RDKit UGM 2025: Biopolymers, bond orders, stereochemistry, bioactivity, text mining and more.

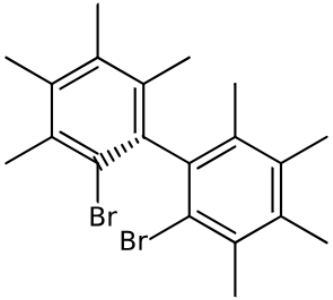
Roger Sayle, Ph.D.
NextMove Software,
Cambridge, UK

14th RDKit UGM, Prague, Czech Republic, 11th September 2025

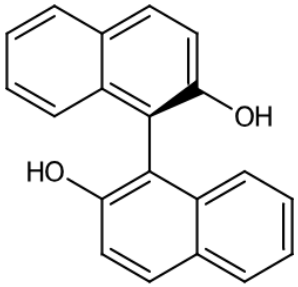
a little about the author…

- Many, many years ago I wrote a molecular graphics program, RasMol, at the time the most popular program of its kind (over 1 million users).
- The two main file formats, PDB files and MDL Mol files each had unique properties and challenges.
- Converting between them lost information both ways! Residue information in PDB files, bond orders, formal charges and hydrogen counts in Mol files.
- This, and more generally 3D↔2D, became a life long study (curse) and defines cheminformatics' core.

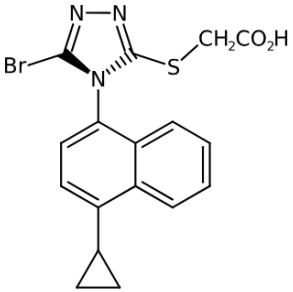
aTROPISOMErs in smiles



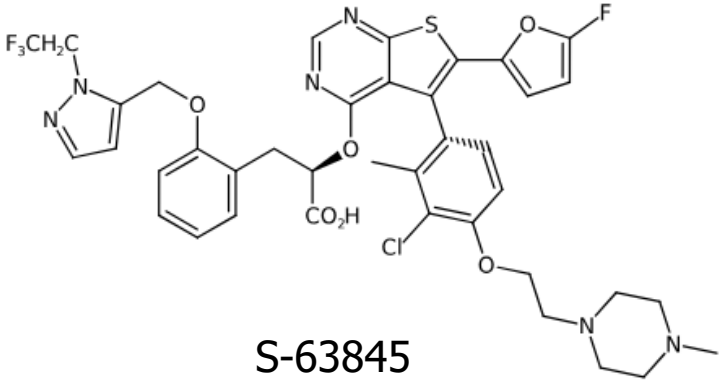
ABEFET



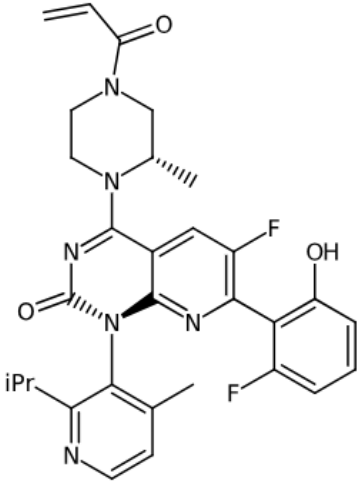
BIRKOC04
(S)-BINOL



TIPQER01
(-)-lesinurad



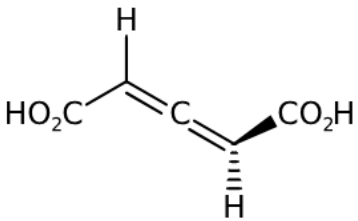
S-63845



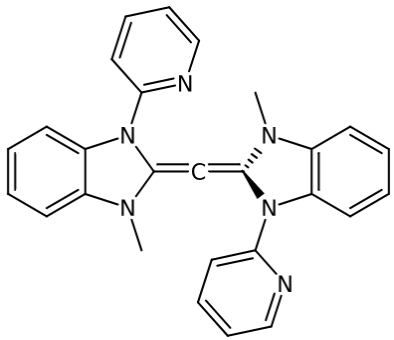
(m)-sotorasib

- Oc1ccc2ccccc2c1-c1c(O)ccc2ccccc12 |wU:10.10| (S)-BINOL
- Tad Hurst's RDKit extension, based upon ChemAxon extension.

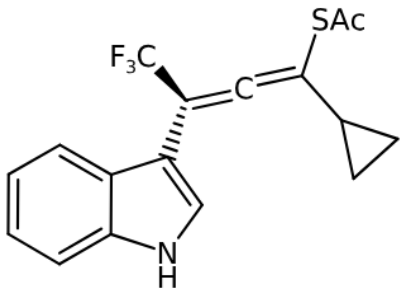
allenic stereochemistry



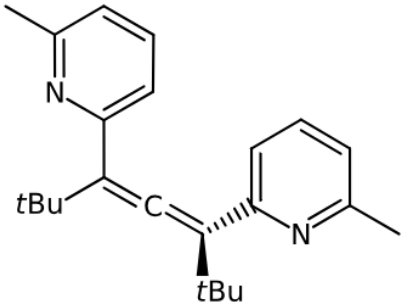
ALLCAM



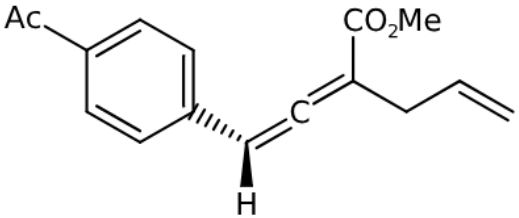
HOHNEZ



ISOZOI



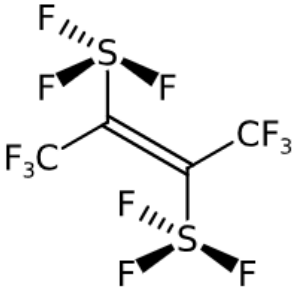
ECUMIZ



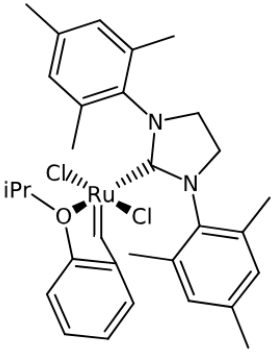
EDUZIQ

- OC(=O)C=[C@AL1]=CC(=O)O.CC(=O)N ALLCAM
- InChI=1S/C5H4O4/c6-4(7)2-1-3-5(8)9/h2-3H,(H,6,7)(H,8,9)/t1-/m0/s1
- Supported by InChI, Daylight, CDK. Ignored by RDKit, OpenBabel.

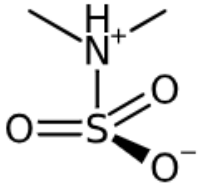
see-saw compounds and friends



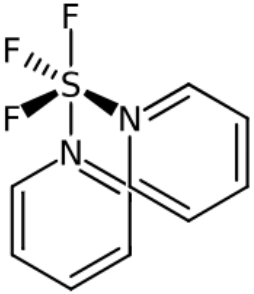
CUVSOB



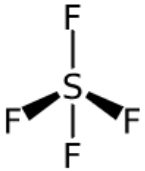
ABEJUM01



BOBGOO



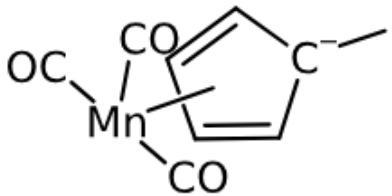
AMOLAS



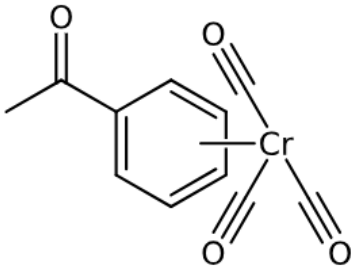
sulfur tetrafluoride
F[S@OH1](F)(F)F

- [S@OH1](F)(F)(F)/C(=C(/[S@OH19](F)(F)F)C(F)(F)F)/C(F)(F)F CUVSOB
- Missing valences considered immediately after parent [like impH].

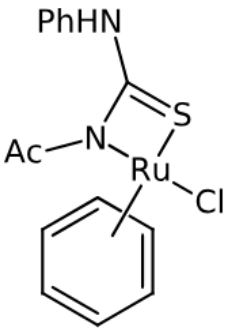
hapticity/piano stool compounds



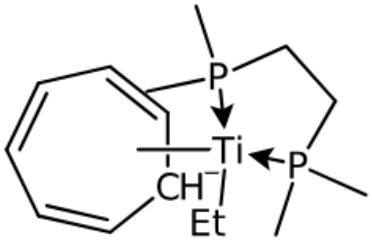
methylcyclopentadienyl manganese tricarbonyl



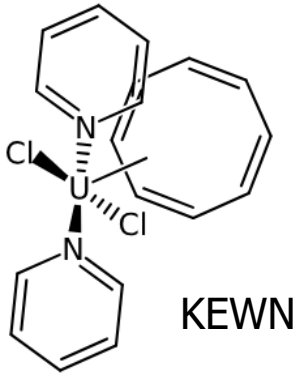
ACTPCR



ADOVOI



BOYZOE



KEWNEF

- *[Mn](C#O)(C#O)(C#O).c1ccc[c-]1C |m:0:8.9.10.11.12|

Ferrocene chirality

Received: 4 June 2022 | Revised: 25 July 2022 | Accepted: 29 July 2022
DOI: 10.1002/elps.202200148
ELECTROPHORESIS

REVIEW

Ferrocene derivatives with planar chirality and their enantioseparation by liquid-phase techniques

Paola Peluso[1] | Victor Mamane[2]

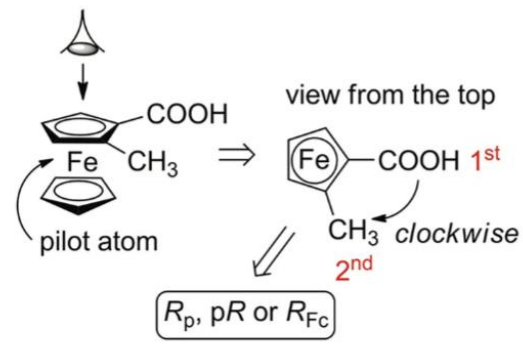
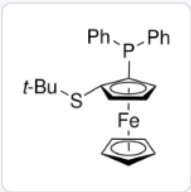
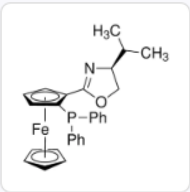


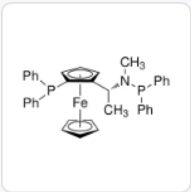
FIGURE 2 Absolute configuration of planar chiral ferrocenes



Sigma-Aldrich
687561
(R_p)-2-(*tert*-Butylthio)-1-(diphenylphosphino)ferrocene
98%



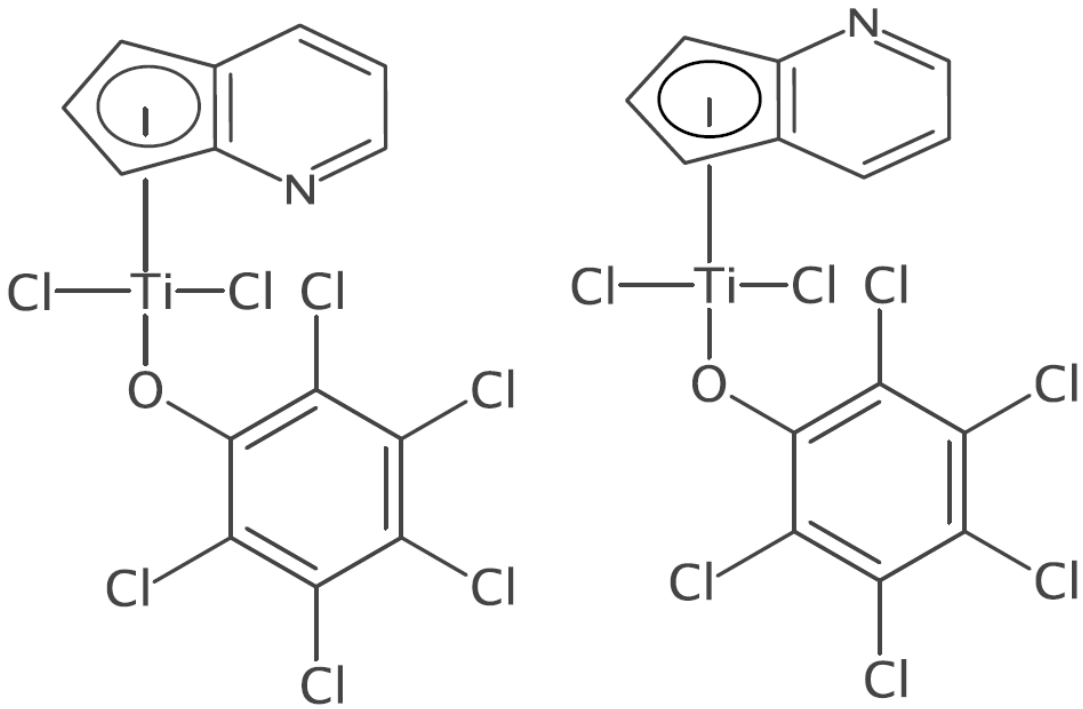
Sigma-Aldrich
717398
(*S*)[(S_p)-2-(Diphenylphosphino)ferrocenyl]-4-isopropyloxazoline
97%



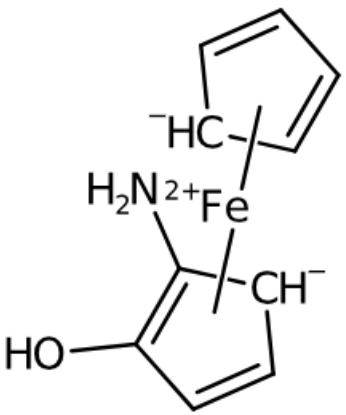
Sigma-Aldrich
682322
(*R*)-*N*-Methyl-*N*-diphenylphosphino-1-[(*S*)-2-diphenylphosphino)ferrocenyl]ethylamine
≥96%

haptic chirality

- Focusing on ferrocene chirality misses the bigger picture, that even half-sandwiches can be chiral.



updating proposals



- May 2023 InChI workshop on stereochemistry

c1cccc1.*[Fe]*.c1c(N)c(O)cc1 |m:5:0.1.2.3.4,7:8.9.11.13.14|

c1cccc1.*[Fe]*.c1c(N)c(O)cc1 |m:5:0.1.2.3.4,7:14.13.11.9.8|

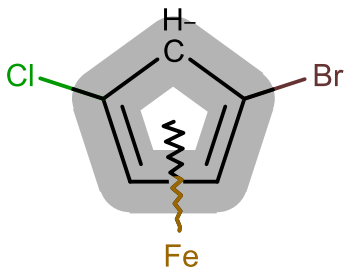
- August 2025 ACS Washington DC

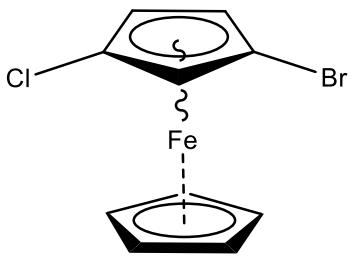
[cH-]1cccc1.*[Fe+2][*@].[cH-]1c(N)c(O)cc1 |m:5:0.1.2.3.4,7:8.9.11.13.14|

[cH-]1cccc1.*[Fe+2][*@@].[cH-]1c(N)c(O)cc1 |m:5:0.1.2.3.4,7:8.9.11.13.14|

unspecified haptic chirality

- A proposal is that for multi-centre bonding where the stereochemistry is unspecified, this should be indicated by a wavy bond.

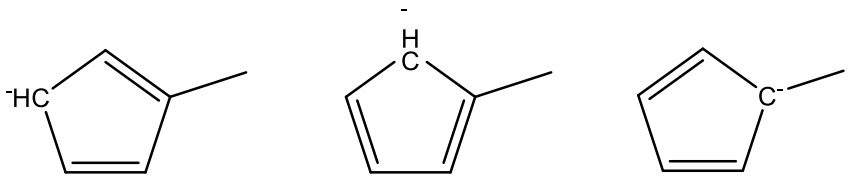




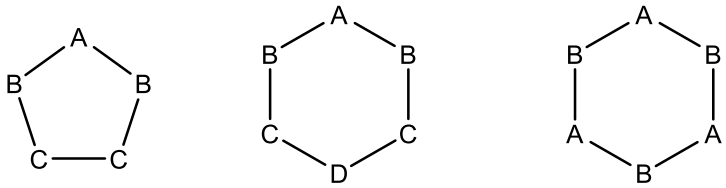
- This is analogous to 2D depictions where double bond stereochemistry is implied by co-ordinates.
- Wavy bonds not required when non-chiral.

Determining haptic chirality

- An “Eta-system” is not chiral if it has a plane of symmetry (ignoring bond orders and formal charges).

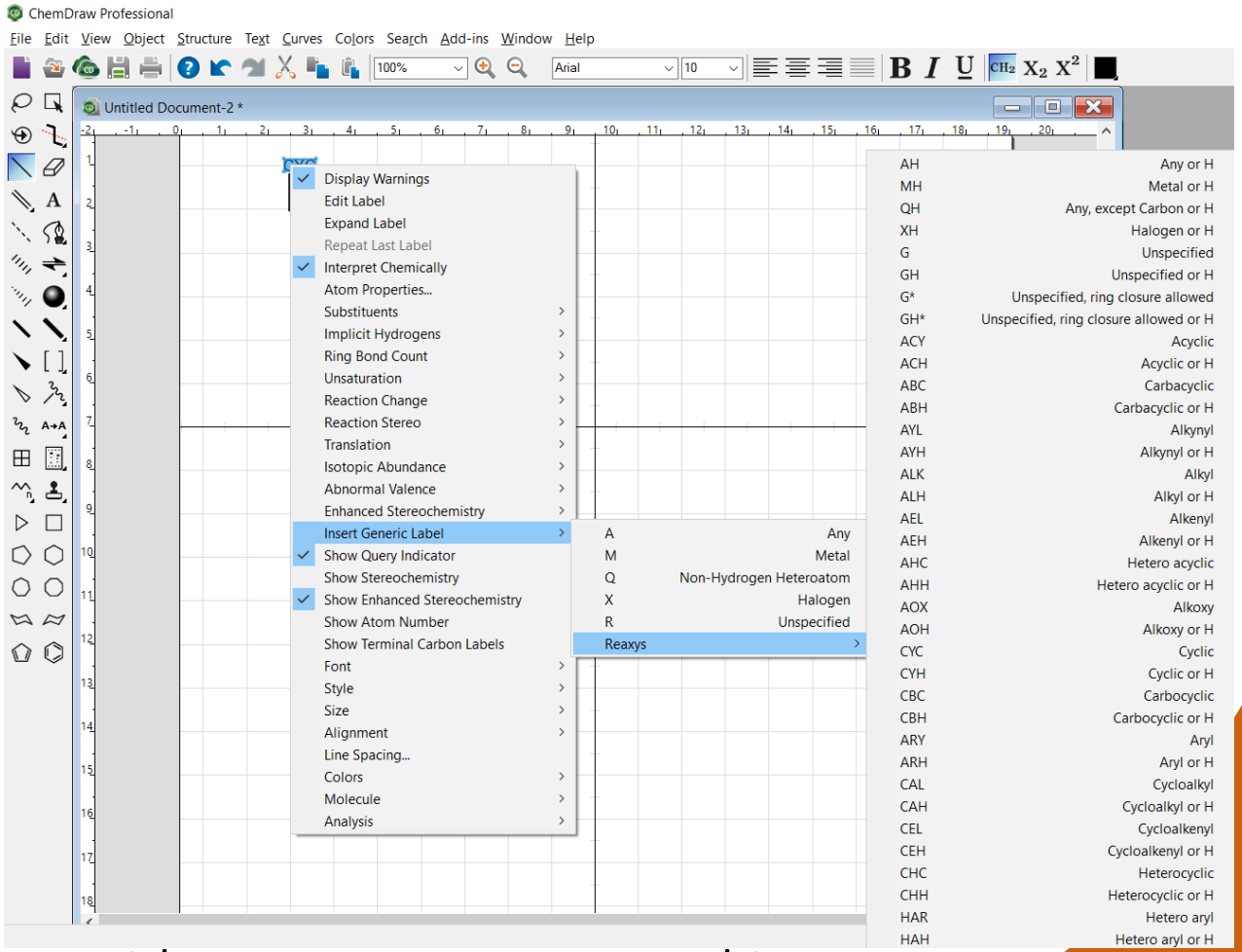


- Number of odd-sized symmetry classes less than 2-(system size mod 2).
  - 5-membered ring, 1 odd-sized symmetry class
  - 6-membered ring, ≤ 2 symmetry class.



chemdraw smarts support in rdkit

- [ACY]
- [ALK]
- [ARY]
- [CBC]
- [CYC]
- [CHC]
- [HAR]



- CXSMARTS: [*]-[*] |$G17_p;AlkaliMetal_p$|

myths of biopoLymers

- There are a small (manageable) number of monomers.
- There’s a one-to-one mapping between monomers and all-atom representations.
- HELM/MonomerMol representations are sufficient.

pubchem contains…

- 122,270,062 compounds (September 2025)
- > 35,811 amino acid monomers
  - > 20,361 β-alanine monomers
- > 6,893 nucleotide monomers
- > 3,699 monosaccharide monomers
- > 2,385 substituent (SAG) monomers
  - > 1,441 Fatty acid monomers
- > 97 lipid monomers

How many Amino acids present?

20 common amino acids

Ala, Cys, Lys, Thr

⇩

87 amino acids

Ala, Cys, Hcy, Lys, 2Nal, Ncy, Thr

⇩

1095 including substituents

Thr, Thr(*t*Bu), Thr(Bn), Thr(PO_3H_2)

⇩

3546 including stereo variants, terminal variants, linker variants, α-methylated

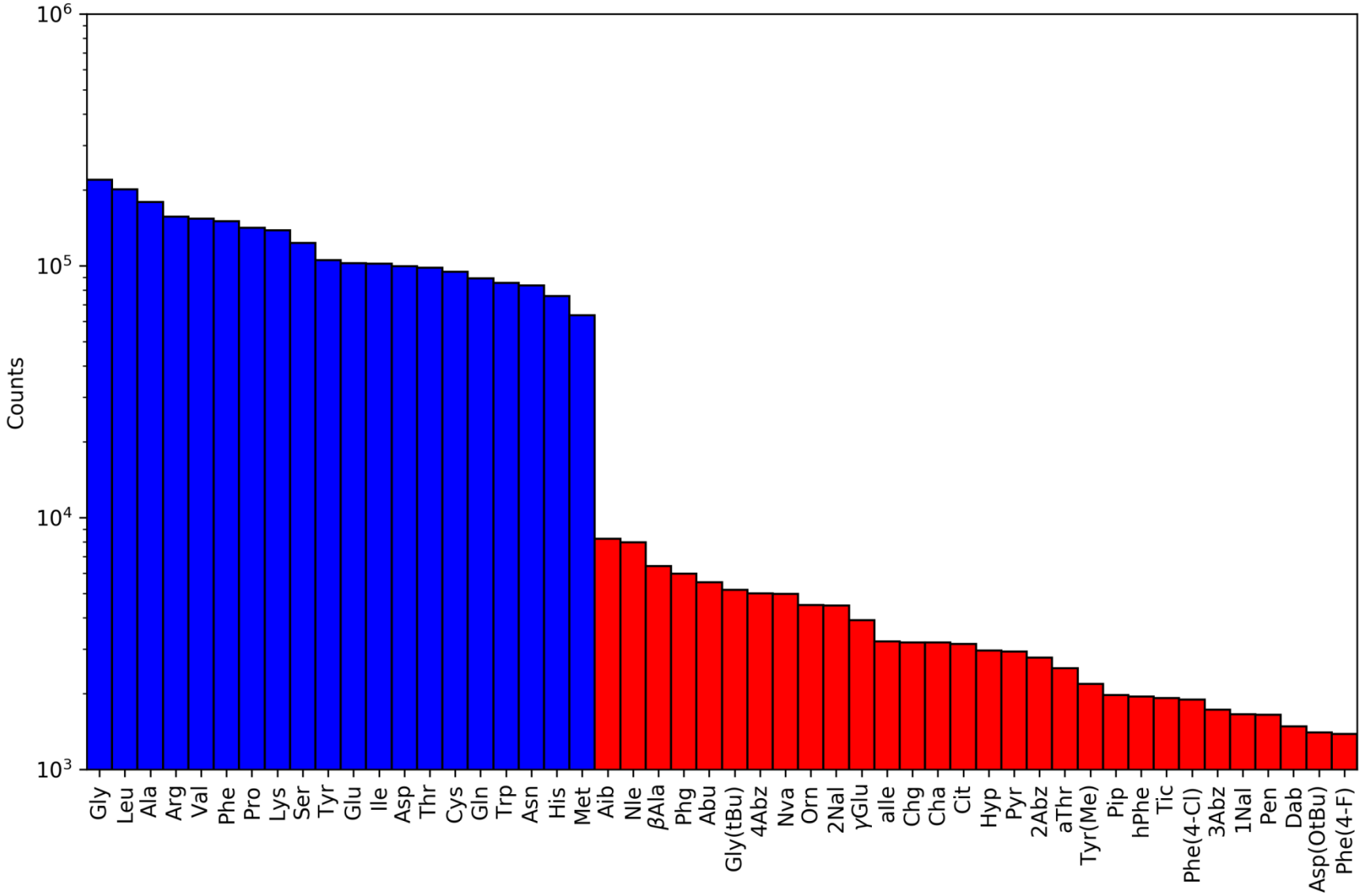
Thr, D-Thr, DL-Thr, aThr, Thr-ol, aMeThr

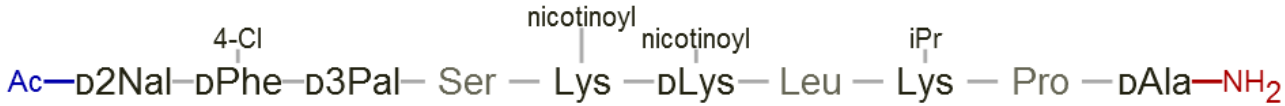
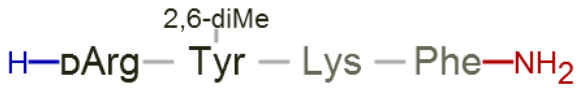
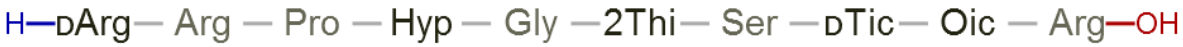
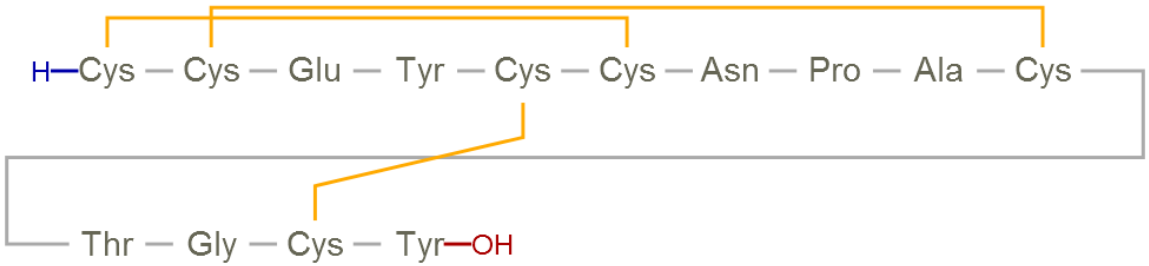
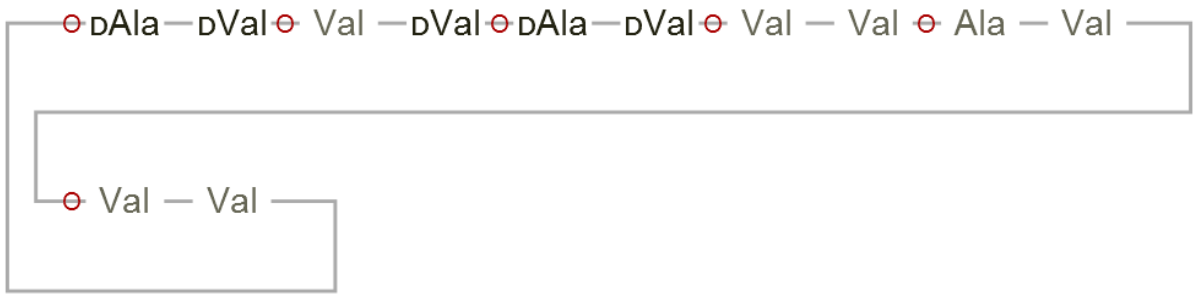
⇩

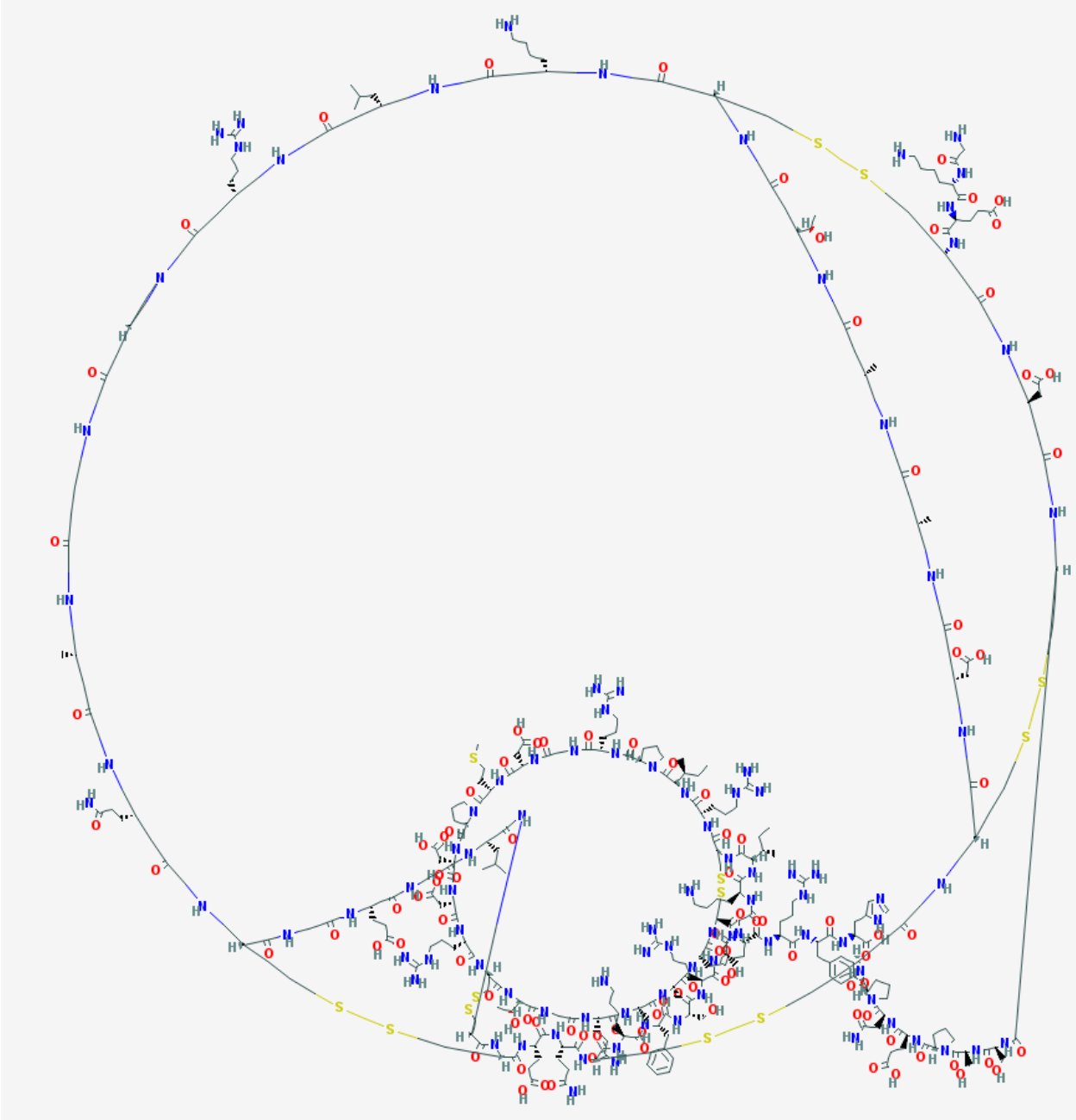
8125 including N-substituted variants

Thr, Me-Thr, Boc-Thr, Me_2-Thr, Fmoc-N(Me)Thr

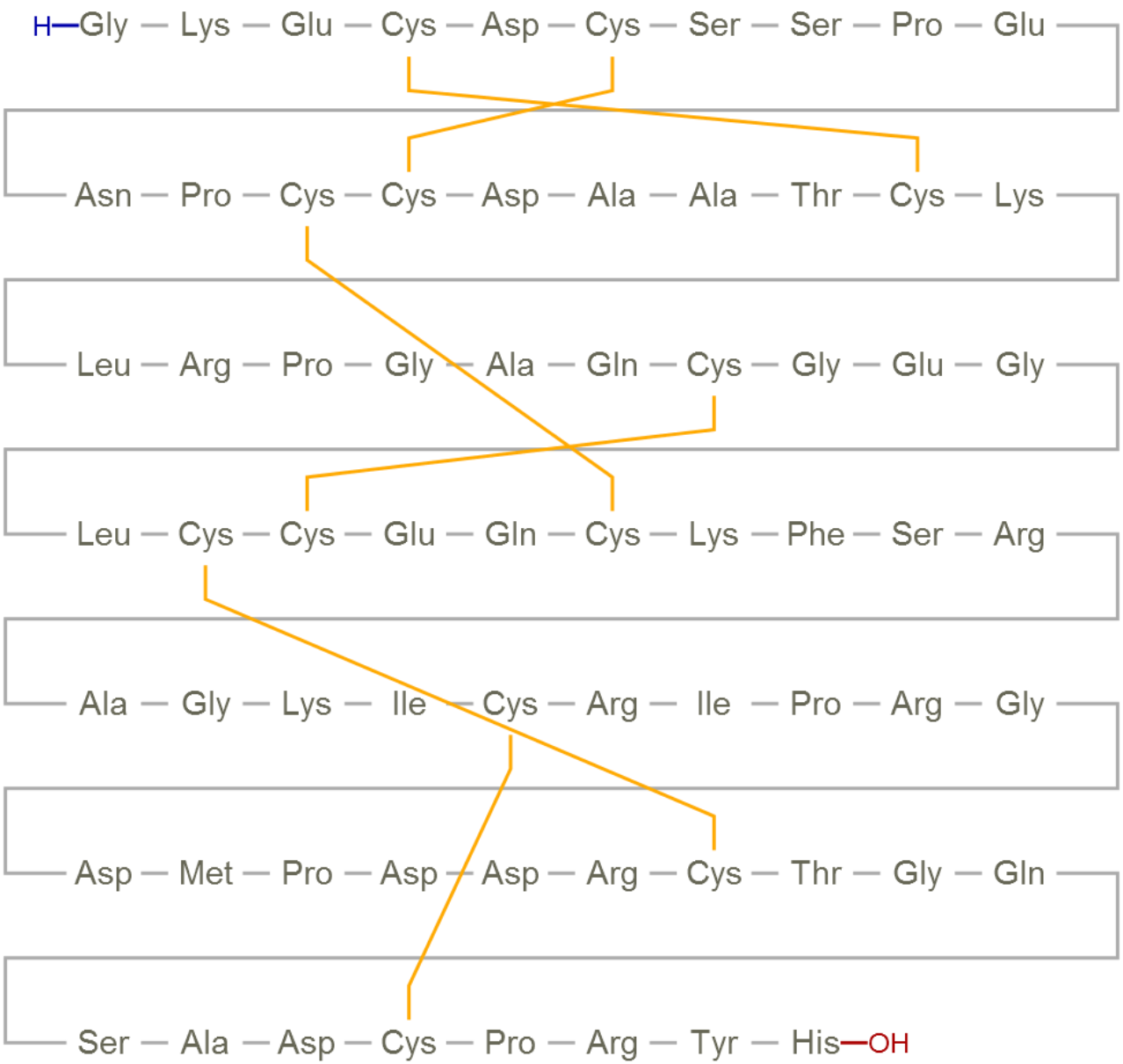
Amino acids in PubChem Structures containing at least three amino acids



iturelix CID16130938	
elamipretide CID11764719	
histrelin CID25077993	
icatibant CID71364	
linaclotide CID16158208	
valinomycin CID5649	



CID56842075
Rhodostomin



CID56842075
Rhodostomin

Lipidated Peptide Dendrimers Killing Multidrug-Resistant Bacteria

Thissa N. Siriwardena[†], Michaela Stach[†], Runze He[†‡], Bee-Ha Gan[†], Sacha Javor[†], Marc Heitz[†], Lan Ma[‡§∥], Xiangju Cai[§], Peng Chen[⊥], Dengwen Wei[⊥], Hongtao Li[⊥], Jun Ma[§], Thilo Köhler[∇], Christian van Delden[∇▽], Tamis Darbre[*†], and Jean-Louis Reymond[*†]

[†] Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

[‡] Shanghai Space Peptides Pharmaceutical Co. Ltd, Shanghai 201210, China

[§] College of Pharmacy, Gansu University of Chinese Medicine, Dingxi East Road 35, Chenguan District, Lanzhou, Gansu Province 730000, China

[∥] Lanzhou Ruibei Pharmaceutical R&D Co., Ltd., Lanzhou, Gansu Province 730000, China

[⊥] Department of General Surgery, Lanzhou General Hospital of Lanzhou Military Region, PLA, 333 South Binhe Road, Qilihe District, Lanzhou, Gansu Province 730046, China

[∇] Department of Microbiology and Molecular Medicine, University of Geneva, CH-1211 Geneva, Switzerland

[▽] Service of Infectious Diseases, University Hospital of Geneva, CH-1205 Geneva, Switzerland

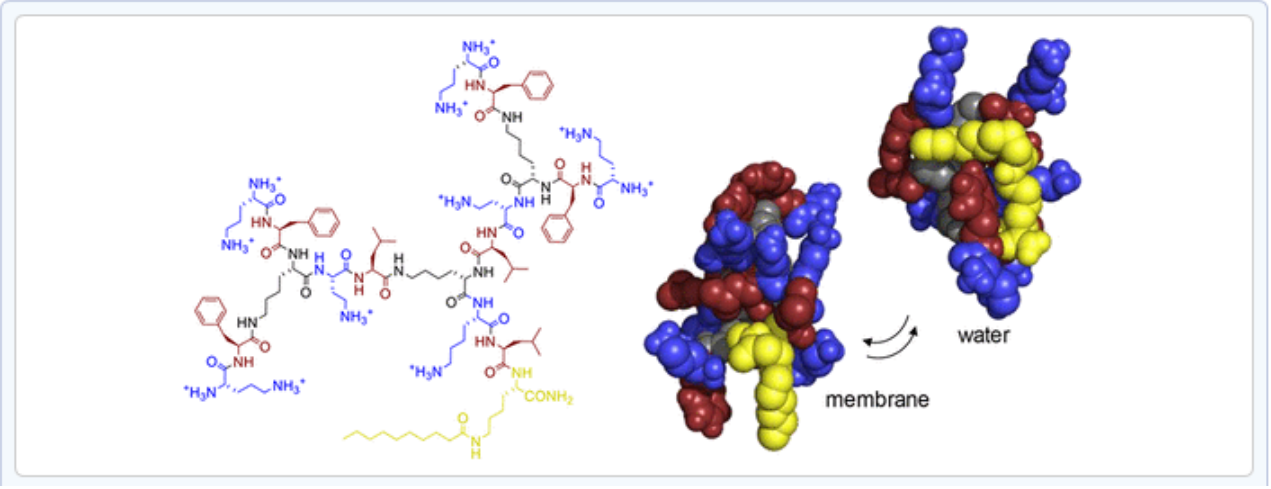
J. Am. Chem. Soc., **2018**, *140* (1), pp 423–432
DOI: 10.1021/jacs.7b11037
Publication Date (Web): December 5, 2017

*jean-louis.reymond@dcb.unibe.ch,
*tamis.darbre@dcb.unibe.ch

Cite this: *J. Am. Chem. Soc.* 140, 1, 423-432

RIS Citation GO

Abstract



Names preferred by MACHINES

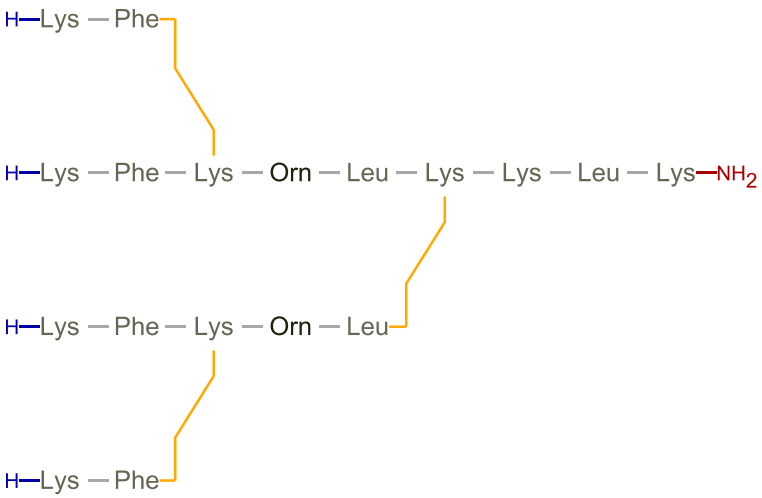
SMILES

CC(C)C[C@H](NC(=O)[C@H](CCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)C(=O)NCCCC[C@H](NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](Cc1ccccc1)NC(=O)[C@@H](N)CCCCN)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CCCCN)C(N)=O

HELM

PEPTIDE1{K.F.K.[Orn].L.K.K.L.K.[am]}|PEPTIDE2{K.F.K.[Orn].L}|PEPTIDE3{K.F}|PEPTIDE4{K.F}$PEPTIDE1,PEPTIDE2,6:R3-5:R2|PEPTIDE2,PEPTIDE3,3:R3-2:R2|PEPTIDE1,PEPTIDE4,3:R3-2:R2$$$

H-Lys-Phe-(1).H-Lys-Phe-Lys(1)-Orn-Leu-Lys(2)-Lys-Leu-Lys-NH2.H-Lys-Phe-Lys(3)-Orn-Leu-(2).H-Lys-Phe-(3)



How many MonoSaccharides present?

113 aldoses, ketoses, aldonic and uronic acids with from 5-9 carbons

AltA, Glc, L-Man, L-Gal, Fru, L-gro-D-glcHept

⇩

407 including deoxy variants, ring variants

L-Glcf, Mans, 2-deoxy-D-manHept, 3-deoxy-D-glcOct2ulo-onic

⇩

971 including anomeric stereo

a-Man, 3,4-deoxy-a-D-eryHex, b-Tyv

⇩

7094 including common substituents at non-anomeric positions

Xylf5Me, a-L-ManNAc3Ac4Ac6Ac, Glc2P3P6P

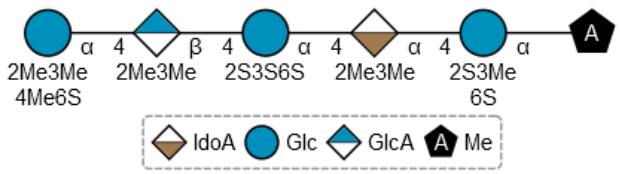
⇩

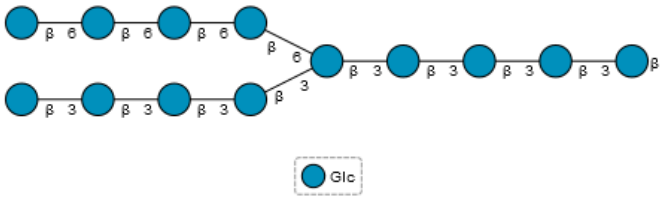
26641 including any substituent anywhere

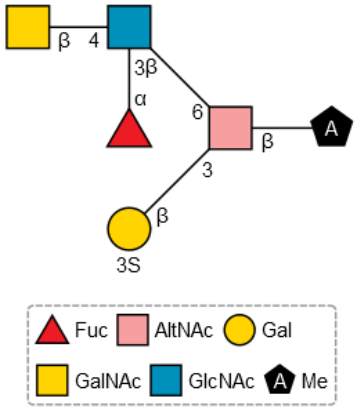
Bz(-2)[Tos(-3)]Ara4Ac(b)-O-Me, TMS(-4)[TMS(-6)]GlcNAc3Me(a)-O-Me

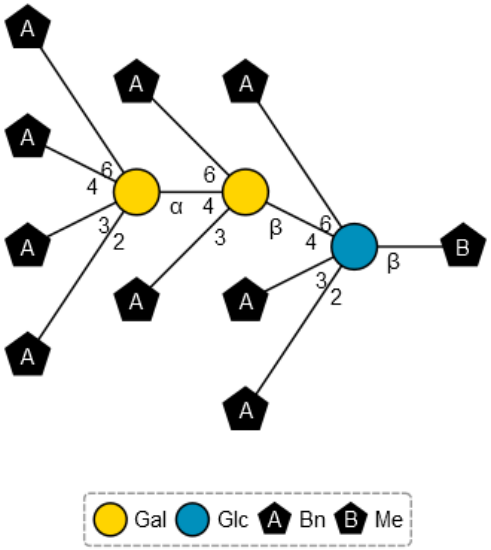
* This analysis excludes monosaccharides with missing stereochemistry

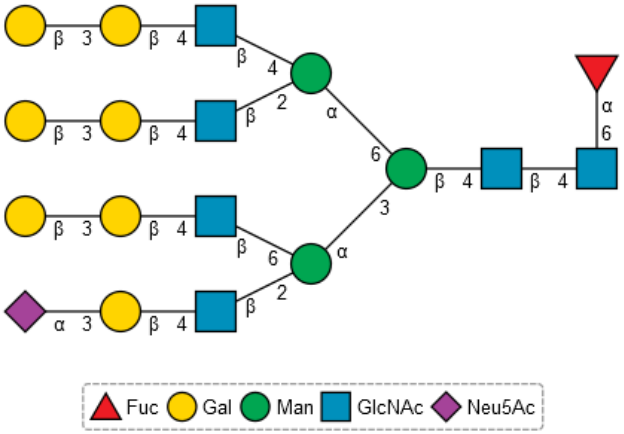
oligosaccharide depictions

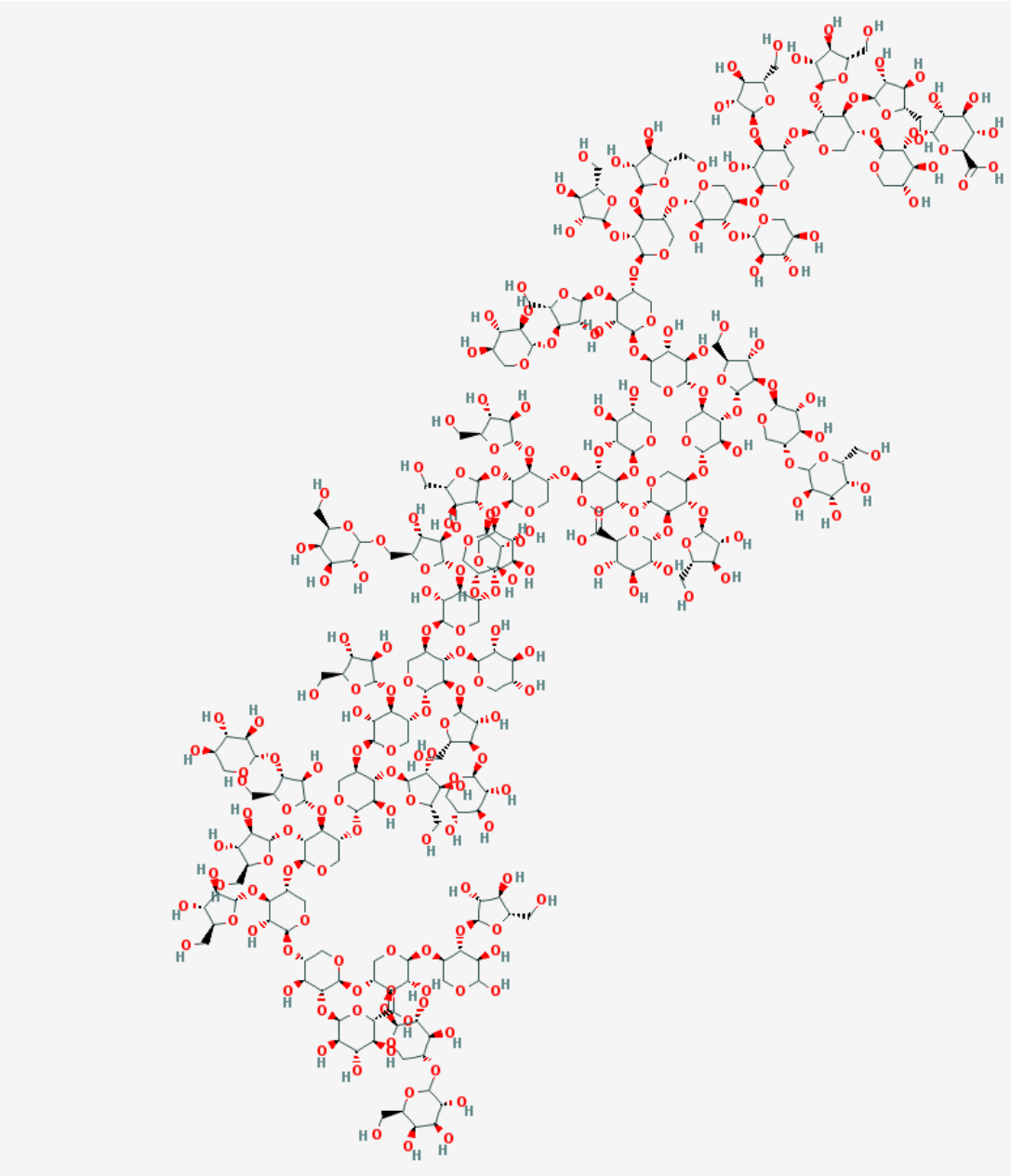






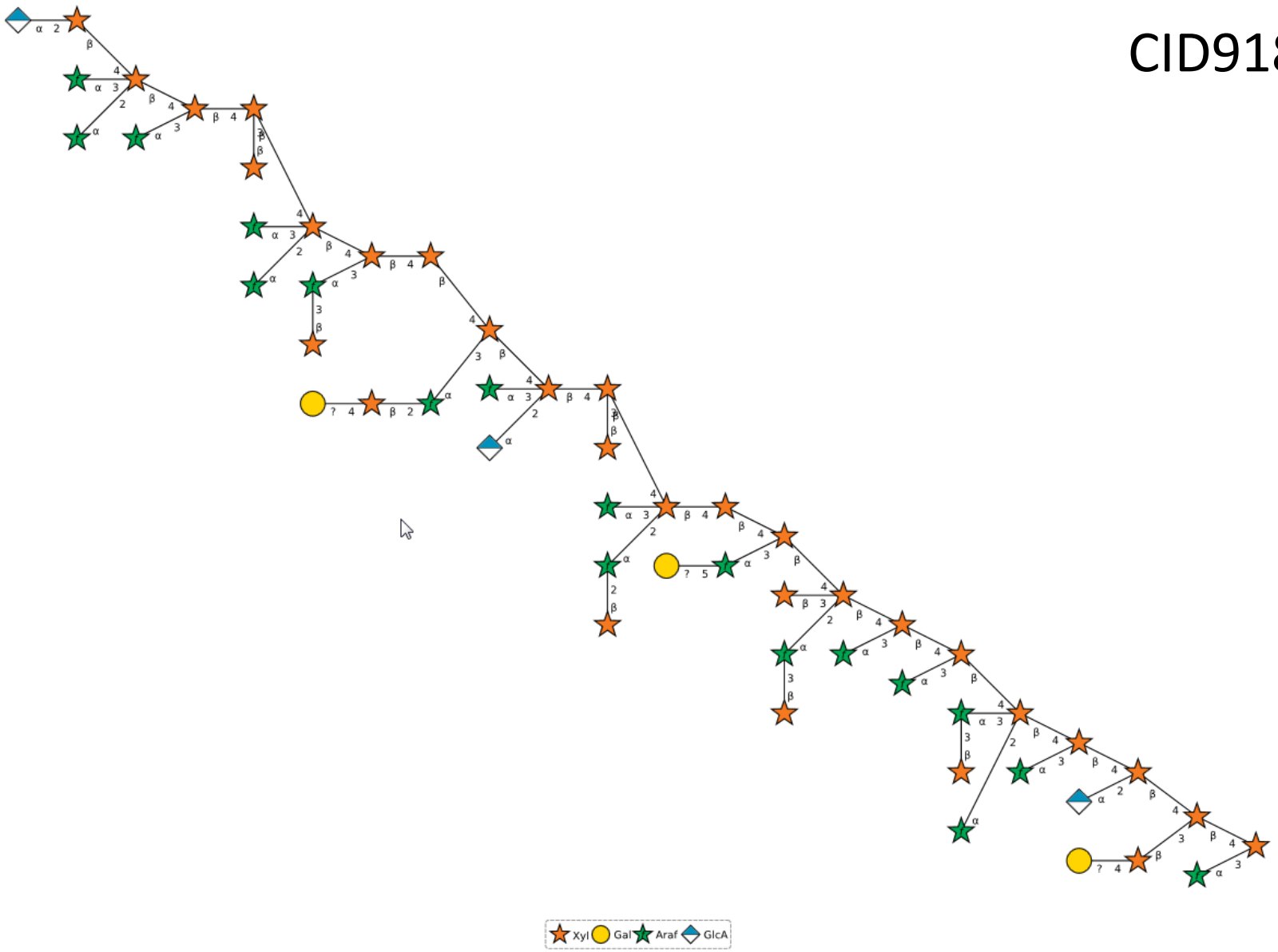






CID91852014

CID91852014



fasta to smiles is the easy bit

- Many monomers exist in alternate forms due to protonation and tautomerization.
- Physiological vs. neutral forms differ.
- Histidine forms HID, HIE and HIP.
- Representation of nitro groups, azides, etc.
- Often overlooked is undefined chirality.

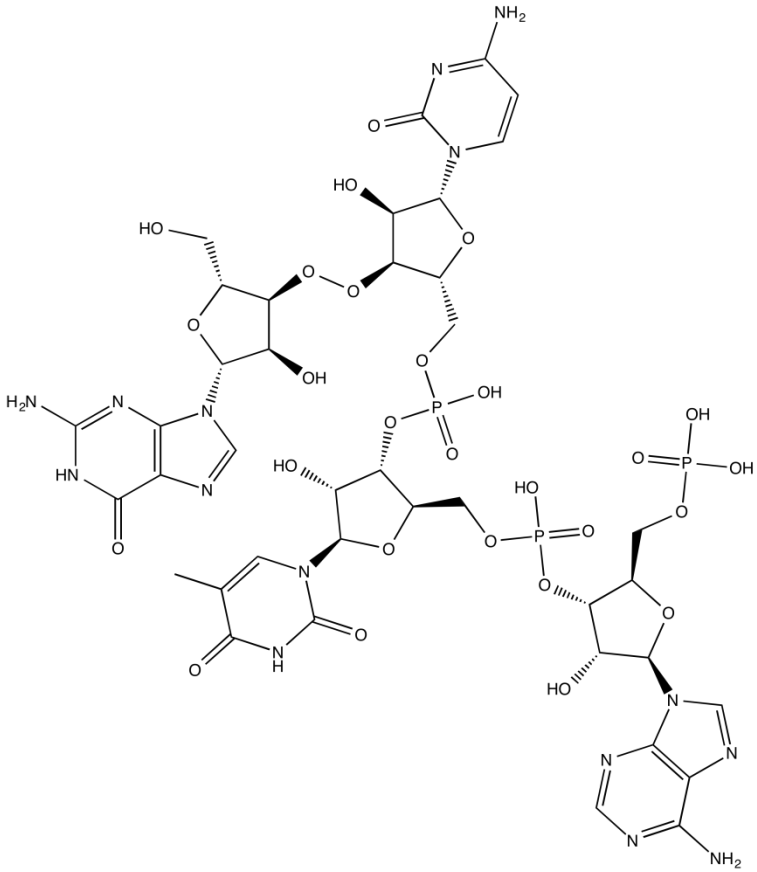
Biovia vs. pistoia helm issues

- RDKit's implementation doesn't have Biovia's bugs.
- The monomer phase problem in RNA registration:
  - Sequence: GATTACA
  - Interpretation: 5'-G-A-T-T-A-C-A-3'
  - Implied phosphates: 5'-G-P-A-P-T-P-T-P-A-P-C-P-A-3'
  - IUMB/PDB/Biovia: G-PA-PT-PT-PA-PC-PA
  - Pistoia HELM: GP-AP-TP-TP-AP-CP-A

- This discrepancy complicates RNA registration and leads to serious bugs in Biovia's HELM support.

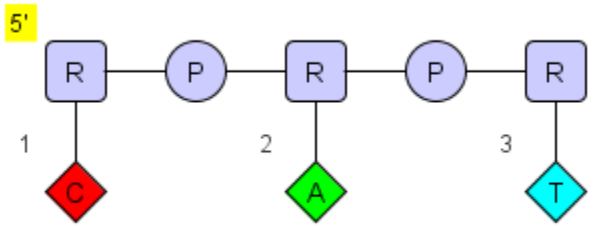
Expanded SCSR file

select molfile(mol('RNA1{R(A)P.R(T)P.R(C)P.R(G)}$$$$')) from dual;

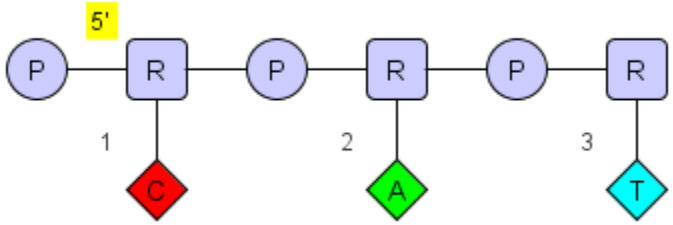


Rdkit helm implementation

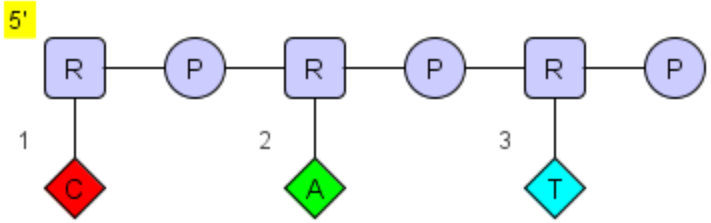
No Caps
Flavor =2



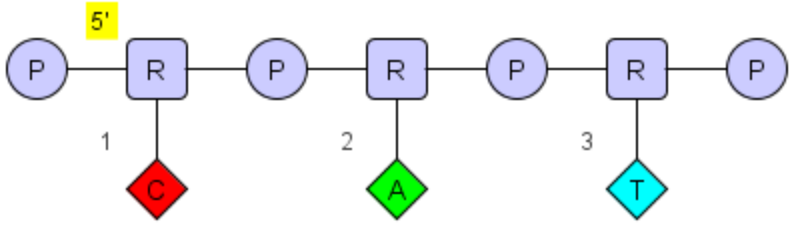
5’-Cap
Flavor = 3



3’-Cap
Flavor = 4



Both Caps
Flavor = 5



Schrödinger HELM

- Rachel Walker's talk this morning introduced in an interesting variant:
  - PEPTIDE1{C.G.D(K)}$$$$V2.0
- Attempting to read this with the Pistoia Alliance's HELM editor results in the error message: Unknown HELM_AA monomer name "D(K)", and the validate HELM call fails.
- Admittedly, this was in a proposal for monomer substructure search.

glycoct bond src/dst qualifiers

- For most amino acids, stereochemistry is within the monomer and can be defined by its definition.
- But for oligosaccharides (and metalloproteins) the stereochemistry occurs between monomers.
- The GlycoCT file format has connection semantics:
  - h exchange of H
  - r exchange of R-prochiral H
  - s exchange of S-prochiral H
  - o exchange of O
  - c exchange of backbone carbon

inline helm (If you've not seen it)

- Ac(1)-Ala-Cys(1)-OH
  - PEPTIDE1{[*CC(=O)* |$_R3;;;;_R2$|].A.C}$ PEPTIDE1,PEPTIDE1,1:R3-3:R3$$$
- H-Ncy(1)-Ala-Cys(1)-OH
  - PEPTIDE1{[N[C@@H](S*)C(=O)* |$;;;_R3;;;_R2$|].A.C}$ PEPTIDE1,PEPTIDE1,1:R3-3:R3$$$
- N4-acetylcytidine (ac4Cyt-Ribf)
  - RNA1{R([CC(=O)Nc1ccn(*)c(=O)n1 |$;;;;;;;;_R1;;;$|])}$$$$

Helm teething problems

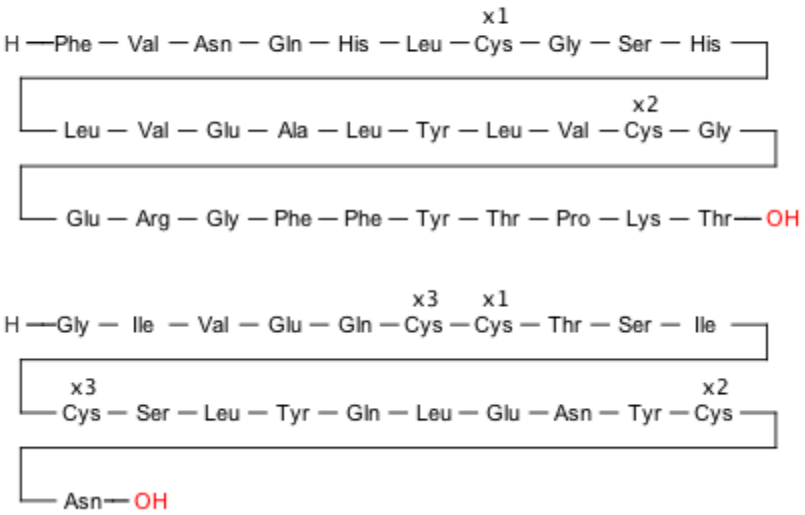
- Pistoia's HELM notation marks a significant advance over the limitations of one-letter bioinformatics.
- Alas, its original goals didn't include data exchange, which has only recently been addressed by the extensions of inlineHELM and XHELM [and fixes from NextMove Software for improved interoperability].
- Alas, this still doesn't address some core limitations:
  - Pistoia Monomer Library: PEPTIDE1{[fmoc].A}$$$$
  - EBI ChEMBL Monomers: PEPTIDE1{[Fmoc_A]}$$$$

25 ACS National Meeting, Boston, MA, Sunday 16th August 2015

helm dialects

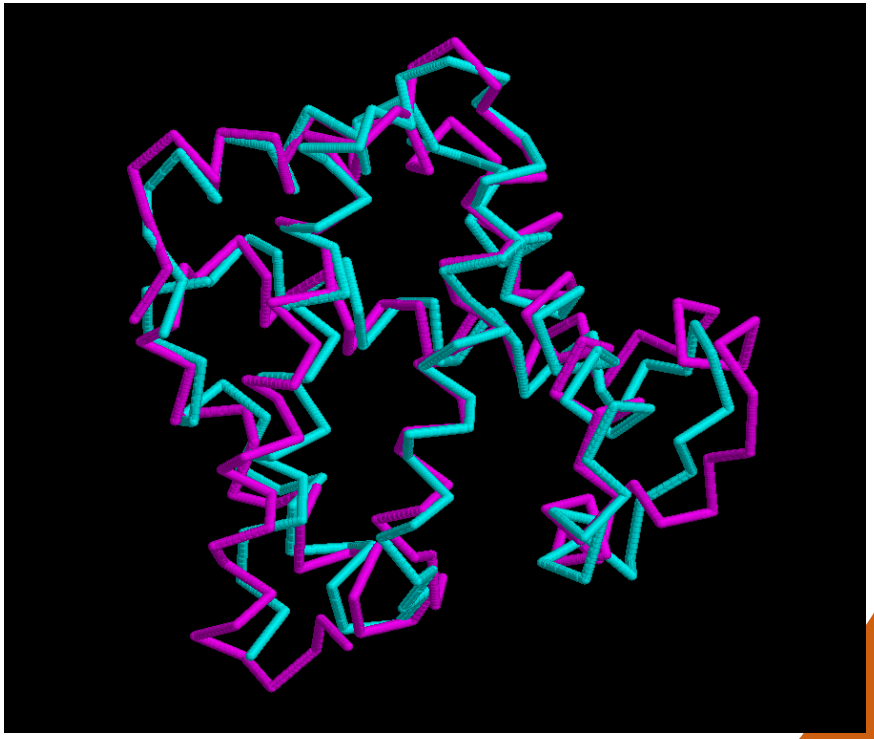
- Heptares/ChemAxon HELM (Conor Scully 20018)
  - CHEM1{[Ac]}|PEPTIDE1{R.K.C.Y.[D-Leu].P.E.C.S.F}|CHEM2{[NH2]}$PEPTIDE1,CHEM1,1:R1-1:R1|PEPTIDE1,CHEM2,10:R2-1:R1$$$V2.0
  - “CHEM{[Ac]}” is usually “PEPTIDE{[ac]}” in HELM.
  - “[D-Leu]” is more commonly “[dL]” in HELM.
  - “CHEM{[NH2]}” which has formula NH_3 is traditionally “PEPTIDE{[am]} in Pistoia/Pfizer HELM.
  - PEPTIDE1{[ac].R.K.C.Y.[dL].P.E.C.S.F.[am]}$$$$

biopolymer depiction



sequence-based superposition

```
Dell% ./pdbfit 1FDH.pdb.gz 1MBN.pdb.gz > out.pdb
PDBfit structural alignment
Roger Sayle, June 1994
Version 1.0

Sequence 1 Length: 141
Sequence 2 Length: 153
Gap penalty -10
Extend penalty -2
Alignment Score = 166

    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH 50
    ||| ::   |   |:|| |: :  | : | |:| | | |   | :|
    VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLK 50

    GSAQVKG------HGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD 94
      |::|:      ||  |  ||   :   ::    | :|:: || | ::
    TEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKI- 99

    PVNF-KLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-- 141
    |:::  ::|: ::  | :: |::| :  :::::| |      :::||:
    PIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKEL 149

    ---- 141

    GYQG 153

Aligning on 140 points
RMS fitting on 140 points
RMS Error: 2.73723
-1.44656 about 0.795543 -0.603704 0.051499
```



pdb-based residue naming

L-N(Me)Ala	MAA	D-N(Me)Ala	33X
L-N(Me)Arg	MMO	D-N(Me)Arg	
L-N(Me)Asp	SOQ	D-N(Me)Asp	OEM
L-N(Me)Cys	NCY	D-N(Me)Cys	
L-N(Me)Gln	GNC	D-N(Me)Gln	HJV
L-N(Me)Glu	EME	D-N(Me)Glu	YBR
L-N(Me)His	E9V		
L-N(Me)Ile	IML	D-N(Me)Ile	
L-N(Me)Leu	MKE	D-N(Me)Leu	MLU
L-N(Me)Met	MME	D-N(Me)Met	
L-N(Me)Phe	MEA	D-N(Me)Phe	ZAE
L-N(Me)Ser	5JP	D-N(Me)Ser	DSE
L-N(Me)Thr	NZC	D-N(Me)Thr	
L-N(Me)Tyr	YNM	D-N(Me)Tyr	
L-N(Me)Val	MVA	D-N(Me)Val	MV9
N(Me)Gly = Sar	SAR	N(Me2)Gly	DMG

antibody representation #1

CCC(C)C(C(=O)NC(CCCNC(=[NH2+])N)C(=O)NC(CC(=O)N)C(=O)NC(Cc1ccc(cc1)O)C(=O)NC(CC(C)C)C(=O)NC(C)C(=O)NC(Cc2c[nH]c3c2cccc3)C(=O)NC(Cc4ccc(cc4)O)C(=O)NC(CCC(=O)N)C(=O)NC(CCC(=O)N)C(=O)NC(CCCC[NH3+])C(=O)N5CCCC5C(=O)NCC(=O)NC(CCCC[NH3+])C(=O)NC(C)C(=O)N6CCCC6C(=O)NC(CCCC[NH3+])C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)CC)C(=O)NC(Cc7ccc(cc7)O)C(=O)NC(C)C(=O)NC(C)C(=O)NC(CO)C(=O)NC(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(CCC(=O)N)C(=O)NC(CO)C(=O)NCC(=O)NC(C(C)C)C(=O)N8CCCC8C(=O)NC(CO)C(=O)NC(CCCNC(=[NH2+])N)C(=O)NC(Cc9ccccc9)C(=O)NC(CO)C(=O)NCC(=O)NC(CO)C(=O)NCC(=O)NC(CO)C(=O)NCC(=O)NC(C(C)O)C(=O)NC(CC(=O)[O-])C(=O)NC(Cc1ccccc1)C(=O)NC(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(C(C)CC)C(=O)NC(CO)C(=O)NC(CO)C(=O)NC(CC(C)C)C(=O)NC(CCC(=O)N)C(=O)N1CCCC1C(=O)NC(CCC(=O)[O-])C(=O)NC(CC(=O)[O-])C(=O)NC(C(C)C)C(=O)NC(C)C(=O)NC(C(C)O)C(=O)NC(Cc1ccc(cc1)O)C(=O)NC(Cc1ccc(cc1)O)C(=O)NC(CS)C(=O)NC(CCC(=O)N)C(=O)NC(CCCNC(=[NH2+])N)C(=O)NC(Cc1ccc(cc1)O)C(=O)NC(CC(=O)N)C(=O)NC(CCCNC(=[NH2+])N)C(=O)NC(C)C(=O)N1CCCC1C(=O)NC(Cc1ccc(cc1)O)C(=O)NC(C(C)O)C(=O)NC(Cc1ccccc1)C(=O)NCC(=O)NC(CCC(=O)N)C(=O)NCC(=O)NC(C(C)O)C(=O)NC(CCCC[NH3+])C(=O)NC(C(C)C)C(=O)NC(CCC(=O)[O-])C(=O)NC(C(C)CC)C(=O)NC(CCCC[NH3+])C(=O)[O-])NC(=O)CNC(=O)C(CCC(=O)N)NC(=O)C(CO)NC(=O)C(C)NC(=O)C(CCCNC(=[NH2+])N)NC(=O)C(CS)NC(=O)C(C(C)O)NC(=O)C(C(C)CC)NC(=O)C(C(C)O)NC(=O)C(C(C)C)NC(=O)C(CCCNC(=[NH2+])N)NC(=O)C(CC(=O)[O-])NC(=O)CNC(=O)C(C(C)C)NC(=O)C(CO)NC(=O)C(C)NC(=O)C(CO)NC(=O)C(CC(C)C)NC(=O)C(CO)NC(=O)C(CO)NC(=O)C1CCCN1C(=O)C(CO)NC(=O)C(CCC(=O)N)NC(=O)C(C(C)O)NC(=O)C(CCSC)NC(=O)C(CCC(=O)N)NC(=O)C(C(C)CC)NC(=O)C(CC(=O)[O-])N Adalimumab_L

antibody representation #2

>Adalimumab_H

EVQLVESGGGLVQPGRSLRLSCAASGFTFDDYAMHWVRQAPGKGLEWVSAITWNSGHIDY

ADSVEGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAKVSYLSTASSLDYWGQGTLVTVSS

>Adalimumab_L

DIQMTQSPSSLSASVGDRVTITCRASQGIRNYLAWYQQKPGKAPKLLIYAASTLQSGVPS

RFSGSGSGTDFTLTISSLQPEDVATYYCQRYNRAPYTFGQGTKVEIK

>Infliximab_H

EVKLEESGGGLVQPGGSMKLSCVASGFIFSNHWMNWVRQSPEKGLEWVAEIRSKSINSAT

HYAESVKGRFTISRDDSKSAVYLQMTDLRTEDTGVYYCSRNYYGSTYDYWGQGTTLTVS

>Infliximab_L

DILLTQSPAILSVSPGERVSFSCRASQFVGSSIHWYQQRTNGSPRLLIKYASESMSGIPS

RFSGSGSGTDFTLSINTVESEDIADYYCQQSHSWPFTFGSGTNLEVK

ANTibody representation #3

- **Adalimumab**

CDR-L1: QGIRNY

CDR-L2: AAS

CDR-L3: QRYNRAPYT

CDR-H1: GFTFDDYA

CDR-H2: ITWNSGHI

CDR-H3: AKVSYLSTASSLDY

- **Infliximab**

CDR-L1: QFVGSS

CDR-L2: YAS

CDR-L3: QQSHSWPFT

CDR-H1: GFIFSNHW

CDR-H2: IRSKSINSAT

CDR-H3: RNYYGSTYDY

Antibody perception examples

```
DavesADC.mol
 SciTegic10231309212D
Courtesy of Keith Taylor, Ladera
 0  0  0  0  0  0            999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 10256 10536 0 0 1
M  V30 BEGIN ATOM
```

Isotyping can be used to check that an antibody has been correctly registered [bridges, glycans, etc.].

DavesADC_H: 450 AA Isotype: Ig gamma-1 [human] IMGT CDR lengths: [8.8.13]

..SCAAS[GFNIKDTY]IHWVR..EWVAR[IYPTNGYT]RYADS..AVYYC[SRWGGDGFYAMDY]WGQGT..

DavesADC_L: 213 AA Isotype: Ig kappa [human] IMGT CDR lengths: [6.3.9]

..TCRAS[QDVNTA]VAWYQ..KLLIY[SAS]FLYSG..ATYYC[QQHYTTPPT]FGQGT..

HC heavy chain: 446 AA Isotype: Ig gamma-1 [human] IMGT CDR lengths: [8.7.10]

..TCTVS[GGSISGYY]WSWIR..EWIGR[IYTSGST]NYNPS..AVYYC[ARGRFTYFDY]WGQGT..

LC light chain: 215 AA Isotype: Ig kappa [human] IMGT CDR lengths: [7.3.9]

..SCRAS[QIVSSAY]LAWYQ..RLLMF[GSS]SRATG..AVYYC[QQYGSSQGT]FGPGT..

Antibody perception examples

```
LC light chain: 215 AA
Isotype: Ig kappa [human]
Aligning against Adalimumab_L (214 AA)
Identity:    81.78% (175/214)
Similarity:  91.59% (196/214)

Ref DIQMTQSPSSLSASVGDRVTITCRASQGIRN-YLAWYQQKPGKAPKLLIYAASTLQSGVP 59
    :| :||||::|| | |:| |::||||| : : ||||||||||:||:||:: :|:  :|:|
Qry EIVLTQSPATLSLSPGERATLSCRASQIVSSAYLAWYQQKPGQAPRLLMFGSSSRATGIP 60

Ref SRFSGSGSGTDFTLTISSLQPEDVATYYCQRYNRAPYTFGQGTKVEIKRTVAAPSVFIFP 119
     |||||||||||||||| |:||| | ||||:|  :  ||| ||||:||||||||||||||
Qry DRFSGSGSGTDFTLTISRLEPEDFAVYYCQQYGSSQGTFGPGTKVDIKRTVAAPSVFIFP 120

Ref PSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTL 179
    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Qry PSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTL 180

Ref TLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 214
    |||||||||||||||||||||||||||||||||||
Qry TLSKADYEKHKVYACEVTHQGLSSPVTKSFNRGEC 215

Anchor positions: 25, 33, 49, 53, 88, 98
IMGT CDR lengths: [7.3.9]
..SCRAS[QIVSSAY]LAWYQ..RLLMF[GSS]SRATG..AVYYC[QQYGSSQGT]FGPGT..
```

Isotyping antibody chains

- Antibody chains may be classified into isotypes using global (NWS) sequence alignment [*n.b.* not blastp].

Light chains	**UNIPROT**	**ChEMBL19 count**
– Ig kappa [human]	IGKC_HUMAN	131
– Ig lambda [human]	LAC2_HUMAN	12
– Ig kappa [mouse]	IGKC_MOUSE	7
Heavy chains		
– Ig gamma-1 [human]	IGHG1_HUMAN	105
– Ig gamma-2 [human]	IGHG2_HUMAN	13
– Ig gamma-4 [human]	IGHG4_HUMAN	21
– Ig gamma-1 [mouse]	IGHG1_MOUSE	2
– Ig gamma-2A [mouse]	GCAA_MOUSE	2
– Ig gamma-2B {secreted} [mouse]	IGG2B_MOUSE	1
– Ig mu {secreted} [human]	IGHM_HUMAN	1 (Panobacumab)

chemical text mining

- From PubMed 5120, May 1976:
  - Brain kininase A hydrolyzes the Phe5-Ser6 peptide bond in bradykinin (Bk), Arg1-Pro2-Pro3-Gly4-Phe5-Ser6-Pro7-Phe8-Arg9. It is isoelectric near pH 5.2 and has a molecular weight of approximately 71 000. The enzyme also hydrolyzes the Phe-Ser peptide bond in Lys-Bk, Met-Lys-Bk, des-Arg1-Bk, Lys9-Bk, **Pro-Gly-Phe-Ser-Pro-Phe-Arg**, and Gly-Pro-Phe-Ser-Pro-Phe-Arg, but does not hydrolyze (0.1%) this bond in des-Phe8-Arg9-Bk.
- OPSIN doesn't handle trivial names (use a dictionary).
- Inorganic (Red Book) nomenclature: vanadium oxytrichloride.
- Phane nomenclature: From US20250223300A1
  - 2^6-methyl-5^6-(9-(oxetan-3-yl)-3,9-diazaspiro[5.5]undecan-3-yl)-5^1H-10-oxa-4-aza-5(2,1)-benzo[d]imidazola-1(3,4),2(2,4)-dipyridinacyclodecaphan-3-one

evaluating bond order perception

- GDB is a poor proxy for chemical space, given the element and hetero-neighbor constraints.
- The MMFF94 validation suite may provide a better coverage of functional groups/relevant chemistry.
- All atom inputs, with explicit hydrogens and perfect geometries, are obviously “easier” than real world data, such as ligands from PDB files.

evaluating bond order perception

- Apologies to Jan and Greg, the truly terrible SMILES in Greg's blog post are from the original QM9 data files, that we're (I believe) supposed to be an implementation of my "Cruft to Content" work.
- I can confirm that as described preserves the correct hydrogen count, and doesn't generate radicals.
- The motivation was that using CSD's bond perception logic (d < Rcov(i)+Rcov(j)+0.45) might perform better or as well as Jan's d < 1.3*(Rcov(i)+Rcov(j)) logic.

Distance geometry

- RDKit's bounds matrix is poorly constrained.
- For acetonitrile, CC#N, with no "freedom"
  - C^0-C^1 1.453-1.473
  - C^1#N^2 1.147-1.167
  - C^0..N^2 2.58-2.64 Trigonometry: > 155.35°
- The "ideal" bond lengths/angles used by molecular mechanics forcefields are misleading, and are not the expected values in the final minimized structure.
- Substructures (sub distance matrices) like thiophene should have no degrees of freedom (rigid bodies).

μÅ atomic accuracy

- A quick word on Patrick's complaints on 4-digit accuracy in V2000 Mol files.
  - RasMol rounded atomic co-ordinates to 1/256th Å.
- The remarkable reproducabilty of (^{19}F-)NMR is the result of statistical mechanics (the central limit theorem) averaging the observations of trillions of molecules, over multi-second time scales, bombarded by solvent at room temperature (295 K).
- Yes, the energy surface is steep (and 2Å RMS as a definition of good is a rant for another day)…

final remarks

- It has been a fantastic RDKit UGM.
- A number of impressive improvements.
- Less to complain about that previous years.

acknowledgements

- The Team at NextMove Software
  - John Mayfield
  - Ingvar Lagerstedt
  - Rachael Pirie
  - Michael Blakey
  - Zayyan Masud

- David Weininger, Greg Landrum and Daniel Lowe.

related talks

- Daylight MUG, 29th February 1996.
- ACS Fall, San Diego, 26th March 2012.
- InChI Biologics Meeting, 27th October 2014.
- 4th RDKit UGM, Zurich, 2nd September 2015.
- 5th RDKit UGM, Basel, 27th October 2016.