Australian National University

Application of deep learning for better batch effect removal allows detection of subtle cellular phenotypes from large flow datasets

Dr. Ben Mashford

(ANU School of Computing & John Curtin School of Medical Research)

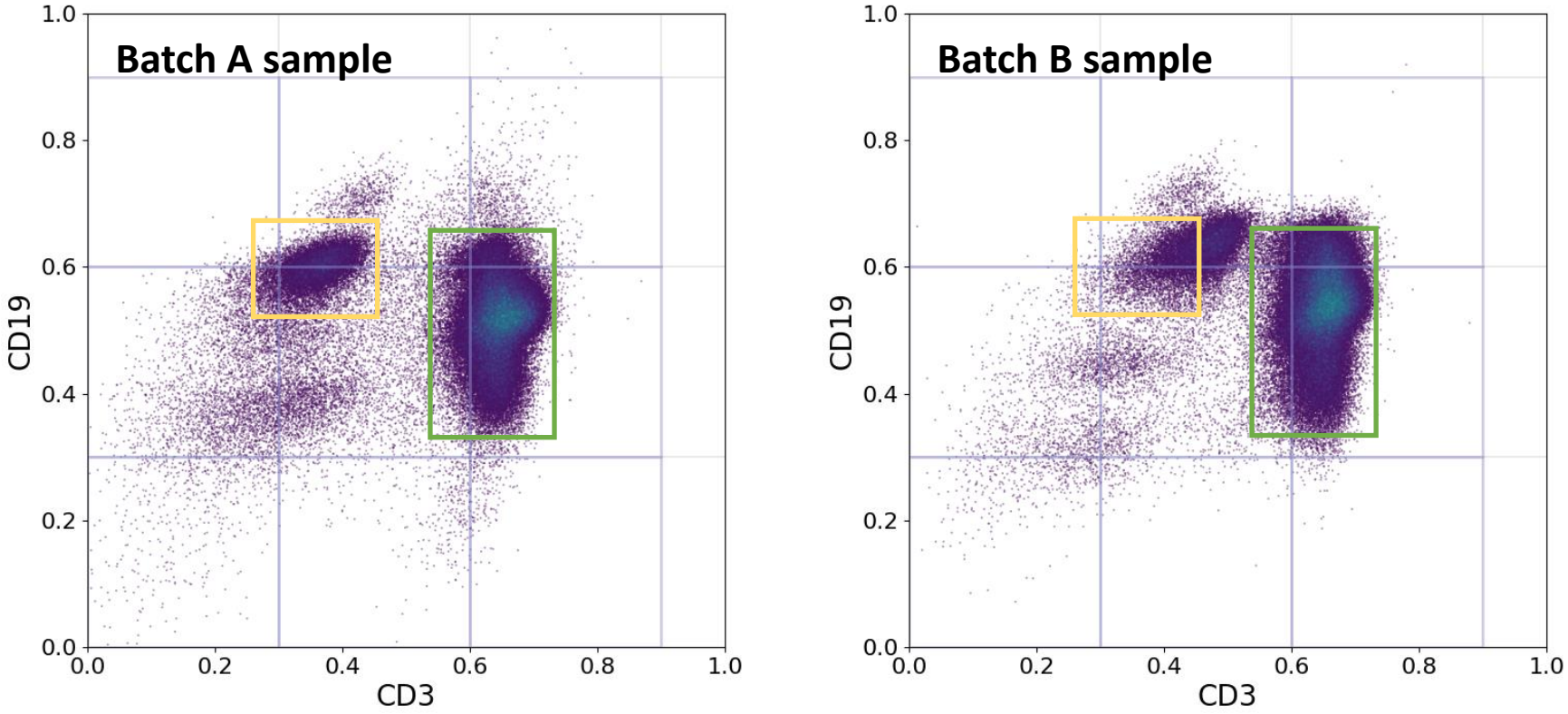
ACS2024

23/10/2024

Batch effects in flow cytometry

- Batch-related variability in marker intensities are inherent to flow cytometry.
- Some major contributions to these batch-effect induced intensity shifts include:
  - **Changes in selection of antibody markers and reagent concentrations**
  - **Operator technique**
  - **Changes in instrument intensity calibration**
- Human operator-defined gates can adjust for this variability, but the process is time-consuming and potentially prone to bias.
- Batch effects can make it difficult to identify and measure subtle phenotypes.

Batch shifts in flow cytometry data



- We can visualise the batch effect by plotting channel pairs.
- The shifts are non-linear - affecting some cell populations more than others.
- It is important to not ‘over-align’ the samples, which may remove biological signal.

Batch alignment methods

- There exist a number of algorithms for batch alignment of flow cytometry samples:

    1) **CytoNorm**
    2) **CytofBatchAdjust**
    3) **(iMUBAC) Multibatch data integration Casanova**
    4) **CyCombine**

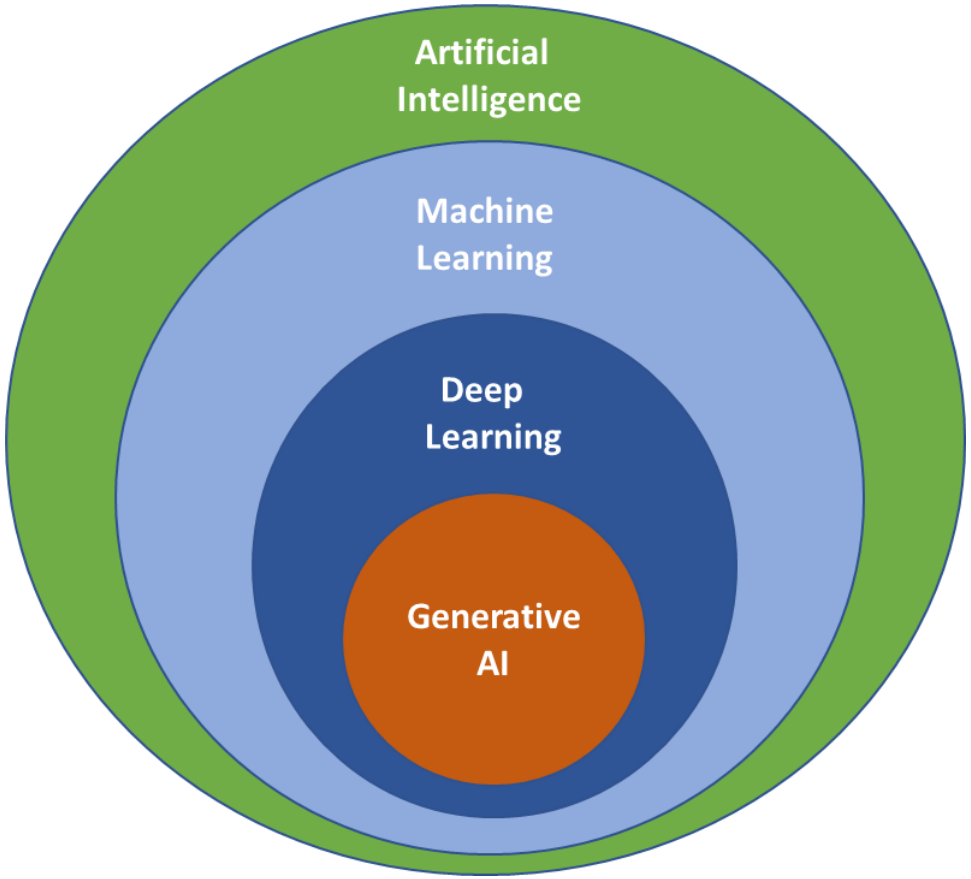
- Methods #1 and #2 require inclusion of a technical replicate across batches for comparison – very often not possible in real world datasets.

- Success of deep learning is due to ability to generalize over noisy, high-dimensional data - demonstrated in: images, video, audio, signal processing.

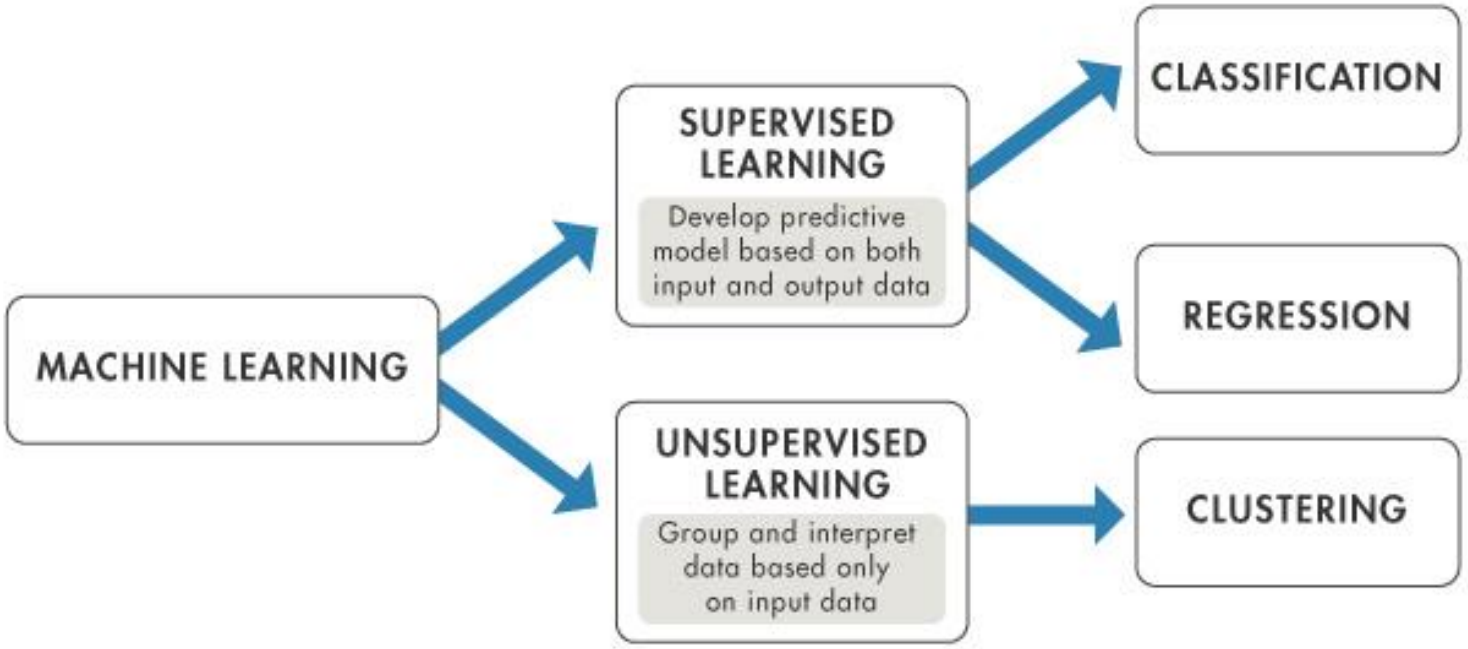
- We have developed a deep learning batch alignment method. Does not require technical replicate.

- Working to validate it against a range of experimental artifacts.

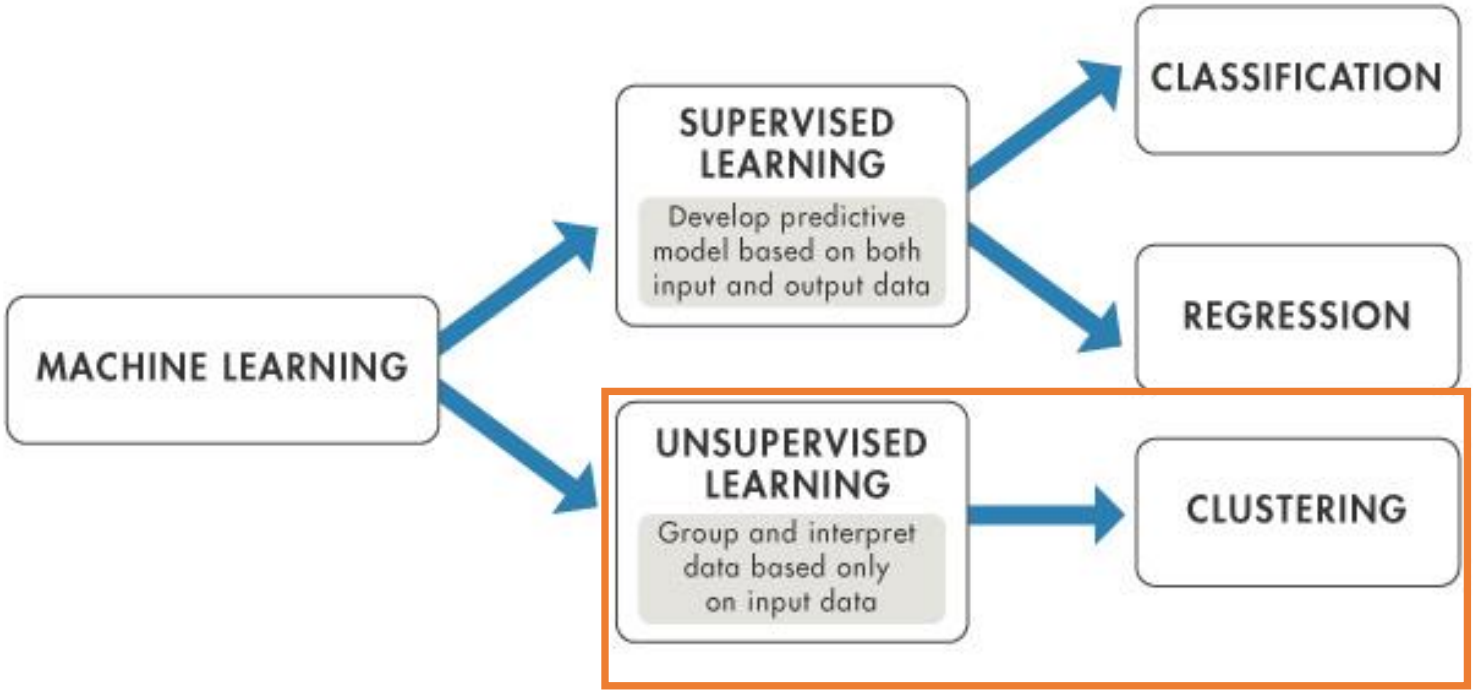
Deep learning vs. Machine Learning?



What can we do with Machine Learning?

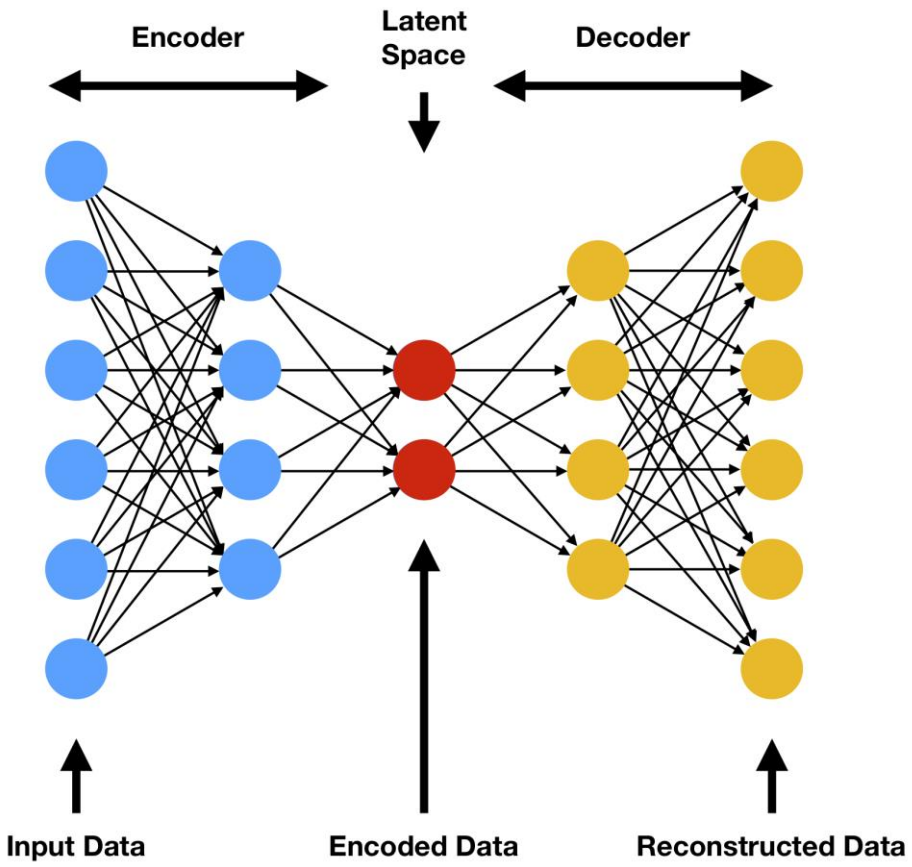


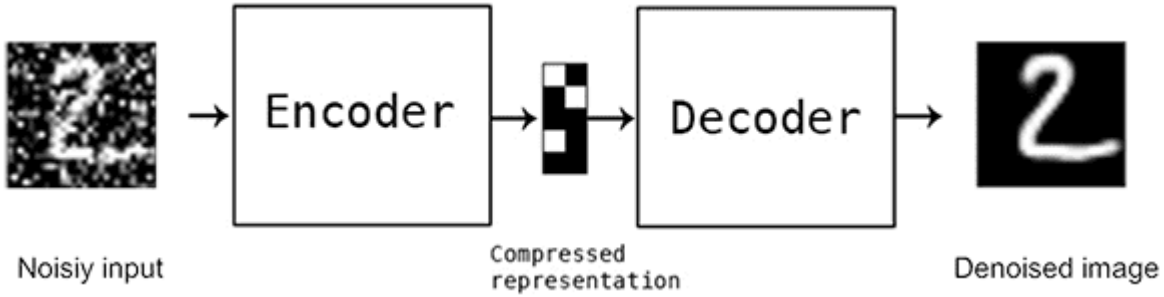
What can we do with Machine Learning?



- **Can we use learn the cell marker distributions in an unsupervised manner?**
- **Can we use a trained ML model to process flow cytometry data for us?**

Autoencoder models for batch alignment

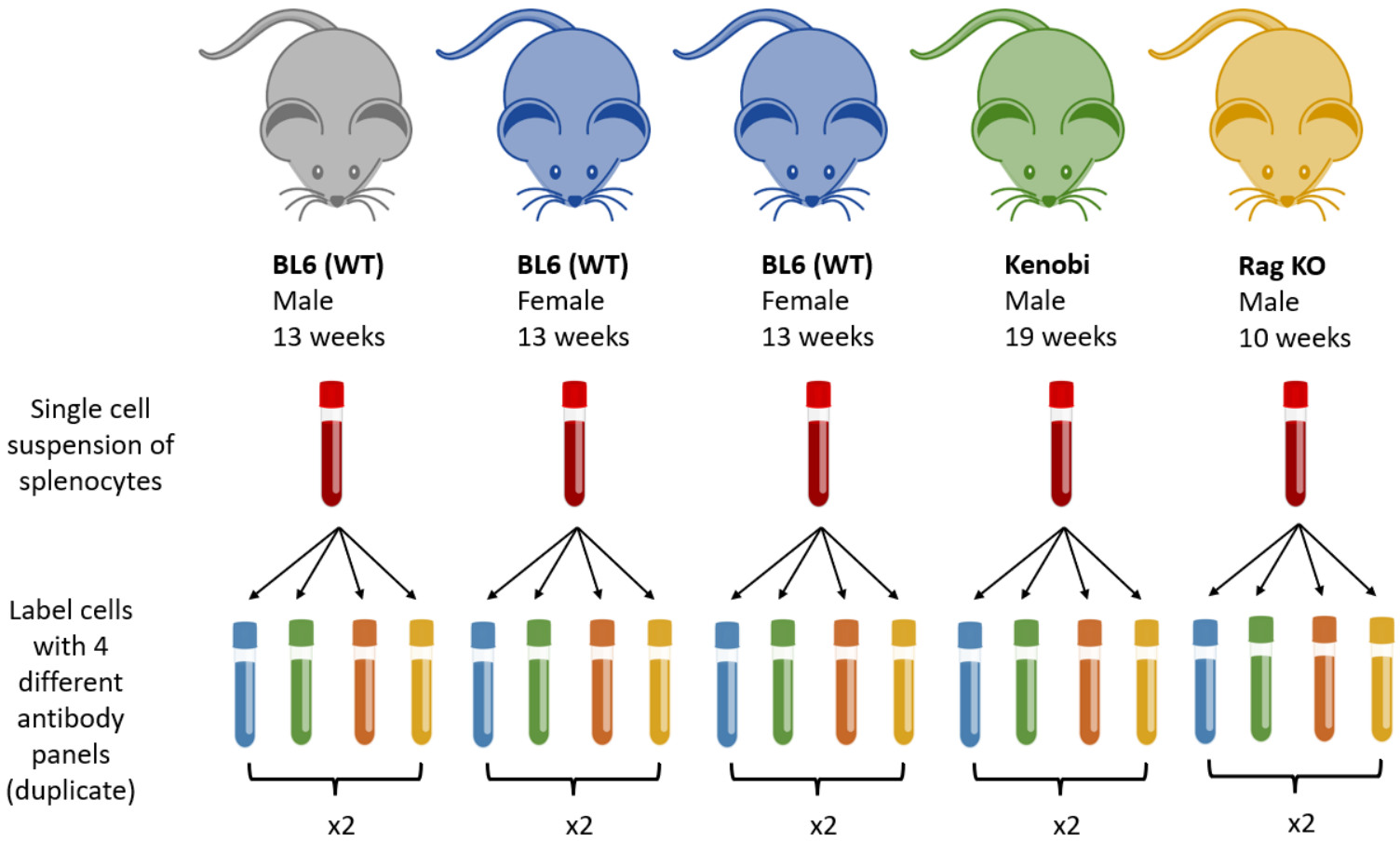




We train an autoencoder to remove batch effects from flow cytometry data. We call our model '**FlowCoder'**.

1. We train a model using **Sample A**.
2. Then feed in other samples (e.g. **Sample B** or **C**) and reconstruct them.
3. The reconstructed data is reconstructed using the latent space features from **Sample A**.

Batch alignment experiment



***Experiment designed and performed by Dillon Hammill**

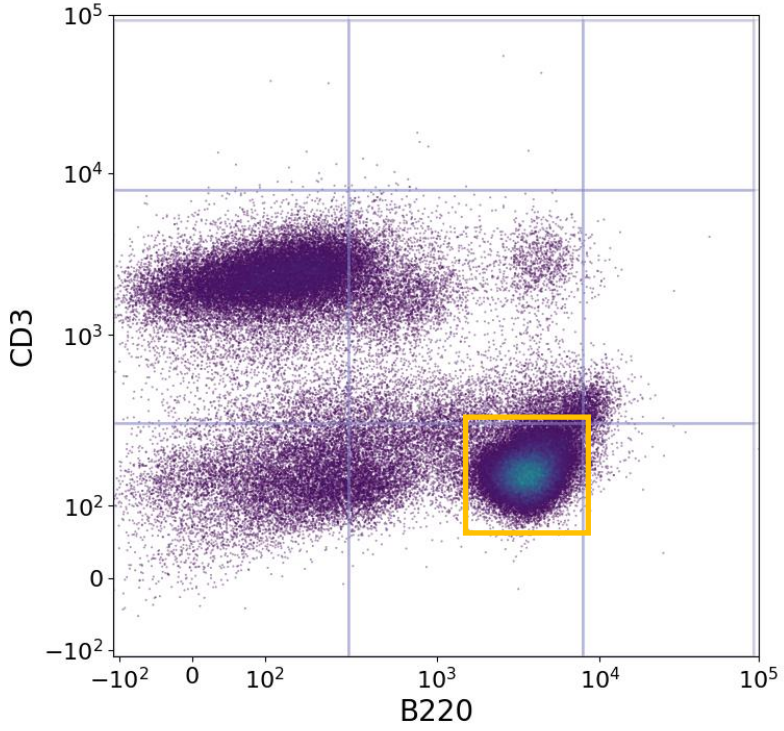
Batch normalisation experiment

Antibody	Panel 1	Panel 2	Panel 3
CD3 FITC	200	200	300
CD25 PE	400	400	400
CD8	100	400	400
CD44 PacBlue	100	400	400
CD62L BV605	400	600	600
CD19 BV510	600	800	800
IgD PerCP	200	800	800
IgM AF700	100	100	600
NK1.1 APC	200	400	600
Ly6C PECy7	200	200	400
B220 BUV737	300	600	800
CD4 AF700	200	300	600
LD APC Cy7	400	600	800

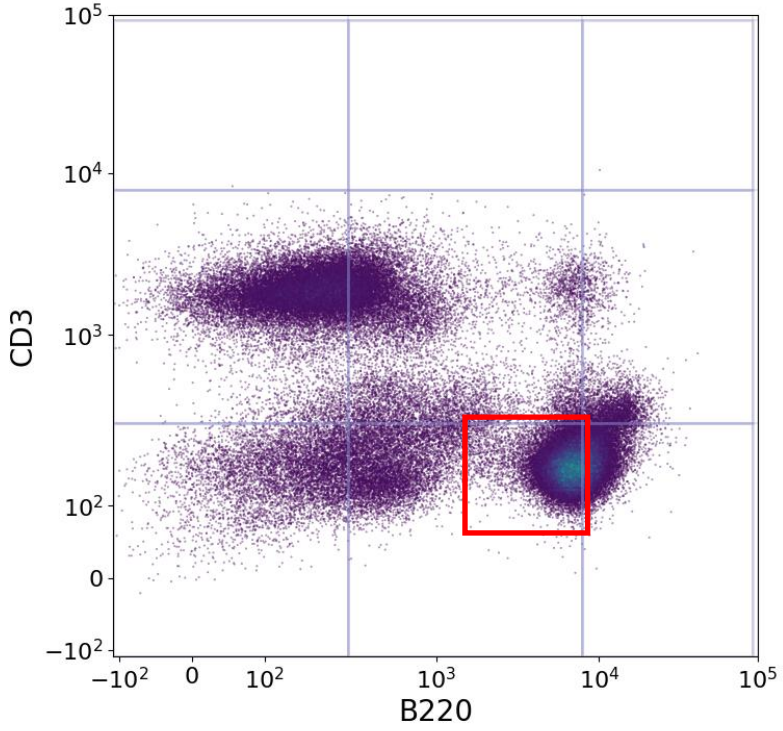
- Experiments were run using 3 flow panels.
- The panels used the same antibody-marker pairs, but the concentration of the marker dye was varied between panels.
- Normally, we don't have ground truth – so this 'synthetic batch effect' dataset is a useful test case for evaluating batch normalization algorithms.

Batch alignment results

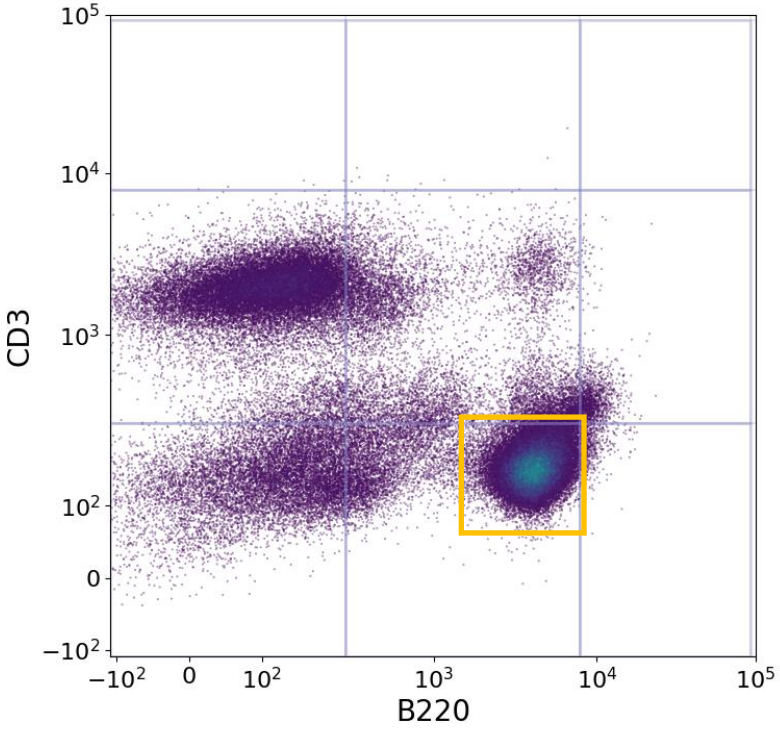
Panel 1 sample



Panel 3 sample



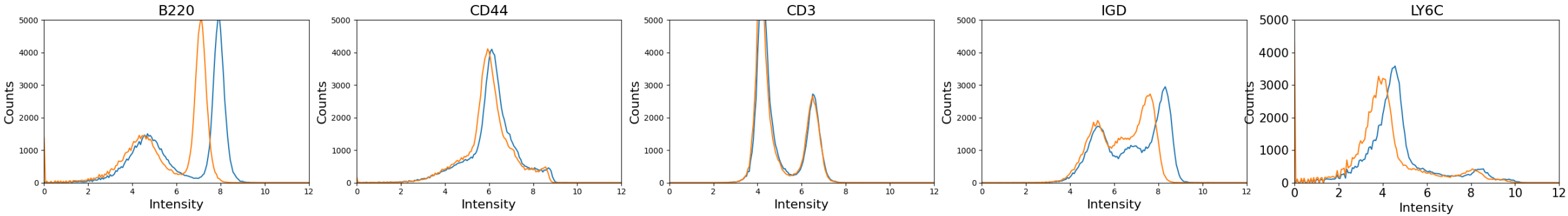
Panel 3 sample – batch normalised



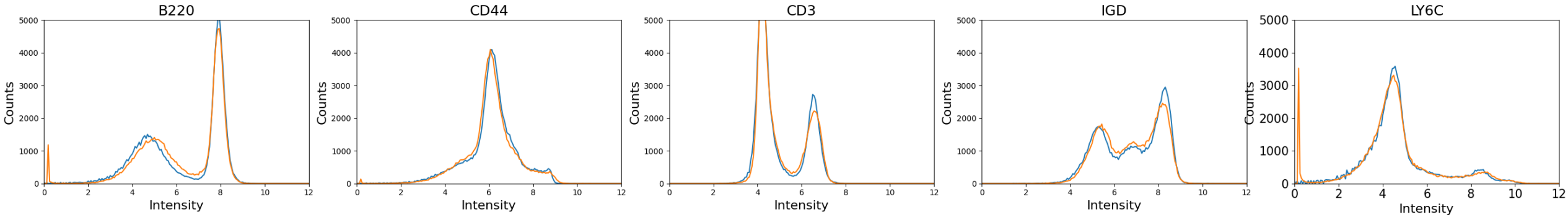
- FlowCoder outputs a batch-corrected FCS data file – compatible with traditional analysis workflows.

Batch alignment results

Before alignment (blue=Panel 1, orange = Panel3)



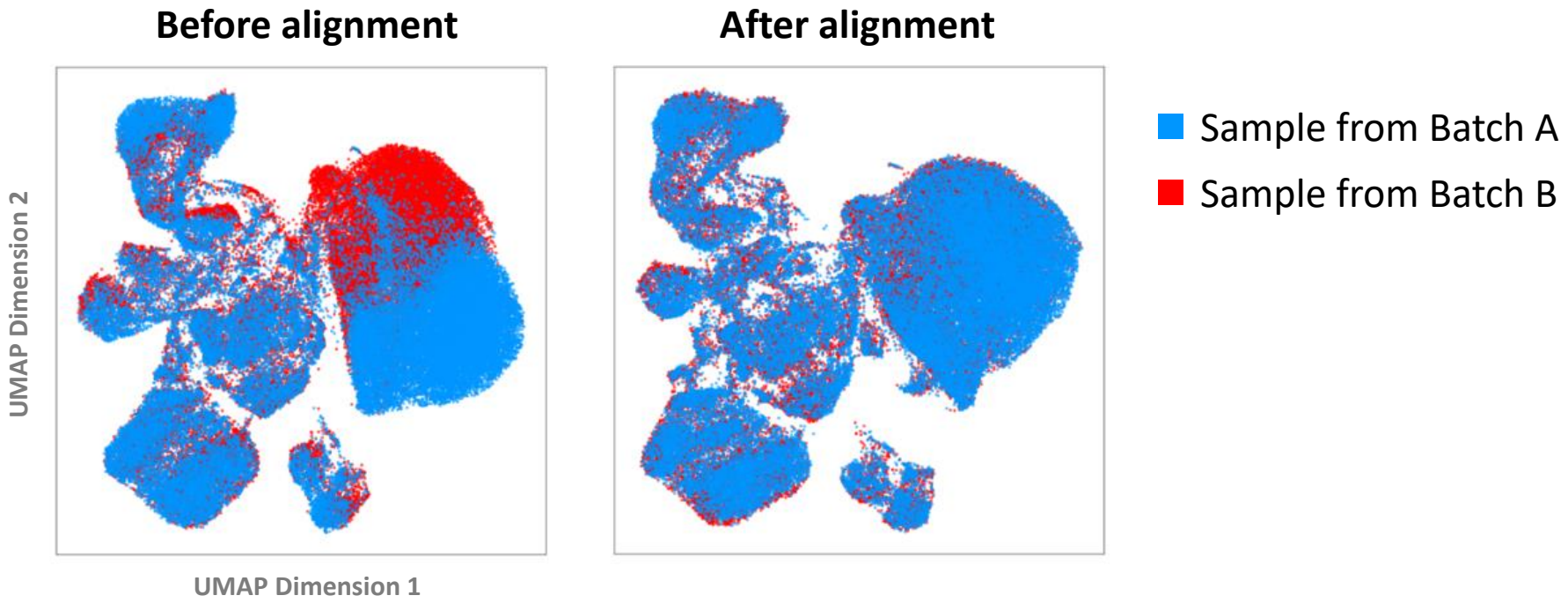
After alignment (blue=Panel 1, orange = Panel3)



- We can use a per-channel histogram to estimate how well the alignment has worked.

High-dimensional alignment metrics

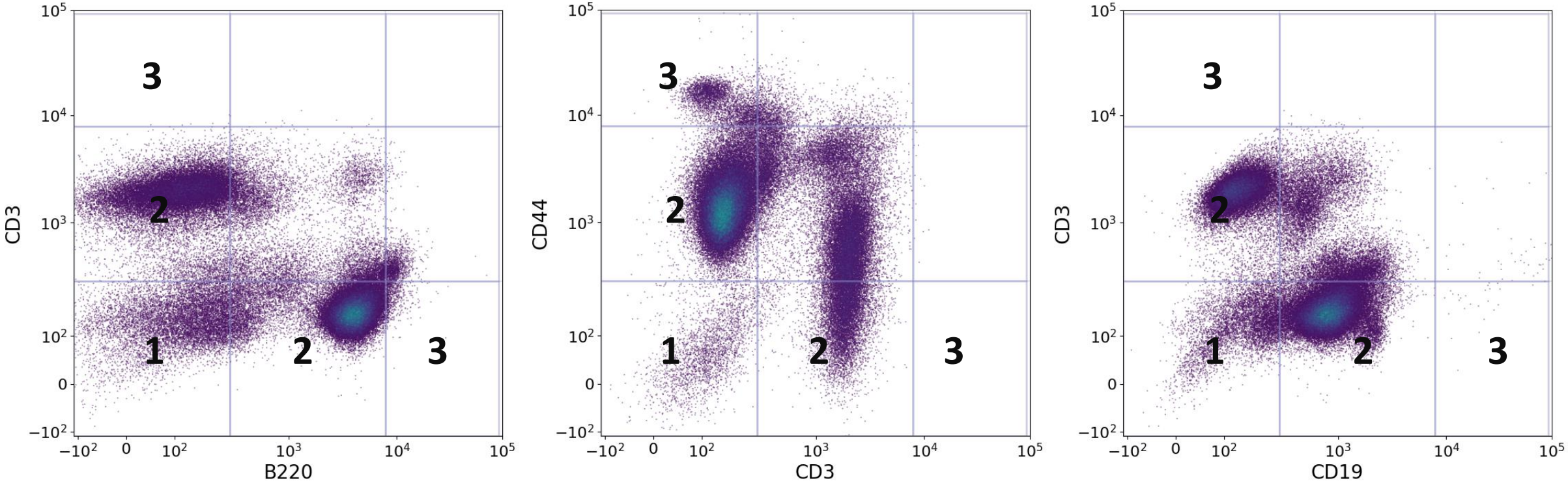
- A UMAP plot projects high dimensional data onto a low dimensional representation
- We compare UMAP plots of the same sample measured across different batches



- UMAP is a useful visualisation, but we also need a way to quantify batch alignment. We need to ensure this metric considers correlations across channel dimensions

High-dimensional alignment metric

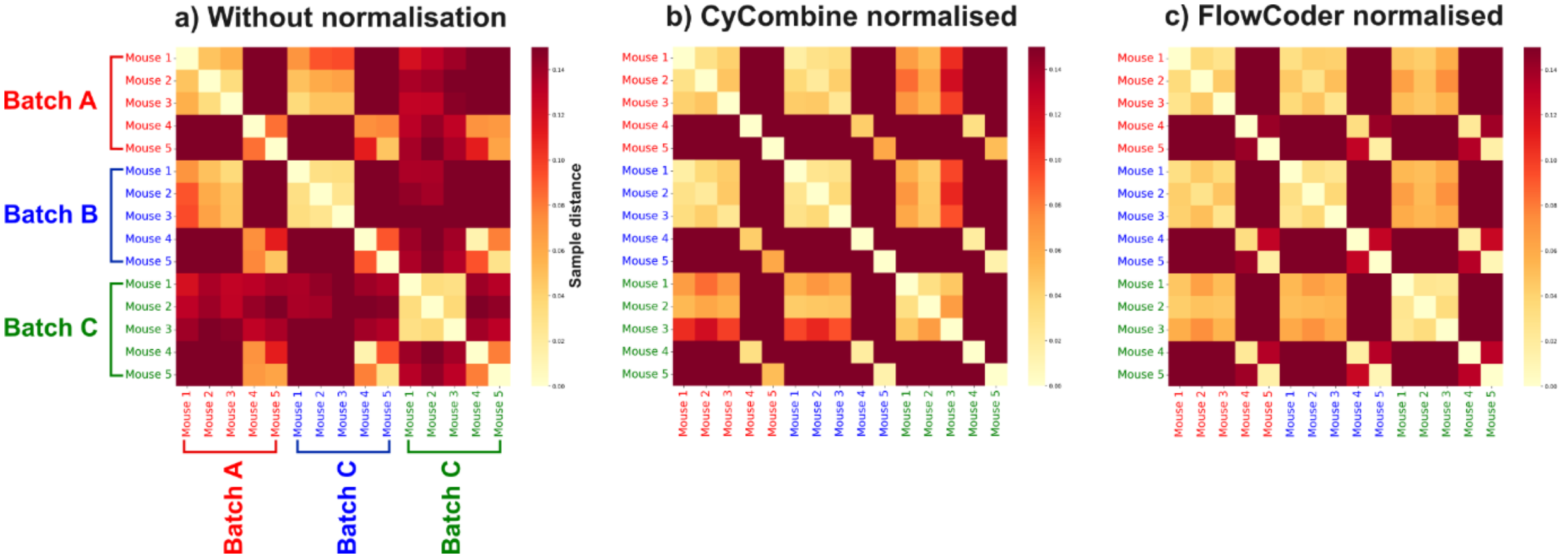
- We begin by filling the entire marker space with uniform voxels (i.e. n-dimensional gates).



- We generate a matrix of size 3^n where n = number of flow channels.
- We then count the number of cells inside each ‘voxel’ within this high-dimensional space – phenotype 'signature’.

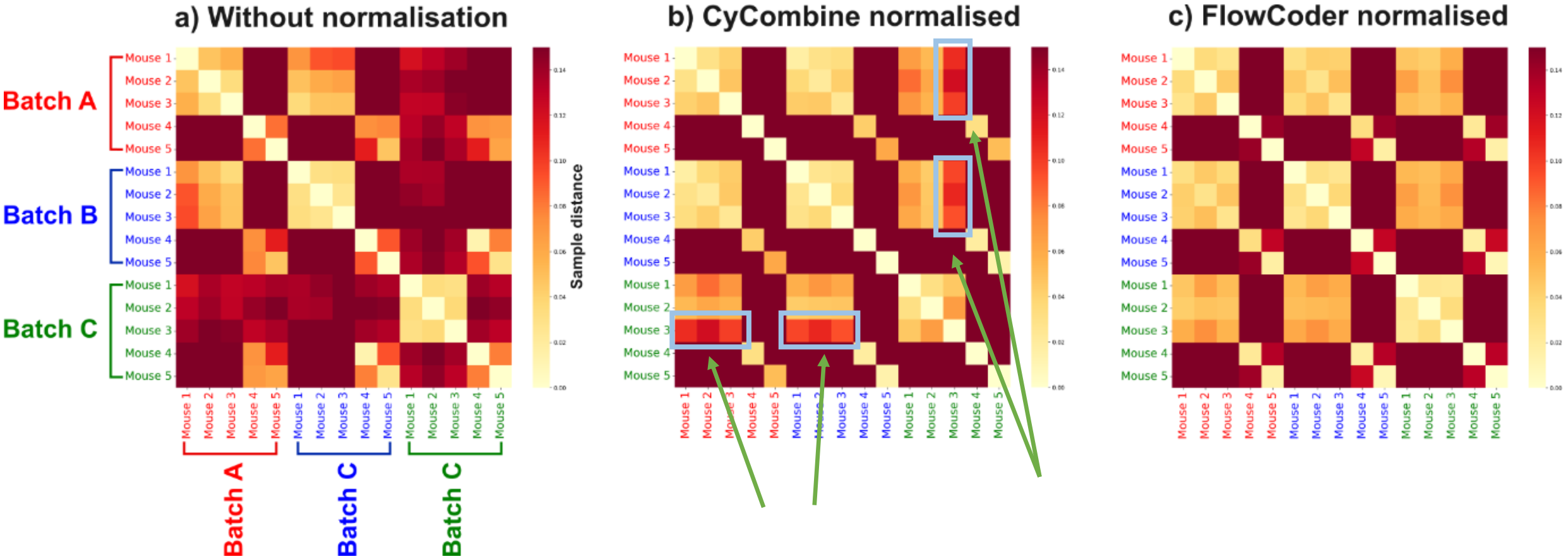
High-dimensional alignment metrics

- We generate a distance heatmap, by looking at absolute difference between all our samples.



- The distance between the same mouse sample (across batches) is reduced in b) and c).

Metrics for batch alignment in high-dimensions



- In this dataset, our ‘FlowCoder’ model is more effective at removing batch effects than other methods.

Using batch-alignment on real-world clinical datasets

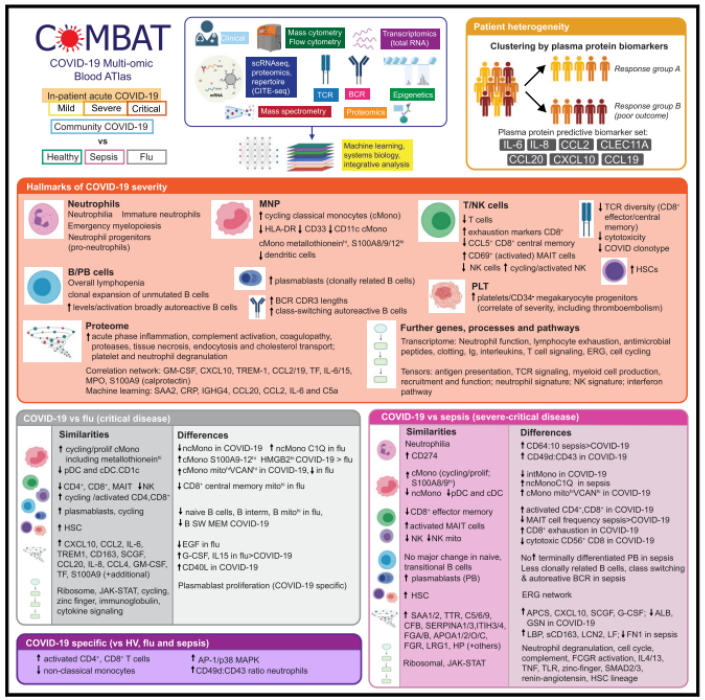
- We are currently evaluating our ML alignment methods on public datasets.

Cell

Resource

A blood atlas of COVID-19 defines hallmarks of disease severity and specificity

Graphical abstract



Authors

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium

Correspondence

julian.knight@well.ox.ac.uk

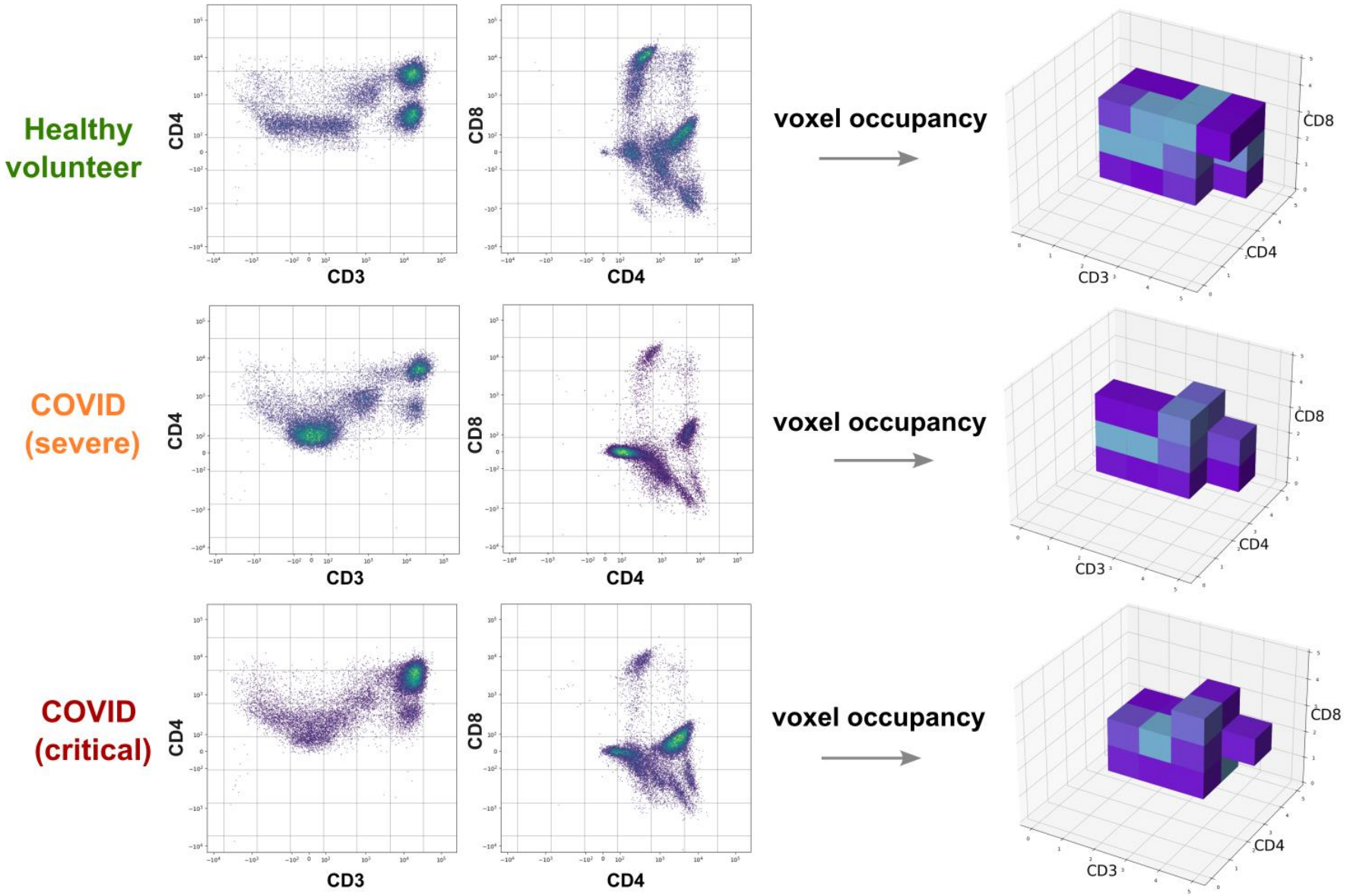
In brief

A multi-omic analysis of patient blood samples reveals both similarities and specific features of COVID-19 when compared with samples obtained from sepsis or influenza patients, which could yield better targeted therapies for severe COVID-19.

Classification	Sample count
COVID (severe)	40
Sepsis	22
COVID (mild)	18
COVID (critical)	18
COVID (health care worker)	13
Flu	11
Healthy volunteer	10
LDN treated	2
Total	**134**

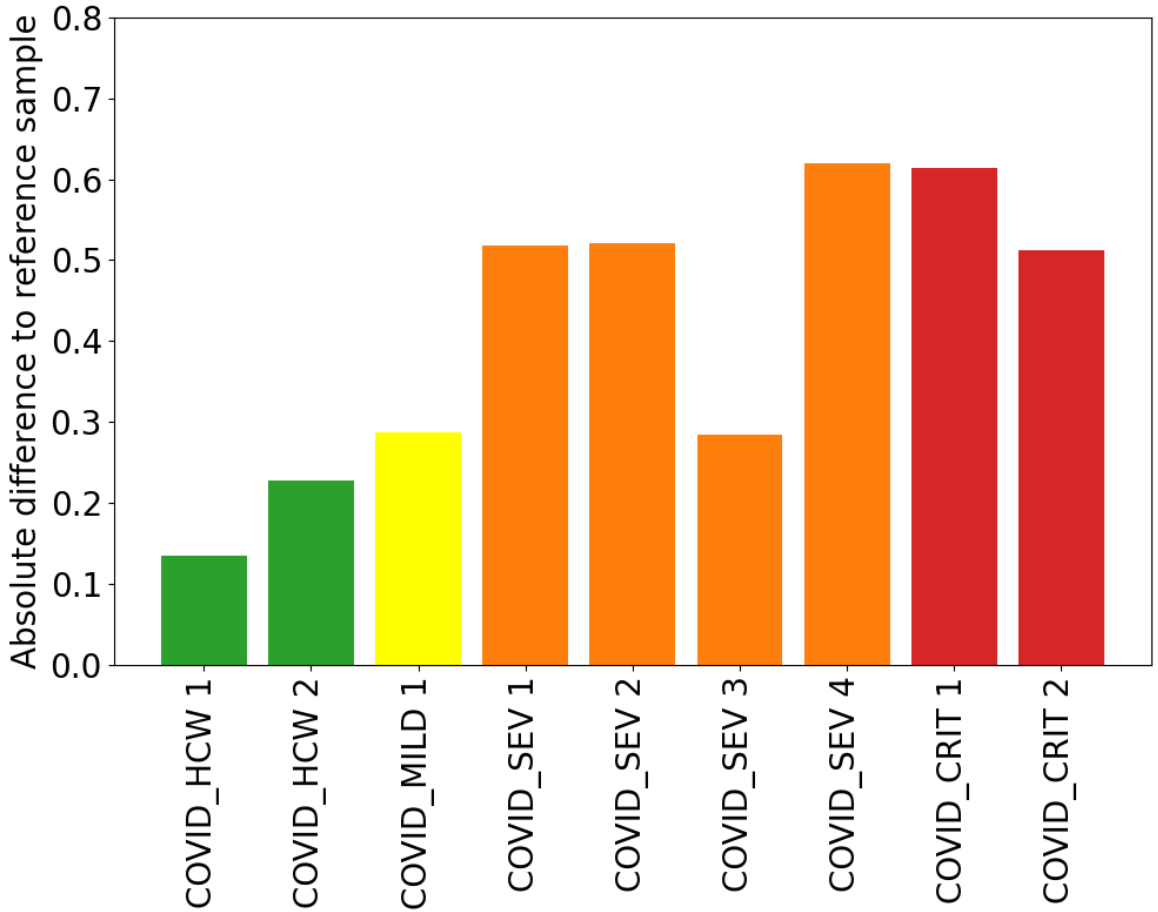
- Here, we compare the flow phenotypes of patients from different COVID infection categories.

High-dimensional alignment metrics



- After alignment, we project each sample into 16-dimensional space and generate a list of cell voxel occupancies.
- Here we illustrate the concept, (showing only 3 of those dimensions).

Stratification of patients by infection severity



- We use a flow sample from a healthy volunteer as our reference.
- The Euclidian distance between 2 samples (e.g. target - healthy) gives us an informative metric.
- Using this method, we see stratification of patients - aligns with observed infection severity.

Summary

- Batch-alignment models offer new opportunities for comparing data across experiments. e.g. More accurate comparison of data from different labs participating in a large study?

- Aligned flow data enables automated workflows, such as fixed gate positions - saves time and also removes a potential source of bias.

- The known strengths of Deep Learning models (i.e. able to handle large dimensionality and their robustness to noise) bring new capabilities to flow cytometry analysis.

- We aim to develop an automated pipeline for identifying subtle phenotypes in high dimensional flow data.

Thank you for listening

- Thanks to Assoc. Prof. Dan Andrews and all the members of the Andrews group.
- Thanks to Dillon Hammill for supplying synthetic batch effect dataset.
- Thanks to ANU and JCSMR.