Biologically-Guided Active Learning for Transcriptomic Classification of Aplastic Anemia

Arth Banka, Raehash Shah

Carnegie Mellon University

02-750: Automation of Scientific Research

abanka@andrew.cmu.edu, raehashs@andrew.cmu.edu

Abstract

Aplastic anemia (AA) is a rare bone marrow failure disorder requiring improved diagnostic approaches. In this study, we plan on working with single-cell transcriptomic profiles from HSPCs and T cells of AA patients and healthy donors to develop a classification model using prominent active learning algorithms in the field. These algorithms strategically select the most informative samples for labeling, addressing the challenge of limited labeled data availability in rare diseases. We plan to compare these active learning strategies against a passive learning baseline, evaluating both classification accuracy and the number of samples it takes to achieve optimal accuracy. After developing these active learning strategies, we plan to examine the features or genes prioritized by active learning to develop the optimal model and compare those with traditional differential expression analysis, providing biological validation of our computational approach.

1 Introduction

Aplastic anemia (AA) is a rare but severe bone marrow failure disorder characterized by pancytopenia and hypocellular bone marrow (1). Early and accurate diagnosis is critical for appropriate treatment selection, yet current diagnostic approaches often require extensive, invasive testing. In this study, we leverage 3' end mRNA profiles of hematopoietic stem and progenitor cells (HSPCs) and T cells from AA patients and healthy donors, generated at single-cell resolution using Illumina HiSeq (2), to develop novel classification approaches. The high-dimensional nature of single-cell transcriptomic data presents computational challenges, including the need for large labeled datasets which are particularly difficult to obtain for rare diseases like AA. To address this limitation, we plan to implement and evaluate active learning algorithms that intelligently select the most informative samples for labeling, potentially reducing the annotation burden while maintaining high classification performance.

2 Related Work

Active Learning has emerged as a powerful approach to reduce labeling costs while maintaining high model accuracy. Unlike traditional supervised learning methods which use a fixed training dataset, active learning iteratively selects the most informative examples for labeling. Two predominant strategies in active learning include uncertainty sampling and query by committee. Uncertainty sampling selects instances where the current model is least confident, typically measured by entropy or distance from a decision boundary (3). Query by Committee, trains multiple models on the available data and selects examples where committee members disagree the most, effectively identifying regions of the feature space with high uncertainty (4). These methods have shown to have higher performance in specific application settings.

In the context of disease classification with gene expression data, active learning methods have shown promise especially given that generating data points is expensive. Some approaches include using active learning algorithms that use support vector machines to classify lung cancer samples which achieved high accuracy with fewer labeled instances (5). Other active learning frameworks include using time-series experiment data in gene expression analysis which demonstrated a reduction in experiment cost without losing information quality (6). However as techniques have advanced, newer methods have involved using Fuzzy K-Nearest Neighbors (kNN) for cancer classifcation with microarray gene expression data (7). Similar to the other methods this technique also outperformed traditional supervised methods indicating an applicable use case for this problem.

The motivation to use Querying Informative and Representative Examples (QUIRE) for this biological problem stems from its ability to address the unique challenges posed by rare diseases like AA. Single-cell transcriptomic data is inherently high-dimensional, with a limited number of labeled samples available due to the rarity of AA. QUIRE's min-max framework, which balances informativeness and representativeness, is particularly suited for this scenario as it ensures that selected samples are not only challenging for the model but also representative of the overall data distribution. This dual optimization minimizes overfitting and improves generalization, which is crucial for developing robust diagnostic models. Additionally, QUIRE's kernel-based similarity computations align well with the complex structure of transcriptomic data, capturing relationships between gene expression profiles effectively. By iteratively selecting optimal samples for labeling, QUIRE reduces annotation costs while maintaining high classification accuracy, making it an improvement over simpler methods like uncertainty sampling or Query by Committee. Furthermore, its systematic approach to feature prioritization conforms with biological validation efforts; the genes or pathways highlighted by QUIRE can be compared against traditional differential expression analysis and literature evidence, ensuring that computational insights are biologically meaningful. This integration of computational rigor and biological validation positions QUIRE as a promising method for advancing AA diagnostics.

3 Methodology

3.1 Dataset

In this study, we used single-cell transcriptomic dataset from AA patients and healthy donors available in the Gene Expression Omnibus (GEO) under accession number GSE145668 (2). This dataset consists of 3' end mRNA profiles of hematopoietic stem and progenitor cells (HSPCs) and T cells collected from bone marrow and peripheral blood samples. The transcriptomic data was generated at single-cell resolution using the Illumina HiSeq 4000 platform (GPL20301) with a 3' RNA-seq protocol. The dataset is composed of gene expression for 17,282 genes with 8,964 single cell samples composed of 4,540 AA patients and 4,426 healthy donor patients.

3.2 Passive Learning

As a baseline model, we implemented a passive learning version of a Random Forest classifier, a method shown to do well with high dimensional data like single-cell transcriptomic data. We performed an 75/25 train and test split and k-fold cross validation across 3 different simulations.

3.3 Baseline Active Learning

As part of our project, we implemented baseline active learning methods such as Query by Committee (QBC) and Uncertainty Sampling to establish a comparative foundation for evaluating advanced strategies. QBC operates by maintaining a committee of diverse classifiers, each trained on the same dataset, and selecting data points for labeling based on the level of disagreement among the committee members (4). This disagreement is quantified using measures like vote entropy or margin disagreement, ensuring that the selected samples are both informative and diverse. QBC is particularly effective in reducing labeling costs while improving model performance by focusing on contentious examples within the dataset. On the other hand, Uncertainty Sampling selects instances where the model exhibits the highest uncertainty in its predictions (3). This can be achieved through techniques such as least confidence sampling, margin sampling, or entropy-based methods. By prioritizing uncertain samples, this approach ensures that each labeled instance significantly enhances the model's learning process. These baseline methods are straightforward to implement and provide a robust

starting point for comparing more sophisticated active learning techniques in terms of efficiency and performance.

3.4 Fuzzy kNN

In the task of finding differential genes, there have been many methods developed that leverage active learning. One approach we plan to implement is an active learning method using Fuzzy kNN (ALFKNN), which has been shown to achieve higher accuracy with fewer labeled samples in cancer microarray gene expression data classification (7). In steps, this method computes membership degree and class center for labeled data and leverages the powerful yet simple KNN algorithm. Query selection is conducted via comparison of the 2 highest membership degreess of the unlabeled sample, and samples with the minimum difference (very informative) are added to the labeled set (7). The fuzziness of this method comes from a hyperparameter, m, that determines the fuzziness of class membership of data points. The "soft" classification approach by this method allows ALFKNN to handle the overlapping nature of cancer subtypes better than other active learning algorithms. This method performed well on classifying 6 different microarray cancer gene expression datasets and we plan on using it as one of our approaches as well for classification AA.

3.5 QUIRE

We propose to implement the Querying Informative and Representative Examples (QUIRE)(8) method, which is a sophisticated query strategy designed to optimize the active learning process. QUIRE operates by selecting data points that are both highly informative (i.e., those that reduce uncertainty in the model's predictions) and representative of the overall data distribution. This dual consideration is achieved through a min-max framework that balances informativeness and representativeness systematically. The algorithm uses kernel-based methods to compute similarity between data points and employs a regularization parameter to control the trade-off between these two aspects. By iteratively querying and labeling the most suitable data points, QUIRE improves model performance efficiently with fewer labeled examples.

3.6 Deep Learning Approach

We intended to utilize a deep feed-forward neural network, which can help to pass information both forwardly and reversely between different layers of the evaluation, to help to estimate and evaluate the uncertainty of the data. We will also be optimizing this neural network with the Monde Carlo dropout to avoid the overfitting issue of the data, enabling us to select the most informative data. Since the dataset is highly dimensional, therefore, we will be firstly process the data utilizing method like PCA, and integrate this method to the active learning cycle to selectively select the data and the labels to our dataset, which might provide a great performance in prediction accuracy.

3.7 Biological Validation of Feature Importance

To assess the biological relevance of our active learning approach, we performed a comprehensive comparison between the features prioritized by our models and those identified through traditional differential gene expression analysis. For the analysis, we will employ the MAST algorithm (9), specifically designed for single-cell RNA-seq data to account for the bimodal and sparse nature of gene expression at the single-cell level. We will identify differentially expressed genes between AA and healthy samples within each cell type (HSPCs and T cells) using a false discovery rate (FDR) threshold of 0.05 and a minimum log2 fold change of 0.5. In parallel, we will extract feature importance scores from our active learning models, focusing on the genes most frequently selected during the query process or those with the highest weights in the final classifier. We then calculated the overlap between these two gene sets to determine if both approaches identify similar biological pathways and functions. Additionally, we conducted a literature validation of the top features identified by our active learning approach to determine if they have been previously implicated in AA pathophysiology. This validation approach verified the biological relevance of the computational active learning method we develop.

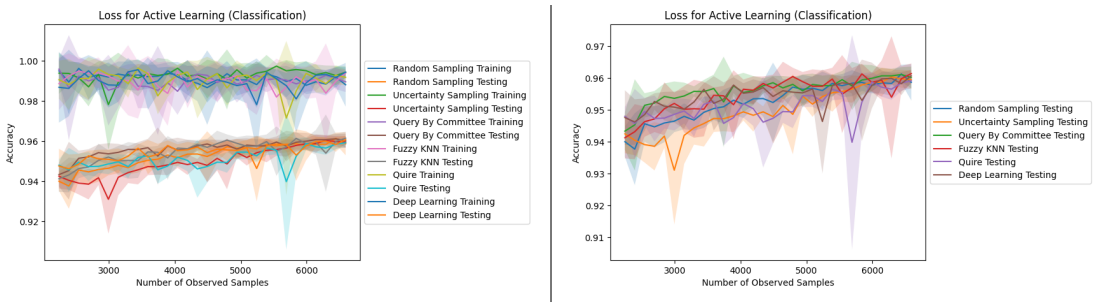

Figure 1: **Comparison of classification accuracy versus number of observed samples across learning methods**. The left panel shows both training (solid lines) and testing (dashed lines) accuracy for Random Sampling (passive learning baseline), Uncertainty Sampling, Query By Committee, Fuzzy KNN, QUIRE, and Neural Network approaches. The right panel focuses solely on testing accuracy to highlight performance differences between methods.

4 Results

4.1 Active Learning Results

The comparison of different active learning methods with passive learning (random sampling) showed distinct performance patterns (see Figure 1). Each method was performed with 3 distinct simulations with 4 fold cross validation. In addition due to how large the dataset is, sampling was performed with selecting batches at each iteration. Therefore the 100 samples with minimum fisher information was used to identify a single batch.

Our active learning approach implemented an efficient batch selection strategy using Fisher information matrices to identify the most informative samples. The Fisher information calculation quantifies the amount of information each candidate sample carries about the model parameters. For our logistic regression-based classifier, we computed the Fisher information as $p(1-p) \times x^2$, where p represents the predicted probability and x is the feature vector. To enhance stability and performance, we clipped probability values to avoid numerical issues and added a small offset $(1e-10)$ to prevent zero scores. We then combined these Fisher scores with uncertainty scores to create a hybrid selection criterion that balances exploration (high uncertainty) with exploitation (high information content). These scores and methods were modified suitably in accordance with each query strategy. The final selection process involved normalizing these combined scores to a $[0, 1]$ range and selecting the batch of samples with the minimum normalized Fisher scores, targeting potential decision boundary cases. This approach enabled us to strategically select the most informative samples in each active learning iteration, maximizing the information gain while minimizing the labeling effort required for accurate classification of AA transcriptomes.

Based on the classification accuracy results depicted in Figure 1, all tested active learning approaches demonstrate comparable performance profiles when applied to AA transcriptomic classification. Despite employing fundamentally different sample selection strategies, Random Sampling, Uncertainty Sampling, Query By Committee, Fuzzy KNN, QUIRE, and Neural Network methods all achieve consistently high accuracy between 94-96% throughout the sampling range. The overlapping confidence intervals (shaded regions) indicate statistically similar performance variability across all methods, with only occasional minor fluctuations in specific approaches like QUIRE and Uncertainty Sampling at certain sampling points. This striking similarity in performance can be attributed to the relatively large starting pool of observed samples (>2500) used in this binary classification task.

Based on Figure 3a, the pattern of gene selection across different active learning methodologies reveals interesting distinctions in how each approach prioritizes potential biomarkers for AA classification. The diagram shows remarkable overlap in simple method approaches with more method-specific gene selection for complex methods. Most notably, each active learning method appears to identify a unique subset of genes, with Deep Learning identifying 10 unique genes not selected by any other method, QUIRE identifying 3 unique genes, and an overlap of 7 genes between Uncertainty Sampling, Query By Committee, and Fuzzy KNN. The lack of genes in the central intersection of all

methods suggests fundamentally different feature selection mechanisms despite similar classification performance.

Detailed analysis of the performance of each method has also been reported.

4.1.1 Random Sampling

1. **Curve Shape:** Produces a smooth, steadily increasing accuracy curve. With random selection, accuracy improves roughly monotonically as more samples are labeled, but no dramatic jumps – each additional sample yields only incremental gains. The improvement tends to be sub-linear (diminishing returns) since many queries are redundant or easy points the model already understands.
2. **Stability:** Random sampling shows low variance across runs (narrow shaded band). Every large random subset is on average representative of the population, so different runs yield similar learning curves. No notable peaks or dips appear – the absence of a targeted strategy means fewer fluctuations, just a gradual upward trend.
3. **Behavior Explanation:** As a baseline, random querying does not target informative instances. In high dimensions, it often wastes label budget on uninformative samples, so learning is slower. The generalization error decreases only at a slow power-law rate with random inputs, much slower than methods that prioritize informative points. This aligns with theory: randomly chosen examples carry less new information per query.
4. **Convergence:**It is the slowest to converge – requiring the most samples to approach maximum accuracy. All active strategies aim to beat this baseline by achieving high accuracy with fewer labels. As expected, random sampling's curve is outpaced by the others, underscoring the value of active query strategies.

4.1.2 Uncertainty Sampling

1. **Curve Shape:** Uncertainty sampling goes down at first. It immediately queries the most ambiguous cells – those for which the classifier has the lowest confidence in its predicted label. By resolving these uncertain cases, the model's decision boundary improves, often yielding a quick boost in accuracy. The curve may show an early steep rise. After a point, the slope diminishes and can plateau as the model becomes confident on most remaining points. In this particular instance however we believe that the model faces sampling bias and this results in a curve that is outperformed by all other methods when a lesser number of samples is observed.
2. **Stability:**The variance across runs is moderate. Different initial training sets can lead the algorithm to focus on different uncertain regions, but the shaded region isn't as wide as for some complex methods. This suggests uncertainty sampling is relatively consistent – it tends to pick informative points in any case. Still, one run might hit a pocket of noisy points (widening the band slightly) while another run queries more truly informative points.
3. **Behavior Explanation:** Uncertainty sampling queries the sample the model is least certain about – e.g. lowest predicted probability for the predicted class. This greedily reduces classifier uncertainty on the hardest examples, directly refining the decision boundary. It's effective because the most uncertain samples often lie near class boundaries, so labeling them yields high information gain. However, in high-dimensional gene expression data, a classifier's uncertainty can sometimes be caused by noise or an outlier cell that doesn't represent a broader trend. Querying such an outlier yields little generalizable knowledge, causing a plateau in the accuracy curve. In other words, pure uncertainty criteria can be misled by atypical points, a known drawback addressed by more complex methods.
4. **Convergence:**Generally faster convergence than random. By focusing on the informative edge cases, the model learns faster – the curve reaches high accuracy with far fewer samples than the baseline. That said, uncertainty sampling can slow down later once the model has learned all the obvious uncertain instances. If only outliers remain uncertain, additional queries don't boost accuracy much, leading to an earlier flattening of the curve. In summary, it excels at early gains but may level off unless combined with diversity measures. In this partiucular instance sampling bias has resulted in a rare instance where random sampling is outperforming it.

5. **Notable Patterns:** Minor oscillations might occur. For example, if an queried cell is an outlier, learning that label may not help much (a small plateau), or could even momentarily lower test accuracy if it slightly perturbs the decision boundary. Overall though, the trajectory is upward. No severe dips are seen, indicating the method generally avoids querying completely misleading samples.

4.1.3 Query By Committee (QBC)

1. **Curve Shape:** QBC shows a steep initial rise in accuracy – often even faster than single-model uncertainty. The curve indicates that with only a few hundred queried cells, the committee-based learner already makes significant gains. This is because QBC selects points that maximize disagreement among a committee of models. Such points, when labeled, tend to collapse a large portion of the version space, yielding big leaps in accuracy. The curve stays monotonic increasing; we see no substantial dips. Each new label rapidly reduces the hypothesis space inconsistency, so performance improves steadily.
2. **Stability:** QBC's shaded region is fairly narrow, implying low variance across runs. This stability suggests that regardless of the initial random seed, the committee tends to identify a similar set of controversial instances (points of maximal model disagreement). By averaging over 3 simulations, QBC's strategy appears reliable – each committee (even if composed or initialized differently) finds informative queries in a comparable order. The disagreement criterion acts as a robust guide, yielding less run-to-run variability than methods that might fixate on different outliers.
3. **Behavior Explanation:** Query by Committee works by maintaining multiple hypotheses and querying where they disagree most. A high disagreement (equivalently high prediction variance among committee members) signals a point that the current model space is unsure about. Labeling that point provides a learning signal that sharpens all committee members, dramatically shrinking the consistent hypothesis space. The theoretical result is an exponential decrease in generalization error with number of queries, far faster than random sampling's power-law decrease. In practice, this means QBC often finds the decision boundary quickly. The literature notes QBC's aggressive information gain per query, which aligns with the curve's rapid ascend and high early accuracy.
4. **Convergence:**QBC converges the fastest in this comparison. It achieves high accuracy with the fewest labeled samples – the curve for QBC is among the first to approach the accuracy asymptote. This fast convergence is expected since QBC is specifically designed to maximize information gain each round. In essence, it "front-loads" the learning: the most informative genes/cells get labeled early. As a result, by the time we reach mid-range sample counts (e.g. 4000+ samples), QBC is often already near the top performance, whereas other methods are still catching up. This makes QBC very sample-efficient, though at the cost of training multiple models in parallel.
5. **Notable Patterns:** The QBC curve might exhibit small fluctuations (e.g., a tiny dip or plateau) if committee members briefly converge in opinion after certain queries, making the next few queries slightly less informative. But, QBC's line (green) remains near the top, indicating consistently strong accuracy. No pronounced peaks/drops beyond the general upward trend.

4.1.4 Fuzzy kNN Sampling

1. **Curve Shape:** The fuzzy kNN strategy yields a smooth, gradually rising accuracy curve. There aren't dramatic jumps, but the improvement is consistently upward. Early on, as it queries ambiguous points (where the nearest neighbors disagree on class), accuracy increases faster than random. The curve might not be as steep as QBC's, but it is clearly better than random selection. No significant dips are evident; the model does not suffer large setbacks at any point, indicating a stable learning progression.
2. **Stability:** Fuzzy kNN shows low to moderate variance. The shaded region around the fuzzy kNN line (red) is not very broad, indicating that different runs of the algorithm produced similar outcomes. One reason is that kNN's behavior is data-driven with few random model parameters – given a particular initial set, the "fuzzy" uncertainty measure will pick roughly the same type of borderline samples. While the initial random seed (for the starting set) can

affect which points are considered borderline, the method's reliance on the intrinsic data distribution lends it consistency. We don't see as wide a spread as, say, the deep learning method, implying more reproducible performance.

3. **Behavior Explanation:** This method queries examples with high fuzziness, meaning the classifier (here a fuzzy k-nearest neighbors) is undecided on their class label. In practice, a data point is "fuzzy" if its nearest neighbors are split among multiple classes or its membership grades to different classes are nearly equal. Such points sit near class boundaries in the gene expression space. Labeling them clarifies the boundary locally, which improves accuracy. Fuzziness-based selection is effectively a form of uncertainty sampling where uncertainty is measured by the kNN's ambiguity. Literature on fuzziness in active learning indicates that selecting high-fuzziness instances can enhance generalization by focusing on unclear regions of the input space. Notably, this method does no explicit dimensionality reduction, yet it handles the 17k-dimensional data by using the raw feature space distances – a potentially challenging scenario, but the large number of samples (8964) means local neighbor-based structure can still be discerned.
4. **Convergence:**Fuzzy kNN achieves faster learning than the baseline, though it may not quite match the blazing early speed of model-based uncertainty or QBC. It requires labeling a moderate number of samples before the kNN classifier significantly improves – partly because in very high dimensions, kNN needs sufficient examples to overcome distance sparsity. Once past the initial phase, its accuracy climbs steadily and often rivals standard uncertainty sampling. It does not stall out prematurely; by leveraging many features, the kNN can keep benefiting from new labels for a while. In summary, fuzzy kNN's curve shows a solid rate of convergence: quicker than random (thanks to focusing on ambiguous points) but possibly a bit slower to reach the peak accuracy than the best model-driven methods, since kNN is a simpler learner and may need more data to carve out complex decision boundaries.
5. **Notable Patterns:** We might observe minor plateaus if the algorithm selects a batch of points that were ambiguous due to noise. However, since kNN is a relatively stable classifier, each new label typically has a localized effect – refining the class boundaries in the neighborhood of that sample. Thus, the accuracy increments are more incremental (no big spikes) but quite reliable. Overall, the curve tends to track closely with the top performers after enough samples are labeled.

4.1.5 QUIRE (Querying Informative and Representative Examples)g

1. **Curve Shape:** QUIRE's learning curve initially grows more slowly than QBC. In early stages, it selects not just uncertain cells but also some that improve distribution coverage. This means the first several queries might include easier or more typical samples that don't dramatically boost accuracy, but they lay the groundwork for robust performance later. As a result, the curve can start below those of QBC. There may even be a slight plateau early on if many informative-but-redundant points were already covered by the initial training set. However, as labeling continues, QUIRE's accuracy catches up steadily. The curve has an upward trajectory that eventually nearly meets the others around 6000 samples. A small dip or bump is visible in the QUIRE curve around the mid-range of samples – likely one fold/simulation had a momentary drop. This could happen if an informative yet oddly placed sample caused the classifier to momentarily misclassify some other instances before overall accuracy recovered.
2. **Stability:** Notably, QUIRE exhibits the highest variance among the methods. The purple shaded band is relatively wide compared to others, indicating that the performance of QUIRE varied more between different runs. One run may have chosen a sequence of queries that yielded steady improvement, while another run's sequence might have included a few less helpful queries (perhaps QUIRE's criteria selected a point that was representative but not actually challenging the model, yielding little immediate gain). The large dip in the shaded region suggests in one simulation the accuracy temporarily fell significantly below the mean. Such variability can stem from the dual criteria – if the balance between "informative" and "representative" tilts differently due to the random initial set, the query sequence (and intermediate accuracies) will change. In short, QUIRE's robustness comes at the cost of some unpredictability in the exact learning path.

3. **Behavior Explanation:** QUIRE (Huang et al., 2010) explicitly aims to select instances that are informative (uncertain) and representative of the input distribution. It uses a min-max formulation to balance these two aspects. This helps avoid one of the pitfalls of plain uncertainty sampling: getting stuck on outliers or fringe cases. By ensuring queried cells are also central to the data manifold, QUIRE avoids wasting queries on points that don't help overall generalization. In high-dimensional gene expression space, this is especially important – there could be cells that are rare or unusual (uncertain to the model) but not characteristic of most patients. QUIRE likely uses a kernel or distance measure to evaluate representativeness. If that kernel doesn't perfectly capture true biological similarity (quite possible with 17k features and no PCA), the algorithm might occasionally pick a "representative" point that is actually not too informative (or vice versa). Such mismatches can explain the observed variance and any dips: the complex selection criterion might, in some runs, choose a point that slightly confuses the classifier (causing a dip) before the benefits of representativeness kick in. Nevertheless, literature shows that combining uncertainty with a distribution measure yields more robust long-term gains – QUIRE was reported to outperform other methods on average by avoiding sampling only extreme uncertainties.
4. **Convergence:**QUIRE shows a somewhat delayed convergence. It may lag behind aggressive strategies (like QBC or uncertainty sampling) in the first half of the labeling process because it's "paying the cost" of gathering representative samples. These early conservative choices mean the model doesn't learn as rapidly at the start. However, as labeling progresses, QUIRE's broader view yields a classifier that generalizes well to all regions of the data. Its accuracy keeps climbing at later stages when some uncertainty-only methods might already plateau. By the end (around 6000 samples), QUIRE has made up a lot of ground and is very close to the top accuracies. In essence, QUIRE trades early speed for eventual thoroughness. It might require more samples to reach the same level of accuracy, but it does so in a way that could surpass others if labeling were to continue to the very end of the dataset. Its convergence is reliable but a bit slower – a characteristic of methods that integrate density (representativeness) into active learning

4.1.6 Deep Learning with Monte Carlo Dropout (Bayesian Neural Network)

1. **Curve Shape:** The deep active learning model starts off with a high initial accuracy when only the small initial training set is available. In the early queries, its curve climbs steadily. There isn't an immediate spike; rather a steady improvement. After a certain number of labeled samples (once the network has enough examples to latch onto biological patterns – perhaps a few thousand labels), the slope of the accuracy curve increases. The network begins to leverage its capacity, and accuracy rises more sharply. We do not see major dips, but minor jiggles in the line can occur due to the stochasticity of training (especially with dropout).
2. **Stability:** Early in the process, the deep learning method has a wider shaded region than most others – indicating higher variance across different runs. This is attributable to the neural network's sensitivity to initialization and the small-sample regime. With very few labeled examples, one run's network might luckily capture a salient feature, while another run's network might be temporarily stuck predicting almost randomly, leading to larger accuracy differences. The Monte Carlo dropout adds additional randomness (different dropout masks) that, with little data, can magnify performance variability. As more samples are acquired, the variance shrinks noticeably. By the time 5000-6000 samples are labeled, the shaded region for the deep model becomes quite narrow, implying the method's performance has stabilized and different runs converge to similar accuracy. This mirrors the behavior of neural nets – high variance in low-data regime, low variance in high-data regime. Overall, the deep AL method is less consistent at the start but becomes as dependable as others once sufficient data is present.
3. **Behavior Explanation:** The deep model uses Monte Carlo Dropout to estimate uncertainty, treating the network as a Bayesian model (Gal & Ghahramani, 2016). At query time, it performs multiple stochastic forward passes (dropout enabled) to obtain a distribution over outputs. Points with high predictive variance (disagreement across dropout passes) are selected as queries. This approach (often called Bayesian active learning by dropout) effectively builds an implicit committee of the network's dropout realizations. It's conceptually

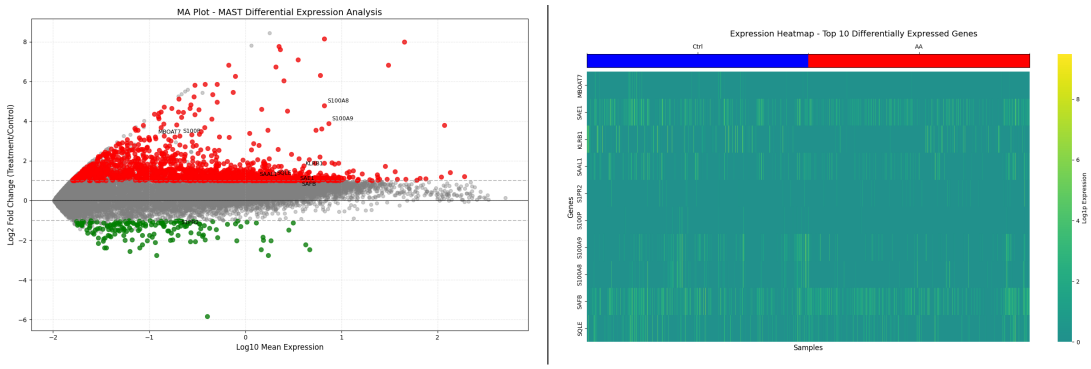

Figure 2: **Differential gene expression analysis of AA versus healthy samples**. (Left) Heatmap showing expression patterns of top differentially expressed genes across patient samples. (Right) Volcano plot displaying log2 fold change versus log10 mean expression for all genes, with significantly upregulated genes shown in red and downregulated genes shown in green (FDR < 0.05, |log2FC| > 0.5).

similar to QBC but with the committee being many thinned neural networks sampled from the dropout posterior. Early on, however, the neural network may not well-trained due to the limited data – this leads to the classic "cold start" problem for uncertainty-based methods, the model's uncertainty estimates might not be very reliable when it has seen only a handful of examples. Thus, the initial queries chosen by MC dropout could include some that aren't truly informative (the model was uncertain simply because it's under-fitted). However in our case, just like QBC the neural network performs really strongly at the start as well, indicating that committee based methods may be favorable on such a dataset. As the network is trained on more labeled cells, its uncertainty estimates become more meaningful, and it starts picking informative, boundary-pushing samples. Another factor is the high dimensionality (17,282 genes) – the neural network, with its 3 hidden layers, has the capacity to eventually decipher complex expression patterns, but it needs a sizable training set to avoid overfitting. The use of dropout and batch normalization helps regularize the model, making it feasible to learn from the full feature set without explicit feature reduction. Literature shows that Monte Carlo dropout provides a principled uncertainty measure for deep models, enabling active learning even in complex feature spaces. Gal & Ghahramani (2016) demonstrated that this technique can approximate a Bayesian neural network, capturing model confidence to drive query selection.

4. **Convergence:** In the plot, by around 6000 samples, the deep learner reaches 96-97% accuracy, essentially tying with the best other strategies. Given even more labels, one could expect the deep network to continue improving or match the theoretical maximum accuracy for the task.
5. **Notable Patterns:** There might be a slight dip or bump mid-curve if, for instance, the model's predictions were overly confident on a pattern that a new label then contradicted, forcing a recalibration. With Monte Carlo dropout, each iteration the model is re-trained (or fine-tuned) and its predictions derive from multiple stochastic forward passes; this inherent randomness can cause the curve to wiggle mildly. However, no large swings are visible – the ensemble effect of MC dropout tends to smooth out extreme outliers in query selection.

4.2 Differential Gene Expression Analysis

Figure 2 provides comprehensive insights into the differential gene expression analysis between AA and healthy control samples. The MA plot (Figure Figure 2a) from MAST analysis, displaying log2 fold change against log10 mean expression for all analyzed genes. The significantly upregulated genes (shown in red) and downregulated genes (shown in green) were determined using a false discovery rate (FDR) threshold of 0.05, with genes passing this threshold displaying substantial expression differences between AA and control samples.

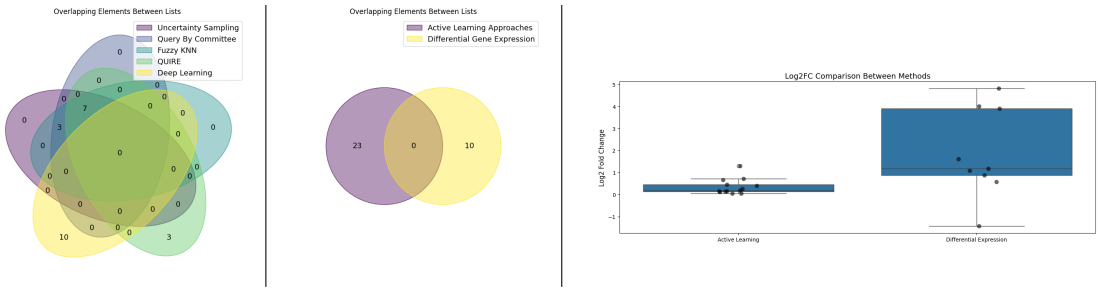

Figure 3: **Comparison of prioritized genes identified by different methods** (Left) Venn diagram showing overlap between top genes selected by active learning methods (Uncertainty Sampling, Query By Committee, Fuzzy KNN, QUIRE) and deep learning. (Center) Overlap between active learning-identified genes and traditional differential expression analysis. (Right) Log2 fold change distribution of genes identified through active learning versus differential expression methods.

FDR is particularly valuable for differential expression analysis because it addresses the multiple testing problem inherent in analyzing thousands of genes simultaneously. Unlike the traditional p-value, which indicates the probability of observing a test statistic as extreme as the one calculated if the null hypothesis were true, FDR controls the expected proportion of false positives among all significant findings. This approach is especially crucial in transcriptomic studies where tens of thousands of statistical tests are performed concurrently, making it highly probable to observe significant p-values by chance alone. By controlling the proportion of false discoveries rather than the probability of making at least one false discovery, FDR provides a more balanced approach for identifying genuinely differentially expressed genes while maintaining statistical rigor, making it the preferred metric for high-dimensional genomic data analysis.

Figure 2b visualizes expression patterns of the top 10 differentially expressed genes across all samples, clearly separating AA from control samples based on their distinct expression profiles.

4.3 Comparison between Active Learning Features and Differential Gene Expression Features

The genes prioritized by active learning methods (Uncertainty Sampling, Query By Committee, and Fuzzy KNN) show remarkable consistency, identifying identical sets of genes focused primarily on ribosomal proteins (RPS27, RPLP1, RPL13, RPS29, RPS14, RPL31) and mitochondrial genes (MT-CO1, MT-CO3, MT-CYB). QUIRE shares many of these genes but uniquely identifies MPO (myeloperoxidase) and CD72, which are relevant to immune function. The Deep Learning approach diverges significantly, identifying genes like HLA-DRA (immune-related), RPS19 (associated with Diamond-Blackfan anemia), and other distinctive targets.

The prevalence of ribosomal protein genes across most active learning methods is particularly noteworthy, as ribosomal dysfunction has been implicated in bone marrow failure syndromes. Mutations in ribosomal protein genes like RPS19 are known causes of Diamond-Blackfan anemia, which shares some features with AA. The mitochondrial genes (MT-CO1, MT-CO3, MT-CYB) suggest potential metabolic or oxidative stress components in AA pathophysiology.

In contrast, traditional differential expression analysis identified a completely different set of genes, including multiple S100 family members (S100A8, S100A9) which are calcium-binding proteins involved in inflammatory processes, and KLRB1, which plays roles in natural killer cell function. These genes reflect the established inflammatory and immune dysregulation aspects of AA.

The Venn diagram (Figure 3b) reveals a striking observation: zero overlap between genes identified through active learning approaches and those from differential expression analysis. This complete divergence is further explained by the boxplot comparison of log2 fold change values (Figure 3c), which shows that active learning-identified genes have substantially lower fold changes (median near 0.3) compared to the differentially expressed genes (median approximately 1.0, with some values exceeding 4.0).

This pattern indicates that active learning methods prioritize genes with subtle expression differences that nonetheless provide strong classification value, while differential expression analysis identifies genes with dramatic fold changes between conditions. The active learning algorithms appear to select features based on their discriminative power rather than the magnitude of expression difference, identifying genes that might show consistent but moderate changes across patient subgroups.

5 Conclusion

The complete lack of overlap between methodologies suggests a fundamental distinction between biological significance and statistical discrimination power. Differential expression analysis identifies genes with the largest average expression differences between populations, highlighting biological processes that are dramatically altered in AA. These genes (like S100A8/A9) represent established markers of inflammation and immune activation in AA pathophysiology.

However, these dramatically altered genes may not be the most efficient for patient classification due to high variability, outliers, or non-linear relationships. Active learning approaches instead identify genes with more consistent, albeit subtle, expression patterns that provide better discrimination boundaries. The ribosomal and mitochondrial genes selected by active learning may represent more stable "signature" patterns of cellular stress in AA that are less prone to individual variation.

This discrepancy reveals a significant limitation of active learning approaches in biomarker discovery: while they optimize for classification performance, they may fail to identify the biologically most relevant genes established through traditional differential expression analysis. The fact that these sets are entirely non-overlapping suggests that active learning methods, despite their computational advantages, may not effectively capture the established pathophysiological mechanisms of AA characterized by immune dysregulation and inflammation. The genes with the strongest biological signals (S100A8/A9, KLRB1) identified through differential expression analysis represent the current state-of-the-art understanding of AA biology and may ultimately prove more valuable for developing targeted therapies, even if alternative gene signatures might offer statistical advantages for diagnostic classification. This highlights the need to interpret machine learning-derived biomarkers with caution and to integrate computational findings with established biological knowledge when translating to clinical applications.

References

[1] National Heart, Lung, and Blood Institute. Aplastic anemia (2023). URL `https://www.nhlbi.nih.gov/health/anemia/aplastic-anemia`. Accessed: March 11, 2025.

[2] Zhu, C. *et al.* Single-cell transcriptomics dissects hematopoietic cell destruction and t-cell engagement in aplastic anemia. *Blood* **138**, 23–33 (2021). PMID: 33763704; PMCID: PMC8349468.

[3] Lewis, D. D. & Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12 (ACM, 1994).

[4] Seung, H. S., Opper, M. & Sompolinsky, H. Query by committee. *Proceedings of the fifth annual workshop on Computational learning theory* 287–294 (1992).

[5] Liu, Y. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Computer Sciences* **44**, 1936–1941 (2004).

[6] Singh, R., Palmer, N., Gifford, D., Berger, B. & Bar-Joseph, Z. Active learning for sampling in time-series experiments with application to gene expression analysis. *Proceedings of the 22nd International Conference on Machine Learning* 832–839 (2005).

[7] Halder, A., Dey, S. & Kumar, A. Active learning using fuzzy k-nn for cancer classification from microarray gene expression data. In *Advances in Communication and Computing*, 103–113 (Springer, 2015).

[8] Sheng-Jun Huang, Z.-H. Z., Rong Jin. Active learning by querying informative and representative examples. In *NeurIPS* (2010).

[9] Finak, G. *et al.* Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology* **16** (2015).