Investigating The Effect of Data Imbalance on Perturbed Single-cell Prediction

Pavlo Bilinskyi
Computer Science
NaUKMA
Kyiv, Ukraine
p.o.bilinskiy@gmail.com

Danyil Orel
Computer Science
NaUKMA
Kyiv, Ukraine
mail.ordan@gmail.com

Mentor: Zeinab Navidi
Computer Science
UofT
Toronto, Canada
zeinab.navidi@mail.utoronto.ca.com

***Abstract*—The advancement of single-cell sequencing technologies has led to fascinating improvements in understanding cell behavior due to the high resolution that they offer in determining biological heterogeneity. Sequenced genomics information can be used in personalized medicine, which has been shown to be the most effective approach for disease treatment. Developing efficient machine learning models is a promising approach toward this goal, and various factors need to be considered for efficient modeling. To the best of our knowledge, this study is the first attempt that investigates the effect of data imbalance in perturbed response prediction in single-cell resolution. We performed many experiments exploring how the characteristics, imbalance, and diversity of training samples affect the machine learning's predictive and generalizability power. We deployed techniques such as up-sampling and down-sampling to modify training data characteristics and reported the R^2 metric to evaluate model performance.**

***Index Terms*—Single-cell, Perturbation, Data Imbalance, Machine Learning, Variational Autoencoder**

I. Introduction

This research was motivated by the personalized medicine approach, which was the subject of many contemporary medical studies. Consider a sick patient who needs treatment. In a traditional "One-Size-Fits-All" approach, all patients with the same diagnosis receive the same treatment. However, treatment can have different effects on patients because of the unique genetic features of the human body. In contrast, in a personalized approach, medicine is prepared by considering each patient's DNA information. Modern scientific methods are used to understand what effect the cure will cause based on the information obtained from a person's unique molecular and genetic profile [1].

Single-cell RNA sequencing(scRNA-seq) technology has enabled obtaining genetic information in individual cell resolution. scRNA-seq is typically used as a matrix of a cell by a gene in computational analysis, with each row indicating the expression level of genes in that cell and each column associated with a gene. Each element indicates the number of RNAs sequenced in a single cell aligned to the position of a known gene, which is indicative of the activity of that gene in the specified cell. Raw values of the cell by gene matric are discrete. However, normalizing and log-transforming the values of the cell by gene matrix is one of the best practices for scRNA-seq analysis. The gene expression profile of a cell is a modality commonly used for indicating a cell's state.

Many single-cell analysis methods underperform in settings where datasets are imbalanced based on cell types [2]. This form of imbalance originates from differences in the available cell types, the number of cells per cell type, and cell-type proportions across samples. Containing an imbalanced cell population is a common feature of many of the public single-cell datasets that may lead to inaccurate biological conclusions. A recent study showed the importance of the problem of data imbalance in the domain of single-cell data analysis, particularly for the task of single-cell integration [3] [4]. However, it has yet to be studied for the perturbation prediction task, which is the topic we got interested in and analyzed in this project.

Here, we present an extensive analysis of the effects of dataset imbalance on the scRNA-seq perturbation prediction task. We performed multiple experiments with different scenarios of cell type imbalance on human peripheral blood mononuclear cell (PBMC) data with seven different cell types [5]. We compared the perturbation prediction performance of models trained on balanced and imbalanced cell populations in each experiment by comparing their R^2 metrics. Finally, we provided our interpretation based on the outcome of each experiment.

II. Method

A. Existing Methods

There are multiple machine learning tools developed for predicting the effect of chemical and genetic perturbations (different formats of treatment) and single-cell states [6] [7] [8] [9]. We used scGen as the reference model and implemented their method in PyTorch as a practice to learn working with deep learning frameworks and gaining hands-on experience [5]. scGen is built upon Variational Autoencoder, which is widely used for this application. Our code is available on the Github page of our project (link: https://github.com/danorel/vaegen).

B. Variational Autoencoder

Variational Autoencoder (VAE) is a generative machine learning model for the dimension reduction task [10]. It

consists of the encoder and decoder components. The encoder is a map from the input space to the latent space, generating latent representations of the input data. The model is trained to reduce the reconstruction error and simultaneously regularize latent distribution. It results in the interpretable and exploitable structure of latent distribution. VAE can generate new data by randomly sampling the latent distribution and decoding it afterward. [10].

In a single-cell data analysis, Variational Autoencoder is used to build interpretable lower-dimension representations of single-cell data [5]. The obtained latent representations of cells can be used for downstream analysis or further processing.

C. scGen

scGen is a method to predict single-cell perturbation response [5]. Given a set of observed cell types in control and stimulation, it aims to predict the perturbation response of a new cell type A by training a model that learns to generalize the response of the cells in the training set (Figure 1). It uses Variational Autoencoder to get the latent space representations of cells. The predictions are obtained using vector arithmetics in the latent space. [5]

Specifically, the cells' gene expression measurements are projected into a latent space using an encoder network and obtain a vector δ that represents the difference between perturbed and unperturbed cells from the training set in latent space. Using δ, unperturbed cells of type A are linearly extrapolated in latent space. The decoder network then maps the linear latent space predictions to highly non-linear predictions in gene expression space [5].

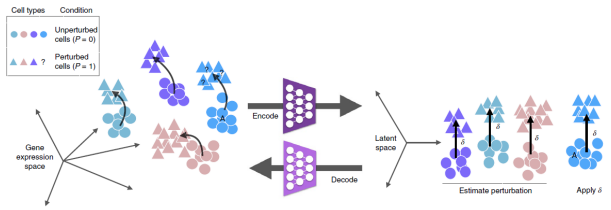

Fig. 1. The scheme of scGen method. Train data is used to learn the perturbation responses for the available cell types and generalize to the new, unobserved types of cells.

D. Dataset

We performed our experiments with different scenarios of cell type imbalance on human peripheral blood mononuclear cell (PBMC) data with seven cell types, each containing a different number of cells and existing in two conditions: control (before treatment) and stimulated (after treated with Interferon (IFN$-\beta$)). The cell type and the number of cells per group is as follows: CD4T (2437 control, 3127 stimulated), CD4T-Mono (1946 control, 615 stimulated), B (818 control, 993 stimulated), CD8T (574 control, 541 stimulated), NK (517 control, 646 stimulated), F-Mono (1100 control, 2501 stimulated), Dendritic (615 control, 463 stimulated), which is displayed in Figure 2. The total number of samples contains 16893 cells, with the gene expression profile of each cell provided for 6998 genes. More specifically, every single cell (either control or stimulated) is represented as a one-dimensional vector with a length of 6998. We downloaded and used preprocessed public data from an existing study, and the data is preprocessed, cleaned, quality controlled, normalized, and log-transformed, following the best practices for scRNA-seq preprocessing [5].

III. Experiment

We designed a series of experiments to investigate the effect of cell type imbalance on perturbed gene expression prediction on a single cell sample. Each experiment involves extracting custom datasets from the original imbalance PBMC dataset to be used for training our implemented VAE model. Each generated training data from the original PBMC dataset includes a different number of cell types number of cells per cell type and is either imbalanced or balanced. We used two main techniques for generating balanced training samples, described as follows:

1) *Downsampling*. In this approach, we performed random sampling without replacement from our dataset, such that the resulting dataset is balanced. The result is shown in Figure 4.
2) *Upsampling*. Let N be the size of the largest class in the original dataset. Then for each cell type, we make a random sample with a replacement of size N. The resulting dataset will be balanced; each class will contain N cells. The result is shown in Figure 3. In this case, we won't lose the data.

We deployed 10 fold cross-validation approach for all of our experiments to increase the robustness and generalizability of the models. In all analyses, we excluded the CD4T cell type from training and held it out as a test set. The evaluation sets were randomly selected cells of B, CD14+Mono, FCGR3A+Mono cell types, 200 cells from each group.

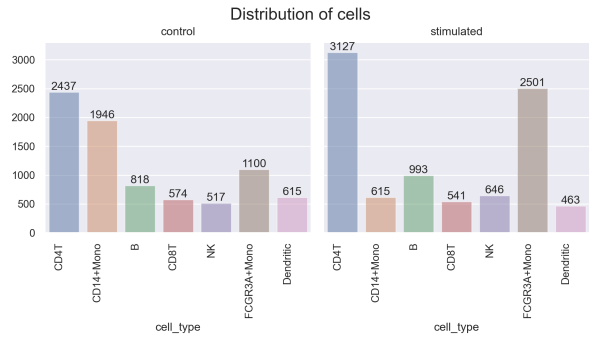

Fig. 2. The distribution of cells by classes in the dataset.

A. Data imbalance extent effect

To assess the impact of data imbalance in terms of cell types, multiple training data were randomly sliced into subsets and compared to their performance on the CD4T cells. An initial

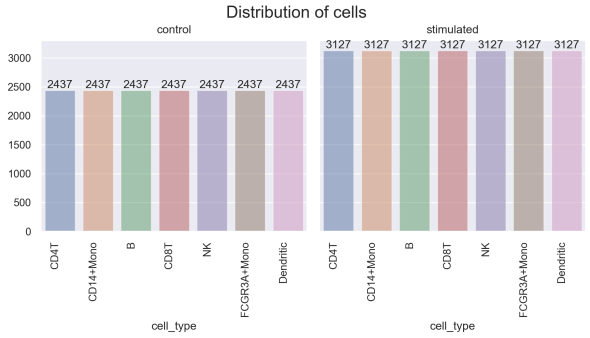

Fig. 3. The distribution of cells by classes after upsampling.

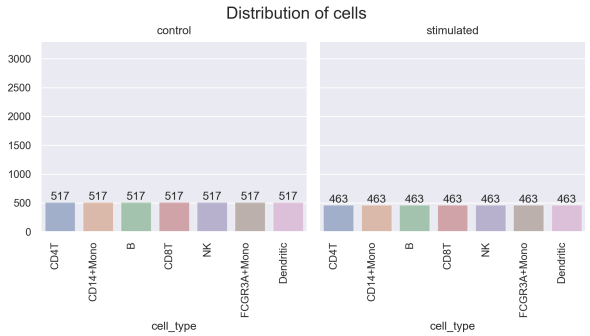

Fig. 4. The distribution of cells by classes after downsampling.

dataset is the greatest extent of training data imbalance. The original dataset was randomly downsampled to the threshold of 1000, 900, 800, 700, 600 for each cell type as separate modeling.

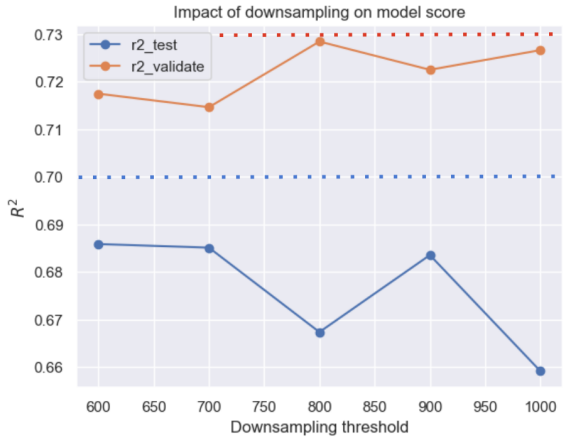

Fig. 5. Blue solid broken line represents measurements of R^2 for test samples (CD4T cells). The red solid broken line represents R^2 for validation samples (B, CD14+Mono, FCGR3A+Mono cells). Red and blue dotted lines represent R^2 benchmarks after applying the downsampling to achieve a fully balanced dataset.

Initially, the first modeling with 1000 cells per each cell type has the most imbalanced dataset. The highest margin between validation and test R^2 scores is observed on the selected threshold.

Next, the original dataset was downsampled to 900 samples per each cell type. The imbalance factor is less than in the previous modeling. The margin between validation and test R^2 scores is much less in comparison to the previous modeling results. Moreover, with a better balance of the initial dataset, a higher R^2 score was obtained for the test dataset.

After several steps, test and validation R^2 scores were measured on the dataset with a downsampling threshold of 600 samples. In this case, the imbalance factor is almost equal to zero - so the classes are almost equally distributed. Distance between test and validation R^2 scores is almost minimal among all experiments with various downsampling thresholds.

Additionally, this work contains an experiment with the least amount of cells (473) found in a cell type called NK. After using a downsampling approach on the initial dataset to that minimum threshold, the model produced the best performance results: R^2 = 0.70 for the test dataset and R^2 = 0.73 for the validation dataset.

After building the plot, you can observe two separate tendencies in Figure 5:

1. R^2 score for the validation dataset gradually worsens with a more balanced initial dataset as the model loses data variability.
2. R^2 score for the test dataset gradually improves with a more balanced initial dataset as the model becomes universal.

B. Data balance method effect

In this experiment, we tried several methods for balancing cell type classes. Each method applies a certain kind of balancing procedure to the original dataset and returns a new, balanced dataset. Using the same validation strategy as in the previous experiment, we wanted to identify the balancing approach, which gives us the most successful dataset in terms of perturbation prediction. As before, each dataset is used for training a prediction model, which is later evaluated on our validation dataset. We are interested in finding the approach which gives us the best dataset.

In addition to upsampling and downsampling, the *N-sampling* balancing method was used. The method is a combination of upsampling and downsampling. First of all, N is chosen between the size of the smallest and the highest classes. Then each class is either upsampled to N if it has fewer instances or downsampled to N if it has more instances.

There are obvious problems associated with downsampling. It requires N to be equal to the size of the smallest class, so we will lose the training data, which leads to worse predictions.

The results of the experiment with evaluation on the validation set are shown in Figure 6. Both upsampling and downsampling datasets showed poor performance. The best R^2-score had a model that was trained on a 1000-sampled dataset.

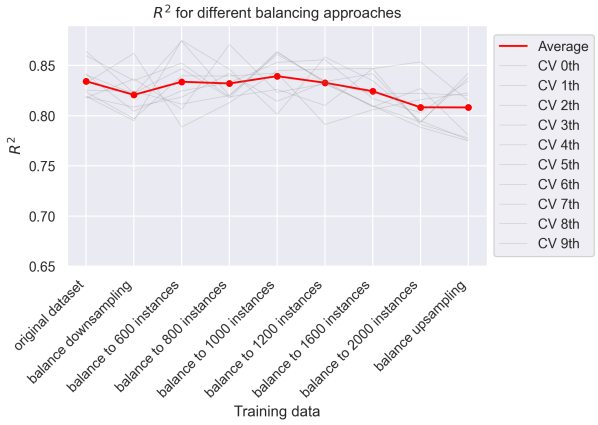

Fig. 6. Impact of data balancing method on model's performance on the validation set. The upsampling, downsampling, and N-sampling approaches were tested. A 10-fold cross-validation was used for evaluation, the red line represents the average R^2 score over CV folds.

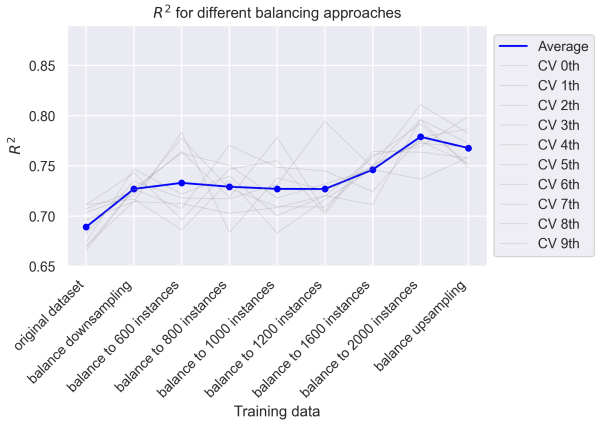

Fig. 7. Impact of data balancing method on model's performance on the test set. The upsampling, downsampling, and N-sampling approaches were tested. A 10-fold cross-validation was used for evaluation, the blue line represents the average R^2 score over CV folds.

However, the results on the test set are totally different (Figure 7). The best score is achieved on the upsampled dataset.

The different behavior on the validation and test sets can be explained by the differences between them. The validation set is a mix of cells from 3 cell type classes, which makes it more diverse. The test set contains only CD4T cells. On the test set, the most successful approach was the upsampling method, and it's probably because CD4T was the largest class, and upsampling doesn't remove the CD4T control cells, which is important for the prediction. All other methods do remove some portion of CD4T data because it's the largest cell type class. However, the original score was worse than upsampling, and it can mean that there was certain important information in other cell-type classes that was not utilized because of class imbalance. In the case of the validation set, the model didn't get any benefits from adding new data after threshold $N = 1000$ samples per class.

To summarize, the recommendations for obtaining the most successful dataset are:

- Examine the validation set and identify the cell classes that are important for predicting the perturbation for the classes in the validation set. Keep these important classes in a maximal volume.
- Apply upsampling on the other train data.

C. Cell population effect

After observing the importance of cell types in the previous experiment, it was crucial to set up a new experiment investigating cell type importance in balanced datasets. What would be the results of predictions after removing certain cell types from the balanced training dataset?

To provide such an experiment, the authors used *leave one out* [11] strategy to achieve dataset isolation from each cell type. After excluding cell types one by one and measuring the model performance, we obtained a plot in Figure 8.

Fig. 8. Blue solid broken line represents measurements of R^2 for test samples (CD4T cells). The red solid broken line represents R^2 for validation samples (B, CD14+Mono, FCGR3A+Mono cells). Red and blue dotted lines represent R^2 scores benchmarks after applying downsampling on the initial dataset achieving a fully balanced dataset including all cell types.

The constructed graph in Figure 8 indicates several helpful inferences:

1. All testing R^2 scores are below the downsampling minimum benchmark. It means that excluding any of the available cell types doesn't help to improve the perturbation prediction results due to reducing data variability. All cell types are important during training the model.

2. Some cell types (like NK and Dendritic) strongly correlate with cells we are trying to predict (CD4T), while the other cell types have a weak correlation with that target cell type. The correlation is measured as the distance between clusters of cells (in Figure 9). The closer the target and excluded cell types are, the worse the predictive results.

D. Diversity effect

The diversity of the data is the presence of cells that are not similar to each other. The more there are dissimilar cells, the

Fig. 9. Latent representation of cells grouped into clusters defined by cell type [5].

more diverse dataset is. In the original dataset, some classes are similar (for example, FCGR3A+Mono and CD4T). The goal of this experiment is to find out whether the diversity of the data has an impact on the model's performance.

For this experiment, we picked two cell type classes (FCGR3A+Mono, B). These classes are very different according to the latent representations shown in Figure 9. Following our experimentation methodology, we construct a series of datasets from these two classes, use them for model training, then evaluate the model. Each dataset has a fixed size of 800 cells.

The following datasets are used:

1) *Highly diverse dataset.* Classes are mixed in $50\% - 50\%$ proportion.
2) *Datasets with zero diversity.* Each consists of 100% cells of a single class.
3) *An intermediate case.* Classes are mixed in $25\% - 75\%$ proportion.

Results on validation and test sets showed that diversity is uncorrelated with the model's performance. The most successful and the most unsuccessful cases are the datasets with zero diversity.

For the validation set, the model's performance strongly correlates with the portion of FCGR3A+Mono cells in the data. For the test set, the opposite correlation is observed: the higher the portion of B cells in the data, the better the score. Again, it can be explained by the differences between train and validation datasets. FCGR3A+Mono cells are important for perturbation prediction of the validation set, whereas B cells are much less important. The opposite picture is observed on the test set.

The possible explanation is that we can estimate the importance of cells for prediction using latent space representations: the closer cells are in a latent space, the more similar the perturbation effect.

IV. Discussion

This study is the first attempt to discover the effect of data imbalance in perturbed response prediction in single-cell resolution. The data imbalance issue of the initial single-cell dataset was explored using upsampling and downsampling techniques.

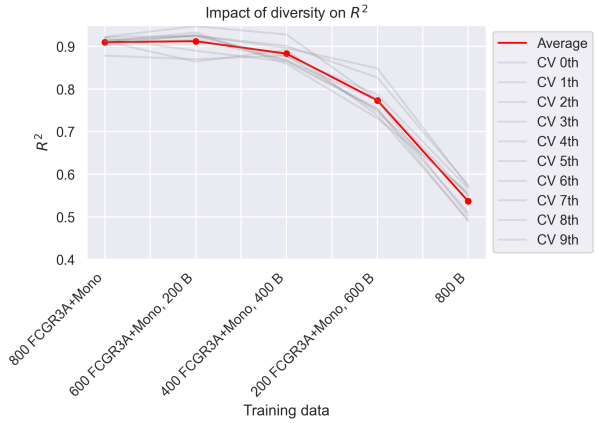

Fig. 10. R^2 score of the model on the validation set, trained on different datasets with different diversity degrees. A 10-fold cross-validation was used for evaluation. The red line represents the average score over CV folds.

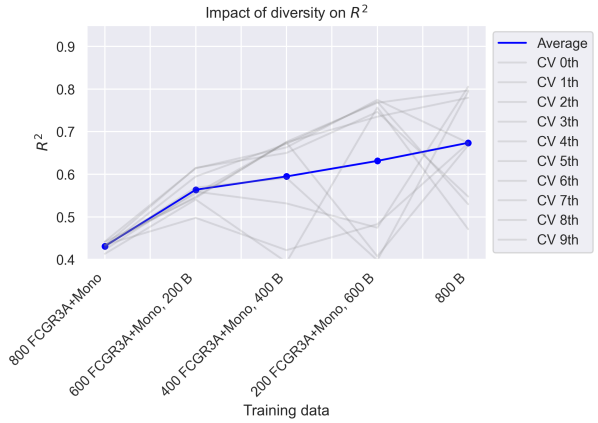

Fig. 11. R^2 score of the model on the test set, trained on different datasets with different diversity degrees. A 10-fold cross-validation was used for evaluation. The blue line represents the average score over CV folds.

Experiments were conducted using a custom Variational Autoencoder generative model implemented using a PyTorch framework. The custom model mocks the scGen model [5] behavior and extends it with custom hyperparameters.

It was decided to measure the model performance using R^2 metric. It compares perturbation responses on balanced and imbalanced datasets.

This study contains four different consecutive experiments:

1) Data imbalance extent effect
2) Data balance method effect
3) Cell population effect
4) Diversity effect

Based on conducted experiments, data imbalance is an issue for perturbation response prediction using single-cell data. Using downsampling and upsampling techniques can slightly improve the prediction quality after reducing cell type imbalance, while training predictions worsen as we reduce data variability.

On the other hand, data balancing is not the universal recipe

to achieve the best prediction outcomes. Experiment with data diversity has found strong results concerning the importance of some cell types over others. Theoretically, an ideal proportion of cell types can be constructed, which will beat the best prediction benchmarks.

Here are important questions that stay uncovered:

1. What is the perfect proportion of cell types? Would this perfect training set be balanced or imbalanced?

2. Can the mathematical correlation be defined between target and excluded cell types in the latent space?

3. Do experiment results generalize to other datasets?

REFERENCES

[1] "What is personalized medicine?," 2008 (accessed December 23, 2022). https://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/pmc_age_of_pmc_factsheet.pdf.

[2] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell rna sequencing data," *Genome Biology*, vol. 21, p. 12, Jan 2020.

[3] H. He and Y. Ma, "Imbalanced learning: Foundations, algorithms, and applications," 2013.

[4] H. Maan, L. Zhang, C. Yu, M. Geuenich, K. R. Campbell, and B. Wang, "The differential impacts of dataset imbalance in single-cell data integration," *bioRxiv*, 2022.

[5] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scgen predicts single-cell perturbation responses," *Nature Methods*, vol. 16, pp. 715–721, 2019.

[6] L. Hetzel, S. Böhm, N. Kilbertus, S. Günnemann, M. Lotfollahi, and F. Theis, "Predicting single-cell perturbation responses for unseen drugs," 2022.

[7] M. Lotfollahi, A. K. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz, "Learning interpretable cellular responses to complex perturbations in high-throughput screens," *bioRxiv*, 2021.

[8] X. Wei, J. Dong, and F. Wang, "scPreGAN, a deep generative model for predicting the response of single-cell expression to perturbation," *Bioinformatics*, vol. 38, pp. 3377–3384, 05 2022.

[9] Y. Roohani, K. Huang, and J. Leskovec, "Gears: Predicting transcriptional outcomes of novel multi-gene perturbations," *bioRxiv*, 2022.

[10] J. Rocca, "Understanding variational autoencoders (vaes)," 2019 (accessed December 23, 2022). https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.

[11] C. Sammut and G. I. Webb, eds., *Leave-One-Out Cross-Validation*. Boston, MA: Springer US, 2010.