Bringing Models to the Domain: Deploying Gaussian Processes in the Biological Sciences

Max Zwießele
Neil D. Lawrence

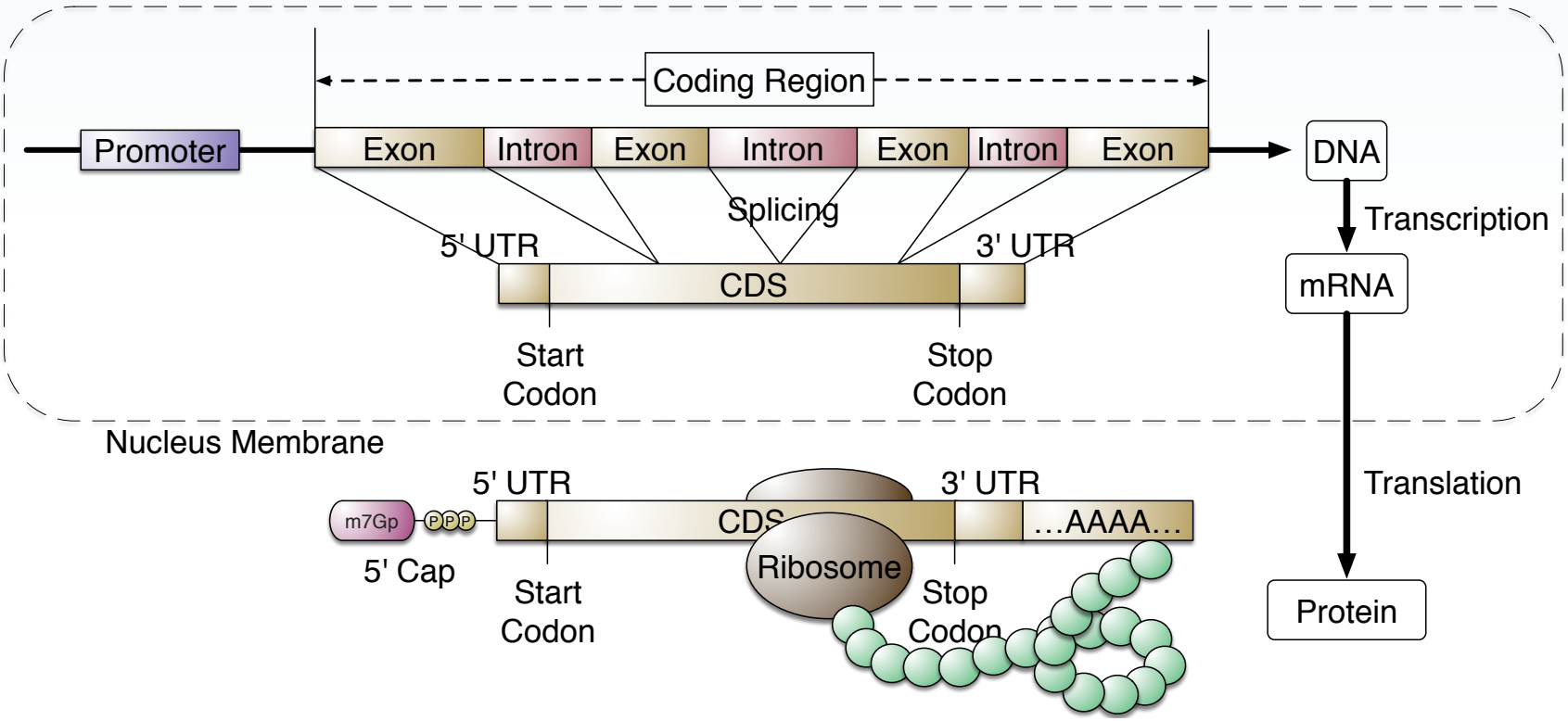
The University of Sheffield
Department of Computer Science

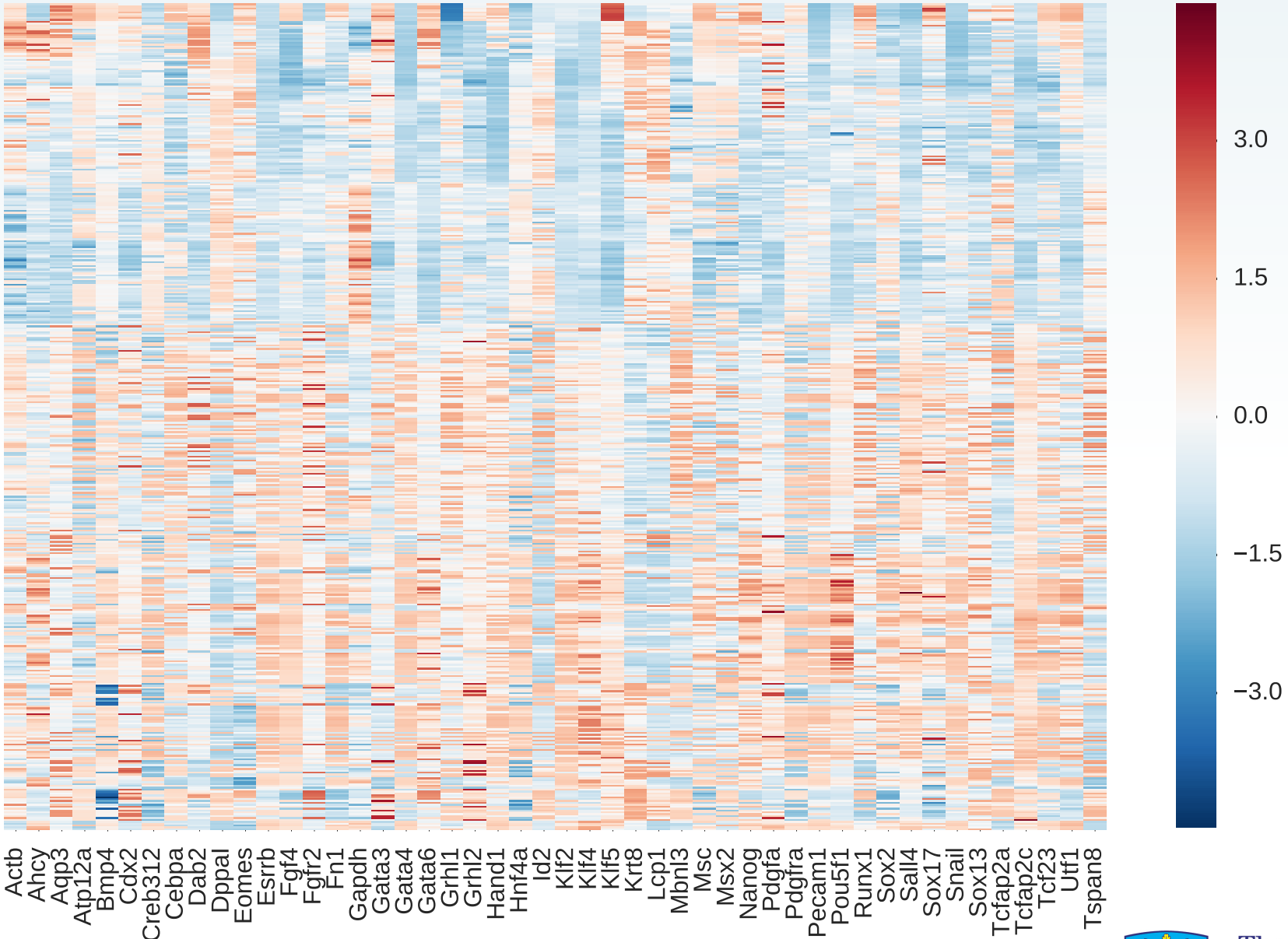
June 21, 2017

The University Of Sheffield.

Gene Expression

From DNA to Proteins

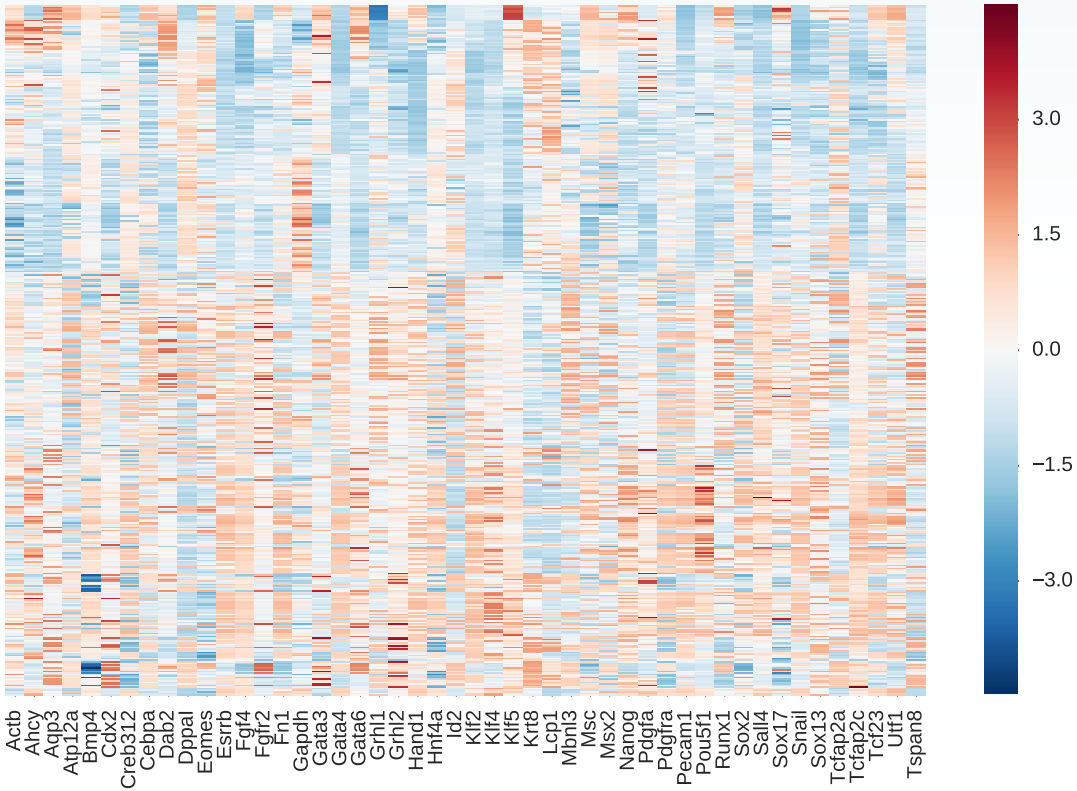


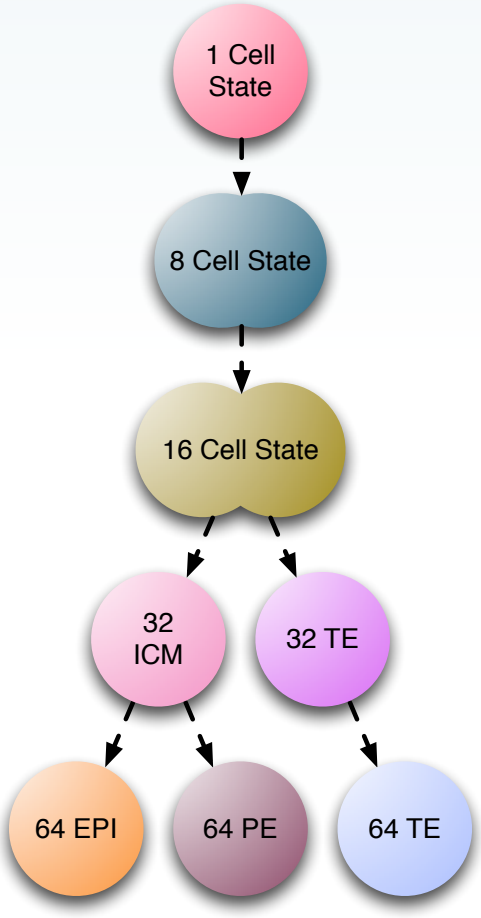


[Guo et al., 2010]

The University Of Sheffield.

Intrinsic Signal Discovery: Pseudo Time





The University Of Sheffield.

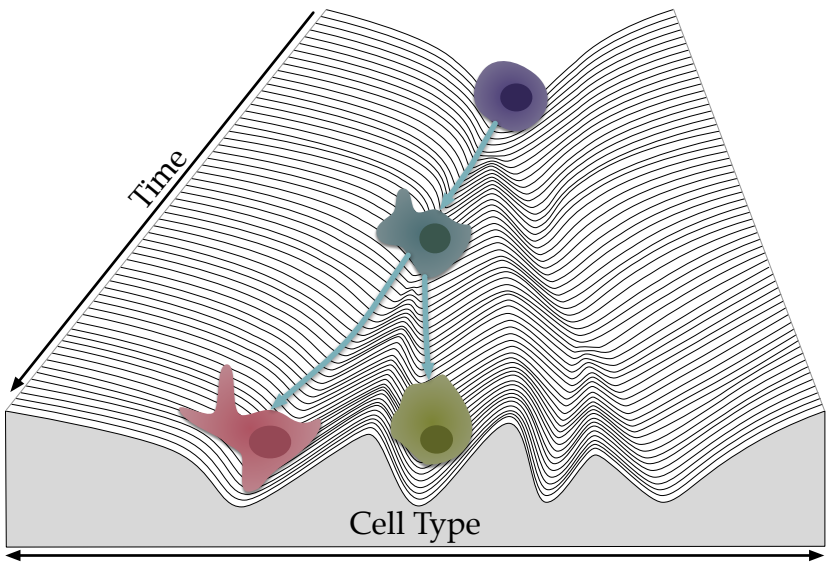
Pseudo Time Ordering: State of the Art

Two step approach:

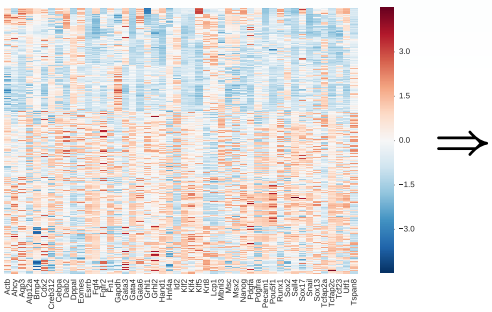
- Lower dimensional representation (usually 2D).
- Given Starting Cell.
- Find ordering in representation, following
  - Minimal spanning Tree (e.g. Monocle [Trapnell et al., 2014]).
  - K-Nearest-Neighbour graph (e.g. Wishbone [Setty et al., 2016]).
- (Post process ordering, smoothing, branching, shortcut detection etc.)

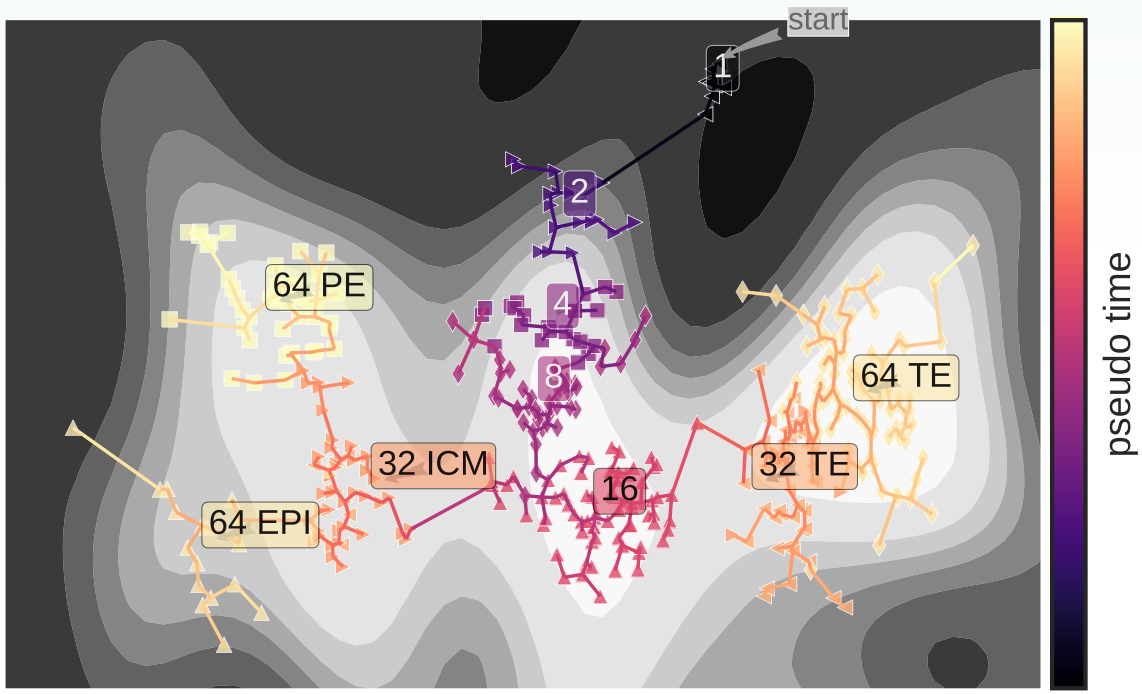
Landscape Learning

- Bayesian Gaussian Process Latent Variable Model (Probabilistic Landscape Modelling) [Titsias, 2009].
- Can be Interpreted as Probabilistic Waddington's Landscape [Waddington, 1966].
- Pairwise Distances in Landscape Distorted by Topography
- Correct Distances for Landscape Ridges and Hills.
- MST (or KNN) to Extract Pseudotime.



Topslam [Zwiessele and Lawrence, 2017]





Joint Modelling of Gene Subsets

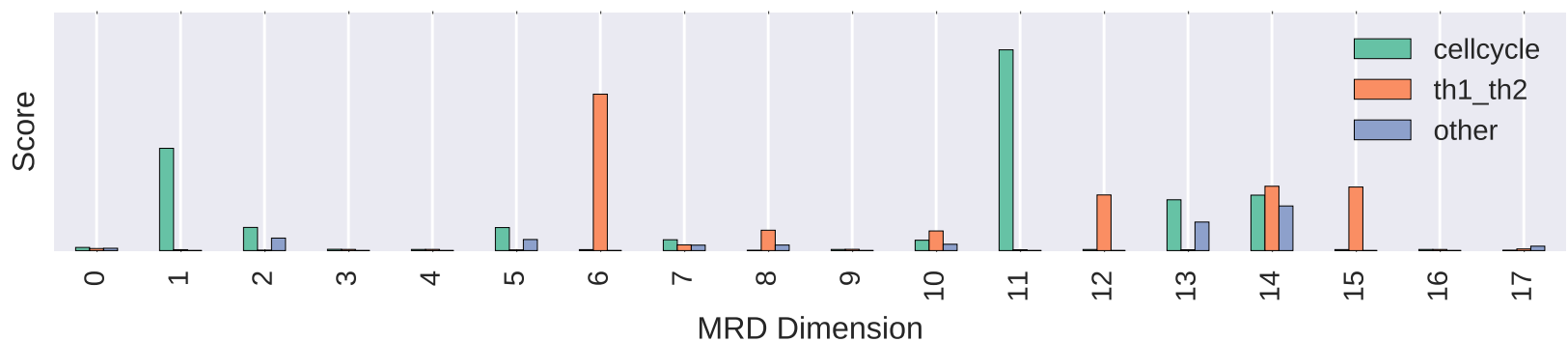
- In collaboration with Aleksandra Kolodziejczyk and Sarah Teichmann
- Singlecell RNASeq
- Naive Thelper to Th1 and Th2
- Samples mixed up during processing
- $\Rightarrow$ No reliable knowledge about time or differentiation state
- PCA (and others) don't reveal differentiation.

Data Cleanup through Prior Knowledge

MRD on gene subsets (similar [Buettner et al., 2015]).

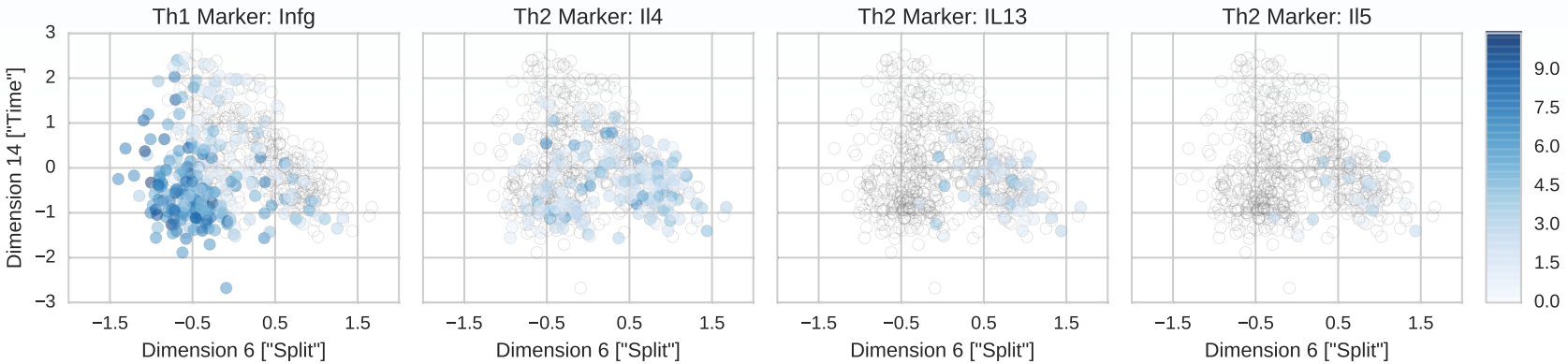
Define gene sets of interest:

Cell cycle related DE in bulk All other



Data Cleanup through Prior Knowledge

Time dimension vs. Split dimension:



Takehome

Pseudo Time Extraction

- Taking landscape into account improves pseudotime extraction.
- Prior knowledge for gene subset detection improves/allows for signal detection.

Takehome

Code

MRD and Bayesian GPLVM [GPy, 2012]:
`https://github.com/SheffieldML/GPy`

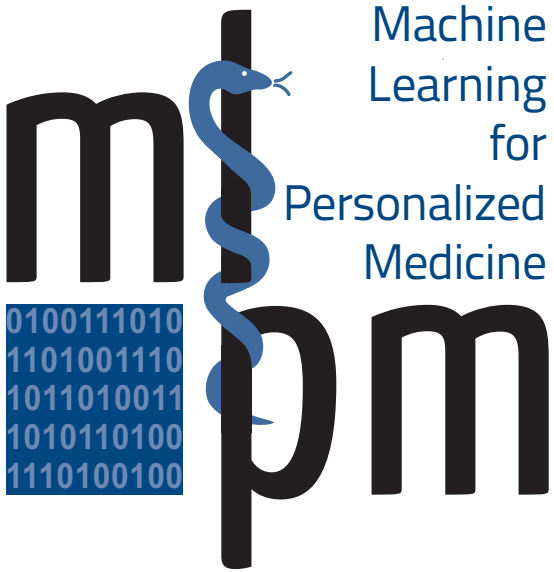
topslam (based on GPy): [Zwiessele and Lawrence, 2017]
`https://github.com/mzwiessele/topslam`

GPclust (based on GPy): [Lázaro-Gredilla et al., 2012]
`https://github.com/SheffieldML/GPclust`

GPfates (based on GPy) [Lönnberg et al., 2017]:
`https://github.com/Teichlab/GPfates`

Acknowledgements

I am grateful for financial support from the European Union 7th Framework Programme through the Marie Curie Initial Training Network "Machine Learning for Personalized Medicine" MLPM2012, Grant No. 316861.



The University Of Sheffield.

Thanks!

References I

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015).
Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.
Nature biotechnology, 33(2):155–160.

GPy (since 2012).
GPy: A Gaussian process framework in python.
`http://github.com/SheffieldML/GPy`.

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., and Robson, P. (2010).
Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst.
Developmental cell, 18(4):675–685.

Lázaro-Gredilla, M., Van Vaerenbergh, S., and Lawrence, N. D. (2012).
Overlapping Mixtures of Gaussian Processes for the data association problem.
Pattern Recognition, 45(4):1386–1395.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U. N., et al. (2017).
Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria.
Science Immunology, 2(9):eaal2192.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. (2016).
Wishbone identifies bifurcating developmental trajectories from single-cell data.
Nature biotechnology, 34(6):637–645.

The University Of Sheffield.

References II

Titsias, M. K. (2009).
Variational Learning of Inducing Variables in Sparse Gaussian Processes.
Artificial Intelligence and Statistics, 5:567–574.

Tosi, A., Hauberg, S., Vellido, A., and Lawrence, N. D. (2014).
Metrics for probabilistic geometries.
In *Uncertainty in Artificial Intelligence*, volume 30, pages 800–808. AUAI Press (Association for Uncertainty in Artificial Intelligence).

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014).
The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.
Nature Biotechnology, 32(4):381–386.

Waddington, C. H. (1966).
Principles of Development and Differentiation.
BioScience, 16(11):821–822.

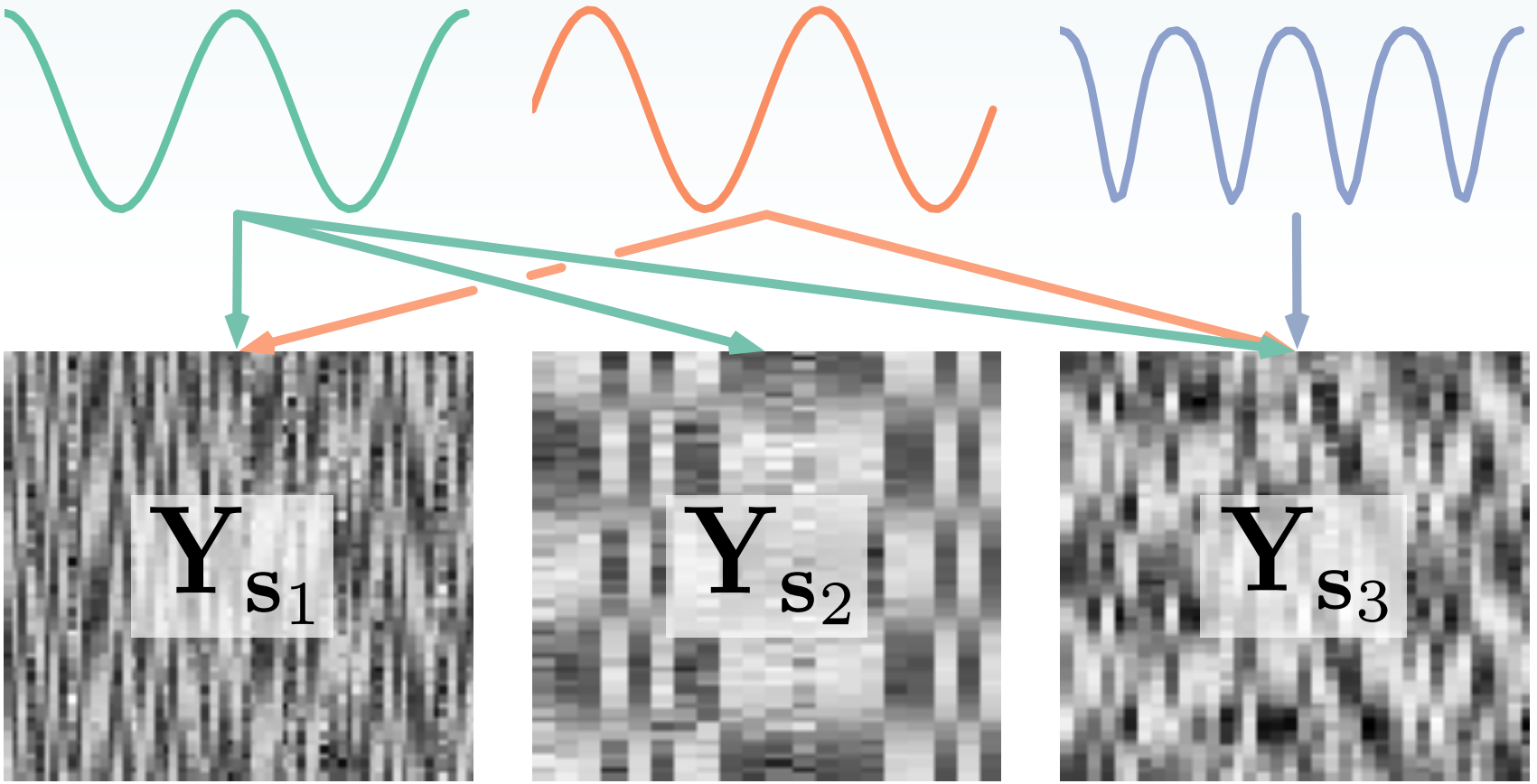
Yamanaka, Y., Ralston, A., Stephenson, R. O., and Rossant, J. (2006).
Cell and Molecular Regulation of the Mouse Blastocyst.
Developmental Dynamics, 235(9):2301–2314.

Zwiessele, M. and Lawrence, N. D. (2017).
Topslam: Waddington landscape recovery for single cell experiments.
bioRxiv.

The University Of Sheffield.

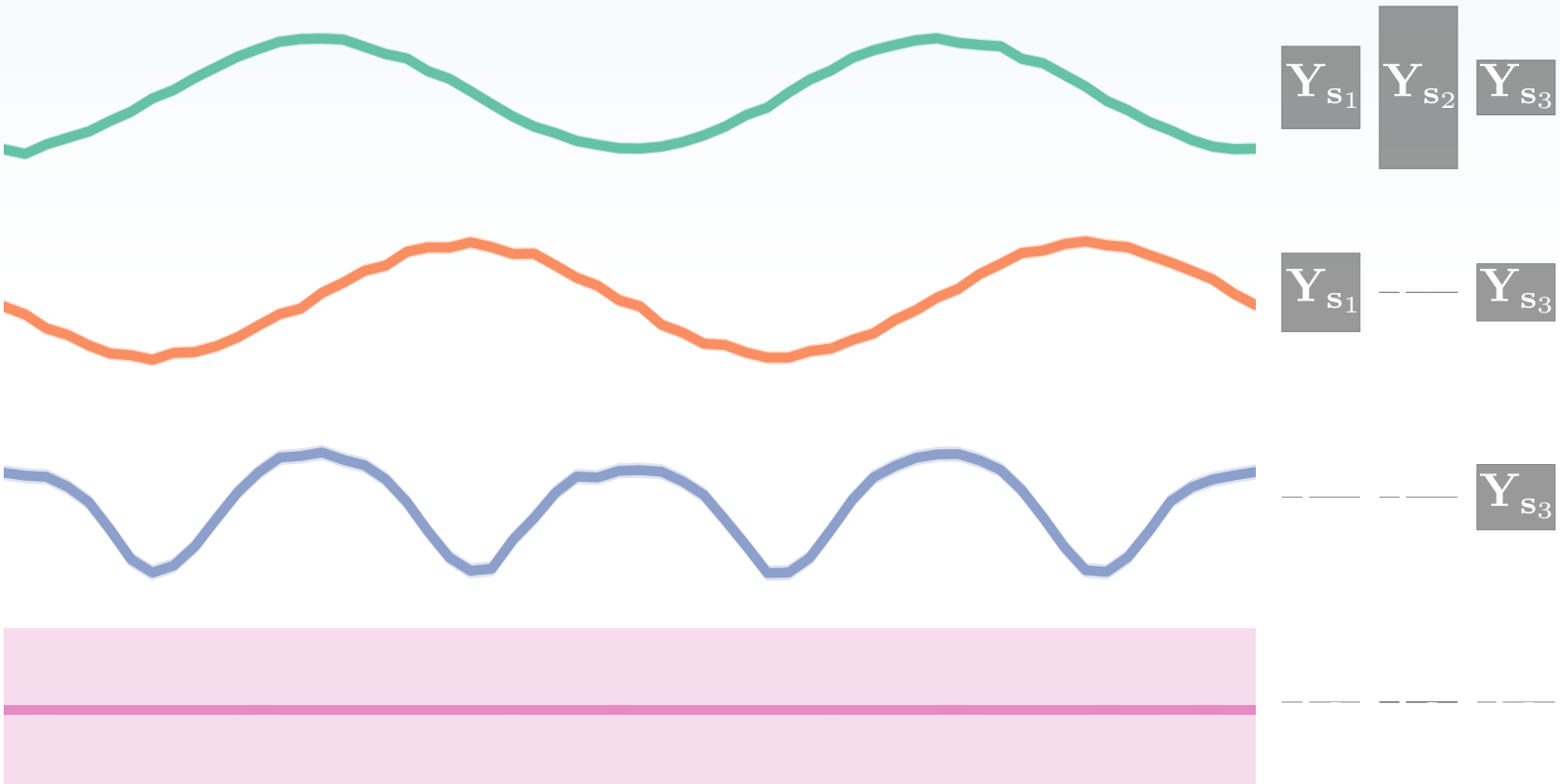
Appendix

MRD



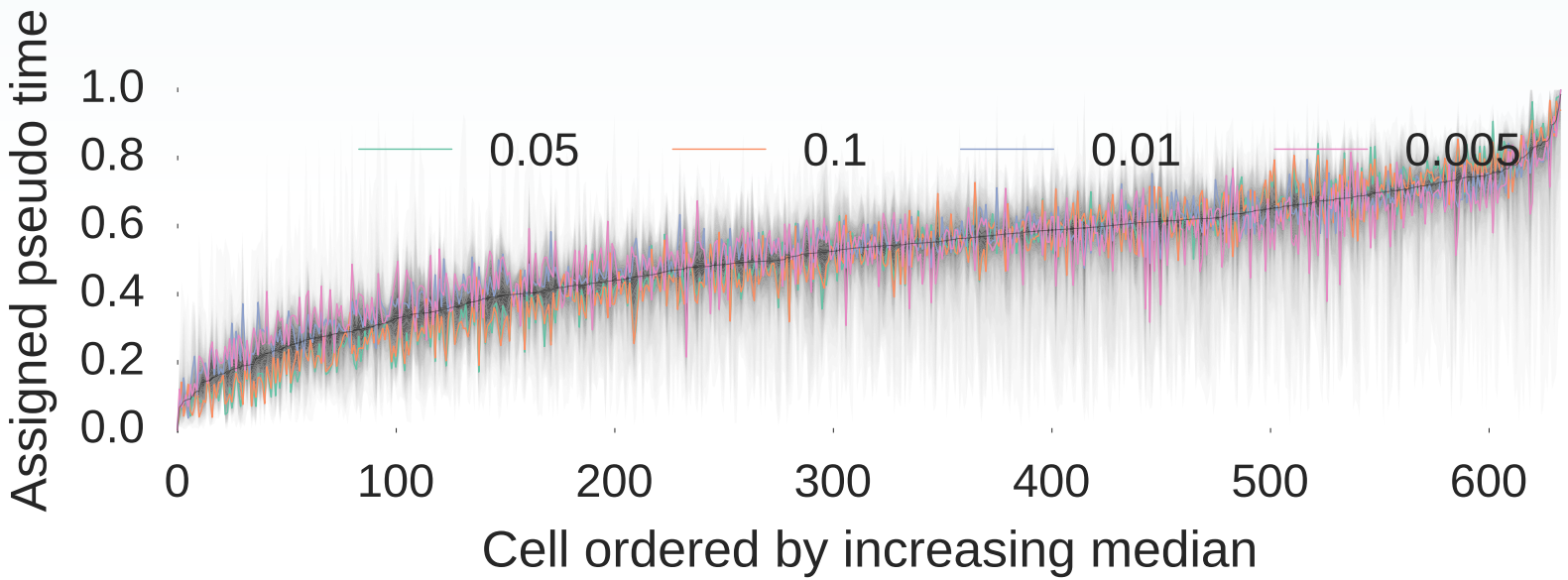
Appendix

MRD



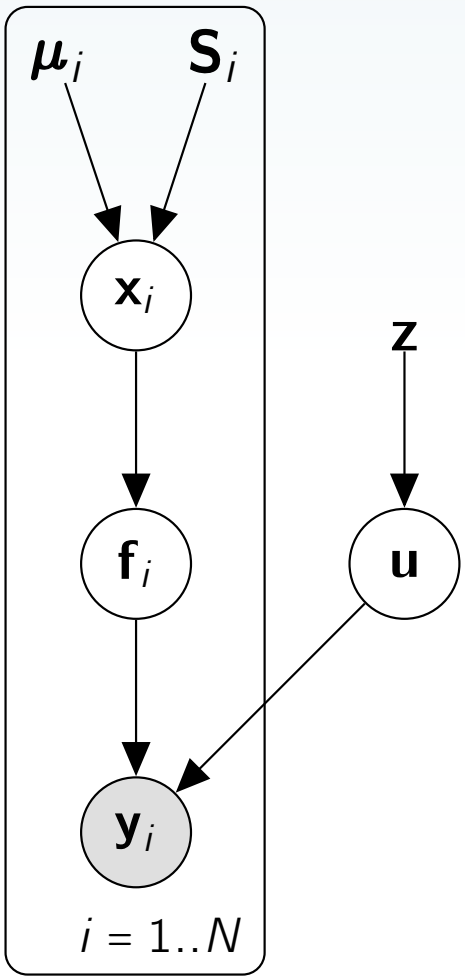
Appendix

Topslam Stability



Appendix

Bayesian GPLVM



$$\log p(\mathbf{y}) = \log \iint p(\mathbf{y}|\mathbf{x},\mathbf{u})p(\mathbf{u})p(\mathbf{x}) \; \mathrm{d}\mathbf{u}\,\mathbf{x}$$

$$= \log \iint p(\mathbf{y}|\mathbf{x},\mathbf{u})p(\mathbf{u}) \; \mathrm{d}\mathbf{u} p(\mathbf{x}) \frac{q(\mathbf{x})}{q(\mathbf{x})} \; \mathrm{d}\mathbf{x}$$

$$\geq \left\langle \langle \log p(\mathbf{y}|\mathbf{x},\mathbf{u}) \rangle_{p(\mathrm{u})} - \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\rangle_{q(\mathrm{x})}$$

$$= \left\langle \langle \log p(\mathbf{y}|\mathbf{x},\mathbf{u}) \rangle_{p(\mathrm{u})} \right\rangle_{q(\mathrm{x})} - \mathrm{KL}\left(q(\mathbf{x})||p(\mathbf{x})\right)$$

$$=: \mathcal{L}_3$$

$$q(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i)$$

The University Of Sheffield.

Appendix

Magnification Factor

[Tosi et al., 2014]'s Magnification Factor:

$$D_{ij}^{\text{euclidean}} = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \mathbf{I} (\mathbf{X}_i - \mathbf{X}_j)}$$

$$\mathbf{J} = \frac{\partial p(\mathbf{y})}{\partial \mathbf{X}}$$

$$\mathbf{M} := \left\langle \mathbf{J}^\top \mathbf{J} \right\rangle = \left\langle \mathbf{J} \right\rangle^\top \left\langle \mathbf{J} \right\rangle + D \cdot \text{cov}(\mathbf{J}, \mathbf{J})$$

$$D_{ij}^{\text{corrected}} = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^\top \frac{\mathbf{M}_i + \mathbf{M}_j}{2} (\mathbf{X}_i - \mathbf{X}_j)}$$