# T-Cell Receptor-Peptide Interaction Prediction with Physical Model Augmented Pseudo-Labeling

Yiren Jian [1], Erik Kruus [2] and Martin Renqiang Min [2]

(1) Dartmouth College (2) NEC Laboratories America Inc.

NEC Laboratories **America**

## Main Contributions

- ✓ We propose a method that trains the deep neural network with physical modeling and data-augmented pseudo-labeling:
  1. Data-augmented pseudo-labeling of TCR-peptide pairs by a model first trained on the labeled dataset.
  2. Physical modeling between TCRs and peptides by Docking.
- ✓ We introduce a new dataset that contains over 80,000 unknown TCR-peptide pairs with docking energy scores.

## Background

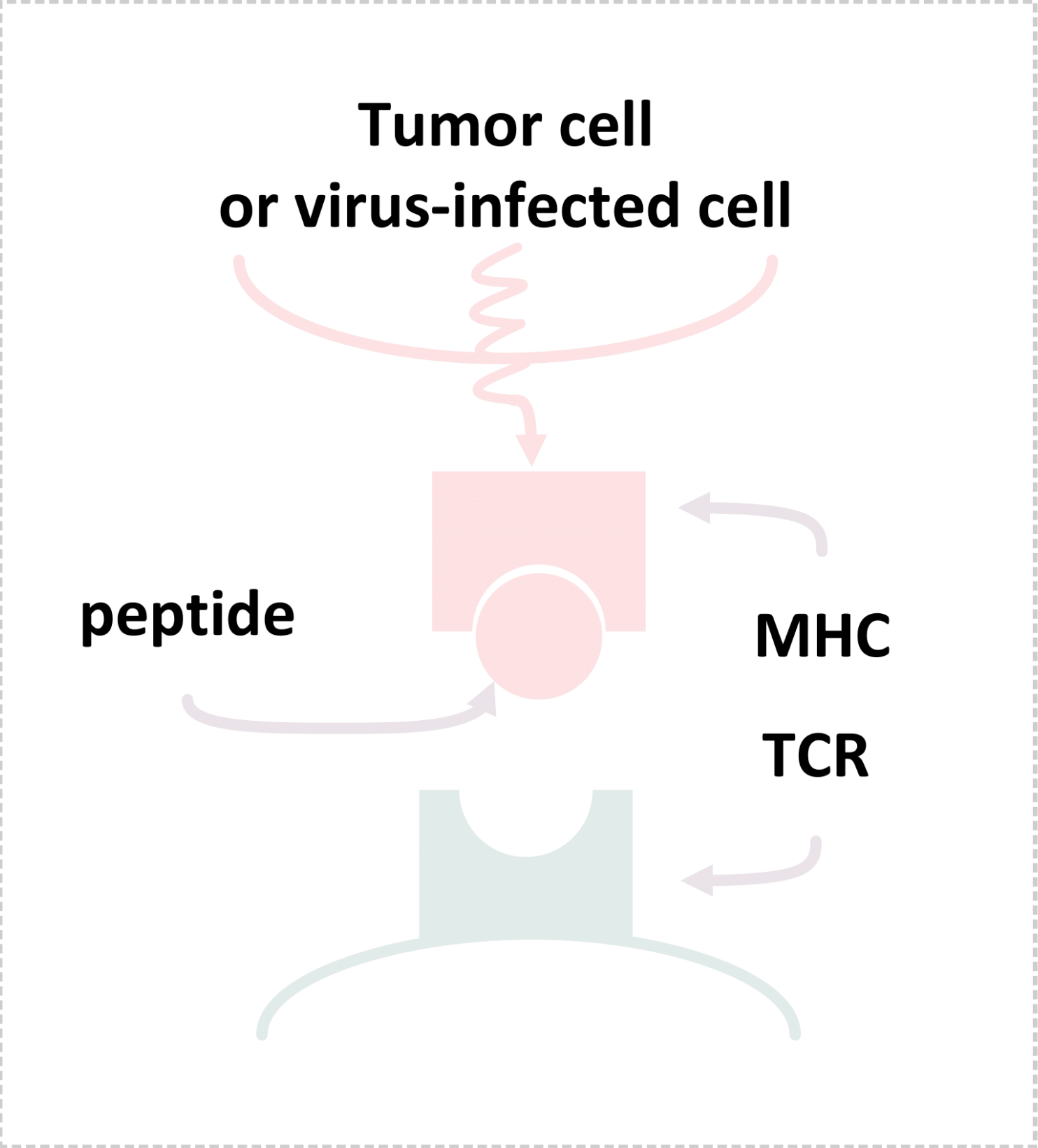


Illustration of T-cell receptors (TCR) and peptide binding: The TCR lies on the surface of the T-cell for recognition of foreign peptides. Peptides are presented by major histocompatibility complex (MHC) found on the surface of tumor cells or virus-infected cells.

## Problem Statement

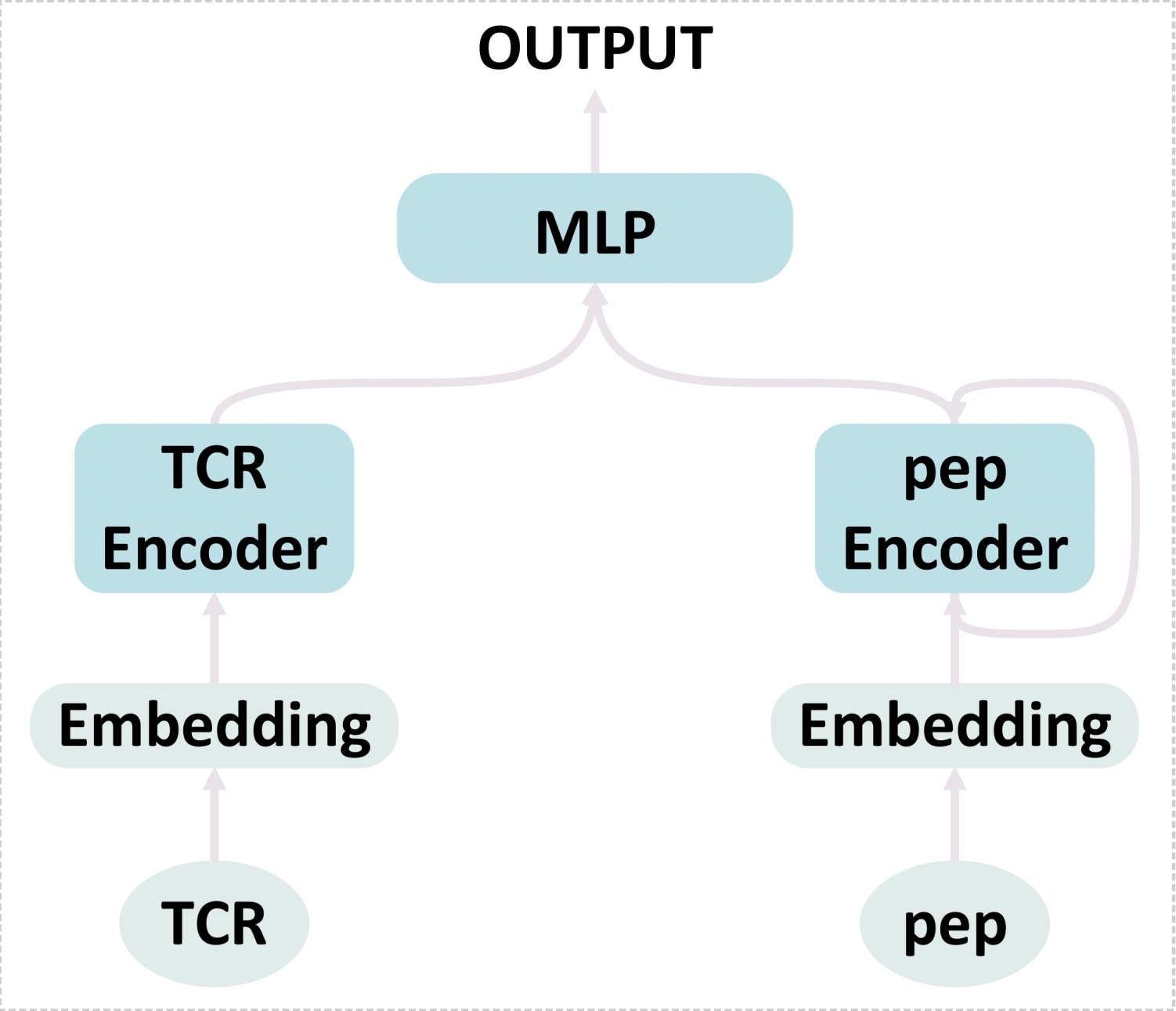
- ✓ Given a sequence of TCR and a sequence of peptide, the goal is to predict the interaction between them.

Challenges:

1. Though datasets like McPAS and VDJdb exists, the data scarcity issue limits the learning of potent deep neural network.
2. The labeled datasets are biased: many peptides/TCRs though different, are similar.
3. Large amount of TCR sequences (without known associated peptides) are available in database that are not being leveraged.

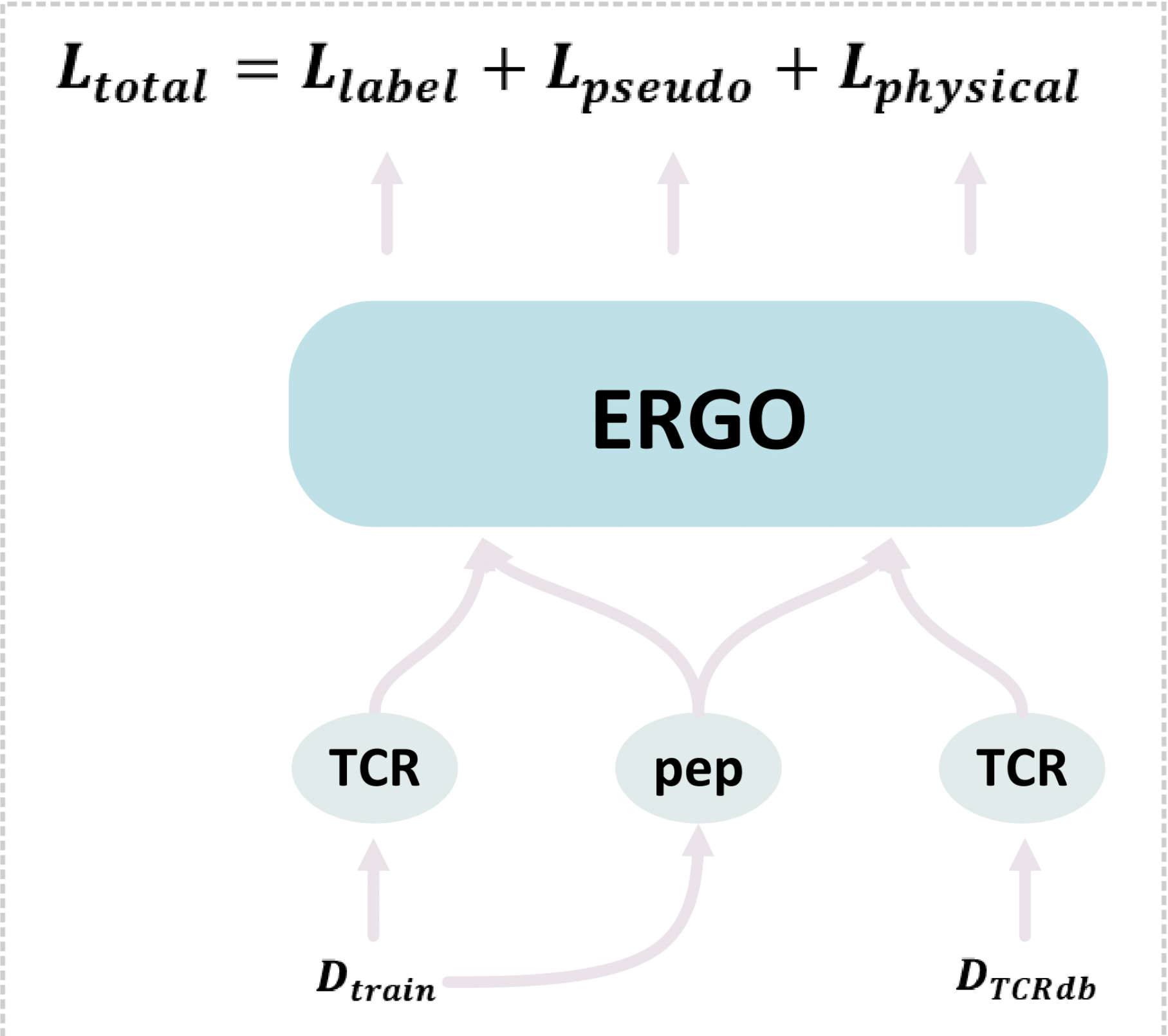
## Methods

### Base Model



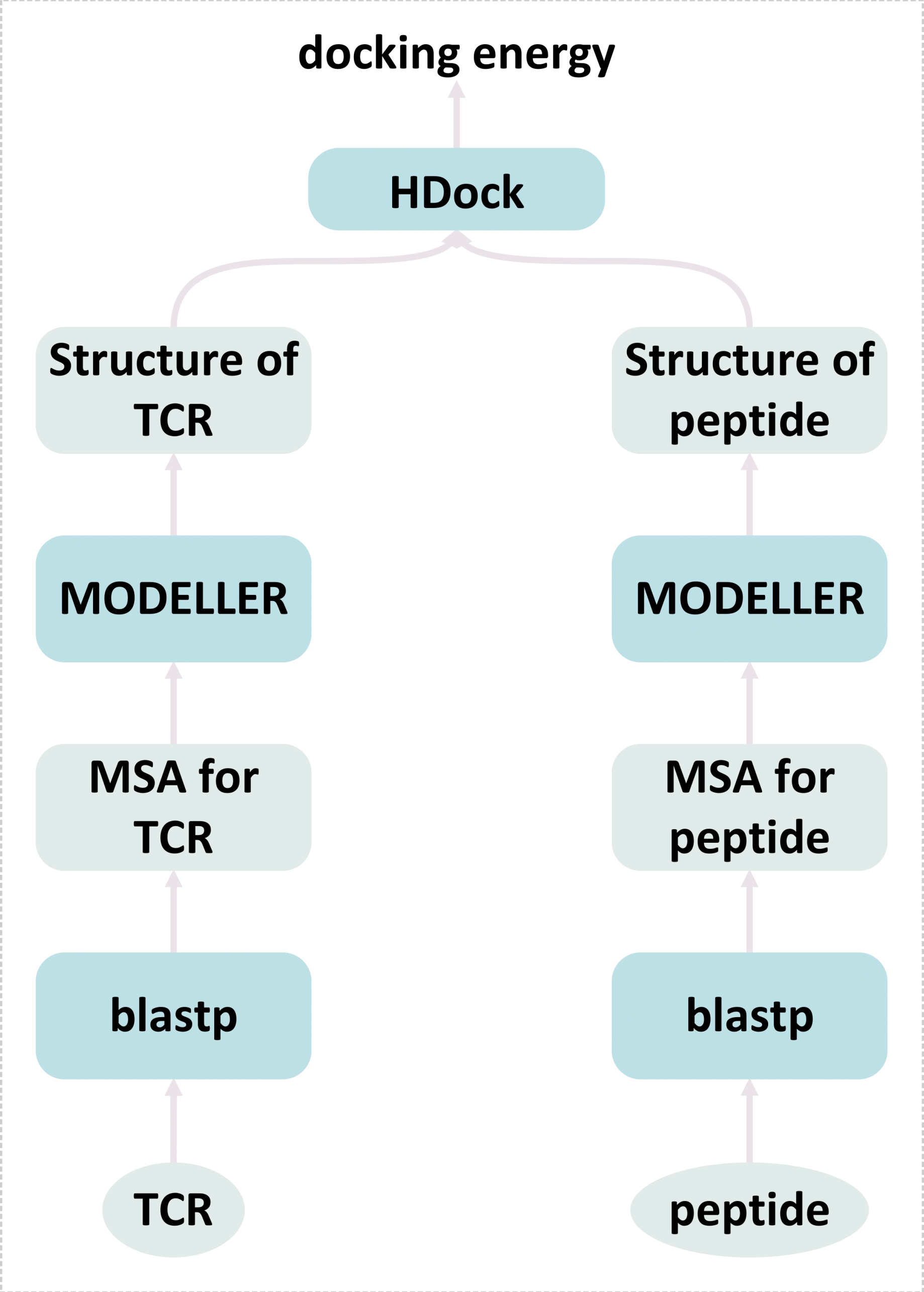
Models used in our study. The model is borrowed from ERGO for fair comparisons. The model has two separate encoders for TCR and peptide

### Training Losses



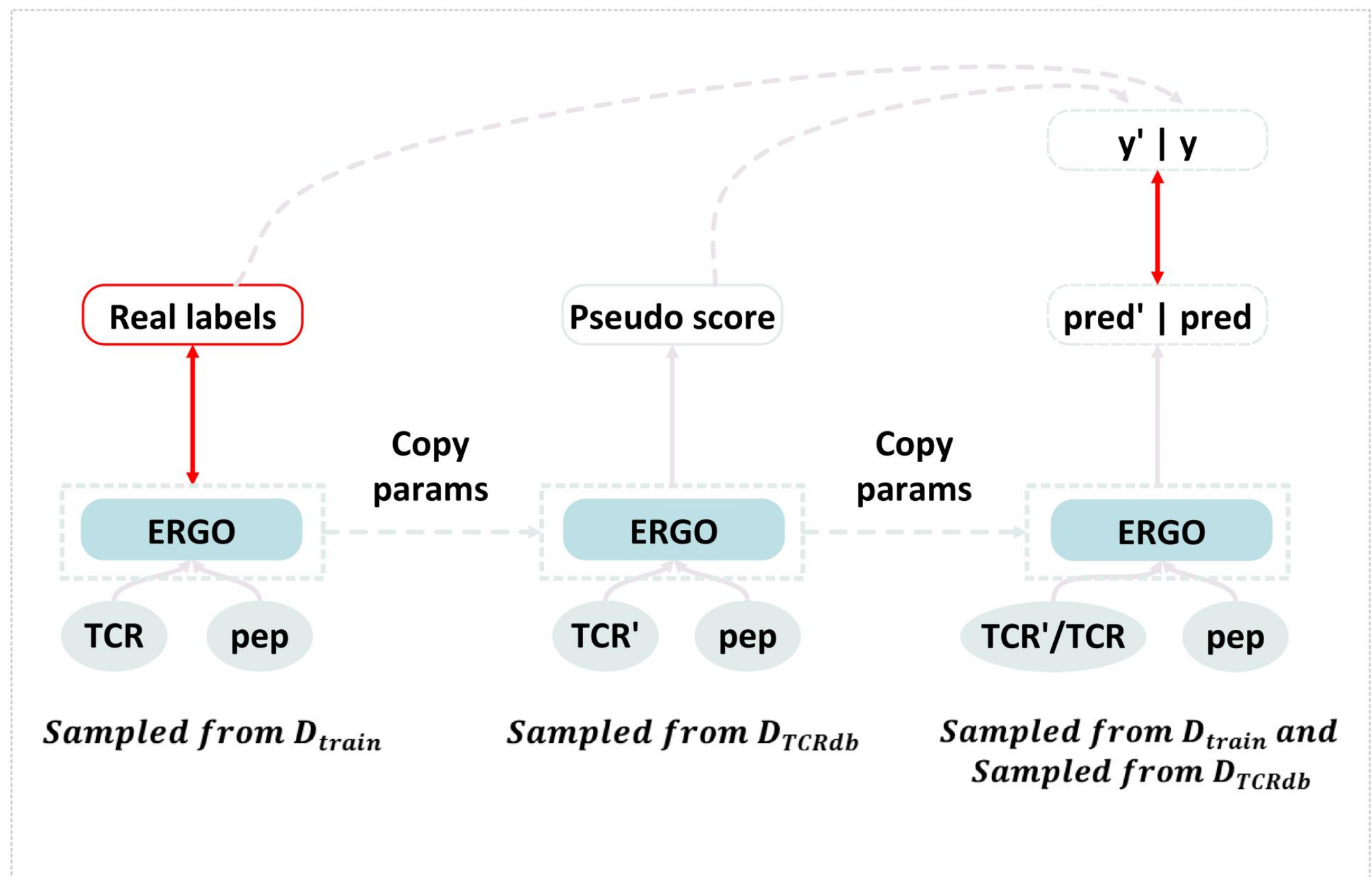
$$L_{total} = L_{label} + L_{pseudo} + L_{physical}$$

### Physical Modeling: Docking by HDOCK



docking using HDOCK. For a given sequence of TCR/peptide, we first use blastp to find the multiple-sequence alignment (MSA) for the sequence. MSA and the corresponding structures from PDB are then used by MODELLER for building the structures of the TCR/peptide. Finally, we call HDOCK with the given structures of the TCR and peptide for computing docking energies.

### Overview of Our Learning Framework



$$\mathcal{L}_{labeled} = \text{BinaryCrossEntropy}(pred, y)$$

Supervised loss from labeled TCR-peptides

$$\mathcal{L}_{\text{physical}} = \text{BinaryCrossEntropy}(pred', y')$$

Learning from pseudo-labels by docking

$$\mathcal{L}_{\text{pseudo-labeled}} = \text{KL-div}(pred', prob')$$

Learning from pseudo-labels by a teacher

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{labeled}} + \beta\mathcal{L}_{\text{physical}} + \gamma\mathcal{L}_{\text{pseudo-labeled}}$$

## Experimental Results

| Data size | 6K | 10K | 20K |
|---|---|---|---|
| ERGO | 67.6 ± 0.4 | 71.9 ± 0.4 | 76.6 ± 0.3 |
| + Pseudo | 69.3 ± 0.4 | 73.6 ± 0.3 | 77.6 ± 0.3 |
| + Docking | 69.4 ± 0.4 | 73.3 ± 0.3 | 77.9 ± 0.2 |
| ours (3 losses) | 70.4 ± 0.3 | 73.7 ± 0.3 | 77.6 ± 0.2 |
| ours + meta-update | **71.5 ± 0.3** | **74.7 ± 0.3** | **78.4±0.2** |

Experimental results on **McPAS** using base model of ERGO-LSTM. Results are collected from 5 different independent experimental runs. In these experiments, ERGO+Psudo and ERGO+Docking perform roughly equally well.

| Data size | 6K | 10K | 20K |
|---|---|---|---|
| ERGO | 68.1± 0.4 | 72.0 ± 0.3 | 73.6 ± 0.4 |
| + Pseudo | 68.4 ± 0.3 | 72.4 ± 0.3 | 73.9 ± 0.3 |
| + Docking | 69.5 ± 0.4 | 73.4± 0.3 | 74.6± 0.3 |
| ours (3 losses) | 70.4± 0.3 | 72.9± 0.3 | 74.6± 0.3 |
| ours + meta-update | **71.5 ± 0.3** | **73.8± 0.3** | **75.2± 0.3** |

Experimental results on **VDJdb** using base model of ERGO-LSTM. Results are collected from 5 different independent experimental runs. In these experiments, ERGO+Pseudo only improves over the baseline marginally, while physical modeling by docking still increase the AUC by significant margins.

| rare peptides | baseline | average | ours |
|---|---|---|---|
| KRWIILGLNK | 52.8 | 54.4 | 68.1 |
| KMVAVFYTT | 48.9 | 54.4 | 65.8 |
| FPRPWLHGL | 50.2 | 54.4 | 58.5 |

Experiments with AE-LSTM model with McPAS dataset of 6K labeled examples. "average" denotes the average AUC for all peptides in this experimental setup.

https://cs.dartmouth.edu/~yirenjian/