A Convention for Peptoid Monomer Naming

Hamish W A Swanson, King Hang Aaron Lau, Tell Tuttle

Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, UK.

1. Abstract

In this document we describe a naming convention for peptoid monomers recently devised at the University of Strathclyde to address the lack of a consistent approach to this topic within the field. Our method is simplified and targeted at assisting those new to the research space, with a view to streamlining communication between those familiar with peptoids and collaborators in adjacent fields. To do this we have linked our convention to pre-existing amino acid nomenclature which is widely taught at undergraduate level in both chemistry and related disciplines.

2. Motivation

Originally, our motivation to develop a peptoid naming scheme was to meet the informational requirements of computational chemistry software packages and file-formats, in which limits on maximum lengths of residue names exist and for which pre-existing naming conventions within the canon are incongruous. One example being the protein data bank (PDB) file format which has specific formatting requirements and can be problematic if residue names are too long. In this effort it soon came to our attention that several naming methodologies seem to be competing within the peptoid research space and that a defined convention would be useful.

Language is, after all, an intrinsic barrier to those who do not speak or read it. In this instance, those who may not *speak the language* are those that lack experience with peptoid nomenclature interpretation; to which the diversity of monomer naming schemes is an initial barrier to understanding, comprehension and communication. In a sense, the peptide research space is only one isomeric step away from that of peptoids and so we were inspired to ground our convention in the pre-existing knowledge of amino acid chemistry, which is widely taught across institutions at the undergraduate level. We have found the following naming convention to be useful in collaborations and attribute this to the inclusion of the single letter code (SLC) of amino acids which draws upon a pre-existing and widely accepted knowledge base.

Unlike natural amino acids, which are limited in their number, the monomer space for peptoids is vast, with more than 250 having been experimentally demonstrated.[1] It would therefore be naïve to think that any one naming convention could capture this huge chemical space in a human readable manner. Despite this fact, we find the method outlined herein to be extensible to a wide breadth of peptoid monomers which are available to the community; thus we believe our approach is best equipped to capture the essence of peptoid diversity which is fundamental to their interesting chemical properties.[2]

Lastly, like all languages, those which are living are not static: therefore, we hope that this convention, which we refer to colloquially as the 'Glasgow Convention', will evolve and be developed further by those working in the peptoid research space. Therefore, we regard this as a starting point and not a fixed set of rules.

3. Monomer Naming Rules

The characteristic peptoid 'N' is retained at the start of any peptoid monomer name. This is to remain consistent with previous naming conventions that have gone before; it is familiar to those within the research space and immediately distinguishes the monomer in question as being an N-substituted glycine.

The monomer name is then tied to pre-existing amino acid chemistry through the incorporation of the most closely related amino acid SLC in the lower-case. Additional lower-case letters are then used to encode further information about the monomer's characteristics. In a sense this method informs a reader of the *degree of chemical difference* from amino acid building blocks; such that if only the SLC is used then it is a direct analogue, while if multiple characters are used then it can be regarded as more chemically distinct from its root amino acid.

The methodology is composite, and information is only included *if it is there*; meaning that if a given functionality is not present then an encoding placeholder is not used. Users should strive to keep names as short as possible for simplicity (*i.e.*, four to five characters in length).

Peptoid Termini

Termini modification, such as acylation, is dealt with in the same manner as in the peptide research space (*e.g.*, Ac- at the start of the sequence). Amidation of the C-termini in peptoids is ubiquitous and so this is regarded as the default unless otherwise stated.[3, 4]

Character 1

The SLC of the *most closely related* amino acid is used here. For example, if the monomer is a direct analouge of lysine, then it would be referred to as Nk. If a monomer is unlike any amino acid, other root letters can be used, for one we would suggest the use of Greek letters for novel or entirely synthetic sidechain and backbone families.

Character 2

- Often peptoid monomers are *pseudo* isomers of amino acids, deviating in the length of the methylene linker.[5, 6] Thus the second position is reserved to describe the sidechain methylene length (*e.g.*, n = no linker, m = methyl, e = ethyl, p = propyl, b = butyl, beyond this numbers are used, such as pentyl as Na5). A summative approach is taken with branches being described by the sum of methylene linker carbons.
- Sidechain ethers can be described here by a letter 'O' with methionine as the amino acid 'root' chemistry. If the sidechain is a sulfoxide methionine analouge, then this would be [SO], *e.g.*, Nm[SO].

Character 3

Further nuance may be captured in the third position:

- If the C_β carbon is *gem*-dimethyl, then a letter *g* is used.
- If an allyl group is present in the sidechain this is indicated with a [yl].
- If an aromatic *o*-nitrobenzyl group protected cysteine analogue is used this is [onb].
- If the sidechain is a naphthalene moiety this is represented as [naph].
- When azide or alkyne functionality is included in sidechains this is represented as lysine root chemistry and [N_3], or an alanine root chemistry and [alk], respectively.
- For substituted amine sidechains the degree of substitution can be described (*e.g.,* s = secondary, t = tertiary and q = quaternary), see the work of Wijaya *et al.*[7]
- Sidechain stereochemistry can be denoted with a letter *r* or *s* as required and this can be put in curved brackets if preferred (*e.g.*, Nfe(S) or Nfes).
- Benzene substituents and their positions within the ring can be included here (*e.g.*, F = fluorine, Cl = chlorine, Br = bromine, I = iodine, m = methyl, Om = methoxy). For clarity a square bracket is used to separate this from the rest of the name (*e.g.*, Nfe[4Br]).

Character 4

Extra chemical differences can be stored in the fourth position:

- An aliphatic *cyclo*-hexyl group is referred to as h = cyclohexyl.
- The length of the methylene group on a substituted amine side chain can also be captured here (*e.g.*, m = methyl, e = ethyl, p = propyl, b = butyl etc.).
- If the sidechain has ethylene-glycol (EG) units, then these are enclosed in square brackets with the number of units in subscript *e.g.*, [EG_3].

4. Additional Information

Use of our nomenclature in the same manner as previous methodologies is encouraged when describing sequences (*e.g.*, Npm-Nae-Npm, would instead be Nf-Nke-Nf, Figure 1). In a computational context it is useful at times to 'fill-in' blank spaces with zeros and omit the N prefix to make all names a consistent length. For example, the analouge of Nf would instead be F000.

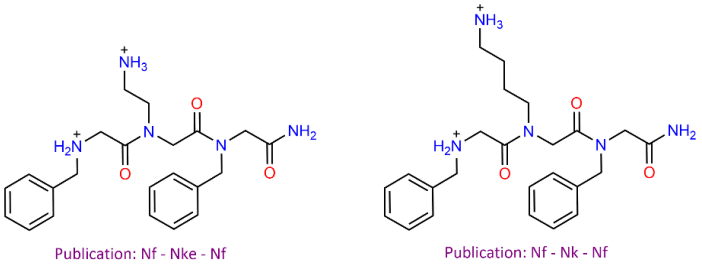

Figure 1. Interpretation of naming scheme for two short peptoid trimers; showing how termini states are implicitly defined unless otherwise stated.

In cases where there is potential overlap with other approaches care must be taken. For example, Nae in previous work refers to N-(2-aminoethyl) glycine,[8] while in this convention it would refer to N-ethyl glycine. Here, other encoding may be needed, e.g., capitalization of the letter at the second position for distinction (*e.g.*, Nae becomes NaE). We leave handling of such instances to user discretion.

An added benefit of this method is that it allows for the organisation of monomer types into families, such as those with chemistry *like* phenylalanine (Figure 2). Previous approaches have not been able to organise functionality in this way and thereby lack the possibility of quick comprehension of sequence chemical character from viewing a sequence (*e.g.*, a series of 'Nfx' monomers would comprise a hydrophobic domain of a sequence).

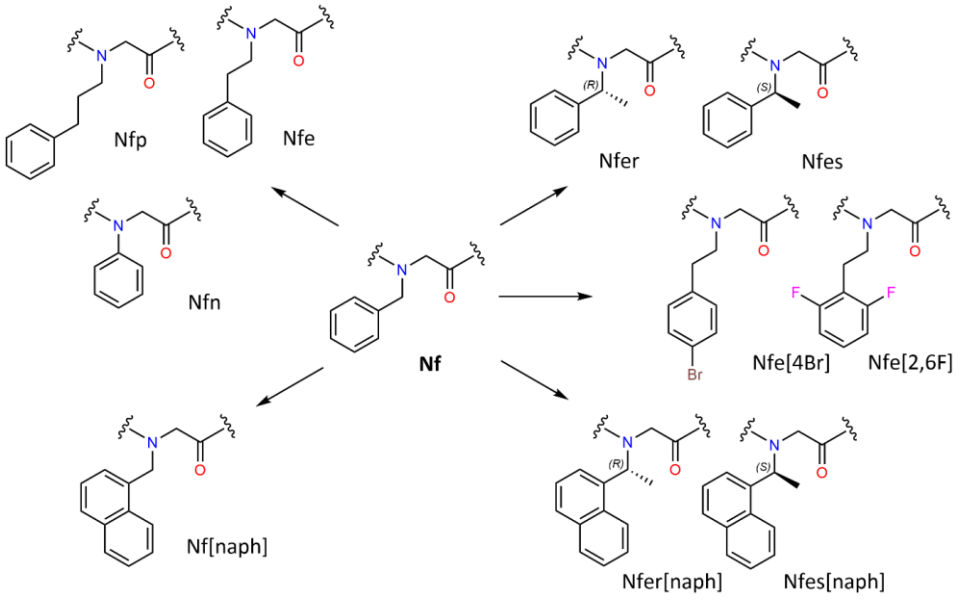

Figure 2. Illustration of the *degree of difference* within a peptoid family centred around the phenylalanine analogue Nf all of which may be captured by a small number of additional letters using our convention.

Finally, this method is subject to change as discussed in section 2. For this reason, we refer the reader to our Wikidot page (http://peptoid-naming-convention.wikidot.com/) where changes or updates to this framework will be outlined.

5. References

1. A. S. Culf and R. J. Ouellette, *Molecules*, 2010, **15**, 5282-5335.
2. Hamish W. A. Swanson, Marwa El Yaagoubi, Aquib Jawed, Varun Saxena, Martyn G.L. Merrilees, Tell Tuttle, Lalit M. Pandey, Ian W. Hamley, King Hang Aaron Lau, in *Supramolecular Nanotechnology: Advanced Design of Self-Assembled Functional Materials*, ed. Martin Conda-Sheridan, Omar Azzaroni, Wiley-VCH GmbH2023, vol. 3, ch. 36, pp. 969-999.
3. R. N. Zuckermann, J. M. Kerr, S. B. H. Kent and W. H. Moos, *J. Am. Chem. Soc.*, 1992, **114**, 10646-10647.
4. A. M. Clapperton, J. Babi and H. Tran, *ACS Polym Au.*, 2022, **2**, 417-429.
5. J. R. B. Eastwood, E. I. Weisberg, D. Katz, R. N. Zuckermann and K. Kirshenbaum, *Pept. Sci.*, 2023.
6. D. Kalita, B. Sahariah, S. P. Mookerjee and B. K. Sarma, *Chem. Asian J.*, 2022, **17**.
7. A. W. Wijaya, A. I. Nguyen, L. T. Roe, G. L. Butterfoss, R. K. Spencer, N. K. Li and R. N. Zuckermann, *J. Am. Chem. Soc.*, 2019, **141**, 19436-19447.
8. K. T. Nam, S. A. Shelby, P. H. Choi, A. B. Marciel, R. Chen, L. Tan, T. K. Chu, R. A. Mesch, B. C. Lee, M. D. Connolly, C. Kisielowski and R. N. Zuckermann, *Nat. Mater.*, 2010, **9**, 454-460.