The Journal of Open Source Software

p2smi: A Python Toolkit for Peptide FASTA-to-SMILES Conversion and Molecular Property Analysis

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Published: unpublished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) often take advantage of the diverse chemistries found in non-canonical amino acids (NCAAs), altered stereochemistry, and backbone modifications. Despite there being several chemoinformatics toolkits, none are tailored to the task of converting a modified peptide from an amino acid representation to the chemical string nomenclature Simplified Molecular-Input Line-Entry System (SMILES), often used in chemical modeling. Here we present p2smi, a Python toolkit with CLI, designed to facilitate the conversion of peptide sequences into chemical SMILES strings. By supporting both cyclic and linear peptides, including those with NCAAs, p2smi enables researchers to generate accurate SMILES strings for drug-like peptides, reducing the overhead for computational modeling and cheminformatics analyses. The toolkit also offers functionalities for chemical modification, synthesis feasibility evaluation, and 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

Several general bioinformatics toolkits exist for chemical representation and cheminformatics workflows (*ChemAxon*, 2025; Landrum, 2025; O'Boyle et al., 2011; OpenEye, 2025); however, many face limitations such as proprietary licensing and lack of specific functionalities for drug-like peptides. These constraints limit high-throughput application of sequence generation and conversion, especially for peptides incorporating noncanonical amino acids (NCAAs), diverse stereochemistry, and common chemical modifications. The development of p2smi was driven by the need to generate large-scale datasets of peptide SMILES strings for pretraining transformer-based models to understand SMILES notation (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. We used p2smi to build a dataset of 10M peptides with NCAAs, backbone modifications, and cyclizations for pretraining a chemical language model that was used for predicting peptide diffusion across an artificial cell membrane (Feller & Wilke, 2025). We have made p2smi available as a pip-installable package, offering both command-line tools and Python functions for seamless integration into larger workflows.

Features

By leveraging the database in SwissSidechain (Gfeller et al., 2012), p2smi accommodates over 100 unnatural amino acid residues. Our package supports multiple cyclization chemistries, including disulfide bonds, head-to-tail, and side-chain cyclizations. Additionally, p2smi offers a SMILES modification tool, allowing users to apply N-methylation and PEGylation—modifications often used to influence peptide-drug stability and bioactivity. An integrated synthetic feasibility check assists researchers in assessing the practical synthesis of natural peptides. Furthermore, p2smi computes key molecular properties such as logP, TPSA, molecular weight, and Lipinski's rule compliance, supporting early-stage drug-likeness evaluation. Collectively, these features position p2smi as a useful tool for both computational peptide modeling and experimental design.

To install p2smi, use the `pip install p2smi` command. Once installed, p2smi offers five primary command-line tools designed to facilitate various aspects of peptide analysis and modification:

- generate-peptides: This tool enables the generation of random peptide sequences based on user-defined constraints and modifications, allowing for the creation of diverse peptide libraries for computational studies.
- fasta2smi: Converts peptide sequences from FASTA format into SMILES notation, facilitating integration with cheminformatics workflows that utilize SMILES strings for molecular representation.
- modify-smiles: Applies specific chemical modifications, such as N-methylation and PEGylation, to existing SMILES strings, enabling the exploration of modified peptides' properties and behaviors.
- smiles-props: Computes molecular properties—including logP, topological polar surface area (TPSA), molecular formula, and evaluates compliance with Lipinski's rules—from provided SMILES strings, assisting in the assessment of peptides' drug-like characteristics.
- synthesis-check: Evaluates the synthetic feasibility of peptides based on defined synthesis rules, aiding researchers in determining the practicality of synthesizing specific peptide sequences.

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. Notably, pyPept (Ochoa et al., 2023) and PepFuNN (Ochoa & Deibler, 2025) have emerged as significant contributions in this area.

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

While both tools offer valuable capabilities, they are not specifically designed for the direct conversion of peptide sequences into SMILES strings—a functionality central to the initial use-case for p2smi of generating a large-scale database. Rather, pyPept and PepFuNN focus on structural representation, analysis, and structure–activity relationship studies of peptides, complementing the sequence-to-SMILES conversion capabilities provided by p2smi.

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.

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.