Survey of Dimensionality Reduction Techniques and their Applications for RNA-seq Data

Nels Blair
Department of Mathematics and Statistics
nels.blair@wsu.edu

Roya Campos
School of Biological Sciences
roya.campos@wsu.edu

Nolan Middleton*
School of Molecular Biosciences
nolan.middleton@wsu.eu

Paul Ayodeji Ola
School of Biological Sciences
paul.ola@wsu.edu

Jehanzeb Saleem
School of Electrical Engineering and Computer Science
jehanzeb.saleem@wsu.edu

Abstract

Biological data pose many challenges for applying machine learning methods, particularly low sample sizes and high dimensionality due to the large number of features being measured. Low sample sizes make the use of recent deep-learning and generative methods infeasible, and as such simpler models are still widely used. Here, we survey the impact of four dimensionality reduction techniques (*a priori* aggregation, principle component analysis, kernelized principle component analysis, and nonnegative matrix factorization) on improving the accuracy of five different simple classification algorithms (decision tree, k-nearest neighbors, naïve Bayes, random forest, and support vector machine) across nine different RNA sequencing datasets from the Gene Expression Omnibus. We find that the impact of dimensionality reduction is highly variable and likely depends on underlying biological context specifics of the dataset. We note that principle component analysis, kernelized principle component analysis, and nonnegative matrix factorization are nearly identical and that the number of reduced dimensions has a much greater impact than the choice of dimensionality reduction method.

1 Introduction

Machine learning techniques have been successfully applied to answer biological questions by analyzing a wide variety of biological data, such as analyzing sequence information to identify molecular interactions between RNA molecules [1], analyzing UV damage patterns to identify transcription factor binding sites [2], and, famously, analyzing structural data to predict protein folding [3]. Of particular interest is gene expression data, which reveals how cells are behaving at a molecular level. Within a cell, the information required to perform cellular activities, such as dividing, taking up nutrients, performing metabolic reactions, etc., is stored in the cell's DNA. The DNA is a molecule made from a series of repeating molecular units called nucleotides. There are four different nucleotides used in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). The information contained in the DNA is encoded in the sequence of nucleotides and is organized into units called genes. Genes are expressed when the cell creates a temporary copy of the gene's nucleotide sequence

*Corresponding author: nolan.middleton@wsu.edu

Preprint.

as a molecule of messenger RNA (mRNA) through a process known as transcription. The RNA molecule is structured similarly to the DNA, consisting of a sequence of the nucleotide bases A, C, G, and U (where U represents uracil, which replaces thymine in RNA for evolutionary and chemical reasons). The mRNA transcript is then translated into a sequence of amino acids which folds into a protein, and the protein's function ultimately confers traits to the organism. Cells very tightly regulate what genes they express and the degree to which they are expressed, and the exact complement of expressed genes determines the cell's behavior, cell type, etc. Numerous diseases, notably cancers, ultimately result from aberrant gene expression. [4]

Perturbations to the cell (i.e. disease or infection) will be alter its gene expression [4]. This begs the question of whether disease state can be inferred from the gene expression, which can be neatly phrased as a classification problem and naturally leads to an exciting application of machine learning techniques. The goal is to gather gene expression data on both diseased and healthy individuals, then train a machine learning model to classify the disease state of individuals. To measure what genes are being expressed by a population of cells, biologists can perform RNA sequencing (RNA-seq), where the RNA is isolated and the exact nucleotide sequences of the isolated RNA molecules are found. Each RNA molecule corresponds to a gene, the identity of which can be deduced from the sequence. The degree to which any given gene is being expressed can then be inferred from the relative abundance of isolated transcripts that correspond to that gene. [5] The National Institutes of Health has also encouraged researchers to deposit their RNA-seq data on the Gene Expression Omnibus (GEO), available at `https://www.ncbi.nlm.nih.gov/sites/GDSbrowser`.

However, like with any biological data, applying machine learning techniques to RNA-seq data poses many challenges. Particularly, the time and cost of conducting biological experiments leads to extremely low sample sizes, especially if the experiment involves human subjects. Additionally, RNA-seq requires the prescription of a set of genes to measure the expression of ahead of time. This results in a wide variety of slightly different methodologies and gene sets, which makes it difficult to meaningfully compare data between RNA-seq datasets or to sensibly combine data across multiple RNA-seq datasets. Often, the expression of tens of thousands of genes are measured, orders of magnitude more than the typical sample size. This results in data with a high dimensionality that far exceeds the number of data points. The advent of single-cell techniques, particularly single-cell RNA sequencing (scRNA-seq), has partially mitigated these problem. scRNA-seq differs from RNA-seq in that, instead of isolating RNA from a population of cells, the RNA is sequenced at the level of individual cells, meaning that every individual cell can act as a separate data point. [6] This allows for the collection of large sample sizes, and these data have been successfully used to create pretrained deep-learning models such as scGPT [7], Tahoe-x1 [8], and CellFM [9]. However, single-cell techniques are expensive, and traditional RNA-seq is still common [10]. While pretrained deep-learning models for RNA-seq data, namely BulkRNABert and DCNet [11, 12], these models are trained for specific applications in cancer utilizing data from the cancer genome atlas and expect a fixed set of genes as features for their inputs [11–13], making them unsuitable for applications to general RNA-seq datasets.

The challenges posed by RNA-seq data, particularly the low sample sizes and the lack of a suitable pretrained model, make it infeasible to use modern generative or deep-learning models. As such, simpler machine learning models are still commonly used in biological contexts. However, the high dimensionality of the data still poses a challenge, motivating the application of dimensionality reduction. Here, we conduct a survey of four dimensionality reduction techniques–*a priori* aggregation, principle component analysis (PCA), kernelized principle component analysis (kPCA), and nonnegative matrix factorization (NMF)–and assess their impact on the accuracy of five simple machine learning models–the decision tree (DT), k-nearest neighbors (k-NN), the naïve Bayes classifier (NB), the random forest (RF), and the support vector machine (SVM)–in classifying disease state from RNA-seq data across nine different RNA-seq datasets (Tables 1, 2).

2 Methods

2.1 Datasets

Nine human datasets from the GEO (`https://www.ncbi.nlm.nih.gov/sites/GDSbrowser`) were surveyed to assess the generalizability of our findings across RNA-seq datasets (Tables 1, 2). The datasets were chosen to encompass a wide range of different sample sizes, number of class labels,

Table 1: Datasets utilized

Abbr.	GEO Accession	Title	Ref.
UCC	GDS1615	Ulcerative colitis and Crohn's disease comparison: peripheral blood mononuclear cells	[16]
SCLC	GDS2373	Squamous cell lung carcinomas	[15]
SEC	GDS2771	Large airway epithelial cells from cigarette smokers with suspect lung cancer	[17]
MDS	GDS3795	Myelodysplastic syndrome: CD34+ hematopoietic stem cells	[18]
ALL	GDS4206	Pediatric acute leukemia patients with early relapse: white blood cells	[19]
HIV	GDS4228	HIV infection and Antiretroviral Therapy effects on mitochondria in various tissues	[20]
JIA	GDS4267	Systemic juvenile idiopathic arthritis and non-systemic JIA subtypes: peripheral blood mononuclear cells	[21]
GBM	GDS5205	Long-term adult survivors of glioblastoma: primary tumors	[22]
MDG	GDS963	Macular degeneration and dermal fibroblast response to sublethal oxidative stress	[14]

Table 2: Dataset parameters

			Classes	
Abbr.	Samples	Features	Num.	Description
UCC	127	22283	3	Normal, ulcerative colitis, Crohn's disease
SCLC	130	22283	6	Type Ia, Ib, IIa, IIb, IIIa, IIIb
SEC	192	22283	3	No cancer, suspected cancer, cancer
MDS	200	54675	2	Healthy, myelodysplastic syndrome
ALL	197	54675	3	Early, late, no relapse
HIV	166	4825	2	HIV-negative, HIV-positive
JIA	154	54675	3	No JIA, systemic JIA, non-systemic JIA
GBM	70	54675	3	Short-term, intermediate, long-term overall survival
MDG	36	12625	2	Healthy, macular degeneration

and number of features. The most limiting case is the MDG dataset, which as only 36 data points in 12625 dimensions [14]. Additionally, the SCLC dataset has 130 data points with six classes that are all similar, each describing a different stage of lung cancer, and poses a challenging classification problem [15].

The datasets were downloaded as "full" SOFT files (text files) from the GEO. Gene expression values were extracted and stored in a tab-delimited tabular format. Features corresponding to sequencing platform controls as opposed to genes were removed from consideration.

2.2 Models

Five simple machine learning strategies were surveyed: DT, k-NN, NB, RF, and SVM. Many combinations of different values for the model hyperparameters were tested. The DT was pruned to maximum depths of 2, 3, and 4, and features were chosen to split on by optimizing the entropy. The Minkowski distance was used with exponents of $p = 1, 2, 3, 4$ for k-NN, and values of k were set to 3, 4, and 5. 100, 200, and 500 estimators were used for RF, and the individual estimators were pruned to maximum depths of 2, 3, and 4. The radial basis function (RBF), linear, and quadratic kernels were tested for the SVM, and values of $C = 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}$ were chosen for the regularization parameter. All models were implemented via `scikit-learn` [23].

2.3 Dimensionality Reduction Strategies

As a baseline, no dimensionality reduction was applied, and the models were trained on the datasets in their full dimensions.

The first dimensionality reduction strategy we applied was aggregation of the data across features based on *a priori* biological knowledge. Genes do not function independently. For instance, some genes encode transcription factors, which are proteins that alter the expression of other genes, and genes whose products all function as part of the same biological pathway or system are often coregulated. [4] The features of the datasets, which correspond to genes, can be therefore be grouped together based on their biological functionality. Genes are assigned a gene ontology (GO) category by the GO consortium [24, 25], and the dimensionality of the data is then reduced from n, the number of individual genes, to m, the number of GO categories. The value corresponding to each GO category is taken to be the mean of the values corresponding to each gene that belongs to that category. There are three different GO categorization schemes that were utilized: component, which corresponds to the subcellular compartment the gene product functions in (e.g. mitochondria, nucleus); function, which corresponds to what the gene product does (e.g. ATP binding, protein binding); and process, which corresponds to what biological pathway the gene product takes part in (e.g. MAPK cascade, inflammatory response).

The other three dimensionality reduction techniques we applied were PCA, kPCA, and NMF, which are all commonly utilized in biological contexts. Briefly, PCA maintains only the components of the data that account for the highest variation, choosing directions . kPCA exploits the fact that PCA can be calculated with only the inner product and utilizes the "kernel trick" (here, the RBF kernel was used), and NMF decomposes the data matrix into two lower-dimensional matrices whose products approximate the original data matrix in a way that, unlike PCA and kPCA, does not re-center the data but can only be approximated (up to 200 iterations were allowed for convergence during implementation). These three techniques were implemented via `scikit-learn`. [23] The dimension of the datasets were reduced to $d = 2, 3, 4, 5, 6, 7, 8, 9$ with each of these methods to allow for visualization ($d = 2$) and to encompass a range of dimensionalities whilst keeping the dimensionality ~one order of magnitude below the number of data points.

A key disadvantage of many traditional dimensionality reduction strategies, including PCA, kPCA, and NMF, is that they obscure the original context of the features. Here, initially, the features correspond directly to biologically-tractable information: each feature represents a gene. However, the reduced set of features after applying PCA, kPCA, or NMF no longer correspond to biologically meaningful attributes. In contrast, the *a priori* aggregation strategy retains biological tractability, as the reduced set of features correspond to gene categories. However, there are a large number (> 1000) different GO categories in each categorization scheme (component, function, and process), meaning that, unlike PCA, kPCA, and NMF, the dimensionality data after *a priori* aggregation will remain high compared to the number of data points. Therefore, surveying all four strategies allows for the assessment of what is more important for the performance of simple classification models: retaining biological context or ensuring the dimensionality is small compared to the number of data points.

2.4 Assessment and Evaluation

Due to limited sample sizes, we utilized leave-one-out validation to assess the models. Models were assessed primarily based on their prediction accuracy because of the heterogeneity in the datasets (i.e. the variable number of classes).

3 Results

3.1 Baseline Performance is Highly Variable Across Datasets

First, we applied no dimensionality reduction and performed leave-one-out validation to assess the baseline accuracy of the five models on all nine datasets (Fig 1). When the hyperparameters of the model are optimized for the dataset, the performance is highly variable. Unsurprisingly, the accuracy of the models on the SCLC dataset was quite low, not exceeding 40% for any model. In contrast, the performance on the MDS dataset was very high, with every model scoring above 90% leave-one-out validation accuracy with optimal hyperparameters. While the baseline performance was highly variable between datasets, the performance between models was more predictable. Generally, the NB model performed worst or near worst while the SVM performed best or near best (Fig 1).

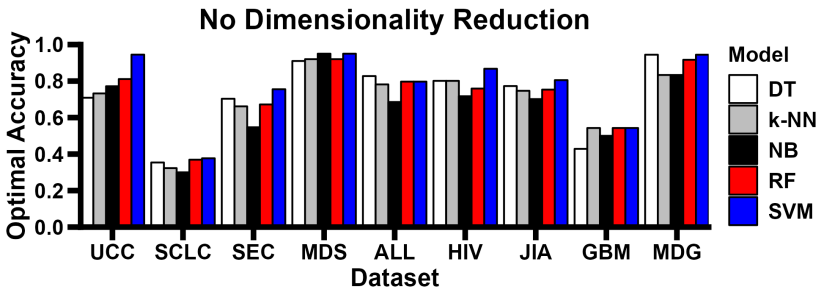

Figure 1: Baseline performance is highly variable. When no dimensionality reduction is applied to the datasets, the classification accuracy varies heavily between the nine different datasets and between the five different models. For each model, the leave-one-out validation accuracy was computed for every dataset across a combination of hyperparameter values. The optimal accuracy for each model on each dataset is the leave-one-out validation accuracy for the combination of hyperparameter values that resulted in the highest leave-one-out validation accuracy for that model on that dataset.

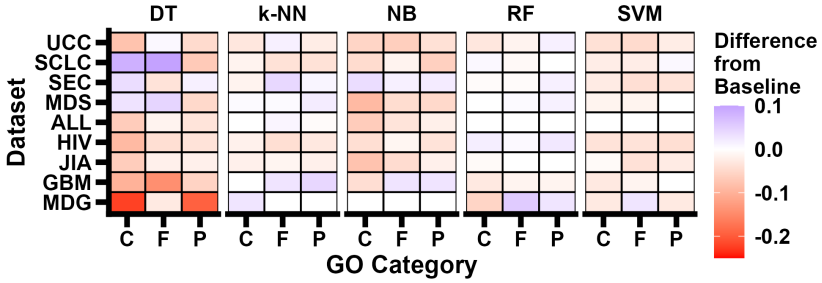

Figure 2: *a priori* aggregation has variable impacts on performance. When the dimensionality of the datasets is reduced by aggregating the data based on GO annotation, the change in classification accuracy from baseline varies heavily between the nine different datasets and between the five different models. For the GO categorization, "C" stands for component, "F" stands for function, and "P" stands for process.

3.2 Impact of *a priori* Grouping is Highly Variable

We next performed dimensionality reduction by aggregating the features based on the GO categorization of the genes. We trained the models on the modified datasets and computed the leave-one-out validation accuracy as above, then compared the accuracy on the reduced dataset to the baseline (Fig 2). The impact of this dimensionality reduction strategy was inconsistent across datasets and models. For the DT model in particular, this dimensionality reduction strategy exhibited a wide range of impacts, providing a large improvement in performance on the SCLC dataset, increasing leave-one-out validation accuracy by ˜10%, but also drastically hindering performance on the MDG dataset, decreasing the accuracy by ˜20%. In contrast, the performance was largely unaltered on any dataset for the RF model.

3.3 Reduced Dimension has a Greater Impact than Choice of Dimensionality Reduction Method

We next applied PCA, kPCA, and NMF dimensionality reduction. If the reduced dimension is fixed, the performance of all three methods are largely similar across most datasets (Fig 3). Notable exceptions are: the NB model, for which kPCA generally performed worse than PCA or NMF; the UCC and MDG datasets, for which kPCA also caused lower performance in each model when compared to PCA and NMF; and the SCLC dataset, for which kPCA generally performed better than PCA and NMF. Figure 3 displays the optimized leave-one-out validation accuracy for reduction to

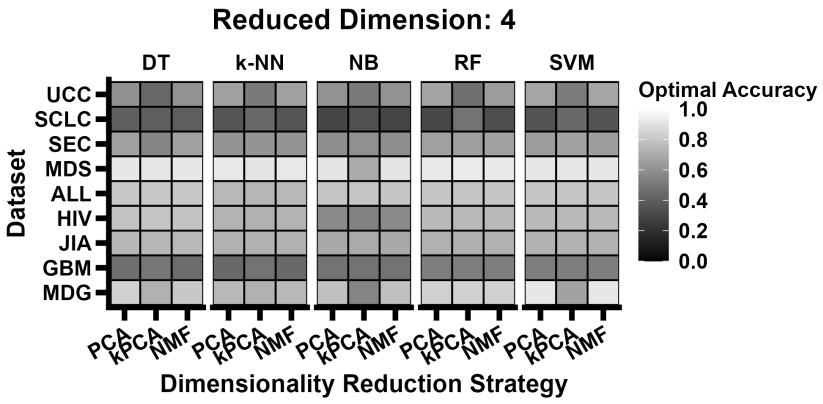

Figure 3: For a fixed dimension ($d = 4$ shown here), the choice of dimensionality reduction method between PCA, kPCA, and NMF has little impact on model performance. Here, the optimal leave-one-out validation accuracy for models trained on each reduced dataset via PCA, kPCA, and NMF is reported. For most datasets, the accuracy is large the same regardless of whether PCA, kPCA, or NMF is used.

four dimensions, but a similar pattern is true across all dimensions tested (supplementary figures are available on GitHub, see below).

Far more impactful than the choice of dimensionality reduction method was the choice of reduced dimension. For some models, particularly the NB model and SVM, the leave-one-out validation accuracy depended strongly on the reduced dimension (Fig 4). Interestingly, the trend was not necessarily monotonic, and peak accuracy for some models on some datasets (e.g. the NB model on the HIV dataset, Fig 4) was achieved at a dimension less than 9. However, the leave-one-out validation accuracy was generally low in very low dimensions such as 2 or 3 across all models. Figure 4 shows the performance of all models across all nine datasets for the PCA reduction strategy, but the trends were similar for kPCA and NMF (supplementary figures are available on GitHub, see below).

In comparison to the baseline performance, the effect of PCA, kPCA, and NMF dimensionality reduction was highly variable (Fig 5). For the RF model, the performance was not highly different from baseline across most datasets, but the NB model and SVM were strongly impacted. While the change in performance was mostly negative, some models saw improved leave-one-out validation accuracy with these dimensionality reduction strategies on some datasets (i.e. the NB model on the SEC and ALL datasets and the DT model on the GBM dataset). Figure 5 shows the performance of all models across all nine datasets for the PCA reduction strategy, but the trends were similar for kPCA and NMF (supplementary figures are available on GitHub, see below).

4 Discussion

The high variability in the baseline performance of the models (Fig 1) is expected given the wide range of sample sizes and numbers of classes represented across the nine datasets. The overall poor performance on the SCLC dataset is likely due to the large number of classes (6) within the dataset and because of how similar the classes are to each other: each class describes a different stage of lung cancer progression rather than describing disease vs. healthy states [15]. The high performance on the MDS dataset is also expected, as the MDS dataset has the highest number of data points, 200, and only two classes: a disease state (myelodysplastic syndrome) and a healthy state [18]. The NB model assumes that each feature is independent to solve the classification problem. This strong assumption may explain the generally poor baseline performance of the NB model (Fig 1). Since each feature here represents the expression level of a given gene, and because some genes influence the expression of others, assuming independence makes little biological sense. In contrast, the generally good performance of the SVM (Fig 1) also makes sense, as SVMs are known to perform well with limited data in high dimension due to the implicit feature mapping of the kernel [26].

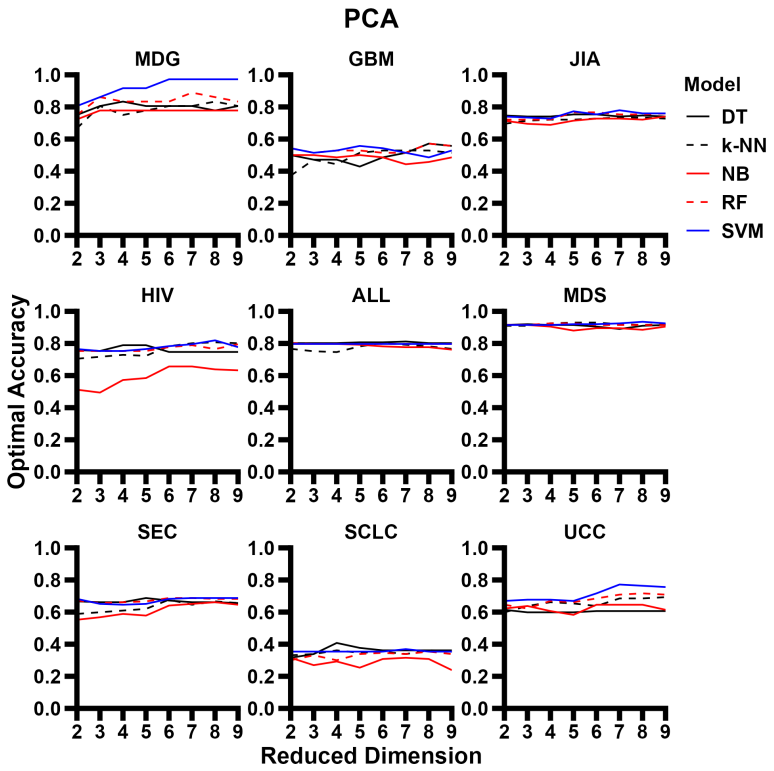

Figure 4: For a fixed dimensionality reduction strategy (PCA shown here), the choice of dimension had a strong impact on model performance across multiple datasets. The optimal dimension varied with dataset and model. Here, the optimal leave-one-out validation accuracy for models trained on each dataset reduced to dimensionality $d = 2 \ldots 9$ is reported.

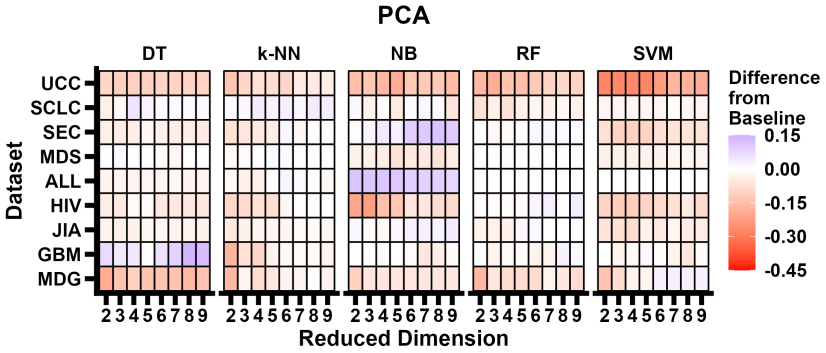

Figure 5: For a fixed dimensionality reduction strategy (PCA shown here), the performance of the model in comparison to the baseline performance is variable. The performance of the RF model was not heavily impacted by this dimensionality reduction for most datasets, whereas the NB and SVM were heavily impacted.

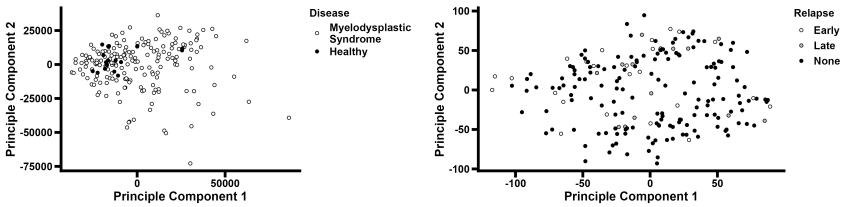

Figure 6: PCA-reduced MDS (left) and ALL (right) datasets. With fewer dimensions, the data points group together and appear "well-mixed," which may make them difficult to separate.

The inconsistencies in the data aggregation strategy (Fig 2) may be due to underlying biological context. In particular, cancer is a complicated disease that arises from mutations tend to co-occur in genes with similar roles or that belong to the same biological pathway [27], so aggregating the features according to their GO categories may result in a reduced set of features that are more informative for cancer datasets, which can aid the DT model. However, for other datasets, if a very small number of individual genes are important, then taking the mean in the aggregation process may instead obscure important the fine-grain features and instead harm performance. Future experiments could consider different aggregation strategies (e.g. taking the max, geometric mean, etc. instead of the arithmetic mean) or other gene categorization/annotation schemes to explore how to incorporate existing biological knowledge into the dimensionality reduction.

Interestingly, the leave-one-validation accuracy of models trained on datasets reduced by PCA, kPCA, and NMF were similar (Fig 3). The results for PCA and NMF in particular were nearly identical across all models and all datasets. One important difference between PCA and NMF is that NMF does not re-center the data. These similarities may then indicate that re-centering the data has little impact on classification problems in RNA-seq data, and that the relative changes in gene expression are more important than the "center point" of the data. kPCA did impact the performance of some datasets and some models, namely the SCLC, UCC, and MDG datasets and the NB model (Fig 3). However, these impacts were mostly found on the most limiting cases of the datasets surveyed: the MDG dataset had the fewest data points (36) [14], the SCLC dataset had the most classes (6) and overall lowest baseline performance [15]), and the NB model performed the worst at baseline out of all models surveyed. This may indicate that for more limiting cases (i.e. when data is especially scarce or the models perform especially poorly) simpler, more direct dimensionality methods may be more appropriate than kPCA, but future research can be directed at testing different kernels for kPCA to assess the impact of kernel choice on model performance in these limiting cases. Nonetheless, for most of the datasets, the choice between PCA, kPCA, and NMF had little impact on model performance for any fixed choice of reduced dimension.

In contrast to the choice between PCA, kPCA, and NMF, the choice of reduced dimension had a large impact on the leave-one-out validation accuracy (Fig 4). The NB and SVM models in particular were very strongly impacted by the choice of reduced dimension. The maximal leave-one-out validation accuracy was not necessarily achieved on the highest dimension (e.g. the NB model on the HIV dataset or the SVM model on the MDG or UCC datasets), which may be indicative of overfitting, even with as few as $d \leq 9$ dimensions. Given the high levels of noise in RNA-seq data [28], this is not surprising: including too many principle components/features may reintroduce some of the noise present in the original dataset, and the models may then overfit to this noise. However, the performance of every model on some datasets, namely the ALL and MDS datasets, were not heavily impacted by the choice of reduced dimension (Fig 4). This may be due to poor spatial separation of the classes (Fig 6) that may have persisted in up to $d = 9$ dimensions. Both of these datasets had higher numbers of data points compared to the other datasets (Table 2), and future studies could look into extending the number of reduced dimensions beyond 10 for other similarly-sized datasets.

The highly variable impact of PCA, kPCA, and NMF dimensionality reduction on the leave-one-out validation accuracy (Fig 5) may also be due to the way the classification models work and the spatial organization of the datasets. The NB model showed improved performance on the SEC and ALL datasets, and this may be because, unlike the individual genes, the principle components or reduced features may behave independently. The DT also saw improved performance on the GBM dataset, and this may be because reducing the dimension to a small number via PCA, kPCA, and NMF reduces the noise present in RNA-seq data [28], which may allow the DT to choose more informative

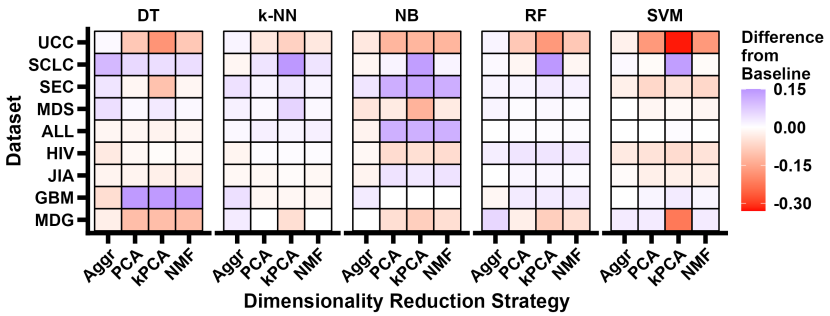

Figure 7: Impact of dimensionality reduction on model performance is variable and specific. The change in classification accuracy from baseline varies heavily between the nine different datasets, the five different models, and the four different dimensionality reduction strategies. Here, "Aggr" refers to the *a priori* aggregation strategy, and the accuracies are reported as the optimal accuracy across all reduced dimensions/GO categories for each dimensionality reduction strategy.

features to split the data on. In contrast, the SVM saw generally lowered performance under PCA, kPCA, and NMF dimensionality reduction across most datasets, and this may be because reducing the dimension may result in data that is more difficult to separate linearly (Fig 6), especially for datasets with more than two classes. Consistent with this notion, the SVM generally saw improved performance with increased dimension.

5 Conclusions

Classifying disease state from human RNA-seq data is a difficult challenge. The lack of large datasets and suitable pretrained models require the use of older, simpler models such as the DT, k-NN, NB classifier, RF, and SVM. RNA-seq data is also very high-dimensional, which can hinder the performance of these models. Here, we surveyed four different dimensionality reduction strategies across nine RNA-seq datasets to assess the impact of dimensionality reduction on the performance of five simple machine-learning models commonly used in biology. The wide variety and inconsistency in the datasets make it difficult to draw general conclusions. The impact of dimensionality reduction appears to be highly specific to the dataset, the model being used, and the underlying biological context (Fig 7).

The performance of the *a priori* aggregation strategy is highly variable across datasets and is likely strongly influenced by underlying biological context, while the PCA, kPCA, and NMF methods had mostly similar impacts on model performance across most datasets. kPCA may have a different impact on datasets in limiting cases with few data points and lots of data classes, but the choice of reduced dimension had a much stronger impact than the choice between PCA, kPCA and NMF, however this impact was variable across datasets and models. Together, these results indicate that dimensionality reduction can improve model performance and is a valuable tool for classifying disease states from RNA-seq data. However, it must be used with caution, and a strong understanding of the underlying biological context, spatial organization of the data, and nature of the classification problem (i.e. number of data points, number of data classes, etc.) is needed to inform the choice of dimensionality reduction strategy, optimal classification model to use, and reduced dimension.

6 Data/Code Availability and Supporting Information

All the data and code used in the developing this report, as well as supplementary figures and supporting information, are available at `https://github.com/nolan-middleton/CPT_S-Group-Project-Fall-2025`.

References

[1] Tingpeng Yang, Yonghong He, and Yu Wang. Introducing tec-lncmir for prediction of lncrna-mirna interactions through deep learning of rna sequences. *Brief Bioinform*, 26(1):bbaf046, 02 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf046. URL https://doi.org/10.1093/bib/bbaf046.

[2] Hannah E Wilson, Scott Stevison, Levi Lamprey, and John J Wyrick. Mapping transcription factor binding sites by learning uv damage fingerprints. *Nucleic Acids Res*, 53(19):gkaf1014, 10 2025. ISSN 1362-4962. doi: 10.1093/nar/gkaf1014. URL https://doi.org/10.1093/nar/gkaf1014.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

[4] Jocelyn Krebs, Elliott Goldstein, and Stephen Kilpatrick. *Lewin's Genes XII*. Jones & Bartlett Learning, 5 Wall Street, Burlington, MA 01803, 12 edition, 2018. ISBN 9781284104493.

[5] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009. doi: 10.1038/nrg2484. URL https://doi.org/10.1038/nrg2484.

[6] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, 6(5):377–382, May 2009.

[7] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods*, 21(8):1470–1480, August 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

[8] Shreshth Gandhi, Farnoosh Javadi, Valentine Svensson, Umair Khan, Matthew G. Jones, Johnny Yu, Daniele Merico, Hani Goodarzi, and Nima Alidoust. Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. *bioRxiv*, 2025. doi: 10.1101/2025.10.23.683759. URL https://www.biorxiv.org/content/10.1101/2025.10.23.683759v1.

[9] Zhuoyi Wei Yun Su Ningyuan Shangguan Shuangyu Yang Chengyang Zhang Wenbing Li Jinbo Zhang Nan Fang Hongyu Zhang Huiying Zhao Yutong Lu Jue Fan Weijiang Yu Yuansong Zeng, Jiancong Xie and Yuedong Yang. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat Commun*, 16(4679), May 2025. doi: 10.1038/s41467-025-59926-5. URL https://doi.org/10.1038/s41467-025-59926-5.

[10] Tracy Boakye Serebour, Adam P. Cribbs, Mathew J. Baldwin, Collen Masimirembwa, Zedias Chikwambi, Angeliki Kerasidou, and Sarah J. B. Snelling. Overcoming barriers to single-cell rna sequencing adoption in low- and middle-income countries. *Eur J Hum Genet*, 32(10):1206–1213, October 2024. doi: 10.1038/s41431-024-01564-4. URL https://doi.org/10.1038/s41431-024-01564-4.

[11] Maxence Gélard, Guillaume Richard, Thomas Pierrot, and Paul-Henry Cournède. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 384–400. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v259/gelard25a.html.

[12] Liu D Wang N He D Wu Z Zhu X Wen X Li X Li J Wang Z Wang X, Wang H. Deep learning using bulk rna-seq data expands cell landscape identification in tumor microenvironment. *Oncoimmunology*, 25(11), February 2022. doi: 10.1080/2162402X.2022.2043662.

[13] Allison P. Heath, Vincent Ferretti, Stuti Agrawal, Maksim An, James C. Angelakos, Renuka Arya, Rosita Bajari, Bilal Baqar, Justin H. B. Barnowski, Jeffrey Burt, Ann Catton, Brandon F. Chan, Fay Chu, Kim Cullion, Tanja Davidsen, Phuong-My Do, Christian Dompierre, Martin L. Ferguson, Michael S. Fitzsimons, Michael Ford, Miyuki Fukuma, Sharon Gaheen, Gajanan L. Ganji, Tzintzuni I. Garcia, Sameera S. George, Daniela S. Gerhard, Francois Gerthoffert, Fauzi Gomez, Kang Han, Kyle M. Hernandez, Biju Issac, Richard Jackson, Mark A. Jensen, Sid Joshi, Ajinkya Kadam, Aishmit Khurana, Kyle M. J. Kim, Victoria E. Kraft, Shenglai Li, Tara M. Lichtenberg, Janice Lodato, Laxmi Lolla, Plamen Martinov, Jeffrey A. Mazzone, Daniel P. Miller, Ian Miller, Joshua S. Miller, Koji Miyauchi, Mark W. Murphy, Thomas Nullet, Rowland O. Ogwara, Francisco M. Ortuño, Jesús Pedrosa, Phuong L. Pham, Maxim Y. Popov, James J. Porter, Raymond Powell, Karl Rademacher, Colin P. Reid, Samantha Rich, Bessie Rogel, Himanso Sahni, Jeremiah H. Savage, Kyle A. Schmitt, Trevar J. Simmons, Joseph Sislow, Jonathan Spring, Lincoln Stein, Sean Sullivan, Yajing Tang, Mathangi Thiagarajan, Heather D. Troyer, Chang Wang, Zhining Wang, Bedford L. West, Alex Wilmer, Shane Wilson, Kaman Wu, William P. Wysocki, Linda Xiang, Joseph T. Yamada, Liming Yang, Christine Yu, Christina K. Yung, Jean Claude Zenklusen, Junjun Zhang, Zhenyu Zhang, Yuanheng Zhao, Ariz Zubair, Louis M. Staudt, and Robert L. Grossman. The nci genomic data commons. *Nat Genet*, 53(3):257–262, March 2021. doi: 10.1038/s41588-021-00791-5. URL `https://doi.org/10.1038/s41588-021-00791-5`.

[14] Nataly Strunnikova, Sara Hilmer, Jessica Flippin, Michael Robinson, Eric Hoffman, and Karl G Csaky. Differences in gene expression profiles in dermal fibroblasts from control and patients with age-related macular degeneration elicited by oxidative injury. *Free Radic Biol Med*, 39(6): 781–796, September 2005.

[15] Mitch Raponi, Yi Zhang, Jack Yu, Guoan Chen, Grace Lee, Jeremy M G Taylor, James Macdonald, Dafydd Thomas, Christopher Moskaluk, Yixin Wang, and David G Beer. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res*, 66(15):7466–7472, August 2006.

[16] Michael E Burczynski, Ron L Peterson, Natalie C Twine, Krystyna A Zuberek, Brendan J Brodeur, Lori Casciotti, Vasu Maganti, Padma S Reddy, Andrew Strahs, Fred Immermann, Walter Spinelli, Ulrich Schwertschlag, Anna M Slager, Monette M Cotreau, and Andrew J Dorner. Molecular classification of crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J Mol Diagn*, 8(1):51–61, February 2006.

[17] Adam M Gustafson, Raffaella Soldi, Christina Anderlind, Mary Beth Scholand, Jun Qian, Xiaohui Zhang, Kendal Cooper, Darren Walker, Annette McWilliams, Gang Liu, Eva Szabo, Jerome Brody, Pierre P Massion, Marc E Lenburg, Stephen Lam, Andrea H Bild, and Avrum Spira. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci Transl Med*, 2(26):26ra25, April 2010.

[18] Petra Gorombei, Fabien Guidez, Saravanan Ganesan, Mathieu Chiquet, Andrea Pellagatti, Laure Goursaud, Nilgun Tekin, Stephanie Beurlet, Satyananda Patel, Laura Guerenne, Carole Le Pogam, Niclas Setterblad, Pierre de la Grange, Christophe LeBoeuf, Anne Janin, Maria-Elena Noguera, Laure Sarda-Mantel, Pascale Merlet, Jacqueline Boultwood, Marina Konopleva, Michael Andreeff, Robert West, Marika Pla, Lionel Adès, Pierre Fenaux, Patricia Krief, Christine Chomienne, Nader Omidvar, and Rose Ann Padua. BCL-2 inhibitor ABT-737 effectively targets leukemia-initiating cells with differential regulation of relevant genes leading to extended survival in a NRAS/BCL-2 mouse model of high risk-myelodysplastic syndrome. *Int J Mol Sci*, 22(19):10658, September 2021.

[19] Lüder Hinrich Meyer, Sarah Mirjam Eckhoff, Manon Queudeville, Johann Michael Kraus, Marco Giordan, Jana Stursberg, Andrea Zangrando, Elena Vendramini, Anja Möricke, Martin Zimmermann, Andre Schrauder, Georgia Lahr, Karlheinz Holzmann, Martin Schrappe, Giuseppe Basso, Karsten Stahnke, Hans Armin Kestler, Geertruy Te Kronnie, and Klaus-Michael Debatin. Early relapse in ALL is identified by time to leukemia in NOD/SCID mice

and is characterized by a gene signature involving survival pathways. *Cancer Cell*, 19(2): 206–217, February 2011.

[20] Caryn G Morse, Joachim G Voss, Goran Rakocevic, Mary McLaughlin, Carol L Vinton, Charles Huber, Xiaojun Hu, Jun Yang, Da Wei Huang, Carolea Logun, Robert L Danner, Zoila G Rangel, Peter J Munson, Jan M Orenstein, Elisabeth J Rushing, Richard A Lempicki, Marinos C Dalakas, and Joseph A Kovacs. HIV infection and antiretroviral therapy have divergent effects on mitochondria in adipose tissue. *J Infect Dis*, 205(12):1778–1787, June 2012.

[21] Claas H Hinze, Ndate Fall, Sherry Thornton, Jun Q Mo, Bruce J Aronow, Gerlinde Layh-Schmitt, Thomas A Griffin, Susan D Thompson, Robert A Colbert, David N Glass, Michael G Barnes, and Alexei A Grom. Immature cell populations and an erythropoiesis gene-expression signature in systemic juvenile idiopathic arthritis: implications for pathogenesis. *Arthritis Res Ther*, 12(3):R123, June 2010.

[22] Guido Reifenberger, Ruthild G Weber, Vera Riehmer, Kerstin Kaulich, Edith Willscher, Henry Wirth, Jens Gietzelt, Bettina Hentschel, Manfred Westphal, Matthias Simon, Gabriele Schackert, Johannes Schramm, Jakob Matschke, Michael C Sabel, Dorothee Gramatzki, Jörg Felsberg, Christian Hartmann, Joachim P Steinbach, Uwe Schlegel, Wolfgang Wick, Bernhard Radlwimmer, Torsten Pietsch, Jörg C Tonn, Andreas von Deimling, Hans Binder, Michael Weller, Markus Loeffler, and German Glioma Network. Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. *Int J Cancer*, 135(8):1822–1831, October 2014.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[25] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard,

Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.

[26] Derek A. Pisner and David M. Schnyer. Chapter 6 - support vector machine. In Andrea Mechelli and Sandra Vieira, editors, *Machine Learning*, pages 101–121. Academic Press, 2020. ISBN 978-0-12-815739-8. doi: https://doi.org/10.1016/B978-0-12-815739-8.00006-7. URL https://www.sciencedirect.com/science/article/pii/B9780128157398000067.

[27] Musalula Sinkala. Mutational landscape of cancer-driver genes across human cancers. *Sci Rep*, 13(1), August 2023. doi: 10.1038/s41598-023-39608-2. URL https://doi.org/10.1038/s41598-023-39608-2.

[28] Ales Varabyou, Steven L Salzberg, and Mihaela Pertea. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. *Genome Res*, 31 (2):301–308, February 2021.