# *peleke-1*: A Suite of Protein Language Models Fine-Tuned for Targeted Antibody Sequence Generation

*Nicholas Santolla*, Trey Pridgen*, Prbhuv Nigam, and Colby T. Ford*

Silico Biosciences
A <Tuple> venture.

## Abstract

Here we present *peleke-1*, a suite of protein language models fine-tuned from state-of-the-art large language models using curated antibody-antigen complex data. These models generate targeted antibody Fv sequences for a given antigen sequence input at-scale. This suite of models provides a reliable, artificial intelligence-driven approach for *in silico* therapeutic antibody discovery along with an open-source framework for future antibody language model tuning.

## Tuning and Inferencing Workflow

**Data Curation**

- Collect antibody data from *SAbDab*
- Download PDB entries from Protein Data Bank
- Retrieve antibody and antigen sequences
- Detect epitope residues

**Model Tuning**

- Retrieve open-source model from Hugging Face
- Update tokenizer with *transformers*
- Set up PEFT adapter
- Tune model for multiple epochs

**Inferencing**

- Generate 50 Fv sequences for each antigen and model
- Fold each generated Fv sequence and antigen as a multimer with *ESM3*

**Evaluation**

- Predict antibody binding affinity with *HADDOCK3*
- Predict protein complex stability with *FoldX*
- Predict Fv sequence humanness with *Humatch*

The *peleke-1* suite consists of multiple protein-language models (PLMs), fine-tuned from existing large language models (LLMs) that span varying architectures and parameter magnitudes. To perform the fine-tuning, copious antibody-antigen sequence information was collected to form a curated training dataset.

Now Available on 🤗 **Hugging Face**:

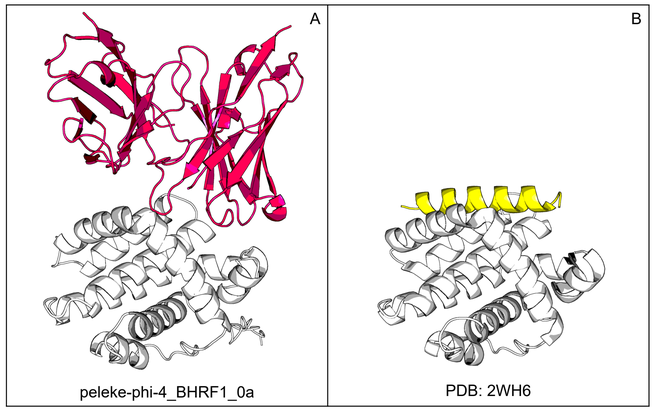
Antibody-Antigen Complex Training Data

- peleke-llama-3.1-8b-instruct
- peleke-mistral-7b-instruct-v0.2
- peleke-phi-4

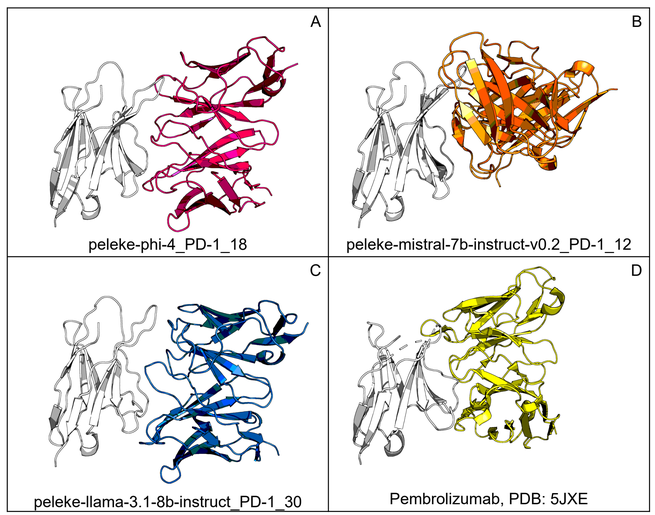
huggingface.co/silicobio/

## Targeted Antibody Generation

Across the 1,050 generated Fv sequences (50 × 7 benchmark antigens), many have epitopes in the resulting predicted structures that closely match the desired epitope residues defined in the model input prompt.




Comparison of (A) peleke-phi-4_BHRF1_0a, a generated anti-BHRF1 Fv structure, in pink; and (B) a BCL-2-like protein 11 (in yellow) bound to BHRF1 (in white, PDB: 2WH6)




Comparison of 3 novel anti-PD-1 candidates: (A) peleke-phi-4_PD-1_18 (in pink); (B) peleke-mistral-7b-instruct-v0.2_PD-1_12 (in orange); and (C) peleke-llama-3.1-8b-instruct_PD-1_30} (in blue); with (D) pembrolizumab (in yellow), all bound to PD-1 (in white, PDB: 5JXE)

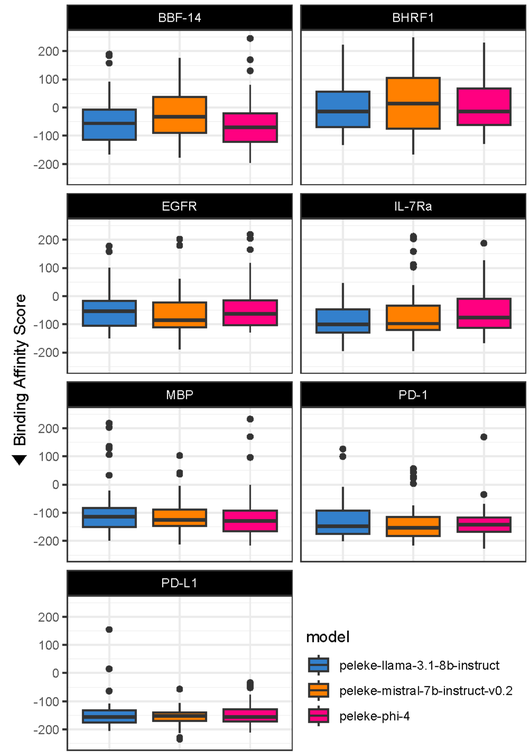
silico.bio/peleke

## Predicted Binding Performance

Predicted antibody-antigen binding affinities were considerably stable between the base models but varied depending on the benchmark antigen.

| Antigen | peleke-1 Fine-Tuned Model | | |
|---|---|---|---|
| | llama-3.1-8b-instruct | mistral-7b-instruct-v0.2 | phi-4 |
| BBF-14 | 22.73% (63.34) | 15% (106.84) | 18.42% (49.88) |
| BHRF1 | 2.63% (218.86) | 10.81% (289.61) | 9.3% (131.62) |
| EGFR | 20% (101.64) | 23.26% (24.66) | 20% (163.88) |
| IL-7Ra | 36.17% (-3.46) | 46.81% (-10.11) | 31.91% (13.64) |
| MBP | 67.39% (-52.56) | 57.78% (-52.52) | 69.39% (-57.98) |
| PD-1 | 73.47% (-66.72) | 79.59% (-82.51) | 76.6% (-70.12) |
| PD-L1 | 79.59% (-71.53) | 70.83% (-62.52) | 72.92% (-73.38) |

Percentage of complexes with van der Waals energies <-25 kcal/mol (median value shown in parentheses) by fine-tuned model and antigen.




Boxplots depicting the predicted binding affinity score distribution by antigen and model. More negative values (HADDOCK scores) indicate better overall binding. Note that the y-axis is showing scores between [-250, 250].

github.com/silicobio/peleke