Toward Canonical HELM Representations for Sequence-Level Peptide Registration

Jan Christopherson[1], Jozsef Kozma[2], Peter Szabad[2], Csaba Peltz[2]
[1]Certara, Radnor, PA 19087; [2]Certara, 1138 Budapest, Hungary; e-mail: csaba.peltz@certara.com

Background and Objective

HELM was built for interoperable peptide representation, not to generate a canonical form: the same molecule can have more than one valid HELM string, arising from ordinary choices like start-point choice, branch/monomer ordering, and R-group assignment order.

As a consequence, practitioners distrust HELM-string equality for duplicate checks and fall back on chemical-structure matching instead, which scales poorly for longer sequences.

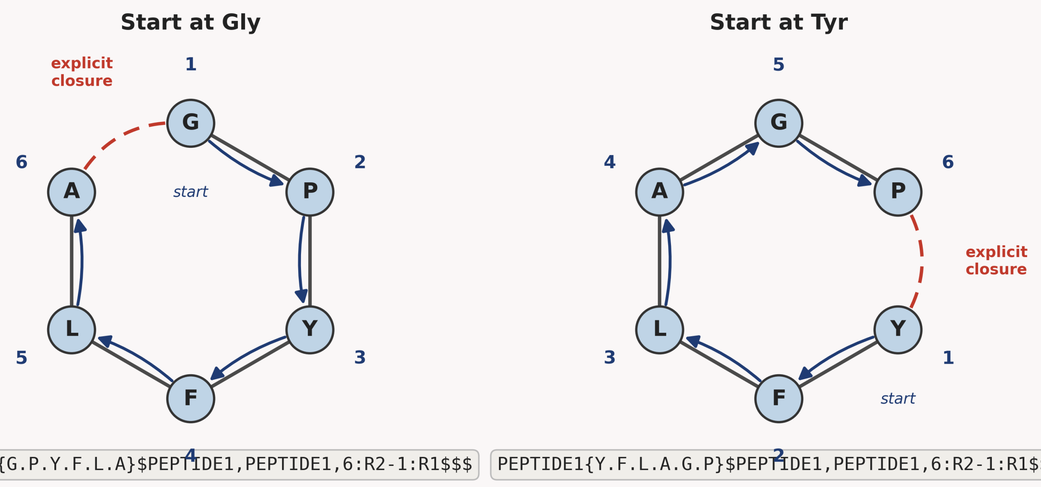
Methods

We illustrate both ambiguity and round-trip information loss using three non-proprietary peptide constructs as examples spanning representative modification types: a head-to-tail cyclic peptide, a branched lipidated peptide, and disulfide-bridged peptide (see Table 1).

Every ambiguity and round-trip loss in Table 1 can turn into a concrete registration or search failure the moment a HELM string moves between systems, tools, or organizations.

The failure mode is easy to imagine; a partner CRO supplies a peptide with its branch polymer serialized differently; a naive string-equality check reports a new entity, fragmenting assay and batch history across two records for one molecule.

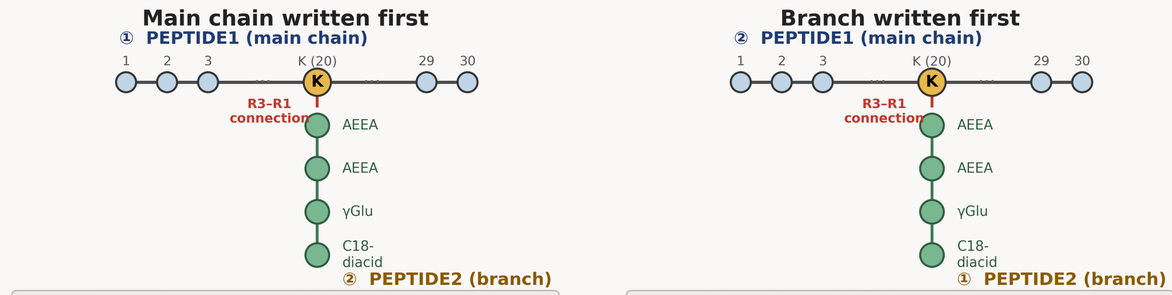
EX-1: same head-to-tail cyclic peptide, two equally valid HELM strings



`PEPTIDE1{G.P.Y.F.L.A}$PEPTIDE1,PEPTIDE1,6:R2-1:R1$$$` `PEPTIDE1{Y.F.L.A.G.P}$PEPTIDE1,PEPTIDE1,6:R2-1:R1$$$`

Same physical ring, same direction of travel (never reversed — that would be a different, retro-connected molecule). Only the chosen start residue differs — and notice the SAME connection text “6:R2-1:R1” closes a DIFFERENT physical bond in each.

EX-2: same molecule, branch/main-chain serialization order swapped



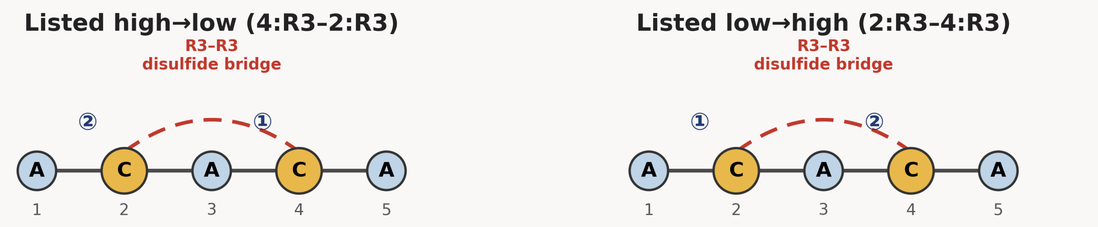
`PEPTIDE1{...K...}|PEPTIDE2{[AEEA].[AEEA].[gGlu].[C18diacid]}$PEPTIDE1,PEPTIDE2,20:R3-1:R1$$$`

`PEPTIDE2{[AEEA].[AEEA].[gGlu].[C18diacid]}|PEPTIDE1{...K...}$PEPTIDE2,PEPTIDE1,1:R1-20:R3$$$`

Same backbone, same branch, same R3–R1 connection — only which block is written first (and the matching order in the connection table) differs. Both strings describe the identical molecule and both pass the HELM spec.

In the lipidated-branch example, terminal capping fidelity — whether a terminus is free or capped (free amine vs. acetyl, free acid vs. amide) — is a common casualty of converting to atomistic structures and back, since these are small details that are easy to drop during atomistic re-perception.

EX-3: same disulfide-bridged peptide, bridge endpoints listed either order



`PEPTIDE1{A.C.A.C.A}$PEPTIDE1,PEPTIDE1,4:R3-2:R3$$$`

`PEPTIDE1{A.C.A.C.A}$PEPTIDE1,PEPTIDE1,2:R3-4:R3$$$`

A Cys–Cys disulfide is a side-chain-to-side-chain bond, not a directional backbone bond like EX-1's — so there's no inherent order to state its two ends. Both strings describe the identical molecule and both pass the HELM spec.

In the disulfide-bridge example, the bond between two cysteine side chains is undirected — unlike a backbone bond, there's no chemically preferred order for stating its two ends, so listing the connection as “4:R3-2:R3” or “2:R3-4:R3” both correctly describe the identical molecule.

Linker-regiochemistry ambiguity can make a similarity search store a near-neighbor rather than an exact match — manufacturing spurious “novel” hits, or missing genuine distinctions that should have been flagged.

Every ambiguity above is a way that matching can silently fail — without an error, without a flag. It shows up as a compound that looks new when it isn't, or two compounds that look the same when they aren't.

Table 1: Overview

Issue Source	Category	Example	What breaks	Consequence	Our Progress
Start-point choice	Ambiguity	EX-1 (head-to-tail cyclic hexapeptide)	Any residue as ring start, same direction	False dedup negative	✅ Lexicographical approach
Branch/monomer order	Ambiguity	EX-2 (lipidated Lys-branch peptide)	Branch listed before or after main chain	False dedup negative	✅ Morgan-esque canonicalization of monomers
R-group assignment	Ambiguity	EX-2 (lipidated Lys-branch peptide)	Multiple valid assignments per bond	Non-unique strings	
Terminal capping	Round-trip loss	EX-2 (lipidated Lys-branch peptide)	Cap details dropped	Incomplete record	*In progress*
Side-chain bridge listing order	Ambiguity	EX-3 (disulfide-bridged pentapeptide)	Bridge ends listed high→low or low→high	False dedup negative	✅ Minimize starting position

Results

Four of the five failure modes stem from HELM-string ambiguity: start-point choice, branch/monomer order, R-group assignment, and side-chain bridge listing order. The fifth reflects genuine round-trip information loss. Our canonicalization approach already resolves the ambiguity cases; closing the round-trip gap is in progress.

Conclusions

This is presented as an open problem, not a solved one. A canonicalization standard for HELM would need a defined traversal rule, a consistent ordering convention for branches and connection points, and an explicit stereo-encoding convention for modified residues and new bonds.

We have an implementation that deals with the problems above: canonicalizing traversal and branch/connection order. It is a first step towards a working canonicalization approach for HELM-based registration, and we are inviting everyone to that conversation. Canonicalization also depends on resolving overlapping monomer definitions across libraries, since even a canonical string is only as trustworthy as the monomer identities it references.

References
- Pistoia Alliance — HELM notation overview (pistoiaalliance.org/helm-notation).
- Milton et al., “HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation,” J. Chem. Inf. Model. 2012.
- Pistoia Alliance HELM Confluence — Complex monomer connections; Annotations and how to use them.
- Wikipedia — Hierarchical editing language for macromolecules.
- Lau et al., “Discovery of the Once-Weekly GLP-1 Analogue Semaglutide,” J. Med. Chem. 2015 (lipidation-strategy background).