Contents

[1. Introduction 3](#_Toc29809407)

[1.1. Document purpose 3](#_Toc29809408)

[1.2. Background 3](#_Toc29809409)

[2. HELM Notation Fundamentals 5](#_Toc29809410)

[2.1. HELM Monomer Definitions 5](#_Toc29809411)

[2.2. Connecting Monomers 6](#_Toc29809412)

[2.3. Connecting Branch and Undefined Monomers 7](#_Toc29809413)

[3. Monomer Core principles 8](#_Toc29809414)

[3.1. Canonicalization 8](#_Toc29809415)

[4. Monomer substitution (peptides) 9](#_Toc29809416)

[4.1. N and C terminal substitutions 10](#_Toc29809417)

[4.1.1. Example C substitution 10](#_Toc29809418)

[4.1.2. Example N substitution 11](#_Toc29809419)

[4.1.3. Midchain N substitution 12](#_Toc29809420)

[4.1.4. Peptide terminal monomers vs CHEMs 13](#_Toc29809421)

[4.2. Side chain substitutions 14](#_Toc29809422)

[4.3. Capping Rules - example structures 16](#_Toc29809423)

[4.4. Backbone monomer side-chains may be connected 17](#_Toc29809424)

[5. Monomer uniqueness rules 20](#_Toc29809425)

[5.1. Capped molecule equivalence 20](#_Toc29809426)

[5.1. R group equivalence – number of R groups 21](#_Toc29809427)

[5.2. R group equivalence –R group positions 23](#_Toc29809428)

[5.3. Recommendations for new monomers 24](#_Toc29809429)

[6. Monomer naming 25](#_Toc29809430)

[6.1. General principles 25](#_Toc29809431)

[6.2. Meaningful names 26](#_Toc29809432)

[6.2.1. Natural amino acids 26](#_Toc29809433)

[6.2.2. Stereoisomerism 26](#_Toc29809434)

[6.2.3. N substitution by simple alkyl groups 27](#_Toc29809435)

[6.2.4. Structural isomerism 28](#_Toc29809436)

[6.2.5. Aldehyde analogues 28](#_Toc29809437)

[6.2.6. Other monomers 29](#_Toc29809438)

[6.3. Common names 31](#_Toc29809439)

[6.4. Monomers with fewer than 2 backbone connection points 34](#_Toc29809440)

**Version History:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Release Date** | **Author** | **History** |
| Version 1\_0 | 6th March 2019 | Claire Bellamy, Anna Gaulton, Evan Bolton, Tianhong Zhang, Roger Sayle, Yohann Potier, Maude Pupin, Greg Adams, Sergio Rotstein. | First released version |
| Version 1\_1 d2 | 4th November 2020 | Claire Bellamy, Anna Gaulton, Evan Bolton, Tianhong Zhang | Updated after discussions with the HELM monomer team. |

# Introduction

## Document purpose

This document details the agreed recommendations for monomer construction and use.

It is not a specification and does not replace any part of the HELM specification but provides guidance for monomer curators and HELM users who want to learn from the experience of other groups.

## Background

Biopolymers contain repeating units; however, they differ from classical polymers since their repeating units are not identical. The monomer unit is identified by common functional groups that connect to form the polymer backbone, which leaves considerable scope for variation in the rest of the molecule. We see this in nature with 22 very diverse amino acids and 5 nucleotide bases.

The HELM notation specifies the process of connecting monomers and the format of the monomer information to be stored, but does not prescribe a set of monomers. Natural monomers are easy to identify, and will be included in all sets, however HELM is specifically designed to be flexible and allow non-natural monomers and even non-standard CHEM monomers that can be any structure at all and do not have to contain backbone functional groups.

There is a wide range of possible approaches to choosing which non-natural monomers to include in a set. At one end of the spectrum, a curator could select a very small set of core skeleton monomers. These would then be substituted by one or more other monomers to create the structure in the final polymer. This would result in a small monomer set, but complex HELM strings. At the other end, all structures could be new monomers resulting in a very large set of monomers and very simple HELM strings.

The HELM team favours defining the variation in the monomer structure itself and not substituting extensively. This keeps the HELM string clear and readable and uses small molecule representation for what it does best. The one exception is for terminal substitutions which should be separated out into new monomers. We also recommend that R groups are only added when needed.

This document also covers canonicalization issues, monomer uniqueness rules and some basic principles for naming monomers.

Most of the examples in this document illustrate peptide monomers. This is because amino acids have the widest variation, however the principles also apply to other polymer types.

HELM has not yet enforced these rules in the open-source tools or created a public monomer set embodies them. The original HELM set is available at <https://github.com/PistoiaHELM/HELMMonomerSets>, but there are some minor departures from these rules. We will aim to update them as soon as we can and will publish the outcome on the openHELM.org and the HELM wiki.

# HELM Notation Fundamentals

The formal specification for HELM and the definition of a monomer is available in the HELM wiki at <https://pistoiaalliance.atlassian.net/wiki/spaces/PUB/pages/13795362/HELM+Notation>

The following is an overview of the main concepts.

## HELM Monomer Definitions

HELM is a hierarchical notation with 4 layers:

* Atom/bonds – used to define monomers
* Monomers – used as the building blocks of simple polymers
* Simple polymers – chains of monomers connected in a predefined and consistent way.
* Complex polymers – chains of simple polymers connected as defined by the user constructing the molecule.

This means that a monomer must contain all the information required to connect it to other monomers in a simple polymer. The information required can be categorised:

|  |  |
| --- | --- |
| **Monomers** | |
| **Identification** |  |
| Symbol | A short name that is used in the HELM string and on graphical representations of the molecule. |
| Name | A more descriptive name based on the monomer structure. |
| Natural analogue | The symbol of the most closely structurally related natural monomer. This is used to write simple sequences and it has no structural function. |
| **Structure definition** |  |
| Molfile and/or SMILES | Chemical structure information about the monomer. It includes the connection point. |
| **Construction codes** |  |
| Polymer type | This defines how the backbone connection points, usually R1 and R2, are attached.  It also defines how branch monomers are attached if they exist within that polymer type. |
| Monomer type | Allowed values = backbone, branch and undefined.  This is used along with the polymer type to control the automatic attachment process. |

Author, Created Date and Internal ID can also be recorded to help manage monomer repositories, but they are not used in the notation itself.

|  |  |
| --- | --- |
| **R groups** | |
| Label | R followed by an integer denoting the R group number. This identifies the R group within the molecule. |
| CapGroupSMILES | SMILES of the leaving group structure. |
| CapGroupName | Description of the atoms in the leaving group. E.g. H, OH etc. |

ID is used to manage records in the monomer repository, and alternate ID (concatenation of label and CapGroupName) is used as an aid for display, but neither are crucial to the notation.

## Connecting Monomers

Backbone monomers connect R1 to R2, so alanine would connect to alanine like this.



However this final structure is not what occurs in reality; we need to know what atoms are really underneath the R1 and R2 labels.

The capping group definition defines what should be assumed if the R rgoups are not used. In this case:

* R1 capping group = H
* R2 capping group = OH

Therefore, the final structure ca be constructed:



## Connecting Branch and Undefined Monomers

Branch monomers are used where the branching is regular. Nucleotide polymers have a branch monomer to connect the base to the sugar.

The R1 group of the base is connected to the R3 of the sugar automatically and the monomer is placed inside brackets in the HELM string. e.g. RNA1{R(A)P}$$$$V2.0 Unused R groups are still capped.

Undefined monomers are not connected automatically, but the connections are defined by the user and recorded in the second section of the HELM string. For example:



# Monomer Core principles

## Canonicalization

HELM monomers do not need to be orthogonal to each other. Monomers can contain substructures of other monomers and it is allowable to define a monomer that is a combination of two or more existing monomers.

Example Asparagine can be constructed from Aspartic acid plus a terminal monomer available in the existing HELM monomer set.



If HELM required that every monomer is independent of all others, it would lead to a set that does not reflect the general practice when synthesizing molecules or the processes involved in biological systems.

The consequence of this policy is that HELM cannot be used to check molecules for uniqueness. HELM recommends that uniqueness checking is performed by converting into a small molecule format and using that format’s toolset.

# Monomer substitution (peptides)

A monomer is an artificial construct, so it is necessary to define what is a core structure and what is a substitution. A peptide monomer can be thought of in this way:



* All peptide monomers must have an amine and carboxylic acid. Most will be alpha amino acids, as shown here, but not all.
* There will be a primary sidechain, which often relates to the natural amino acid.
* The primary side chain can be substituted, often by a small range of common small molecules.
* The terminal positions can be substituted, often by a small range of common small molecules.

Allowing substituents in more places leads to a smaller monomer set, but more complex HELM strings. After analysing public data sets, the project decided that terminal substituents should be allowed, but monomers with sidechain substituents should be new monomers.

This means that there are two subsets of peptide monomers:

* Standard backbone peptide monomers – must include at least 2 R groups and can be used to continue the peptide polymer chain.
* Terminal peptide monomers – only have 1 R group and will terminate the polymer chain if used.

These types are not defined in the monomer\_type, but are identified through the number of R groups in the structure.

These rules are outlined in more detail in this section.

## N and C terminal substitutions

*Substitutions at the terminus of a peptide chain should be separate monomers.*

*Monomers used for terminal N and C substitutions will be known as terminal monomers.*

Many peptides have a simple structure, but include substitution at the terminus. Like this:



These substitutions use the backbone R1 or R2 connection points and the terminal monomer ensures that the chain cannot be continued at that point.

C and N substitutions are very common and account for a lot of the variation in public sets. By separating these substitutions from the parent monomers, we can reduce the number of monomer variants.

Terminal monomers can be identified by the fact that they have a R1 or R2 group, but not both.

### Example C substitution

A molecule with a C terminal substitution should be written as separate monomers. For example:



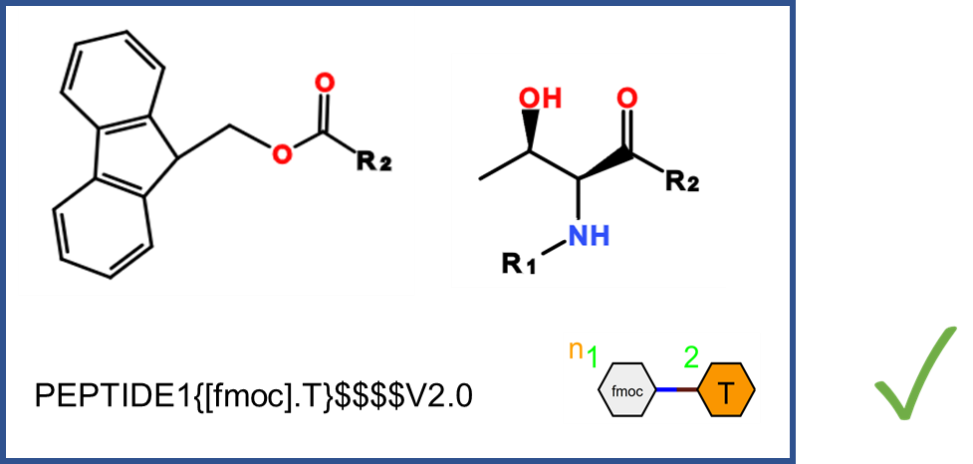
And not



The chain can be continued from the N terminus of the polymer but there is no substitution point available at the C terminus once the ‘am’ monomer has been added.

### Example N substitution

In this example the N terminus is substituted by fmoc. The two monomers should be defined separately as shown below:

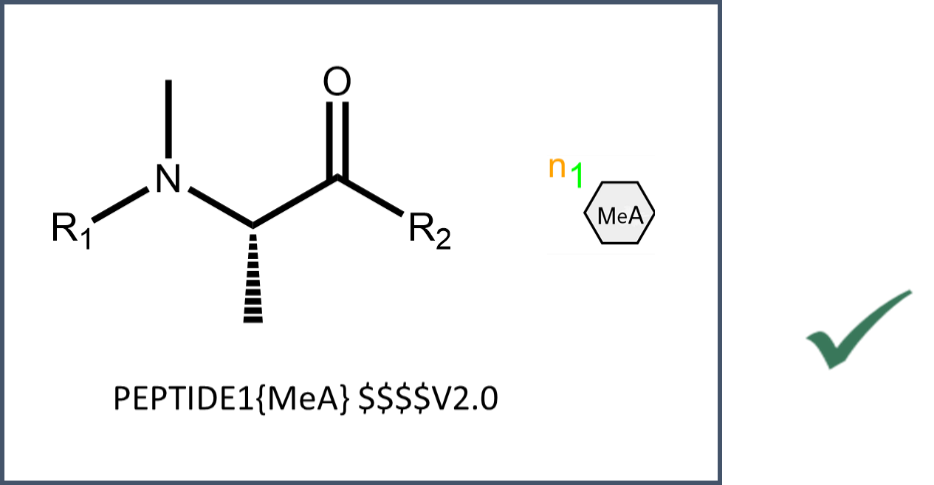


### Midchain N substitution

The primary amine in most amino acids means that it is possible to N substitute and still have a free R group to continue the polymer chain.

Mid-chain N substitution should be part of the monomer structure and not separated into two structures since mid-chain substitution results in complex and difficult to read HELM strings.



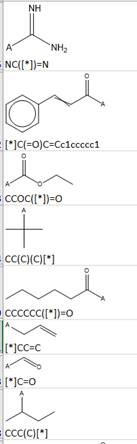


### Peptide terminal monomers vs CHEMs

It is possible to define terminal monomers as CHEMs.

If a structure is available in the monomer.org core peptide monomer set the peptide terminal monomer should instead of the CHEM. If the community agrees that a new terminal monomer is required, it will be added to this set.

*Example common terminal monomers*



## Side chain substitutions

*Terminal substitutions on the side chain should be included in the monomer structure and not be formed by attaching a second monomer.*

Avoid the “hedgehog” approach to HELM where terminal monomers are attached to the main chain simply to encode minor changes to common monomers.



Structural variation should be included in the monomer structure itself.

For example, OMe Tyr should not be Me plus Tyr, but a new monomer with the methyl group already attached.



And not



Similarly, more complex substitutions on known amino acids should still be new monomers.



This rule applies even if the monomer includes a natural amino acid with a standard protecting group.

## Capping Rules - example structures

|  |  |
| --- | --- |
| **Structure** | **Preferred representation** |
|  | One monomer – side chain substitution should be included in the structure |
|  | Two monomers. N capped structure so the cap should be separated out. |
|  | Two monomers. N capped structure so the cap should be separated out. |
|  | Single monomer. The hexyl group does not replace the connection point, so it should not be separated. |
|  | Two monomers. N capped structure so the cap should be separated out. |

## Backbone monomer side-chains may be connected

*Side-chain substitution is allowed for connections between backbone monomers.*

This allows branching and more complex structures to be built in HELM.

These connections can be via any R group. So we can have Cys-Cys connections connected from R3 to R3:



This type of connection can be used to form rings within peptides and cross links between strands. For example this cyclic structure is formed by a R3 to R3 connection but the main chain is a single strand.



You can also have R2 connected to R3, which also allows you to branch a peptide chain.



There is no restriction on connections between backbone monomers.

# Monomer uniqueness rules

The following must be true for monomers to be identical:

* The fully capped molecule structure is the same.
* The uncapped structure is the same. i.e. all R groups including their numbering are in the same position on the molecule.

There is one exception:

* A structure may have additional R groups so long as when those R groups are capped: the capped structure is the same.

The following sections illustrate these rules in more detail.

## Capped molecule equivalence

*R group caps must result in the same capped structure for the monomer to be equivalent.*

HELM does not require all R groups to be attached to other monomers in a HELM string. When R groups are not used, the cap determines the final structure. Therefore, a different cap will result in a different structure.

For example, a monomer with an R3 like this could have different capping groups:



If R1=H, R2=OH and R3=OH then the underlying monomer is tyrosine.

If R1=H, R2=OH and R3=Cl then the underlying monomer is 4-chlorophenyl alanine.

These monomers are not equivalent.

## R group equivalence – number of R groups

*A monomer can be equivalent to another even if it has different numbers of R groups.*



These monomers are equivalent.

This means:

*A curator can add R groups to an existing monomer if the R group cap results in the same fully capped structure.*

Adding R groups will not affect existing use of the monomer since the new R groups will simply be capped as they were not used. However:

*It is not possible to remove R groups from an existing monomer, since you may invalidate HELM strings where those R groups have been used.*

## R group equivalence –R group positions

*Monomers with R groups in different positions are different monomers.*

Even if the capped structure is the same, monomers with the R groups in different positions will connect to other monomers at different locations when used within a HELM string. The automatic connection rules will change the way these monomers are used.

For example, Tyr and Gly would normally connect like this:



PEPTIDE1{Y.G}$$$$V2.0

However, if the R2 and R3 groups are swapped the connected structure would be



also PEPTIDE1{Y.G}$$$$V2.0

It is not possible to change any R group position as all R groups behave in the same way. Non-backbone R groups are not connected automatically, but the connections are still specified within the HELM string, so any change in position will still result in a different chemical structure.

## Recommendations for new monomers

*The number of R groups should be small initially and expanded on demand.*

Since it is possible to add R groups, curator should only add the R groups required at the time of use.

# Monomer naming

In creating these recommendations, where possible, the HELM monomer discussion group has considered the IUPAC Biochemical Nomenclature (White Book) recommendations.

*Biochemical Nomenclature and Related Documents,* 2nd edition, Portland Press, 1992. Edited C Liébecq. [ISBN 1-85578-005-4]

The relevant sections on amino acids can be found at

* <https://www.qmul.ac.uk/sbcs/iupac/AminoAcid/AA17.html>
* <https://www.qmul.ac.uk/sbcs/iupac/AminoAcid/A1819.html#AA182>

However, there are some differences between these HELM recommendations and IUPAC in the interests of compatibility with the HELM notation and readability of the monomer symbol when displayed.

These recommendations are preliminary and there will be cases that are not covered. The team welcome comments and suggestions for improvement.

## General principles

Monomer names are used within the HELM string and also shown in graphical representations of the macromolecule. A meaningful, short name is most useful, like

* Asu(Ph)
* Dab(Bn)

However, the complexity of chemical structures means not all names can be kept short. Meaning should be prioritised over length for medium length names, e.g.

* Lys(CONHPh(4-Me))
* D-aMePhe(4-EtO2H)
* Phe(4-N(EtCl)2)

However, full IUPAC names are discouraged e.g.

* L-Gly(1-carboxymethyl-4-oxoquinolin-3-yl)
* Nva(OMe)-benzothiazol-2-yl

Since the name becomes too long and unwieldy

* In-line HELM should be used for single-use monomers.
* A number sequence, e.g. X001, should be used if either of the following applies:
  + For significant single-use monomers, for example, if they are approved products.
  + In-house, if there is a specific need, for example, to facilitate analysis of a data set.

Vendors may produce short abbreviations for common substances in the future and we would want to adopt them where possible. In-line HEM would enable us to keep our options open for rarely used monomers.

### Allowed Character sets

The monomer symbol is used as an ID and appears in the HELM string. Expert users often read HELM strings and deduce information about the structures of the monomer from the symbol. Nevertheless, HELM’s primary use is as a machine-readable notation. These rules attempt to balance these competing priorities.

Monomer symbols may contain:

* Alphanumeric characters (HELM is case insensitive, but both upper and lower case letters are allowed)
* - and \_

No other characters are allowed.

## Meaningful names

### Natural amino acids

Single letter codes should only be used when there are no modifications or where stated explicitly in this document. Otherwise three letter codes should be used. e.g.

* Phe\_4F and not F\_4F
* Pro\_OBn and not P\_OBn

### **Stereoisomerism**

Where there are no other modifications, single letter codes may be used with stereochemistry identifiers.

L is implied and therefore not written.

Lower case d or dl must be used when the single letter code is used.

|  |  |  |
| --- | --- | --- |
| A | Alanine |  |
| dA | D-Alanine |  |
| dlA | DL-Alanine |  |

Upper case D or DL followed by a ‘–‘ must be used when the three-letter code is used.

|  |  |  |
| --- | --- | --- |
| Phe\_4Cl | 4-Chloro Phenylalanine |  |
| D-Phe\_4Cl | D-4-Chloro Phenylalanine |  |

### N substitution by simple alkyl groups

Where there are no other modifications, single letter codes may be used with simple alkyl N substitutions.

|  |  |  |
| --- | --- | --- |
| MeA | N-Methyl-Alanine |  |
| EtA | N-Ethyl-Alanine |  |

### **Structural** **isomerism**

Beta and gamma amino acids should be prefixed by a lower case b or g followed by the 3 letter code.

|  |  |  |
| --- | --- | --- |
| bAla | Beta alanine |  |
| gGlu | Gamma glutamic acid |  |

### Aldehyde and alcohol analogues

Aldehydes and alcohol analogues of natural amino acids should be indicated via a post script.

|  |  |  |
| --- | --- | --- |
| Leu-al | Leucine-al |  |
| Leu-ol | Leucine-ol | Note that the capping group for R2 is OH. |

### Amine replacement

Replacing the amine group results in a monomer with only one connection point. Therefore all these examples are terminal monomers.

The amine replacement is named before the 3 letter code to differentiate it from a side-chain modification.

|  |  |
| --- | --- |
| Deamino-Phe |  |
| O-Val |  |

### Alkyl rings at the alpha position

Alkyl rings will be named using the prefix Ac followed by the ring size and c since it is an entirely carbon ring. This is in line with the convention adopted by Sugar and Splice.

|  |  |  |
| --- | --- | --- |
| Ac3c | |  | | --- | | C1CC1(C(=O)O)N | |
| Ac5c | [H:1]NC1(CCCC1)C([OH:2])=O |
| Ac6c | C1CCC(CC1)(C(=O)O)N |

### Other monomers

A monomer name should be derived from the largest core monomer that can be identified within it. For example: monomers with 4 substituted phenyl groups should be named in relation to tyrosine and not phenylalanine.

When the natural amino acid is substituted, the structure of the name consists of the following elements:

N\_subs Stereo - parent three-letter code\_backbone/sidechain subs

Examples

|  |  |
| --- | --- |
| **Short name** | **Structure** |
| N\_Bu-D-Phe\_4Cl |  |
| Lys\_5Non |  |
| Pro\_3OBz |  |
| Ser\_Et2ol |  |

## Common names

These are non-systematic names that are in common use. They should be used in preference to the standard rules wherever there is a conflict. Minor variations such as the d form or minor substitutions may use these names as a root.

|  |  |  |  |
| --- | --- | --- | --- |
| **Symbol** | **Smiles** | **Name** | **NaturalAnalog** |
| Aad | [H:1]N[C@@H](CCCC([OH:3])=O)C([OH:2])=O | 2-aminoadipic acid | X |
| Abu | CC[C@H](N[H:1])C([OH:2])=O | 2-aminobutanoic acid | X |
| Aca | CCCCCCCC[C@H](N[H:1])C([OH:2])=O | 2-aminocapric acid | X |
| Aib | CC(C)(N[H:1])C([OH:2])=O | alpha-aminoisobutyric acid (2-aminoalanine) | X |
| Apm | [H:1]N[C@@H](CCCCC([OH:3])=O)C([OH:2])=O | 2-aminopimelic acid | X |
| App | O[C@H](CC([OH:2])=O)[C@H](Cc1ccccc1)N[H:1] | gamma-amino-beta-hydroxybenzenepentanoic acid | X |
| Asu | [H:1]N[C@@H](CCCCCC([OH:3])=O)C([OH:2])=O | 2-aminosuberic acid | X |
| Aze | [H:1]N1CC[C@H]1C([OH:2])=O | 2-carboxyazetidine | X |
| Bal | [H:1]NCCC([OH:2])=O | beta-Alanine | A |
| Bux | O[C@H](CN[H:1])CC([OH:2])=O | 4-amino-3-hydroxybutanoic acid | X |
| Cap | O[C@H]([C@H](CCC1CCCCC1)N[H:1])C([OH:2])=O | gamma-amino-beta-hydroxycyclohexanepentanoic acid | X |
| Cha | [H:1]N[C@@H](CC1CCCCC1)C([OH:2])=O | 3-cyclohexylalanine | A |
| Chg | [H:1]N[C@@H](C1CCCCC1)C([OH:2])=O | (S)-2-amino-2-cyclohexylacetic acid | X |
| Cit | NC(=O)NCCC[C@H](N[H:1])C([OH:2])=O | citrullin | X |
| Cya | OS(=O)(=O)C[C@H](N[H:1])C([OH:2])=O | 3-sulfoalanine | A |
| Dab | [H:1]N[C@@H](CCN[H:3])C([OH:2])=O | 2,4-diaminobutanoic acid | X |
| Dpm | NC(CCCCC(O)=O)(N[H:1])C([OH:2])=O | diaminopimelic acid | X |
| Dpr | [H:1]N[C@@H](CN[H:3])C([OH:2])=O | 2,3-diaminopropanoic acid | X |
| Dsu | NC(CCCC[C@H](N[H:1])C([OH:2])=O)C(O)=O | 2,7-diaminosuberic acid (2,7-diaminooctanedioic acid) | X |
| Edc | CCSSC[C@H](N[H:1])C([OH:2])=O | S-ethylthiocysteine | C |
| Ggu | [H:1]N[C@@H](CCC([OH:2])=O)C([OH:3])=O | gamma-glutamic acid | E |
| Gla | OC(=O)C(C[C@H](N[H:1])C([OH:3])=O)C([OH:2])=O | gamma-carboxyglutamic acid | E |
| Har | NC(N)NCCCC[C@H](N[H:1])C([OH:2])=O | homoarginine | R |
| Hcy | [H:1]N[C@@H](CCS[H:3])C([OH:2])=O | homocysteine | C |
| Hhs | [H:1]N[C@@H](CCc1c[nH]cn1)C([OH:2])=O | homohistidine | H |
| Hpr | [\*]N1CCCC[C@H]1C([\*])=O | (S)-piperidine-2-carboxylic acid | X |
| Hse | OCC[C@H](N[H:1])C([OH:2])=O | homoserine | S |
| Hyl | OC(CC[C@H](N[H:1])C([OH:2])=O)CN[H:3] | 5-hydroxylysine | K |
| Hyp | O[C@@H]1C[C@H](N([H:1])C1)C([OH:2])=O | 4-hydroxyproline | P |
| Iva | CC[C@](C)(N[H:1])C([OH:2])=O | isovaline | V |
| Mhp | CC1CN([H:1])[C@@H](C1O)C([OH:2])=O | 4-methyl-3-hydroxyproline | P |
| Nal | [H:1]N[C@@H](Cc1ccc2ccccc2c1)C([OH:2])=O | 3-naphthylalanine | A |
| Nle | CCCC[C@H](N[H:1])C([OH:2])=O | norleucine | L |
| Nva | CCC[C@H](N[H:1])C([OH:2])=O | norvaline | V |
| Oic | [H:1]N1C2CCCCC2C[C@H]1C([OH:2])=O | 2-carboxyocthydroindole | X |
| Orn | [H:1]N[C@@H](CCCN[H:3])C([OH:2])=O | L-ornithine | K |
| Pen | CC(C)(S[H:3])[C@H](N[H:1])C([OH:2])=O | penicillamine (3-mercaptovaline) | V |
| Phg | [H:1]N[C@H](C([OH:2])=O)c1ccccc1 | 2-phenylglycine | G |
| Pip | [H:1]N1CCCC[C@H]1C([OH:2])=O | (S)-piperidine-2-carboxylic acid | X |
| Pqa | [H:1]N1CCN(CC1)c1ccc2ccn(CC([OH:2])=O)c(=O)c2c1 | Piperazine quinazolinone acetic acid, 2-(4-oxo-6-piperazin-1-yl-quinazolin-3-yl)acetic acid | X |
| Sar | CN([H:1])CC([OH:2])=O | sarcosine (N-methylglycine) | G |
| seC | [SeH]C[C@H](N[H:1])C([OH:2])=O | SelenoCysteine | C |
| Spg | [H:1]NC1(CCCC1)C([OH:2])=O | 1-amino-1-carboxycyclopentane | X |
| Sta | CC(C)C[C@H](N[H:1])[C@@H](O)CC([OH:2])=O | statin (4-amino-3-hydroxy-6-methylheptanoic acid) | X |
| Thi | [H:1]N[C@@H](Cc1cccs1)C([OH:2])=O | 3-thienylalanine | A |
| Tic | [H:1]N1Cc2ccccc2C[C@H]1C([OH:2])=O | 1,2,3,4-tetrahydro-3-isoquinolinecarboxylic acid | X |
| Tle | CC(C)(C)[C@H](N[H:1])C([OH:2])=O | 3-methylvaline | V |
| Tml | C[N+](C)(C)CCCC[C@H](N[H:1])C([OH:2])=O | epsilon-N-trimethyllysine | K |
| Tza | [H:1]N[C@@H](Cc1cscn1)C([OH:2])=O | 3-thiazolylalanine | A |
| Wil | [H:1]N[C@@H](Cn1ccc(=O)[nH]c1=O)C([OH:2])=O | alpha-amino-2,4-dioxopyrimidinepropanoic acid | X |

## Monomers with fewer than 2 backbone connection points

These are terminal monomers but have established abbreviations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Symbol** | **Smiles** | **Name** | **NaturalAnalog** |
| ac | CC([OH:2])=O | N-Terminal Acetic Acid | X |
| am | N[H:1] | C-Terminal amine | X |
| Bua | CCCC([OH:2])=O | butanoic acid | X |
| fmoc | [OH:2]C(=O)OCC1c2ccccc2-c2ccccc12 | fmoc N-Terminal Protection Group | X |
| Glc | OCC([OH:2])=O | glycolic acid | X |
| Glp | [OH:2]C(=O)[C@@H]1CCC(=O)N1 | pyroglutamic acid | E |
| Hiv | CC(C)[C@H](O)C([OH:2])=O | 2-hydroxyisovaleric acid | X |
| Hva | CCC[C@H](O)C([OH:2])=O | 2-hydroxypentanoic acid | X |
| Lac | C[C@H](O)C([OH:2])=O | lactic acid | X |
| Maa | [OH:2]C(=O)CS[H:3] | mercaptoacetic acid | X |
| Mba | CC[C@H](S[H:3])C([OH:2])=O | mercaptobutanoic acid (GMBA) | X |
| Mpa | C[C@H](S[H:3])C([OH:2])=O | mercaptopropanoic acid | X |
| Nty | Oc1ccc(C[C@H](N([H:1])N(=O)=O)C([OH:2])=O)cc1 | nitrotyrosine | Y |