# HELM Monomer Guidelines

Contents

[1 HELM Monomer Guidelines 1](#_Toc524449486)

[1.1 Introduction 2](#_Toc524449487)

[1.2 Monomer Concepts 2](#_Toc524449488)

[1.2.1 Cannonicalization. 2](#_Toc524449489)

[1.2.2 Variation and the exploding monomer set 4](#_Toc524449490)

[1.2.3 Monomers and the synthetic route to the polymer 6](#_Toc524449491)

[1.3 HELM specific decisions 7](#_Toc524449492)

[1.3.1 R group variation 7](#_Toc524449493)

**Version History:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Release Date** | **Author** | **History** |
| 0.1 | 11th September 2018 | Claire Bellamy | Kick off discussion |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Introduction

The aim of this document is to record the questions and decisions of the HELM monomer discussion group. We will aim to clarify the concepts of

* What is a monomer and how should one be recognised?
* how should monomers be classified particularly when they could be used with more than one polymer type?
* How should monomers be named?

Other questions will arise as work progresses and later discussions may influence earlier decisions. When this happens, the appropriate section will be updated.

This group will follow conventions where they exist and can be applied, however, most only cover natural monomers, e.g. natural amino acids and nucleotide bases and not the wider monomers HELM users typically need. It is hoped than existing nomenclature bodies like IUPAC will take up the challenge and provide guidelines this group can adopt.

We should remember that HELM requires more information about a monomer than the simple structure and a name and will continue to be responsible for creating guidelines for managing these aspects.

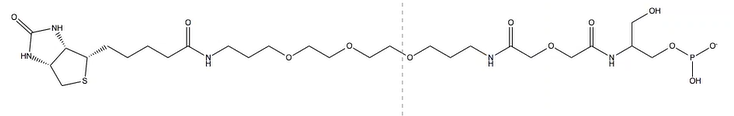
## Monomer Concepts

This section will record our thoughts on what a monomer is and what rules a required to define a “good” monomer set?

Initially it will outline the decisions that must be made. Once decisions have been made, the definitions will be built from them.

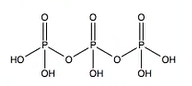
### Cannonicalization.

If we want the HELM notation to be fully canonicalizable then monomers must fulfil strict criteria. It cannot be possible to create a monomer from two other monomers in the set. This is challenged by commercially available monomers used in practice. For example:

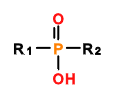


Although clear-cut amino acids are difficult to find in this structure, it contains such a lot of functionality that it is difficult to see how it could be guaranteed to be independent of all other monomers in the set.

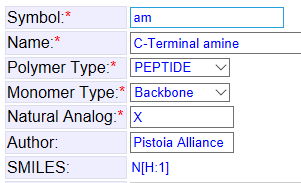
The current phosphate definition also requires HELM users to create new monomers where there are multiple phosphates connected to each other. E.g. a triphosphate



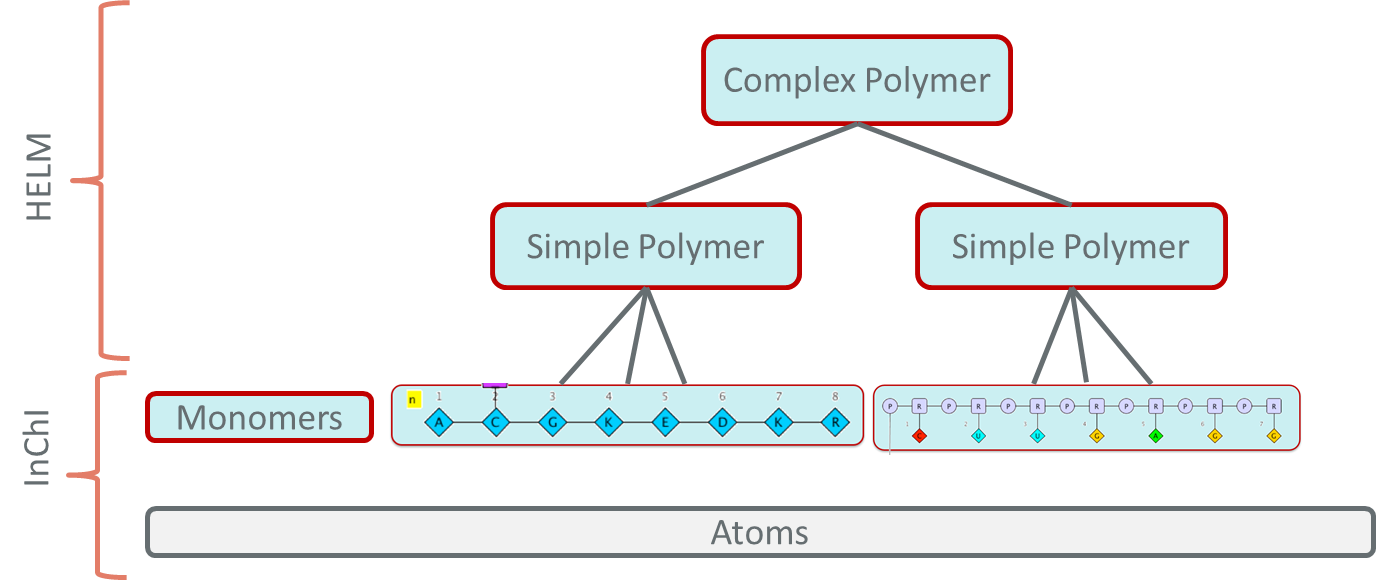
Cannot be constructed from the existing individual phosphate monomer:



And amine is defined as a peptide monomer in the current HELM demo set which will clearly appear in many other monomers.



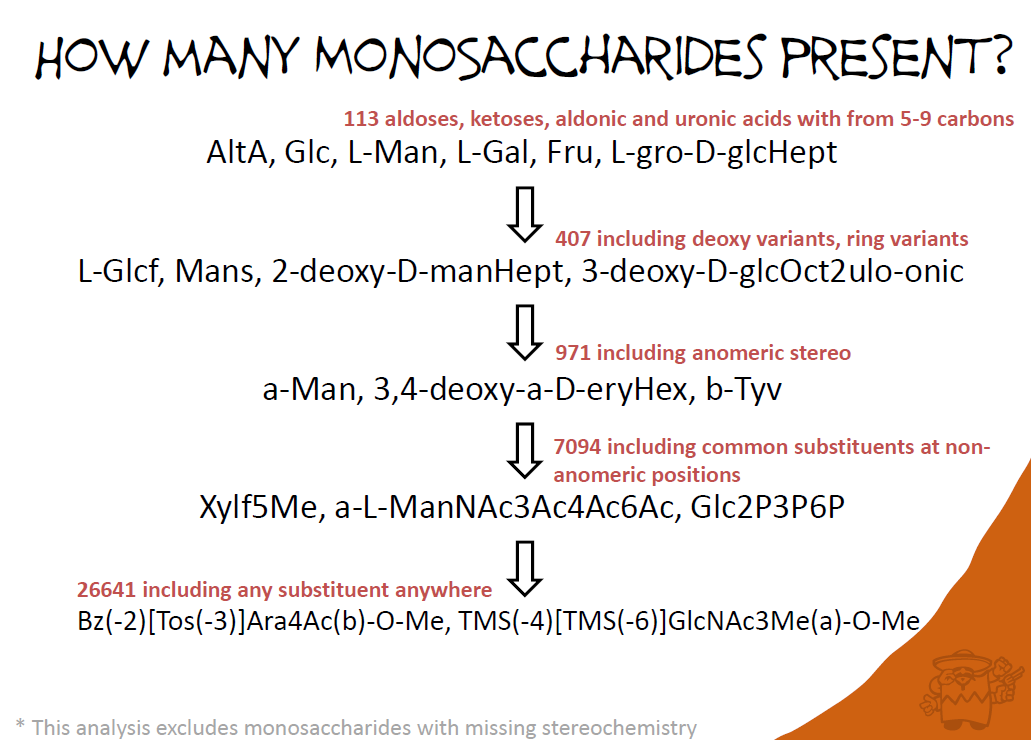
Canonicalization could be performed via small molecule representations i.e. mol, SMILES and particularly InChI. HELM encodes monomer and higher-level information not captured by atom/bond representation and would rely on small molecule representations for canonicalization.

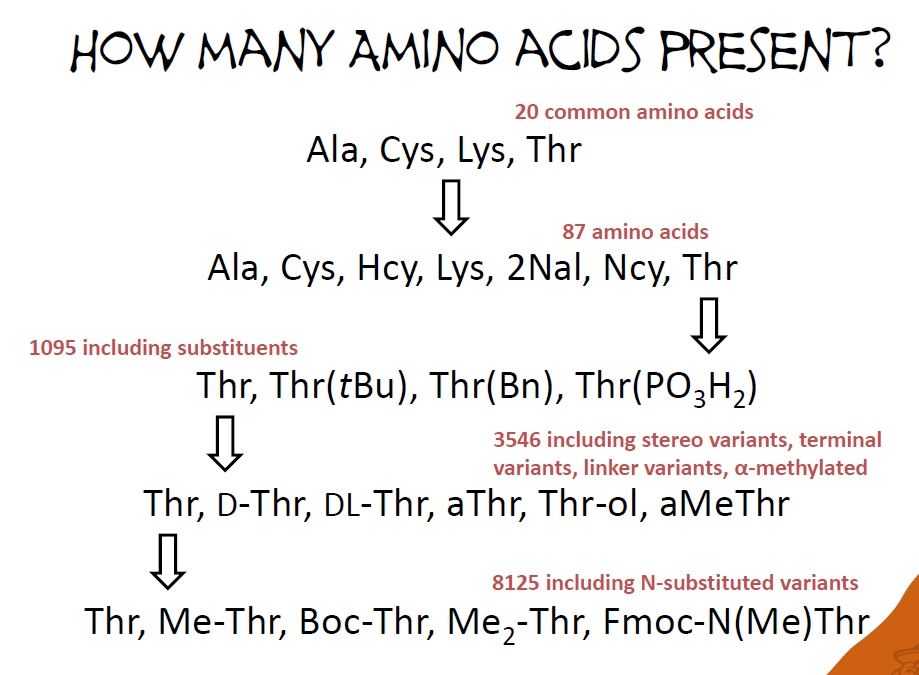


**QUESTION 1: must monomers be entirely independent of each other, so you can canonicalize HELM directly or should we rely on conversion to a small molecule representation for canonicalization?**

### Variation and the exploding monomer set

Non-natural monomers with structures that are very different to the natural set, need to be defined individually. However, there are many that are a minor variation on naturally occurring monomers. These minor variations risk the set getting very large, very quickly as shown in Noel O’Boyd’s presentation.





Large monomer sets are not necessarily easy to manage for scientists who are only interested in a limited subset of them. The HELM editor relies upon a list where the user can select from a manageable set



However, this is a UI challenge not a conceptual one. Readability could be improved by presenting a pre-filtered preferred subset, for example.

Readability of the HELM string itself could be improved by allowing caps to be included:

1. PEPTIDE1{[meA].A}$$$$V2.0

Is easier to read than

1. PEPTIDE1{[Me].A.A}$$$$V2.0 although not by much. (Note that I have made up the methyl peptide capping group [Me] for this example.)

And the following is even less easy to read

1. CHEM1{[Me]}|PEPTIDE1{A.A}$CHEM1,PEPTIDE1,1:R1-1:R1$$$V2.0

If we take approach 1, we need to decide which variants are allowed. For example, simple hydrocarbon chains appear straightforward: Me, Et, propyl, isopropyl, butyl etc. but how many should be in a reference set? Is 5 enough? 20? 100? Also, once you start adding functionality, even for common and simple examples like Boc, Fmoc, Tos, the variation becomes even larger.

Options:

* variations on natural monomers should be described as the natural monomer connected to a separate ‘capping group’ monomer.
* a short, recommended list of modified monomers are allowed.
* All variations should be a new monomer.

**QUESTION 2: which of these options should we take?**

### Monomers and the synthetic route to the polymer

A polymeric structure reflects the starting materials which it was made from. A monomer could be defined as the starting materials of the polymer.

***Issues***

* The synthetic route is not always (often) known.
* The starting materials can differ for the same product.
* The capping group would need to reflect the starting material, which would require monomers that only differ in capping groups. Is this acceptable?
* How would we represent multi-step syntheses?

***Pros***

Provides a way of identifying what the monomers should be.

**QUESTION 3 should monomers reflect the reactants in the synthetic path to the polymer?**

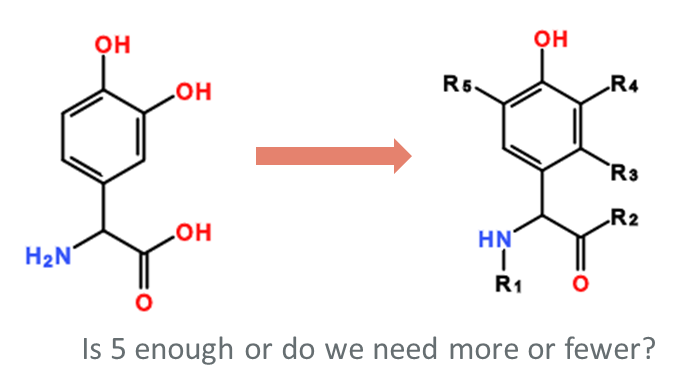
## HELM specific decisions

These decisions are a result of the requirements of the HELM notation itself. They do not concern the definition of what a monomer is and how it can be identified within a polymer, but are issues we need to address.

At present this section is a holding area for questions. More detail will be added as discussions progress.

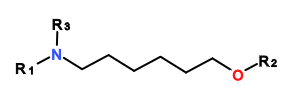
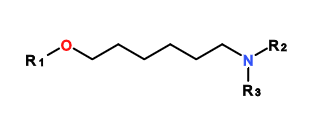
### R group variation

How many R groups should a monomer have? In the Norine set monomers have a lot of connections.

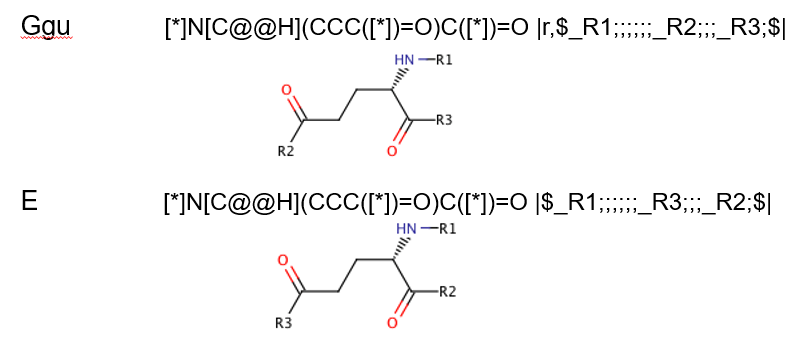


***Question*** - Should all monomers include large numbers of R groups in order to cover possible uses?

The same monomer can have connection points that are numbered differently:



Also



***Question –*** How many variations of R groups should be created?