Unsupervised physics-informed disentanglement of multimodal data for high-throughput scientific discovery

Nathaniel Trask¹ Carianne Martinez² ³ Kookjin Lee³ Brad Boyce⁴

Abstract
We introduce physics-informed multimodal autoencoders (PIMA) - a variational inference framework for discovering shared information in multimodal scientific datasets representative of high-throughput testing. Individual modalities are embedded into a shared latent space and fused through a product of experts formulation, enabling a Gaussian mixture prior to identify shared features. Sampling from clusters allows cross-modal generative modeling, with a mixture of expert decoder imposing inductive biases encoding prior scientific knowledge and imparting structured disentanglement of the latent space. This approach enables discovery of fingerprints which may be detected in high-dimensional heterogeneous datasets, avoiding traditional bottlenecks related to high-fidelity measurement and characterization. Motivated by accelerated co-design and optimization of materials manufacturing processes, a dataset of lattice metamaterials from metal additive manufacturing demonstrates accurate cross-modal inference between images of mesoscale topology and mechanical stress-strain response.

1. Motivation
Many scientific and engineering datasets are multimodal, necessitating the fusion of disparate sources and datatypes for informed analysis. For example, in the realm of process optimization for materials manufacturing, processes ranging from microelectronic fabrication to metal additive manufacturing involve a myriad of process settings along with in-process and post-process measurements (Liu et al., 2012; Sochol et al., 2018). Moreover, automated high-throughput characterization methods are increasingly generating large, rich, multimodal datasets, fueled by advances in robotics and automation (Boyce & Uchic, 2019). Many scientific datasets admit fingerprints: easily measurable signals which correlate with a difficult to measure underlying physical process. The hunt for exploitable fingerprints extends beyond material science (Isayev et al., 2015), and can be found in all science and engineering domains ranging from quantum mechanics (Chakraborty et al., 2021) to climate change (Hasselmann, 1997; Hegerl et al., 2007). Rapid datasets designed to detect fingerprints may potentially serve as a surrogate for, or in conjunction with, bespoke experiments capturing high-fidelity modalities. Accordingly, we aim to discover comprehensive fingerprints constructed from the weighted integration of several disparate data sources, each with unique fidelity, sparsity, and spatiotemporal resolution. The multimodal scientific data under consideration includes both physical and simulated data and differs from the text/audio/video modalities commonly considered in the multimodal literature (Baltrušaitis et al., 2018), providing the opportunity to impose physics-based inductive biases and move beyond purely data-driven linear techniques such as principle component analysis typically used for fingerprint detection.

Herein, we present a novel variational inference framework for synthesizing multimodal scientific data with the aim of cross-modal inference; if one can reliably perform generative modeling of a high-fidelity but slow measurement from a low-fidelity but fast fingerprint, high-throughput experimentation and material characterization are no longer infeasible. Such applications however mandate an unsupervised approach, as costly human-in-the-loop data labelling precludes high-throughput testing.

Concretely, cross-modal inference corresponds to training jointly across modalities $X_1, ..., X_D$ in a manner that supports generative sampling of individual modalities $p(X_i | X_j)$ for $i \neq j$. We achieve this in a variational inference setting by combining the following algorithmic contributions (Figure 1): 1. encoding data into unimodal embeddings $q(Z | X_i)$ and applying a product of experts (PoE) model to fuse

---
¹Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA ²Applied Information Sciences Center, Sandia National Laboratories, Albuquerque, NM, USA ³School of Computing and Augmented Intelligence, Arizona State University, USA ⁴Materials, Physical, and Chemical Sciences Center, Sandia National Laboratories, Albuquerque, NM, USA. Correspondence to: Nathaniel Trask <natrask@sandia.gov>.
data into a multimodal posterior $q(Z|X) = \Pi_i q(Z|X_i)$; 2. adopting a Gaussian mixture prior $p(Z|c)$ to determine latent clusters shared across modalities; and 3. decoding with a physics-informed mixture of experts (MoE) model $p(X_i|c, Z)$ to impose inductive biases. For scientific settings, the expert model provides a critical new means of fusing experimental audiovisual data with traditional scientific models; rather than considering generalized linear models commonly used in MoE (Jordan & Jacobs, 1994), we may incorporate parameterized physical models, surrogates or simulators for the system under consideration. These ingredients are designed to yield an ELBO loss with closed form expressions for requisite integrals and is amenable to a novel expectation maximization strategy to fit clusters and experts. In concert, this architecture produces fingerprints in the form of latent clusters spanning modalities, with cross-modal estimators allowing inference of cluster membership for a single modality.

For unsupervised learning, several works apply variational autoencoders (VAE) to seek latent disentangled representations of data which admit efficient separation into meaningful classes (Burgess et al., 2018; Chen et al., 2018; Locatello et al., 2019; Kim & Mnih, 2018). While desirable from an interpretability and accuracy perspective, such representations are often challenging to reliably discover in the absence of labels. The complementary information available in multimodal data has been shown to provide multiple pathways to disentanglement; a human may be unable to distinguish an image of a one and a seven, but if the digit is read aloud there is no confusion. However, the fact that scientific data is governed by physical models potentially allows the expert model to extract more information than purely data-driven approaches - known physics encodes the generative process and therefore imposing even a low-fidelity physical model as inductive bias may provide substantial disentanglement. Our tests demonstrate that the combination of PoE, GMM, and expert models provide not only disentangled clusters, but also an ordering of clusters in latent space reflecting information shared across modalities. For examples with unambiguous class ownership, we demonstrate 100% test accuracy when classifying data into clusters. We anticipate these physics-based disentangled representations will enable future causal analysis of large high-throughput datasets.

1.1. Relationship to prior literature

This work draws from several thematic bodies of literature. The non-exhaustive list below denotes those works which have most informed our approach as well as provide overviews for the recent state-of-the-art.

Gaussian mixture embeddings For deep unsupervised clustering, several works replace the standard normal prior from (Kingma & Welling, 2014; Rezende et al., 2014) with a Gaussian mixture model (GMM) to facilitate disentanglement and provide an explicit parameterization of clusters (Dilokthanakul et al., 2016; Jiang et al., 2017; Rao et al., 2019; Lee et al., 2020). Each modality in the multi-modal prior distribution is expected to provide disentangled latent representations of data which admit an explicit parameterization of class distributions. The current work is most similar to VaDE (Jiang et al., 2017) in its use of mean-field distributions to obtain a separable ELBO, and Bayesian estimator for $q(c|X)$. This work builds upon VaDE by incorporating multimodal data inputs while maintaining computational tractability of the ELBO, as well as employing clusters to decode into physics-informed MoE models.

Disentanglement Another line of research is to extract latent disentangled representations into different factors of variations in data using VAEs. Earlier works such as β-VAE (Higgins et al., 2017) and Annealed VAE (Burgess et al., 2018) introduce additional weighting parameters to the KL divergence term of the original VAE ELBO loss. In Factor VAE (Kim & Mnih, 2018) and β-TCVAE (Chen et al., 2018) the ELBO is further decomposed to derive and penalize the total correlation to promote disentanglement in learned representations. For our purposes however, without an explicit parameterization of the cluster distributions to condition off of, it is not possible to introduce a physics-informed expert MoE model.

Multimodal inference Generative modeling from multimodal data can be broadly categorized into either conditional generative models (Sohn et al., 2015; Pu et al., 2016) which directly learn conditional cross-modal distributions $p(X_i|X_j)$, or joint models (Suzuki et al., 2017; Vedantam et al., 2018; Wu & Goodman, 2018), which explicitly learn joint distributions that learn $p(Z, X_1, ..., X_D)$. We pursue the later as (Wu & Goodman, 2018) has been shown to provide better description of the underlying data distribution. We pursue the strategy used by works such as joint multimodal VAE (Suzuki et al., 2017) and joint VAE (Vedantam et al., 2018), where a joint inference network $q(Z|X_1, X_2)$ is trained, followed by training of two additional unimodal inference networks $q(Z|X_1)$ and $q(Z|X_2)$ which handle missing data at test time. The unimodal inference networks are trained to either match the joint inference network or to maximize an ELBO derived to perform unimodal variational inference. More recently, MVVAE (Wu & Goodman, 2018) and MMVVAE (Shi et al., 2019) were proposed to model the joint posterior as a product of experts (PoEs) and a mixture of experts (MoEs). Most recently, MoPoE-VAE (Sutter et al., 2021) proposed a new ELBO formulation, which generalizes ELBO formulations derived from PoEs and MoEs. Our encoder bears similarities to MMVVAE, MoPoE-VAE, and PoE, while preserving a computationally tractable closed form ELBO when combined with the GMM.
Physics-informed disentanglement of multimodal data

Figure 1. Individual modalities are encoded into Gaussian distributions in a shared latent space. During training the posterior is sampled from a product of experts distribution fusing complementary information into a shared multimodal Gaussian distribution. A Gaussian mixture prior parameterizes clusters encoding cross-modal shared information. Sampling from mixture components provides generative models using either black-box decoders or expert physics models encoding prior physics knowledge. To facilitate cross-modal generative inference, unimodal embeddings are trained to reproduce the multimodal embedding, allowing inference of $p(c | X_i)$. Shown here, an expert strain-hardening plasticity model allows two types of cross-modal inference: costly measurements of stress-strain response may be inferred from high-throughput imaging of lattice topology, or generative microstructural images can be provided to suggest microstructure which correlate with a given stress/strain measurement.

Physics-informed ML and fingerprinting
Substantial works in recent years have focused on introducing physics into either solving partial differential equations (PDEs) or for building surrogates, typically introducing a PDE residual regularizer in physics-informed neural networks (Lagaris et al., 1998; Raissi et al., 2019) or by embedding physics directly into network architecture in structure-preserving ML (Trask et al., 2020). Such tools can be combined to provide parametric surrogates of simulations which can perform real-time inference over a database of parameterized PDE solutions (Lu et al., 2019; Wang et al., 2021; Mao et al., 2021). This paper provides a framework to fuse either these physics-informed surrogates or simpler empirical models together with experimental data. In contrast to traditional tools for fingerprinting which rely on purely data-driven techniques like PCA (Hasselmann, 1997; Hegerl et al., 2007), the current framework provides a means to incorporate domain expertise into fingerprints tailored toward a scientific task.

Major contributions:

- Novel fusion of PoE w/ Gaussian mixture to obtain parameterized cluster “fingerprints” for downstream data analysis and high-throughput diagnostic tasks.

- Multimodal embedding allows cross-modal inference while preserving closed form expressions for expectations in ELBO.

- Mixture of experts decoding allows incorporation of interpretable inductive biases by assuming model form describing scientific processes. Potential for embedding physics-informed surrogates or simulators.

- Improvements over SotA unimodal unsupervised techniques.

- Disentanglement of clusters into structured latent space exposing relationships across modalities.

2. Framework construction

Given individual modalities $X_i \subset \mathbb{R}^d_i$, we partition the set of all modalities $\mathcal{M} = \{X_1, ..., X_D\}$ into $\mathcal{M}_{DD}$ consisting of images/videos/audio amenable to purely data-driven modeling, and $\mathcal{M}_S$ consisting of scientific modalities amenable
to expert modeling. For each $X_i \in M$ we seek an embedding $Z \in \mathbb{R}^l$ in latent dimension $l << d_i$. Assuming a categorical variable $c$ clustering data into $C$ clusters in latent space, our variational autoencoder amounts to introducing parameterized prior $p$ and posterior $q$ distributions that maximize the following ELBO loss:

$$
\mathcal{L} = \mathbb{E}_{q(Z,c|X_1,\ldots,X_D)} \left[ \log \frac{p(X_1,\ldots,X_D, Z)}{q(Z,c|X_1,\ldots,X_D)} \right].
$$

(1)

We further assume separability of both prior and posterior:

$$
p(X_1,\ldots,X_D, Z, c) = (\Pi_{i=1}^D p(X_i|Z, c)) \ p(Z|c) p(c),
$$

(2)

$$
q(Z|X_1,\ldots,X_D) = q(Z|X_1,\ldots,X_D)q(c|X_1,\ldots,X_D).
$$

(3)

Our framework consists of four components: 1. unimodal deep encodings with a product of experts (PoE) multimodal fusion, 2. a mixture of Gaussians prior, 3. a mixture of experts decoding of modalities $X \in M_S$, and 4. unimodal encoders for cross-modal inference. We introduce each component sequentially, derive a closed form expression for the ELBO, and introduce an expectation maximization assignment of clusters and expert models.

### 2.1. Multi-modal embedding

Assuming the unimodal embeddings may be modeled as multivariate Gaussians with diagonal covariance, we obtain posterior probabilities $q(Z_i|X_i) = \mathcal{N}(Z_i; \mu_i, \sigma_i^2 \mathbf{I})$, with mean and covariance provided by the set of neural networks

$$
[\mu_i, \sigma_i^2] = F_i(X_i; \hat{\theta}_i),
$$

(4)

where $\hat{\theta}_i$ denotes trainable weights and biases. For this work we consider a simple class of 1D/2D convolutional encoders, whose architecture is provided in Appendix A.

To estimate $q(Z|X_1,\ldots,X_D)$ in the ELBO, it follows from Bayes’ rule and pairwise independence that

$$
q(Z|X_1,\ldots,X_D) \propto \Pi_{i=1}^D q(Z_i|X_i),
$$

(5)

so that the posterior is a scaled product of individual modalities. To obtain closed form expressions for the ELBO later, we assume

$$
q(Z|X_1,\ldots,X_D) \propto \Pi_{i=1}^D q(Z_i|X_i).
$$

(6)

The product of Gaussian distributions is again Gaussian, yielding the multimodal distribution:

$$
q(Z|X_1,\ldots,X_D) = \mathcal{N}(\mu, \sigma^2 \mathbf{I}),
$$

(7)

$$
\sigma^{-2} = \sum_{i=1}^D \sigma_i^{-2}, \quad \frac{\mu}{\sigma^2} = \sum_{i=1}^D \frac{\mu_i}{\sigma_i^2},
$$

(8)

which may be sampled during training using the reparameterization trick: by sampling $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and calculating $Z = \mu + \epsilon \odot \sigma$, we may back-propagate through the random node $Z$ into the unimodal encoders, where $\odot$ denotes the Hadamard product.

### 2.2. Gaussian mixture prior and expert decoding

We adopt a simple Gaussian mixture prior, modeling

$$
p(c) = \text{Cat}(c|\pi),
$$

(9)

$$
p(Z|c) = \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I}),
$$

(10)

$$
p(X_i|Z, c) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2 \mathbf{I}).
$$

(11)

To ensure a positive $\pi$ that sums to unity, we parameterize it as the softmax of a trainable vector. To decode $X_i \in M_{DD}$, we employ a neural network with parameters $\hat{\theta}_i$

$$
[\hat{\mu}_i, \hat{\sigma}_i^2] = G_i(X_i; \hat{\theta}_i).
$$

(12)

For $X_i \in M_S$, we assume an expert model $p(X_i|c) = \mathcal{N}(\mathbf{t}; \hat{\epsilon}_c, \hat{\sigma}_c^2 \mathbf{I})$, where $\mathbf{t}$ is an independent variable and $\hat{\epsilon}_c$ denotes expert parameters associated with each cluster. The specific choice of $\mathcal{E}$ will be problem dependent and specified in the experiment section.

Because $p(X_i|Z, c)$ admits interpretation as a mixture of experts model (Jordan & Jacobs, 1994), we obtain closed form expressions for the mean and variance to facilitate postprocessing and uncertainty quantification:

$$
\mathbb{E}[X_i] = \sum_c \pi_c \mathbf{t}; \hat{\epsilon}_c,
$$

(13)

$$
\text{Var}[X_i] = \left( \sum_c \pi_c (\hat{\sigma}_c^2 - \mathcal{E}(\mathbf{t}; \hat{\epsilon}_c)^2) \right) - \mathbb{E}[X_i]^2.
$$

(14)

In practice, $\mathcal{E}$ may take a variety of forms and its judicious selection imparts significant prior knowledge. We consider in this work simple generalized linear models and, for the mechanical data, an empirical linear strain-hardening model. In general, these could range from empirical engineering correlations obtained from e.g. dimensionless analysis or singular perturbation theory, to analytic parametric solutions to PDE based models, or to parametric physics-informed ML surrogates/reduce order models (see e.g. (Lu et al., 2019; Wang et al., 2021; Mao et al., 2021; Trask et al., 2020)).

### 2.3. ELBO loss and EM minimizer

A modification of the derivation in (Jiang et al., 2017) to account for multimodality yields the closed form for the
Physics-informed disentanglement of multimodal data

A weighted least squares problem for the optimal expert
\[ \mu \]
where
\[ \gamma \]
(\text{Jiang et al., 2017}), and we have taken
\[ ||\cdot|| \]
to facilitate batched access to large datasets, this may be variation of the ELBO with respect to
\[ \theta \]:

\[ \frac{\text{d}L}{\text{d}\theta} = -\sum_{c \in M} \sum_{t_n} \gamma_c \left( \log \sigma_{c,n}^2 + \frac{\langle X_{i,n} - \mathcal{E}(t_n; \hat{\theta}) \rangle^2}{\sigma_c^2} \right) - \sum_c \sum_j \gamma_c \left( \log \sigma_{c,j}^2 + \frac{(\mu_j - \mu_{c,j})^2}{\sigma_{c,j}^2} \right) + 2 \sum_c \gamma_c \log \frac{\pi_c}{\gamma_c} + \sum_j (1 + \log \sigma_j^2), \]

where \( ||\cdot|| \) denotes the \( \ell_2 \)-norm, subscripts denote scalar components of tensors, \( \gamma_c \) is the posterior distribution

\[ \gamma_c = p(c|Z) = \frac{\pi_c p(Z|c)}{\sum_c \pi_c p(Z|c')}, \tag{15} \]

we estimate \( q(c|X_1, ..., X_d) = p(c|Z) = \gamma_c \) following (\text{Jiang et al., 2017}), and we have taken \( \sigma_c = 1 \). The derivation may be found in Appendix B. We seek the minimizer of this loss over the entire data set:

\[ L = -\sum_d L_d. \]

In standard expectation-maximization fashion, we note that for fixed value of \( \gamma_c \), the variation of \( L \) with respect to the cluster centers \( \mu_c \) yields the global minimizer

\[ \mu_c = \frac{\sum_d \gamma_{cd} \mu_d}{\sum_d \gamma_{cd}}, \tag{16} \]

where \( \mu_d \) and \( \gamma_{cd} \) denote the encoded mean and posterior \( p(c|X_1, ..., X_D) \) of the \( d \)th data point, respectively.

To facilitate batched access to large datasets, this may be calculated incrementally following the streaming algorithm outlined in Algorithm 1, alternating between an EM update for \( \mu_c \) followed by an Adam update (\text{Kingma & Ba, 2014}) of the remaining variables.

A weighted least squares problem for the optimal expert model parameters may be similarly obtained by taking the variation of the ELBO with respect to \( \hat{\theta} \):

\[ \hat{\theta}_c = \argmin_{\theta'} \sum_d \sum_{t_n} \gamma_{cd} \left( X_{i,n} - \mathcal{E}(t_n; \hat{\theta}') \right)^2. \tag{17} \]

Efficient solution of this nonlinear least squares problem at each epoch will be dependent upon the problem-specific expert model and data stream, and to perform batching may require a streaming technique such as recursive least squares or Kalman filtering (\text{Cioffi & Kailath, 1984}). For simplicity, we update \( \hat{\theta}_c \) with Adam in this work but note that solving this at each epoch to ensure the expert model provides a best fit to the current partitions is likely to provide substantial improvement.

\begin{algorithm}[h]
\begin{algorithmic}
\caption{Training with streaming EM for cluster centers}
\begin{algorithmic}
\Input{data \( x = \{X_1, ..., X_d\} \) in batches \( B \)}
\For{\( i = 1 \) \textbf{to} \( N_{\text{epochs}} \)}
\State Calculate \( \gamma = p(c|x) \) for all \( x \)
\State Let \( W_{\text{new}} = 0, M = 0 \)
\For{\( b \in B \)}
\State \( W^{\text{old}} \leftarrow W^{\text{new}} \)
\State \( W^{\text{new}} \leftarrow W^{\text{new}} + \sum x \in b \gamma_c(x) \)
\EndFor
\State \( M^{\text{old}} \leftarrow M + \frac{\sum b \gamma_c(x)}{N_{\text{epochs}}} \)
\EndFor
\State Calculate Adam update on ELBO
\end{algorithmic}
\end{algorithm}

\end{algorithm}

\subsection{2.4. Cross-modal inference}

The primary objective of this work is to perform cross-modal inference: sampling from \( q(Z|X_i) \) allows: 1. generative modeling by decoding \( p(X_i|z) \) for \( i \neq j \), and 2. an estimate of \( p(c|Z) \) via Eqn. (15). Unfortunately, sampling from the unimodal encoders \( q(Z|X_i) \) provides poor embeddings far from the multimodal embedding \( q(Z|X_1, ..., X_d) \).

To remedy this, we introduce a second set of unimodal encoders \( q(Z|X_i) \sim \mathcal{N}(Z; \hat{\mu}_i, \hat{\sigma}_i^2 \mathbf{1}) \) with identical architecture to \( q(Z|X_i) \). After the multimodal network is trained, we minimize the KL-divergence between \( q(Z|X_i) \) and \( q(Z|X_1, ..., X_D) \) so that the unimodal embeddings reproduce the multimodal one. In this sense, the unsupervised multimodal training provides labels which allows supervised training of unimodal embeddings. The KL loss admits the closed form expression

\[ L_i = \frac{1}{2} \sum_{d,k} \left[ \log \frac{\hat{\sigma}_{kd}^2}{\sigma_{kd}^2} + \frac{\hat{\mu}_{kd}^2 - \mu_{kd}^2}{\sigma_{kd}^2} \right], \tag{18} \]

which may be sequentially optimized with Adam for each modality \( i \) after the multimodal model has been fit. An additional possibility not pursued here is to perform a Bayesian estimate to identify either the cluster most likely to generate the data

\[ p(c|X_i) = \frac{p(X_i|c)p(c)}{\sum_c p(X_i|c')p(c')} \text{ for } X_i \in M, \tag{19} \]

or the cluster centroid most likely to have generated the data

\[ p(\mu_c|X_i) = \frac{p(X_i|\mu_c)p(\mu_c)}{\sum_{\mu_c} p(X_i|\mu_c')p(\mu_c')} \text{ for } X_i \in M_{DD}. \tag{20} \]
3. Experiments

Hyper-parameters for both training and architecture were selected using the Weights and Biases experiment tracking tool (Biewald, 2020) and were based on the MNIST dataset (LeCun et al., 1998) using a 90/10 train/validation split of the training data. MNIST results are reported on the standard 10,000 held out test examples (LeCun et al., 2010). All parameters used in the study, code and data to reproduce experiments, and discussion of hardware and training time for each experiment, may be found in Appendix A. We consider a non-physical and physical pair of experiments described in Figure 2.

![Figure 2. Experimental setup for unsupervised multimodal MNIST (left) and multimodal mechanical test (right). For MNIST, we replace the labels on digits $c \in \{0, ..., 9\}$ by a sample of the function $X_2 = ct + c$, for $t \in [0, 1]$ and Gaussian noise $c$; $c$ is therefore encoded either as an image in $X_1$ or as the slope of a 1D function in $X_2$. For the mechanical problem, a high-throughput compression test is performed on two populations of additively manufactured lattices corresponding to gyroid and octet microstructure (Garland et al., 2020). We supplement the resulting stress-strain curves $X_2$ with images of the microstructure $X_1$ to obtain low- and high-throughput modalities, respectively.](image)

3.1. Unsupervised multimodal MNIST

In the first experiment, we take the traditional MNIST dataset consisting of images $X_1$ and labels $c \in \{0, ..., 9\}$ and manufacture a synthetic 1D modality $X_2 = ct + c$, where $t \in [0, 1]$ and $c \sim \mathcal{N}(0, 0.01)$ for a clean dataset or $c \sim \mathcal{N}(0, 0.5)$ for a noisy data set. We adopt the affine expert model $E(t; \theta_c) = \theta_c t$, and perform unsupervised clustering of the multimodal dataset $(X_1, X_2)$, and additionally perform cross-modal inference. For this artificial problem, the labels are thinly veiled as the slope of second modality, and so we expect that if we successfully perform multimodal inference we should obtain accuracy comparable to a supervised MNIST benchmark.

We define unsupervised clustering accuracy (acc) as in (Xie et al., 2016), (Jiang et al., 2017):

$$acc = \max_{m \in M} \frac{\sum_{i=1}^{N} \mathbb{1}\{l_i = m(c_i)\}}{N}, \quad (21)$$

where $N$ is the number of examples, $M$ is the set of all possible mappings from a cluster to a label assignment, $l_i$ is the true label and $c_i$ is the cluster assignment by the model for example $i$. Calibration of the latent dimension (Figure 3) revealed $l = 3$ to be an optimal embedding dimension. In Table 1 we provide a comparison to classification accuracy against state of the art supervised and unsupervised models trained on images only. For multimodal inference with low noise we obtain perfect classification outperforming the current state-of-the-art in supervised classification and outperforming the state-of-the-art in unsupervised learning.

In addition to observing improved accuracy, we investigate in Figure 4 the disentangled representation of clusters and provide examples of generative models. Surprisingly, the sequential ordering of clusters in $X_2$ induces an sequential embedding of corresponding clusters in latent space. As noise is increased, the data in $X_2$ is sufficiently large to prevent distinct clustering into disentangled classes, however the ordering of digits is roughly preserved. When generative modeling is performed for $X_2$, this is reflected by samples from the tail of the Gaussian mixture which generate images of adjacent digits. For example, the cluster of 2’s contains images of 3’s in its periphery. This suggests that the sequential ordering of data in $X_2$ may induce generative models for $X_1$ which reflect cross-modal ordering. Confusion matrices

![Table 1: MNIST clusters and resulting accuracies for multimodal clean dataset, varying encoding dimension of size 1, 2, 3, and 4 and selecting model from hyperparameter training with lowest validation loss. Test accuracy achieved by clusters shown: 99.9%, 100.0%, 99.9%, and 100.0%, respectively. Statistics correspond to 10 runs using optimal hyperparameters. 3 of the 4 encoding dimensions from the 4D model are shown.](image)
Physics-informed disentanglement of multimodal data

| Method           | Notes       | CNN SotA\* | VAE+GMM† | DEC† | VaDE† | GMVAE\†† | GMVAE\†† |
|------------------|-------------|------------|----------|------|-------|----------|----------|
| Accuracy (max)   | Supervised | 99.91      | 72.94    | 84.30| 94.46 | 88.54    | 96.92    |
| Accuracy (mean±stdev) | n/a         | n/a        | n/a      | n/a  | n/a   | 82.31 (3.75) | 87.82 (5.33) |

Table 1. Unsupervised classification accuracy for MNIST. Results gathered from (An et al., 2020), (Jiang et al., 2017) and (Dilokthanakul et al., 2016) denoted by *,† and ††, respectively. If statistics were not provided we assume maximum accuracy was reported. While the data augmentation offered by \(X_2\) is not incorporated in comparisons to unimodal unsupervised benchmarks, a comparison to the supervised setting is valid. For all experiments we do not overparameterize and keep clusters equal to the number of digits.

Figure 4. Generative modeling for multimodal MNIST experiment. Clustering in latent space with generative digit images \(X_1\) at cluster centroids (left) and expert fit to \(X_2\) (center) for clean (top) and noisy (bottom) \(X_2\) data. In both cases, data is embedded into disentangled clusters adjacent to digits to obtain ordered clusters; for noisy data clusters overlap to reflect class confusion in \(X_2\) but preserve ordering. In this noisy case, this is reflected when sampling \(X_1\) from \(Z_i \in \{\mu_c \pm 3\sigma_c\}_i\) for clusters 1-9 (right): at the periphery of each cluster a digit \(±1\) is generated, reflecting the fuzzy class boundary of \(X_2\).

for multimodal and cross-modal classification in Figure 5 reveal an approximately banded structure, whereby misclassified modalities primarily occur between adjacent digits. While this single experiment is insufficient to remark on the generality of this result, it suggests the potential for learning generative models of images which reflect information from the expert model, a particularly exciting prospect for scientific datasets.

### 3.2. Metal additive lattice fingerprinting

The lattice dataset consists of \(X_1\) images of 3D printed lattices split between two types of printed metamaterials and corresponding \(X_2\) stress/strain curves performed in a high-throughput uniaxial compression machine. Even with high-throughput testing, only 91 pairs of images were able to be generated, highlighting the utility of high-throughput photographs \(X_1\) as surrogates to infer \(X_2\). We select as expert model a linear strain-hardening model partitioning the stress-strain response into two piecewise linear regions (Jones, 2009). Even a simple model like this succinctly encodes a large number of interpretable quantities of interest: a yield stress, together with elastic and plastic moduli. We gather in Figure 6 results from generative modeling and include in the appendix additional results demonstrating 94.74%/94.74%/94.74% classification accuracy of the two clusters for multimodal/\(X_1/X_2\)-cross-modal inferences, respectively.
Physics-informed disentanglement of multimodal data

4. Conclusions and future work

The present approach provides an abstract variational inference for discovering fingerprints in an unsupervised manner while incorporating physical model biases. This framework is widely applicable to a range of scientific disciplines where detection of fingerprints are crucial for tasks ranging from predicting and attributing climate change to designing biochemical pathways at a molecular level. In addition to the application focus on fingerprint generation, this framework may be used for a variety of general purpose downstream tasks based on multimodal processing of scientific data. For this work we have focused on a simple MNIST example to probe dynamics for an easily replicable and understandable dataset, along with a simple high-throughput manufacturing example to illustrate feasibility for facilitating high-throughput experimentation.

In future work, we will employ more sophisticated physics-informed surrogates as expert models for processes in additive manufacturing and semiconductor device design bridging multiscale and multifidelity information from data sources corresponding to both physical experiment (transmission electron microscopy, atom probe tomography, micro x-ray computerized tomography, or synchotron x-ray diffraction) and high-fidelity simulations (quantum density functional theory, molecular dynamics, crystal plasticity, and continuum mechanics). To date, these costly but rich sources of information are antagonistic to high-throughput testing and simulation. This framework provides an exciting platform for discovering data-driven scientific fingerprints which may be combined with advances in automated experimentation to accelerate scientific discovery.

Software and Data

Pending acceptance, all data and code used to generate results will be hosted on a Github page. For now, we include a subset of the code to reproduce the MNIST examples through the anonymous Github:\footnote{https://anonymous.4open.science/r/PIMA-5D6B/} unfortunately anonymous github does not offer enough storage to host the accompanying dataset.
Acknowledgements

The authors thank Warren Davis, Anthony Garland and Lekha Patel for providing guidance on the variational inference framework and review of the manuscript and Kat Reiner and Greg Geller for providing computing support. All authors acknowledge funding under the Beyond Fingerprinting Sandia Grand Challenge Laboratory Directed Research and Development program. N. Trask acknowledges funding under the Beyond Fingerprinting Sandia Grand Challenge Laboratory Directed Research and Development program. N. Trask acknowledges funding under the Collaboratory on Mathematics and Physics-Informed Learning Machines for Multiscale and Multiphysics Problems (PhILMs) project funded by DOE Office of Science (Grant number DE-SC001924) and the DOE Early Career program. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc. for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND number: SAND2022-1159 O

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation, pp. 265–283, 2016.

An, S., Lee, M., Park, S., Yang, H., and So, J. An ensemble of simple convolutional neural network models for MNIST digit recognition. arXiv preprint arXiv:2008.10400, 2020.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2):423–443, 2018.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Boyce, B. L. and Uchic, M. D. Progress toward autonomous experimental systems for alloy development. MRS Bulletin, 44(4):273–280, 2019. ISSN 0883-7694.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in $\beta$-VAE. arXiv preprint arXiv:1804.03599, 2018.

Chakraborty, A., Nandi, P., and Chakraborty, B. Fingerprints of the quantum space-time in time-dependent quantum mechanics: An emergent geometric phase. arXiv preprint arXiv:2110.04370, 2021.

Chen, R. T., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in vaes. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 2615–2625, 2018.

Cioffi, J. and Kailath, T. Fast, recursive-least-squares transversal filters for adaptive filtering. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2):304–337, 1984.

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. Deep unsupervised clustering with Gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648, 2016.

Garland, A. P., White, B. C., Jared, B. H., Heiden, M., Donahue, E., and Boyce, B. L. Deep convolutional neural networks as a rapid screening tool for complex additively manufactured structures. Additive Manufacturing, 35:101217, 2020.

Hassellmann, K. Multi-pattern fingerprint method for detection and attribution of climate change. Climate dynamics, 13(9):601–611, 1997.

Hegerl, G., Zwiers, F., Braconnot, P., Gillett, N. P., Luo, Y. M., Orsini, J. M., Nicholls, N., Penner, J. E., and Stott, P. A. Understanding and attributing climate change, 2007.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In 5th International Conference on Learning Representations, ICLR 2017, 2017.

Hunter, J. D. Matplotlib: A 2d graphics environment. Computing in science & engineering, 9(03):90–95, 2007.

Isayev, O., Fourches, D., Muratov, E. N., Oses, C., Rasch, K., Tropsha, A., and Curtarolo, S. Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chemistry of Materials, 27(3):735–743, 2015.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. Variational deep embedding: an unsupervised and generative approach to clustering. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 1965–1972, 2017.
Physics-informed disentanglement of multimodal data

Jones, R. M. Deformation theory of plasticity. Bull Ridge Corporation, 2009.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. Neural computation, 6(2): 181–214, 1994.

Kim, H. and Mnih, A. Disentangling by factorising. In International Conference on Machine Learning, pp. 2649–2658. PMLR, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Kuhn, H. W. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2): 83–97, 1955.

Lagaris, I. E., Likas, A., and Fotiadis, D. I. Artificial neural networks for solving ordinary and partial differential equations. IEEE transactions on neural networks, 9(5): 987–1000, 1998.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

Lee, D. B., Min, D., Lee, S., and Hwang, S. J. Meta-GMVAE: Mixture of Gaussian VAE for unsupervised meta-learning. In International Conference on Learning Representations, 2020.

Liu, A., Zhu, W., Tsai, D., and Zheludev, N. I. Micro-machined tunable metamaterials: a review. Journal of Optics, 14(11):114009, 2012.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In International Conference on Machine Learning, pp. 4114–4124. PMLR, 2019.

Lu, L., Jin, P., and Karniadakis, G. E. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv preprint arXiv:1910.03193, 2019.

Mao, Z., Lu, L., Marxen, O., Zaki, T. A., and Karniadakis, G. E. Deepm&net for hypersonics: Predicting the coupled flow and finite-rate chemistry behind a normal shock using neural-network approximation of operators. Journal of Computational Physics, 447:110698, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research, 12:2825–2830, 2011.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. Variational autoencoder for deep learning of images, labels and captions. Advances in neural information processing systems, 29:2352–2360, 2016.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707, 2019.

Rao, D., Visin, F., Rusu, A., Pascanu, R., Teh, Y. W., and Hadsell, R. Continual unsupervised representation learning. Advances in Neural Information Processing Systems, 32:7647–7657, 2019.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In International Conference on Machine Learning, pp. 1278–1286. PMLR, 2014.

Shi, Y., Paige, B., Torr, P., et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in Neural Information Processing Systems, 32:15718–15729, 2019.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28:3483–3491, 2015.

Sutter, T. M., Daunhawer, I., and Vogt, J. E. Generalized multimodal ELBO. In 9th International Conference on Learning Representations, ICLR, 2021.

Suzuki, M., Nakayama, K., and Matsuo, Y. Joint multimodal learning with deep generative models. In 5th International Conference on Learning Representations, ICLR 2017, 2017.

Trask, N., Huang, A., and Hu, X. Enforcing exact physics in scientific machine learning: a data-driven exterior calculus on graphs. arXiv preprint arXiv:2012.11799, 2020.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. Generative models of visually grounded imagination. In 6th International Conference on Learning Representations, ICLR, 2018.
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

Wang, S., Wang, H., and Perdikaris, P. Learning the solution operator of parametric partial differential equations with physics-informed deeponets. *arXiv preprint arXiv:2103.10974*, 2021.

Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *arXiv preprint arXiv:1802.05335*, 2018.

Xie, J., Girshick, R., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pp. 478–487. PMLR, 2016.
A. Architecture, hyperparameters, and implementation

Model Architectures

We employ relatively small convolutional architectures to serve as encoders for both modalities. The image modality encoder consists of 2 2D convolutional layers with 32 and 64 channels respectively, each with 3x3 kernels. We apply the exponential linear unit (ELU) activation function as well as batch normalization after each convolutional layer, then pass the output to a fully connected layer of size $encoding_{im} \times 2$ to enable an embedding into a representation of the mean and standard deviation of the input in the latent space. For the 1D modality encoder, in place of 2D convolutional layers, we use 1D convolutions with 8 and 16 channels in the respective layers, but with an otherwise identical architecture.

The image decoder begins with a fully connected layer of appropriate size to be reshaped into 32 channels of 2D arrays, with each dimension having a length $\frac{1}{4}$ of the length of the number of pixels per side of the original image. The reshaped output of the initial dense layer is passed into a series of 3 deconvolutional layers with 64, 32, and 1 channel respectively, each with a kernel size of 3. The first 2 deconvolutional layers use a stride of 3 and a Rectified Linear Unit (ReLU) activation function, and the final deconvolutional layer uses a stride of 1. Zero padding is used to retain the input shape while traversing these layers.

Hyperparameters

We used the MNIST dataset along with the Weights and Biases tool (Biewald, 2020) to perform a hyperparameter search over learning rates and encoding dimension size. For the multimodal models, we found that a learning rate of $1.97e^{-5}$ gave the best result as measured by validation loss. For the MNIST dataset, an encoding dimension of size 3 gave the most consistent results, and we used a 2D encoding dimension for the lattice dataset. For the unimodal models, we found that a learning rate of $4.398e^{-5}$ for the image modality and a learning rate of $4.398e^{-3}$ for the 1D modality supported learning, but we leave a rigorous hyperparameter tuning for unimodal models for future work.

Implementation details

Each of the input data modalities was normalized to have values in $[0, 1]$. The 1D data was sampled to generate an array of length 20 for MNIST and length 100 for the lattice stress-strain data. The lattice images were cropped and subsampled into quadrants resulting in images of dimension 152x152. We further augmented the dataset by flipping the images along each axis.

To train the MNIST multimodal models, we used 10% of the standard MNIST train set for validation, and selected the model with the lowest validation loss. We did not observe any indication of overfitting the multimodal model to the training data. All results reported are on the held out test dataset. For the MNIST unimodal models, we again did not observe overfitting when measuring model performance by accuracy to the cluster labels generated from the trained multimodal model, and we report results on the test set from the model resulting from the final training epoch.

We split the lattice dataset into a train/test 80/20 split. Since we did not observe overfitting during multimodal training, we selected the model with the lowest training loss and applied that model to the held out test set, and we report those results. We select the model resulting from the final training epoch for the unimodal tasks.

Our models are implemented in Python using Tensorflow (Abadi et al., 2016), and we leverage the Scikit-learn library (Pedregosa et al., 2011) for data preparation and accuracy metrics, Scipy (Virtanen et al., 2020) for data preparation and the linear_sum_assignment implementation of the Hungarian method (Kuhn, 1955) for efficient computation of the unsupervised cluster accuracy. We visualize our results using the Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) libraries. Training was performed on NVIDIA DGX-1 and DGX-2 machines with each run executed on 1 GPU. A combination of P100 and V100 GPUs were used in this work. Training runs took on average approximately (machine dependent) 10-14 hours with our longest run on lattice data taking approximately one day. We made no attempt to optimize parallel training to improve run time or training efficiency.

B. Derivation of ELBO

To derive a closed form expression for the single sample ELBO

$$L_d = \mathbb{E}_{q(Z, c | X_1, ..., X_D)} \left[ \log \frac{p(X_1, ..., X_D, Z)}{q(Z, c | X_1, ..., X_D)} \right]$$

(22)
we apply the separability assumptions
\begin{align}
p(X_1, ..., X_D, Z, c) &= (\Pi_{i=1}^D p(X_i|Z, c)) p(Z|c)p(c), \\
q(Z, c|X_1, ..., X_D) &= q(Z|X_1, ..., X_D)q(c|X_1, ..., X_D).
\end{align}

providing the additive decomposition
\begin{equation}
\mathcal{L} = \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log p(X_1, ..., X_D, Z) \right] - \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log q(Z,c|X_1, ..., X_D) \right]
\end{equation}
\begin{equation}
= \sum_{i=1}^{D} \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log p(X_i|Z, c) \right] + \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log p(Z|c) \right]
\end{equation}
\begin{equation}
+ \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log p(c) \right] - \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log q(Z|X_1, ..., X_D) \right]
\end{equation}
\begin{equation}
- \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log q(c|X_1, ..., X_D) \right].
\end{equation}

For convenience we denote $\mathbb{E}_{q(Z,c|X_1, ..., X_D)} = \mathbb{E}_q$. The separability assumptions therefore decompose the ELBO into constituent expectations of the form
\begin{equation}
\mathbb{E}_q \left[ \log f(Z, c) \right] = \sum_c \int_{\mathbb{R}^l} f(Z, c) \log g(Z, c) \, dZ
\end{equation}
which may be integrated exactly for the Gaussian/categorical $f$ and $g$ appearing in the ELBO. The only term which may not be immediately computed is $\mathbb{E}_q \left[ \log q(c|X_1, ..., X_D) \right]$. The lack of a reparameterization trick for the categorical distribution precludes backpropagation into the encoder, forcing us to consider an encoder which only provides predictions for $Z$. While there are options to use e.g. a regularized Gumbel-softmax approximation to the categorical distribution (Jang et al., 2016), we would lose the tractability of the closed form expression for the ELBO. Instead we follow (Jiang et al., 2017) and approximate $q(c|X_1, ..., X_D) = p(c|Z)$ using the following justification.

Rewriting the ELBO
\begin{equation}
\mathcal{L}_d = \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log \frac{p(X_1, ..., X_D, Z)}{q(Z,c|X_1, ..., X_D)} \right]
\end{equation}
\begin{equation}
= \mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log \frac{p(X_1, ..., X_D|Z)p(Z)}{q(Z|X_1, ..., X_D)} + \log \frac{p(c|Z)}{q(c|X_1, ..., X_D)} \right]
\end{equation}
\begin{equation}
= \int_{\mathbb{R}^l} \log \frac{p(X_1, ..., X_D|Z)p(Z)}{q(Z|X_1, ..., X_D)} \, dZ + D_{KL}(q(c|X_1, ..., X_D) || p(c|Z)),
\end{equation}
we seek extremal points with respect to $c$. The first term is independent of $c$, and the positive second term takes zero value when $q(c|X_1, ..., X_D) = p(c|Z)$, providing the desired maximum. We caution however that this holds only at local minima of the loss landscape, but empirically has been shown to perform well as an estimator.

Finally for completeness, we gather from (Jiang et al., 2017) the various integral formulas required to compute the expectations in closed form with modifications for our multimodal setting.

**Lemma B.1.** Given Gaussian distributions $f(Z) = \mathcal{N}(Z; \mu_1, \sigma_1 I)$ and $g(Z) = \mathcal{N}(Z; \mu_2, \sigma_2 I)$
\begin{equation}
\int_{\mathbb{R}^l} f(Z) \log g(Z) \, dZ = -\frac{1}{2} \sum_j \log 2\pi \sigma^2_{2,j} + \frac{\sigma^2_{2,j}}{\sigma^2_{1,j}} + \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma^2_{2,j}}.
\end{equation}

**Lemma B.2.**
\begin{equation}
\mathbb{E}_{q(Z,c|X_1, ..., X_D)} \left[ \log p(Z|c) \right] = \sum_c q(c|X_1, ..., X_D) \int_{\mathbb{R}^l} q(Z|X_1, ..., X_D) \log p(Z|c) \, dZ,
\end{equation}
where the integrand may be computed from Lemma B.1.
Lemma B.3.
\[
\mathbb{E}_{\mathbf{q}(\mathbf{z}, \mathbf{c}|\mathbf{x}_1, \ldots, \mathbf{x}_D)} \left[ \log p(\mathbf{z}|\mathbf{c}) \right] = \int_{\mathbb{R}^\ell} q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) d\mathbf{z} \sum_c q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D) \pi_c, \tag{30}
\]
\[= \sum_c q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D) \pi_c.\]

Lemma B.4.
\[
\mathbb{E}_{\mathbf{q}(\mathbf{z}, \mathbf{c}|\mathbf{x}_1, \ldots, \mathbf{x}_D)} \left[ \log \mathbf{q}(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) \right] = \sum_c q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D) \int_{\mathbb{R}^\ell} q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) \log q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) d\mathbf{z}, \tag{31}
\]
\[= \int_{\mathbb{R}^\ell} q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) \log q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) d\mathbf{z},\]
where the integrand may be computed from Lemma B.1.

Lemma B.5.
\[
\mathbb{E}_{\mathbf{q}(\mathbf{z}, \mathbf{c}|\mathbf{x}_1, \ldots, \mathbf{x}_D)} \left[ \log \mathbf{q}(\mathbf{c}|\mathbf{x}_1, \ldots, \mathbf{x}_D) \right] = \int_{\mathbb{R}^\ell} q(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_D) d\mathbf{z} \sum_c q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D) \log q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D), \tag{32}
\]
\[= \sum_c q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D) \log q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D).\]

For all terms, \(q(c|\mathbf{x}_1, \ldots, \mathbf{x}_D)\) is calculated via the posterior estimator \(\gamma\) given in Equation (15).