# Peptide Featurization Methods in Drug Discovery

Grace Adalia Rico

Master of Science in Bioinformatics
Bioinformatics Research Center
Aarhus University

June 2026

In collaboration with Raven Biosciences

# Abstract

Predicting the biological properties of peptides from sequences is a central problem in drug discovery in silico, and the quality of those predictions depends critically on the numerical representation provided by featurizers. This thesis benchmarks new and existing featurization methods spanning small molecule descriptors, peptide specific encodings, protein language model embeddings, and structure-based descriptors. These are evaluated on major histocompatibility complex binding affinity and hemolysis tasks. The novel methods include a circular fingerprint adaptation and a semi-supervised self-distillation model. Of the featurizers tested, small molecule and simple peptide featurizers performed on par with deep learning methods, contrary to expectations that deep learning methods would dominate. Featurizers were less successful on binding affinity than hemolysis which aligns with the comparative complexities of these two tasks. The novel deep learning model tested here was as successful as, if not more successful than other featurizers. Findings suggest that featurizers' performance is task dependent, and there remain tasks such as binding affinity where even the most sophisticated models are yet to capture the relevant features of peptides.

# Acknowledgements

I would like to thank my colleagues and mentors at Raven Biosciences for their guidance and expertise. I would especially like to thank Lucy for her insight and support.

# Contents