Deep Learning Interpretability for the Discovery of Biomedical Patterns

Mara Graziani, mgr@zurich.ibm.com, @mormontr

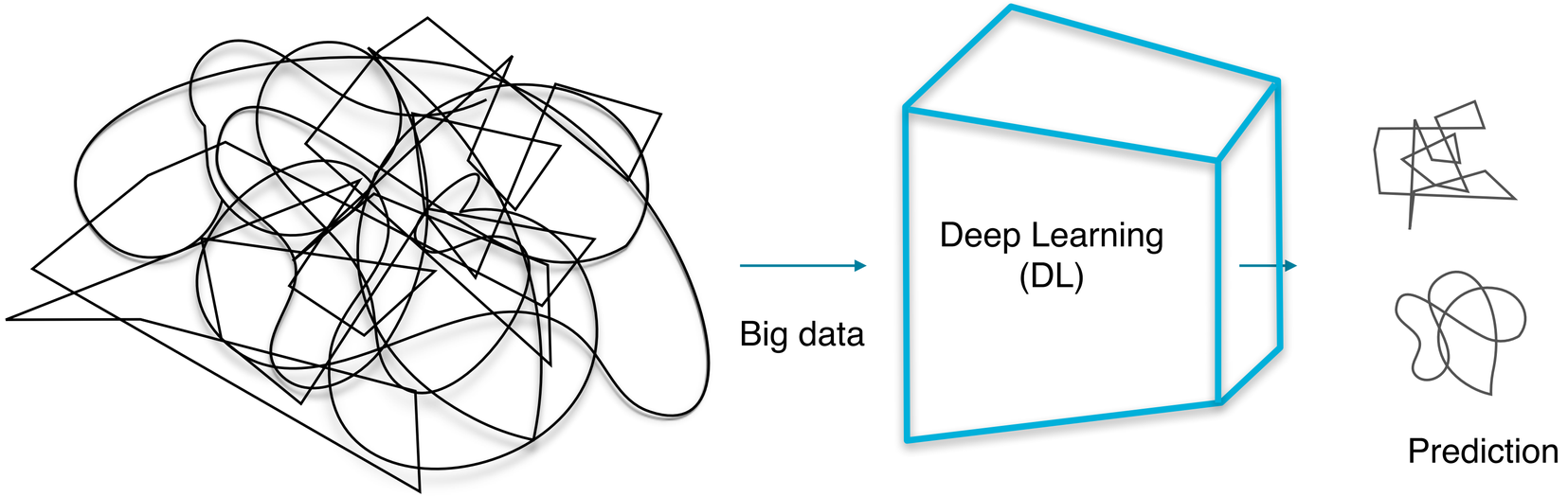
postdoc researcher at IBM Research Zurich / ZHAW / HES-SO Valais

Funding: SNF Sinergia for Characterising Colorectal Cancer Molecular Subtypes and EU H2020 AI4Media

Ongoing work in collaboration and with the valuable help of Niccolo' Marini, Nicolas Deutschmann, Nikita Janakarajan, Henning Müller and Maria Rodriguez Martinez

And also thanks to the important feedback given by the Institute of Pathology at the Univ. of Bern: Inti Zloebec and Huu-Giao Nguyen

Deep Learning Interpretability: What have we learned?



- "Models are approximations, never exactly true" [Box, 1997]
- The sole optimization performance is not sufficient to ensure reliability [Doshi-Velez and Kim, 2017]
- Performance drops, little robustness, hidden biases [Arvidsson et al., 2015; Nguyen et al., 2015; Zou et al., 2018]

Deep Learning Interpretability: What have we learned?

Feature Attribution (i)

Grad-CAM (Selvaraju et al., 2017)

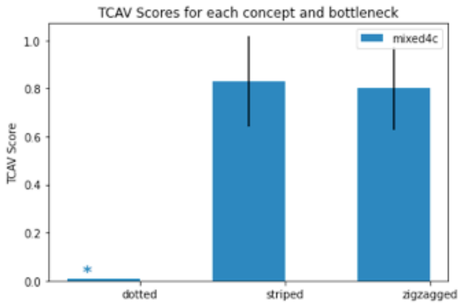
Extremal perturbation masks as in Fong et al., 2019

Feature Visualization (ii)

Lucid toolbox by Olah et al., 2017

Deep Dream (Mordvintsev et al., 2015)

Concept Attribution (iii)

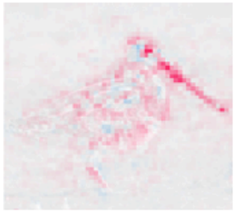


CAV-based explanations (Kim et al., 2018a)

Surrogates (iv)

LIME (Ribeiro et al., 2016)

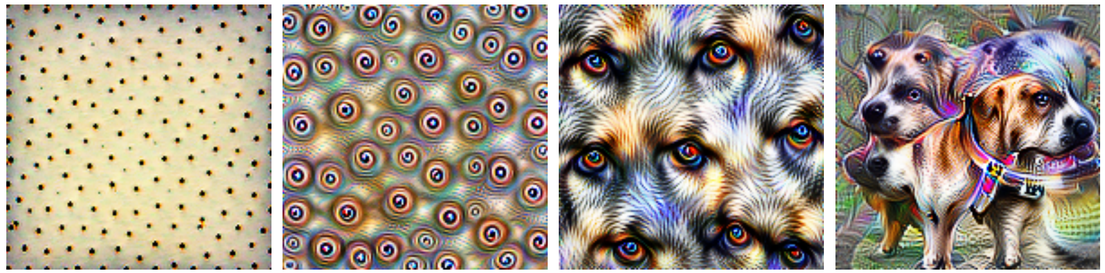
SHAP as in Lundberg et Lee, 2017



- Previous work and current research efforts teach us about
  - Model evaluation, debugging
  - Some data correlations reflect the reality but may be harmful [Lengherich et al., 2022]
  - Interpretability for performance improvement [Graziani et al., 2021]
  - Our general understanding about DL generalisation (and memorisation patterns) [Graziani et al., 2019]

We learned: architectural biases in CNNs

- Low layers extract simple features of color and texture [Olah et al., 2016].
- Complex (high-level) concept representations appear at deep layers [Kim et al., 2018; Graziani et al., 2018]
- ImageNet[7] pre-trained CNNs are biased towards texture [Gheiros et. al., 2018]

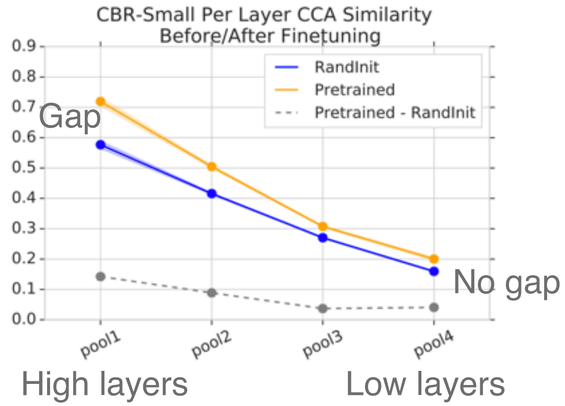


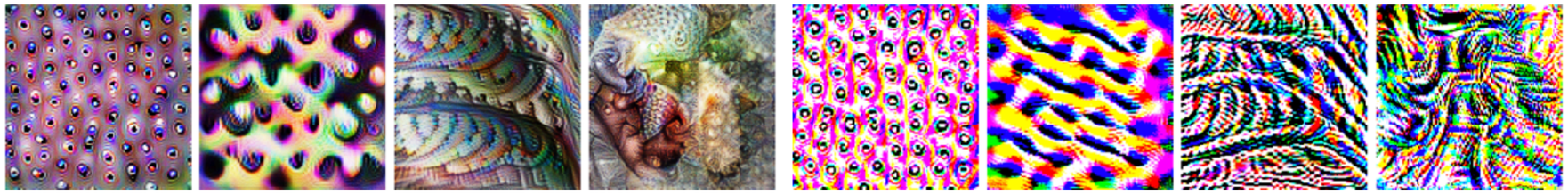
(a) Texture image
81.4% Indian elephant
10.3% indri
8.2% black swan

(b) Content image
71.1% tabby cat
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% Indian elephant
26.4% indri
9.6% black swan

- Gabor-like filters not crucial for medical images [Raghu et al., 2019], feature reuse at low layers [Graziani et al., 2018]

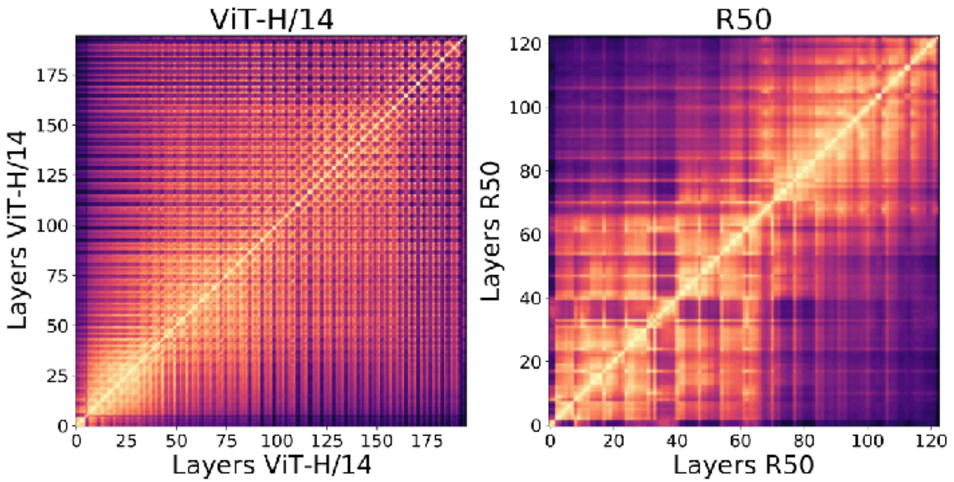




(a) pretrained (b) finetuned

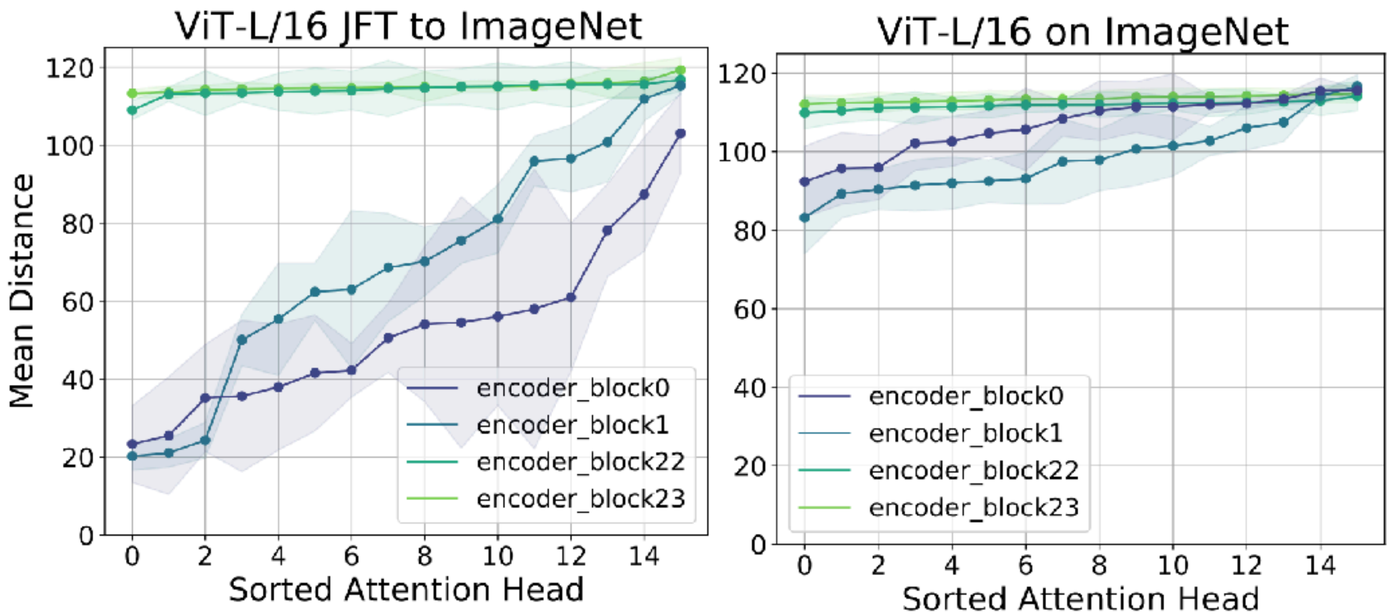
We are now learning about transformers

- Transformer's pay attention globally and locally [Raghu et al., 2022]



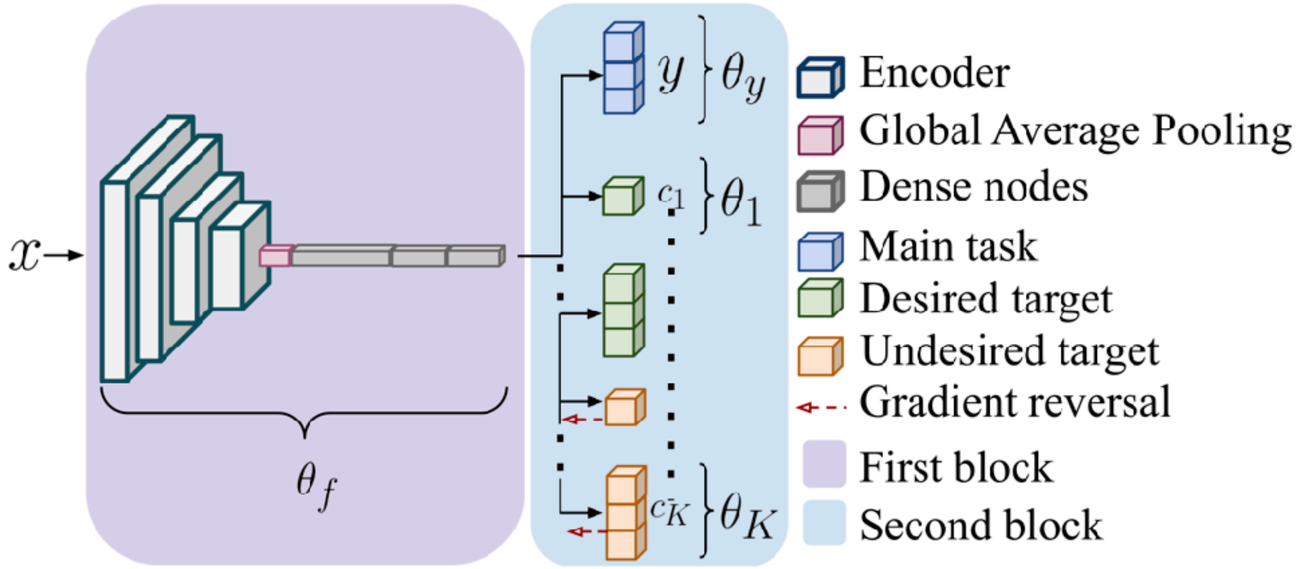
Large dataset

Small dataset



We learned: interpretability and domain-expertise improve our models

- Linear probing representations shows the learning dynamics [Kim et al., 2018; Graziani et al., 2018] and it can be used to improve model optimization [Graziani et al., 2021]
  - Multi-task adversarial architecture to learn desired patterns and forget undesired ones

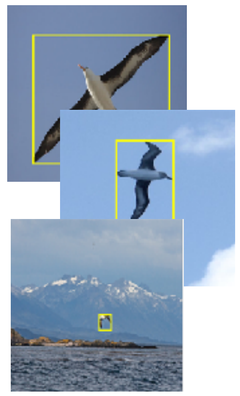


$$E = \underbrace{\lambda_m \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y)}_{\text{main task}} + \underbrace{\sum_{k=1}^{K} \lambda_k \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_k^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k)}_{\text{K extra tasks}}$$

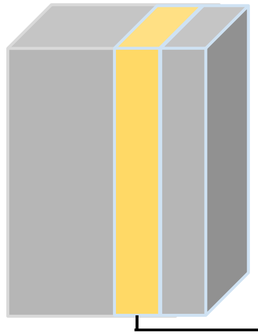
Weighting of losses nontrivial:
Vanilla sum and uncertainty-based approach
[Graziani et al., 2021]

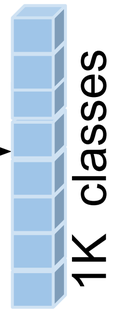
We learned: interpretability and domain-expertise improve our models

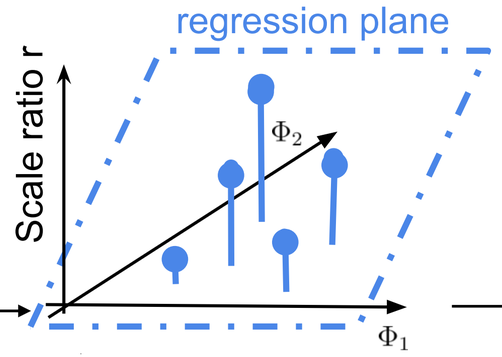
- Another example of “useful” interpretability: adapting transfer learning with domain-knowledge [Graziani et al., 2020]

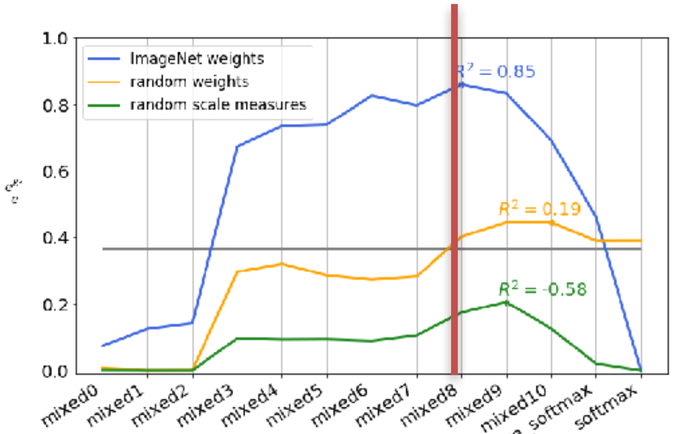


Bounding-boxes to measure scale



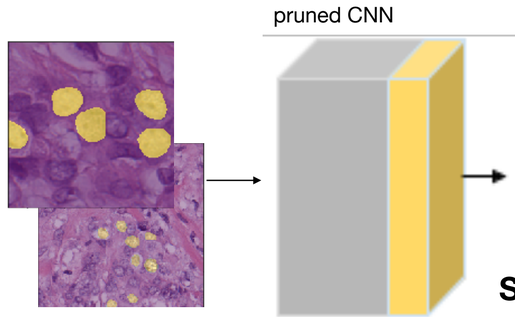






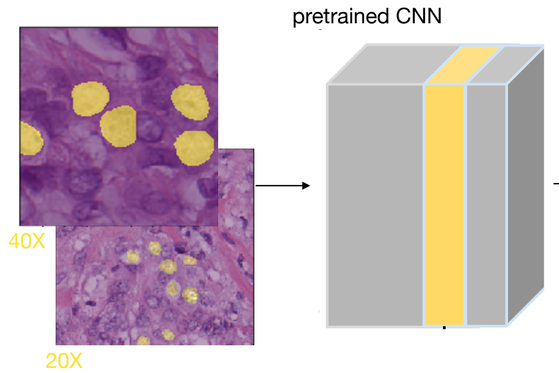
Transfer to histopathology

Pruning strategy that removes the layers where invariance is learned to improve the transfer



MAE = 54.93

Significant improvements in performances after pruning



MAE = 81.85

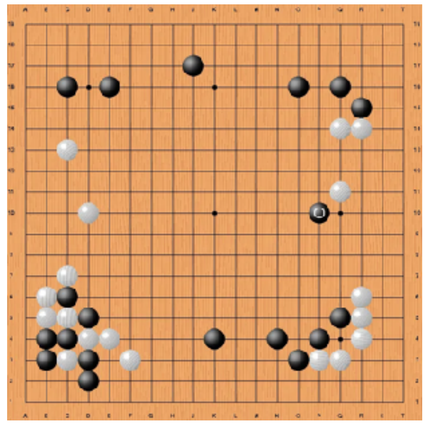
Task: Regression of average nuclei

What we know :

So much we do not know yet!

Deep Learning Interpretability

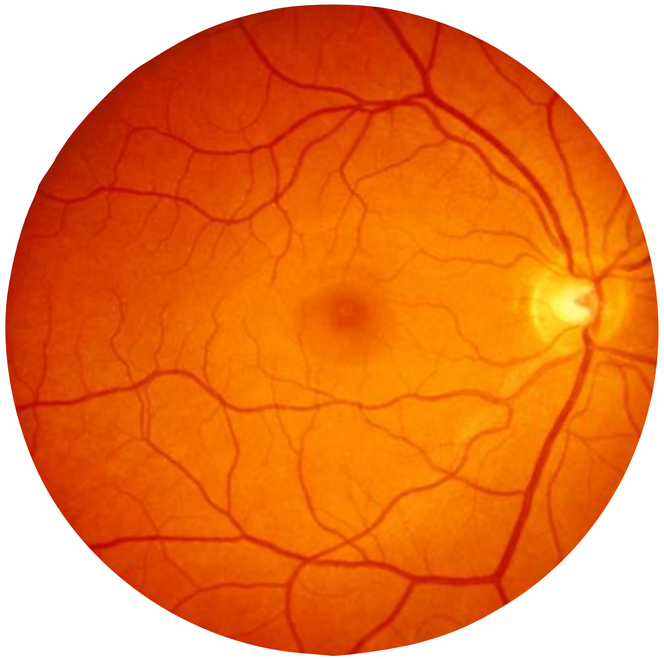
- We do not yet understand the full picture
  - AlphaGo beats world Go champion in 2016 by opting for an *unusual* move [Silver et al., 2016]



- What is in between what DL can achieve and humans cannot?
- Can interpretability can help us find out? YES!

1 Minute Trivia

Extreme trivia

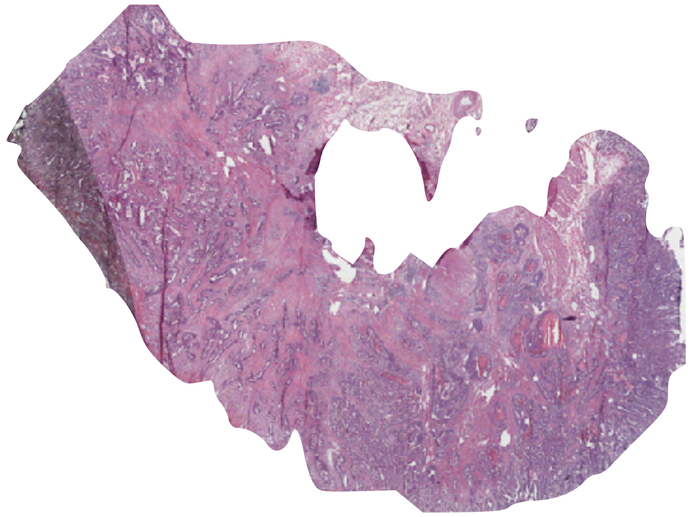


- Could you predict the gender of the patient?
- Its risk of cardiovascular issues?

- CNNs predict the gender [Korot et al., 2021], smoking habits and the risk of cardiovascular diseases [Poplin et al., 2018] from eye fundus imaging

Extreme trivia

- Could you predict the current distribution of pollen in Switzerland?

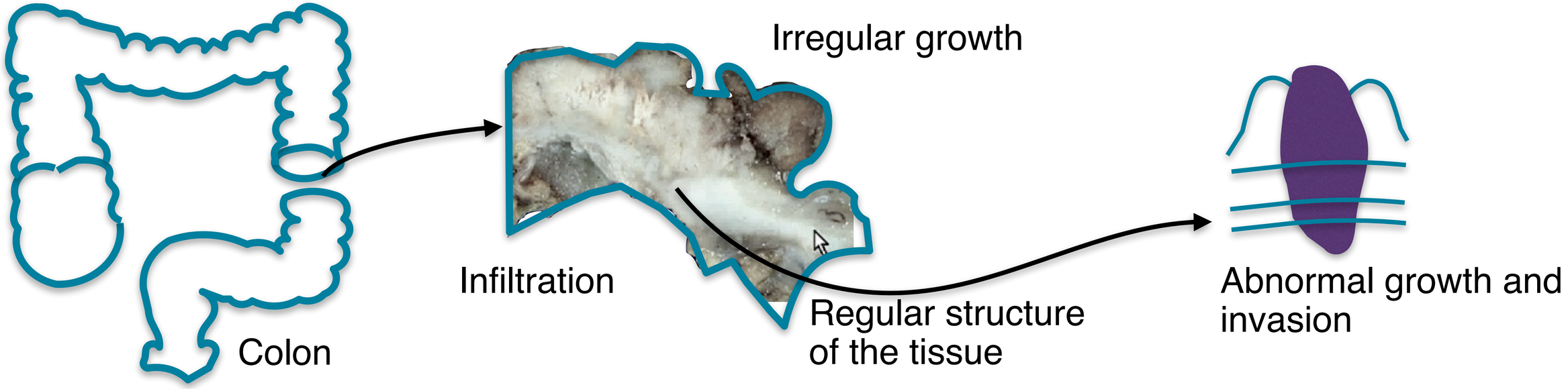


- Could you predict gene mutations and expressions in human tissue?

- These are tasks above the capabilities of many of us…and of many domain-experts too
- Patterns of gene mutation can be predicted from human tissue microscopy [Kather et al., 2020]

Interpretable DL modelling for colorectal cancer

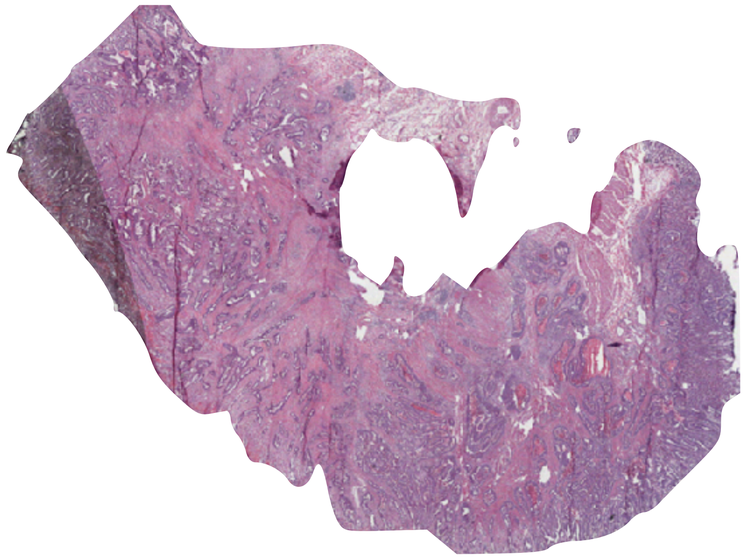
Learning biomedical patterns in colorectal adenocarcinoma

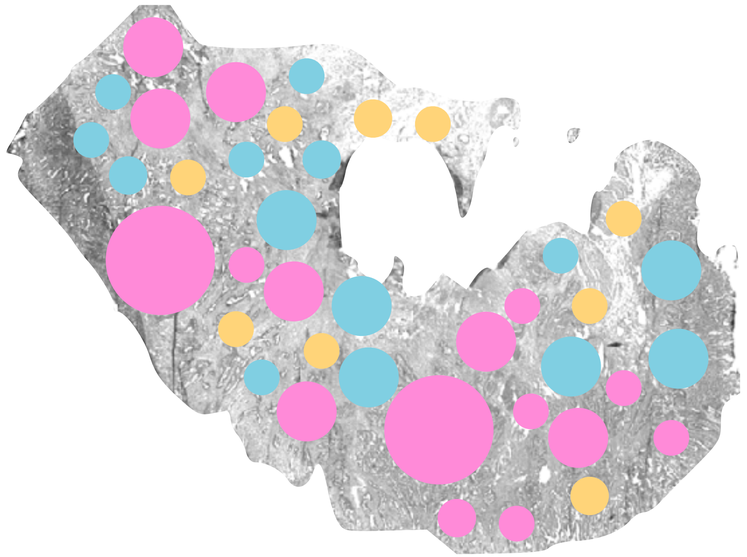


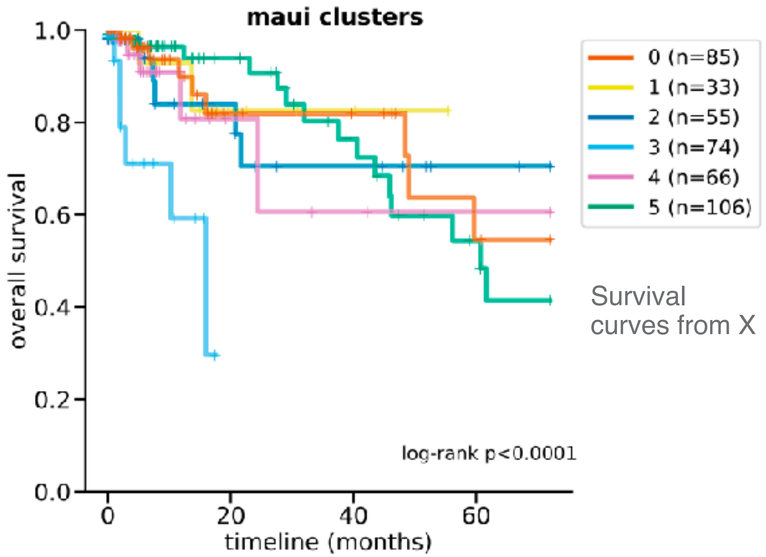
- > 80% of colon cancer is adenocarcinoma [Xi et al., 2021]
- Molecular sub-types differentiate prognosis [Ronen et al., 2019], but RNA-seq rarely integrated
- Previous work to predict gene mutations [Kather et al., 2020] and molecular subtypes (CMS) [imCMS]

- DL interpretability to facilitate scientific discovery:
  - relationship between tissue microscopy and molecular patterns

Objective: identify histologic appearance of molecular subtypes





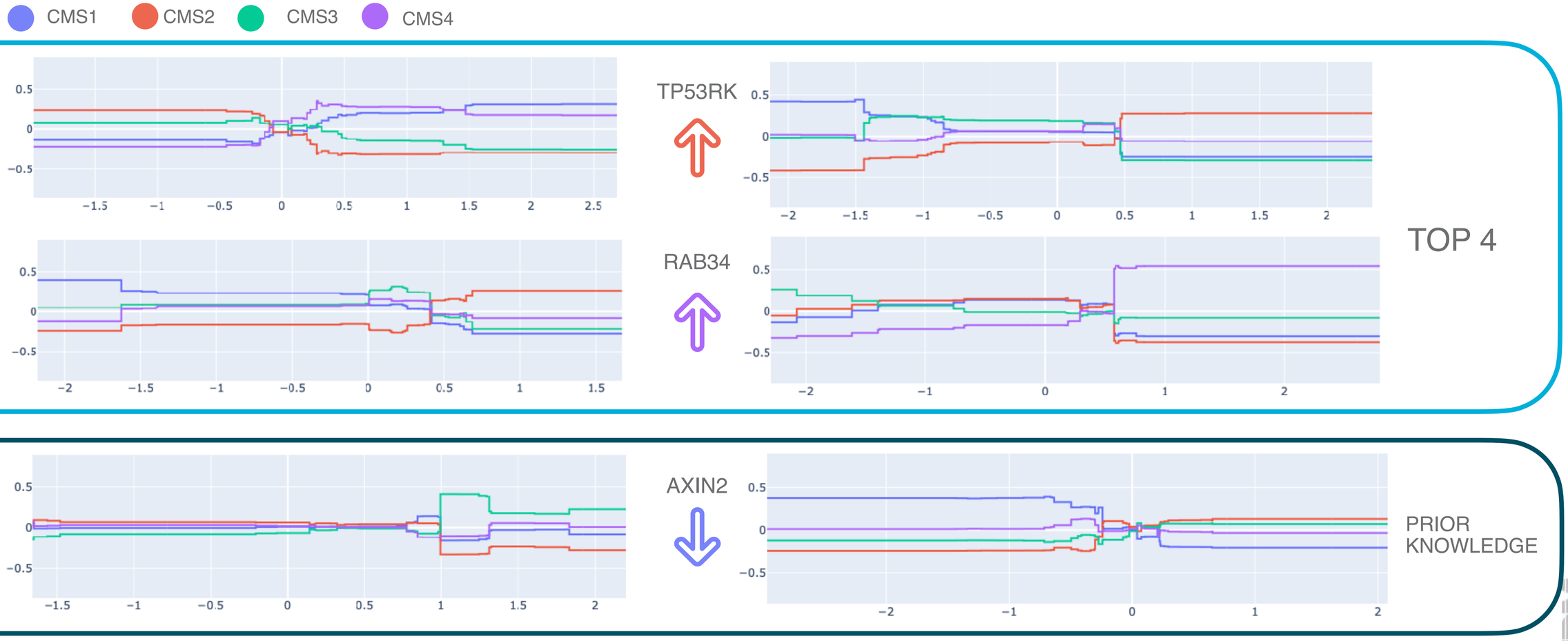


Survival curves from X

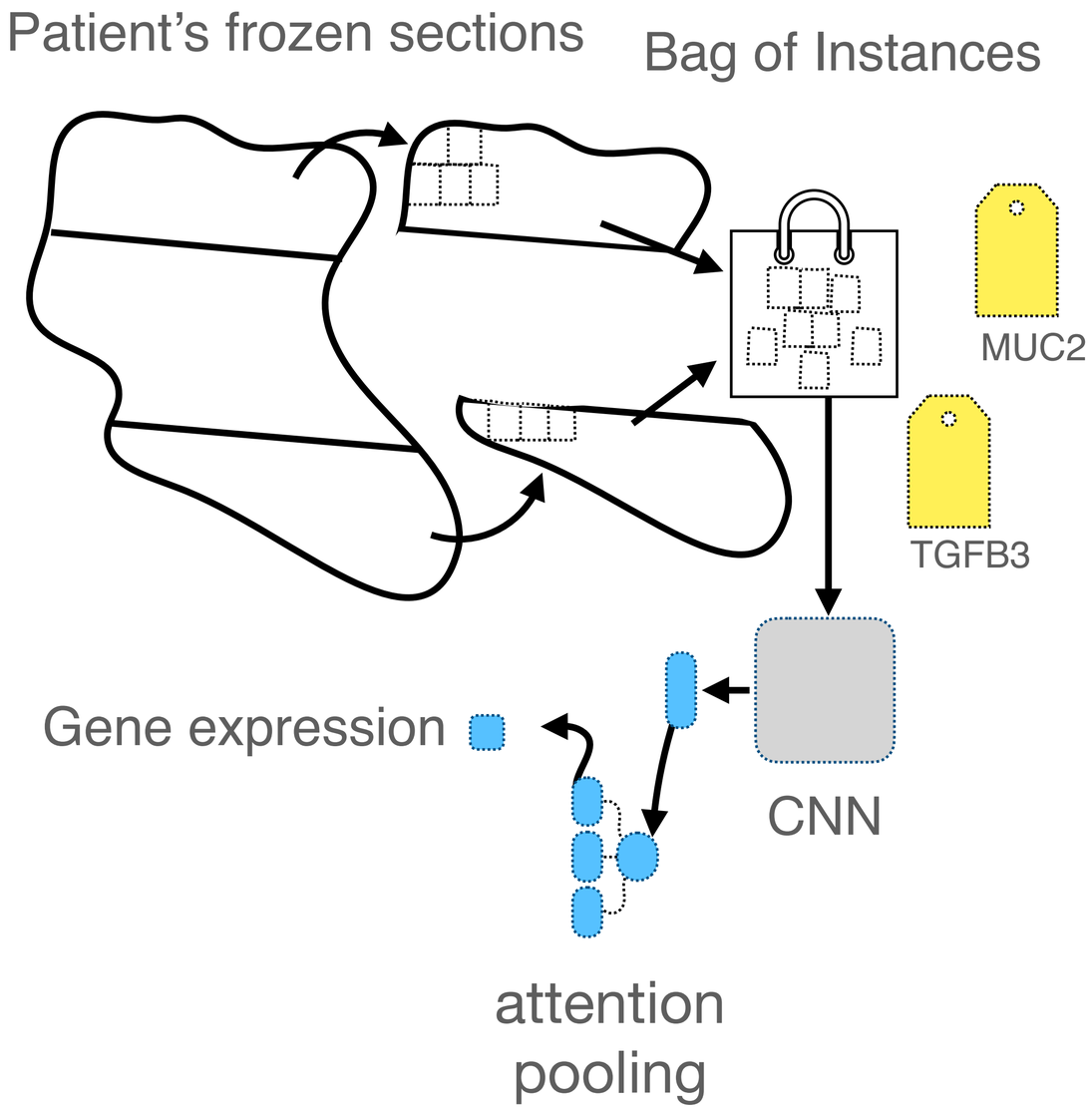
- Use DL interpretability to discover relationships between
  - Histologic tissue appearance and molecular subtypes
  - Tumor heterogeneity and molecular analyses

Step back: gene selection and upper bound

- 40-gene signature of CMS [Buechler et al., 2020] + 7 additional genes identified as biomarkers [Pan et al., 2019; PMC3635192]
- [!!!] Preliminary analysis with Explainable Boosting Machines (EBMs)[Caruana et al., 2015]
- Test AUC (One-Vs-Rest) UB **0.94**; vs. vanilla XGBoost AUC 0.88 (thinking about Context after Ben's talk…)



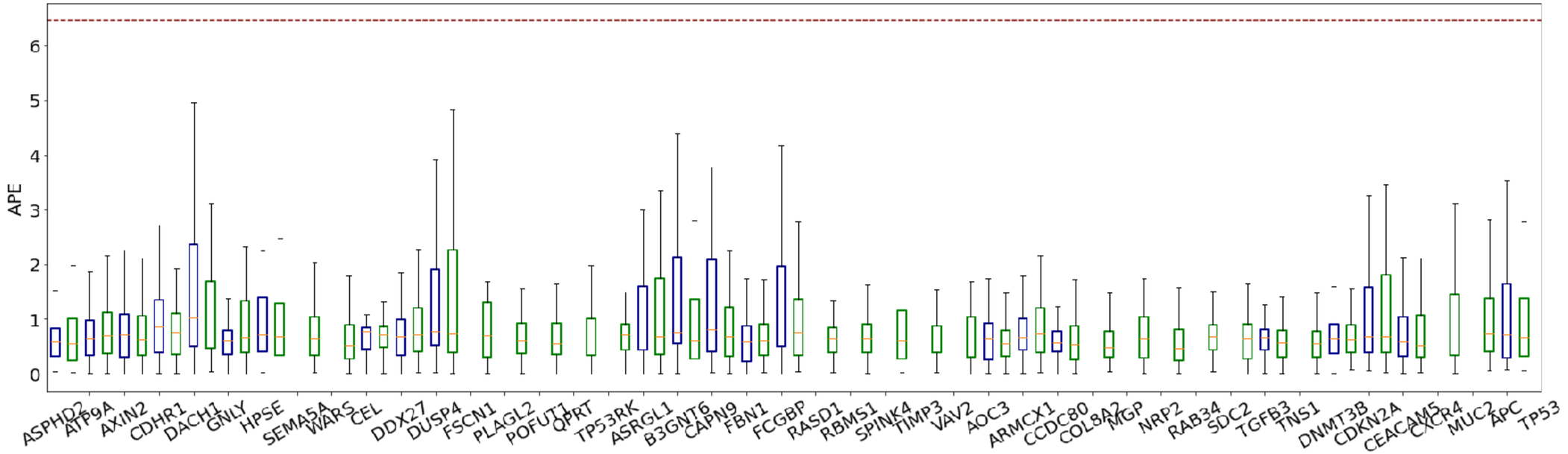
How? Paying attention as the model does



- Attention-based Multiple Instance Learning [Ilse et al., 2018]:
  - Learns a single label from a bag of instances
  - Attention-based pooling
  - 1 model/gene; multi-task combination of genes
- Very hard for a pathologist, but not random DL predictions for 47 genes:
  - Mean Average Percentage Error < 60%
  - CMS prediction AUC (OVR) UB 0.94
  - CMS prediction from images AUC 0.67

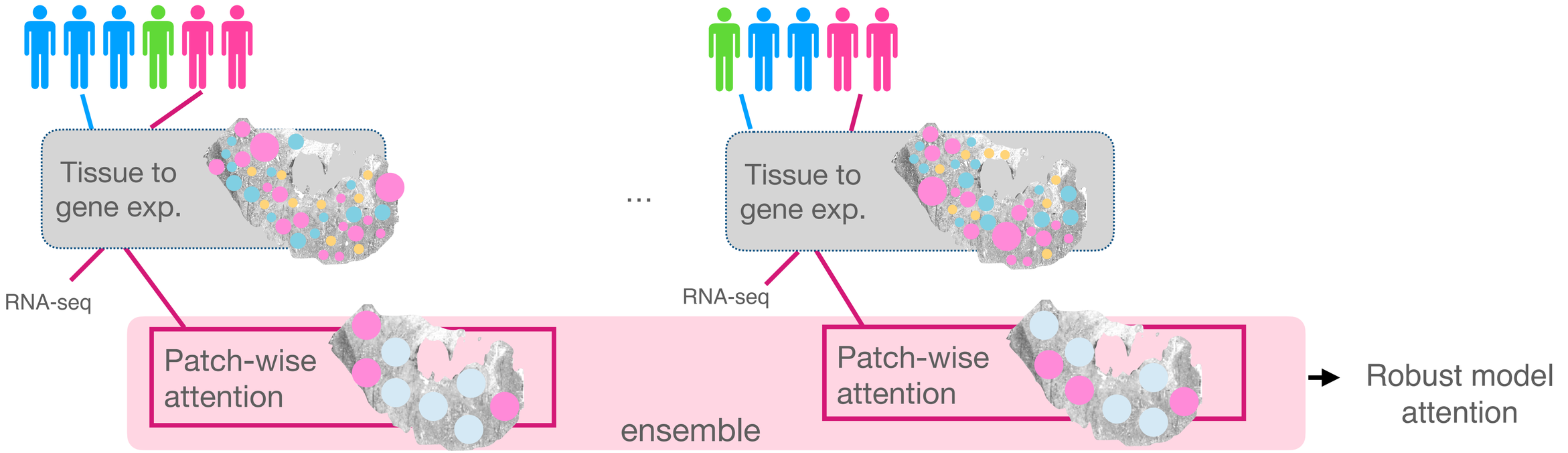
Results

- Test Median Absolute Percentage Error (MAPE) avg. all genes = **0.66** vs. 0.80 baseline and 0.93 random
- Top 4 genes from EBMs not too well learned:
  - FSNC1 ρ = 0.26; MAPE = 0.70 // TP53RK ρ = 0.58; MAPE = 0.71 // QPRT ρ = 0.43; MAPE = 0.74 // RAB34 ρ = 0.63; MAPE = 0.68
- Best gene model is AOC3: **0.54** MAPE (vs. 60% baseline), ρ=0.63 (vs. 0.50 baseline)*
  - AOC3 is over expressed (+) in CMS4 and plays a role in adipogenesis NRP2,
- COL8A2, TGFB3 also have ρ at 0.57, 0.57 and 0.50 (vs. 0.65, 0.51, 0.43 baseline)*



* all p-values < 0.001

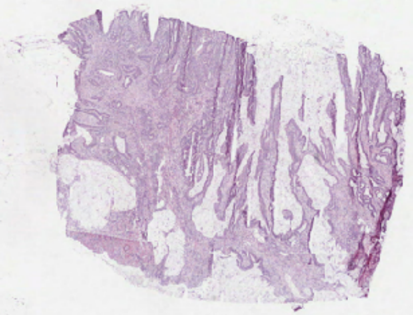
Interpreting the model



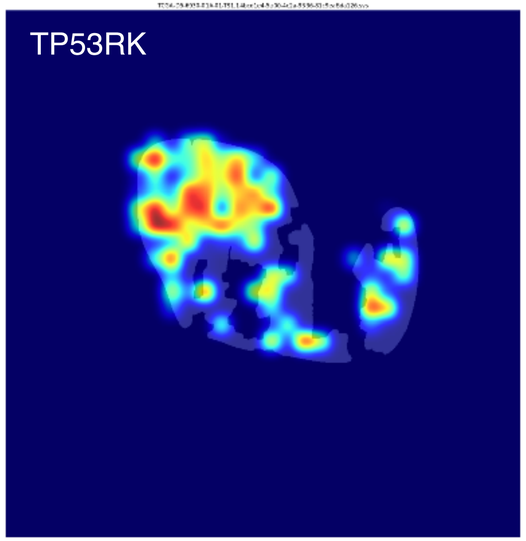
*Based on the WIP of Nicolas Deutschman and Johnatan Hub

Interpreting the model … by cherry picking

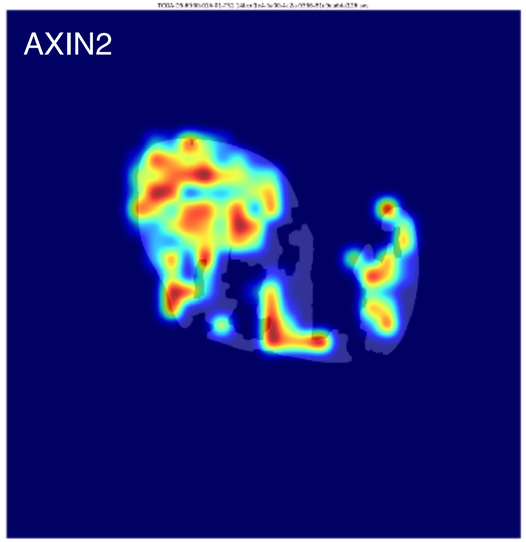
TOP SECTION



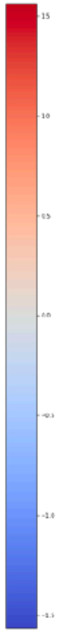
TP53RK



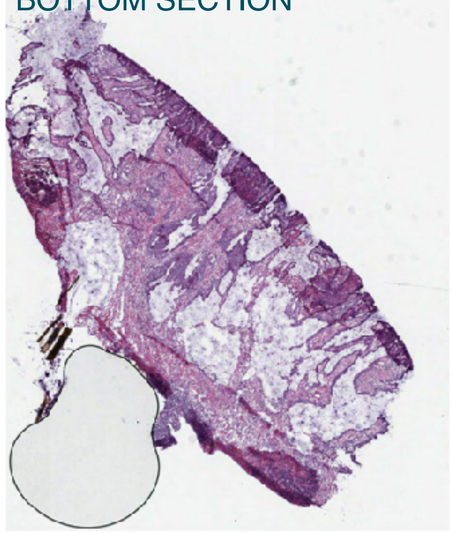
AXIN2



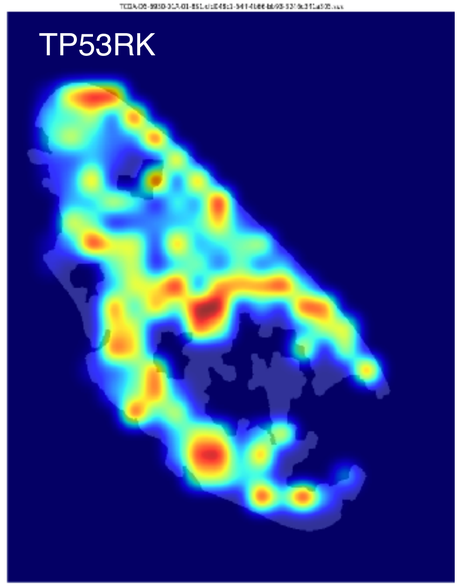
AXIN2 GEX



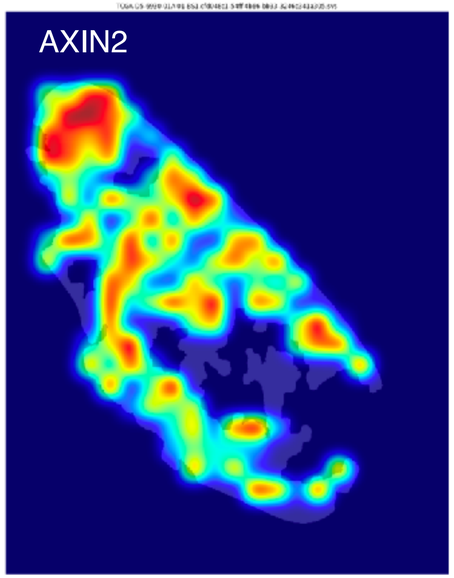
BOTTOM SECTION



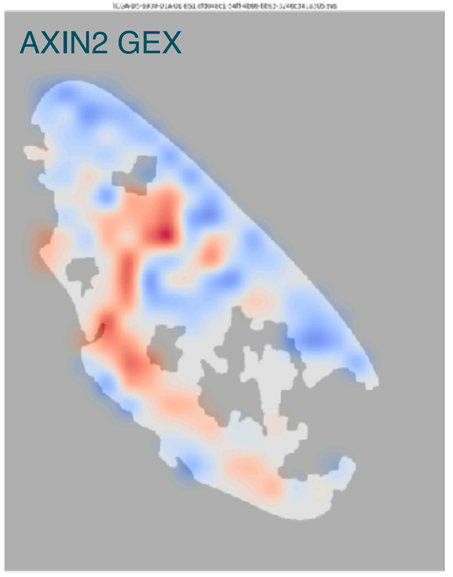
TP53RK



AXIN2



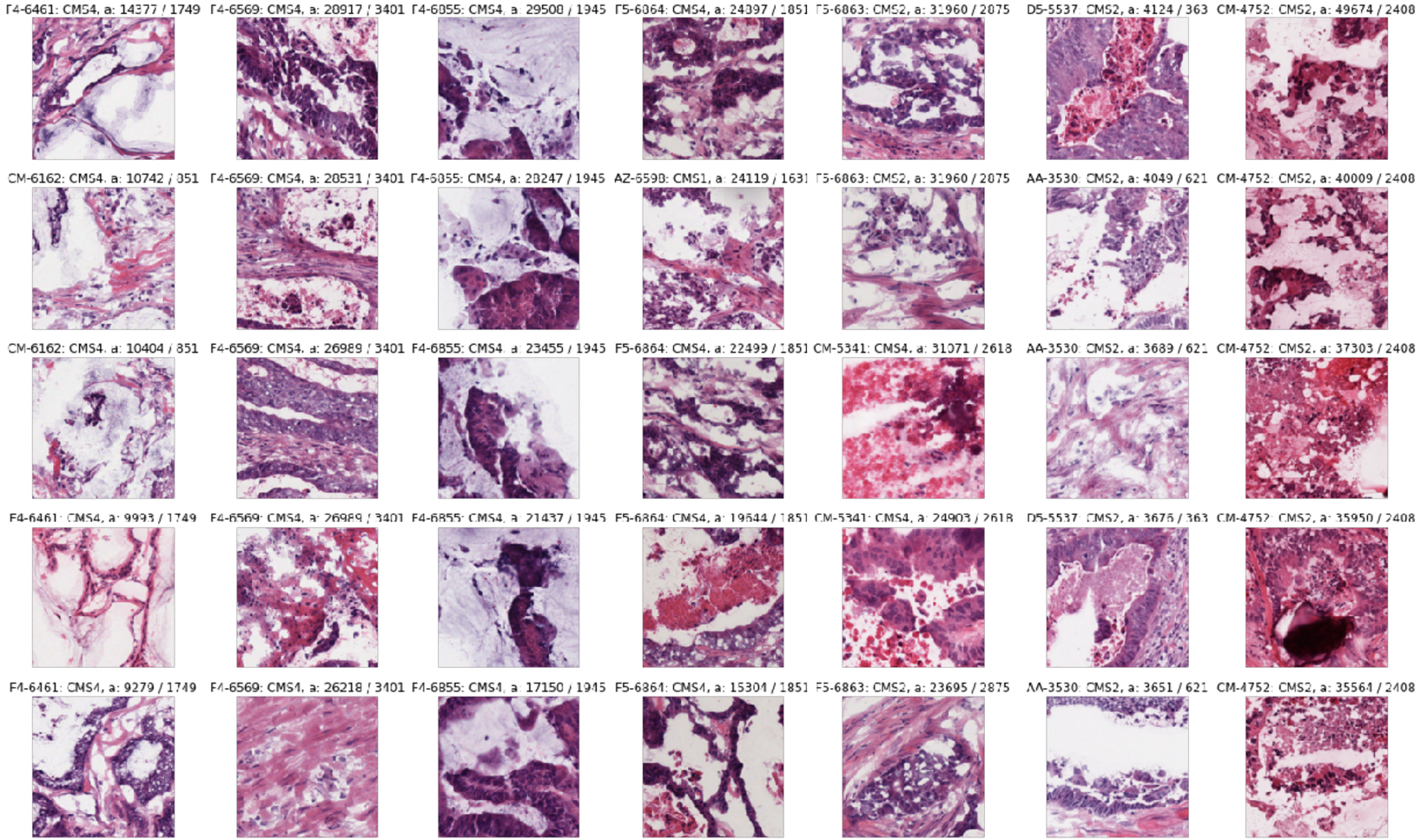
AXIN2 GEX



Come talk to me for more results!

Interpreting the model … by measuring

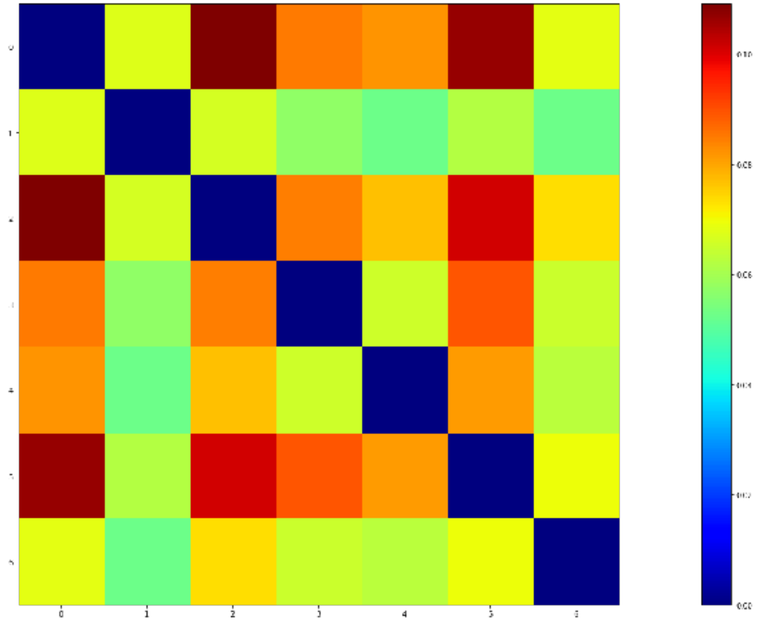
TP53RK ⇩ CMS2



LOW GEX

HIGH GEX

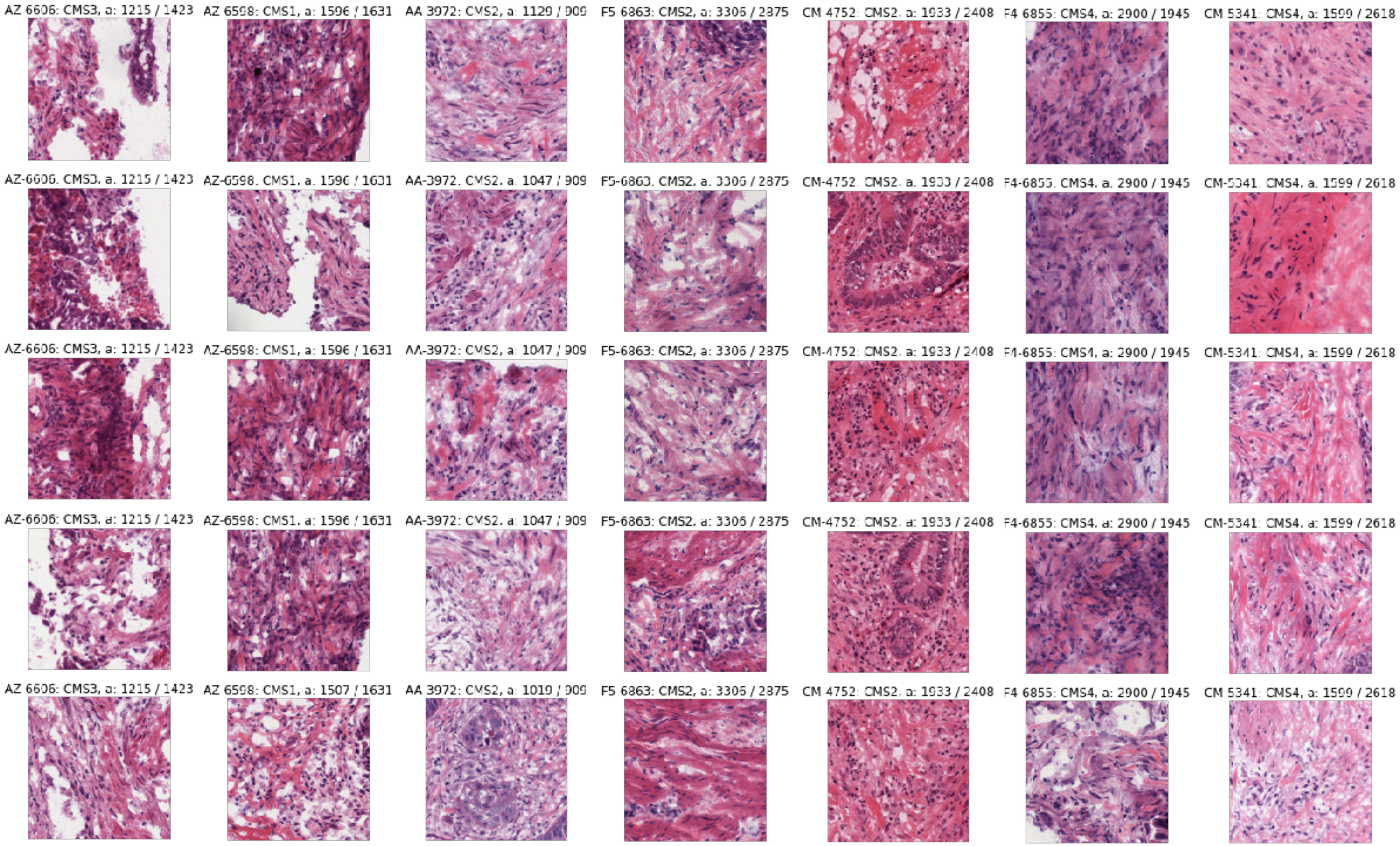
Structural Similarity Index



Model seems confused…

Interpreting the model … by measuring

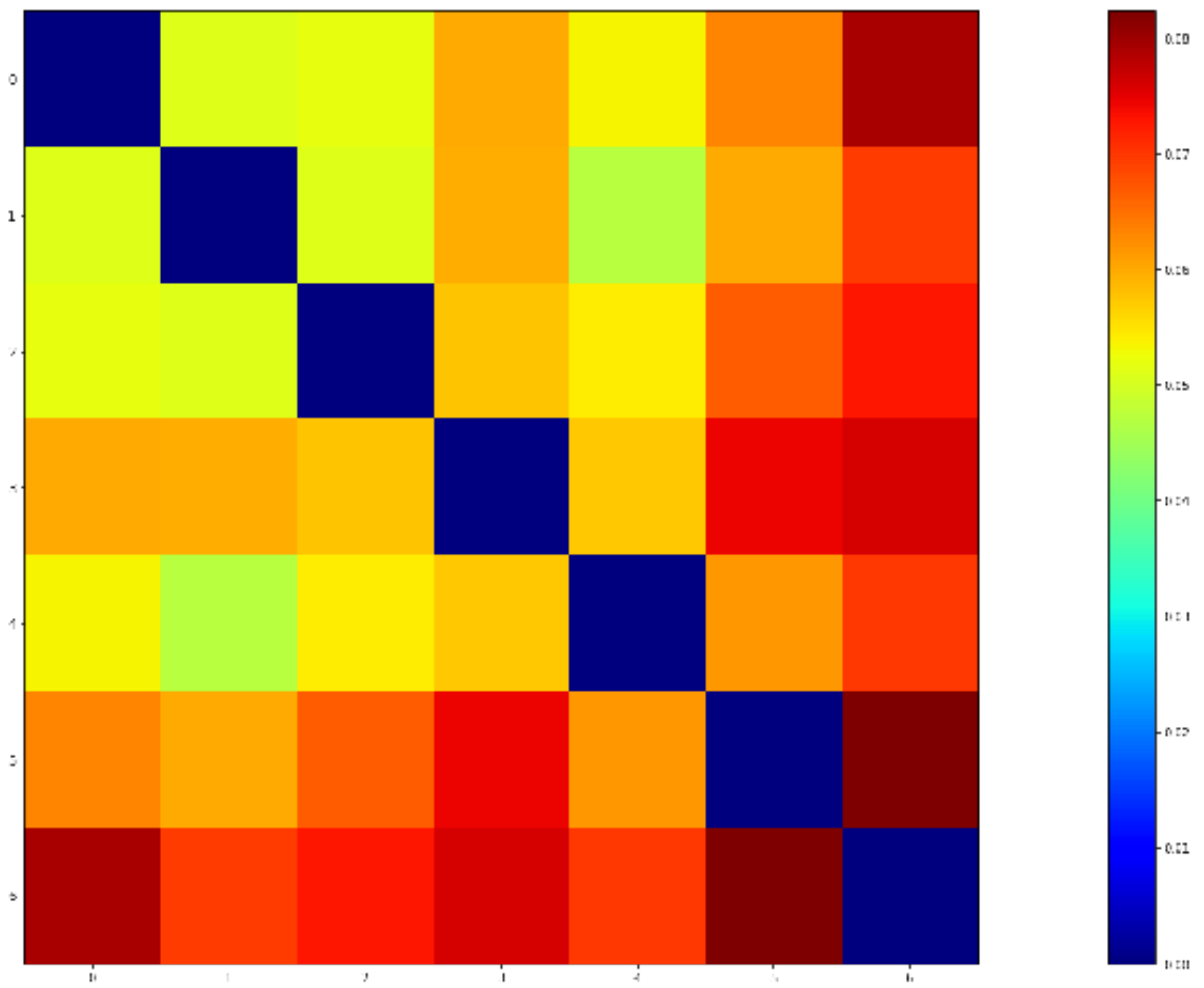
RAB34 ↑CMS4



LOW GEX

HIGH GEX

Better job at separating visual features



… but more metrics should be used!

Final remarks

- There is so much we do not know and understand yet
- Interpretability may be a means to fill the gap between what DL can achieve and humans cannot

- In biomedical research, it can uncover new patterns
  - How do we assess, verify and test new knowledge?
  - How do we disentangle real relationships from spurious ones?

  - Attention mechanisms in MIL can teach us about where to pay attention, yet we need to understand how and when we can translate the discovered information into new knowledge

  - Yet preliminary work, lots to extend further