MACHINE LEARNING FOR THERAPEUTIC PEPTIDES

Sequence and Structure Representations for Peptide Property Prediction

Fusing sequence and chemical structure representations and sharing one model across four property prediction tasks

Tarek Ibrahim, James Green

Department of Systems and Computer Engineering Faculty of Engineering and Design, Carleton University

Internship, Carleton University Research Experience for Undergraduate Students (I-CUREUS) 2026-27

1 INTRODUCTION

Therapeutic peptides are short amino acid chains whose suitability depends on properties that are slow and costly to evaluate. Prediction from sequence alone would let candidate libraries be triaged before any bench work.

PeptiVerse (2026) is the state of the art for four such properties: permeability, solubility, nonfouling and hemolysis. Its data and partitions are public, so every result here is directly comparable. A peptide can be read as a sequence by a protein language model or as a chemical structure by a chemical language model; whether the two views are complementary remains unresolved.

2 OBJECTIVES AND HYPOTHESES

H1 Representation fusion. Jointly encoding sequence and chemical structure yields better results than either representation alone.

H2 Multitask learning. One shared model trained on all four related properties outperforms four single task models.

3 MATERIALS AND METHODS

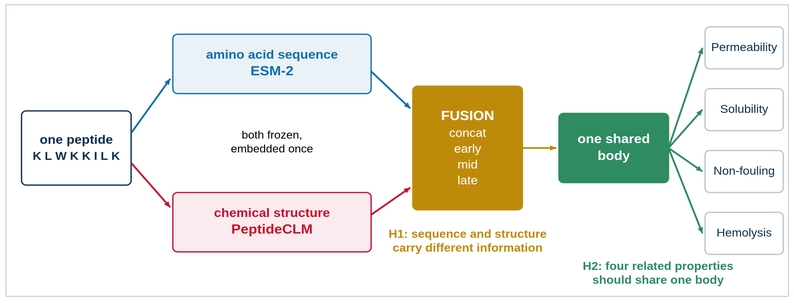

Figure 1. Model architecture. Both encoders remain frozen; only the fusion block, the shared body and the four task heads are trained.

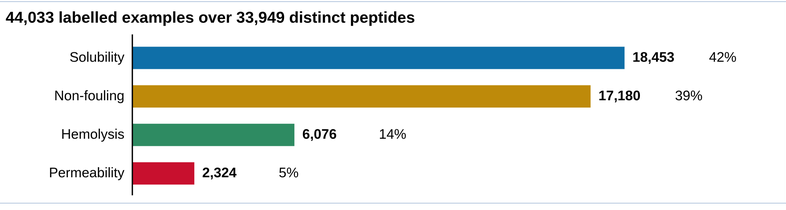

Figure 2. Composition of the four benchmark datasets.

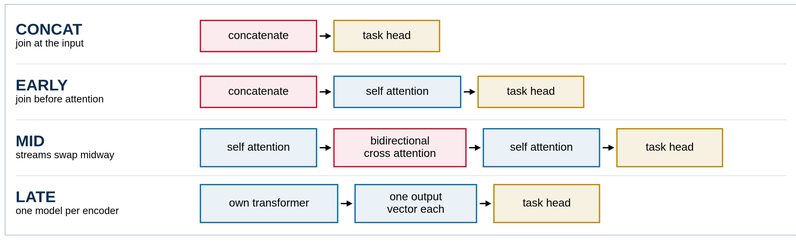

Figure 3. The four fusion designs, distinguished by where the two token streams join: at the input, before attention, midway through cross attention, or after separate transformers.

Encoders. ESM-2 embeds the sequence, PeptideCLM the chemical structure. Both stay frozen, so each run trains only the fusion block and task heads.

4 RESULTS

Trained one model per task, this work reaches test F1 of **0.893** on permeability, **0.734** on solubility, **0.726** on nonfouling and **0.557** on hemolysis. Neither hypothesis was supported.

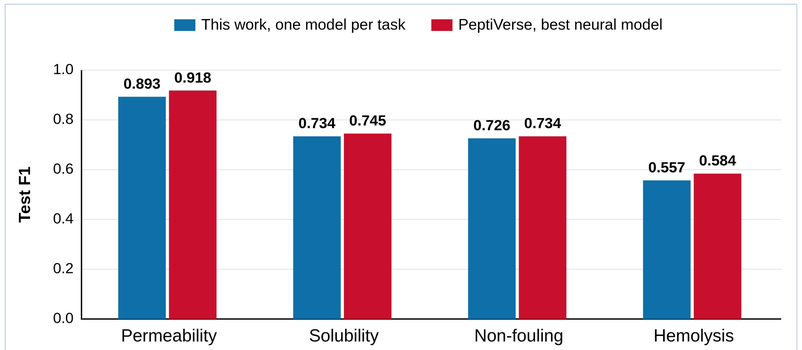

Figure 4. This work against the state of the art, property by property.

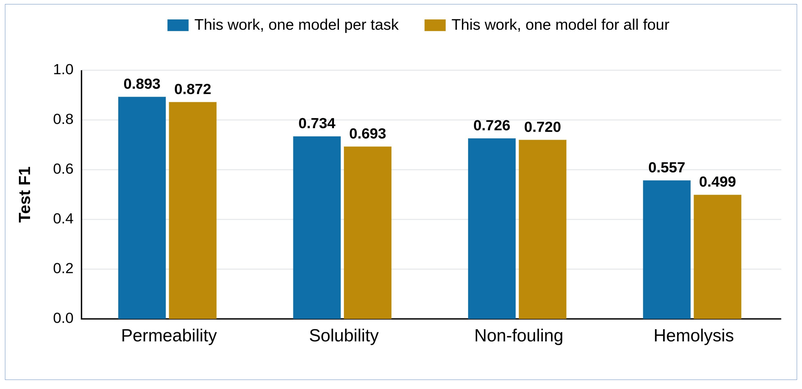

Figure 5. Test of H2. One shared model never beat four separate ones. It comes close on some properties and falls clearly behind on others.

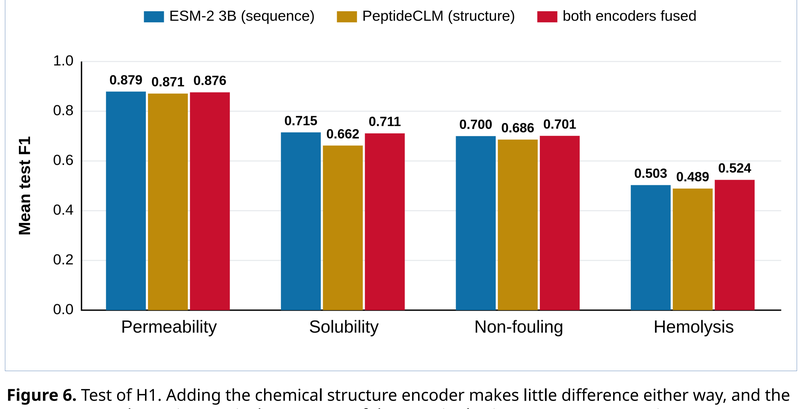

Figure 6. Test of H1. Adding the chemical structure encoder makes little difference either way, and the sequence encoder on its own is the stronger of the two single views on most properties.

5 GENERALIZATION AND FUSION DEPTH

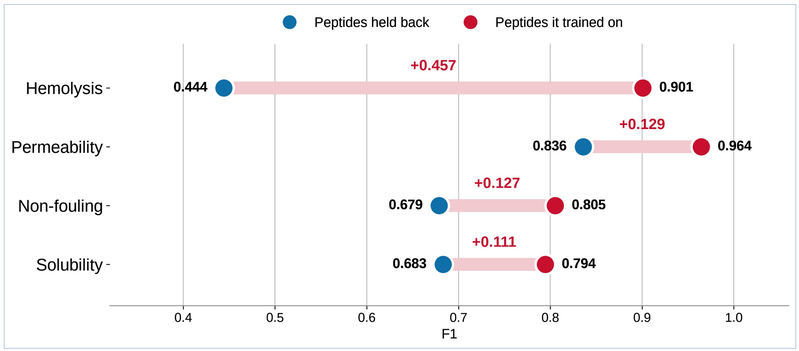

Figure 7. Generalization gap by property. Every model does better on the peptides it trained on than on ones it has not seen, and the size of that drop is not the same across the four properties.

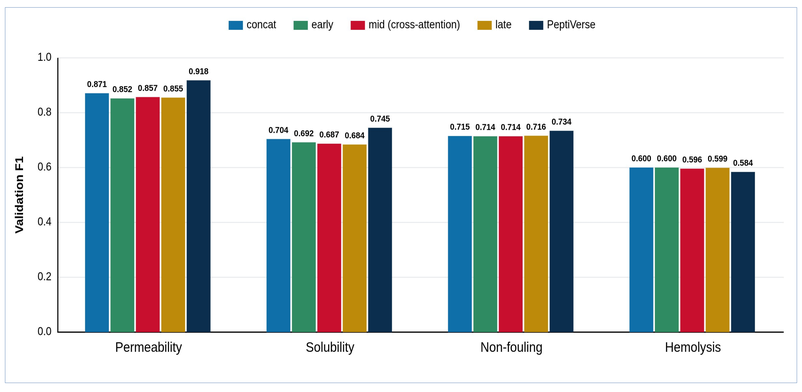

Figure 8. Where the two encoders join. The four join points fall within 0.020 validation F1 of one another on every property. The choice of fusion depth does not make a significant impact.

7 LIMITATIONS

The likeliest explanation is that the two encoders are telling us the same thing. Both describe the same molecule, and both were trained on overlapping collections of peptides, so the second view adds little value.

8 CONCLUSIONS

- Both encoders were frozen, so these results bound the pretrained representations rather than the architectures.
- Class support is uneven: hemolysis is 14 percent of labelled examples and shows the largest gap.
- Evaluation is confined to one public benchmark, so transfer to other corpora and to wet lab validation is untested.

8 CONCLUSIONS

Fusing sequence and structure embeddings, and sharing one model across the four properties, both fail to improve on a single frozen sequence encoder. Establishing which choices carry no measurable effect bounds where further search is warranted: adapting the encoders is more promising than adding representations or sharing heads.

Acknowledgements: Compute was provided by the Google TPU Research Cloud. Datasets are from PeptiVerse (2026). Supported by the I-CUREUS program, Carleton University. Carried out in the Carleton University Biomedical Informatics Co-laboratory (cuBIC).