Contextualized Transfer Learning: Transforming Heterogeneity into Predictive Power with Generative Latent Structures in Resource-Limited Settings

Siddharth Nirgudkar
Acton-Boxborough Regional High School, Massachusetts, United States of America
MIT PRIMES

Supervised by:
Dr. Benjamin Lengerich, Computer Science and Artificial Intelligence Lab, MIT, Massachusetts, United States of America

January 2, 2025

Contextualized Transfer Learning: Transforming Heterogeneity into Predictive Power with Generative Latent Structures in Resource-Limited Settings

Siddharth Nirgudkar

January 2, 2025

Abstract

Predicting biomedical outcomes in resource-limited settings is challenging due to data scarcity and patient variability: retraining models locally lacks power, while borrowing models fails to capture context-specific causes. Current approaches frame these challenges as a tradeoff: transfer learning enhances generalization by leveraging data from other settings but sacrifices patient-specific adaptation, while contextualized learning adapts to specific contexts but struggles with limited data. We introduce Contextualized Transfer Learning (CTL) as a novel approach that reconciles these conflicting goals by modeling the joint distribution of predictors and outcomes, $p(x, y \mid c) \sim f(z(c))$, where $f(z(c))$ represents the latent structure shared across contexts. This enables information sharing across disparate outcomes, patients, and predictors, introducing a new dimension to transfer learning: generalizing across tasks while simultaneously tailoring predictions to individual patient contexts. Data scarcity and patient variability is an especially prominent problem in neurological diseases. We apply CTL to predicting Alzheimer's disease and show that CTL reduces mean square error by 22.9% compared to contextualized regression (CR) and boosts classification accuracy by 8%, outperforming population-based methods by 30%. We also show the interpretibility of CTL, which places heavy emphasis on a select few predictors which is critical for understanding biological insight. These results highlight CTL's potential as a powerful tool for precision diagnostics, particularly in resource-limited settings.

Keywords

Machine Learning, Contextualized Learning, Transfer Learning, Computational Biology, Precision Prediction, Modeling Population Heterogeneity, Generative Latent Structures

Contents

Introduction

Challenges of Biomedical Predictions in Resource-Limited Settings

In the field of biomedical predictions, one of the foremost challenges is the need for patient-specific understanding due to the heterogeneity of diseases [1]. For example, due to advancements in genomic sequencing, we know know that there are various types of breast cancer, such as hormone-positive, HER2-postive, and triple negative breast cancers, each requiring distinct treatment approaches [2]. This complexity is especially critical in resource limited settings where it is imperative that the correct and precise treatment is selected to avoid unnecessary costs and ensure optimal patient outcomes. Traditional methods, such as population-based or cluster modeling, fall short in providing the granularity required for personalized treatment [3–5]. This gap has been addressed by Contextualized Learning (CL), which by taking account of patient context (for example, medical covariates) enables sample-specific modeling without a loss of statistical power [6, 7].

However, there is an equally significant problem CL does not address: the inability to transfer information across diverse tasks. A practical use case is during a pandemic, such as COVID-19, where hospitals face overwhelming patient loads and must prioritize who receives limited resources like ICU beds and ventilators [8, 9]. In such triage scenarios, having the ability to share information across hospitals is critical in order for adaptability - sharing information allows for patterns to be elucidated across hospitals that can then be used to make better decisions for patient outcomes. Furthermore, transfer learning itself is important inside a hospital. Having a down stream predictive model that can use information from upstream models about various other variables - environmental factors, and health indicators can then make better decisions for patient outcomes.

Transferring information is not only needed for high stakes environments but also for exploratory research for treatments, specifically when it comes to neurological diseases. Unlike other tissues where multiple samples can be taken over time, the transcriptomic state of the brain can only be observed once—post-mortem [10]. This fundamental limitation severely restricts the amount of data available for analysis, complicating the ability to make accurate predictions [11]. Neurological diseases, which are already among the most debilitating and complex to treat, are further hindered by this scarcity of data.

From making decisions in triage medical tasks to learning information about neurological diseases, there is a critical need to share information across disparate situations.

A New Hope: Generative Modeling Perspective

Generative modeling has revolutionized the field of natural language processing (NLP) through the development of large language models (LLMs), which demonstrate the ability to perform well across disparate tasks [12]. These models, which learn to predict the next word in a sequence or fill in a masked token, can share their knowledge across a wide range of tasks, creating emergent behaviors that were not explicitly trained for: translating, summarizing, or writing creatively [13]. They are able to do this by uncovering underlying patterns and structures in the data that are consistent across different tasks [14].

We aspire to do this in biomedical settings, to harness the associated benefits. However, unlike NLP where there is an obvious pretext task such as predicting masked words, biomedical data lacks a straightforward equivalent [15]. In NLP the masked token task helps the model learn relationships within language data [16], but in biomedicine the main challenge is to find a task that can enable generlizability across different datasets and conditions.

Generative modeling provides a potential solution. Instead of directly estimating the conditional distribution $p(y \mid x)$ where x are the predictors (features) and y are the outcomes, generative models aim to estimate the joint distributions $p(x, y)$ [17]. By doing so, we can uncover latent structures within the data that are consistent across disparate tasks, contexts, and observed variables. This approach can lead to more generalizable biomedical predictions.

Our Contribution: Contextualized Transfer Learning (CTL)

We propose Contextualized Transfer Learning (CTL) to leverage this generative modeling perspective. The core idea for CTL is that $p(x, y, c) \sim z$ with $y, x \perp c|z$. Here, z represents a latent space whose distribution $p(z)$ is conserved across tasks (outcomes, predictors, and contexts). The intuition behind this assumption is prevalent in the biological space. We all share physical processes and we all have a finite set of regulators expressed in different ways

- creating unique genotypes and phenotypes. If we gain understanding of this shared space, we can use it across diverse tasks and environments. This perspective allows us to decompose the conditional probability $p(x, y \mid c)$, into separable terms that rely on a shared z instead of experiment specific variables: $p(y|x, c) = \int_z dZp(y|x, z)p(z|c)$. By being able to model y solely based on x and z, information based on one context c, can be represented through z and used for disparate contexts. Moreover, $p(z \mid c)$ will be retained across observations and tasks allowing us to have a common transfer medium even in scenarios where predictors and outcomes are not shared.

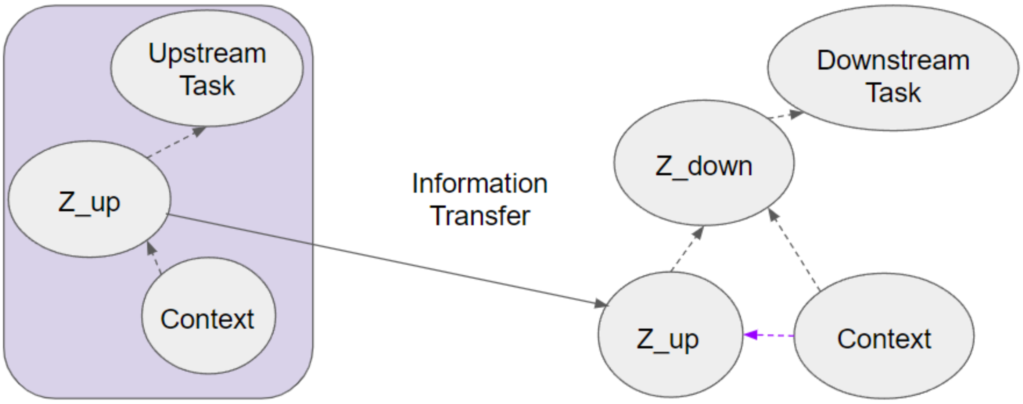


Figure 1: Motivating idea for CTL: Information from upstream tasks can be shared to downstream tasks through latent information z even if observations, predictions, and context vary. *Created by Sid Nirgudkar

Similar Heterogeneity as Generative Consistency Within this generative framework, "similar heterogeneity" refers to the consistency in the distribution $p(z)$ across different tasks and patients groups, due to the common underlying structure. This consistency implies that while outcomes and patient-specific factors may vary, the fundamental distribution of the predictors remains stable across different tasks. This generative consistency is crucial for CTL's ability to transfer knowledge across diverse settings.

Advantages of Generative Consistency With this assumption of generative consistency, CTL solves the two apparently-conflicting goals:

- **Transferability:** Since $p(z)$ is similar across tasks, CTL can transfer learned models or parameters from one task to another, under the assumption that the new task will operate under a similar generative process for the predictors.
- **Adaptability:** CTL can also adapt to specific patient contexts within this generative framework by adjusting the model to account for variations in $p(y|x)$, the conditional distribution of outcomes given predictors, across different patient groups.

Contributions

In this paper, we introduce Contextualized Transfer Learning (CTL) as a novel approach that integrates generative modeling principles to address the challenges of transfer learning in complex, resource-limited biomedical settings. Our method offers a new perspective on how generalization and customization can be harmonized through the shared latent structures of $p(x)$, ultimately enabling more effective and adaptable predictive models across varied contexts. This framework requires several innovations:

- To facilitate the sharing of information, specifically the learned latent parameters from upstream models, we needed insight on how that should be done. We created a new idea formulation that allows for adding this shared heterogeneity information to contextualized models.
- We develop a deep learning architecture to efficiently perform CTL.

- We introduce a method for efficient archetype dictionary creation. Archetype dictionaries are the 'core ingredient' models that every sample-specific model is created from [6, 7, 18]. Currently, the size of the archetype dictionary (number of core ingredients) is guessed by the user, but we propose a method of estimating the size of the archetype dictionary through ϵ convex hull approximation. Giving an upper bound on the number of archetypes needed will reduce unnecessary computational overhead and is useful not only for CTL but general contextualized use cases as well.

- Finally, we demonstrate the usefulness of CTL through its application in neurological scenarios, specifically in increasing the accuracy of predicting Alzheimer's in patients.

Related Work

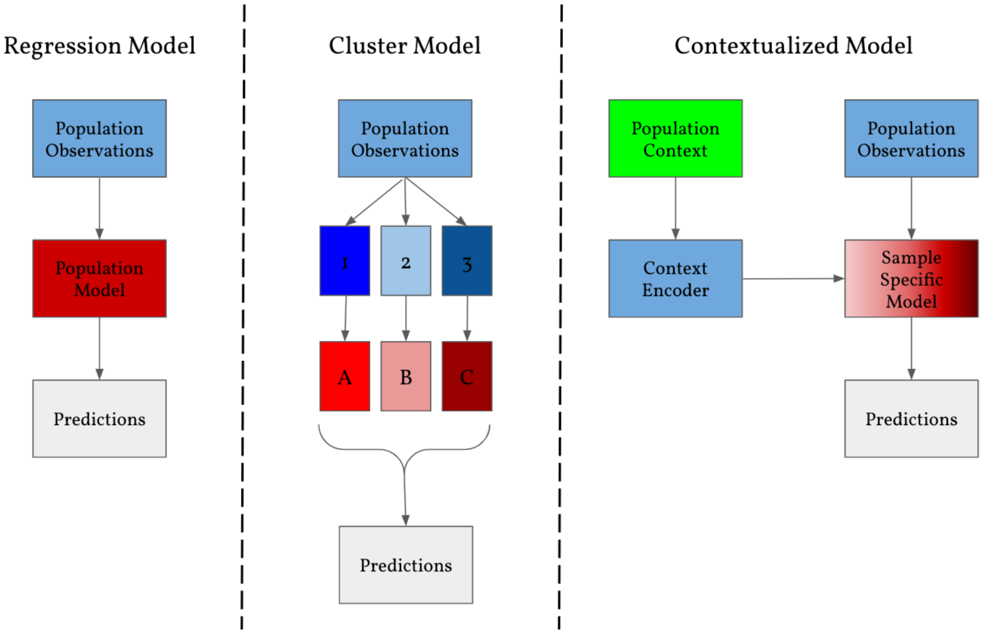


Figure 2: A Comparison of Population based modeling, Cluster based modeling, and Context based modeling. In population modeling there is only one model per population cohort. For cluster based modeling, the population is differentiated into sub groups where each group is homogeneous and each subgroup uses it's own model. Contextualized model creates sample specific models based of context. (image credit: [19])

Contextualized Learning CL creates individual models for each patient based on their specific context [6, 18]. It uses deep learning to understand the complex relationships between a patient's context (e.g. their medical history) and the model parameters needed to make accurate predictions [7, 18]. First data is collected, which includes observations (e.g., symptoms, test results, genetic levels) and contextual information (e.g., clinical background, environmental factors). The context encoder is a deep learning model that translates the contextual information (also called covariates) into parameters for a context-specific model [6, 7].

Transfer Learning Traditional Transfer Learning (TTL) effectively transfers knowledge between similar tasks, like various types of image classification [20–22]. However, TTL does not work effectively when different source and target tasks are involved, limiting its scope [23–25]. Heterogeneous Transfer Learning attempts to bridge this gap by creating an intermediate feature space through complex mappings [26, 27]. While this allows for transfer across disparate tasks, it results in black-box models that are difficult to interpret. This lack of transparency is problematic, especially in medical contexts, where the model's interpretability is critical.

Contextualized Transfer Learning

Mechanism for Information Transfer

Information from upstream models will create more accurate sample-specific models that better reflect the true population heterogeneity, and this will, in turn, improve model performance. In traditional contextualized learning, one set of contexts - or a single context modality - is used to create sample-specific models. For example, these could be medical information for a patient. We expand this framework, incorporating latent information from upstream models as additional context modalities. All of these context modalities are then used to create sample-specific models for our current task.

Problem Statement

Given a labeled dataset consisting of targets $y \in \mathbf{Y}$, observations $x \in X$ and a set of n context modalities $\{c_1, c_2, \ldots c_n\}$ (for one patient) $\rightarrow \{C_1, C_2, \ldots C_n\} = \mathbb{C}$ where $C_n = [c_{n_1}, c_{n_2}, \ldots c_{n_i}]$ (i patients), we would like to learn a model: $\mathrm{P}(\mathbf{Y} \mid X, \mathbb{C}, \theta)$. Our targets are the outcomes we aim to predict, and our observations are used to make predictions about the targets. Context modalities are called covariates - they represent contextual info that provider background and influence the target variable but are not directly used for prediction. For example C_1 could represent the clinical background for our patient group, and C_2 could be information regarding their environment that was learned from an upstream model. We represent every sample specific model as θ_i.

Problem Solution

We approach this problem by enabling information sharing across heterogeneous tasks as shown in the graphical model below. For each patient, every context modality c_n can be represented as a lower complexity refined context r_n. Every upstream task (n) has a latent factor (subtype) s_n that itself is generated from the super subtype $\mathbb{S}$. This super subtype generates θ, X, Y for every task.

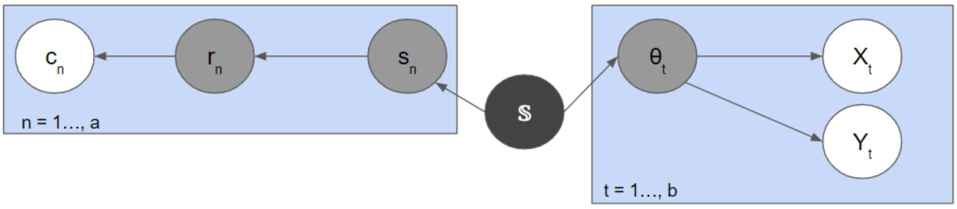


Figure 3: Graphical Model for CTL. The white colored circles signify information that is observed (raw context c, observations X and output Y. The shaded circles represent latent variables. The graphical model represents our assumption of some shared information in a latent space which is $\mathbb{S}$ in the figure that also produces θ, X, and Y. The innovation here lies in the utilization of shared latent information across different tasks, which traditional models do not account for. a represents the number of context modalities that are present, and b represents the number of tasks for the current model. For example, predicting the presence of Alzheimer's and the type (early or regular) would be considered as two tasks. *Created by Sid Nirgudkar

From this graphical model, we can define a specific probabilistic model and create a differentiable loss function.

$$P(\mathbf{Y} \mid X, \mathbb{C}) = \int_{\theta} P(\mathbf{Y} \mid X, \theta) \cdot P(\theta \mid X, \mathbb{S}) \cdot P(\mathbb{S} \mid \mathbf{R}) \cdot P(\mathbf{R} \mid \mathbb{C}) d\theta \tag{1}$$

To build this model, we ultimately seek to create a differentiable loss function ℓ_u for unsupervised tasks (Markov Networks, Bayesian Networks, Neighborhood Networks) or ℓ_s for supervised tasks (Regression/Classification). To define ℓ as a differentiable loss function, we create an end-to-end encoder - comprised of the shared preparation functions F that map $C_n \rightarrow R_n$, context encoders E that map $R_n \rightarrow s_n$, the subtype weightage function β that create the super subtype, and the archetype dictionary $\mathcal{A}$ that is weighted with the super subtype to create the sample specific model. Our end-to-end encoder $\phi(\mathbb{C}; F, E, \beta, \mathcal{A})$ is defined below where N is patient cohort size.

$$\phi(\mathbb{C}; F, E, \beta, \mathcal{A}) = \theta \tag{2}$$

Our single loss function for unsupervised models can be encapsulated by our differentiable loss function ℓ_U. This would be used to study heterogeneity in our patient group and look for patterns in our data instead of trying to predict. Our process for learning all the parameters is as follows:

$$\hat{F}, \hat{E}, \hat{\beta}, \hat{\mathcal{A}} = \underset{F,E,\beta,\mathbf{A}}{\operatorname{argmin}} \sum_{i=1}^{N} \ell_U\left(\phi(\mathbb{C}; F, E, \beta, \mathcal{A})X_i)\right) \tag{3}$$

For supervised models, we use a more well-defined loss function (Mean Squared Error with an L1 regularization term) as ℓ_s:

$$\ell_s = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{i=1}^{N} |\theta_i| \tag{4}$$

Substituting with the defined parameters yields:

$$\ell_s = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i \cdot \phi(\mathbb{C}; F, E, \beta, \mathcal{A}))^2 + \lambda \sum_{i=1}^{N} |\phi(\mathbb{C}_i; F, E, \beta, \mathcal{A})| \tag{5}$$

Deep Learning Architecture for CTL

The CTL framework expands on the CL framework by incorporating multiple context modalities and allowing for heterogeneous tasks as shown in Fig.4. The overall learning structure will be a feed forward network where the training data is processed through the network, while the validation set is utilized to prevent over-fitting. A standard backpropogation is used to train the model.

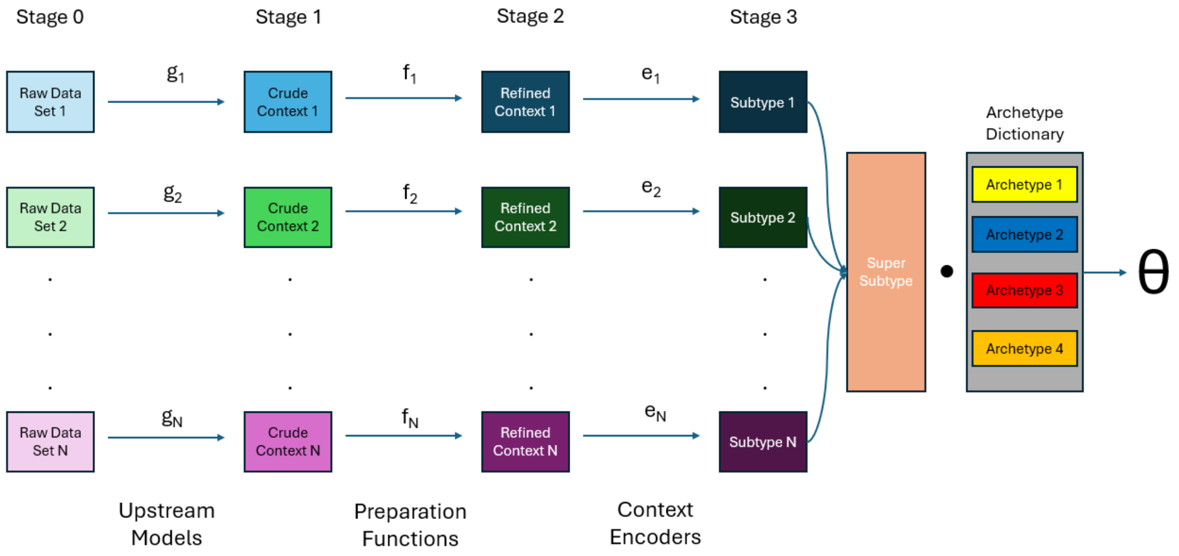


Figure 4: Architecture for transfer learning paradigm. *Created by Sid Nirgudkar

Stage 0: The initial stage involves raw data collection, including both context and observations. However, only the contexts are used to create the sample-specific model. The observations are used to train the feed forward network. This raw data is passed through upstream models represented as the g functions, and the data extractions from that make context modalities (a different modality for every upstream model). It is important to note that one g function will be the identity for the associated context with the current data for the task. The crude context is the direct output from the upstream models. We do not have access to or the ability to optimize these upstream models.

Stage 1: We then refine each crude context modality, using a set of shared transformation functions (across samples) for every context modality. These transformation functions simplify the crude context to the refined context (r_n) that has an order of 1. This is important because simplifying the complexity of the data makes it easier for the context encoders to perform well and integrate additional information into the model, because $\dim(r_n) \leq \dim(c_n)$. Generally, $f_n(c_n) = r_n$. If $\text{ord}(c_n) = 1$, then its f_n is linear:

$$r_n = f_n(c_n) = wc_n + b$$

where w are the weights and b is the bias, $w_i, b_i \sim \mathcal{U}(-1, 1)$. If $\text{ord}(c_n) \geq 1$, then we utilize a Multi-Layer-Perceptron:

$$r_n = f_n(c_n) = W_L \cdot \phi\left(W_{L-1} \cdot \phi\left(\ldots \phi\left(W_1 \cdot c_n + b_1\right) + b_2 \ldots\right) + b_{L-1}\right) + b_L$$

where the number of layers L depends on the complexity of the crude context modality and is a hyper-parameter, and ϕ is a non-linear activation function (ReLU).

Stage 2: Once the refined context $r_n \in R$ is created, a shared set of context encoders $e_n \in E$ acts on them. Through prior work [6, 7], we have had the best success with Neural Generative Additive Models (NGAMs) as context encoders [28]. Every encoder e_n maps r_n to a subtype $s_n \in \mathbb{R}^k$, where k is the size of the archetype dictionary. A subtype is a vector that will weigh the archetypes to create personalized models; they can be viewed as patient-specific IDs that reflect their context. The specific structure and process of the NGAM context encoder is as follows: Every feature $m = [m_1, m_2, \ldots, m_i] \in r$ is processed by a sub-encoder a—a single-layer neural network:

$$\forall i,\ o_i = a_i(m_i) = \phi(w_{1,i} \cdot m_i + b_{1,i})$$

Generally, our context encoder e_n that operates on a single r_n can be described as:

$$s_n = e_n(r_n) = \sum_{i=1}^{I} a_i(m_i) + b_n$$

More specifically, $s_n = w[o_1, o_2, \ldots, o_i]^T + b$, where $w \in \mathbb{R}^{k \times i}$. Once again for both the sub-encoder and the context encoders. $w_i, b_i \sim \mathcal{U}(-1, 1)$.

Stage 3: Every s_n is finally linearly weighted $\Sigma_{i=1}^{n} w_i s_i$ to produce the super subtype $\mathbb{S}$. This process ensures that the relevance of each r_n and its associated s_n is appropriately weighted depending on its importance to the main task. For example, if the refined context came from a function mapping Gene Regulatory Networks (GRNs) for a foot disease, when predicting a brain disease outcome, it should hold less weight compared to a context mapping GRNs for a brain disease. The archetypes serve as the vertices of a convex hull, representing extrema models - all sample-specific models are convex combinations of these archetypes. The archetype dictionary $[A_1, A_2, \ldots A_k] \in \mathcal{A}$ stores all archetypes. The shape of an archetype A_i is $\forall i \in k$ is $\mathbb{R}^{d_{\text{observations}} \times d_{\text{output}}}$. All weights w in the archetype dictionary are uniformly initialized [-1,1]. To produce each sample-specific model θ, we apply the sigmoid function to $\mathbb{S}$ to constrain the weightage:

$$\theta = \sigma(\mathbb{S}) \cdot \mathcal{A}$$

where $\sigma(\mathbb{S}) \in [0, 1]^k$ ensures the convex combination of archetypes. It is important to note that $\mathcal{A}$ is learned - nothing has to be known *a priori*.

Analysis of Framework

Glass Box Transfer Learning In contrast to heterogeneous transfer learning, Contextualized Transfer Learning (CTL) offers a transparent 'glass box' model, allowing users to interpret how context is integrated and predictions are made. Unlike heterogeneous transfer learning, which increases data complexity, CTL reduces it by projecting data into a lower-dimensional space, improving interpretability and practicality. CTL also avoids the computational burden of creating complex intermediate spaces by leveraging shared heterogeneity and archetype-based modeling, and thereby efficiently integrating diverse data sources. This makes CTL particularly suitable for resource-limited settings, enhancing predictive accuracy even with limited data and computational resources.

Increased Consistency in Predictions With CTL, we also observe better constant predictions with less variability as compared to CL - due to the added information.

Proof: We assume that every C_i is drawn from a common distribution $\mathbb{P}(C)$ with a μ_c and σ_c^2. Given that every $r_i = f_i(c_i)$ we assign μ_r, σ_r^2 as the mean and variance of a refined context modality. We assume that all r_i have been taken independently (from different upstream models), and we average all of them to produce R, parallel to what happens with our subtypes. From here we see that $\mathbb{E}(R) = \mu_r$ and $\text{VAR}(R) = \frac{\sigma_r^2}{n}$ We can connect this to the model predictions because we can define the whole model as a smooth function ω so $\hat{Y} = \omega(\text{R})$. From the delta method we get $\text{VAR}(\hat{Y}) \approx (\frac{d\omega}{dR})^2 \text{VAR}(R)$ [29]. Substituting $\text{VAR}(R)$ it is trivial to see that as n (context modalities) increase, the variation of the predictions decrease.

Flexible Archetype Dictionary CL requires users to estimate the size of the archetype dictionary, which can be trial end error at best. CTL handles a lot of information, so to increase unnecessary computational overhead, we introduce a method to determine the optimal size of the archetype dictionary using ϵ convex hull approximation.

Method: *Suppose contexts C are contained in a polytope P_0 with k_0 vertices which can be approximated by a polytope P_1 with $k_1 < k_0$ vertices with Γ information loss. Then, the k_1 archetypes can be used to represent sample-specific models with no more than $f(\Gamma)$ information loss.*

Rationale: The raw context space inherently contains unnecessary complications that hinder in determining the archetype dictionary size. To overcome that, we devise a framework to create a 'complexity' representation for every sample in a dataset. Every sample i has n context modalities. For every n we calculate the distance between the sample value and the mean value $C_{n_i} - C_{n_\mu}$ that is computed through the Mahalanobis distance [30, 31]. Doing this for all i we create a dataset $D \in \mathbb{R}^{i \times n}$. The model space θ is a lower complexity reflection of the co-variate space [18]. If the optimal convex hull of D represented as $\text{OTP}(D, 0)$ has k_0 vertices then $\text{OTP}(\theta, 0)$ by principle will need **at most** k_0 vertices [32, 33]. Following the approach of [32, 33], we can create a ϵ-approximate convex hull T, $T \subseteq D$ such that the Hausdorff Distance between T and D should be at most ϵ. Finding the optimal number of vertices of a ϵ-approximate convex hull $\text{APP}(D, \epsilon)$ is very hard, and only [32] algorithm comes close to reaching that boundary **which is what we used**. Conventional approximation formulas that do not take into account of D, state that there exists an ϵ-approximate convex hull with $\frac{1}{\epsilon^{\frac{n-1}{2}}}$ vertices, however in most cases that value is far greater than $\text{OTP}(D, \epsilon)$ [32]. Given that θ is a lower-complexity representation of $\mathcal{C}$, obtained through a distance-matching regularizer, the approximate convex hull $\text{APP}(\theta, \epsilon)$ retains the essential structure of the approximate convex hull $\text{APP}(D, \epsilon)$. However, due to the reduced complexity of θ, fewer points are required to approximate $\text{APP}(\theta, \epsilon)$ as compared to $\text{APP}(D, \epsilon)$.

Mathematically, the set $\mathbf{U}$ containing points that approximate the convex hull $\text{APP}(D, \epsilon)$ is defined iteratively by selecting the point $\mathbf{q} \in D$ that has the greatest distance from the current approximation $\text{APP}(\mathbf{U}, \epsilon)$:

$$\mathbf{U} = \mathbf{U} \cup \{\mathbf{q}\}, \quad \text{where } \mathbf{q} = \arg\max_{\mathbf{p} \in D} \text{distance}(\mathbf{p}, \text{APP}(\mathbf{U}, \epsilon)),$$

$$\text{and } \mathbf{q} \in D \text{ is selected if } \text{distance}(\mathbf{q}, \text{APP}(\mathbf{U}, \epsilon)) \geq \frac{\epsilon' \cdot \delta_0}{2}.$$

This process continues until no point $\mathbf{q}$ satisfies the condition:

$$\text{distance}(\mathbf{q}, \text{APP}(\mathbf{U}, \epsilon)) \geq \frac{\epsilon' \cdot \delta_0}{2}.$$

Here, δ_0 is the approximate diameter of D, calculated as:

$$\delta_0 = \max_{\mathbf{p}_i, \mathbf{p}_j \in D} \|\mathbf{p}_i - \mathbf{p}_j\|,$$

where δ_0 has the same dimensionality as the space in which D resides (i.e., δ_0 is in units of distance in $\mathbb{R}^n$). The parameter ϵ' is an adjusted tolerance factor, calculated as:

$$\epsilon' = 8\epsilon^{1/3} + \epsilon,$$

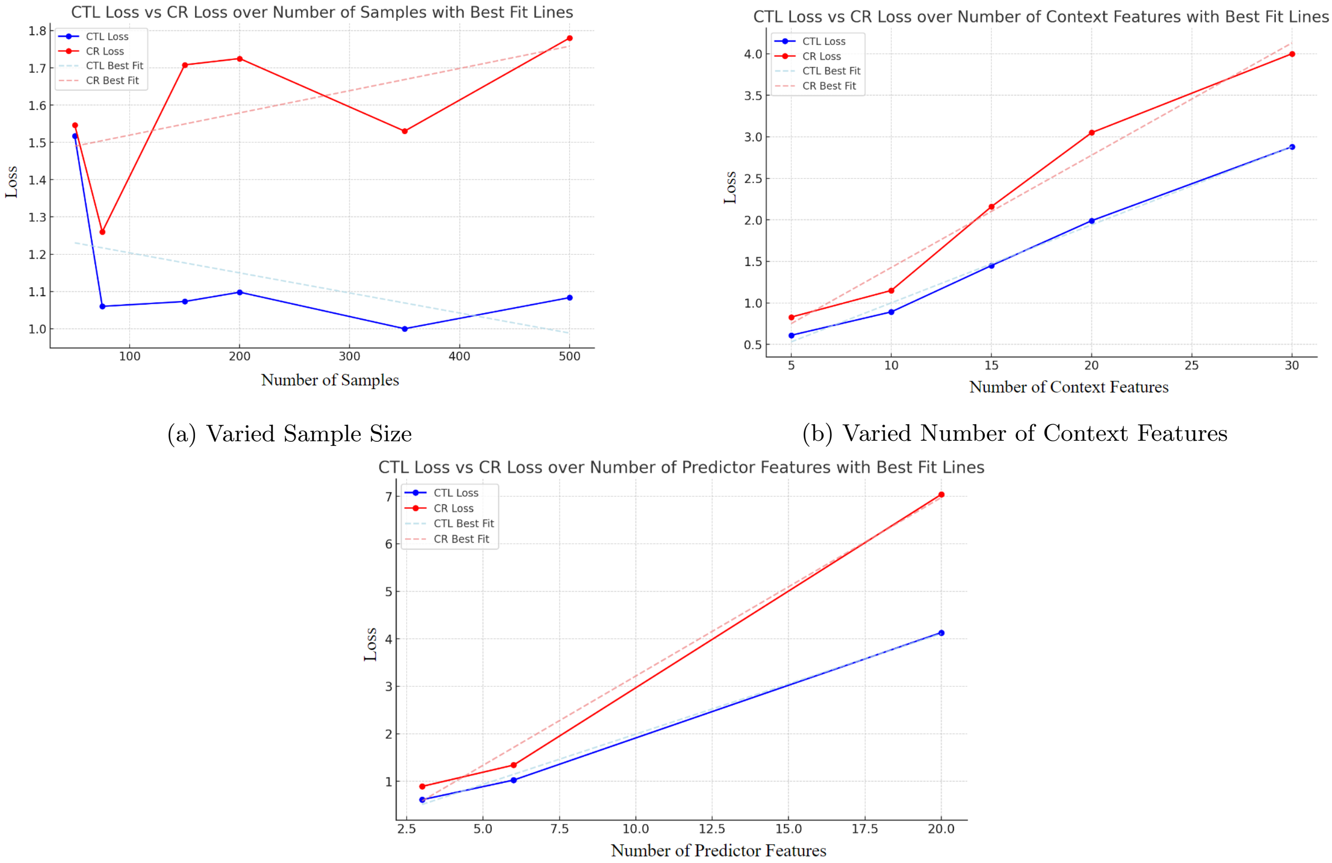
where ϵ' is dimensionless, reflecting the degree of approximation tolerated in the convex hull construction. Furthermore, we propose a metric Γ that is defined as the percentage of points that are not covered in the ϵ-approximate convex hull based off [34] . Γ serves as a concise method for measuring information loss that is meaningful to the average user. If $\text{APP}(D, \epsilon)$ has Γ information loss then $\text{APP}(\theta, \epsilon)$ has $f(\Gamma)$ information loss where $f(\Gamma) \leq \Gamma$. We can define this upper bound is because θ is a lower dimensionality projection of D.

Case Study

Preliminary Simulation Data

As a proof of concept, we test CTL in various complexities and environments, synthetic data was created and used due to the difficulty for procuring real medical datasets.

The simulation dataset took a variety of sizes for X and C. We tested in supervised settings, and the baseline was a traditional CL model that had access to X, C, Y. CTL also had the learning's from one upstream unsupervised model that created genetic correlation networks. This experiment would model use case scenarios where genetic information is learnt upstream and then utilized for a different task.



(a) Varied Sample Size

(b) Varied Number of Context Features

(c) Varied Number of Predictor Features

Figure 5: Performance Comparison of Contextualized Transfer Learning verses Contextualized Regression in various test cases. *Created by Sid Nirgudkar

As can be seen from Fig.5, CTL consistently outperforms CL in a variety of scenarios. We also observe some other interesting trends. From 5a, we see an example of theorem 2 - there is less variance in CTL as compared to CR when sample size is changed. We also see that CTL accumulates loss at a slower rate as compared to CL, which could prove to be advantageous in scaling up to difficult machine learning tasks.

ROSMAP Alzheimer's Dataset

Alzheimer's disease (AD) is a progressive neurodegenerative disorder marked by cognitive decline, memory loss, and behavioral changes [35, 36]. It is highly heterogeneous, with variability in symptoms, progression, and underlying biology, influenced by genetic, environmental, and lifestyle factors [37]. This heterogeneity makes accurate prediction and diagnosis challenging. The Kellis Lab [38] provided us access to the ROSMAP dataset which includes extensive genomic data and clinical context of individuals with and without Alzheimer's. On this real medical dataset, CTL outperformed CL and the population baseline model in regression and classification as well as uncovered significant latent context and genetic factors that played a heavy role in the differentiation of sample specific models.

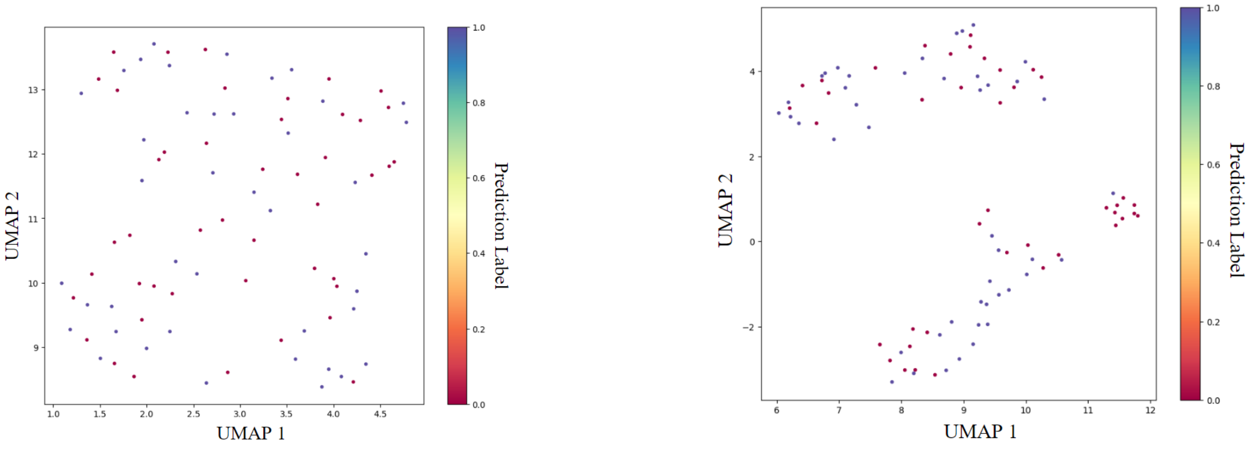
The dataset consisted of 427 samples, over 40000 genes and 89 context markers. We were constrained by our computing power available, so the data was normalized and dimensionality reduction was performed (through PCA), simplifying the dataset to 427 samples, 50 genetic features, and 50 context features. From there, 272 samples were used for training, 69 were chosen for validation, and 86 were used to test. Random seed selection was chosen to split the dataset.

The baseline regression and classification model worked only with the genes and was asked to predict the probability of a patient developing Alzheimer's. CL worked with the genes and the context for the patient, and contextualized regression and classification were performed. Through this dataset we showed one of the innovations with CTL, to use Genetic Correlation Networks (GRNs) side by side with clinical context - the 50 genes were fed to an upstream contextualized correlation network model to output a GRN matrix that was used as the second context set downstream alongside the clinical context, predicting the probability of a patient developing Alzheimer's.

	Classification		**Regression**
	Correct Classifications	Incorrect Classifications	Mean Squared Error Loss
Population	30	56	0.4531
Contextualized	47	39	0.3652
CTL (ours)	**56**	**30**	**0.2817**

Table 1: Results of classification and regression experiments on ROSMAP Alzheimer's dataset, comparing population and contextualized models to CTL. By transferring across tasks and predictors, CTL outperforms both the population models and contextuailzed models. *Created by Sid Nirgudkar

It can be seen that CTL outperformed CC, getting around 8% more correct classifications, while significantly surpassing PC (Population Classification Baseline), getting a 30% improvement. CTL has reduced the MSE loss by around 20 % as compared to CC, and Contextualized Regression (CR) itself offers around a 20 % reduction in MSE loss as compared to the population regression (PR).



(a) Contextualized Classification

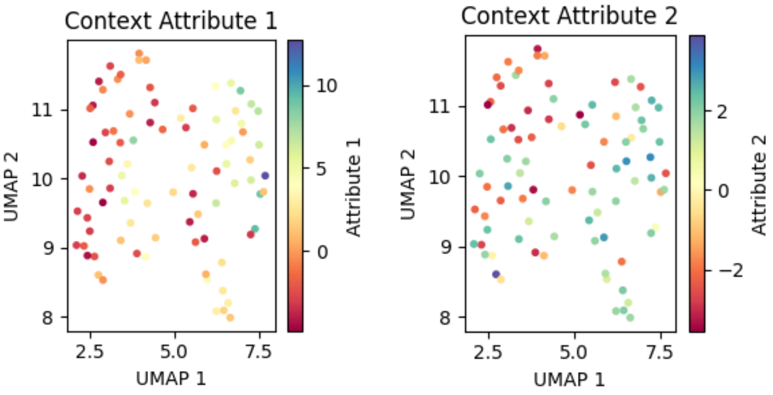
(b) Contextualized Transfer Learning

Figure 6: Comparison of heterogeneity of sample specific models between CC and CTL. A blue dot represents a positive Alzheimer's prediction while a red dot represents a negative Alzheimer's prediction. *Created by Sid Nirgudkar

Through a general comparison of the UMAP projections of θ we can uncover some useful information about the predictions. As shown in Fig. 6a, CC seems to have almost no clustering of groups, indicating a high degree of variability in the regression parameters for individual patients. While Alzheimer's is a very heterogeneous disease, we expect models to give us some further insight by grouping patients together, hinting at common latent factors that can then be analyzed in the future by doctors. In contrast, in Fig. 6b we see that CTL has a more structured arrangement of data points, with clear and distinct clusters. This shows how CTL is more effective at grouping patients with similar regression parameters together. The presence of more defined groups shows that CTL can better leverage the shared heterogeneity across different tasks, leading to personalized predictions based of shared factors and can learn these latent discriminatory factors. While both of the bigger groups are semi-heterogeneous, the bottom shape seems to have a more defined pattern - the negative predictions for Alzheimer's are at the end. We also see a distinct group that is almost all negative predictions, something that is not present in contextualized classification. It is important to realize that these differences are due to the extra information that was passed down from an unsupervised model, speaking to the power of sharing information across disparate tasks.

In seeking to understand the role context played in predictions, we wanted to analyze the regression parameters to notice some trends or patterns.

CC does not show many influential context features as seen in Fig. 11 however we can see clear patterns in two of the context features analyzed more closely below:

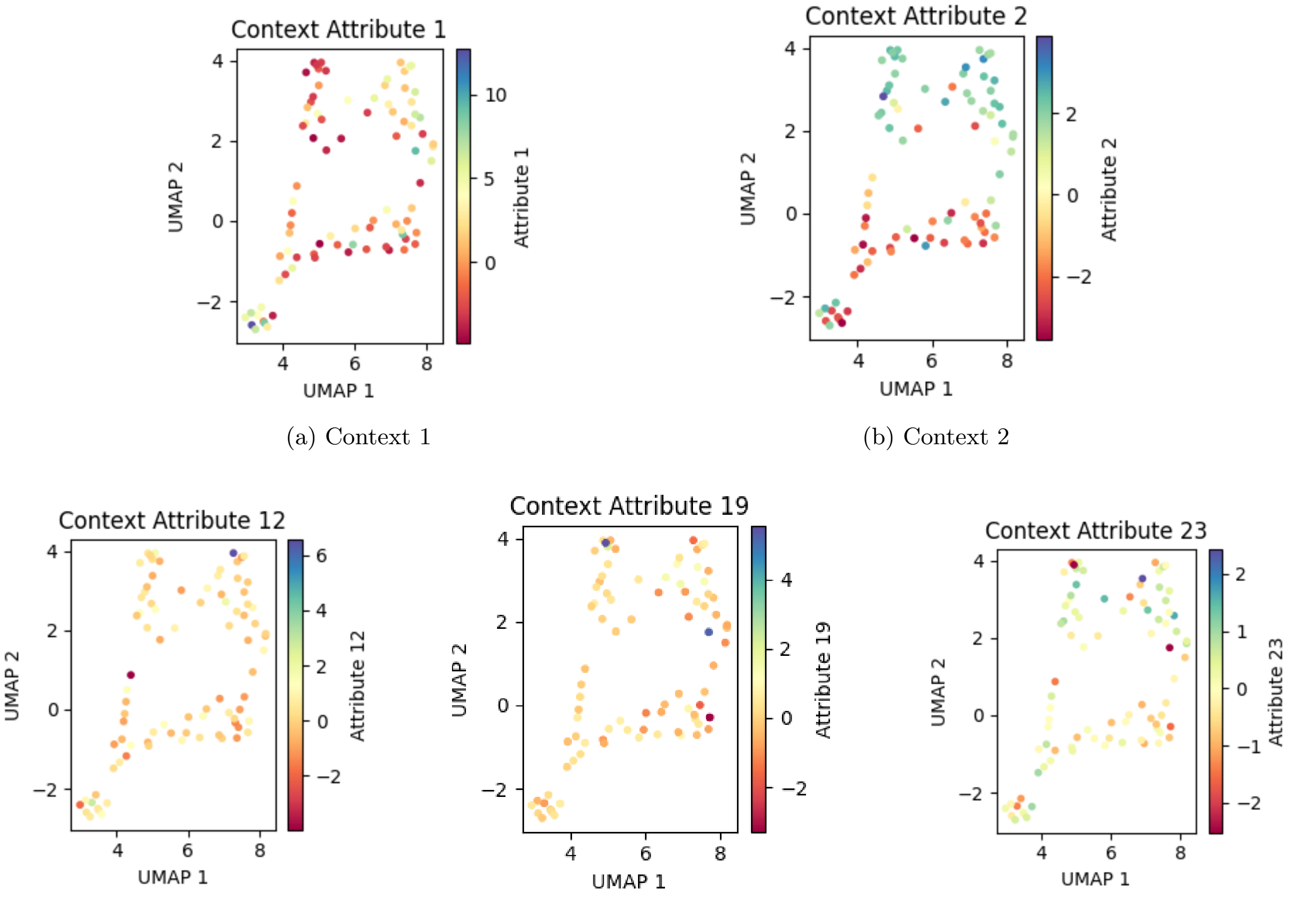


(a) Context 1

(b) Context 2

Figure 7: The two most influential contexts for creating sample specific models in contextualized classification models. a-b) clear horizontal gradient present. *Created by Sid Nirgudkar

In contrast to CL, CTL provides us with more information about influential context features. From Fig. 12 five are especially prominent - that can be seen in Fig. 8 below:



(a) Context 1

(b) Context 2

(c) Context 12

(d) Context 19

(e) Context 23

Figure 8: Illustration of deterministic contexts in forming sample-specific regression parameters. (a) Distinct top-left oval-like cluster of points in terms of color. (b) Clear horizontal separation is visible. (c-d) Context 12 shows a darker color gradient at the bottom progressing to a lighter gradient, and context 19 displays a horizontal gradient from left to right. (e) A vertical gradient is present, more pronounced than in (c) and (d), but less so than in (a) and (b). *Created by Sid Nirgudkar

As compared to CC, we firstly see more influential contexts. The most influential contexts features - in Fig. 8a and Fig. 8b, show more patterns in CTL as compared to CC. There is not only a gradient present but some localized regions in the UMAP space that correspond to the context feature values. We also see semi-influential context features in Fig. 8c, 8d, 8e. While the gradients are not as pronounced we see darker regions on certain sides of the figures as well as areas that contain a greater concentration of a different color. These sorts of patterns that CTL allows us to see are especially critical when it comes to predictions as it provides meaningful information to doctors about what context could plays a large role in differentiating people from the perspectives of disease prognosis.

Lastly, we sought to understand which of the genetic features played critical roles in creating sample specific predictions. While it is known that there is significant interplay between various genes and their resultant phenotypic effects, it is important to have glass box models that have some interpretive components to ensure that doctors can easily interpret the model.

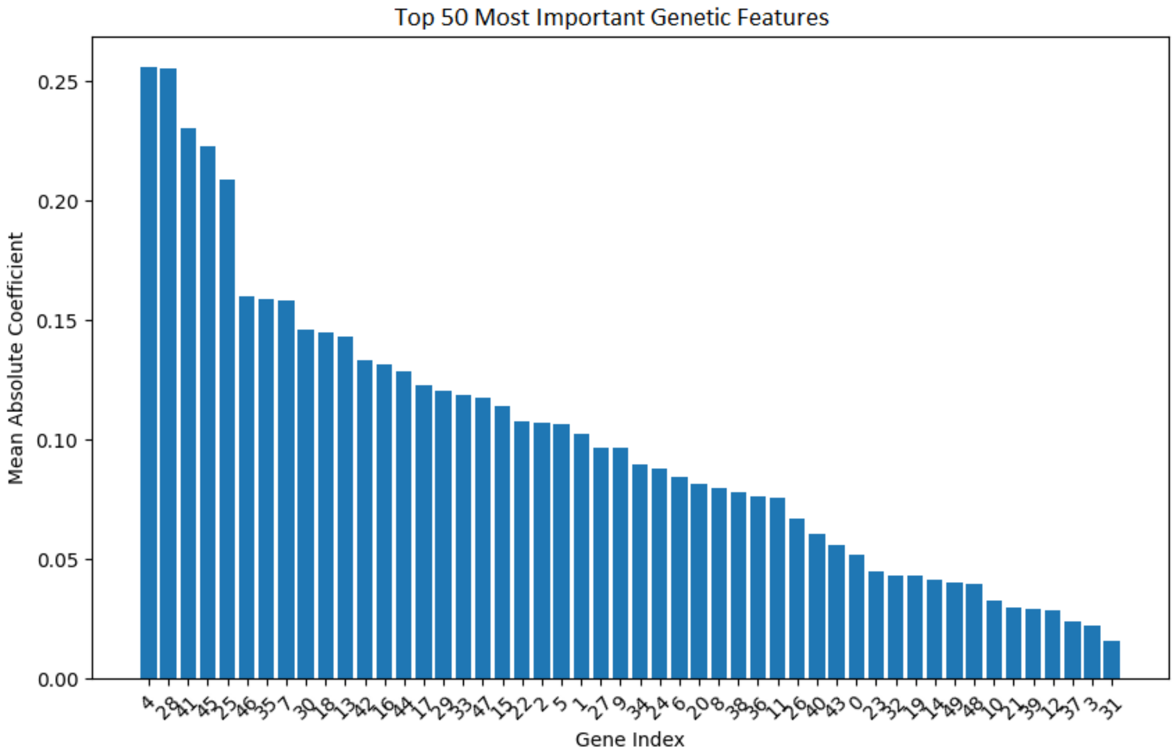


Figure 9: Average associated regression coefficient in contextualized correlation models $n = 86$. *Created by Sid Nirgudkar

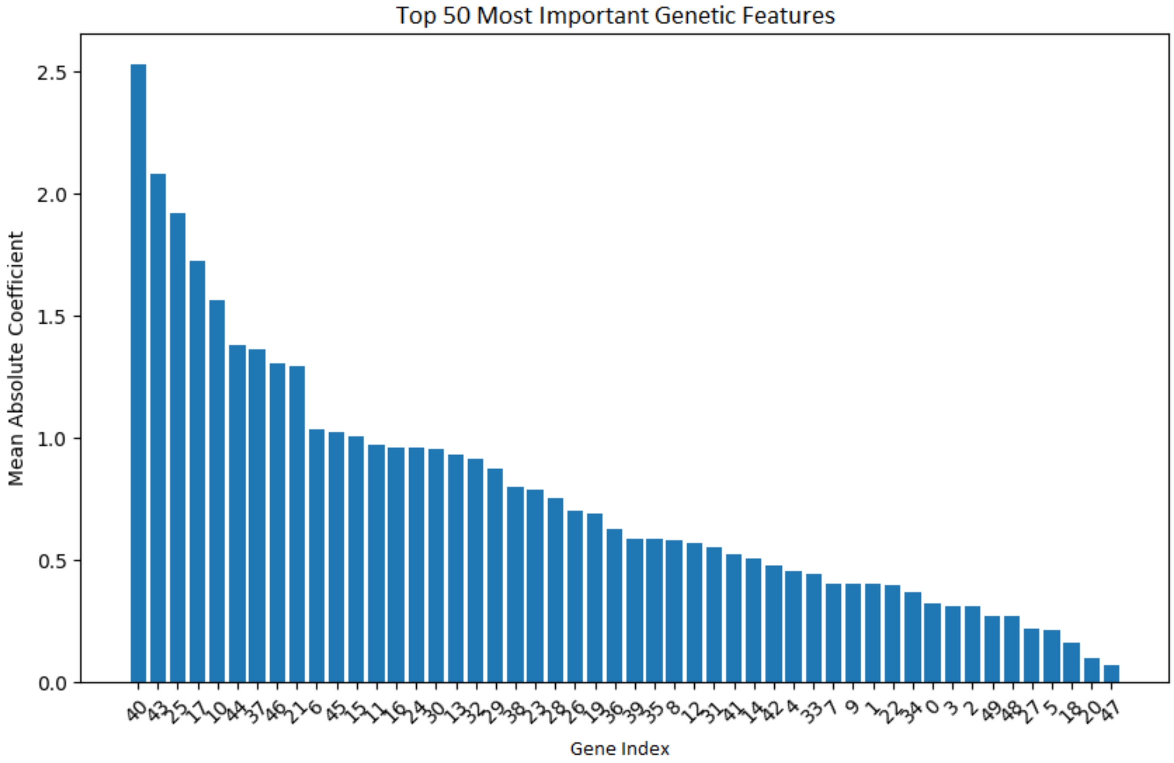


Figure 10: Average associated regression coefficient in contextualized transfer learning models $n = 86$. *Created by Sid Nirgudkar

From analyzing the average weights associated with the genetic features we can explore some interesting trends. As shown in Fig. 9 and 10 both CC has have a handful of genetic feature that have a mean absolute coefficient (mac) that is much higher than the rest of the genetic features. However CTL has one genetic feature that is much higher than the next few next highest. In addition, the actual genetic features each model places a high importance on are almost completely different except genetic feature 46 which both models seem to place a relatively high level of importance on. Additionally, the average mac level is much higher in CTL as compared to CC, which shows how CTL is deterministic and significantly increases the interpretability of these models. Once again, this is vital in medical scenarios where models should be as glass-box as possible.

Due to computational restrictions, we are only able to elucidate which genetic features are important, however this similar process can be done with more computational power to identify individual genes. Each genetic feature is created by adding many genes together to create a larger group, so it shows us that the collections of the genes that created genetic feature 40 plays a critical role in prediction. Obviously, if we had enough computational power, then each gene could be used allowing us to elucidate individual genes.

Discussion

CTL as Generative Transfer Learning

CTL can be viewed as a form of generative transfer learning. The key insight is the recognition of a stable or similar $p(x)$ across various contexts. This allows the model to transfer knowledge by generating predictions or insights based on the shared structure of the data. For example, in predicting Alzheimer's, CTL leverages the generative similarity in predictors such as genetic markers or clinical features x, even though the outcomes y (e.g., disease progression or Genetic Correlation Networks) differ. By modeling $p(x)$ generatively, CTL can adapt the learned model to new tasks or patient populations without the need for extensive retraining.

Implication of Findings

The findings from this study on Contextualized Transfer Learning (CTL) reveal significant advancements in predictive modeling, particularly in resource-limited settings. CTL enhances predictive accuracy by leveraging upstream information, enabling accurate predictions even with sparse data. This capability is crucial in low-resource healthcare environments where comprehensive patient data collection is challenging. Additionally, CTL's integration of diverse contextual information supports personalized medicine, offering precise predictions tailored to individual patient profiles and leading to more effective and targeted treatments. Demonstrated through the Alzheimer's case study, CTL significantly outperforms traditional methods, and can aid clinicians in making better-informed decisions. Moreover, CTL's ability to uncover latent factors contributing to disease variability provides deeper insights into disease mechanisms, facilitating improved treatments and interventions.

Limitations of CTL

While we have shown that CTL has numerous benefits to preexisting algorithms and functions, there are some limitations.

- Upstream models have to be able to pass the current data through them, which means that they cannot be completely unrelated - there has to be some shared features to provide a meaningful additional input for the downstream task.
- It has been widely observed that in neural networks, the solution space becomes non-convex [39]. Through CTL we observe that additional context modalities provide more information which then creates the solution space more non-convex. This causes many saddle points which reduces model performance and requires re-runs.

Future Research Directions

Contextualized Transfer Learning (CTL) is opening new doors for researching transfer learning across heterogeneous tasks. While CTL itself is one approach to this research, it also serves as an inspiration for other derivatives and the new notion of learning common heterogeneity across diverse tasks to elucidate latent generation parameters. CTL catalyzes deeper research in the field, with implications in both medical contexts and difficult machine learning tasks in resource-constrained settings.

Locally, we aim to continue our work with CTL by:

- Continuing research to elucidate common differentiating groups in the general Alzheimer's population and Alzheimer's patients with Down syndrome.

- Tuning CTL hyper-parameters further to improve results and make CTL more user-friendly.
- Creating a function to estimate the approximate convex hull coverage.

Conclusions

Contextualized Transfer Learning (CTL) represents a transformative advancement in modeling, particularly beneficial in resource-constrained settings and personalized medicine. By effectively leveraging diverse upstream information, CTL significantly enhances accuracy and offers more personalized, context-specific medical predictions. The application of CTL in clinical settings, as evidenced by the Alzheimer's case study, demonstrates its superiority over traditional models in both accuracy and practical utility, increasing the accuracy of CL by 8% which itself is 30% better than the baseline. Furthermore, CTL's ability to provide insights into disease heterogeneity underscores its potential for improving our understanding and treatment of complex diseases. Continued research and development in CTL are essential to fully realize its benefits and extend its applications across various medical and predictive tasks.

Acknowledgments

First and foremost I would like to thank Dr. Lengerich of the Kellis Lab, (MIT Computer Science and Broad Institute) for mentoring me through this journey. This experience would not have been possible without Prof. Gerovitch, whom I thank sincerely for letting me participate in the MIT PRIMES Program.

References

[1] X Zhang, Y Zhou, et al. “Unraveling patient heterogeneity in complex diseases through individualized co-expression networks: a perspective”. In: *Frontiers in Genetics* 12 (2021), pp. 1234–1245.

[2] Bianca Nogrady. “How cancer genomics is transforming diagnosis and treatment”. In: *Nature* (2020).

[3] Sergios Theodoridis. *Machine learning: A Bayesian and Optimization Perspective.* Cambridge, MA: Academic Press, 2015.

[4] Pengfei Wei and Michael Beer. “Regression Models for Machine Learning”. In: *Computational Methods in Applied Mechanics and Engineering.* N/A, 2023, pp. 113035–113046.

[5] Kamalaker Dadi et al. “Population modeling with machine learning can enhance measures of mental health”. In: *GigaScience* 5 (2023), pp. 30–42.

[6] Ben Lengerich et al. “NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters”. In: *arXiv preprint arXiv:2111.01104* (2021).

[7] Benjamin Lengerich et al. “Contextualized Machine Learning”. In: *arXiv preprint arXiv:2310.11340* (2023).

[8] Charles L. Sprung et al. “How should ICU beds be allocated during a crisis? Evidence from the COVID-19 pandemic”. In: *PLOS ONE* 16.5 (2021), e0250918.

[9] Ezekiel J. Emanuel et al. “A Framework for Rationing Ventilators and Critical Care Beds During the COVID-19 Pandemic”. In: *JAMA* 323.18 (2020), pp. 1773–1774.

[10] Melissa Lamar and Daniel Biel. “Post-mortem tissue: An underutilized resource for studying brain diseases in humans”. In: *Molecular Psychiatry* 25 (2020), pp. 2721–2729.

[11] Sara Mariani et al. “Neurodegenerative diseases: from cell biology to systems biology”. In: *Neuroscience* 414 (2019), pp. 1–15.

[12] Tom B Brown, Benjamin Mann, Nick Ryder, et al. “Language models are few-shot learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017), pp. 5998–6008.

[14] Alec Radford, Jeffrey Wu, Rewon Child, et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[16] Hangbo Bao, Li Dong, and Furu Wei. “BEiT: BERT Pre-Training of Image Transformers”. In: *arXiv preprint arXiv:2106.08254* (2021).

[17] Christopher M Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[18] Benjamin J Lengerich. “Personalized Regression Enables Sample-Specific Pan-Cancer Analysis”. In: *Bioinformatics* (2018).

[19] Ben Lengerich. “Contextualized Machine Learning”. In: *arXiv preprint* (2023).

[20] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2010), pp. 1345–1359.

[21] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems.* 2014, pp. 3320–3328.

[22] Chuanqi Tan et al. “A survey on deep transfer learning”. In: *International conference on artificial neural networks* (2018), pp. 270–279.

[23] Lisa Torrey and Jude Shavlik. “Transfer learning”. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques.* IGI global. 2009, pp. 242–264.

[24] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3.1 (2016), pp. 1–40.

[25] Fuzhen Zhuang et al. “A comprehensive survey on transfer learning”. In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[26] Jing Wang, Rong Jin, and Yang Zhou. “Heterogeneous transfer learning for image classification”. In: *2011 International Conference on Computer Vision* (2011), pp. 1565–1572.

[27] Chenxia Li, Sijia Zhang, and Yong Liu. “Heterogeneous transfer learning for image classification: A survey”. In: *2019 International Joint Conference on Neural Networks (IJCNN).* IEEE. 2019, pp. 1–8.

[28] Rishabh Agarwal et al. “Neural Additive Models: Interpretable Machine Learning with Neural Nets”. In: *arXiv preprint arXiv:2004.13912* (2020).

[29] George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2002.

[30] Prasanta Chandra Mahalanobis. “On the generalized distance in statistics”. In: *Proceedings of the National Institute of Sciences of India* 2 (1936), pp. 49–55.

[31] Maz Jamilah Masnan et al. “Understanding Mahalanobis distance criterion for feature selection”. In: *AIP Conference Proceedings* 1660.1 (2015), p. 050075.

[32] Avrim Blum, Sariel Har-Peled, and Benjamin Raichel. “Sparse approximation via generating point sets”. In: *arXiv preprint arXiv:1712.04564* (2017).

[33] Ananya Kumar. “Streaming Algorithms for Approximate Convex Hulls”. MA thesis. Carnegie Mellon University, 2018.

[34] Evangelos Anagnostopoulos. “Algorithms for Deciding Membership in Polytopes of General Dimension”. In: *Combinatorial Optimization - 5th International Symposium ISCO* (2018).

[35] Young Kim and Hyun Ji Ko. “Biological Markers for Alzheimer’s Disease”. In: *Dementia and Neurocognitive Disorders* 19.3 (2020), pp. 83–91.

[36] Victor L Villemagne et al. “Imaging tau and amyloid- proteinopathies in Alzheimer disease and other conditions”. In: *Nature Reviews Neurology* 17.4 (2021), pp. 225–236.

[37] Philip Scheltens et al. “Alzheimer’s disease”. In: *The Lancet* 388.10043 (2016), pp. 505–517.

[38] Kellis Lab. *Kellis Lab at MIT: Computational Biology and Genomics*. Accessed: 2024-08-17. 2024.

[39] Razvan Pascanu. “On the saddle point problem for non-convex optimization”. In: *arXiv* (2014).

Appendix

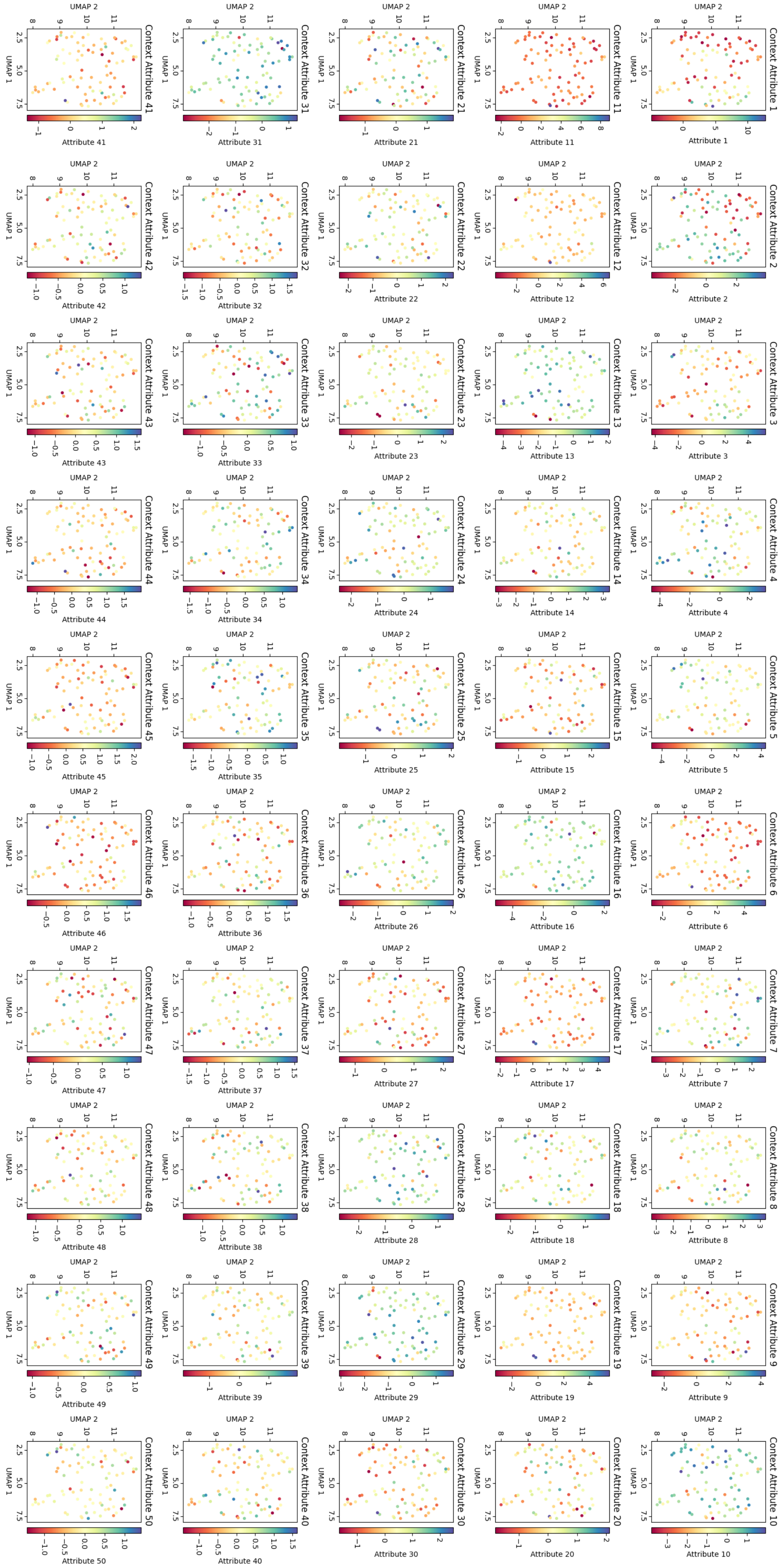


Figure 11: UMAP array for CL showing all 50 context feature's influence on creating sample-specific-models.

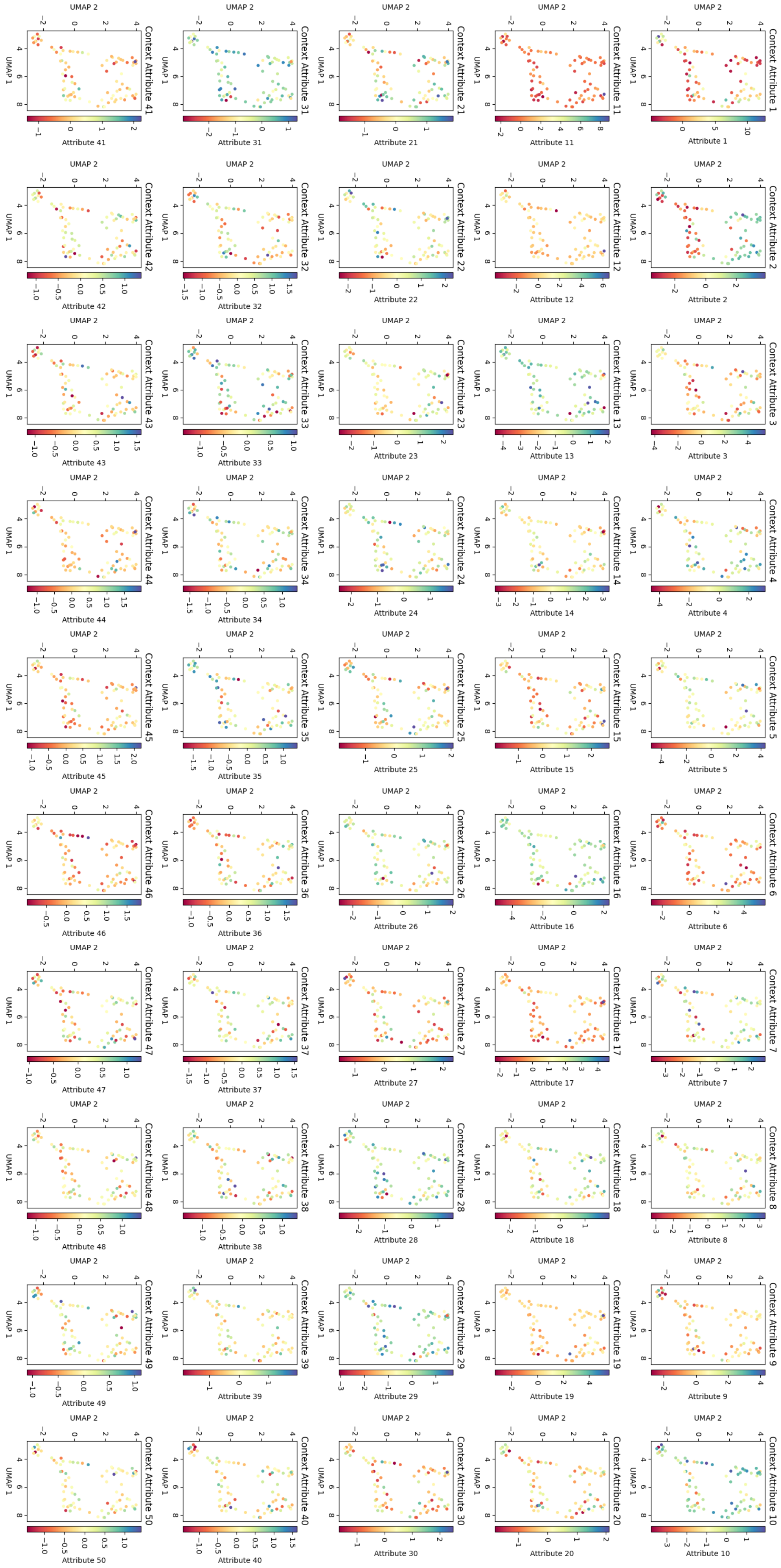


Figure 12: UMAP array for CTL showing all 50 context feature's influence on creating sample-specific-models.