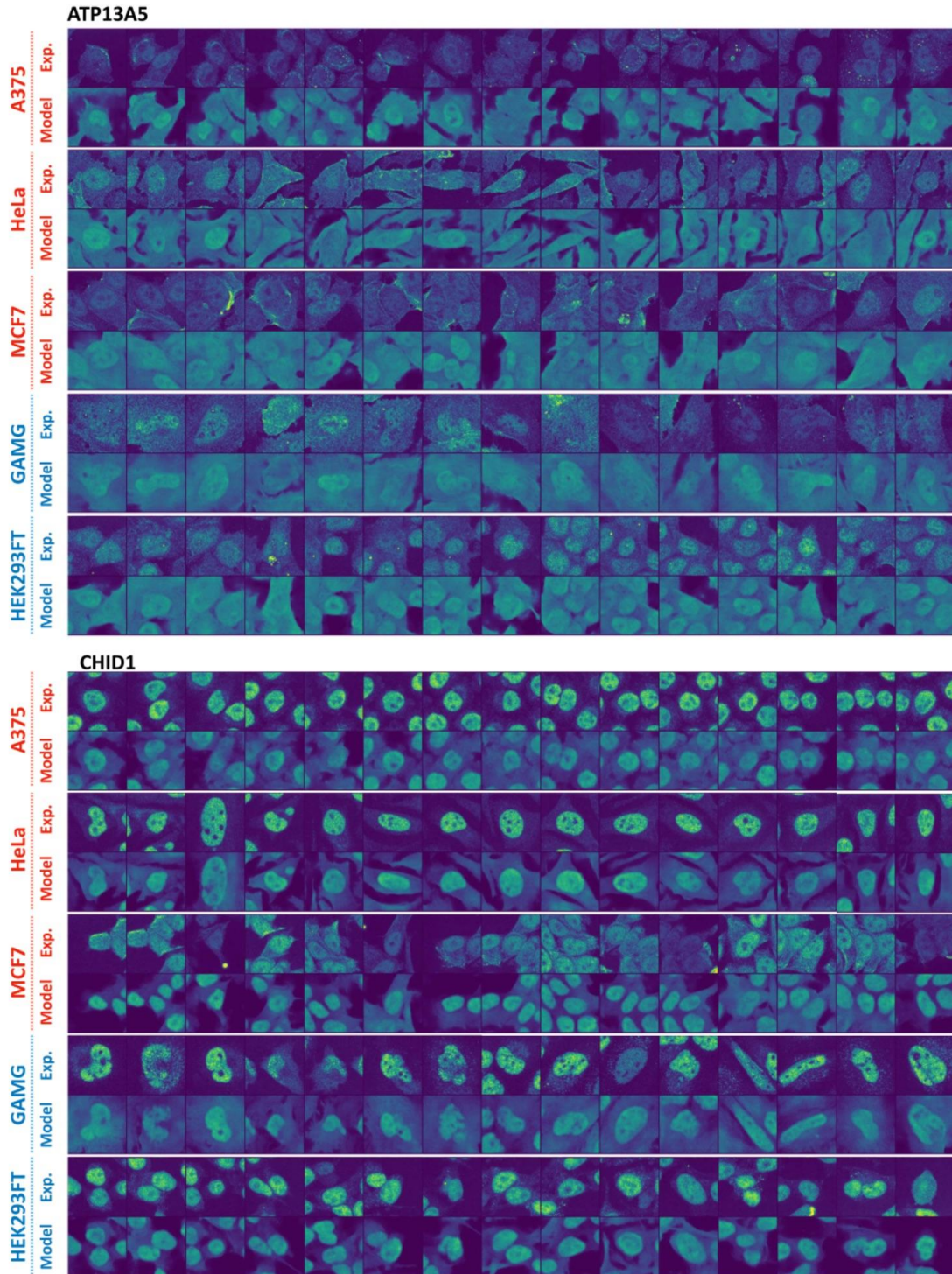

Prediction of protein subcellular localization in single cells

In the format provided by the
authors and unedited

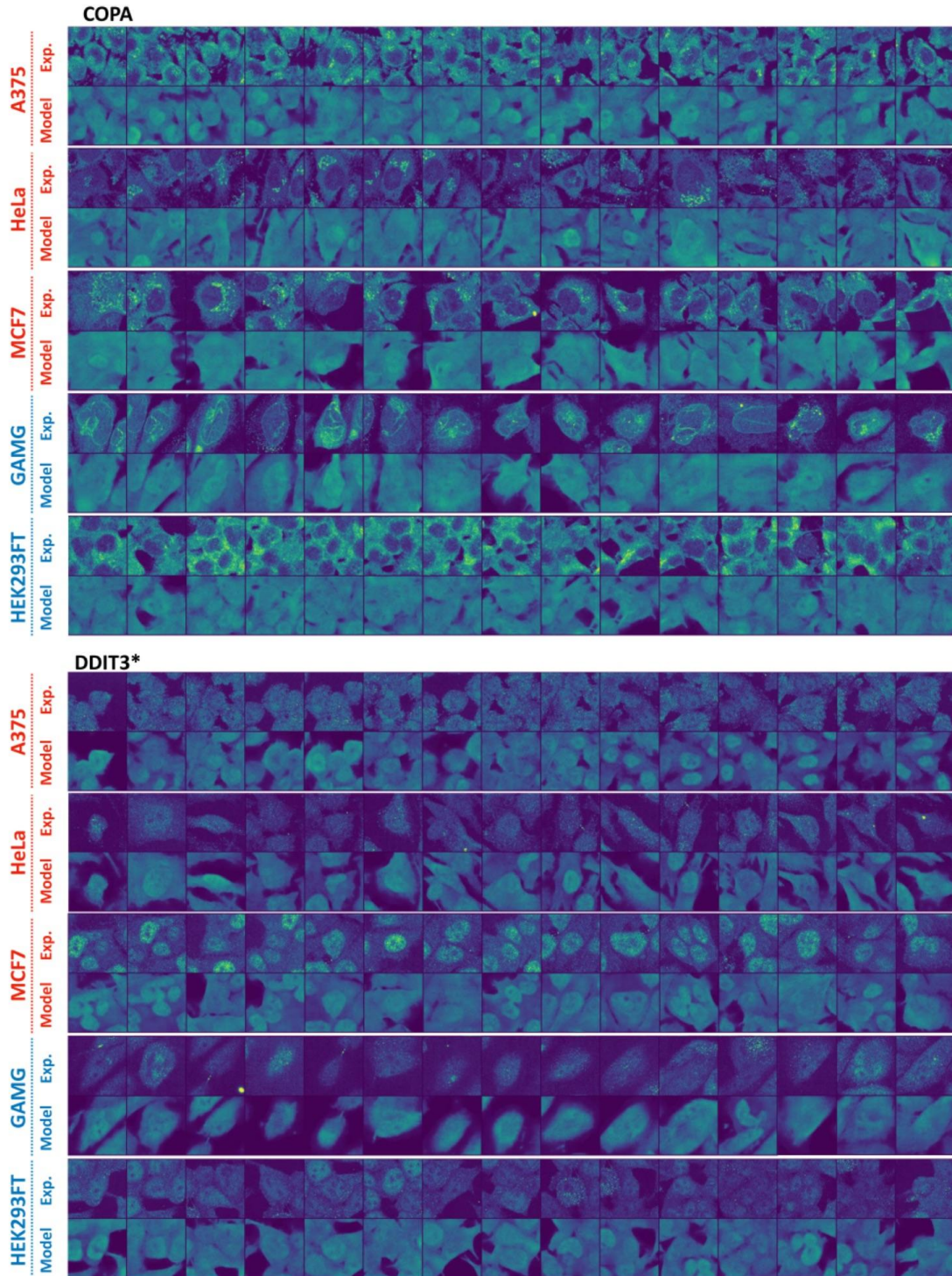
Table of Contents:

Supplementary Figures 1-8

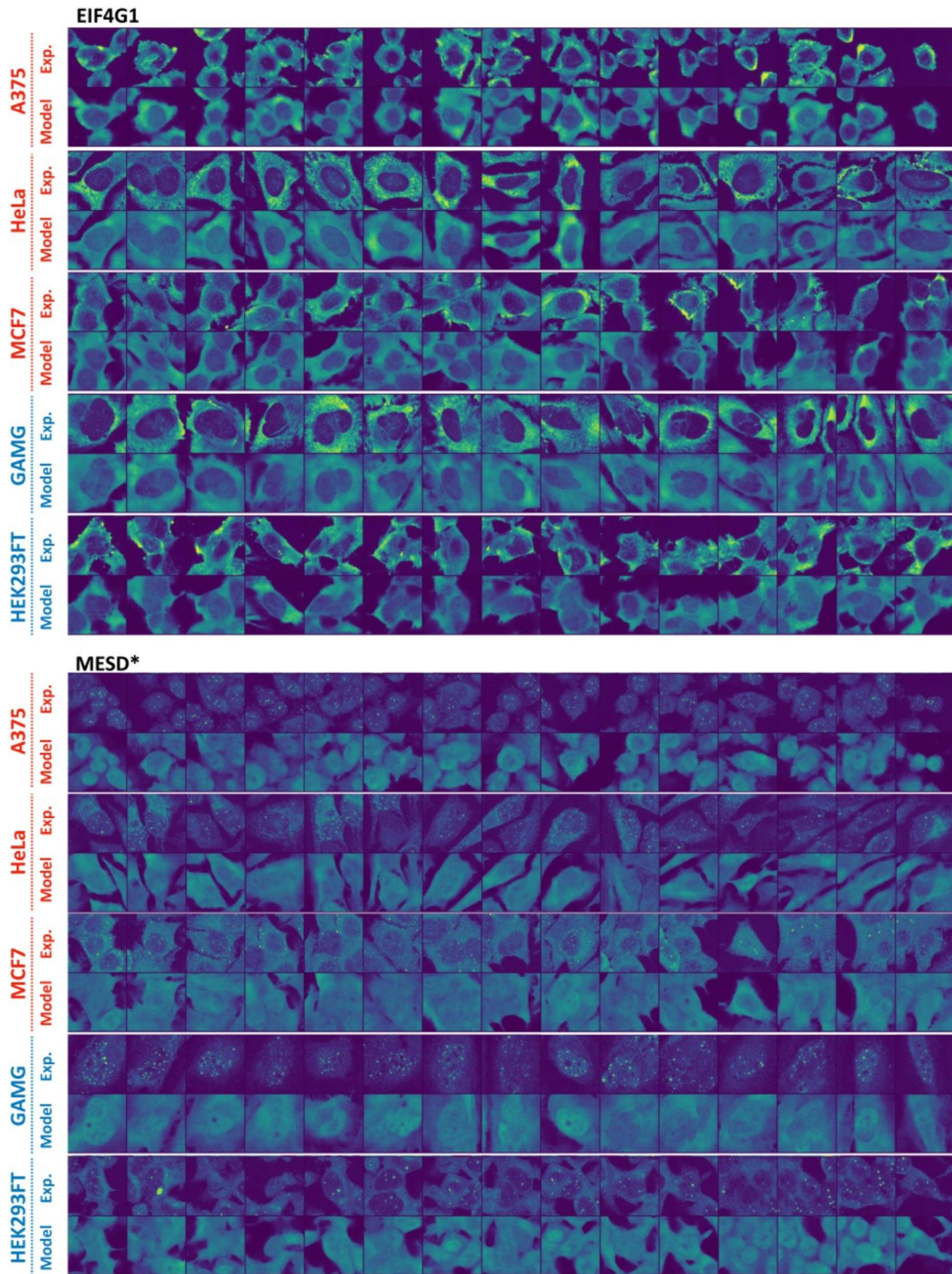
Supplementary Tables 1-4



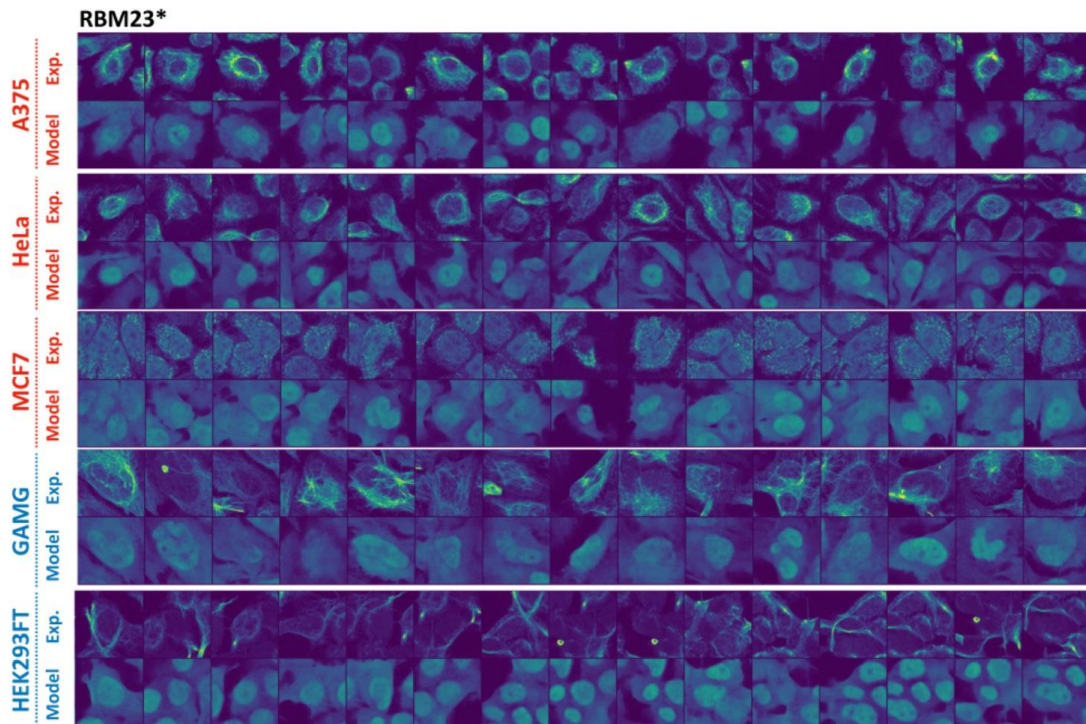
Supplementary Fig. 1. Experimental validation of our model on 5 cell lines stained for ATP13A5 and CHID1. 3 out of the 5 cell lines are not used in training the model (red). ATP13A5 and CHID1 were selected as the most variable proteins across cell lines (Methods). Each cell crop is 50µm x 50µm randomly selected from six 205.8µm x 205.8µm field of views.



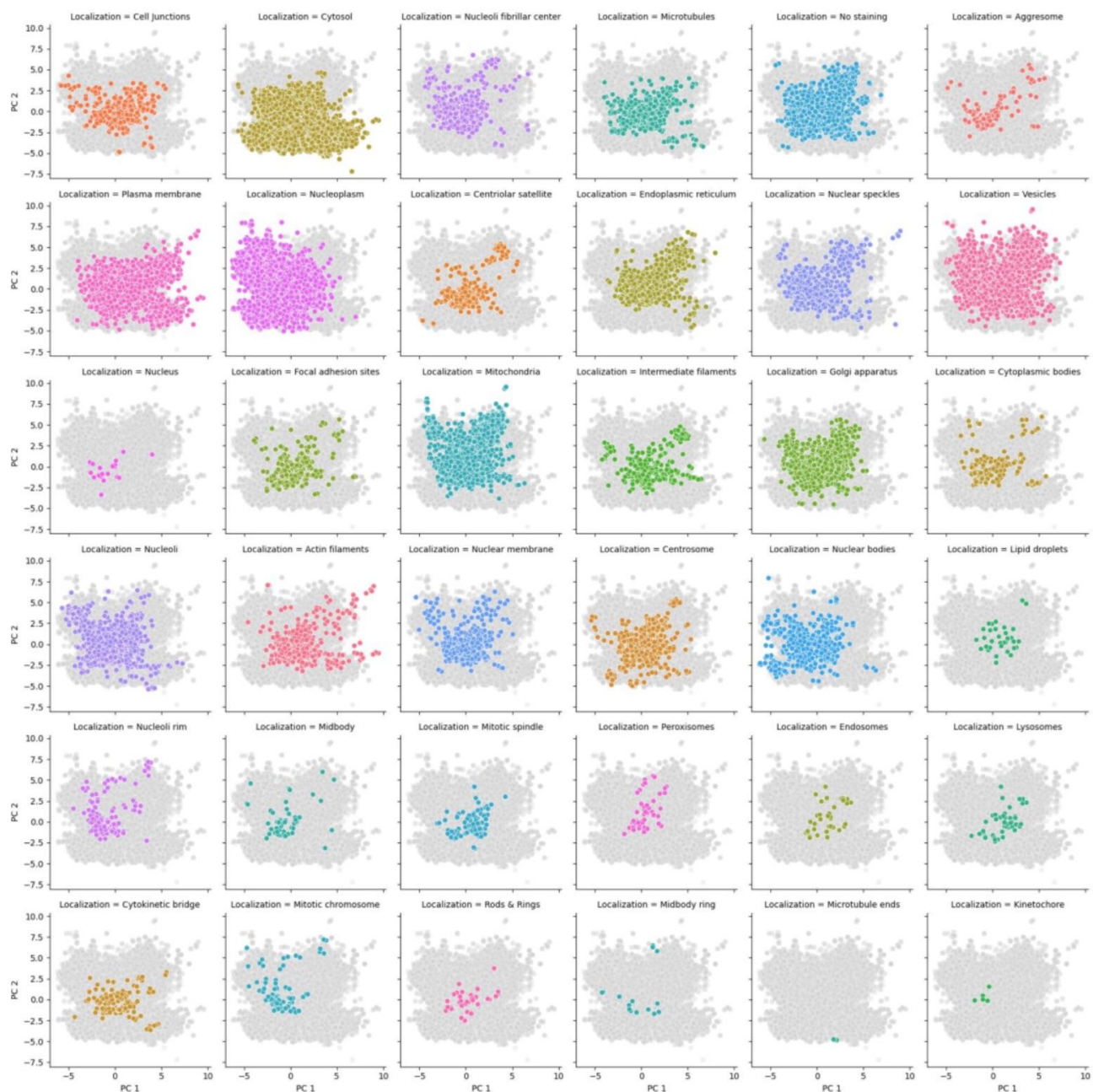
Supplementary Fig. 2. Experimental validation of our model on 5 cell lines stained for COPA and DDIT3. 3 out of the 5 cell lines are not used in training the model (red). *: DDIT3 is in the test set of Holdout 1. COPA was selected as one of the most variable proteins across cell lines (Methods). DDIT3 was selected as one of the most variable proteins across single cells of the same cell line (Methods). Each cell crop is 50 μ m x 50 μ m randomly selected from six 205.8 μ m x 205.8 μ m field of views.



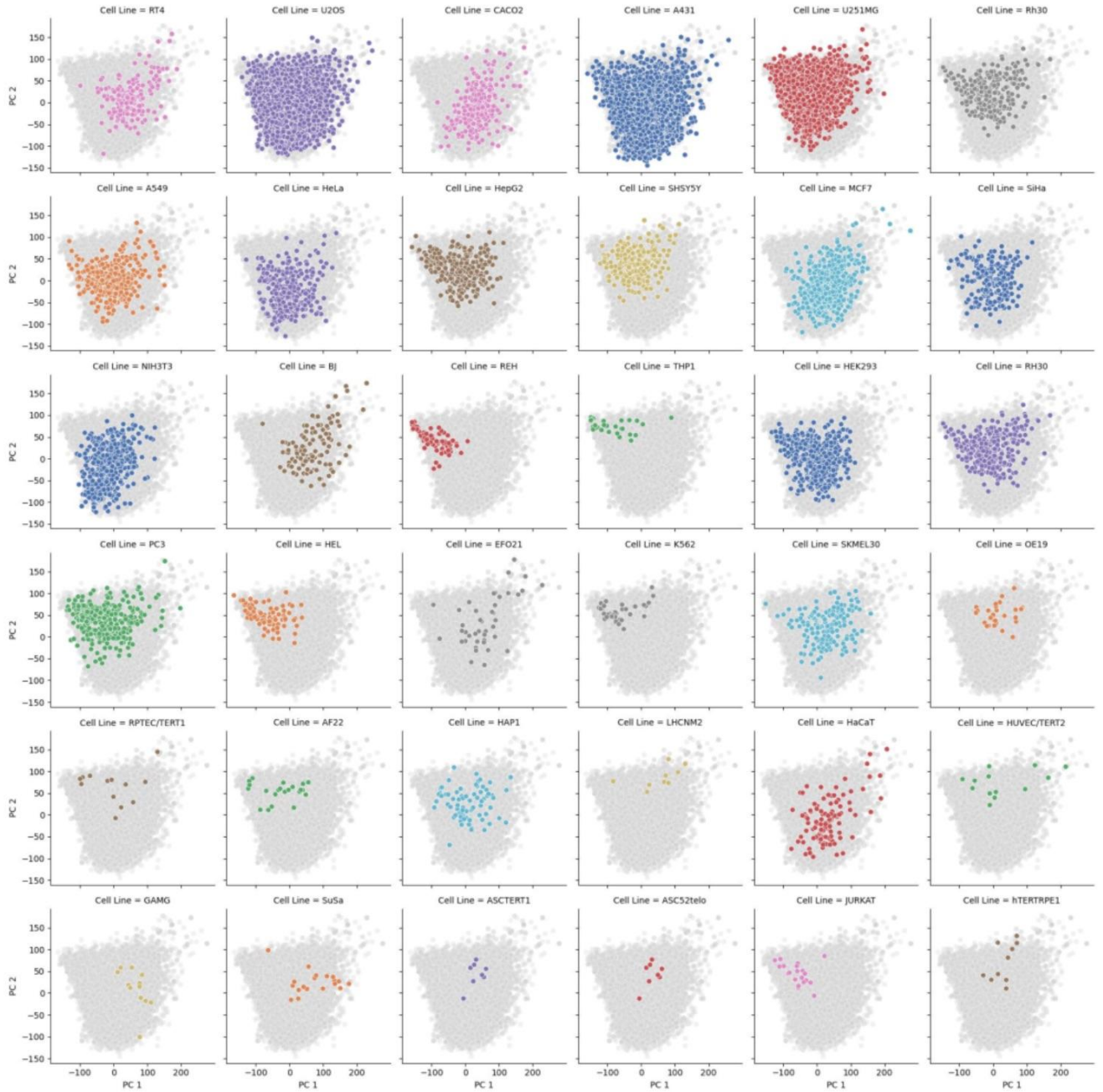
Supplementary Fig. 3. Experimental validation of our model on 5 cell lines stained for EIF4G1 and MESD. 3 out of the 5 cell lines are not used in training the model (red). *: MESD is in Holdout 2. MESD was selected as one of the most variable proteins across cell lines (Methods). EIF4G1 was selected as one of the least variable proteins across cell lines and is predicted to be localized mainly outside of the nucleus. Each cell crop is 50µm x 50µm randomly selected from six 205.8µm x 205.8µm field of views.



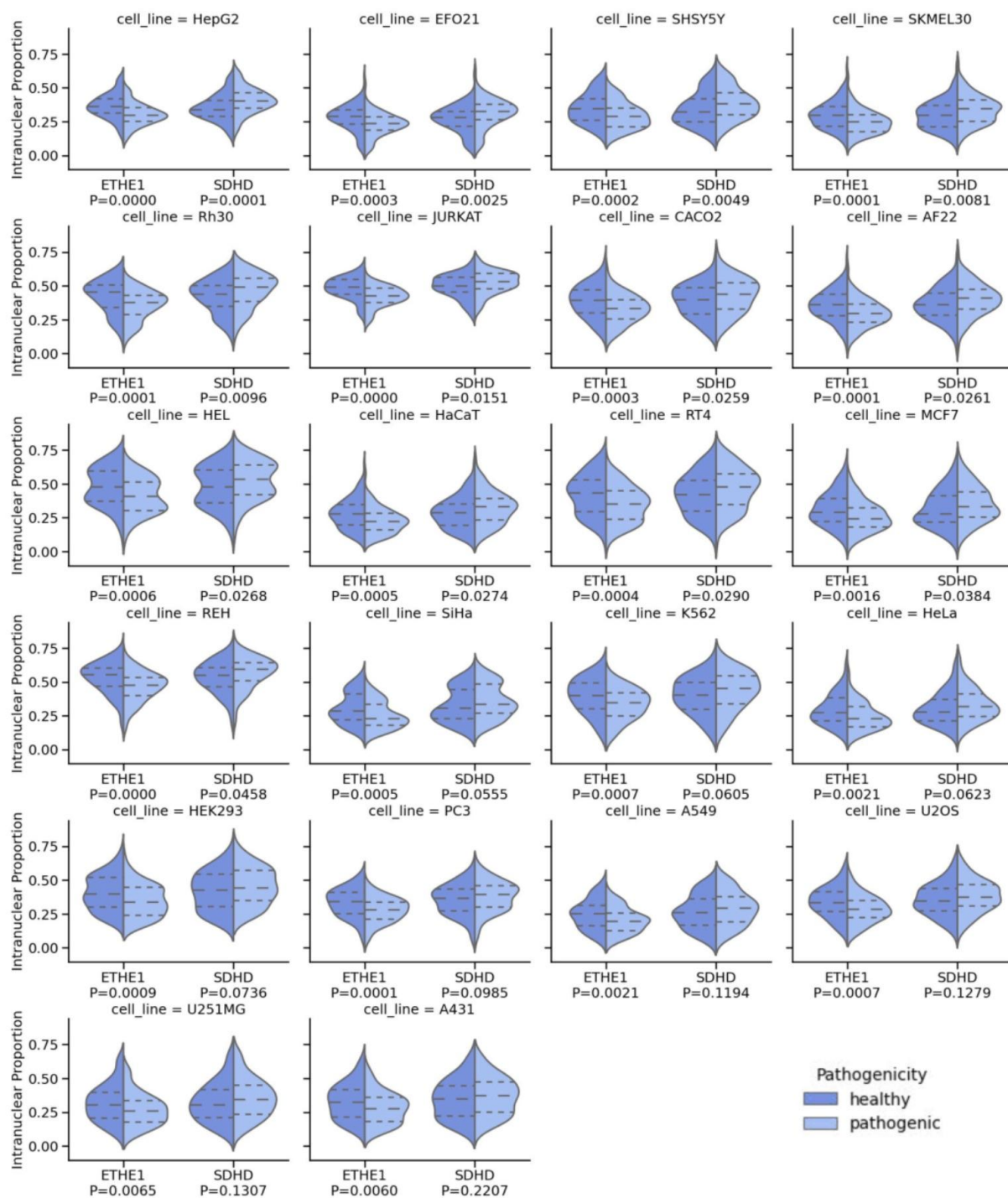
Supplementary Fig. 5. Experimental validation of our model on 5 cell lines stained for RBM23. 3 out of the 5 cell lines are not used in training the model (red). *: RBM23 is in Holdout 2. RBM23 was selected as one of the most variable proteins across cell lines. Each cell crop is 50 μ m x 50 μ m randomly selected from six 205.8 μ m x 205.8 μ m field of views.



Supplementary Fig. 6. PCA of protein sequence representation of 40,622 proteoforms corresponding to 12,614 genes. We used the same PCA embeddings of protein sequences as shown in Fig. 6a. Each panel shows the sequence representation of proteoforms annotated with a particular localization, while all other proteoforms are colored in gray.



Supplementary Fig. 7. PCA of 36 cell lines. We computed PCA of the image representations (Fig. 1b) of the landmark stains from 36 cell lines. Each panel shows the image representations of cells in a particular cell line, while all other cell lines are colored in gray.



Supplementary Fig. 8. Comparison of PUPS's predictions using wild type and pathogenic protein sequences. Violin plots showing the distribution of intranuclear proportions of predicted wild type and pathogenic SDHD and ETHE1 images in 22 cell lines. Dashed lines indicate the quartiles.

	protein_cell line	counts		protein_cell line	counts
0	AP1B1-202_u2os	66	25	NOL11-201_a431	24
1	C9orf40-201_u2os	34	26	OXR1-211_a431	40
2	CADPS2-205_mcf7	38	27	PDZD7-207_u2os	34
3	CYP20A1-203_caco2	32	28	PFDN5-202_u2os	24
4	DOP1A-202_u2os	32	29	PI4K2A-201_u2os	36
5	DSN1-207_u251mg	54	30	POLR3D-201_u2os	94
6	F11R-204_mcf7	70	31	PPP2R5C-203_u251mg	38
7	GLRX2-201_u2os	90	32	PRPF40A-202_hel	82
8	GMIP-204_a431	50	33	RHAG-205_k562	104
9	HLF-202_hepg2	66	34	RNGTT-201_u2os	36
10	HNF1B-204_u2os	28	35	SEPTIN9-220_u251mg	38
11	HPDL-201_mcf7	118	36	SERINC2-202_hacat	36
12	IKZF3-216_rt4	62	37	SGSM3-201_u2os	32
13	ING3-204_hel	84	38	SLC17A1-203_u2os	38
14	KEL-202_k562	158	39	SLC51B-201_caco2	108
15	KYNU-202_siha	30	40	SYBU-209_caco2	46
16	LARP7-209_u2os	58	41	THYN1-204_u2os	56
17	LARS2-201_u2os	28	42	TMEM176B-201_u2os	44
18	LIMD2-202_rh30	60	43	TNNI2-202_htcepi	30
19	MED17-227_efo21	42	44	TSEN2-202_a431	82
20	MEOX1-201_hela	52	45	VCAN-201_a431	28
21	MLLT3-210_hel	114	46	Z82206.1-201_u2os	30
22	MPDU1-218_a431	22	47	ZC3H12D-201_hepg2	68
23	N4BP2-201_u2os	72	48	ZNF510-203_rh30	46
24	NEXN-201_u2os	24	49	ZNF83-211_u251mg	44

Supplementary Table 1. Number of images used for each protein_cell line pair in Fig. 4c.

	antibody_cellline	counts		antibody_cellline	counts
0	A4GALT-201_mcf7	194	25	CTNNBIP1-201_shsy5y	44
1	AARS2-201_mcf7	98	26	CXorf56-204_pc3	24
2	ABCA4-201_mcf7	76	27	DDIT3-203_pc3	58
3	AC118549.1-202_pc3	82	28	DDX20-209_rh30	24
4	ANKRA2-202_rh30	70	29	DLL3-201_shsy5y	112
5	ANKRD17-201_rh30	76	30	DNAJB13-201_pc3	28
6	AP002495.1-206_rh30	112	31	DOK7-206_mcf7	48
7	ATOH8-201_rh30	76	32	DOP1B-202_mcf7	60
8	BCAS3-222_pc3	42	33	DOPEY2-202_mcf7	60
9	BCL7B-201_mcf7	18	34	DRGX-201_rh30	104
10	BIN2-205_hel	36	35	DSEL-201_rh30	44
11	BTBD10-201_mcf7	64	36	EGLN3-201_rh30	76
12	BTBD7-201_rh30	92	37	EMC6-202_pc3	52
13	C11orf71-201_mcf7	62	38	ESRRG-213_shsy5y	50
14	C16orf71-201_mcf7	54	39	EXOSC2-203_pc3	22
15	CABLES1-208_pc3	34	40	FAM133A-202_mcf7	212
16	CCDC124-202_pc3	44	41	FAM90A1-201_mcf7	104
17	CDCA4-201_pc3	22	42	FAM90A10P-201_mcf7	104
18	CDIPT-201_mcf7	44	43	FAM90A18P-201_mcf7	104
19	CDK10-201_mcf7	108	44	FAM90A7P-203_mcf7	104
20	CENPC-201_rh30	184	45	FBP2-201_mcf7	138
21	CENPI-201_rh30	84	46	FBXL8-201_mcf7	50
22	CENPT-208_mcf7	88	47	FOSB-204_bj[humanfibroblast]	32
23	CEP95-203_rh30	88	48	FPGT-TNNI3K-202_rh30	60
24	CETN1-201_pc3	40	49	FZD7-201_bj[humanfibroblast]	28

Supplementary Table 2. Number of images used for each protein_cell line pair in Extended Data Fig. 7b.

	antibody_cellline	counts		antibody_cellline	counts
0	ADCYAP1R1-203_shsy5y	118	25	PPP1R13L-209_a431	32
1	AMDHD2-219_a431	26	26	PRRG1-209_a431	26
2	AMMECR1L-213_a431	22	27	PTGIS-201_u2os	26
3	ARL13B-212_a431	32	28	PTPN20-213_a549	52
4	CACYBP-202_u2os	42	29	RC3H2-205_hela	22
5	DTNBP1-203_a431	30	30	SCARB2-223_a431	22
6	F11R-204_caco2	22	31	SEC14L4-201_a549	48
7	GAGE12G-201_a431	44	32	SH3GLB1-201_u2os	46
8	GHR-212_u2os	32	33	SLC27A1-201_u2os	26
9	GIMAP4-202_rh30	60	34	SLC37A4-201_a431	30
10	GSTM2-201_a549	54	35	SMAD3-201_a431	22
11	ITGB1BP2-205_u2os	62	36	SNAP47-221_u2os	32
12	KCNK7-203_hacat	24	37	SRGAP2-207_a431	24
13	LRRFIP2-205_a549	36	38	TIPARP-201_a431	42
14	LYPLAL1-201_a431	22	39	TMEM106B-202_u251mg	24
15	MDK-205_u2os	40	40	TMEM164-201_u251mg	22
16	MFF-201_u2os	32	41	TNFAIP8L3-201_u251mg	24
17	MON1B-209_a431	28	42	TOLLIP-203_u2os	28
18	NARFL-202_a431	28	43	TP53I11-203_mcf7	58
19	NECAP1-210_hela	30	44	TPPP-201_u2os	26
20	NEXN-201_bj	24	45	TRIM3-206_hela	26
21	NPEPPS-227_a549	32	46	TUBB2A-201_a549	24
22	OVGP1-201_u2os	28	47	TUBB4A-203_a549	24
23	POLG-227_siha	32	48	UGT8-201_hacat	28
24	PPM1A-207_a431	26	49	ZNF213-208_u251mg	32

Supplementary Table 3. Number of images used for each protein_cell line pair in Extended Data Fig. 7c.

	Name	Clone	Vendor	Catalog #	Host	Stock	Final concentration
1st Ab	a-Tubulin	DM1A	Abcam	ab7291	Mouse	1.0 mg/mL	1.0 µg/mL
1st Ab	calreticulin	poly	Abcam	ab2908	Chicken	1.0 mg/mL	1.25 µg/mL
1st Ab	EIF4G1	HPA028487	Atlas Antibodies		Rabbit	0.1 mg/mL	2.0 µg/mL
1st Ab	DDIT3	HPA058416	Atlas Antibodies		Rabbit	0.3 mg/mL	2.0 µg/mL
1st Ab	N4BP2	HPA042607	Atlas Antibodies		Rabbit	0.2 mg/mL	2.0 µg/mL
1st Ab	CHID1	HPA039374	Atlas Antibodies		Rabbit	0.05 mg/mL	2.0 µg/mL
1st Ab	COPA	HPA028024	Atlas Antibodies		Rabbit	0.1 mg/mL	2.0 µg/mL
1st Ab	ATP13A5	HPA031772	Atlas Antibodies		Rabbit	0.3 mg/mL	2.0 µg/mL
1st Ab	MESD (MESDC2)	HPA039414	Atlas Antibodies		Rabbit	0.1 mg/mL	2.0 µg/mL
1st Ab	RBM23	HPA004144	Atlas Antibodies		Rabbit	0.1 mg/mL	2.0 µg/mL
1st Ab	PSME3IP1 (FAM192A)	HPA054382	Atlas Antibodies		Rabbit	0.1 mg/mL	2.0 µg/mL
2nd Ab	Anti-rabbit Alexa488	Poly	Thermo Fisher Scientific	A11034	Goat	2.0 mg/mL	2.5 µg/mL
2nd Ab	Anti-mouse Alexa555	Poly	Thermo Fisher Scientific	A21424	Goat	2.0 mg/mL	2.5 µg/mL
2nd Ab	Anti-chicken Alexa647	Poly	Thermo Fisher Scientific	A21449	Goat	2.0 mg/mL	2.5 µg/mL
DAPI	DAPI	N/A	Sigma-Aldrich	F0895	N/A	0.4 µg/mL	0.2 µg/mL

Supplementary Table 4. Antibodies used for experimental validation.