Computational Molecular Biology and Bioinformatics

SWING

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

October, 2025

What is SWING?

Protein language models embed protein sequences for different tasks. However, these are suboptimal toward learning the language of protein interactions.

Sliding Window Interaction Grammar (SWING) is an interaction language model (iLM) that models differences in amino-acid properties to generate an interaction vocabulary [1]. SWING successfully predicted both class I and class II peptide–major histocompatibility complex interactions.

SWING outperformed passive uses of protein language model embeddings, demonstrating the value of the unique iLM architecture. Overall, it is a generalizable zero-shot iLM that learns the language of protein–protein interactions.

The SWING framework

Language generation for an interaction has the following three crucial components:

- **A sliding window:** The length of it can span from 3 to 36 AA.
- **The target protein sequence:** A protein sequence of interest
- **A biochemical metric:** The Grantham polarity and Miyazawa hydrophobicity scales were used as the two metrics in the study. The calculation of biochemical difference is done as $\mathrm{biochemical}_{\mathrm{difference}} = |\mathrm{metric}(aa_i) - \mathrm{metric}(aa_j)|$, where 'metric' denotes the value for a particular AA for a specific biochemical metric, i denotes the position on the sliding window and j is the position on the target protein. The rounded-off biochemical difference is the encoding for the pair of AA at a given position on the two sequences.

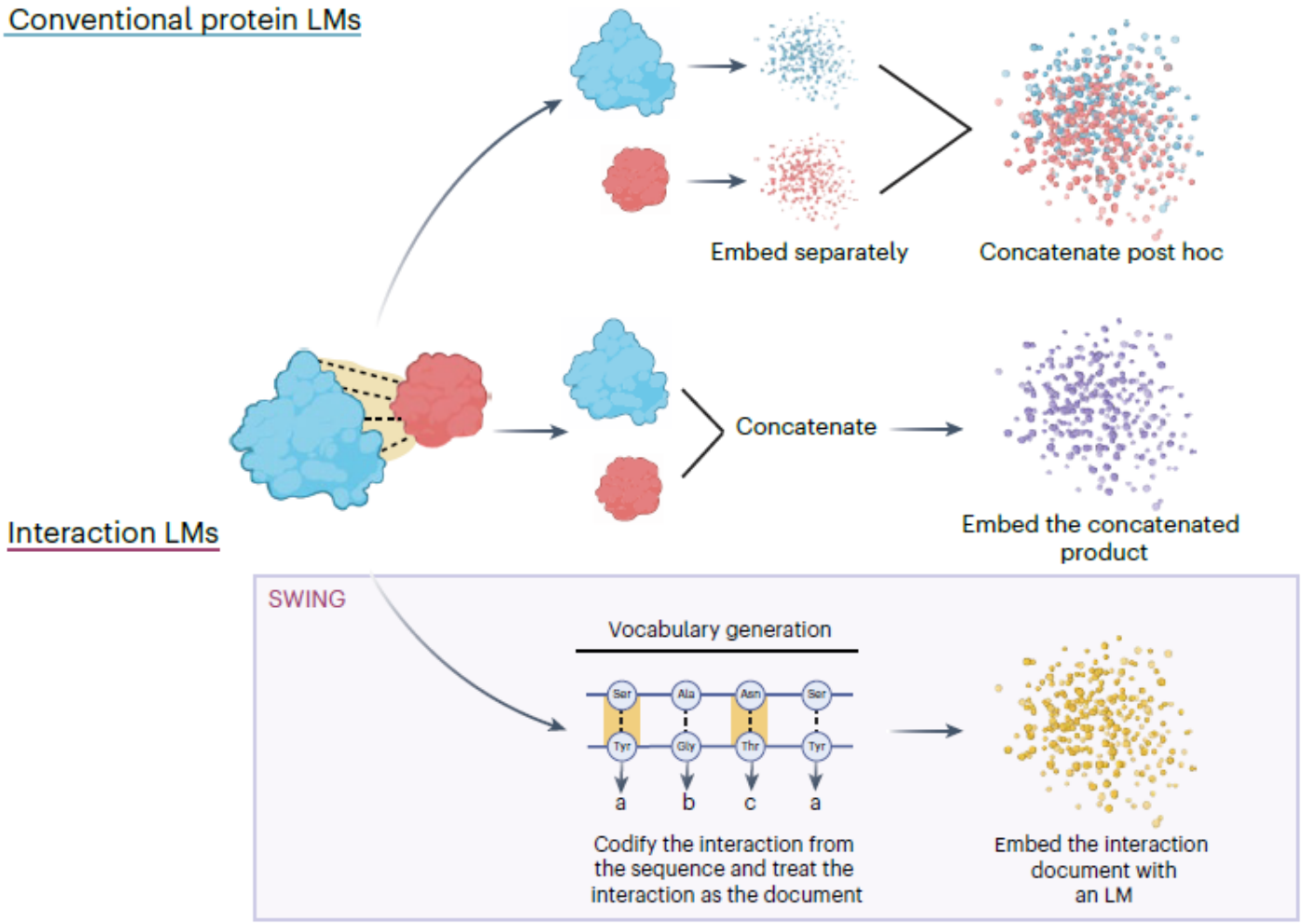
The encodings are concatenated sequentially to generate the interaction language.

Novelty of SWING

The interaction language sequence is divided into k-sized sub-sequences, also called k-mers. The hyperparameter k is either tuned or chosen based on biological context. The k-mers are the input to the embedding architecture for generating the low-dimensional representations for each interaction. The Doc2Vec model is used for generating the interaction level embeddings.

A target biochemical k-mer is predicted using the embeddings of surrounding k-mers (also called the context window) along with the interaction embedding for a particular interaction. In the distributed bag of words, the k-mers in a fixed length context window are predicted from the interaction embedding alone.

Novelty of SWING

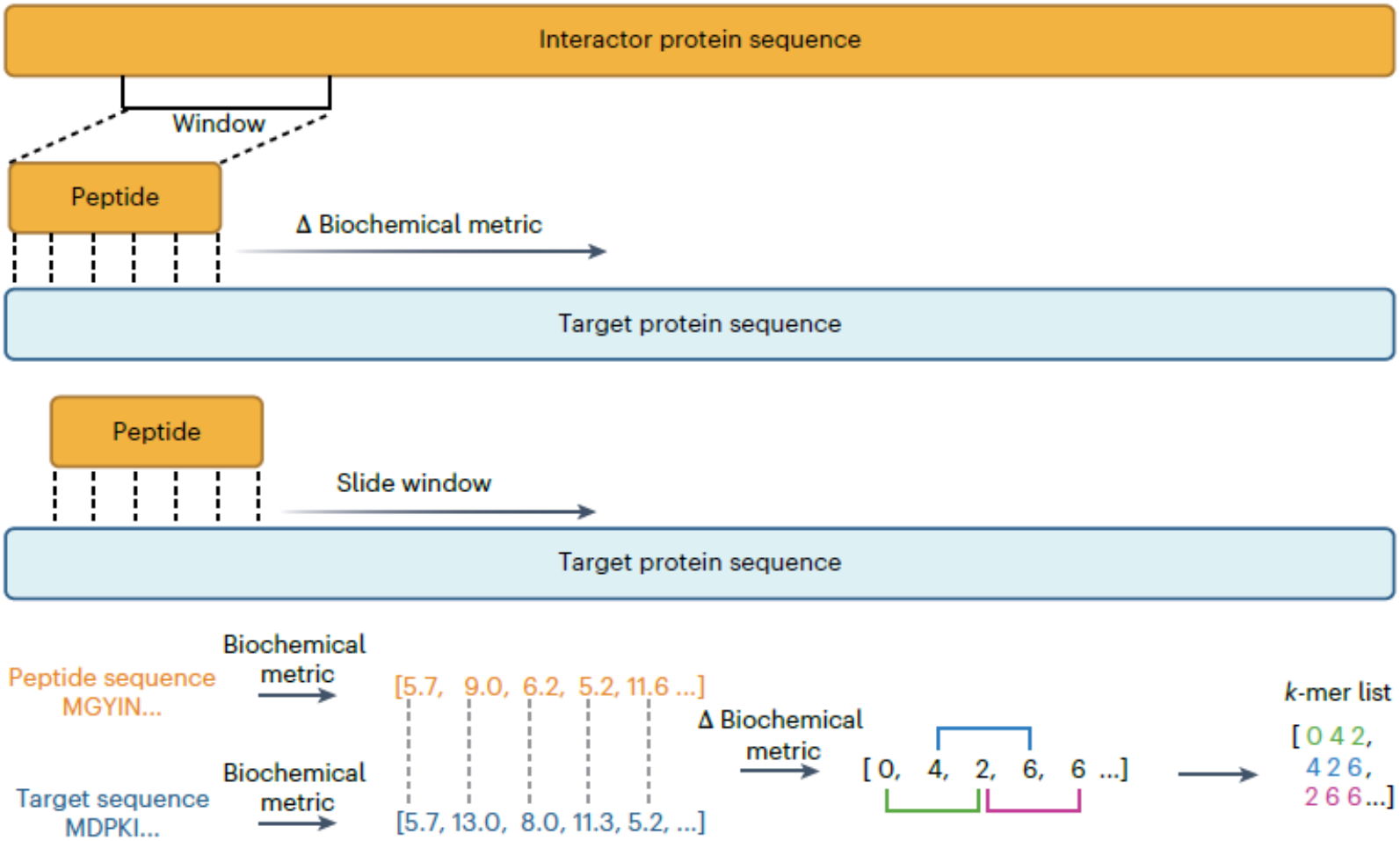


Vocabulary generation of SWING

Encoded sequences were split by character using a custom word-level tokenizer from the HuggingFace Tokenizers library. The tokenizer was trained with a vocabulary of 16 tokens consisting of the digits 0–9 and special tokens. After tokenization, sequences were truncated or padded to a length of 4,500 tokens with a random 15% of the sequences held out as an evaluation set.

The sequence embeddings were extracted from a custom RoBERTa model built and trained on a masked language modeling task with the HuggingFace Transformers library. The hidden layer and intermediate layer dimensions were 128 and 64, respectively, and the model was built with four hidden layers and four attention heads. The GeLu activation function, an initial learning rate of 1×10^{-4}, layer dropout probabilities of 0.1 and a mask probability of 0.15 were used. Training was done for 40 epochs and aimed to minimize a cross-entropy loss.

Vocabulary generation of SWING

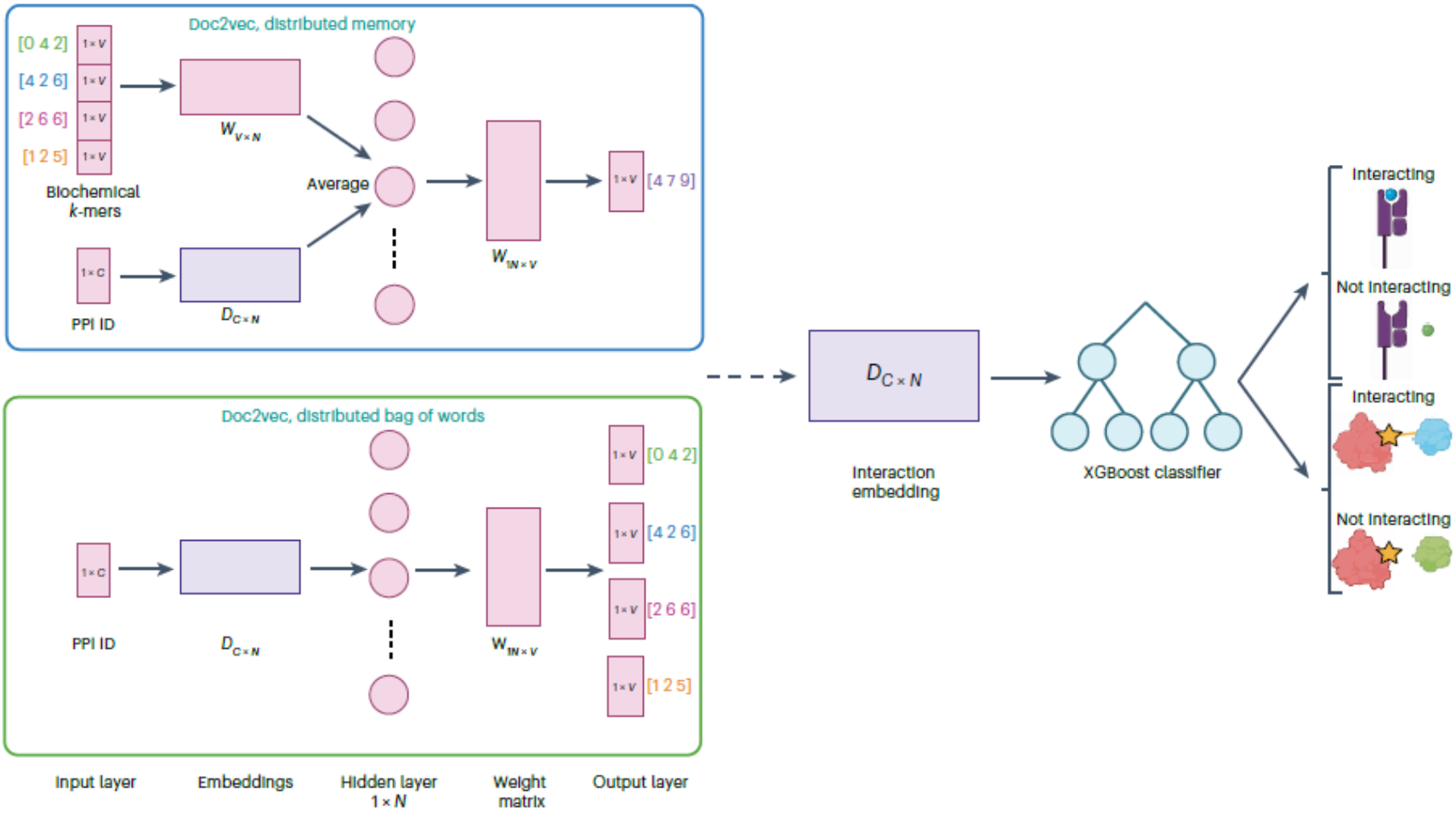


Embedding and classification functions of SWING

We trained an XGBoost Classifier model directly on the SWING encodings and used 5-fold stratified cross-validation to report the mean AUROC for the different prediction tasks.

In our CEmbed-pLM, the whole target and interactor sequences are concatenated end to end, then encoded with Grantham scores. The encoding is k-merized and used to train a Doc2Vec model from gensim v.4.2.0 to create embedded representations of the PPI. The final embeddings are used to train a XGBClassifier() from XGBoost to predict whether the interaction is maintained or not.

Embedding and classification functions of SWING



Comparative analysis

	SWING	NetMHCPan	MixMHCPred	AF	ESM1b	EVE
Embeds the interaction	✓	✗	✗	✗	✗	✗
Works across PPI contexts	✓	✗	✗	✗	✗	✗
Predicts class I pMHC interactions	✓	✓	✓			
Predicts class II pMHC interactions	✓	✓	✓			
Predicts class I and class II interactions with the same model	✓	✗	✗			
Predicts across MHC classes	✓	✗	✗			
Generalizes across species while trained on one species	✓	✗	✗			
Predicts variant effects on the phenotype	✓			✓	✗	✓
Predicts edgotype effects on the phenotype	✓			✗	✓	✗

References

1. Siwek, J.C., Omelchenko, A.A., Chhibbar, P., Arshad, S., Rosengart, A., Nazarali, I., Patel, A., Nazarali, K., Rahimikollu, J., Tilstra, J.S. and Shlomchik, M.J., Sliding Window Interaction Grammar (SWING): a generalized interaction language model for peptide and protein interactions. Nature Methods, 22:1707-1719, 2025.