



can you summarize the detailed training steps for mBLM

## OI Gemma3

The detailed training steps for mBLM involved pre-training and fine-tuning.

**Pre-training:** mBLM was pre-trained to predict masked amino acid residues using a total of 253,808 unique paired antibody sequences from GenBank and Observed Antibody Space.

**Fine-tuning:** The pre-trained mBLM was fine-tuned using a final-layer embedding, followed by a multi-head self-attention block and a multi-layer perceptron (MLP) block. This fine-tuning was performed for antibody specificity prediction based on seven categories: influenza HA head, influenza HA stem, HIV, SARS-CoV-2 S NTD, SARS-CoV-2 S RBD, SARS-CoV-2 S S2, and others. The training and test sets had a minimum pairwise Levenshtein distance of 10 and an average of 68 to ensure sequence divergence.

B. Paper1.pdf

please provide more details

## OI Gemma3

Here's a more detailed breakdown of the mBLM training steps, based on the provided text:

### 1. Dataset Collection for Pre-training:

- **Sources:** Antibody sequences were gathered from two primary databases: GenBank and Observed Antibody Space.
- **Sequence Number:** A total of 253,808 unique paired antibody sequences were used. "Paired" means both the heavy and light chain sequences were available.

### 2. Pre-training Phase:

- **Objective:** The model was trained to predict masked amino acid residues within the antibody sequences. This is a common technique in natural language processing (NLP) known as masked language modeling. The model learns the relationships between amino acids by trying to "fill in the blanks" in the sequences.
- **Process:** The model processes the antibody sequence and attempts to predict the identity of amino acids that have been randomly masked (hidden).

### 3. Fine-tuning Phase:

- **Starting Point:** The pre-trained mBLM model (which already understands basic antibody

Here's a more detailed breakdown of the mBLM training steps, based on the provided text:

### 1. Dataset Collection for Pre-training:

- **Sources:** Antibody sequences were gathered from two primary databases: GenBank and Observed Antibody Space.
- **Sequence Number:** A total of 253,808 unique paired antibody sequences were used. "Paired" means both the heavy and light chain sequences were available.

### 2. Pre-training Phase:

- **Objective:** The model was trained to predict masked amino acid residues within the antibody sequences. This is a common technique in natural language processing (NLP) known as masked language modeling. The model learns the relationships between amino acids by trying to "fill in the blanks" in the sequences.
- **Process:** The model processes the antibody sequence and attempts to predict the identity of amino acids that have been randomly masked (hidden).

### 3. Fine-tuning Phase:

- **Starting Point:** The pre-trained mBLM model (which already understands basic antibody sequence "grammar") was used as a foundation.
- **Layers Added:**
  - **Final-layer Embedding:** This likely extracts a feature representation of the antibody sequence from the pre-trained model.
  - **Multi-head Self-Attention Block:** This allows the model to focus on different parts of the antibody sequence when making predictions.
  - **Multi-layer Perceptron (MLP):** This is a standard neural network layer used for classification.