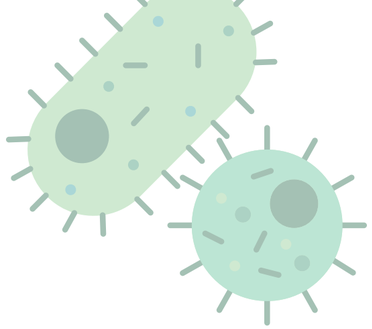
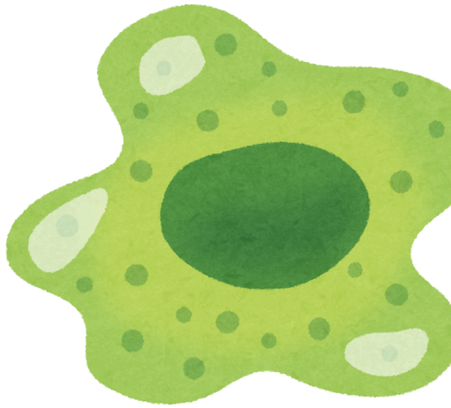
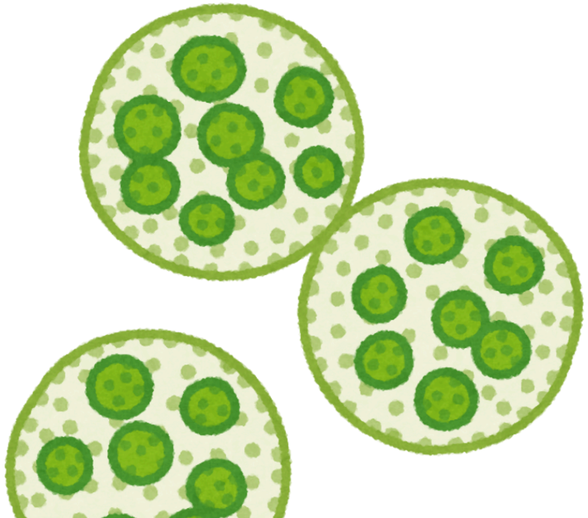


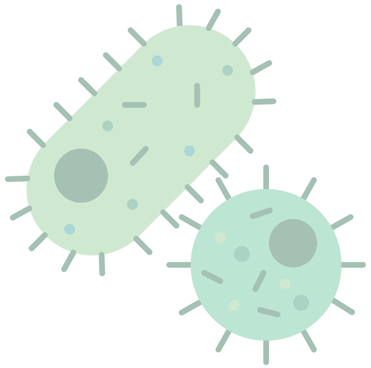


A dataset for machine learning (ML)-guided antibody discovery

https://doi.org/10.1016/j.cels.2024.11.005

Long DO-Pham-The





Cell Systems

CellPress
OPEN ACCESS

Report

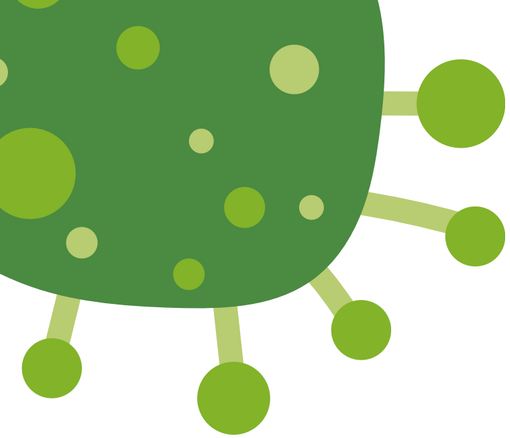
Evaluating predictive patterns of antigen-specific B cells by single-cell transcriptome and antibody repertoire sequencing

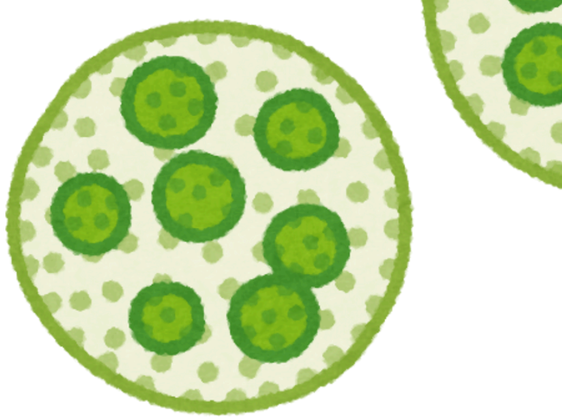
Lena Erlach,[1] Raphael Kuhn,[1] Andreas Agrafiotis,[1,2] Danielle Shlesinger,[1] Alexander Yermanos,[1,3,4] and Sai.T. Reddy[1,4,5,*]
[1]Department of Biosystems Science and Engineering, ETH Zurich, 4057 Basel, Switzerland
[2]Institute of Microbiology, ETH Zurich, 8049 Zurich, Switzerland
[3]Center for Translational Immunology, University Medical Center Utrecht, 3584 CX Utrecht, the Netherlands
[4]Botnar Institute of Immune Engineering, 4056 Basel, Switzerland
[5]Lead contact
*Correspondence: sai.reddy@ethz.ch
https://doi.org/10.1016/j.cels.2024.11.005

SUMMARY

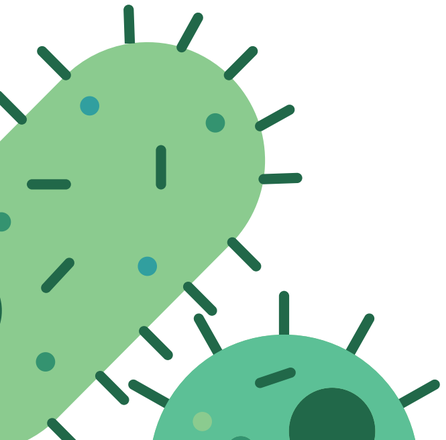
The field of antibody discovery typically involves extensive experimental screening of B cells from immunized animals. Machine learning (ML)-guided prediction of antigen-specific B cells could accelerate this process but requires sufficient training data with antigen-specificity labeling. Here, we introduce a dataset of single-cell transcriptome and antibody repertoire sequencing of B cells from immunized mice, which are labeled as antigen specific or non-specific through experimental selections. We identify gene expression patterns associated with antigen specificity by differential gene expression analysis and assess their antibody sequence diversity. Subsequently, we benchmark various ML models, both linear and non-linear, trained on different combinations of gene expression and antibody repertoire features. Additionally, we assess transfer learning using features from general and antibody-specific protein language models (PLMs). Our findings show that gene expression-based models outperform sequence-based models for antigen-specificity predictions, highlighting a promising avenue for computationally guided antibody discovery.

Contents





1. Problem: The need of a dataset about antigen-specific labeling



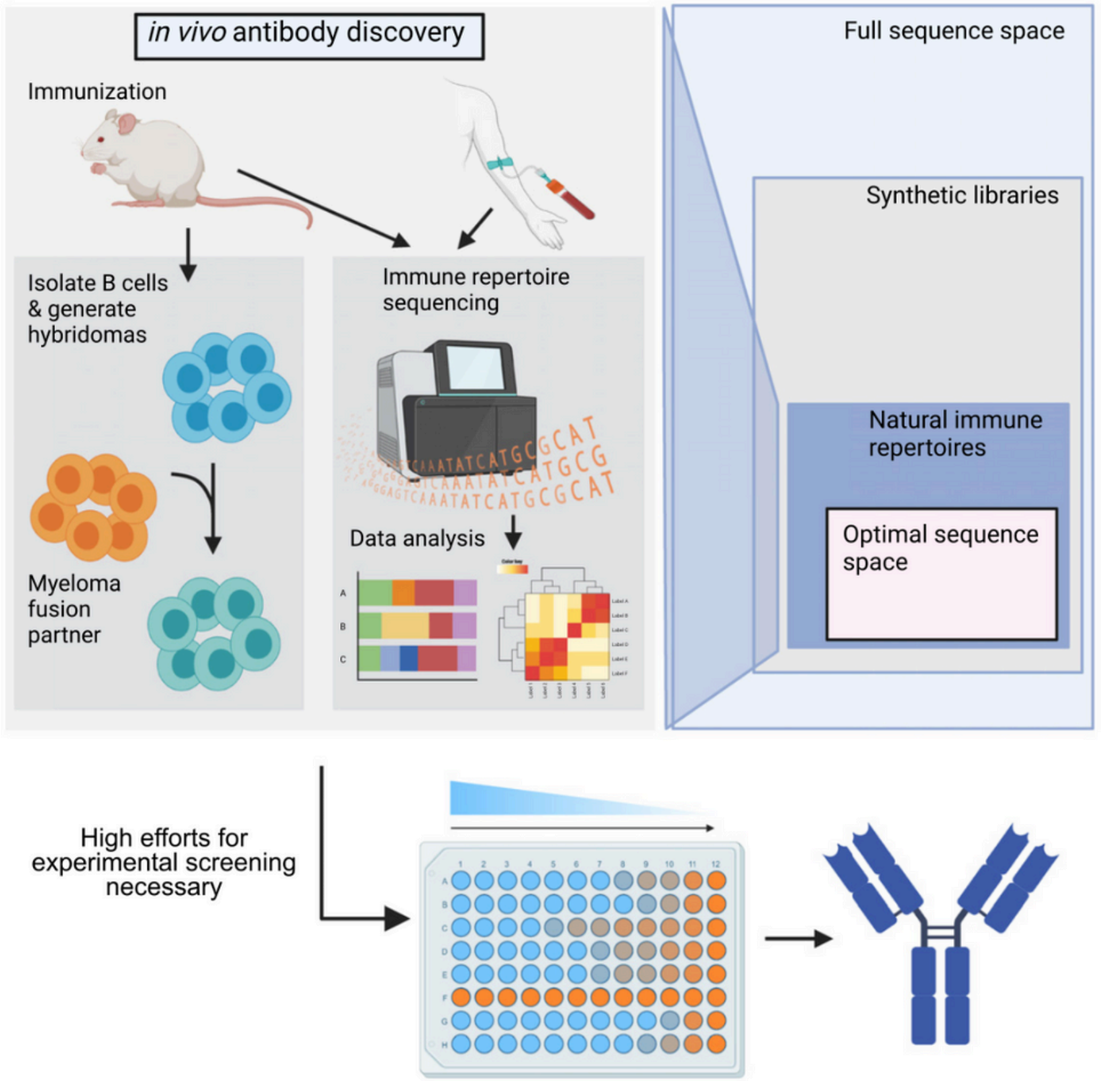
We have multiple antibody clones inside our body.

But which ones should be chosen for therapeutic development?

=> The **antigen-specific antibodies** (the ones that respond to the antigen of the invaders)

Antigen-specific antibody
=> therapeutic candidate for drug development

Non-Antigen-specific antibody
=> Ignore



How to **identify suitable therapeutic antibodies** (antigen-specific ones) among thousands of candidates?

Traditional methods require a lot of laboratory work with many screening steps to identify “the chosen ones”

=> expensive, exhaustive, tedious

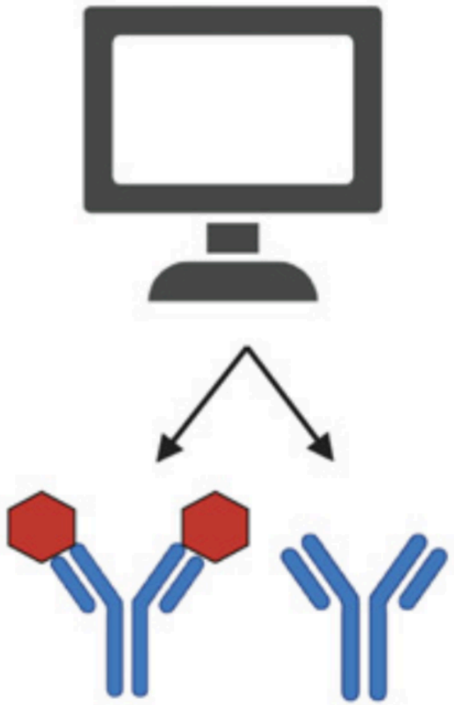
Classification models

Logistic regression

Kernel SVM

Random Forest

Gradient boosting



How about training a Machine Learning (ML) model to help us predict:

- which antibody clones are antigen-specific?
- which ones are not?

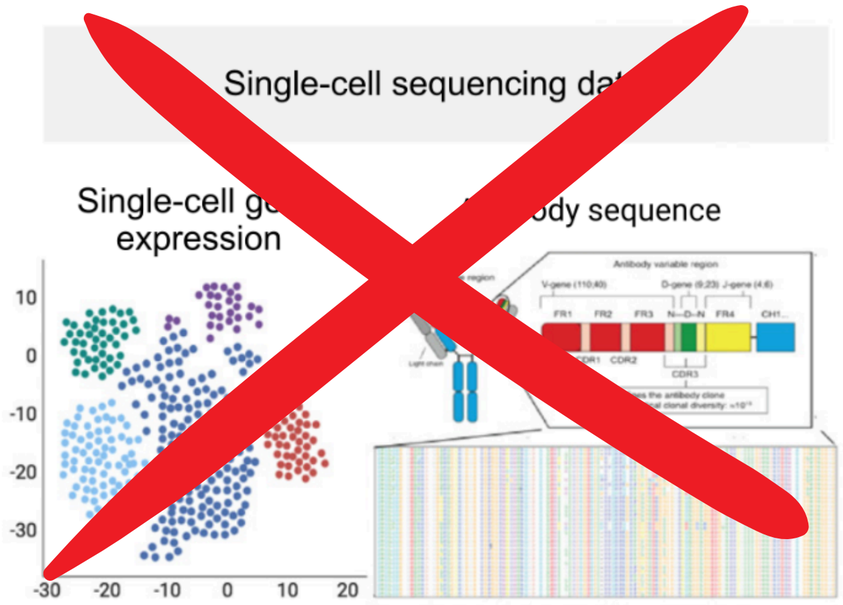
Identifying antigen-specific antibodies from the others is a **classification task**

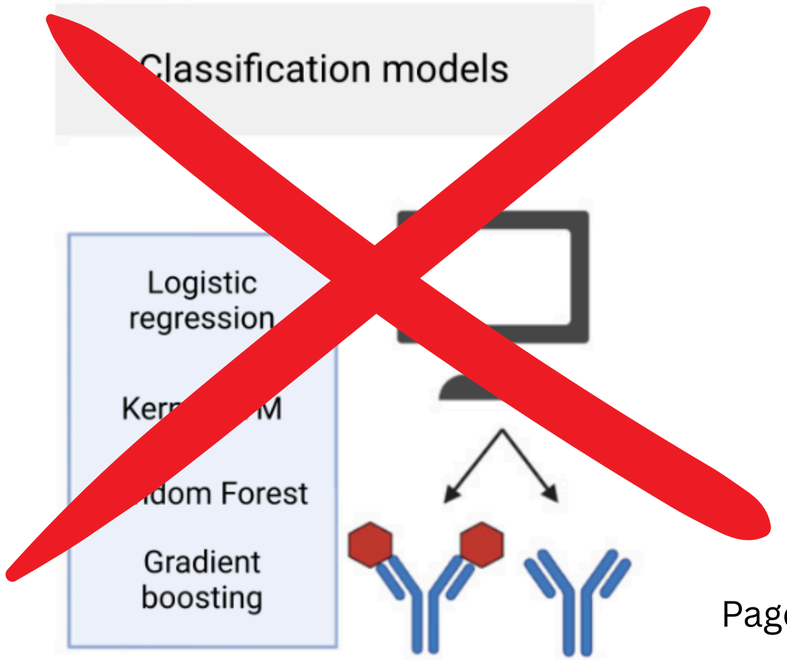
=> **need train a Supervised Learning model**

To train a Supervised Learning model

=> **need a labeled dataset**

PROBLEM: we don't have that much labeled datasets about antigen-specific antibodies yet





The best way to predict the future
is to create it.

- Peter Drucker -

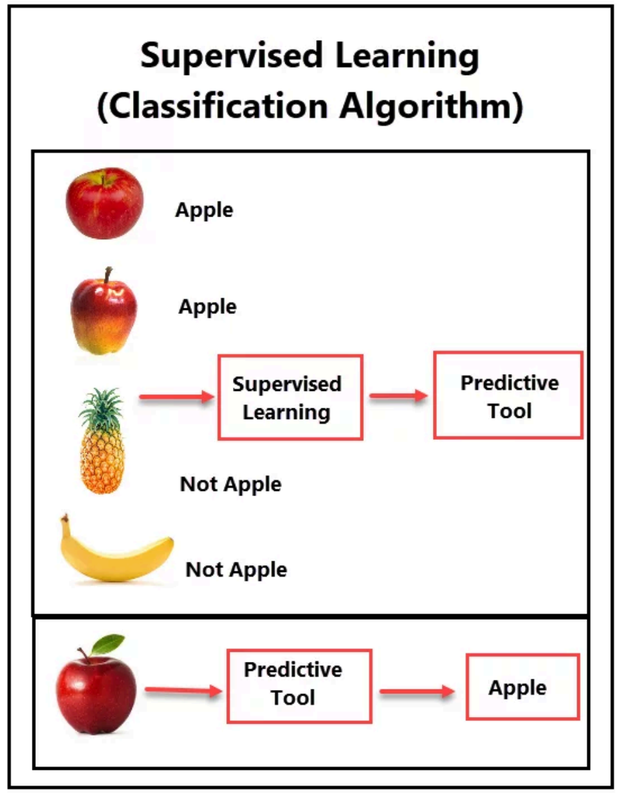
The best way to get the data
is to GENERATE the DATA

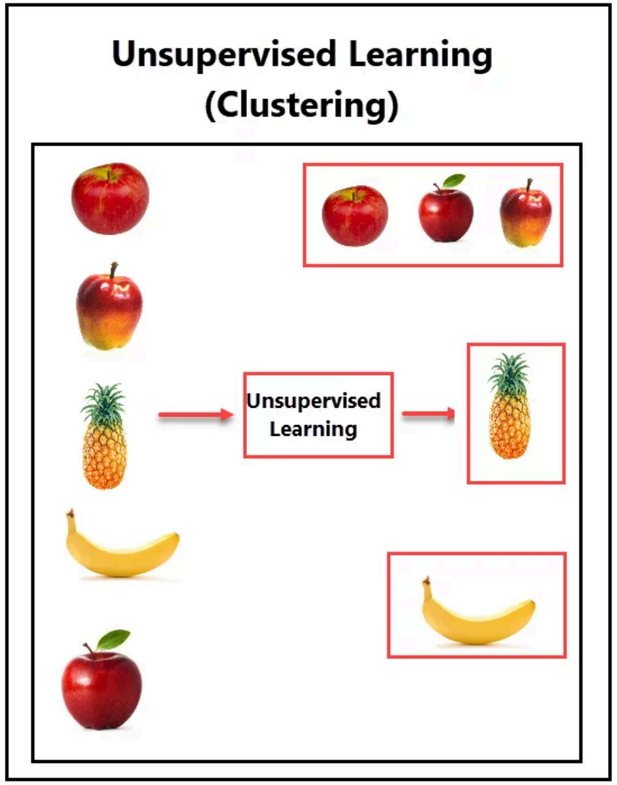
Supervised learning:

- The inputs are images of fruits with 2 labels: “Apple” and “Not apple”
- Use these images to train a predictive tool to distinguish apples from not-apple fruits
- Next time, we give it a new fruit and ask “is this an apple?” -> it will predict and give the result as “apple” or “not-apple”

Unsupervised learning

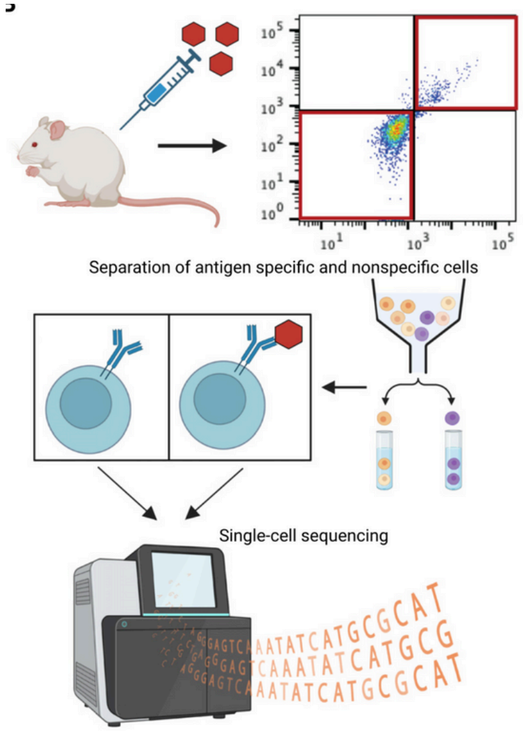
- The images of the fruits are not labeled (unlabeled)
- We can still train a model to distinguish one fruit from the others
- The difference here is that the model can only split the fruits into different groups, but cannot call out the name (or the label). This is also called clustering.



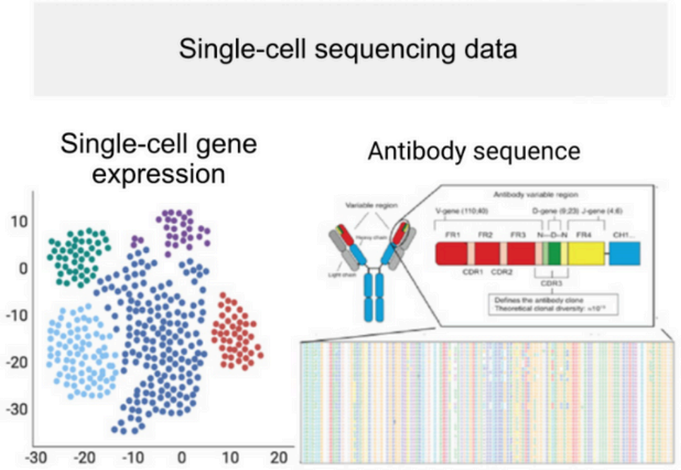


2. Experiment design:

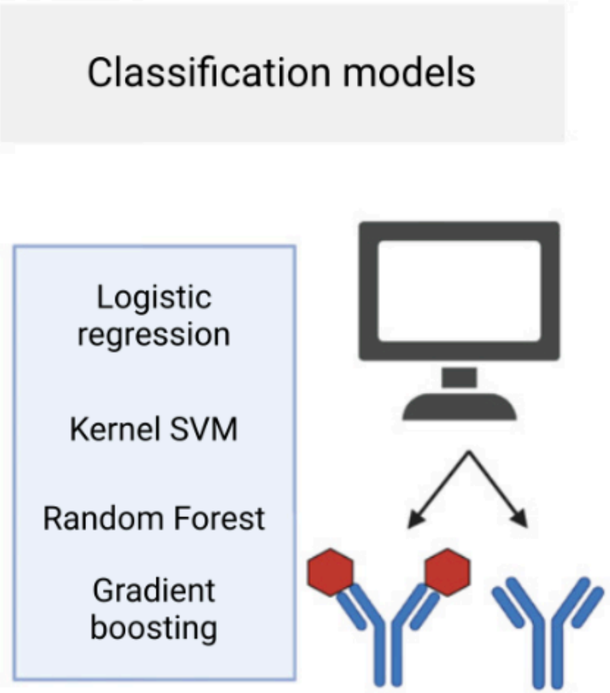
The use of single B cell repertoire analysis



Experimental works



Data generation



Training

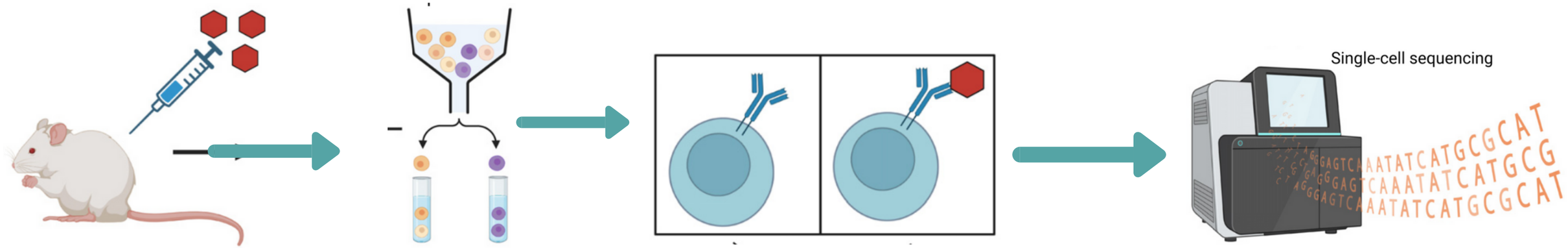
Step 1: Experimental works

Inject Ovalbumin (OVA) and SARS-CoV-2 RBD (RBD) into mice to trigger immune response

Collect the samples, then use FACS technology to identify antigen-specific cells from others (labeling)

Run single-cell sequencing

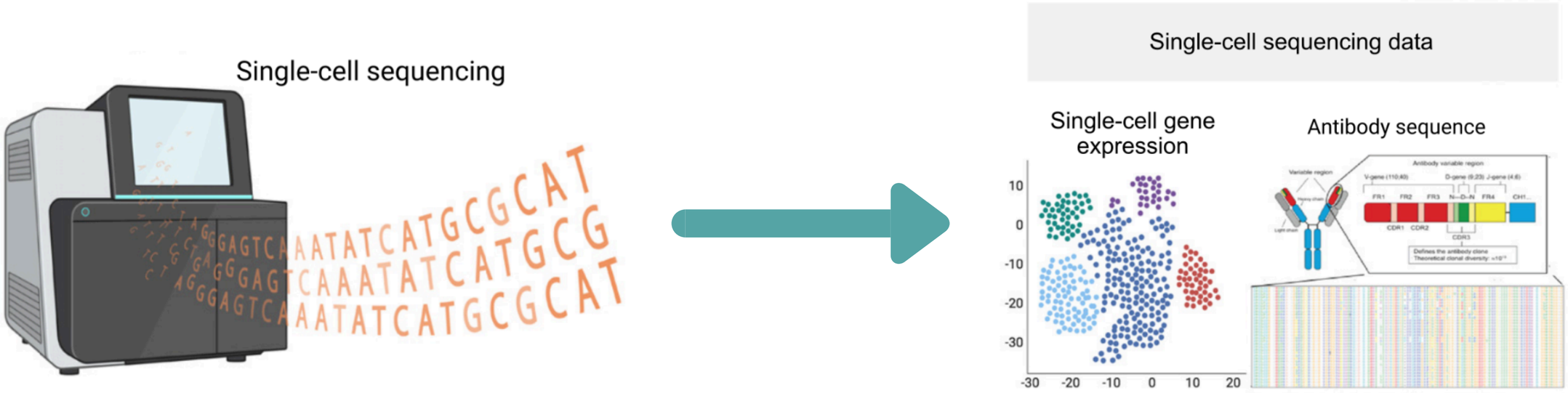
OVA and RBD (antigens to trigger immune response)



Step 2: Data Generation

After the single-cell sequencing, they performed downstream analysis to generate useful data for training.

Particularly, that is **gene expression** and **antibody sequence**.



Data generation

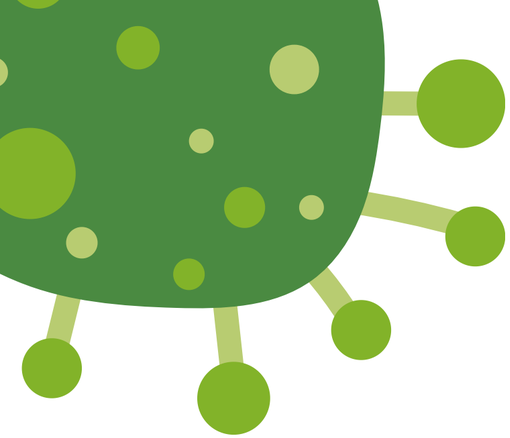
Step 3: Training

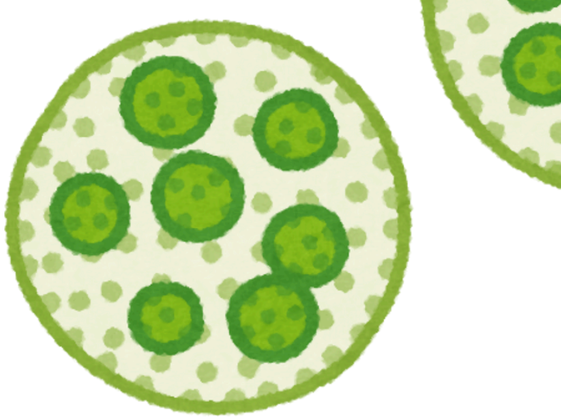
After obtaining the data,
they used it to train some classification models
=> To see if their data can be used to train a good model or not

Two datasets: gene expression and antibody sequence
Used them both to train the models
=> To see which dataset is better for training

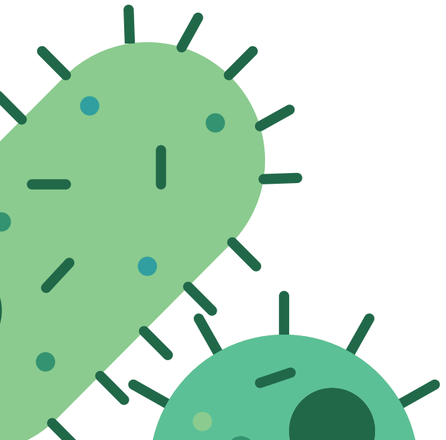
Data generation

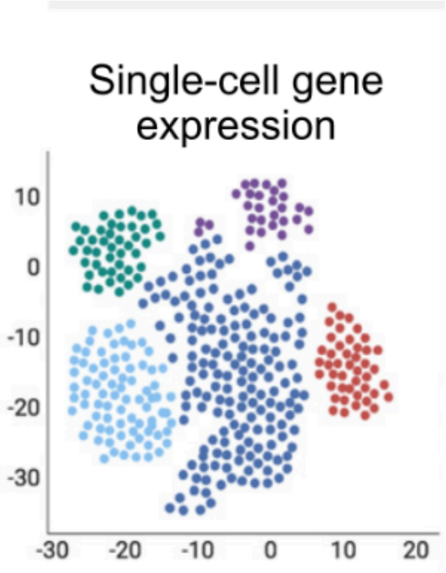
Training





3. Results

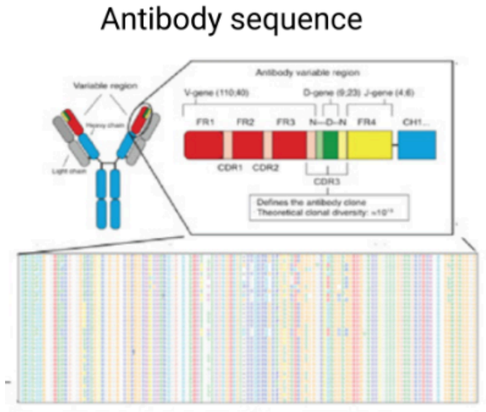




The training using gene expression data

- kSVC: yielded **0.856** F1 score for OVA
- GBoost: yielded **0.94** F1 score for RBD

=> Good result (high F1 score)

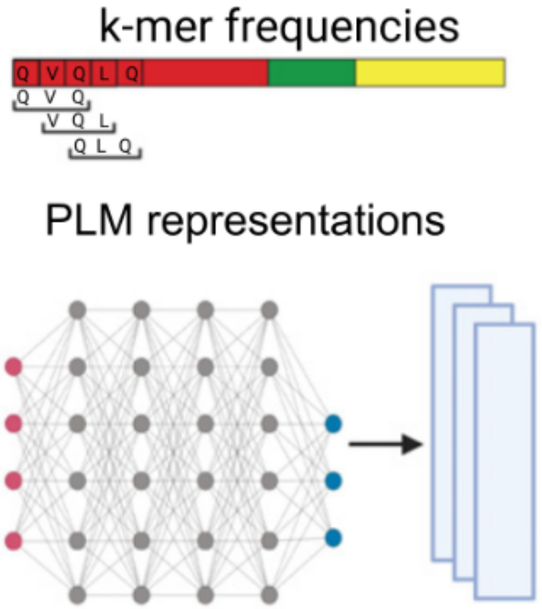


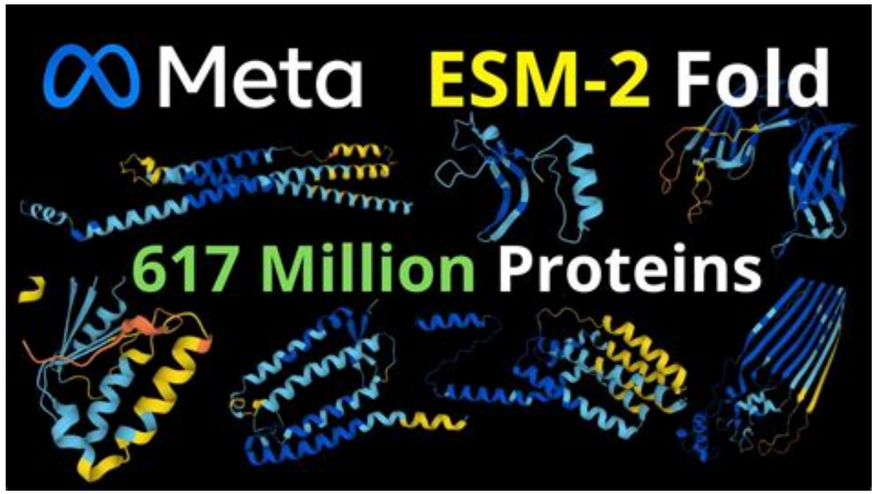
The training using antibody sequence data

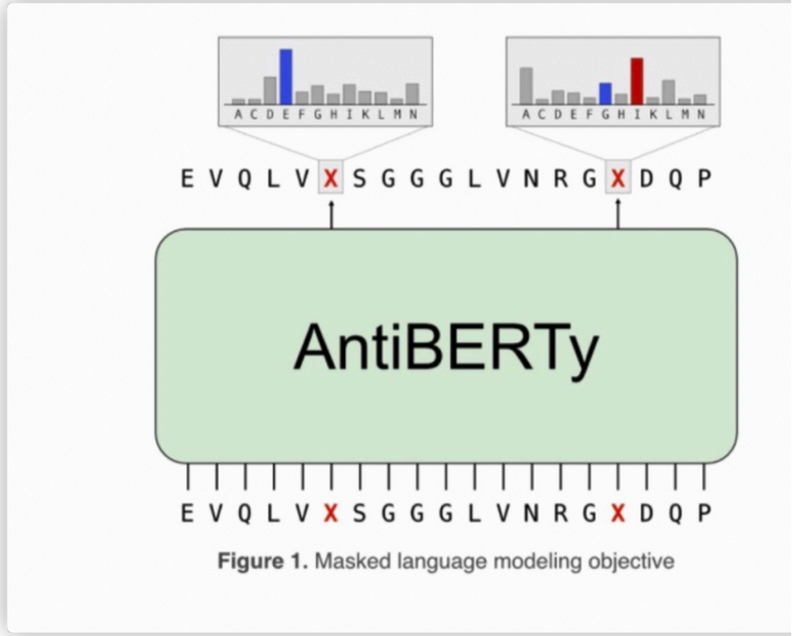
- Random Forest: yielded **0.807** F1 score for OVA
- Random Forest: yielded **0.68** F1 score for RBD

Gene Expression turns out to be **the better dataset** for training antigen-specific antibody classification model

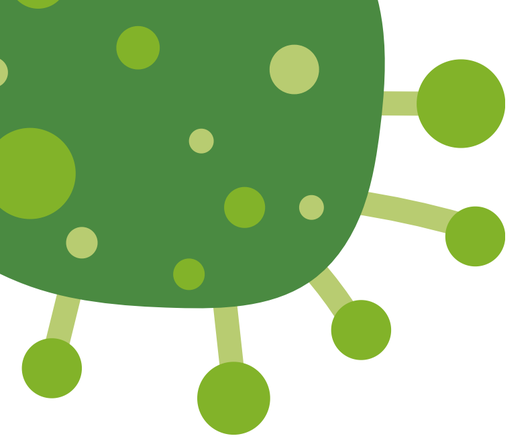
Training the data with some Protein Language Model (deep learning) like ESM2, ESM3 and AntiBERTY **DOES NOT** improve the predictions

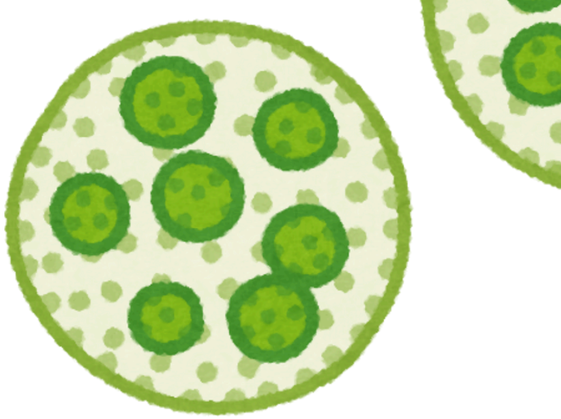




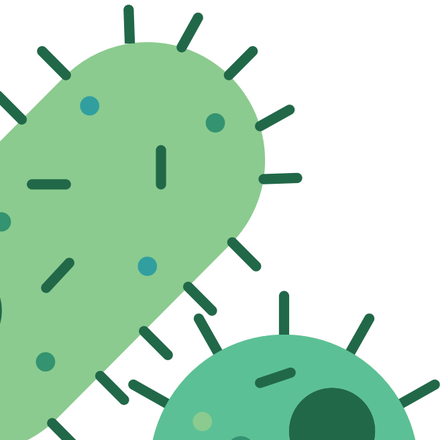


Deep Learning is not always the best solution. Sometimes, **basic machine learning models can outperform deep learning ones.**





4. Summary



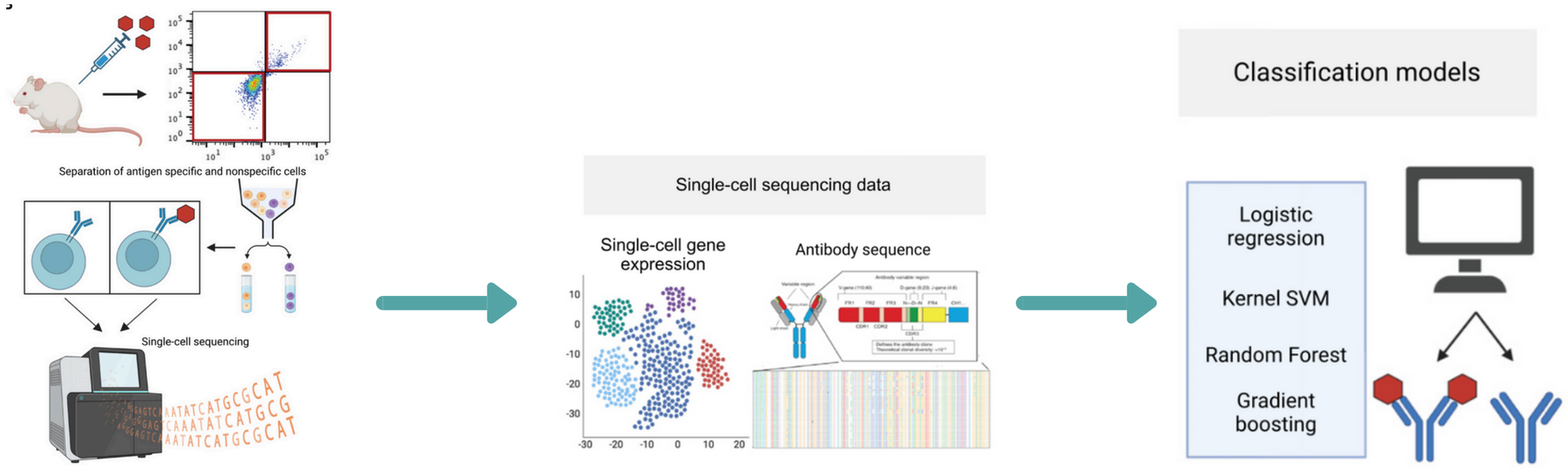
The authors of this paper performed some experimental works to generate a single-cell sequencing dataset **(they focused on providing us with a dataset for training**, not the models**)**.

This dataset is **antigen-specific labeling**
(it tells us which antibody responds to which antigen, which antibody does not)

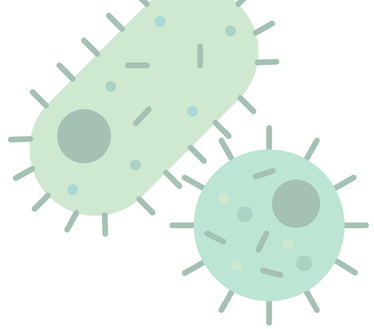
Then, they train some classification models to see if they can classify antigen-specific antibodies from others

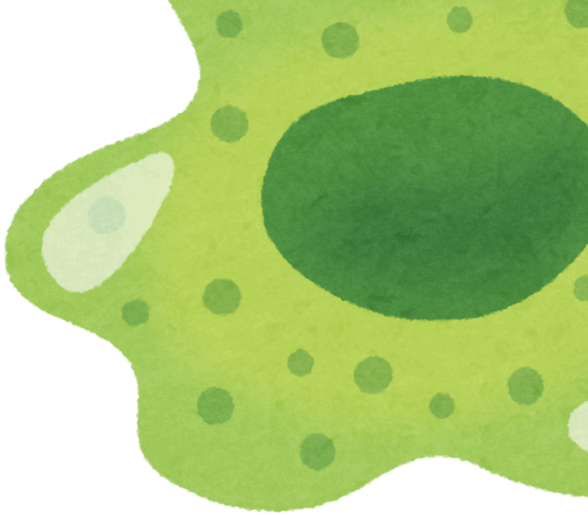
Outcome:

- The dataset is useful for training antigen-specific antibody classification model
- Gene expression is the better dataset/feature for training



Experimental works | Data generation | Training





Thank You