bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Synthesizing developmental trajectories

Paul Villoutreix[a,1], Joakim Andén[b,1], Bomyi Lim[a,c], Hang Lu[d,e], Ioannis G. Kevrekidis[b,c], Amit Singer[b,f], and Stanislav Y. Shvartsman[a,b,2]

[a]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; [b]Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544; [c]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; [d]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [e]Interdisciplinary Program in Bioengineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [f]Department of Mathematics, Princeton University, Princeton, NJ 08544, USA; [1]P.V. and J.A. contributed equally to this work.; [2]To whom correspondence should be addressed. E-mail: stas@princeton.edu.

Dynamical processes in biology are studied using an ever-increasing number of techniques, each of which brings out unique features of the system. One of the current challenges is to develop systematic approaches for fusing heterogeneous datasets into an integrated view of multivariable dynamics. We demonstrate that heterogeneous data fusion can be successfully implemented within a semi-supervised learning framework that exploits the intrinsic geometry of high-dimensional datasets. We illustrate our approach using a dataset from studies of pattern formation in Drosophila. The result is a continuous trajectory that reveals the joint dynamics of gene expression, subcellular protein localization, protein phosphorylation, and tissue morphogenesis. Our approach can be readily adapted to other imaging modalities and forms a starting point for further steps of data analytics and modeling of biological dynamics.

1. Introduction

The need to synthesize data from different observations into coherent multivariable trajectories is discussed in multiple contexts, from physics to social sciences, but systematic approaches for accomplishing this task have yet to be established (1–5). Here we address this task for imaging studies of developing tissues, where patterns of cell fates are established by complex regulatory networks (6–8). Advances in live imaging continue to provide new insights into the dynamics of individual components in these networks, but imaging more than three reporters at the same time is still challenging and limited to model genetic organisms (9, 10). Furthermore, in the absence of reliable live reporters, dynamics of some state variables can only be inferred from fixed tissues. Because of these limitations, extracting the multivariable dynamics from the heterogeneous datasets collected by imaging of live and fixed tissues becomes a non-trivial task (11, 12).

The problem can be illustrated by an imaging dataset from the early Drosophila embryo (Fig. 1A-B), a model system in which a graded profile of the nuclear localization of transcription factor Dorsal (Dl) establishes the dorsoventral (DV) stripes of gene expression that control cell fates and tissue deformations (13–15). Current mechanisms of the DV patterning system invoke multiple state variables, such as the levels of gene expression and protein phosphorylation (16) (Fig. 1C). These mechanisms were elucidated in studies that reveal only a small subset of the full state space, most commonly 2-3 variables per experiment. Can these partial views be fused into a consistent multivariable trajectory? This is a general question that applies to essentially all developmental systems.

We realized that this question can be addressed by casting the task of data fusion as a matrix completion problem (Fig. 1D). Specifically, an image of a fixed embryo or a frame from a live imaging movie can be viewed as a column in a matrix where rows correspond to the relevant variables, such as developmental time or the level of gene expression at a given position. Because of limitations in the number of states that can be accessed simultaneously, the matrix is incomplete. For example, live imaging of gastrulation provides information about nuclear positions as a function of time, but

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

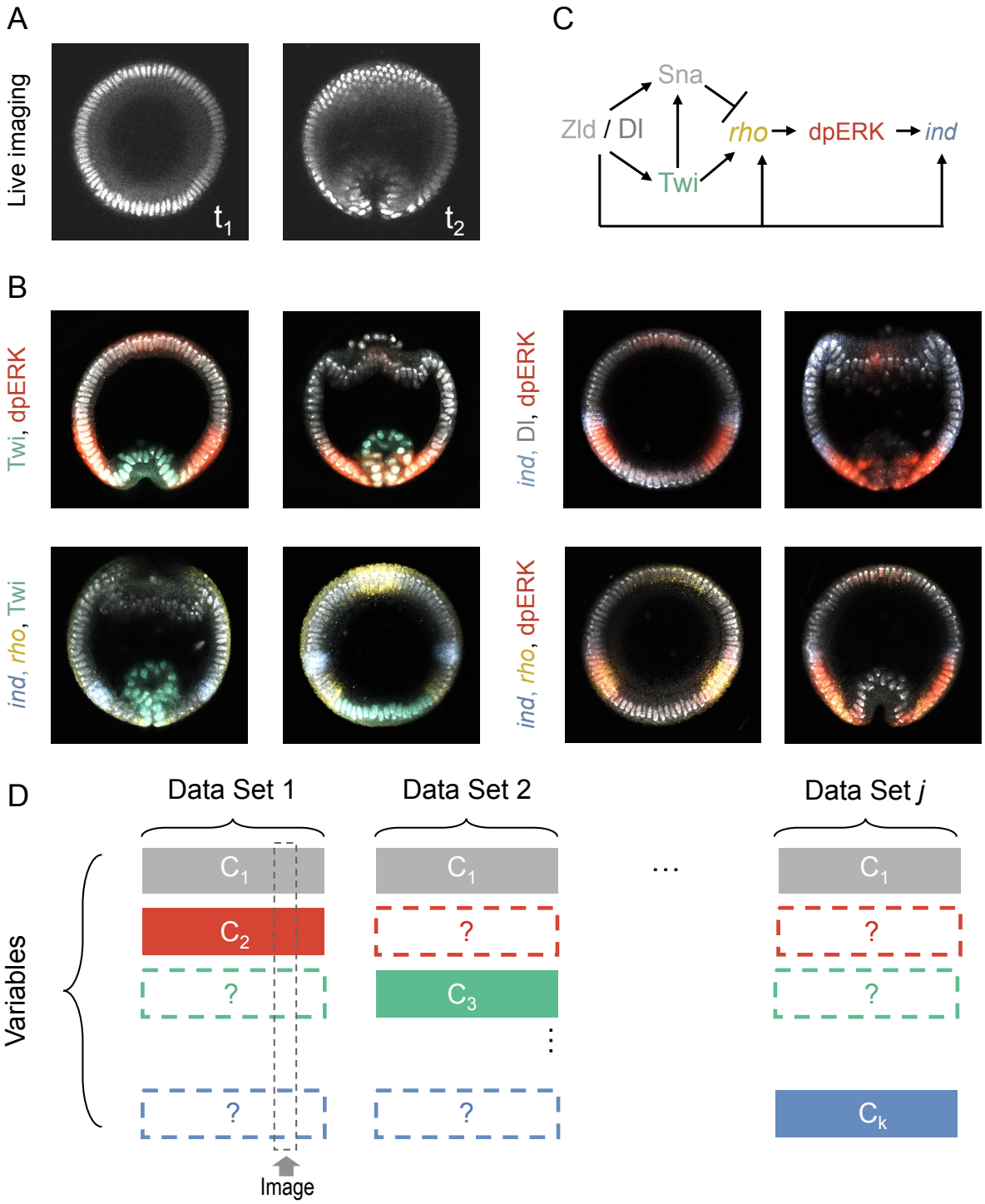

Fig. 1. Stating the problem of data fusion. (A-B) Example datasets of molecular signals and morphology during the DV patterning of Drosophila embryo; all images are collected from optical cross-section along the DV axis, $\sim 15\%$ from the posterior pole of the embryo. (A) Frames from a live imaging movie, showing positions of nuclei during the early stages of gastrulation. (B) Images of fixed embryos, stained with probes and antibodies revealing the spatial patterns of nuclear Dl (pink), Twi (green), dually phosphorylated ERK (red), and transcripts of rho (yellow) and ind (blue). (C) A fragment of a DV patterning network in the early Drosophila embryo. (D) Data fusion as a matrix completion problem: Each row corresponds to a variable, e.g. nuclear positions, gene expression levels, time stamp, revealed by visualizing different molecular or cellular components, nuclei, transcripts, or protein phosphorylation. Each column of the matrix corresponds to an image giving access to some of the states through various channels. The remaining states, labeled with a question mark, must be estimated from other datasets.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

is silent about the levels of gene expression. On the other hand, an image of a fixed embryo reveals the distribution of an active enzyme but has no direct temporal information. Thus, multivariable data fusion requires completing this matrix, filling in the missing components by estimates informed by the rest of the data. Below we show how this task can be accomplished by solving a suitably posed semi-supervised learning problem. We first provide a closed-form solution to this problem and then demonstrate its successful performance on synthetic and experimental datasets.

2. Results

Semi-supervised learning framework for matrix completion. We assume here that all experiments contain a common variable, which is sufficient to determine all other variables that can be measured or to be predicted. For instance, this variable is revealed by a signal that reports positions of nuclei. This means that the first row in the matrix are complete. To complete other rows, we must establish the mappings between the common variable and each of the target variables. These mappings can be found within a semi-supervised learning framework, in which the values of the variables in the incomplete rows are estimated from a training dataset (17, 18).

As an example, consider images from fixed embryos that are stained to reveal the spatial pattern of an active enzyme, visualized using a phosphospecific antibody (Fig. 2). They provide labeled data points that contain information about the common variable and a specific target variable. On the other hand, images without this staining, such as the frames from live imaging of morphogenesis, provide unlabeled data points with only the common variable. By finding a mapping between the common and target variables, we can essentially "color" the frames of a live imaging movie by snapshots of molecular patterns from fixed embryos.

A critical assumption in finding the mappings is that the multivariable dynamics of the patterning process are both low-dimensional and smooth. This assumption is supported by studies with mathematical models of specific biological systems and by computational analysis of datasets from imaging studies of development (19, 20). More formally, we consider a set of data points $(x_1, ..., x_l, x_{l+1}, ..., x_{l+u})$ belonging to a space $\mathcal{X}$. These points correspond to the values of the common variable in the complete row. On the other hand, a row corresponding to any one of the target variables is incomplete. The values in the filled columns of this row are called labels. These are denoted $(y_1, ..., y_l)$ and belong to a target space $\mathcal{Y}$. The semi-supervised learning techniques transfer the information contained in the labeled data points $((x_1, y_1), ..., (x_l, y_l))$ to the unlabeled points $(x_{l+1}, ..., x_{l+u})$, while preserving the intrinsic structure of the dataset(18). Stated otherwise, these techniques *learn* the mapping $y = f(x)$ assuming that the considered process is smooth, which means that similar values of x give rise to similar values of $f(x)$.

The missing values, corresponding to the unlabeled data points in each of the incomplete rows, are found by solving the following optimization problem:

$$\vec{f} = \underset{\substack{\vec{f}\in\mathcal{Y}^{l+u} \\ \forall i\leq l, f_i=y_i}}{\operatorname{argmin}} \sum_{i,j=1}^{l+u} w_{i,j} \|f(x_i) - f(x_j)\|^2 \qquad [1]$$

where $\vec{f} = (f_1, ..., f_{l+u})$ are the values of the target variable on the data points $(x_1, ..., x_{l+u})$, the considered norm in $\mathcal{Y}$ is the Euclidean distance and the weights $w_{i,j}$ represent the similarity between two data points x_i and x_j. The norm in the space $\mathcal{X}$ can for example be the Euclidean distance in

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

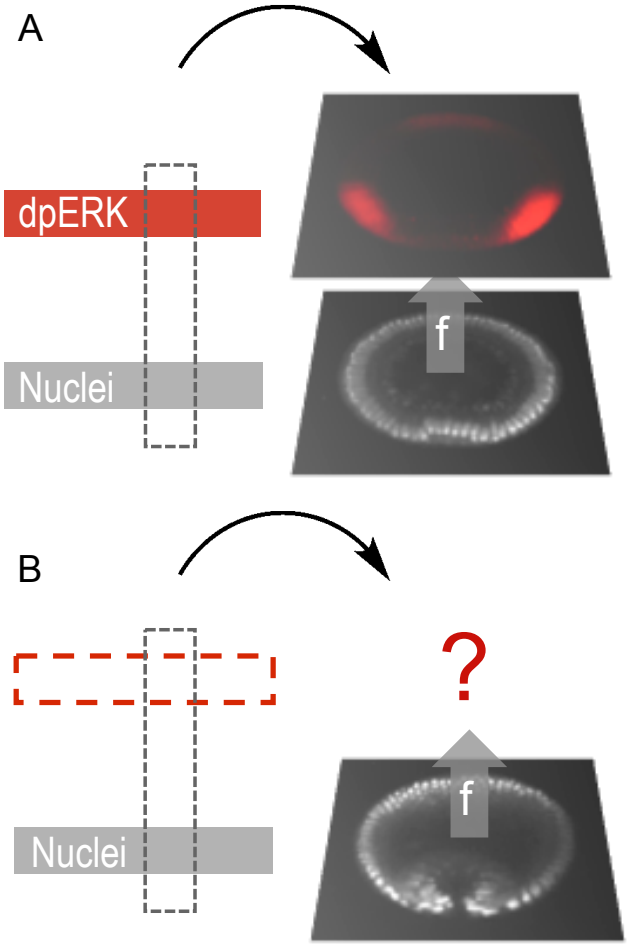

Fig. 2. Learning a mapping from a common channel. An experimental image can be decomposed into various channels. E.g., red: dpERK, visualized with a phosphospecific antibody, gray: nuclei, visualized through either DAPI (in fixed images) or Histone-RFP (in live imaging). The training ensemble of labeled images (A) is used to predict the labels on a set of unlabeled images (B) using common information, the morphology obtained through the nuclei signal in this case. Morphological proximity yields similar labels.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the space where each dimension corresponds to an image pixel or some coordinate in an arbitrary feature transform of that image.

This quadratic optimization problem, known as harmonic extension, has a unique solution that relates the unlabeled data points $f_{l+1}, ..., f_{l+u}$ to the labels $y_1, ..., y_l$ where $\mathcal{Y} = \mathbb{R}$ (17, 21). The explicit solution reads:

$$\vec{f^u} = (D_u - W_{uu})^{-1} W_{ul} Y \qquad [2]$$

where $Y = (y_1, ..., y_l)$ and $\vec{f^u} = (f_{l+1}, ..., f_{l+u})$, and $d_i = \sum_{j=1}^{l+u} w_{i,j}$, $D_u = \mathrm{diag}(d_{l+1}, ..., d_{l+u})$, $W_{uu} = (w_{i,j})_{l+1 \leq i,j \leq l+u}$, and $W_{ul} = (w_{i,j})_{\substack{l+1 \leq i \leq l+u \\ 1 \leq j \leq l}}$ (*Methods and Materials*).

Illustrative example. To illustrate our method, we considered a one-dimensional nonlinear trajectory in a three-dimensional space. The trajectory is given by the set of equations

$$\begin{cases} x^{(1)}(t) & = at\,(\cos(bt) + \epsilon^{(1)}) \\ x^{(2)}(t) & = at\,(\sin(bt) + \epsilon^{(2)}) \\ y(t) & = ct \exp(-d(t-e)^2) \end{cases} \qquad [3]$$

where a, b, c, d, e are constants, $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are Gaussian noise sources and t is a real-valued parameter. The set of points $(x^{(1)}(t), x^{(2)}(t))$ forms a one-dimensional non-linear manifold embedded in the two dimensional plane and it is parameterized by t. These points are analogs of the embryo morphology. In the absence of noise, this mapping from t to the 2D plane can be inverted as $t = \frac{1}{|a|}\sqrt{(x^{(1)})^2(t) + (x^{(2)})^2(t)}$. The signal $y(t)$ is a smooth function of t and is thus a smooth function of $(x^{(1)}, x^{(2)})$ by composition. In this example, y corresponds to the target modality that we would like to estimate.

To mimic the setting of data fusion with three modalities, $((x^{(1)}, x^{(2)}), t, y)$, we consider the following situation: suppose that one acquire a set of labeled points, i.e. a set of l triplets, $(((x^{(1)}(t_1), x^{(2)}(t_1)), y(t_1)), ..., ((x^{(1)}(t_l), x^{(2)}(t_l)), y(t_l)))$ and a set of u unlabeled, but timestamped, points, $(((x^{(1)}(t_{l+1}), x^{(2)}(t_{l+1})), t_{l+1}), ..., ((x^{(1)}(t_{l+u}), x^{(2)}(t_{l+u})), t_{l+u}))$, as shown in Fig. 3A-B. The pairwise similarity measures $w_{i,j}$ are computed using Euclidean norm between pairs of data points $(x^{(1)}(t_i), x^{(2)}(t_i))$ and $(x^{(1)}(t_j), x^{(2)}(t_j))$. Then, using Equation (2) it is possible to estimate $y = f(x)$ on the set of unlabeled data points using the harmonic extension algorithm. The results are shown in Fig. 3C. We then directly obtain y as a function of t by composition using the known time stamps $(t_{l+1}, ..., t_{l+u})$.

The accuracy of the estimated multivariable dynamics can be assessed using a K-fold validation strategy on the labeled samples (Fig. S1 and *Methods and Materials*). For chosen set of parameters and the size of the dataset, the error is $\sim 1\%$. As expected for the semi-supervised learning framework, the error decreases with the addition of new unlabeled data points. This example demonstrates how the proposed approach successfully recovers multivariable dynamics from heterogeneous datasets that combine continuous views for part of the state variables and snapshots that report several states without direct temporal information.

Fusion of imaging datasets. As a representative dataset from imaging studies of multivariable dynamics in living systems, we use a collection of ~ 1000 images each of which reveals the spatial position of the nuclei and either a timestamp or the distribution of one or several components of the DV patterning network (Fig. 1D). To apply the semi-supervised learning approach to data fusion to this dataset we need to compute pairwise similarities between the images using the common channel. Prior to this, we took several preprocessing steps that aim to minimize image variability associated

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

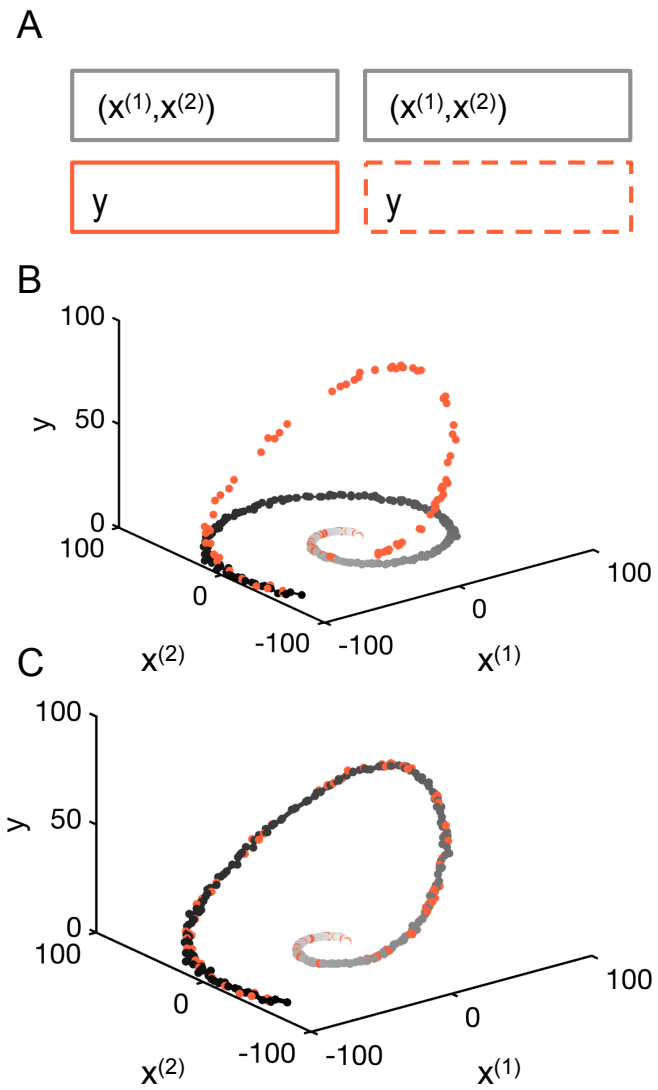

Fig. 3. Illustrative example. (A) Matrix formulation of the problem with 120 labeled samples, $(((x^{(1)}(t_1), x^{(2)}(t_1)), y(t_1)), ..., ((x^{(1)}(t_l), x^{(2)}(t_l)), y(t_l)))$, and 300 unlabeled samples $(((x^{(1)}(t_{l+1}), x^{(2)}(t_{l+1})), t_{l+1}), ..., ((x^{(1)}(t_{l+u}), x^{(2)}(t_{l+u})), t_{l+u}))$. (B) The points are distributed on a non-linear 1-dimensional manifold in the -plane. Some points, the snapshots, contain a value for the signal . (C) Result of the interpolation on the nonlinear manifold using the harmonic extension algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

to sample handling, microscope calibration and imaging. First, the images were registered to align their ventral-most points. The images were then resized and cropped such that the embryos occupy 80% of the image. All images were resized to 100 by 100 pixels. To overcome local variations of image intensity, we computed a local average using a Gaussian kernel, and then renormalized the image by that value. We also applied a logistic function to the images to handle contrast variability, Fig. S2.

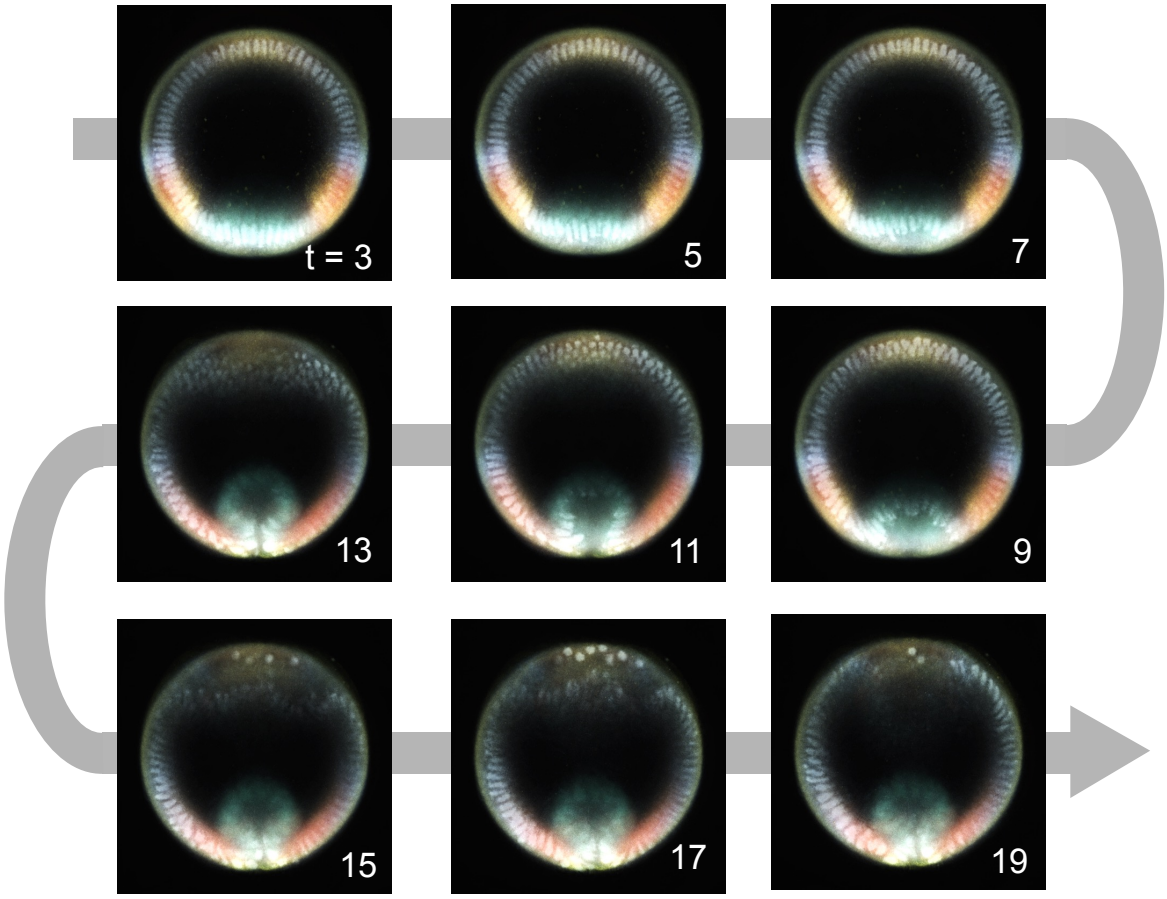

Fig. 4. Colored movie frames obtained with our data fusion algorithm. The temporal resolution is 2 min, extracted from a 30 s resolution movie. The time stamps on the images are in min and indicate elapsed time from the start of the live movie. The colors correspond to dpERK (red), Dl (pink), rho (yellow), ind (blue), Twi (green).

Most importantly, to ensure that pairwise differences between images are insensitive to small translations or deformations, we applied the scattering transform (22) and compared the resulting transform vectors. The scattering transform of an image is a signal representation obtained by alternating wavelet decompositions and pointwise modulus operators. We found that second-order scattering coefficients with an averaging scale of 64 pixels provided sufficient invariance. These are computed using the ScatNet toolbox (23, 24). The result is a vector of dimension 784 for each image. The point clouds corresponding to each of the 11 datasets were centered separately. The corresponding low-dimensional manifold on which the data points lie is shown on Fig. S4.

For each of the 512x512 pixels of each live movie frames, there is a common channel reporting the nuclei spatial position and there are 5 channels that we would like to complete. These channels contain the information about the spatial distributions of one enzyme (dpERK), two transcription

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

factors (Twist and Dorsal), and transcripts of two genes (ind and rho). We thus solved the data fusion problem for each pixel and each channel, leading to 5x512x512 semi-supervised learning solutions. The combination of labeled and unlabeled datasets is described on Table S2. The result is a multivariable trajectory for the joint dynamics of tissue shape and five molecular components within the regulatory network that patterns the DV axis of the embryo (Fig. 4). To evaluate the accuracy of the method, we computed the cross-validation error for each pixel and averaged over the entire images. We found that the normalized absolute error is of 0.9-2.5% of the signal range when considering the various modalities of the entire experimental datasets (Table S3). We show how the algorithm performs on several examples in Fig. 5.

Discussion. To summarize, we presented a formal approach to synthesizing developmental trajectories. By posing the task of data fusion as a semi-supervised learning problem, we obtained a closed form expression for the estimated values of all variables using harmonic extension. The reconstructed trajectories provide the basis for the more advanced mechanistic studies of multivariable processes responsible for the highly reproducible dynamics of developmental pattern formation. Our approach can also be extended to other semi-supervised learning methods (25), if the dimensionality of the intrinsic geometry is greater than one or if there is no unique common channel among all experiments. Most of the previous attempts to accomplishing this task explored specific features of developmental systems, such as the expression level of a particular gene, and used a discrete number of temporal classes, usually defined in ad hoc way (16, 26). Our approach reconstructs continuous time dynamics and relies on the intrinsic geometry of multidimensional datasets.

We conclude by pointing two directions for the future extensions and applications of the presented approach. First, while there are no conceptual limitations in using the presented matrix completion framework to studies of pattern formation and morphogenesis problems in three-dimensions(27), it is important to increase the computational efficiency of our approach, which can be done at multiple levels, starting with dimensionality reduction at the preprocessing step. At the same time, for a large class of patterning processes that happen on the surfaces of epithelial sheets, one can use the recently developed "tissue cartography" approach to first flatten the three-dimensional images (5), which should make our approach directly applicable. Second, following the step of data fusion, one can attempt to model the observed multivariable dynamics. Here one can employ several modeling methodologies, from mechanistic modeling of specific molecular and tissue-level processes (28–32), to equation-free approaches, which aim to deduce the underlying mechanisms directly from data (33, 34). Going beyond working with already existing datasets, it will be important to establish an adaptive data assimilation framework, whereby data collection itself is guided by the future need to reconstruct the multivariable trajectories. These ideas are already being discussed in the context of transcriptional profiling studies of cell differentiation and can be extended to synthesis of multivariable trajectories from imaging studies (35).

Methods and Materials

Extended methods and materials are presented in *SI Appendix*.

Image datasets. All images are cross-sections of Drosophila embryos taken at $\sim 90\mu$m from the posterior pole. Time-lapse movies were obtained using a Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective. The nuclei were stained with Histone-RFP. A total of 7 movies was acquired with a time resolution of 30 seconds per frame. All movies start about 2.5 hr after fertilization and end after about 20 min after gastrulation starts (about 3.3 hr after fertilization). Four

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

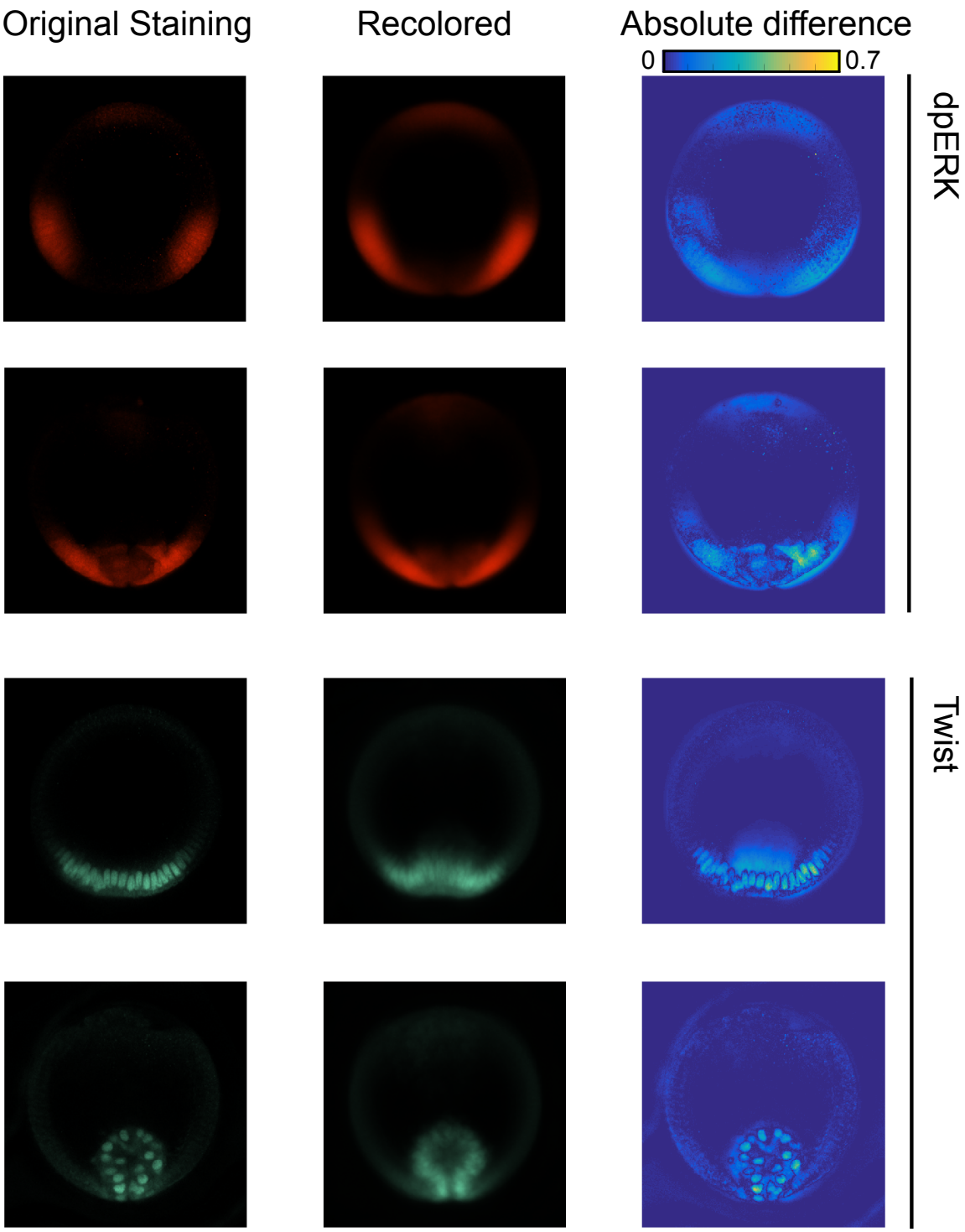

Fig. 5. Examples showing how the algorithm performs on four fixed samples. The first column shows the original measurement, the second column shows the result of recoloring the snapshot through K-fold cross validation, and the third column shows the absolute difference between the original and recolored images, normalized by the signal range.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

datasets of fixed images were acquired to visualize nuclei, protein expression of dpERK, Twist, and Dorsal, and mRNA expression of ind and rho. Immunostaining and fluorescent in situ hybridization protocols were used as described before13. DAPI (1:10,000; Vector laboratories) was used to visualize nuclei. Rabbit anti-dpERK (1:100; Cell Signaling), mouse anti-Dorsal (1:100; DSHB), rat anti-Twist (1:1000; gift from Eric Wieschaus, Princeton University), sheep anti-digoxigenin (1:125; Roche), and mouse anti-biotin (1:125; Jackson Immunoresearch) were used as primary antibodies. Alexa Fluor conjugates (1:500; Invitrogen) were used as secondary antibodies. Stained embryos were imaged using Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective. Embryos were mounted in a microfluidic device for end-on imaging, as described previously (16, 36). The first dataset contains 108 images stained with rabbit anti-dpERK and rat anti-Twist antibodies. The second dataset contains 59 images stained with mouse anti-Dorsal antibody, rabbit anti-dpERK antibody, and ind-DIG probe. The third dataset contains 58 images stained with ind-biotin probe, rho-DIG probe, and rabbit-dpERK antibody. The fourth dataset contains 30 images stained with rat anti-Twist antibody, ind-biotin probe, and rho-DIG probe. The distribution of the datasets as labeled and unlabeled data depending on the considered variable is summarized on Table S2. Raw images are can be found in Supplementary Files 1 or on the public github repository https://github.com/paulvill/data-fusion-images.

The affinity matrix. The affinity matrix $W = (w_{i,j})$ is computed using a Gaussian kernel $w_{i,j} = \exp(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j})$ with scaling parameters σ_i and σ_j computed locally as the average of the distance with respect to the 10 closest neighbors as described in SI appendix. We used the Euclidean norm in the space of scattering transformed data. The resulting affinity matrix is shown on Fig. S3. The corresponding underlying one-dimensional manifold is shown on Fig. S4.

Computing the cross validation error. The K-fold cross validation error was computed by extracting subsamples of the labeled data points and the semi-supervised learning framework was used to predict the value of the labels on them. For the image datasets, we computed the absolute error between the actual value of pixel intensity to the predicted one. The absolute error was then normalized by the range of the signal computed from the entire set of images for a given channel. The number of bins K was chosen so that the number artificially unlabeled data points was about 20. The results for each dataset are shown in Table S3 and described in SI appendix.

Movie coloring. The result of data fusion led to multimodal time lapses of developing embryo showing nuclei and the spatio-temporal dynamics of dpERK, Dl, rho, ind, and Twi. The images were colored using the color code shown in Table S4, i.e. dpERK (red), Dl (pink), rho (yellow), ind (blue), Twi (green). A resulting colored movie is provided in Supplementary Files 2.

Code implementation. The semi-supervised framework used to accomplish the task of data fusion is completely implemented in the open-source MATLAB library and fully runs in GNU Octave. It is available as Supplementary Software or on the public github repository https://github.com/paulvill/data-fusion.

Acknowledgments

PV was partially supported by the WIN program between Princeton University and the Weizmann Institute. SYS and YGK were supported by the NSF award MCB 1516970. YGK and AS were supported by NSF award CMMI-1310173. JA and AS are supported by Simons Investigator Award

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and Simons Collaboration on Algorithms and Geometry from Simons Foundation, and the Moore Foundation Data-Driven Discovery Investigator Award to AS. HL was supported by the GM088333 award from the NIH. We thank Carmeline Dsilva, Mahim Misra, Roy Lederman and David Sroczynski for helpful discussions.

1. Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin JC (2016) Imagining the future of bioimage analysis. *Nature Biotechnology* 34(12):1250–1255.
2. Crosetto N, Bienko M, Van Oudenaarden A (2015) Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics* 16(1):57–66.
3. de Bakker BS, et al. (2016) An interactive three-dimensional digital atlas and quantitative database of human development. *Science* 354(6315):aag0053.
4. Glasser MF, et al. (2016) A multi-modal parcellation of human cerebral cortex. *Nature*.
5. Heemskerk I, Streichan SJ (2015) Tissue cartography: compressing bio-image data by dimensional reduction. *Nature methods*.
6. Trapnell C, et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* 32(4):381–386.
7. Haghverdi L, Buettner M, Wolf FA, Buettner F, Theis FJ (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* 13(10):845–848.
8. Rizvi AH, et al. (2017) Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*.
9. Cutrale F, et al. (2017) Hyperspectral phasor analysis enables multiplexed 5d in vivo imaging. *Nature Methods*.
10. Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* 33(5):495–502.
11. Thompson CL, et al. (2014) A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. *Neuron* 83(2):309–323.
12. Castro-González C, et al. (2014) A digital framework to build, visualize and analyze a gene expression atlas with cellular resolution in zebrafish early embryogenesis. *PLoS Comput Biol* 10(6):e1003670.
13. Lim B, Levine M, Yamakazi Y (2017) Transcriptional pre-patterning of drosophila gastrulation. *Current Biology*.
14. Martin AC, Kaschube M, Wieschaus EF (2009) Pulsed contractions of an actin–myosin network drive apical constriction. *Nature* 457(7228):495–499.
15. Gilmour D, Rembold M, Leptin M (2017) From morphogen to morphogenesis and back. *Nature* 541(7637):311–320.
16. Lim B, et al. (2015) Dynamics of inductive erk signaling in the drosophila embryo. *Current Biology* 25(13):1784–1790.
17. Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions in *Proceedings of the 20th International conference on Machine learning (ICML-03)*. pp. 912–919.
18. Belkin M, Niyogi P (2004) Semi-supervised learning on riemannian manifolds. *Machine learning* 56(1-3):209–239.
19. Surkova S, et al. (2009) Canalization of gene expression in the drosophila blastoderm by gap gene cross regulation. *PLoS Biol* 7(3):e1000049.
20. Dsilva CJ, et al. (2015) Temporal ordering and registration of images in studies of developmental dynamics. *Development* 142(9):1717–1724.
21. Zhu X (2005) Semi-supervised learning literature survey.
22. Mallat S (2012) Group invariant scattering. *Comm. Pure Appl. Math.* 65(10):1331–1398.
23. Sifre L, Mallat S (2013) Rotation, scaling and deformation invariant scattering for texture discrimination in *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1233–1240.
24. Andén J, Mallat S (2014) Deep scattering spectrum. *IEEE Transactions on Signal Processing* 62(16):4114–4128.
25. Moscovich A, Jaffe A, Nadler B (2016) Minimax-optimal semi-supervised regression on unknown manifolds. *arXiv preprint arXiv:1611.02221*.
26. Fowlkes CC, et al. (2008) A quantitative spatiotemporal atlas of gene expression in the drosophila blastoderm. *Cell* 133(2):364–374.
27. Royer LA, et al. (2016) Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. *Nature Biotechnology*.
28. Goyal Y, et al. (2017) Divergent effects of intrinsically active mek variants on developmental ras signaling. *Nature Genetics* 49(3):465–469.
29. Félix MA, Barkoulas M (2015) Pervasive robustness in biological systems. *Nature Reviews Genetics* 16(8):483–496.
30. Lei J, Levin SA, Nie Q (2014) Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. *Proceedings of the National Academy of Sciences* 111(10):E880–E887.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

31. Kicheva A, et al. (2014) Coordination of progenitor specification and growth in mouse and chick spinal cord. *Science* 345(6204):1254927.
32. Wunderlich Z, DePace AH (2011) Modeling transcriptional networks in drosophila development at multiple scales. *Current opinion in genetics & development* 21(6):711–718.
33. Yair O, Talmon R, Coifman RR, Kevrekidis IG (2016) No equations, no parameters, no variables: data, and the reconstruction of normal forms by learning informed observation geometries. *arXiv preprint arXiv:1612.03195*.
34. Brunton SL, Proctor JL, Kutz JN (2016) Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* 113(15):3932–3937.
35. Kleyman M, et al. (2017) Selecting the most appropriate time points to profile in high-throughput studies. *eLife* 6:e18541.
36. Levario TJ, Zhan M, Lim B, Shvartsman SY, Lu H (2013) Microfluidic trap array for massively parallel imaging of drosophila embryos. *Nature protocols* 8(4):721–736.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Synthesizing developmental trajectories - Supporting Information Appendix

Paul Villoutreix[a,1], **Joakim Andén**[b,1], **Bomyi Lim**[a,c], **Hang Lu**[d,e], **Ioannis G. Kevrekidis**[b,c], **Amit Singer**[b,f], **and Stanislav Y. Shvartsman**[a,b,2]

[a]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544; [b]Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544; [c]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544; [d]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [e]Interdisciplinary Program in Bioengineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [f]Department of Mathematics, Princeton University, Princeton, NJ 08544, USA; [1]P.V. and J.A. contributed equally to this work.; [2]To whom correspondence should be addressed. E-mail: stas@princeton.edu.

1. Semi-supervised learning

Problem Setting. We define the problem of data fusion using semi-supervised learning in general terms. A classical setting is to have a modality x belonging to some space $\mathcal{X}$ which is common to all experiments and M other modalities $(y^{(1)}, \ldots, y^{(M)})$ that vary across experiments. The difficulty here is that only a certain number of modalities can be measured simultaneously, so our goal is to fill in the missing modalities. The critical assumption that we use is that the common modality x is sufficient to determine the other modalities. In other words, there exists a set of functions $f_*^{(1)}, \ldots, f_*^{(M)}$ such that $y^{(m)} = f_*^{(m)}(x)$ for all $1 \leq m \leq M$ and all x.

Let us represent the ith experiment, here containing the common modality x and the modalities m_1, and m_2 as an example by

$$\begin{pmatrix} x_i \\ \times \\ \vdots \\ \times \\ y_i^{(m_1)} \\ \times \\ \vdots \\ \times \\ y_i^{(m_2)} \\ \times \\ \vdots \\ \times \end{pmatrix},$$

where $\times$ denotes an empty, or unmeasured, modality. The corresponding completed vector would then be

$$\begin{pmatrix} x_i \\ f^{(1)}(x_i) \\ \vdots \\ f^{(M)}(x_i) \end{pmatrix}$$

where $f^{(1)}, \ldots, f^{(M)}$ are our estimates of the underlying functions $f_*^{(1)}, \ldots f_*^{(M)}$ with $f^{(m_1)}(x_i) = y_i^{(m_1)}$ and $f^{(m_2)}(x_i) = y_i^{(m_2)}$.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Problem formulation: The problem of fusing multiple modalities can be summarized as completing the following matrix using the information on the row x while preserving the continuity of the underlying manifold.

$$\begin{pmatrix} x_1 & \dots & x_n \\ y_1^{(1)} & \dots & y_n^{(1)} \\ & \vdots & \\ y_1^{(M)} & \dots & y_n^{(M)} \end{pmatrix}.$$

For the mth modality, we denote the set of experiments where it is measured by $\Omega(m)$. This is a subset of $\{1,\dots,n\}$, in which it has the complement $\overline{\Omega(m)} = \{1,\dots,n\}\backslash\Omega(m)$. We shall refer to $\Omega(m)$ as the set of labeled datapoints, while $\overline{\Omega(m)}$ is the set of unlabeled data points for the mth modality. Note that we do not impose any particular structure on the sets of labeled points $\Omega(m)$. They can occur in any order and vary in size between the different modalities.

Multiple semi-supervised learning by harmonic extension The problem of completing the mth row of the matrix using the measured elements $Y^{(m)} = \{y_i^{(m)} : i \in \Omega(m)\}$ can be posed as a semi-supervised learning problem, which consists in learning the mapping $f^{(m)}(x) = y^{(m)}$ on each of the points $(x_1, ..., x_n)$ (1). To simplify the notation, we shall consider the n values of $f^{(m)}$ on these points as a column vector in $\mathbb{R}^n$. In addition, we denote the subset of its values when restricted to $\Omega(m)$ and $\overline{\Omega(m)}$ by $f^{(m)}|_{\Omega(m)} \in \mathbb{R}^{|\Omega(m)|}$ and $f^{(m)}|_{\overline{\Omega(m)}} \in \mathbb{R}^{|\overline{\Omega(m)}|}$, respectively, where $|A|$ denotes the number of elements in the set A. Since $Y^{(m)}$ is also a column vector in $\mathbb{R}^{|\Omega(m)|}$, we formulate our problem as

$$f^{(m)} = \underset{\substack{f\in\mathbb{R}^n \\ f|_{\Omega(m)}=Y^{(m)}}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\|f(x_i) - f(x_j)\|^2 \qquad [1]$$

where $w_{i,j}$ is a measure of the affinity between x_i and x_j that varies between 0 and 1, we compute it in 6.

This problem can be rewritten in terms of a matrix for each row (modality) m. First we introduce the Laplacian matrix L, given by

$$L = D - W$$

where $D = \operatorname{diag}(d_1,\dots,d_n)$ is the diagonal matrix with entries $d_1,\dots,d_n$ and $d_i = \sum_{j=1}^{n} w_{i,j}$ and $W = (w_{i,j})_{1\le i,j\le n}$. With this, our optimization problem can be rewritten as

$$f^{(m)} = \underset{\substack{f\in\mathbb{R}^n \\ f|_{\Omega(m)}=Y^{(m)}}}{\operatorname{argmin}} \quad f^{\mathrm{T}}Lf. \qquad [2]$$

This is a quadratic optimization problem with linear constraints and can be solved by differentiating and setting the derivative to zero. The solution $f^{(m)}$, on the subsets $\Omega(m)$ and $\overline{\Omega(m)}$, is given by

$$\begin{aligned} f^{(m)}|_{\Omega(m)} &= Y^{(m)} \\ f^{(m)}|_{\overline{\Omega(m)}} &= \left(D|_{\overline{\Omega(m)},\overline{\Omega(m)}} - W|_{\overline{\Omega(m)},\overline{\Omega(m)}}\right)^{-1} W|_{\overline{\Omega(m)},\Omega(m)} Y^{(m)} \end{aligned}, \qquad [3]$$

where we have adopted the notation of $M|_{A,B}$ to denote the submatrix of M obtained by extracting the rows in the index set A and the columns in index set B.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The solution Eq. (3) can be rewritten as

$$f^{(m)}(i) = Y^{(m)}(i) \quad \forall i \in \Omega(m) \qquad [4]$$

and

$$f^{(m)}(i) = \frac{1}{D(i,i)} \sum_{j=1}^{n} w_{i,j} f^{(m)}(j) \quad \forall i \in \overline{\Omega(m)}. \qquad [5]$$

Note that Eq. (5) is an implicit equation satisfied by the values. To calculate the desired values of $f^{(m)}$, Eq. (3) must be used.

The above calculations are performed separately for each modality $m = 1, \ldots, M$, each solution completing a different row in the overall measurement matrix.

Computing the Affinity Matrix. Given a set of point $(x_1, ..., x_n)$, we compute the affinity matrix $W = (w_{i,j})$ with

$$w_{i,j} = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i \sigma_j}\right) \qquad [6]$$

where the scaling parameters σ_i and σ_j are defined as $\sigma_t = \frac{1}{|\mathcal{N}(x_t)|} \sum_{k \in \mathcal{N}(x_t)} \|x_t - x_k\|$ for every $t = 1, \ldots, n$, where $\mathcal{N}(x_t)$ represents the neighborhood composed of the $|\mathcal{N}(x_t)| = 10$ closest points of x_t (2). The norms here are Euclidean norms (3). This gives us a proximity scale that varies across the manifold if the sampling of the points is not uniform.

K-Fold Cross Validation. To evaluate the accuracy of the reconstruction obtained by data fusion we employ K-fold cross validation (4). This approach estimates the expected prediction error. The idea is to predict the value of labels on a subset of the labeled data points and compare the prediction to the known labels.

Let us consider labeled data points $x|_{\Omega(m)}$ with labels $Y^{(m)}$ and unlabeled data points $x|_{\overline{\Omega(m)}}$. The labeled data points can be uniformly assigned to K bins randomly using an indexing function $\kappa : \Omega(m) \to \{1, ..., K\}$. For each value k of κ, we thus have the index set $\kappa^{-1}(k) = \{i \in \Omega(m) : \kappa(i) = k\}$, a subset of $\Omega(m)$. We can thus define reduced set of labeled measurements $\Omega_k(m) = \Omega(m) \backslash \kappa^{-1}(k)$, with the corresponding set of unlabeled measurements $\overline{\Omega_k(m)} = \overline{\Omega(m)} \cup \kappa^{-1}(k)$.

For $i \in \Omega(m)$, we denote by $f^{(m),-\kappa(i)}(i)$ the label prediction on data point i where $\Omega_{\kappa(i)}(m)$ and $\overline{\Omega_{\kappa(i)}(m)}$ are used as the set of labeled and unlabeled points, respectively. Since $y_i^{(m)}$ is known, we can evaluate the performance of our method by comparing the prediction $f^{(m),-\kappa(i)}(i)$ to the known value $y_i^{(m)}$. Repeating this for all $i \in \Omega(m)$, we calculate the average absolute error normalized by the range for each modality m

$$\mathrm{Err}^{(m)}(f) = \left(\frac{1}{|\Omega(m)|} \sum_{i \in \Omega(m)} |f^{(m),-\kappa(i)}(i) - y_i^{(m)}|\right) / \left(\max(Y^{(m)}) - \min(Y^{(m)})\right). \qquad [7]$$

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2. Illustrative example

To demonstrate the efficiency of this technique as an interpolation method on a non-linear manifold, we first consider a toy example with a non-linear 1-dimensional trajectory embedded in 3-dimensional space. Specifically, we have

$$\begin{cases} x^{(1)}(t) &= at\left(\cos(bt) + \epsilon^{(1)}\right) \\ x^{(2)}(t) &= at\left(\sin(bt) + \epsilon^{(2)}\right) \\ y(t) &= ct\exp(-d(t-e)^2) \end{cases} \quad [8]$$

where, a, b, c, d, e are constants, $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are Gaussian noise sources and t is a real-valued parameter. The set of points $(x^{(1)}(t), x^{(2)}(t))$ forms a 1-dimensional non-linear manifold embedded in the 2-dimensional space as it is parameterized by t. These points correspond to the embryo morphology in our toy example. In the absence of noise, this mapping from t to the 2D plane can be inverted as $t = \frac{1}{|a|}\sqrt{(x^{(1)})^2(t) + (x^{(2)})^2(t)}$. The signal $y(t)$ is a smooth function of t and is thus a smooth function of $(x^{(1)}(t), x^{(2)}(t))$ by composition. In this example, y corresponds to the target modality that we would like to estimate.

To reproduce the setting of data fusion with three modalities, $((x^{(1)}, x^{(2)}), t, y)$, we consider the following situation. Suppose that one can acquire a set of labeled points, i.e. a set of l triplets, $((x^{(1)}(t_1), x^{(2)}(t_1)), y(t_1)), ..., ((x^{(1)}(t_l), x^{(2)}(t_l)), y(t_l))$ and a set of u unlabeled, but timestamped, points, $((x^{(1)}(t_{l+1}), x^{(2)}(t_{l+1})), t_{l+1}), ..., ((x^{(1)}(t_{l+u}), x^{(2)}(t_{l+u})), t_{l+u})$, as shown in Fig. 1A and B. We therefore have $\Omega = \{1, \ldots, l\}$ and $\overline{\Omega} = \{l+1, \ldots, l+u\}$ and $n = l + u$ (since we only consider y as labels, and hence have only one modality to fill, we have suppressed the superscript (m)). The pairwise similarity measures $w_{i,j}$ are computed using Euclidean norm between pairs of data points $(x^{(1)}(t_i), x^{(2)}(t_i))$ and $(x^{(1)}(t_j), x^{(2)}(t_j))$ as presented in 6. Then using the framework presented above, in particular equations Eq. (4) and Eq. (5), it is possible to estimate $y = f(x)$ on the set of unlabeled data points using the harmonic extension algorithm. The results are shown in Fig. 1C. We then directly obtain y as a function of t by composition using the known time stamps $(t_{l+1}, ..., t_{l+u})$.

For this example, we now quantify the evolution of the error as a function of the number of unlabeled samples. The constants used were $a = 1$, $b = 0.1$, $c = 1$, $d = 0.005$, $e = 65$, $l = 120$, and $u = 300$ while t varies from 1 to 100 and $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are sampled according to centered Gaussian distributions with standard deviation 2.02. Using a K-fold validation strategy on the labeled samples, we characterize the accuracy of the reconstruction using the normalized absolute error as described by equation Eq. (7) (Fig. S1A). The results are shown on Fig. S1B. We found that with no additional data points, the absolute error was on average 1.05% of the signal range, and decreased uniformly to 0.53% when considering the problem with an additional 300 unlabeled data points.

3. Experimental Data Sets

Data Sets description. To explore the capabilities of our data fusion strategy, we considered 11 different datasets, each of which consisting of a set of images, obtained either as live movies or fixed samples, and containing a channel recording the spatial distribution of nuclei.

Live Movies. Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective was used to obtain time-lapse movies of Drosophila embryonic development. Live imaging of Histone-RFP embryos was used to visualize nuclei. A total of 7 movies were obtained, with a time resolution of 30 seconds per frame. All the movies start during nuclear cycle 14 (about 2.5 hr after fertilization) and end after about 20 min after gastrulation starts (about 3.3 hr after fertilization). The number of frames for each movie is 156, 67, 53, 55, 50, 86, and 90.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fixed samples. In our experiments, 4 datasets were acquired to visualize nuclei, protein expression of dpERK, Twist, and Dorsal, and mRNA expression of ind and rho. Immunostaining and fluorescent in situ hybridization protocols were used as described before (5). DAPI (1:10,000; Vector laboratories) was used to visualize nuclei. Rabbit anti-dpERK (1:100; Cell Signaling), mouse anti-Dorsal (1:100; DSHB), rat anti-Twist (1:1000; gift from Eric Wieschaus, Princeton University), sheep anti-digoxigenin (1:125; Roche), and mouse anti-biotin (1:125; Jackson Immunoresearch) were used as primary antibodies. Alexa Fluor conjugates (1:500; Invitrogen) were used as secondary antibodies. Stained embryos were imaged using Nikon A1-RS confocal microscope with a 60x Plan-Apo oil objective. Embryos were mounted in a microfluidic device for end-on imaging, as described previously (5, 6). All the images were taken at $\sim$90 μm from the posterior pole of the embryo.

Dataset 1 Stained with rabbit anti-dpERK and rat anti-Twist antibodies. Number of samples: 108.

Dataset 2 Stained with mouse anti-Dorsal antibody, rabbit anti-dpERK antibody, and ind-DIG probe. Number of samples: 59.

Dataset 3 Stained with ind-biotin probe, rho-DIG probe, and rabbit-dpERK antibody. Number of samples: 58.

Dataset 4 Stained with rat anti-Twist antibody, ind-biotin probe, and rho-DIG probe. Number of samples: 30.

Preprocessing Steps. In order to compare the images of morphology (nuclei channel) among all the datasets, we used a certain number of preprocessing steps to overcome measurement-dependent variability.

Image Rotation. Images were oriented so that the highest Dorsal signal is located at the ventral-most point. Some late-stages embryos were oriented manually.

Image Resizing. The images were resized and cropped such that the embryo would occupy 80% of the image. This was achieved by thresholding the images and computing their bounding box, then resizing and translating the image so that the bounding box occupied the square stretching from 10% to 90% of the image width, both horizontally and vertically. The images were resized to 100 by 100 pixels.

Intensity Renormalization. Since there is some local variation of image intensity that is due to the device used to acquire data and the configuration of the microscope, we wish to remove this from the images before comparing them. This is done by computing a local average using a Gaussian kernel, and then renormalizing the image by that value. Specifically, given an image x, we calculate its local average $y(u) = x \star \phi(u)$, where $\phi(u) = \mathrm{e}^{-u^2/2c^2}$ for some kernel width $c > 0$ and $\star$ denotes convolution. Then the renormalized image is given, for each pixel u, by

$$\frac{y(u)}{y(u)^2 + \epsilon^2} x(u), \qquad [9]$$

where $\epsilon > 0$ is some threshold that prevents the renormalization from exploding when the image is locally close to zero.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Contrast Increase. Another consequence of the experimental setup and the configuration of the microscope is that the contrast differs between measurements. As a result, the range of pixel intensities is not the same for different images of embryos at the same developmental stage. To fix this, we pass the pixel intensities through a logistic function with two parameters. Given an input image x, it is adjusted to give, for each pixel u,

$$\frac{1}{1+\mathrm{e}^{-a(x(u)-b)}}. \qquad [10]$$

Scattering Transform. To compare images without being sensitive to small translations or deformations, we applied the scattering transform (7) and compared the resulting transform vectors. Differences between images in the scattering space are meant to reflect changes in the underlying morphology as opposed to variation due to the imaging method.

The scattering transform of an image is a signal representation obtained by alternating wavelet decompositions and pointwise modulus operators. The resulting coefficients are locally invariant to translations and stable to deformations. The similarity of the signals can then be measured by computing the Euclidean distance of their scattering transforms. These transforms have enjoyed significant success in classification of images (8, 9), but also time series (10), since these tasks often requires insensitivity to translation and deformation.

For our embryo images, we have found that second-order scattering coefficients with an averaging scale of 64 pixels allow for the proper amount of invariance. These are computed using the ScatNet toolbox (11).

The result is a vector of dimension 784 for each image. The point clouds corresponding to each of the 11 datasets were centered separately.

Computing the affinity matrix. The affinity matrix was computed according to the definition of the Gaussian kernel with an adaptive scaling factor, as described in 6. The result is shown on Fig. S3.

Outlier filtering. Even after applying all the previous steps of image preprocessing some data points were considerably away from the main manifold. First we calculated the closest neighbor of each point. We then kept the points whose closest neighbor distance was less than 3 times the median minimal neighbor distance.

Results. To fuse the channels of the various datasets described in 3, we applied the harmonic extension algorithm described in 1. The data points are preprocessed images of nuclei spatial distribution using the steps presented in 3 to 3 with parameter values as shown in Table S1 and the modalities correspond to the images obtained from the various fluorescent reporters described in section 3 of the SI appendix. The corresponding low-dimensional manifold on which the data points x_i lie is shown on Fig. S4.

We denote by x_i the preprocessed image i of the nuclei channel, it is an element of $\mathbb{R}^{784}$ as the result of scattering transformation applied to a 100×100 image. We denote by $y_i^{(m)}$ the image corresponding to modality m associated to x_i and by $y_i^{(m)}(p)$ its restriction to the pixel p, $y_i^{(m)}(p)$ is an element of $\mathbb{R}$.

For each of the 10000 pixels in the live movie, there are 5 modalities that we would like to fuse: dpERK ($m = 1$), Twist ($m = 2$), Dorsal ($m = 3$), ind ($m = 4$), rho ($m = 5$). We therefore solve the data fusion problem for each pixel and each modality, leading to 50000 semi-supervised learning solutions. The combination of labeled and unlabeled datasets is described on Table S2.

To evaluate the accuracy of the method, we computed the cross-validation error as described by equation Eq. (7) for each pixel and averaged over the entire images. The range of the signal is

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

calculated based on the entire images. Table S3 shows the results for each of the datasets where the number of unlabeled data points is 309 (the number of live movie frames), the unlabeled data points are chosen randomly among the live movie frames and the fixed images from the 4 datasets which serve as unlabeled data points. For each of the datasets, we chose K such the number of points within each bin was about 20. We have $K = 6, 3, 3, 2$ for datasets 1 to 4 respectively.

Coloring movies. The resulting fused dataset led to the construction of a multimodal movie, where a different color was attributed to each modality and the RGB values were added in the resulting movie. We used the set of colors presented on Table S4.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions in *Proceedings of the 20th International conference on Machine learning (ICML-03)*. pp. 912–919.
2. Zelnik-Manor L, Perona P (2004) Self-tuning spectral clustering. *Advances in neural information processing systems* 17(1601-1608):16.
3. Lederman RR, Talmon R (2014) Common manifold learning using alternating-diffusion. *Yale University, New Haven, CT, USA, Tech. Rep. YALEU/DCS/TR-1497*.
4. Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning*. (Springer series in statistics Springer, Berlin) Vol. 1.
5. Lim B, et al. (2015) Dynamics of inductive erk signaling in the drosophila embryo. *Current Biology* 25(13):1784–1790.
6. Levario TJ, Zhan M, Lim B, Shvartsman SY, Lu H (2013) Microfluidic trap array for massively parallel imaging of drosophila embryos. *Nature protocols* 8(4):721–736.
7. Mallat S (2012) Group invariant scattering. *Comm. Pure Appl. Math.* 65(10):1331–1398.
8. Bruna J, Mallat S (2013) Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8):1872–1886.
9. Sifre L, Mallat S (2013) Rotation, scaling and deformation invariant scattering for texture discrimination in *IEEE Conf. on Comput. Vis. and Pattern Recognit.* (IEEE), pp. 1233–1240.
10. Andén J, Mallat S (2014) Deep scattering spectrum. *IEEE Trans. Sig. Proc.* 62:4114–4128.
11. Andén J, et al. (2014) Scatnet. *Computer Software. Available: http://www.di.ens.fr/data/software/scatnet/*.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Files 1: This compressed folder contains all the raw images organized into folders: movie1-7, and data_set1-4. The folders movie1-7 correspond to the live movies, while the folders data_set1-4 correspond to stained fixed images. The description of the datasets can be found within each folder or in section 3 of the SI appendix.

Supplementary Files 2: Colored movie. The compressed folder contains frames of a colored movie, with, in the main folder, all the channels combined, nuclei (gray), dpERK (red), Dl (pink), rho (yellow), ind (blue), Twi (green), using the color map presented in Table S4. The number of each image corresponds to a time step, and they have been extracted from movie 5. The subfolder "grayscale" contains the actual value of the variables in separate frames, the name of each image indicates the variable considered, while the number of each image corresponds to the time step.

Supplementary Software: The compressed folder contains the MATLAB library implementing the data fusion algorithm on an illustrative example and on experimental cross-sections of developing fly embryos. The file README.md contains instructions on how to run the code. The main scripts are found in the examples folder and consist of:

- spiral_example.m: implements the data fusion algorithm on a non-linear 1-dimensional trajectory in a 3-dimensional space, nicknamed "spiral".
- spiral_example_cross_validation.m: implements the K-fold cross-validation on the "spiral".
- experimental_dataset.m: implements the data fusion on the experimental datasets leading to a multimodal movie containing the spatio-temporal dynamics of 5 chemical species on top of morphological changes.
- experimental_dataset_cross_validation.m: implements the K-fold cross validation on the experimental datasets.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

	Live Movies	Dataset 1	Dataset 2	Dataset 3	Dataset 4
ϵ	0.01	0.01	0.01	0.05	0.01
c	10	10	10	5	10
a	10	10	10	15	10
b	0.8	1.0	1.0	0.5	1.0

Table S1. Values of the parameters for intensity renormalization (Section 3) and contrast increase (Section 3) for each of the experimental datasets.

	Live Movies	Dataset 1	Dataset 2	Dataset 3	Dataset 4
$m = 1$ dpERK	$\overline{\Omega(1)}$	$\Omega(1)$	$\Omega(1)$	$\Omega(1)$	$\overline{\Omega(1)}$
$m = 2$ Twist	$\overline{\Omega(2)}$	$\Omega(2)$	$\overline{\Omega(2)}$	$\overline{\Omega(2)}$	$\Omega(2)$
$m = 3$ Dorsal	$\overline{\Omega(3)}$	$\overline{\Omega(3)}$	$\Omega(3)$	$\overline{\Omega(3)}$	$\overline{\Omega(3)}$
$m = 4$ ind	$\overline{\Omega(4)}$	$\overline{\Omega(4)}$	$\Omega(4)$	$\Omega(4)$	$\Omega(4)$
$m = 5$ rho	$\overline{\Omega(5)}$	$\overline{\Omega(5)}$	$\overline{\Omega(5)}$	$\Omega(5)$	$\Omega(5)$

Table S2. Distribution of the datasets into labeled and unlabeled sets depending on the modality. We refer to $\Omega(m)$ as the set of labeled datapoints, while $\overline{\Omega(m)}$ is the set of unlabeled data points for the mth modality.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

	Normalized Absolute Error (in %) 309 unlabeled data points
Dataset 1	
dpERK	2.37
Twist	2.03
Dataset 2	
dpERK	2.01
Dorsal	1.10
ind	1.45
Dataset 3	
dpERK	1.57
ind	0.9
rho	1.11
Dataset 4	
Twist	2.18
ind	1.63
rho	1.84

Table S3. Normalized Absolute Error obtained by K-fold cross-validation for each modality of each dataset. In each case, we performed 10 repetitions, where the labeled samples are distributed randomly among the K bins, and the 309 unlabeled data points are chosen randomly. The error is then averaged over 10 repetitions. More details about the Normalized Absolute Error can be found in section 1 of the SI appendix and about how it is computed on the real datasets in section 3 of the SI appendix.

	R	G	B
dpERK	213	36	2
Twist	92	188	146
Dorsal	229	195	207
ind	103	139	191
rho	225	223	75

Table S4. Color scheme used to color the final movie.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

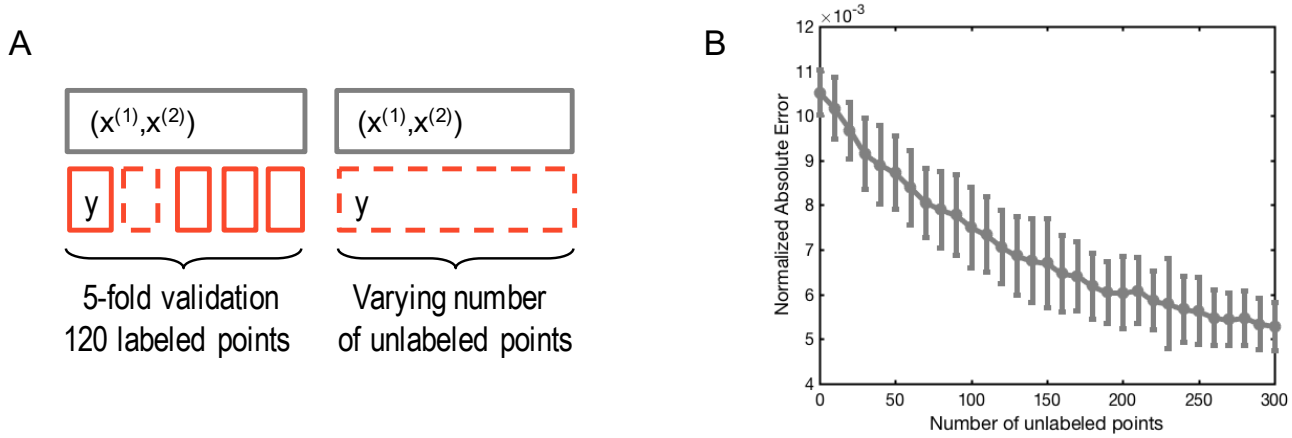

Fig. S1. K-fold cross validation on the illustrative example A) Setting with $K = 5$, there are 120 labeled points and the number of unlabeled points varies from 0 to 300. B) The normalized absolute error as a function of the number of unlabeled points. There are 100 repetitions for each number of unlabeled data points.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

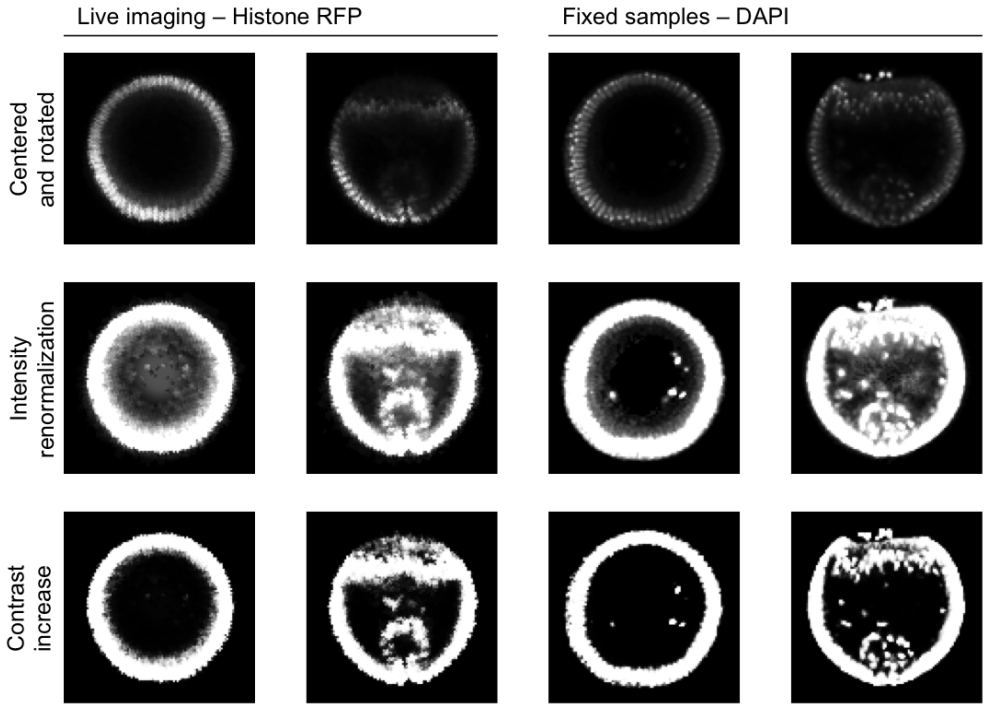

Fig. S2. Illustration of the image preprocessing steps applied on the nuclei channel and described in section S2 of the SI appendix. The first line shows images resulting from rotation and centering steps. The second line shows images resulting from intensity renormalization. The third line shows images resulting from contrast increase. The first two columns show early and later stages from movie frames stained with Histone-RFP. The last two columns represent early and later stages from fixed samples stained with DAPI.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

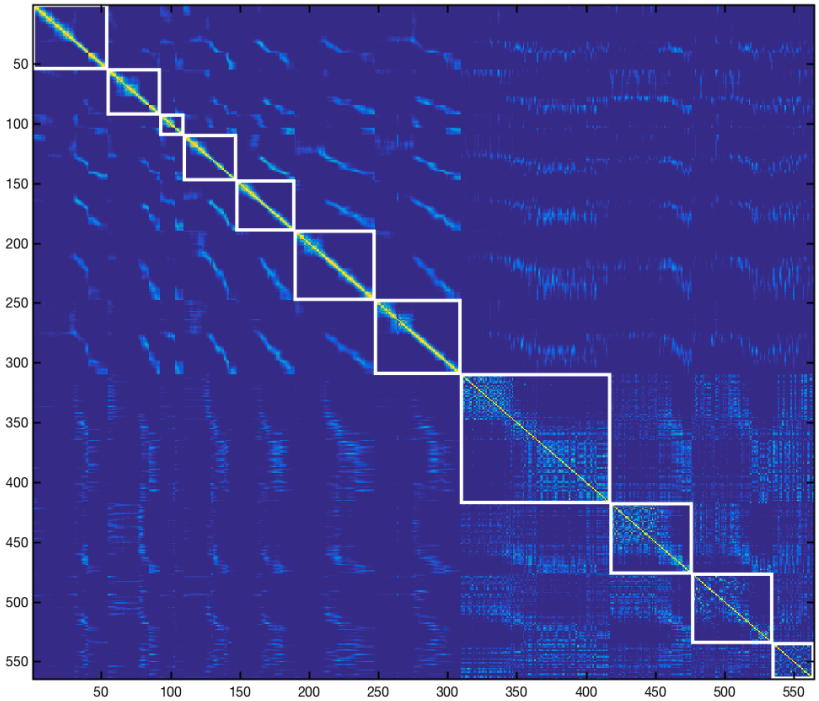

Fig. S3. The Affinity Matrix $W = (w_{i,j})$ obtained by comparing images as described by equation 6 in SI appendix is shown as a heatmap. The white squares identify each of the 11 datasets. The first 7 correspond to live movies, the last 4 correspond to the datasets of fixed images.

bioRxiv preprint doi: https://doi.org/10.1101/157834; this version posted June 30, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

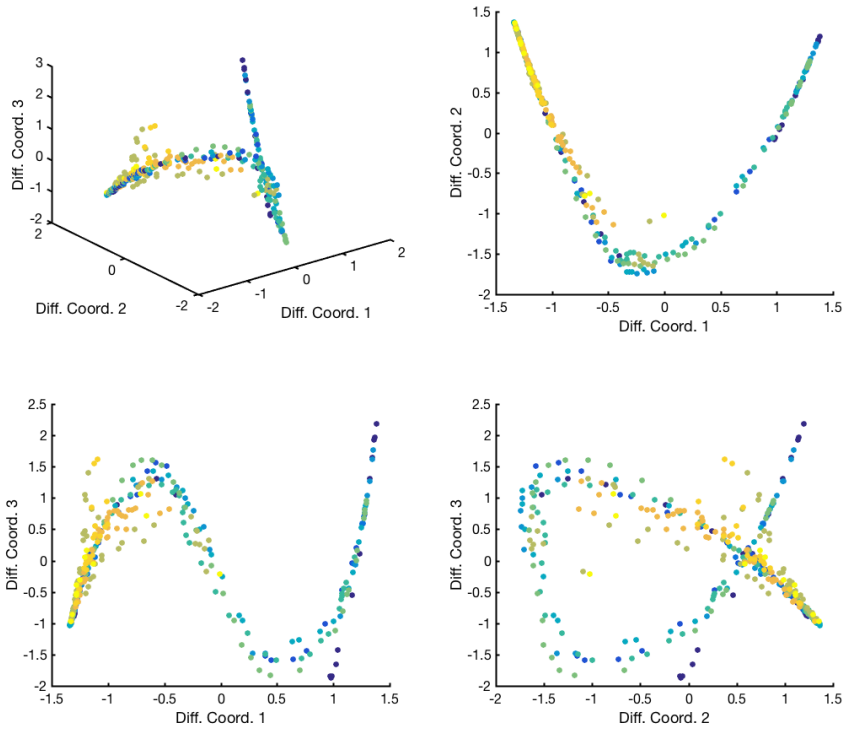

Fig. S4. Low-dimensional embedding of the 11 datasets obtained by diffusion maps. Each dot is a point and each color is a different dataset. The top left panel shows the points obtained by embedding the points in the first three diffusion map coordinates. The top right panel shows the data points in the plane formed by the first two diffusion map coordinates, while the two bottom panels show the embedding in the planes obtained with the first and third (left) or second and third (right) diffusion map coordinates. Some outliers were filtered out for visualization purposes if their closest neighbor distance was at least twice the median closest neighbor distance, leading to a very well-defined 1-dimensional manifold.