The Journal of Open Source Software

p2smi: A toolkit enabling SMILES generation and property analysis for noncanonical and cyclized peptides

Aaron L. Feller [1¶] **and Claus O. Wilke** [1,2]

1 Department of Interdisciplinary Life Sciences, The University of Texas at Austin, Austin, TX, United States **2** Department of Integrative Biology, The University of Texas at Austin, Austin, TX, United States ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Converting peptide sequences into useful representations for downstream analysis is a common step in computational modeling and cheminformatics. Furthermore, peptide drugs (e.g. Semaglutide, Degarelix) include chemistries beyond natural amino acids and standard backbone structure. Common modifications used include noncanonical amino acids, alternate stereochemistry (D- vs L- amino acids), modified chemistrie such as N-methylation and PEGylation. Here we present p2smi, a Python toolkit with CLI, designed to facilitate the generation of drug-like peptides and their conversion into chemical SMILES strings. By supporting peptide cyclizations, unnatural amino acid incorporation, and chemical modifications, p2smi enables generation of accurate SMILES strings for drug-like peptides, providing a missing link for computational modeling and cheminformatics analyses. The toolkit offers functionalities for chemical modification of peptides and for calculation of molecular properties such as hydrophobicity, topological polar surface area, molecular weight, and adherence to Lipinski's rules for drug-likeness.

Statement of need

The development of p2smi was driven by the need to generate large-scale datasets of drug-like peptide SMILES strings for pretraining transformer-based models to predict membrane permeation from chemical structure (Feller & Wilke, 2025). Built on the core concepts from the CycloPs (Duffy et al., 2011) method for FASTA-to-SMILES conversion, p2smi has evolved into a stand-alone resource to support peptide-focused machine learning pipelines and peptide design workflows. While several bioinformatics toolkits exist for chemical representation and cheminformatics workflows (*ChemAxon*, 2025; Cock et al., 2009; Landrum, 2025; O'Boyle et al., 2011; OpenEye, 2025);, many face limitations such as proprietary licensing and lack in ability to interpret or encode noncanonical amino acids (NCAAs). These constraints limit high-throughput application of sequence generation and conversion, especially for drug-like peptides containing diverse stereochemistries. In, addition, there are several python tools that focus on structure generation and cyclization (Tien et al., 2013; Yang et al., 2025), however, these are not able to incorporate all necessary modifications. We used p2smi to build a dataset of 10M peptides with NCAAs, backbone modifications, and cyclizations for pretraining a chemical language model (Feller & Wilke, 2025). To support the community, have made p2smi available as an open source package on PyPI, offering both command-line tools and Python functions for seamless integration into larger workflows.

The Journal of Open Source Software

Features

p2smi offers five core command-line tools to support peptide sequence generation, conversion, modification, and analysis:

- **generate-peptides:**
  Generates random peptide sequences with customizable parameters; number of peptides, minimum and maximum length, percentage of unnatural amino acids, rate of D-stereochemistry, and cyclization types (randomly chosen). Currently accommodates over 100 unnatural amino acid residues described in SwissSidechain (Gfeller et al., 2012).
  Input: Settings and output filename.
  Output: FASTA file with expanded single character notation.

- **fasta2smi:**
  Converts peptide sequences from FASTA format (with the expanded set of NCAAs) into SMILES notation. Conducts cyclization reactions from notation in the FASTA header, supporting five types of cyclization reactions; disulfide bonded, head-to-tail, sidechain-to-sidechain, sidechain-to-head, and sidechain-to-tail.
  Input: Protein FASTA file. Optional FASTA header notation for cyclization reaction.
  Output: File in novel .p2smi format that includes the single character amino acid representation, type of cyclization reaction, and the resulting SMILES string.

- **modify-smiles:**
  Applies N-methylation and PEGylation to existing SMILES strings. Rates of modification are defined by CLI arguments with peptides and sites randomly selected. Changes are recorded when input is in the .p2smi format.
  Input: Text file with single SMILES per line or .p2smi file.
  Output: Single SMILES per line or .p2smi format (if input is .p2smi).

- **smiles-props:**
  Computes molecular properties from SMILES strings including: molecular weight, TPSA, MolLogP, hydrogen donor/acceptor count, rotatable bond count, ring count, fraction Csp3, heavy atom count, formal charge, molecular formula, and compliance with Lipinski's rules.
  Input: Text file with single SMILES per line or .p2smi format.
  Output: Text file with JSON formatted dictionary of properties.

- **synthesis-check:**
  Synthetic feasibility of natural peptides including several forbidden motifs (N/Q at N-terminus, proline/glycine runs, DG/DP motifs, cysteine count, terminal P/C), a maximum length restriction, hydrophobicity check, and minimum charge distribution.
  Input: Protein FASTA file.
  Output: Protein FASTA file with modified header (PASS/FAIL).

For detailed usage instructions and options for each command, users can append the –help flag to any command (e.g., generate-peptides –help). This will provide guidance on the command's functionality and available parameters.

State of the field

In the realm of peptide informatics, several tools have been recently developed to facilitate the analysis and representation of peptides, particularly those incorporating NCAAs and complex modifications including cyclization. Notably, pyPept (Ochoa et al., 2023), PepFuNN (Ochoa & Deibler, 2025), and cyclicpeptide (Yang et al., 2025) have emerged as significant contributions in this area.

pyPept is a Python library that generates 2D and 3D representations of peptides. It converts se-

quences from formats like FASTA, HELM, or BILN into molecular graphs, enabling visualization and physicochemical property calculations. Notably, pyPept allows customization of monomer libraries to accommodate a wide range of peptide modifications. It also offers modules for rapid peptide conformer generation, incorporating user-defined or predicted secondary structure restraints, which is valuable for structural analyses.

PepFuNN is an open-source Python package designed to explore the chemical space of peptide libraries and conduct structure–activity relationship analyses. It includes modules for calculating physicochemical properties, assessing similarity using various peptide representations, clustering peptides based on molecular fingerprints or descriptors, and designing peptide libraries tailored to specific requirements. Additionally, PepFuNN provides tools for extracting matched pairs from experimental data, aiding in the identification of key mutations for subsequent design iterations.

The cyclicpeptide package provides a unified framework for converting between cyclic peptide sequences and structures, aligning cyclic peptides via graph methods, and analyzing their properties to support drug design. It supports multiple cyclization types and monomer libraries, validates its conversions on large cyclic peptide datasets with high accuracy and stability, and enables efficient cyclic peptide generation. By integrating these modular tools, it fills a gap in peptide informatics by facilitating standardized representations and transformations specifically for cyclic peptides, complementing existing tools focused more on linear peptides or structural analyses.

While these tools offer valuable capabilities, they are not specifically designed for the direct conversion of drug-like peptides into SMILES strings, a functionality central to the initial use-case for p2smi of generating a large-scale database. Rather, these recent additions in the field focus on structural representation, analysis, and structure–activity relationship studies of peptides, complementing the sequence-to-SMILES conversion capabilities provided by p2smi.

Code availability

We have provided p2smi as a pip-installable package, available on PyPI at https://pypi.org/project/p2smi. The source code, including documentation and example notebooks, is openly available on GitHub at https://github.com/aaronfeller/p2smi.

Data availability

The dataset of 10M cyclic peptides with noncanonical amino acids and chemical modifications, generated using p2smi, can be found at https://zenodo.org/records/15042141.

Acknowledgements

This work was supported by NIH grant 1R01 AI148419. C.O.W. was also supported by the Blumberg Centennial Professorship in Molecular Evolution at The University of Texas at Austin.

References

ChemAxon. (2025). https://www.chemaxon.com.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & Hoon, M. de. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422.

Duffy, F. J., Verniere, M., Devocelle, M., Bernard, E., Shields, D. C., & Chubb, A. J. (2011). CycloPs: Generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. *Journal of Chemical Information and Modeling*, *51*(4), 829–836.

Feller, A. L., & Wilke, C. O. (2025). Peptide-aware chemical language model successfully predicts membrane diffusion of cyclic peptides. *Journal of Chemical Information and Modeling*, *65*(2), 571–579.

Gfeller, D., Michielin, O., & Zoete, V. (2012). SwissSidechain: A molecular and structural database of non-natural sidechains. *Nucleic Acids Research*, *41*(D1), D327–D332.

Landrum, G. (2025). *RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling*. https://www.rdkit.org.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*, 1–14.

Ochoa, R., Brown, J., & Fox, T. (2023). pyPept: A python library to generate atomistic 2D and 3D representations of peptides. *Journal of Cheminformatics*, *15*(1), 79.

Ochoa, R., & Deibler, K. (2025). PepFuNN: Novo nordisk open-source toolkit to enable peptide in silico analysis. *Journal of Peptide Science*, *31*(2), e3666.

OpenEye, C. M. S. (2025). *OEChem*. https://www.eyesopen.com.

Tien, M. Z., Sydykova, D. K., Meyer, A. G., & Wilke, C. O. (2013). PeptideBuilder: A simple python library to generate model peptides. *PeerJ*, *1*, e80.

Yang, L., Cao, S., Liu, L., Zhu, R., & Wu, D. (2025). Cyclicpeptide: A python package for cyclic peptide drug design. *Briefings in Bioinformatics*, *26*(1), bbae714.