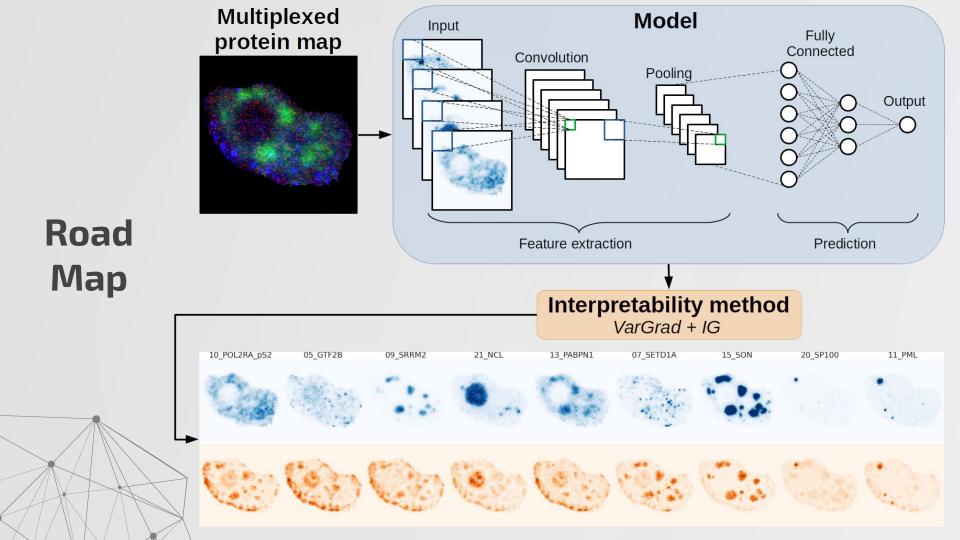


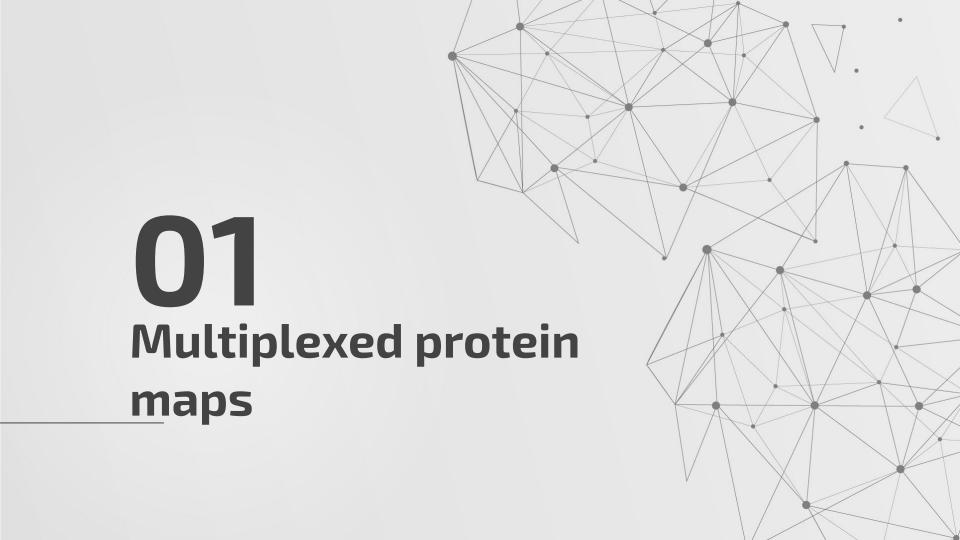


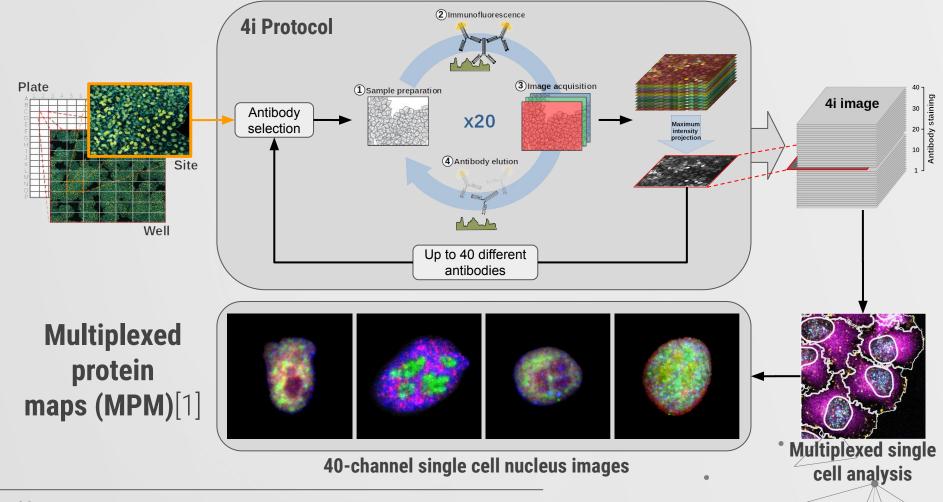
ICB

Predicting transcription rate from multiplexed protein maps using deep learning

Andres Becker May 25, 2021 Advisor: Dr. Hannah Spitzer



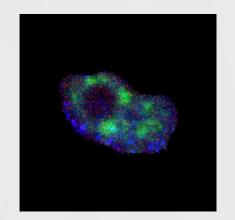


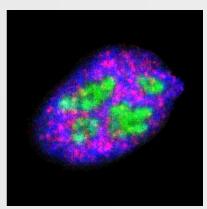


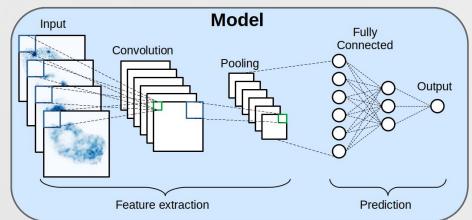
Multiplexed protein maps

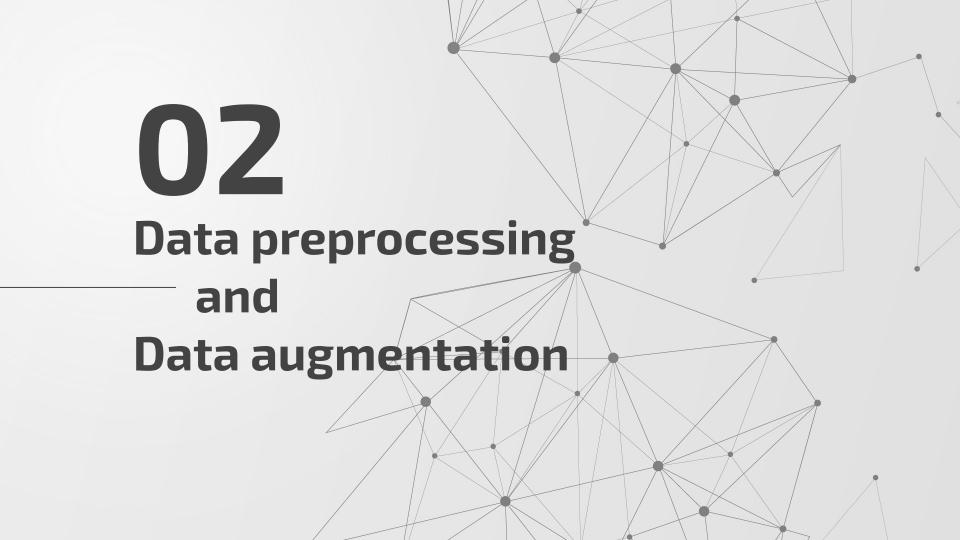
The data provided by **Scott Berry** from Pelkmans Lab (at the University of Zurich)

- **3703** images of cell nucleus
- **33** channels (proteins markers)
- Cell **transcription rate** (TR) is used as target variable
 - Average of channel 5-Ethynyl Uridine (EU),
 which marks newly transcribed RNA
 molecules in 30 minutes









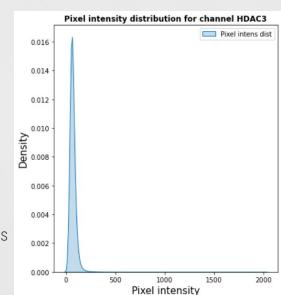
Data preprocessing

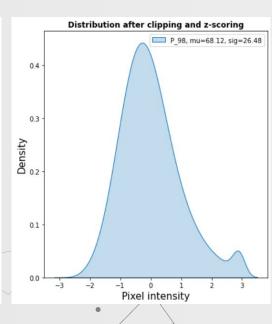
To improves model performance and training stability:

- Per-channel 98% percentile clipping
- Per-channel standardization

$$z_c = \frac{x_c - \mu_c}{\sigma_c}$$

 After preprocessing, the cell images are stored as a TensorFlow dataset

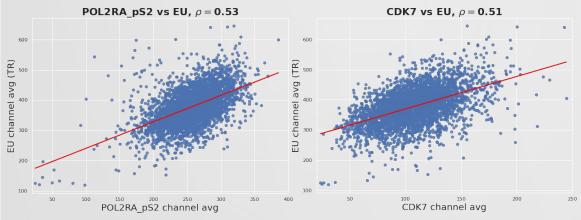


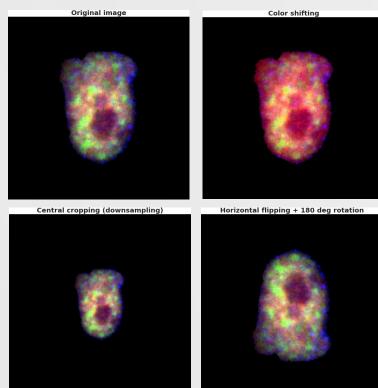


Data augmentation

To alleviate overfitting [1] and encourage the model to focus on spatial information:

- Random color shifting
- Random upsampling and downsampling (center zoom in/out)
- Random horizontal flipping + 90 degree rotations





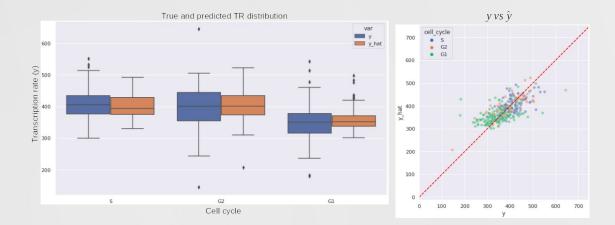


Model performance comparison

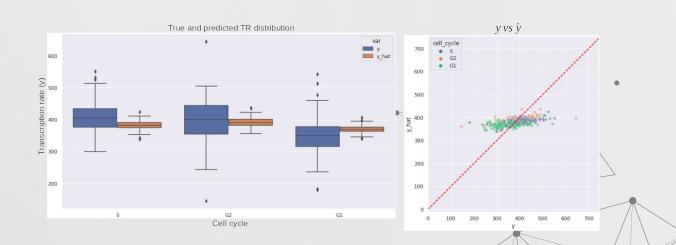


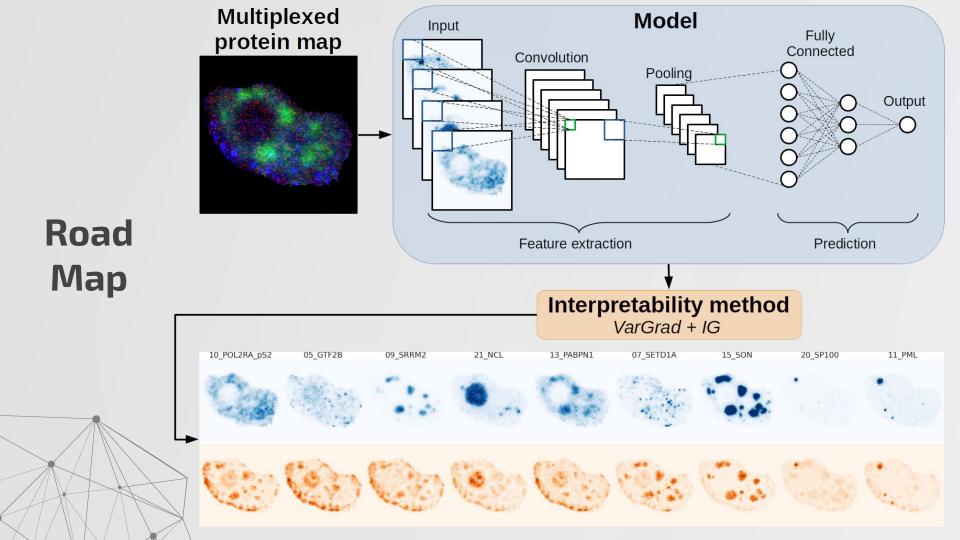
Model performance comparison

 Simple CNN model trained only with spatial information



Linear model
 trained only with
 spatial
 information







Interpretability methods

- Attribution methods assign a score to each input feature based on how much it contribute to the output of the model [1]
- In gradient-based methods the score assignment is based on the gradient [2]

$$\phi(f, x) := \nabla_x f$$
$$= \frac{\partial f}{\partial x}$$

However, as the model learns the relationship between inputs and outputs, the gradients can approximate to 0 (saturation) [3]

[1] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. "How to Explain Individual Classification Decisions".

[2] P. Sturmfels, S. Lundberg, and S.-I. Lee. "Visualizing the Impact of Feature Attribution Baselines".

[3] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks".



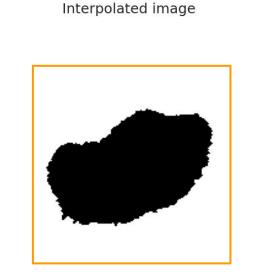
Interpretability methods: Integrated Gradients (IG)

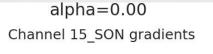
To alleviate saturation, IG accumulates the gradient over a progression from a baseline to the input image [1]

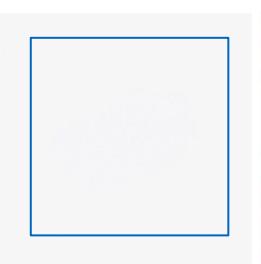
$$\phi_i^{IG}(f, x, x') := (x_i - x_i') \int_{\alpha=0}^1 \underbrace{\frac{\partial f(x' + \alpha(x - x'))}{\partial x_i}} dx$$

-0.005

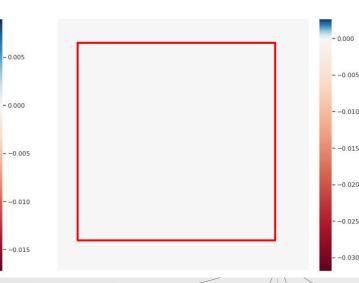
-0.000







Channel 15 SON IG



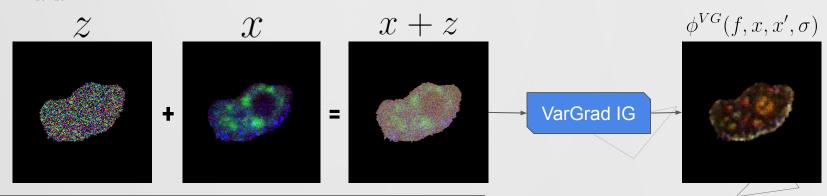
Interpretability methods: VarGrad IG

- However, attribution methods like IG are usually noisy and in some cases not better than a random designation of feature importance [1]
- Ensemble interpretability method like VarGrad [2] overcome simple methods like IG [3]

$$\phi^{VG}(f, x, x', \sigma) := Var(\Phi)$$

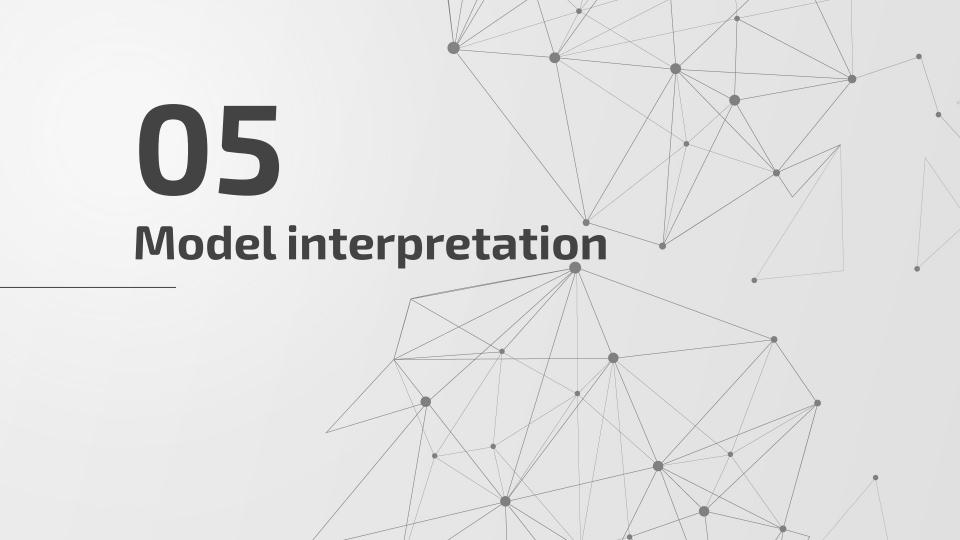
where
$$\Phi := \{\phi^{IG}(f, x + z, x') \mid z \in Z\}$$
, $Z := \{z_1, \dots, z_M\} \subseteq \mathbb{R}^{D \times D \times C}$

and $z_{d_1,d_2,c} \sim \mathcal{N}(\mu,\sigma^2)$ are the i.i.d entrances of the noise image \mathcal{Z}



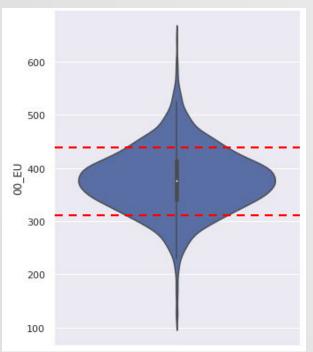
- [1] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. "A benchmark for interpretability methods in deep neural networks".
- [2] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. "Smooth-Grad: removing noise by adding noise".
- [3] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim. "Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values"

$$\phi^{VG}(f, x, x', \sigma) := Var(\{\phi^{IG}(f, x + z, x) \mid z \in \{z_1, \dots, z_M\} \subseteq \mathbb{R}^{D \times D \times C}\}$$
where $z_{d_1, d_2, c} \sim \mathcal{N}(\mu, \sigma^2)$, with $d_1, d_2 \in \{1, \dots, D\}$ and $c \in \{1, \dots, C\}$



Experimental setup

- To classify the cells as low, medium and high TR, we use one standard deviation away from the mean TR
- Sample 3 cells with low, medium and high TR at random



High: if 438.02 <= TR

Medium: if 316.56 < TR < 438.02

Low: if TR <= 316.56

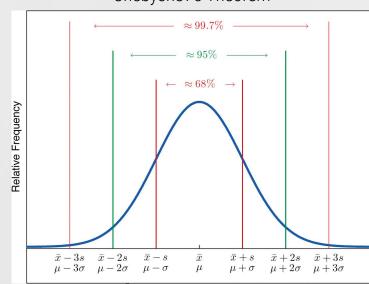
0r

High: 14.7% / 544 cells

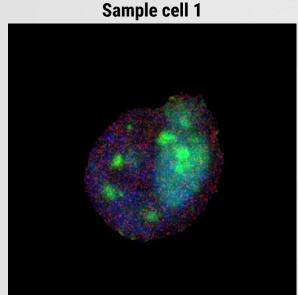
Medium: 71% / 2627 cells

Low: 14.3 % / 532 cells

Chebyshev's Theorem



Experimental setup

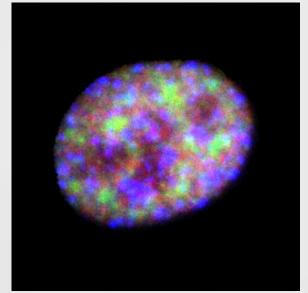


Cell id: 277417

TR: 133.04 (Low TR)

Cell Cycle: G1

Sample cell 2

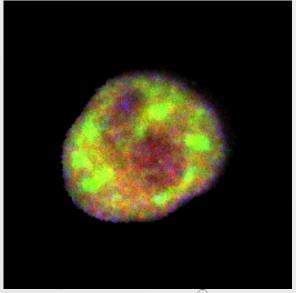


Cell id: 321001

TR: 378.19 (Medium TR)

Cell Cycle: S

Sample cell 3

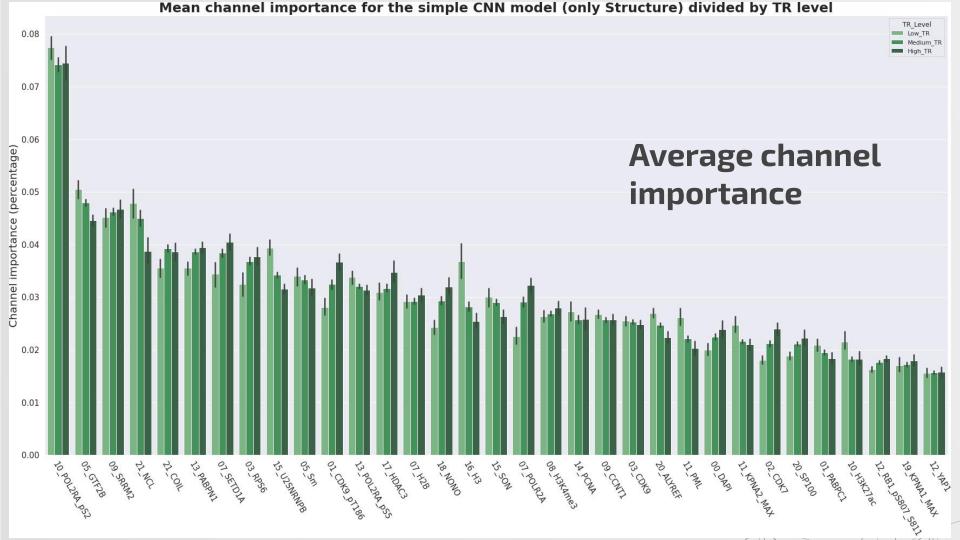


Cell id: 195536

TR: 540.09 (High TR)

Cell Cycle: G2

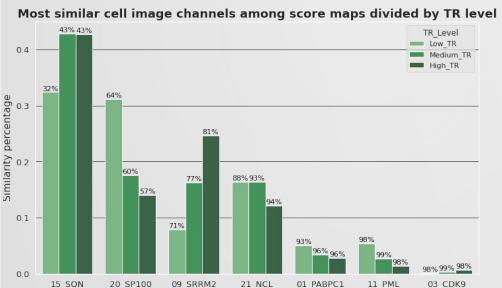
	10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100			
Cell id : 277417 TR : 133.04 (Low TR)				score mar	channels						
Cell Cycle: G1				Score map	Chamileis						
och oyolc. Of			0	(3)							
	10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100			
Cell id : 321001	7		3	and ?	3)		530				
TR : 378.19 (Medium TR)	score map channels										
Cell Cycle: S				o/# '	3	1					
	10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100			
Cell id : 195536	0		5		0		5				
TR : 540.09 (High TR)				score map	channels						
Cell Cycle: G2	致为	· 500	\$ (3)	1	4.5	5.5					



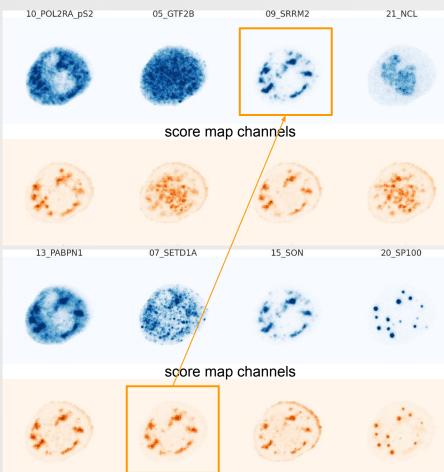


Similarity between Score Maps and Cell image

- Measure the similarity between score maps channels and cell image channels.
- Then, for each score map channel, select the cell image chanel most similar to it.



Cell image channels

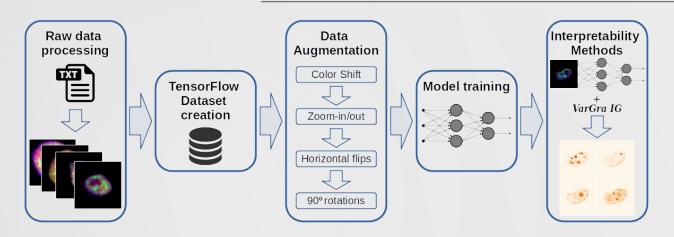


Most similar cell image channels to score map channels divided by transcription level

				Low	_TR								Mediu	m_TR								High	_TR				1.0
15_SON	84.0%	6.0%	8.0%	0%	0%	2.0%	0%	0%	15_SON	94.3%	<1%	4.1%	0%	<1%	<1%	0%	0%	15_SON	94.7%	0%	3.5%	0%	0%	0%	1.7%	0%	2.0
20_SP100	0%	90.0%	6.0%	0%	0%	4.0%	0%	0%	20_SP100	0%	98.8%	0%	1.1%	0%	0%	0%	0%	20_SP100	0%	100.%	0%	0%	0%	0%	0%	0%	
09_SRRM2	64.0%	12.0%	20.0%	0%	2.0%	2.0%	0%	0%	09_SRRM2	53.4%	1.8%	43.1%	0%	<1%	<1%	0%	0%	09_SRRM2	35.0%	0%	63.1%	0%	0%	0%	0%	0%	
21_NCL	0%	10.0%	0%	86.0%	0%	4.0%	0%	0%	21_NCL	0%	3.7%	0%	95.4%	0%	<1%	0%	0%	21_NCL	0%	8.7%	0%	91.2%	0%	0%	0%	0%	
01_PABPC1	2.0%	38.0%	2.0%	44.0%	14.0%	0%	0%	0%	01_PABPC1	4.5%	24.2%	0%	59.4%	10.9%	<1%	<1%	0%	01_PABPC1	3.5%	28.0%	1.7%	38.5%	28.0%	0%	0%	0%	
11_PML	0%	57.9%	0%	0%	0%	42.0%	0%	0%	11_PML	0%	50.7%	<1%	0%	0%	48.8%	0%	0%	11_PML	0%	54.3%	0%	0%	0%	45.6%	0%	0%	
03_CDK9	72.0%	10.0%	12.0%	0%	0%	4.0%	0%	0%	03_CDK9		4.5%	14.3%	0%	<1%	<1%	7.5%	0%	03_CDK9	56.1%	7.0%	21.0%	0%	0%	0%	15.7%	0%	- 0.8
00_DAPI	44.0%	24.0%	8.0%	12.0%	0%	0%	0%	12.0%	00_DAPI	56.4%	3.4%	30.3%	5.6%	0%	<1%	0%	3.4%	00_DAPI	54.3%	0%	31.5%	0%	0%	0%	0%	14.0%	
19_KPNA1_MAX	42.0%	26.0%	6.0%	8.0%	0%	4.0%	0%	0%	19_KPNA1_MAX		6.8%	14.3%	9.0%	0%	0%	0%	0%	19_KPNA1_MAX		0%	15.7%	1.7%	0%	0%	0%	0%	
09_CCNT1		4.0%	12.0%	8.0%	0%	0%	0%	0%	09_CCNT1	69.6%	<1%	25.0%	2.6%	0%	0%	0%	0%	09_CCNT1		0%	33.3%	0%	0%	0%	0%	0%	
16_H3	0%	66.0%	6.0%	0%	16.0%	12.0%	0%	0%	16_H3	1.5%	63.2%	0%	1.1%	25.0%	6.4%	0%	0%	16_H3	0%	63.1%	0%	7.0%	28.0%	0%	0%	0%	
03_RPS6	0%	14.0%	0%	82.0%	0%	0%	0%	0%	03_RPS6	3.0%	2.6%	<1%	93.1%	0%	<1%	0%	0%	03_RPS6	3.5%	5.2%	0%	91.2%	0%	0%	0%	0%	
11_KPNA2_MAX		38.0%	6.0%	0%	36.0%	6.0%	0%	0%	11_KPNA2_MAX	25.7%	28.4%	3.7%	3.0%	35.6%	<1%	0%	0%	11_KPNA2_MAX		21.0%	7.0%	1.7%	19.2%	0%	0%	0%	0.6
<u>0</u> 17_HDAC3	0.0000000000000000000000000000000000000	22.0%	8.0%	0%	0%	2.0%	0%	0%	17_HDAC3	60.2%	5.6%	31.8%	0%	0%	1.5%	0%	0%	17_HDAC3	49.1%	1.7%	49.1%	0%	0%	0%	0%	0%	-0.6
14_PCNA		68.0%	6.0%	0%	12.0%	10.0%	0%	0%	14_PCNA	26.1%	46.9%	17.4%	1.5%	2.6%	4.1%	0%	0%	14_PCNA	28.0%	35.0%	29.8%	3.5%	0%	0%	0%	0%	
10_POL2RA_pS2		22.0%	22.0%	0%	0%	8.0%	0%	0%	10_POL2RA_pS2		6.0%	40.5%	0%	0%	1.1%	0%	0%	10_POL2RA_pS2	47.3%	3.5%	49.1%	0%	0%	0%	0%	0%	
O7_POLR2A		12.0%	8.0%	26.0%	0%	0%	0%	0%	07_POLR2A		1.5%	34.4%	11.7%	0%	0%	0%	0%	2022-120-722-722-722	43.8%	0%	52.6%	1.7%	0%	0%	0%	0%	
E 18_NONO		12.0%	16.0%	22.0%	0%	0%	0%	0%	18_NONO	58.3%	4.1%	25.0%	11.7%	0%	<1%	0%	0%	18_NONO	54.3%	0%	40.3%	3.5%	0%	0%	0%	0%	
ŭ	56.0%	Administration of the Parket o	12.0%	18.0%	0%	2.0%	0%	0%	12_YAP1	52.6%	8.3%	25.0%	12.8%	0%	<1%	0%	0%	12_YAP1	56.1%	5.2%	31.5%	5.2%	0%	0%	0%	0%	
O5_GTF2B		20.0%	4.0%	48.0%	4.0%	8.0%	0%	0%	tona—om-ma		3.7%	1.8%	47.7%	<1%	<1%	0%	0%	Service Control of		5.2%	3.5%	45.6%	0%	0%	0%	0%	- 0.4
02_CDK7	1000 STORY	36.0%	10.0%	24.0%	2.0%	0%	0%	0%	02_CDK7		10.6%	30.6%	12.5%	0%	0%	0%	0%	02_CDK7		3.5%	45.6%	1.7%	0%	0%	0%	0%	
20_ALYREF		52.0%	6.0%	2.0%	10.0%	6.0%	2.0%	0%	20_ALYREF		39.0%	2.6%	5.3%	3.7%	3.4%	0%	0%	20_ALYREF		38.5%	7.0%	3.5%	1.7%	0%	0%	0%	
07_SETD1A		0%	22.0%	6.0%	0%	0%	0%	0%	07_SETD1A	71.9%	<1%	26.5%	1.1%	0%	0%	0%	0%	07_SETD1A	68.4%	0%	29.8%	0%	0%	0%	1.7%	0%	
13_PABPN1	Construction (Construction)	14.0%	14.0%	2.0%	0%	4.0%	0%	0%	13_PABPN1	69.3%	1.1%	27.6%	<1%	<1%	<1%	0%	0%	13_PABPN1	68.4%	1.7%	29.8%	0%	0%	0%	0%	0%	
08_H3K4me3	5-1000 CV/507 MS	22.0%	8.0%	4.0%	0%	8.0%	0%	0%	08_H3K4me3		5.6%	34.0%	6.8%	0%	0%	0%	0%	08_H3K4me3	33.3%	3.5%	63.1%	0%	0%	0%	0%	0%	
01_CDK9_pT186		60.0%	6.0%	6.0%	0%	10.0%	0%	0%	01_CDK9_pT186		20.8%	21.9%	7.9%	<1%	<1%	0%	0%	01_CDK9_pT186	56.1%	5.2%	36.8%	0%	0%	0%	1.7%	0%	0.7
		72.0%	4.0%	0%	10.0%	10.0%	0%	0%	10_H3K27ac		56.0%	7.9%	1.5%	3.0%	7.1%	0%	0%	10_H3K27ac	35.0%	35.0%	24.5%	0%	3.5%	0%	0%	0%	-0.2
0.000	14.0%	44.0%	12.0%	0%	24.0%	4.0%	0%	0%	07_H2B		17.8%	28.0%	2.6%	5.3%	<1%	0%	0%	07_H2B	40.3%	3.5%	54.3%	0%	1.7%	0%	0%	0%	
950-000	OWERSON	18.0%	0%	54.0%	0%	2.0%	0%	0%	05_Sm	22.7%	8.3%	1.5%	67.4%	0%	0%	0%	0%	05_Sm	22.8%	7.0%	3.5%	66.6%	0%	0%	0%	0%	
			4.0%	2.0%	10.0%	10.0%	0%	0%	13_POLZRA_pS5	52.2%	24.6%	6.4%	9.4%	3.4%	3.4%	0%	0%	13_POL2RA_pS5	63.1%	10.5%	19.2%	3.5%	3.5%	0%	0%	0%	
10,000 T 1500 DO	26.0%	10.0%	0% 6.0%	62.0%	0% 20.0%	8.0%	0%	0%	21_COIL	51.8%	1.5% 8.7%	9.8%	36.7% 11.3%	0% 18.1%	1.1%	0%	0% 0%	21_COIL	63.1% 77.1%	7.0%	19.2%	17.5% 7.0%	0% 5.2%	0%	0% 0%	0% 0%	
15_U2SNRNPB		44.0%	6.0%	18.0%	6.0%	4.0%	0%	0%	15_U2SNRNPB 12 RB1 pS807 S811	33.7%	18.9%	23.8%	20.0%	2.2%	1.1%	0%	0%	15_U2SNRNPB	40.3%	7.0%	43.8%	8.7%	0%	0%	0%	0%	
12_RB1_pS807_S811	22.0%	100000000	0.0%	۵.0%	0.0%	4.0%	0%	90	12_ND1_P3007_5011	33.170		23.0% O ₀	20.0%	2.270	1.176	0%	90	12_RB1_pS807_S811	40.3%		43.076 O ₀	0.7%	0,	₹,	0%	90	- 0.0
		70									70									70							

Cell image channels

Outlook



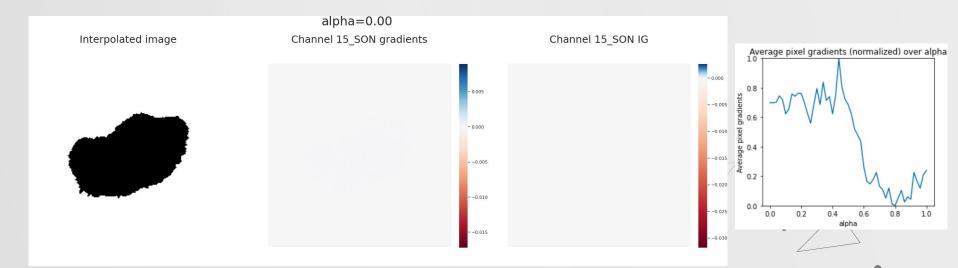
- It is possible to predict TR using only localization information from the cell nucleus.
- Spatial organization of proteins inside the cell nucleus can be as important as the overall protein aboundas.
- It is important to generate data with subnuclear localization information, and develop methods able to use this data.
- Interpretability methods allow us know to which proteins and nuclear bodies were most relevant for the prediction of TR, and how this changes as the TR decreases or increases
- Interpretability methods allow us to learn from CNN models, which has the potential to provide guidance for new discoveries in the field of biology.



Interpretability methods: Integrated Gradients (IG)

To alleviate saturation, IG accumulates the gradient over a progression from a baseline to the input image [1]

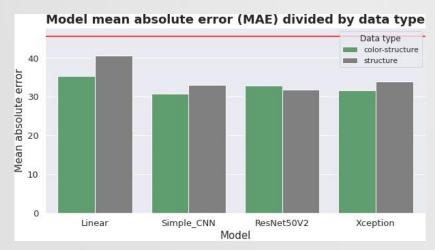
$$\phi_i^{IG}(f, x, x') := (x_i - x_i') \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

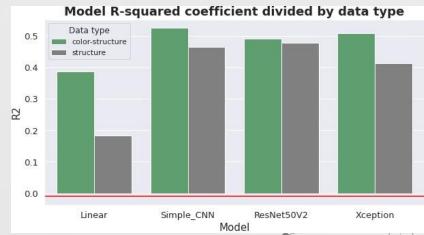




Model performance comparison

Model	Number of parameters	Data type	$ar{e}$	s(e)	R^2	MAE	MSE	Huber
\bar{y} (baseline)	0	targets avg	4.86	59.99	-0.01	45.56	3622	45.07
Linear	34	color-structure	4.06	46.83	0.38	35.26	2203	34.77
	94	structure	4.03	54.15	0.18	40.52	2941	40.02
Simple CNN	160k	color-structure	3.00	41.27	0.52	30.68	1708	30.18
	100K	structure	0.77	43.94	0.46	33.08	1926	32.59
ResNet50V2	24m	color-structure	1.49	42.81	0.49	32.73	1830	32.24
	24111	structure	0.45	43.38	0.47	31.83	1877	31.33
Xception	21m	color-structure	6.69	41.57	0.50	31.66	1768	31.16
	21111	structure	7.23	45.50	0.41	33.92	2117	33.42





Most active channels during TR prediction for the simple CNN model

- Cell id: 277417
- TR: 133.04 (Low TR)
- Cell Cycle: G1

• Cell C	ycie. G i						
10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100
S.			. 7				
8.19%	5.71%	3.92%	3.52%	2.79%	2.58%	2.03%	1.75%
Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap

Most active channels during TR prediction for the simple CNN model

- Cell id: 321001
- TR: 378.19 (Medium TR)

	Cycle: S	,					
10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100
47		530	org ?	53		53	
6.04%	4.65%	4.85%	4.5%	3.82%	3.36%	2.54%	2.08%
			- · ·			330	
Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap
77			OFR 9			330	

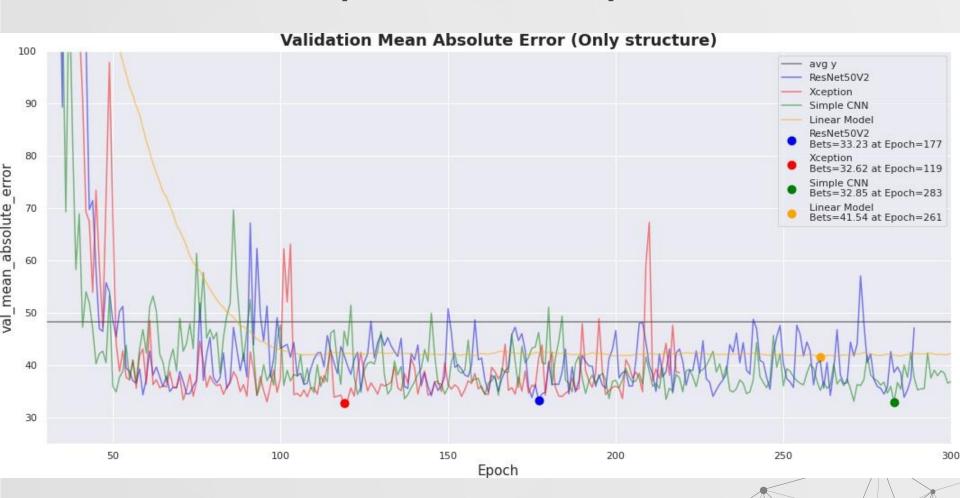
Most active channels during TR prediction for the simple CNN model

- Cell id: 195536
- TR: 540.09 (High TR)
- Call Cycle: G'

• Cell C	ycle: G2						
10_POL2RA_pS2	05_GTF2B	09_SRRM2	21_NCL	13_PABPN1	07_SETD1A	15_SON	20_SP100
		5		63		53	
7.01%	4.01%	4.46%	3.48%	3.8%	4.46%	2.18%	2.84%
W. Co		5		4.3	5		
Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap	Overlap
No. Way		4 34		677	1 × 2	(P)	



Model performance comparison



Cell cycle phase classification (G1, S, G2 and M)

Cell cycle phase was determined by means of a Support Vector Machine (SVM) classifier and k-means clustering.

- 1. A SVM classifier is trained to identify M phase cells based on the nuclear information.
- 2. A second SVM classifier is trained to identify cells in phase S.
- 3. Cells in phase G1 and G2 are classified using a k-means algorithm (excluding cells in S and M phase).



Linear model definition

- To see how much information is encoded only in the cell shape (and not in the color), we also used Linear models to predict the TR and compare the results with the CNN model results.
- Recall that a linear model is defined as:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_d x_{i,d} + \alpha ||\boldsymbol{\beta}||$$

where β_j (with $j \in \{1, \ldots, d\}$) are the coefficients for each of the input variables (in this case the average of each input channel), β_0 is the intercept, $\alpha||\boldsymbol{\beta}||$ the regularization term and $\hat{y_i}$ the target variable (in this case the average of the 00_EU channel, i.e. the TR).



Top 10 most active channels during TR prediction without random color shifting









