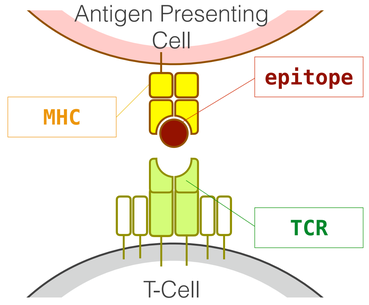
Iterative Attack-and-Defend Framework for Improving TCR-Epitope Binding Prediction Models

Pengfei Zhang[1,2], Hao Mei[1,2], Seojin Bang[3], and Heewook Lee[1,2*]
[1]School of Computing and Augmented Intelligence, [2]Biodesign Institute, Arizona State University
[3]Deepmind, Google

National Institutes of Health

Background.

- **Role of TCRs in Immunity.** T cell receptors (TCRs) plays a critical roles in adaptive immune systems as they enable killer T cells to recognize abnormal cells from healthy cells.
- **Data**
  - Positive Pairs: Readily available from databases online.
  - **Negative Pairs**: Manually generated either by random recombinations or pairing with background TCRs.



Motivation.

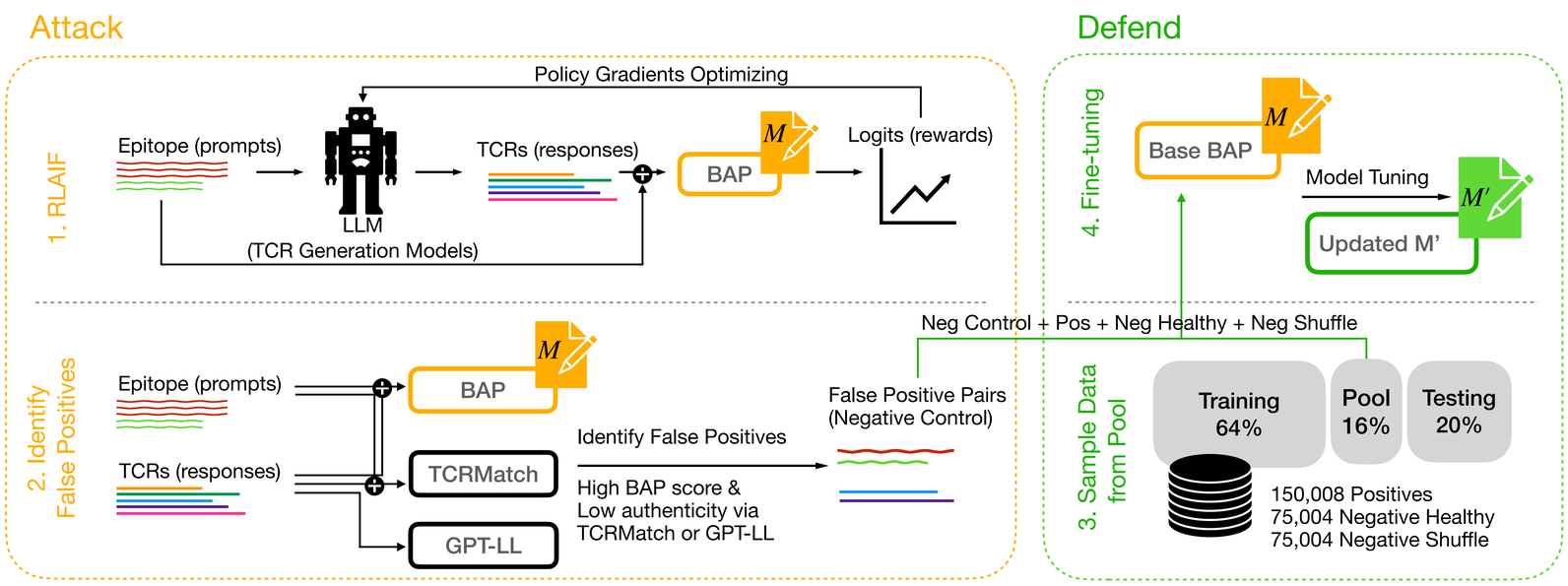
Addressing limitations of negative sample data coverage to enhance the precision and reliability of TCR-epitope binding predictions.

Contributions.

1. Identified data coverage problem of TCR-epitope binding prediction models.
2. Proposed an attack-and-defend framework for learning the limitations of data coverage.
3. Built and open-sourced a robust and easy-to-use negative control data.

Our attack-and-defend framework.

- **Attack Prediction Models**: Expose model vulnerabilities by generating and identifying biologically implausible TCR sequences.
- **Defend by Fine-Tuning**: Improve prediction models by incorporating these identified false positives, enhancing detection accuracy.
- **Iterative Refinement**: Repeat the attack and defend steps to continuously improve model robustness, reducing false positives to a minimum.

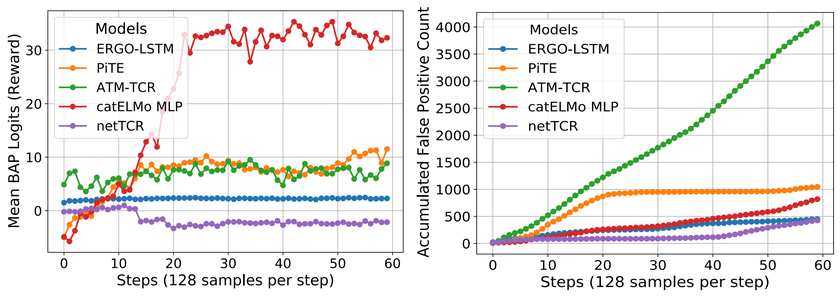


BAP attack is a universal problem.

- Five prediction models were studied, covering a diverse range of embeddings and neural network structures.

Table 1. List of Base BAP Models covered in our study.

Model Name	Architecture	Embeddings	Original Negative Set
ATM-TCR [1]	Self-attention	BLOSUM	Negative shuffle pairs
ERGO-LSTM [22]	LSTM	BLOSUM	Negative shuffle pairs
PiTE [29]	Self-attention	catELMo	Negative healthy pairs
catELMo MLP [28]	MLP	catELMo	Negative healthy pairs
netTCR-2.0 [13]	CNN	BLOSUM	Validated Negative Pairs



Our negative control dataset improve BAP robustness without additional attack.

Table 2. Change in model accuracy in percentage with augmented negative control dataset compared to base Model. Larger is better. The model benefited the most in each group is underlined.

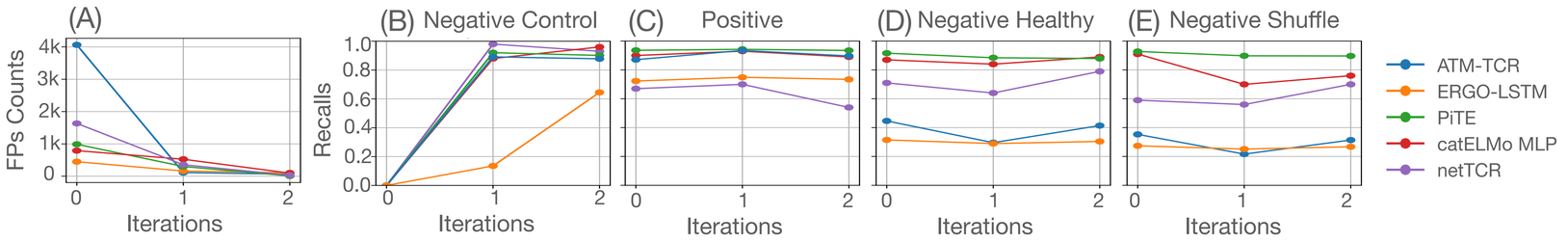
Model	Change in Model Accuracy (%)			
	Negative control	Positive	Negative healthy	Negative shuffle
ATM-TCR [1]	74.93	6.55	-1.78	-1.03
ERGO-LSTM [22]	92.21	3.83	20.53	39.75
PiTE [29]	95.33	1.29	3.16	7.66
catELMo MLP [28]	98.02	1.06	1.26	10.45
netTCR-2.0 [13]	97.19	2.46	-1.62	4.63

Acknowledgments

This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM155417.

Fine-tuning improves robustness against generated false positives

- False positives drastically reduced from thousands to under one hundred.
- Dramatically performance improvement on negative control group, while maintaining performance for other original groups.



tSNE of latent space and SeqLogo from identified false positives from each prediction model.

- Fewer false positives are detected, with prediction models becoming more robust to attack.