DETECTION AND CLASSIFICATION OF CANCER AND OTHER NONCOMMUNICABLE DISEASES USING NEURAL NETWORK MODELS

Steven Lee Gore

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

July 2023

APPROVED:

Rajeev K. Azad, Major Professor
Ron Mittler, Committee Member
Pamela Padilla, Committee Member
Vladimir Shulaev, Committee Member
Armin Mikler, Committee Member
Jyoti Shah, Chair of Department of Biological Sciences
John Quintanilla, Dean of the College of Science
Victor Prybutok, Dean of the Toulouse Graduate School

Gore, Steven Lee. *Detection and Classification of Cancer and Other Noncommunicable Diseases Using Neural Network Models*. Doctor of Philosophy (Biology), July 2023, 79 pp., 3 tables, 11 figures, 188 numbered references.

Here, we show that training with multiple noncommunicable diseases (NCDs) is both feasible and beneficial to modeling this class of diseases. We first use data from the Cancer Genome Atlas (TCGA) to train a pan cancer model, and then characterize the information the model has learned about the cancers. In doing this we show that the model has learned concepts that are relevant to the task of cancer classification. We also test the model on datasets derived independently of the TCGA cohort and show that the model is robust to data outside of its training distribution such as precancerous legions and metastatic samples.

We then utilize the cancer model as the basis of a transfer learning study where we retrain it on other, non-cancer NCDs. In doing so we show that NCDs with very differing underlying biology contain extractible information relevant to each other allowing for a broader model of NCDs to be developed with existing datasets. We then test the importance of the samples source tissue in the model and find that the NCD class and tissue source may not be independent in our model. To address this, we use the tissue encodings to create augmented samples. We test how successfully we can use these augmented samples to remove or diminish tissue source importance to NCD class through retraining the model. In doing this we make key observations about the nature of concept importance and its usefulness in future neural network explainability efforts.

ACKNOWLEDGEMENTS

I am very grateful for the support that I have received. To my professor, Dr. Rajeev K. Azad for believing in my abilities from the day we met and for your patience and skill in mentorship. My deepest gratitude for your commitment to your students and your belief in my potential. To Dr Pamela Padilla, Dr. Ron Mittler, Dr. Armin Mikler and Dr. Vladimir Shulaev for their membership on my dissertation committee and continued support and confidence in me throughout my Ph.D. program. I have valued their mentorship throughout this time. They have been a wonderful example. I would also like to thank Dr. Jyoti Shah for his role as my first scientific mentor when I was an undergraduate; his kindness, support and early mentorship set the tone for my academic and scientific endeavors. To the University of North Texas, Toulouse Graduate School, and the Department of Biological Sciences for their financial support through teaching assistantships, and summer research fellowships.

To my lab mates for the conversations and time. I would like to say thank you to Dr. Soham Sengupta, Dr. Janak Sunuwar, Dr. David Burks and Danyang Shao for being kind, professional colleagues throughout my Ph.D. program and for their continued friendship after.

Finally, I would like to thank my family. My wife Alexandra, children Isabella and Lucas who have sacrificed immensely during this time. It is their love and support that has gotten me through. Without them I could not have accomplished this and it is for their sake that I have pursued this in the first place. To my parents, Karen and Richard, for their endless support and love, for all the dinners where I was too tired to talk, for picking up the kids and watching them on short notice, for everything.

TABLE OF CONTENTS

Page

LIST OF TABLES AND FIGURES

Page

Tables

Figures

CHAPTER 1

INTRODUCTION

1.1 Noncommunicable Diseases and Epigenetics

Noncommunicable diseases (NCDs) are a group of diseases that are not caused by the acute infection of a pathogen. They are responsible for an estimated 41 million deaths each year, this accounts for approximately 74% of all deaths worldwide. Of these, 9 million are caused by cancer globally.[1] Many NCDs have complex risk factors that stem from genetic, epigenetic and environmental sources. The complex nature of these diseases makes detection and treatment difficult. The discovery of cell free circulating DNA exposed a new potentially powerful method of minimally invasive cancer detection.[2,3] The increased load of cell free circulating DNA in cancer patients, regardless of cancer type or tissue of origin, meant that tumor may be diagnosed earlier than has been the normal.[4,5]

Early events in tumorigenesis are attributable to epigenetic changes making epigenetic biomarkers attractive for early tumor detection.[6] Epigenetic changes also continue to be prevalent through all stages of tumor progression.[7-10] The same holds true for many other non-cancer NCDs as well. [11-17] A broader understanding of the epigenetic changes that underpin these diseases may also provide druggable targets and may indicate an environmental or developmental risk factor for later development of a given NCD.

1.2 Epigenetics

The term *epigenetic* refers, broadly, to changes in gene expression and chromatin structure independent of the genetic sequence itself that are heritable from parent to daughter cells.[18-21] Epigenetic mechanisms orchestrate complex developmental lineages in multicellular

organisms and generate functional diversity among cells and tissues. They integrate extracellular signals allowing the cell to respond to environmental stimuli, changing gene expression and therefore cell function without altering the genome of that cell. Due to this function, the role of epigenetics is best described as an interface between the static genome and a dynamic environment.[18,21]

Of the various mechanisms that are covered by the field of epigenetics, methylation has garnered much attention as a potentially important diagnostic and therapeutic target.[22-26] Changes in DNA methylation are observed in many NCDs and are prevalent in all cancers.[12-15,27-40] While the contribution of methylation to disease development and progression is still not well understood, it is clear that methylation plays a key role in early disease development. Additionally, methylation is persistent on cell-free circulating DNA.[2,3,41,42] The prevalence of detectable signal in the blood is vital for next generation minimally invasive diagnostics. This signal also can serve as target or proxy reporter for therapeutics. Together these qualities make methylation both a quantifiable diagnostic, a therapeutic target, and a rich source of data for monitoring therapy.

1.3 Methylation

Methylation refers to the addition of a methyl group to a cytosine residue. Often these modifications occur at cytosine-phosphate-guanine (CpG) dinucleotides.[14,43,44] The density of CpGs and their proximity to promoters of genes defines those regions as islands, shores, and shelves.[45,46] CpG islands are regions over 500 bp in length whose composition is >55% CpG and are typically found at or very close to a gene promoter. CpG shores are areas found within 2 kilobases of a CpG island and have a lower frequency than associated islands. CpG shores often

lie within genes or intergenic regions. CpG shelfs are found within 4 kilobases of a CpG island and have a lower frequency of CpGs within them.[47,48] There is an observed linear dependence on methylation status based on inter CpG distance. Generally, the greater the distance (in bp), the less likely two sites will share methylation status. Within islands, CpGs are often < 25 bps apart with islands containing up to 60 CpG sites.[49]

In development, approximately 20% of all methylated sites are dynamically regulated while the remaining are static. Genomic regions are generally referred to as hypomethylated or hypermethylated when they exhibit decreased or increased methylation levels relative to a control. When methylation changes do occur, the 'state' of the cell is altered.

1.4 Methylation and NCD Etiology

While changes in methylation are generally tightly regulated, it is possible for cells to become predisposed to disease onset through dysregulation of methylation.[6-8,10] Hypomethylation is associated with greater access to promoter regions and may turn on or upregulate downstream genes. If an oncogene is erroneously hypomethylated, this may increase a cell's likelihood of becoming precancerous.[50]

Hypermethylation is associated with expression silencing through decreased transcription factor binding efficiency and chromatin formation and may downregulate genes such as tumor suppressor genes.[51,52] The tumor suppressor p16 indirectly regulates p53 through MDM2.[51-54] When p16 is hypomethylated, the regulatory effects of the MDM2-p53 cycle are negated resulting in a build-up of p53 at which point aggregates of p53 protein may be found rendering them functionally insignificant. Hypomethylation of p16 is therefore

considered a significant cancer risk factor for some tumors such as esophageal squamous cell carcinoma (ESCC). This is without the need for a second strike to the expressed copy of p16.[51-53]

1.5 Methylation Detection

There are several methods used in DNA methylation studies, however, we focus here on array-based methods as they are still the largest source of methylation data available today.

1.5.1 Methylation Detection by Array

The most common arrays utilize the Infinium systems.[55,56] In each, the selected CpG sites are detected as unmethylated or methylated. The Infinium I system uses a two probe system while the Infinium II uses a single probe system. In both cases a CpG site is measured as methylated, M, or unmethylated, U.[55,56]

The array system relies on bisulfite conversion of unmethylated cytosines. DNA is first denatured and exposed to sodium bisulfite which deaminates unmethylated cytosines leaving a uracil in its place.[55-57] Converted DNA is then PCR amplified where the uracil is replaced by thymine.[57] Probes to detect methylated sites contain a guanine at the location of the methylated cytosine allowing the probe to form a base pair and extension to continue. At an unmethylated site a thymine exists in place of the cytosine preventing base pairing and stopping the extension.[55,56]

The methylated versus unmethylated probes may then be quantified to determine methylation level at a given site. Beta value, a common metric to quantify methylation level, is calculated as the ratio of methylated over total probes (methylated plus unmethylated) for a given site,

$$\mathrm{Beta_j} = \frac{\mathrm{M_j}}{\mathrm{M_j} + \mathrm{U_j}}$$

where U_j is the number of unmethylated probes at a site j, and M is the number of methylated probes at site j. This metric is convenient for use in neural networks as they have a range of [0,1] which is optimal in these models.[55,56,58]

1.5.2 Methylation Detection Technologies

The illumina 450k array measures 480,000 CpG sites from across the genome including CpG islands, CpG island shores, gene promoters, gene bodies, and intergenic regions, enabling researchers to examine methylation profiles at different genomic regions. Its broad range of curated sites made it attractive for a large number of studies and has resulted in a high volume of 450k array data.[55,56,58] The wide adoption of the 450k array has led to extensive validation of the array and a large number of studies and tools dedicated to its use, processing and limitations.

1.6 Modeling NCDs

Modeling of NCDs has largely been done on individual diseases in individual tissues and generally is based on small sample sizes.[59-61] The volume of information present in online repositories makes it possible to aggregate data into a larger set for use in more comprehensive models covering more tissues and more diseases. To date, no large-scale study has been conducted of any NCD outside of cancer that utilizes methylation data.

Many studies did not have the number of samples to support a statistical analysis of such a large number of sites and focus on very few sites that are the result of feature selection and dimensionality reduction pipelines. While studies so far have yielded important information

about the role of methylation in NCD development,[7-17,24,25,49,51,62-66] large-scale studies other than of cancer have not been pursued. Though a lack of a coordinated large-scale effort for non-cancerous NCDs is evident, a large volume of NCD data has become available in recent years and been accessible through various online repositories, which can be leveraged for NCD model development.

1.7 Challenges Associated with Aggregating Biological Data from Various Sources

1.7.1 Batch Effect

Batch effect refers to systematic variation in experimental data that arises from non-biological sources, such as technical variability introduced during sample processing, differing experimental conditions, or variations in reagents or equipment.[67] Such variations in individual studies are noise and may be filtered out without much consequence. However, across different studies, the noise signature introduced by each study may become a detectable signal that a model could utilize for the task it is trained on. This may impact data interpretation and introduce false associations or artifacts.[67,68]

Models trained on datasets that suffer from batch effect tend to perform poorly outside of the training data due to the reliance on information that is specific to that data including the noise introduced by batch effect intrinsic to that data. Care must be taken to identify potential batch effects and mitigate their effects before model training.[67,68]

1.7.2 Bias

Medical and biological data often contain missing or underrepresented areas in their datasets. Bias is an issue related to batch effect but is often used to describe the composition of

the dataset rather than differences in trends in each component set.[67-72] In disease studies, there could be many sources of bias that arise from clinical practices, choice of assay, location or economics. Many of the factors that create bias are often outside of the direct control of researchers. For example, they cannot force someone to miss time at work to participate in a study. While aggregating smaller datasets together may improve statistical power, improper care in selecting the datasets may introduce bias into the larger set. This reality is reflected in large biological datasets of all types.[73-75] These biases can be damaging to people who fall into the segment of the population that is missed in the dataset making the set unfair in its usefulness and outcomes. Additionally, poor sampling coverage can lead to undertrained models in those regions meaning they may seem to perform well overall but are inconsistent.[73,75] When it comes to health outcomes and therapy development, these variations in performance can be deadly.

1.7.3 Metadata

Metadata is data that describes aspects of the data of interest not discernible from the given data[76]. In modeling it is vitally important as it serves as information upon which decisions will be made during dataset selection and construction and may be used for downstream labeling of samples and post hoc analysis of a trained model.[76,77] In methylation data this is often a set of descriptors of the methylation data source such as 'Age at collection', 'Species', or 'Disease status'. In virtually all cases metadata is human entered and follows loose standards on format or necessary information. In many cases datasets have varying entries for the same field of metadata that arise from human error, such as spelling errors, or discrepancies of entry across multiple individuals, such as abbreviation being used by one individual but not by

another. Across samples there is virtually no standard of entry outside of the standards imposed by some data repositories on format such as those imposed by the National Institutes of Health: Genome Expression Omnibus (NIH GEO).

The non-uniformity and general messiness of human entries make it challenging for a data engineer to understand the initial set of data they are presented with. Standardizing and processing non-uniform metadata is incredibly time consuming, slowing down all other analysis until it is accomplished. In some cases, the importance of including metadata is ignored completely or is encoded into nonstandard locations such as sample name; as an example, a sample may be named 'SCZ_sample_age_34_male' rather than having entries for disease status, age and gender. Discovering and extracting this information is often more difficult and time consuming and generally yields far less information than standard metadata sources.

1.8 NCDs and Neural Networks

Traditional pipelines such have been widely used to identify methylation sites that are valuable as diagnostic markers or regulatory sites in NCDs. This type of statistical analysis provides significance measures for CpGs and the model's confidence in that significance estimation in a reproducible manner. Though widely used, the differential analysis approach is limited in its scope in a few important ways. When analyzing thousands to millions of CpG sites simultaneously, these methods often face the challenge of multiple hypothesis testing which can lead to an increased risk of false positives. Furthermore, they usually focus on individual CpG sites or regions causing them to potentially miss more complex patterns or interactions. Linear models used in traditional approaches may fail to capture complex, non-linear relationships between methylation patterns and experimental conditions.

Neural networks overcome many of these issues. Neural networks can intake a large number of features and are capable of learning complex, non-linear relationships. In addition, the information learned may be used to train new models through transfer learning or fine-tuning.

1.8.1 General Overview of Neural Networks and Their Function

Deep learning is based on a mathematical model called the artificial neural network (ANN),[78] which is composed of interconnected nodes (w) that process input data. The nodes are organized into layers (l) that perform a specific computation, f(). The output of one layer is fed as input to the next layer until the final layer produces the outcome prediction.

The fundamental equation of an artificial neuron is the dot product of an input vector X and an input vector, W, which is added to a bias term, b, and then the application of an activation function, f() to this term. Mathematically, this can be expressed as:

$$a = f(\sum_{l=1}^{L} w_l x_{l-1} + b_l)$$

where $x_1, x_2, ..., x_n$ are the input features, $w_1, w_2, ..., w_n$ are the corresponding weights, b is the bias term, f() is the activation function, and a is the output of the node.[78]

The input data, X, is paired with their corresponding labels denoted by Y. The neural network is trained by computing the distance between labels $\hat{Y}$ predicted by it and Y using a loss function. Stochastic gradient descent (SGD) is a commonly used algorithm used to train neural networks, which entails iteratively updating the weight vector and biases with the goal to minimize the loss function. This is done using the formula $\vartheta = \vartheta - \alpha * \nabla L(\vartheta)$ where θ is the complete set of parameters, i.e. ($w_1, w_2, \ldots w_n$) and ($b_1, b_2 \ldots b_n$), for all layers from 1st to nth in the

network. The variable α is the learning rate and $\nabla L(\theta)$ is the partial derivative of the loss function L, referred to as the gradient of the loss function.[79-81]

The learning rate, α, controls the size of the updates to the parameters and is an important hyperparameter that must be carefully chosen to ensure the algorithm converges to a good solution. If the learning rate is too large, the algorithm may overshoot the minimum of the loss function and fail to converge, while if it is too small, the algorithm may take too long to converge.

Stochastic gradient descent as described above has several weaknesses. It can fall into one of the local minima on the optimization landscape causing models to underperform and may be very slow to converge. Additionally, where the gradient is flat or saddle shaped the model may stop making meaningful updates and become stuck in an area of the gradient that is not a minimum. Variations on SGD have since been developed that attempt to address these issues such as adam,[82] mini-batch SGD,[83] RMSprop,[84] and adagrade.[85]

1.8.2 Unsupervised and Supervised Machine Learning

Neural network algorithms can be broadly categorized into two types: supervised and unsupervised learning. While the supervised learning is driven by learning of patterns inherent in a set of labeled data by the algorithm, unsupervised learning does not require characterized or labeled data to learn the patterns. In the supervised learning, the performance of the neural network is heavily dependent on the quality and quantity of labeled data available for training. In contrast, unsupervised learning can be performed on large amounts of unlabeled data, which is often easier to obtain in many domains.

In the supervised learning the neural network is trained using labeled data, wherein, Y, a

set of ground truth labels for the data, is supplied to the model. This type of learning is commonly used in tasks such as image classification, speech recognition, natural language processing, as well as many biological or biomedical domains, where the goal is to classify or predict a specific output given an input.

In the unsupervised learning of a neural network, the input X is used in place of a supplied Y in the loss function, and thus, the output is a generated sample. In general, unsupervised neural networks are used for generative tasks. Variational autoencoders (VAEs) and generative adversarial networks (GANS) are used for these tasks. Transformer networks and graph neural networks (GNNs) are common as well but not exclusive to unsupervised learning tasks. The work in this dissertation focuses on unsupervised methods, specifically VAEs.

1.8.3 Variational Autoencoders

The goal of a VAE is to compress and regenerate an input sample.[86]. The VAE architecture uses an encoder/decoder with a bottleneck in the middle. The encoder , e(), is a set of layers that gradually decrease in size compressing information from the input to a latent layer , z = e(x). The layers of the decoder, g(), increase in size from that of the latent layer to the input size to generate the input sample $\hat{x}$ = g(z) = g(e(x)) [86]. Formally, the encoder finds a set of weights that approximates a function to compress the input X into a latent vector, Z. Given Z, g() then learns a set of weights to approximate a reconstruction function such that $\widehat{X} \approx X$.

The VAE loss attempts to optimize the evidence lower bound (ELBO) which consists of two parts: a reconstruction loss and a regularization loss. The reconstruction loss is as defined above and measures the distance between the output and the input. It may be any of the loss

functions such as mean squared error or cross-entropy depending on the specific needs of the modelling task.

The regularization loss defines the VAE from a normal autoencoder, which uses only reconstruction loss, as it forces the latent distribution to be gaussian in nature. This is done by calculating the loss between the latent distribution and a gaussian prior in terms of the Kullback-Leibler loss (KL). The regularization loss is defined as:

$$\mathrm{KL} = -\frac{1}{2}\sum_{j=1}^{J}(1 - \log((\varrho_j)^2) - (\mu_j)^2 - (\varrho_j)^2)$$

where μ_j is the standard deviation of a gaussian and ϱ_j is the mean of a gaussian. Both are provided by a sampling layer in the network that is only connected to the latent layer and has no input. The full VAE loss is,

$\mathrm{VAE} = (\mathrm{KL}) + \text{regularization loss}$, where KL is the regularization loss [86].

Plain autoencoders (AEs) do not use a regularization loss and therefore do not explicitly constrain the structure of the latent space. This only penalizes deviation from the original input and allows for highly nonlinear encodings of samples in the latent space. Given this, AEs often cluster dissimilar samples nearby and may leave large regions between sample clusters unutilized and devoid of useful latent encodings causing the generator to output noisy or nonsense output when sampling that area of the latent space. Ideally, a latent space should be smooth, arranged with similar samples closer to each other, and continuous, with samples arranged so that there is a coherent interpolation between two points. In VAEs, smoothness and continuity are achieved by constraining the latent space through the additional regularization loss term.

1.8.4 Concepts

As mentioned in the previous section, concepts are high level abstract features that arise from the combination of two or more lower level features such as the input features. They are useful in interpreting a model's performanc. [86-89] By understanding the model at a conceptual level changes may be made in the parameters, hyperparameters or training data to improve model quality. In contrast to feature level explanations, such as feature saliency maps,[86,90] concept analysis allows researchers to gain an understanding of human level concepts and provides a means to manipulate the network. Feature based methods exist outside of the model's inner representation of the data and may be difficult to interpret, especially in an omics setting where human interpretation is often not possible at the nucleotide or gene level without some aids.

Concepts may arise at any layer of a neural network in either supervised or unsupervised settings. They capture underlying structure or patterns in the data that may be used for downstream tasks such as generation or classification. In the image classification tasks, concepts in lower layers (closer to the input) tend to be fine grained such as fine lines and dots. In the later layers, concepts are much larger and may include things such as noses, hands, or other large structures depending on the images being classified. It does not always hold that earlier layers have finer grained concepts but they are generally less complex.

Concepts are detectable in a latent layer of a network. In many cases any layer except the input and output layer may be considered a latent layer. In the case of generative models, there is often a specific layer that is bottleneck, or the most compressed space, from which a generator network samples, which is referred to as the latent layer. In VAEs, this layer is the

sampling layer that generates the encoding z and is regulated by the regularization loss.

All layers of a neural network produce vectors of some size, as such a concept may be described by a concept activation vector (CAV) that is the size and shape of the layer being investigated. The most commonly used method for detecting and quantizing a CAV is to subtract the mean vector of a class of samples from another class. Often the other class is simply a random subset of samples from the training set or the training set as a whole. Other methods focus on probability distributions of samples in the two sets. Every method produces a vector that may be traversed upon which at least one concept is encoded.

1.8.5 Concept Level Model Explanation

Testing with Concept Activation Vectors (TCAV)[86,91] is a technique that measures the importance concepts in a neural network. It evaluates the degree to which a concept is represented in a latent layer and how important the concept is to the output of the model.

This is done by first labeling data with known concepts. The method uses these labels to train a binary classifier on the CAVs. The boundary between the positive and negative classes is interpreted as the normal to the CAV. The CAV is then traversed, and a loss calculated for each step in the traversal. The gradient of the loss is calculated and the integral of the gradient gives the importance of each concept to a given output class.[86,91]

TCAV is capable of evaluating concept importance in a model agnostic way. Due to this it is valuable for assessing different architectures on a given task and for identifying how a model may be improved. It is also capable of finding erroneously entangled concepts allowing the data set to be modified to better disentangle these concepts in subsequent training rounds.

TCAV is a powerful method but has some disadvantages as well. Primarily, TCAV is

dependent upon human labeling. As datasets grow, good labels could be more challenging and expensive to generate. Additionally, TCAV does not give a complete explanation of the model's performance or of learned concepts that impact that performance. This means that erroneously learned concepts may go undiscovered and biases in datasets may be difficult to understand, limiting model improvements.

1.9 Hypothesis and Aims

We hypothesize that cells exist on a methylation landscape that may be exploited to extract biologically significant information about that space using a neural network model. If this is true, then a neural network could be used to map all cell methylation states to a latent space regardless of the type or source of the cells. A pan-methylation model would fundamentally shift NCD research by unifying previously unconnected efforts across multiple diseases into a single space. This would mean that shared risk factors, drug targets and biomarkers could be investigated across diseases more easily. Moreover, complex relationships among CpGs could be more easily discovered by latent space analysis, giving rise to new hypotheses that have to be validated in the wet lab and thus uncover novel drug targets or biomarkers.

Modeling multiple diseases across multiple tissues is a difficult task and requires care in data collection, data engineering, model choice and model analysis. To our knowledge, no effort, as of the publication of this dissertation, has been made to develop such a model. Therefore, we propose a series of steps to accomplish this task and demonstrate, through our experimental plan to accomplish the specific aims set forth, that it is possible.

1.9.1 Aim 1: Develop a Neural Network Model to Map the Landscape of the Cancer Methylome

The Cancer Genome Atlas is a large repository of cancer data that spans multiple types of omics data. It contains over 20,000 samples representing 33 cancer types and normal matched tissue samples. It is the largest NCD repository in the world with over 2.5 petabytes of well curated data that was procured as a concerted effort by the National Cancer Institute and the National Human Genome Research Institute. The size and quality of TCGA data is unmatched and provides strong a basis to begin modeling NCDs. Because the data is sourced from multiple tissues with matched normal samples, a model derived from this data would have a broad representation of not only different cancers but normal tissues as well. With both disease and normal states represented, it should be capable of being retrained for other disease states or new tissues without the risk of overfitting.

To this end, we proposed to develop a neural network based model that is capable of retaining the natural structure of the data while rendering low level feature representations into detectable concepts. While pursuing the goal of developing a pan disease model, it is necessary to retain as much usable information in the latent space as possible. In some training paradigms, such as pure supervised learning, information not relevant to the task may be lost. If this is allowed to happen, it would not serve the purpose of transfer learning well as the new task may require learning disease domain spaces that could have been discarded by the model trained to address a specific task. If information is lost that may be useful to future transfer learning goals, the model will appear brittle for new tasks and may require more data than is available to train.

In this pursuit of developing a pan-cancer or pan-disease model, it important to analyze

the model's performance and determine whether the presence of certain concepts can confound the interpretation of the model. Though neural networks are generally thought of as 'black box' models, interpretability and explainability techniques have progressed far enough that models may be examined in a reliable fashion and their outputs explained with respect to their inputs. Saliency maps are input level importance scores. They indicate the importance of a feature for a given task (class label in the case of supervised learning) and provide a basis to better understand the model performance. Additionally, latent space analysis by visual inspection is a subjective way to interpret the latent representations of samples. On the other hand, concept level importance scores (such as TCAV) give researchers a deeper understanding at how human level concepts influence model behavior in an objective manner. These will be explored in this and later aims.

1.9.2 Aim 2: Utilize the Previously Trained Pan-Cancer Model as a Foundation for a Model of Non-Cancer Noncommunicable Diseases (NCDs)

To test whether a model trained on one type of NCD may be used to improve training of another type of NCD, we propose the use of transfer learning on several NCDs that are known to have methylation signatures. If the new model, derived via transfer learning, is capable of generalizing on this new data better than a model trained from scratch, this provides evidence that there is a shared landscape upon which NCDs lie, which may be exploited for future modeling purposes.

To end this, we planned to compile a new dataset of NCDs for use in model training. This was accomplished by collecting methylation samples from the GEO database and minimizing specific sources of bias such as age, geographic location, gender, and tissue source. We

anticipated a significantly smaller dataset than the TCGA data used to train the model in Aim 1. We then planned to use a transfer learning approach to retrain the model while allowing it to rely on concepts learned previously. We used a two-round transfer learning scheme where the weights of model from Aim 1 were frozen and a new output layer was added to the model with random weights. This new layer was trained first, so that the random weights do not obliterate the learned information in the rest of the model during backpropagation. In the second round, all weights were unfrozen and a small learning rate was used to allow the model to fully adjust to the new task while not diminishing too much of its previously learned information. The model's performance was assessed and concept based approaches to understand how to improve the performance were utilized. Similar methods as used in Aim 1 (section b) were used here with more emphasis on concept level analysis. Because of the smaller dataset, we anticipated incomplete sampling, leaving areas where the new model may not generalize well. We anticipated biases in the new dataset that could not get corrected. We posited that concept level analysis was better suited at illuminating those missing areas.

1.9.3 Aim 3: Use Augmented Data to Fine Tune the Previously Trained Model

The field of NCD modeling has been slow to progress and siloed among specific diseases and their tissue sources. This is symptomatic of a greater issue that is particularly vexing in omics data, the inability to generate high fidelity artificial samples for improved model training. This issue causes omic modelling efforts to rely on the availability of new usable data. Unless the task of data generation from a wet lab and the modeling task in a dry lab are unified, this means the modeling task is fundamentally slowed and beholden to the needs of wet lab experiments, which may not generate the kinds of samples or the volume of samples necessary

for modeling. Using the concept level analysis, we may directly apply the concept vectors to existing samples to produce augmented samples that smooth our datasets sampling over areas that were not present in the original dataset. Although these samples are expected to be of lower quality than real experimental data, they provide an opportunity to quickly iterate over models that can generalize better than those using only the existing data. We expect this to provide important knowledge about omics modeling and elucidate important mechanisms in NCD biology.

CHAPTER 2

CANCERNET: A UNIFIED DEEP LEARNING NETWORK FOR PAN-CANCER DIAGNOSTICS*

2.1 Introduction

Survival rates of cancer patients dramatically improve when diagnosed in early stages as tumors may not have spread yet. However, detection rates in early stages are inconsistent across cancers. As an example, ~63% of breast cancer cases are diagnosed in stage 1 while only ~17% lung cancer cases are diagnosed in the same stage (https://seer.cancer.gov/csr/1975_2017/). This is owing, in part, to the fact that diagnostic development has historically focused on detecting individual cancers. Many cancers are detected only when the symptoms appear, which most often occur in later stages. The development of pan-cancer diagnostics would enable detection of more cancer types, including rare cancers that are not typically the focus of individual biomarker research, thus dramatically improving the prognosis and survival of cancer patients. Such a tool would allow clinicians to diagnose more patients earlier and guide more informed treatment decisions. Additionally, successful application of such a tool to pre-symptomatic patients would necessitate further efforts to locate the tumor to a specific body site with greater resolution. Here, we present a unified cancer diagnostic capable of both, robust cancer diagnosis and tissue of origin detection, for 33 different cancers.

Approximately 60% of genes in humans are found in genomic regions dense with CpG

* This entire chapter is reproduced from Gore, S., & Azad, R. K. (2022). CancerNet: a unified deep learning network for pan-cancer diagnostics, BMC Bioinformatics, 23(1), 1-17. Originally published under CC-BY; authors retain copyright.

dinucleotides called CpG islands which may be methylated.[92] The degree of methylation influences expression of downstream genomic regions. Tissue specific patterns of methylation arise through development and limit the possible changes to the cell state during development or carcinogenesis.[93,94] Methylation has been shown to be significantly altered in many cancers making it promising as a pan-cancer biomarker, and furthermore, as patterns of alteration vary by cancer types or subtypes, methylation is being exploited to distinguish different cancer types or subtypes.[92,95-100] Methylation data have previously been used to successfully develop classifiers for individual cancer types and cancers derived from tissues with common developmental lineages.[101-115]

The high dimensional, real value data obtained from high-throughput methylation arrays, such as those archived in The Cancer Genome Atlas (TCGA; https://www.cancer.gov/tcga), are well suited for use with machine learning classifiers, including neural network. Detection of tissues of origin of cancers can be cast as a supervised task within the realm of machine learning. Supervised methods may artificially separate samples based only on pre-defined classes; however, unsupervised methods may generate a latent space which can be leveraged for many downstream tasks while retaining the underlying structure of the data. Among neural network architectures, unsupervised methods have seen growing use in biological data analysis, particularly for dimensionality reduction with high degrees of success.[116-120]

Briefly, a class of unsupervised methods attempts to regenerate realistic samples from some low dimensional representation.[86,121] Variational autoencoders (VAEs),[86] an unsupervised method, have been used as a basis for downstream regression or classification in a host of

applications, including methylation or transcriptional data analysis. This has been done by passing the latent mapping of a sample to a classifier such as a support vector machine.[120] However, this does not allow for features in the latent space to be modified to improve the classification task. The unsupervised and supervised tasks may be used to constrain each other, resulting in an unsupervised latent space that retains the natural distribution of the data but is optimized for the classification task.

Here we propose a model where both the generative (unsupervised) and the classification (supervised) trainings take place at the same time. This hybrid generator/classifier architecture enables learning of discriminative features intrinsic to input data in tandem with producing a robust classifier. Tuned for and trained on cancer tissues of origin and normal/non-cancerous tissues, our proposed neural network, CancerNet, is currently capable of detecting 33 different cancers. CancerNet was assessed on multiple independent datasets including samples that were not used in training, and metastatic and early cancer samples.

2.2 Materials and Methods

2.2.1 Methylation Data

Illumina 450k methylation array data were downloaded from The Cancer Genome Atlas (TCGA) GDC portal for all cancer types. Metastatic and recurrent samples were removed. This resulted in total 13,325 samples. Each sample was labeled by its tissue of origin and TCGA cancer type designation. Rather than creating a distinct class for each normal tissue, all samples that were from non-cancerous tissues were included in the normal class. This was done due to the extremely low numbers of normal samples available for some tissue types. Additional validation sample sets were downloaded from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/).

Details of specific GEO datasets used are provided in Supplementary Table 1.

2.2.2 Data Preparation

We relied on the CpG density clustering approach implemented in CancerLocator to process the methylation data before inputting to CancerNet.[107] CpGs that were not assigned to a CpG island were first removed. The remaining methylation data were scanned for Illumina 450k probes that map to within 100 bp of each other, which were then concatenated. These clusters were then filtered to eliminate those with 3 CpGs or less.[107] The beta values for the resulting clusters were then averaged. This resulted in 24,565 clusters that map to CpG islands. These average beta values were used as input to CancerNet. The dataset was then randomly split into training/test/validation sets with 80% in training set and 10% each in the test and validation sets. We ensured that the training set did not include more than one sample per patient by removing one of any matched pairs present in the dataset and replacing it with a random sample from the same class.

2.2.3 Performance Assessment

Held-out test data from TCGA and GEO datasets were used to assess CancerNet's performance measured in terms of recall, precision, and F-measure. For a specific class (e.g. a cancer tissue of origin or normal), recall defines the fraction of samples belonging to this class that are correctly identified by a classification method. Precision is the fraction of predictions for this class that are correct, and F-measure is the harmonic mean of recall and precision. Unless otherwise noted, the F-measure presented in this work is weighted F-measure due to large class imbalance among tumor classes. Weighted F-measure is the weighted average of F-

measure values with weight proportional to the number of true instances for each class. The F-measure function in the scikit learn python library (https://scikit-learn.org/stable/index.html) was used to calculate this.

2.2.4 Neural Network

The CancerNet program was written in Python using the keras package (version 2.0.8) [122] with a tensorflow (version 1.12.0)[123] backbone. The neural network architecture of CancerNet consists of an encoder, decoder and classifier (Fig. 2.1). The encoder has an input layer of 24,565 nodes, which is fully connected to a dense layer of 1,000 nodes that uses a relu nonlinearity and two dense activation free layers that are passed to a probabilistic layer, also called the latent layer, characteristic of a VAE architecture with 100 nodes. The decoder has a single dense layer of 1,000 nodes that uses a relu activation and is fully connected to an output layer of 24,565 nodes that uses a sigmoid activation. The classifier takes the latent layer as an input to a dense 100 node layer that uses a relu activation and is fully connected to the classifier output layer that has 34 nodes and uses a softmax activation (Fig. 2.1). CancerNet was trained using the Adam optimizer with a learning rate of 0.001. All layers were randomly initialized and then trained until convergence. Early stopping was used to limit training time and prevent overfitting and was limited to 50 epochs without validation accuracy improvement. The final loss of the network was the sum of the VAE loss and the categorical cross-entropy loss, which are applied to the generative output and the classification output respectively.

VAE loss is composed of two terms. The first term quantifies the divergence between the output of the generator and the input to the model using categorical cross-entropy. The second term is used to enforce gaussian distributions in the latent layer by calculating the Kullbeck-

Leibler divergence of the encoders' distribution and a standard normal. The VAE loss beta term can be used to create a disentangled VAE. When beta is greater than one, features are forced to disentangle and become easier to interpret. Beta is set to 1 in CancerNet.

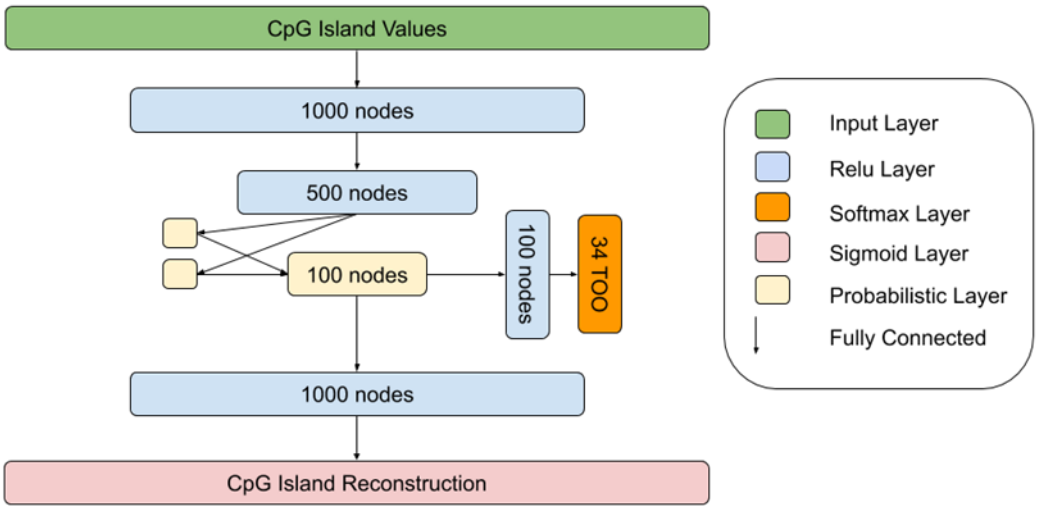

Figure 2.1: The CancerNet architecture

Methylation data are input to the encoder. The encoder is composed of two dense feedforward layers using the Relu activation function. Output of the encoder is passed to the probabilistic layer, which passes its output to the classifier and generator/decoder. The classifier is two dense feedforward layers, the first with the ReLu activation function and the second with the softmax activation function. The decoder is two dense feedforward layers, the first using the Relu activation and the second using the sigmoid activation.

Cross-entropy is applied to the classifier output to estimate a loss based on the difference between the classifier output and the class labels. This is distinct from the cross-entropy for quantifying the VAE loss based on the difference between the generative output and the sample itself. Weights of 0.01 and 1 are applied to the VAE loss and classifier loss, respectively. The generator and classification losses together enforce the latent space representation of samples to preserve information about samples' natural distribution while also creating an easily classifiable distribution of samples. In doing so the latent space acts as a prior in the classifier.

2.2.5 Prevention of Leakage

Leakage is a phenomenon in machine learning where information about the task is inadvertently added to the data on which the task is being performed [124]. This can lead to very brittle models or even completely useless models when used outside the test and training data. Tasks such as normalizing datasets prior to splitting into training/testing/validation sets can introduce information present in the test and validation sets into the trained model, thus artificially inflating the performance of the model in validation and test phases.[124] The beta values of the Illumina 450k array were normalized on a sample by sample basis and bounded in the range [0, 1], preventing information from crossing among samples. The validation set is then used as a sanity check to confirm the model performance on unseen data. We also demonstrate further that the model is robust by using independent datasets retrieved from GEO.

2.3 Results

2.3.1 Model Performance

CancerNet's parameters were learnt from training data obtained from The Cancer Genome Atlas (TCGA) for 33 different cancers and a normal class. The cancers investigated were adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML),

brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), uveal melanoma (UVM).

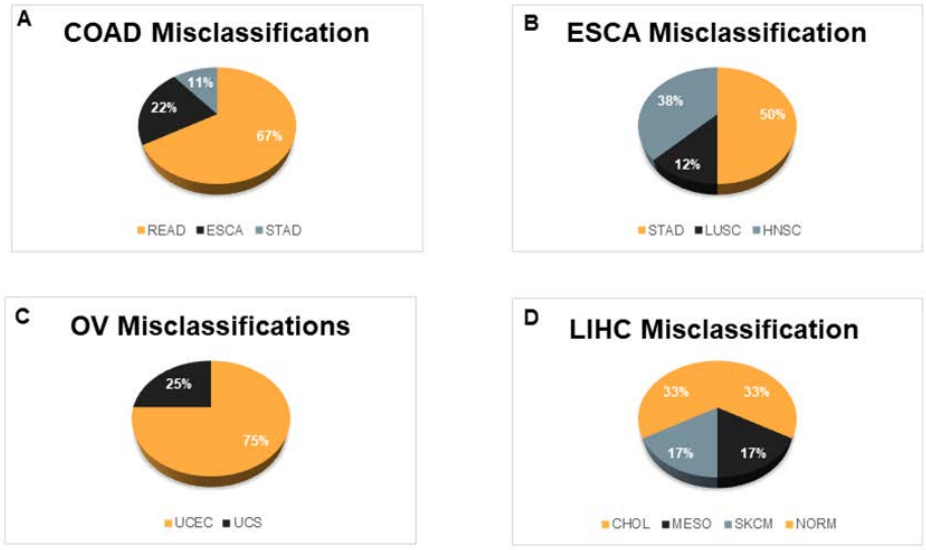

Figure 2.2: Misclassification rates for 4 cancer types to illustrate trends observed in CancerNet.

A.COAD misclassifies primarily to READ with fewer misclassifications in ESCA and STAD. B. ESCA misclassifies to HNSC, LUSC and STAD. Lung misclassifications occur often among some sample types. C. OV samples misclassify as the two uterine cancer types present in CancerNet; UCEC and UCS. D. LIHC misclassifies as CHOL, MESO, SKCM and NORM. (refer to Abbreviations for cancer types indicatedr different cancers indicated on the X-axis; normal is abbreviated NORM)).

The overall performance of CancerNet, as quantified through F-measure (see Methods), is ~99.6% (Supplementary Table 2). Many of the misclassifications occurred among cancers from the same or similar organs and tissue classes that share developmental lineages (Fig. 2.2 and

Supp. Figs. 1-28). Where this did not hold true, we found a pattern of misclassification among adenocarcinomas and squamous carcinomas (Fig. 2.3, Supplementary Figs. 1 -28). Examination of the latent space (Fig. 2.4), along with misclassification rates, shows that misclassifications occurred among closely neighboring classes (Fig. 2.2) or for individual samples of a class that are singletons and very far from the rest of the class (Fig. 2.4 and Supplementary Figs. 29 - 63).

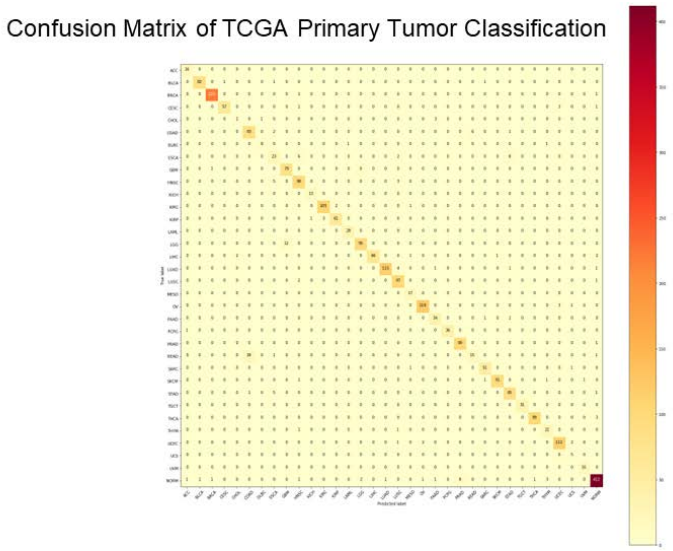

Figure 2.3: Confusion matrix of TCGA primary tumor classification

Primary tumors across 33 TCGA cancer types were classified. The correct class is shown by the Y-axis and the predicted class is shown by the X-axis (refer to Abbreviations for different cancers indicated on the X-axis; normal is abbreviated NORM).

2.3.2 Latent Space Evaluation

We confirmed that the latent space of CancerNet maintains the natural distribution of the sample data by comparing it to the latent space generated through a multi-omic clustering algorithm in a flagship paper from TCGA consortium.[108] The latent space of CancerNet shows high concordance with the latent space of TCGA data presented in Hoadley et al. study[108] (Fig. 2.4). Similar to this study, we observed clustering of the samples by tissue of origin and position in a specific organ in the CancerNet's latent space (Fig. 2.4). Similar distributions of various

subtypes of cancers were also observed. These observations suggest that CancerNet's latent space maintains the natural distribution of the sample data. Note that Hoadley et al. used a highly curated set of methylation sites, devoid of any tissue specific promoter sites and chosen based on hypomethylation status, in order to perform unsupervised clustering of methylation data samples to establish that cancer type specific signatures are present in the tumor samples.[108] In contrast, CancerNet obtains similar results but with far less preprocessing of the data and in a manner that facilitates integration with other data types such as those used in the Hoadley et al. study (e.g. mRNA, aneuploidy, miRNA, and RPPA) by way of a latent space that is vector encoded.

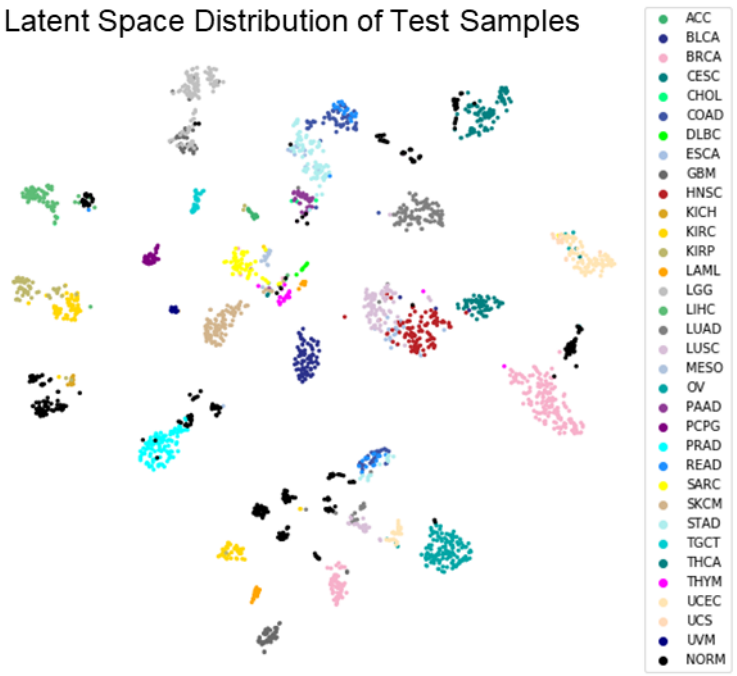

Figure 2.4: Visualization of test samples in the latent space

T-SNE was used to reduce the latent space dimension from 100 to 2. Samples originating from the same tissue form cluster(s) and are close to sample groups of similar tissues. Those tissues that often misclassify among each other, such as UCS/UCES and COAD/READ, appear intermingled in the latent space. For abbreviations, refer to the full abbreviation list. Normal samples are abbreviated NORM and are displayed in gray.

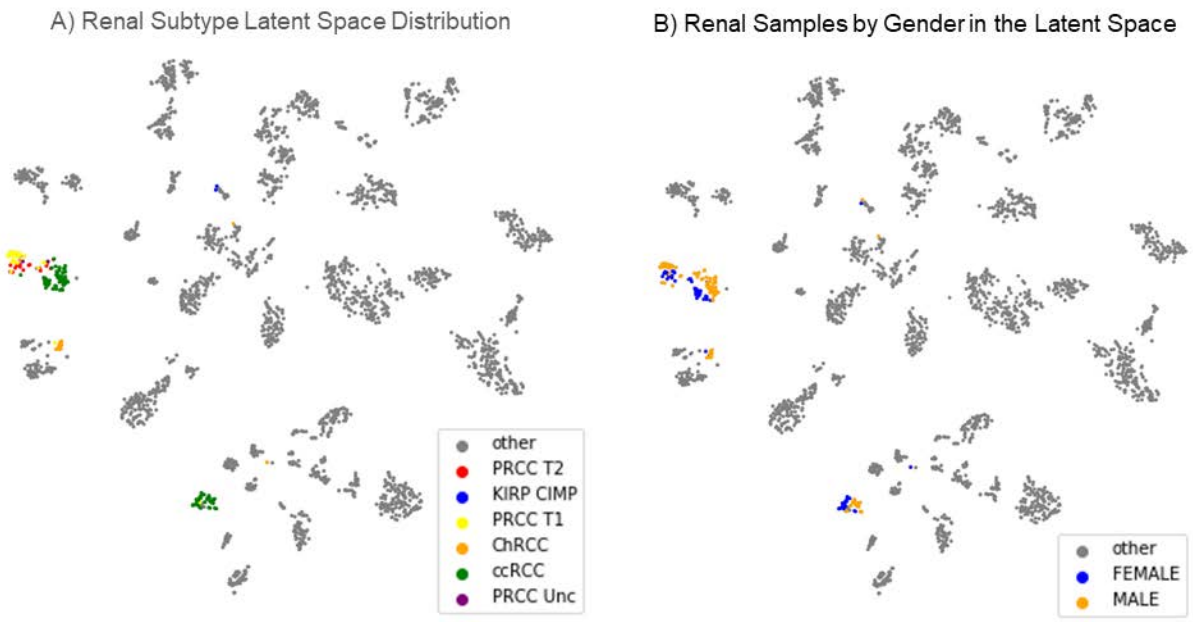

Figure 2.5: Renal subtype latent space distribution

(A) Samples representing different renal subtypes, as determined by the TCGA analysis of renal cancers, are mapped onto the latent space. Clear separation of subtypes PRCC T1 and T2 and ChRCC indicates that the neural network has learned features for discriminating between these renal subtypes. (B) Separation of renal samples in the latent space by gender.

Renal cancer samples (KIRC, KICH, and KIRP) were apportioned into 3 clusters in the latent space (Fig. 2.4), very similar to those described in Hoadley et al. study.[108] The largest cluster consists of two distinct subclusters connected by a streak (Fig. 2.5A); while one subcluster is primarily composed of clear cell renal cell carcinoma (ccRCC) samples, the other is populated with papillary renal clear cell type1 (PRCC T1) samples with type 2 (PRCC T2) connecting these two (Fig. 2.5A). The remaining clusters are primarily composed of ccRCC and chromophobe renal cell carcinoma (chRCC) samples, respectively (Fig. 2.5A); chRCC is a rare subtype of renal cell carcinoma (RCC) found in only 5% of all renal cancer patients with a distinct etiology.[125] The presence of this RCC subtype as a distinct cluster in the latent space is encouraging as it could indicate the presence of detectable and therapeutically important features in the network. Among renal cancers, a distinct separation of samples by gender was

also observed in the latent space (Fig. 2.5B).

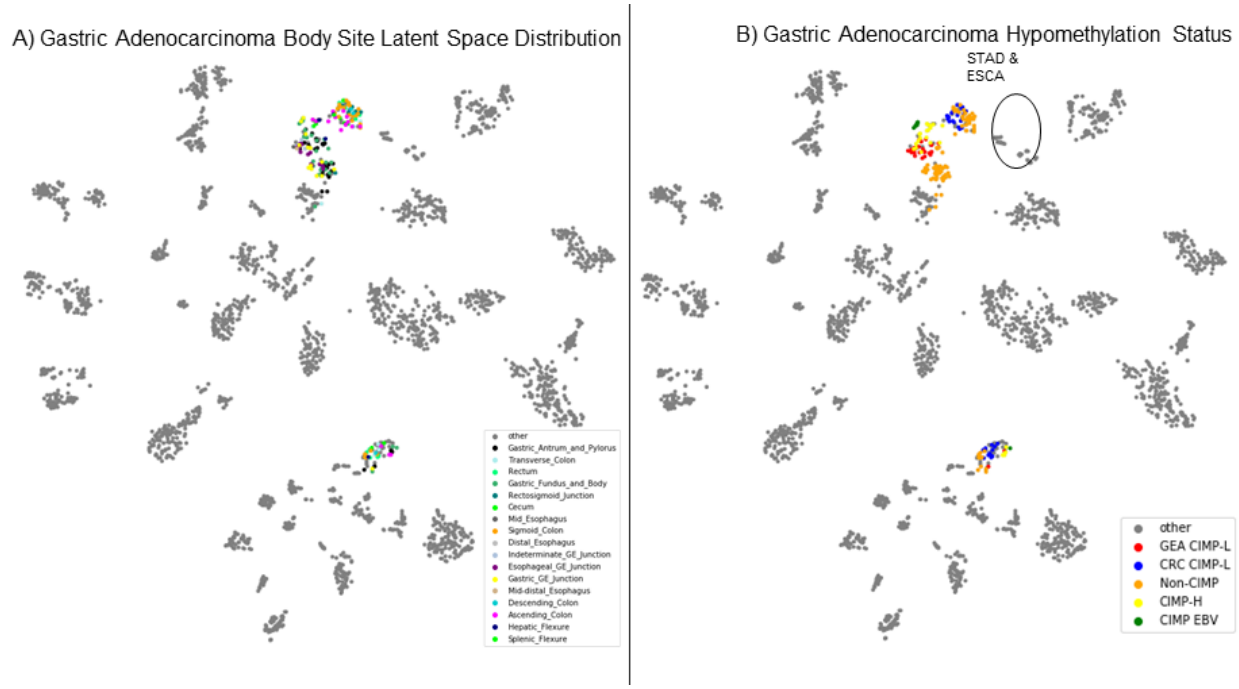

Figure 2.6: Gastric adenocarcinoma latent space distribution.

Samples representing different gastric adenocarcinomas cluster in the latent space by (A) body site of tumor and (B) hypomethylation status. The linear kernel is used to test the separability of each in the full 100 dimensional latent space. The high performance of these models shows that the body sites and methylation statuses are not overlapping in the latent space in the way that the t-SNE plot appears to show in 2 dimensions.

Gastrointestinal adenocarcinoma samples also arrange in a similar way as in the latent space of Hoadley et al. study.[108] Esophageal samples are split among the larger gastrointestinal cluster and a cluster of HNSC in the latent space (Fig. 2.4). Gastrointestinal adenocarcinomas show strong organ site signatures (Fig. 2.6A) and are best explained by hypomethylation status (Fig. 2.6B). Groupings correspond to CpG island methylator phenotype (CIMP) status as described by Ang et al.[126] Non-CIMP separates from CIMP-high (CIMP-H) and CIMP-low (CIMP-L) (Fig. 2.6B). Stomach adenocarcinoma (STAD) and esophageal carcinoma (ESCA) group together with CIMP-H and gastroesophogeal (GEA) CIMP-L status. Epstein-Barr virus (EBV) positive samples form their own cluster. Similarly, molecular subtypes follow the same pattern as in

Hoadley et al. study[108] (Fig. 2.6B). In Figure 2.6 the latent space has been reduced from 100 dimensions to 2 for visualization purposes. To verify that distinct clusters did form by body sit and hypomethylation status in gastrointestinal adenocarcinomas we trained a linear SVM on these classes. These models achieve high (greater that 10 fold cross validation accuracy scores demonstrating the separability of these classes in the latent space.

Squamous cell carcinoma samples (CESC, ESCA, HNSC, and LUSC) segregate by human papillomavirus (HPV) status in the latent space (Fig. 2.7A), which is in concordance with Campbell et al. study.[127] However, CancerNet did not show sensitivity to smoking status (Fig. 2.7B).

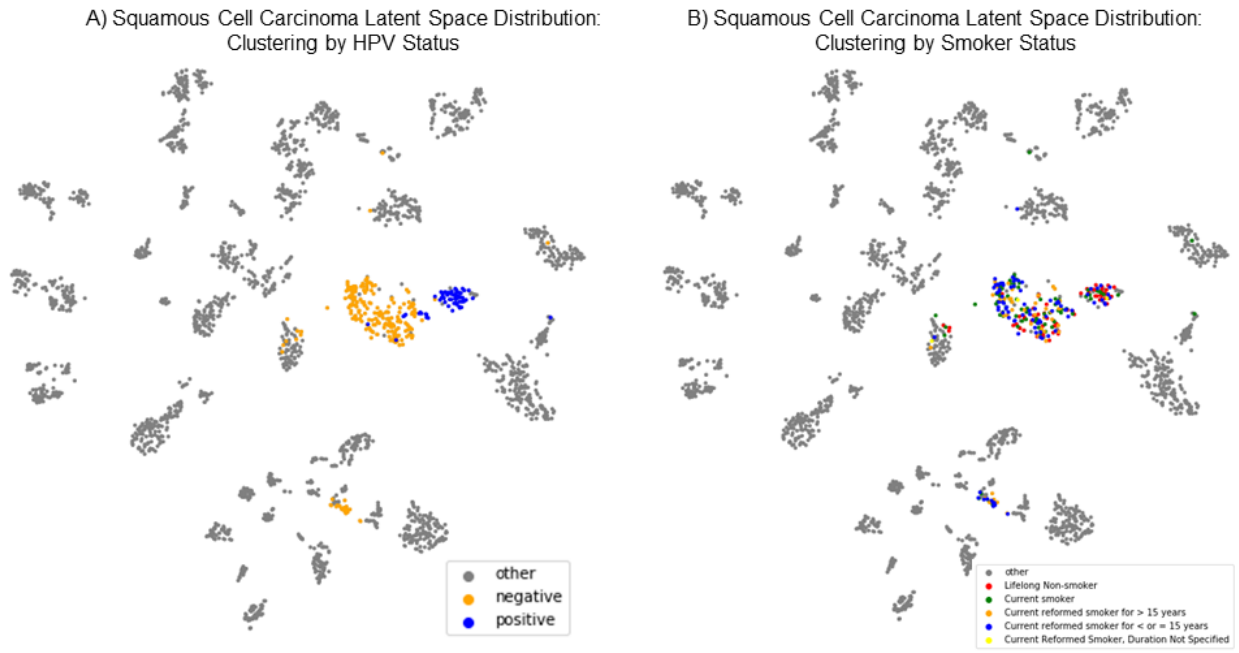

Figure 2.7: Squamous cell carcinoma latent space distribution.

Squamous cell carcinoma samples tend to cluster in the latent space by their tissues of origin and by A) HPV status but not by B) smoker status.

2.3.3 Assessment on Metastatic Cancers, Precancerous Lesions, and Age-related Methylation Drift

When trained on a narrow range of data neural networks may catastrophically fail on

unseen conditions; these models are said to be brittle. To assess how brittle CancerNet is, we evaluated across 3 “untrained” conditions: metastatic tumors, precancerous lesions, and age stratified data. Our results demonstrate that CancerNet performs well across all stages of cancer and is robust to age related epigenetic drift.

2.3.4 Metastasis and Precancerous Lesions

Metastatic cancer is the cause of death in 66% of solid tumor cases.[128] Identification of a second cancer occurrence as a metastatic or second primary tumor is important to inform treatment. In 3-5% of all cases, cancers of unknown primaries (CUPs) are also found;[129] these tumors arise as the metastasis of previously undiscovered primary tumors and are the fourth most deadly cancer.[129,130] Detecting the tissues of origin in both of these scenarios can assist in critical treatment decisions. We demonstrate that CancerNet is capable of robust and highly accurate metastatic tissue of origin classification and this performance is maintained in early cancer samples as well.

To assess CancerNet’s performance on metastatic cancer datasets, we first predicted tissue of origin for metastatic samples available in TCGA for BRCA, CESC, COAD, HNSC, PAAD, PCPG, PRAD, SARC, SKCM, THCA. The tissues of origin for all these metastatic cancers were predicted with an overall unweighted F-measure of 91%.

TCGA data were processed by the different labs and so it is possible that uninformative variance in noise could be introduced due to small but predictable variance in human error, reagent preparation or some other part of the sample processing pipeline. This is known as batch effect. Batch effect can provide a source of information about sample classes that, if learned, could make the model brittle in real world applications where the same effect is not

present. We used several Gene Expression Omnibus (GEO) datasets to assess whether this was the case and further validate the model on non-TCGA derived data. These datasets also gave us the opportunity to test CancerNet's performance on cancer stages that were not represented in TCGA, such as precancerous lesions. These, along with primary, metastatic, and recurrent samples, made possible the assessment of CancerNet's performance across all stages of cancer for several tissue and cancer types.

The first dataset (GEO accession: GSE58999) contained paired metastatic and primary tumors in breast cancer patients. CancerNet achieved an unweighted F-measure of 99% on this dataset. The second dataset (GEO accession: GSE113019) contained triplets of liver samples from each patient, namely, non-tumorous, primary tumor and recurrent samples, respectively. CancerNet achieved an unweighted F-measure of 100% for all primary tumors, 100% for metastatic samples, and 85% for the normal samples. The third dataset (GEO accession: GSE38240) contained 4 normal samples and 8 PRAD metastatic samples; CancerNet attained unweighted F-measure of 88% and 93% for these classes, respectively.

The final dataset (GEO accession: GSE67116) consisted of 96 uterine samples that were stratified across cancer stages with precancerous endometrial hyperplasia, primary tumor and metastasis represented in addition to two cell lines. Samples were harvested from various tissue sites within the uterus. Because endometrial hyperplasia increases a patient's risk of developing uterine cancer by 30%,[131] we used these hyperplasia samples as putative cancer samples. We then labeled them as uterine cancer and checked CancerNets output. CancerNet achieved an unweighted F-measure of 85% on this dataset. On all other sites CancerNet achieved an unweighted F-measure of 92%. CancerNet produced an unweighted F-measure of 66% on

hyperplasia samples, predicting them as uterine cancer in most cases. This indicates that CancerNet may be capable of cancer detection even when just precancerous lesions are present. However, cancer progression for the hyperplasia patients was not documented in the database and we cannot say for certain if this was a correct cancer prediction or not for these precancerous legions.

To further assess the predictive capability of CancerNet on precancerous samples, we used a dataset derived from 55 precancerous ductal carcinoma *in situ* samples (GEO accession: GSE66313). Forty of these samples later developed malignant forms of breast cancer. CancerNet identified the "future" cancer samples (40 of 55) with an unweighted F-measure of ~91% and "non-future" cancer samples (15 of 55) with an unweighted F-measure of ~66%, demonstrating that the model is capable of not only detecting cancer and its tissue of origin but has a reasonably high level of predictive capacity for pre-cancers as well without being explicitly trained to do so.

2.3.5 Age-Related Methylation Drift

Age-related CpG methylation drift is the normal global hypomethylation associated with aging[132]. Some cancer etiologies may be associated with age-related methylation drift [132,133]. CancerNet may be classifying based on background age-related methylation drift rather than methylation changes relevant to carcinogenesis. To verify that this was not the case, we used a dataset (GEO accession: GSE113904) with 232 age-stratified normal colon tissue samples. Samples were from individuals of age ranging from 29 to 81 years. CancerNet classified all of these samples correctly as normal regardless of age.

2.4 Discussion

Here we developed and validated an end-to-end unified model for diagnosing multiple cancers. We achieved high performance, 99.6% fmeasure, through all cancer stages with robustness to possible confounding factors such as age. Previously published neural network classifiers performed approximately as well,[115] ~96% f-measure, as CancerNet but lack a latent space that can encode complex features already present in the data that lend robustness to our model and allow for its possible extension outside the initial use. In addition the model presented by Zeng (24) ingests fewer features and predicts on fewer cancer types. While we lack the tools to fully characterize the trained latent space it can serve as a foundation for future research to develop explainability methods and potentially for discovery of new complex combinations of features that may be important for cancer etiology.

The overall performance of CancerNet on metastatic samples exceeds that of pathologists; the correct tissue was the first choice 49% of the time by pathologists,[134] in contrast, correct tissue was the first choice 91% of the time by CancerNet when evaluated on TCGA metastatic cancer samples. CancerNet also substantially outperformed other models that perform cancer tissue of origin classification based on DNA methylation[106,107] (for all 3 cancer types and a normal type investigated by CancerLocator[106,107] and 12 of 14 cancer types investigated by a model based on random forests,[106,107] Supplementary Table 3). Strong results in both the metastatic and normal categories demonstrate that the model has learned reliable cancer signatures and is capable of tissue of origin detection in cancers that have undergone metastasis. Precancerous lesion classification does not fall neatly under the classification task for which CancerNet was trained. Due to the transitional nature of precancerous lesions, they could

be classified as normal tissue, or predicted as cancerous, which they may become. The performance of CancerNet on precancerous samples is promising and is likely the result of the latent space prior for the classification task. If more precancerous samples for which the progression is known are made available, it may then be possible to add a predictive task to the model and train the model for that specific task. Together, the performance across the cancer spectrum is consistent and demonstrates the robustness of CancerNet.

Where efforts have been made to focus on cancer tissue-of-origin detection, some studies, surprisingly, have done so without determining whether a sample is cancerous or not. Without non-cancerous classes incorporated within a model framework, the model may actually learn tissue specific signatures due to the retention of cell specific methylation signatures even in carcinogenesis. This approach may thus lead to a model learning normal tissue signatures rather than cancer signatures. Therefore, it is pertinent to include normal samples to allow the model to learn to discriminate between normal tissue specific signatures and tumor tissue specific signatures. We, therefore, included normal samples in CancerNet training and classification and ensured that CancerNet's performance is not an artifact of tissue specific signatures. Assessment on different datasets demonstrates that CancerNet is able to robustly diagnose cancer and detect the cancer tissue of origin as well.

Robust tissue level classification is a huge step forward for early cancer detection. Indeed, many cancers have no early diagnostic whatsoever. The clinical use of such a model will benefit from inclusion of information about tumor evolution and tumor subtypes. Such information would aid in treatment decisions and prognosis determination. It is our belief that clinical diagnostic is not the only significant use of such a model. Research in cancer biology may

be aided by investigating the learned features in the model's latent space. Such features may illuminate complex interactions between multiple mutations and methylation dysregulation in a given cellular context. This could provide valuable information about new drug targets. The value of this information coming from a unified model cannot be understated as it provides the opportunity to find potential targets present in multiple cancers and subdivide tumors in feature space rather than in anatomical space allowing discernment of yet unknown aspects of the tumor microenvironment and its effects on oncogenic pathways by way of epigenetics.

Detecting cancer in asymptomatic patients or screening population for cancers requires minimally invasive procedures. Current methods of screening body fluids for biomarkers have been proposed for use with circulating cell-free DNA, cfDNA.[41,135-137] Several studies have shown that methylation persists on the fragments of circulating tumor DNA (ctDNA) and is stable enough to provide cancer diagnosis and tissue of origin classification.[3,138-141] Several key steps must be taken to adapt CancerNet for use with ctDNA. Primarily the number of CpG islands present in a sample at different stages must be assessed. If the model relies on far more features than can feasibly be found in a typical sample, then the model must be adapted to that reality. Additionally, circulating cfDNA may come from multiple sources. Presumably the majority of DNA fragments could come from cells such as macrophages or other normal tissues with good access to the blood that are turned over at a fair rate. Filtering samples to identify the ctDNA fragments of interest is a necessary preprocessing step. We expect technological advances in cfDNA processing will make possible non-invasive, robust early diagnosis of cancers and tissue of origin determination using emerging tools from the field of artificial intelligence such as CancerNet.

2.5 Abbreviations

- ACC - Adrenocortical carcinoma
- BLCA - Bladder urothelial carcinoma
- BRCA - Breast invasive carcinoma
- CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma
- CHOL – Cholangiocarcinoma
- CIMP - CpG island methylator phenotype
- COAD - Colon adenocarcinoma
- DLBC - Lymphoid neoplasm diffuse large B-cell lymphoma
- ESCA - Esophageal carcinoma
- GBM - Glioblastoma multiforme
- HNSC - Head and neck squamous cell carcinoma
- HPV - Human papillomavirus
- KICH - Kidney chromophobe
- KIRC - Kidney renal clear cell carcinoma
- KIRP - Kidney renal papillary cell carcinoma
- LAML - Acute myeloid leukemia
- LGG - Brain lower grade glioma
- LIHC - Liver hepatocellular carcinoma
- LUAD - Lung adenocarcinoma
- LUSC - Lung squamous cell carcinoma
- MESO – Mesothelioma
- NORM - Normal (non-cancer)
- OV - Ovarian serous cystadenocarcinoma

- PAAD - Pancreatic adenocarcinoma
- PCPG - Pheochromocytoma and paraganglioma
- PRAD - Prostate adenocarcinoma
- READ - Rectum adenocarcinoma
- SARC – Sarcoma
- SKCM - Skin cutaneous melanoma
- STAD - Stomach adenocarcinoma
- TGCT - Testicular germ cell tumors
- THCA - Thyroid carcinoma
- THYM – Thymoma
- UCEC - Uterine corpus endometrial carcinoma
- UCS - Uterine carcinosarcoma
- UVM - Uveal melanoma

CHAPTER 3

DISEASENET: A TRANSFER LEARNING APPROACH TO NCD NONCOMMUNICABLE DISEASE CLASSIFICATION MODEL BUILDING

3.1 Introduction

Noncommunicable diseases (NCDs) are responsible for approximately 7 in 10 deaths worldwide and the total number of deaths due to NCDs in 2021 exceeded all deaths attributable to communicable diseases combined. Even when the NCDs are not fatal or terminal they contribute to a significant loss in quality of life for individuals affected by them[142]. Among these, asthma, arthritis and schizophrenia (SCZ) can have devastating effects on the quality of life, even leading to premature death. Asthma is the most common NCD in children worldwide and is the cause of an estimated 455,000 deaths every year[143]. Ongoing efforts to understand the molecular underpinnings of these diseases has relied heavily on genetic data. Epigenetics has more recently been implicated in the etiology and identification of NCDs[27,31,36,38-40,61,144-146]. The use of next generation omics based methods in these efforts has produced rich datasets that may be leveraged to develop powerful neural network based models of these NCDs. While growing, the volume of data available for many NCDs is still too low to train neural networks from the available data alone. However, large amount of omics and other data are available for some NCD families, such as cancer, which may allow the use of transfer learning to generate neural network models of NCDs that lack enough data to build such models. One such type of data of interest for disease classification is methylation data, procured from DNA methylation sites in the human genome. However, it is yet to be shown that model trained on methylation data from one class of NCD, such as cancer, can be used as the basis for models of other NCDs.

Machine learning models have been applied to many types of omics data to address a variety of biological questions [59,60,147-151]. Cancer has been a focal point for many researchers and has been powered by some of the largest and well-curated omics datasets available. Non-cancer NCDs, while actively and passionately researched, generally make up a much smaller share of the available disease omics data as of this publication. Due to this, model based NCD research is generally limited to a small set of features making it difficult to find interactions outside of those features. Additionally, models of non-cancer NCDs tend to be limited to the specific disease, or disease family, being researched rather than incorporation of multiple diseases within an integrated model. Ideally, emerging models should incorporate larger numbers of disease types and rely on larger feature sets.

We posit that NCDs exist on a landscape. If it is so, then a model trained on one NCD, or NCD family, could be extended or retrained on another due to the transferability of information learned about one disease to another. Models such as this would allow researchers to find common risk factors or understand complex risk factors, which could aid in the discovery of biomarkers or development of novel therapies.

Previously, we had trained a model, CancerNet, on DNA methylation data [148]. Here, we use transfer learning to train the CancerNet model to identify 3 NCDs; Asthma, arthritis and SCZ. To our knowledge, no other models have been produced that incorporate multiple NCDs or non-cancer NCDs with cancer samples.

3.1.1 Noncommunicable Diseases and Methylation

GWAS studies have revealed genetic loci associated with Asthma [35-37], arthritis [27-34], and SCZ [39,152-161] but these loci are responsible for a fraction of the risk[61]. Epigenetic studies

involving gene expression, histone modifications, and methylation have revealed evidence that implicates epigenetic risk factors, which may work in concert with genetic risk factors[34,61,150,162].

Each of the three diseases manifests in or affects different tissues. This could make modeling even more complicated as the model may learn tissue specific signatures rather than disease specific ones. We selected methylation data from peripheral blood samples as these were available for all three diseases we chose for this study; further previous studies have reported the prevalence of epigenetic signatures of these diseases in peripheral blood samples [33,36,38,154,160,161,163]. Additionally, early risk factors for SCZ were found to be prevalent in methylation of peripheral blood cells[155,161].

Other than the large volume of data available for cancer, already known links between cancer and other NCDs, such as SCZ, provide evidence that there may be overlapping risk factors. This indicates that cancer may exist on the same (epi)genetic landscape as other non-cancer NCDs. It has been reported that methylation aberrations along with tumor suppressor regulatory changes appear to directly link SCZ with cancer rate[164]. The glucocorticoid receptor gene NR3C1 is also implicated in multiple neurological NCDs including SCZ[146] and is a predictor of poor prognosis in ER- breast cancer[165]. Indeed, longitudinal studies with second generation antipsychotics show that their mechanism is strongly linked to renormalizing methylation changes associated with SCZ [159]. Links such as these serve to illustrate why the development of a pan NCD model through transfer learning is biologically feasible and likely to generate a clinically useful model as well as a rich, biologically meaningful, latent space from the data.

3.1.2 Noncommunicable Disease Detection with Machine Learning

Research that exploits differential NCD methylation pattern within a neural network

framework often focuses on one disease or family of disease[29,31,32,34,59,60,146,147,149,166]. The datasets used in each model are small, requiring heavy use of feature selection. This unnecessarily limits the scope of the model and limits the information learned by the model. Transfer learning that has been underutilized as a modeling tool in this field provides significant advantages in that the disease of interest may be understood in the context of other NCDs and within a richer information space.

Transfer learning is a machine learning technique that utilizes information learned by a model trained for one task on another task the model has not been explicitly trained for[167]. Generally, the two tasks are expected to be in a similar domain and share low level features[167-170]. Because the two tasks are related, the pretrained model is expected to have parameters much closer to those needed to perform well in the new task. This means the search space is greatly reduced and the new task may be learned with fewer examples[171].

Previously, a model trained on expression data, MultiPLIER, was successfully transferred to model other rare NCDs in the expression space[118]. This was done using an unsupervised PLIER network[151] and resulted in a model with a rich latent space. Our method differs mostly due to the use of methylation as our input data. Methylation is easily detected in the blood due to the longer half life of DNA in circulating blood as compared to freely circulating RNA. Additionally, RNA is best assessed from a whole cell environment making it limited in its use as a diagnostic or predictive dataset in these cases. The methylome, however, is not as well characterized and has less prior information to rely on, making it a poor candidate for PLIER model training which requires prior information as input to the model. For the reasons mentioned above, we here sought to exploit DNA methylation data for NCD classification within

a transfer learning framework. To our knowledge this is the first attempt to utilize transfer learning to train a multi non-cancer NCD model.

3.2 Materials and Methods

3.2.1 Data and Preprocessing

We obtained methylation datasets of NCDs (accession numbers: GSE36054, GSE41169, GSE56553, GSE69270, GSE71841, GSE89251, GSE99863, GSE111942, GSE121192, GSE152027, GSE174422) from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). We used the MethylSuite package to download and prepare raw methylation data for each dataset. Each sample was put through our processing pipeline to generate 24,565 CpG islands. These islands were determined using the method outlined in the CancerNet paper [148] and are as follows. When two CpGs were found within 100 base pairs of each other, they were grouped in a cluster. If a CpG already in a cluster is within 100 bp of another CpG, the new CpG is added to that cluster, i.e. the two clusters are merged. The average beta-value representing the methylation intensity was computed for each cluster and then used as the input.

3.2.2 Model

We used here the same model architecture as was used in CancerNet [148]which utilizes a variational autoencoder (VAE) that has a classification task trained at the same time as the generative task (Fig. 3.1). All layers are dense layers unless otherwise noted.

The encoder has 24565 input nodes. The input features are the mean beta values for 24565 CpG islands that were calculated as described in the data preprocessing section. The

encoder is made up of two hidden layers with 1000 and 500 nodes each. This is followed by a sampling layer made up of 100 nodes.

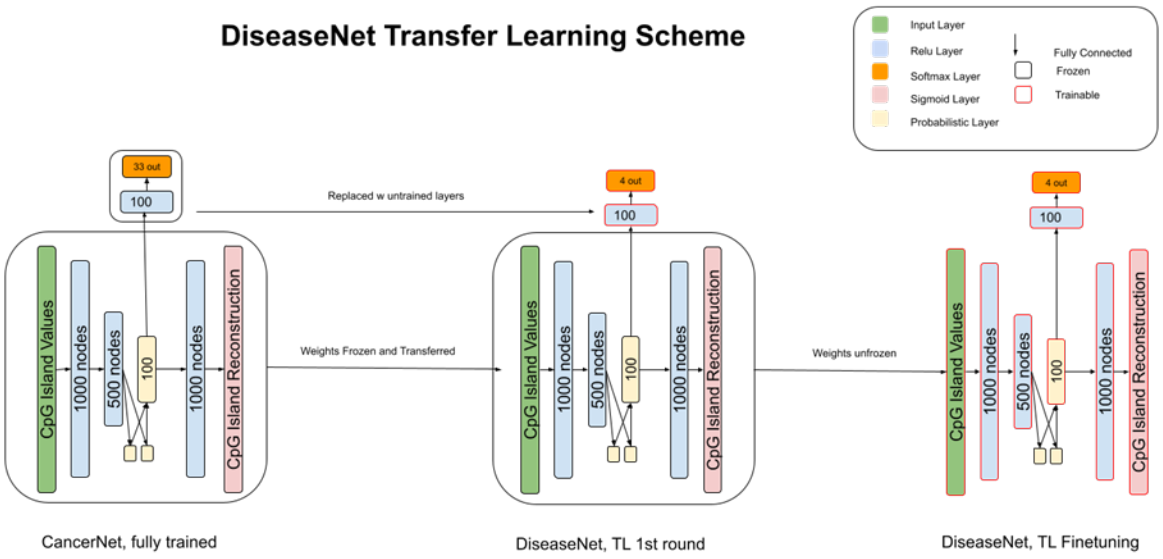

Figure 3.1: DiseaseNet Architecture and Transfer Learning Scheme

From left to right: The process starts with the fully trained CancerNet. The weights are frozen for the encoder and decoder. The classification layers are replaced with randomly initialized weights. The center picture is the first round of training where only the classifier is trained. Weights of the entire model are unfrozen and allowed to train in the last round until convergence.

The sampling layer's output is then used as input to a classifier and a decoder. The classifier has a 100-node layer followed by an output layer that is either 4 or 37 nodes for the transfer learning task or the retraining task respectively. The output layer uses a softmax activation function.

The decoder has a 1000-node layer followed by a 24565-node output layer. The output of the decoder uses a sigmoid activation function.

The loss function is the weighted sum of the classifier and decoder individual losses. The losses used are categorical cross entropy loss and VAE loss for the classifier and decoder, respectively. Weights of 1 and .001 were used for classifier and decoder, respectively.

The model was implemented in Keras [122]. For a list of dependencies, refer to the readme file at https://github.com/Sgore83/DiseaseNet

For the transfer learning task, the model was initialized with the trained weights from CancerNet while the full retrain was initialized with the initial random weights from CancerNet to minimize starting state variability.

3.2.3 Transfer Learning

The complete data set was divided into training, validation and test sets. The training set was 60% of the total data, the test and validation sets were 20% each. CancerNet's architecture and weights were loaded and the final two layers of the classifier's output were replaced by randomly initialized weights for the new classification output size of 4. We then froze the weights of the original layers and trained the last two classifier layers until convergence. Early stopping with a patience of 100 epochs was used to determine convergence. The weight file was only updated when the model improved.

Fine tuning was then done by making the whole model trainable and training the whole model with a learning rate of 1×10^{-6}. This second round of training was allowed to train until convergence with an early stopping patience of 200 epochs. The weight file was only updated when the model improved.

3.2.4 Binary Classifier Models

Binary Models for each class were trained using the same architecture as DiseaseNet except the classifier output layer which only had a single binary node. For each class, samples were labeled using a 1 vs all scheme where any sample not belonging to the target class was

labeled as the negative class.

3.2.5 Performance Metrics

We assessed the models performance using f-measure which is the harmonic mean of precision and recall according to the following formulas

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Fmeasure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

where TP is the count of true positives, TN is the count of true negative and FP is the count of false positives.

3.3 Results

3.3.1 Transfer Learning

Our approach to a robust NCD classification was based on the premise that a model trained on cancer methylation data contains information that could be leveraged to train a new model of non-cancer NCD based on the methylation data. With CancerNet as our initial model, we randomly initialized a new classification layer with 4 output nodes (SCZ, Arthritis, Asthma, and Normal) and trained only the new layer. We then unfroze the rest of the model and trained it with a very low learning rate.

The resulting model, DiseaseNet, produced a classification f-measure of 94% on average and an f-measure of 95%, 98%, 97%, 87% for SCZ, Arthritis, Asthma and Normal classes, respectively (Fig. 3.2).

As a sanity check, we trained binary classifiers for each NCD and the normal class, excluding the step of transfer learning in this process. The performance of these individual models serves as a class specific lower bound that is derived from the information contained in the non-cancer NCD dataset only. If these models performed as well or better than the transfer learning approach, then we could attribute performance largely to the variation between individual NCD classes itself rather than to the transfer learning process. The f-measure values for the single NCD class models—SZD, asthma, arthritis, and normal models—were 6%, 13%, 24%, and 86%, respectively (Fig. 3.2). This indicates that in most classes, the transfer learning approach added significant information not previously contained in the non-cancer NCD data alone.

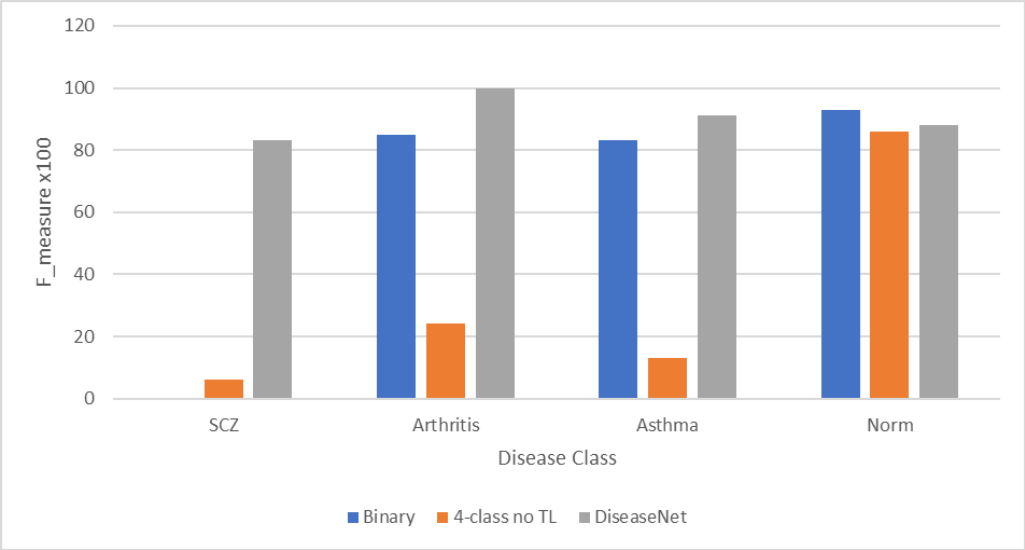

Figure 3.2: Comparison of Classification F measure for Different Training Schemes

The blue, orange and grey bars are the f measures of the binary models, 4class model without transfer learning and fully trained DiseaseNet (with transfer learning), respectively for each class. In almost all cases DiseaseNet did better than the other two but overall DiseaseNet outperformed the other training options.

We then trained a 4 class NCD model without transfer learning. This served as a fine-grained test on the information overlap among classes in the non-cancer NCD dataset. The

values of f-measure for the classes were; SCZ: 0%, Asthma: 83%, Arthritis: 85%, Normal: 93% (Fig. 3.2). Compared to the binary classification without transfer learning, large improvements were observed in all but 1 classes, which demonstrates there is contrastive information allowing some classes to improve when the model is provided with this information. The f-measure value was highest for the normal and further, this was the highest overall accuracy achieved for the normal among the three scenarios considered here (binary without transfer learning, 4-class without transfer learning, and 4-class with transfer learning). Surprisingly, with 4-class without transfer learning, the model produced 0% overall accuracy (f-measure) for SCZ, that is failed to learn at all for this class. The overall performance of the model, is likely a due to learning a suboptimal weight set during model training that sacrifices SCZ performance for higher NORM performance. Here, the transfer learning approach performs approximately as well as the lower bound for the normal class that was established with the binary classifier but improves in all other classes significantly. This indicates that the parameters from a model trained on cancer methylation data may be transferred to create an improved model of other non-cancer NCDs with less support.

3.3.2 Model Characterization

Visual inspection of the latent space shows disease segregation with some overlap between normal and SCZ samples, indicating the likely source of misclassifications by this model between these two classes (Fig. 3.3a). The segregation observed between classes may be attributable to the source type of the different datasets. SCZ and asthma were composed from only a single source type, whole blood and peripheral blood mononucleate cells (PMBCs), respectively. Normal and arthritis were composed of multiple cell types (Supp. Table 1).

Inspecting the latent space based on the sample sources, we found that the source types segregated well and did not intermingle (Fig. 3.3b). This is not surprising as samples with different cellular makeup are expected to have differing methylomic signatures and should occupy different areas of the latent space. The primary question is how important the cell-specific signatures are to the classification output of the model.

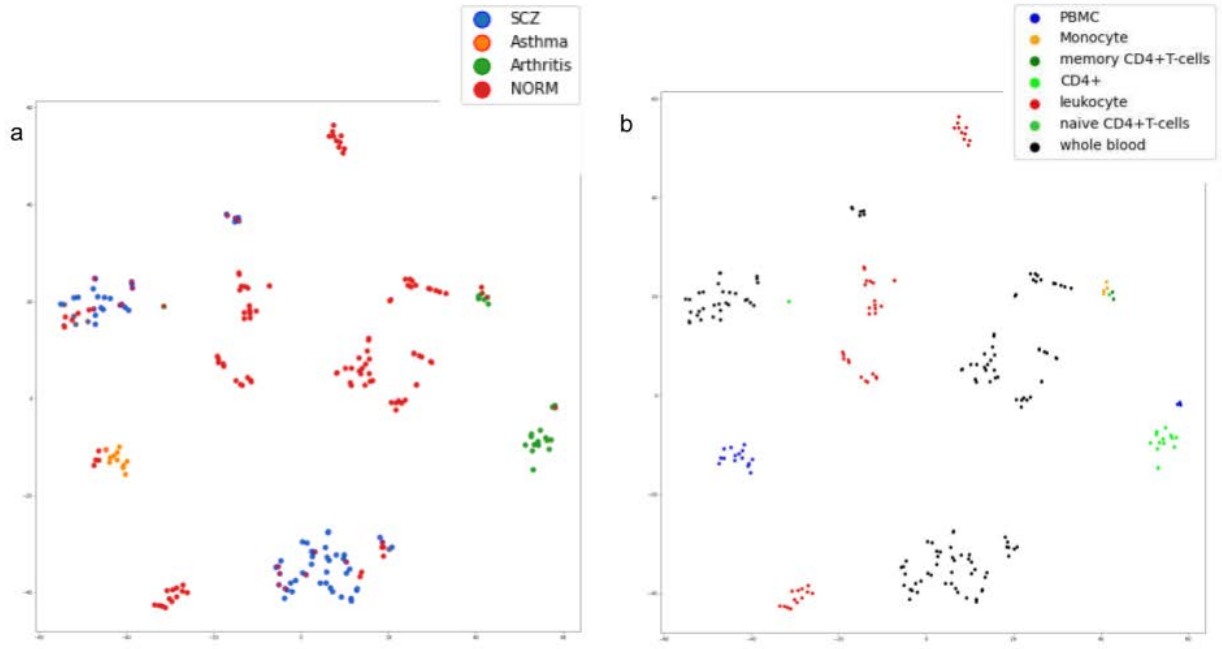

Figure 3.3: Distribution of Samples in DiseaseNet Latent Space

(a) SCA, Asthma, Arthritis and Norm are colored blue, yellow, green and red respectively. They are well separated in the sace with only a few places of overlap by class. (b) Samples are colored by their tissue source. This shows a very clear separation between sources leading us to believe that the tissue source plays an important role in the classification of the diseases in this model

To investigate this, we leveraged the fact that 'cell type' is encoded as a concept vector within DiseaseNet. Concepts are high level features of an input that may or may not be explicitly trained into the model. A class label in a classification task is a trained concept, though not the only one learned, whereas a generative model learns many concepts without the need for a label. They are represented within neural networks as vectors in some latent layer. These

concept vectors can be detected, and an importance score assigned by a method called TCAV. TCAV results indicate how significant a concept, cell source in this case, is to a given class.

TCAV scores for each class are provided in Supp. Tables 2-5 and source distributions are provided in Supp. Table 1. We found that classes have higher TCAV scores (importance) for sources from which they were derived, in general. An exception to this is the peripheral mononuclear blood cells (pmbc) source. It has low importance in every class for which it occurs; arthritis, asthma and normal (Supp. Tables 2-4). In contrast, when classes have samples from multiple sources this does little to decrease the importance of the source concept, as can be observed in the arthritis TCAV scores where monocyte and CD4+ concepts have high TCAV scores, as opposed to the PBMC score which is low in this class (Supp. Table 2). In opposition to this, Asthma is the only class for which the only source is PBMC, however PBMC has a low TCAV score for this class. Oddly the CD4+ concept is highly important to the asthma class despite not being a source for asthma samples. The consistently low TCAV score for PBMC across all classes, along with most classes being partly sourced from PBMC, demonstrates the tendency for confounding conceptual/contextual information to be minimized when stratified across classes.

3.4 Discussion

This study demonstrates the effectiveness of transfer learning in the generation of generalizable NCD models. In particular we highlight the potential that such an approach has in improving modeling of NCDs with small support sets. DiseaseNet, as described with 4 classes here, is a proof of concept that may spur further applications of transfer leaning to disease diagnosis and classification. Increasing tissue sources and disease classes is the primary focus of our ongoing effort to model the NCD landscape further.

We also demonstrated concept-based explanation of our model. It is important to note that high TCAV scores are indicative of greater importance of the concept to a class, there may be many concepts that are important. Individual concepts do not demonstrate completeness of the class concept.

The TCAV results demonstrate that models such as ours may benefit from a diverse tissue source among different diseases. The implication here is that larger and more diverse datasets should be obtained, however, this is a known issue among many NCD studies and the retrieval of such datasets. Instead, we believe data augmentation of biological datasets would greatly improve such models without incurring high costs and efforts/time invested in gathering further data from wet lab experiments. Data augmentation has proven highly effective in model training in the field of computer vision and it stands to reason that it would positively benefit omics modeling as well.

The field of computer vision is easily understood by human vision and augmentations are obvious (such as rotation, cropping and brightness). Omics data, broadly, does not benefit from easy human interpretation and the primary issue here is the difficulty in understanding which augmentations would be most pertinent and how to apply them in an omics setting. As an example, our samples have 24,565 input features. Each is real-valued and many features are dependent upon others or at least correlated. It is difficult to understand how to change those samples so that we may augment a concept and minimally affect the information that is important. We suggest that concept vectors be used in a generative model, such as DiseaseNet, to create an augmented sample set. Such sample sets could be easily produced and would result in better models. The limitation is that the concept being augmented must be

understood well and explained in the context. The explainability methods such as TCAV and SHAP may greatly benefit augmentation efforts.

The VAE portion of models such as DiseaseNet may be used in the training of diffusion models in order to generate high fidelity omics samples for improved model training. Here, omics models may have a significant advantage over computer vision models in several ways. Primarily, the input size is smaller. The input size of computer vision models could be an order of magnitude larger than DiseaseNet. Second, the features may not be consistent in a computer vision model as the subject content and placement of objects in images changes from picture to picture creating a very high degree of variability within image datasets. In DiseaseNet, the same input feature always represents the same CpG island. This increased feature complexity and input feature size means the training sets for image recognition tasks must be vastly larger than those used to train omics models. However, larger omics datasets would still benefit models trained on them, if they are well constructed. Thus, while diffusion models in computer vision may need trillions of examples to generate reliable outputs, omics diffusion models may need far less, due to the lower complexity of the inputs. While this remains to be seen and no diffusion model based on omics data has been produced as of the publication of this report, we remain optimistic that these models will be the future of model building for the biological and medical diagnostic space.

CHAPTER 4

APPLICATION OF LATENT VECTORS FOR DATA AUGMENTATION

4.1 Introduction

Data augmentation is the creation of new samples by either iterative optimization or sampling of a latent variable [172]. Data augmentation is a long practice in model building [173,174]. As more biological models have been created so have data augmentation techniques that deal with the challenges present in biological datasets [175]. Omics datasets have presented difficult combinations of challenges stemming from inconsistencies in data collection or intrinsic issues stemming from data complexity and unobserved hidden dependencies among features.

With the growing use of neural networks, data augmentation has grown in importance as these models thrive of larger datasets. Notably, the computer vision field has utilized data augmentation with great success [176-178]. Efforts to use neural networks on biological data have seen limited use of data augmentation techniques to improve model performance with research focusing on single cell RNA, methylation, and SNP data [179-182]. All of these efforts focus on generating new samples from generative models such as variational autoencoders, generative adversarial networks or deep Boltzmann machines. They rely on the variable and imperfect nature of the generative process of these models to produce new samples that are somewhat different from their authentic counterparts [179,180,183,184]. This effort is excellent at building datasets that normalize sources of confounding variations such as batch effects but do not address the need for missing samples in the population.

A recent study by Treppner et. al focuses on single cell RNA seq data and isolates latent vectors of cell types in the latent space of a variational autoencoder to estimate their

contribution to the models training [185]. Here they employ the concept of 'cell type' to perform an explainability study. Studies using concepts found in generative models trained on omics data are exceedingly rare[185]. The success of Treppner et. al demonstrates the power of concepts in the omics space. Many other concepts in virtually all omics fields are left unexplored.

The goal of this study is to understand how concepts may be used to augment methylation datasets to further the modeling of non-communicable diseases. In doing so neural network models may be improved upon for diseases that are rare, understudied, or difficult to obtain samples for. All of these scenarios lead to small sample sizes with limited statistical power.

We have previously trained DiseaseNet (chapter 3) by transferring learned latent representations of cancer to three other non-communicable diseases; schizophrenia, asthma, and arthritis. Here we use that model's latent space to detect cell type in order to minimize the effect that the source of the sample has on the model's classification output.

4.2 Methods

4.2.1 Concept Activation Vectors

Concept activation vectors (cavs) are directions in the latent space upon which a specific concept varies. To determine these, we first labeled all samples that contained a given concept using a binary label. We then trained a logistic regression classification model on the activations of these samples at the latent layer of DiseaseNet. The cav was defined as the normal of that model's decision boundary.

4.2.2 Latent Data Augmentation

Augmented samples were created by the summation of several vectors derived from the latent space; sample activation vectors and one or more concept vector. To change the cell type for a given sample, the cav of the sample's original cell type was subtracted from the sample's activation vector. The cav of the desired cell type was added to the modified sample vector.

4.2.3 Augmented Data Generation

DiseaseNet was modified so that the latent layer could be fed inputted directly. Then modified sample activations produced in 4.2.2 were used as input and the generative output of DiseaseNet was used as augmented samples.

4.2.4 Augmented Data Validation

We used cosine similarity to determine the similarity of augmented samples and their real counterparts when this made sense. We also used cosine similarity to determine how similar augmented cavs were to their unaugmented counterparts.

4.3 Results

4.3.1 The Impact of TCAV Score on Sample Modification

Using the cavs for sample source generated (Chapter 3), we altered the latent vectors of the training samples. This was done by subtracting the cav of the original sample source and adding the target sample source cav. We chose to attempt this in arthritis first because it was the only non-normal class that came from multiple sources, allowing us to compare modified samples to their authentic counterparts. We first investigated the success of the latent vector

modifications by visually inspecting how the modified samples positions in the latent space had changed. We found that the distance the samples moved was related to the importance of the target sample source to the samples class. The arthritis samples that were modified to contain the monocyte vector (Fig. 4.1), which had high importance in arthritis, moved much further than the modified pbmc samples, which had very low importance to arthritis.

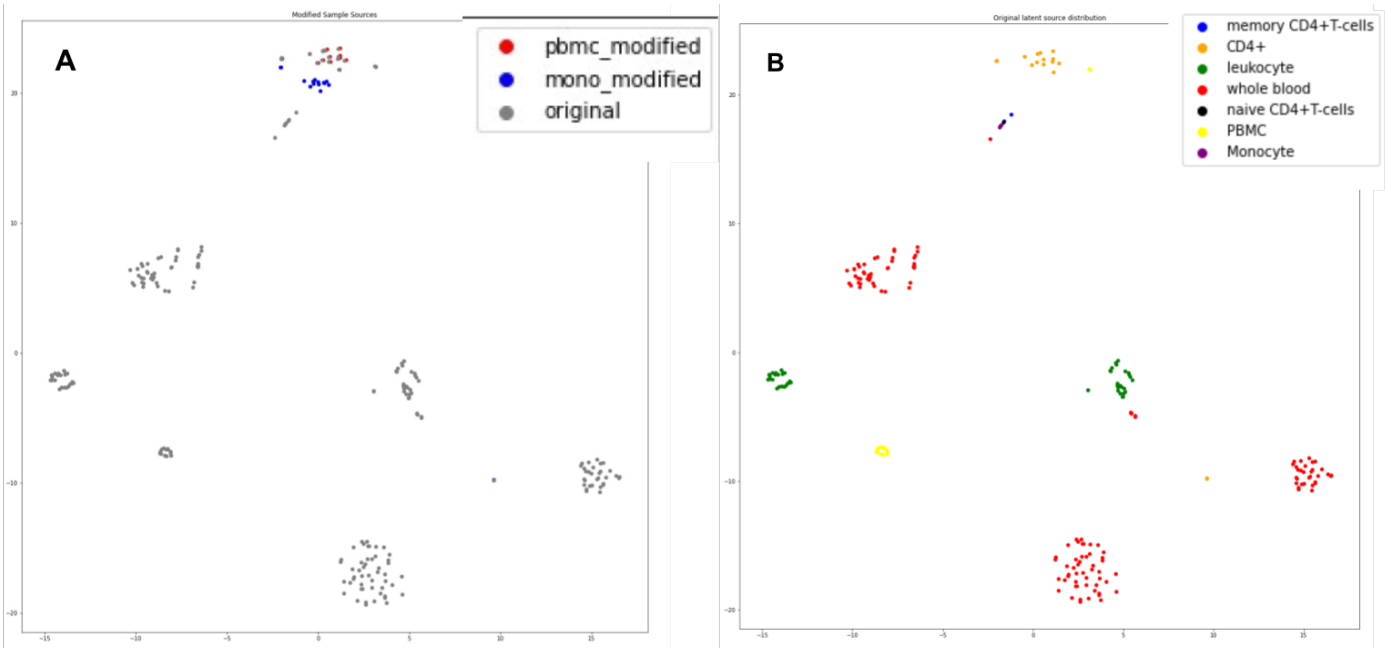

Figure 4.1: Distribution of Modified Samples in Latent Space

Latent vectors of arthritis samples arising from CD4+ cells were modified to have monocyte or PBMC signatures. The modified samples are plotted in A; the original samples are potted in B.

We observe that the modified samples for monocytes moved much farther than the pbmc modified samples did. Their starting point was the CD4 samples, as shown in Figure 4.1B.

4.3.2 Similarity of Modified and Authentic Samples

We measured the cosine similarity between augmented samples and authentic samples. Again, this was only done for arthritis samples as this class is the only one with multiple sample sources. For example, an arthritis sample derived from CD4+ cells that was modified to have the

pbmc vector was compared to an arthritis sample derived from pbmc using cosine similarity. Modified samples were highly similar to authentic samples from that sample source. We observed high cosine similarity, above 80%, regardless of how far an augmented sample moved in the latent space. Modified pmbc samples moved the least and had the highest similarity to their authentic counterparts. This seems to indicate that the cav vector for pbmc is not present in arthritis samples encodings despite pbmc derived arthritis samples being present in the training data. This is further evidence that the TCAV score is a measure of magnitude of a cav given a sample class.

4.3.3 Results of Training with Augmented Samples

To test whether augmenting cell type would affect model performance we trained two different models. The first was initialized with DiseaseNets trained weights. We used a transfer learning approach as described in 3.2.2 and trained on the same training data used to train DiseaseNet with augmented samples added to it. Table 4.2 shows the detailed performance of this model. When compared with the original DiseaseNet performance, Table 4.1, no significant change in performance can be observed. We felt this was not a surprising result due DiseaseNet already being trained on a very similar dataset as the augmented set. We decided to investigate the difference in performance from when CancerNet was used to initialize training. Using CancerNets trained weights to initialize the model we trained the new model using the same dataset as the model initialized with DiseaseNets weights. We used the same transfer learning protocol to train this model. This resulted in a new model, DiseaseNet-Aug, that performed slightly better on Arthritis and Normal classes but had lower performance in SCZ and asthma, Table 4.3, when compared to DiseaseNet, Table 4.1.

Table 4.1: Performance of DiseaseNet

	SCZ	Asthma	Arthritis	NORM
precision	0.72	1	0.84	0.99
Recall	0.98	1	1	0.80
f-measure	0.83	1	0.91	0.89
support	59	11	21	131

Table 4.2: Performance of DiseaseNet initialized transfer learning with augmented data

	SCZ	Asthma	Arthritis	NORM
precision	0.72	1	0.86	0.96
recall	0.94	1	.95	0.82
f-measure	0.82	1	0.90	0.88
support	59	11	21	131

Table 4.3: Performance of DiseaseNet-Aug

	SCZ	Asthma	Arthritis	NORM
precision	0.68	0.73	0.87	1
recall	1	1	1	0.74
f-measure	0.81	0.84	0.93	0.85
support	59	11	21	131

4.4 Discussion

The lack in significant performance improvement is somewhat concerning as it relates to the loss in model performance for arthritis and SCZ in DiseaseNet-Aug. It is worth noting that SCZ and asthma, the classes that saw lower performance after training with augmented data, did not contain any samples from pbmcs or monocytes while arthritis and normal did. The calculation of the concept vectors for cell types was done using the difference between a random sampling of training data and the samples coming from the cell type class of interest. It is possible that due to some cell types being represented in only one disease class the cell

source vector may only be represented through those samples. So while they may not be entangled, the vectors of disease and cell source may be necessarily similar. In this case, the addition of cell source to a different class may introduce information about the class it is most represented in. We believe that the similarity scores of augmented and original samples in the latent space and between generated and original samples is evidence that the effect of such a case may be minimal. However, this is best explored by restricting augmentation to only classes with more than one cell source, normal and arthritis in this case, and training from CancerNet. By exploring the TCAV scores in these cases we may better understand how to improve concept driven latent space augmentation in generative models for omics data.

The finding that TCAV score is related to the amount a sample moves when a given cav is applied to it suggests that the TCAV score is a measure of magnitude for a given cav. If true, this would mean that a generalized, unsupervised version of TCAV could decompose clusters of samples into their complete complement of concept vectors. The completeness of a set of cavs is very important in neural network explainability studies. When a set of concepts is incomplete the model may be interpreted incorrectly leading users or engineers to make misguided decisions based on that information. If two concepts are entangled but one of those concepts is unknown, the model may appear to be less biased or more robust than it actually is. In this scenario, the model's weakness may not be known and could cause catastrophic failure in very rare use cases. In medical applications of neural networks this is especially alarming as it has implications about health outcomes and general medical trust, which are not entirely independent of each other.

In addition to these concerns, concept based understanding of models can help answer questions about the data needed to improve a model. Such information can be used to guide studies on sample selection in a specific and quantifiable manner. By doing so new models may be made, or old models improved, in a more efficient manner. Where samples are difficult or impossible to collect in the real world, augmented samples may be used in their place to improve a model. While this is not the most ideal case, it can be used to incrementally improve a model in a directed way while the real world samples are obtained. This is a cheap and fast way to address undesirable behavior in a model once it is detected.

CHAPTER 5

DISCUSSION AND OUTLOOK

Through the work on modeling cancer within a single model we have shown that modeling NCDs through neural networks is feasible. In transferring the learned information to the problem of modeling other NCDs with less sample support, we were able to improve upon weaknesses in existing NCD modeling efforts. By doing so we have provided a pipeline for modeling diseases through neural networks. This is significant as neural networks can be used to model complex feature interactions, may be easily extended to other problems, and may be integrated with other data types more easily than other models.

Crucially, neural network's latent representation of samples allows for detection of concepts that can be manipulated to improve model function. We proposed a method for data augmentation that takes advantage of the rich latent encodings present in neural networks. Due to incomplete understanding of omics data, efforts to model such data are hampered by biases or errors in the datasets. Correcting such issues may be prohibitively expensive or time consuming. In using the latent encodings in neural networks, powerful generative models can be utilized to produce high fidelity, realistic samples. While these samples may not fully represent missing samples from the population distribution, they can act as a stand in for them and are most useful in making erroneous concepts less important to a model's output.

While we used methylation data in this study, the concepts and techniques employed in this work are not limited to only methylation data and could impact different biological fields with varied data types, particularly where synthetic or augmented samples need to be used. Fields such as metagenomics often use synthetic datasets as a way to benchmark algorithms.

Mixture models are commonplace in metagenomics but are beholden to assumptions about the sample source being modeled. Neural networks can learn hidden or unknown dynamics of these data and could generate better samples to benchmark on.

Our use of VAEs opens the possibility of using multiple encoders to intake multi-omic datasets. Traditionally, integrating different types of omic data has been a non-trivial task and requires heavy data manipulation. Incorporating data types such as SNP, copy number variation, methylation, RNA-seq and/or sequence data requires the integration of three different data types (binary, ordinal, and continuous) with different data dependencies (spatial, graph, longitudinal). To integrate these data into a non-neural network machine learning framework, such as a decision trees or support vector machine, it would need to be transformed into one of the datatypes. To do so requires multiple data transformations and loss of information. Neural networks and especially VAEs can be used to intake each data type independently and then bring the vectorizations of these disparate data into a single latent space.

VAEs are also at the heart of very powerful generative models called latent diffusion models [186] that are capable of generating highly realistic images from texts. The ability to cross domains, such as from text to image, could be used in biology to generate samples of one type from samples of another, such as RNA-seq from methylation samples, exploiting the underlying mechanisms that associate these data. This ability could power greater data augmentation challenges. The use of such models might also be used for data imputation. Biological data are prone to machine errors making it difficult to utilize all the data from a study. When these errors occur, researchers are faced with the decision to either repair the data programmatically

or remove the data. Imputation techniques of various complexity may be employed but are based on methods that could miss long range or complex interactions among features. Diffusion models, while not perfect, can handle such long range, complex interactions and produce highly realistic outputs when trained. Their use could make imputed and repaired data of much higher fidelity. Additionally, diffusion models are capable of being conditioned by multiple inputs, it is possible that phenotypic data could be used to refine imputations. It is also possible that samples could be generated from clinical description alone, which would give researchers a high level of control over synthetic samples with the added advantage of speedy generation.

With the recent publication of the human pangenome [187,188], and in light of the significance of epigenomics in NCD detection in this and other works, the development of a human pan-epigenome may be an important step forward. Here we have mapped multiple diseases and disease subtypes to a latent space. Similar work to detect genetic lineage using the pangenome was recently published [187]. Development of a pan-epigenome may be important to understand disease etiology using similar techniques as heredity studies on the pangenomes [187]. Additionally, the use of such models may form the basis of improved understanding of epigenetic regulation.

In summary, the development and application of neural network based biological tools, such as those accomplished in this dissertation study, will empower rapid and deep understanding of previously difficult and intractable biological and biomedical problems, and we believe, in the context of this study with promising outcomes, the development of such tools will accelerate and usher in new ways of preventing, detecting, and treating diseases.

REFERENCES

1. Noncommunicable diseases fact sheet. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases Web site. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Updated 2022. Accessed 5/24/, 2023.

2. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer—a survey. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*. 2007;1775(1):181-232.

3. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early-and late-stage human malignancies. *Science translational medicine*. 2014;6(224):224ra24.

4. Laird PW. The power and the promise of DNA methylation markers. *Nature Reviews Cancer*. 2003;3(4):253-266.

5. Lima S, Hernandez-Vargasl H, Hercegl Z. Epigenetic signatures in cancer: Implications for the control of cancer. *Curr Opin Mol Ther*. 2010;12(3):316-324.

6. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*. 2002;3(6):415-428.

7. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews genetics*. 2007;8(4):286-298.

8. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nature Reviews Cancer*. 2004;4(2):143-153.

9. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. *Nature reviews genetics*. 2002;3(6):415-428.

10. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. *Nature reviews genetics*. 2006;7(1):21-33.

11. Reszka E, Jabłońska E, Lesicka M, et al. An altered global DNA methylation status in women with depression. *J Psychiatr Res*. 2021;137:283-289.

12. Toperoff G, Aran D, Kark JD, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Hum Mol Genet*. 2012;21(2):371-383.

13. Liebold I, Grützkau A, Göckeritz A, et al. Peripheral blood mononuclear cells are hypomethylated in active rheumatoid arthritis and methylation correlates with disease activity. *Rheumatology*. 2021;60(4):1984-1995.

14. Bell CG, Teschendorff AE, Rakyan VK, Maxwell AP, Beck S, Savage DA. Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC medical genomics*. 2010;3(1):1-11.

15. Chavez-Valencia RA, Chiaroni-Clarke RC, Martino DJ, et al. The DNA methylation landscape of CD4 T cells in oligoarticular juvenile idiopathic arthritis. *J Autoimmun*. 2018;86:29-38.

16. Chavez-Valencia RA, Chiaroni-Clarke RC, Martino DJ, et al. The DNA methylation landscape of CD4 T cells in oligoarticular juvenile idiopathic arthritis. *J Autoimmun*. 2018;86:29-38.

17. Ciechomska M, Roszkowski L, Maslinski W. DNA methylation as a future therapeutic and diagnostic target in rheumatoid arthritis. *Cells*. 2019;8(9):953.

18. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. *Genes Dev*. 2009;23(7):781-783.

19. Suzuki MM, Bird A. DNA methylation landscapes: Provocative insights from epigenomics. *Nature reviews genetics*. 2008;9(6):465-476.

20. Bird A. Perceptions of epigenetics. *Nature*. 2007;447(7143):396.

21. Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33(3):245-254.

22. Muller HM, Widschwendter A, Fiegl H, et al. DNA methylation in serum of breast cancer patients: An independent prognostic marker. *Cancer Res*. 2003;63(22):7641-7645.

23. Skvortsova TE, Rykova EY, Tamkovich SN, et al. Cell-free and cell-bound circulating DNA in breast tumours: DNA quantification and analysis of tumour-related gene methylation. *Br J Cancer*. 2006;94(10):1492-1495.

24. Ponomaryova AA, Rykova EY, Cherdyntseva NV, et al. Potentialities of aberrantly methylated circulating DNA for diagnostics and post-treatment follow-up of lung cancer patients. *Lung Cancer*. 2013;81(3):397-403.

25. Akirav EM, Lebastchi J, Galvan EM, et al. Detection of β cell death in diabetes using differentially methylated circulating DNA. *Proceedings of the National Academy of Sciences*. 2011;108(47):19018-19023.

26. Cheuk IWY, Shin VY, Kwong A. Detection of methylated circulating DNA as noninvasive biomarkers for breast cancer diagnosis. *Journal of breast cancer*. 2017;20(1):12-19.

27. Richardson B, Scheinbart L, Strahler J, Gross L, Hanash S, Johnson M. Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*. 1990;33(11):1665-1673.

28. Nakano K, Boyle DL, Firestein GS. Regulation of DNA methylation in rheumatoid arthritis synoviocytes. *The journal of immunology*. 2013;190(3):1297-1303.

29. Cribbs A, Feldmann M, Oppermann U. Towards an understanding of the role of DNA methylation in rheumatoid arthritis: Therapeutic and diagnostic implications. *Therapeutic advances in musculoskeletal disease*. 2015;7(5):206-219.

30. Zhu H, Wu L, Mo X, et al. Rheumatoid arthritis–associated DNA methylation sites in peripheral blood mononuclear cells. *Ann Rheum Dis*. 2019;78(1):36-42.

31. Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*. 2013;31(2):142-147.

32. Ai R, Hammaker D, Boyle DL, et al. Joint-specific DNA methylation and transcriptome signatures in rheumatoid arthritis identify distinct pathogenic processes. *Nature communications*. 2016;7(1):1-9.

33. Liebold I, Grützkau A, Göckeritz A, et al. Peripheral blood mononuclear cells are hypomethylated in active rheumatoid arthritis and methylation correlates with disease activity. *Rheumatology*. 2021;60(4):1984-1995.

34. de la Rica L, Urquiza JM, Gómez-Cabrero D, et al. Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. *J Autoimmun*. 2013;41:6-16.

35. Hudon Thibeault A, Laprise C. Cell-specific DNA methylation signatures in asthma. *Genes*. 2019;10(11):932.

36. Yang IV, Pedersen BS, Liu A, et al. DNA methylation and childhood asthma in the inner city. *J Allergy Clin Immunol*. 2015;136(1):69-80.

37. Perera F, Tang W, Herbstman J, et al. Relation of DNA methylation of 5′-CpG island of ACSL3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PloS one*. 2009;4(2):e4488.

38. Walton E, Hass J, Liu J, et al. Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophr Bull*. 2016;42(2):406-414.

39. Auta J1, Smith RC, Dong E, et al. DNA-methylation gene network dysregulation in peripheral blood lymphocytes of schizophrenia patients. *Schizophr Res*. 2013;150(1):312-318.

40. Grayson DR, Guidotti A. The dynamics of DNA methylation in schizophrenia and related psychiatric disorders. *Neuropsychopharmacology*. 2013;38(1):138-166.

41. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. *Clin Chem*. 2015;61(1):112-123.

42. Luo H, Zhao Q, Wei W, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Science translational medicine*. 2020;12(524):eaax7533.

43. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. *Science*. 1975;187(4173):226-232.

44. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*. 1975;14(1):9-25.

45. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. DNA motifs associated with aberrant CpG island methylation. *Genomics*. 2006;87(5):572-579.

46. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. Predicting aberrant CpG island methylation. *Proceedings of the National Academy of Sciences*. 2003;100(21):12253-12258.

47. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. 2011;6(6):692-702.

48. Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*. 2009;41(2):178-186.

49. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res*. 2016;44(11):5123-5132.

50. Ehrlich M. DNA hypomethylation in cancer cells. *Epigenomics*. 2009;1(2):239-259.

51. Taghavi N, Biramijamal F, Sotoudeh M, et al. p16 INK4a hypermethylation and p53, p16 and MDM2 protein expression in esophageal squamous cell carcinoma. *BMC Cancer*. 2010;10:1-9.

52. Zöchbauer-Müller S, Fong KM, Virmani AK, Geradts J, Gazdar AF, Minna JD. Aberrant promoter methylation of multiple genes in non-small cell lung cancers. *Cancer Res*. 2001;61(1):249-255.

53. Vaissière T, Hung RJ, Zaridze D, et al. Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer Res*. 2009;69(1):243-252.

54. Kersting M, Friedl C, Kraus A, Behn M, Pankow W, Schuermann M. Differential frequencies of p16INK4a promoter hypermethylation, p53 mutation, and K-ras mutation in exfoliative

material mark the development of lung cancer in symptomatic chronic smokers. *Journal of Clinical Oncology*. 2000;18(18):3221-3229.

55. Bibikova M, Barnes B, Tsan C, et al. High density DNA methylation array with single CpG site resolution. *Genomics*. 2011;98(4):288-295.

56. Sandoval J, Heyn H, Moran S, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*. 2011;6(6):692-702.

57. Jurkowska RZ, Jurkowski TP, Jeltsch A. Structure and function of mammalian DNA methyltransferases. *Chembiochem*. 2011;12(2):206-222.

58. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium methylation 450K technology. *Epigenomics*. 2011;3(6):771-784.

59. Gunasekara CJ, Hannon E, MacKay H, et al. A machine learning case–control classifier for schizophrenia based on DNA methylation in blood. *Translational psychiatry*. 2021;11(1):1-10.

60. Moghadam BT, Etemadikhah M, Rajkowska G, et al. Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods. *J Psychiatr Res*. 2019;114:41-47.

61. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753.

62. Zhang J, He X, Liu Y, Cai Q, Chen H, Qing L. Multi-modal cross-attention network for alzheimer's disease diagnosis with multi-modality data. *Comput Biol Med*. 2023:107050.

63. Wild L, Flanagan JM. Genome-wide hypomethylation in cancer may be a passive consequence of transformation. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*. 2010;1806(1):50-57.

64. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*. 2006;103(5):1412-1417.

65. Amatya VJ, Naumann U, Weller M, Ohgaki H. TP53 promoter methylation in human gliomas. *Acta Neuropathol*. 2005;110:178-184.

66. Woodson K, Mason J, Choi S, et al. Hypomethylation of p53 in peripheral blood DNA is associated with the development of lung cancer. *Cancer Epidemiology Biomarkers & Prevention*. 2001;10(1):69-74.

67. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol*. 2017;35(6):498-507.

68. Wang Y, LêCao K. Managing batch effects in microbiome data. *Briefings in bioinformatics*. 2020;21(6):1954-1970.

69. Noor P. Can we trust AI not to further embed racial bias and prejudice? *BMJ*. 2020;368.

70. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA dermatology*. 2021;157(11):1362-1369.

71. Cho MK. Rising to the challenge of bias in health care AI. *Nat Med*. 2021;27(12):2079-2081.

72. Belenguer L. AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*. 2022;2(4):771-787.

73. Flynn E, Chang A, Altman RB. Large-scale labeling and assessment of sex bias in publicly available expression data. *BMC Bioinformatics*. 2021;22(1):1-23.

74. Williamson CW, Nelson TJ, Thompson CA, et al. Bias reduction through analysis of competing events (BRACE) correction to address cancer treatment selection bias in observational data. *Clinical Cancer Research*. 2022;28(9):1832-1840.

75. Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and class imbalance in oncologic Data—Towards inclusive and transferrable AI in large scale oncology data sets. *Cancers*. 2022;14(12):2897.

76. Jones MB, Berkley C, Bojilova J, Schildhauer M. Managing scientific metadata. *IEEE Internet Comput*. 2001;5(5):59-68.

77. Wang Z, Lachmann A, Ma'ayan A. Mining data and metadata from the gene expression omnibus. *Biophysical reviews*. 2019;11:103-110.

78. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys*. 1943;5:115-133.

79. McClelland JL, Rumelhart DE, Hinton GE. The appeal of parallel distributed processing. *MIT Press, Cambridge MA*. 1986;3:44.

80. Werbos P. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA*. 1974.

81. Anderson JA, Rosenfeld E, Pellionisz A. *Neurocomputing.* Vol 2. MIT press; 1988.

82. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2014.

83. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*. 2012;14(8):2.

84. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*. 2012;14(8):2.

85. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*. 2011;12(7).

86. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 2013.

87. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. 2013.

88. Olah C, Mordvintsev A, Schubert L. Feature visualization. *Distill*. 2017;2(11):e7.

89. Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*. 2016.

90. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*. 2017.

91. Kim B, Gilmer J, Wattenberg M, Viégas F. Tcav: Relative concept importance testing with linear concept activation vectors. . 2018.

92. Yang Z, Jones A, Widschwendter M, Teschendorff AE. An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol*. 2015;16(1):140.

93. Lokk K, Modhukur V, Rajashekar B, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*. 2014;15(4):3248.

94. Salas LA, Wiencke JK, Koestler DC, Zhang Z, Christensen BC, Kelsey KT. Tracing human stem cell lineage during development using DNA methylation. *Genome Res*. 2018;28(9):1285-1295.

95. Sahnane N, Magnoli F, Bernasconi B, et al. Aberrant DNA methylation profiles of inherited and sporadic colorectal cancer. *Clinical epigenetics*. 2015;7(1):131.

96. Ross JP, Rand KN, Molloy PL. Hypomethylation of repeated DNA sequences in cancer. *Epigenomics*. 2010;2(2):245-269. https://doi.org/10.2217/epi.10.2. doi: 10.2217/epi.10.2.

97. Lee S, Wiemels JL. Genome-wide CpG island methylation and intergenic demethylation propensities vary among different tumor sites. *Nucleic Acids Res*. 2016;44(3):1105-1117.

98. Liggett TE, Melnikov A, Yi Q, et al. Distinctive DNA methylation patterns of cell-free plasma DNA in women with malignant ovarian tumors. *Gynecol Oncol*. 2011;120(1):113-120.

99. Stefansson OA, Moran S, Gomez A, et al. A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular oncology*. 2015;9(3):555-568.

100. Bormann F, Rodríguez-Paredes M, Lasitschka F, et al. Cell-of-origin DNA methylation signatures are maintained during colorectal carcinogenesis. *Cell reports*. 2018;23(11):3407-3418.

101. Capper D, Jones DT, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469-474.

102. Mundbjerg K, Chopra S, Alemozaffar M, et al. Identifying aggressive prostate cancer foci using a DNA methylation classifier. *Genome Biol*. 2017;18(1):1-15.

103. Robles AI, Arai E, Mathé EA, et al. An integrated prognostic classifier for stage I lung adenocarcinoma based on mRNA, microRNA, and DNA methylation biomarkers. *Journal of Thoracic Oncology*. 2015;10(7):1037-1048.

104. Brentnall AR, Vasiljević N, Scibior-Bentkowska D, et al. A DNA methylation classifier of cervical precancer based on human papillomavirus and human genes. *International journal of cancer*. 2014;135(6):1425-1432.

105. Melnikov AA, Scholtens DM, Wiley EL, Khan SA, Levenson VV. Array-based multiplex analysis of DNA methylation in breast cancer tissues. *The Journal of Molecular Diagnostics*. 2008;10(1):93-101.

106. Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics*. 2018;34(3):398-406.

107. Kang S, Li Q, Chen Q, et al. CancerLocator: Non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol*. 2017;18(1):1-12.

108. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*. 2018;173(2):291-304. e6.

109. Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PloS one*. 2020;15(5):e0226461.

110. Wei J, Haddad A, Wu K, et al. A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. *Nature communications*. 2015;6(1):1-11.

111. Tian Z, Meng L, Long X, et al. DNA methylation-based classification and identification of bladder cancer prognosis-associated subgroups. *Cancer cell international*. 2020;20(1):1-11.

112. Wu SP, Cooper BT, Bu F, et al. DNA methylation–based classifier for accurate molecular diagnosis of bone sarcomas. *JCO precision oncology*. 2017;1:1-11.

113. Chen W, Zhuang J, Wang PP, et al. DNA methylation-based classification and identification of renal cell carcinoma prognosis-subgroups. *Cancer cell international*. 2019;19(1):1-14.

114. Capper D, Jones DT, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature*. 2018;555(7697):469-474.

115. Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PloS one*. 2020;15(5):e0226461.

116. Way GP, Greene CS. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *BioRxiv*. 2017:174474.

117. Amodio M, Van Dijk D, Srinivasan K, et al. Exploring single-cell data with deep multitasking neural networks. *Nature methods*. 2019:1-7.

118. Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: A transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell systems*. 2019;8(5):380-394. e4.

119. Wang Z, Wang Y. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinformatics*. 2019;20(18):1-7.

120. Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life science alliance*. 2019;2(6).

121. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. . 2014:2672-2680.

122. Chollet Fcao. Keras. . 2015.

123. Mart\'inAbadi, AshishAgarwal, PaulBarham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. . 2015. https://www.tensorflow.org/.

124. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2012;6(4):1-21.

125. Davis CF, Ricketts CJ, Wang M, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer cell*. 2014;26(3):319-330.

126. Ang PW, Loh M, Liem N, et al. Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features. *BMC Cancer*. 2010;10(1):227.

127. Campbell JD, Yau C, Bowlby R, et al. Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell reports*. 2018;23(1):194-212. e6.

128. Dillekås H, Rogers MS, Straume O. Are 90% of deaths from cancer caused by metastases? *Cancer medicine*. 2019;8(12):5574-5576.

129. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: Useful in patient management. *Current treatment options in oncology*. 2013;14(4):634-642.

130. Pavlidis N, Briasoulis E, Hainsworth J, Greco FA. Diagnostic and therapeutic management of cancer of an unknown primary. *Eur J Cancer*. 2003;39(14):1990-2005.

131. Lacey Jr JV, Chia VM. Endometrial hyperplasia and the risk of progression to carcinoma. *Maturitas*. 2009;63(1):39-44.

132. Moran S, Martinez-Cardús A, Boussios S, Esteller M. Precision medicine based on epigenomics: The paradigm of carcinoma of unknown primary. *Nature Reviews Clinical Oncology*. 2017;14(11):682.

133. Ehrlich M. DNA methylation in cancer: Too much, but also too little. *Oncogene*. 2002;21(35):5400-5413.

134. Sheahan K, O'Keane JC, Abramowitz A, et al. Metastatic adenocarcinoma of an unknown primary site: A comparison of the relative contributions of morphology, minimal essential clinical data and CEA immunostaining status. *Am J Clin Pathol*. 1993;99(6):729-735.

135. van der Heijden AG, Mengual L, Ingelmo-Torres M, et al. Urine cell-based DNA methylation classifier for monitoring bladder cancer. *Clinical epigenetics*. 2018;10(1):71.

136. Viet CT, Schmidt BL. Methylation array analysis of preoperative and postoperative saliva DNA in oral cancer patients. *Cancer Epidemiology and Prevention Biomarkers*. 2008;17(12):3603-3611.

137. Sun K, Jiang P, Chan KA, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proceedings of the National Academy of Sciences*. 2015;112(40):E5503-E5512.

138. Diehl F, Schmidt K, Choti MA, et al. Circulating mutant DNA to assess tumor dynamics. *Nat Med*. 2008;14(9):985-990.

139. Teschendorff AE, Menon U, Gentry-Maharaj A, et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS one*. 2009;4(12):e8274.

140. Shen SY, Singhania R, Fehringer G, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563(7732):579-583.

141. Chan KA, Jiang P, Chan CW, et al. Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences*. 2013;110(47):18761-18768.

142. Vos T, Lim SS, Abbafati C, et al. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019. *The Lancet*. 2020;396(10258):1204-1222.

143. Roth GA. Global burden of disease collaborative network. global burden of disease study 2017 (GBD 2017) results. seattle, united states: Institute for health metrics and evaluation (IHME), 2018. *The Lancet*. 2018;392:1736-1788.

144. Wockner LF, Noble EP, Lawford BR, et al. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Translational psychiatry*. 2014;4(1):e339.

145. Hanson M, Godfrey KM, Lillycrop KA, Burdge GC, Gluckman PD. Developmental plasticity and developmental origins of non-communicable disease: Theoretical considerations and epigenetic mechanisms. *Prog Biophys Mol Biol*. 2011;106(1):272-280.

146. Liu L, Wu J, Qing L, et al. DNA methylation analysis of the NR3C1 gene in patients with schizophrenia. *Journal of Molecular Neuroscience*. 2020;70(8):1177-1185.

147. Zhang M, Pan C, Liu H, Zhang Q, Li H. An attention-based deep learning method for schizophrenia patients classification using DNA methylation data. . 2020:172-175.

148. Gore S, Azad RK. CancerNet: A unified deep learning network for pan-cancer diagnostics. *BMC Bioinformatics*. 2022;23(1):1-17.

149. Zhou J, Chen Q, Braun PR, et al. Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders. *Proceedings of the National Academy of Sciences*. 2022;119(34):e2206069119.

150. Tang M, Huang T, Yang J, Guo C. *Integrative multi-omics for diagnosis, treatments, and drug discovery of aging-related neuronal diseases.* Frontiers Media SA; 2022.

151. Mao W, Zaslavsky E, Hartmann BM, Sealfon SC, Chikina M. Pathway-level information extractor (PLIER) for gene expression data. *Nature methods*. 2019;16(7):607-610.

152. Walton E, Hass J, Liu J, et al. Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophr Bull*. 2016;42(2):406-414.

153. Grayson DR, Guidotti A. The dynamics of DNA methylation in schizophrenia and related psychiatric disorders. *Neuropsychopharmacology*. 2013;38(1):138-166.

154. Alfimova MV, Kondratiev NV, Golov AK, Golimbet VE. Methylation of the reelin gene promoter in peripheral blood and its relationship with the cognitive function of schizophrenia patients. *Mol Biol (N Y)*. 2018;52(5):676-685.

155. Mak M, Samochowiec J, Frydecka D, et al. First-episode schizophrenia is associated with a reduction of HERV-K methylation in peripheral blood. *Psychiatry Res*. 2019;271:459-463.

156. Nishioka M, Bundo M, Koike S, et al. Comprehensive DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia. *J Hum Genet*. 2013;58(2):91-97.

157. Murata Y, Ikegame T, Koike S, et al. Global DNA hypomethylation and its correlation to the betaine level in peripheral blood of patients with schizophrenia. *Prog Neuro-Psychopharmacol Biol Psychiatry*. 2020;99:109855.

158. Wockner LF, Noble EP, Lawford BR, et al. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. *Translational psychiatry*. 2014;4(1):e339.

159. Hu M, Xia Y, Zong X, et al. Risperidone-induced changes in DNA methylation in peripheral blood from first-episode schizophrenia patients parallel changes in neuroimaging and cognitive phenotypes. *Psychiatry Res*. 2022;317:114789.

160. Li M, Li Y, Qin H, et al. Genome-wide DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia in the chinese han population. *Mol Psychiatry*. 2021;26(8):4475-4485.

161. Nishioka M, Bundo M, Koike S, et al. Comprehensive DNA methylation analysis of peripheral blood cells derived from patients with first-episode schizophrenia. *J Hum Genet*. 2013;58(2):91-97.

162. Gautam Y, Johansson E, Mersha TB. Multi-omics profiling approach to asthma: An evolving paradigm. *Journal of Personalized Medicine*. 2022;12(1):66.

163. Fikri RMN, Norlelawati AT, El-Huda ARN, et al. Reelin (RELN) DNA methylation in the peripheral blood of schizophrenia. *J Psychiatr Res*. 2017;88:28-37.

164. Zhuo C, Wang D, Zhou C, et al. Double-edged sword of tumour suppressor genes in schizophrenia. *Frontiers in Molecular Neuroscience*. 2019;12:1.

165. Pan D, Kocherginsky M, Conzen SD. Activation of the glucocorticoid receptor is associated with poor prognosis in estrogen receptor-negative breast cancer. *Cancer Res*. 2011;71(20):6360-6370.

166. Guidotti A, Auta J, Davis JM, et al. Toward the identification of peripheral epigenetic biomarkers of schizophrenia. *J Neurogenet*. 2014;28(1-2):41-52.

167. Bozinovski S, Fulgosi A. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. . 1976;3:121-126.

168. Dietterich TG, Pratt L, Thrun S. Special issue on inductive transfer. *Mach Learning*. 1997;28(1).

169. Pratt LY. Discriminability-based transfer between neural networks. *Advances in neural information processing systems*. 1992;5.

170. West J, Ventura D, Warnick S. Spring research presentation: A theoretical foundation for inductive transfer. *Brigham Young University, College of Physical and Mathematical Sciences*. 2007;1(08).

171. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Advances in neural information processing systems*. 2014;27.

172. Van Dyk DA, Meng X. The art of data augmentation. *Journal of Computational and Graphical Statistics*. 2001;10(1):1-50.

173. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*. 1987;82(398):528-540.

174. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*. 1977;39(1):1-22.

175. Weissbrod O, Rahmani E, Schweiger R, Rosset S, Halperin E. Association testing of bisulfite-sequencing methylation data via a laplace approximation. *Bioinformatics*. 2017;33(14):i325-i332.

176. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90.

177. Simard PY, Steinkraus D, Platt JC. Best practices for convolutional neural networks applied to visual document analysis. . 2003;3(2003).

178. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324.

179. Díez López C, Montiel González D, Vidaki A, Kayser M. Prediction of smoking habits from class-imbalanced saliva microbiome data using data augmentation and machine learning. *Frontiers in Microbiology*. 2022;13:2576.

180. Zheng Z, Le NQK, Chua MCH. MaskDNA-PGD: An innovative deep learning model for detecting DNA methylation by integrating mask sequences and adversarial PGD training as a data augmentation method. *Chemometrics Intellig Lab Syst*. 2023;232:104715.

181. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177(7):1888-1902. e21.

182. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nature methods*. 2018;15(12):1053-1058.

183. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*. 1987;82(398):528-540.

184. Cui X, Goel V, Kingsbury B. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2015;23(9):1469-1477.

185. Treppner M, Salas-Bastos A, Hess M, Lenz S, Vogel T, Binder H. Synthetic single cell rna sequencing data from small pilot studies using deep generative models. *Scientific reports*. 2021;11(1):9403.

186. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. . 2022:10684-10695.

187. Eizenga JM, Novak AM, Sibbesen JA, et al. Pangenome graphs. *Annual review of genomics and human genetics*. 2020;21:139-162.

188. Wang T, Antonacci-Fulton L, Howe K, et al. The human pangenome project: A global resource to map genomic diversity. *Nature*. 2022;604(7906):437-446.