HELM Specification

"The limits of my language mean the limits of my world." Ludwig Wittgenstein

1 Contents

2 USE OF SPECIFICATION - TERMS, CONDITIONS & NOTICES 4
2.1 License Grant 4
2.2 NO WARRANTIES 4
2.3 THIRD PARTY RIGHTS 4
2.4 Ownership 5
2.5 Issue reporting 5
3 Introduction 6
3.1 Audience 6
3.2 Scope 6
4 HELM Concepts 7
4.1 General Notation Characteristics 7
5 HELM Component Specification 8
5.1 Monomer 8
5.1.1 Structure 8
5.1.2 ID 8
5.1.3 Name 9
5.1.4 Attachment Points 9
5.1.5 Natural Analog 10
5.1.6 Polymer Type 10
5.1.7 Monomer Type 10
5.1.8 Monomer Ambiguity 11
5.2 Simple Polymer 11
5.2.1 Specific Simple Polymer 12
5.2.2 Non-specific Simple Polymer 13
5.2.3 Unknown Polymer 13
5.2.4 Simple polymer notation 13
5.2.5 Monomer annotation 14
5.3 Complex Polymer 14
5.3.1 ListOfSimplePolymers$ 14
5.3.2 ListOfConnections$ 15
5.3.3 ListOfPolymerGroups$ 16
5.3.4 ExtendedAnnotation$ 16
5.4 HELM version 16
5.5 Reserved characters 17
5.6 In-line HELM notation 18
6 File Formats 18

6.1 HELM File Extensions 19
6.2 Exchangeable HELM 19
6.2.1 Description of XML tags 20
7 Appendix 1: Ambiguous HELM Quick Reference 22
8 Appendix 2: Example Monomer Definitions 23
8.1 PEPTIDE Monomers 23
8.2 Nucleotide Monomers 23
8.3 CHEM Monomers 24
9 Appendix 3: HELM examples 25
9.1 HELM example with in-line notation 28
9.2 HELM 2.0 examples 28
10 Appendix 4: XHELM examples 29

Version History:

Version	Release Date	Author	History
1	20 – Sept -2013	Claire Bellamy, Rama Bhamidipati, Stefan Klostermann, Roland Knispel, Matthias Nolte, Akos Papp, Tianhong Zhang	First released version.
1.1	15 – May-2014	Markus Weisser	Added In-line notation and exchangeable HELM notation.
2.0	4-April-2016	Markus Weisser	Added ambiguity notation
2.01	24-May-2016	Claire Bellamy, Markus Weisser and Tianhong Zhang.	Added atom-mapped SMILES, clarified the process of encoding molfiles, clarified numbering where there are ambiguous or repeating monomers and clarified the correct use of .helm files. Other minor corrections.
2.02	1- Jan-2016	Claire Bellamy	Changed definition of HELM to be case insensitive.

2 USE OF SPECIFICATION - TERMS, CONDITIONS & NOTICES

The material in this document details a Pistoia Alliance specification in accordance with the terms, conditions and notices set forth below. This document does not represent a commitment to implement any portion of this specification in any company's products. The information contained in this document is subject to change without notice.

2.1 License Grant

Licensor hereby grants you the right, without charge, on a perpetual, non-exclusive and worldwide basis, to utilize the Specification for the purpose of developing, making, having made, using, marketing, importing, offering to sell or license, and selling or licensing, and to otherwise distribute products complying with the Specification, in all cases subject to the conditions set forth in this Agreement and any relevant patent and other intellectual property rights of third parties (which may include members of Licensor). This license grant does not include the right to sublicense, modify or create derivative works based upon the Specification. For the avoidance of doubt, products implementing this Specification are not deemed to be derivative works of the Specification.

2.2 NO WARRANTIES

THE SPECIFICATION IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, ACCURACY, COMPLETENESS AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL LICENSOR, ITS MEMBERS OR ITS CONTRIBUTORS BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THE SPECIFICATION.

2.3 THIRD PARTY RIGHTS

WITHOUT LIMITING THE GENERALITY OF SECTION“NO WARRANTIES” ABOVE, LICENSOR ASSUMES NO RESPONSIBILITY TO COMPILE, CONFIRM, UPDATE OR MAKE PUBLIC ANY THIRD PARTY ASSERTIONS OF PATENT OR OTHER INTELLECTUAL PROPERTY RIGHTS THAT MIGHT NOW OR IN THE FUTURE BE INFRINGED BY AN IMPLEMENTATION OF THE SPECIFICATION IN ITS CURRENT, OR IN ANY FUTURE FORM. IF ANY SUCH RIGHTS ARE DESCRIBED ON THE SPECIFICATION, LICENSOR TAKES NO POSITION AS TO THE VALIDITY OR INVALIDITY OF SUCH ASSERTIONS, OR THAT ALL SUCH ASSERTIONS THAT HAVE OR MAY BE MADE ARE SO LISTED.

2.4 Ownership

This specification is the property of the Pistoia Alliance. The Pistoia Alliance is a global, not-for-profit, precompetitive alliance of life science companies, vendors, publishers, and academic groups that aims to lower barriers to innovation by improving the interoperability of R&D business processes. The Pistoia Alliance HELM project has developed this specification through the efforts of the member companies involved.

More information about the Pistoia Alliance can be obtained at http://www.pistoiaalliance.org/

More information about HELM can be obtained at http://www.OpenHELM.org.

Future updates to this specification will be published on www.OpenHELM.org.

2.5 Issue reporting

The reader is encouraged to report any technical or editing issues/problems with this specification to info@OpenHELM.org or via the HELM website http://www.OpenHELM.org.

Enhancement requests may be made in the same way. The technical and scientific suitability of an enhancement request will be reviewed by the HELM notation panel and suitable requests will be implemented as resources allow.

3 Introduction

HELM (Hierarchical Editing Language for Macromolecules) enables the representation of a wide range of biomolecules (e.g. proteins, nucleotides, antibody drug conjugates) whose size and complexity render existing small-molecule and sequence-based informatics methodologies impractical or unusable.

HELM is a hierarchical notation that represents complex macromolecules as polymeric structures, termed complex polymers, which are built from simple polymers comprised of predefined monomers. HELM supports unnatural components (e.g. unnatural amino acids) and chemical modifications as well as conventional natural components.

This document is the formal definition of the notation and will be used to verify the conformance of implementations of HELM in software.

For further details of how HELM came to be created and the software that is associated with it see:

> Zhang, T., et. al., (2012), 'HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation', *J. Chem. Inf. Model*., vol 52,pp 2796–2806
>
> The HELM website http://www.openhelm.org

3.1 Audience

This specification is intended for cheminformaticians, bioinformaticians, and developers that require precise definition of HELM to build software tools. It assumes some familiarity with basic chemical concepts such as atoms, different types of bonds and common biomolecules such as peptides and nucleotides.

3.2 Scope

This document defines the HELM notation; it is the authoritative definition of the HELM concept and contains examples of the HELM notation for biological and chemical molecules to illustrate the concise and coherent grammatical framework.
It does not contain any information about how HELM can be used in software tools and is not a descriptive introduction to HELM.

4 HELM Concepts

HELM is a hierarchical representation of components at four levels: Complex Polymer, Simple Polymer, Monomer, and Atom.

-------------------------Increasing granularity------------------------->

Complex Polymer <-> Simple Polymer <-> Monomer <-> Atom

<-----------------------Increasing abstraction----------------------------

The lower levels are familiar to anyone involved with chemical or biological sciences. A monomer is simply a small molecule which comprises a number of atoms connected together by bonds. The atom-bond representation of the monomer does not form part of a HELM string and so monomers (in theory) can be stored as any available form e.g. SMILES with atom-maping CXSMILES or Molfile.

HELM requires that whichever chemical structure representation is used at the monomer level supports the unambiguous definition of the monomer's attachment points. Simple polymers are made up from monomers that are of a single type e.g. PEPTIDE, RNA (includes DNA), or CHEM (chemical compounds). Each type has its own rules governing how monomers are linked.

Complex polymers are made up from simple polymers of same or different types. Thus chemical modifiers can be added to peptides or any other combination can be realized. The HELM notation specifies the simple polymer to be added and the location of the attachment point. In this way any complex polymer can be built up easily in any way required by the science.

There are a number of cases in which many structural features of a biomolecule are not known. HELM can cover this by also represent ambiguous macromolecules. This ambiguity can be on all four levels: monomers, simple polymer, connections and grouping.

4.1 General Notation Characteristics

The HELM notation is a line notation, which is case insensitive. It contains no whitespace between the notation elements. Whitespace can be used in names or annotations, however.

Float values within HELM notation are represented with the "." as their decimal separator.

5 HELM Component Specification

5.1 Monomer

A Monomer is comprised of atoms and bonds, and can be represented by a known chemical structure format such as SMILES with atom-mapping, CXSMILES or Molfile. Each monomer belongs to a polymer type, and has to have a unique ID in that polymer type.

Monomers have the following set of properties:

- structure
- name
- ID
- attachment points (primary and optional)
- natural analog
- polymer type
- monomer type

Example monomers of the three main polymer types can be found in Appendix 2.

New implementations can use the monomers in the reference implementation as their initial set. The full set can be found in:

https://github.com/PistoiaHELM/HELMNotationToolkit/blob/master/source/org/helm/notation/resources/MonomerDBGZEncoded.xml

The *exchangeable HELM format* (see section 6) should be used when information is transferred between organizations.

5.1.1 Structure

The atom-bond and connection point representation of the monomer can be stored as a SMILES with atom-mapping, CXSMILES, or Molfile.

5.1.2 ID

The monomer ID is a symbol of one or more characters which must be unique within a given polymer type. Whitespace characters in monomer ID are not supported.

Although the characters used for the monomer ID are not defined within the HELM specification, HELM recommends that the IDs in the reference set or specified in the recommendations on the HELM website are used as far as possible. Where a monomer is not present in the reference set, abbreviations in common use are preferred.

5.1.3 Name

The name of the monomer is a more complete description of the component unlike its corresponding shorthand ID. For example ‘Cysteine’ = Name and ‘C’ = ID.

5.1.4 Attachment Points

An attachment point is a specified location on a monomer where one monomer can be linked to another. Attachment points are specified using R groups, each of which needs to have a unique name (e.g. R1, R2, and R3) and a corresponding leaving group that is cleaved to free the attachment point for bonding. The number of attachment points on a monomer is the maximum number of bonds that a monomer can form with other components.

A leaving group is a chemical fragment, such as a hydroxyl group that is defined as part of the monomer if the attachment point is unused. For example, R1-OH represents R1 attachment with -OH as leaving group. It will be used to determine the exact mass of a complex polymer.

There are two attachment point subtypes:

5.1.4.1 Primary attachment points

Primary attachment points are attachment points on Backbone and Branch monomers within specific polymer types. There are recommendations for the assignment of R groups for monomers of a specific simple polymer type. For clarity the R group nomenclature is used below, however the same principle applies to atom maps.

Amino acids

> R1 replaces one of the hydrogens on the amino group and R2 to the hydroxyl of the carboxylic acid.

Nucleotides

> Base – R1 replaces the hydrogen of the amino group that is involved in the formation of the nucleoside.
>
> Sugar – R1 replaces the hydrogen on the 5’ hydroxyl group, R2 replaces the hydrogen on the 3’ hydroxyl group and R3 replaces the hydroxyl group involved in the formation of the nucleoside.
>
> Phosphate – R1 and R2 replace 2 hydroxyl groups on the phosphate.

Assigning the attachment points in this order results in simple polymers where peptides are written from N terminus to C terminus (left to right) and nucleic acids from 5’ to 3’, in accordance with standard conventions.

5.1.4.2 Optional attachment points

Optional attachment points are all attachment points whose connection rules to other monomers are not predefined. It includes all attachment points on monomers with undefined monomer type (see section 5.1.7.3), and those attachments on backbone and branch monomers outside the primary attachment points defined above.

5.1.4.3 Attachment point technology specific definition

There are a variety of methods that could be used to define an attachment point. The recommendation from the project team is that the following methods are used.

- Molfiles: R groups
- SMILES: atom maps

5.1.5 Natural Analog

A natural analog may be assigned to a monomer. A natural analogue is used to generate the single letter sequence equivalent for the macromolecule.

For example, the non-natural amino acid selenocysteine can have the ID ‘Sec’ and the natural analog ‘C’ (Cysteine).

5.1.6 Polymer Type

Each monomer must belong to a single “simple polymer” type. Allowed polymer types are peptide (PEPTIDE), nucleotide (RNA), and chemical (CHEM) at this time, but additional polymer types such as saccharide could be added in the future. The polymer type specifies the monomer subset, and dictates how the primary attachment points of the monomers are used.

5.1.7 Monomer Type

The type of the monomer specifies its place in the polymer skeleton. There are three types in HELM: ‘backbone’ (the main repetitive elements of a sequence), ‘branch’ (linked to the backbone monomers, but essential to define the biological context, where it is not given by the backbone monomers alone), and ‘undefined’ (used for monomers, which are usually not part of a sequence).

5.1.7.1 Backbone

A backbone monomer is included in the main linear polymer chain. Backbone monomers generally have two attachment points ‘R1’ and ‘R2’ that form the main chain of the simple polymer. For example, in RNA polymers, sugar and phosphate linkers are of ‘backbone’ monomer type. In peptides all amino acids are of backbone monomer type.

Backbone monomers may have a single R group when they act as terminal monomers e.g. N-methylated amino acids.

Backbone monomers may have one or more branch attachment points, for example ‘R3’ to which the ‘R1’ of a ‘branch’ monomer can be connected.

5.1.7.2 Branch

Branch monomers are monomers that do not form part of the main polymeric chain. For example, in RNA polymers the nucleobase is a ‘branch’ monomer type.

5.1.7.3 Undefined

All other monomers are classified as undefined. There are no predefined rules for how attachment points are used for this type.

5.1.8 Monomer Ambiguity

Unknown monomers can represented by four different characters:

- *
  The character “*” represents 0..n unknown monomers.
- X
  The character “X” represents one single unknown amino acid in a PEPTIDE polymer.
- N
  The character “N” represents one single unknown base in a RNA polymer.
- _
  The character “_” represents a deleted or missing single monomer. This is typically used for list elements (see section List Elements)

5.2 Simple Polymer

A simple polymer is comprised of one or more monomers of the same polymer type. By definition, simple polymers are linear chains.

Branching in the simple polymer repeating unit e.g. nucleobases in nucleotides is handled by the simple polymer level – this type of branch is a monomer branch.

Bonds between simple polymers are defined in the complex polymer level, therefore cyclisation and chain branching are not handled by the simple polymer notation itself.

Simple Polymers can be divided into three categories:

- Specific
- Non-specific
- Unknown

5.2.1 Specific Simple Polymer

There are two types of specific simple polymers: PEPTIDE, and RNA. Each type has specific rules that define how the backbone and branch monomers are connected. Additional polymer types with different backbone chemistries can be added to the notation language, but have not been defined here.

5.2.1.1 Specific Simple Polymer Rules

All specific simple polymers are chains which are built by adding monomers to the right of the initial monomer. By convention ‘R1’ is the left attachment point and ‘R2’ the right. Adding a monomer to the right links ‘R2’ of the left-monomer to ‘R1’ of the right-monomer and thus establishes directionality. Directionality means AB is different from BA, AB has ‘R2’ of A linked to ‘R1’ of B whereas BA has ‘R2’ of B linked to ‘R1’ of A.

‘R1’ of the left most and ‘R2’ of the right most monomer maintain their leaving group.

Specific simple polymer notation always starts with a backbone monomer but can end with either a backbone monomer or a branch monomer.

The ID of a branch monomer should be enclosed in parentheses ‘()’ and placed to the immediate right of the backbone monomer to which it is attached. When a backbone monomer is connected to branch monomer as well as another backbone monomer to its right, in the notation, the right backbone monomer should be written to the right of the closed parenthesis of the branch monomer. For example, A(B)C represents ‘R3’ of backbone monomer ‘A’ linked to ‘R1’ of branch monomer ‘B’ and ‘R2’ of A linked to ‘R1’ of C.

Using the attachment point recommendation in section 5.1.4.1 and these connection rules results in simple polymers where peptides are written from N terminus to C terminus (left to right) and nucleic acids from 5’ to 3’, in accordance with standard conventions.

5.2.2 Non-specific Simple Polymer

There is one non-specific simple polymer type CHEM which is used to represent chemical structures which are not part of a specific simple polymer type.

Monomers of the CHEM polymer type are of monomer type “undefined”. In this type the connection rules are not defined. A CHEM simple polymer can contain only one monomer; to connect two undefined monomers each monomer needs to be defined as a CHEM simple polymer and their connection defined at the complex polymer level.

5.2.3 Unknown Polymer

Unknown Polymers are marked as BLOB type polymers. These polymers do not contain a list of monomers but they specify their type inside the curly braces. The polymer BLOB1{Bead} for example represents a polymer with the type “Bead”.

5.2.4 Simple polymer notation

In a simple polymer notation the sequence is defined using monomer IDs.

Monomers with multi-character IDs must be enclosed in square brackets “[]”.

“.” is used between connected monomer units which are groups that represent the repetitive functional unit of the given polymer type. E.g. the phosphate, sugar and nucleobase of RNA are a monomer unit.

The ID of a branch monomer should be enclosed in parentheses “()”.

5.2.4.1 Monomer List

Monomers can be put into a list that is grouped using parentheses “()” to represent a pseudo monomer.

“+” as the separator within this list represents an AND relationship of the monomers. All elements in this list are possible and thus form a mixture The ratio of each element can be given as a numerical value after the monomer separated by the colon character. The default value is 1.

“,” as the separator within this list represents an XOR (excluding OR) relationship of the monomers. Only one single element of the list is present, i.e. it is not a mixture. The probability of each element can be given as a numerical value after the monomer separated by the colon character.

5.2.4.2 Monomer repeating units

A single monomer or a group of monomers can be repeated multiple times. The repeating unit is marked as the repeat unit count enclosed in single quote characters immediately after the monomer or a group of monomers that are enclosed in brackets. The repeat unit count can be a single value or a range of repeats. The range is defined with a “-“ between the two numbers.

5.2.5 Monomer annotation

Inline annotations of monomers are marked with quotation marks “”. They are always located after the monomer, i.e. before the separator of the next monomer.

5.3 Complex Polymer

A Complex Polymer is comprised of Simple Polymers, which are optionally connected to each other. A branched or cyclic peptide is also defined as a Complex Polymer (where the e.g. the cyclization is defined as a connection between two monomers of the same Simple Polymer). Note that any Atom or single Monomer, which is meant to be part of a Complex Polymer is ultimately promoted to a Simple Polymer by applying the rules of the notation, even if that means, that the Simple Polymer consists of only a single Monomer/Atom.

The Complex Polymer Notation has 4 distinct sections, each of them terminated by a ‘**$**’ sign and concatenated as follows:
ListOfSimplePolymers**$**ListOfConnections**$**ListOfPolymerGroups**$**ExtendedAnnotation**$**

In HELM specification 1 and 1.1, the section number 3 was used for the list of hydrogen pairings instead of the polymer groups. The hydrogen pairings are now stored in the connection section (number 2) since hydrogen bonds are now recognized as a special form of connection.

5.3.1 ListOfSimplePolymers$

List of all Simple Polymer components with each component identified by Polymer ID, which is formed by its Polymer Type and an integer suffix to distinguish it from other members of the same polymer type within a given Complex Polymer. e.g. PEPTIDE1, PEPTIDE2.

Format: PolymerID{SimplePolymerNotation}

Curly brackets around are used around each simple polymer notation: “{}”

A pipe is used as the separator between simple polymers: “|”

Example:
PEPTIDE1{A.R.G.[dF].C.K.[ahA].E.D.A}|RNA1{R(A)P.[mR](U)[sP].R(G)P.R([5meC])P.[dR](T)P.[dR](T)P.[dR](T)P}|CHEM1{SS3}

5.3.2 ListOfConnections$

List of connected monomer-pairs with 3 parts to each connection description. Cyclic polymers are represented using the same polymerID for source and target.

Format: SourcePolymerID,TargetPolymerID,SourceMonomerPosition:SourceAttachmentPoint-TargetMonomerPosition:TargetAttachmentPoint

- MonomerPosition is the position of the monomer (not monomer unit) in the simple polymer as counted from left to right. Where there are ambiguous elements or repeating units, use the literal monomer index. In other words, each ambiguous monomer or monomer with repeating units will be counted as a single monomer for the monomer position.
- AttachmentPoint is the attachment point of the monomer referred by R# (with # being an integer)
- Separator between monomer position and attachment point: “:”
- Connection indicator: “-“
- Separator between each of the 3 parts of connection description: “,”
- Separator between each connection description: “|”

Example: **RNA1,CHEM1,21:R2-1:R1** where R2 attachment point of 21st monomer in RNA1 and R1 attachment point of 1st monomer in CHEM1 are connected.

If the exact connection is unknown, it can be described with the a monomer ID instead of the exact monomer position. If this is also unknown, it can be described with a “?”. The “?” can also be used for unknown R groups.

Example: **PEPTIDE1,CHEM1,C:R3-1:?** where R3 attachment point of any cysteine in PEPTIDE1 and an unknown attachment point of 1st monomer in CHEM1 are connected.

5.3.2.1 Hydrogen pairings

The hydrogen pairings are defined as a special form of connection. It is not a bond between two atoms, but rather a set of hydrogen bonds between two monomers. Therefore, the representation uses the word 'pair' instead of the attachment point symbol 'R#'.

Example: RNA1,CHEM1,21:pair-1:pair

5.3.3 ListOfPolymerGroups$

A simple polymer group can contain two or more simple polymers and will be assigned a group ID such as G1, G2…, which can be referenced in HELM2 notation. Grouping annotation follows the same syntax as for monomer or polymer annotations.

Example of a mixture of four PEPTIDE polymers: **G1(PEPTIDE1+PEPTIDE2+PEPTIDE3+PEPTIDE4)**

The different elements of a group have a default ratio of 1:1:1:… Other ratios can be defined individually following the same syntax as for monomer mixtures.

Different groups are separated with the "|" symbol between each group.

A group can be an element of another group. In the following example, G1 is an element of G2.

G1(PEPTIDE1+PEPTIDE2)|G2(CHEM1+G1)

5.3.4 ExtendedAnnotation$

This section can be used for any additional annotation in valid JSON format.

Example: **{"PEPTIDE1":{ "ChainType":"hc"},"PEPITDE2":{"ChainType":"lc"}}**

5.3.4.1 Simple polymer annotations

Inline annotations of polymers are marked with quotation marks "". They are always located after the polymer, i.e. before the separator of the next polymer.

5.4 HELM version

HELM 2.0 represents a significant extension of the HELM notation. As such HELM strings which make use of the additional functionality such as ambiguity representation or annotations, should include the string

“V2.0” immediately after the fourth $. HELM strings that only contain functionality specified in HELM V1 or V1.1 will have nothing following the fourth $.

5.5 Reserved characters

The following symbols have special meaning in HELM, and are considered reserved characters.

Reserved characters should not be used in monomer IDs.

Dollar sign: “$”; to separate the major sections of the complex polymer notation

Curly brackets: “{‘, ‘}”; to enclose Simple Polymer notation in the Simple Polymer list.

Vertical pipe: “|”; to separate simple polymers, connections and groups within their respective sections of the complex polymer notation

Period: “.”;to separate monomers or monomer units in the simple polymer notation

Comma: “,”; to separate the three parts of a connection or hydrogen pair

Dash: “-“; to separate connection source and target, and to separate the min and max of repeating units.

Colon: “:”; to separate monomer position from attachment point or the key word pair designating hydrogen pair

Brackets: “[ ]”; to enclose multi-character monomer IDs

Parentheses: “( )”; to enclose branch monomer IDs

5.6 In-line HELM notation

Not all situations require the registration of all monomers of the polymer chains in local databases or files. Conditions might arise, where existing monomers are just temporarily modified and it is not intended to store this modified structure with a new ID and name.

In-line notation does not require monomer names or any other monomer information beyond the atom/bond definition. The unregistered monomers will be inserted in the HELM code as an extended SMILES string with RGroups in square brackets or SMILES with atom maps.

The extended SMILES (cxsmiles) format is a format specified by ChemAxon (https://www.chemaxon.com/marvin/help/formats/cxsmiles-doc.html) for storing special features of molecules within the SMILES string. The attachment point information is written explicitly as ANY atom in the SMILES sting and encoded with an alias in a semicolon separated list of elements of the SMILES notation.

Example of extended SMILES for in-line notation: [*]NCC([*])=O |$_R1;;;;_R2;$|
This monomer has two RGroups at positions 1 and 5 of the molecule. The RGroup description uses the notation convention for attachment points.

Example HELM notation of a peptide consisting of 3 monomers with one of them modified. The in-line notation is highlighted:

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

Atom mapped SMILES defines the attachment point information with asterisks and a numeric identifier:
PEPTIDE1{G.**[[*:1]N[C@@H](C=O)C([*:2])=O]**.C}$$$$

This format is widely used by open source toolkits such as CDK and is supported by many commercial software tools including Marvin.

The in-line notation can be used for all types of polymers.

6 File Formats

HELM is a line notation for a single macromolecule structure, and the project does not intend to define a proprietary file format for multiple structures with or without additional information. There are a large number of existing formats such as SDF, XML, CSV and JSON which can and should be used for that purpose. If SDF is used the project's recommendation is to add an "HELM" tag immediately after the MOLFILE section for each molecule, and MOLFILE could be empty if the atom/bond representation is unknown or unavailable. The guidelines published for the use of SDFile should be consulted and their recommendations should take priority.

6.1 HELM File Extensions

The HELM notation toolkit and editor supports the following HELM file extensions:

.helm for standard HELM

.chelm for canonical HELM

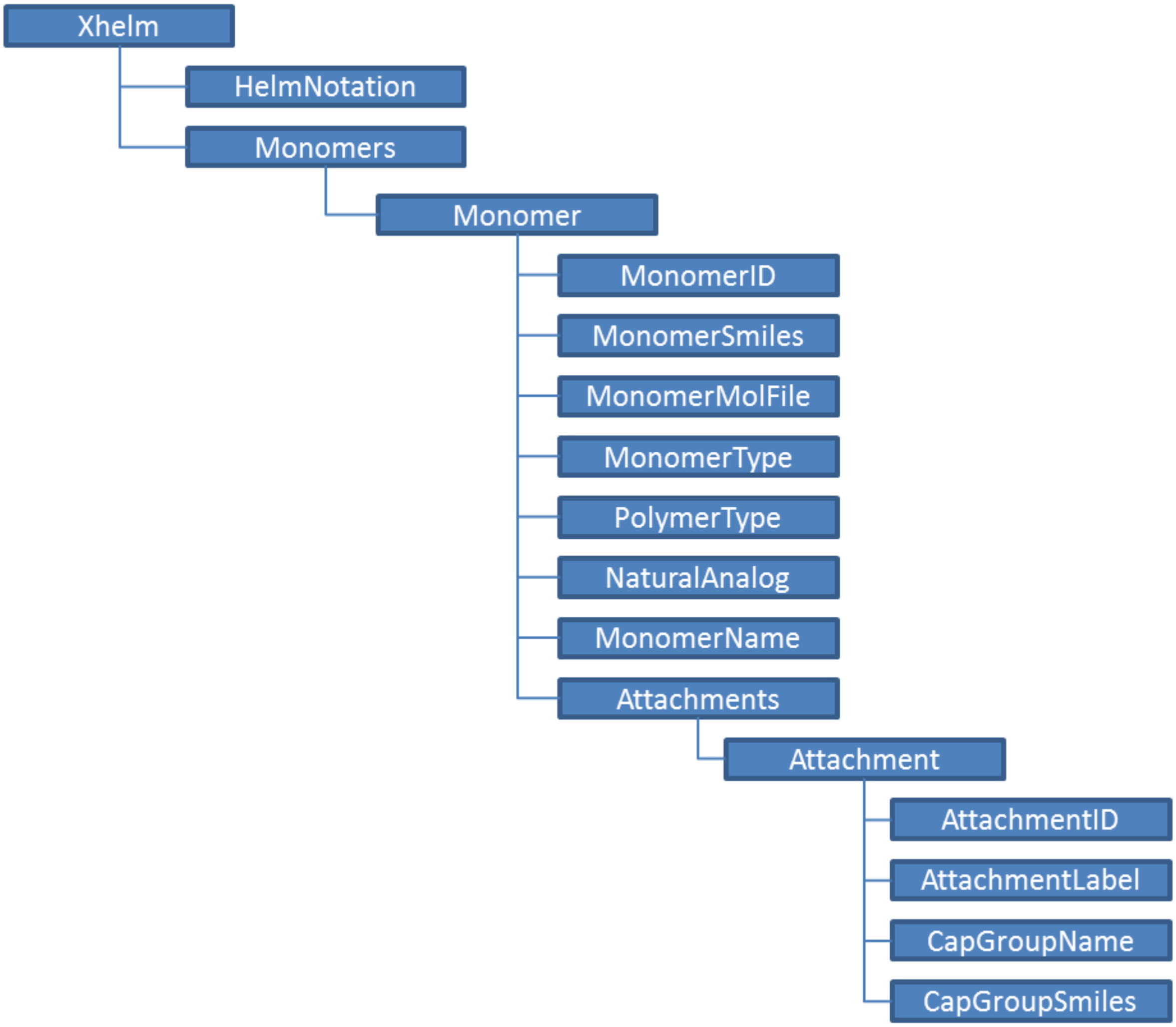
.xhelm for exchangeable HELM

all of which contain a single macromolecule structure. In other words, the entity of the file content represents a single HELM structure. We highly recommend that you don't use them for multiple HELM structures.

6.2 Exchangeable HELM

The purpose of exchangeable HELM is to allow organizations to exchange the complete and unequivocal structural definition of macromolecules in a single operation without requiring them to have a common set of monomer definitions. HELM itself relies on externally defined monomer definitions that are usually held within central databases within organizations.

The exchangeable HELM (XHELM) encapsulates the HELM notation in an XML document together with the complete list of monomers that are used in polymer definitions. The structure of the XML format is defined by an XML schema document (XHelmSchema.xsd).



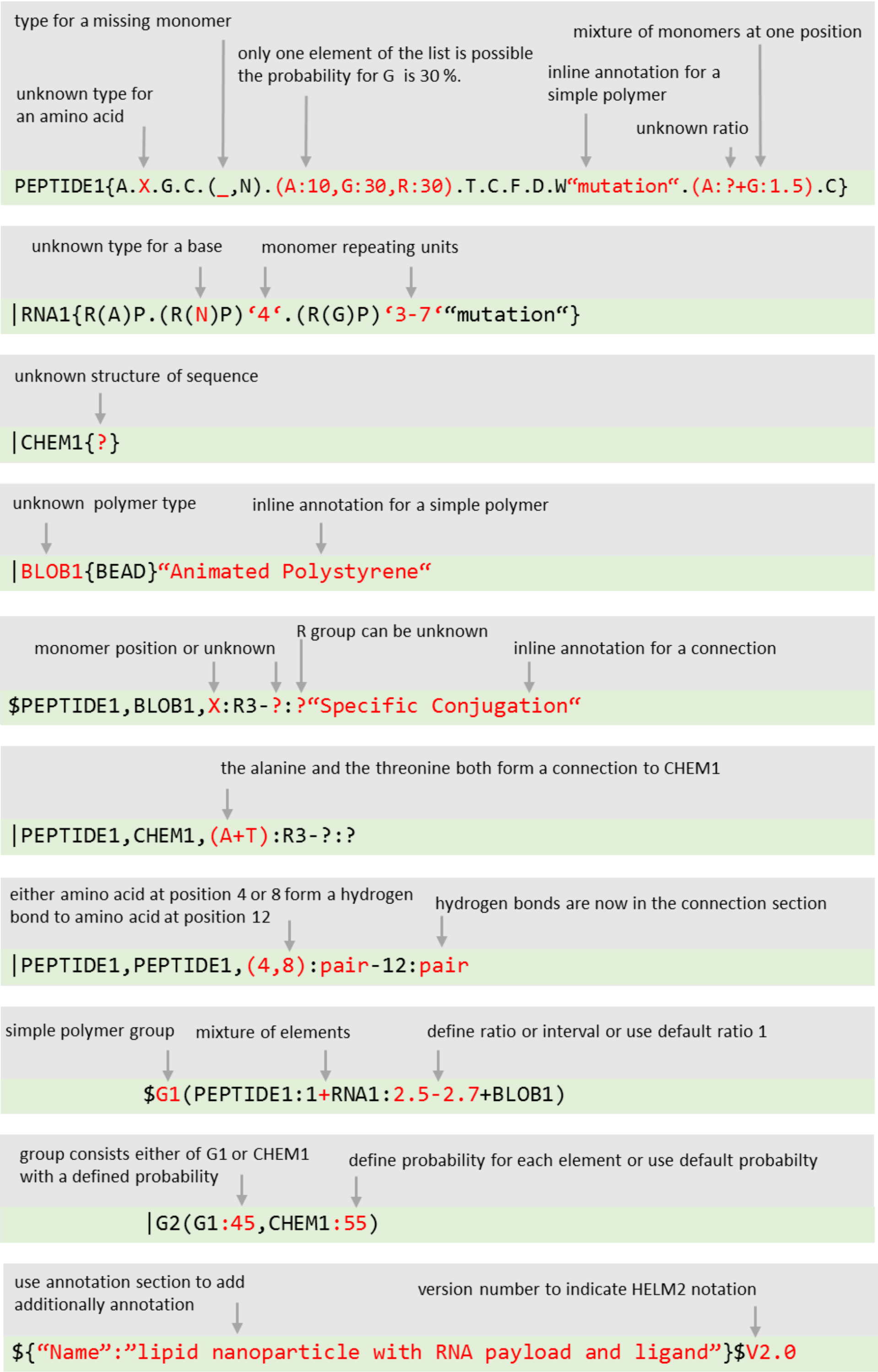
6.2.1 Description of XML tags

Xhelm	Root tag
HelmNotation	Contains the original HELM notation as it is described in chapter "HELM Component Specification"
Monomers	Collection of monomers that are used in the original HELM notation of this XHELM document
Monomer	Grouping tag of a single monomer
MonomerID	Unique ID of monomers that are listed in polymer description.
MonomerSmiles	SMILES used to define the structure of the monomer with attachment points.

	It may be an atom-mapped SMILES or the ChemAxon CXSMILES.
MonomerMolFile	gzipped then Base-64 encoded mol file representation of the monomer structure.
MonomerType	Monomer type (backbone, branch or unspecified)
PolymerType	Polymer type for each monomer (PEPTIDE, RNA, CHEM)
NaturalAnalog	Natural analog for generating single letter code in polymers
MonomerName	Name of the monomer
Attachments	Collection of attachment points on the monomers where each monomer can be linked to other monomers.
Attachment	Grouping tag of attachments
AttachmentID	Attachment identifier following the convention that is described in chapter "Primary attachment points", such as R1-H
AttachmentLabel	Label of attachment, such as R1
CapGroupName	Name of the capping group, such as H
CapGroupSmiles	Atom mapped SMILES or CXSMILES representation of the capping group.

Examples can be found in the appendix 4 section.

7 Appendix 1: Ambiguous HELM Quick Reference



type for a missing monomer

unknown type for an amino acid

only one element of the list is possible the probability for G is 30 %.

inline annotation for a simple polymer

unknown ratio

mixture of monomers at one position

```
PEPTIDE1{A.X.G.C.(_,N).(A:10,G:30,R:30).T.C.F.D.W“mutation“.(A:?+G:1.5).C}
```

unknown type for a base

monomer repeating units

```
|RNA1{R(A)P.(R(N)P)‘4‘.(R(G)P)‘3-7‘“mutation“}
```

unknown structure of sequence

```
|CHEM1{?}
```

unknown polymer type

inline annotation for a simple polymer

```
|BLOB1{BEAD}“Animated Polystyrene“
```

monomer position or unknown

R group can be unknown

inline annotation for a connection

```
$PEPTIDE1,BLOB1,X:R3-?:?“Specific Conjugation“
```

the alanine and the threonine both form a connection to CHEM1

```
|PEPTIDE1,CHEM1,(A+T):R3-?:?
```

either amino acid at position 4 or 8 form a hydrogen bond to amino acid at position 12

hydrogen bonds are now in the connection section

```
|PEPTIDE1,PEPTIDE1,(4,8):pair-12:pair
```

simple polymer group

mixture of elements

define ratio or interval or use default ratio 1

```
$G1(PEPTIDE1:1+RNA1:2.5-2.7+BLOB1)
```

group consists either of G1 or CHEM1 with a defined probability

define probability for each element or use default probabilty

```
|G2(G1:45,CHEM1:55)
```

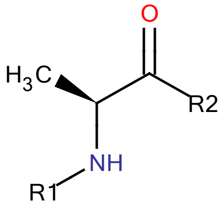
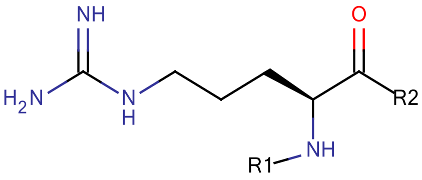
use annotation section to add additionally annotation

version number to indicate HELM2 notation

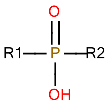
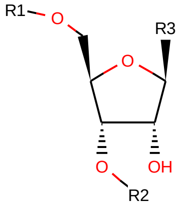
```
${“Name”:”lipid nanoparticle with RNA payload and ligand”}$V2.0
```

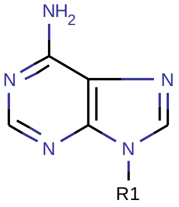
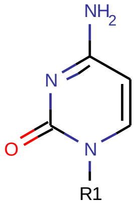
8 Appendix 2: Example Monomer Definitions

8.1 PEPTIDE Monomers

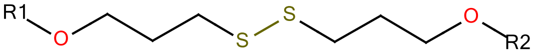
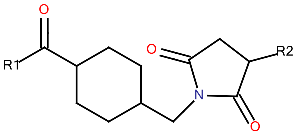
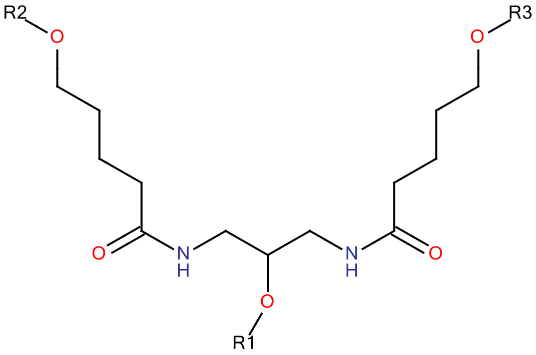
ID	Name	Structure	Attachment Points	Natural Analog
A	Alanine		R1-H R2-OH	A
R	Arginine		R1-H R2-OH	R

8.2 Nucleotide Monomers

ID	Name	Structure	Monomer Type	Attachment Points	Natural Analog
P	Phosphate		Backbone	R1-OH R2-OH	P
R	Ribose		Backbone	R1-H R2-H R3-H	R

A	Adenine		Branch	R1-H	A
C	Cytosine		Branch	R1-H	C

8.3 CHEM Monomers

ID	Name	Structure	Attachment Points
SS3	Dipropanol Disulfide		R1-H R2-H
SMCC	SMCC Linker		R1-OH R2-H
sDBL	Symmetric Doubler		R1-H R2-H R3-H

9 Appendix 3: HELM examples

Sample	Format	Notation
1	HELM	PEPTIDE1{A.R.G.[dF].C.K.[meA].E.D.A}$$$$
	SMILES	NCCCC[C@H](NC(=O)[C@H](CS)NC(=O)[C@@H](Cc1ccccc1)NC(=O)CNC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](C)N)C(=O)N(C)[C@@H](C)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](C)C(=O)O
	InChI	1S/C45H72N14O15S/c1-23(47)36(65)54-27(14-10-18-50-45(48)49)38(67)51-21-33(60)53-30(19-26-11-6-5-7-12-26)41(70)58-32(22-75)42(71)56-29(13-8-9-17-46)43(72)59(4)25(3)37(66)55-28(15-16-34(61)62)39(68)57-31(20-35(63)64)40(69)52-24(2)44(73)74/h5-7,11-12,23-25,27-32,75H,8-10,13-22,46-47H2,1-4H3,(H,51,67)(H,52,69)(H,53,60)(H,54,65)(H,55,66)(H,56,71)(H,57,68)(H,58,70)(H,61,62)(H,63,64)(H,73,74)(H4,48,49,50)/t23-,24-,25-,27-,28-,29-,30+,31-,32-/m0/s1
2	HELM	RNA1{R(A)P.[mR](U)[sP].R(G)P.R([5meC])P.[dR](T)P.[dR](T)}$$$$
	SMILES	OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](OC)[C@@H]1OP(=O)(S)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H](C[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H](C[C@@H]1O)n1cc(C)c(=O)[nH]c1=O)n1cc(C)c(=O)[nH]c1=O)n1cc(C)c(N)nc1=O)n1cnc2c1nc(N)[nH]c2=O)n1ccc(=O)[nH]c1=O)n1cnc2c1ncnc2N
	InChI	1S/C60H78N19O39P5S/c1-21-9-77(58(90)69-45(21)61)52-38(84)41(116-120(95,96)104-14-28-25(8-34(109-28)76-11-23(3)50(87)73-60(76)92)114-119(93,94)103-13-27-24(81)7-33(108-27)75-10-22(2)49(86)72-59(75)91)29(111-52)15-105-122(99,100)117-42-30(112-54(39(42)85)79-20-67-36-48(79)70-56(63)71-51(36)88)17-107-123(101,124)118-43-31(113-55(44(43)102-4)74-6-5-32(82)68-57(74)89)16-106-121(97,98)115-40-26(12-80)110-53(37(40)83)78-19-66-35-46(62)64-18-65-47(35)78/h5-6,9-11,18-20,24-31,33-34,37-44,52-55,80-81,83-85H,7-8,12-17H2,1-4H3,(H,93,94)(H,95,96)(H,97,98)(H,99,100)(H,101,124)(H2,61,69,90)(H2,62,64,65)(H,68,82,89)(H,72,86,91)(H,73,87,92)(H3,63,70,71,88)/t24-,25-,26+,27+,28+,29+,30+,31+,33-,34-,37+,38+,39+,40+,41+,42+,43+,44+,52-,53-,54-,55-,123?/m0/s1
3	HELM	PEPTIDE1{A.R.C.A.A.K.T.C.D.A}$PEPTIDE1,PEPTIDE1,8:R3-3:R3$$$
	SMILES	NCCCC[C@@H]1NC(=O)[C@H](C)NC(=O)[C@H](C)NC(=O)[C@H](CSSC[C@H](NC(=O)[C@@H](NC1=O)[C@@H](C)O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](C)C(=O)O)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](C)N
	InChI	1S/C38H66N14O14S2/c1-16(40)28(56)47-22(10-8-12-43-38(41)42)31(59)50-24-14-67-68-15-25(35(63)49-23(13-26(54)55)33(61)46-19(4)37(65)66)51-36(64)27(20(5)53)52-32(60)21(9-6-7-11-39)48-30(58)18(3)44-29(57)17(2)45-34(24)62/h16-25,27,53H,6-15,39-40H2,1-5H3,(H,44,57)(H,45,62)(H,46,61)(H,47,56)(H,48,58)(H,49,63)(H,50,59)(H,51,64)(H,52,60)(H,54,55)(H,65,66)(H4,41,42,43)/t16-,17-,18-,19-,20+,21-,22-,23-,24-,25-,27-/m0/s1
4	HELM	PEPTIDE1{A.R.C.D.K.A}|PEPTIDE2{G.A.K.A}$PEPTIDE1,PEPTIDE2,4:R3-1:R1$$$
	SMILES	NCCCC[C@H](NC(=O)[C@H](C)NC(=O)CNC(=O)C[C@H](NC(=O)[C@H](CS)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](C)N)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](C)C(=O)O)C(=O)N[C@@H](C)C(=O)O
	InChI	1S/C39H71N15O13S/c1-19(42)30(57)50-25(12-9-15-45-39(43)44)34(61)54-27(18-68)36(63)53-26(35(62)52-24(11-6-8-14-41)33(60)49-22(4)38(66)67)16-28(55)46-17-29(56)47-20(2)31(58)51-23(10-5-7-

		13-40)32(59)48-21(3)37(64)65/h19-27,68H,5-18,40-42H2,1-4H3,(H,46,55)(H,47,56)(H,48,59)(H,49,60)(H,50,57)(H,51,58)(H,52,62)(H,53,63)(H,54,61)(H,64,65)(H,66,67)(H4,43,44,45)/t19-,20-,21-,22-,23-,24-,25-,26-,27-/m0/s1
5	HELM	RNA1{R(A)P.R(G)P.R(C)P.R(U)P.R(C)P.R(C)P.R(C)}|RNA2{R(U)P.R(G)P.R(G)P.R(G)P.R(G)P.R(A)P.R(G)}$$RNA1,RNA2,17:pair-11:pair|RNA1,RNA2,20:pair-8:pair|RNA1,RNA2,14:pair-14:pair|RNA1,RNA2,11:pair-17:pair|RNA1,RNA2,8:pair-20:pair$RNA2{as}\|RNA1{ss}$
	SMILES	OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1O)n1ccc(N)nc1=O)n1ccc(N)nc1=O)n1ccc(N)nc1=O)n1ccc(=O)[nH]c1=O)n1ccc(N)nc1=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1ncnc2N.OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1ncnc2N)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1nc(N)[nH]c2=O)n1ccc(=O)[nH]c1=O
	InChI	1S/C69H84N32O47P6.C65H84N24O47P6/c70-45-25-46(77-10-76-45)96(11-78-25)59-34(107)40(144-149(118,119)130-4-18-31(104)32(105)57(137-18)97-12-79-26-47(97)85-64(71)90-52(26)112)19(138-59)5-132-151(122,123)146-42-21(140-61(36(42)109)99-14-81-28-49(99)87-66(73)92-54(28)114)7-134-153(126,127)148-44-23(142-63(38(44)111)101-16-83-30-51(101)89-68(75)94-56(30)116)9-135-154(128,129)147-43-22(141-62(37(43)110)100-15-82-29-50(100)88-67(74)93-55(29)115)8-133-152(124,125)145-41-20(139-60(35(41)108)98-13-80-27-48(98)86-65(72)91-53(27)113)6-131-150(120,121)143-39-17(3-102)136-58(33(39)106)95-2-1-24(103)84-69(95)117;66-28-1-6-83(61(101)76-28)53-36(93)35(92)22(125-53)12-118-137(106,107)132-44-23(126-54(38(44)95)84-7-2-29(67)77-62(84)102)13-120-139(110,111)133-45-24(127-55(39(45)96)85-8-3-30(68)78-63(85)103)14-121-141(114,115)135-47-26(129-57(41(47)98)87-10-5-32(91)80-65(87)105)16-122-140(112,113)134-46-25(128-56(40(46)97)86-9-4-31(69)79-64(86)104)15-123-142(116,117)136-48-27(130-59(42(48)99)89-20-75-34-51(89)81-60(71)82-52(34)100)17-119-138(108,109)131-43-21(11-90)124-58(37(43)94)88-19-74-33-49(70)72-18-73-50(33)88/h1-2,10-23,31-44,57-63,102,104-111H,3-9H2,(H,118,119)(H,120,121)(H,122,123)(H,124,125)(H,126,127)(H,128,129)(H2,70,76,77)(H,84,103,117)(H3,71,85,90,112)(H3,72,86,91,113)(H3,73,87,92,114)(H3,74,88,93,115)(H3,75,89,94,116);1-10,18-27,35-48,53-59,90,92-99H,11-17H2,(H,106,107)(H,108,109)(H,110,111)(H,112,113)(H,114,115)(H,116,117)(H2,66,76,101)(H2,67,77,102)(H2,68,78,103)(H2,69,79,104)(H2,70,72,73)(H,80,91,105)(H3,71,81,82,100)/t17-,18-,19-,20-,21-,22-,23-,31-,32-,33-,34-,35-,36-,37-,38-,39-,40-,41-,42-,43-,44-,57+,58+,59+,60+,61+,62+,63+;21-,22-,23-,24-,25-,26-,27-,35-,36-,37-,38-,39-,40-,41-,42-,43-,44-,45-,46-,47-,48-,53+,54+,55+,56+,57+,58+,59+/m11/s1
6	HELM	RNA1{P.R(A)P.R(G)P.R(C)P.R(U)P.R(T)P.R(T)P.R(T)P.R(T)}|CHEM1{SS3}$RNA1,CHEM1,1:R1-1:R1$$$
	SMILES	OCCCSSCCCOP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1O)n1cc(C)c(=O)[nH]c1=O)n1cc(C)c(=O)[nH]c1=O)n1cc(C)c(=O)[nH]c1=O)n1cc(C)c(=O)[nH]c1=O)n1ccc(=O)[nH]c1=O)n1ccc(N)nc1=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1ncnc2N
	InChI	1S/C84H113N23O62P8S2/c1-30-15-102(81(126)96-65(30)119)70-47(111)46(110)34(155-70)19-148-171(132,133)165-57-38(158-73(50(57)114)103-16-31(2)66(120)97-82(103)127)23-152-174(138,139)167-59-40(160-75(52(59)116)105-18-33(4)68(122)99-84(105)129)25-153-175(140,141)166-58-39(159-74(51(58)115)104-17-32(3)67(121)98-83(104)128)24-151-173(136,137)164-56-37(157-72(49(56)113)101-10-8-43(109)93-80(101)125)22-149-172(134,135)163-55-36(156-71(48(55)112)100-9-7-42(85)92-79(100)124)21-150-176(142,143)169-61-41(162-77(54(61)118)107-29-91-45-64(107)94-78(87)95-69(45)123)26-154-177(144,145)168-60-35(20-147-170(130,131)146-12-6-14-179-178-13-5-11-108)161-76(53(60)117)106-28-90-44-62(86)88-27-89-63(44)106/h7-10,15-18,27-29,34-41,46-61,70-77,108,110-118H,5-6,11-14,19-26H2,1-4H3,(H,130,131)(H,132,133)(H,134,135)(H,136,137)(H,138,139)(H,140,141)(H,142,143)(H,144,145)(H2,85,92,124)(H2,86,88,89)(H,93,109,125)(H,96,119,126)(H,97,120,127)(H,98,121,128)(H,99,122,129)(H3,87,94,95,123)/t34-,35-,36-,37-,38-,39-,40-,41-,46-,47-,48-,49-,50-,51-,52-,53-,54-,55-,56-,57-,58-,59-,60-,61-,70+,71+,72+,73+,74+,75+,76+,77+/m1/s1

7	HELM	RNA1{R(A)P.R(A)P.R(G)P.R(G)P.R(C)P.R(U)P.R(A)P.R(A)P}|RNA2{R(A)P.R(A)P.R(G)P.R(G)P.R(C)P.R(U)P.R(A)P.R(A)P}|CHEM1{sDBL}$RNA2,CHEM1,24:R2-1:R3|RNA1,CHEM1,24:R2-1:R2$$$
	SMILES	OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OC[C@H]1O[C@@H]([C@H](O)[C@@H]1OP(=O)(O)OCCCCC(=O)NCC(O)CNC(=O)CCCCOP(=O)(O)O[C@@H]1[C@@H](COP(=O)(O)O[C@@H]2[C@@H](COP(=O)(O)O[C@@H]3[C@@H](COP(=O)(O)O[C@@H]4[C@@H](COP(=O)(O)O[C@@H]5[C@@H](COP(=O)(O)O[C@@H]6[C@@H](COP(=O)(O)O[C@@H]7[C@@H](COP(=O)(O)O[C@@H]8[C@@H](CO)O[C@@H]([C@@H]8O)n8cnc9c8ncnc9N)O[C@@H]([C@@H]7O)n7cnc8c7ncnc8N)O[C@@H]([C@@H]6O)n6cnc7c6nc(N)[nH]c7=O)O[C@@H]([C@@H]5O)n5cnc6c5nc(N)[nH]c6=O)O[C@@H]([C@@H]4O)n4ccc(N)nc4=O)O[C@@H]([C@@H]3O)n3ccc(=O)[nH]c3=O)O[C@@H]([C@@H]2O)n2cnc3c2ncnc3N)O[C@@H]([C@@H]1O)n1cnc2c1ncnc2N)n1cnc2c1ncnc2N)n1cnc2c1ncnc2N)n1ccc(=O)[nH]c1=O)n1ccc(N)nc1=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1nc(N)[nH]c2=O)n1cnc2c1ncnc2N)n1cnc2c1ncnc2N
	InChI	1S/C169H216N72O111P16/c170-72-7-11-226(166(269)214-72)146-92(251)108(58(323-146)21-311-361(289,290)349-118-70(335-158(102(118)261)238-51-210-86-138(238)218-162(180)222-142(86)265)33-319-367(301,302)351-120-68(333-160(104(120)263)240-53-212-88-140(240)220-164(182)224-144(88)267)31-317-365(297,298)345-114-64(329-154(98(114)257)234-47-206-82-126(176)190-39-198-134(82)234)27-307-355(277,278)337-106-56(19-242)321-150(90(106)249)230-43-202-78-122(172)186-35-194-130(78)230)341-357(281,282)309-23-60-110(94(253)148(325-60)228-13-9-76(247)216-168(228)271)343-359(285,286)315-29-66-116(100(259)156(331-66)236-49-208-84-128(178)192-41-200-136(84)236)347-363(293,294)313-25-62-112(96(255)152(327-62)232-45-204-80-124(174)188-37-196-132(80)232)339-353(273,274)305-15-3-1-5-74(245)184-17-55(244)18-185-75(246)6-2-4-16-306-354(275,276)340-113-63(328-153(97(113)256)233-46-205-81-125(175)189-38-197-133(81)233)26-314-364(295,296)348-117-67(332-157(101(117)260)237-50-209-85-129(179)193-42-201-137(85)237)30-316-360(287,288)344-111-61(326-149(95(111)254)229-14-10-77(248)217-169(229)272)24-310-358(283,284)342-109-59(324-147(93(109)252)227-12-8-73(171)215-167(227)270)22-312-362(291,292)350-119-71(336-159(103(119)262)239-52-211-87-139(239)219-163(181)223-143(87)266)34-320-368(303,304)352-121-69(334-161(105(121)264)241-54-213-89-141(241)221-165(183)225-145(89)268)32-318-366(299,300)346-115-65(330-155(99(115)258)235-48-207-83-127(177)191-40-199-135(83)235)28-308-356(279,280)338-107-57(20-243)322-151(91(107)250)231-44-203-79-123(173)187-36-195-131(79)231/h7-14,35-71,90-121,146-161,242-244,249-264H,1-6,15-34H2,(H,184,245)(H,185,246)(H,273,274)(H,275,276)(H,277,278)(H,279,280)(H,281,282)(H,283,284)(H,285,286)(H,287,288)(H,289,290)(H,291,292)(H,293,294)(H,295,296)(H,297,298)(H,299,300)(H,301,302)(H,303,304)(H2,170,214,269)(H2,171,215,270)(H2,172,186,194)(H2,173,187,195)(H2,174,188,196)(H2,175,189,197)(H2,176,190,198)(H2,177,191,199)(H2,178,192,200)(H2,179,193,201)(H,216,247,271)(H,217,248,272)(H3,180,218,222,265)(H3,181,219,223,266)(H3,182,220,224,267)(H3,183,221,225,268)/t56-,57-,58-,59-,60-,61-,62-,63-,64-,65-,66-,67-,68-,69-,70-,71-,90-,91-,92-,93-,94-,95-,96-,97-,98-,99-,100-,101-,102-,103-,104-,105-,106-,107-,108-,109-,110-,111-,112-,113-,114-,115-,116-,117-,118-,119-,120-,121-,146+,147+,148+,149+,150+,151+,152+,153+,154+,155+,156+,157+,158+,159+,160+,161+/m1/s1
8	HELM	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$$$
	SMILES	NCCCC[C@H](NC(=O)CNC(=O)[C@H](CSC1CC(=O)N(CC2CCC(CC2)C(=O)NCCCCCCOP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2OP(=O)(O)OC[C@H]2O[C@@H]([C@H](O)[C@@H]2O)n2cnc3c2nc(N)[nH]c3=O)n2cnc3c2nc(N)[nH]c3=O)n2cnc3c2ncnc3N)n2cnc3c2nc(N)[nH]c3=O)n2ccc(=O)[nH]c2=O)n2ccc(=O)[nH]c2=O)n2ccc(N)nc2=O)C1=O)NC(=O)[C@H](C)N)C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CC(=O)O)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CCCNC(=N)N)C(=O)O
	InChI	1S/C120H173N42O67P7S/c1-48(123)95(179)145-56(97(181)133-33-68(165)140-51(11-4-6-23-121)98(182)142-53(18-19-70(167)168)100(184)144-55(31-71(169)170)101(185)141-52(12-5-7-24-122)99(183)143-54(113(190)191)13-10-26-132-114(126)127)42-237-64-32-69(166)158(105(64)189)34-49-14-16-50(17-15-49)96(180)131-25-8-2-3-9-30-209-230(195,196)210-36-58-84(78(173)107(218-58)155-27-20-65(124)146-118(155)192)224-232(199,200)212-37-59-85(79(174)108(219-59)156-28-21-66(163)147-119(156)193)225-233(201,202)213-38-60-86(80(175)109(220-60)157-29-22-67(164)148-120(157)194)226-234(203,204)215-41-63-89(83(178)112(223-63)162-47-139-75-94(162)151-117(130)154-104(75)188)229-236(207,208)214-39-61-87(81(176)110(221-61)159-44-136-72-90(125)134-

		43-135-91(72)159)228-235(205,206)216-40-62-88(82(177)111(222-62)161-46-138-74-93(161)150-116(129)153-103(74)187)227-231(197,198)211-35-57-76(171)77(172)106(217-57)160-45-137-73-92(160)149-115(128)152-102(73)186/h20-22,27-29,43-64,76-89,106-112,171-178H,2-19,23-26,30-42,121-123H2,1H3,(H,131,180)(H,133,181)(H,140,165)(H,141,185)(H,142,182)(H,143,183)(H,144,184)(H,145,179)(H,167,168)(H,169,170)(H,190,191)(H,195,196)(H,197,198)(H,199,200)(H,201,202)(H,203,204)(H,205,206)(H,207,208)(H2,124,146,192)(H2,125,134,135)(H4,126,127,132)(H,147,163,193)(H,148,164,194)(H3,128,149,152,186)(H3,129,150,153,187)(H3,130,151,154,188)/t48-,49?,50?,51-,52-,53-,54-,55-,56-,57+,58+,59+,60+,61+,62+,63+,64?,76+,77+,78+,79+,80+,81+,82+,83+,84+,85+,86+,87+,88+,89+,106-,107-,108-,109-,110-,111-,112-/m0/s1

9.1 HELM example with in-line notation

Sample	Format	Notation
1	HELM	PEPTIDE1{A.R.G.[dF].C.K.[meA].E.D.A}$$$$
	HELM with one monomer in in-line notation (cxsmiles)	PEPTIDE1{A.[NC(=N)NCCC[C@H](N[*])C([*])=O|$;;;;;;;;;_R1;;_R2;$|].G.[dF].C.K.[meA].E.D.A}$$$$
	HELM with one monomer in in-line notation (atom map)	PEPTIDE1{A.[NC(=N)NCCC[C@H](N[*:1])C([*:2])=O].G.[dF].C.K.[meA].E.D.A}$$$$V2.0
	SMILES	[H]NCCCC[C@H](NC(=O)[C@H](CS[H])NC(=O)[C@@H](CC1=CC=CC=C1)NC(=O)CNC(=O)[C@H](CCCNC(N)=N)NC(=O)[C@H](C)N[H])C(=O)N(C)[C@@H](C)C(=O)N[C@@H](CCC(O)=O)C(=O)N[C@@H](CC(O)=O)C(=O)N[C@@H](C)C(O)=O
	InChI	1S/C45H72N14O15S/c1-23(47)36(65)54-27(14-10-18-50-45(48)49)38(67)51-21-33(60)53-30(19-26-11-6-5-7-12-26)41(70)58-32(22-75)42(71)56-29(13-8-9-17-46)43(72)59(4)25(3)37(66)55-28(15-16-34(61)62)39(68)57-31(20-35(63)64)40(69)52-24(2)44(73)74/h5-7,11-12,23-25,27-32,75H,8-10,13-22,46-47H2,1-4H3,(H,51,67)(H,52,69)(H,53,60)(H,54,65)(H,55,66)(H,56,71)(H,57,68)(H,58,70)(H,61,62)(H,63,64)(H,73,74)(H4,48,49,50)/t23-,24-,25-,27-,28-,29-,30+,31-,32-/m0/s1

9.2 HELM 2.0 examples

Sample	Description	Notation
1	Monomer ambiguity with missing monomer	PEPTIDE1{A.C.D.E.(_,K)}$$$$V2.0
2	Monomer ambiguity with monomer mixture	PEPTIDE1{A.A.A.A.(A:1+G:1+[Aha]:1+X:1).A}$$$$V2.0
3	Connection ambiguity with monomer mixture	PEPTIDE1{A.C.D.E}|PEPTIDE2{G.C.S.P.K}|CHEM1{[[*]SCCCc1ccccc1 |$_R1;;;;;;;;;;$|]}$PEPTIDE2,CHEM1,(C+K):R3-1:R1$$$$V2.0

4	Connection ambiguity with undefined connection partner	PEPTIDE1{A.C.D.E}|PEPTIDE2{G.C.S.P.K}|CHEM1{[[*]SCCCc1ccccc1 |$_R1;;;;;;;;;;$|]}$PEPTIDE2, CHEM1,?:R3-1:R1$$$$V2.0
5	Composition ambiguity	PEPTIDE1{A.C.D.E}|PEPTIDE2{G}|CHEM1{[Dig]}|CHEM2{[Dig]}$PEPTIDE1,CHEM1,C:R3-1:R1\|PEPTIDE2,CHEM2,C:R3-1:R1$G1(PETPDIE1+CHEM1:2.5)|G2(PEPTIDE2+CHEM2:1.5)$$V2.0
6	Inline Annotations for monomer and polymer	PEPTIDE1{A.G"mutated"}"LC"|PEPTIDE2{L.C}"HC"$$$$V2.0

10 Appendix 4: XHELM examples

Sample	Format	Notation
1	HELM	PEPTIDE1{A.A.G.K}$PEPTIDE1,PEPTIDE1,1:R1-4:R2$$$

	XHELM	<?xml version="1.0" encoding="UTF-8"?>
		<Xhelm>
		  <HelmNotation>PEPTIDE1{A.A.G.K}$PEPTIDE1,PEPTIDE1,1:R1-4:R2$$$</HelmNotation>
		  <MonomerList>
		    <Monomer>
		      <MonomerID>K</MonomerID>
		      <MonomerSmiles>[*]N[C@@H](CCCCN[*])C([*])=O |$_R1;;;;;;;;_R3;;_R2;$|</MonomerSmiles>
		
		<MonomerMolFile>H4sIAAAAAAAAAKWUP0/EMAzF93wKS7Bi2c5fzxxiugPdwM7IwsDA5z+7B22vqcTRizI0v9ivL+prAsD+/ev74xOAhYSJcxbZwThCsA1gBqDZnIaqwpsQUbDFA6NJiHNBUlZ/IrRdgke4lFifPyoqqXkvY6maVlXkLxXBEnX0IrTNS8SY4+Qlb1NJSKXd7CVhbXKzl4xJ6+SFZyqHLd+IsOWUtqpMeVEu87y8XK9CdqK8npfj3dUqQ17iuTeli+z+Q6XYibiOXsoWFXs1LwIefhF3lKmnVthWaRwoz6mh1Ncayj01VHpqqPa0DbfGkla/ThZ0D3B8fj2b81/DCzyW3u7hiF7xdNiFcG8jnADxHCCFtAQAAA==</MonomerMolFile>
		      <MonomerType>Backbone</MonomerType>
		      <PolymerType>PEPTIDE</PolymerType>
		      <NaturalAnalog>K</NaturalAnalog>
		      <MonomerName>Lysine</MonomerName>
		      <Attachments>
		        <Attachment>
		          <AttachmentID>R2-OH</AttachmentID>
		          <AttachmentLabel>R2</AttachmentLabel>
		          <CapGroupName>OH</CapGroupName>
		          <CapGroupSmiles>O[*] |$;_R2$|</CapGroupSmiles>
		        </Attachment>
		        <Attachment>
		          <AttachmentID>R1-H</AttachmentID>
		          <AttachmentLabel>R1</AttachmentLabel>
		          <CapGroupName>H</CapGroupName>
		          <CapGroupSmiles>[*][H] |$_R1;$|</CapGroupSmiles>
		        </Attachment>
		        <Attachment>
		          <AttachmentID>R3-H</AttachmentID>
		          <AttachmentLabel>R3</AttachmentLabel>

```
          <CapGroupName>H</CapGroupName>
          <CapGroupSmiles>[*][H] |$_R3;$|</CapGroupSmiles>
        </Attachment>
      </Attachments>
    </Monomer>
    <Monomer>
      <MonomerID>C</MonomerID>
      <MonomerSmiles>[*]N[C@@H](CS[*])C([*])=O |$_R1;;;;;_R3;;_R2;$|</MonomerSmiles>

<MonomerMolFile>H4sIAAAAAAAAAJ2TPQ/CIBCGd37Fm+hqc3dAC7MaJz+iibuji4ODv1+gtdKPxNY
LA7zcPbxciwL2t+fr/gBYSJjYWpIN2lAK8IADKBvf8N7jKkSk4ooKZxzXM8Ni6lnYJazRRYyPhsJal1ntGE
V+UFZUWEeu9VL+5WXFhXgrbS1nlMvMG9nalRPjM8phDiV0N3kJrnSVd/c4ncKFLaXx0rvReTG5L7G
7TE2Hul5mUCRU+Gr0G02mhCLu/Q7qI/FQrYZqSLSjqk4q52qQzDDXphfSV016Ol11D5x3p8SJVuIps
TRmxWTomLE9bJRahlBvGBYktKADAAA=</MonomerMolFile>
      <MonomerType>Backbone</MonomerType>
      <PolymerType>PEPTIDE</PolymerType>
      <NaturalAnalog>C</NaturalAnalog>
      <MonomerName>Cysteine</MonomerName>
      <Attachments>
        <Attachment>
          <AttachmentID>R2-OH</AttachmentID>
          <AttachmentLabel>R2</AttachmentLabel>
          <CapGroupName>OH</CapGroupName>
          <CapGroupSmiles>O[*] |$;_R2$|</CapGroupSmiles>
        </Attachment>
        <Attachment>
          <AttachmentID>R1-H</AttachmentID>
          <AttachmentLabel>R1</AttachmentLabel>
          <CapGroupName>H</CapGroupName>
          <CapGroupSmiles>[*][H] |$_R1;$|</CapGroupSmiles>
        </Attachment>
        <Attachment>
          <AttachmentID>R3-H</AttachmentID>
          <AttachmentLabel>R3</AttachmentLabel>
          <CapGroupName>H</CapGroupName>
          <CapGroupSmiles>[*][H] |$_R3;$|</CapGroupSmiles>
        </Attachment>
      </Attachments>
    </Monomer>
    <Monomer>
      <MonomerID>G</MonomerID>
      <MonomerSmiles>[*]NCC([*])=O |$_R1;;;;_R2;$|</MonomerSmiles>

<MonomerMolFile>H4sIAAAAAAAAAKWSPw+CQAzF9/sUL9FV0t4/uFmME2gc3B1dHBz8/PbuRA9xg
NgUeLy2vxwNCugu98f1BlDDQS6ng27xDqUADzipF/mJEALOmoikDxuqam5CVp58VKBKqoQtxojf+aI
QWZsVB+v+oJg86w3ZgnJYQqllLVFxZQ03BaWfTUGkEA97qbmgnFazzxJPYFxWjR2+bRlF1sApR668
66krlkn3b9dNe236TcZuB5z2xzQRR9JTmsCxsutbpdYS6gmdUi5fhAIAAA==</MonomerMolFile>
      <MonomerType>Backbone</MonomerType>
      <PolymerType>PEPTIDE</PolymerType>
      <NaturalAnalog>G</NaturalAnalog>
      <MonomerName>Glycine</MonomerName>
      <Attachments>
        <Attachment>
          <AttachmentID>R2-OH</AttachmentID>
          <AttachmentLabel>R2</AttachmentLabel>
          <CapGroupName>OH</CapGroupName>
          <CapGroupSmiles>O[*] |$;_R2$|</CapGroupSmiles>
        </Attachment>
        <Attachment>
          <AttachmentID>R1-H</AttachmentID>
          <AttachmentLabel>R1</AttachmentLabel>
          <CapGroupName>H</CapGroupName>
          <CapGroupSmiles>[*][H] |$_R1;$|</CapGroupSmiles>
        </Attachment>
      </Attachments>
    </Monomer>
    <Monomer>
      <MonomerID>A</MonomerID>
      <MonomerSmiles>C[C@H](N[*])C([*])=O |$;;;_R1;;_R2;$|</MonomerSmiles>

<MonomerMolFile>H4sIAAAAAAAAAKWSuw7CMAxF93yFJVhrOc57poipBXVgZ2RhYOD7SYIg6UOi
CCtS1Hudo1snAqC73B/XGwBZNuSlZsstfEoIAAdgo1+tUiEEODMRifRlUHsfe6FRSNpz0gijS7CDMWJ
5ZYpFJiVfFG3JL1Lkd0qQrmRRf2RJf9xoZG/qLP0vlJjFJQojs6qzHNdTHFrFZS6hogyb1ZRyRxptGM1l
NYXz/GV9C1HVb2Paq+ZqlEz2pqqd9+r8BMdqBzAcTpmQjuQ9DVgmZ9+3QmxjiSe9Zxcz4AIAAA==</
MonomerMolFile>
      <MonomerType>Backbone</MonomerType>
      <PolymerType>PEPTIDE</PolymerType>
      <NaturalAnalog>A</NaturalAnalog>
      <MonomerName>Alanine</MonomerName>
      <Attachments>
        <Attachment>
          <AttachmentID>R2-OH</AttachmentID>
          <AttachmentLabel>R2</AttachmentLabel>
          <CapGroupName>OH</CapGroupName>
          <CapGroupSmiles>O[*] |$;_R2$|</CapGroupSmiles>
        </Attachment>
        <Attachment>
          <AttachmentID>R1-H</AttachmentID>
          <AttachmentLabel>R1</AttachmentLabel>
          <CapGroupName>H</CapGroupName>
          <CapGroupSmiles>[*][H] |$_R1;$|</CapGroupSmiles>
        </Attachment>
      </Attachments>
    </Monomer>
  </MonomerList>
</Xhelm>
```