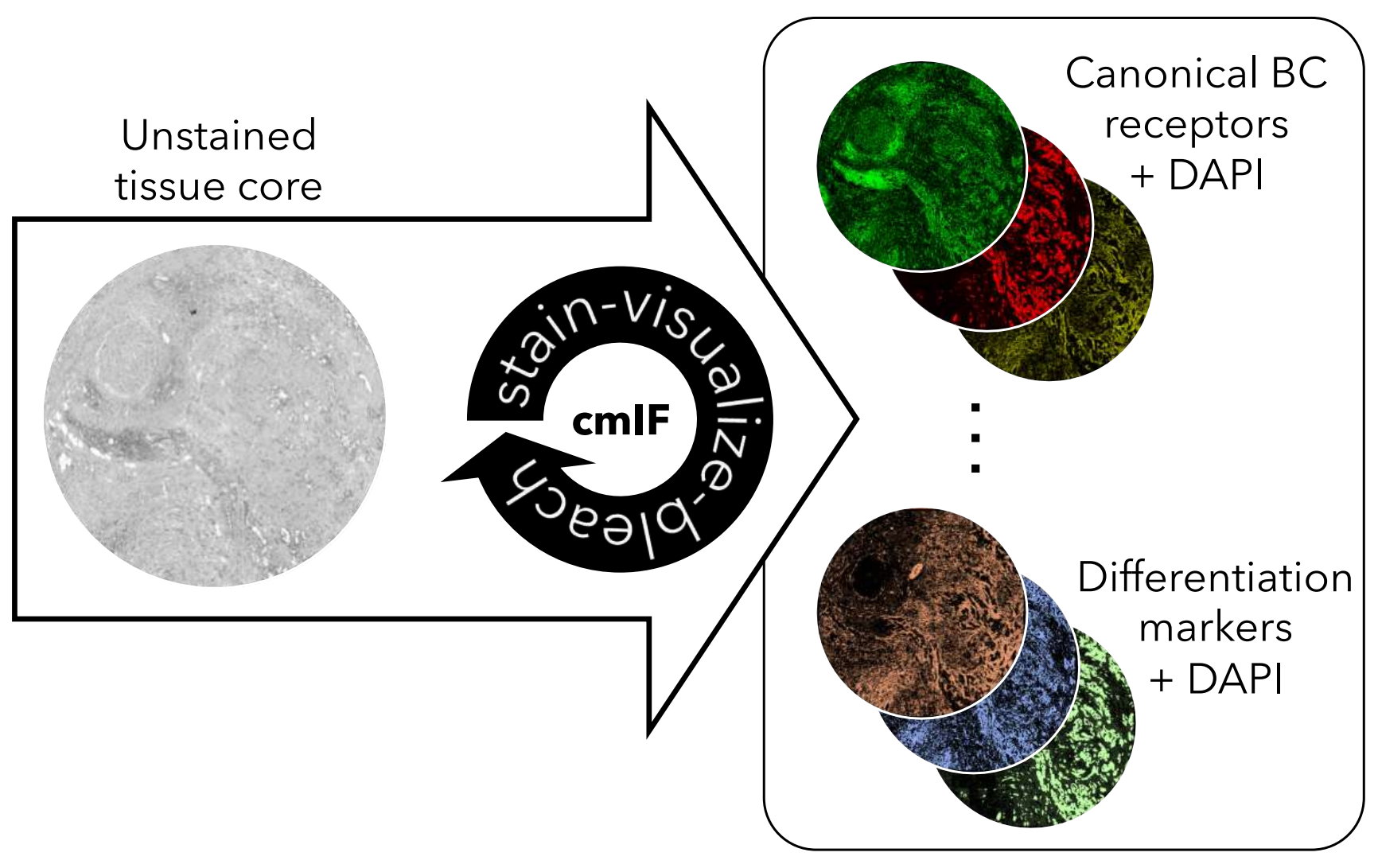
Balanced learning of cell state representations

Erik Burlingame,[*,†] Jennifer Eng,[†] Guillaume Thibault,[†] Geoffrey Schau,[*,†] Koei Chin,[†] Joe W. Gray,[†,‡] Young Hwan Chang[*,†]

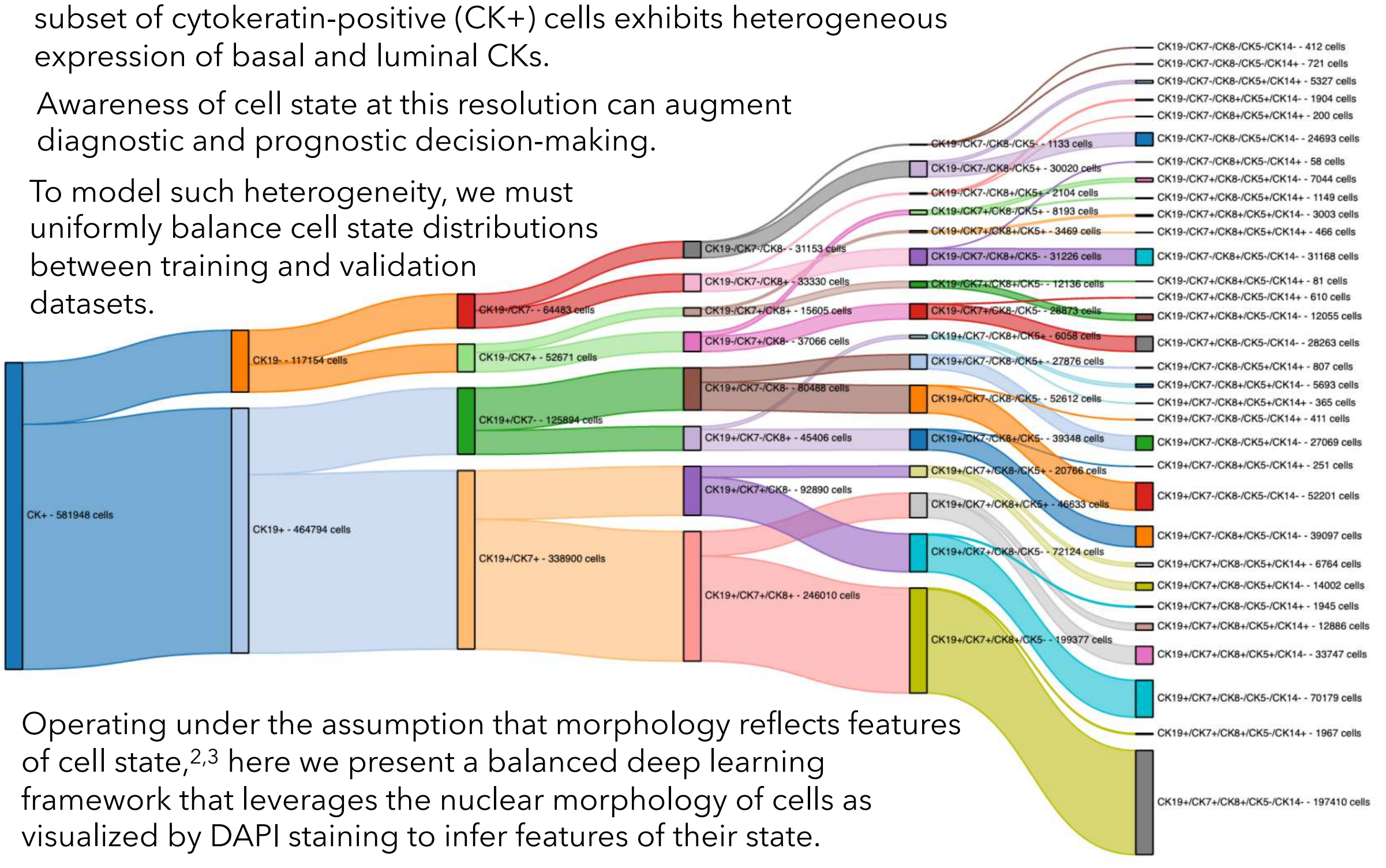
[*]Computational Biology Program, Department of Biomedical Engineering, Oregon Health & Science University
[†]Oregon Center for Spatial Systems Biomedicine, Department of Biomedical Engineering, Oregon Health & Science University
[‡]Knight Cancer Institute, Oregon Health & Science University

Cyclic multiplexed immunofluorescence (cmIF) enables deep cell state characterization of breast cancer tissue microarrays (TMAs)

- Cell state characterization is essential to patient diagnosis and treatment and can be defined by a cell's morphology or the markers it expresses.
- High-dimensional imaging methods like cmIF[1] enable unprecedented *in situ* cell state characterization through iterative labeling of tens of markers within the same tissue.

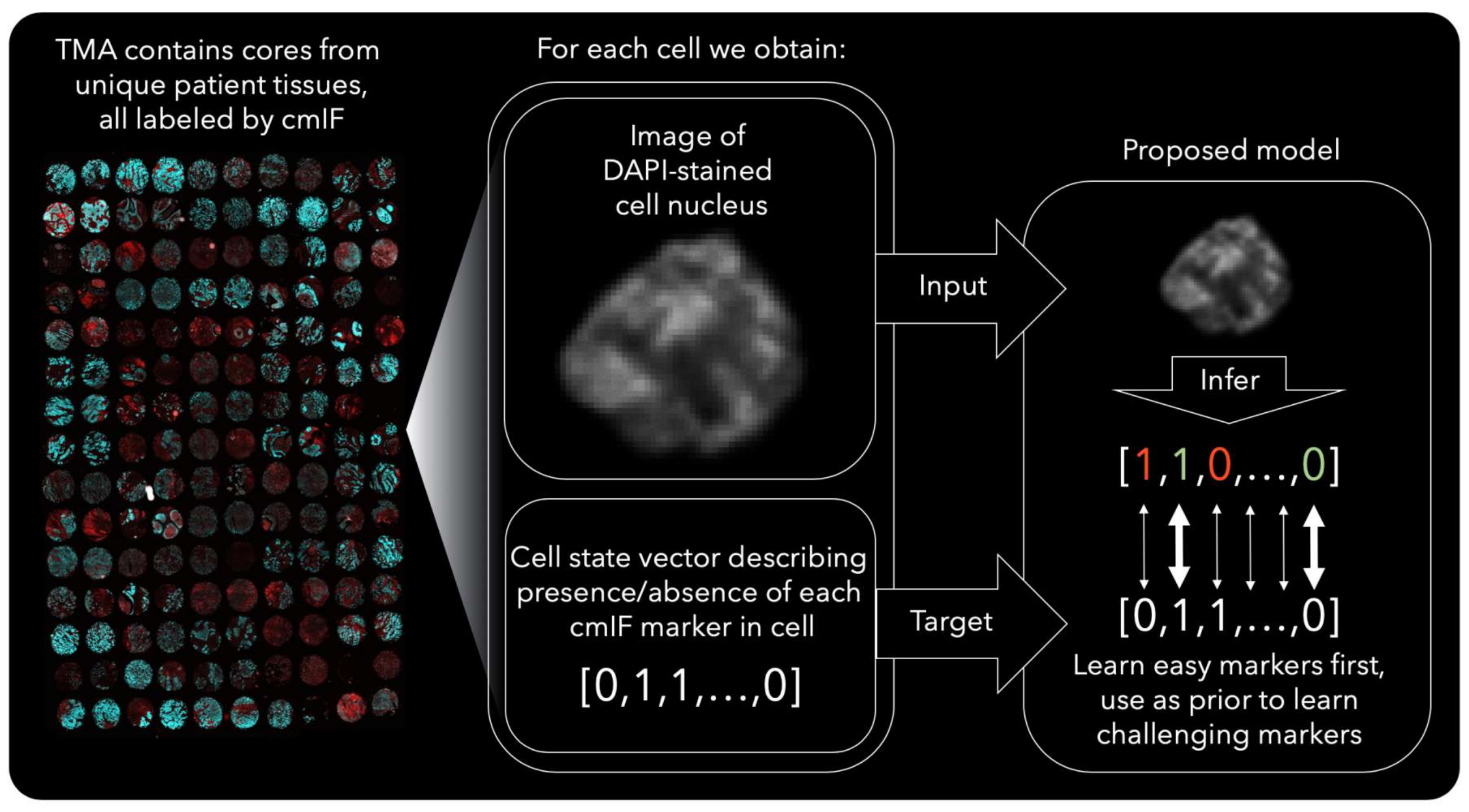


- For example, when applied to breast cancer TMAs, cmIF reveals that the subset of cytokeratin-positive (CK+) cells exhibits heterogeneous expression of basal and luminal CKs.
- Awareness of cell state at this resolution can augment diagnostic and prognostic decision-making.
- To model such heterogeneity, we must uniformly balance cell state distributions between training and validation datasets.



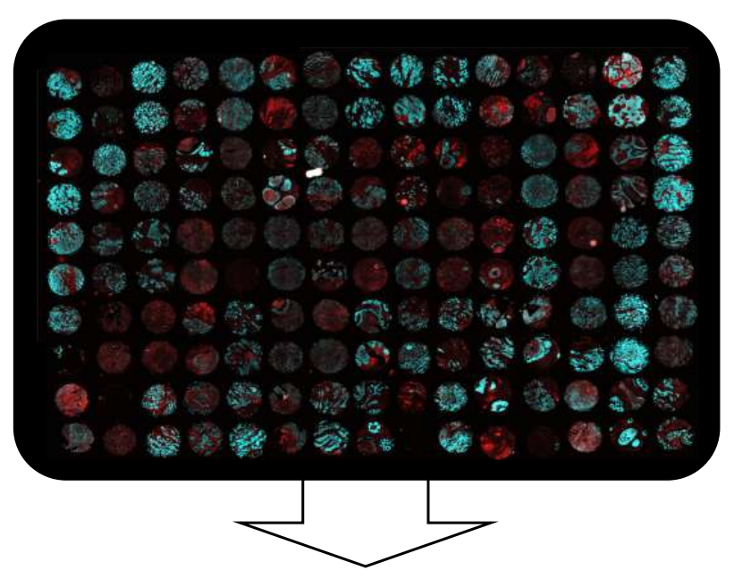
- Operating under the assumption that morphology reflects features of cell state,[2,3] here we present a balanced deep learning framework that leverages the nuclear morphology of cells as visualized by DAPI staining to infer features of their state.

- cmIF lends itself to a multi-label learning paradigm, but training/validation stratification is not trivial.

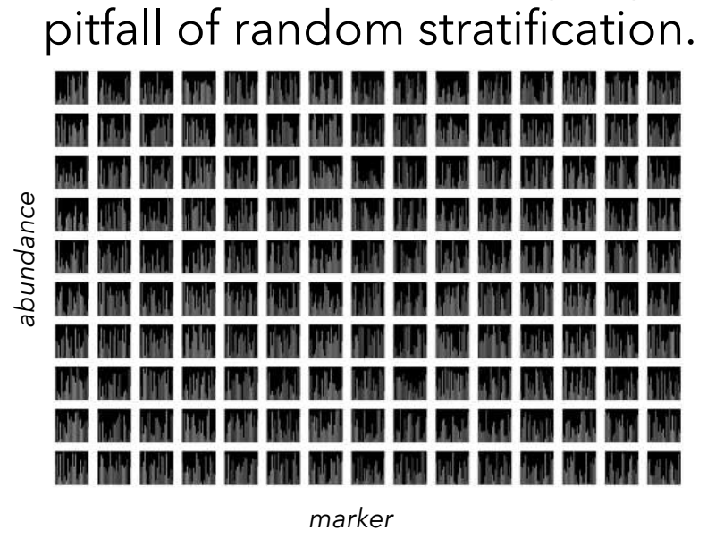

TMA contains cores from unique patient tissues, all labeled by cmIF
For each cell we obtain:
Image of DAPI-stained cell nucleus
Cell state vector describing presence/absence of each cmIF marker in cell
[0,1,1,…,0]
Input
Target
Proposed model
Infer
[1,1,0,…,0]
[0,1,1,…,0]
Learn easy markers first, use as prior to learn challenging markers

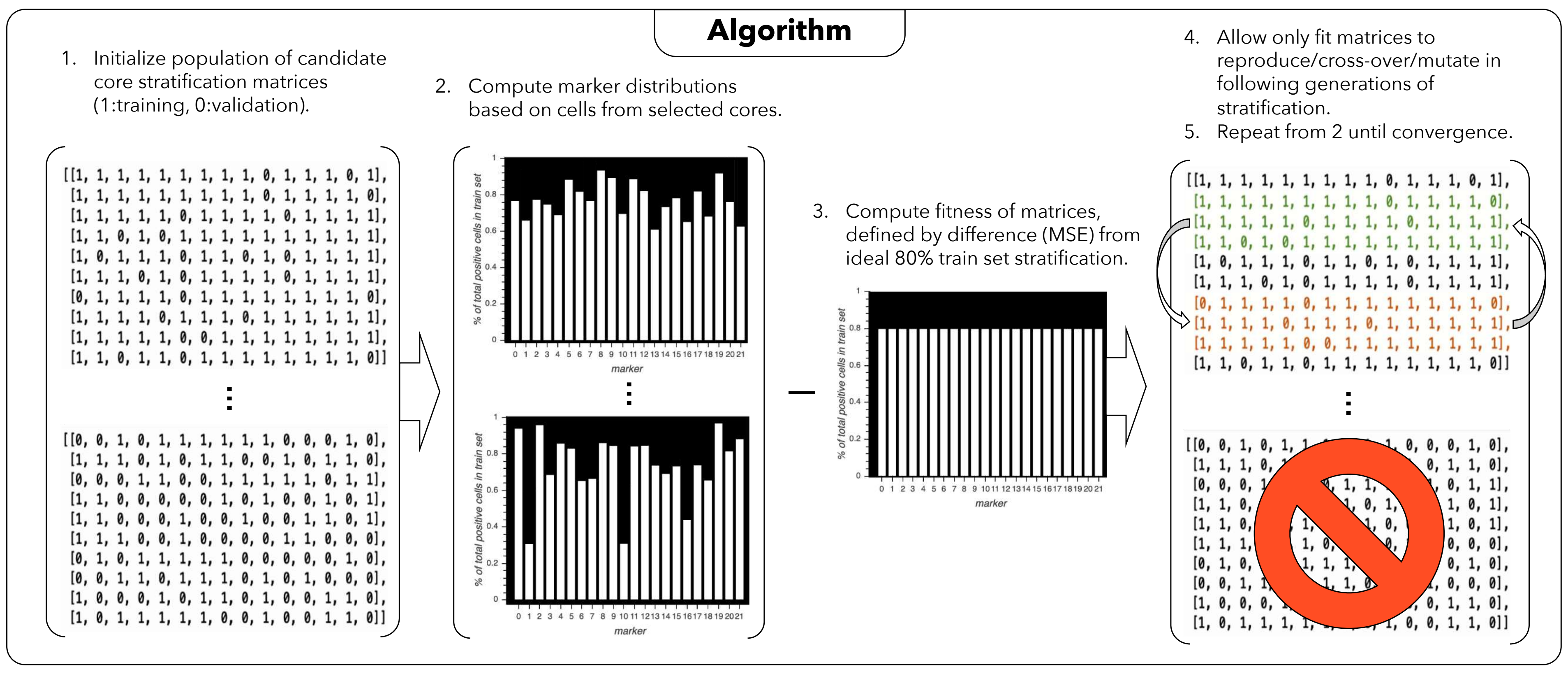
A simple genetic algorithm ensures balanced training/validation stratification of TMA cores for cmIF representation learning

- Cell populations vary widely between TMA cores, necessitating principled stratification of cores.



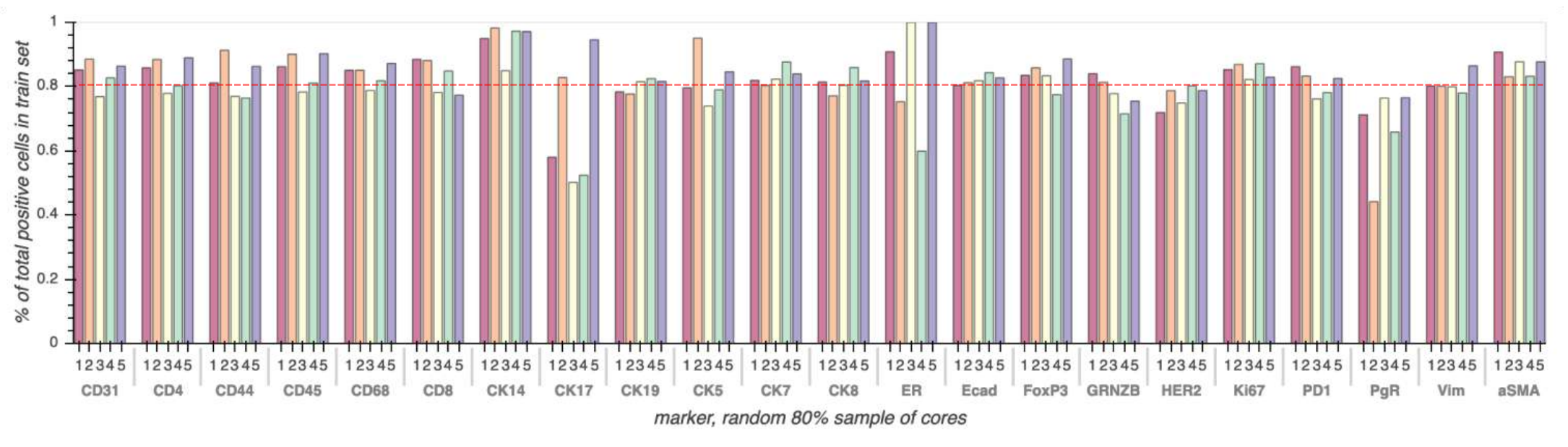
- Marker distributions highlight pitfall of random stratification.



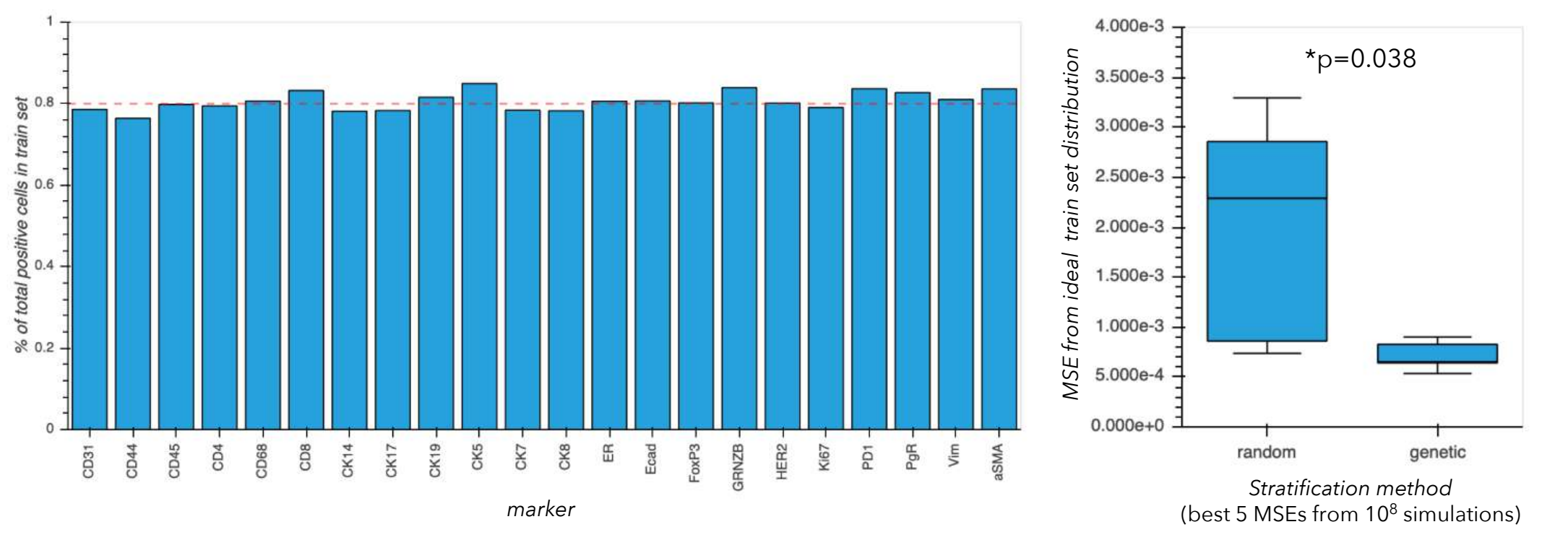


1. Initialize population of candidate core stratification matrices (1:training, 0:validation).
2. Compute marker distributions based on cells from selected cores.
3. Compute fitness of matrices, defined by difference (MSE) from ideal 80% train set stratification.
4. Allow only fit matrices to reproduce/cross-over/mutate in following generations of stratification.
5. Repeat from 2 until convergence.

- Random core stratification is prone to over- and under-sampling of markers, as highlighted by these 5 simulations (red line is ideal train set stratification).

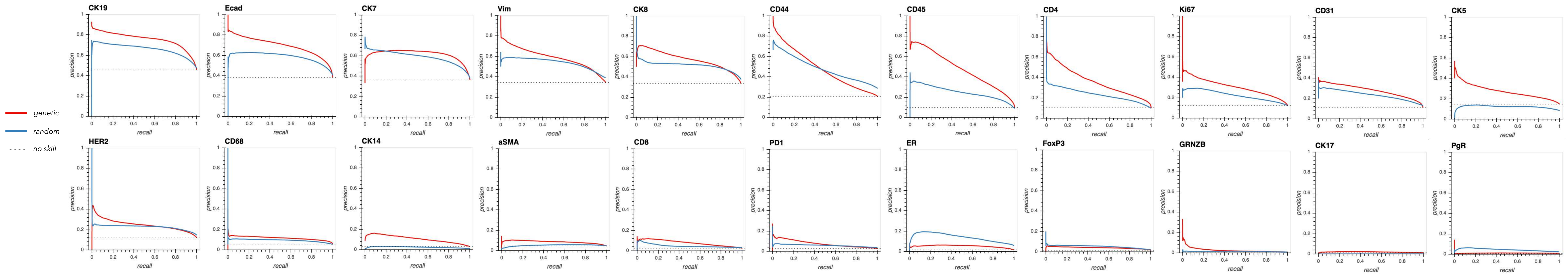


- Genetic core stratification balances training/validation sets across all markers.

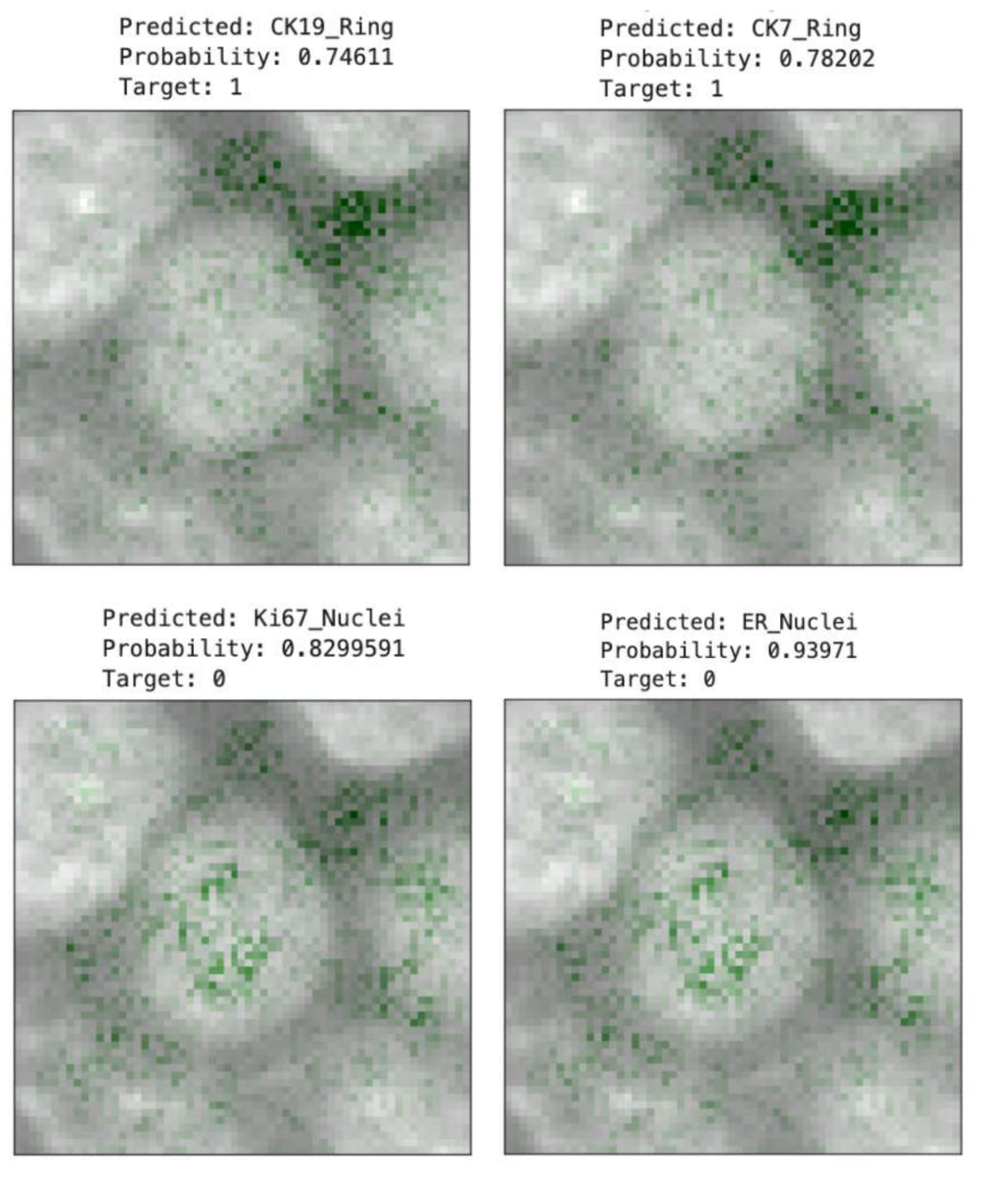


Genetic stratification of TMA cores into training/validation sets yields a more generalizable cell state inference model

- Following either *genetic* or *random* data stratification, Resnet18 models[4] are trained to infer a 22-marker target vector given an input image of a DAPI-stained nucleus; the model trained using *genetic* stratification generalizes better on 19 of 22 markers.



Model attention reveals salient features



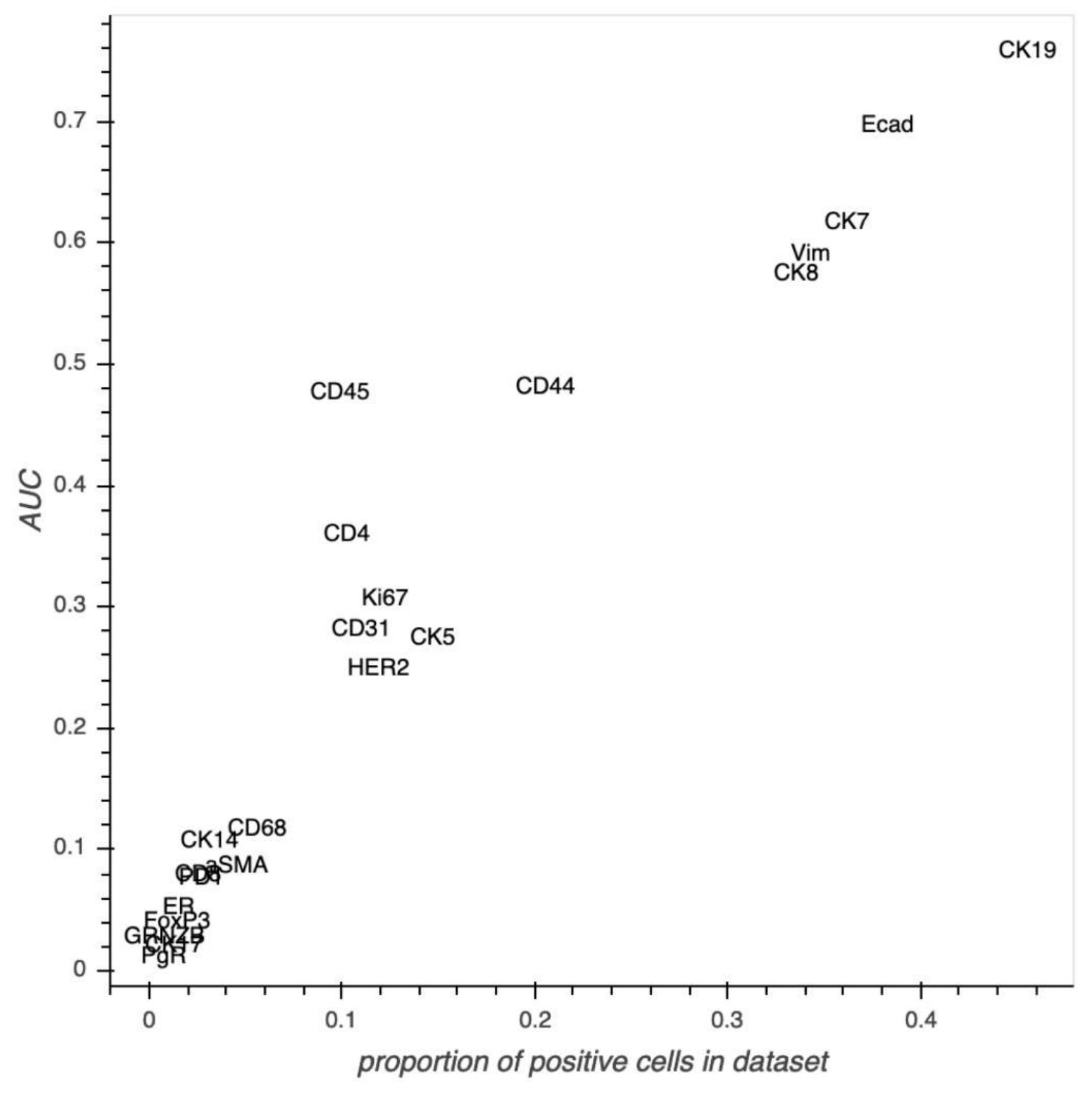
```
Predicted: CK19_Ring
Probability: 0.74611
Target: 1
```

```
Predicted: CK7_Ring
Probability: 0.78202
Target: 1
```

```
Predicted: Ki67_Nuclei
Probability: 0.8299591
Target: 0
```

```
Predicted: ER_Nuclei
Probability: 0.93971
Target: 0
```

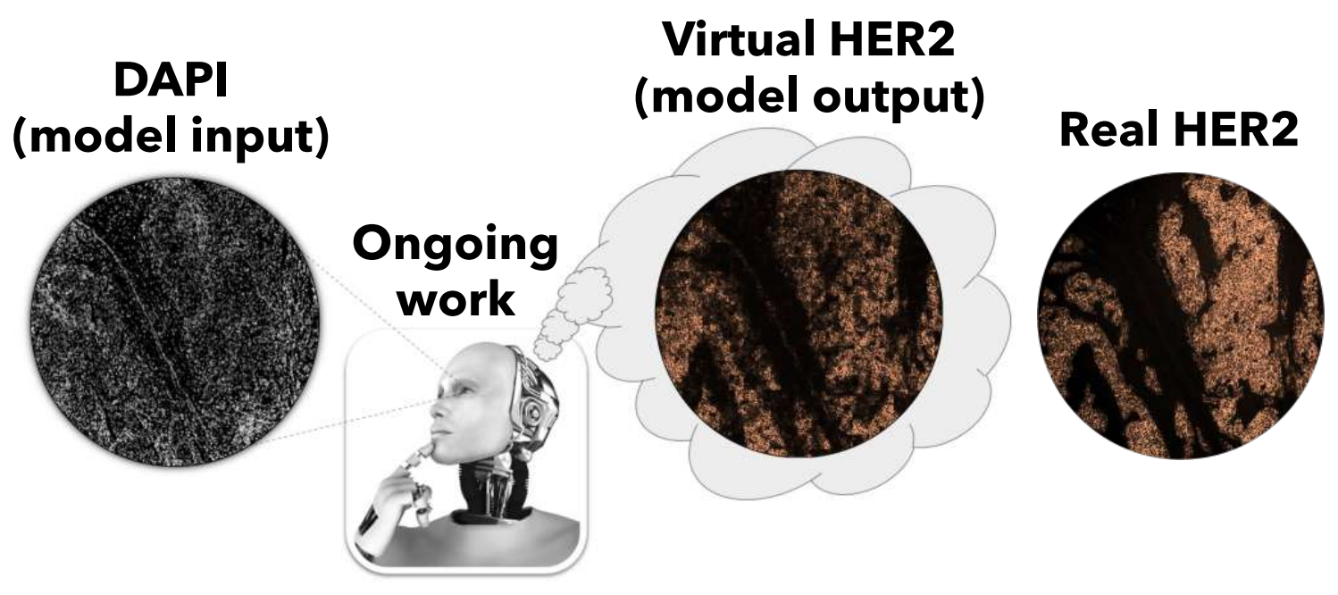
Model performance is correlated with marker prevalence in dataset



- The model performs best on the most prevalent markers.
- Pre-conditioned models trained on independent cell subtypes—e.g. immune, cancer, stromal—may yield improvements, especially for cell subtypes with exclusive and rare markers, e.g. FoxP3+ or GRNZB+ immune cells.

Conclusions and ongoing work

- Here we present a proof-of-concept framework optimized for learning generalizable representations of cell state and which objectively measures the information content of nuclear morphology as visualized by DAPI.



- Learned cell state representations can facilitate virtual staining of human biopsy tissues based on hematoxylin and eosin[2,3] and DAPI stains alone.
- A model which infers cell state using low-cost and widely available reagents like DAPI—even if only a limited number of cell state features—could bring the benefits of cmIF to more patients and in a clinically relevant timeframe.

Acknowledgements: We gratefully acknowledge the OHSU Histopathology Shared Resource, the OHSU Advanced Multiscale Microscopy Shared Resource, and the OHSU Advanced Computing Center.
Funding: This work was supported by the National Cancer Institute (U54CA209988, U2CCA233280), the Human Tumor Atlas Network (1U2CCA233280-01), the OHSU Center for Spatial Systems Biomedicine, the Prospect Creek Foundation, the OHSU Foundation, the Susan G. Komen Foundation, and the Oregon Clinical & Translational Research Institute. Erik Burlingame is funded in part by a scholarship from Achievement Rewards for College Scientists (ARCS).

[1] Eng *et al*. (2020). Cyclic multiplexed-immunofluorescence, a highly multiplexed method for single-cell analysis. *Methods Mol Biol*. 2055:521-562.
[2] Burlingame *et al*. (2018). SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. *Proc. SPIE 10581, Medical Imaging 2018: Digital Pathology*, 1058105.
[3] Burlingame *et al.* (2019). SHIFT: speedy histological-to-immunofluorescent translation of whole slide images enabled by deep learning. *bioRxiv* 730309. doi: https://doi.org/10.1101/730309
[4] He *et al*. (2015). Deep residual learning for image recognition. *arXiv*:1512.03385v1