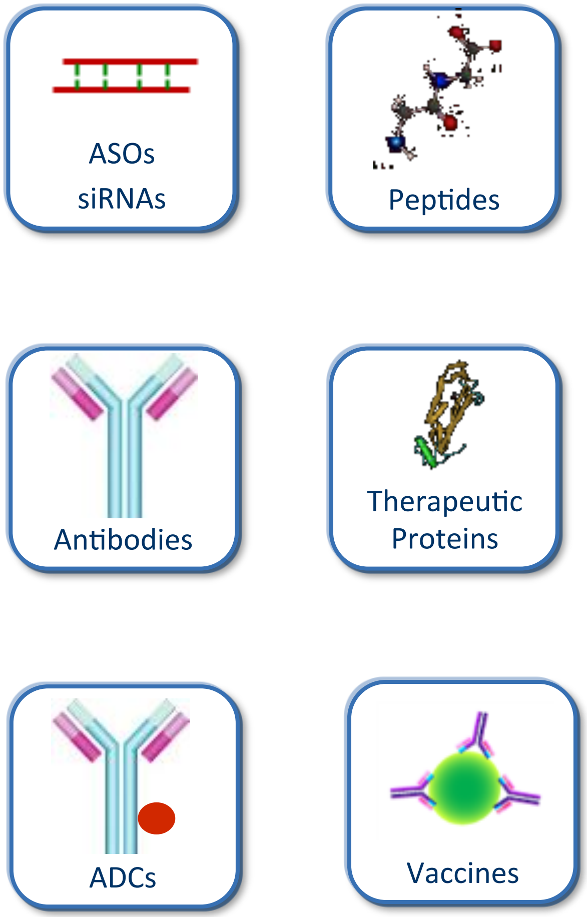
HELM: an Open Standard for Complex Polymeric Structures

Tianhong Zhang, Ph.D.
Research Informatics, Pfizer Inc.
HELM Project, Pistoia Alliance

August 16, 2015
ACS Boston

Biotherapeutics



ASOs
siRNAs

Peptides

Antibodies

Therapeutic
Proteins

ADCs

Vaccines

- The pharm/biotech industry has been shifting from small molecules to biologics

- The types of biologics are quite diverse

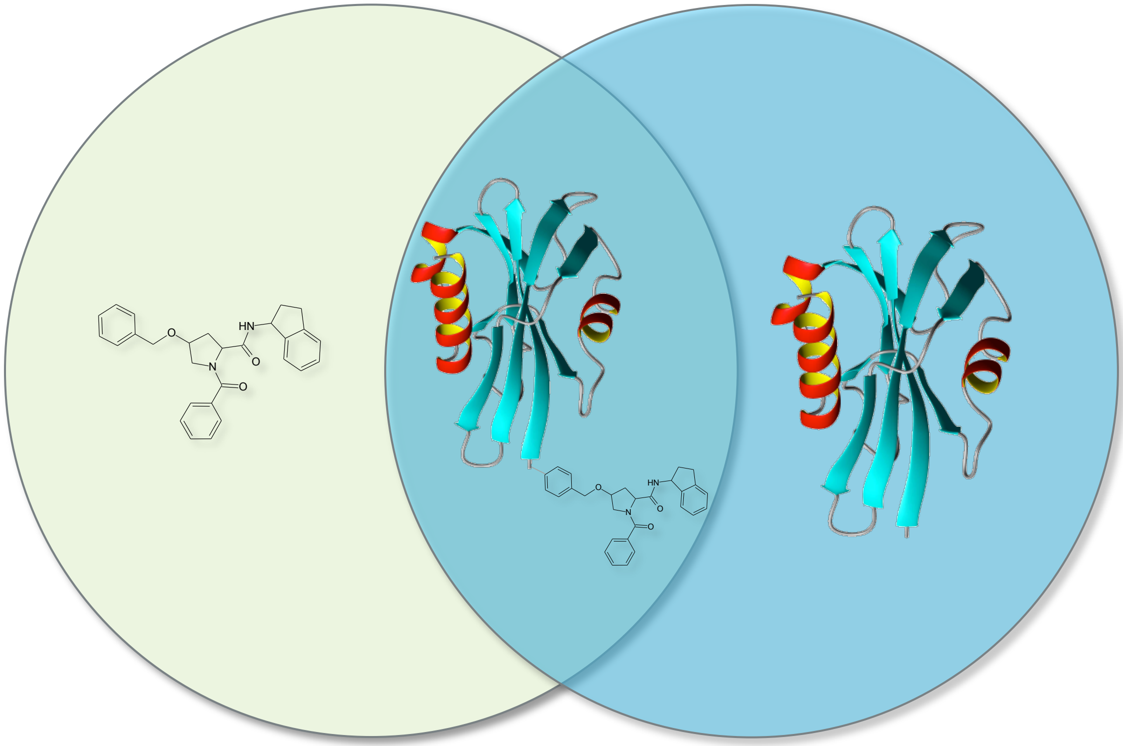
- Most biologics are chemically modified

The Informatics Challenge

Cheminformatics Tools

Bioinformatics Tools

Small Molecules



Sequences

How to represent chemically modified biopolymers so that they are machine readable, and we can build informatics tools to facilitate their research and development?

Biomolecule Structure Formats

- CHUCKLES
  - Siani et al, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588-593
  - Peptide and analogs
- Protein Line Notation (PLN)
  - Jensen et al, *J. Chem. Inf. Model.* **2008**, *48,* 2404–2413
  - Protein and Peptide
- Self-Contained Sequence Representation (SCSR)
  - *Chen et al, J. Chem. Inf. Model.*, **2011**, *51(9),* 2186-2208
  - Enhanced MOLFILE V3000 format
  - All biologics

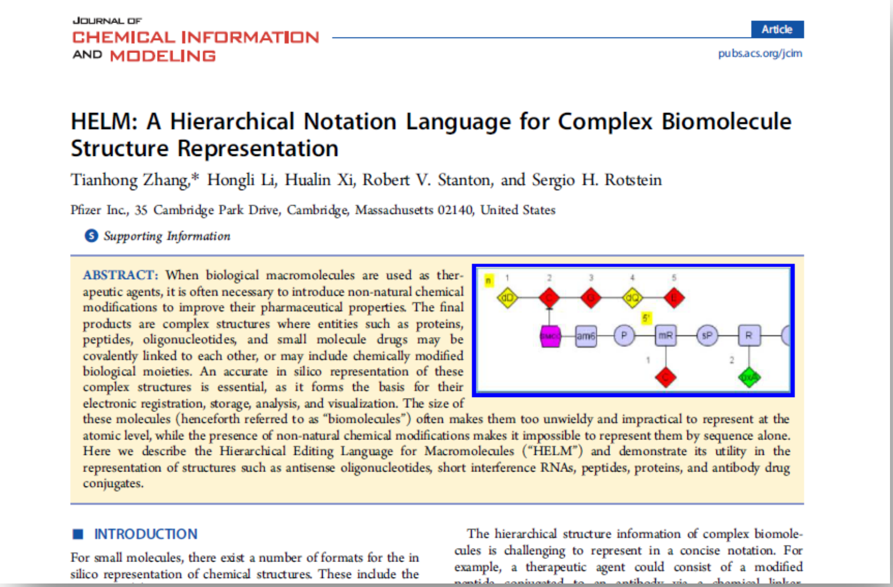
HELM Line Notation

- Hierarchical Editing Language for Macromolecules
  - *J. Chem. Inf. Model* **2012**, 52, 2796-2806

- Notation Language for polymers
  - Vocabulary (Monomers)
  - Grammar (Syntax)

- Hierarchical
  - Complex Polymer
  - Simple Polymer
  - Monomer
  - Atom

- HELM to macromolecules
  - SMILES or InChi to small molecules



JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation

Tianhong Zhang,* Hongli Li, Hualin Xi, Robert V. Stanton, and Sergio H. Rotstein

Pfizer Inc., 35 Cambridge Park Drive, Cambridge, Massachusetts 02140, United States

Supporting Information

ABSTRACT: When biological macromolecules are used as therapeutic agents, it is often necessary to introduce non-natural chemical modifications to improve their pharmaceutical properties. The final products are complex structures where entities such as proteins, peptides, oligonucleotides, and small molecule drugs may be covalently linked to each other, or may include chemically modified biological moieties. An accurate in silico representation of these complex structures is essential, as it forms the basis for their electronic registration, storage, analysis, and visualization. The size of these molecules (henceforth referred to as "biomolecules") often makes them too unwieldy and impractical to represent at the atomic level, while the presence of non-natural chemical modifications makes it impossible to represent them by sequence alone. Here we describe the Hierarchical Editing Language for Macromolecules ("HELM") and demonstrate its utility in the representation of structures such as antisense oligonucleotides, short interference RNAs, peptides, proteins, and antibody drug conjugates.

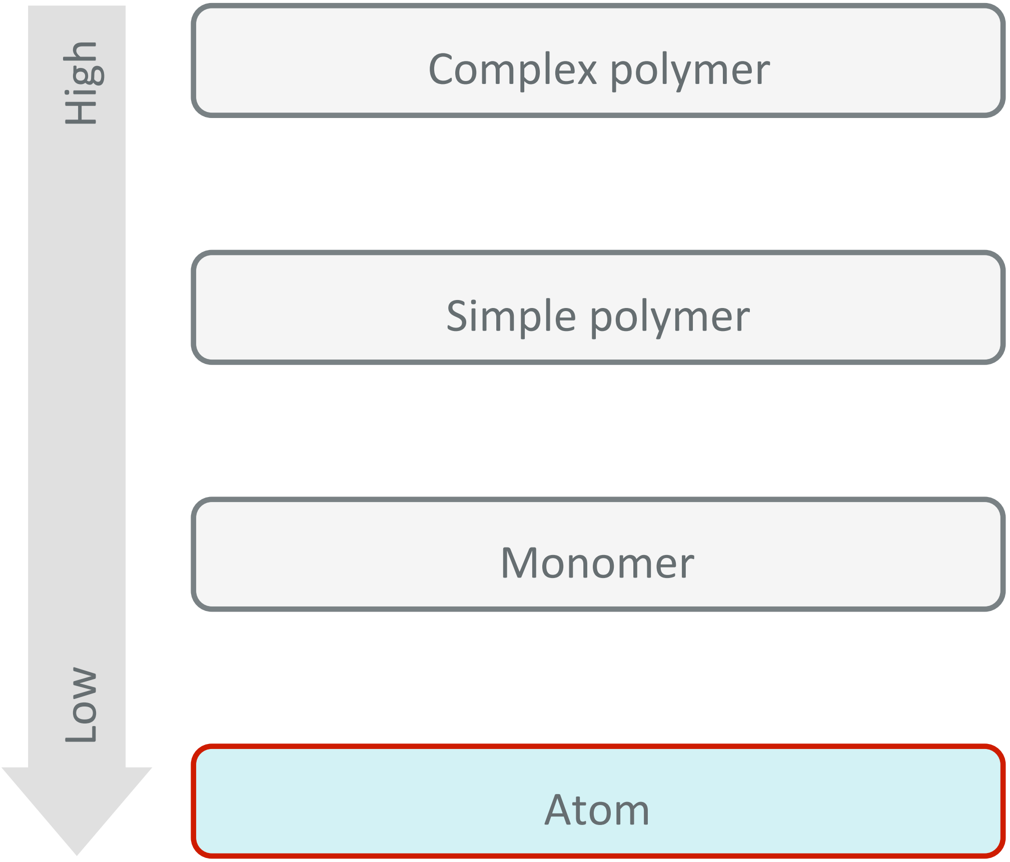
INTRODUCTION

For small molecules, there exist a number of formats for the in silico representation of chemical structures. These include the

The hierarchical structure information of complex biomolecules is challenging to represent in a concise notation. For example, a therapeutic agent could consist of a modified

Structure hierarchy

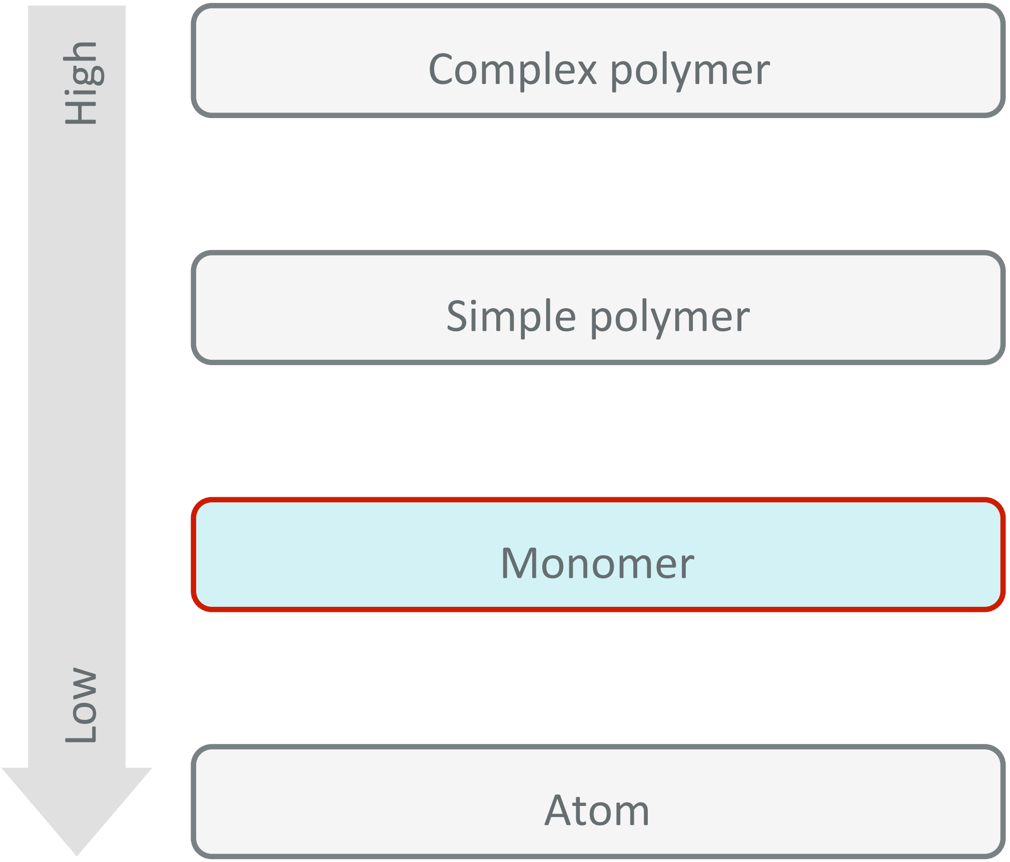
- Higher level components are a combination of lower level components



- Molecules described by a HELM notation consist of atoms and bonds

Structure hierarchy

- Higher level components are a combination of lower level components



High

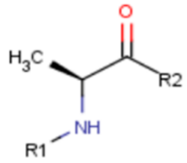
Low

Complex polymer

Simple polymer

Monomer

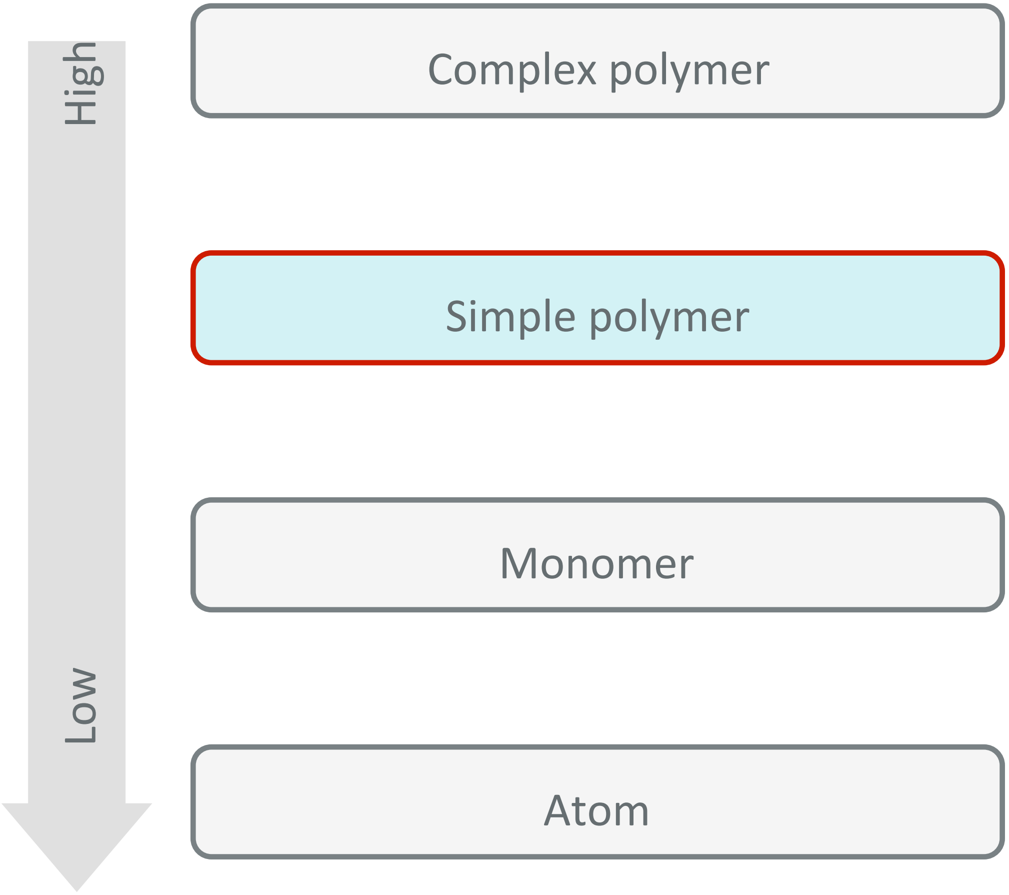
Atom

Structure	
SMILES	C[C@H](N[*])C([*])=O |r,$;;;_R1;;_R2;$|
ID	A
Attachment Points	R1-H R2-OH
Natural Analog	A
Polymer Type	PEPTIDE
Monomer Type	Backbone
Name	L-Alanine

- Each monomer is given a unique ID, and is backed by its structure and attachment points

Structure hierarchy

- Higher level components are a combination of lower level components

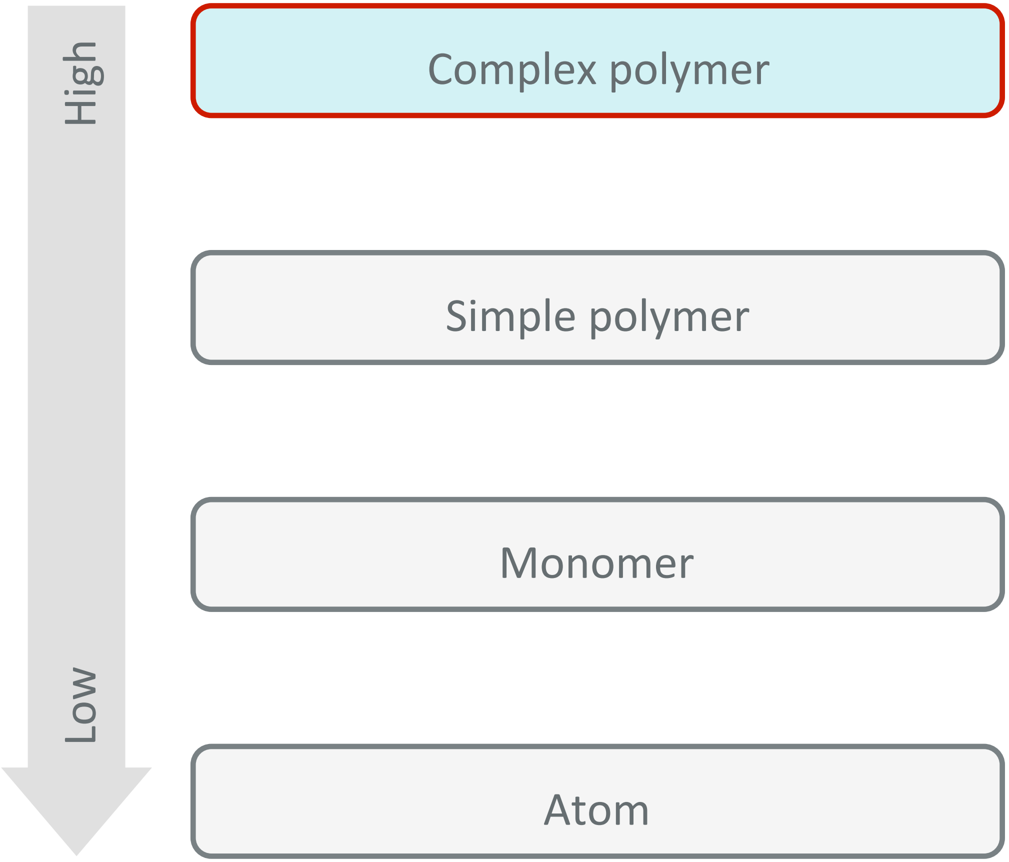


- Linear chains of monomers for a single polymer type (e.g. peptide chain, singular nucleic acid strand)

Type	Monomer (unit)
Peptide	A - Alanine
Nucleic acid	R(A)P R – Ribose A – Adenine P – Phosphate
Chem	[PEG3] – Pegylation

Structure hierarchy

- Higher level components are a combination of lower level components



High

Complex polymer

Simple polymer

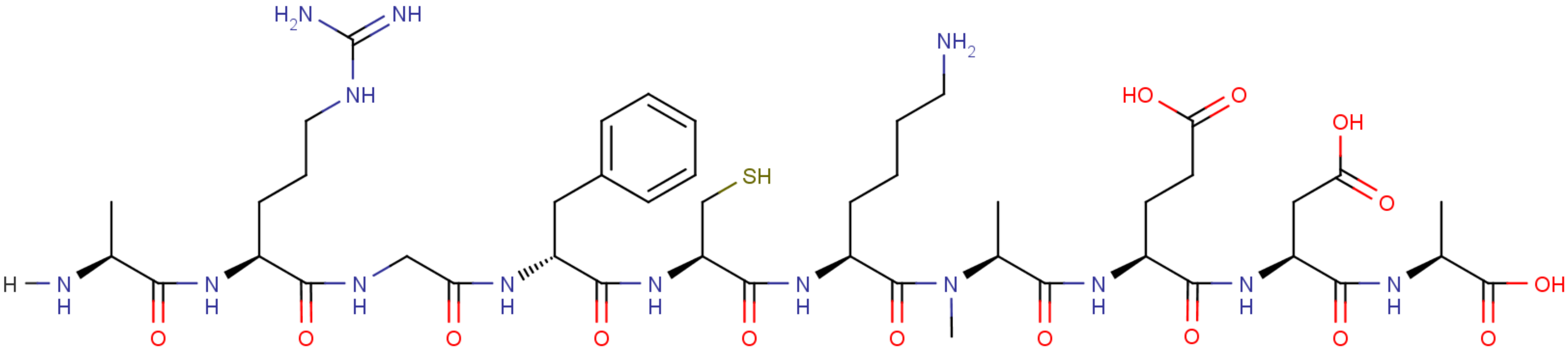
Monomer

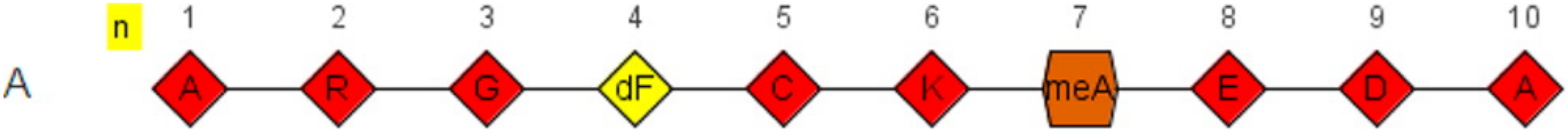
Low

Atom

- Entire chemical structure information of the macromolecule
- List of simple polymers
- List of connections
- List of hydrogen bonds
- List of annotations

Linear Peptide

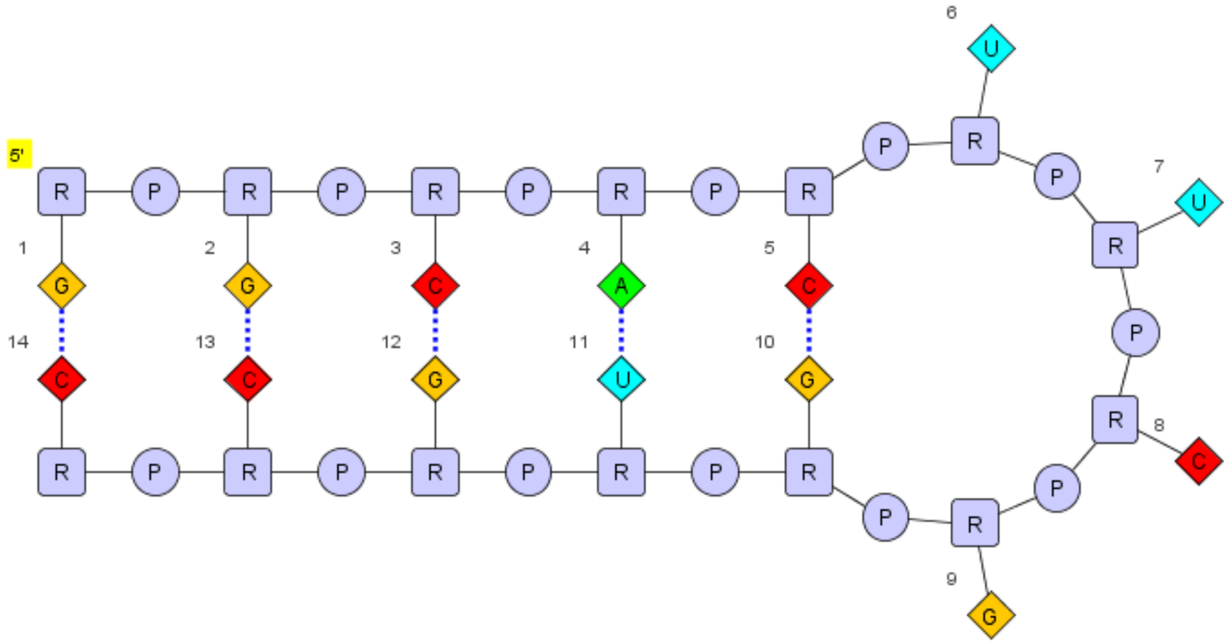




B PEPTIDE1{A.R.G.[dF].C.K.[meA].E.D.A}$$$$

A. Monomer Graph View B. HELM Notation

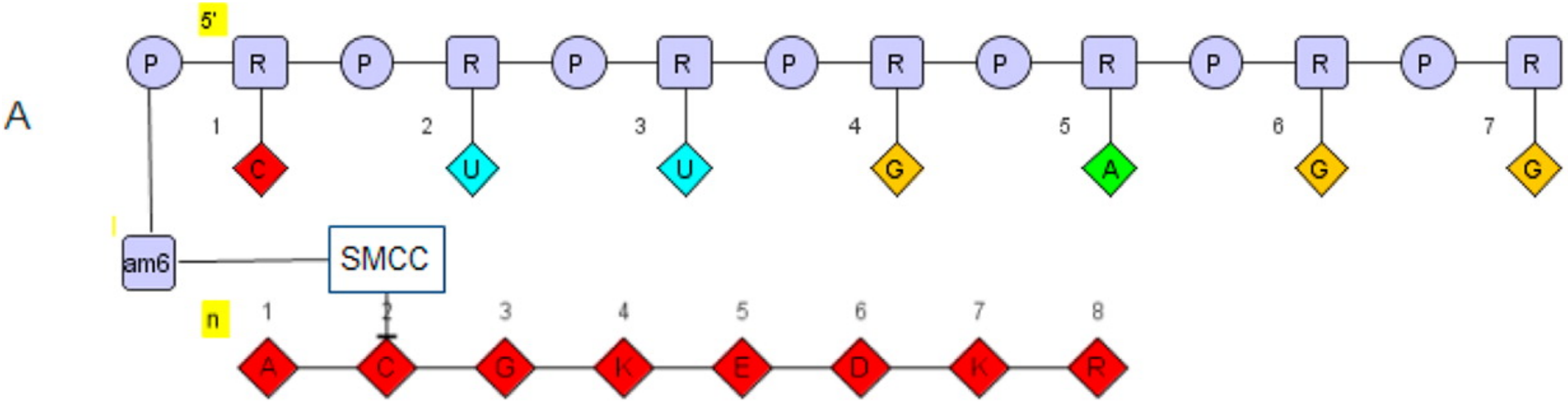
Short Hairpin RNA



HELM notation:

```
RNA1{R(G)P.R(G)P.R(C)P.R(A)P.R(C)P.R(U)P.R(U)P.R(C)P.R(G)P.R(G)P.R(U)P.R(G)P.R(C)P.
R(C)}$$RNA1,RNA1,11:pair-32:pair|RNA1,RNA1,5:pair-38:pair|
RNA1,RNA1,14:pair-29:pair|RNA1,RNA1,8:pair-35:pair|RNA1,RNA1,2:pair-41:pair$$
```

Oligonucleotide Peptide Conjugate



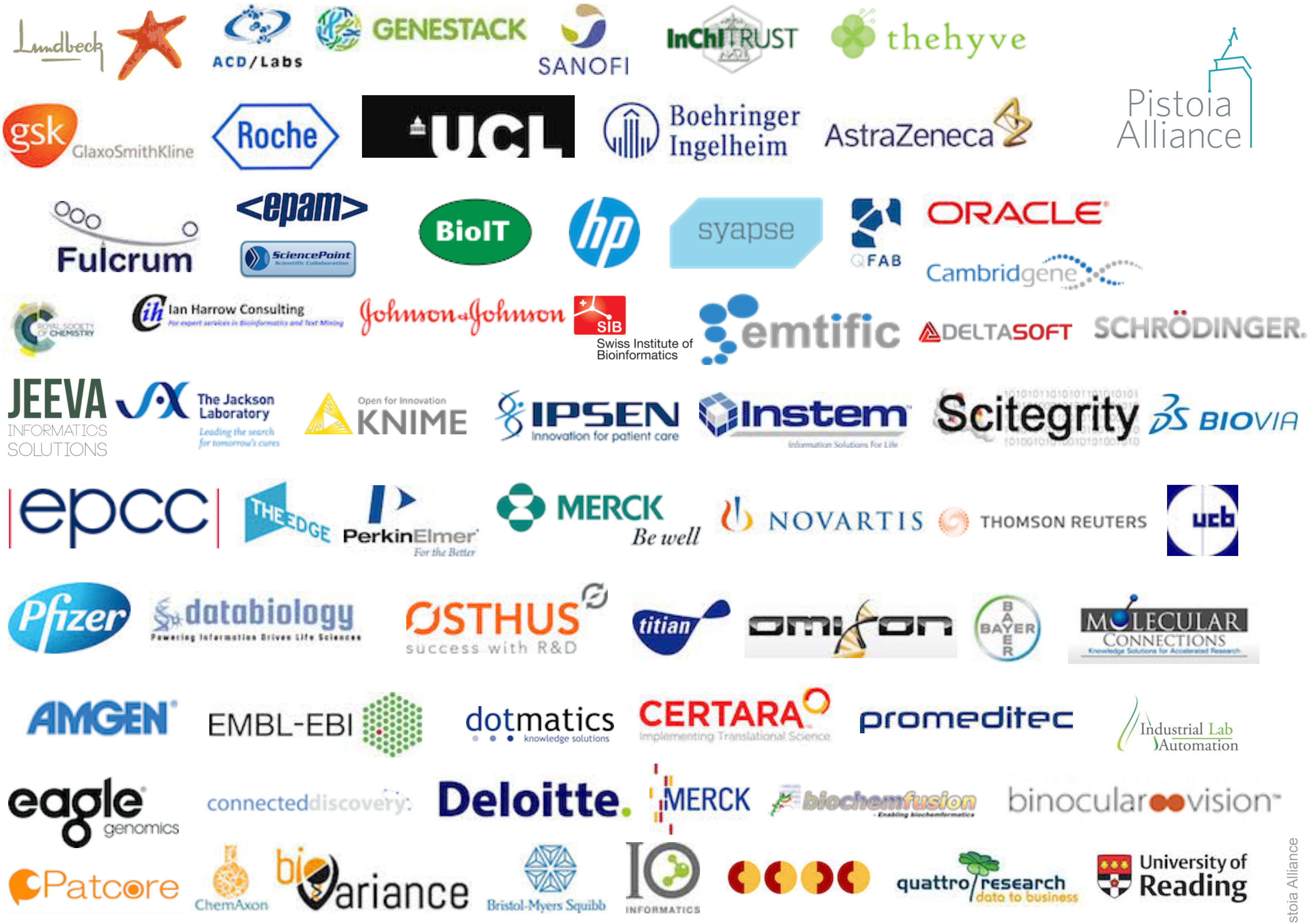
B RNA1{[am6]P.R(C)P.R(U)P.R(U)P.R(G)P.R(A)P.R(G)P.R(G)}|PEPTIDE1{A.C.G.K.E.D.K.R}|CHEM1{SMCC}$PEPTIDE1,CHEM1,2:R3-1:R2|RNA1,CHEM1,1:R1-1:R1$$$

A. Monomer Graph View B. HELM Notation

Pistoia Alliance HELM Project

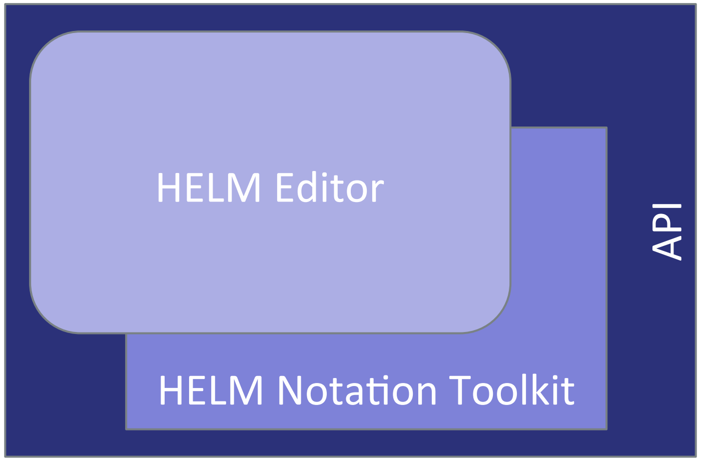
- The Pistoia Alliance is a global, non-profit alliance of life science companies, vendors, publishers, and academic groups that work together to solve common problems and lower barriers to innovation in R&D (http://www.pistoiaalliance.org)

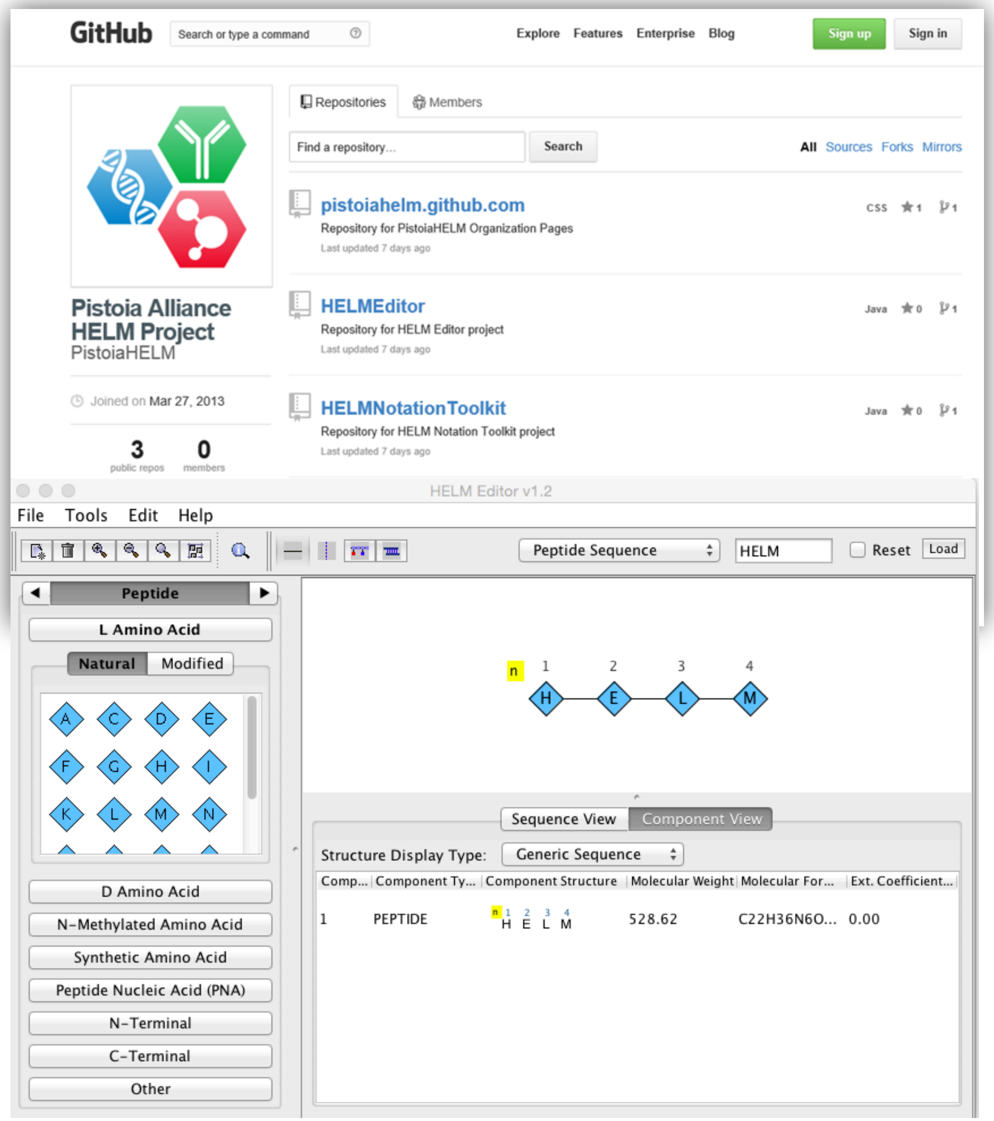
- Transition HELM technology from Pfizer proprietary to Open Source (http://www.openhelm.org)
  - Provide an industry-wide standard for data exchange within and between organizations
  - Reduce software development costs by minimizing the need for companies to develop similar functionality



Pistoia Alliance

Open Source





Pistoia Alliance HELM Project
PistoiaHELM

Joined on Mar 27, 2013

3 public repos   0 members

pistoiahelm.github.com — CSS ★1 1
Repository for PistoiaHELM Organization Pages
Last updated 7 days ago

HELMEditor — Java ★0 1
Repository for HELM Editor project
Last updated 7 days ago

HELMNotationToolkit — Java ★0 1
Repository for HELM Notation Toolkit project
Last updated 7 days ago

HELM Editor v1.2

Structure Display Type: Generic Sequence

Comp...	Component Ty...	Component Structure	Molecular Weight	Molecular For...	Ext. Coefficient...
1	PEPTIDE	H E L M	528.62	C22H36N6O...	0.00

- MIT license
- Source Code
- Binary Distribution
- Editor Demo

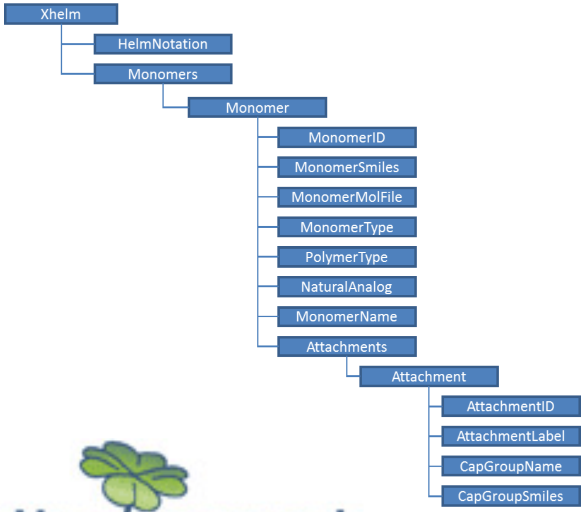
https://github.com/PistoiaHELM

HELM Extension (1.1)

- In-line HELM
  - Enables the incorporation of ad-hoc monomer into HELM notation

```
PEPTIDE1{G.[[*]N[C@@H](C=O)C([*])=O |$_R1;;;;;;_R2;$|].C.D.E.H}$$$$
```

- Exchangeable HELM
  - Enables the exchange of biomolecule structure across organizations without sharing monomer database



Xhelm
HelmNotation
Monomers
Monomer
MonomerID
MonomerSmiles
MonomerMolFile
MonomerType
PolymerType
NaturalAnalog
MonomerName
Attachments
Attachment
AttachmentID
AttachmentLabel
CapGroupName
CapGroupSmiles

```
<?xml version="1.0" encoding="UTF-8"?>
<Xhelm>
  <HelmNotation>PEPTIDE1{H.[dE].[dL].M}$$$$</HelmNotation>
  <Monomers>
    <Monomer>
      <MonomerID>M</MonomerID>
      <MonomerSmiles>CSCC[C@H](N[*])C([*])=O |$;;;;;;_R1;;_R2;$|</MonomerSmiles>
      ....
      ....
    </Monomer>
</Monomers>
</Xhelm>
```

Beyond Specific Structures

- While HELM 1.x enables the unambiguous representation of complex polymeric structures where all monomers, simple polymers and connections are known, there are many examples where at least one of the structure features is not fully known.
  - ADC from random lysine conjugation

- There is a need to extend HELM for ambiguous polymeric structures

- Furthermore, there is a desire to enhance HELM's annotation capability.

HELM 2 (Draft): Ambiguity

- Monomer
  - Unknown monomer
    - Single: X for Peptide, N for DNA/RNA
    - Multiple: * for 0-n monomers
  - Missing monomer: _
- Simple Polymer
  - Unknown polymer type: BLOB#{PolymerType}
  - Unknown structure: SimplePolymeID{*}
- Connection
  - Unknown monomer position: MonomerIDs or *
  - Unknown attachment point: *
- List (Monomer/Simple Polymer)
  - OR relation with optional probability:
    - (Element1:Prob1,Element2:Prob2)
  - AND relation with optional ratio:
    - (Element1:Ratio1+Element2:Ratio2)
  - Monomer group: implicit (inline)
  - Simple polymer group: explicit (predefined) in section 3, G1,G2…

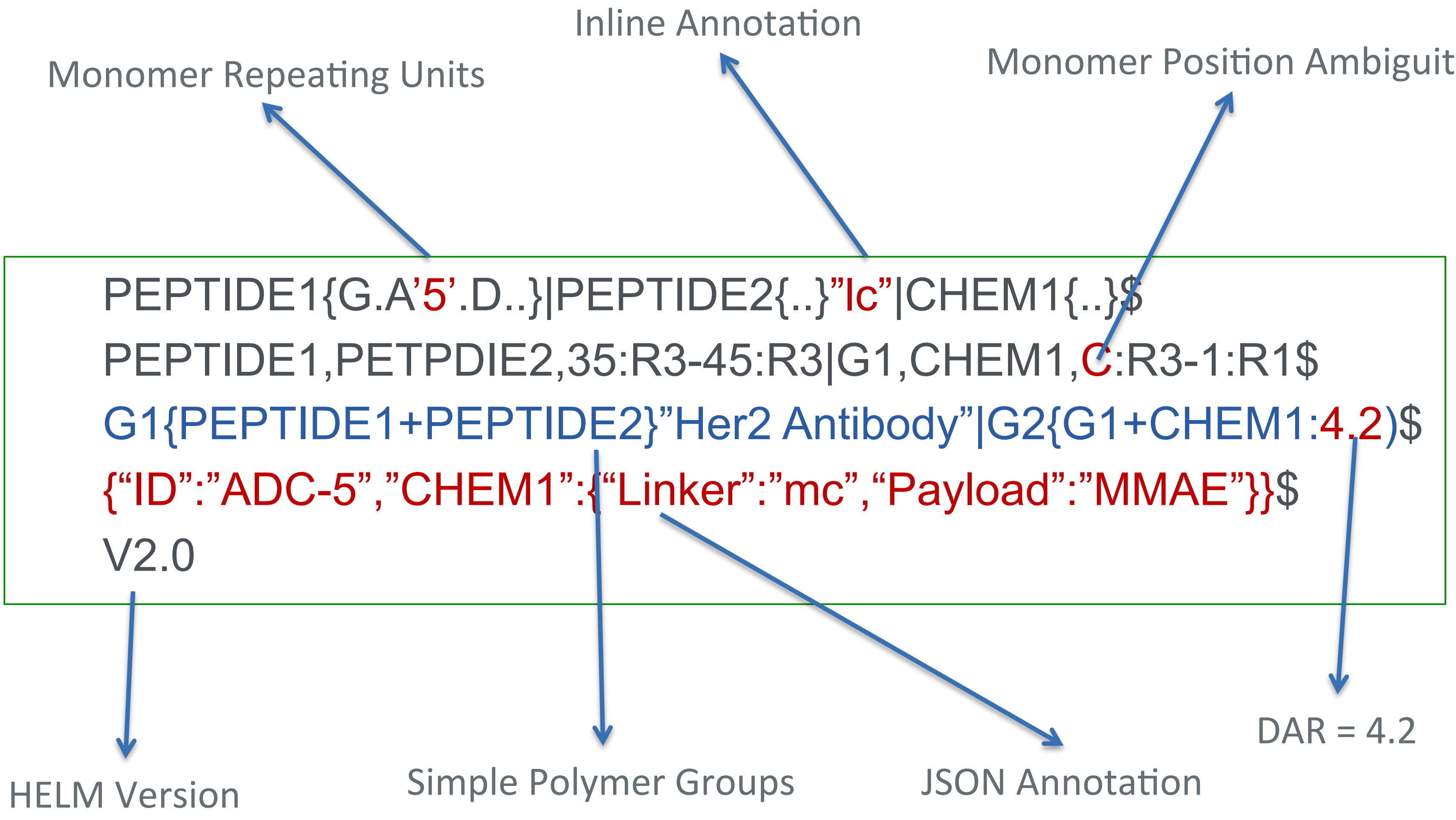
HELM 2 (Draft): Annotation

- Inline Annotation: Double Quoted Text “Annotation”
  - Applicable after Monomer, Simple Polymer, Grouped Polymer and Connection
  - A.G.C”mutation of A”.E => C is a mutation of A
  - PEPTIDE1{*}”IL6”   => PEPTIDE1 has no sequence, but has the name of IL6

- Monomer Repeating Units: Single Quoted Number ‘RepeatingUnits’
  - Apply after Monomer
  - D.R.E.A’5’ == D.R.E.A.A.A.A.A

- Extended Annotation: Section 4, JSON format
  - Can reference elements in HELM structure hierarchy
  - PEPTIDE1{A.G.C.D.E.F}$$${“PEPTIDE1”:{“target”:”jak3”}}$

HELM 2 (Draft): ADC Example



Inline Annotation

Monomer Repeating Units

Monomer Position Ambiguity

```
PEPTIDE1{G.A’5’.D..}|PEPTIDE2{..}”lc”|CHEM1{..}$
PEPTIDE1,PETPDIE2,35:R3-45:R3|G1,CHEM1,C:R3-1:R1$
G1{PEPTIDE1+PEPTIDE2}”Her2 Antibody”|G2{G1+CHEM1:4.2)$
{“ID”:”ADC-5”,”CHEM1”:{“Linker”:”mc”,“Payload”:”MMAE”}}$
V2.0
```

DAR = 4.2

HELM Version

Simple Polymer Groups

JSON Annotation

The HELM Ecosystem

- Pharma / Biotech / Institutes
  - BMS, GSK, Lundbeck, Merck, Novartis, Pfizer, Roche
- Software vendors
  - ACD/Labs, Arxspan, Biochemfusion, BioMax, Biovia, ChemAxon, NextMove, Scilligence
- Content / Service Providers
  - EBI (ChEMBL), eMolecules, quattro
- Active discussions on-going with others
  - e.g. FDA

Acknowledgements

Pfizer Colleagues

- Peter Henstock
- David Klatte
- Christine Lawrence
- Frank Loganzo
- Hongli Li
- Sergio Rotstein
- Simone Sciabola
- Rob Stanton
- Nathan Tumey
- Simon Xi
- Tianhong Zhang

The Pistoia Alliance HELM Project Team, especially:

- Sergio Rotstein (Pfizer) – Domain Lead
- Claire Bellamy (Pistoia Alliance) – Project Manager

Active Team Members:

- Roland Knispel (ChemAxon)
- Matthias Nolte (BMS)
- Jan Holst Jensen (Chembiofusion)
- Thomas Gan (Merck)
- Stefan Klostermann (Roche)
- Sven Neumeyer (Novartis)
- Yohann Potier (Novartis)
- Tianhong Zhang (Pfizer)

Steering Committee Members:

- John Wise (Pistoia Alliance)
- Margret Assfalg (Roche)
- Leah O'Brien (GSK)
- Ramesh Durvasula (BMS)
- Sergio Rotstein (Pfizer)
- Alex Drijver (ChemAxon)
- Chris Waller (Merck)
- Quan Yang (Novartis)

Pistoia Alliance

www.OpenHelm.org
info@openhelm.org

A Pistoia Alliance Innovation

HELM

Hierarchical Editing Language for Macromolecules