Delft University of Technology

Master's Thesis

MSc Programme in Computer Science

# VAE-based Multivariate Normative Modeling: An Investigation of Covariate Modeling Methods

**Author:**
Remy Duijsens

**Supervisors:**
Dr. C. Lofi (TU Delft)
Dr. H. Schnack (Erasmus MC, Utrecht University)
Dr. R. Muetzel (Erasmus MC)

June 19, 2025

## Abstract

Normative modeling is a promising statistical framework in clinical neuroscience that characterizes individual deviations from population-based reference distributions. While traditional approaches focus on univariate modeling of individual brain measures, multivariate normative modeling using deep generative models, particularly Variational Autoencoders (VAE), has recently emerged as a powerful alternative. These models capture high-dimensional dependencies across brain features and enable the detection of subtle deviations that are difficult to observe with univariate methods. However, current multivariate approaches lack systematic evaluation of covariate modeling methods and remain underexplored in handling batch effects and clinical applicability. For this work, an experimental platform is developed to train and evaluate VAE-based multivariate normative models. The models are trained on structural MRI data from the Generation R Study and the Healthy Brain Network, incorporating key covariates such as age, sex, and acquisition site. A wide set of covariate modeling methods is systematically evaluated in terms of reconstruction quality, latent space covariate invariance, and alignment with normative priors. This work also investigates whether VAE-based multivariate normative models can accommodate batch effects, such as site variation, and compares them to traditional data harmonization techniques, like ComBat. Finally, the model is applied to the clinically relevant task of brain age estimation. The results show that incorporating covariate modeling into the VAE architecture can significantly improve covariate invariance. When examining the influence of batch effects, covariate modeling methods and ComBat data harmonization both reduce site-related information in low-dimensional latent spaces. However, when the latent dimensionality increases, ComBat data harmonization outperforms all covariate modeling methods. In a proof-of-concept application, the model was successfully extended for brain age estimation, capturing age-related deviations while preserving an age-invariant latent space. Altogether, these findings show the potential of VAE-based multivariate normative models for clinical neuroscience applications.

**Keywords:** normative modeling; multivariate normative modeling; variational autoencoder; covariate modeling; batch effects; brain age estimation; Generation R; clinical neuroscience.

# Preface

This TU Delft Computer Science master's thesis is based on an external collaboration with ErasmusMC under the supervision of Hugo Schnack and Ryan Muetzel at the Department of Child and Adolescent Psychiatry/Psychology. This thesis builds upon previous work involving the predictive modeling of various psychological disorders using brain imaging data and research on normative modeling techniques. This effort demonstrates how the intersection of computer science and applied fields, such as predictive behavioral modeling in the medical domain, can lead to valuable theoretical and practical insights. Using recent machine learning techniques, particularly in developing multivariate normative models, this work aims to provide new insights that can improve current clinical practices in normative modeling. In addition to the possible impact of the findings of this work, this collaboration is also a very valuable personal learning experience.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **ACVAE** | Adversarial Conditional Variational Autoencoder |
| **ADHD** | Attention-Deficit/Hyperactivity Disorder |
| **AE** | Autoencoder |
| **BAG** | Brain Age Gap |
| **cVAE** | Conditional Variational Autoencoder |
| **ELBO** | Evidence Lower Bound |
| **ENIGMA** | Enhancing NeuroImaging Genetics through Meta-Analysis (consortium) |
| **GenR** | Generation R |
| **GPR** | Gaussian Process Regression |
| **GRL** | Gradient Reversal Layer |
| **HBN** | Healthy Brain Network |
| **HBR** | Hierarchical Bayesian Regression |
| **HCV** | HSIC-Constrained VAE |
| **HSIC** | Hilbert–Schmidt Independence Criterion |
| **IQ** | Intelligence Quotient |
| **KLD** | Kullback–Leibler Divergence |
| **MAE** | Mean Absolute Error |
| **MI** | Mutual Information |
| **MLP** | Multi-Layer Perceptron |
| **MMD** | Maximum Mean Discrepancy |
| **mmVAE** | Multimodal Variational Autoencoder |
| **MRI** | Magnetic Resonance Imaging |
| **MSE** | Mean Squared Error |
| **NPM** | Normative Probability Map |
| **PCNtoolkit** | Predictive Clinical Neuroscience Toolkit |
| **PNC** | Philadelphia Neurodevelopmental Cohort |
| **RBF** | Radial Basis Function |
| **RF** | Random Forest |
| **SFCN** | Simple Fully Convolutional Network |
| **SMSE** | Standardized Mean Squared Error |
| **SVM** | Support Vector Machine |
| **t−SNE** | t-distributed Stochastic Neighbor Embedding |
| **VAE** | Variational Autoencoder |
| **VFAE** | Variational Fair Autoencoder |

# 1 Introduction

## 1.1 Motivation and Relevance

For a long time, predictive clinical neuroscience research has relied on case-control studies that average data from patient groups and compare them to matched controls. While this approach has shown important group-level differences, it implicitly assumes that individuals sharing a diagnostic label are biologically homogeneous and that the group means can fully represent the disorder in question. Large multisite studies have shown these assumptions to be problematic [1, 2]. Advances in machine learning and the growing demand for individual-level predictions enabled the development of the normative modeling framework. Rather than grouping individuals based on diagnostic labels and examining average group differences, normative modeling assesses individuals against a reference population. This allows for statistical inferences at the individual level, enabling precise characterization of deviations from expected trajectories [3].

In normative modeling, only the 'normal' distribution is modeled, and patients may be detected as outliers. This has long been used in everyday clinical settings, most notably through growth charts for height, weight, and head circumference in pediatrics [4]. Normative modeling has recently gained attention within neuroscience, particularly due to its potential to explain complex brain data. Applying normative modeling to this type of data presents additional challenges, as brain measures are influenced by biological factors such as age, sex, and intelligence. These nuisance factors are often denoted as covariates in this context. If a normative model ignores these sources of variance, it risks misclassifying healthy variation and masking true abnormalities [5, 6]. Consider cortical thickness in the brains of a five-year-old: compared with an adult reference distribution, the child appears to be an extreme outlier. However, relative to age-matched peers, the measurement may fall within the normal range. By explicitly accounting for covariates, normative modeling can isolate nuisance variability and improve the detection of clinically meaningful deviations. Effective covariate modeling is, therefore, a prerequisite for accurate individual-level predictions and remains an active area of research in many domains [7].

Another important factor to consider in normative modeling, beyond biological covariates such as age and sex, is the presence of batch effects. Batch effects are systematic variations that arise due to differences in data acquisition conditions or protocols. In the context of neuroimaging, a notable example is site variance. Site variance is a batch effect that arises from variations across different scanning locations or equipment. Site variance typically combines sample-level discrepancies (e.g., variations in population characteristics, sampling procedures, and inclusion criteria across sites) with measurement-level differences (e.g., differences in scanner models, calibration methods, and acquisition parameters). These sources are often correlated, making it difficult to distinguish their contributions to the total site variance. If these batch effects are left unaddressed, they can introduce systematic noise, which negatively impacts the performance, reliability, and generalizability of normative models across different sites. Therefore, effectively accounting for these batch effects is crucial for providing accurate individual-level predictions in real-world settings.

Traditionally, normative modeling approaches have operated primarily on a univariate level, considering single variables independently [8, 9]. While these methods have proven effective, these univariate models overlook the inherent relationships in more complex data (e.g., multiple brain regions). For instance, accurate estimation of brain age has been shown to benefit from approaches that integrate multiple brain measures, with such models outperforming single-variable approaches [10, 11]. This limitation has generated interest in transitioning towards multivariate normative modeling, which considers multiple correlated variables simultaneously. By leveraging the coherence among multiple brain measures, these models offer a more comprehensive representation of brain patterns, potentially improving the accuracy and robustness of individual-level predictions. Another important benefit of the multivariate approach is that it reduces the impracticality of interpreting numerous individual univariate predictions, thus improving clinical applicability.

Recent advances in machine learning and deep learning, primarily through Variational Autoencoders (VAE), have further accelerated developments in multivariate normative modeling [12, 13, 14, 15, 16]. VAEs are well-suited for capturing normative latent relationships within high-dimensional brain data, providing representations that can capture small biological differences. Recent works using VAE-based normative models have demonstrated promising results in various clinical contexts, such as identifying individual deviations in neurodegenerative diseases like Alzheimer's disease [14].

Despite these advancements, multivariate normative modeling remains a relatively novel field, with many open questions and methodological challenges yet to be addressed. There is a compelling need to further analyze these models, particularly in methods for modeling covariates, accommodating batch effects, and applying the model to more clinically relevant tasks. This thesis benefits from a unique opportunity to use the high-quality neuroimaging and phenotypic data from the Generation R (GenR) cohort. Data that, to our knowledge, has not previously been analyzed within a multivariate normative modeling setting. Using this dataset alongside an open-source dataset, such as the Healthy Brain Network (HBN), provides a good testbed for systematically evaluating normative models.

## 1.2 Problem Definition

Despite promising initial results, current VAE-based normative modeling approaches show several important limitations that withhold broader clinical and scientific adoption. A significant shortcoming is the lack of clear justifications for selecting covariate modeling methods. Recent studies typically condition on covariates such as age, sex, or intracranial volume without systematically comparing or sufficiently explaining their selected modeling methods. Some works fail to specify the method used and only mention that some conditioning on the covariates has been applied. As a result, it remains unclear how effective the presented covariate modeling techniques are.

Furthermore, the ability of current VAE-based models to handle other sources of variability, such as batch effects, remains largely unexplored. Site accommodation has been extensively studied and addressed in univariate normative modeling [17, 9]. However, these insights have not been sufficiently examined in the multivariate case. Effectively accommodating these batch effects is important to ensure robustness, particularly when normative models are intended for clinical use across unseen sites.

Another problem is that evaluations of existing VAE-based normative models lack depth. The analysis focuses predominantly on deviation scores, prediction accuracy, or loss metrics. This does not sufficiently address the models' normative properties and the effectiveness of covariate modeling techniques. Mitigating covariate effects is critical in ensuring that detected deviations genuinely reflect clinically meaningful patterns rather than residual confounding effects due to these underlying biological factors. Equally important, current works rarely systematically assess how varying the latent dimension size affects model performance.

A more practical shortcoming is that the field lacks a structured and standardized evaluation approach tailored to (VAE-based) multivariate normative models. While existing tools provide an extensive experimental platform for univariate normative modeling, a similar platform does not yet exist for (VAE-based) multivariate approaches. The absence of such experimental tools complicates comparisons across studies, hinders reproducibility, and limits the systematic exploration and validation of new types of models.

## 1.3 Research Questions

This research aims to advance the theoretical understanding and practical application of multivariate normative models, specifically using VAEs, by addressing critical gaps in current multivariate normative modeling efforts. Specifically, this thesis systematically evaluates techniques for modeling clinically relevant covariates, such as age and sex, within multivariate normative models. These covariate modeling techniques are based on various approaches previously introduced in the multivariate normative modeling literature and other relevant works. Furthermore, this systematic evaluation explores to what extent these multivariate models can effectively address challenges such as accommodating batch effects. Finally, the thesis compares these multivariate normative modeling methods to each other and state-of-the-art univariate approaches.

The following research questions guide this investigation:

RQ1 **What is the effect of different covariate modeling techniques on the reconstruction, covariate invariance, and normative properties of VAE-based multivariate normative models?**

In univariate normative modeling, adjusting for age, sex, and other biological factors is a well-established practice. Multivariate models face the same challenge but lack a systematic evaluation of how different covariate modeling strategies perform. This thesis compares several established modeling techniques, including explicit conditioning, latent-space disentanglement, and adversarial penalties, across a range of latent dimensionalities. By quantifying their impact on the reconstruction quality, the residual covariate information in the latent space, and the alignment of learned latent distributions with normative expectations, the strengths and limitations of each method are investigated to guide future research and clinical applications.

RQ2 **To what extent can VAE-based multivariate normative models accommodate site variations, and how does this compare to existing data harmonization methods?**

Batch effects pose a significant challenge for normative modeling, especially when deploying models across diverse clinical or research settings. Traditional harmonization approaches such as ComBat [18] correct for these effects by adjusting the data prior to modeling. While effective, these methods require access to all data during preprocessing and risk removing clinically relevant variation if this variation is confounded with the batch effect. This research investigates whether VAE-based multivariate normative models can accommodate batch effects, specifically site variation, without requiring data preprocessing. Three strategies are evaluated: (1) a baseline VAE model without explicit site handling, (2) VAE models extended with covariate modeling techniques applied to the site variable, and (3) traditional preprocessing using ComBat prior to training. All methods are evaluated using a multisite dataset by quantifying their impact on the reconstruction quality, the residual covariate information in the latent space, and the alignment of learned latent distributions with normative expectations. This allows for a systematic comparison of their ability to reduce batch effects while preserving clinically meaningful variation in the learned representations.

RQ3 **How does a VAE-based multivariate normative model compare to an univariate normative model's ability to capture complex normative patterns in brain data, and how does this affect clinical tasks such as brain age prediction?**

Comparing univariate and multivariate normative models is challenging because univariate approaches treat each brain feature independently, whereas multivariate models capture joint patterns across features. To enable a direct comparison, deviation scores from both models are translated into z-scores and summarized per subject. For the univariate model, z-scores are computed per feature and averaged over all features. For the multivariate VAE, z-scores are computed per latent dimension and averaged over the latent space size. The average z-score per subject reflects the deviation from the normative population. Multivariate normative models offer a significant advantage by modeling dependencies between features. This ability is particularly important in clinical tasks where deviations across brain regions may be diagnostically relevant. Brain age prediction is a task that greatly benefits from this coherent representation. A proof-of-concept is presented in which the multivariate VAE-based model is applied to brain age prediction. This model is trained to predict brain ages within a biologically plausible range.

## 1.4 Contributions

This thesis makes the following contributions:

1. An experimental framework for (VAE-based) multivariate normative models.
2. A VAE-based multivariate normative model that encodes high-dimensional brain measure data trained using the GenR and HBN datasets.
3. An evaluation of the effect of different covariate modeling techniques in VAE-based multivariate normative models (addresses RQ1).
4. An evaluation of the ability of multivariate normative models to accommodate site variation compared to traditional harmonization techniques (addresses RQ2).

5. A comparison of current state-of-the-art univariate models to the VAE-based multivariate model in terms of deviation scores and reconstruction quality (addresses RQ3).

6. The application of this model to a clinical task, namely brain age estimation (addresses RQ3).

## 1.5 Structure Overview

The structure of this thesis is as follows. Section 2 provides an overview of related work, establishing the context and current state-of-the-art within the field of normative modeling. First, the concept and motivation for normative modeling are introduced, followed by the multivariate extension of normative modeling, which includes recent advancements using VAEs. Furthermore, covariate encoding and modeling techniques from recent literature are discussed. Finally, the brain age estimation task is described.

Section 3 outlines the methodological framework used in this work. The datasets and preprocessing procedures are first described, followed by the proposed model architecture and the examined covariate modeling techniques. Subsequently, the evaluation metrics and the experimental setup are presented.

In Section 4, the outcomes of the experiments are reported. Section 5 provides an interpretation and analysis of the results, including a reflection on the study's limitations. Finally, Section 6 summarizes the main findings and contributions of this thesis and discusses potential directions for future research.

# 2 Background

## 2.1 From Case-Control to Normative Thinking

For a long time, predictive clinical neuroscience research has been dominated by case-control studies that average patient data and compare it to a matched control group. Although statistically convenient, this methodology rests on two strong assumptions: (1) that individuals sharing a diagnostic label are biologically homogeneous, and (2) that the mean of the patient group is an accurate representation of the disorder. Large multi-site studies have now shown that neither assumption holds. For example, meta-analyses of brain volume differences in schizophrenia and bipolar disorder reveal inconsistent regional abnormalities across cohorts, with many effects failing to replicate when a single site is excluded [1, 2]. Similar results are observed in attention-deficit/hyperactivity disorder (ADHD), major depression, and autism spectrum disorder. These findings highlight a significant limitation of group averages, namely that they do not account for clinically relevant individual variability. Figure 1 illustrates this limitation. It shows that clinical populations often do not fit the clean separation assumed by classical case-control designs. Patient data may have overlapping distributions, hidden subgroups, or even variations that also exist in healthy individuals. These patterns show the need for individual-level modeling approaches. A wide array of machine learning strategies has been developed to address this heterogeneity. As reviewed by Schnack [19], researchers have investigated methods ranging from nonlinear transformations and kernel-based classifiers to clustering-driven and generative approaches.

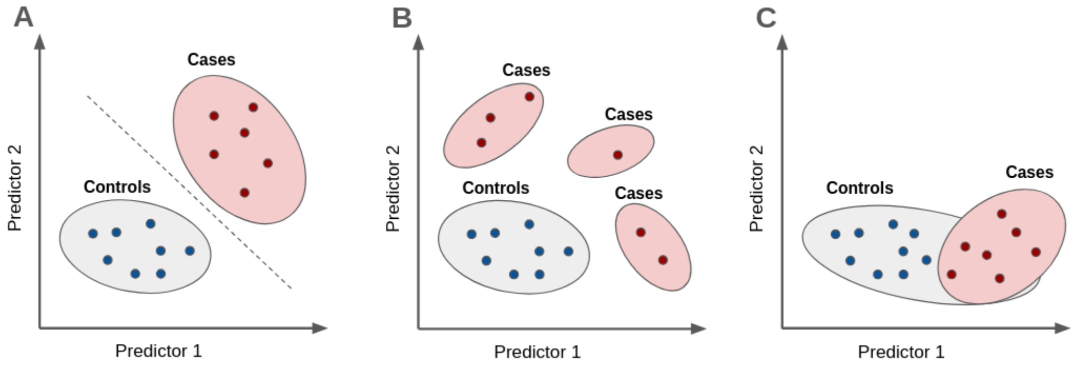

Figure 1: Illustration of different assumptions in case-control studies. (A) The classic case-control approach assumes that cases and controls form separated groups. (B) The case group may consist of multiple subgroups, each with its own pattern. (C) Deviations might not be exclusive to the case group but can also exist within healthy variation.

**Normative modeling** emerged as a response to case-control studies by reframing the analysis around the individual. Instead of asking *"How does the average patient differ from the average control?"* the question becomes *"How does a specific individual deviate from what is expected in the healthy population, given their age, sex, and other covariates?"*. Here, only the 'normal' distribution is modeled, and patients may be detected as outliers. From a modeling perspective, the earliest attempts translated this question into an outlier–detection problem. Mourão-Miranda et al. trained a one-class support vector machine (SVM) on feature vectors extracted from control fMRI scans. Here, the SVM learned a decision surface that captures normal variation, and patients were flagged whenever their distance to this surface exceeded a threshold [20]. Although conceptually simple, the approach produced only a binary abnormal/normal label and could not explain the features driving the decision.

Rezek and Beckmann took a different, more statistical view of the problem. They modeled the full probability density of healthy variation and treated disease as the low-probability tails of that density [21]. Their work framed the problem in the language of generative modeling. Although the paper did not include an empirical implementation, it described that a normative model should output a complete predictive distribution, not just a point estimate.

The first large-scale empirical implementation was introduced by Ziegler et al. with their voxel-wise Gaussian Process Regression (GPR) [22]. GPR is a non-parametric Bayesian regressor that returns, for each covariate vector $x$, both a mean prediction $\mu(x)$ and a predictive variance $\sigma^2(x)$. By fitting an independent GPR at every voxel, the authors could transform a new brain scan into a three-dimensional z-score map, which they referred to as a Normative Probability Map (NPM). The key modeling insight was that the kernel trick underlying GPR enables the learning of complex, nonlinear covariate effects. However, the approach was computationally intense and required large, healthy datasets for training.

An early adaption to these fully probabilistic frameworks was the study by Erus et al. [23], who analyzed structural MRI data from a large youth cohort (PNC) to map normative developmental trajectories of brain volumes and their relationship to cognitive performance. By applying mixed-effect regression models to a large, demographically diverse dataset, they generated reference curves for brain maturation across age. Individual deviations from these curves were then linked to cognitive metrics, demonstrating the clinical utility of subject-specific estimates. The work laid an important foundation by showing that deviations from typical developmental trajectories, rather than group averages, are useful in understanding individual cognitive function.

Taken together, these studies established three modeling principles that remain fundamental:

1. **Reference distribution** Train on a demographically diverse set of healthy subjects so that the learned representation spans the (covariate) space encountered at test time.

2. **Probabilistic output** Use probabilistic algorithms like SVMs, density estimators, Bayesian regressors, or neural networks that can return a predictive distribution for accurate deviation scores instead of hard labels.

3. **Subject specific deviation metric** Compute an easily interpretable statistic (e.g., distance from the class boundary, negative log-likelihood, or standardized residual) that can be used in clinical classification and prediction tasks.

Marquand and colleagues have integrated these principles into a coherent statistical framework. The following section examines that framework in detail.

## 2.2 Univariate Normative Modeling

Univariate normative modeling is a framework for quantifying individual deviations in single features relative to a healthy population distribution. In this approach, population-level trajectories of a brain measure (e.g., the volume of a brain region or voxel-wise cortical thickness) are modeled as a function of important covariates (e.g., age, sex), as shown in Figure 2A. Marquand et al. formalized this framework by generalizing the concept of pediatric growth charts to neuroimaging data. Comparable to plotting a child's height against age to assess growth, a brain feature can be mapped against relevant covariates to establish its normative range across the population [8, 1]. Importantly, this enables probabilistic inference at the individual level. Each individual can be assigned a personalized deviation score (e.g., a z-score or outlier probability).

Mathematically, a normative model estimates $p(\mathbf{y} \mid \mathbf{x})$, the full conditional distribution of neuroimaging features $\mathbf{y}$ given covariates $\mathbf{x}$ (e.g., age, sex, site). The individual deviation is then quantified as a centile score or $Z$-value reflecting the probability of observing $\mathbf{y}$ under this distribution. A basic schematic of a univariate normative model is shown in Figure 3. By working at the individual level, normative modeling addresses biological heterogeneity and provides a new route toward predictive medicine. To demonstrate this, Marquand et al. applied GPR to a large healthy sample to model the relationship between trait impulsivity and reward-related brain activation [8]. This allowed them to identify individuals who deviated significantly from the normative range. These deviations occurred even in participants without a formal diagnosis. This demonstrated that symptoms can be mapped to individuals independently of their categorical clinical labels. Rutherford et al. have taken these ideas and materialized them through open-source tools. The Predictive Clinical Neuroscience Toolkit (PCNtoolkit) is a framework with a set of normative algorithms that can be used for fitting normative models and computing subject-specific deviation scores [9]. Furthermore, it provides a comprehensive methodology, spanning data preprocessing, evaluation, and visualization.

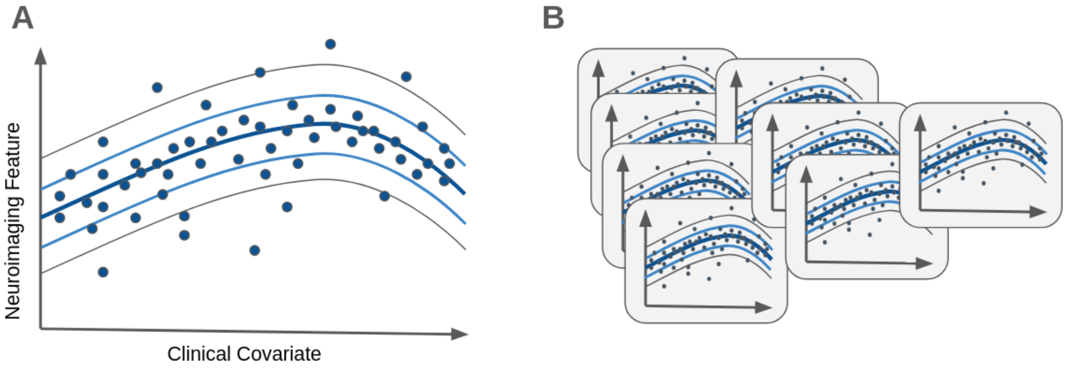

Figure 2: Illustration of univariate normative models. (A) A typical normative curve mapping clinically relevant covariates to biological features. (B) Highlighting the necessity to interpret many univariate normative models when multiple biological features are considered.

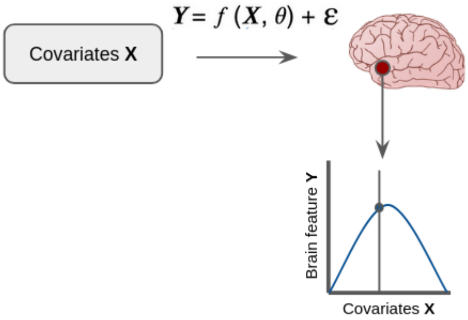

Figure 3: A single brain feature $\boldsymbol{Y}$ is modeled as a function of covariates $\boldsymbol{X}$, such as age or sex, using a regression function $\boldsymbol{Y} = f(\boldsymbol{X}, \Theta) + \epsilon$, where $\Theta$ are the model parameters, and $\epsilon$ are residuals. The output of an individual is then compared against the estimated normative distribution to compute a subject-specific deviation score.

Recent evidence highlights the benefits of normative modeling in different predictive tasks. Rutherford et al. demonstrated that individual deviation measures derived from univariate normative models outperform the original imaging features in various predictive contexts [3]. In their benchmarks using structural MRI and functional MRI data, normative predictions showed improved performance for clinical classification (distinguishing patients with schizophrenia from healthy controls) and regression (predicting continuous cognitive scores).

However, univariate normative models have important limitations that limit their usability for complex neuroimaging data:

- **Lacks representation of feature coherence:** Modeling one feature at a time ignores important dependencies and covariance between features. Brain measures often vary in correlated ways (e.g., coordinated atrophy across regions). However, univariate models cannot detect when an individual's multivariate pattern of deviations is abnormal since each feature's deviation is assessed in isolation. In other words, the joint distribution of the data is factored into separate univariate distributions, losing information.
- **Limited model complexity:** While univariate models can use nonlinear regression, they ultimately map a low-dimensional covariate space to a single output. Computational considerations have led to these relatively simple normative models (e.g., linear or Gaussian process) per feature. They struggle to capture complex patterns and higher-order effects in high-dimensional neuroimaging data.

- **Clinical Interpretability:** As illustrated in Figure 2B, interpreting univariate normative models becomes increasingly impractical as the number of features grows. Each feature requires its own independent model, leading to a fragmented interpretation of an individual's deviation from the norm. Interpreting these deviations across dozens or even hundreds of isolated models in clinical settings can be time-consuming, error-prone, or even undoable. More importantly, it becomes difficult to grasp the overall pattern of deviation, which is essential for meaningful clinical interpretation.

These shortcomings motivate the development of multivariate normative modeling approaches. Multivariate methods aim to account for feature coherence by modeling multiple brain features jointly and identifying deviation patterns that univariate models may miss. This could benefit tasks that rely on these coherent brain patterns, such as estimating brain age. The following section will explore recent efforts in multivariate normative modeling.

## 2.3 Multivariate Normative Modeling

Traditional normative models treated each biological feature independently by modeling a univariate distribution for each feature. This univariate approach fails to capture the coherence between many biological features (e.g., different brain regions) and may overlook subtle but coherent deviations [15]. For example, in disorders, pathological changes present themselves as distributed patterns. Here, several connected brain regions, each showing mild atrophy, might collectively represent an abnormal profile even if no single region is an extreme outlier. This motivated a shift beyond univariate norms to truly multivariate normative models that learn the joint distribution of brain features.

Autoencoders were a natural choice for modeling the high-dimensional joint distributions of brain data. An autoencoder is an unsupervised neural network that compresses data into a low-dimensional code (encoder) and then reconstructs the data from this code (decoder). When trained on a large healthy control cohort, an autoencoder learns to accurately reproduce normative brain patterns while implicitly modeling their covariance structure. Individuals with unusual brain patterns (e.g., patients) reconstruct poorly, resulting in larger errors that signify a deviation from the norm. This framework was first applied to brain MRI data by Pinaya et al. [15], who trained a deep autoencoder on healthy subjects to derive a normative model of structural brain features. They then used this model to evaluate healthy individuals and patients with schizophrenia and autism. The autoencoder-based normative approach successfully detected individual deviations, giving each individual a deviation score based on reconstruction error or latent distance. Notably, their model revealed distinct patterns for schizophrenia and autism consistent with known disease-specific patterns. Pinaya et al. [16] extended this work to neurodegenerative diseases and tested its generalizability across datasets. The model performed consistently across external cohorts, demonstrating generalization capabilities comparable to those of traditional supervised classifiers. This initial work demonstrated the ability of deep autoencoders to capture deviations from the norm that traditional univariate methods would miss.

More recent work built on this idea using VAEs instead of regular Autoencoders. These models added a probabilistic latent space to the autoencoder framework. VAEs learn a low-dimensional latent representation $z$ of the data that is constrained by a prior distribution (typically $z \sim \mathcal{N}(0, I)$). The VAE objective combines a reconstruction term with a Kullback–Leibler (KL) divergence regularizer $D_{\mathrm{KL}}(q_\phi(z \mid x) \,\|\, p(z))$ that penalizes deviations of the learned posterior $q_\phi(z \mid x)$ from the chosen prior $p(z)$. A schematic of a VAE-based multivariate model is shown in Figure 4. In a normative modeling context, this means the latent codes of healthy brains are enforced to follow a standard Gaussian distribution. The normative prior acts as a regularizer, preventing overfitting to individual abnormalities and providing a reference distribution of the input data. This VAE framework inherently provides each individual's deviation scores (e.g., distance measures in latent space).

Aguila et al. [13] introduced a conditional VAE (cVAE) framework to model covariate effects within normative modeling explicitly. In their approach, the autoencoder's encoding and decoding are conditioned on variables such as age so that age-related variation is accounted for in the normative model. This prevents typical aging effects from being misrepresented in what the model considers pathological deviations. They also proposed a latent-space deviation metric. Rather than using reconstruction error alone, they computed each subject's deviation as a standardized distance in the latent space of the

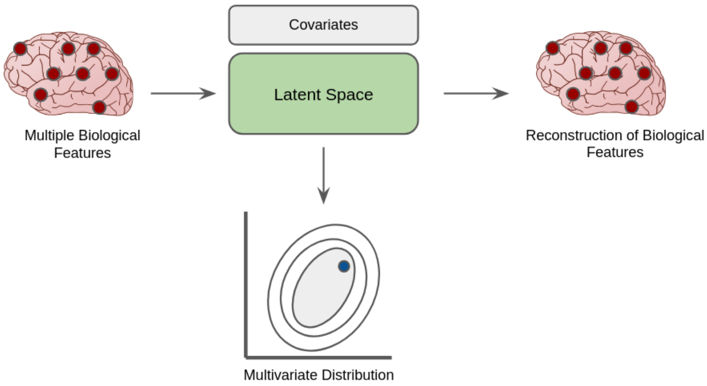
Covariates
Latent Space
Multiple Biological Features
Reconstruction of Biological Features
Multivariate Distribution

Figure 4: Multiple biological features are encoded into a latent space conditioned on covariates. The learned latent representation is regularized to follow a multivariate Gaussian prior, enabling the model to learn normative variation. The latent code is used to reconstruct the input, and deviations from the normative latent distribution indicate abnormality.

VAE. This method successfully identified patients as outliers and correlated with disease severity across heterogeneous datasets.

Wang et al. [24] proposed an adversarial conditional VAE (ACVAE) to improve sensitivity to disease-related deviations. They combined a cVAE with an adversarial training objective that encourages the latent space to be discriminative yet generalizable. In experiments on an Alzheimer's dataset, this ACVAE produced more sensitive deviation maps and better distinguished patients from controls than standard VAEs.

Kumar et al. [25, 26] extended this work to multimodal data by developing a multimodal VAE (mmVAE) to capture the joint distribution of T1-weighted and T2-weighted MRI features. Their model demonstrated improved detection of Alzheimer's disease by integrating multiple data sources, revealing patterns that univariate models missed, and achieving a better correlation with clinical severity scores.

Even though this progress is promising, several challenges remain. First, it remains unclear how to effectively model all relevant covariates and ensure that normative scores accurately reflect actual variations rather than demographic ones. Second, accounting for batch effects remains an issue that warrants further study, as many models are trained on multi-site datasets in practice. The following sections provide additional background on current efforts on these issues.

## 2.4 Covariate Effect Modeling

In normative modeling, covariates such as age, sex, or intelligence quotient (IQ) introduce variability in the training data. If not properly accounted for, they can bias the model, hide true deviations, and reduce the accuracy and interpretability of individual predictions [5, 6]. Therefore, properly encoding and integrating covariates into the modeling process is crucial for creating normative models that accurately capture individual deviations from the norm.

For quantitative covariates, z-score standardization is a widely adopted strategy in deep learning [27]. While continuous encoding supports smooth modeling of effects, alternative strategies exist. For instance, some studies, including Aguila et al. [13] and Wang et al. [24], discretize age into bins based on quantiles and then apply one-hot encoding. Although this approach captures nonlinear effects in a piecewise manner, it introduces artificial thresholds and may sacrifice predictive granularity and generalizability [7].

Qualitative covariates, such as sex or site, are typically encoded using one-hot encoding. This method avoids imposing any artificial order among categories and allows models to differentiate between groups without implying specific distances between them. One-hot encoding is effective when the number of categories is small. However, when dealing with sparse inputs and high-cardinality categorical variables, such as datasets with dozens of scanning sites, one-hot encoding becomes inefficient due to the resulting high-dimensional feature space. In these cases, learned embedding vectors are an alternative. Guo and Berkhahn [28] demonstrated that entity embeddings of categorical variables can capture latent similarities between categories, improving model performance while simultaneously reducing dimensionality.

Once covariates have been encoded into a numerical representation, they need to be incorporated into a multivariate normative model. VAEs have become a leading tool in multivariate normative modeling research. However, a standard VAE does not inherently distinguish between variation due to covariates and other sources of variability. If covariate effects are not explicitly modeled, the latent variables may entangle those effects. This entanglement could significantly hinder capturing patterns of interest. For example, a VAE trained on brain data without incorporating age information could misclassify normal age-related changes as anomalies, as brain structure varies systematically with age. To address this, several approaches have been proposed to integrate covariate information into VAE-based models. The simplest method is architectural conditioning, whereby the covariate vector $c$ is concatenated to the input of the encoder, decoder, or both. For example, Aguila et al. [13] conditioned both the encoder (by inputting $[x, c]$) and decoder (by inputting $[z, c]$), allowing the VAE to model the conditional distribution $p(x|c)$ implicitly. While encoder and/or decoder conditioning is straightforward, it does not impose any explicit constraint on the relationship between the latent representation $z$ and $c$. Consequently, the latent space $z$ may still carry information about the covariates if doing so improves reconstruction. In particular, encoder-only conditioning may encourage the encoder to embed covariate effects into $z$, thereby facilitating accurate reconstruction. Decoder-only conditioning, where $c$ is provided only to the decoder, may partially decouple $z$ from $c$ by allowing the decoder to reconstruct covariate-related variation directly, but without guarantees of independence. The most common implementation feeds $c$ to both encoder and decoder, which is also known as a cVAE architecture. Several studies, including Aguila et al. [13] and Wang et al. (2023) [24], have demonstrated that conditioning on covariates using a cVAE improves the detection of disease-related abnormalities compared to an unconditioned VAE. However, these works incorporate covariates only through architectural conditioning. In both studies, the covariate vector $c$ is concatenated to the encoder and decoder inputs, while the optimization objective remains the standard VAE loss. Because the loss contains no term that penalizes statistical dependence between the latent variables $z$ and $c$, residual covariate information can still leak into $z$ and restrict full disentanglement. Research in nuisance factor invariant latent space representation shows that adding an independence regulariser to the loss, such as a maximum mean discrepancy penalty [29] or an adversarial discriminator [30], pushes the latent space toward stronger covariate invariance. Whether these ideas also translate to normative models remains an open question that needs testing.

Despite the presence of some forms of covariate modeling, the literature lacks systematic evaluations of different approaches. Furthermore, the currently used methods have been introduced across different datasets and experimental settings, making direct comparisons difficult. Moreover, some studies, such as Kumar et al. [26], mention covariate modeling without clearly specifying how covariates are encoded or incorporated, hindering reproducibility and interpretation. This thesis systematically compares and evaluates different covariate modeling strategies for VAE-based normative modeling.

## 2.5 Accommodating Batch Effects

Batch effects are systematic differences arising from variations in data measurement conditions. These effects can pose a significant challenge in biomedical studies. In neuroimaging, one common form of batch effect is site variance, arising from differences in data collection locations or scanner types, which introduces systematic noise that can degrade model performance and challenge the robustness of normative modeling.

A common strategy is pre-harmonizing data using methods like ComBat as a data preprocessing step [18]. However, it requires access to all data at training time and can remove unknown clinically relevant variance if that variance is confounded with site [31, 32]. In normative modeling, batch effects are typically incorporated into the model rather than removed a priori. For example, the PCNtoolkit implements HBR, a univariate normative modeling algorithm that can account for batch effects [33]. HBR delivers site-agnostic deviation scores while preserving all shared variance and supports transfer to new sites via hierarchical priors. However, it can be computationally intensive and requires careful specification of priors and model structure [34, 33].

Deep generative models such as variational autoencoders (VAEs) offer an alternative multivariate approach by capturing joint patterns across features. Deep learning methods have delivered robust results in modeling multi-site brain imaging data [35], raising the question of whether site variance can be implicitly accounted for within VAE-based normative models. Current multivariate normative modeling studies have not explicitly addressed batch effects. However, VAEs have been applied successfully in harmonization tasks outside normative modeling. DeepComBat implements a conditional VAE to disentangle and correct site-specific biases in imaging features [36]. Similarly, style-transfer VAEs and deep residual networks have been used to harmonize MRI features across sites and scanners [37]. These studies show that deep generative models have the ability to learn representations agnostic to batch effects.

## 2.6 Brain Age Estimation

One prominent application of multivariate normative modeling is the estimation of brain age. Brain age estimation uses machine learning techniques to predict an individual's brain age from biological data (e.g., neuroimaging data). Chronological age refers to a person's actual age in years, whereas predicted brain age is the age estimated by the model based on biological features (e.g., MRI scans). The difference between predicted and chronological age, termed the brain age gap (BAG), is defined as:

$$BAG = \text{Predicted Brain Age} - \text{Chronological Age}. \tag{1}$$

A positive BAG indicates accelerated brain aging, whereas a negative BAG suggests a held or delayed aging process. A large body of evidence supports brain age as a biomarker of brain health. Individuals with higher positive BAG tend to show worse cognitive performance and an increased risk for neurodegenerative diseases. In aging cohorts, increased BAG has been associated with advanced physiological aging and a higher mortality risk [38, 39]. Contrarily, a lower (negative) BAG may reflect resilience or other protective factors in brain aging. Clinical studies using brain age estimation have shown deviations in different disorders. Patients with Alzheimer's disease, mild cognitive impairment, schizophrenia, and traumatic brain injury often show elevated BAG relative to healthy, age-matched controls [40, 41, 42]. In line with this work, Brouwer et al. applied brain age prediction to a healthy population within the same age range as the GenR and HBN cohorts. They reported a sex difference of approximately one year in BAG, with males showing higher BAG than females [43].

Model evaluation in brain age studies uses the Mean Average Error (MAE) between predicted and true age as the primary accuracy metric. Pearson correlation (r) and coefficient of determination (R$^2$) assess rank-order preservation and explained variance [44]. The standard deviation of BAG in normative samples quantifies typical variability and helps define clinically significant deviations. Regression analyses of predicted true age are used to evaluate bias (with an ideal slope of 1 and intercept of 0). Researchers also examine whether BAG correlates with chronological age, which is a sign of residual bias, and whether the BAG distribution in healthy controls centers on zero. Ideally, the BAG and chronological age correlation should also be close to zero in the training sample or a healthy test sample, although the latter is not known in advance. The reason for this is that BAG is designed to be age-independent in normative populations. Once clinical or non-normative samples, such as elderly individuals with dementia, are considered, BAG and age can be correlated.

Early brain age models used univariate biological features such as a specific brain region, total brain volume, or mean cortical thickness, which capture broad aging trends. These univariate models learned a normative aging curve where an individual's BAG represents how much their brain deviates from that curve to their chronological age [38, 45]. Sun et al. further demonstrate that predicted ages for different brain regions can be compared to investigate correlating patterns of developmental trajectories across brain networks [45].

Advancements in data availability and algorithmic methods have significantly improved prediction accuracy. Traditional models with single features achieved MAE of approximately 5–6 years in adult cohorts [38, 46]. Deep learning approaches have reduced MAE to around 2–4 years. For example, Bashyam et al. trained a 3D convolutional network on MRI scans and reported an MAE of 3.5 years [35]. Peng et al. introduced a lightweight convolutional network based on a Simple Fully Convolutional Network (SFCN) that achieved an MAE of 2.1 years on the UK Biobank dataset, including techniques such as data augmentation, model ensembling, and bias correction to optimize performance [47].

Multivariate normative modeling has great potential for brain age estimation. Sun et al. implemented univariate normative BAG trajectories for individual brain regions. However, this univariate approach does not capture the coherence between multiple brain features [45]. Deep learning models can use high-dimensional inputs to achieve superior MAE (2–4 years) yet lack an explicit normative framework for interpreting individual deviations [35, 47]. A multivariate normative model could, in theory, jointly estimate normative aging distributions across all features, combining the best aspects of univariate normative trajectories with the state-of-the-art predictive accuracy of deep learning approaches.

# 3 Methodology

## 3.1 Model Architecture

This section describes the construction of a VAE-based multivariate normative model. First, a VAE-based model is defined and optimized to establish a stable and effective baseline. This model serves as the foundation for all subsequent models and experiments. Next, several covariate modeling techniques are introduced to account for the influence of present covariate information. These methods modify different parts of the baseline model (e.g., the loss function). Finally, a proof-of-concept extension of the model for brain age estimation is presented, where the VAE is adapted to estimate a brain age gap in a normative modeling context.

### 3.1.1 Baseline Model

The baseline multivariate normative model is implemented as a VAE that directly learns the representation of brain features without explicitly modeling covariates. Figure 5 shows a schematic diagram of the baseline model. The encoder and decoder are both fully connected (feedforward) neural networks. Key architectural components include:

- **Input Vector** $(\mathbf{x})$: Standardized brain feature vector.
- **Reconstruction Vector** $(\mathbf{x}')$: Reconstructed brain feature vector.
- **Latent Space Dimensionality** $(d)$: The size of the latent space.
- **Encoder** $(f_\phi^{\mathrm{enc}})$: Maps $\mathbf{x}$ to the parameters $(\mu_\phi, \sigma_\phi)$ of a standard normal distribution from which the latent representation $\mathbf{z} \in R^d$ is sampled using the reparameterization trick.
- **Decoder** $(f_\theta^{\mathrm{dec}})$: Maps $\mathbf{z}$ to a reconstruction $\mathbf{x}' = f_\theta^{\mathrm{dec}}(\mathbf{z})$, typically representing the mean of a Gaussian output distribution.
- **Prior Assumption** $(p(\mathbf{z}))$: A fixed standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in $R^d$ based on the assumption that the biological input data has such a distribution.

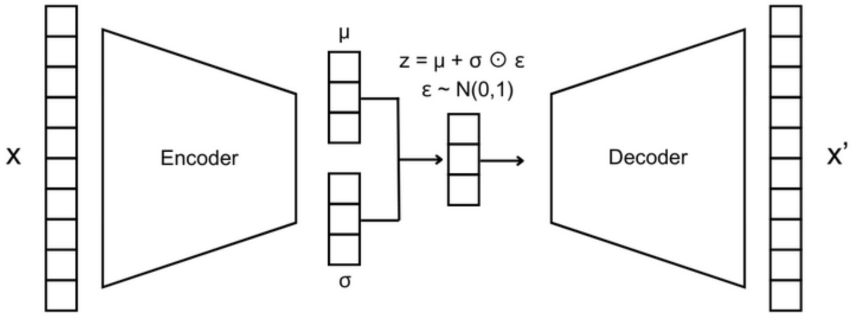

Figure 5: Schematic diagram of the baseline multivariate normative model. The model is a variational autoencoder that maps the input feature vector $\mathbf{x} \in R^d$ to a latent representation $\mathbf{z} \in R^k$ using an encoder network. The decoder reconstructs the input as $\hat{\mathbf{x}}$ from the latent code. The latent space is regularized to follow a standard normal prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. No covariates are used in this baseline model.

Training minimizes the negative Evidence Lower Bound (ELBO). For each input $\mathbf{x}$, the loss function consists of a reconstruction term and a regularization term via KL divergence:

$$\mathcal{L}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \beta\, D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})),$$

where:

- $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ is the mean-squared reconstruction error.
- $D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z}))$ is the Kullback-Leibler divergence between the approximate posterior and the standard normal prior.
- $\beta$ is a weighting coefficient that controls the relative importance of the KL term in the loss function.

The model is trained for 200 epochs using stochastic gradient-based optimization to minimize $\mathcal{L}(\mathbf{x})$. A KL annealing schedule is applied to stabilize training and encourage effective representation learning. The coefficient $\beta$ is linearly increased from $0.1$ to $1.0$ between epochs 10 and 60. The schedule results in the following training phases:

- Epochs 0–9: Low KL weight ($\beta = 0.1$) prioritizes reconstruction, allowing the model to learn an initial encoding.
- Epochs 10–60: Gradual increase in $\beta$ introduces stronger regularization, guiding the latent distribution toward the prior.
- Epochs 61–200: Full KL regularization is enforced with $\beta = 1.0$, encouraging the latent space to form a standard normal distribution.

The training strategy first prioritizes accurate reconstruction, then progressively enforces alignment with the prior distribution. As training progresses, increasing $\beta$ shifts the focus to regularization, ensuring that the latent space adheres to the standard Gaussian prior. By the end of the training, the latent space should be structured to follow $\mathcal{N}(\mathbf{0}, \mathbf{I})$. New samples whose latent codes deviate significantly from this distribution can be identified as outliers or atypical instances relative to the learned normative model.

A hyperparameter optimization was conducted to select model parameters, such as the number of layers and network depth, to improve the baseline model's performance. The hyperparameter study used a latent space dimensionality of $d = 8$. The hyperparameter search was performed using Optuna [48]. The following search space is explored:

- **Batch size**: {8, 16, 32, 64, 128}
- **Normalization layers**: {False, True}
- **Gradient clipping**: {False, True}
- **Optimizer**: {Adam, AdamW}
- **Activation function**: {SiLU, ReLU, LeakyReLU}
- **Network depth**: {2, 3, 4} hidden layers
- **First hidden layer size**: {512, 256, 128, 64, 32}
- **Learning rate**: {0.01, 0.001, 0.0001}
- **Learning rate scheduler**: {Linear step, Cosine annealing, Reduce-on-plateau}

The size of the first hidden layer, in combination with the selected network depth, defines the architecture of both the encoder and the decoder. Subsequent hidden layers are constructed by recursively halving the size of the previous layer. Other hyperparameters, such as the choice of activation functions and optimizers, are commonly found in popular VAE model architectures.

Each trial was trained for a fixed number of 500 epochs to ensure convergence. The best-performing configuration determined with this search was:

- **Batch size**: 32
- **Normalization layers**: None

- **Gradient clipping**: Not applied
- **Optimizer**: AdamW
- **Activation function**: SiLU (Sigmoid Linear Unit)
- **Network architecture**: Three hidden layers with sizes 256, 128, and 64.
- **Learning rate**: 0.001
- **Learning rate scheduler**: Reduce-on-plateau, configured as:
  - Mode: `min`
  - Factor: 0.5 (reduce learning rate by half)
  - Patience: 10 epochs (waiting period before reduction)

The selection of the AdamW optimizer is supported by its improved generalization performance over the original Adam optimizer, achieved by decoupling weight decay from the gradient update process [49]. The SiLU activation function, also known as the Swish function, has been shown to outperform traditional activation functions such as ReLU in various deep-learning settings, including specific VAE configurations. Its smooth, non-monotonic properties facilitate better gradient flow and convergence during training [50]. The complete architecture, featuring three hidden layers of sizes 256, 128, and 64 for both encoder and decoder, has a total of 100452 learnable model parameters.

### 3.1.2 Covariate Modeling Methods

This work considers three covariates, namely age, sex, and site. These covariates are used to evaluate different covariate modeling methods. Additionally, the site variable is also treated as a batch effect and plays a central role in the experiments concerning the accommodation of site variance.

Covariates are numerically represented in a form suitable for inclusion in the statistical model while preserving their semantic meaning. This work handles covariates as follows:

- **Age** Chronological age is $z$–score normalized to zero mean and unit variance, aligning its scale with that of the other network inputs and stabilizing gradient-based optimization in the VAE loss [51]. Alternative encodings reported in the literature, such as quantile binning followed by one-hot vectors [13, 24], are not employed in the models because they introduce arbitrary thresholds and discard the natural ordering of age.
- **Sex and Site** Given the low cardinality of these variables (two sexes, one site in GenR, and three sites in HBN), one-hot encoding is used. One-hot encoded vectors avoid implied ordinality and do not explode the feature space when the cardinality is low.

The covariates are represented by the covariate vector $\mathbf{c}$. The covariate vector is made available to the network, where it can be used according to the needs of the specific covariate modeling method. Depending on the approach, $c$ may be concatenated to the encoder input, to the decoder input, provided as additional latent nodes, or used within penalty terms in the loss function designed to encourage covariate invariance. All modeling methods are applied on top of the previously defined baseline model.

The investigated modeling methods include those used in prior work on VAE-based normative models and popular methods for creating covariate-invariant latent spaces in VAEs. Furthermore, some methods that explicitly force the model to learn covariate structure in the latent space are included to test whether the model behaves in line with the intuition that this increases the dependency on the covariates. To enable clear and consistent referencing throughout this thesis, each covariate modeling method is assigned a unique identifier using the naming convention `CM-<ID>`, where "CM" stands for Covariate Method. These identifiers are used in subsequent sections (e.g., evaluation and results) to refer to each method. The following covariate modeling techniques are evaluated in this work:

- **CM-1: Decoder-Only Modeling** In this method, the covariates are included only during the decoding phase. The architecture, shown in Figure 6, is as follows:

$$\mathbf{x} \rightarrow \text{Encoder} \rightarrow \mathbf{z}, \quad \{\mathbf{z}, \mathbf{c}\} \rightarrow \text{Decoder} \rightarrow \mathbf{x}'$$

  The idea is that the latent representation $\mathbf{z}$ should become independent of the covariates, as the decoder directly receives the covariate information and does not require $\mathbf{z}$ to encode it.

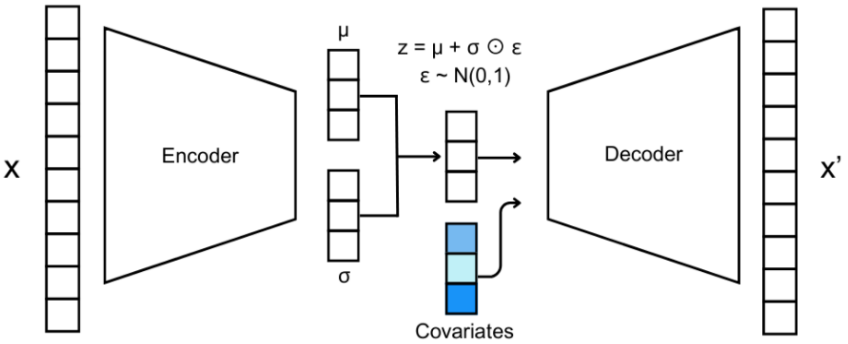

Figure 6: Architecture of the VAE-based multivariate normative model with covariates modeled in the decoder.

- **CM-2: Encoder-Only Modeling** In this method, the covariates are concatenated only to the input of the encoder. The encoder maps the combined input to the latent space, while the decoder reconstructs solely from $\mathbf{z}$. The architecture, shown in Figure 7, is as follows:

$$\{\mathbf{x}, \mathbf{c}\} \rightarrow \text{Encoder} \rightarrow \mathbf{z}, \quad \mathbf{z} \rightarrow \text{Decoder} \rightarrow \mathbf{x}'$$

  This method is included for completeness but is not expected to effectively achieve covariate invariance, as there is no constraint preventing the latent space from encoding covariate information.

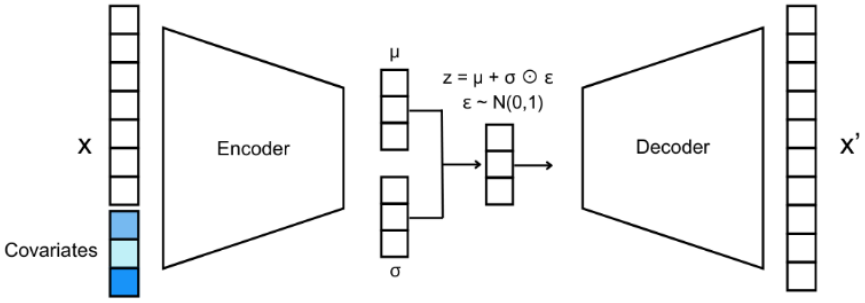

Figure 7: Architecture of the VAE-based multivariate normative model with covariates modeled in the encoder.

- **CM-3: Encoder-Decoder Modeling (cVAE)** In this method, the covariates are provided to both the encoder and the decoder. The encoder receives the concatenated input of brain features and covariates, while the decoder reconstructs the brain features based on the latent representation and the covariates. The architecture, shown in Figure 8, is as follows:

$$\{\mathbf{x}, \mathbf{c}\} \rightarrow \text{Encoder} \rightarrow \mathbf{z}, \quad \{\mathbf{z}, \mathbf{c}\} \rightarrow \text{Decoder} \rightarrow \mathbf{x}'$$

This design enables the model to condition the full generative process on the covariates and models the conditional distribution $p(\mathbf{x} \mid \mathbf{c})$. This method is very relevant, as it is the standard practice in many recent VAE-based multivariate normative modeling studies (e.g., Aguila et al. [13] and Wang et al. [24]).

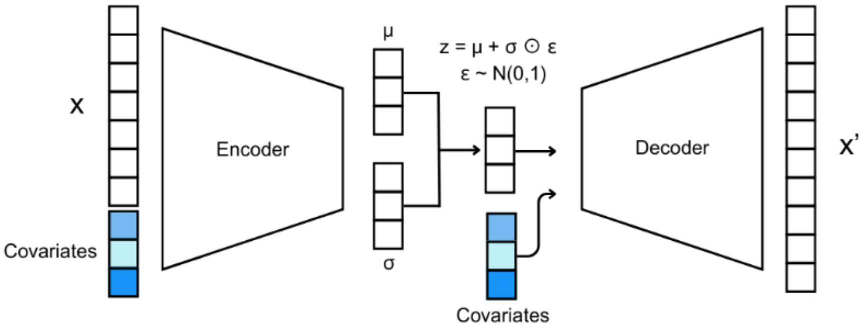

Figure 8: Architecture of the VAE-based multivariate normative model with covariates modeled in both encoder and decoder.

- **CM-4: Covariate Reconstruction** In this method, the covariates are concatenated to the input alongside the brain measurements. The decoder is trained to reconstruct both the brain features and the covariates. This method effectively treats the covariates as additional input features without any special treatment. The architecture, shown in Figure 9, is as follows:

$$\{\mathbf{x}, \mathbf{c}\} \rightarrow \text{Encoder} \rightarrow \mathbf{z}, \quad \mathbf{z} \rightarrow \text{Decoder} \rightarrow \{\mathbf{x}', \mathbf{c}'\}$$

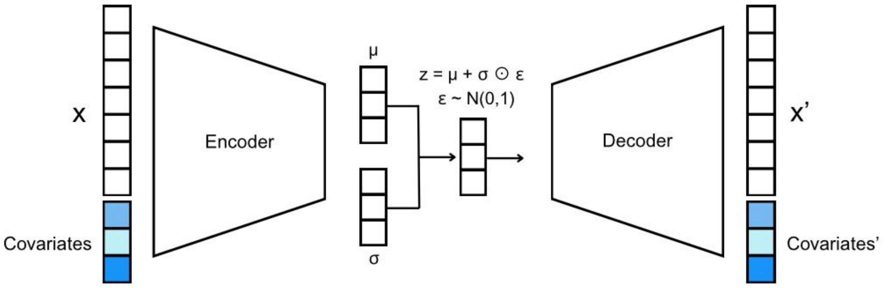

Figure 9: Architecture of the VAE-based multivariate normative model with reconstructed covariates.

- **CM-5: Conditional Loss Term** This method extends CM-4 by introducing a loss penalty specifically designed to target covariate reconstruction. In this variant, the covariates are treated as explicit reconstruction targets, and an additional loss term is used to optimize their reconstruction alongside the brain features. The new loss function follows the same structure as the one used in the baseline model, including an MSE reconstruction loss and a KL divergence regularisation term for the covariates. It is expected that this configuration forces the latent distribution to include the covariate information, resulting in a strong (and unwanted) dependence on the covariates. This model also serves as a helpful reference to verify whether the observed behavior aligns with theoretical expectations. This means that this method can be used to determine if it indeed causes a measurable dependence on those covariates in the latent space and whether the evaluation metrics can show this.

- **CM-6: Adversarial Loss Term** This method introduces an adversarial objective to the VAE training process, which aims to discourage the encoding of covariate information in the latent space. The idea is to optimize the encoder such that it becomes difficult for a downstream classifier (the adversary) to predict covariates from the latent representation. An auxiliary adversary network is attached to the mean vector of the latent distribution, $\mathbf{z}_\mu$, through a Gradient Reversal Layer (GRL) [30]. The GRL inverts the gradient signal during backpropagation, forcing the encoder to learn latent features that are predictive of the input $\mathbf{x}$ but uninformative of the covariates $\mathbf{c}$. The overall architecture can be described as follows:

  $$\mathbf{x} \rightarrow \text{Encoder} \rightarrow \{\mathbf{z}_\mu, \mathbf{z}_{\log \sigma^2}\}, \quad \mathbf{z} \rightarrow \text{Decoder} \rightarrow \mathbf{x}'$$

  $$\mathbf{z}_\mu \xrightarrow{\text{GRL}} \text{Adversary} \rightarrow \hat{\mathbf{c}}$$

  The adversary consists of multi-layer perceptrons (MLPs) that try to predict the covariates. Separate heads are used for continuous and categorical variables. The adversarial loss includes a mean squared error term for continuous covariates and a cross-entropy loss for categorical covariates. These losses are scaled and added to the total training loss.

- **CM-7: cVAE + Adversarial Loss Term** This method combines the covariate conditioning strategy of CM-3 (cVAE) with the adversarial penalty approach of CM-6. Here, covariates are included in both the encoder and decoder pathways, allowing the model to learn the conditional distribution $p(\mathbf{x} \mid \mathbf{c})$ while simultaneously applying an adversarial loss to encourage the latent representation to become more invariant to the covariates. A Gradient Reversal Layer (GRL) is attached to the mean vector $\mathbf{z}_\mu$ of the latent distribution. The adversary network attempts to predict the covariates from this representation, while the encoder is trained to make this prediction task difficult. The combined architecture is defined as follows:

  $$\{\mathbf{x}, \mathbf{c}\} \rightarrow \text{Encoder} \rightarrow \{\mathbf{z}_\mu, \mathbf{z}_{\log \sigma^2}\}, \quad \{\mathbf{z}, \mathbf{c}\} \rightarrow \text{Decoder} \rightarrow \mathbf{x}'$$

  $$\mathbf{z}_\mu \xrightarrow{\text{GRL}} \text{Adversary} \rightarrow \hat{\mathbf{c}}$$

  While the adversarial cVAE has been explored in prior work (e.g., Wang et al. [24]), its focus has been on improving brain feature reconstruction by applying adversarial losses to the input space. This method investigates whether applying the adversarial loss directly to the covariates improves covariate invariance in the latent space.

- **CM-8: cVAE + MMD Loss Term (VFAE)** This method is based on the Variational Fair Autoencoder (VFAE) proposed by Louizos et al. [29]. It aims to enforce statistical independence between the latent representation $\mathbf{z}$ and the covariates $\mathbf{c}$ by introducing a Maximum Mean Discrepancy (MMD) penalty term during training. The idea is that if the latent distributions for different groups (e.g., males vs. females) are similar, then the model has learned to ignore group-specific information in $\mathbf{z}$. The approach differs from, for example, adversarial methods like CM-6 and CM-7 in how independence is enforced. Instead of using an adversary to predict covariates and penalize predictability explicitly, the MMD-based approach compares latent distributions across groups and penalizes dissimilarity directly using kernel-based distance metrics. The architecture follows a conditional VAE (CM-3), where covariates are included in both the encoder and decoder inputs. However, an additional MMD penalty is applied to the latent code to promote invariance. During training, the MMD is computed between all pairwise combinations of subgroup-specific latent samples (e.g., $\mathbf{z}_{\text{male}}$ vs $\mathbf{z}_{\text{female}}$), using a radial basis function (RBF) kernel. This encourages the latent distribution to become invariant to the covariates without requiring the training of an adversarial network. The penalty is added to the total loss, weighted by a hyperparameter $\lambda_{\text{MMD}}$. This method is especially useful for one-hot encoded categorical covariates.

- **CM-9: cVAE + HSIC Loss Term (HCV)** This method uses the Hilbert-Schmidt Independence Criterion (HSIC) as a regularization term to increase statistical independence between the latent representation $\mathbf{z}$ and the covariates $\mathbf{c}$. Inspired by the HSIC-constrained VAE (HCV) [52], the model follows the same conditional VAE structure as CM-3, in which the covariates are provided to both the encoder and decoder. However, an additional penalty is introduced during training to discourage the encoding of covariate-specific information in the latent space. The HSIC is a non-parametric measure of dependence between two random variables based on kernel embeddings of their joint distribution. In this method, it is used to quantify the degree of dependence between latent samples and covariates. An RBF kernel is applied to both $\mathbf{z}$ and $\mathbf{c}$, and the HSIC value is computed based on the resulting kernel matrices. A lower HSIC value indicates greater independence. The penalty is added to the overall training loss, weighted by a hyperparameter $\lambda_{\text{HSIC}}$. This method applies to both continuous covariates (e.g., age) and categorical covariates (e.g., sex, site), making it a flexible choice for enforcing covariate invariance.

- **CM-10: Disentangled Subspace** This method partitions the latent space into two subspaces: a sensitive subspace that captures covariate-related information and an uninformative subspace meant to remain invariant to covariates. The model allocates $d_s$ latent dimensions for covariates and $d_u$ dimensions for covariate-invariant representations, with $d = d_s + d_u$. The sensitive part $\mathbf{z}_s$ is trained to predict the covariates, while the decoder reconstructs only from the uninformative part $\mathbf{z}_u$ combined with the covariates. Additionally, an HSIC penalty is applied between $\mathbf{z}_u$ and $\mathbf{c}$ to discourage residual covariate information in the uninformative subspace. Note that other types of penalties may also be used for this purpose. The architecture is defined as follows:

$$\mathbf{x} \to \text{Encoder} \to \{\mathbf{z}_s, \mathbf{z}_u\}, \quad \{\mathbf{z}_u, \mathbf{c}\} \to \text{Decoder} \to \mathbf{x}'$$

$$\mathbf{z}_s \to \text{Covariate Predictor} \to \hat{\mathbf{c}}$$

  To ensure comparability with other covariate modeling methods to this alternative definition of the latent space, the size of the uninformative subspace, $d_u$, is set equal to the latent space dimensionality used in all other methods. This ensures that the mapping of samples to the normative latent space occurs in the same dimensionality. While this method increases the overall model capacity through a larger latent space, this design choice provides the fairest comparison to the other covariate modeling strategies.

### 3.1.3 Brain Age Model

To perform brain age estimation, the VAE architecture is modified to learn a BAG. The goal is to predict the BAG, defined as the difference between the predicted and chronological age. The BAG is constrained to follow a standard normal distribution using a KL divergence term. The predicted brain age is computed by adding the estimated BAG to the known chronological age, and this value is passed to the decoder (instead of the chronological age) to reconstruct the original brain features. This setup ensures that the model cannot simply encode age directly into the latent space. Figure 10 shows a schematic of this architecture.

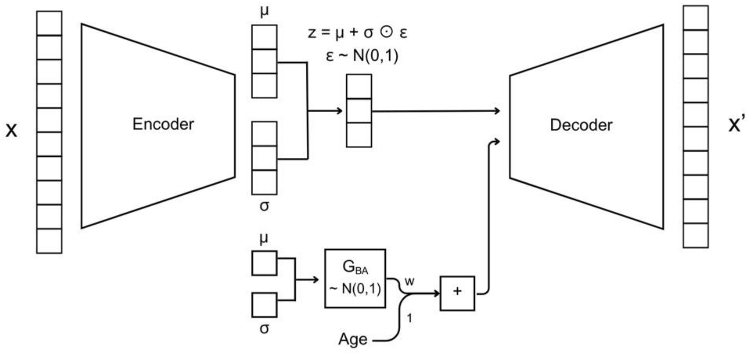

Figure 10: Architecture of the multivariate normative model adjusted for brain age estimation.

The architecture is designed to store age-related information needed for reconstruction in a dedicated latent variable, denoted $G_{BA}$. During training, the encoder takes the brain features, the subject's chronological age, and other covariates as input. It outputs a multi-dimensional latent vector $\mathbf{z}$, which captures brain variation, and a one-dimensional value $G_{BA}$, representing the BAG. $G_{BA}$, like the other latent dimensions, is forced to follow a standard normal distribution. The $G_{BA}$ is then added to the chronological age and passed to the decoder. A scaling factor, $w$, is introduced to scale the z-scored brain age gap and control its effect on chronological age. Because both quantities are expressed in z-score units rather than years, an appropriate $w$ is necessary to keep the calculated brain age within a biologically plausible range. The decoder reconstructs the scan based on $\mathbf{z}$ and age $+ G_{BA}$. When the brain appears older or younger than expected, the decoder relies on a positive or negative value of $G_{BA}$, respectively, to match the observed features.

One important aspect to consider is the interpretation of the BAG. Since $G_{BA}$ is trained to follow a standard normal distribution, its values represent standardized deviations, i.e., z-scores. To translate these values into meaningful age differences in years, a rescaling step is necessary. Another important aspect is that the learned deviations must reflect true individual differences rather than variation due to demographic factors or dataset biases. Covariate modeling methods, as presented in 3.1.2, should be applied to the brain age model to ensure that the latent representation $\mathbf{z}$ becomes invariant to the covariates. This allows the BAG to capture meaningful deviations in brain aging.

## 3.2 Model Evaluation

To evaluate the multivariate normative models, three key properties are assessed: **reconstruction quality**, **covariate invariance**, and **normative alignment**. These properties reflect different but complementary aspects of model performance, each important for measuring the performance of multivariate normative models. In addition to these quantitative evaluation metrics, qualitative visual inspection methods are also used to gain a deeper understanding of the structure of the learned latent space. These visual tools help assess whether covariate information is entangled with the latent representation, providing additional support for interpreting the quantitative results.

### 3.2.1 Reconstruction Quality

Reconstruction quality assesses how accurately the model can reproduce the original input data from its latent representation. High reconstruction quality indicates that the model has effectively captured the central tendencies and variability of the brain features. The following metrics will be used to represent the reconstruction quality:

- **Mean Squared Error (MSE)**: Quantifies the average squared difference between the input and its reconstruction. It provides an absolute measure of reconstruction error in the same units as the input data, reflecting how closely the reconstruction matches individual values. MSE is non-negative and unbounded from above. A lower value indicates a better reconstruction quality, with zero representing a perfect reconstruction.

- **Coefficient of Determination ($R^2$)**: Measures the proportion of variance in the input data that is explained by the reconstruction. It is computed as:

  $$R^2 = 1 - \frac{\text{MSE}}{\mathrm{Var}(\mathbf{x})}$$

  where $\mathrm{Var}(\mathbf{x})$ is the variance of the original data. $R^2$ ranges from $-\infty$ to 1, where higher values indicate a better reconstruction. An $R^2$ of 1 denotes perfect reconstruction, and an $R^2$ of 0 is equivalent to predicting the mean of the input. Unlike MSE, $R^2$ is a scale-independent, relative measure that enables direct comparison across models and datasets, including those from other studies.

Together, these metrics provide complementary insights: MSE offers a direct measure of reconstruction error, while $R^2$ explains this accuracy relative to the variability in the original data. The reconstruction quality metrics form the foundation for evaluating overall model quality. Reconstruction performance should remain stable under different experimental and modeling conditions. Stability in reconstruction quality is important because it reflects how well the model represents the input data in its latent representation. A model that shows large fluctuations in its reconstructions across different experimental setups may fail to capture the underlying structure of the data consistently. In this work, methods are investigated that incorporate covariates into the modeling process. These extensions should not compromise the model's ability to reconstruct brain feature representations. A substantial drop in reconstruction quality is therefore used as a key indicator that a particular model architecture or configuration may not be learning meaningful or generalizable latent representations.

### 3.2.2 Covariate Invariance

Covariate invariance evaluates to what extent known covariates (e.g., age, sex, site) are present in the learned latent representation. Normative models aim to be invariant to such covariates so that detected deviations are not confounded by variation in demographic variables. The following metrics are used to assess the degree to which residual covariate information remains in the latent space:

- **Random Forest Prediction Accuracy**: A Random Forest (RF) model is used for detecting covariate information present in the learned latent representation of the model. This type of model is chosen because it is a strong, non-parametric model that performs well out-of-the-box and is robust to overfitting [53]. An RF model can capture both linear and nonlinear relationships, making it well-suited for detecting covariate dependencies that may still be encoded in the latent space. The latent representations of the test samples are used as input features for the RF, with the corresponding covariate values serving as prediction targets. The model's predictions are then compared to the true covariate values to evaluate residual information in the latent space. The RF model is trained to predict covariates from the latent space using 70% of the test set. The remaining 30% is used for evaluation. Poor predictive performance (i.e., low classification accuracy or high regression error) indicates stronger covariate invariance. The procedure for each covariate is as follows:

  – *Age (continuous)*: An RF regressor is used and evaluated using the MSE between the predicted age and the true age. A lower MSE indicates that more age information is present in the latent space, and a higher MSE indicates stronger invariance. MSE is non-negative and has no fixed upper bound. Under full invariance, the optimum is $MSE = Var(age)$, representing prediction by the mean only.

  – *Sex (binary)*: An RF classifier is used and evaluated using classification accuracy. Accuracy ranges from 0 to 1, with 1 indicating perfect prediction. Lower values imply stronger invariance. For a balanced dataset, random guessing yields an expected accuracy of 0.5. This serves as a baseline under full invariance.

  – *Site (categorical)*: An RF classifier is used and evaluated using classification accuracy. Like sex, these metrics range from 0 to 1. Lower accuracy indicates reduced site information in the latent space and, thus, greater invariance. For multiple classes, the baseline performance depends on the class distribution (e.g., the chance level for a uniform distribution is 1/number of classes).

- **Mutual Information (MI)**: Quantifies the statistical dependency between the latent representation and each covariate. MI measures how much knowing one variable reduces uncertainty about the other [54]. In this context, it reflects how much information about a covariate (e.g., age, sex, or site) is retained in the latent space. MI is non-negative and unbounded from above. A value of zero indicates complete statistical independence, which means perfect covariate invariance. Lower MI values, therefore, correspond to stronger disentanglement. MI is computed separately for each covariate and each latent dimension and then aggregated by averaging over all dimensions.

Together, these metrics provide complementary views on the presence of covariate information in the latent space. RF prediction assesses whether a covariate can be accurately recovered from the latent representation. In contrast, MI quantifies statistical dependence in a model-free way and is sensitive to all forms of dependency.

### 3.2.3 Normative Alignment

Normative alignment evaluates how well the latent space aligns with the assumed prior distribution (usually a standard multivariate normal distribution). This is essential for defining a meaningful normative range and detecting outliers. The following metric is used for the evaluation of this property:

- **Kullback–Leibler (KL) Divergence**: Measures how much the learned posterior $q_\phi(z \mid x)$ diverges from the prior $p(z) = \mathcal{N}(0, I)$. Lower KL divergence indicates better alignment with the normative prior. This metric is computed per latent dimension and then averaged across all dimensions.

KL divergence is also a central component of the loss function in VAEs, where it acts as a regularizer to encourage the approximate posterior to match the prior. Since it directly shapes the structure of the latent space during training, it also serves as a valuable metric for evaluating whether the resulting representation maintains the desired normative properties. A low final KL divergence indicates that the model has successfully learned a latent space agreeing with the normative prior.

### 3.2.4 Qualitative Analyses

In addition to quantitative evaluation metrics, a range of qualitative analysis methods is used to assess the behavior and structure of the multivariate normative models. These methods provide insight into how information is encoded, whether latent representations adhere to the modeling assumptions, and how demographic variables influence the learned space. The following visual inspection tools are used:

- **Latent Space Scatter Plots**: All pairwise combinations of latent dimensions are visualized using 2D scatter plots. These plots reveal the geometric structure of the latent space and highlight statistical dependencies between latent variables. Strong directional patterns, skewness, or clustering indicate correlations between dimensions, which can reduce disentanglement. This method also allows the identification of outliers in specific latent dimensions. If the deviation is substantial for a sample in the training set, it may indicate that this sample is problematic. These samples should be carefully reviewed in a case-by-case manner.

- **t-SNE Projection of Latent Codes**: To qualitatively assess how well the model separates or entangles known covariates in the latent space, 2D and 3D t-distributed stochastic neighbor embedding (t-SNE) are applied to the latent codes of the test set. These projections reduce the latent space to a size that is interpretable for visualization. Points are color-coded according to covariate values: continuous covariates, such as age, are represented using a gradient, and categorical covariates, like sex or site, are visualized using discrete colors or markers. These plots reveal whether samples cluster based on demographic attributes, which would suggest covariate leakage.

- **Brain Difference Map**: To improve the interpretability of the latent space, the generative property of the VAE is used by feeding synthetic latent codes into the decoder. For each latent dimension, a high negative and positive value (e.g., a z-score of $\pm 3$) is assigned while keeping all other dimensions fixed at zero. Covariates, such as age, are fixed at their mean (e.g., zero after standardization), and sex is arbitrarily set to male. The resulting reconstructed brain data can then be used to generate difference maps that highlight which brain regions are most influenced by each latent factor. This method shows how specific latent dimensions correspond to anatomical variation and helps identify which features are being encoded in each dimension of the learned representation.

Together, these qualitative methods allow for a better understanding of the model's latent structure and help confirm whether the models meet normative modeling requirements. They can help interpret the quantitative metrics reported in Section 3.2.

## 3.3 Data

This work evaluates the VAE-based normative modeling framework using structural brain imaging data from children and adolescents. Two datasets are used: a proprietary dataset from the GenR study, which is available through collaboration with Erasmus MC, and an open-access dataset from the Healthy Brain Network (HBN) study. Together, these datasets provide a variation in age, sex, and imaging sites. This section gives an overview of the data characteristics and explains the preprocessing pipeline.

### 3.3.1 Datasets

This work uses two large, high-quality neuroimaging datasets, namely the GenR and the HBN datasets.

**Generation R Study (GenR)** The Generation R Study is a population-based prospective cohort study based in Rotterdam, the Netherlands, designed to identify early environmental and genetic determinants of growth, development, and health from fetal life onward [55, 56]. The cohort initially recruited 9,778 pregnant mothers and has collected extensive longitudinal data, including neuroimaging and health-related covariates.

Specifically, T1-weighted structural MRI images were collected across three waves:

- **Wave 1:** 1070 children aged 6–9 years.
- **Wave 2:** 4087 children aged 9–11 years.
- **Wave 3:** 3725 adolescents aged 13–17 years.

MRI acquisition in this cohort was performed at a single site in Rotterdam using the same scanner for consistency across participants.

**Healthy Brain Network (HBN)** The Healthy Brain Network is an open-science initiative launched by the Child Mind Institute to collect comprehensive neuroimaging, cognitive, and behavioral data from 10,000 children and adolescents (aged 5–21 years) in the New York City area [57]. Unlike GenR, the HBN dataset includes participants recruited across multiple sites:

- **RUBIC:** Rutgers University Brain Imaging Center
- **CBIC:** Citigroup Biomedical Imaging Center
- **CUNY:** City University of New York Advanced Science Research Center

MRI acquisition in HBN was performed using different scanner models, providing site variability that is ideal for testing model generalizability under multi-site conditions.

An important distinction between the datasets is that GenR contains longitudinal data with repeated measures from the same individuals across different developmental stages (i.e., waves). This allows potential longitudinal analyses. In contrast, HBN primarily contains cross-sectional data with mostly unique participants. Table 1 summarizes key demographic characteristics of the GenR and HBN datasets used in this work. Note that the SI site is not used since it was incomplete.

Table 1: Summary statistics for the GenR and HBN datasets

| **Characteristic** | **GenR** | **HBN** |
|---|---|---|
| # Samples | 8529 | 2689 |
| # Subjects | 5145 | 2373 |
| Age Range (years) | 6.1 – 17.1 | 5.0 – 21.8 |
| Sex (% Female) | 50.3% | 35.9% |
| # Sites | 1 (Rotterdam) | 3 (RUBIC, CBIC, CUNY) |

In addition to the brain measurements, multiple covariates are available for both datasets. Table 2 details the covariates and how they are encoded. Encoding site information is important when assessing the model's ability to accommodate multi-site variability. Likewise, encoding age and sex is important for modeling covariate effects.

Table 2: Overview of available covariates for the GenR and HBN datasets

| Covariate | Type | Encoding | Datasets |
|---|---|---|---|
| Site | Categorical | 0: Rotterdam (GenR)<br>1: RUBIC (HBN)<br>2: CBIC (HBN)<br>3: CUNY (HBN) | GenR, HBN |
| Age | Continuous | Measured in years | GenR, HBN |
| Sex | Binary | 0: Male<br>1: Female | GenR, HBN |

### 3.3.2 Preprocessing Pipeline

The preprocessing pipeline applied to the datasets consists of several steps to ensure the data is clean, standardized, and suitable for training and evaluating normative models. Figure 11 shows the complete preprocessing pipeline. Table 3 shows the final dataset sizes after applying the complete preprocessing pipeline.

Table 3: Dataset split sizes after preprocessing

| Dataset | Total Samples | Training Samples (70%) | Test Samples (30%) |
|---|---|---|---|
| GenR | 8529 | 5991 | 2538 |
| HBN | 2369 | 1659 | 710 |

**Extraction of Brain Measurements** Structural MRI data (T1-weighted scans) were processed using FreeSurfer version 6.0.0 [58]. This software automatically performs cortical and subcortical segmentation and provides regional brain measurements. Specifically, features were extracted from:

- **Cortical Parcellation:** Measurements for left and right hemispheres, including cortical surface area (mm$^2$), mean cortical thickness (mm), and cortical volume (mm$^3$) across anatomically predefined regions.
- **Subcortical Segmentation:** Volumetric measures (mm$^3$) for subcortical structures such as the hippocampus, amygdala, caudate, and thalamus.

This extraction yields a high-dimensional tabular representation of the brain for each individual.

**Quality Control** Quality control was performed to identify and exclude poor-quality scans. GenR had already undergone extensive quality control by the GenR research group. Therefore, additional quality control steps based on the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium protocols [59] were applied only to the HBN dataset. These protocols involve statistical outlier detection based on standardized feature values. The ENIGMA-based quality control was applied as follows:

- Each feature was checked for outliers, defined as exceeding 2.698 standard deviations from the mean. A sample was excluded if 10 or more features were flagged as outliers.
- To prevent over-penalization due to a few highly sensitive features, additional checks were implemented. Features that were flagged as outliers significantly more often than expected (exceeding 2.698 standard deviations compared to the mean outlier rate across features) were identified as sensitive features and excluded from the total count when evaluating sample quality.

Using this procedure, 10 samples were flagged and excluded from the HBN dataset.

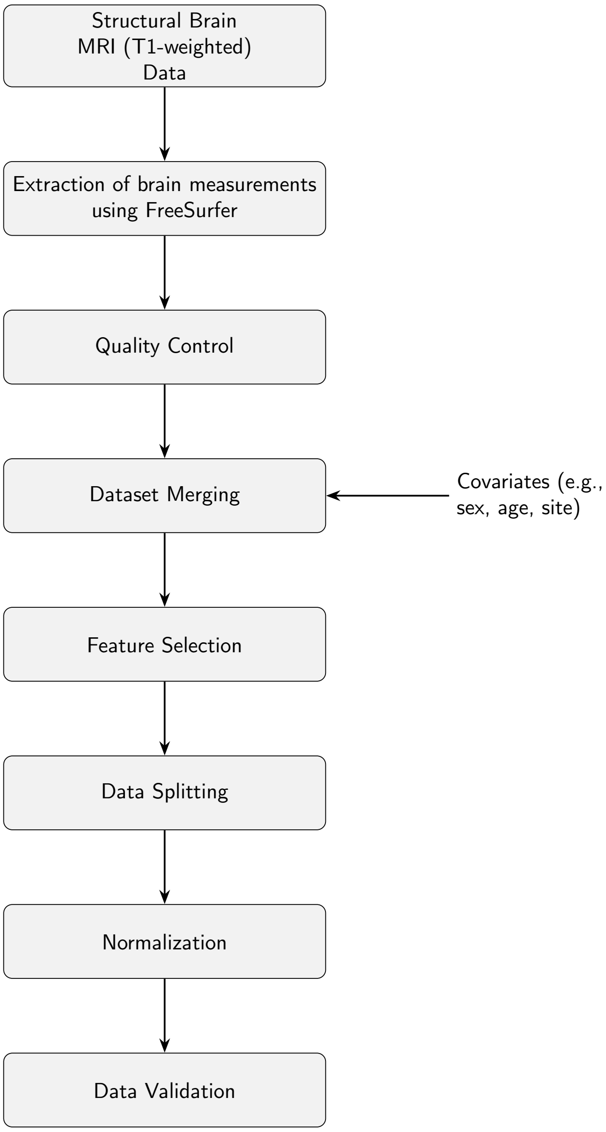
Structural Brain MRI (T1-weighted) Data
Extraction of brain measurements using FreeSurfer
Quality Control
Dataset Merging
Covariates (e.g., sex, age, site)
Feature Selection
Data Splitting
Normalization
Data Validation

Figure 11: Overview of the data preprocessing pipeline

**Dataset Merging** All brain region feature names and covariate feature names were aligned between GenR and HBN to ensure consistency between datasets. Covariate data (e.g., age, sex, site) were then merged with the extracted brain measurements into a unified dataset for each cohort, simplifying downstream processing and modeling.

**Feature Selection** Feature selection was applied to focus on a consistent and interpretable set of brain measures. Only cortical volumes from the left and right hemispheres were selected. These left and right versions of the volumetric features were then averaged into a single feature per brain region. This choice reflects practices in previous VAE-based normative modeling works [13, 25], which aim to balance information richness with model complexity. The total number of brain features then amounts to 34. By using only cortical volumes, the dimensionality of the data is kept small while still capturing a complete representation of brain structure.

**Data Splitting** The dataset was partitioned into training and test sets, with 70% of the samples allocated to the training set and the remaining 30% to the hold-out test set. Within the training set, a further split was performed: 85% for training and 15% for validation.

Special attention was paid to prevent data leakage. GenR contains longitudinal data (multiple scans from the same individuals across different waves). Random splits could accidentally place scans from the same individual in both training and test sets. A grouping-based splitting function was applied to prevent this, ensuring that all scans from a single individual were assigned to the same data split. After splitting, datasets were shuffled within each split.

**Normalization** The brain measurement features were standardized to z-scores based on statistics computed from the training set only. Standardizing features to a normative scale is important for the interpretability of deviations and facilitating model convergence during training. The normalization parameters (means and standard deviations) calculated on the training data were stored and applied unchanged to the validation and test sets to prevent data leakage. Covariate encoding was applied depending on the covariate modeling technique, as described in Section 3.1.2.

**Data Validation** After preprocessing, all datasets were checked for invalid or missing entries. Samples containing invalid values were excluded accordingly to ensure clean model input.

## 3.4 Experiments

This section presents the experimental design and evaluation procedures to answer the research questions defined in Section 1.3. Each experiment is introduced with a description of its objective and relevance to the research questions. The model training setup is explained in detail, including dataset configuration, architectural choices, and parameter selection. This is followed by an explanation of the evaluation approach, highlighting both the quantitative metrics and qualitative analyses used to assess the model's behavior.

### 3.4.1 Experimental Platform

While univariate normative modeling has benefited from structured toolkits such as the PCNtoolkit [9], current multivariate normative modeling studies lack a similarly comprehensive and reusable platform. To address this, an open-source experimental platform is introduced as part of this thesis [1]. This framework enables researchers to define VAE-based multivariate normative models, select covariate embedding techniques, and systematically evaluate performance using a wide range of metrics and criteria. By making the platform publicly available, this work aims to promote reproducibility in the field.

### 3.4.2 Baseline Experiments

The baseline experiment establishes a reference point for evaluating multivariate normative models. In this experiment, no covariate information is included in the model. This enables the assessment of the default behavior of a VAE trained solely on brain imaging data without any explicit modeling of age, sex, or other covariates. Doing so provides a benchmark against which more advanced models (e.g., those incorporating covariates or accounting for batch effects) can be compared.

**Model Training** The baseline model is trained on both the GenR dataset and the HBN dataset separately. The model architecture for the baseline model is presented in Section 3.1.1. An important part of this experiment is explaining the role of latent space size. The number of dimensions in the latent space affects how much information the model can compress and reconstruct. A tiny latent space may oversimplify the data, resulting in poor reconstructions. A larger latent space can improve reconstruction but may also increase the risk of overfitting, reduced regularization, or entanglement with demographic variables. To explore this, the model is trained with a range of latent sizes: $1, 2, 3, 4, 5, 8, 12, 16$. Each model configuration is trained five times using different random seeds. This helps account for variation due to model initialization and the stochastic nature of training.

[1] https://github.com/remdui/MultivariateNormativeModeling

**Model Evaluation** After training, model performance is assessed based on the key properties defined in Section 3.2: reconstruction quality, covariate invariance, and alignment with the normative prior. Results are averaged across five training runs using different random seeds, and standard deviations are reported to reflect variability. In addition to these quantitative metrics, a set of qualitative analyses is applied to understand better the structure of the learned latent space and the presence of covariate information. These analyses, including latent space visualizations and reconstruction comparisons, are described in detail in Section 3.2.4.

The baseline model results play a central role in the other experiments. It shows how the model behaves without explicitly modeling any covariate information. It also explains how the size of the latent space influences the trade-offs between reconstruction performance, regularisation, and the impact of covariates. This makes it easier to understand whether later improvements are due to better modeling choices or simply differences in model capacity.

### 3.4.3 Covariate Modeling Experiments

This experiment investigates the impact of various strategies for incorporating covariates into multivariate normative models on key properties, including reconstruction quality, latent space invariance, and alignment with the normative prior. This experiment addresses RQ1, which focuses on evaluating the effect of covariate modeling techniques on important model properties.

**Model Training** The experiment builds on the baseline setup by introducing covariate information into the model in various ways, as described in Section 3.1.2. Each approach represents a different assumption about how covariates should influence the learned representation. Models are trained and evaluated on two datasets. For GenR, the covariates age and sex are modeled. For HBN, age, sex, and site are included. Each covariate modeling strategy is tested across the same range of latent dimensionalities as in the baseline experiment $(1, 2, 3, 4, 5, 8, 12, 16)$ and repeated five times using different random seeds. This setup allows for a fair comparison across conditions and model capacities.

**Model Evaluation** Performance is evaluated using the metrics described in Section 3.2, enabling direct comparison with the baseline models. Models that effectively remove the influence of covariates are expected to show reduced covariate dependence in the latent space, as indicated by lower mutual information scores and improved covariate prediction accuracy. These improvements should ideally come without compromising reconstruction performance. To complement the quantitative metrics, a consistent set of qualitative analyses is applied as described in Section 3.2.4. These visual inspection methods provide insight into how covariates influence the latent space and whether disentanglement assumptions are met.

In summary, this experiment provides insight into which covariate modeling approaches best mitigate the influence of covariates while maintaining strong reconstruction and normative alignment across different model capacities. Additionally, these experiments aim to identify which covariate modeling strategies are particularly effective for different types of covariates, such as categorical (e.g., sex, site) and continuous (e.g., age) variables.

### 3.4.4 Site Accommodation Experiments

This experiment evaluates whether VAE-based multivariate normative models can effectively accommodate site-related batch effects, which are common in multi-site neuroimaging studies. Site variance can originate from differences in scanner hardware, acquisition protocols, or population sampling between sites, and if not properly accounted for, it may introduce systematic bias into model predictions. Handling such variance is particularly important for normative models intended for clinical use, where robustness and generalizability across unseen data sources are critical. This experiment addresses RQ2, which examines to what extent multivariate VAEs can handle site variability and how their performance compares to existing harmonization methods such as ComBat.

**Model Training** The experiment uses a combined variant of the GenR and HBN datasets. GenR is a single-site cohort, whereas HBN includes data from three distinct sites. A subset of 750 samples is randomly selected from all sites to ensure a balanced representation across sites. One site from HBN is excluded due to insufficient data, ensuring that the comparison is not complicated by class imbalance or underpowered subsets. The sampling procedure is stratified to ensure a fair distribution of age and sex across the selected subset, reducing potential biases introduced by demographic imbalances.

Three model configurations are considered:

- **Baseline VAE** trained directly on the raw brain features without covariate information, using the baseline model defined in Sections 3.1.1.
- **ComBat Harmonization** where ComBat is applied as a preprocessing step to remove site-related variance while preserving covariate effects (e.g., age and sex), followed by training the baseline VAE architecture on the harmonized data.
- **VAE + Covariate Modeling** trained directly on the original (non-harmonized) data but with explicit modeling of covariates, including age, sex, and site. The covariate modeling follows the best-performing strategy outlined in Section 3.1.2.

Each model is trained five times. The training setup for all three configurations is kept identical to ensure comparability.

**Model Evaluation** A leave-one-site-out evaluation is applied to assess how well the model generalizes to unseen sites. In each evaluation fold, the model is trained on data from all but one site and then evaluated on a test set that includes two components: (1) a set of 100 samples from the held-out (unseen) site and (2) a randomly selected subset of 100 samples from each training (seen) site. This setup enables a direct comparison of how the model maps data from seen and unseen sites in the latent space. Performance is measured using the quantitative metrics described in Section 3.2. Metrics are averaged across the different folds and random seeds to assess robustness and generalization. This evaluation assesses whether the VAE architecture, with and without covariate modeling methods, can match or outperform ComBat, a traditional data harmonization technique.

This experiment demonstrates the capabilities of VAE-based multivariate normative models to handle site-related batch effects, a key challenge in multi-site neuroimaging studies.

### 3.4.5 Comparison to Univariate Model

This experiment evaluates how the proposed multivariate normative model compares to a traditional univariate normative approach and addresses RQ3. While univariate models construct a dedicated model for each brain feature independently, multivariate models are designed to capture joint structure across features in a shared latent representation. These fundamental differences in modeling assumptions, training objectives, and outputs make direct comparison challenging. In particular, multivariate models do not guarantee one-to-one reconstructions for each feature but instead aim to preserve the overall statistical structure. Despite these limitations, this experiment compares the models by evaluating the average deviation scores of the test samples. However, such comparisons cannot completely capture the benefits of using multivariate models. Downstream clinical tasks offer a more appropriate setting for a more practical inspection. Section 3.4.6 explores one such task as a proof of concept of multivariate normative models.

**Model Training** All 34 cortical volume features from the GenR and HBN datasets are selected for analysis. Two model types are compared. The first is a univariate model trained using the PCN-toolkit [9], where Hierarchical Bayesian Regression (HBR) is applied independently to each feature. Age and sex are included as covariates in every univariate model, resulting in 34 feature-specific models. The second is a multivariate VAE with a covariate modeling method applied that is trained using the best-performing strategy identified in the covariate modeling experiments (Section 3.4.3).

**Model Evaluation** Deviation scores from univariate and multivariate normative models are translated into z-score units and then summarized across measures for direct comparison. In the univariate case, each brain feature yields a z-score. In contrast, the multivariate VAE provides a z-score for each latent dimension. For each test subject, the mean of all feature-wise z-scores (univariate) and the mean of all latent-dimension z-scores (multivariate) are computed. These per-subject average z-scores are used to compare both models.

### 3.4.6 Brain Age Estimation Experiment

Section 3.4.5 highlighted the difficulty in comparing multivariate and univariate normative models due to their different assumptions and designs. Univariate models treat each feature independently, whereas multivariate models are designed to capture interdependencies between features. This structural difference complicates direct comparison. As a result, evaluating a clinically meaningful downstream task can provide additional insights into the performance of multivariate normative models. One such task is brain age estimation, which is explained in detail in Section 2.6.

This experiment is an important proof of concept for the applicability of multivariate normative models in clinical settings. Many clinically relevant problems rely on complex patterns that cannot be adequately modeled when features are treated in isolation. In such contexts, capturing interactions across multiple brain regions becomes necessary. Brain age estimation is a clear example. Here, the biological process of brain aging presents itself in complex patterns across many brain regions, making it a poor fit for univariate modeling.

**Model Training** The model is trained on both the GenR and HBN datasets using the brain age architecture described in Section 3.1.3. The best-performing covariate embedding strategy, as identified in the covariate modeling experiments, is used to make the model invariant to covariates.

**Model Evaluation** After training, BAG values are computed for all samples in the test set. These scores are added to the chronological age to produce an adjusted brain age estimate. Since the BAG values are modeled as standardized z-scores, they are rescaled back to age units using the mean and standard deviation obtained from the preprocessing step. This step ensures that the results are interpretable in years. Finally, the predicted brain age is compared to the actual chronological age using the MAE over the entire test set.
This experiment shows how multivariate normative modeling can be used for clinically meaningful tasks like brain age estimation. It demonstrates the model's capability to capture biologically relevant variation across multiple brain features. Furthermore, it serves as an example for other clinical or predictive tasks for evaluating multivariate normative models.

## 3.5 Available Resources

Several computational resources were available to facilitate data processing, model development, and evaluation for this research project. The ResearchSuite environment at ErasmusMC hosts the GenR datasets and provides sufficient computational power for initial data exploration and analysis. ResourceSuite is a Linux environment that includes the software and tools necessary for handling GenR data. Additionally, access to the Snellius supercomputer [60] in the Netherlands has been provided through the Child and Adolescent Psychiatry/Psychology NeuroImaging (CAPPNI) research group. Snellius is the most powerful supercomputer in the Netherlands. It features multiple GPU nodes (NVIDIA A100 and H100) that can be utilized for training models. Documentation is available for both computing resources and the Gen R dataset. To comply with the data policies of ErasmusMC, a mandatory data security session was held.

# 4 Results

## 4.1 Baseline Model

The baseline multivariate normative model, as defined in Section 3.1.1, was implemented without explicit covariate modeling. The model was trained and tested on both the GenR and HBN datasets using various latent space dimensionalities. Each configuration was trained five times with different random seeds, and results were averaged to account for variability due to random initialization and stochastic training dynamics. The evaluation was conducted using the three main quantitative properties defined in Section 3.2: reconstruction quality, covariate invariance, and normative alignment. The results for the baseline model across different latent dimensions are presented in Tables 4 and 5 for GenR and HBN, respectively. Note that for the HBN dataset, covariate invariance is also assessed for the site covariate, as this dataset includes multiple acquisition locations.

Table 4: Evaluation results of the baseline model across different latent space dimensions ($d$) trained using the GenR dataset.

| **Latent Dim.** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| 1 | 0.55 | 0.44 | 1.91 | 0.03 | 0.56 | 0.06 | 1.79 |
| 2 | 0.45 | 0.54 | 1.81 | 0.10 | 0.59 | 0.05 | 1.52 |
| 3 | 0.42 | 0.57 | 1.79 | 0.17 | 0.61 | 0.07 | 1.21 |
| 4 | 0.40 | 0.59 | 1.77 | 0.18 | 0.62 | 0.08 | 1.03 |
| 5 | 0.39 | 0.61 | 1.77 | 0.17 | 0.63 | 0.08 | 0.88 |
| 8 | 0.36 | 0.63 | 1.76 | 0.25 | 0.66 | 0.08 | 0.63 |
| 12 | 0.35 | 0.64 | 1.76 | 0.30 | 0.65 | 0.13 | 0.44 |
| 16 | 0.35 | 0.64 | 1.76 | 0.34 | 0.66 | 0.13 | 0.33 |

Table 5: Evaluation results of the baseline model across different latent space dimensions ($d$) trained using the HBN dataset.

| **Latent Dim.** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Covariate Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | KL Divergence |
| 1 | 0.47 | 0.53 | 1.36 | 0.02 | 0.59 | 0.05 | 0.36 | 0.01 | 1.80 |
| 2 | 0.42 | 0.58 | 0.93 | 0.11 | 0.65 | 0.09 | 0.38 | 0.01 | 1.29 |
| 3 | 0.40 | 0.60 | 0.90 | 0.20 | 0.61 | 0.09 | 0.42 | 0.04 | 1.04 |
| 4 | 0.38 | 0.62 | 0.81 | 0.18 | 0.66 | 0.09 | 0.45 | 0.02 | 0.87 |
| 5 | 0.38 | 0.62 | 0.82 | 0.18 | 0.66 | 0.10 | 0.46 | 0.07 | 0.72 |
| 8 | 0.37 | 0.63 | 0.77 | 0.23 | 0.65 | 0.12 | 0.45 | 0.11 | 0.48 |
| 12 | 0.37 | 0.63 | 0.81 | 0.25 | 0.66 | 0.15 | 0.49 | 0.17 | 0.31 |
| 16 | 0.37 | 0.63 | 0.80 | 0.39 | 0.66 | 0.24 | 0.47 | 0.22 | 0.23 |

Several patterns can be observed from these results. First, reconstruction quality improves with increasing latent dimension: MSE decreases and $R^2$ increases, indicating that more model capacity enables better reconstruction of the input data. At the same time, covariate invariance metrics show a tradeoff to the improved reconstruction quality. As the latent space grows, covariate prediction performance improves, prediction MSE for age decreases, classification accuracy for sex (and site for HBN) increases, and MI scores increase. This means the latent space encodes more covariate-related information, suggesting a drop in covariate invariance as model complexity increases. Normative alignment, measured via the KL divergence, improves with larger latent spaces. The KL scores decrease steadily as the latent space size increases, indicating that the posterior distribution learned by the model increasingly aligns with the assumed standard normal prior.

These trends are visualized in Figures 12 and 13, which plot the evaluation metrics as functions of latent dimensionality for GenR and HBN, respectively. Each plot groups the metrics by evaluation type, with the y-axis representing the metric values and the x-axis representing latent size. These figures clearly illustrate the tradeoff between reconstruction quality and covariate invariance as latent size increases, as well as the improved prior alignment.

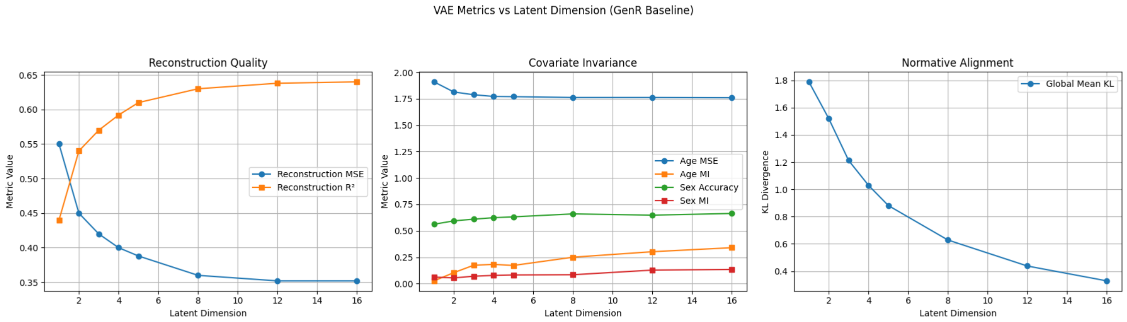

Figure 12: Combined view of (1) reconstruction quality (MSE & R²), (2) covariate invariance metrics (MSE for age, accuracy for sex) & mutual information), and (3) normative alignment (mean KL divergence) as functions of latent dimension for the GenR baseline model.

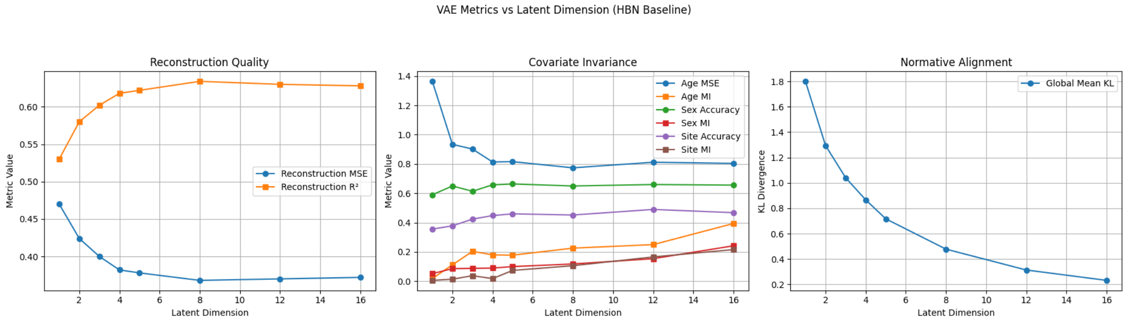

Figure 13: Combined view of (1) reconstruction quality (MSE & R²), (2) covariate invariance metrics (MSE for age, accuracy for sex and site) & mutual information), and (3) normative alignment (mean KL divergence) as functions of latent dimension for the HBN baseline model.

Most patterns are consistent between GenR and HBN, except for one notable exception: the age prediction error. While both datasets span a similar age range, GenR consistently shows higher prediction MSE for age compared to HBN. One possible explanation is that the GenR data is collected in distinct measurement waves, which may introduce age clusters into the data. This violates the assumption of a uniform age distribution, potentially weakening the model's ability to predict age and thereby increasing error. While this limitation is relevant, it does not undermine the validity of GenR for comparative experiments across modeling methods, as all models are evaluated on the same distribution.

To complement the quantitative analysis, a qualitative inspection is performed for the baseline models with a latent dimensionality of $d = 5$. This latent space size offers a balance between interpretability and sufficient capacity for reconstruction. Figure 14 shows 2D t-SNE projections of the latent space for the GenR test set, color-coded by age and sex. Age is visualized using a continuous color gradient representing the z-scored chronological age, while sex is represented using discrete color labels. A clear dependence on age is visible: older individuals tend to cluster near the center of the space, while younger individuals are distributed towards the edges. For sex, a similarly structured clustering can be observed. In an ideally invariant representation, covariate values would be randomly spread throughout the latent space. These results indicate that age and sex information are encoded to some extent in the model.

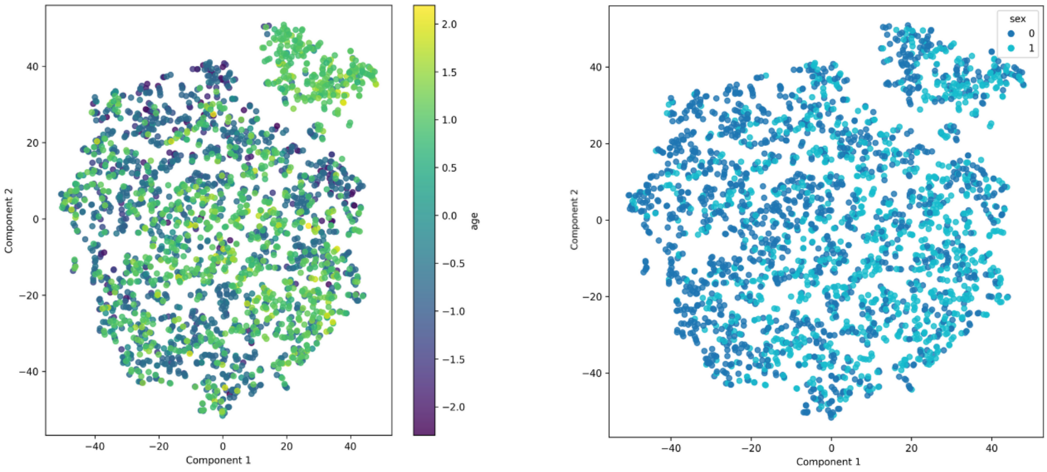

Figure 14: 2D t-SNE projection of the latent space of the baseline model ($d = 5$), trained on GenR, color-coded by covariates: age (left) and sex (right).

A similar analysis is presented in Figure 15 for the HBN dataset, which includes the site covariate. The t-SNE plots show similar dependency patterns related to age and sex. However, the distribution of site labels appears more uniformly spread across the latent space rather than forming distinct clusters. This suggests that the baseline model may already partially account for site-related variance or that site effects in HBN are relatively subtle. The experiments described in Section 3.4.4 provide a more comprehensive analysis of the site covariate.

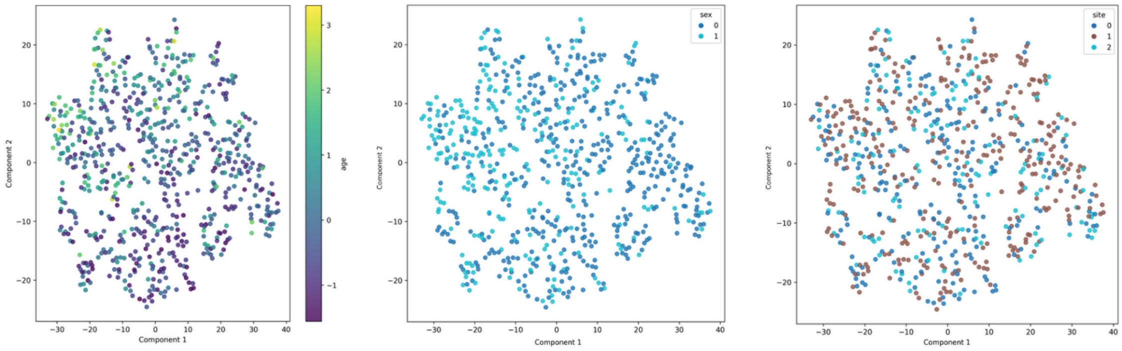

Figure 15: 2D t-SNE projection of the latent space of the baseline model ($d = 5$), trained on HBN, color-coded by covariates: age (left), sex (middle), and site (right).

Figure 16 presents latent space scatter plots for the GenR baseline model ($d = 5$). Each subplot shows the relationship between two latent dimensions. The distributions appear approximately Gaussian and centered, with no strong linear or non-linear dependencies across dimensions. A few outliers can be observed, but no extreme deviations are present. This suggests that the latent dimensions are mostly independent and that the model does not exhibit strong entanglement across latent dimensions. The baseline model trained on the HBN dataset shows similar results.

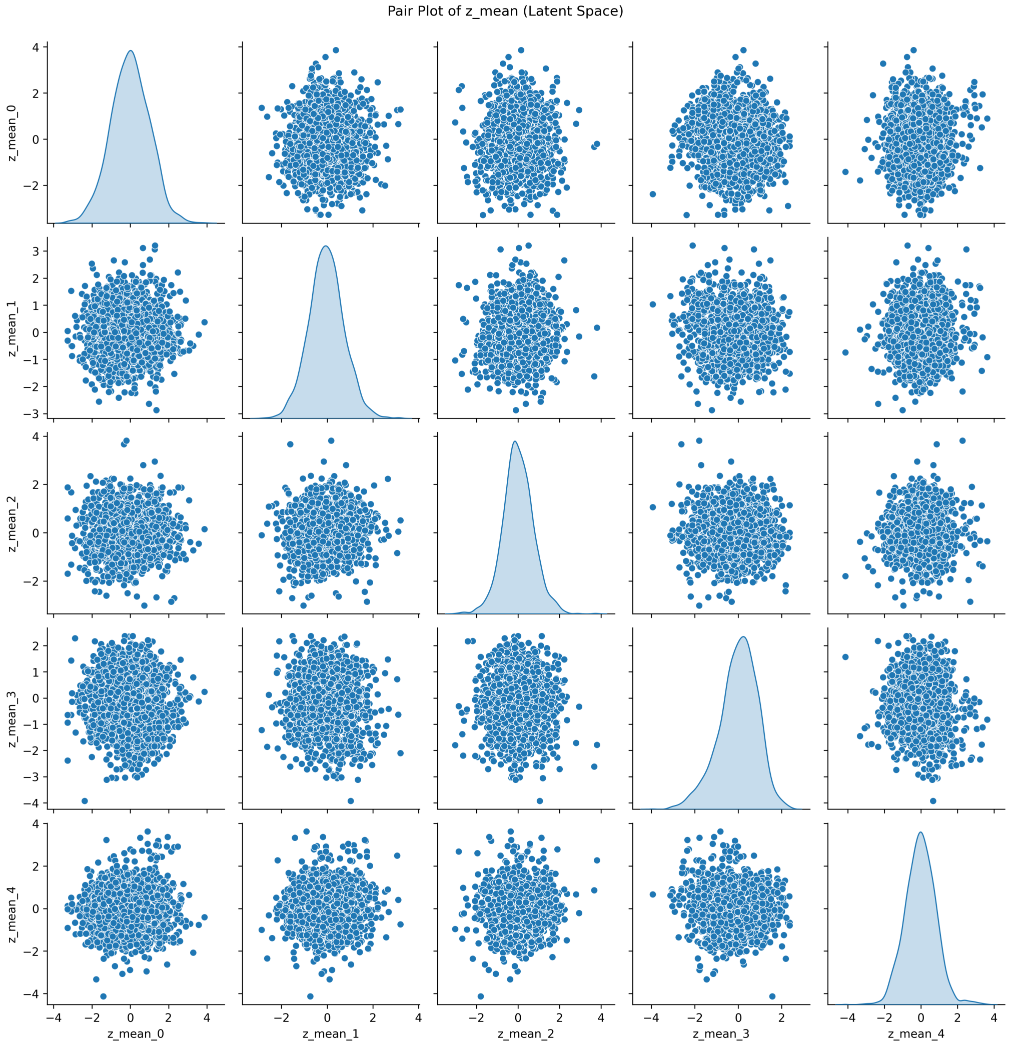

Figure 16: Pairwise scatter plots of latent dimensions for GenR baseline model ($d = 5$), showing approximate Gaussian structure and minimal correlation across dimensions.

Finally, the interpretability of the latent space is examined by identifying the anatomical variation encoded in each latent dimension. This is achieved using the brain visualization approach described in Section 3.2.4, which uses the generative capabilities of the variational autoencoder. For each latent variable, an outlier value ($z = \pm 3$) is assigned while keeping all other dimensions fixed at zero. Covariates such as age and sex are held constant (e.g., age at the standardized mean, sex set to male). The resulting reconstructions are subtracted to produce difference maps, highlighting which brain regions are most affected by changes in the corresponding latent dimension. The visualizations in Figure 17 are based on a model with a latent dimensionality of $d = 4$. The figure shows that each latent factor encodes specific information:

- Z_0 primarily affects volume in the midbrain and particularly in frontal regions.
- Z_1 captures differences involving the inferior brain areas, especially the temporal and occipital lobes.
- Z_2 shows characteristics of a global volume modulator, affecting most brain areas in a relatively uniform manner.
- Z_3 shows a contrast between frontal and temporal regions versus posterior regions, primarily the occipital lobe.

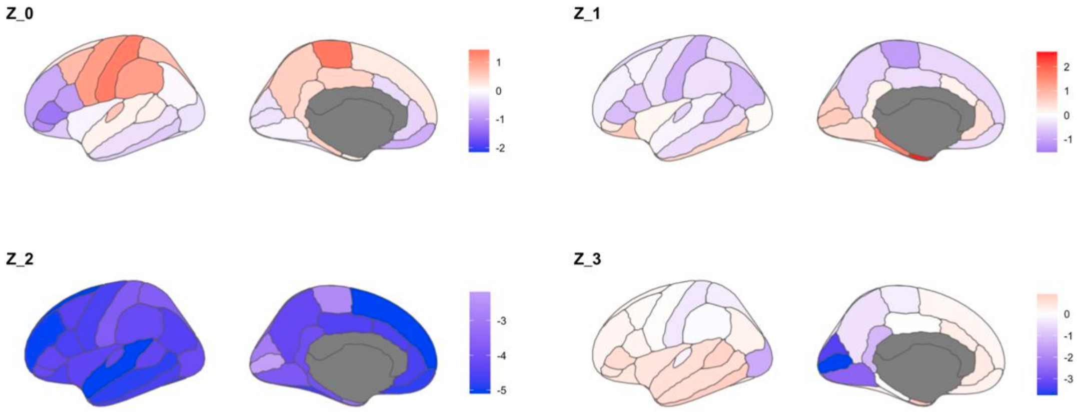

Figure 17: Brain difference maps for each latent dimension ($d = 4$) trained on the GenR dataset. Reconstructions corresponding to $z = +3$ and $z = -3$ were generated per dimension and subtracted to visualize the brain regions most influenced by each latent factor.

Together, these results establish a clear understanding of the baseline model's behavior across datasets and latent sizes. They provide important reference points for interpreting the effectiveness of more advanced modeling strategies introduced in later experiments.

## 4.2 Covariate Modeling Methods

This section presents the evaluation results of various covariate modeling strategies for the GenR and HBN datasets. The detailed descriptions and mathematical formulations of all evaluated methods are provided in Section 3.1.2. Tables 6 and 7 show the performance metrics for each method when applied to the GenR and HBN datasets, respectively, at a fixed latent dimensionality of $d = 5$. Results for additional latent dimensionalities are provided in the supplementary results in the Appendix, in Sections A.1 and A.2.

Table 6 presents the results for covariate modeling methods applied to the GenR dataset, which includes age and sex as covariates. The baseline model serves as a reference point. The best-performing methods in terms of reducing covariate dependence in the latent space are CM-3 (Encoder-Decoder Modeling, or cVAE), CM-8 (cVAE with an MMD Loss Term), and CM-9 (cVAE with an HSIC Loss Term). These methods show the lowest mutual information scores and lowest prediction accuracy for sex, alongside increased age prediction error, suggesting that the latent representation is less informative of the covariates, as intended. CM-9, for example, reduces the mutual information for sex to 0.02 and for age to 0.06 while slightly improving KL divergence compared to the baseline. In contrast, CM-4 (Covariate Reconstruction) and CM-5 (Conditional Loss Term) show significantly worse performance with respect to covariate invariance. These methods exhibit the highest mutual information scores and lowest age prediction errors, clearly indicating that the covariate information is encoded in the latent space. This is expected behavior, as these methods explicitly encourage the model to learn covariate representations.

Table 6: Evaluation of different covariate modeling methods when modeling age and sex at latent dimensionality $d = 5$ trained using the GenR dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Normative Alignment |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI Age | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.39 | 0.61 | 1.77 | 0.17 | 0.63 | 0.08 | 0.88 |
| CM-1: Decoder-only | 0.38 | 0.61 | 1.77 | 0.12 | 0.63 | 0.06 | 0.84 |
| CM-2: Encoder-only | 0.39 | 0.61 | 1.76 | 0.13 | 0.61 | 0.07 | 0.88 |
| CM-3: Encoder-Decoder (cVAE) | 0.39 | 0.61 | 1.78 | 0.04 | 0.58 | 0.01 | 0.82 |
| CM-4: Covariate reconstruction | 0.39 | 0.61 | 1.75 | 0.59 | 0.63 | 0.07 | 0.93 |
| CM-5: Conditional loss term | 0.39 | 0.61 | 1.75 | 0.56 | 0.65 | 0.11 | 0.93 |
| CM-6: Adversarial loss term | 0.39 | 0.61 | 1.76 | 0.21 | 0.64 | 0.08 | 0.89 |
| CM-7: Conditional Adversarial loss term | 0.38 | 0.61 | 1.79 | 0.06 | 0.58 | 0.04 | 0.82 |
| CM-8: MMD loss term (VFAE) | 0.38 | 0.61 | 1.79 | 0.03 | 0.58 | 0.04 | 0.83 |
| CM-9: HSIC loss term (HCV) | 0.38 | 0.61 | 1.78 | 0.06 | 0.59 | 0.02 | 0.82 |
| CM-10: Disentangled subspace | 0.38 | 0.61 | 1.77 | 0.10 | 0.62 | 0.10 | 0.84 |

Table 7 presents the corresponding results for the HBN dataset, where age, sex, and site are all treated as covariates. Similar to the GenR results, CM-3, CM-8, and CM-9 again show the strongest performance in achieving covariate invariance. These methods consistently report low mutual information values across all three covariates and reduced prediction accuracies, indicating reduced dependence on demographic and site-related information in the latent space. CM-8, for instance, achieves MI scores of 0.06 for age, 0.05 for sex, and 0.03 for site while maintaining high reconstruction quality. As with GenR, CM-4 and CM-5 perform the worst in terms of covariate invariance, which aligns with expectations. These methods are specifically designed to encode covariate information into the latent space, thereby revealing strong dependencies between the covariates and the learned latent representation. Overall, the evaluation confirms that CM-3 (cVAE), CM-8 (cVAE + MMD), and CM-9 (cVAE + HSIC) are effective strategies for promoting covariate invariance in multivariate normative models without sacrificing reconstruction quality or alignment with the normative prior. These methods represent promising candidates for scenarios where removing demographic and acquisition-related confounds from the latent space is critical.

Table 7: Evaluation of different covariate modeling methods when modeling age, sex, and site at latent dimensionality $d = 5$ trained using the HBN dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI Age | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.38 | 0.62 | 0.81 | 0.17 | 0.65 | 0.10 | 0.46 | 0.06 | 0.72 |
| CM-1: Decoder-only | 0.37 | 0.63 | 0.96 | 0.13 | 0.64 | 0.13 | 0.46 | 0.07 | 0.70 |
| CM-2: Encoder-only | 0.37 | 0.63 | 0.85 | 0.19 | 0.66 | 0.07 | 0.48 | 0.08 | 0.72 |
| CM-3: Encoder-Decoder (cVAE) | 0.37 | 0.63 | 1.12 | 0.06 | 0.58 | 0.02 | 0.45 | 0.09 | 0.68 |
| CM-4: Covariate reconstruction | 0.37 | 0.63 | 0.31 | 0.52 | 0.68 | 0.09 | 0.47 | 0.08 | 0.77 |
| CM-5: Conditional loss term | 0.37 | 0.63 | 0.32 | 0.52 | 0.68 | 0.14 | 0.53 | 0.14 | 0.80 |
| CM-6: Adversarial loss term | 0.37 | 0.63 | 0.82 | 0.21 | 0.65 | 0.10 | 0.47 | 0.06 | 0.75 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 1.09 | 0.09 | 0.59 | 0.05 | 0.48 | 0.08 | 0.72 |
| CM-8: MMD loss term (VFAE) | 0.37 | 0.63 | 1.10 | 0.06 | 0.59 | 0.05 | 0.45 | 0.03 | 0.68 |
| CM-9: HSIC loss term (HCV) | 0.37 | 0.63 | 1.11 | 0.06 | 0.58 | 0.02 | 0.45 | 0.07 | 0.68 |
| CM-10: Disentangled subspace | 0.37 | 0.63 | 0.95 | 0.12 | 0.67 | 0.12 | 0.44 | 0.05 | 0.71 |

The results presented in the appendix (Sections A.1 and A.2) allow for an inspection of how different covariate modeling methods perform across a range of latent dimensionalities ($d = 1, 2, 3, 4, 5, 8, 12, 16$). This section describes the patterns and trends found across these dimensions.

Across all methods and dimensionalities, reconstruction quality remains largely unaffected. Both MSE and $R^2$ scores are highly consistent across methods and dimension sizes, reaffirming that the different covariate modeling strategies do not hinder the model's ability to reconstruct brain features. This suggests that tradeoffs between covariate invariance and reconstruction performance are minimal, and efforts to promote invariance do not substantially degrade the reconstructions.

To complement these quantitative metrics, two models showing the most significant deviation from the baseline were evaluated via 2D t-SNE projections of their $d = 5$ latent spaces using the GenR test set. First, CM-5 (Conditional Loss Term), which explicitly reconstructs covariates and therefore enforces their encoding, is considered. Figure 18 shows its t-SNE embedding color-coded by z-scored age (continuous gradient) and sex (discrete labels). A strong separation by both age and sex is immediately apparent, exceeding the baseline dependence illustrated in Figure 14, in line with CM-5's higher MI, higher sex prediction accuracy, and lower age prediction MSE. Next, CM-8 (cVAE + MMD loss), one of the methods achieving the most significant reduction in covariate invariance according to the metrics, is examined. Its t-SNE embedding in Figure 19, again color-coded by age and sex, reveals points that are far less organized by covariate, matching the obtained results. These visuals confirm that the quantitative scores can correctly show the covariate dependence in the latent space, underscoring the impact of covariate modeling.

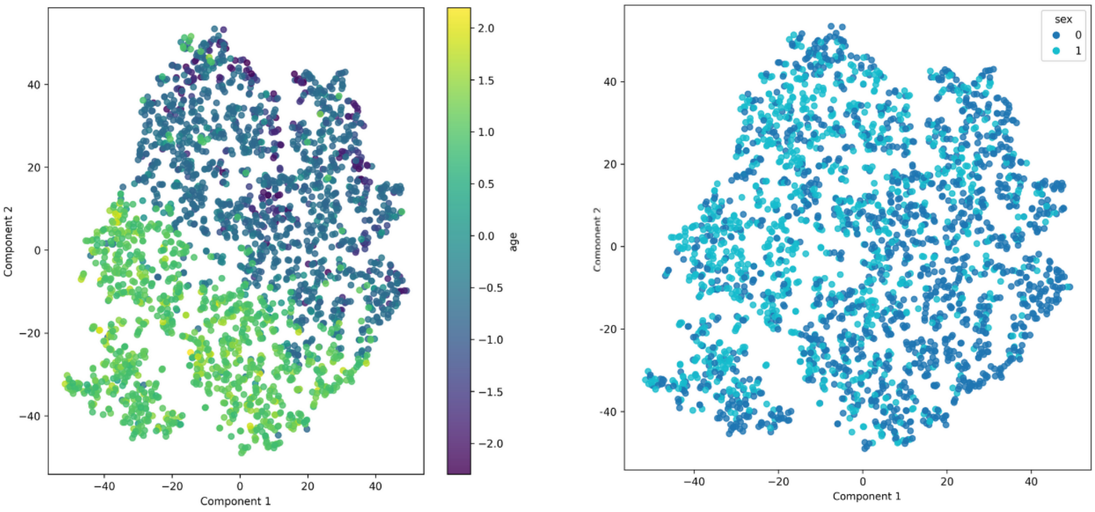

Figure 18: 2D t-SNE projection of the latent space of CM-5 (Conditional Loss Term, $d = 5$), trained on GenR, color-coded by covariates: age (left) and sex (right).

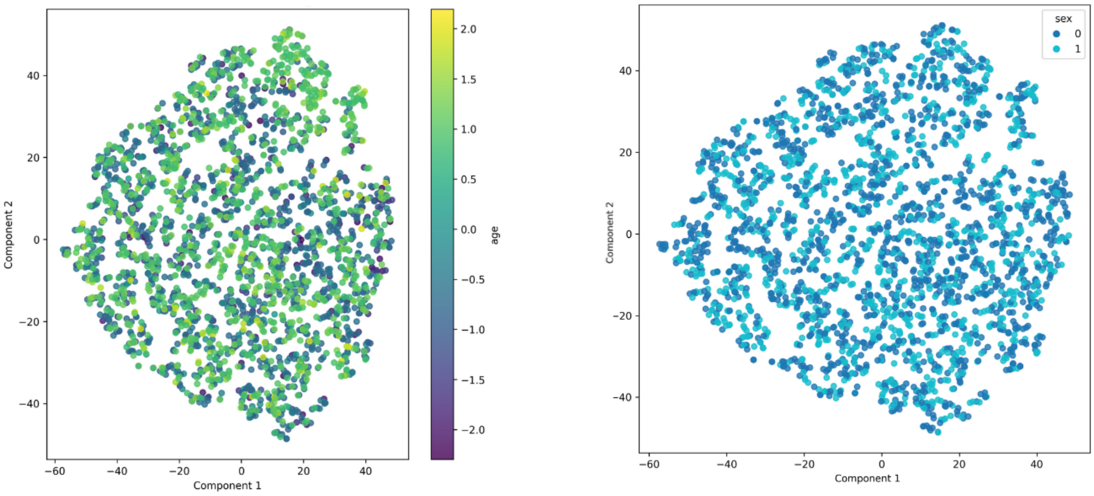

Figure 19: 2D t-SNE projection of the latent space of CM-8 (cVAE + MMD loss, $d = 5$), trained on GenR, color-coded by covariates: age (left) and sex (right).

Overall, CM-3 (Encoder-Decoder/cVAE), CM-8 (cVAE + MMD), and CM-9 (cVAE + HSIC) demonstrate the strongest performance across GenR and HBN in terms of reducing covariate-related information in the latent space, while preserving high reconstruction quality and normative alignment. These methods are especially effective at lower to moderate latent dimensions ($d = 3$ to $d = 8$), making them strong candidates for building robust and invariant multivariate normative models. In contrast, methods like CM-4 and CM-5 serve as negative controls, confirming that explicit covariate reconstruction leads to strong entanglement with the latent code. The results demonstrate the impact of various covariate modeling methods.

## 4.3 Site Accommodation

To assess the ability of multivariate normative models to accommodate site-related variance, three modeling strategies were evaluated on a balanced, multi-site dataset composed of data from GenR and HBN (see Section 3.4.4). The selected models include (1) the baseline model, (2) the baseline model trained on ComBat-harmonized data, and (3) the covariate-aware CM-3 and CM-8 models. Evaluation metrics for the model trained with a latent dimensionality of $d = 5$ are presented in Table 8. Results for additional latent dimensionalities are provided in Appendix A.3.

Table 8: Site accommodation evaluation of different covariate modeling methods and ComBat harmonization, conditioned on age, sex, and site at latent dimensionality $d = 5$.

| **Model** | **Test Site** | **Reconstruction Quality** | | **Invariance (age)** | | **Invariance (sex)** | | **Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.42 | 0.60 | 0.77 | 0.26 | 0.61 | 0.16 | 0.36 | 0.09 | 0.62 |
| | RUBIC | 0.38 | 0.63 | 1.02 | 0.14 | 0.64 | 0.16 | 0.37 | 0.09 | 0.77 |
| | CBIC | 0.38 | 0.60 | 0.83 | 0.07 | 0.58 | 0.04 | 0.58 | 0.27 | 0.76 |
| | **Average** | 0.39 | 0.61 | 0.87 | 0.16 | 0.61 | 0.12 | 0.44 | 0.15 | 0.72 |
| ComBat Harmonized Baseline | GenR | 0.42 | 0.60 | 0.76 | 0.26 | 0.61 | 0.15 | 0.30 | 0.09 | 0.62 |
| | RUBIC | 0.38 | 0.63 | 0.87 | 0.11 | 0.63 | 0.10 | 0.28 | 0.03 | 0.77 |
| | CBIC | 0.39 | 0.59 | 0.86 | 0.29 | 0.66 | 0.12 | 0.33 | 0.10 | 0.73 |
| | **Average** | 0.40 | 0.61 | 0.83 | 0.22 | 0.63 | 0.12 | 0.30 | 0.07 | 0.71 |
| CM-3 | GenR | 0.42 | 0.60 | 0.99 | 0.05 | 0.54 | 0.10 | 0.39 | 0.08 | 0.63 |
| | RUBIC | 0.38 | 0.63 | 1.11 | 0.08 | 0.56 | 0.07 | 0.33 | 0.19 | 0.69 |
| | CBIC | 0.41 | 0.58 | 1.27 | 0.13 | 0.48 | 0.04 | 0.39 | 0.05 | 0.58 |
| | **Average** | 0.40 | 0.60 | 1.12 | 0.09 | 0.53 | 0.07 | 0.37 | 0.11 | 0.63 |
| CM-8 | GenR | 0.42 | 0.60 | 0.99 | 0.05 | 0.57 | 0.11 | 0.40 | 0.10 | 0.63 |
| | RUBIC | 0.38 | 0.63 | 1.14 | 0.13 | 0.57 | 0.08 | 0.36 | 0.16 | 0.69 |
| | CBIC | 0.41 | 0.57 | 1.20 | 0.19 | 0.48 | 0.04 | 0.42 | 0.02 | 0.56 |
| | **Average** | 0.40 | 0.60 | 1.11 | 0.12 | 0.54 | 0.08 | 0.39 | 0.09 | 0.63 |

When examining the results of Table 8 (latent dimensionality $d = 5$), the first notable finding is that the ComBat-harmonized baseline model outperforms all other models in terms of site prediction accuracy and MI related to site. Lower values on both metrics indicate that less site-specific information is retained in the latent space. This trend is consistent across all three sites and is reflected in the average site prediction accuracy of 0.30 and MI of 0.07, compared to 0.44 and 0.15 for the non-harmonized baseline. The CM-3 and CM-8 models also show improved site invariance compared to the baseline model, although not as strong as the ComBat harmonized model.

However, the influence of different latent space dimensionalities is significant in this experiment. At lower latent dimensionalities, both the CM-3 and CM-8 models outperform the baseline model and the ComBat harmonized model in terms of site invariance. As the latent dimensionality increases (e.g., $d \geq 5$), the ComBat-harmonized model begins to outperform the covariate-aware models. At even higher dimensions, the baseline model eventually surpasses both CM-3 and CM-8 in terms of site prediction accuracy, indicating that these covariate-aware models lose their advantage as model capacity increases. A visualization of the average prediction accuracy for the site covariate across latent dimensionalities is presented in Figure 20. The plot confirms that the advantage of covariate modeling methods diminishes as dimensionality increases, while the ComBat-harmonized baseline remains consistently low across larger latent space sizes.

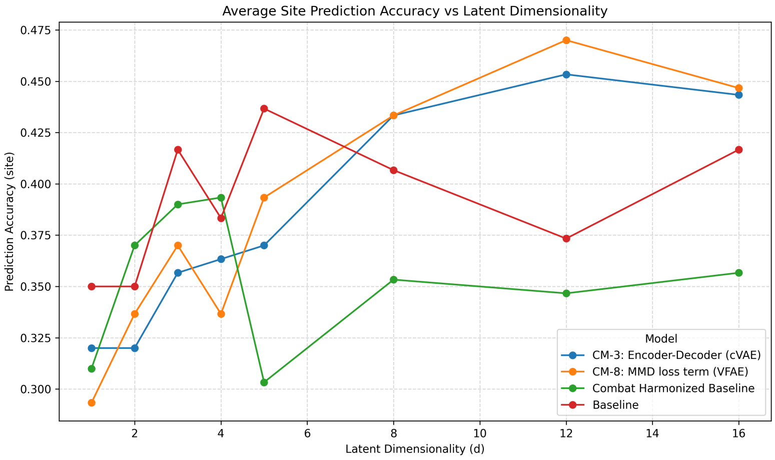

Figure 20: Average prediction accuracy for site across all test sets (GenR, RUBIC, CBIC) as a function of latent dimensionality. Each line corresponds to one model configuration. Lower is better.

Interestingly, the comparison across other covariates reveals that the baseline and ComBat models perform similarly in terms of age and sex invariance. In contrast, the CM-3 and CM-8 models achieve substantially better invariance, even in a multi-site setup with unseen sites. These results highlight the strength of covariate-aware models in learning disentangled representations for age and sex, even when site variability is present. Normative alignment also improves when using CM-3 and CM-8. In contrast, the ComBat-harmonized baseline performs comparably to the baseline, showing limited improvement from harmonization in this regard.

## 4.4 Brain Age Estimation

For the brain age estimation task, the CM-8 covariate method was used with latent space dimensionality $d = 5$. This model extends the CM-3 model by incorporating an MMD loss to further mitigate the influence of covariates in the latent space. The brain age model is evaluated on both the GenR and HBN datasets. For each, a different scaling factor was determined to ensure biologically plausible predictions of brain age.

For the GenR dataset, a scaling factor of $w = 0.5547$ was found through repeated experiments. This value resulted in an MAE of approximately one year between the predicted brain age and chronological age across the test set. The average chronological age in the GenR test set was 11.48 years, and the average predicted brain age was 11.45 years, indicating that the brain age model preserves the mean of the age distribution with a mean average shift of 1 year across all subjects.

For the HBN dataset, a scaling factor of $w = 0.3672$ was found. The average chronological age in the HBN test set was 10.39 years, and the average predicted brain age was 10.41 years. As in the GenR results, the predicted brain age closely mirrors the chronological age. The smaller scaling factor is likely a result of the difference in data distribution between the GenR and HBN test sets. The standard deviations of age in the GenR and HBN test sets were 2.37 and 3.45, respectively.

To assess the effect of modifying the architecture to include brain age estimation, various properties of the normative model are evaluated. These include reconstruction quality, covariate invariance (age, sex, and site for HBN), and normative alignment. The evaluation results for GenR are shown in Table 9 and for HBN in Table 10.

For GenR, the inclusion of the BAG component did not significantly alter model behavior compared to the CM-8 variant without brain age estimation. The metrics remain comparable across reconstruction loss, covariate invariance, and KL divergence. This suggests that the normative structure and disentanglement achieved by CM-8 are preserved in the brain age model.

Table 9: Evaluation of the brain age model using the GenR dataset (latent dimensionality $d = 5$), compared with the baseline and covariate-controlled CM-8 models.

| Model | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Normative Alignment |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI Age | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.39 | 0.61 | 1.77 | 0.17 | 0.63 | 0.08 | 0.88 |
| CM-8: MMD loss term | 0.38 | 0.61 | 1.79 | 0.03 | 0.58 | 0.04 | 0.83 |
| Brain Age Model | 0.38 | 0.62 | 1.78 | 0.05 | 0.57 | 0.02 | 0.76 |

For HBN, the results are largely consistent with those of GenR. A slight decrease in prediction error for chronological age is observed in the brain age model compared to CM-8, alongside a slight increase in MI for the age covariate. This may suggest leakage of age-related information from the explicitly modeled BAG component into the other latent dimensions. However, the differences are small and do not significantly affect the overall disentanglement or normative alignment. Other metrics, including sex and site invariance, as well as KL divergence, remain comparable across models.

Table 10: Evaluation of the brain age model using the HBN dataset (latent dimensionality $d = 5$), compared with the baseline and covariate-controlled CM-8 models.

| Model | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI Age | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.38 | 0.62 | 0.81 | 0.17 | 0.65 | 0.10 | 0.46 | 0.06 | 0.72 |
| CM-8: MMD loss term | 0.37 | 0.63 | 1.10 | 0.06 | 0.59 | 0.05 | 0.45 | 0.03 | 0.68 |
| Brain Age Model | 0.37 | 0.63 | 1.00 | 0.11 | 0.61 | 0.04 | 0.43 | 0.06 | 0.60 |

## 4.5 Univariate Comparison

To evaluate the differences between univariate and multivariate normative modeling approaches, a multivariate model with latent dimensionality $d = 5$ was trained using the CM-8 covariate modeling method and compared against a univariate HBR model. Both models were trained on the same standardized dataset, ensuring consistency in preprocessing and input distributions.

Figure 21 presents a scatter plot comparing the averaged z-scores derived from the univariate normative model (UNM) and the multivariate normative model (MNM) for each subject. Each point represents one subject, with the UNM average z-score on the x-axis and the MNM average z-score on the y-axis. The dashed diagonal line indicates the identity line ($y = x$), where both models produce the same z-score.

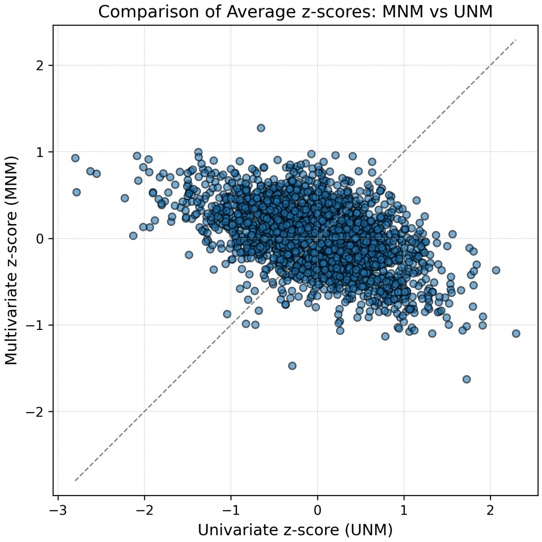

Figure 21: Comparison of average z-scores between multivariate and univariate normative models.

No clear relationship is observed between the z-scores obtained from the two models. This discrepancy could be attributed to differences in the interpretation and construction of the z-scores. In the univariate model, each feature is modeled independently, and deviations reflect direct differences from the group mean. In contrast, the multivariate model captures the joint distribution of multiple features and reduces the high-dimensional data to a lower-dimensional latent representation. As such, a deviation in the multivariate space may not align with deviations seen in any single univariate feature. Furthermore, interesting deviations might be lost, especially in the univariate case, because all the z-scores are averaged over a total of 34 features. Not all features contribute equally in explaining notable deviations in test subjects, and this negatively impacts the comparison to the multivariate model.

# 5 Discussion

This study aimed to systematically assess VAE-based multivariate normative modeling techniques, with a focus on their ability to model covariates. This was done by using two distinct, high-quality neuroimaging datasets, namely HBN and GenR. This work aimed to provide new insights into the properties and limitations of VAE-based multivariate normative models across various experimental setups. In addition, this work developed the necessary training and evaluation tools to perform such experiments in a consistent and reproducible manner. These tools form the basis of an open experimental framework that enables future research to define VAE-based multivariate normative models.

The general findings of this work show that VAE-based multivariate normative models are a promising approach for normative modeling. The learned latent representations follow the assumed normal prior distribution and successfully capture distinct, meaningful aspects of brain anatomy. Brain region reconstructions from generated latent vectors confirm that each latent dimension encodes unique anatomical patterns, such as localized volume effects or global modulation. However, the baseline model retains covariate information in its latent space. This is evident both quantitatively, through MI scores and prediction accuracy metrics, and qualitatively via t-SNE plots. Such entanglement of covariate information (e.g., age or sex) can obscure true deviations and compromise the interpretability of deviations from the norm. This finding underscores the importance of explicitly modeling covariates when building normative models.

To address this, a variety of covariate-aware modeling strategies were evaluated. Among these, three models, namely CM-3 (cVAE), CM-8 (cVAE + MMD), and CM-9 (cVAE + HSIC), consistently outperformed the baseline and other approaches in reducing covariate information in the latent space without sacrificing reconstruction quality or normative alignment. These methods achieved the lowest mutual information and prediction accuracy scores for covariates across both datasets. These results were further supported by t-SNE projections, showing minimal covariate structure in the latent space for these models. Other strategies, such as CM-4 (covariate reconstruction) and CM-5 (conditional loss term), which explicitly enforce the encoding of covariates, showed poor performance in achieving invariance. These served as valuable control experiments, confirming that the latent space became entangled with covariates as expected.

The study also examined whether VAE-based models could handle batch effects, particularly those related to site variation. The baseline model already showed acceptable site invariance, especially on the HBN dataset, as suggested by relatively low MI scores and t-SNE plots with no distinct site clustering. However, when combining GenR and HBN, additional variance is introduced, and model performance varies depending on the strategy. Interestingly, the benefit of using models with explicit covariate modeling was most substantial at lower latent dimensionalities. As model capacity increased (e.g., $d \geq 5$), the ComBat harmonized baseline began to outperform the covariate-aware models in terms of site invariance. This suggests that larger latent spaces could naturally absorb more variability, including unwanted covariate effects. This highlights the importance of carefully considering latent dimensionality when designing models for multi-site studies.

As a proof-of-concept clinical application, a brain age estimation model was developed using the CM-8 covariate-aware model. This task demonstrates the multivariate normative model's ability to predict a biologically plausible brain age variable while maintaining a covariate invariant latent space for age and other covariates. The inclusion of an explicit scaling factor ensured that the outputs were biologically plausible. Importantly, brain age prediction did not compromise the core normative modeling properties of the VAE. Finally, the multivariate model was compared to a traditional univariate normative modeling approach. This appeared to be a very challenging task since the models are fundamentally different in design. While the univariate model estimates deviations for individual features independently, the multivariate model captures complex joint deviations across brain measures. The univariate model provides a straightforward measure of deviation per feature. In contrast, the multivariate model could also offer insights into why a deviation occurs by analyzing latent dimensions. These differences might explain why the z-scores from both models were uncorrelated.

This work provides strong evidence that VAE-based multivariate normative models, particularly those using covariate-aware strategies like CM-3, CM-8, and CM-9, are well-suited for modeling normative brain variation. However, several limitations are identified, some of which could be investigated in future work.

Although the HBN and GenR are high-quality datasets, some concerns related to the data were identified in the work. The GenR dataset is structured in waves, which results in clustered age distributions around specific ages. As a result, certain age ranges, such as between 12 and 13 years, are not represented in the data. The datasets are also primarily composed of healthy children. While this is ideal for training normative models, it limits the direct evaluation of the model's sensitivity to clinical pathology. Future work should include patient data with clear diagnostic labels to test the model's ability to detect clinically meaningful deviations. This also raises the question of what contributes to a 'healthy' individual, which, in itself, is a challenging question that requires clinical expertise and remains debatable even then. The age range could be considered in future work. The datasets in this work consist of children aged 5 to 22. This type of data is not often used in other related VAE-based multivariate normative modeling literature. Furthermore, normative modeling for this age group is more difficult than for adults. One reason is the greater variation in brain region sizes, as children's brains are still developing. In future research, this model could also be applied to data from adults.

Ethnicity is also a known factor that influences brain structure, but it was not formally modeled in this study. Although the GenR and HBN datasets include ethnically diverse participants, the current analysis did not address potential biases related to demographic variability beyond age and sex. Another limitation is that the current evaluation of site variation was conducted on only three sites. The analysis could benefit from a more diverse range of available sites. This could be useful for detecting different types of site-related patterns, such as scanner-specific differences or population-based differences.

Future studies could also investigate how various subsets of brain features, such as cortical thickness, surface area, volume, subcortical regions, or combinations, affect model performance. Because these features capture different aspects of brain structure, their influence on reconstruction, normative deviation, and covariate sensitivity may vary. Understanding these differences could help design models more effectively for specific clinical tasks. Also, the individual treatment of features from the left and right hemispheres can be considered. This work used the average of both hemispheres to reduce the feature dimensionality. Given larger datasets, a larger feature space can also be considered, with, for example, more types of brain measures and without averaging measures over both hemispheres.

In the brain age estimation task, explicit scaling was needed to produce biologically plausible age estimates. In future models, this scaling factor should ideally be learned implicitly. Additionally, brain age predictions could be validated against other clinical markers or development scores to establish clinical relevance. Correlating brain age gap (BAG) with diagnoses or behavioral outcomes could reveal patterns of accelerated or delayed neurodevelopment.

Finally, in this study, the primary focus was on researching multivariate normative models from a methodological perspective. A first step towards clinical application was made through the brain estimation task. However, to better understand the strengths and difficulties of the modeling approach mentioned in this study, the model should be evaluated in a broader range of clinically relevant tasks and datasets. In addition, future research should focus on model explainability by improving the interpretability and transparency of multivariate normative models. Developing new methods that make sure clinicians understand and trust the model's output is necessary to secure its practical use in real-world healthcare settings.

# 6 Conclusion

This study used VAE-based multivariate normative models to investigate methods for modeling covariate effects. Many of the properties of VAE-based multivariate normative models remain unexplored, particularly in relation to modeling covariates. Additionally, the problem of accommodating batch effects was explored and compared to an existing data harmonization method. This was done by applying and comparing the best-performing covariate modeling methods to a baseline model and a more traditional preprocessing method, ComBat data harmonization. Furthermore, given that a model was created that successfully captures brain characteristics while accounting for important covariates, such as age and sex, this study examined a proof-of-concept for a downstream task. Brain age estimation, being inherently multivariate, was chosen as the task of interest. Lastly, the performance of the VAE-based model was compared to that of a univariate alternative.

First, when evaluating how different covariate modeling techniques affect core properties of VAE-based multivariate normative models, the effects of covariate methods are clearly demonstrated. For example, the baseline VAE alone does not sufficiently account for the covariates, as it shows some dependence between the covariate and the learned latent representation. This is noticeable both through the use of covariate invariance metrics and qualitative analysis of the latent space. Different covariate modeling methods were tested to learn about their effects on the model. Most notable are the encoder-decoder architecture, CM-3, also known as a cVAE, which is used in related works, the cVAE combined with an MMD loss term, CM-8, and the cVAE with an HSIC loss term, CM-9. All three models outperformed the baseline model. The differences among these covariate modeling methods were small, indicating that the cVAE, on its own, is a powerful covariate modeling method. Although the three models showed promising results, it is worth noting that these methods are not the most advanced available. Future work can explore more advanced solutions for covariate modeling.

Second, this study investigated to what extent VAE-based multivariate normative models can accommodate site-related variation. The baseline VAE model, when applied to the HBN sites, already showed good performance, demonstrating its ability to accommodate batch effects. When combining the GenR and HBN datasets, the baseline model showed similar site prediction performance as when trained on the multisite HBN dataset alone. Both the CM-3 and CM-8 models, as well as the model using ComBat harmonized data, outperform the baseline model when the latent space dimensionality is low. However, when the latent space dimensionality is increased (5 and higher), the ComBat harmonized model outperforms the covariate-aware models. Furthermore, when the latent space dimensionality is increased further (8 and higher), even the baseline model outperforms the covariate-aware models. These results show the potential of covariate-aware methods in small latent spaces. Additionally, it raises the question of methods that can be designed that also perform well in higher-dimensional latent spaces.

Finally, this study compared the VAE-based multivariate normative model to a univariate normative modeling approach in terms of their ability to capture complex normative patterns in brain data. It is challenging to compare these modeling approaches directly. A key strength of the multivariate normative model is that it does more than reconstruct brain measures. It enables us to understand the biological factors that influence the brain. Because the model is fitted to all brain measures simultaneously, the latent dimensions can capture coherent patterns across all brain measures. Inspecting these latent dimensions tells us why individuals deviate, not just by how much they deviate. By contrast, a univariate normative model essentially learns little about the deviations. The covariate effects are regressed out of the model, and the model only provides an estimate of the spread of the brain metrics. This means the univariate model is primarily appropriate for specifying how atypical a single measure is. Lastly, this work examined the effect of the multivariate model on a clinical application, specifically brain age estimation. The results show that the multivariate model has great potential, as it learns brain age in a controlled way while keeping the latent space age-invariant. However, the current brain age model has some issues, for example, with scaling the brain age gap so that the brain age falls within a biologically plausible range. This can be corrected by using a scaling factor. Ideally, a model learns this scaling implicitly. Future work can explore more accurate brain age models based on this research. Additionally, future work can examine different clinical tasks to assess the strengths and weaknesses of the VAE-based multivariate normative model in various contexts.

# References

[1] A. F. Marquand, S. M. Kia, M. Zabihi *et al.*, "Conceptualising mental disorders as deviations from normative functioning," *Molecular Psychiatry*, vol. 24, pp. 1415–1424, 2019.

[2] T. Wolfers, N. T. Doan, T. Kaufmann *et al.*, "Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models," *JAMA Psychiatry*, vol. 75, no. 11, pp. 1146–1154, 2018.

[3] S. Rutherford, P. Barkema, I. F. Tso *et al.*, "Evidence for embracing normative modeling," *eLife*, vol. 12, p. e85082, 2023.

[4] "A health professional's guide for using the new who growth charts," *Paediatrics & Child Health*, vol. 15, no. 2, pp. 84–98, 2010.

[5] S. Abbasi-Sureshjani, R. Raumanns, B. E. J. Michels, G. Schouten, and V. Cheplygina, "Risk of training diagnostic algorithms on data with demographic bias," in *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, ser. Lecture Notes in Computer Science, vol. 12444. Springer, 2020, pp. 183–192.

[6] K. A. Weber, Z. Teplin, T. D. Wager *et al.*, "Confounds in neuroimaging: A clear case of sex as a confound in brain-based prediction," *Frontiers in Neurology*, vol. 13, p. 960760, 2022.

[7] K. Sanghavi, J. Ribbing, J. Rogers *et al.*, "Covariate modeling in pharmacometrics: General points for consideration," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 13, 2024.

[8] A. F. Marquand, I. Rezek, J. Buitelaar, and C. F. Beckmann, "Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies," *Biological Psychiatry*, vol. 80, no. 7, pp. 552–561, 2016.

[9] S. Rutherford, S. M. Kia, T. Wolfers *et al.*, "The normative modeling framework for computational psychiatry," *Nature Protocols*, vol. 17, pp. 1711–1734, 2022.

[10] J. Rokicki, T. Wolfers, W. Nordhøy *et al.*, "Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders," *Human Brain Mapping*, vol. 42, no. 6, pp. 1714–1726, Apr. 2021.

[11] S. A. Valizadeh, J. Hänggi, S. Mérillat, and L. Jäncke, "Age prediction on the basis of brain anatomical measures," *Human Brain Mapping*, vol. 38, no. 2, pp. 997–1008, Feb. 2017.

[12] S. Kumar, P. R. O. Payne, and A. Sotiras, "Normative modeling using multimodal variational autoencoders to identify abnormal brain volume deviations in alzheimer's disease," *Proceedings of SPIE*, vol. 12465, p. 1246503, 2023, pMID: 38130873, PMCID: PMC10731988.

[13] L. Aguila, J. Chapman, M. Janahi, and A. Altmann, "Conditional vaes for confound removal and normative modelling of neurodegenerative diseases," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, ser. Lecture Notes in Computer Science, vol. 13431. Springer, 2022, pp. 435–445.

[14] L. Aguila, J. Chapman, and A. Altmann, "Multi-modal variational autoencoders for normative modelling across multiple imaging modalities," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, ser. Lecture Notes in Computer Science, vol. 14220. Springer, 2023, pp. 410–420.

[15] W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study," *Human Brain Mapping*, vol. 40, no. 3, pp. 944–954, 2019.

[16] W. Pinaya, C. Scarpazza, R. Garcia-Dias *et al.*, "Using normative modelling to detect disease progression in mild cognitive impairment and alzheimer's disease in a cross-sectional multi-cohort study," *Scientific Reports*, vol. 11, p. 95098, 2021.

[17] J. M. M. Bayer, R. Dinga, S. M. Kia *et al.*, "Accommodating site variation in neuroimaging data using normative and hierarchical bayesian models," *NeuroImage*, vol. 264, p. 119699, 2022.

[18] M. Reynolds, T. Chaudhary, M. E. Torbati *et al.*, "Combat harmonization: Empirical bayes versus fully bayes approaches," *NeuroImage: Clinical*, vol. 39, p. 103472, 2023.

[19] H. G. Schnack, "Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases)," *Schizophrenia Research*, vol. 214, pp. 34–42, 2019, machine Learning in Schizophrenia. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920996417306497

[20] J. Mourão Miranda, D. Hardoon, A. Marquand *et al.*, "Patient classification as an outlier detection problem: An application of the one-class support vector machine," *NeuroImage*, vol. 58, pp. 793–804, 2011.

[21] I. Rezek and C. F. Beckmann, "Models of disease spectra," arXiv preprint arXiv:1207.4674, 2012. [Online]. Available: https://arxiv.org/abs/1207.4674

[22] G. Ziegler, G. R. Ridgway, R. Dahnke, C. Gaser, and Alzheimer's Disease Neuroimaging Initiative, "Individualized gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects," *NeuroImage*, vol. 97, pp. 333–348, 2014.

[23] G. Erus, H. Battapady, T. D. Satterthwaite *et al.*, "Imaging patterns of brain development and their relationship to cognition," *Cerebral Cortex*, vol. 25, no. 6, pp. 1676–1684, 2015.

[24] X. Wang, R. Zhou, K. Zhao, A. Leow, Y. Zhang, and L. He, "Normative modeling via conditional variational autoencoder and adversarial learning to identify brain dysfunction in alzheimer's disease," in *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–4.

[25] S. Kumar and A. Sotiras, "Normvae: Normative modeling on neuroimaging data using variational autoencoders," *CoRR*, vol. abs/2110.04903, 2021. [Online]. Available: https://arxiv.org/abs/2110.04903

[26] V. Kumar, S. Zhang *et al.*, "Multimodal variational autoencoder for normative modeling of alzheimer's disease using structural mri," in *Proceedings of SPIE Medical Imaging*, 2023.

[27] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 173–10 196, 2023.

[28] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," 2016. [Online]. Available: https://arxiv.org/abs/1604.06737

[29] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," in *International Conference on Learning Representations*, 2016.

[30] Y. Ganin, E. Ustinova, H. Ajakan *et al.*, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[31] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[32] J.-P. Fortin, N. Cullen, Y. I. Sheline *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *NeuroImage*, vol. 167, pp. 104–120, 2018.

[33] S. M. Kia, H. Huijsdens, R. Dinga *et al.*, "Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture Notes in Computer Science. Springer, 2020, vol. 12902, pp. 207–217.

[34] J. E. Villalón-Reina, C. A. Moreau, T. M. Nir *et al.*, "Multi-site normative modeling of diffusion tensor imaging metrics using hierarchical bayesian regression," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, ser. Lecture Notes in Computer Science. Springer, 2022, vol. 13431, pp. 207–217.

[35] V. M. Bashyam, G. Erus, J. Doshi *et al.*, "Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14,468 individuals worldwide," *Brain*, vol. 143, no. 7, pp. 2312–2324, 2020.

[36] F. Hu, A. Lucas, A. A. Chen *et al.*, "Deepcombat: A statistically motivated, hyperparameter-robust, deep learning approach to harmonization of neuroimaging data," *Human Brain Mapping*, vol. 45, no. 11, p. e26708, 2024.

[37] N. Russkikh, D. V. Antonets, D. Shtokalo *et al.*, "Style transfer with variational autoencoders is a promising approach to rna-seq data harmonization and analysis," *Bioinformatics*, vol. 36, no. 20, pp. 5076–5085, 2020.

[38] J. H. Cole and K. Franke, "Predicting age using neuroimaging: Innovative brain ageing biomarkers," *Trends in Neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.

[39] J. H. Cole, R. E. Marioni, S. E. Harris, and I. J. Deary, "Brain age and other artificial intelligence–based neuromarkers: Theory, applications, and pitfalls," *Neuroscience & Biobehavioral Reviews*, vol. 84, pp. 45–56, 2018.

[40] K. Franke and C. Gaser, "Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and alzheimer's disease," *GeroPsych*, vol. 25, no. 4, pp. 235–245, 2012.

[41] T. Kaufmann, D. van der Meer, N. T. Doan *et al.*, "Common brain disorders are associated with heritable patterns of apparent aging of the brain," *Nature Neuroscience*, vol. 22, no. 10, pp. 1617–1623, 2019.

[42] H. G. Schnack, N. E. van Haren, M. Nieuwenhuis, H. E. Hulshoff Pol, W. Cahn, and R. S. Kahn, "Accelerated brain aging in schizophrenia: A longitudinal pattern recognition study," *American Journal of Psychiatry*, vol. 173, no. 6, pp. 607–616, 2016, pMID: 26917166. [Online]. Available: https://doi.org/10.1176/appi.ajp.2015.15070922

[43] R. M. Brouwer, J. Schutte, R. Janssen, D. I. Boomsma, H. E. Hulshoff Pol, and H. G. Schnack, "The speed of development of adolescent brain age depends on sex and is genetically determined," *Cerebral Cortex*, vol. 31, no. 2, pp. 1296–1306, 10 2020. [Online]. Available: https://doi.org/10.1093/cercor/bhaa296

[44] A.-M. G. de Lange, M. Anatuürk, J. Rokicki *et al.*, "Mind the gap: Performance metric evaluation in brain-age prediction," *Human Brain Mapping*, vol. 43, p. e25837, 2022.

[45] Y. Sun, H. Nolan, K. S. Button *et al.*, "Brain age prediction incorporates normative modeling to enhance clinical interpretability," *Nature Communications*, vol. 15, no. 1, p. 1251, 2024.

[46] I. Beheshti, S. Nugent, O. Potvin, and S. Duchesne, "Bias adjustment in neuroimaging-based brain age framework: A robust scheme," *NeuroImage: Clinical*, vol. 24, p. 102063, 2019.

[47] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[48] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 2623–2631.

[49] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1711.05101

[50] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[51] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114

[52] R. Lopez, J. Regier, M. I. Jordan, and N. Yosef, "Information constraints on auto-encoding variational bayes," 2018. [Online]. Available: https://arxiv.org/abs/1805.08672

[53] L. Breiman, "Random forests," vol. 45, no. 1, pp. 5–32.

[54] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[55] V. W. V. Jaddoe, C. M. van Duijn, A. J. van der Heijden *et al.*, "The generation r study: Design and cohort profile," *European Journal of Epidemiology*, vol. 21, no. 6, pp. 475–484, 2006.

[56] M. Kooijman, C. Kruithof, C. Duijn *et al.*, "The generation r study: Design and cohort update 2017," *European Journal of Epidemiology*, vol. 31, 2016.

[57] L. M. Alexander, J. Escalera, L. Ai *et al.*, "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Scientific Data*, vol. 4, p. 170181, 2017.

[58] B. Fischl, "Freesurfer," *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[59] P. M. Thompson, J. L. Stein, S. E. Medland *et al.*, "The enigma consortium: Large-scale collaborative analyses of neuroimaging and genetic data," *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 153–182, 2014.

[60] SURF.nl, "Snellius: The national supercomputer," https://www.surf.nl/en/services/snellius-the-national-supercomputer.

# A Supplementary Results

## A.1 Covariate Modeling Methods Results - GenR

Table 11: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 1$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.55 | 0.44 | 1.88 | 0.03 | 0.57 | 0.06 | 1.79 |
| CM-1: Decoder-only | 0.53 | 0.46 | 1.93 | 0.00 | 0.56 | 0.05 | 1.71 |
| CM-2: Encoder-only | 0.55 | 0.44 | 1.90 | 0.01 | 0.55 | 0.06 | 1.72 |
| CM-3: Encoder-Decoder (cVAE) | 0.53 | 0.46 | 1.89 | 0.00 | 0.52 | 0.01 | 1.62 |
| CM-4: Covariate reconstruction | 0.54 | 0.44 | 1.92 | 0.03 | 0.56 | 0.06 | 1.90 |
| CM-5: Conditional loss term | 0.54 | 0.45 | 1.91 | 0.04 | 0.57 | 0.06 | 1.90 |
| CM-6: Adversarial loss term | 0.55 | 0.45 | 1.93 | 0.01 | 0.57 | 0.06 | 1.84 |
| CM-7: Conditional Adversarial loss term | 0.53 | 0.46 | 1.91 | 0.00 | 0.52 | 0.01 | 1.64 |
| CM-8: MMD loss term (VFAE) | 0.53 | 0.46 | 1.88 | 0.00 | 0.55 | 0.01 | 1.62 |
| CM-9: HSIC loss term (HCV) | 0.53 | 0.46 | 1.91 | 0.00 | 0.51 | 0.00 | 1.62 |
| CM-10: Disentangled subspace | 0.53 | 0.46 | 1.93 | 0.01 | 0.56 | 0.06 | 1.71 |

Table 12: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 2$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.45 | 0.54 | 1.82 | 0.11 | 0.59 | 0.06 | 1.51 |
| CM-1: Decoder-only | 0.44 | 0.55 | 1.82 | 0.07 | 0.60 | 0.06 | 1.43 |
| CM-2: Encoder-only | 0.45 | 0.54 | 1.82 | 0.12 | 0.59 | 0.06 | 1.49 |
| CM-3: Encoder-Decoder (cVAE) | 0.44 | 0.55 | 1.82 | 0.03 | 0.55 | 0.01 | 1.34 |
| CM-4: Covariate reconstruction | 0.45 | 0.54 | 1.79 | 0.31 | 0.59 | 0.05 | 1.64 |
| CM-5: Conditional loss term | 0.45 | 0.54 | 1.78 | 0.33 | 0.60 | 0.07 | 1.64 |
| CM-6: Adversarial loss term | 0.45 | 0.54 | 1.82 | 0.09 | 0.59 | 0.07 | 1.53 |
| CM-7: Conditional Adversarial loss term | 0.45 | 0.55 | 1.84 | 0.04 | 0.60 | 0.05 | 1.36 |
| CM-8: MMD loss term (VFAE) | 0.44 | 0.55 | 1.82 | 0.04 | 0.53 | 0.01 | 1.35 |
| CM-9: HSIC loss term (HCV) | 0.45 | 0.55 | 1.82 | 0.04 | 0.54 | 0.01 | 1.33 |
| CM-10: Disentangled subspace | 0.45 | 0.54 | 1.81 | 0.08 | 0.60 | 0.06 | 1.43 |

Table 13: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 3$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.42 | 0.57 | 1.79 | 0.18 | 0.61 | 0.07 | 1.21 |
| CM-1: Decoder-only | 0.41 | 0.58 | 1.80 | 0.07 | 0.61 | 0.06 | 1.19 |
| CM-2: Encoder-only | 0.42 | 0.57 | 1.79 | 0.15 | 0.60 | 0.08 | 1.21 |
| CM-3: Encoder-Decoder (cVAE) | 0.41 | 0.58 | 1.79 | 0.04 | 0.57 | 0.01 | 1.11 |
| CM-4: Covariate reconstruction | 0.42 | 0.57 | 1.76 | 0.38 | 0.63 | 0.06 | 1.29 |
| CM-5: Conditional loss term | 0.42 | 0.57 | 1.77 | 0.42 | 0.62 | 0.07 | 1.29 |
| CM-6: Adversarial loss term | 0.42 | 0.57 | 1.79 | 0.14 | 0.61 | 0.07 | 1.22 |
| CM-7: Conditional Adversarial loss term | 0.41 | 0.58 | 1.79 | 0.01 | 0.58 | 0.04 | 1.12 |
| CM-8: MMD loss term (VFAE) | 0.41 | 0.58 | 1.79 | 0.04 | 0.56 | 0.03 | 1.11 |
| CM-9: HSIC loss term (HCV) | 0.41 | 0.58 | 1.79 | 0.05 | 0.56 | 0.01 | 1.11 |
| CM-10: Disentangled subspace | 0.41 | 0.58 | 1.81 | 0.10 | 0.62 | 0.07 | 1.16 |

Table 14: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 4$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.40 | 0.59 | 1.77 | 0.19 | 0.63 | 0.07 | 1.04 |
| CM-1: Decoder-only | 0.40 | 0.60 | 1.79 | 0.11 | 0.61 | 0.07 | 0.99 |
| CM-2: Encoder-only | 0.40 | 0.59 | 1.78 | 0.16 | 0.63 | 0.07 | 1.02 |
| CM-3: Encoder-Decoder (cVAE) | 0.39 | 0.60 | 1.78 | 0.06 | 0.58 | 0.01 | 0.95 |
| CM-4: Covariate reconstruction | 0.40 | 0.59 | 1.75 | 0.53 | 0.64 | 0.08 | 1.07 |
| CM-5: Conditional loss term | 0.40 | 0.59 | 1.76 | 0.50 | 0.65 | 0.10 | 1.08 |
| CM-6: Adversarial loss term | 0.40 | 0.59 | 1.76 | 0.15 | 0.64 | 0.05 | 1.05 |
| CM-7: Conditional Adversarial loss term | 0.39 | 0.60 | 1.80 | 0.07 | 0.57 | 0.03 | 0.96 |
| CM-8: MMD loss term (VFAE) | 0.39 | 0.60 | 1.79 | 0.05 | 0.55 | 0.01 | 0.95 |
| CM-9: HSIC loss term (HCV) | 0.39 | 0.60 | 1.78 | 0.06 | 0.55 | 0.01 | 0.95 |
| CM-10: Disentangled subspace | 0.40 | 0.60 | 1.77 | 0.10 | 0.61 | 0.07 | 0.98 |

Table 15: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 5$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.39 | 0.61 | 1.77 | 0.17 | 0.63 | 0.08 | 0.88 |
| CM-1: Decoder-only | 0.38 | 0.61 | 1.77 | 0.12 | 0.63 | 0.06 | 0.84 |
| CM-2: Encoder-only | 0.39 | 0.61 | 1.76 | 0.13 | 0.61 | 0.07 | 0.88 |
| CM-3: Encoder-Decoder (cVAE) | 0.39 | 0.61 | 1.78 | 0.04 | 0.58 | 0.01 | 0.82 |
| CM-4: Covariate reconstruction | 0.39 | 0.61 | 1.75 | 0.59 | 0.63 | 0.07 | 0.93 |
| CM-5: Conditional loss term | 0.39 | 0.61 | 1.75 | 0.56 | 0.65 | 0.11 | 0.93 |
| CM-6: Adversarial loss term | 0.39 | 0.61 | 1.76 | 0.21 | 0.64 | 0.08 | 0.89 |
| CM-7: Conditional Adversarial loss term | 0.38 | 0.61 | 1.79 | 0.06 | 0.58 | 0.04 | 0.82 |
| CM-8: MMD loss term (VFAE) | 0.38 | 0.61 | 1.79 | 0.03 | 0.58 | 0.04 | 0.83 |
| CM-9: HSIC loss term (HCV) | 0.38 | 0.61 | 1.78 | 0.06 | 0.59 | 0.02 | 0.82 |
| CM-10: Disentangled subspace | 0.38 | 0.61 | 1.77 | 0.10 | 0.62 | 0.10 | 0.84 |

Table 16: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 8$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.36 | 0.63 | 1.76 | 0.25 | 0.66 | 0.08 | 0.63 |
| CM-1: Decoder-only | 0.36 | 0.63 | 1.77 | 0.12 | 0.65 | 0.09 | 0.60 |
| CM-2: Encoder-only | 0.36 | 0.63 | 1.76 | 0.26 | 0.64 | 0.09 | 0.63 |
| CM-3: Encoder-Decoder (cVAE) | 0.36 | 0.64 | 1.78 | 0.08 | 0.60 | 0.06 | 0.59 |
| CM-4: Covariate reconstruction | 0.36 | 0.63 | 1.74 | 0.54 | 0.67 | 0.11 | 0.66 |
| CM-5: Conditional loss term | 0.36 | 0.63 | 1.74 | 0.53 | 0.77 | 0.21 | 0.66 |
| CM-6: Adversarial loss term | 0.36 | 0.63 | 1.76 | 0.26 | 0.66 | 0.12 | 0.63 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 1.78 | 0.06 | 0.62 | 0.03 | 0.59 |
| CM-8: MMD loss term (VFAE) | 0.36 | 0.64 | 1.78 | 0.07 | 0.61 | 0.05 | 0.59 |
| CM-9: HSIC loss term (HCV) | 0.36 | 0.64 | 1.78 | 0.08 | 0.61 | 0.06 | 0.59 |
| CM-10: Disentangled subspace | 0.36 | 0.63 | 1.78 | 0.08 | 0.65 | 0.08 | 0.60 |

Table 17: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 12$ trained on the GenR dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.35 | 0.64 | 1.76 | 0.33 | 0.64 | 0.12 | 0.44 |
| CM-1: Decoder-only | 0.35 | 0.64 | 1.77 | 0.21 | 0.66 | 0.13 | 0.42 |
| CM-2: Encoder-only | 0.36 | 0.64 | 1.76 | 0.31 | 0.68 | 0.14 | 0.43 |
| CM-3: Encoder-Decoder (cVAE) | 0.35 | 0.64 | 1.77 | 0.09 | 0.64 | 0.07 | 0.40 |
| CM-4: Covariate reconstruction | 0.35 | 0.64 | 1.73 | 0.68 | 0.69 | 0.17 | 0.46 |
| CM-5: Conditional loss term | 0.35 | 0.64 | 1.74 | 0.64 | 0.89 | 0.42 | 0.46 |
| CM-6: Adversarial loss term | 0.35 | 0.64 | 1.76 | 0.30 | 0.68 | 0.13 | 0.44 |
| CM-7: Conditional Adversarial loss term | 0.35 | 0.65 | 1.76 | 0.18 | 0.78 | 0.23 | 0.42 |
| CM-8: MMD loss term (VFAE) | 0.35 | 0.64 | 1.77 | 0.10 | 0.64 | 0.08 | 0.40 |
| CM-9: HSIC loss term (HCV) | 0.35 | 0.64 | 1.77 | 0.10 | 0.65 | 0.07 | 0.40 |
| CM-10: Disentangled subspace | 0.35 | 0.64 | 1.77 | 0.20 | 0.67 | 0.14 | 0.42 |

Table 18: Evaluation results of different covariate modeling methods when modeling age, sex at latent dimensionality $d = 16$ trained on the GenR dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Normative Alignment |
|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | KL Divergence |
| Baseline | 0.35 | 0.64 | 1.76 | 0.36 | 0.66 | 0.13 | 0.33 |
| CM-1: Decoder-only | 0.35 | 0.64 | 1.77 | 0.23 | 0.65 | 0.13 | 0.31 |
| CM-2: Encoder-only | 0.35 | 0.64 | 1.76 | 0.30 | 0.67 | 0.15 | 0.33 |
| CM-3: Encoder-Decoder (cVAE) | 0.35 | 0.65 | 1.76 | 0.14 | 0.65 | 0.10 | 0.31 |
| CM-4: Covariate reconstruction | 0.35 | 0.64 | 1.73 | 0.69 | 0.71 | 0.21 | 0.34 |
| CM-5: Conditional loss term | 0.35 | 0.64 | 1.74 | 0.67 | 0.97 | 0.69 | 0.35 |
| CM-6: Adversarial loss term | 0.35 | 0.64 | 1.76 | 0.40 | 0.67 | 0.22 | 0.34 |
| CM-7: Conditional Adversarial loss term | 0.35 | 0.65 | 1.75 | 0.31 | 0.94 | 1.38 | 0.31 |
| CM-8: MMD loss term (VFAE) | 0.35 | 0.65 | 1.77 | 0.14 | 0.64 | 0.10 | 0.31 |
| CM-9: HSIC loss term (HCV) | 0.35 | 0.65 | 1.77 | 0.14 | 0.65 | 0.10 | 0.31 |
| CM-10: Disentangled subspace | 0.35 | 0.64 | 1.76 | 0.27 | 0.66 | 0.18 | 0.32 |

## A.2 Covariate Modeling Methods Results – HBN

Table 19: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 1$ trained on the HBN dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.47 | 0.53 | 1.30 | 0.02 | 0.59 | 0.05 | 0.36 | 0.00 | 1.80 |
| CM-1: Decoder-only | 0.45 | 0.55 | 1.31 | 0.03 | 0.58 | 0.05 | 0.36 | 0.00 | 1.80 |
| CM-2: Encoder-only | 0.47 | 0.53 | 1.36 | 0.02 | 0.59 | 0.06 | 0.41 | 0.01 | 1.80 |
| CM-3: Encoder-Decoder (cVAE) | 0.45 | 0.55 | 1.44 | 0.03 | 0.50 | 0.01 | 0.37 | 0.01 | 1.74 |
| CM-4: Covariate reconstruction | 0.47 | 0.53 | 1.36 | 0.02 | 0.59 | 0.07 | 0.35 | 0.00 | 1.81 |
| CM-5: Conditional loss term | 0.47 | 0.53 | 1.36 | 0.01 | 0.61 | 0.09 | 0.37 | 0.01 | 1.81 |
| CM-6: Adversarial loss term | 0.47 | 0.53 | 1.41 | 0.02 | 0.59 | 0.07 | 0.37 | 0.00 | 1.81 |
| CM-7: Conditional Adversarial loss term | 0.45 | 0.55 | 1.43 | 0.01 | 0.59 | 0.05 | 0.34 | 0.00 | 1.79 |
| CM-8: MMD loss term (VFAE) | 0.45 | 0.55 | 1.44 | 0.02 | 0.55 | 0.01 | 0.37 | 0.00 | 1.74 |
| CM-9: HSIC loss term (HCV) | 0.45 | 0.55 | 1.43 | 0.03 | 0.51 | 0.01 | 0.36 | 0.01 | 1.74 |
| CM-10: Disentangled subspace | 0.46 | 0.55 | 1.34 | 0.02 | 0.58 | 0.05 | 0.34 | 0.00 | 1.80 |

Table 20: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 2$ trained on the HBN dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.42 | 0.58 | 0.93 | 0.13 | 0.65 | 0.09 | 0.37 | 0.01 | 1.30 |
| CM-1: Decoder-only | 0.42 | 0.58 | 1.08 | 0.10 | 0.61 | 0.07 | 0.36 | 0.01 | 1.23 |
| CM-2: Encoder-only | 0.42 | 0.58 | 0.93 | 0.12 | 0.64 | 0.08 | 0.39 | 0.01 | 1.29 |
| CM-3: Encoder-Decoder (cVAE) | 0.42 | 0.58 | 1.15 | 0.05 | 0.54 | 0.00 | 0.41 | 0.01 | 1.19 |
| CM-4: Covariate reconstruction | 0.42 | 0.58 | 0.45 | 0.42 | 0.67 | 0.07 | 0.40 | 0.02 | 1.41 |
| CM-5: Conditional loss term | 0.42 | 0.58 | 0.45 | 0.42 | 0.64 | 0.06 | 0.41 | 0.01 | 1.41 |
| CM-6: Adversarial loss term | 0.43 | 0.58 | 0.98 | 0.12 | 0.64 | 0.06 | 0.38 | 0.02 | 1.32 |
| CM-7: Conditional Adversarial loss term | 0.42 | 0.58 | 1.12 | 0.04 | 0.64 | 0.10 | 0.41 | 0.05 | 1.26 |
| CM-8: MMD loss term (VFAE) | 0.42 | 0.58 | 1.16 | 0.07 | 0.57 | 0.01 | 0.39 | 0.01 | 1.19 |
| CM-9: HSIC loss term (HCV) | 0.42 | 0.58 | 1.14 | 0.05 | 0.56 | 0.00 | 0.40 | 0.01 | 1.19 |
| CM-10: Disentangled subspace | 0.42 | 0.58 | 1.06 | 0.07 | 0.61 | 0.07 | 0.38 | 0.02 | 1.23 |

Table 21: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 3$ trained on the HBN dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.40 | 0.60 | 0.92 | 0.20 | 0.61 | 0.09 | 0.45 | 0.04 | 1.04 |
| CM-1: Decoder-only | 0.39 | 0.61 | 1.00 | 0.13 | 0.65 | 0.09 | 0.40 | 0.01 | 0.98 |
| CM-2: Encoder-only | 0.40 | 0.60 | 0.91 | 0.15 | 0.64 | 0.07 | 0.40 | 0.04 | 1.04 |
| CM-3: Encoder-Decoder (cVAE) | 0.39 | 0.61 | 1.11 | 0.05 | 0.58 | 0.05 | 0.41 | 0.03 | 0.96 |
| CM-4: Covariate reconstruction | 0.40 | 0.60 | 0.41 | 0.44 | 0.66 | 0.08 | 0.40 | 0.03 | 1.08 |
| CM-5: Conditional loss term | 0.40 | 0.60 | 0.39 | 0.50 | 0.66 | 0.07 | 0.39 | 0.02 | 1.09 |
| CM-6: Adversarial loss term | 0.40 | 0.60 | 0.87 | 0.18 | 0.63 | 0.07 | 0.41 | 0.03 | 1.05 |
| CM-7: Conditional Adversarial loss term | 0.39 | 0.61 | 1.11 | 0.03 | 0.61 | 0.03 | 0.47 | 0.07 | 1.00 |
| CM-8: MMD loss term (VFAE) | 0.39 | 0.61 | 1.12 | 0.05 | 0.59 | 0.06 | 0.42 | 0.02 | 0.95 |
| CM-9: HSIC loss term (HCV) | 0.39 | 0.61 | 1.11 | 0.05 | 0.59 | 0.06 | 0.42 | 0.03 | 0.96 |
| CM-10: Disentangled subspace | 0.40 | 0.61 | 0.99 | 0.13 | 0.63 | 0.07 | 0.39 | 0.01 | 1.00 |

Table 22: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 4$ trained on the HBN dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Covariate Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.38 | 0.62 | 0.79 | 0.14 | 0.66 | 0.09 | 0.45 | 0.02 | 0.86 |
| CM-1: Decoder-only | 0.38 | 0.62 | 0.93 | 0.16 | 0.65 | 0.11 | 0.43 | 0.04 | 0.84 |
| CM-2: Encoder-only | 0.38 | 0.62 | 0.86 | 0.16 | 0.65 | 0.10 | 0.47 | 0.07 | 0.87 |
| CM-3: Encoder-Decoder (cVAE) | 0.38 | 0.62 | 1.13 | 0.05 | 0.59 | 0.04 | 0.43 | 0.04 | 0.81 |
| CM-4: Covariate reconstruction | 0.38 | 0.62 | 0.36 | 0.53 | 0.68 | 0.07 | 0.47 | 0.06 | 0.92 |
| CM-5: Conditional loss term | 0.38 | 0.62 | 0.37 | 0.51 | 0.68 | 0.10 | 0.48 | 0.07 | 0.92 |
| CM-6: Adversarial loss term | 0.38 | 0.62 | 0.89 | 0.17 | 0.64 | 0.08 | 0.44 | 0.02 | 0.89 |
| CM-7: Conditional Adversarial loss term | 0.37 | 0.63 | 1.09 | 0.07 | 0.60 | 0.06 | 0.44 | 0.05 | 0.83 |
| CM-8: MMD loss term (VFAE) | 0.38 | 0.62 | 1.12 | 0.06 | 0.55 | 0.02 | 0.43 | 0.04 | 0.81 |
| CM-9: HSIC loss term (HCV) | 0.38 | 0.62 | 1.12 | 0.05 | 0.58 | 0.04 | 0.42 | 0.03 | 0.81 |
| CM-10: Disentangled subspace | 0.38 | 0.62 | 0.95 | 0.13 | 0.64 | 0.08 | 0.44 | 0.04 | 0.84 |

Table 23: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 5$ trained on the HBN dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Covariate Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.38 | 0.62 | 0.81 | 0.17 | 0.65 | 0.10 | 0.46 | 0.06 | 0.72 |
| CM-1: Decoder-only | 0.37 | 0.63 | 0.96 | 0.13 | 0.64 | 0.13 | 0.46 | 0.07 | 0.70 |
| CM-2: Encoder-only | 0.37 | 0.63 | 0.85 | 0.19 | 0.66 | 0.07 | 0.48 | 0.08 | 0.72 |
| CM-3: Encoder-Decoder (cVAE) | 0.37 | 0.63 | 1.12 | 0.06 | 0.58 | 0.02 | 0.45 | 0.09 | 0.68 |
| CM-4: Covariate reconstruction | 0.37 | 0.63 | 0.31 | 0.52 | 0.68 | 0.09 | 0.47 | 0.08 | 0.77 |
| CM-5: Conditional loss term | 0.37 | 0.63 | 0.32 | 0.52 | 0.68 | 0.14 | 0.53 | 0.14 | 0.80 |
| CM-6: Adversarial loss term | 0.37 | 0.63 | 0.82 | 0.21 | 0.65 | 0.10 | 0.47 | 0.06 | 0.75 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 1.09 | 0.09 | 0.59 | 0.05 | 0.48 | 0.08 | 0.72 |
| CM-8: MMD loss term (VFAE) | 0.37 | 0.63 | 1.10 | 0.06 | 0.59 | 0.05 | 0.45 | 0.03 | 0.68 |
| CM-9: HSIC loss term (HCV) | 0.37 | 0.63 | 1.11 | 0.06 | 0.58 | 0.02 | 0.45 | 0.07 | 0.68 |
| CM-10: Disentangled subspace | 0.37 | 0.63 | 0.95 | 0.12 | 0.67 | 0.12 | 0.44 | 0.05 | 0.71 |

Table 24: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 8$ trained on the HBN dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Covariate Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.37 | 0.63 | 0.78 | 0.23 | 0.65 | 0.12 | 0.44 | 0.08 | 0.48 |
| CM-1: Decoder-only | 0.36 | 0.64 | 0.91 | 0.16 | 0.67 | 0.10 | 0.50 | 0.08 | 0.47 |
| CM-2: Encoder-only | 0.37 | 0.63 | 0.81 | 0.23 | 0.67 | 0.13 | 0.49 | 0.13 | 0.48 |
| CM-3: Encoder-Decoder (cVAE) | 0.37 | 0.63 | 1.01 | 0.14 | 0.62 | 0.06 | 0.45 | 0.07 | 0.45 |
| CM-4: Covariate reconstruction | 0.37 | 0.63 | 0.31 | 0.56 | 0.68 | 0.16 | 0.49 | 0.14 | 0.49 |
| CM-5: Conditional loss term | 0.37 | 0.63 | 0.35 | 0.57 | 0.69 | 0.20 | 0.52 | 0.15 | 0.49 |
| CM-6: Adversarial loss term | 0.37 | 0.63 | 0.80 | 0.26 | 0.67 | 0.13 | 0.48 | 0.13 | 0.48 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 1.05 | 0.10 | 0.73 | 0.18 | 0.62 | 0.20 | 0.47 |
| CM-8: MMD loss term (VFAE) | 0.37 | 0.63 | 1.03 | 0.11 | 0.62 | 0.08 | 0.47 | 0.08 | 0.45 |
| CM-9: HSIC loss term (HCV) | 0.37 | 0.63 | 1.01 | 0.15 | 0.62 | 0.05 | 0.46 | 0.06 | 0.45 |
| CM-10: Disentangled subspace | 0.36 | 0.64 | 0.88 | 0.19 | 0.65 | 0.12 | 0.47 | 0.12 | 0.48 |

Table 25: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 12$ trained on the HBN dataset.

| **Covariate Modeling Method** | **Reconstruction Quality** | | **Covariate Invariance (age)** | | **Covariate Invariance (sex)** | | **Covariate Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.37 | 0.63 | 0.83 | 0.24 | 0.66 | 0.16 | 0.51 | 0.20 | 0.31 |
| CM-1: Decoder-only | 0.37 | 0.63 | 0.90 | 0.21 | 0.65 | 0.16 | 0.47 | 0.17 | 0.29 |
| CM-2: Encoder-only | 0.38 | 0.63 | 0.79 | 0.23 | 0.67 | 0.21 | 0.49 | 0.16 | 0.31 |
| CM-3: Encoder-Decoder (cVAE) | 0.37 | 0.63 | 0.96 | 0.16 | 0.63 | 0.07 | 0.46 | 0.06 | 0.29 |
| CM-4: Covariate reconstruction | 0.38 | 0.62 | 0.28 | 0.68 | 0.74 | 0.24 | 0.53 | 0.18 | 0.31 |
| CM-5: Conditional loss term | 0.38 | 0.62 | 0.30 | 0.63 | 0.72 | 0.19 | 0.52 | 0.18 | 0.32 |
| CM-6: Adversarial loss term | 0.36 | 0.64 | 0.78 | 0.26 | 0.67 | 0.19 | 0.51 | 0.16 | 0.33 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 0.84 | 0.15 | 0.79 | 0.36 | 0.67 | 0.34 | 0.30 |
| CM-8: MMD loss term (VFAE) | 0.37 | 0.63 | 0.95 | 0.14 | 0.63 | 0.06 | 0.45 | 0.14 | 0.29 |
| CM-9: HSIC loss term (HCV) | 0.37 | 0.63 | 0.96 | 0.16 | 0.62 | 0.07 | 0.47 | 0.06 | 0.29 |
| CM-10: Disentangled subspace | 0.36 | 0.64 | 0.90 | 0.19 | 0.66 | 0.17 | 0.49 | 0.14 | 0.31 |

Table 26: Evaluation results of different covariate modeling methods when modeling age, sex, site at latent dimensionality $d = 16$ trained on the HBN dataset.

| Covariate Modeling Method | Reconstruction Quality | | Covariate Invariance (age) | | Covariate Invariance (sex) | | Covariate Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Error | MI Site | KL Divergence |
| Baseline | 0.37 | 0.63 | 0.82 | 0.37 | 0.66 | 0.25 | 0.46 | 0.21 | 0.24 |
| CM-1: Decoder-only | 0.36 | 0.64 | 0.90 | 0.30 | 0.66 | 0.19 | 0.52 | 0.18 | 0.23 |
| CM-2: Encoder-only | 0.37 | 0.63 | 0.79 | 0.26 | 0.66 | 0.25 | 0.48 | 0.18 | 0.23 |
| CM-3: Encoder-Decoder (cVAE) | 0.37 | 0.63 | 0.99 | 0.18 | 0.65 | 0.13 | 0.47 | 0.16 | 0.22 |
| CM-4: Covariate reconstruction | 0.38 | 0.62 | 0.28 | 0.76 | 0.69 | 0.27 | 0.51 | 0.18 | 0.24 |
| CM-5: Conditional loss term | 0.37 | 0.63 | 0.27 | 0.72 | 0.70 | 0.35 | 0.57 | 0.28 | 0.24 |
| CM-6: Adversarial loss term | 0.37 | 0.63 | 0.79 | 0.41 | 0.69 | 0.26 | 0.51 | 0.20 | 0.24 |
| CM-7: Conditional Adversarial loss term | 0.36 | 0.64 | 0.77 | 0.21 | 0.82 | 0.45 | 0.64 | 0.46 | 0.24 |
| CM-8: MMD loss term (VFAE) | 0.37 | 0.63 | 1.00 | 0.19 | 0.64 | 0.16 | 0.49 | 0.16 | 0.22 |
| CM-9: HSIC loss term (HCV) | 0.37 | 0.63 | 0.99 | 0.18 | 0.65 | 0.13 | 0.48 | 0.17 | 0.22 |
| CM-10: Disentangled subspace | 0.36 | 0.64 | 0.88 | 0.35 | 0.68 | 0.19 | 0.49 | 0.18 | 0.24 |

## A.3 Site Accommodation Results - GenR + HBN

Table 27: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 1$.

| Model | Test Site | Reconstruction Quality | | Invariance (age) | | Invariance (sex) | | Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.49 | 0.53 | 1.65 | 0.00 | 0.51 | 0.09 | 0.34 | 0.03 | 1.81 |
| | RUBIC | 0.48 | 0.53 | 1.49 | 0.04 | 0.53 | 0.08 | 0.37 | 0.00 | 1.80 |
| | CBIC | 0.49 | 0.49 | 1.46 | 0.00 | 0.64 | 0.08 | 0.34 | 0.00 | 1.75 |
| | **Average** | 0.49 | 0.52 | 1.53 | 0.01 | 0.56 | 0.08 | 0.35 | 0.01 | 1.79 |
| Combat Harmonized Baseline | GenR | 0.49 | 0.53 | 1.46 | 0.02 | 0.53 | 0.07 | 0.29 | 0.00 | 1.81 |
| | RUBIC | 0.48 | 0.53 | 1.37 | 0.06 | 0.60 | 0.09 | 0.32 | 0.00 | 1.79 |
| | CBIC | 0.49 | 0.49 | 1.60 | 0.00 | 0.61 | 0.12 | 0.32 | 0.00 | 1.74 |
| | **Average** | 0.49 | 0.52 | 1.48 | 0.03 | 0.58 | 0.09 | 0.31 | 0.00 | 1.78 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.47 | 0.55 | 1.32 | 0.09 | 0.59 | 0.00 | 0.28 | 0.04 | 1.71 |
| | RUBIC | 0.47 | 0.54 | 1.83 | 0.00 | 0.59 | 0.00 | 0.41 | 0.04 | 1.63 |
| | CBIC | 0.46 | 0.52 | 1.63 | 0.06 | 0.51 | 0.00 | 0.27 | 0.02 | 1.64 |
| | **Average** | 0.47 | 0.54 | 1.59 | 0.05 | 0.56 | 0.00 | 0.32 | 0.03 | 1.66 |
| CM-8: MMD loss term (VFAE) | GenR | 0.47 | 0.55 | 1.46 | 0.09 | 0.50 | 0.00 | 0.29 | 0.04 | 1.70 |
| | RUBIC | 0.47 | 0.54 | 1.67 | 0.00 | 0.56 | 0.00 | 0.32 | 0.03 | 1.64 |
| | CBIC | 0.47 | 0.52 | 1.67 | 0.06 | 0.49 | 0.01 | 0.27 | 0.02 | 1.63 |
| | **Average** | 0.47 | 0.54 | 1.60 | 0.05 | 0.52 | 0.00 | 0.29 | 0.03 | 1.66 |

Table 28: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 2$.

| Model | Test Site | Reconstruction Quality | | Invariance (age) | | Invariance (sex) | | Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.44 | 0.58 | 0.91 | 0.06 | 0.62 | 0.10 | 0.43 | 0.10 | 1.31 |
| | RUBIC | 0.43 | 0.58 | 1.03 | 0.08 | 0.58 | 0.09 | 0.32 | 0.04 | 1.29 |
| | CBIC | 0.44 | 0.54 | 1.13 | 0.07 | 0.60 | 0.01 | 0.30 | 0.11 | 1.28 |
| | **Average** | 0.44 | 0.57 | 1.02 | 0.07 | 0.60 | 0.07 | 0.35 | 0.08 | 1.29 |
| Combat Harmonized Baseline | GenR | 0.44 | 0.58 | 0.99 | 0.14 | 0.59 | 0.12 | 0.42 | 0.00 | 1.31 |
| | RUBIC | 0.43 | 0.58 | 1.10 | 0.07 | 0.62 | 0.16 | 0.39 | 0.03 | 1.29 |
| | CBIC | 0.44 | 0.54 | 1.16 | 0.19 | 0.64 | 0.15 | 0.30 | 0.04 | 1.27 |
| | **Average** | 0.44 | 0.57 | 1.08 | 0.13 | 0.62 | 0.14 | 0.37 | 0.02 | 1.29 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.44 | 0.58 | 1.26 | 0.07 | 0.57 | 0.00 | 0.33 | 0.02 | 1.15 |
| | RUBIC | 0.44 | 0.58 | 1.34 | 0.06 | 0.56 | 0.06 | 0.32 | 0.07 | 1.11 |
| | CBIC | 0.44 | 0.55 | 1.35 | 0.19 | 0.60 | 0.05 | 0.31 | 0.02 | 1.13 |
| | **Average** | 0.44 | 0.57 | 1.32 | 0.11 | 0.58 | 0.04 | 0.32 | 0.04 | 1.13 |
| CM-8: MMD loss term (VFAE) | GenR | 0.44 | 0.58 | 1.18 | 0.06 | 0.58 | 0.00 | 0.34 | 0.04 | 1.16 |
| | RUBIC | 0.44 | 0.58 | 1.29 | 0.06 | 0.46 | 0.06 | 0.33 | 0.06 | 1.11 |
| | CBIC | 0.44 | 0.55 | 1.41 | 0.19 | 0.59 | 0.04 | 0.34 | 0.03 | 1.13 |
| | **Average** | 0.44 | 0.57 | 1.29 | 0.10 | 0.54 | 0.03 | 0.34 | 0.04 | 1.13 |

Table 29: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 3$.

| **Model** | **Test Site** | **Reconstruction Quality** | | **Invariance (age)** | | **Invariance (sex)** | | **Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.41 | 0.60 | 0.86 | 0.11 | 0.62 | 0.08 | 0.42 | 0.08 | 1.04 |
| | RUBIC | 0.41 | 0.60 | 0.78 | 0.21 | 0.59 | 0.12 | 0.41 | 0.04 | 1.02 |
| | CBIC | 0.42 | 0.57 | 0.88 | 0.26 | 0.61 | 0.05 | 0.42 | 0.00 | 1.03 |
| | **Average** | 0.41 | 0.59 | 0.84 | 0.19 | 0.61 | 0.08 | 0.42 | 0.04 | 1.03 |
| Combat Harmonized Baseline | GenR | 0.41 | 0.60 | 0.89 | 0.17 | 0.59 | 0.10 | 0.36 | 0.00 | 1.02 |
| | RUBIC | 0.41 | 0.60 | 0.83 | 0.22 | 0.61 | 0.10 | 0.41 | 0.07 | 1.05 |
| | CBIC | 0.42 | 0.57 | 0.86 | 0.30 | 0.59 | 0.05 | 0.40 | 0.00 | 1.02 |
| | **Average** | 0.41 | 0.59 | 0.86 | 0.23 | 0.60 | 0.08 | 0.39 | 0.02 | 1.03 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.41 | 0.60 | 1.07 | 0.08 | 0.50 | 0.02 | 0.38 | 0.03 | 0.92 |
| | RUBIC | 0.41 | 0.60 | 1.15 | 0.04 | 0.51 | 0.01 | 0.36 | 0.11 | 0.92 |
| | CBIC | 0.41 | 0.58 | 1.22 | 0.11 | 0.53 | 0.03 | 0.33 | 0.03 | 0.91 |
| | **Average** | 0.41 | 0.59 | 1.15 | 0.08 | 0.51 | 0.02 | 0.36 | 0.06 | 0.92 |
| CM-8: MMD loss term (VFAE) | GenR | 0.42 | 0.60 | 1.06 | 0.07 | 0.57 | 0.01 | 0.41 | 0.02 | 0.91 |
| | RUBIC | 0.41 | 0.60 | 1.11 | 0.03 | 0.50 | 0.06 | 0.33 | 0.03 | 0.92 |
| | CBIC | 0.41 | 0.58 | 1.18 | 0.12 | 0.53 | 0.02 | 0.37 | 0.06 | 0.91 |
| | **Average** | 0.41 | 0.59 | 1.12 | 0.07 | 0.53 | 0.03 | 0.37 | 0.04 | 0.91 |

Table 30: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 4$.

| **Model** | **Test Site** | **Reconstruction Quality** | | **Invariance (age)** | | **Invariance (sex)** | | **Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.41 | 0.61 | 0.77 | 0.23 | 0.60 | 0.09 | 0.30 | 0.10 | 0.78 |
| | RUBIC | 0.39 | 0.62 | 0.93 | 0.22 | 0.63 | 0.09 | 0.44 | 0.08 | 0.88 |
| | CBIC | 0.39 | 0.59 | 0.88 | 0.18 | 0.56 | 0.01 | 0.41 | 0.05 | 0.87 |
| | **Average** | 0.40 | 0.61 | 0.86 | 0.21 | 0.60 | 0.06 | 0.38 | 0.08 | 0.84 |
| Combat Harmonized Baseline | GenR | 0.39 | 0.62 | 0.83 | 0.18 | 0.62 | 0.11 | 0.36 | 0.05 | 0.88 |
| | RUBIC | 0.39 | 0.61 | 0.97 | 0.18 | 0.57 | 0.15 | 0.34 | 0.00 | 0.86 |
| | CBIC | 0.40 | 0.59 | 0.86 | 0.20 | 0.62 | 0.06 | 0.48 | 0.07 | 0.87 |
| | **Average** | 0.39 | 0.61 | 0.89 | 0.19 | 0.60 | 0.11 | 0.39 | 0.04 | 0.87 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.44 | 0.58 | 1.18 | 0.11 | 0.53 | 0.15 | 0.33 | 0.04 | 0.66 |
| | RUBIC | 0.39 | 0.62 | 1.10 | 0.14 | 0.49 | 0.07 | 0.34 | 0.05 | 0.77 |
| | CBIC | 0.40 | 0.59 | 1.19 | 0.14 | 0.57 | 0.04 | 0.42 | 0.12 | 0.77 |
| | **Average** | 0.41 | 0.60 | 1.16 | 0.13 | 0.53 | 0.09 | 0.36 | 0.07 | 0.73 |
| CM-8: MMD loss term (VFAE) | GenR | 0.43 | 0.58 | 1.05 | 0.06 | 0.52 | 0.05 | 0.31 | 0.10 | 0.65 |
| | RUBIC | 0.40 | 0.61 | 1.15 | 0.06 | 0.52 | 0.05 | 0.36 | 0.10 | 0.76 |
| | CBIC | 0.40 | 0.59 | 1.13 | 0.09 | 0.59 | 0.04 | 0.34 | 0.03 | 0.77 |
| | **Average** | 0.41 | 0.59 | 1.11 | 0.07 | 0.54 | 0.05 | 0.34 | 0.08 | 0.73 |

Table 31: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 5$.

| **Model** | **Test Site** | **Reconstruction Quality** | | **Invariance (age)** | | **Invariance (sex)** | | **Invariance (site)** | | **Normative Alignment** |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.42 | 0.60 | 0.77 | 0.26 | 0.61 | 0.16 | 0.36 | 0.09 | 0.62 |
| | RUBIC | 0.38 | 0.63 | 1.02 | 0.14 | 0.64 | 0.16 | 0.37 | 0.09 | 0.77 |
| | CBIC | 0.38 | 0.60 | 0.83 | 0.07 | 0.58 | 0.04 | 0.58 | 0.27 | 0.76 |
| | **Average** | 0.39 | 0.61 | 0.87 | 0.16 | 0.61 | 0.12 | 0.44 | 0.15 | 0.72 |
| Combat Harmonized Baseline | GenR | 0.42 | 0.60 | 0.76 | 0.26 | 0.61 | 0.15 | 0.30 | 0.09 | 0.62 |
| | RUBIC | 0.38 | 0.63 | 0.87 | 0.11 | 0.63 | 0.10 | 0.28 | 0.03 | 0.77 |
| | CBIC | 0.39 | 0.59 | 0.86 | 0.29 | 0.66 | 0.12 | 0.33 | 0.10 | 0.73 |
| | **Average** | 0.40 | 0.61 | 0.83 | 0.22 | 0.63 | 0.12 | 0.30 | 0.07 | 0.71 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.42 | 0.60 | 0.99 | 0.05 | 0.54 | 0.10 | 0.39 | 0.08 | 0.63 |
| | RUBIC | 0.38 | 0.63 | 1.11 | 0.08 | 0.56 | 0.07 | 0.33 | 0.19 | 0.69 |
| | CBIC | 0.41 | 0.58 | 1.27 | 0.13 | 0.48 | 0.04 | 0.39 | 0.05 | 0.58 |
| | **Average** | 0.40 | 0.60 | 1.12 | 0.09 | 0.53 | 0.07 | 0.37 | 0.11 | 0.63 |
| CM-8: MMD loss term (VFAE) | GenR | 0.42 | 0.60 | 0.99 | 0.05 | 0.57 | 0.11 | 0.40 | 0.10 | 0.63 |
| | RUBIC | 0.38 | 0.63 | 1.14 | 0.13 | 0.57 | 0.08 | 0.36 | 0.16 | 0.69 |
| | CBIC | 0.41 | 0.57 | 1.20 | 0.19 | 0.48 | 0.04 | 0.42 | 0.02 | 0.56 |
| | **Average** | 0.40 | 0.60 | 1.11 | 0.12 | 0.54 | 0.08 | 0.39 | 0.09 | 0.63 |

Table 32: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 8$.

| Model | Test Site | Reconstruction Quality | | Invariance (age) | | Invariance (sex) | | Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.40 | 0.62 | 0.79 | 0.31 | 0.60 | 0.18 | 0.38 | 0.25 | 0.44 |
| | RUBIC | 0.38 | 0.63 | 0.79 | 0.25 | 0.59 | 0.16 | 0.34 | 0.16 | 0.46 |
| | CBIC | 0.39 | 0.60 | 0.85 | 0.36 | 0.60 | 0.11 | 0.50 | 0.20 | 0.47 |
| | **Average** | 0.39 | 0.62 | 0.81 | 0.31 | 0.60 | 0.15 | 0.41 | 0.20 | 0.46 |
| Combat Harmonized Baseline | GenR | 0.40 | 0.62 | 0.83 | 0.27 | 0.57 | 0.15 | 0.38 | 0.12 | 0.44 |
| | RUBIC | 0.38 | 0.62 | 0.75 | 0.30 | 0.61 | 0.35 | 0.28 | 0.22 | 0.44 |
| | CBIC | 0.39 | 0.60 | 0.80 | 0.15 | 0.61 | 0.14 | 0.40 | 0.06 | 0.48 |
| | **Average** | 0.39 | 0.61 | 0.79 | 0.24 | 0.60 | 0.21 | 0.35 | 0.13 | 0.45 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.41 | 0.61 | 1.00 | 0.27 | 0.48 | 0.12 | 0.44 | 0.18 | 0.37 |
| | RUBIC | 0.38 | 0.63 | 1.06 | 0.25 | 0.49 | 0.02 | 0.42 | 0.24 | 0.42 |
| | CBIC | 0.39 | 0.60 | 1.06 | 0.17 | 0.60 | 0.08 | 0.44 | 0.19 | 0.41 |
| | **Average** | 0.39 | 0.61 | 1.04 | 0.23 | 0.52 | 0.07 | 0.43 | 0.20 | 0.40 |
| CM-8: MMD loss term (VFAE) | GenR | 0.41 | 0.61 | 0.95 | 0.24 | 0.54 | 0.10 | 0.42 | 0.24 | 0.37 |
| | RUBIC | 0.38 | 0.63 | 1.00 | 0.26 | 0.57 | 0.04 | 0.47 | 0.27 | 0.42 |
| | CBIC | 0.39 | 0.60 | 1.07 | 0.17 | 0.63 | 0.02 | 0.41 | 0.10 | 0.41 |
| | **Average** | 0.39 | 0.61 | 1.01 | 0.22 | 0.58 | 0.05 | 0.43 | 0.20 | 0.40 |

Table 33: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 12$.

| Model | Test Site | Reconstruction Quality | | Invariance (age) | | Invariance (sex) | | Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.39 | 0.63 | 0.79 | 0.42 | 0.63 | 0.31 | 0.29 | 0.29 | 0.31 |
| | RUBIC | 0.37 | 0.64 | 0.76 | 0.38 | 0.61 | 0.22 | 0.37 | 0.29 | 0.32 |
| | CBIC | 0.39 | 0.60 | 0.79 | 0.24 | 0.53 | 0.11 | 0.46 | 0.32 | 0.30 |
| | **Average** | 0.38 | 0.62 | 0.78 | 0.35 | 0.59 | 0.21 | 0.37 | 0.30 | 0.31 |
| Combat Harmonized Baseline | GenR | 0.40 | 0.62 | 0.75 | 0.40 | 0.59 | 0.25 | 0.33 | 0.15 | 0.29 |
| | RUBIC | 0.38 | 0.63 | 0.75 | 0.28 | 0.61 | 0.33 | 0.37 | 0.28 | 0.30 |
| | CBIC | 0.39 | 0.60 | 0.79 | 0.31 | 0.57 | 0.21 | 0.34 | 0.24 | 0.33 |
| | **Average** | 0.39 | 0.62 | 0.76 | 0.33 | 0.59 | 0.26 | 0.35 | 0.22 | 0.31 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.40 | 0.62 | 1.14 | 0.24 | 0.52 | 0.20 | 0.38 | 0.43 | 0.27 |
| | RUBIC | 0.38 | 0.63 | 1.14 | 0.14 | 0.53 | 0.17 | 0.60 | 0.47 | 0.28 |
| | CBIC | 0.39 | 0.59 | 1.21 | 0.35 | 0.54 | 0.06 | 0.38 | 0.21 | 0.28 |
| | **Average** | 0.39 | 0.61 | 1.16 | 0.24 | 0.53 | 0.14 | 0.45 | 0.37 | 0.28 |
| CM-8: MMD loss term (VFAE) | GenR | 0.40 | 0.62 | 1.15 | 0.25 | 0.52 | 0.23 | 0.40 | 0.39 | 0.27 |
| | RUBIC | 0.38 | 0.63 | 1.13 | 0.09 | 0.51 | 0.22 | 0.60 | 0.48 | 0.28 |
| | CBIC | 0.39 | 0.59 | 1.21 | 0.38 | 0.54 | 0.05 | 0.41 | 0.20 | 0.28 |
| | **Average** | 0.39 | 0.61 | 1.16 | 0.24 | 0.52 | 0.17 | 0.47 | 0.36 | 0.28 |

Table 34: Site accommodation evaluation results of different covariate modeling methods and combat harmonization conditioned on Age, Sex, Site at latent dimensionality $d = 16$.

| Model | Test Site | Reconstruction Quality | | Invariance (age) | | Invariance (sex) | | Invariance (site) | | Normative Alignment |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | Pred. MSE | MI | Pred. Accuracy | MI | Pred. Accuracy | MI | |
| Baseline | GenR | 0.39 | 0.62 | 0.77 | 0.28 | 0.67 | 0.37 | 0.30 | 0.28 | 0.24 |
| | RUBIC | 0.38 | 0.63 | 0.90 | 0.42 | 0.59 | 0.29 | 0.49 | 0.44 | 0.23 |
| | CBIC | 0.40 | 0.59 | 0.85 | 0.43 | 0.58 | 0.17 | 0.46 | 0.25 | 0.22 |
| | **Average** | 0.39 | 0.61 | 0.84 | 0.38 | 0.61 | 0.28 | 0.42 | 0.32 | 0.23 |
| Combat Harmonized Baseline | GenR | 0.39 | 0.62 | 0.73 | 1.01 | 0.57 | 0.31 | 0.29 | 0.24 | 0.22 |
| | RUBIC | 0.38 | 0.63 | 0.85 | 0.33 | 0.56 | 0.33 | 0.36 | 0.17 | 0.24 |
| | CBIC | 0.39 | 0.60 | 0.77 | 0.46 | 0.54 | 0.21 | 0.42 | 0.30 | 0.24 |
| | **Average** | 0.39 | 0.62 | 0.78 | 0.60 | 0.56 | 0.28 | 0.36 | 0.24 | 0.23 |
| CM-3: Encoder-Decoder (cVAE) | GenR | 0.40 | 0.62 | 1.07 | 0.41 | 0.70 | 0.23 | 0.41 | 0.24 | 0.20 |
| | RUBIC | 0.37 | 0.64 | 1.00 | 0.35 | 0.47 | 0.17 | 0.42 | 0.40 | 0.22 |
| | CBIC | 0.39 | 0.60 | 1.00 | 0.19 | 0.60 | 0.21 | 0.50 | 0.43 | 0.20 |
| | **Average** | 0.39 | 0.62 | 1.02 | 0.32 | 0.59 | 0.20 | 0.44 | 0.36 | 0.21 |
| CM-8: MMD loss term (VFAE) | GenR | 0.40 | 0.62 | 1.07 | 0.39 | 0.68 | 0.19 | 0.41 | 0.25 | 0.20 |
| | RUBIC | 0.37 | 0.64 | 1.02 | 0.38 | 0.51 | 0.09 | 0.41 | 0.39 | 0.22 |
| | CBIC | 0.39 | 0.60 | 0.99 | 0.20 | 0.63 | 0.21 | 0.52 | 0.41 | 0.20 |
| | **Average** | 0.39 | 0.62 | 1.03 | 0.32 | 0.61 | 0.16 | 0.45 | 0.35 | 0.21 |