

1 Machine Learning Approaches Identify Genes  
2 Containing Spatial Information from Single-Cell  
3 Transcriptomics Data.

4

5 Phillipe Loher, PL, Computational Medicine Center, Thomas Jefferson University, Philadelphia,  
6 PA, 19107, USA

7 Nestoras Karathanasis, NK, Computational Medicine Center, Thomas Jefferson University,  
8 Philadelphia, PA, 19107, USA

9

10 NK and PL equally contributed to this work.

11 Corresponding Author: Nestoras Karathanasis

12

13

14

15

## 16 **Abstract**

17 **Motivation:** We participated in the DREAM Single Cell Transcriptomics Challenge. The  
18 challenge's focus was two-fold; a) to identify the top 60, 40 and 20 genes that contain the most  
19 spatial information, and b) to reconstruct the 3-D arrangement of the *D. melanogaster* embryo  
20 using information from those genes.

21 **Results:** We developed two independent approaches, leveraging machine learning models from  
22 Lasso and Deep Neural Networks, that we successfully apply to high-dimensional single-cell  
23 sequencing data. Our methods allowed us to achieve top performance when compared to the  
24 ground truth. Among ~40 participating teams, the resulting solutions placed 10th, 6th, and 4th in  
25 the three DREAM sub-challenges #1, #2 and #3, respectively. Notably, for the Lasso approach  
26 we introduced a feature selection technique, Lasso-TopX, that allows a user to define a specific  
27 number of features they are interested in and the Neural Network approach utilizes weak  
28 supervision for linear regression to accommodate for uncertain or probabilistic training labels.  
29 Furthermore, we identified novel *D. melanogaster* genes that carry important positional  
30 information and were not previously suspected. Lastly, we show how the indirect use of the full  
31 datasets' information can lead to data leakage and generate bias in overestimating the model's  
32 performance.

33 **Availability:** <https://github.com/TJU-CMC-Org/SingleCell-DREAM/>.

34 **Contact:** [Nestoras.Karathanasis@jefferson.edu](mailto:Nestoras.Karathanasis@jefferson.edu)

35

36

## 37 Introduction

38 Single-cell RNA sequencing (scRNA-seq) has been rapidly gaining popularity and allows  
39 biologists to gain knowledge about the abundance of genes for thousands of cells, individually,  
40 from a given tissue. Such an approach does not suffer from the drawback of standard approaches  
41 where the aggregation of a large starting population of cells obscure the ability to detect cell-to-  
42 cell variation. Unfortunately, scRNA-seq approaches do not typically maintain the spatial  
43 arrangement of the cells (Karaikos *et al.*, 2017).

44 For more than a decade, the DREAM (Stolovitzky *et al.*, 2007) initiative has driven crowd-  
45 sourced open science scientific contests in different areas of biology and medicine. Recently, the  
46 DREAM Single Cell Transcriptomics Challenge (Tanevski *et al.*, 2019), in which we  
47 participated, focused on tackling the reconstruction of the 3-D arrangement of cells using  
48 predefined number of genes. Specifically, the goal of this DREAM challenge was to use *D.*  
49 *melanogaster* embryo as a model system and seek to determine whether one can reconstruct the  
50 spatial arrangement of cells from a stage 6 embryo by only using a limited number of genes. The  
51 challenge piggy backed off previously published scRNA-seq datasets and a computational  
52 mapping strategy called DistMap, that leveraged in-situ hybridization data from 84 genes of the  
53 Berkeley Drosophila Transcription Network Project (BDTNP), which was shown to uniquely  
54 classify almost every position of the *D. melanogaster* embryo (Karaikos *et al.*, 2017). Out of  
55 these 84 genes (herein referred to as “inSitu genes”) and without using hybridization data, the  
56 participants were asked to identify the *most informative* 60, 40, and 20 genes for subchallenges  
57 #1, #2, and #3 respectively. In addition to gene selection, each subchallenge also required  
58 participants to submit 10 locations predictions (X, Y, Z coordinates) for each of the cells using  
59 only the selected genes (Tanevski *et al.*, 2019).

60 In order to identify the most informative genes, we describe two independent feature selection  
61 strategies. The first, we name Lasso-TopX, leverages linear models using the Least Absolute  
62 Shrinkage and Selection Operator (Lasso) (Friedman *et al.*, 2010; Tibshirani, 1996). Lasso has a  
63 few important characteristics that made it desirable to use. Specifically, the models are easy to  
64 interpret because each feature gets assigned a coefficient and the coefficients are combined  
65 linearly. It is also useful for dimensionality reduction because the resulting coefficients can be  
66 exactly zero, essentially eliminating features (James et al., 2013; Friedman et al., 2010). Our  
67 second feature selection strategy leverages Deep Neural Networks (NN). NNs are making major  
68 advances in problem solving by allowing computers to better discover structure in high-  
69 dimensional data (LeCun *et al.*, 2015). By linking multiple non-linear layers together, we sought  
70 to use Deep Learning in order to discover subsets of genes that would not have otherwise been  
71 possible with more traditional linear approaches.

72 In what follows, we describe our methods and the novel elements that allowed us to meet the  
73 objectives of the DREAM challenge. Notably, Lasso-TopX allows a user to specify the exact  
74 number of key features they are interested in. And to take advantage of Distmap's probabilistic  
75 mapping where a cell's location is not always unique, we also describe how NNs can be trained  
76 using weak supervision (Zhou, 2018) for use in linear regression. Importantly, while not an  
77 objective of the DREAM challenge, we extend our techniques to other genes by looking for *non-*  
78 *inSitu* genes that also carry spatial information.

79

80

81

## 82 **Methods**

83 In summary, we used two methodologies to identify the most informative features (*D.*  
84 *melanogaster* genes); an approach based on Deep Neural Networks (NN) models, and an  
85 approach based on Lasso models, which we call Lasso-TopX. Both are supervised approaches  
86 that use training data. We then utilized inference techniques on the trained models to get a list of  
87 the most important 60 / 40 / 20 inSitu genes. In order to help baseline our results prior to the end  
88 of the competition, we also leveraged a process (herein named Random) that picked genes  
89 randomly. For the selected genes using NN, Lasso-TopX, and Random, we passed only those  
90 genes into DistMap (Karaiskos *et al.*, 2017) to get the spatial predictions.

### 91 **Data made available by competition organizers**

92 Below is a summary of the data provided to us by the DREAM challenge:

- 93 • *Reference database:* The reference database comes from the expression patterns  
94 (Fowlkes *et al.*, 2008) of the in-situ hybridizations of 84 genes from the Berkeley  
95 Drosophila Transcription Network Project project. The *in-situ* expression of 84 genes  
96 is quantified across the 3039 *D. melanogaster* embryonic locations.
- 97 • *Spatial coordinates:* X, Y, and Z coordinates were supplied for the 3039 locations of  
98 the *D. melanogaster* embryo.
- 99 • *Single cell RNA sequencing:* Three expression tables were provided; the raw,  
100 normalized, and binarized expression of 8924 genes across 1297 cells (Karaiskos *et*  
101 *al.*, 2017).
- 102 • *DistMap source code* was provided and it was used to identify the cell locations in the  
103 initial publication (Karaiskos *et al.*, 2017).

## 104 **Cell Locations**

105 For each cell (n=1297) available in the RNA sequencing data, we generated training labels  
106 representing their 3-D positions by running DistMap (Karaiskos *et al.*, 2017) with the following  
107 inputs:

- 108 - single cell RNA sequencing expression data, both raw and normalized
- 109 - the Reference database
- 110 - the spatial coordinates

111 Briefly, DistMap calculates several parameters, a quantile value and one threshold per inSitu  
112 gene, to predict the cell locations. It employs these values to binarize the expression of the genes'  
113 and calculate the Matthews Correlation Coefficients (MCC) for every cell-bin (embryo location)  
114 combination. By doing this, DistMap maps a cell to multiple likely positions. Lasso-TopX and  
115 NN approaches (described below) use these MCC values to determine the training labels.

## 116 **Feature selection approaches**

### 117 **Random**

118 As a baseline approach, among the 84 inSitu genes from which we were allowed to pick, we  
119 randomly selected 60, 40 and 20 of them for the respective subchallenges. This random selection  
120 allowed us to benchmark (see Results) how Lasso-TopX and NN feature selection approaches  
121 compared against a random process. We performed this selection step 10 times, one for each  
122 outer cross validation fold, see “Post Challenge Outer Cross Validation” section, below. The  
123 importance of this comparison is to evaluate if the cost of building a method, both timewise and

124 computationally, has any advantage over a simple approach, that does not leverage machine  
125 learning (Karathanasis *et al.*, 2014).

## 126 **Lasso-TopX**

127 We introduce a method, Lasso-TopX, that is implemented in the R programming language and  
128 leverages the glmnet package (Friedman *et al.*, 2010) to build generalized linear models with  
129 Lasso (Tibshirani, 1996). This method allows for the identification of the most informative N  
130 features, where N is 60, 40, and 20 for sub challenges 1, 2, and 3 respectively.

### 131 PRE-PROCESSING

132 We used the following data to identify the most important genes:

- 133 - *Single cell RNA sequencing*: we subset the provided normalized single cell RNAseq  
134 dataset to include only the 84 *in-situ* genes.
- 135 - *Top cell locations*: for training labels, we identified the locations of the cells using  
136 distMap with the code provided from the challenge's organizers, see section 'Cell  
137 Locations' above. For each cell, we use the bin (embryo locations) corresponding to the  
138 maximum MCC. In our feature selection process, we employed only the cells that are  
139 mapped uniquely to one location (1010 out of 1297 cells), Supplementary Figure S1.

### 140 TRAINING FLOW AND FEATURE SELECTION

141 We performed the following steps to identify the important features employing Lasso-TopX:

- 142 1) In order to identify the most important 60 / 40 / 20 features we performed a repeated five-  
143 fold cross validation (CV) process. The CV was repeated 20 times for 300 different values of  
144 lambda. Lambda is Lasso's hyperparameter which the user needs to optimize. Intuitively,

145 fewer features will be selected as lambda increases. The range of the lambda values was  
146 defined manually, using 70% of the data and only one time, in order for models with 60 / 40 /  
147 20 features to be produced. In relation to the competition, we retrieved lambda ranges from  
148 glmnet packages, 100 values, and tripled the density to include 300 values. In total, we fitted  
149  $5 \times 20 \times 300 = 30,000$  models. Importantly, in order to avoid overfitting, during each CV fold  
150 only the training data corresponding to this fold are standardized and the resulting model is  
151 applied to the test data (Friedman *et al.*, 2010).

152 2) For each model, we extracted the following information

153 a) *The error from the model*: the Euclidean distance of the predicted XYZ location to the  
154 top location.

155 b) The number of features which were used and their corresponding coefficients.

156 3) We selected the *best lambda* value by calculating the mean error per lambda across the  
157 repeated five-fold CV (Figure 1). The *best lambda* was producing the minimum mean error  
158 and models with the desired number of features. We retained only the models corresponding  
159 to this lambda value. One lambda value and one hundred models (5-fold CV \* 20 times)  
160 were selected per sub-challenge.

161 4) For each one of the selected models we extracted their features and calculated two metrics.

162 a) *Stability*: the number of times a feature was selected as important across the repeated  
163 cross validation procedure (Figures 1b left, for sub-challenge 3), and

164 b) *Mean Coefficient*: the mean value of the coefficients that a feature was assigned across  
165 all coordinates (Figure 1b right), for sub-challenge 3.

166 5) Finally, we utilized the RankSum statistic to combine these two metrics and calculate the  
167 overall importance of the selected features. In cases where we had more than the desired



168 number of features in our final feature list we kept the features with the higher RankSum  
169 statistic (Figure 1). Having more than the desired number of features is possible as models  
170 with the same lambda may select different features during the repetitive cross validation  
171 process.

172

173

174

175

176

177

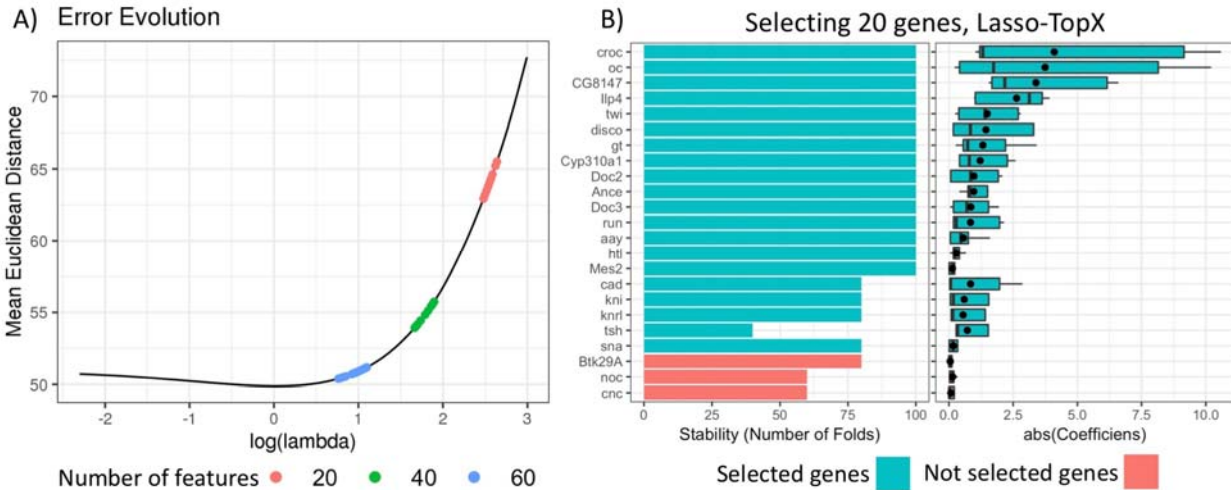
178

179

180

181

182



183  
184 **Figure 1. Selecting 20 genes – Lasso-TopX.**

185 **A) Lambda Selection:** The mean Euclidean distance error across the repeated fivefold CV in  
 186 relation to the natural logarithm of lambda is presented for one out of 10 outer Cross Validation  
 187 runs. Several models with 20 features (red dots) are produced with their lambda values to range  
 188 from 14.00083,  $\log(14.00083) = 2.64$ , to 11.93777,  $\log(11.93777) = 2.48$ . The lambda that  
 189 produced the models with the minimum Mean Euclidean Distance error was selected. The same  
 190 process was followed for both 40 (green dots) and 60 (light blue dots) features. **B) Left):**  
 191 Stability, the number of times a feature was selected across the repeated cross validation  
 192 procedure. 20 genes are shown in one out of the 10 outer Cross Validation runs. Each gene was  
 193 selected at least one time by the best performing models during the repeated cross validation  
 194 ( $\lambda = 11.93777$ ). 20 genes with the higher RankSum statistic were selected (Blue), the last  
 195 three genes (red), had the lowest RankSum statistic and left out from the final list. **B right) The**  
 196 distribution of the absolute value of the coefficients of the selected genes. The mean value of the  
 197 coefficients is shown with a black point character.

198

## 199 **Neural network-based approach using weak supervision**

200 In this approach, we perform weakly supervised learning (Zhou, 2018) using Neural Networks.  
201 After training the models, we calculate variable importance scores to rank each gene. We  
202 describe several techniques that we used to help eliminate overfitting and make sure our model  
203 generalizes well. Because the training labels were not given to us directly and because we could  
204 not assume the max MCC from DistMap was always correct, we devised a technique that is able  
205 to use multiple training labels for the same set of input neuron values.

### 206 PRE-PROCESSING

207 All genes (n=8,924) from the normalized RNAseq dataset, 'dge\_normalized.txt' were used as  
208 predictor variables. For generating the training labels for the 1,297 cell locations, we used the  
209 MCC based procedure also used by Lasso-TopX, but with one modification. Instead of using  
210 only the location (X, Y, Z) from the max MCC score, we used all locations that had an MCC  
211 score greater or equal than 95% of the max MCC score.

212

### 213 NN ARCHITECTURE AND LOSS FUNCTION

214 Model training and inference were all performed using Python 3.5, the PyTorch (Paszke, Adam,  
215 Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin,  
216 Alban Desmaison, Luca Antiga, 2017) machine learning library, and the numpy and pandas  
217 packages. We used a fully-connected NN architecture described below.

- 218 • *Input neurons*: one input neuron per RNAseq gene
- 219 • *Hidden layers*: two hidden layers were used, each with 100 neurons

- 220 • *Output neurons*: three output neurons were used (X position, Y position, and Z
- 221 position)
- 222 • *Loss function*: Euclidean Loss using all three output neurons
- 223 • *Activation function*: Rectified Linear Units (Nair and Geoffrey E., 2017; Hahnloser *et*
- 224 *al.*, 2000)
- 225 • *Optimizer*: Adadelata (an adaptive learning rate method) was used for gradient descent
- 226 (Zeiler, 2012)
- 227 • *Other*: To help avoid overfitting and allow the model to generalize better, we used
- 228 Hinton Dropout (Srivastava *et al.*, 2014), set at 10% for both the hidden and input
- 229 layers.

230

## 231 TRAINING FLOW

232 Using the pre-processed data, we performed 5-fold cross validation 40 times for a total of 200  
233 models. Before training, the genes were first standardized to have a mean of 0 and unit variance.  
234 For each model and to prevent bias (Ambroise and McLachlan, 2002), the parameters used for  
235 this adjustment were determined from the training splits only, and then applied to the validation  
236 split.

237 Similarly, using the training split only to avoid selection bias (Ambroise and McLachlan, 2002;  
238 Smialowski *et al.*, 2009), we removed correlated genes by first identifying genes that (1) were  
239 not an inSitu-gene and (2) had a Pearson correlation with at least one inSitu-gene of  $\geq 0.6$  or  
240  $\leq -0.6$ . We then removed these RNAseq genes from all splits. This allowed us to remove non-  
241 inSitu genes that might prevent variable importance scores of inSitu genes from showing up

242 which we thought might be helpful because the subchallenges only asked us to report inSitu  
243 genes.

244 At every 80/20 split, we made sure that a cell's gene expression values were never found in both  
245 the training and validation splits. During training, minibatch sizes of 100 were used. To help  
246 prevent the model from overfitting on the training data, we also performed early-stopping by  
247 stopping the training process once the Euclidian loss on the validation fold did not improve after  
248 50 epochs and keeping the model with the lowest loss. Because looking at the validation fold's  
249 loss function could potentially overestimate a model's performance, we evaluated our  
250 subchallenge scores on an external hold-out set from a separate outer cross-validation fold (see  
251 Results).

## 252 FEATURE SELECTION

253 Variable importance (VIP) scores were calculated and ranked for each of the 200 NN models  
254 used in the training process. We implemented (and include in our source code) Gedeon's  
255 method (Gedeon, 1997) to come up with Variable Importance Scores for each model. For each  
256 model, we only kept the genes with the highest 60/40/20 VIP scores depending on the  
257 subchallenge. We then sorted the lists by consensus vote to get one list per subchallenge. For  
258 the subchallenges, only inSitu-genes were selected.

## 259 **Location predictions**

260 The following steps were used by Lasso-TopX, NN and Random approaches in order to predict  
261 10 locations per cell:

- 262 1. We subset the *in-situ* expression database, keeping only the 60, 40, 20 genes as they were  
263 identified by the feature selection steps above.

- 264 2. We calculated MCC applying DistMap with the following modifications:
- 265 2.1. We employed only the cells belonging in the training sets and the selected genes to
- 266 calculate DistMap's parameters, which are used to binarize the genes' expression data
- 267 (Karaiskos *et al.*, 2017).
- 268 2.2. We used the same parameters to binarize the expression data of the cells belonging in the
- 269 test set.
- 270 2.3. Finally, similar to "Cell Locations" section above, we calculated the MCC for every cell-
- 271 bin combination and we selected the 10 bins that correspond to the top 10 highest MCC
- 272 scores.

273

## 274 **Post Challenge Outer Cross Validation**

275 In the post-challenge phase, the organizers split the data in 10 folds, on which our approaches

276 were re-run for stability and overfitting evaluation (Tanevski *et al.*, 2019). Separately for each of

277 the 10 iterations, only the respective 9 training folds were used for feature selection and to train

278 Distmap. The cells' locations were predicted for the remaining validation fold. We refer to this

279 post challenge cross validation as the 'outer cross validation' because any cross-validations

280 described in our feature selection methods occurred using data only *within* the training-folds of

281 this outer cross-validation.

## 282 **Blind evaluation metric**

283 Prior to the competition ending, in which contestants did not have access or insight into the

284 challenge organizers' scoring functions, we evaluated our location predictions by calculating for

285 each cell the mean Euclidean distance of the top 10 predicted locations from the cell location

286 with the maximum MCC (*MeanEuclDistPerCell*). For the cells that did not map uniquely, we  
287 used the first bin among the ties as returned by R.

288 Then, we calculated the mean of the *MeanEuclDistPerCell* per outer cross-validation fold across  
289 all cells which we refer to as *MeanEuclDistPerFold*. Finally, the mean of the  
290 *MeanEuclDistPerFold* across all 10 outer cross-validation folds was calculated and referred to as  
291 *MeanEuclDistAllFold*.

292

## 293 **Results**

### 294 **Challenge submission (Lasso-TopX or NN)**

295 We ran both the Lasso-TopX and NN approaches for all three subchallenges. Because the  
296 challenges final scoring algorithms were not available to any participants until after the  
297 competition concluded, we compared our two feature selection approaches using a blind  
298 evaluation metric (see Methods) we devised and thought might be a proxy to a good leadership  
299 score. Because the teams were only allowed one final submission to each subchallenge, we used  
300 this blind metric to determine if the results from either NN or Lasso-TopX would be submitted to  
301 each subchallenge. This blind metric was calculated individually for both feature selection  
302 methods and for each subchallenge. Our evaluation metric suggested that Lasso-TopX may  
303 perform slightly better than NN for some subchallenges (data not shown). Based on this, our  
304 final submission used results based on NN for subchallenge 2 and Lasso-TopX for the other two.  
305 Our submitted results ranked 10th, 6th, and 4th in the three sub-challenges, respectively, among  
306 ~40 participating teams (Tanevski *et al.*, 2019).

## 307 **Evaluation Post-Challenge**

308 After the challenge ended, the organizers devised a post-challenge cross-validation scheme (see  
309 Methods and (Tanevski *et al.*, 2019) for more detail) to evaluate the robustness of the methods.  
310 It was only after this resubmission phase did the organizers make the true scoring functions  
311 (“s1”, “s2” and “s3” scores) publicly available. Supplementary Figure S2 and Figure 2 show the  
312 results of our blind, s1, s2, and s3 metrics across the outer 10-fold cross-validation. As expected,  
313 both Lasso-TopX and NN behaved better than Random. The results of the three scoring schemes  
314 (Figure 2) agreed with our previous findings using our blind metric, and in agreement with the  
315 challenge paper (Tanevski *et al.*, 2019) show that our scores have little variability and that our  
316 methods generalize well. Because of the higher s2 score in subchallenge #1 for NN (Figure 2),  
317 we note the possibility that our NN approach could have ranked more favorably in subchallenge  
318 #1 when compared to the submitted Lasso-TopX predictions.

319

320

321

322

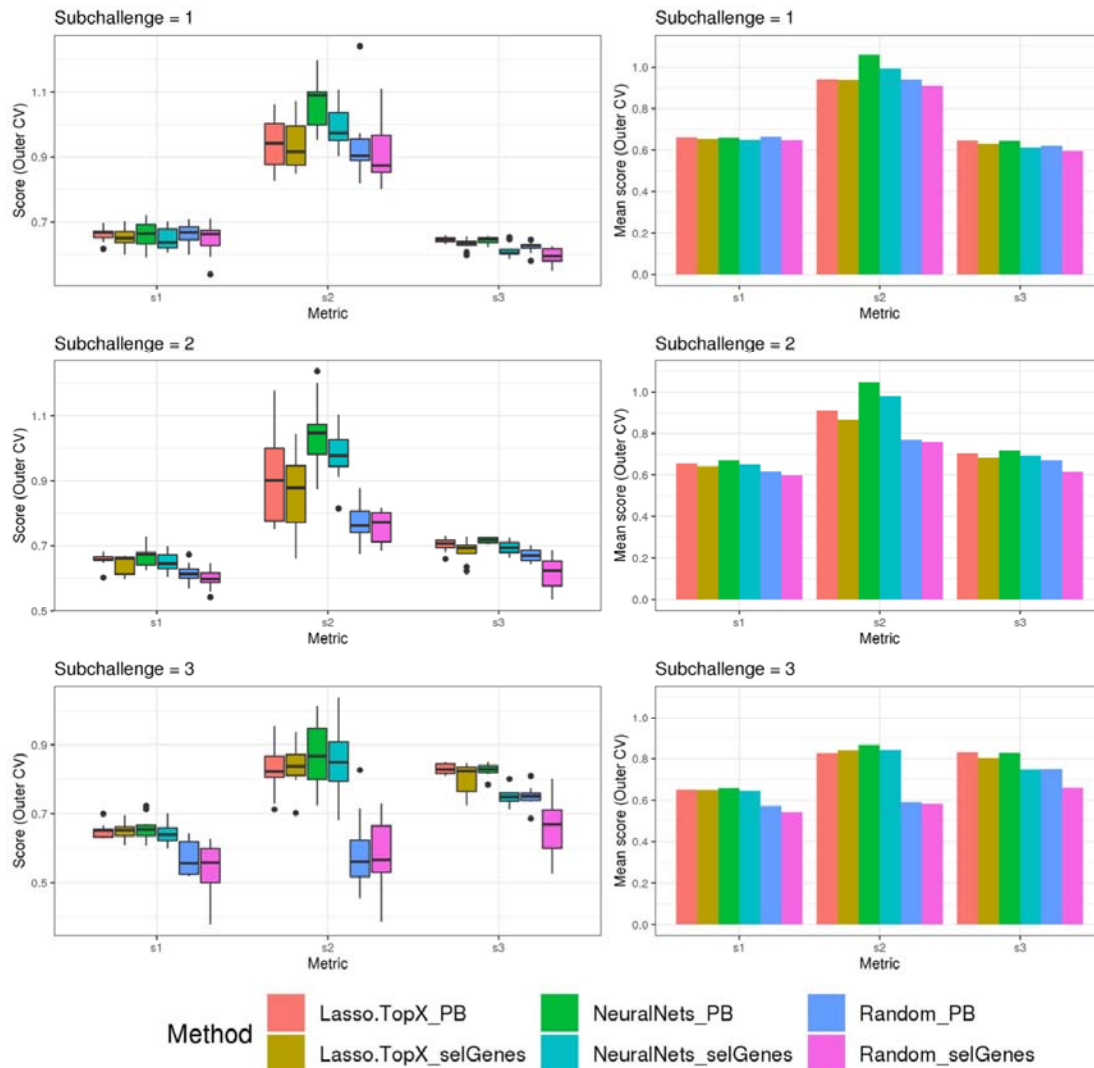
323

324

325

326





327

328 **Figure 2. Comparison of different methods, organizers scoring functions.**

329 *Lasso.TopX (Lasso.TopX\_selGenes), performed better for sub-challenges 1 and 3, and*

330 *NeuralNets (NeuralNets\_selGenes) performed better for sub-challenge 2. For sub-challenge 1,*

331 *Lasso.TopX performed better than NeuralNets for s1 and s3. For sub-challenge 2, NeuralNets*

332 *performed better for all scores and for sub-challenge 3 Lasso.TopX performed better for s1 and*

333 *s3 scores. In all sub-challenges, both methods performed better than Random*

334 *(Random\_selGenes). The binarized expression data that were produced using all expression*

335 *data, \_PB extension, showed an extreme bias in overestimation of performance, across all*  
336 *metrics, methods and sub-challenges. Lasso.TopX\_PB performed always better than*  
337 *Lasso.TopX\_selGenes, NeuralNets\_PB performed always better than NeuralNets\_selGenes and*  
338 *Random\_PB performed always better than Random\_selGenes.*

339

## 340 **Considerations for Weak Supervision**

341 When generating training labels during the pre-processing step of NNs (see Methods), there were  
342 several reasons why we allowed multiple training labels for the same cell. First, it allowed all  
343 locations of the 287 cells (Supplementary Figure S1) that did not uniquely map to be used during  
344 training. Also, we could not be certain that the max MCC was always the right value to use and  
345 wanted to better leverage the probabilistic mapping strategy enabled by DistMap.

346 Figure 3a shows that the vast majority of the time, there exists more than one spatial location for  
347 a cell when using the 95% cutoff. The most common number of selected training labels per cell  
348 location is 5, with a mean of 7. When allowing multiple training labels per cell, our dataset  
349 became much larger: 11,491 observations instead of only 1,297 observations when only the  
350 value with the max MCC was used. One pleasant consequence of having more training data is  
351 that it makes it harder to overfit a neural network which is especially problematic in high-  
352 dimensionality settings (Verleysen *et al.*, 2003; Bellman, 1961).

353 In Figure 3b we also show that the surviving training labels per cell generally represent similar  
354 spatial coordinates when compared with randomly shuffled locations in the training data. This  
355 suggested that allowing multiple training labels per cell during training could guide the model to

356 generalize to less-specific spatial regions without being pegged to any one location that could  
357 have been incorrectly classified.

358 Importantly, as mentioned in Methods, when generating training and validation sets we made  
359 sure that a cell's gene expression values were never found in both sets. We accomplished this by  
360 splitting on cell names versus the row indices. This is especially important because the pre-  
361 processing steps allowed the same cell to be found multiple times but with different training  
362 labels (Figure 3a). We found that if we did not split this way, we would have indirect data  
363 leakage (Luo *et al.*, 2016) and significantly overestimate the performance of our models because  
364 the validation split could contain identical predictor variables (gene expression levels) as the  
365 training splits but with training labels that had similar (though not identical) spatial locations  
366 (Figure 3b).

367

368

369

370

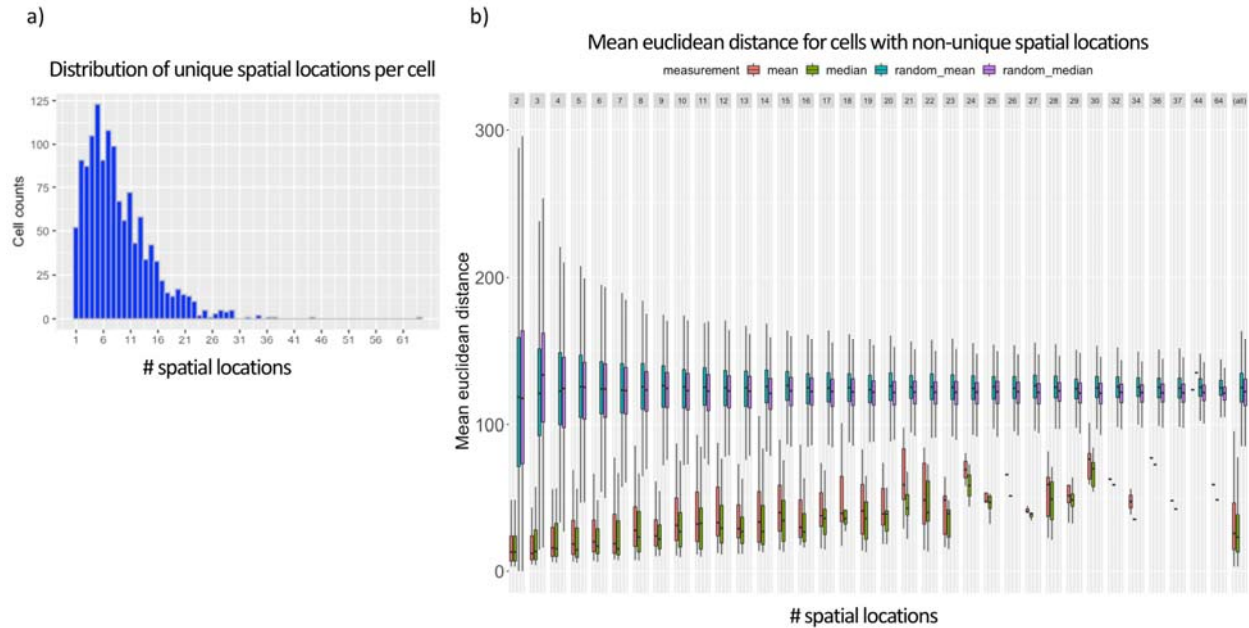
371

372

373

374

375



376  
377 **Figure 3. Ambiguous training labels leveraged in NN approach.**

378 (a) Shows that most *D. melanogaster* cells contain more than one spatial location (mean of 7)  
379 when using *DistMap*'s predictions and thresholding at 95% of the max MCC. (b) Shows the  
380 mean and median Euclidean distances between the cell's with non-unique training labels that  
381 meet the same 95% threshold. Cell's that are selected at this threshold have mean and median  
382 Euclidean distances much lower than random and highlight the probabilistic nature of the  
383 training data.

384

385

386

387

388 **Measuring and avoiding data leakage during location prediction**

389 We also sought to determine what our scores would have looked like if data leakage occurred  
390 during the location prediction stage. In machine learning and statistics, *data leakage* can lead to  
391 inflated performance estimates when data from the validation or test set are used during training  
392 (Luo *et al.*, 2016). Overfitting because of data leakage would have been easy to do by mistake  
393 because the provided binarized expression data, generated by DistMap, were produced using all  
394 expression data and consequently should never be used at any step of training or testing. For  
395 example, one might think that instead of modifying DistMap to perform the two-step approach  
396 described in Methods, a contestant could have used the provided binarized data to directly  
397 calculate the MCC scores and the 10 cell positions. However, as is evident from Supplementary  
398 Figure S2 and Figure 2 this will lead to overestimation of performance irrespective of the scoring  
399 functions (blind, s1, s2, s3) or the methods (NNs, Lasso-TopX, Random) used. In both figures  
400 we present bars and boxplots which correspond to the overfitted location predictions using the  
401 unmodified and provided binarized data (extension “PB”) and compare it to the approach we  
402 used (extension “selGenes”).

403

#### 404 **InSitu Genes with Spatial Information**

405 We observed that the genes selected across the outer 10 cross validation folds were stable,  
406 (Supplementary table S1). More specifically, in the Lasso-TopX case, 74, 54 and 27 genes were  
407 selected in total for sub-challenge 1, 2 and 3, respectively, with 44 (60%), 25 (46%) and 15  
408 (56%) of them to be selected across all folds, Supplementary Figure S3a. Similarly, in the NNs  
409 case, 64, 47 and 23 genes were selected, with 58 (91%), 33 (70%) and 13 (57%) of them to be  
410 selected across all folds, for sub-challenge 1, 2 and 3, respectively (Supplementary Figure S3b).

411 As expected, Random did not show the same trend (Supplementary Figure S3c) with zero genes  
412 selected across all folds. We observed agreement in the in-situ genes selection between our two  
413 distinct feature selection strategies despite differences in pre-processing, features used during  
414 training, and inference models. Specifically, we observed a mean of 80%, 76% and 64%  
415 agreement for sub-challenge 1, 2 and 3 respectively, across the outer cross validation folds  
416 (Supplementary Figure S4).

417

### 418 **Non-inSitu Genes with Spatial Information**

419 While not a focus on this competition, we additionally ran both our methodologies using  
420 RNASeq data information from *both* the non-inSitu and inSitu genes and we were able to  
421 discover many informative non-inSitu genes that also contain positional information. The list of  
422 20/40/60 genes using the NN methodology would have been composed of 56%, 50%, 52%,  
423 respectively, of non-inSitu genes (Supplementary table S1). In the Lasso-TopX case, 67% 66%  
424 and 70% of the selected genes were non-inSitus when selecting the most informative 20 / 40 / 60  
425 genes (Supplementary table S1). Similar to the inSitu genes analysis, we calculated the stability  
426 of the selected genes across the outer cross validation folds. In the Lasso case, 36.6%, 28.7%  
427 and 23.7% of genes were selected across all folds of the outer CV, when selecting for 20, 40 and  
428 60 genes, respectively (Supplementary Figure S5a). In the NNs case, 46%, 59% and 67% of  
429 genes were selected across all folds when selecting 20, 40 and 60 genes, respectively,  
430 Supplementary Figure S5b. Furthermore, we observed that on average 52%, 57% and 51% of  
431 genes identified by NN and Lasso-TopX were in common across the 10-fold outer cross  
432 validation folds, when selecting 60, 40 and 20 genes, respectively, (Supplementary Figure S4b).

433 Interestingly, we observed (Figure 4) that several non-inSitu genes were selected consistently  
434 across all 60 feature selection runs ( $60 = 2 \text{ methods} * 3 \text{ subchallenges} * 10 \text{ outer CV folds}$ ).  
435 Specifically, 142 genes were identified across all runs consisting of 38 inSitus and 104 non-  
436 inSitus. As expected, due to the fact that inSitu genes contain spatial information, they were  
437 selected on average more often, 25 times out of 60, than non-inSitus, 14 out of 60. However,  
438 focusing on the most stable genes, genes that were selected in at least 30 out of the 60 runs, 15  
439 out of 29 are non-inSitus, (Figure 4).

440

441

442

443

444

445

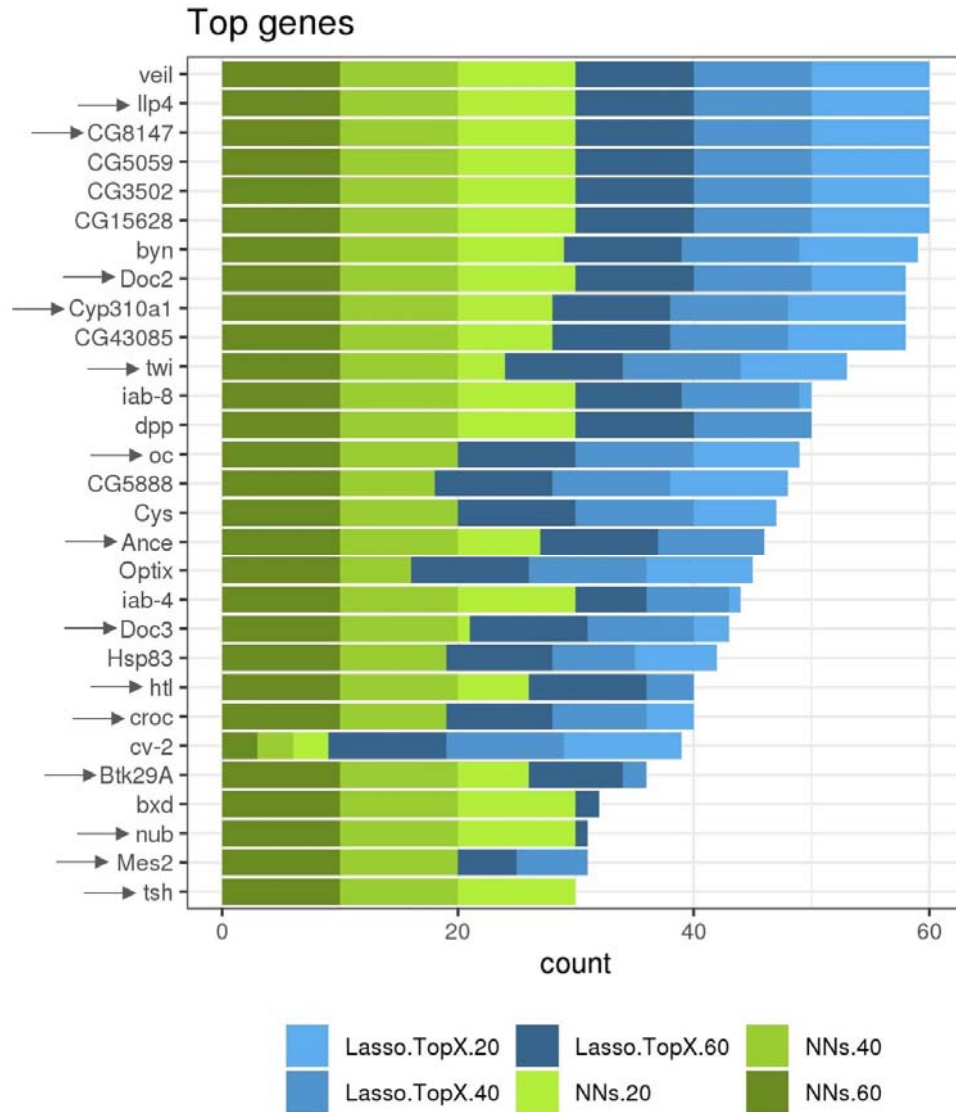
446

447

448

449

450



451

452 **Figure 4. More frequent identified genes.**

453 *Several inSitu and non-inSitu genes were selected across the sub-challenges, methods and outer*  
 454 *cross validation folds. 29 genes, 14 inSitu (arrow) and 15 non-inSitu were selected in at least 30*  
 455 *out of a total of 60 feature selection runs, xaxis. Lasso-TopX and NNs are shown in shades of*  
 456 *blue and green, respectively. Dark, regular and light shades correspond to selecting 60, 40 and*  
 457 *20 genes, respectively.*

458



## 459 Discussion

460 In this paper we present a modified Lasso workflow, named Lasso-TopX, that is able to extract  
461 the most important user defined number of features and a NN approach that uses weak  
462 supervision that is able to use multiple training labels per cell for linear regression. We  
463 developed these methods as part of our participation in the DREAM single cell transcriptomics  
464 challenge (Tanevski *et al.*, 2019). The challenge had two main objectives, a) to identify the top  
465 60, 40 and 20 genes that contain the most spatial information, and b) to reconstruct the 3-D  
466 arrangement of the *D. melanogaster* embryo using information from those genes. The 20 and 60  
467 genes were identified using Lasso-TopX and the 40 genes using NN. In all cases, the 3-D  
468 locations were predicted using Matthew Correlation Coefficients based methodology. Our team,  
469 named DeepCMC, ranked 10th, 6th, and 4th in the three sub-challenges, respectively, among  
470 ~40 participating teams from all around the world (Tanevski *et al.*, 2019).

471 A typical Lasso workflow consists of first identifying the best lambda, using cross validation and  
472 then employing that lambda to train on the full dataset to identify the most informative features  
473 (James *et al.*, 2013). This process does not allow the user to specify a discrete number of features  
474 they are interested in because the selected lambda is not tied to a user-defined number of features  
475 (Friedman *et al.*, 2010). Also, running the typical workflow multiples times could lead to slightly  
476 different optimal lambda values as the data splits during the cross validation could differ and  
477 thus this could lead to slightly different features and different number of features. Another  
478 approach that could be employed to meet the subchallenges requirements, is to have performed  
479 the typical workflow and then select the top 20 / 40 / 60 genes from the resulting list of genes.  
480 This approach is not optimal; for example, if someone could select the best 20 features, these 20  
481 features would not necessarily be a subset of the best 40 features (James *et al.*, 2013).

482 Considering the above, we developed Lasso-TopX, that leverages Lasso and is able to identify  
483 the most important user-defined number of features, employing repeated cross validation to  
484 make the results less dependent on any particular choice of data split. Lasso-TopX can be applied  
485 to classification or regression problems where finding the most important, stable, and user-  
486 defined number of features is important.

487 We also note that Lasso-TopX will provide the most value if the user-defined number of features  
488 is less than what the traditional Lasso workflow would have chosen. Taking as an example,  
489 figure 1A, we see that Lasso's error is decreasing in the beginning as we move to higher values  
490 of lambda (from left to right), then there is a local minimum (close to  $\log(\lambda) = 0$ ), and then  
491 the Lasso error increases again. At the local minimum, Lasso provides the best features given its  
492 underlying assumptions. We suggest the user to run regular Lasso, to identify Lasso's optimal  
493 performance point and then to define the desired number of features on the right-hand side of  
494 that point.

495 For our NN approach, we show that a cell's training labels do not have to be unique. This is  
496 especially useful to take advantage of training data generated from DistMap's probabilistic  
497 mapping output. We demonstrate how to properly split training and validation data when non-  
498 unique (Figure 3a) but correlated (Figure 3b) training labels are used in order to prevent data  
499 leakage. We hope that our approach will be helpful in the active research field of Weak  
500 Supervision (Zhou, 2018) and as probabilistic training labels become more commonplace. A lot  
501 of research in weak supervision for NNs has been around logistic regression and we hope that  
502 our techniques will help facilitate more applications in linear regression.

503 Not all decisions were consistent between Lasso-TopX and NN feature-selection approaches.  
504 For instance, the number of features (# of genes from RNASeq data used) and training labels  
505 (max MCC versus 95%) used during training differ between the approaches. Therefore, any  
506 differences in performance (Fig 3) and feature stability (Sup Figure S3, Sup Figure S5) also  
507 reflect various decisions made during the pre-processing stages. This was intentional as our goal  
508 was not to determine if NN or Lasso-TopX was better using the exact same data, but which  
509 independent approach would allow us to best address the sub-challenges.

510 One attribute of our models that is worth pointing out is that both the NN and Lasso approach  
511 could have also been used to predict X, Y, and Z spatial locations directly (not just for feature  
512 selection) because they were already optimized to predict a cell's spatial position. However, we  
513 chose to use DistMap for this step to make the location prediction step consistent between the  
514 NN and Lasso approach and because the challenge organizers wanted 10 different location  
515 predictions per cell. Leveraging DistMap's probabilistic mapping approach, using the MCC  
516 values made the later portion convenient because we were able to rank all spatial bins using the  
517 value of the coefficients.

518 Furthermore, we show that extra caution should be taken when deciding which data to be used in  
519 the location prediction process. We illustrated that in the case where DistMap/MCC approach  
520 was employed to predict cells' 3-D locations, the provided genes' expression binarized table  
521 should not be used as it leads to overestimation of performance. We observed this behavior for  
522 all feature selection methods (Lasso-TopX, Neural Nets, Random) and all score metrics, one of  
523 our own and three generated by the challenge's organizers. We believe that this is happening due  
524 to data leakage because the binarized table supplied to the contestants was generated using the  
525 full *Single cell RNA sequencing* and *Reference Database* datasets. We show quantitatively that

526 even if the data were split into training and testing, using this table during location prediction  
527 transferred information from the test data to training (through calculating DistMap parameters),  
528 leading to an overestimation of performance.

529 Lastly, while identifying non-insitu genes was not a focus of the competition, we show that our  
530 methods were able to identify non-insitu genes that also contain spatial information. We show  
531 that the Lasso-TopX and NN approaches both reported similar genes. Surprisingly, when  
532 focusing on the most stable genes, slightly more than half (15 or 29) were non-insitu genes  
533 (Figure 4, Supplementary table S1). We believe that these *D. melanogaster* genes would be  
534 good candidates for exploring in future work involving spatial information.

535

## 536 **References**

- 537 Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of  
538 microarray gene-expression data. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 6562–6566.
- 539 Bellman,R.E. (1961) Adaptive Control Processes A Guided Tour Princeton Legacy Library.
- 540 Fowlkes,C.C. *et al.* (2008) A Quantitative Spatiotemporal Atlas of Gene Expression in the  
541 *Drosophila* Blastoderm. *Cell*, **133**, 364–374.
- 542 Friedman,J. *et al.* (2010) Regularization Paths for Generalized Linear Models Via Coordinate  
543 Descent. *JournalofStatisticalSoftware*, **33**.
- 544 Gedeon,T.D. (1997) Data mining of inputs: analysing magnitude and functional measures. *Int. J.*  
545 *Neural Syst.*, **8**, 209–218.

- 546 Hahnloser, R.H.R. *et al.* (2000) Digital selection and analogue amplification coexist in a cortex-  
547 inspired silicon circuit. *Nature*, **405**, 947–951.
- 548 Karaiskos, N. *et al.* (2017) The *Drosophila* embryo at single-cell transcriptome resolution.  
549 *Science* (80-. ), **358**, 194–199.
- 550 Karathanasis, N. *et al.* (2014) Don't use a cannon to kill the ... miRNA mosquito. *Bioinformatics*,  
551 **30**, 1047–1048.
- 552 LeCun, Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.
- 553 Luo, W. *et al.* (2016) Guidelines for developing and reporting machine learning predictive  
554 models in biomedical research: A multidisciplinary view. *J. Med. Internet Res.*, **18**, 1–10.
- 555 Nair, V. and Geoffrey E., H. (2017) Rectified Linear Units Improve Restricted Boltzmann  
556 Machines. *J. Appl. Biomech.*, **33**, 384–387.
- 557 Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito,  
558 Zeming Lin, Alban Desmaison, Luca Antiga, A.L. (2017) Automatic differentiation in  
559 PyTorch. In, *Conference on Neural Information Processing Systems (NIPS 2017)*. Long  
560 Beach, pp. 1–4.
- 561 Smialowski, P. *et al.* (2009) Pitfalls of supervised feature selection. *Bioinformatics*, **26**, 440–443.
- 562 Srivastava, N. *et al.* (2014) Dropout: A Simple Way to Prevent Neural Networks from  
563 Overfitting. *J. Mach. Learn. Res.*, 1929–1958.
- 564 Stolovitzky, G. *et al.* (2007) Dialogue on reverse-engineering assessment and methods: The  
565 DREAM of high-throughput pathway inference. *Ann. N. Y. Acad. Sci.*, **1115**, 1–22.

- 566 Tanevski,J. *et al.* (2019) Predicting cellular position in the Drosophila embryo from Single-Cell  
567 Transcriptomics data. *bioRxiv*.
- 568 Tibshirani,R. (1996) Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B*,  
569 267–288.
- 570 Verleysen,M. *et al.* (2003) On the effects of dimensionality on data analysis with neural  
571 networks. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes*  
572 *Bioinformatics)*, **2687**, 105–112.
- 573 Zeiler,M.D. (2012) ADADELTA: An Adaptive Learning Rate Method.
- 574 Zhou,Z.-H. (2018) A brief introduction to weakly supervised learning. *Natl. Sci. Rev.*, **5**, 44–53.
- 575