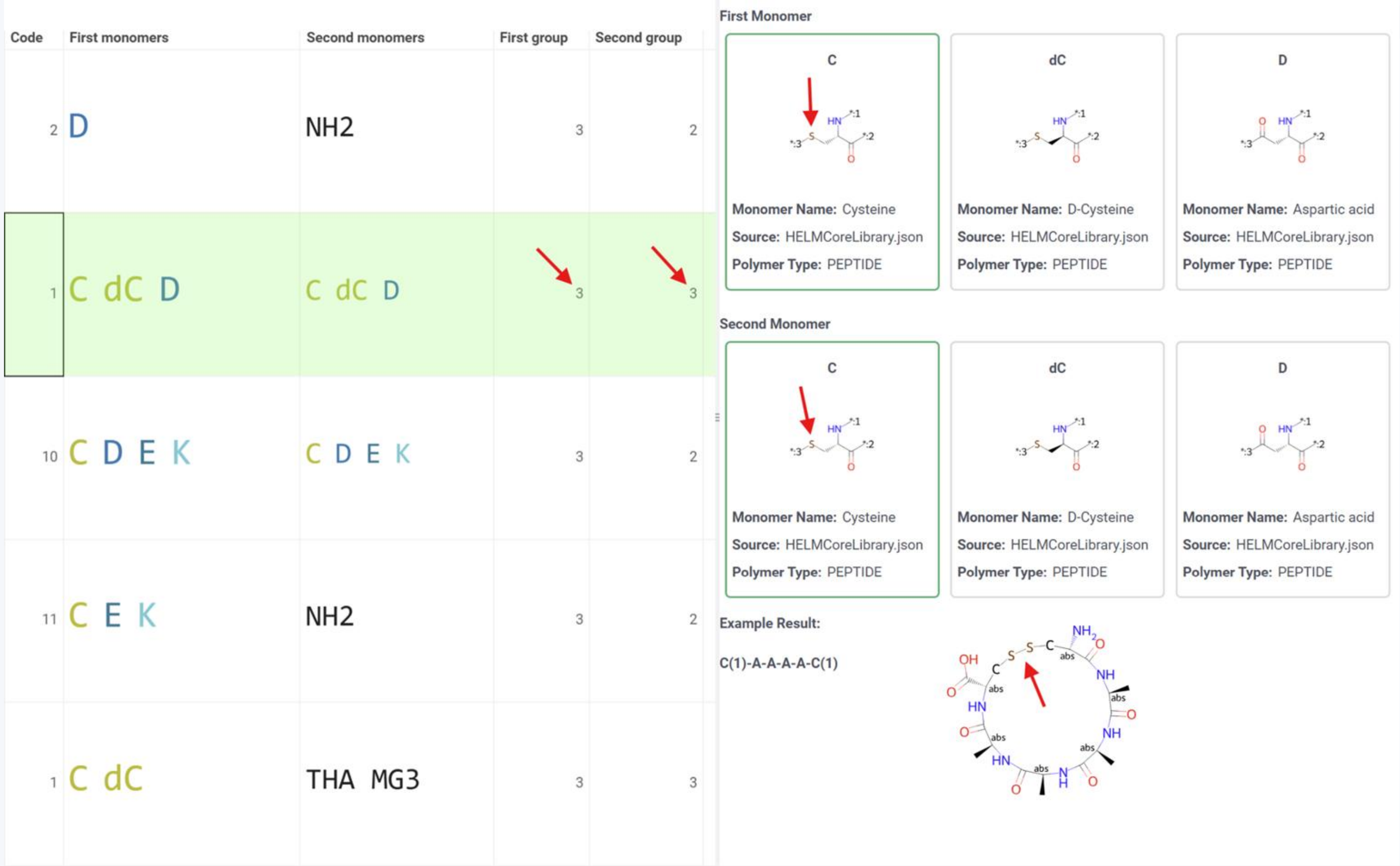
Rules - make sequences readable again



PEPTIDE1{R.F.R.I.F.D.C.T.G.H.F.Y.G.H.F.Y.G.H.F.Y.P.[THA].[meI].G.T}$PEPTIDE1,PEPTIDE1,7:R3-22:R3$$$V2.0

R-F-R-I-F-D-C(1)-T-G-H-F-Y-G-H-F-Y-G-H-F-Y-P-THA(1)-meI-G-T

Link Rules



Code	First monomers	Second monomers	First group	Second group
2	D	NH2	3	2
1	C dC D	C dC D	3	3
10	C D E K	C D E K	3	2
11	C E K	NH2	3	2
1	C dC	THA MG3	3	3

First Monomer

C

Monomer Name: Cysteine
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

dC

Monomer Name: D-Cysteine
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

D

Monomer Name: Aspartic acid
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

Second Monomer

C

Monomer Name: Cysteine
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

dC

Monomer Name: D-Cysteine
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

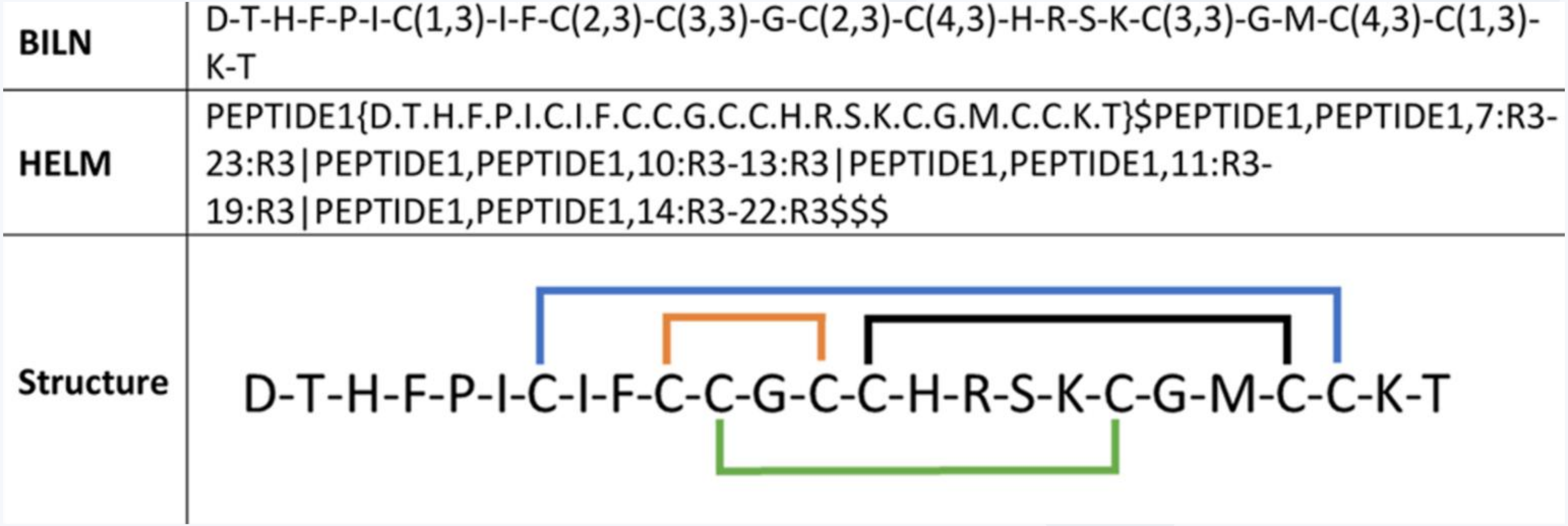
D

Monomer Name: Aspartic acid
Source: HELMCoreLibrary.json
Polymer Type: PEPTIDE

Example Result:

C(1)-A-A-A-A-C(1)

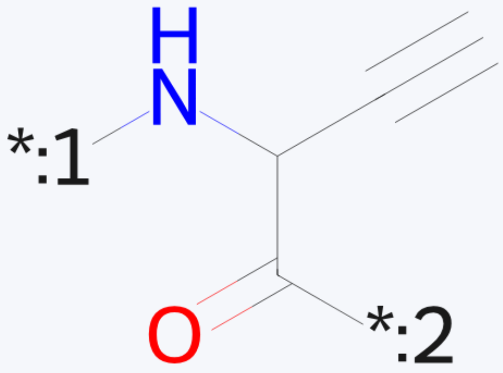
BILN - How is it different?

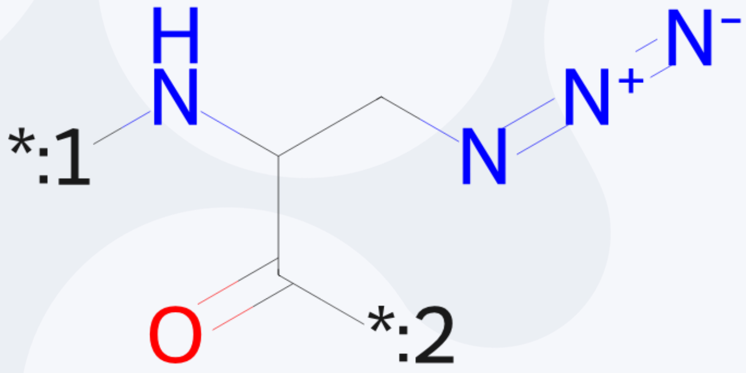


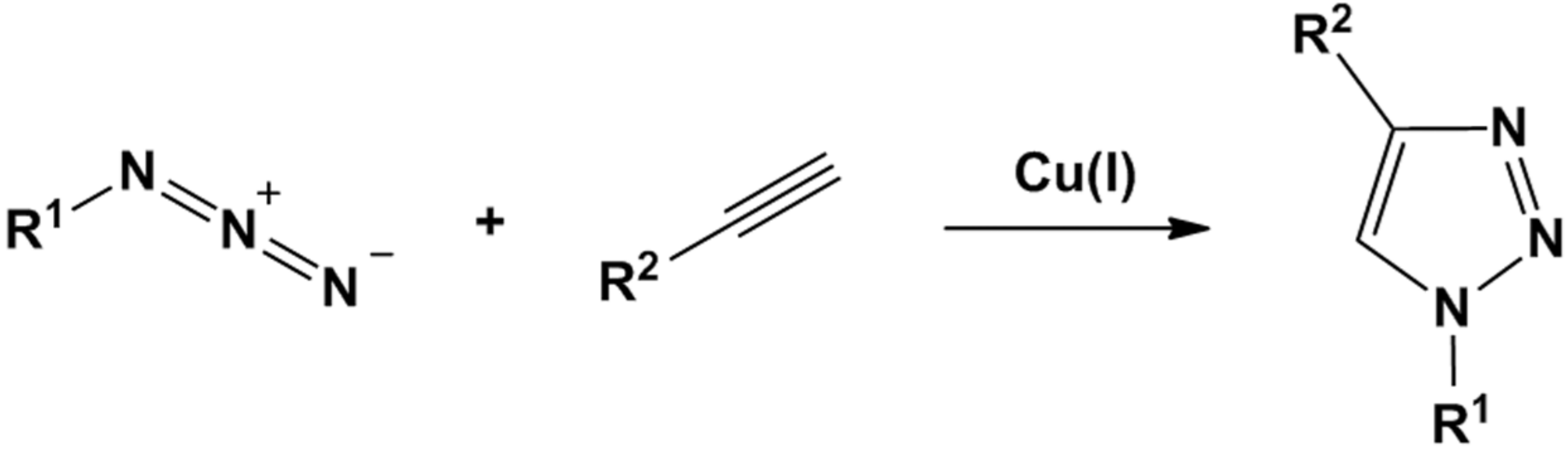
BILN	D-T-H-F-P-I-C(1,3)-I-F-C(2,3)-C(3,3)-G-C(2,3)-C(4,3)-H-R-S-K-C(3,3)-G-M-C(4,3)-C(1,3)-K-T
HELM	PEPTIDE1{D.T.H.F.P.I.C.I.F.C.C.G.C.C.H.R.S.K.C.G.M.C.C.K.T}$PEPTIDE1,PEPTIDE1,7:R3-23:R3|PEPTIDE1,PEPTIDE1,10:R3-13:R3|PEPTIDE1,PEPTIDE1,11:R3-19:R3|PEPTIDE1,PEPTIDE1,14:R3-22:R3$$$
Structure	D-T-H-F-P-I-C-I-F-C-C-G-C-C-H-R-S-K-C-G-M-C-C-K-T

Fox, Thomas, et al. "BILN: a human-readable line notation for complex peptides." Journal of Chemical Information and Modeling 62.17 (2022): 3942-3947.

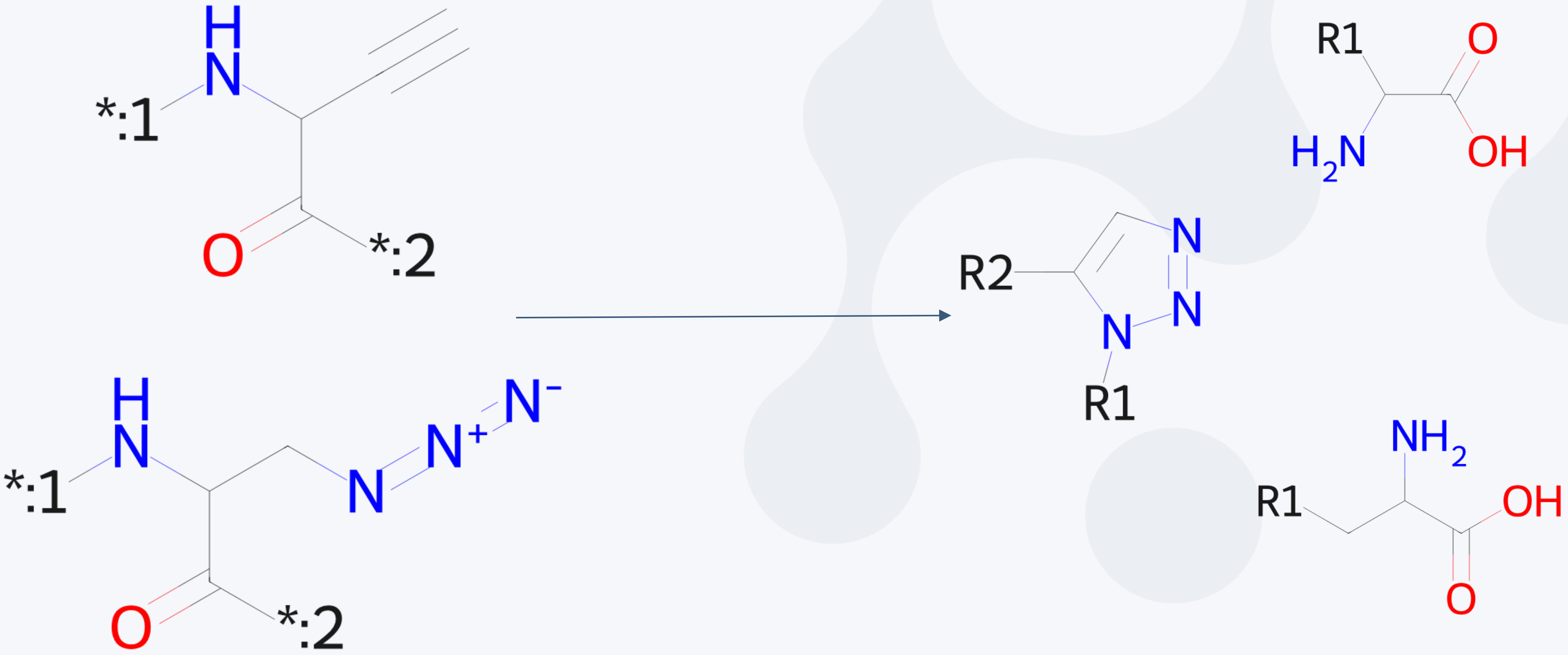
Reaction Rules



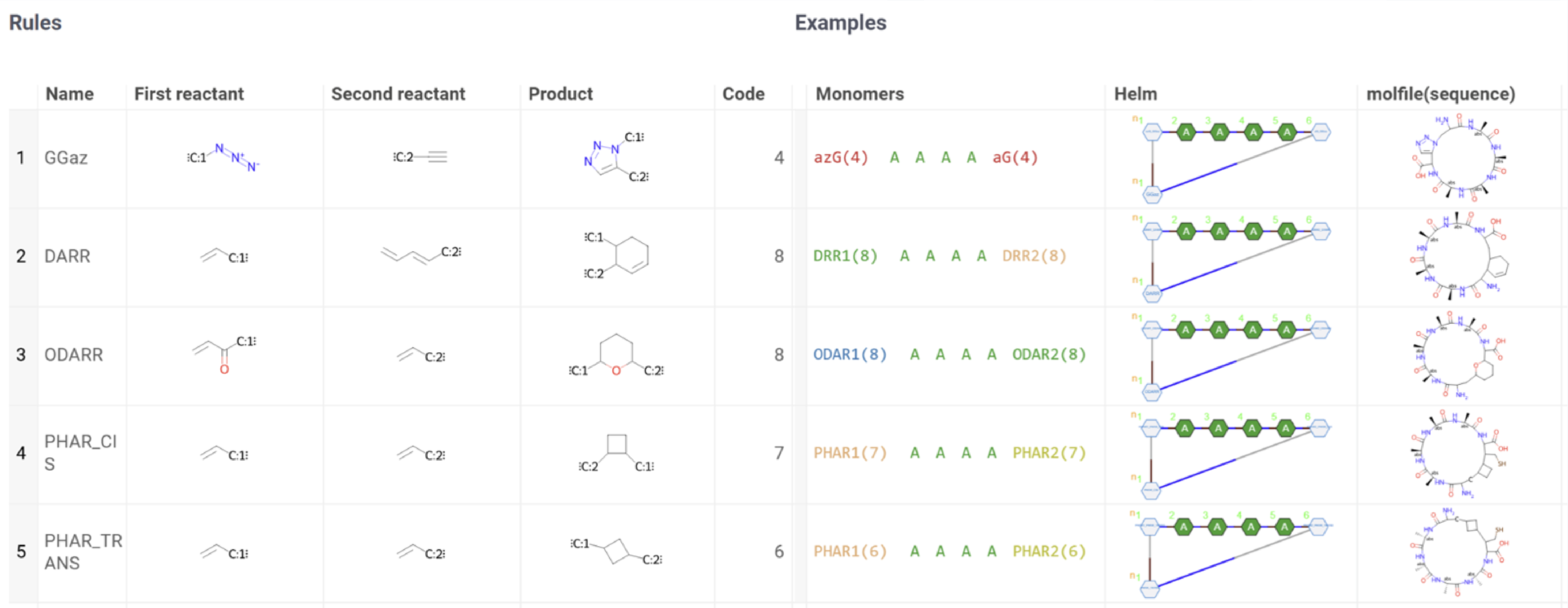




Reaction Rules



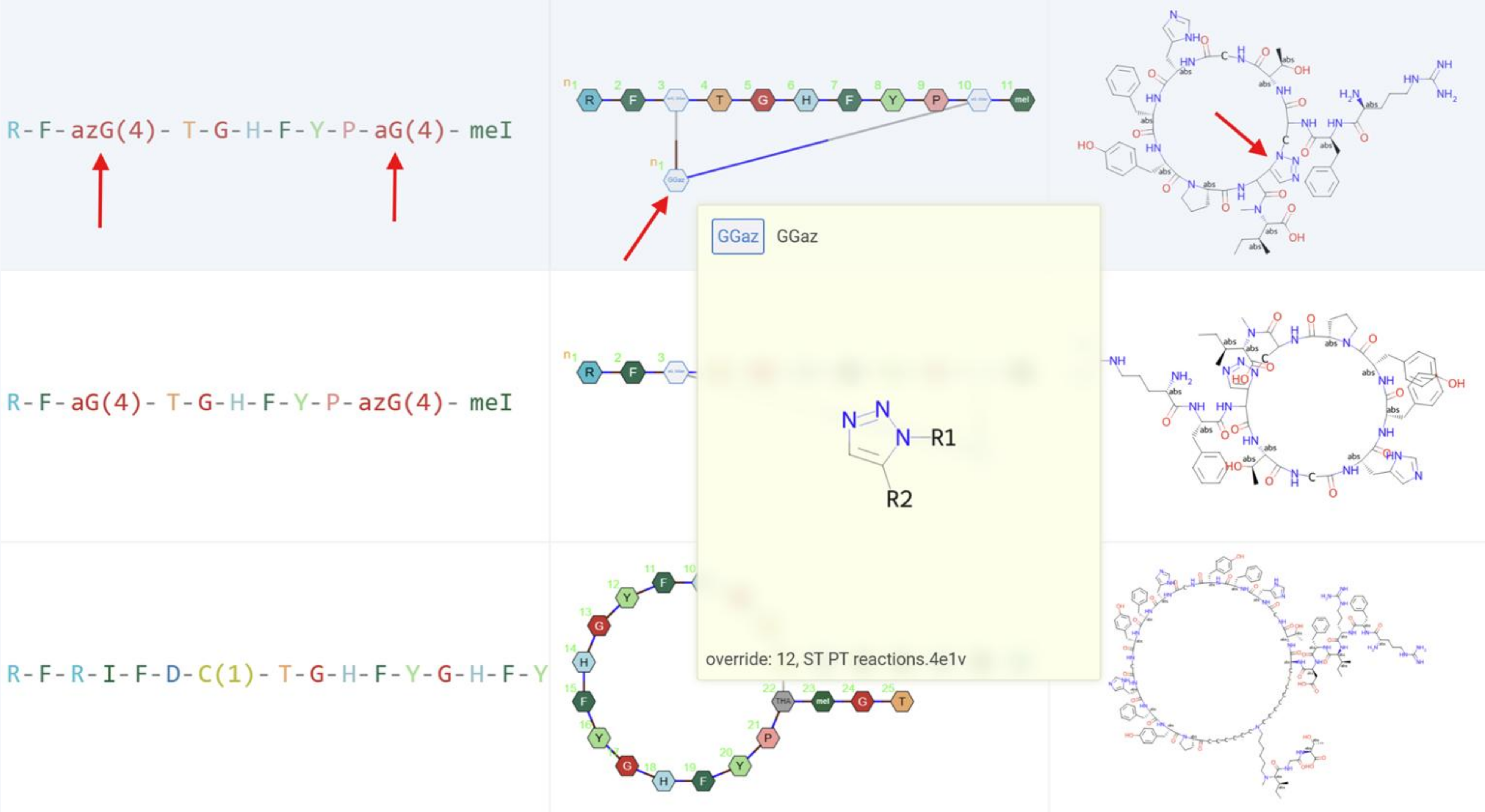
Reaction Rules



	Rules					Examples		
	Name	First reactant	Second reactant	Product	Code	Monomers	Helm	molfile(sequence)
1	GGaz				4	azG(4) A A A A aG(4)		
2	DARR				8	DRR1(8) A A A A DRR2(8)		
3	ODARR				8	ODAR1(8) A A A A ODAR2(8)		
4	PHAR_CIS				7	PHAR1(7) A A A A PHAR2(7)		
5	PHAR_TRANS				6	PHAR1(6) A A A A PHAR2(6)		

Reaction Rules

Automatically created reaction product monomers



R-F-azG(4)-T-G-H-F-Y-P-aG(4)-meI

R-F-aG(4)-T-G-H-F-Y-P-azG(4)-meI

R-F-R-I-F-D-C(1)-T-G-H-F-Y-G-H-F-Y

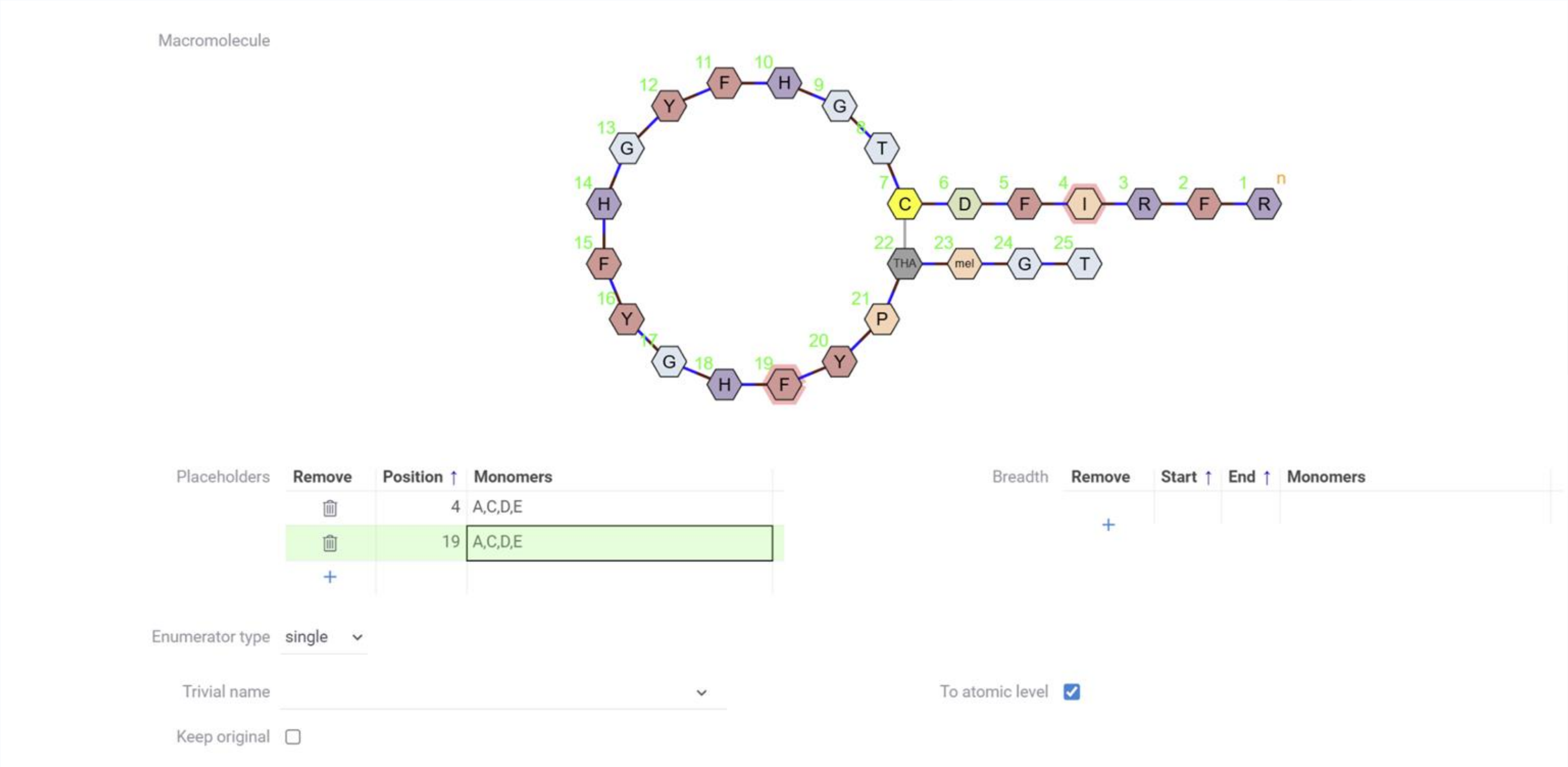
GGaz GGaz

override: 12, ST PT reactions.4e1v

Conversion with rules

The picture can't be displayed.

Enumerator



Macromolecule

Placeholders

Remove	Position ↑	Monomers
🗑	4	A,C,D,E
🗑	19	A,C,D,E
+		

Breadth

Remove	Start ↑	End ↑	Monomers
+			

Enumerator type: single

Trivial name:

To atomic level ☑

Keep original □

Enumerator

The picture can't be displayed.

Thank You!

Acknowledgements

- RDKit maintainers and community
- Datagrok team
- Rhitankar Pal
- All our users ☺

Run the platform right now in your browser: https://public.datagrok.ai