

A Machine Learning Framework for Short Peptide Sequence Optimization

Anh Trinh

Georgia Institute of Technology, School of Chemical and Biomolecular Engineering

ABSTRACT

Designing peptides plays an important role in applications ranging from therapeutics and biomaterials to diagnostics. However, due to the large combinatorial sequence space and the high cost and time required for experimental screening, experimental trial and error approaches are prohibitively expensive. Furthermore, peptide design is inherently a multi-objective problem that requires simultaneous optimization of different properties such as biological activity, stability, solubility, and safety. These challenges motivate the use of computational design strategies. Traditional physics-based and sequence-alignment methods often struggle to handle variable length sequences and often rely on structural information that is unavailable for many peptides. More recently, deep learning models such as AlphaFold, ESM, and diffusion-based approaches have transformed protein modeling. However, their large data requirements, high computational cost, and black-box nature reduce their practicality for deterministic multi-objective optimization in limited-data settings. This work proposes a data-driven, multi-objective peptide design framework that integrates sequence-to-feature transformations using Fast Fourier Transform (FFT) – based representations, interpretable feature attribution through GroupSHAPLEY, and metric-learning based optimization strategies. A bidirectional mapping between sequence space and feature space is introduced to identify critical feature contributions and

improve interpretability during peptide optimization. The primary focus of this poster is the optimization component of the framework. Specifically, distance metric learning methods, including Neighborhood Component Analysis (NCA) and Large Margin Nearest Neighbor (LMNN) are investigated to maximize class separation between peptide groups and identify discriminative feature representations. Comparative analyses were performed to evaluate the robustness, tunability, and optimization behavior of these methods on both synthetic and peptide-representative datasets. Using Support Vector Machine (SVM) as the predictive model, the proposed methods demonstrate improved optimization efficiency through dimensionality reduction in synthetic data experiments. Multiple optimization constraints can be incorporated to assess the robustness, scalability, and adaptability of the framework across different design scenarios. The proposed framework aims to provide an interpretable and computationally efficient alternative for peptide design under limited-data constraints.

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See <https://creativecommons.org/licenses/by-sa/4.0/>

