UCLouvain

Where and with whom: Using Bayesian inference and deep learning to study protein localisation and protein-protein interactions

Laurent Gatto

21 December 2020

Abstract

In biology, localisation of and interaction among proteins define their functions and activity. These information can be assayed experimentally and extracted from public databases. In this talk, I will present two use-cases, using Bayesian inference and deep learning, to infer protein sub-cellular localisation from experimental data and publicly available annotation.

Acknowledgements

- Bayesian spatial proteomics: Dr Oliver Crook (U Cambridge, now U Oxford). Funded by the Wellcome Trust.
- Deep learning, data integration: Aayush Grove (International Institute of Information Technology, Bangalore and CBIO, UCLouvain).

Outline

Scientific question: where, with whom

Experimental and annotation data

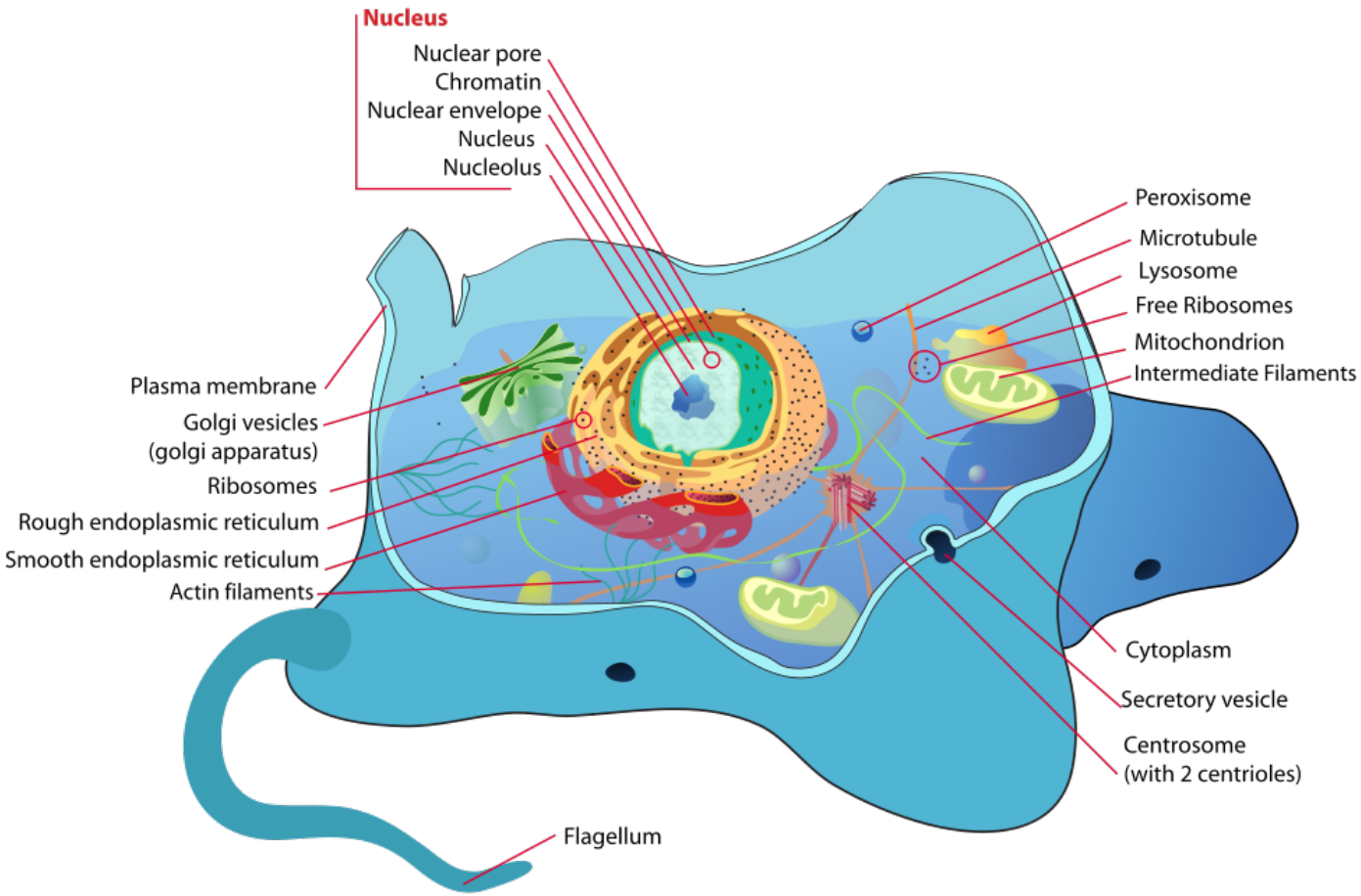
Data analysis (overview)

Bayesian inference

Deep learning

Conclusions

Cell organisation - **localisation is function**



Spatial proteomics is the systematic study of protein localisations.

Localisation – interactions – re-localisation – mis-localisation

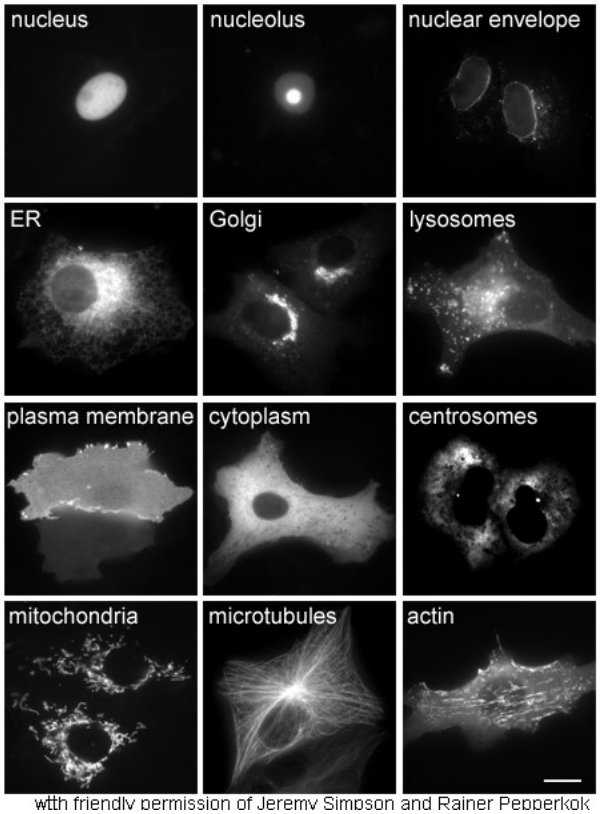
Image from Wikipedia `http://en.wikipedia.org/wiki/Cell_(biology)`.

Outline

Experimental data

- On protein localisation
  - targetted microscopy-based
  - global protein localisation map
- On protein-protein interactions (PPI)

Fusion proteins and immunofluorescence



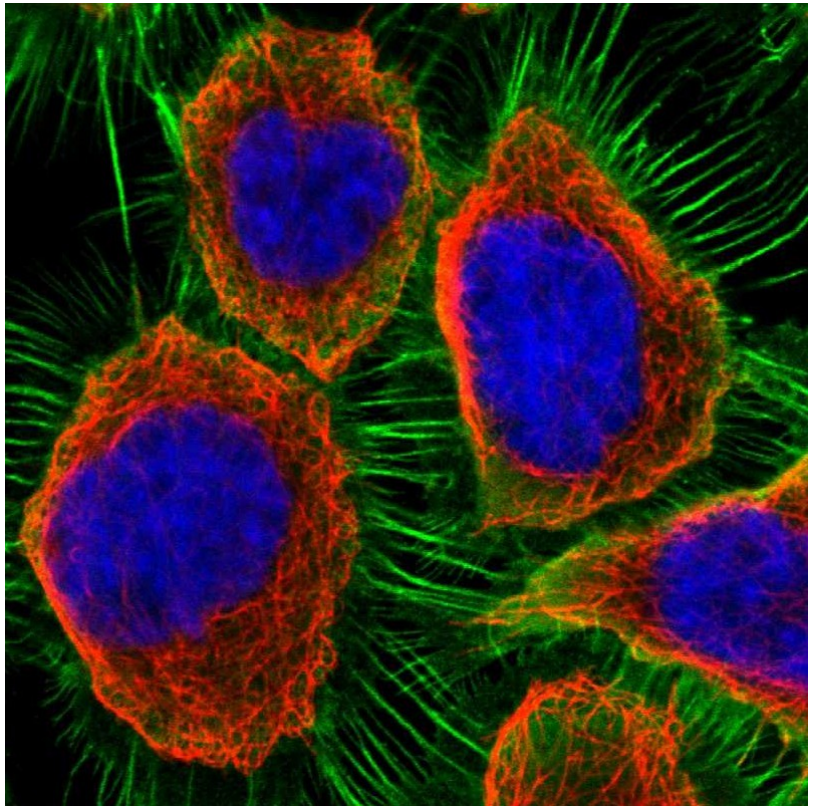
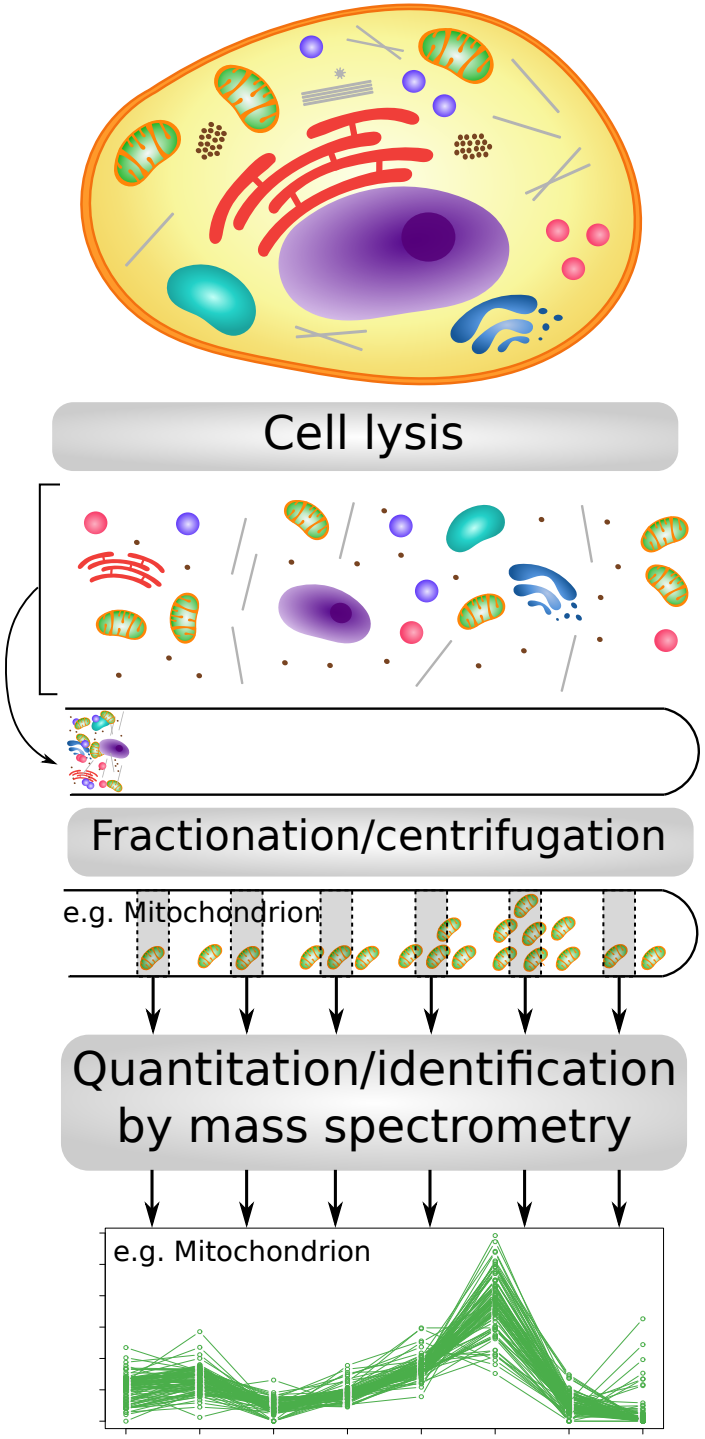


Figure: Targeted protein localisation. Example of discrepancies between IF and FPs as well as between FP tagging at the N and C termini (Stadler et al., 2013).

Explorative/discovery approaches, steady-state **global localisation maps** (as opposed to targeted microscopy-based approaches).

Density gradient: PCP (Dunkley et al., 2006), LOPIT (Foster et al., 2006), hyperLOPIT (Christoforou et al., 2016; Mulvey et al., 2017) and **Differential centrifugation** Itzhak et al. (2016), LOPIT-DC (Geladaki et al., 2019).



Quantitation data

	Fraction_1	Fraction_2	$\dots$	Fraction_L
$\mathbf{x}_1$	$x_{1,1}$	$x_{1,2}$	$\dots$	$x_{1,L}$
$\mathbf{x}_2$	$x_{2,1}$	$x_{2,2}$	$\dots$	$x_{2,L}$
$\mathbf{x}_3$	$x_{3,1}$	$x_{3,2}$	$\dots$	$x_{3,L}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_i$	$x_{i,1}$	$x_{i,2}$	$\dots$	$x_{i,L}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_N$	$x_{N,1}$	$x_{N,2}$	$\dots$	$x_{N,\ L}$

Quantitation data and organelle markers

	Fraction_1	Fraction_2	$\dots$	Fraction_L	markers
$\mathbf{x}_1$	$x_{1,1}$	$x_{1,2}$	$\dots$	$x_{1,L}$	unknown
$\mathbf{x}_2$	$x_{2,1}$	$x_{2,2}$	$\dots$	$x_{2,L}$	loc_1
$\mathbf{x}_3$	$x_{3,1}$	$x_{3,2}$	$\dots$	$x_{3,L}$	unknown
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_i$	$x_{i,1}$	$x_{i,2}$	$\dots$	$x_{i,L}$	loc_k
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_N$	$x_{N,1}$	$x_{N,2}$	$\dots$	$x_{N,\ K}$	unknown

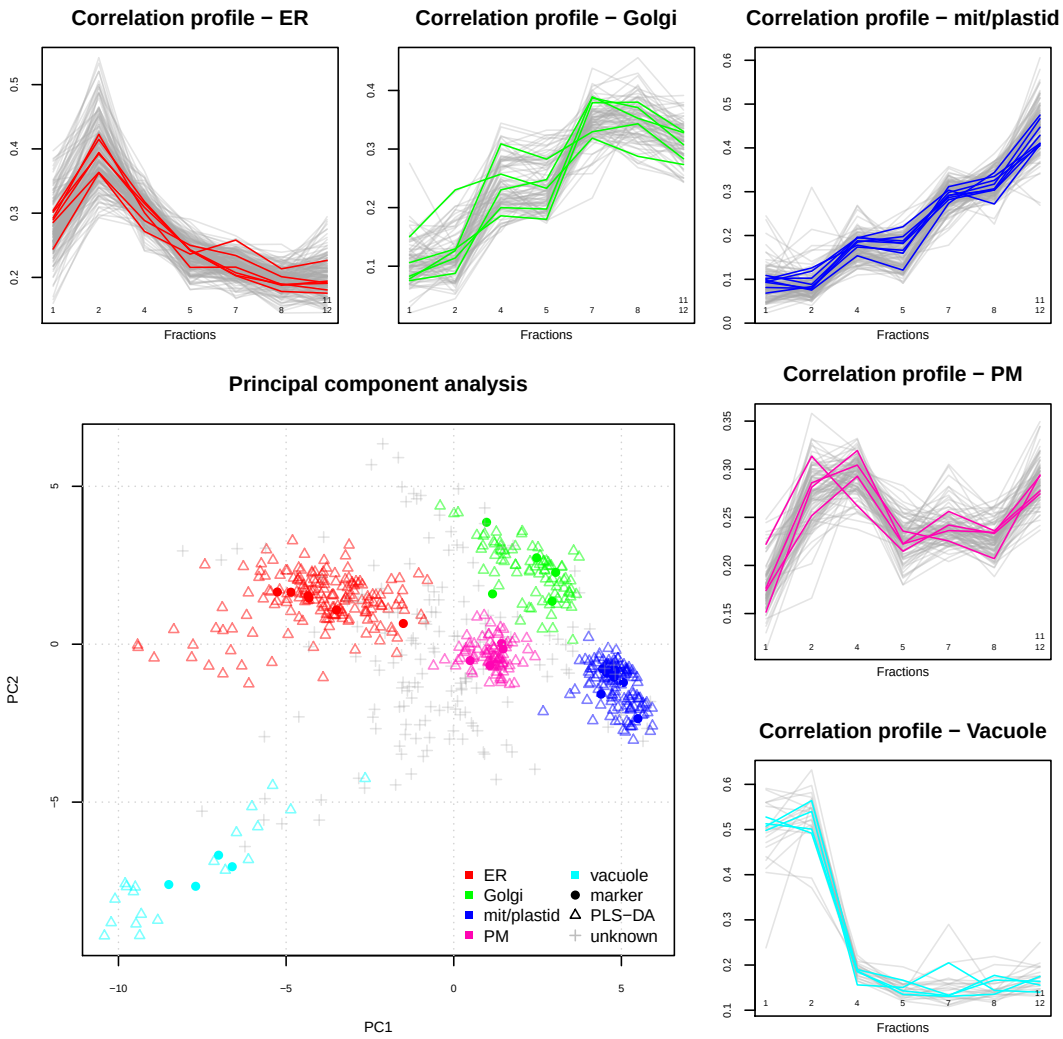
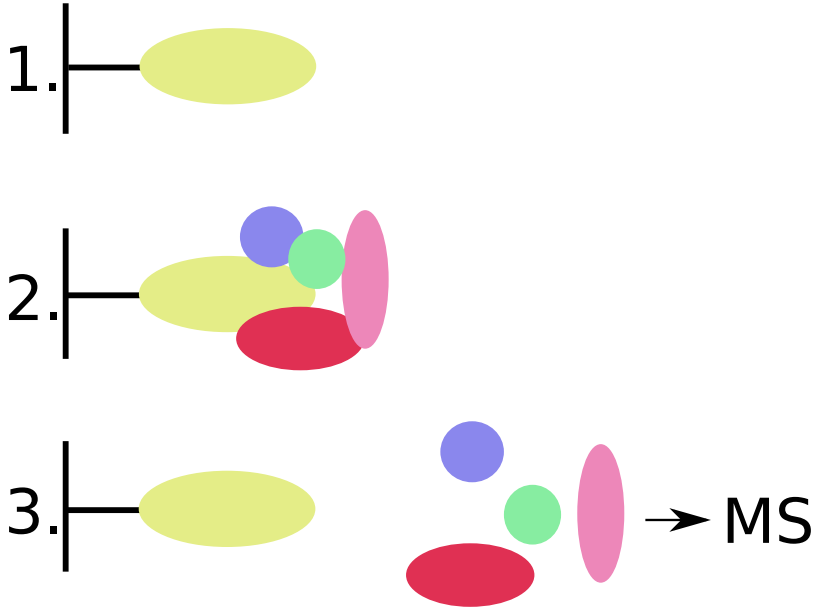


Figure: From Gatto et al. (2010), *Arabidopsis thaliana* data from Dunkley et al. (2006)

Protein-protein interactions

Affinity purification mass spectometry (AP-MS). The bait protein (yellow) is immobilised on a matrix (1). A protein mixture is passed through and only the interacting partners (prey) are retained (2). Then the prey proteins are eluded, digested and analysed by mass spectrometry (3).



We have used the Bioplex (Huttlin et al., 2020) data (`https://bioplex.hms.harvard.edu/`).

Annotation data

Extracted from public repositories:

- **Gene Ontology** (GO) knowledgebase is the world's largest source of information on the functions of genes. (Ashburner et al., 2000). Terms (biological process, molecular signature, **cellular compartment**) represented as directed acyclic graphs (DAG).
- **Human Proteome Atlas** (HPA) maps *all* the human proteins in cells, tissues and organs using an integration of various omics technologies, including **antibody-based imaging**, mass spectrometry-based proteomics, transcriptomics and systems biology. (Uhlén et al., 2005; Uhlen et al., 2010) The **Cell Atlas** provides high-resolution insights into the expression and spatio-temporal distribution of RNA and proteins in human cell lines.

Annotation data

- **STRING** is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. (Szklarczyk et al., 2018)

Experimental vs annotation

Experimental data $\neq$ annotation data

Specific $\neq$ generic

(Expected) high quality $\neq$ unknown quality

Expensive to produce $\neq$ free to consome

Generally small $\neq$ big data

Outline

Computational challenges

- Visualisation (unsupervised learning) (Gatto et al., 2014, 2019)
- Classification (Gatto et al., 2014)
- Novelty detection (semi-supervised learning) (Breckels et al., 2013; Crook et al., 2020)
- **Uncertainty quantification** (Crook et al., 2018)
- **Multi-localisation** (Crook et al., 2018)
- Spatial dynamics
- Data integration (transfer learning) (Breckels et al., 2016)
- **Deep learning**

To uncover and understand biology

Outline

Supervised Machine Learning

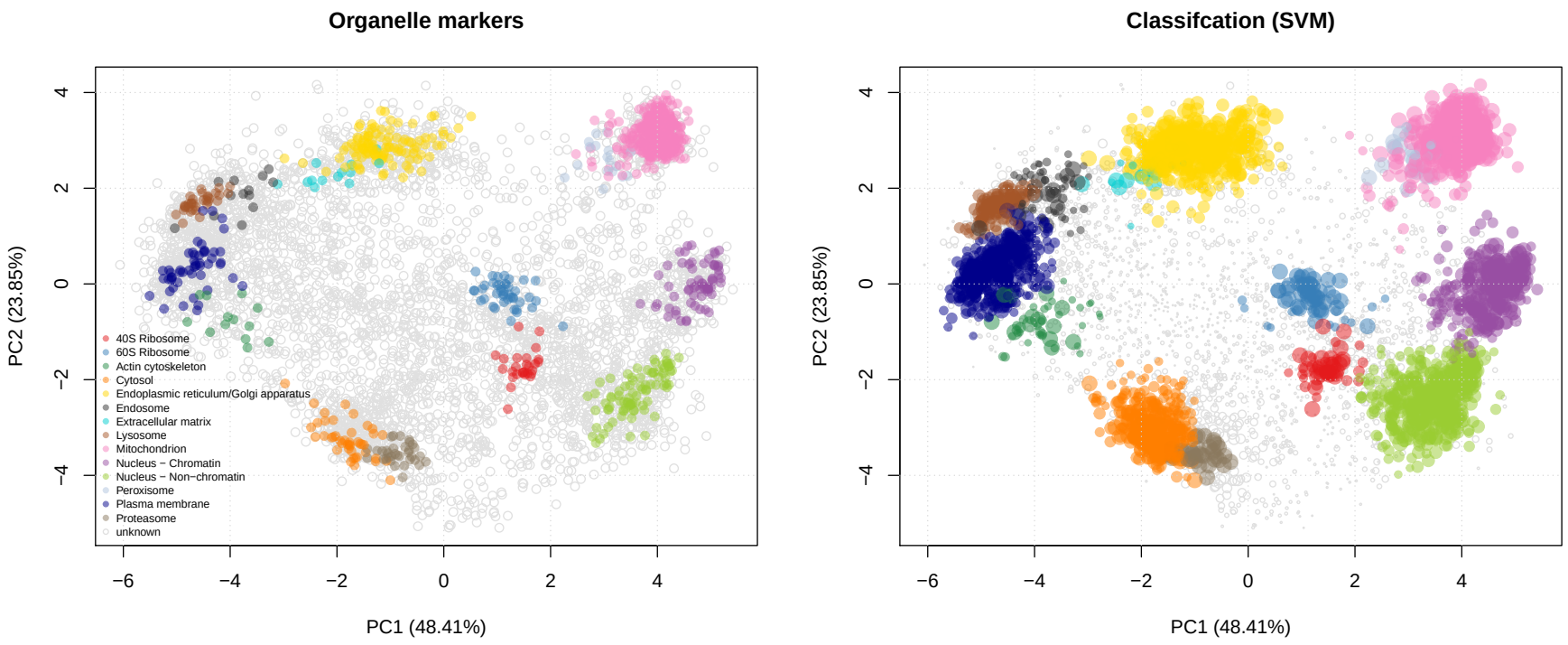
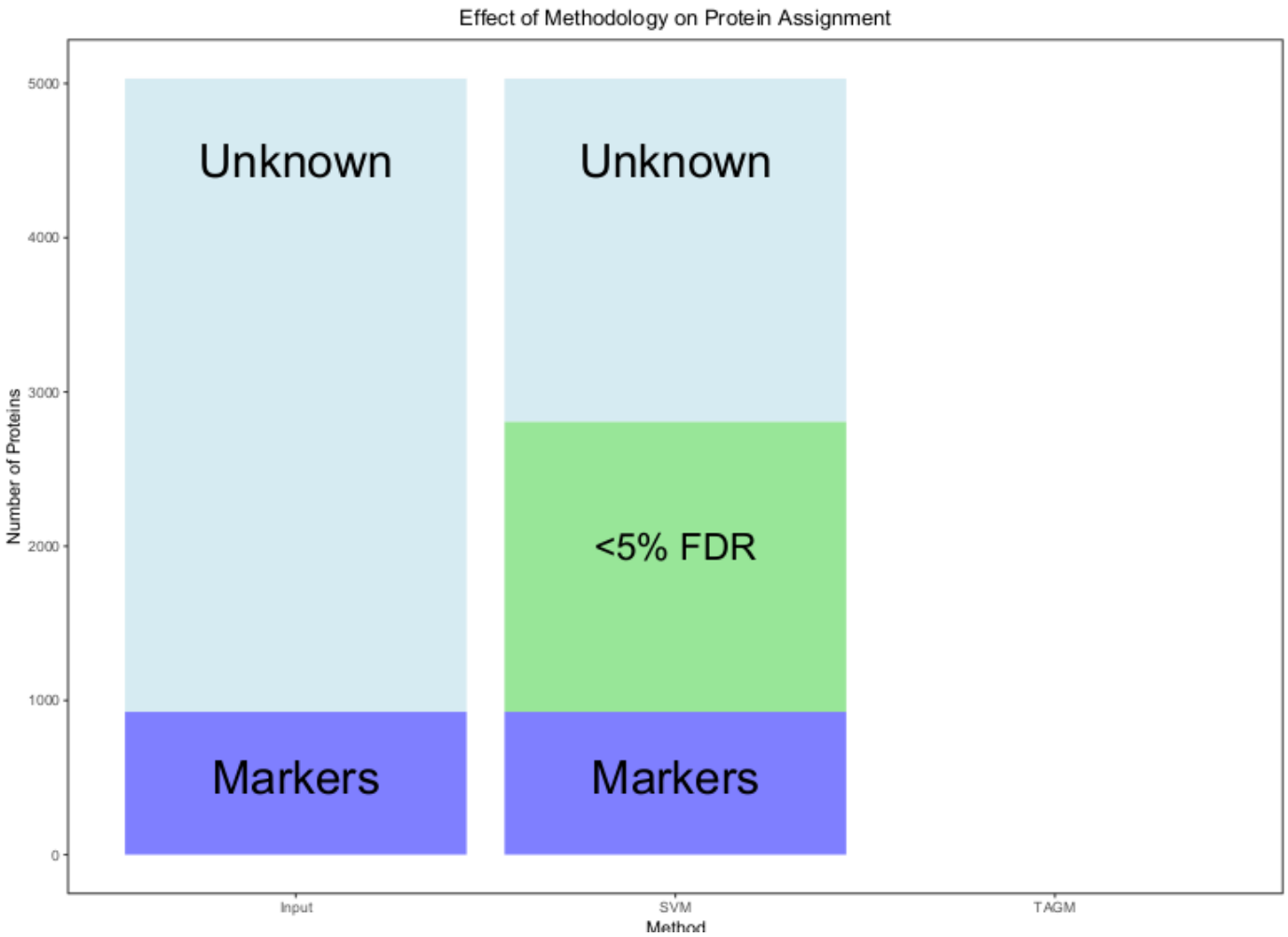


Figure: Support vector machines classifier (after 5% FDR classification cutoff) on the embryonic stem cell data from Christoforou et al. (2016).

How much do we learn? How much do we miss?



Bayesian Mixture Modelling For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.

Bayesian Mixture Modelling For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.
- With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an **outlier component**. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a *T Augmented Gaussian Mixture model* (Crook et al., 2018, 2019).

Bayesian Mixture Modelling For Spatial Proteomics

- *T Augmented Gaussian Mixture model (TAGM)* is a **multivariate Gaussian generative model** for MS-based spatial proteomics data. It posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution.
- With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an **outlier component**. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a *T Augmented Gaussian Mixture model* (Crook et al., 2018, 2019).
- This methodology allows proteome-wide **uncertainty quantification**, thus adding a further layer to the analysis of spatial proteomics.

$$\mathbf{x}_i | z_i = k \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

$$\mathbf{x}_i | z_i = k \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k) \tag{1}$$

$$\mathbf{x}_i | z_i = k, \phi_i \quad \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)^{\phi_i} \mathcal{T}(\kappa, \boldsymbol{M}, V)^{1-\phi_i} \tag{2}$$

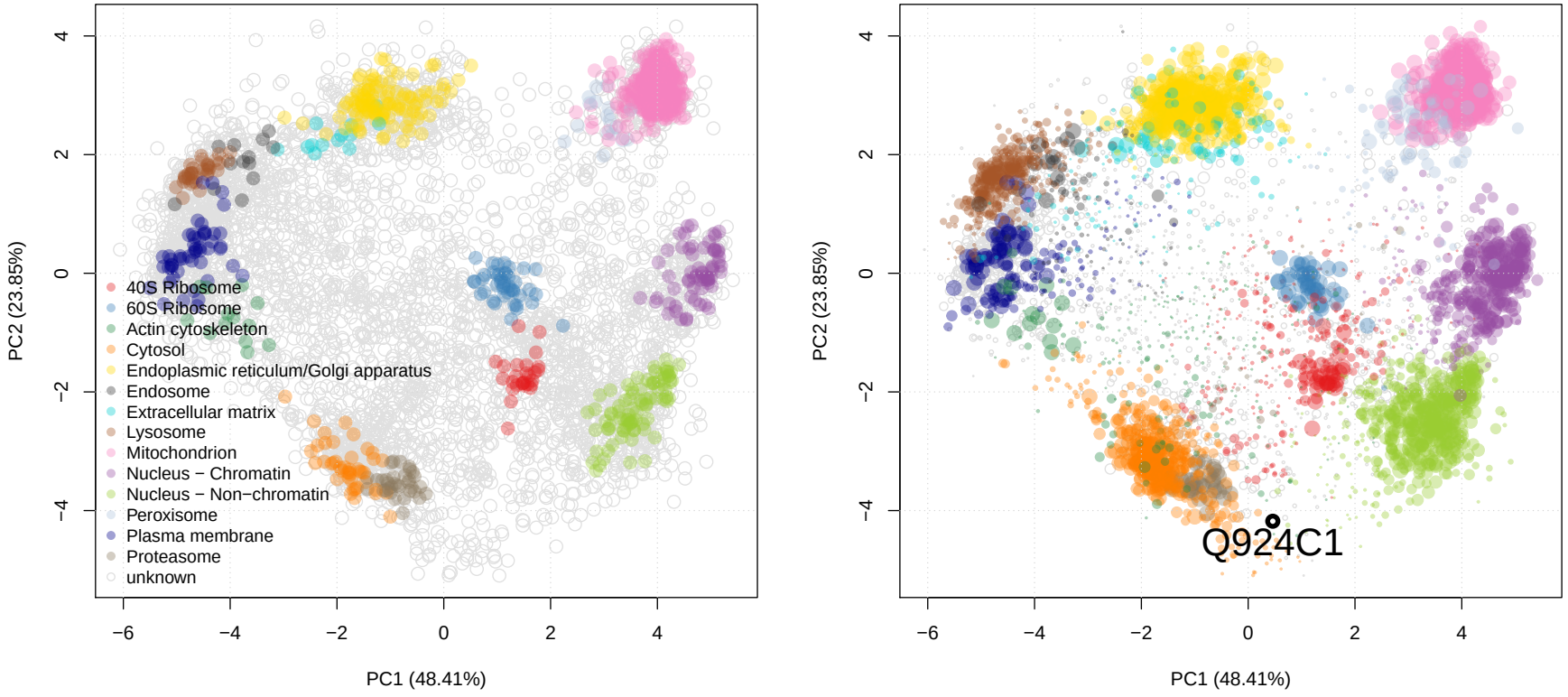
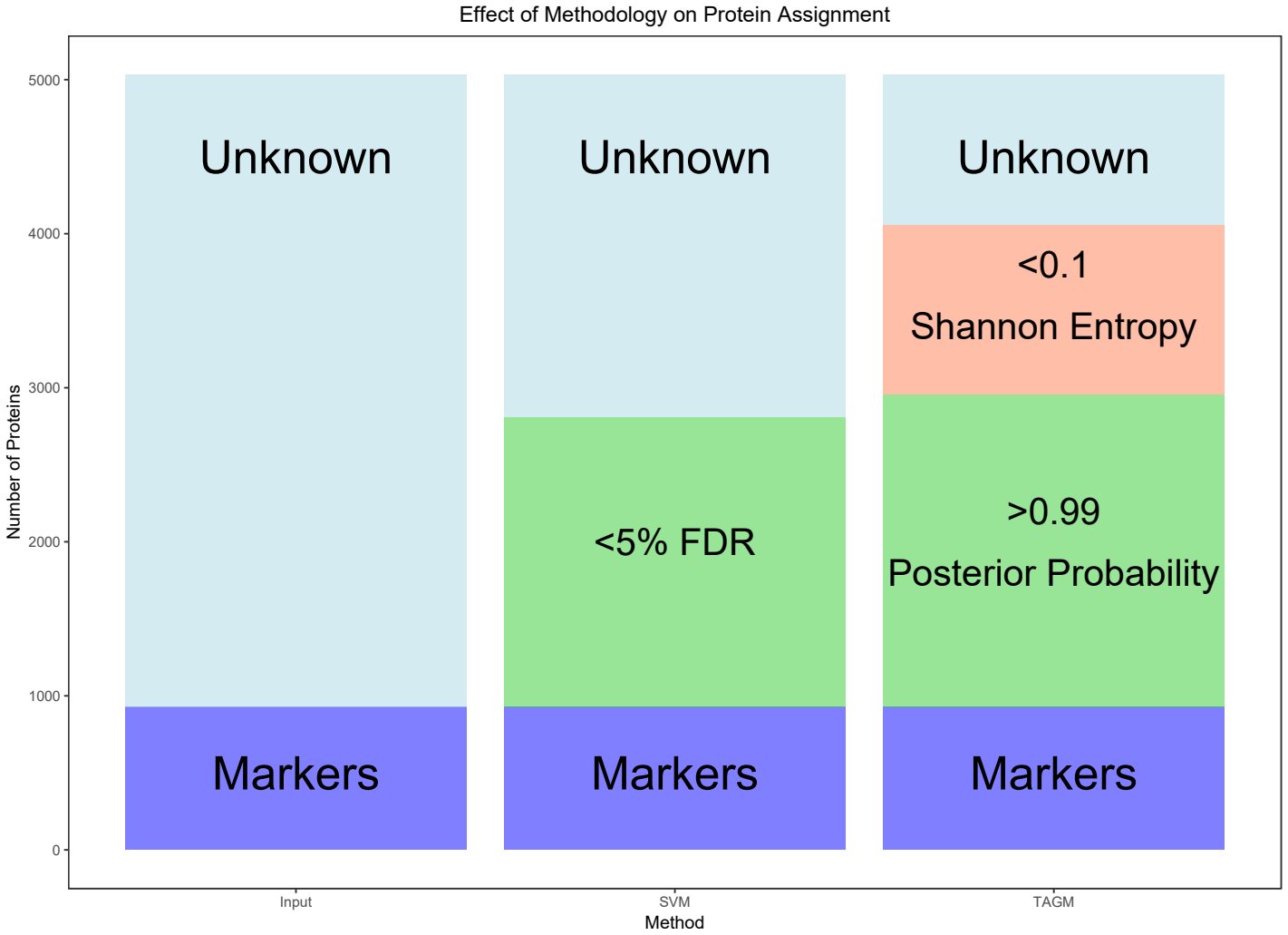


Figure: Assignment of proteins of *unknown* location to one of the annotated classes. The dots are scaled according to the protein assignment probabilities.


Effect of Methodology on Protein Assignment
Number of Proteins
5000
4000
3000
2000
1000
0
Unknown
Unknown
Unknown
<0.1
Shannon Entropy
<5% FDR
>0.99
Posterior Probability
Markers
Markers
Markers
Input
SVM
TAGM
Method

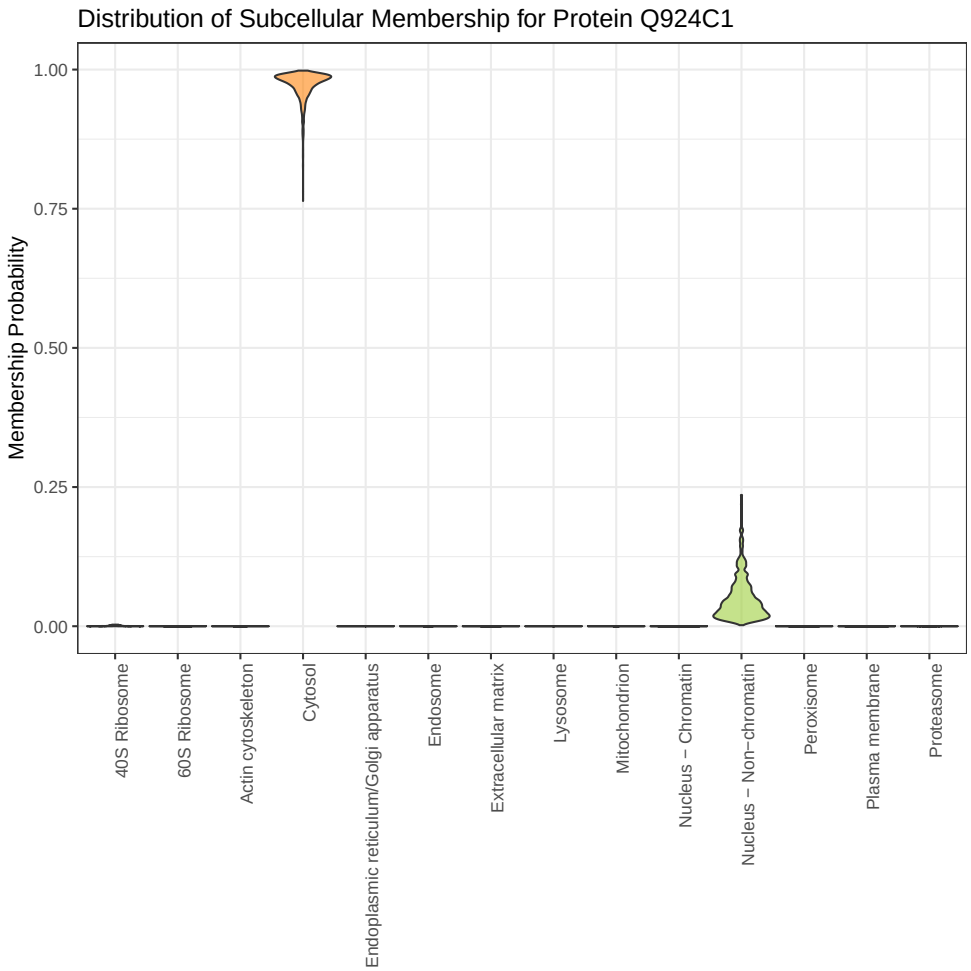
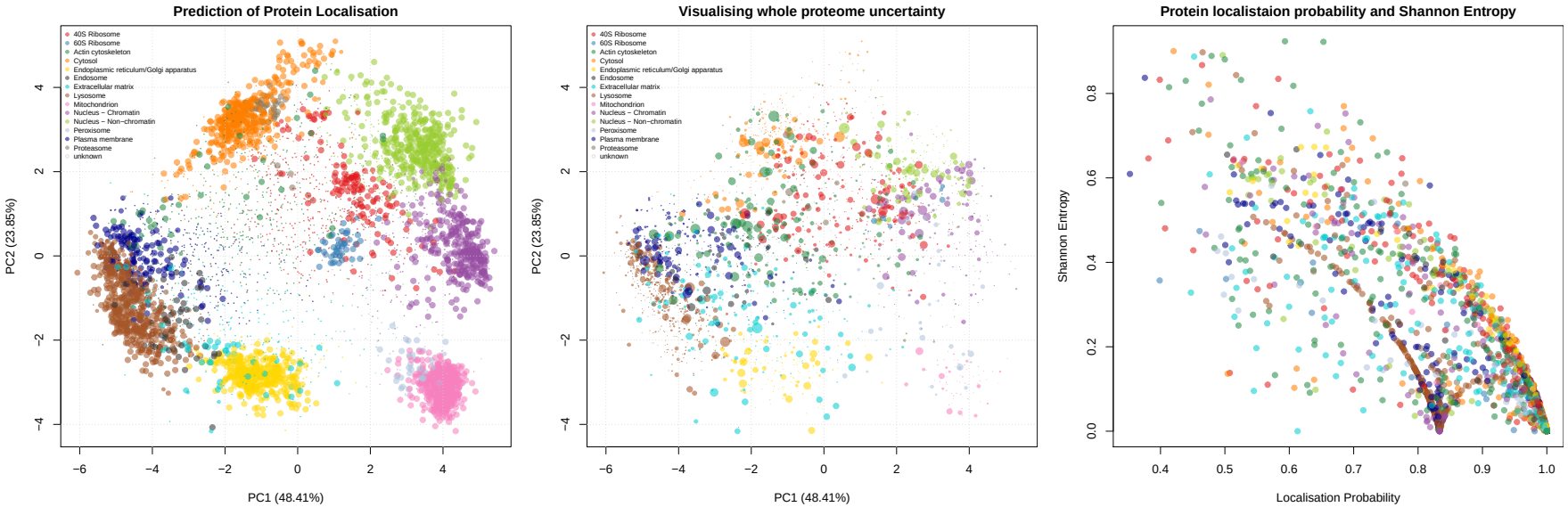


Figure: Exportin 5 (Q924C1) forms part of the micro-RNA export machinery, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus to mediate further transport between nucleus and cytoplasm. The model correctly infers that it most likely localises to the cytosol but there is some uncertainty with this assignment. This uncertainty is reflected in possible assignment of Exportin 5 to the nucleus non-chromatin and reflects the multi-location of the protein.

Whole sub-cellular proteome uncertainty



Outline

Data

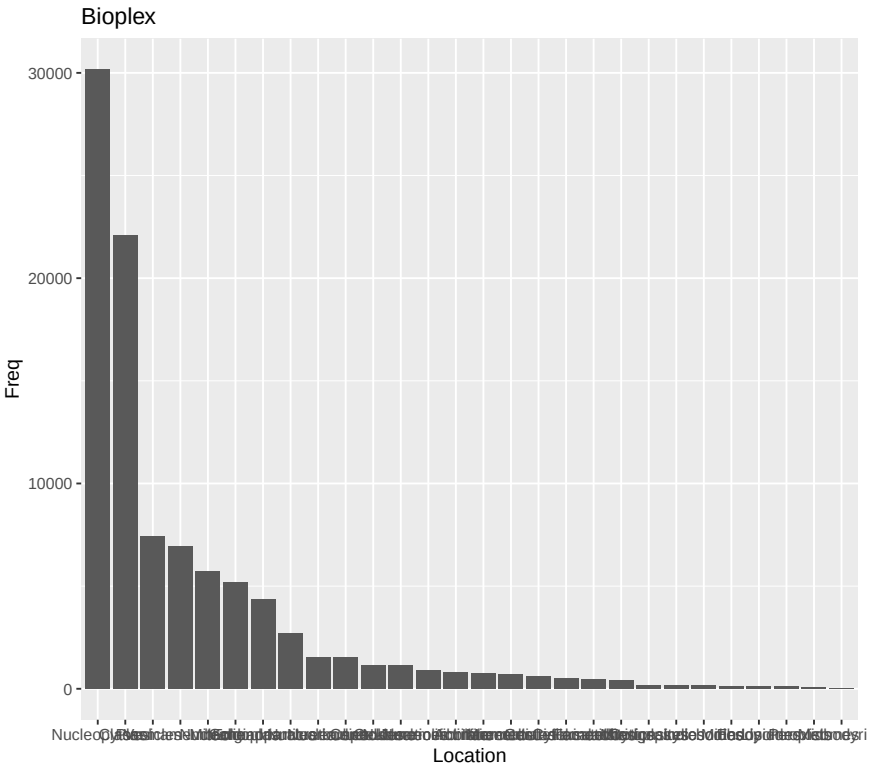
- STRING database (PPI, annotation): 3,121,320 pairwise interactions.
- Bioplex (PPI, large experimental resource): 64,861 pairwise interactions.
- Human protein atlas (localisation, large experimental resource): assigned to 28 reliable locations.

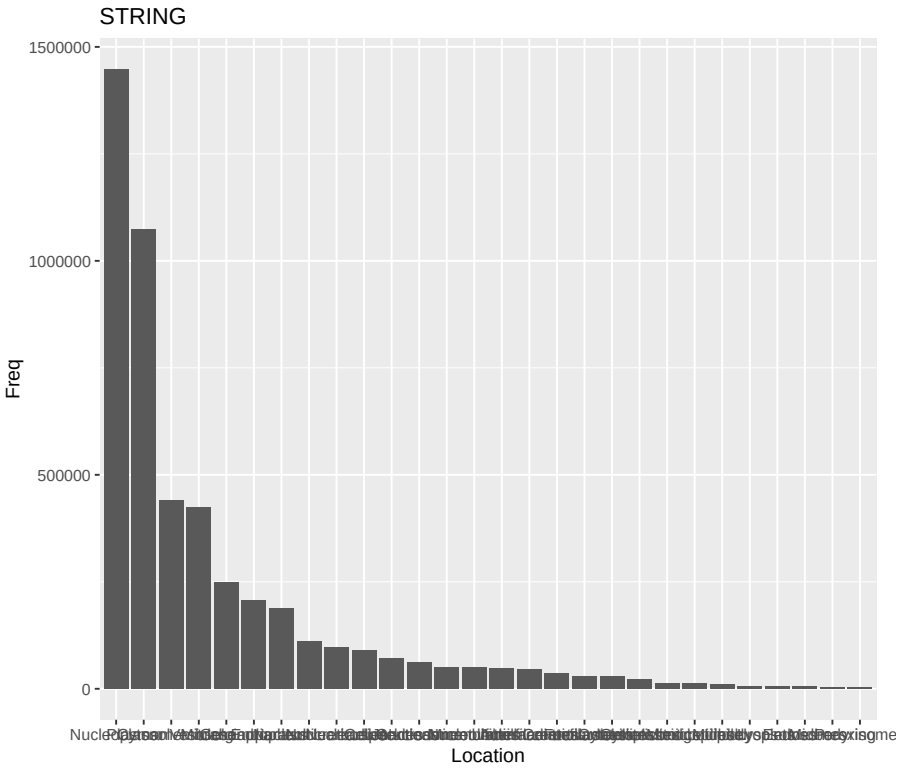
Data cleaning (1)

	source	protein1	protein2
1	string	ENSP00000310301	ENSP00000388940
2	string	ENSP00000379658	ENSP00000394560
3	string	ENSP00000360112	ENSP00000360312
	...		
4	bioplex	ENSP00000294889	ENSP00000304586
5	bioplex	ENSP00000261531	ENSP00000252011

	location1	location2
1		
2	{'Nucleoli', 'Cytosol'}	
3	{'Nucleoplasm'}	
	...	
4		{'Nucleoplasm', 'Cytosol'}
5	{'Nucleoplasm', 'Cytosol'}	{'Nucleoplasm', 'Nuclear bodies', 'Centrosome'}

	combined_score	locations ($loc_1 \cap loc_2$)	
1	0.23		new finding
2	0.21	{'Nucleoli', 'Cytosol'}	filtering
3	0.25	{'Nucleoplasm'}	filtering
	...		
4	1.00	{'Nucleoplasm', 'Cytosol'}	filtering
5	1.00	{'Nucleoplasm'}	always included

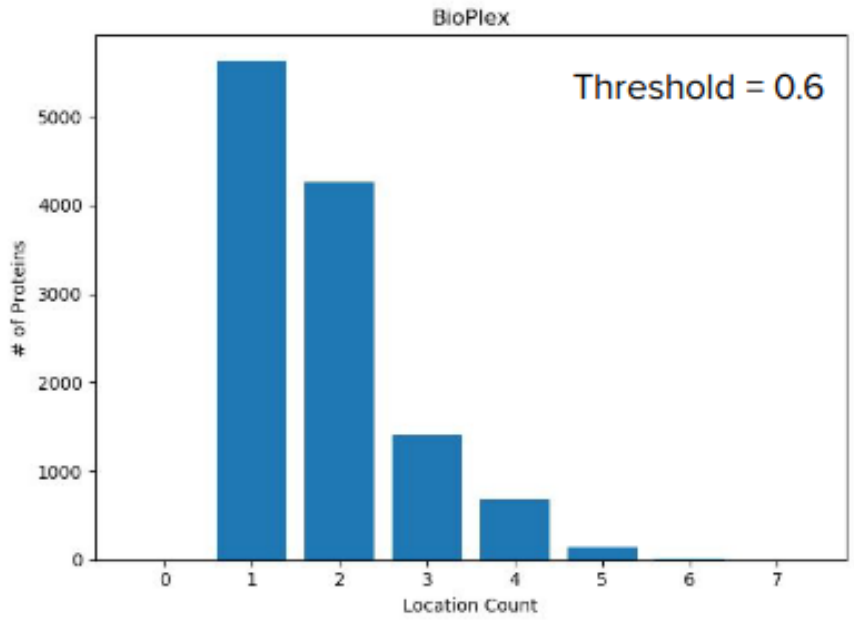

Bioplex
Freq
30000
20000
10000
0
Location

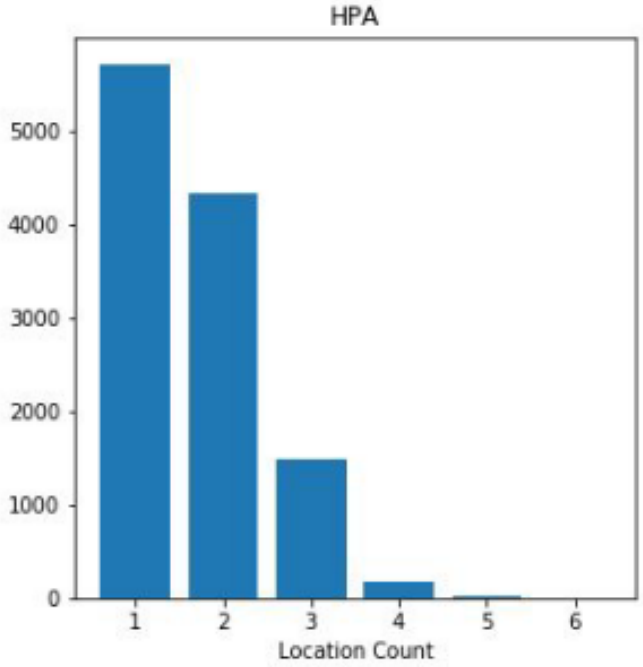

STRING
Freq
1500000
1000000
500000
0
Location

Data cleaning (2)

Calculate the normalized frequency count and retain the annotations up to the cumulative threshold.

- Location observations: p_i: (loc_1, loc_2),(loc_1) and (loc_1, loc_2, loc_3).
- Location frequencies: loc_1: 0.5, loc_2: 0.333 and loc_3: 0.167.
- Retained locations for a threshold of 0.85: loc_1 and loc_2


BioPlex
Threshold = 0.6
# of Proteins
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
Location Count


HPA
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
Location Count

Multilabel classification

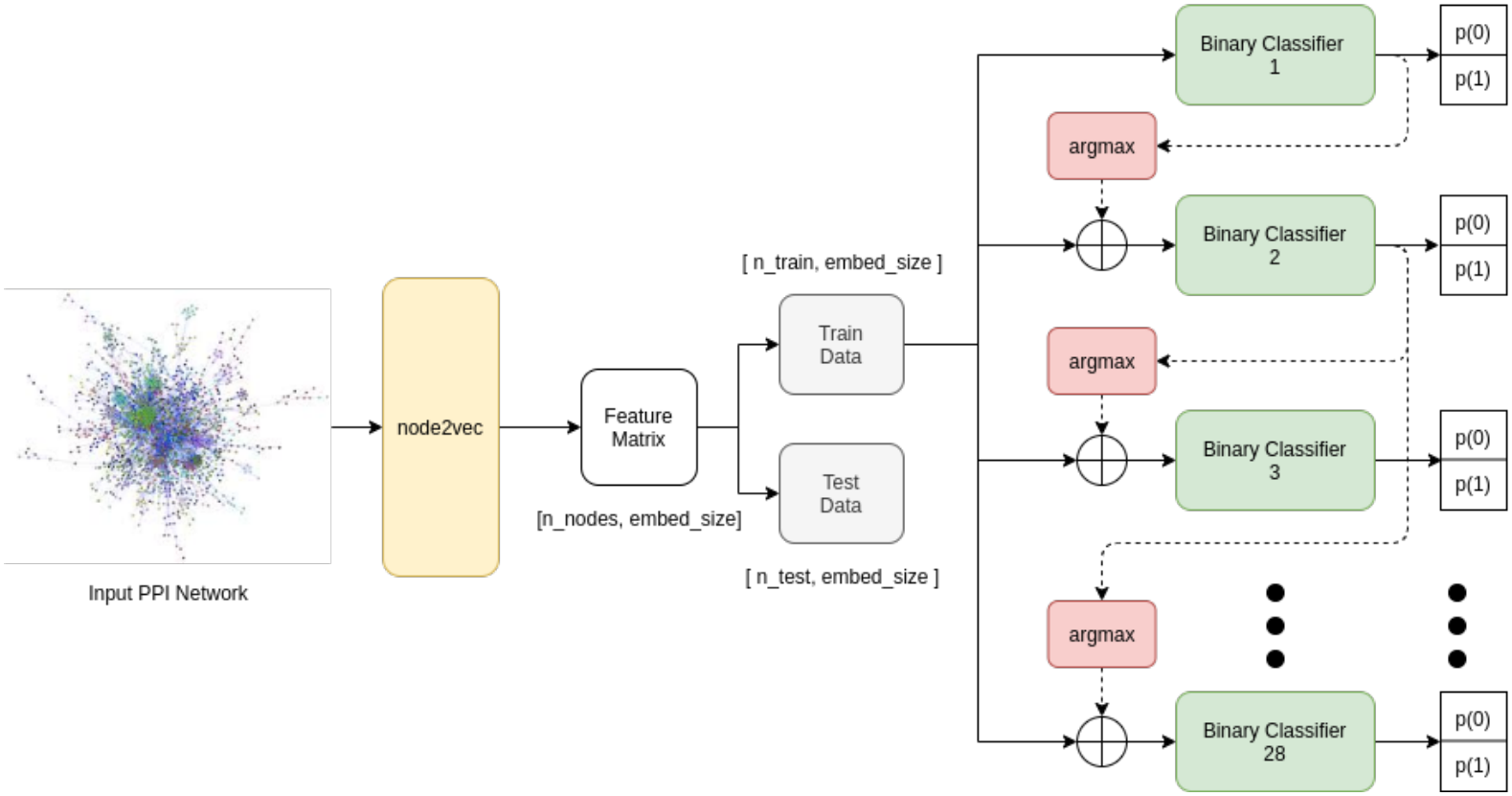


Figure: 28 binary classifiers each of which predicts the output of a single location.

Binary classifier

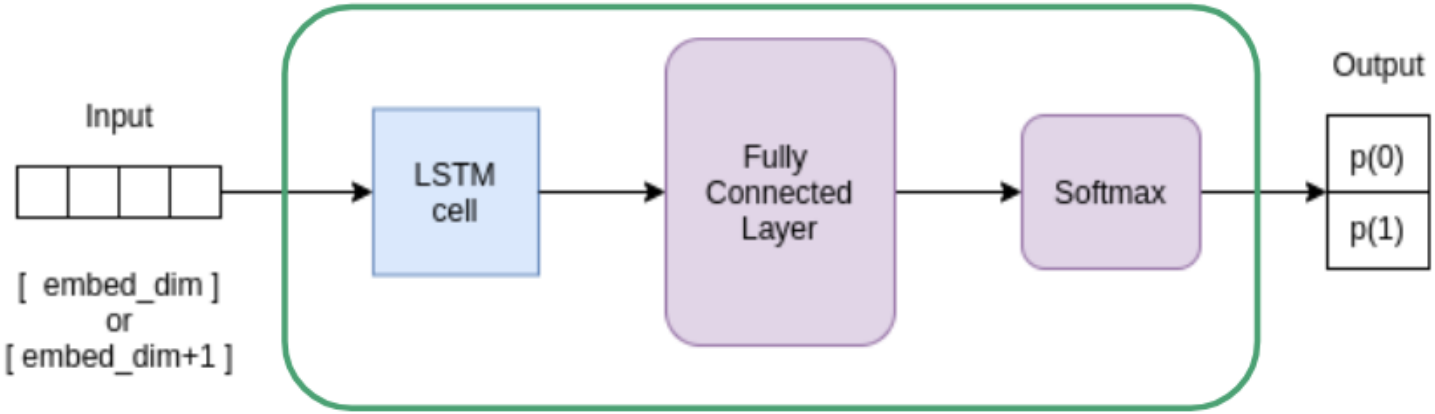


Figure: LSTM cell followed by the fully-connected layer that reduces the hidden dimension to 2 with the softmax activation function (based on Pan et al. (2019)).

Multilabel classification (DAG)

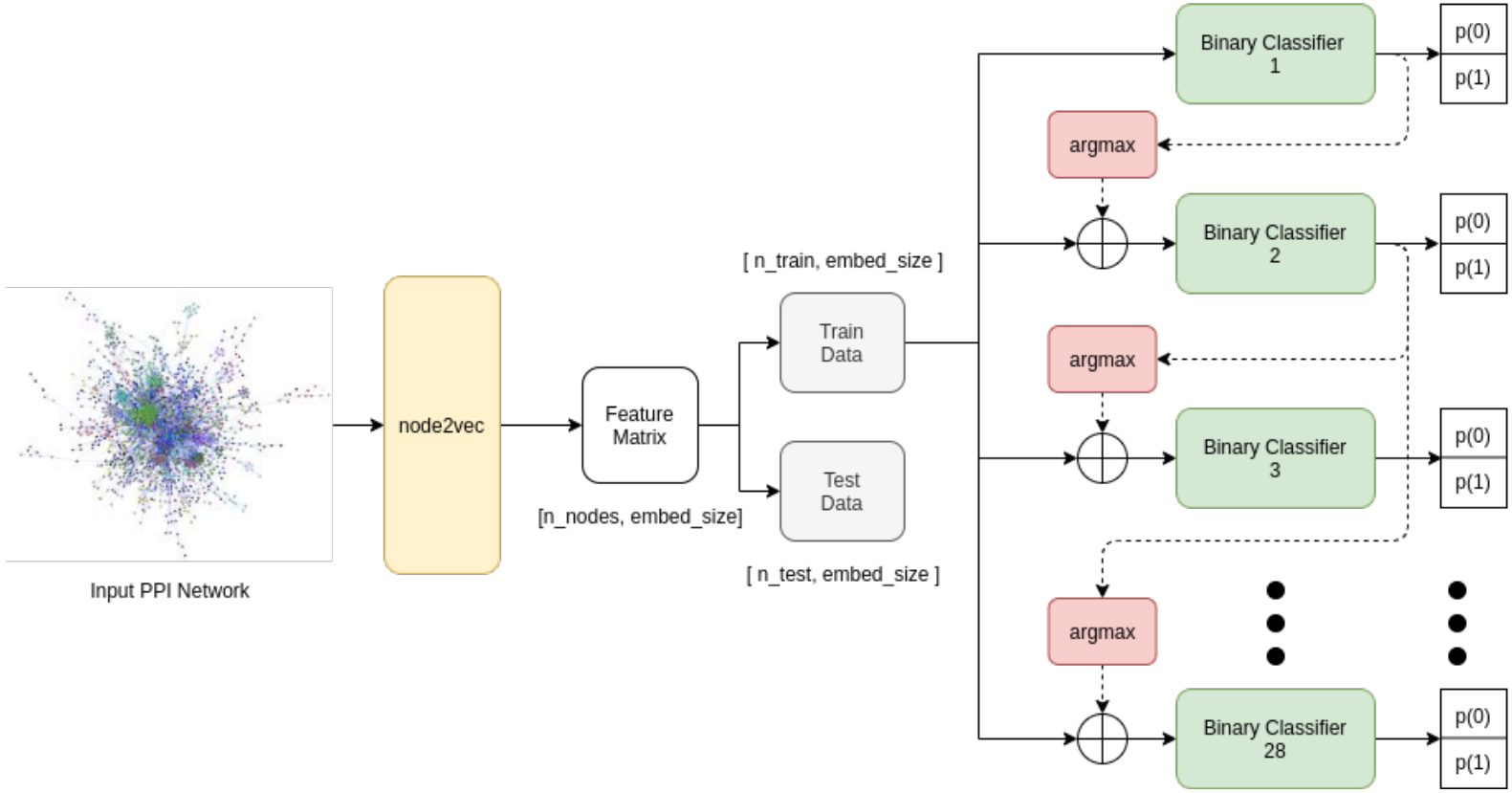
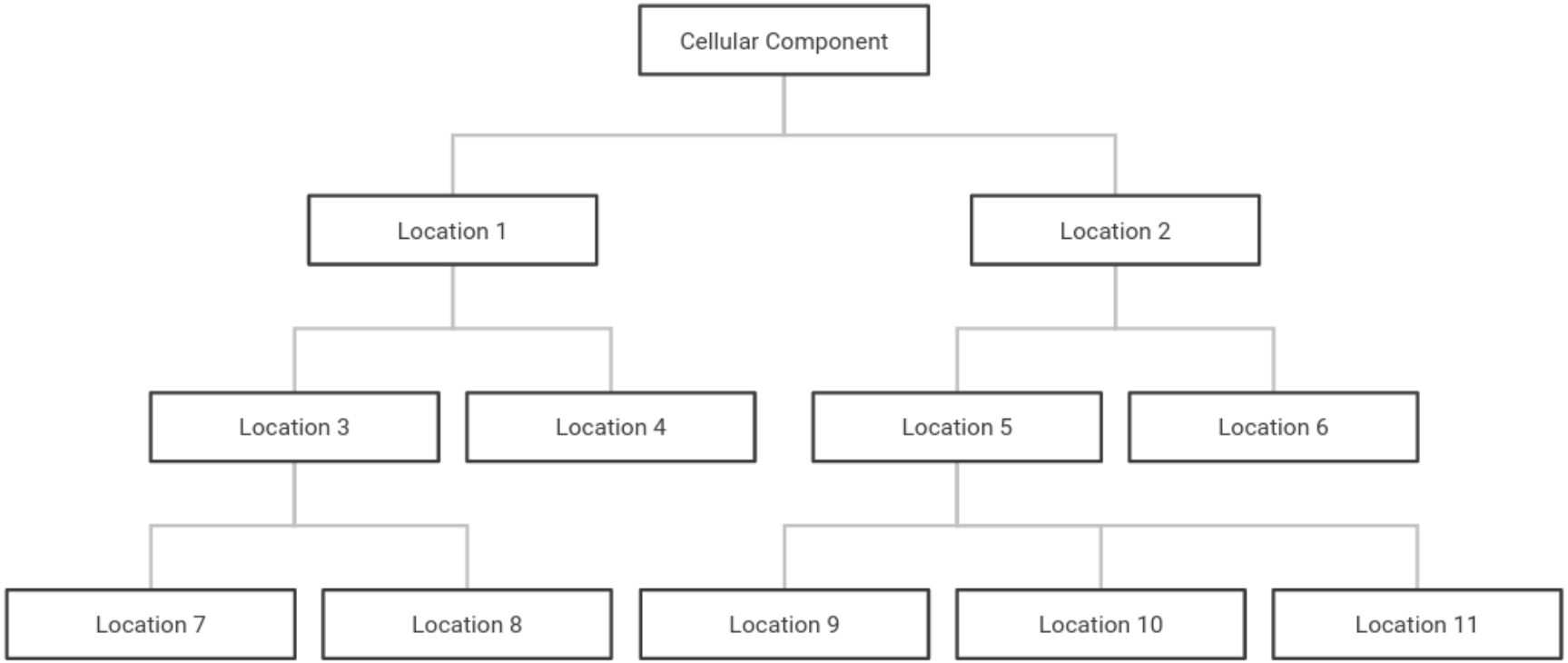
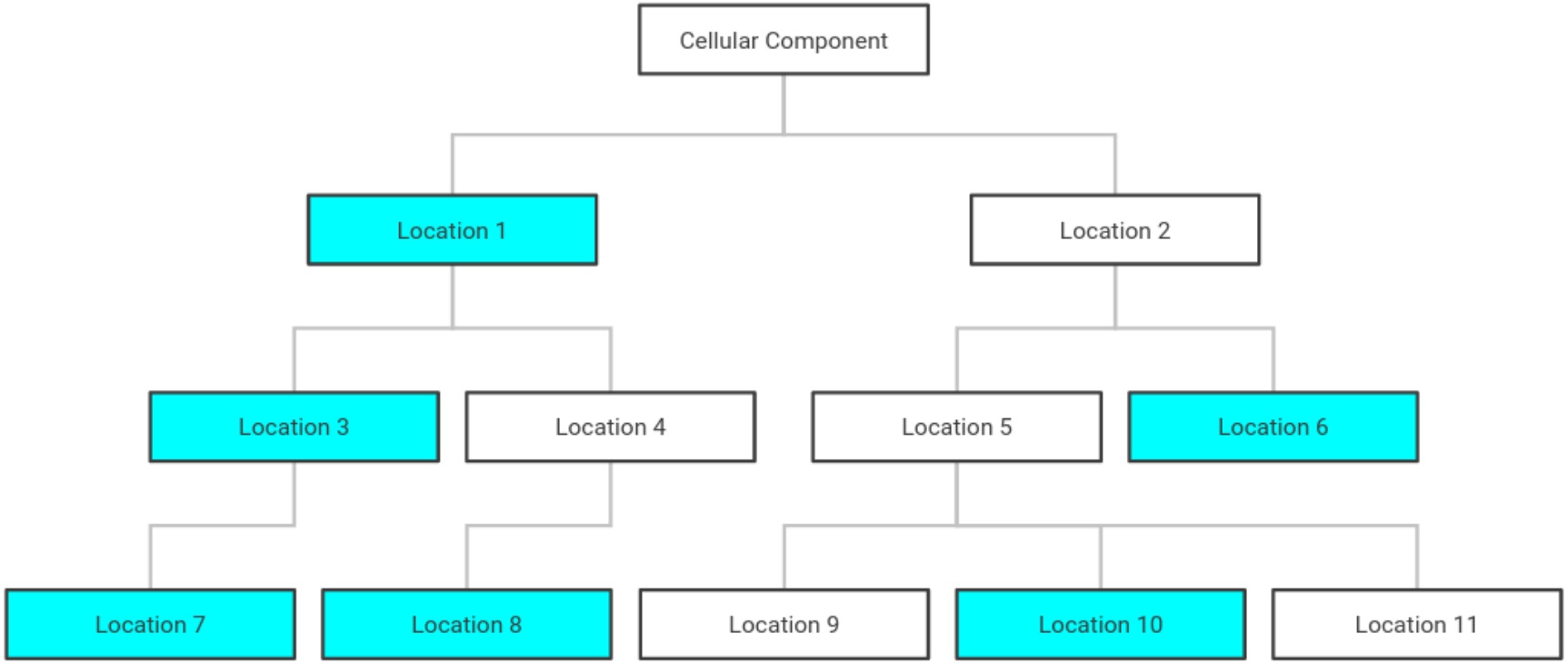


Figure: 28 locations of interest are not independent! These dependencies, based on the GO DAGS are incorporated while making predictions (dotted lines).

DAG structure (1)



DAG structure (2)



DAG structure (3)

Model tuning

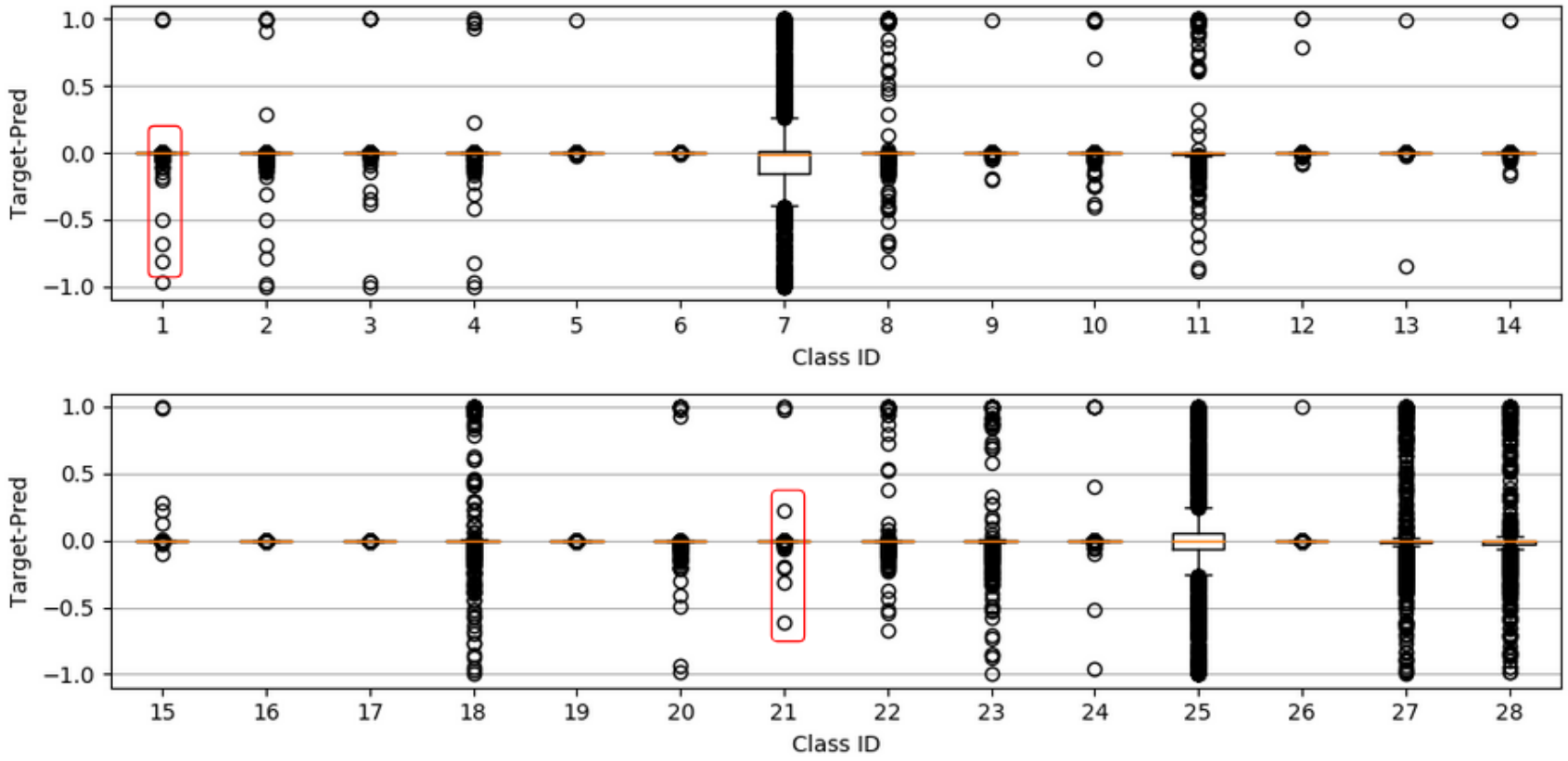


Figure: Model hyperparameters optimisation: number of epochs and 28 class thresholds (aggregated binary cross-entropy loss).

Preliminary results

Data	Filtering	Thresholding (% retained)	Loss	Accuracy	Sensitivity	Specificity	AUC-ROC	# data points (approx)
Combined	-	100	0.09211198312	0.8958986402	0.9571145481	0.6061253268	0.7816199375	65000
Combined	After node2vec	100	0.06054629228	0.970810473	0.9872313536	0.6738782051	0.8305547794	65000
Combined	Before node2vec	100	0.03273588772	0.9793367386	0.9918602378	0.8043154762	0.898087857	65000
Combined	-	70	0.08751676821	0.9372128248	0.9737043101	0.6011392776	0.7874217939	65000
Combined	After node2vec	70	0.05432904263	0.9745306969	0.9885046844	0.687366453	0.8379355687	65000
Combined	Before node2vec	70	0.02413146035	0.9863733053	0.9937820852	0.8583333333	0.9260577093	65000
Combined	-	100	0.08841074915	0.8633737564	0.9440426985	0.611838943	0.7779408208	150000
Combined	After node2vec	100	0.0515132316	0.9673402309	0.9871013833	0.6383223684	0.8127118758	150000
Combined	Before node2vec	100	0.03110835513	0.9817546606	0.9938010343	0.8235960145	0.9086985244	150000
Combined	-	70	0.08060763591	0.9333102703	0.9724473643	0.6090694232	0.7907583938	150000
Combined	After node2vec	70	0.0303748032	0.9744595885	0.9905288649	0.657127193	0.8238280289	150000
Combined	Before node2vec	70	0.01893794582	0.9898679852	0.9955202527	0.8930253623	0.9442728075	150000

- First iteration with the full dataset, after filtering, threshold of 0.6: AUC-ROC of 0.92.
- Biological findings: make **inferences on the unannotated** dataset.

Outline

Conclusions

- From data to biology and medicine.
- Interpretability is important; experimental vs. annotation.
- Interdisciplinary and collaborative work.

Thank you for your attention

laurent.gatto@uclouvain.be – de Duve Institute
`lgatto.github.io/about`

References I

M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000. doi: 10.1038/75556. URL `http://www.ncbi.nlm.nih.gov/pubmed/10802651`.

L M Breckels, S B Holden, D Wojnar, C M Mulvey, A Christoforou, A Groen, M W Trotter, O Kohlbacher, K S Lilley, and L Gatto. Learning from heterogeneous data sources: An application in spatial proteomics. *PLoS Comput Biol*, 12(5):e1004920, May 2016. doi: 10.1371/journal.pcbi.1004920.

LM Breckels, L Gatto, A Christoforou, AJ Groen, KS Lilley, and MW Trotter. The effect of organelle discovery upon sub-cellular protein localisation. *J Proteomics*, 88:129–40, Aug 2013.

A Christoforou, C M Mulvey, L M Breckels, A Geladaki, T Hurrell, P C Hayward, T Naake, L Gatto, R Viner, A Martinez Arias, and K S Lilley. A draft map of the mouse pluripotent stem cell spatial proteome. *Nat Commun*, 7:8992, Jan 2016. doi: 10.1038/ncomms9992.

Oliver M. Crook, Claire M. Mulvey, Paul D. W. Kirk, Kathryn S. Lilley, and Laurent Gatto. A bayesian mixture modelling approach for spatial proteomics. *PLOS Computational Biology*, 14(11):1–29, 11 2018. doi: 10.1371/journal.pcbi.1006516. URL `https://doi.org/10.1371/journal.pcbi.1006516`.

References II

Oliver M. Crook, Aikaterini Geladaki, Daniel J. H. Nightingale, Owen Vennard, Kathryn S. Lilley, Laurent Gatto, and Paul D. W. Kirk. A semi-supervised bayesian approach for simultaneous protein sub-cellular localisation assignment and novelty detection. *PLOS Computational Biology*, 16(11):1–21, 11 2020. doi: 10.1371/journal.pcbi.1008288. URL `https://doi.org/10.1371/journal.pcbi.1008288`.

OM Crook, LM Breckels, KS Lilley, PDW Kirk, and L Gatto. A bioconductor workflow for the bayesian analysis of spatial proteomics [version 1; peer review: 1 approved, 2 approved with reservations]. *F1000Research*, 8(446), 2019. doi: 10.12688/f1000research.18636.1.

TPJ Dunkley, S Hester, IP Shadforth, J Runions, T Weimar, SL Hanton, JL Griffin, C Bessant, F Brandizzi, C Hawes, RB Watson, P Dupree, and KS Lilley. Mapping the Arabidopsis organelle proteome. *PNAS*, 103(17):6518–6523, Apr 2006.

LJ Foster, CL de Hoog, Y Zhang, Y Zhang, X Xie, VK Mootha, and M Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, Apr 2006.

L Gatto, JA Vizcaino, H Hermjakob, W Huber, and KS Lilley. Organelle proteomics experimental designs and analysis. *Proteomics*, 2010.

L Gatto, LM Breckels, T Burger, DJ Nightingale, AJ Groen, C Campbell, N Nikolovski, CM Mulvey, A Christoforou, M Ferro, and KS Lilley. A foundation for reliable spatial proteomics data analysis. *MCP*, 13(8):1937–52, Aug 2014.

References III

Laurent Gatto, Lisa M Breckels, and Kathryn S Lilley. Assessing sub-cellular resolution in spatial proteomics experiments. *Current Opinion in Chemical Biology*, 48:123–149, 2019. ISSN 1367-5931. doi: https://doi.org/10.1016/j.cbpa.2018.11.015. URL `http://www.sciencedirect.com/science/article/pii/S1367593118301339`.

Aikaterini Geladaki, Nina Kočevar Britovšek, Lisa M Breckels, Tom S Smith, Owen L Vennard, Claire M Mulvey, Oliver M Crook, Laurent Gatto, and Kathryn S Lilley. Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.*, 10(1):331, January 2019.

Edward L. Huttlin, Raphael J. Bruckner, Jose Navarrete-Perea, Joe R. Cannon, Kurt Baltier, Fana Gebreab, Melanie P. Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, John Szpyt, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Guha Thakurta, Ramin Rad, Joshua Pan, David P. Nusinow, Joao A. Paulo, Devin K. Schweppe, Laura Pontano Vaites, J. Wade Harper, and Steven P. Gygi. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *bioRxiv*, 2020. doi: 10.1101/2020.01.19.905109. URL `https://www.biorxiv.org/content/early/2020/01/19/2020.01.19.905109`.

D N Itzhak, S Tyanova, J Cox, and G H Borner. Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, 5, Jun 2016. doi: 10.7554/eLife.16950.

C M Mulvey, L M Breckels, A Geladaki, N K Britovek, DJH Nightingale, A Christoforou, M Elzek, M J Deery, L Gatto, and K S Lilley. Using hyperlopit to perform high-resolution mapping of the spatial proteome. *Nat Protoc*, 12(6):1110–1135, Jun 2017. doi: 10.1038/nprot.2017.026.

References IV

Xiaoyong Pan, Lei Chen, Min Liu, Tao Huang, and Yu-Dong Cai. Predicting protein subcellular location using learned distributed representations from a protein-protein network. *bioRxiv*, 2019. doi: 10.1101/768739. URL https://www.biorxiv.org/content/early/2019/09/15/768739.

C Stadler, E Rexhepaj, V R Singan, R F Murphy, R Pepperkok, M Uhlén, J C Simpson, and E Lundberg. Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat Methods*, 10(4):315–23, Apr 2013.

Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1131. URL https://doi.org/10.1093/nar/gky1131.

References V

Mathias Uhlén, Erik Björling, Charlotta Agaton, Cristina Al-Khalili Szigyarto, Bahram Amini, Elisabet Andersen, Ann-Catrin Andersson, Pia Angelidou, Anna Asplund, Caroline Asplund, Lisa Berglund, Kristina Bergström, Harry Brumer, Dijana Cerjan, Marica Ekström, Adila Elobeid, Cecilia Eriksson, Linn Fagerberg, Ronny Falk, Jenny Fall, Mattias Forsberg, Marcus Gry Björklund, Kristoffer Gumbel, Asif Halimi, Inga Hallin, Carl Hamsten, Marianne Hansson, My Hedhammar, Görel Hercules, Caroline Kampf, Karin Larsson, Mats Lindskog, Wald Lodewyckx, Jan Lund, Joakim Lundeberg, Kristina Magnusson, Erik Malm, Peter Nilsson, Jenny Ödling, Per Oksvold, Ingmarie Olsson, Emma Öster, Jenny Ottosson, Linda Paavilainen, Anja Persson, Rebecca Rimini, Johan Rockberg, Marcus Runeson, Åsa Sivertsson, Anna Sköllermo, Johanna Steen, Maria Stenvall, Fredrik Sterky, Sara Strömberg, Mårten Sundberg, Hanna Tegel, Samuel Tourle, Eva Wahlund, Annelie Waldén, Jinghong Wan, Henrik Wernérus, Joakim Westberg, Kenneth Wester, Ulla Wrethagen, Lan Lan Xu, Sophia Hober, and Fredrik Pontén. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & Cellular Proteomics*, 4(12):1920–1932, 2005. ISSN 1535-9476. doi: 10.1074/mcp.M500279-MCP200. URL `https://www.mcponline.org/content/4/12/1920`.

References VI

Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–1250, December 2010. ISSN 1546-1696. doi: 10.1038/nbt1210-1248. URL http://dx.doi.org/10.1038/nbt1210-1248.