***Actions***

|  |  |  |
| --- | --- | --- |
| **Who** | **What**  | **By when** |
| **All** | Review the agreed principles and email Claire stating whether you agree or would like to see changes.  | 6th November 2018 |
| **All** | Should a public monomer set include numbered monomers – or should we discourage this and either require a (semi) systematic name or ask users to use in-line HELM? | 6th November 2018 |

# HELM Monomer Discussion Notes

The notes below capture the decisions and principles that are the output from the monomer meeting. See slides for the topic background and questions.

## Principles (tentatively) agreed

## Canonicalization

The set we are creating will not be bound by rules on canonicalization.

* A separate set could be created whose purpose is for canonicalization and computational analysis but work to do this is outside the scope of this discussion.

## Substitution

Capping monomers are defined as monomers with only one connection point.

Capping monomers should only be connected to backbone monomers at the R1 or R2 points. This will terminate the polymer chain.

Capping monomers should be as small a functional group as possible and should not include functionality that could be separated into an amino acid. For example:

 and not 

Capping monomers should be the same polymer type as the main chain of the polymer to keep the HELM string as readable as possible. It is acceptable to have the same structure in multiple polymer types.

Backbone monomers must have at least two connection points.

Only backbone or branch monomers may be connected to additional connection points. Capping monomers may only be connected to the R1 and R2 connection points.

When non-standard connection points are substituted, the connection will create a branch. This branch will be represented as a new polymer chain in the HELM string.

Non-terminal N substitutions and substitutions elsewhere on the amino acid, L, D and positional isomers should be defined as new monomers.

## Monomer naming

Am agreed naming approach would reduce variation between sets and aid information exchange between groups using HELM. The following are the HELM recommendations (containing what we have so far and very much a draft).

### Natural amino acids

Single letter codes can be used alone or alongside the following, limited, modifications:

* l, d and dl amino acids
* Me, Et, Ph for N substitution

Otherwise three letter codes should be used.

### **Stereoisomerism**

L is implied and therefore not written.

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

Upper case D or DL when the three-letter code is used.

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

Beta and gamma amino acids should be prefixed by a lower case b or g.

The following abbreviations (when applicable) followed by the three letter code.

* Iso (IUPAC recommended)
* nor (NB not IUPAC recommended)
* allo
* tert

Unless an alternative abbreviation is in common use.

## Common non-natural amino acids

See appendix for common names that should be used.

### 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:

Stereo – structural isomer - (N subs) parent three-letter code (backbone/sidechain subs)

* E.g. d-N(Bu)Phe(4-Cl)

## Common names to be used with HELM monomers

This is a proposed initial set. Please send Claire monomers you think should be added.

|  |  |  |  |
| --- | --- | --- | --- |
| **symbol** | **smiles** | **name** | **naturalAnalog** |
| Bal | [H:1]NCCC([OH:2])=O | beta-Alanine | A |
| Cha | [H:1]N[C@@H](CC1CCCCC1)C([OH:2])=O | 3-cyclohexylalanine | A |
| Cya | OS(=O)(=O)C[C@H](N[H:1])C([OH:2])=O | 3-sulfoalanine | A |
| Nal | [H:1]N[C@@H](Cc1ccc2ccccc2c1)C([OH:2])=O | 3-naphthylalanine | A |
| Thi | [H:1]N[C@@H](Cc1cccs1)C([OH:2])=O | 3-thienylalanine | A |
| Tza | [H:1]N[C@@H](Cc1cscn1)C([OH:2])=O | 3-thiazolylalanine | A |
| Edc | CCSSC[C@H](N[H:1])C([OH:2])=O | S-ethylthiocysteine | C |
| Hcy | [H:1]N[C@@H](CCS[H:3])C([OH:2])=O | homocysteine | C |
| seC | [SeH]C[C@H](N[H:1])C([OH:2])=O | SelenoCysteine | 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 |
| Phg | [H:1]N[C@H](C([OH:2])=O)c1ccccc1 | 2-phenylglycine | G |
| Sar | CN([H:1])CC([OH:2])=O | sarcosine (N-methylglycine) | G |
| Hhs | [H:1]N[C@@H](CCc1c[nH]cn1)C([OH:2])=O | homohistidine | H |
| Hyl | OC(CC[C@H](N[H:1])C([OH:2])=O)CN[H:3] | 5-hydroxylysine | K |
| Orn | [H:1]N[C@@H](CCCN[H:3])C([OH:2])=O | L-ornithine | K |
| Tml | C[N+](C)(C)CCCC[C@H](N[H:1])C([OH:2])=O | epsilon-N-trimethyllysine | K |
| Nle | CCCC[C@H](N[H:1])C([OH:2])=O | norleucine | L |
| Hyp | O[C@@H]1C[C@H](N([H:1])C1)C([OH:2])=O | 4-hydroxyproline | P |
| Mhp | CC1CN([H:1])[C@@H](C1O)C([OH:2])=O | 4-methyl-3-hydroxyproline | P |
| Har | NC(N)NCCCC[C@H](N[H:1])C([OH:2])=O | homoarginine | R |
| Hse | OCC[C@H](N[H:1])C([OH:2])=O | homoserine | S |
| Iva | CC[C@](C)(N[H:1])C([OH:2])=O | isovaline | V |
| Nva | CCC[C@H](N[H:1])C([OH:2])=O | norvaline | V |
| Pen | CC(C)(S[H:3])[C@H](N[H:1])C([OH:2])=O | penicillamine (3-mercaptovaline) | V |
| Tle | CC(C)(C)[C@H](N[H:1])C([OH:2])=O | 3-methylvaline | V |
| 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 |
| 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 |
| Cit | NC(=O)NCCC[C@H](N[H:1])C([OH:2])=O | citrullin | X |
| 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 |
| Oic | [H:1]N1C2CCCCC2C[C@H]1C([OH:2])=O | 2-carboxyocthydroindole | 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 |
| 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 |
| Tic | [H:1]N1Cc2ccccc2C[C@H]1C([OH:2])=O | 1,2,3,4-tetrahydro-3-isoquinolinecarboxylic acid | X |
| 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 capping monomers but have established abbreviations.

|  |  |  |  |
| --- | --- | --- | --- |
| Glc | OCC([OH:2])=O | glycolic acid | X |
| 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 |
| Bua | CCCC([OH:2])=O | butanoic acid | X |
| Glp | [OH:2]C(=O)[C@@H]1CCC(=O)N1 | pyroglutamic acid | E |
| Nty  | Oc1ccc(C[C@H](N([H:1])N(=O)=O)C([OH:2])=O)cc1 | nitrotyrosine | Y |
| ac | CC([OH:2])=O  | N-Terminal Acetic Acid  | X |
| am  | N[H:1]  | C-Terminal amine  | X |
| fmoc | [OH:2]C(=O)OCC1c2ccccc2-c2ccccc12  | fmoc N-Terminal Protection Group | X |