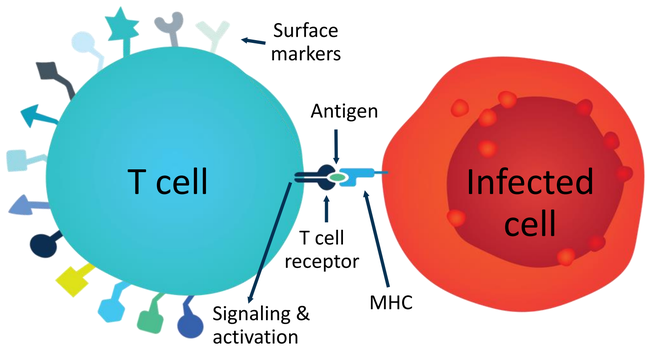
Machine learning to predict viral specificity of CD8+ T cells from high throughput multi-omics data

Florian Schmidt [1], Christine Eng[1], Kanxing Wu[1], Katja Fink[1], Michael Fehlings[1], Dan MacLeod[2], Andreas Wilm[1, #]

1: ImmunoScape Pte Ltd, 1 Scotts Road, #24-10, Singapore 228208, 2: ImmunoScape, Inc., 11575 Sorrento Valley Road, Suite 204, San Diego, CA 92121, United States, #andreas.wilm@immunoscape.com

ImmunoScape

Motivation



CD8+ T cells can interact with a wide variety of antigens.

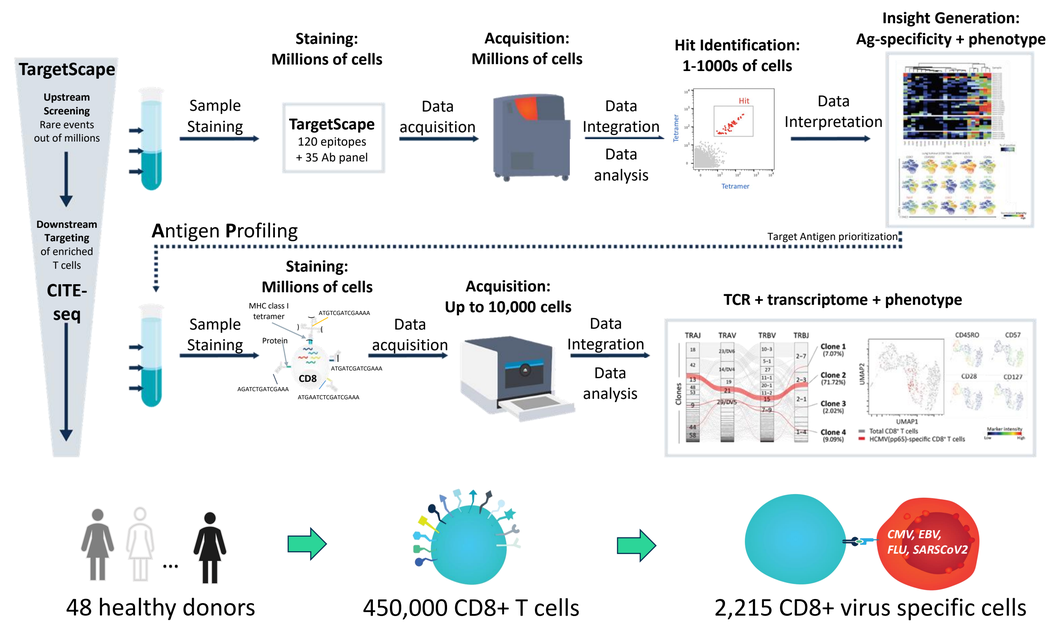
Inferring T cell specificity remains challenging in-silico.

Antigen specificity is determined by TCR sequences, which are determined during T cell maturation (VDJ recombination).

Goal

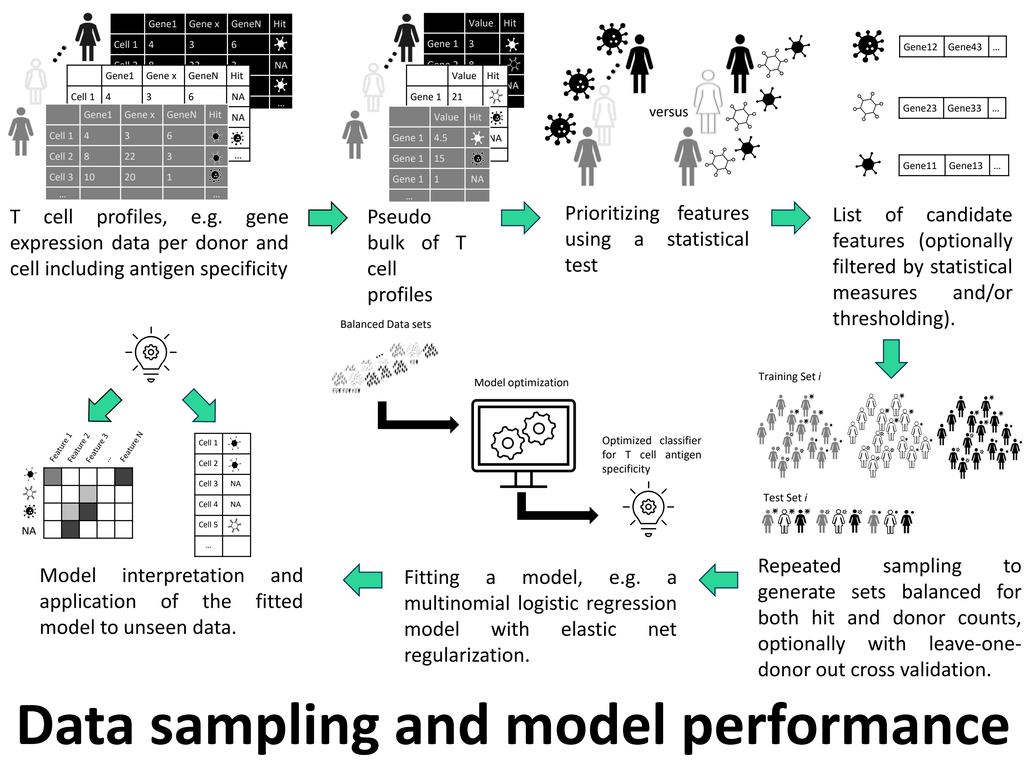
Phenotype: Mass cytometry + Single cell *omics* → Machine learning → Antigen specificity

Data generation



48 healthy donors → 450,000 CD8+ T cells → 2,215 CD8+ virus specific cells

Machine learning pipeline



T cell profiles, e.g. gene expression data per donor and cell including antigen specificity

Pseudo bulk of T cell profiles

Prioritizing features using a statistical test

List of candidate features (optionally filtered by statistical measures and/or thresholding).

Repeated sampling to generate sets balanced for both hit and donor counts, optionally with leave-one-donor out cross validation.

Fitting a model, e.g. a multinomial logistic regression model with elastic net regularization.

Model interpretation and application of the fitted model to unseen data.

Model fitting and optimization

Data sampling:

1) Randomly sample ***x*** number of cells from each donor
2) Sampling at least ***y*** cells per source

Elastic net parameter:

Perform a grid search for the elastic net α parameter (0 to 1 with step-size 0.05)

Monte Carlo Cross Validation:

10 outer folds (20% test, 80% training) with 6-fold inner cross validation for tuning of regression coefficients and regularization parameter λ.

Average training and test error across 10 outer folds

Average regression coefficients across 10 outer folds

Retrain final model with the identified parameters on the entire dataset for future predictions

Performance estimates and model Interpretation

Different feature sets

	Gene 1	Gene x	Gene N	*Marker 1*	*Marker x*	*Marker N*	Hit
Cell 1	4	3	6	*3*	*9*	*4*	NA
Cell 2	8	22	3	*2*	*12*	*3*	NA
Cell 3	10	20	1	*7*	*8*	*5*	
...							...

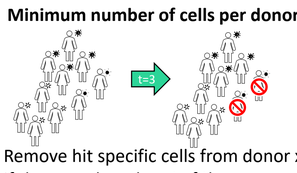
	Gene 1	Gene x	Gene N	Hit
Cell 1	4	3	6	NA
Cell 2	8	22	3	NA
Cell 3	10	20	1	
...				...

	Marker 1	*Marker x*	*Marker N*	Hit
Cell 1	*3*	*9*	*4*	NA
Cell 2	*2*	*12*	*3*	NA
Cell 3	*7*	*8*	*5*	
...				...

Feature set: Combined, RNAonly, ADT only

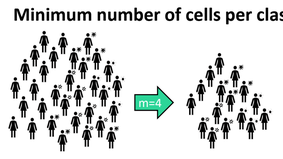
Data sampling and model performance

Minimum number of cells per donor

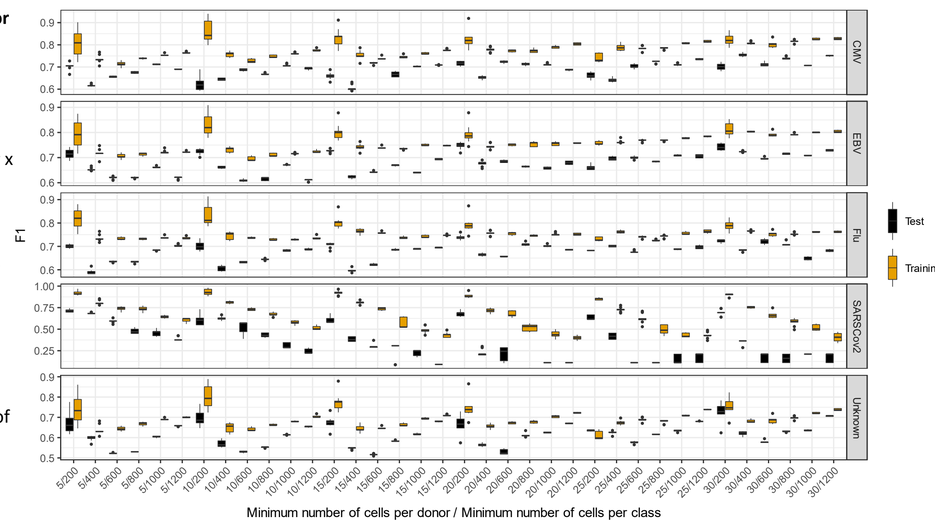


Remove hit specific cells from donor x if there are less than *t* of them.

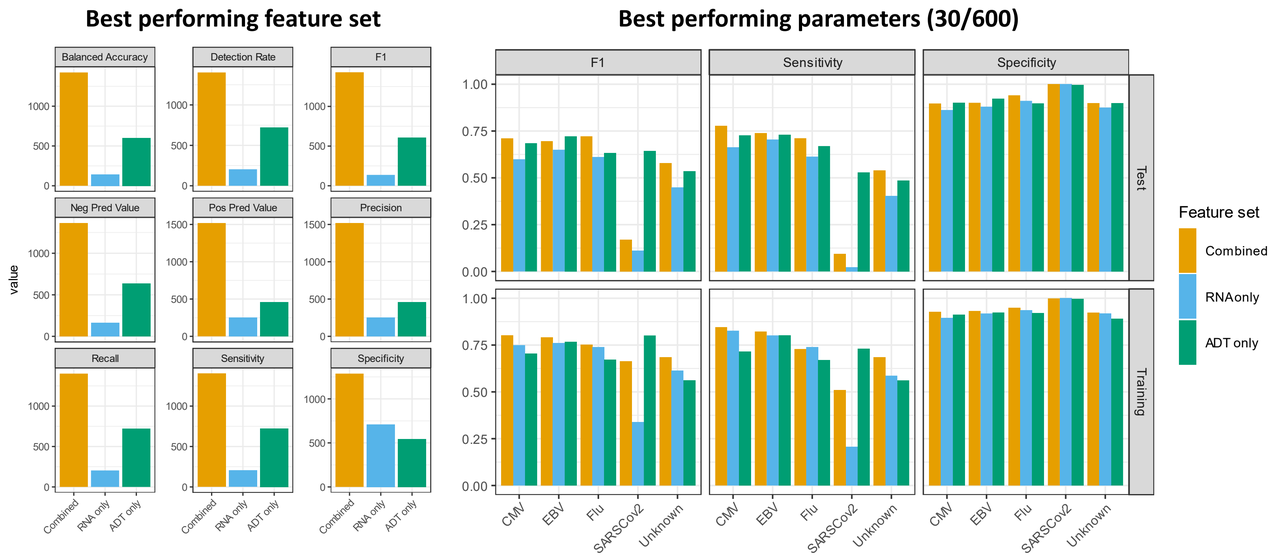
Minimum number of cells per class



Attempt to sample equal number of cells *m* per category.

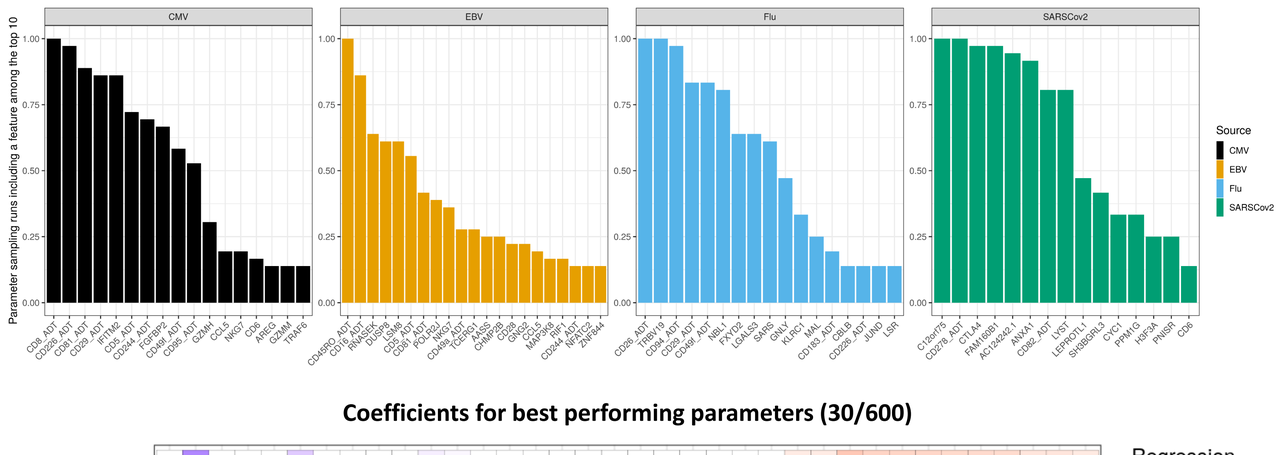


Performance of different feature sets

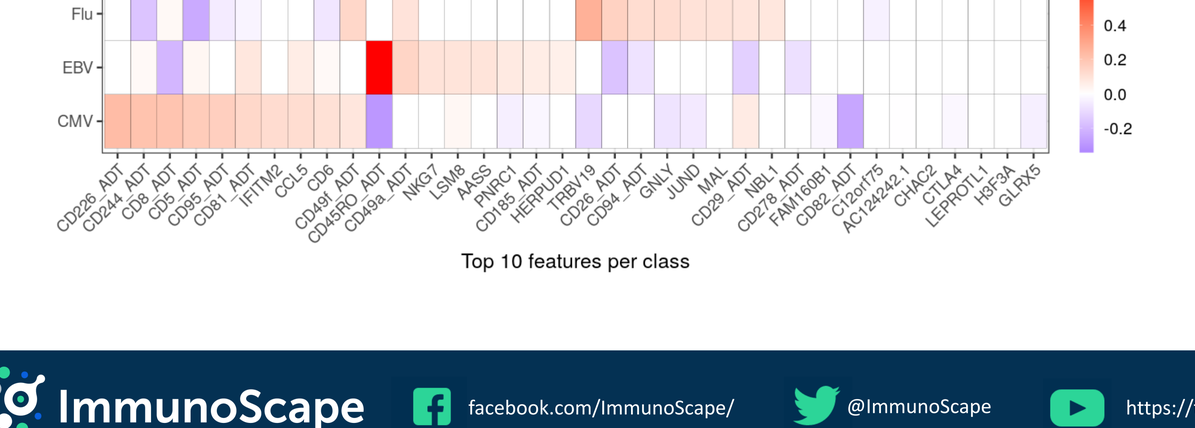


Model coefficients

Coefficient stability against data sampling

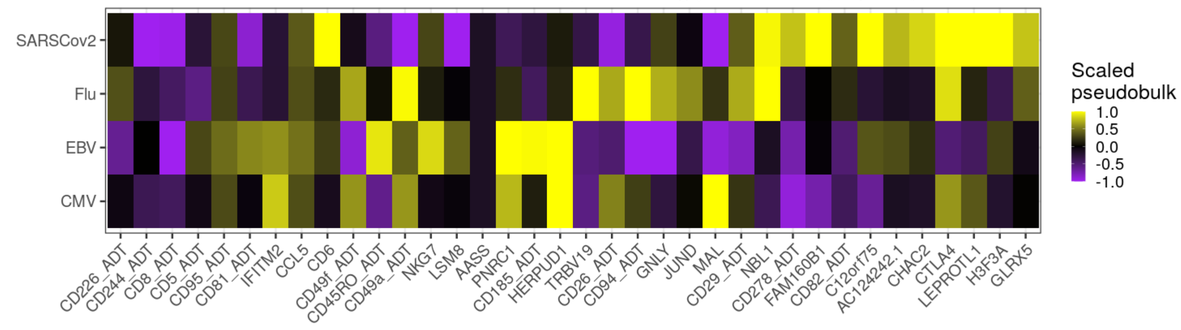


Coefficients for best performing parameters (30/600)

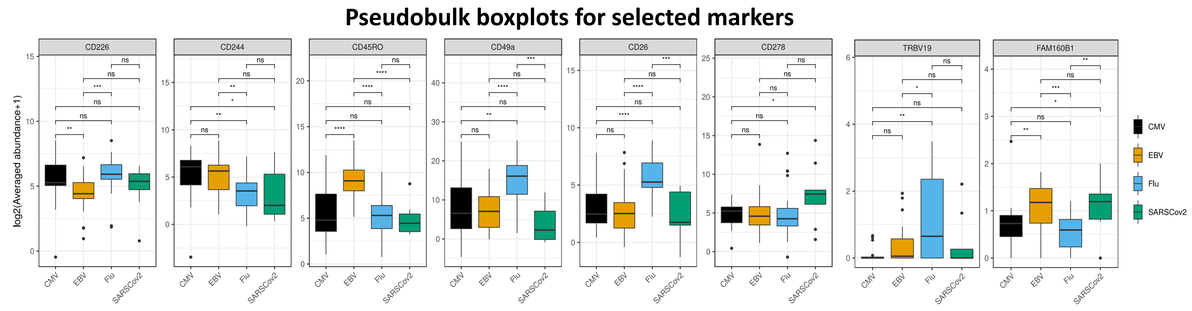


Phenotype

Abundance of features for best performing parameters (30/600)



Pseudobulk boxplots for selected markers



Conclusions

Antigen specificity can be well predicted from phenotypic single cell data.

Data balancing and sampling are important for generalization.

Features based on both scRNA-seq and surface markers are relevant.