Learning Invariant Representations

With Applications to High-Energy Physics

Justin Tan

orcid.org/0000-0003-2555-0853

School of Physics

The University of Melbourne

Submitted in partial fulfilment of
the requirements of the degree of
Master of Philosophy

January, 2020

ABSTRACT

In searches for new physics in high-energy physics, experimental analyses are primarily concerned with physical processes which are rare or hitherto unobserved. To claim a statistically significant discovery or exclusion of new physics when studying such decays, it is necessary to maintain an appropriate signal to noise ratio. This makes systems capable of efficient discrimination of signal from datasets overwhelmingly dominated by background events an important component of modern experimental analyses. However, naïve application of these methods is liable to raise poorly understood systematic effects which may ultimately degrade the significance of the final measurement. To understand the origin of these systematic effects, we note that there are certain protected variables in experimental analyses which should remain unbiased by the analysis procedure. Variables that the input parameters of models of new physics are strongly dependent upon and variables used to model background contributions to the total measured event yield fall into this category. Systems responsible for separating signal from background events achieve this by sampling events with signal-like characteristics from all candidate events. If this procedure introduces sampling bias into the distribution of protected variables, this introduces systematic effects into the analysis which are difficult to characterize. Thus it is desirable for these systems to distinguish between signal and background events without using information about certain protected variables.

Beyond high-energy physics, building systems that make decisions independent of certain protected or sensitive information is an important theme in the real-world application of machine learning and statistics. We address this task as an optimization problem of finding a representation of the observed data that is invariant to the given protected quantities. This representation should satisfy two competing criteria. Firstly, it should contain all relevant information about the data so that it may be used as a proxy for arbitrary downstream tasks, such as inference of unobserved quantities or prediction of target variables. Secondly, it should not be informative of the given protected quantities, so that downstream tasks are not influenced by these variables. If the protected quantities to be censored from the intermediate representation contain information that can improve the performance of the downstream task, it is likely that removing this information will adversely affect this task. The challenge lies in balancing both objectives without significantly compromising either requirement.

The contribution of this thesis is a new set of methods for addressing this problem. This thesis is divided into two parts, which are largely independent of one another. The first part of this thesis is about constraining the optimization procedure by which the representation is learnt to reduce the informativeness of the representation of the given protected quantities, such that the representation is invariant to changes in these quantities. The second part of this thesis approaches the problem from a latent variable model perspective, in which additional unobserved (latent) variables are introduced which explain the interaction between different attributes of the observed data. These latent variables can be interpreted as a more fundamental, compact lower-dimensional representation of the original high-dimensional unstructured data. By constraining the structure of this latent space, we demonstrate we can isolate the influence of the protected variables into a latent subspace. This allows downstream tasks to only access a relevant subset of the learned representation without being influenced by protected attributes of the original data. The feasibility of our proposed methods is demonstrated through application to a challenging experimental analysis in precision flavor physics at the Belle II experiment - the study of the $b \to s\gamma$ transition, a sensitive probe of potential new physics.

DECLARATION

I declare that this thesis compromises only my original work and that all work contained herein is my own unless explicitly indicated otherwise. Due acknowledgement has been made in the text to all other material used. This thesis is fewer than the maximum word limit in length, exclusive of tables, maps, bibliographies and appendices.

JANUARY, 2020

JUSTIN TAN

CONTENTS

1 INTRODUCTION **8**

1.1 Theoretical Preliminaries . . . 9

1.1.1 Flavor-Changing Neutral Currents . . . 9

1.1.2 Theoretical Predictions . . . 11

1.1.3 Experimental Techniques . . . 12

1.1.4 Experimental Results . . . 14

1.1.5 Constraints on New Physics from $\mathcal{B}_{s\gamma}$ . . . 14

1.1.6 Lepton Flavor Universality . . . 15

1.2 The Belle II Experiment at SuperKEKB . . . 16

1.2.1 Previous B-physics experiments . . . 16

1.2.2 The SuperKEKB Accelerator . . . 16

1.2.3 The Belle II Detector . . . 16

2 INVARIANCE THROUGH REGULARIZATION **19**

2.1 Introduction . . . 19

2.1.1 Machine Learning . . . 21

2.2 Mutual Information Regularization . . . 25

2.2.1 Enforcing Invariance in Learning Algorithms . . . 25

2.2.2 A divergence minimization perspective . . . 26

2.2.3 A density ratio estimation perspective . . . 27

2.2.4 Equivalence to Adversarial Training . . . 28

2.3 Experiments . . . 29

2.3.1 Bivariate Gaussian Classification . . . 29

2.3.2 Application to $b \to s\gamma$ . . . 30

2.4 Application to Lepton ID at Belle II . . . 37

2.4.1 Comparing Approaches for Particle Identification at Belle II . . . 38

2.5 Conclusion . . . 41

3 INVARIANCE THROUGH DISENTANGLEMENT **44**

3.1 Motivation . . . 44

3.1.1 Generative Models . . . 44

3.1.2 Variational Inference . . . 46

3.1.3 Variational Autoencoders . . . 47
3.2 Latent Space Disentanglement . . . 49
3.2.1 Informal Definition . . . 49
3.2.2 Unsupervised Learning of Disentangled Representations . . . 50
3.2.3 Introducing Supervision . . . 52
3.2.4 Quantifying Disentanglement . . . 53
3.3 Background / Related Work . . . 53
3.3.1 Disentangled Representation Learning . . . 54
3.3.2 Invariant Representations . . . 55
3.4 Invariance through Disentanglement . . . 56
3.4.1 Representation Definition . . . 57
3.4.2 Latent-sensitive Correspondence . . . 58
3.4.3 General Objective . . . 58
3.4.4 Optimizing for Disentanglement of Sensitive Factors . . . 58
3.4.5 Final Objective . . . 60
3.5 Experiments . . . 60
3.5.1 DSprites Variants . . . 60
3.5.2 Application to High Energy Physics . . . 64
3.5.3 Future Work . . . 67
3.6 Conclusion . . . 70

A APPENDIX: INVARIANCE THROUGH REGULARIZATION **78**
A.1 Proof of Proposition 1 . . . 78
A.2 Training features . . . 79
A.3 Maximum Likelihood Fit . . . 80
A.4 Hyperparameter Optimization . . . 81

B APPENDIX: INVARIANCE THROUGH DISENTANGLEMENT **82**
B.1 Disentanglement Hyperparameters . . . 82
B.1.1 Latent Dimension . . . 82
B.1.2 Mutual Information Weighting . . . 83
B.1.3 Total Correlation Weighting . . . 83
B.1.4 Supervised Regularization Weighting . . . 84
B.1.5 Hyperparameter Sensitivity . . . 84
B.2 Signal/Background M_{bc} plots . . . 84
B.2.1 Using the Full Latent Representation . . . 85
B.3 Observed Dataset . . . 86

LISTING OF FIGURES

1.1 Leading order Feynman diagrams for the radiative decay $b \to s\gamma$. 10
1.2 Schematic of typical $B \to X\gamma$ event. 13
1.3 Hypothetical heavy scalar $H^{\pm}$ contribution to Z and γ penguin diagrams in the a2HDM. 15
1.4 Photon energy resolution and PID efficiency at Belle II. 18

2.1 Example of background sculpting in M_{bc} for the $b \to s\gamma$ decay. 20
2.2 Decision surface in Cartesian space for bivariate Gaussian classification. 30
2.3 (1/2) 2D correlation plots between M_{bc} and features with high M_{bc} dependence. 32
2.3 (2/2) 2D correlation plots between M_{bc} and features with high M_{bc} dependence. 33
2.4 Classification performance on $b \to s\gamma$ dataset after baseline feature removal. 34
2.5 Illustration of mass sculpting after baseline feature removal. 34
2.6 Mutual information training dynamics (left) and signal efficiency vs. background rejection (right) for different values of λ. 35
2.7 Background M_{bc} distribution shift post-selection for different values of λ with signal M_{bc} distribution overlaid. 35
2.8 Background M_{bc} distribution shift post-selection for different values of λ with background M_{bc} distribution overlaid. 36
2.9 Maximum likelihood fit of signal and background M_{bc} distributions for different values of metric $N_S/\delta N_S$. 37
2.10 Variance of metrics for different models across hyperparameter search. 38
2.11 Binary Belle II PID perforance using univariate E/p likelihood at $\langle \epsilon_S \rangle = 0.9$. 39
2.12 Distributions of input features to supervised learning algorithms for ECL PID. 41
2.13 Distributions of input features to supervised learning algorithms for ECL PID. 42
2.14 Binary Belle II PID perforance comparison between feedforward neural network and MI-regularized feedforward neural network at $\langle \epsilon_S \rangle = 0.9$. 43
2.15 Classification performance of each model in low-momentum region below 1.0 GeV. 43

3.1 Example images from the DSprites and DSprites-Scream datasets. 62
3.2 Latent traversals for the DSprites dataset. 63
3.3 Latent traversals for the DSprites-Scream dataset. 64
3.4 DSprites-Scream example instances. 65
3.5 Decision surfaces for unbiased classification on DSprites-Scream. 65
3.6 False positive rate versus true positive rate when trained on an M_{bc}-invariant representation obtained via optimizing the β-TCVAE-Sensitive, β-TCVAE and Factor-VAE losses. 68

3.7 Signal efficiency versus background rejection when trained on an M_{bc}-invariant representation obtained via optimizing the β-TCVAE-Sensitive, β-TCVAE and Factor-VAE losses. 69
3.8 M_{bc} distributions after enforcing a multidimensional correlation. 70
3.9 ΔE distributions after enforcing a multidimensional correlation. 70
3.10 Baseline discriminator M_{bc} and ΔE distributions. 71
3.11 Visualization of invariance of the downstream classification task to the sensitive variable $s = M_{bc}$. . . 72

B.1 Dependency of metrics on latent space dimension. 83
B.2 Dependency of metrics on $I_q(x; z)$ weighting α. 83
B.3 Dependence of metrics on total correlation weighting β. 84
B.4 Dependence of metrics on supervised regularization parameter λ_s. 84
B.5 Sensitivity to hyperparameters of different disentanglement losses. 85
B.6 Baseline discriminator signal/background prior/post selection plots. 86
B.7 Signal/background prior/post selection plots using representation obtained via β-TCVAE-Sensitive loss. 87
B.8 Signal/background prior/post selection plots using representation obtained via Factor-VAE loss. . . . 88
B.9 Signal/background prior/post selection plots using representation obtained via β-TCVAE loss. 89
B.10 Marginal distributions of observed data x considered in disentanglement experiments. 92

List of Tables

2.1 Variables x_k with greatest normalized mutual information $I(x_k; z)/H(x_k)$ with protected variable $z = M_{bc}$. 31

2.2 Model comparison on AUC and $N_S/\delta N_s$ metrics. 36

3.1 Convolutional network architectures used for DSprites experiments. 62

3.2 Metrics for different disentangling losses under invariant representation. 67

A.1 Hyperparameter ranges considered in search. 81

B.1 Range of hyperparameters for each model considered in hyperparameter scans described in Appendix B.1. 82

B.2 Metrics when the full latent representation is used in lieu of the invariant latent representation. 85

CHAPTER 1

INTRODUCTION

Building systems that make decisions independent of certain protected or sensitive information is an important theme in the real-world application of machine learning and statistics. We address this task as an optimization problem of finding a representation of the observed data that is invariant to the given protected quantities. This representation should satisfy two competing criteria. Firstly, it should contain all relevant information about the data so that it may safely be used as a proxy for arbitrary downstream tasks, such as inference of unobserved quantities or prediction of target variables. Secondly it should not be informative of the given protected quantities, so that downstream tasks are not influenced by these variables.

If the protected quantities to be censored from the intermediate representation contain information that can improve the performance of the downstream task, it is likely that removing this information will adversely affect this task. The main challenge in learning invariant representations lies in balancing both objectives without significantly compromising downstream performance or invariance. It is instructive to consider why we might consider it acceptable to sacrifice downstream performance to enforce this criterion. Systems for automated decision-making are typically part of a larger framework and subordinate to a larger objective. For example, in the field of high-energy physics the primary goal of an experimental analysis is to make an accurate and precise measurement of some physical observable. This observable is related to some physical signal process of interest and obscured by noise from background events of no interest. Modern experimental analyses make use of learning algorithms which use information about event kinematics and geometry measured by the experimental detector to distinguish signal from background events. This is achieved by sampling events with signal-like characteristics based off the available event information from the total collection of events. These systems are only useful to the extent that they decrease the signal-to-noise ratio and therefore increase the statistical significance of the measurement by preferentially selecting signal events over background events. It is possible for such systems to compromise the statistical significance of the measurement by introducing a sampling bias into the distributions of certain physically important variables. If these variables are used to model the signal-plus-background/background hypotheses, or if the parameters of proposed models of New Physics depend on these variables, this can introduce systematic errors into the final measurement of the analysis which are difficult to characterize (Section 2.2). Thus analysts may wish for these systems to make decisions which are independent of these physically important variables, potentially sacrificing signal-background discrimination power for a net increase in measurement significance. Another example comes from the field of algorithmic fairness (Dwork et al., 2012). The motivation behind this field stems from the growing trend of delegating consequential real-world decisions to automated systems which base their logic on historical statistics. For example healthcare provision, recidivism prediction, or approval of finance. Systems trained on historical data are likely to inherit the past bias inherent in the dataset - resulting in these systems discriminating against demographic subgroups in the population. To prevent these systems from exhibiting discriminatory behaviour, it may be necessary to compromise some metrics of performance to force decisions to be invariant to certain demographic attributes of the data.

This contribution of this thesis is a new set of methods for addressing this problem. This thesis is divided into two parts, which are largely independent of one another. The first part of this thesis is about constraining the optimization procedure by which the representation is learnt to reduce the informativeness of the representation of the given protected quantities, such that the representation is invariant to changes in these quantities. The second part of this thesis approaches the problem from a latent variable model perspective. In a latent variable model, additional unobserved (latent) variables are introduced which seek to explain the interaction between different attributes of the observed data. These latent variables can be interpreted as a more fundamental, compact lower-dimensional representation of the original high-dimensional unstructured data - one that is more amenable to manipulation to achieve the desired outcome of representation invariance. In the second section, we focus our attention to the field of *disentangled* representation

learning in latent-variable models. Here, every latent variable should capture a different semantic aspect of the observed data, and be independent of the other factors of variation, forming a minimal but informative representation of the original data. This allows downstream tasks to only access a relevant subset of the learned representation without being influenced by protected attributes of the original data.

Both parts of this thesis approach invariant representation learning as a machine learning problem, and heavily rely on the ability of neural networks to approximate an arbitrary function. We do not focus on neural network models per se, but their application to the approximation of quantities which are difficult to estimate, or computationally intractable (Section 3.1.2). This thesis assumes familiarity of the reader with basic concepts from probability and neural networks, of which a sufficient overview may be found in Bishop (2006).

We demonstrate the feasibility of both approaches through application to a challenging classification task in precision flavor physics. The objective of this task is to distinguish between signal events of interest and background events while making such decisions independent of certain physically important variables. Experimental analyses in high-energy physics employ an event selection strategy to separate signal events of interest and background events. If this strategy makes decisions conditional on the value of certain physically important variables, this can lead to a bias in the distribution of these variables under optimal event selection, ultimately leading to degradation in the measurement significance of physical observables (Section 3.3.2). As such, we would like to ensure event selection procedures do not rely on these observables. The main difficulty to overcome is that the protected variables are highly informative of the target variables, making it difficult to withhold information about these variables from the event selection procedure without significantly decreasing the signal efficiency to background acceptance ratio of the final event selection. We demonstrate that it is possible to learn a representation which is invariant to the given protected physical variables, while remaining informative of the original observed data, so it may be fruitfully used in downstream classification tasks. We note that our attempts to resolve this issue formed the provenance of this thesis.

1.1 Theoretical Preliminaries

The Standard Model (SM) is a theory developed during the second half of the last century that aims to describe the fundamental particles and their interactions. However, there are many experimentally observed phenomena in high-energy physics that the SM fails to address. As a consequence, it is widely accepted that the Standard Model emerges as the low-energy limit of a more fundamental ultraviolet-complete theory that is well-defined at arbitrarily high energy scales. In this viewpoint, the SM is an effective field theory at the electroweak scale characterized by the magnitude of the vacuum expectation value of the Higgs, $v = 247$ GeV.

Current efforts at particle colliders to uncover the structure of physics at high energy scales can be divided into two categories, the energy frontier and the intensity frontier. The first involves the direct production of new particles and is limited by the achievable beam energy. The second indirectly probes high-energy scales through precision measurements of specific SM processes that may reveal new physics (NP) contributions through indirect interference effects in rare, or suppressed, decays that occur at second order in perturbation theory in the Standard Model. The work of this thesis falls into the latter category. The advantage of this category is that indirect searches for NP may be sensitive to phenomena which occur at significantly higher mass scales than the beam energies currently achievable at the energy frontier. At the intensity frontier, signatures of potential NP are conjectured to be found in deviations of measurements of observables associated with such suppressed decays from the Standard Model prediction. This makes accurate and precise measurements of such observables of paramount importance.

1.1.1 Flavor-Changing Neutral Currents

This thesis is primarily concerned with the analysis of flavor-changing neutral current (FCNC) decays of B mesons. These processes have a low SM branching fraction due to multiple suppression factors:

- The GIM mechanism (Glashow et al., 1970) prohibits FCNC processes from occurring at first order of perturbation theory.
- Smallness of off-diagonal CKM matrix elements in the associated effective Hamiltonian: $V_{tq} \sim 10^{-2}, q = d, s$.
- Helicity suppression in radiative and leptonic decays due to the pseudoscalar B meson.

Suppression of the FCNC transition is a consequence of the particle content and the particular Yukawa couplings found in the SM. As generic NP processes may be unaffected by these considerations, it is possible for NP to enter physical observables at a similar magnitude to the SM contribution, making NP contributions potentially easier to detect than in

tree-level charged flavor changing processes. This makes FCNCs an important probe of new physics. FCNC transitions are permitted to occur at second-order, or at loop-level in the SM. Two quark-level FCNC transitions are possible: $b \to s$ and $b \to d$, proceeding through two charged current interactions involving an up-type quark and the exchange of a virtual W boson (Figure 1.1, left). These processes are commonly referred to as penguin decays, ostensibly due to the superficial resemblance of their corresponding Feynman diagrams to a penguin. In a wide range of extensions to the SM, hypothetical new TeV-scale particles capable of coupling to quarks may enter the loop with amplitudes comparable with the SM contribution, pushing experimental observables away from their SM predictions (e.g. Figure 1.1, right). Furthermore, as the hypothetical NP contributions to the internal loop may be off-shell, FCNCs provide a window to search for NP at mass scales currently inaccessible to searches reliant on direct production.

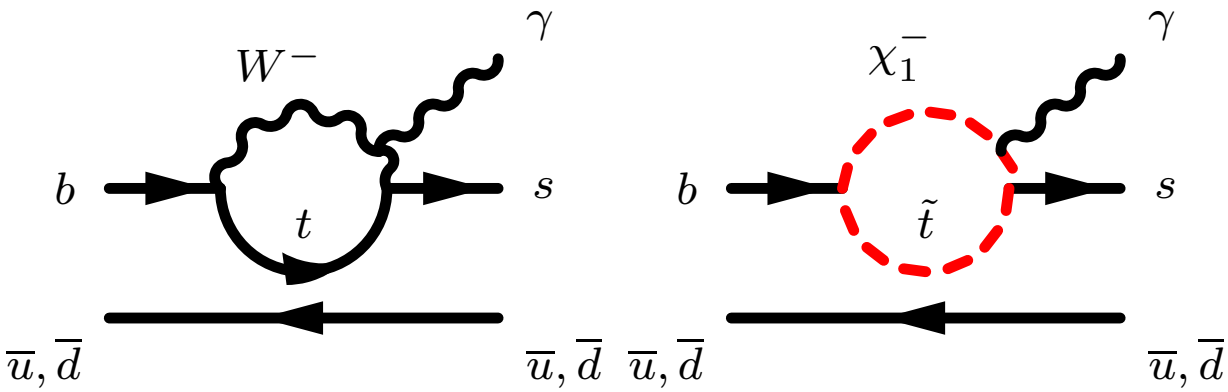


Figure 1.1: Leading order SM contribution (right) and hypothetical non-SM supersymmetric contribution (left) to the radiative decay $b \to s\gamma$. In the left diagram a heavy chargino and light scalar top (stop) is present in the virtual loop. Reused with permission from Hurth and Nakao (2010).

This thesis is primarily concerned with measurement of the rate of radiative penguin decays, involving the emission of a high-energy photon in the final state and electroweak penguin decays, associated with production of a dilepton pair $\ell^+\ell^-, \ell = e, \mu$ in the final state. Radiative and electroweak penguin decays provide exceptionally theoretically clean predictions as their final states contain color singlet photons and leptons, respectively. Both categories of decays also exhibit characteristic experimental signatures in the final state - namely the energetic radiative photon in the $b \to s(d)\gamma$ transition and the high-momentum dilepton pair in the $b \to s\ell\ell$ transition, aiding experimental detection.

Effective Field Theory

In B-physics, we wish to study the phenomenology of weak processes with characteristic energies much lower than the electroweak scale, while remaining sensitive to new physics at high energy scales. To do so, it is convenient to introduce an effective low-energy theory obtained by removing the heavy particles in the SM as dynamical degrees of freedom. The virtual effects of the excluded heavy particles are now incorporated into the modified couplings between the remaining light fields. This allows us to parameterize the effect of new physics at high energy on low-energy observables that can be measured at collider experiments. As a concrete example, flavor processes can receive contributions from charged Higgs scalars proposed by the Two-Higgs Doublet Model of Type II. Here the charged Higgs may appear as a mediator in the loop process which only receives a nominal contribution from a charged W boson in the Standard Model. The ratio of the $b \to s\gamma$ branching fraction to the leptonic decay branching fraction $R_H = \mathcal{B}(B \to X_s\gamma)/\mathcal{B}(B \to X\ell\nu)$ can be used to constrain the mass of the charged Higgs.

Application to B decays

B-decays are electroweak processes with a characteristic energy scale $\mu \ll v$. The effective theory is obtained by removing the heavy $W^\pm, Z^0$ fields, the Higgs field, and the top quark as dynamical degrees of freedom. The dynamics of this effective theory are encoded in the following effective Hamiltonian (Donoghue et al., 2014):

$$\mathcal{H}_{\text{eff}} = \frac{G_F}{\sqrt{2}} \left(\sum_i \lambda^i_{\text{CKM}} C_i(\mu) Q_i(\mu) + \text{h.c.} \right). \tag{1.1}$$

Here G_F is the Fermi constant, $\lambda_{\rm CKM}$ denotes products of CKM matrix elements, Q_i are local effective operators constructed from light fields, and C_i are the corresponding Wilson coefficients. Non-local interactions arising from virtual exchange of heavy particles are replaced with a set of local interactions, encoded in the operators Q_i.

Depending on the NP scenario, the heavy degrees of freedom excluded from the effective theory may be hypothetical high-mass scale particles such as charged Higgs or supersymmetric particles. As a result, the Wilson coefficients may receive NP contributions:

$$C_i(\mu) = C_i^{\rm SM}(\mu) + C_i^{\rm NP}(\mu). \tag{1.2}$$

NP may also enter through the addition of new operators $Q_i^{\rm NP}$ with different Dirac structure.

Effective Hamiltonian for penguin decays

The effective Hamiltonian relevant to the $b \to q\gamma$ and $b \to q\ell^+\ell^-$ transitions, $q = s, d$ reads

$$\mathcal{H}_{\rm eff}^{b\to q\gamma} = -\frac{4G_F}{\sqrt{2}} \left[\lambda_t^{(q)} \sum_{i=1}^{10} C_i Q_i - \lambda_p^{(q)} \sum_{i=1}^{2} C_i Q_i^p + \text{h.c.} \right]. \tag{1.3}$$

With the CKM factors $\lambda_r^{(q)} = V_{rb}V_{rq}^*$. For a large class of proposed extensions to the SM, significant NP effects in radiative/electroweak transitions appear in the Wilson coefficients of the following dimension-6 operators.

- Electromagnetic and chromomagnetic dipole operators

$$Q_7 = \frac{e^2}{16\pi^2} m_b (\bar{q_L} \sigma^{\mu\nu} b_R) F_{\mu\nu} \qquad Q_8 = \frac{\alpha_s^2}{16\pi^2} m_b \left(\bar{q_L} \sigma^{\mu\nu} \lambda^a b_R\right) G^a_{\mu\nu} \tag{1.4}$$

- Vector and axial-vector leptonic operators, which only contribute to $b \to s(d)\ell^+\ell^-$ modes.

$$Q_9 = \frac{e^2}{16\pi^2} (\bar{q_L} \gamma_\mu b_L) \sum_\ell \left(\bar{\ell}\gamma^\mu \ell\right) \qquad Q_{10} = \frac{e^2}{16\pi^2} (\bar{q_L} \gamma_\mu b_L) \sum_\ell \left(\bar{\ell}\gamma^\mu \gamma_5 \ell\right) \tag{1.5}$$

Where λ^a are the $SU(3)$ color generators, $F_{\mu\nu}, G_{\mu\nu}$ are the electromagnetic and chromomagnetic field strength tensors, respectively, and the subscripts L, R refer to the left- and right-handed fermion field components. NP contributions to FCNC decays manifest most strongly in the dipole and leptonic operators for a large class of SM extensions - making the coefficients C_7 and C_8 of special interest.

Constraints on the Wilson coefficients are obtained by performing a global fit to observables associated with the $b \to s(d)$ transitions. The aim is to determine permissible values for NP contributions to the Wilson coefficients $C_i^{\rm NP}(\mu)$, constraining the parameter space of NP models by requiring that model predictions should lie within a certain range of the experimentally measured values, typically 1-2σ (Blake et al., 2017). Currently measurement of the inclusive branching ratio of the $b \to s\gamma$ decay represents the strongest constraint on C_7 and NP in the $b \to s\gamma$ transition (Paul and Straub, 2017), highlighting the importance of a high-statistics experimental measurement of $\mathcal{B}(b \to s\gamma)$. Measurement of observables associated with the $b \to s\ell\ell$ transition can provide constraints on the lepton current coefficients C_9, C_{10} in addition to C_7, C_8.

1.1.2 Theoretical Predictions

Radiative decays based on the $b \to s(d)\gamma$ transitions are an important test of SM extensions. To leading order in the SM, these processes are generated by the electromagnetic dipole operator $Q_7^{(q)}$. However, NP effects may modify their respective Wilson coefficients.

$b \to s\gamma$

The branching ratio of the inclusive decay $B \to X_s\gamma$[1] strongly constrains the $b \to s\gamma$ transition. The $B \to X_s\gamma$ decay width is given by the quark-level transition in the heavy b quark limit. The leading nonperturbative corrections to the

[1] Let q denote the spectator quark of the B meson, then X_s denotes a charmless hadronic final state with $s\bar{q}$ flavor quantum numbers

decay rate from this partonic approximation are suppressed by factors of $\mathcal{O}\left(\Lambda_{\text{QCD}}/m_b\right)$ - where $\Lambda_{QCD} \approx 200$ Gev is the characteristic QCD energy scale:

$$\Gamma(B \to X_s\gamma) = \Gamma(b \to s\gamma) + \mathcal{O}(\frac{\Lambda_{\text{QCD}}}{m_b}) \tag{1.6}$$

As the b quark is heavy relative to the QCD scale, nonperturbative effects are suppressed, making radiative B decays theoretically clean. The perturbative part of the calculation, determining $\Gamma(b \to s\gamma)$ is highly nontrivial and results in the next-to-next-to-leading-order (NNLO) prediction of the $B \to X_s\gamma$ branching fraction for photon energy $E_\gamma > 1.6$ GeV Misiak (2015)

$$\mathcal{B}(B \to X_s\gamma)_{\text{SM}} = (3.36 \pm 0.23) \times 10^{-4} \quad (E_\gamma > 1.6 \text{ GeV}) \tag{1.7}$$

The total uncertainty is given by the sum in quadrature of nonperturbative, higher-order, interpolation and model-parametric uncertainties, at $(\pm 5, \pm 3, \pm 3, \pm 2)\%$ respectively.

$b \to d\gamma$

As the values of the Wilson coefficients in the effective Hamiltonian are equal up to CKM factors, the decay rate of the $b \to d\gamma$ transition is expected to be suppressed by an approximate factor of $|V_{td}/V_{ts}|^2$[2] at leading order.[3]

In evaluation of the branching fraction for $B \to X_d\gamma$ at NNLO, the tree-level transition $b \to u\bar{u}d\gamma$ leads to a non-negligible contribution to the inclusive photon energy spectrum and must be taken into account. Misiak et. al. Misiak (2015) find:

$$\mathcal{B}(B \to X_d\gamma)_{\text{SM}} = (1.73^{+0.12}_{-0.22}) \times 10^{-5} \quad (E_\gamma > 1.6 \text{ GeV}) \tag{1.8}$$

1.1.3 Experimental Techniques

The rarity of FCNC decays makes them experimentally challenging to observe, as background events misidentified as signal events form the overwhelming majority of any collected experimental sample. To obtain an measurement with acceptable statistical significance, selection criteria based on variables built from kinematic and geometric properties of the event must be carefully constructed to excise background while efficiently retaining signal events.

Fully inclusive methods

Measurements of the $B \to X_s\gamma$ branching fraction have been performed by the B-factory experiments (Altmannshofer et al., 2018) using both fully inclusive and semi-inclusive approaches. Each method provides access to different observables. The fully inclusive measurement eschews reconstruction of the signal hadronic final state and is based upon detection of a hard photon with $E_\gamma \sim m_b/2$. To control the background level, the non-signal B-meson (the *tag* B) is fully reconstructed in a hadronic decay mode (*hadronic* tag) or from a high-momentum lepton from semileptonic decay of the tag B meson (*leptonic* tag). This permits measurement to the CP flavor tag for the combined $B \to X_q\gamma$ decays and provides access to the direct CP and isospin asymmetries. Fully inclusive methods cannot separate the contributions from $b \to s\gamma$ and $b \to d\gamma$, and instead estimate the relative contribution from the CKM expectation.

Fully inclusive methods tend to be theoretically cleaner due to freedom from uncertainties associated with fragmentation of the hadronic system, however the measured value of E_γ is subject to the energy resolution of the calorimeter of the experiment, while the statistical uncertainty and minimum permitted photon energy heavily depend on the background rejection capability of the experiment. In these studies continuum background is suppressed using features based on event-topology and the dominant background arises from generic B decay modes that produce photons through secondary meson decays (predominantly $\pi^0 \to \gamma\gamma$) which mimic the radiative photon.

Semi-inclusive methods

In the semi-inclusive, or *sum-of-exclusives* method, the hadronic system is reconstructed from a set of exclusive hadronic final states. Limits on the number of species of particles reconstructed[4] and the maximum multiplicity of the final state

[2] As the large top quark mass makes it the dominant contribution within the $W^\pm$-loop

[3] Subtleties arise from the effect of CKM hierachies on the relative contribution of the current-current operators $Q^{u,c}_{1,2}$, leading to larger theoretical uncertainties compared to the $b \to s$ transition

[4] Neutral decays, such as $\pi^0 \to \gamma\gamma$ and $K^0_S \to \pi^+\pi^-$ introduce a larger background upon reconstruction

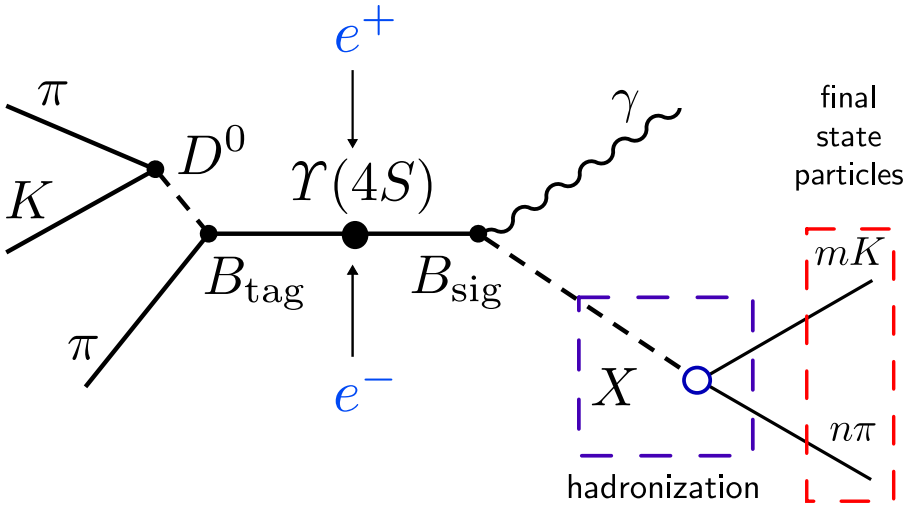


Figure 1.2: Schematic of typical $B \to X\gamma$ event. The signal B meson B_{sig} decays into a hadronic portion X and energetic photon γ. The hadronic portion can decay into an indefinite number of kaons K (for the $b \to s\gamma$ transition) and pions π in the final state. Semi-inclusive methods explicitly reconstruct the signal as the sum of several different exclusive hadronic final states. High-multiplicity states or states with a large number of neutral intermediate particles tend to be challenging to reconstruct reliably. The tag B meson produced alongside the signal B is typically not reconstructed in a semi-inclusive analysis.

are dictated by the statistics and background rejection capabilities of the experiment. The photon energy in the B rest frame may be determined from the total momentum of the hadronic final state via $E_\gamma = (m_B^2 - m_{X_s}^2)/2m_B$ with a resolution significantly better than direct photon energy measurement with the calorimeter. This method results in a high signal efficiency relative to inclusive methods and can distinguish the $b \to s$ and $b \to d$ contributions. This method also determines the flavor and charge of the signal B meson and hence can be used to extract the direct CP and isospin asymmetries. Both Belle (Saito, 2015) and BaBar (Lees, 2012) have performed semi-inclusive analyses of the $b \to s\gamma$ transition, using 711 fb^{-1} and 429 fb^{-1} of e^+e^- collision data collected at the $\Upsilon(4S)$ resonance, respectively. The most recent methods published by both experiments reconstruct the hadronic final state as a sum of 38 exclusive decay modes, suppress continuum background via event-topology features and extract the branching fraction via a fit to the beam-constrained mass m_{bc}. The 38 final states considered are primarily $B \to Kn(\pi)\gamma$ modes, where the K stands for K^+ or K_S^0 and $n(\pi)$ refers to 1-4 pions of which up to two can be a π^0. The considered decay modes are estimated to account for approximately half of the inclusive rate for $B \to X_s\gamma$ according to simulation based on known decay rates. Assuming isospin symmetry between K_L^0 and K_S^0, the proportion of measured final states is approximately 75% for both analyses. A schematic of a typical $B \to X\gamma$ event is shown in Figure 1.2.

The main limitation of this method stems from the systematic uncertainty associated with hadronization models used when simulating the physical decay process and estimating the fraction of missing decay modes excluded from the analysis. It will be possible to exploit the increase in luminosity at Belle II to include higher-multiplicity heavy decay modes in the analysis. In particular, the improved particle identification (PID) capabilities further increase the sensitivity of $b \to d\gamma$ analyses. This will enable a better understanding of the hadronic spectrum and the fragmentation of the X_q system, reducing the associated systematic uncertainties of the analysis.

After reconstruction and application of preliminary selection criteria in analyses of FCNC decays, the signal-to-noise ratio of the resulting sample is still typically too low to be of use in measurements of observables to claim discovery or exclusion of new physics at an acceptable level of statistical significance. Therefore, experimental analyses employ some sort of supervised learning algorithm (Section 2.1.1) to learn the structure of the joint distributions of a selected collection of discriminative variables in an automated manner. Events are categorized as signal or background on the basis of this learnt model. Such techniques are commonly referred to as machine learning or multivariate analysis (MVA) techniques, as they exploit the interaction between multiple different variables instead of standard selection criteria, which typically make use of linear decision boundaries on individual variables. Variants of boosted decision trees (Keck, 2017) have seen heavy usage for this purpose owing to their high degree of interpretability, however 'black-box' algorithms such as neural networks have seen increasing adoption in modern analyses due to apparently superior performance levels (Keck (2017), Apostolakis et al. (2018)). MVA methods have installed themselves as an integral part of modern analyses and will become increasingly important in the continued absence of an unambiguous signature of new physics and the required sensitivity of experimental analyses increases. However, there are important

caveats to the application of machine learning to experimental high-energy physics, which we discuss in depth in Chapter 2.

1.1.4 Experimental Results

Theoretical predictions are typically given with $E_\gamma > 1.6$ GeV to control non-perturbative uncertainties. However, in experimental measurements the minimum E_γ is restricted by the large combinatorial background at low values of the radiative photon energy. Experiments at the B-factories currently involve measurements with $E_\gamma \geq 1.7$ GeV, requiring a model-dependent extrapolation down to the full energy range, performed by the Heavy Flavor Averaging Group. The quoted world average of extrapolated values is (Amhis et al., 2016):

$$\mathcal{B}(B \to X_s\gamma) = (3.43 \pm 0.21 \pm 0.07) \times 10^{-4} \quad (E_\gamma > 1.6 \text{ GeV}). \tag{1.9}$$

Where the first error is from statistical and systematic error combined in quadrature and the second is due to the model dependent extrapolation. This is fully consistent with the SM expectation (Misiak, 2015).

Relative to the $b \to s\gamma$ transition, the $b \to d\gamma$ transition remains largely experimentally untested. The $b \to d\gamma$ decay proceeds through a similar loop diagram to $b \to s\gamma$, but carries an additional suppression factor given by the squared ratio of CKM matrix elements $|V_{td}/V_{ts}|^2$. Measurement of the $b \to d\gamma$ branching fraction is an avenue to determination of $|V_{td}|$, the length of the least-known side of the unitarity triangle. A fully-inclusive measurement is not possible due to the dominant $b \to s\gamma$ background; the sum-of-exclusives method has exclusive access to $\mathcal{B}(B \to X_d\gamma)$. BaBar (del Amo Sanchez, 2010) has performed a measurement of $\mathcal{B}(B \to X_d\gamma)$ using a sum of 7 exclusive X_d final states with hadronic mass threshold 2.0 GeV/c^2 using the full 429 fb^{-1} dataset collected by BaBar. The final states involve up to 4 pions with at most one neutral pion, with the neutral meson decays $\pi^0 \to \gamma\gamma$ and $\eta \to \gamma\gamma$ decays explicitly reconstructed. After correcting for unobserved decay modes, they obtain the ratio

$$\frac{\mathcal{B}(B \to X_d\gamma)}{\mathcal{B}(B \to X_s\gamma)} = 0.040 \pm 0.009 \text{ (stat.)} \pm 0.010 \text{ (sys.)} \tag{1.10}$$

, which is consistent with the value of $|V_{td}/V_{ts}|$ derived from $B^0_{(s)} - \bar{B^0_{(s)}}$ mixing as well as the SM expectation. Here the dominant uncertainty is systematic due to the low fraction of modes considered. The dominant systematic uncertainty from missing modes can be reduced, potentially down to 10% , by adding higher-multiplicity exclusive final states to the analysis, which will be possible with higher 'luminosity achievable by Belle II. Furthermore, the upgraded particle identification capabilities at the Belle II detector will improve separation of charged pions and kaons, suppressing the large $b \to s\gamma$ background.

1.1.5 Constraints on New Physics from $\mathcal{B}_{s\gamma}$

Both the SM prediction and the most recent experimental average for $\mathcal{B}_{s\gamma} = \mathcal{B}(B \to X_s\gamma)$ possess similar levels of uncertainty and are consistent at the 1σ level, placing strong phenomenological constraints on the space of NP models (Blake et al., 2017). One of the most closely linked extensions of the SM to the $b \to s$ transition is the two-Higgs-Doublet-Model (2HDM), a phenomenologically rich model that arises in the minimal supersymmetric extension of the SM. Within the 2HDM, the scalar spectrum consists of two charged scalars $H^\pm$, one neutral pseudoscalar A^0 and the neutral scalars H^0, h^0, of which one is to be identified with the SM-like Higgs boson. The most generic 2HDM class permits tree-level FCNCs. In the aligned-2HDM (a2HDM), the alignment in flavor space of the Yukawa matrices for each type of right-handed fermion is assumed, enforcing the absence of tree-level FCNCs. The neutral components of the two complex scalar $SU(2)_L$ doublets $\phi_i(x), i = 1, 2$ introduced acquire respective VEVs $\langle 0| \phi_i(x) |0\rangle = v_i/\sqrt{2}$; $\quad v^2_{SM} = v^2_1 + v^2_2$. The rotation angle $\tan\beta = v_2/v_1$ defines the $SU(2)$ transformation that diagonalizes the mass matrices of the charged scalar and pseudoscalar fields, intepreted as a rotation from the flavor to Higgs basis.

$$\begin{pmatrix} \Phi_1 \\ -\Phi_2 \end{pmatrix} = \mathcal{R}(\beta) \begin{pmatrix} \phi_1 \\ e^{-i\theta}\phi_2 \end{pmatrix} \tag{1.11}$$

With θ the relative phase between the doublets. The charged bosons $H^\pm$ may couple to SM fermions through a generalization of the Yukawa interactions - hence $H^\pm$ scalar exchanges in the a2HDM lead to NP contributions to the Wilson coefficients $C_{7,9,10}$ at the matching scale through interference in the Z^0, γ penguin diagrams, Figure 1.3; $C_{7,9,10} = C^{\text{SM}}_{7,9,10} + C^{H^\pm}_{7,9,10}$. The magnitude of this contribution is dependent on the coupling strength of the charged

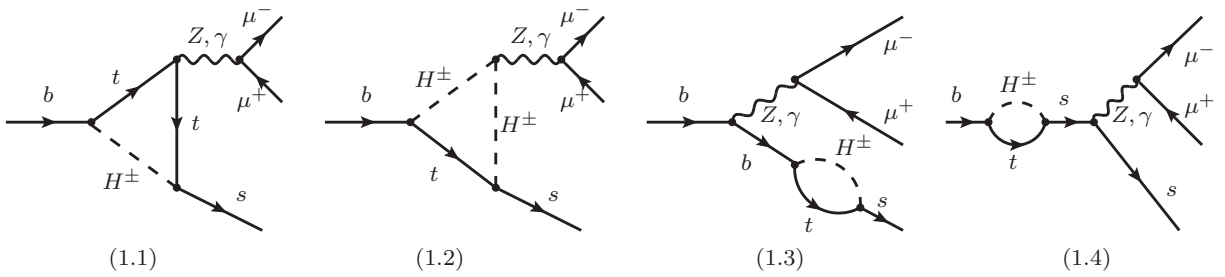


Figure 1.3: Hypothetical heavy scalar $H^{\pm}$ contribution to Z and γ penguin diagrams in the a2HDM. Reused with permission from Hu et al. (2017)

scalar with the fermion mass-eigenstate fields. The combined experimental average 1.9 severely constrains the Wilson coefficient $C_7^{H^{\pm}}$, enforcing the constraint at the matching scale $\mu = 160$ GeV (Hu et al., 2017):

$$-0.0634 \leq C_7^{H^{\pm}}(\mu) + 0.242\, C_8^{H^{\pm}}(\mu) \leq 0.0464. \tag{1.12}$$

The NP contributions to $C_{7,8}$ then place the following $\tan\beta$ independent bound on the charged Higgs mass in the a2HDM (Misiak and Steinhauser, 2017):

$$M_{H^{\pm}} > 480 \text{ GeV} \quad (95\%\ CL). \tag{1.13}$$

Global fits based on observables associated with $b \to s\ell^+\ell^-$ decays can further constrain the a2HDM parameters (Hu et al., 2017). Through measurement of the $b \to s\gamma$ transition, powerful constraints on the Higgs sector can be derived from flavor physics.

1.1.6 Lepton Flavor Universality

The $b \to s\ell\ell$ transition provides a further probe for NP through the ratio of the branching fraction between muon and electron final states:

$$R(X) = \frac{\mathcal{B}(B \to X\ \mu^+\mu^-)}{\mathcal{B}(B \to X\ e^+e^-)}. \tag{1.14}$$

Where X is the hadronic portion of the semi-leptonic decay. Such measurements test universality of gauge couplings in the lepton sector. In the SM, the electroweak gauge bosons $W^{\pm}$ and Z^0 have flavor-blind couplings to leptons, hence this ratio $R(X)$ is expected to be close to unity under assumption of lepton flavor universality (LFU). LFU violation would be indicative of NP effects preferentially coupling to one flavor over the other. The LHCb collaboration has performed measurements of $R(K)$ and $R(K^*)$ that exhibit a 2.6σ and $2.1 - 2.5\sigma$ deficit from the SM prediction, respectively (Aaij et al., 2017). Measurement of $R(K^*)$ and $R(K)$ at Belle II would provide important independent verification of the anomalies reported by LHCb, whereas the inclusive measurement of $R(X_s)$ will be uniquely accessible at Belle II. Numerous NP models have been proposed to explain the observed discrepancy between the SM prediction and measurements of observables associated with the $b \to s\ell\ell$ transition. These generally contain input parameters which are strongly dependent on the dilepton invariant mass $m^2_{\ell\ell} = q^2$ and $z = \cos\theta$, the cosine of the angle between the 3-momentum of the positively charged lepton and parent B meson in the dilepton center of mass frame. For example, the double differential decay width is directly parameterized by z and q^2 (Section 9.4, Altmannshofer et al. (2018)):

$$\frac{d^2\Gamma}{dq^2dz} = \frac{3}{8}\left[(1+z^2)H_T(q^2) + 2zH_A(q^2) + 2(1-z^2)H_L(q^2)\right]. \tag{1.15}$$

Where H_T, H_A, H_L are independent observables which are functions of q^2. As the Wilson coefficients describing the transition (Eq. 1.1) are extracted from experimental measurements of such observables, introducing sampling bias into the distribution of z and q^2 ultimately leads to increased systematic uncertainty in the determination of the associated Wilson coefficients. Therefore, the efficiency of methods for lepton identification at particle colliders should remain constant across momentum space to avoid introducing systematic effects into these measurements, a problem we discuss in Chapters 2 and 3.

1.2 THE BELLE II EXPERIMENT AT SUPERKEKB

The Belle II physics program revolves around precision measurement of experimental observables in the flavor and heavy quark sector. Deviations of measured observables from theoretical predictions can be interpreted in terms of new physics. Searches for NP at Belle II are complementary to direct production searches at the LHC which rely on a high center of mass energy to produce otentially undiscovered heavy particles. Measurements of processes sensitive to quantum corrections such as FCNC decays will form an integral part of the physics program at Belle II. Such processes involve internal loops which may receive non-negligible contributions from NP at high-energy scales. With the expected dataset at Belle II, searches for FCNC processes are capable of probing energy scales up to 100 TeV, far above achievable beam energies at the LHC.

1.2.1 Previous B-physics experiments

The Belle II experiment is a second generation B factory. The Belle experiment at the KEKB collider in KEK together with the BaBar experiment at the PEP-II collider in SLAC formed the first generation of B factories. Both experiments utilized an asymmetric e^+e^- collider and collected a combined integrated luminosity of 1.5 ab^{-1} at the $\Upsilon(4S)$ resonance. The Belle and BaBar experiments provided important insights into the flavor structure of the Standard Model, validating with good precision the theoretical predictions of the Standard Model. Notably, the observation of CP violation in the B meson system prompted the award of the 2008 Nobel Prize to M. Kobayashi and T. Maskawa for their theory of CP violation. However, the significance of many interesting measurements of flavor observables remain limited by statistics, motivating the construction of the SuperKEKB Accelerator. It should be mentioned that the LHCb experiment at CERN has been taking data since 2009. The LHC produces high-energy hadron-hadron collisions and cannot exploit the clean environment of the $\Upsilon(4S)$ resonance, however, abundant $pp \to b\bar{b}$ pair production counteracts the lower reconstruction efficiency relative to the B-factories.

1.2.2 The SuperKEKB Accelerator

The Belle II experiment is located at the SuperKEKB accelerator complex in Tsukuba, Japan. SuperKEKB is an asymmetric-energy, double ring e^+e^- collider that represents a significant upgrade over its predecessor, the KEKB accelerator. The accelerator consists of two rings 3 km in circumference where electron and positron beams are accelerated to 4 and 7 GeV, respectively, and subsequentally crossed to collide at the center of the Belle II detector. The beams are tuned to exhibit a center of mass energy $E_{\text{CMS}} = \sqrt{s} = 10.58$ GeV, corresponding to the mass of the $\Upsilon(4S)$ resonance. The $\Upsilon(4S)$ subsequently decays into an entangled $B\bar{B}$ pair with a branching ratio above 96%. As $\sqrt{s}$ lies just above the threshold for $B\bar{B}$ production, the B meson pair is produced without accompanying particles, leading to a lower multiplicity event environment relative to hadron colliders.

The upgraded collider is expected to deliver an integrated luminosity of 50 ab^{-1}, which will contain approximately 5×10^{10} events at the $\Upsilon(4S)$ resonance, a factor of 50 larger than the combined dataset of the Belle and BaBar experiments. The accelerator is designed to achieve an instantaneous luminosity 40 times greater with respect to KEKB, achieved by a more focused beam crossing and an increase in the beam current. The higher statistics collected at the $\Upsilon(4S)$ resonance at Belle II will enable higher-precision checks against Standard Model predictions and further investigation into the flavor anomalies observed at previous B-factories and LHCb.

1.2.3 The Belle II Detector

The Belle II detector has been upgraded from its Belle counterpart to handle the greater luminosity and beam background, approximately 20 times higher than the Belle experiment. Importantly for the study of rare B decays, the redesigned tracking system and particle identification systems enhance kaon/pion separation power at low momenta and increases the neutral particle reconstruction efficiency relative to the Belle detector. A detailed overview of the components of the Belle II detector can be found in Altmannshofer et al. (2018).

Tracking Systems

The tracking system at Belle II consists of the vertex detector (VXD) and the central drift chamber (CDC). The VXD consists of a two-layer depleted field effect transistor pixel detector (PXD) and a four-layer silicon strip detector (SVD)

around the collision point. It is designed to provide precise measurements of the impact parameter and secondary vertex of charged particles. Positioning the detector closer to the interaction point, combined with the upgraded vertex detectors improves the impact parameter resolution in the beam direction by a factor of approximately 2 for sub-GeV momenta. Importantly for inclusive reconstruction of the $B \to X_s\gamma$ decay, the SVD improves the reconstruction efficiency of $K^0_S \to \pi^+\pi^-$ decays due to the larger outer radius coverage.

The CDC is a cylindrical wire drift chamber surrounding the vertexing detectors. Upgraded drift cells near the beam pipe, extended outer radius coverage and faster readout electronics improve the momentum and dE/dx resolution relative to Belle, which aids track reconstruction and measurement of long-lived particles.

Particle Identification Systems

Rare decay channels place stringent requirements on the charged π/K separation capabilities of the detector. In particular, reliable π/K separation is crucial for studying the charmless $b \to d\gamma$ transition. Relative to the $b \to s$ penguin, the $b \to d$ transition is suppressed by the CKM factor $|V_{td}/V_{ts}|^2$ - an approximate 25-fold reduction. This makes the background originating from misidentified kaons in $b \to s$ processes extremely dangerous. Furthermore, exclusive modes such as $B^0 \to K^{*0}\gamma$ and $B^0 \to \rho^0\gamma$ are kinematically similar and topologically identical, with the former possessing a branching ratio approximately 50 times larger than the latter. This makes good kaon rejection fundamental to rejection of the large $b \to s$ background in analysis of $b \to d$ transitions. It is anticipated that the reduction in the kaon misidentification rate at the same pion efficiency relative to Belle I will help to improve the significance of the signal for the $b \to d\gamma$ decay.

At Belle II, the PID systems in both the central (barrel) region and forward endcap regions of the detector are based on Cherenkov ring imaging detectors. PID is performed by the Time-Of-Propagation (Top) counter in the central region and the Aerogel Ring Imaging Cherenkov (ARICH) counter in the forward endcap region. Cherenkov photons are emitted by a charged particle moving with velocity β exceeding the speed of light in a dielectric medium with refractive index $n(\lambda)$ at a characteristic angle $\cos\theta_c = \beta/n(\lambda)$. The mass of the particle is inferred by combining the Cherenkov angle with the track momentum measured in the central tracker, identifying it as a kaon, pion or proton. Monte Carlo simulations estimate the kaon selection efficiency to exceed 95%, with a pion misidentification probability below 3% inside the barrel and forward endcap region.

Electromagnetic Calorimeter

High energy resolution and detection efficiency for photons across a wide range of momenta is extremely important as π^0 and other neutral particles decaying to photons form a third of decay products of the B meson. For radiative penguin decays, reliable reconstruction of the characteristic high-energy photon signature is crucial to signal efficiency and background rejection. The electromagnetic calorimeter (ECL) at Belle II uses thallium-doped caesium iodide crystals in the barrel region and caesium iodide crystals in the endcap regions, with a total mass of about 43 tons. To compensate for the higher beam backgrounds, the ECL is equipped with faster readout electronics that will permit improved rejection of fake calorimeter clusters based on online analysis of the cluster waveform. This will suppress beam background contamination and improve photon energy resolution. The ECL is also an important instrument for electron and muon identification, necessary for study of semi-leptonic and leptonic B decays at Belle II. These latter decays are of interest for studies of lepton flavor universality violation or precise measurement of CKM elements.

K^0_L and Muon Detector (KLM)

Identification of muons and K^0_L is important in the study of semi-leptonic and leptonic B decays at Belle II. The KLM detector at Belle II is an iron support structure which serves as the return path of the magnetic flux generated by the superconducting solenoid of the Belle II detector and an absorber to decelerate particles with a relatively long mean free path which escape the inner detectors. The KLM is split into two sections: a barrel part arranged parallel to the beam and an end-cap region oriented normally to the beam. The barrel region is instrumented with glass-electrode resistive plate counters while the end-cap region is instrumented with scintillator strips for K^0_L and muon detection.

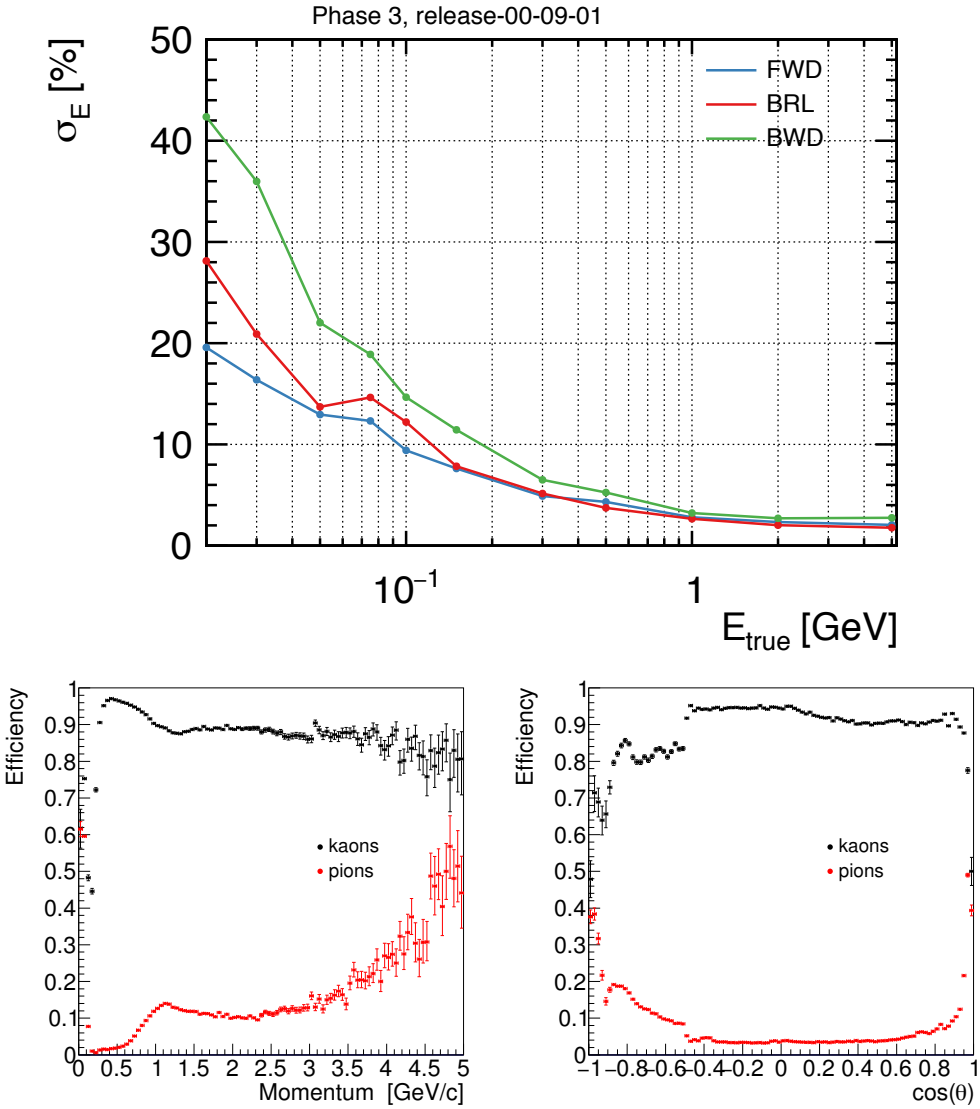


Figure 1.4: **Top:** Photon energy resolution at Belle II over individual detector regions (Ferber, 2017). At the lower threshold on radiative photon energy, $E_\gamma^{\text{min}} = 1.7$ GeV, the energy resolution is approximately 2%. **Bottom:** K/π efficiency across all detector regions (Bennett, 2017). Excellent K/π separation is possible in the barrel and forward endcap region, but is limited in the backward endcap. Note that the efficiencies in the left-hand plot have been integrated over the full θ distribution.

CHAPTER 2

INVARIANCE THROUGH REGULARIZATION

2.1 INTRODUCTION

Precision measurements in experimental high-energy physics seek to maximize the statistical significance level of measurements of physical observables. It is frequently the case that observables of interest when testing models of new physics are associated with decays which are either highly suppressed or conjectured to exist but have not been observed yet. In either case, events corresponding to these decays will be heavily obscured by background processes which closely mimic the experimental signature of the signal process. This makes the problem of isolating rare signals from data samples overwhelmingly dominated by background events a common theme in experimental high-energy physics.

However, discrimination between signal and background events is not the only desiderata for event classification algorithms in HEP. In many cases the statistical significance of the measurement - the main focus of an analysis - and the classification power of the supervised model may not be completely parallel objectives. In measurements of physical observables, the background-only and signal-plus-background hypotheses are modelled through a parametric likelihood fit on the spectrum of one or more 'fit-variables' capable of discriminating between signal and background structure by virtue of their characteristic shape. The background spectrum in such cases is typically smooth and featureless while the signal spectrum exhibits some well-understood physical characteristics, such as resonant structure. Examples of fit-variables include:

- The beam-constrained mass M_{bc} in measurements associated with hadronic B decays, defined as the invariant mass of the B meson with the energy of the B meson constrained to the colliding beam energy:
$$M_{bc}^2 = E_{\text{beam}}^2 - p_B^2 \tag{2.1}$$
Where $E_{\text{beam}} = \sqrt{s}/2$ is the beam energy in the center of mass (CM) frame of the $\Upsilon(4S)$ resonance, and p_B is the momentum of the B meson in the CM frame. The reconstructed invariant mass of the B meson candidate should be equal to the nominal B meson mass m_B, hence $M_{bc} = m_B \approx 5.28$ GeV for a correctly reconstructed B decay.
- The energy difference ΔE, defined as the difference between the beam energy in the CM frame and the energy of the B meson.
$$\Delta E = E_B - E_{\text{beam}} \tag{2.2}$$
The energy of the B meson should be equal to the beam energy when reconstructed in the CM frame, hence $\Delta E = 0$ for a correctly reconstructed B decay.
- The decay length difference Δz in CP measurements, defined as the distance between the decay positions of the B mesons originating from decay of the $\Upsilon(4S)$ resonance.

To properly characterize the systematic uncertainties involved in these measurements, the distributions of the fit-variables should remain unchanged when some learning algorithm selects a signal-enriched sample of events. In other words, the learning algorithm should not introduce a selection bias into the final sample. However, supervised learning algorithms tend to exhibit a strong dependency on these fit-variables, as these variables are constructed to clearly distinguish between signal and background events and therefore are highly predictive of the event class. Unconstrained optimization of these learning algorithms tend to result in a highly non-uniform selection efficiency in the fit-variables. In the worst case scenario, the background surviving the selection procedure is dominated by events with signal-like characteristics,

resulting in artificial sculpting of the shape of the surviving background spectrum to mimic the signal peak. An example of this is shown in Figure 2.1. This is problematic for the following reasons:

- This can lead to uncontrollable systematic uncertainties if the background model cannot be independently checked in a control region.
- A non-uniform selection efficiency in both signal and background events makes the shape of the empirical distributions of physically important variables dependent on the cut on the classifier output. This increases model reliance on Monte Carlo simulation - as the signal model usually cannot be independently checked in a control region.

These effects combine to degrade the final significance level of the measurement - a highly undesirable outcome. This illustrates the desirability of learning a model which is 'blind', or invariant to certain observables, and the need for experimental analyses to take objectives beyond discrimination power into account. Ultimately, we would like to build a classifier that makes decisions independently of certain physically important observables - such that the original distribution of these observables is preserved for any subsample of the data. In particular, when a signal-rich subsample of the collected dataset is selected by the classifier to increase the signal-to-noise ratio, we would like the distribution of these variables to remain unchanged. In this chapter we propose one method of achieving this outcome based on penalizing the dependency of the classifier output on the variables in question.

A further reason for learning classifiers that make decisions independent of certain variables arises because proposed models of new physics may contain input parameters that are strongly dependent on specific kinematic variables. An example of this is the inclusive $b \to s\ell\ell$ transition. The two main kinematic variables associated with this decay are the dilepton invariant mass $m^2_{\ell\ell} = q^2$ and $z = \cos\theta$, the cosine of the angle between the 3-momentum of the positively charged lepton and parent B meson in the dilepton center of mass frame. Important observables associated with this decay are parameterized in terms of these kinematic variables. As the Wilson coefficients of the transition (Eq. 1.1) are inferred from measurements of these observables, introducing sampling bias into the distribution of z and q^2 ultimately leads to increased systematic uncertainty in the determination of the associated Wilson coefficients. In general, if such important physical observables are affected with sampling bias introduced by classifiers intended to raise the signal-to-nosie ratio, this can introduce systematic effects into the analysis that make it difficult to assess the legitimacy of proposed extensions to the Standard Model.

The first variable listed above, the beam-constrained mass M_{bc}, is especially relevant in the analysis of the radiative penguin $b \to s(d)\gamma$ transition as the signal and background yields necessary in the measurement of physical observables associated with this process are extracted through a maximum likelihood fit to the shape of the M_{bc} distribution. As this decay has a very low branching fraction ($\mathcal{O}(10^{-4})$), it is important for learning algorithms to be powerful enough to increase the signal-to-noise ratio to a level that results in an acceptable statistical uncertainty in associated measurements while ensuring that the M_{bc} distribution remains invariant under selection. We demonstrate that our proposed method may be applied to remove classifier dependence on M_{bc} while balancing these competing requirements.

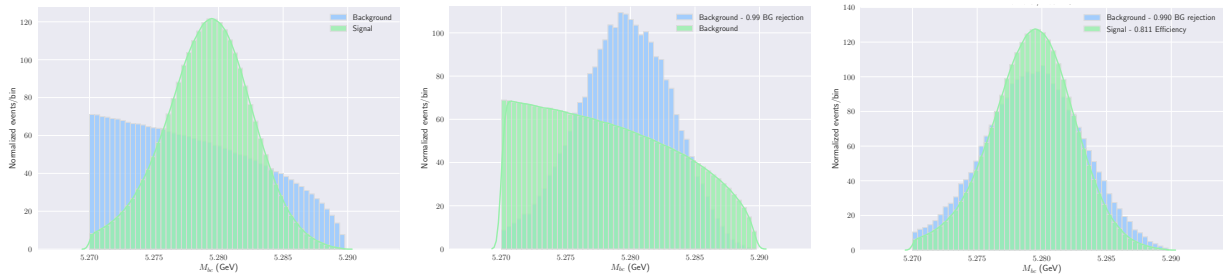


Figure 2.1: Example of background sculpting in M_{bc} for the $b \to s\gamma$ decay after selection of a signal-rich subsample by rejecting 99.5% of background events based on the output of a shallow neural network naively trained to distinguish $b \to s\gamma$ events from background. **Left:** Original signal (green) and background distributions (blue) prior to any selection. We would like the signal and background distributions to retain these distinct shapes post-selection. **Middle:** Background distribution post-selection (blue) has been significantly changed, or *sculpted* from the original distribution (green), peaking at the nominal B meson mass $m_B \approx 5.28$ GeV, as with correctly reconstructed signal events. **Right:** Sculpted background peak (blue) is degenerate with characteristic signal peak (green), making it difficult to separate the signal and signal-plus-background hypotheses.

This chapter is structured as follows. In the introduction, we briefly discuss preliminaries which lay the foundation for our main contribution in Section 2.2. We discuss the connection between our work and the adversarial procedure

outlined in Louppe et al. (2017) for achieving invariance in Section A.1. Sections 2.3 and 2.4 contain experimental results and two applications of our approach to precision high-energy physics.

2.1.1 Machine Learning

Machine learning methods aim to learn properties of the underlying structure of observed data x directly from the data itself. By learning some statistical model describing the distribution of the observed data, this information may be used to predict some unobserved target variable (classification and regression), make decisions in a dynamic environment (reinforcement learning), or retain this information as a compact representation of the data (dimensionality reduction), among many other applications. This is to be placed in contrast with traditional expert systems, which require an external oracle to encode logic in the process by hand and do not model the observed data itself.

The increased availability of high-quality data and computing power has driven interest and development in the field of machine learning, such that methods based on machine learning and neural networks have achieved state-of-the-art performance in a variety of fields - among them computer vision (He et al., 2015), natural language processing (Devlin et al., 2019), reinforcement learning (Hassabis, 2016) and generative modelling (Kingma and Dhariwal, 2018). A comprehensive overview of neural networks and machine learning may be found in Bishop (2006) and Goodfellow et al. (2016).

The subfield of machine learning that this thesis will chiefly be concerned with is supervised learning, where we wish to predict a set of target variables given some input data. We provide a short overview under the statistical learning framework below. A more comprehensive introduction to statistical learning theory may be found in Mohri et al. (2018).

Supervised Learning

In this paradigm, we wish to predict some random variable y from another random variable x. The learner has access to a finite sequence of pairs $S = \{(x_1, y_1), \dots (x_n, y_n)\}$, referred to as the *training set*. Each instance $x_i \in X \subset \mathbb{R}^D$ is a D-dimensional vector, and $y_i \in Y \subseteq \mathbb{R}$ is the corresponding instance target, or label. In the case of classification, the labels can only assume values from a finite set, whereas for regression, the target variables may be continuous. We assume the following simple generative model of the training set S: an instance x is sampled from the true underlying data distribution $p^*(x)$ and mapped to y using some unknown function $f : X \to Y$:

$$x \sim p^*(x), \tag{2.3}$$
$$y = f(x). \tag{2.4}$$

We assume the examples in the training set are independently and identically distributed (i.i.d.) according to $p^*(x)$. That is, each instance is independently sampled from $p^*(x)$ and labelled according to f. The objective of the learner is to learn some hypothesis function $h : X \to Y$ mapping instances x to the predicted output $\hat{y}$. The hypothesis function may be used to predict the label of new instances. Note that the learner has no knowledge of the true underlying data distribution $p^*(x)$ or the true labelling function f. The only information available to the learner to form the hypothesis function is the training set. Intuitively, the larger the training sample gets, the more accurately the learner will be able to model the true distribution $p^*(x)$ and labelling function f used to generate the data.

In order for the learnt hypothesis function h to be useful, it should make accurate predictions on previously unseen data. Let us assume that the possible hypothesis functions the learner can consider is restricted to some set $\mathcal{H}$ - we will discuss this motivation behind this shortly. To evaluate the quality of the learned hypothesis, the loss function $\ell : \mathcal{H} \times X \times Y \to \mathbb{R}^+$ is introduced, which maps the predicted value $h(x)$ and the true label y to some non-negative number. Popular choices of loss function include the $0/1$ loss for classification of discrete targets:

$$\ell_{0/1}(h(x), y) = \begin{cases} 0, & \text{if } h(x) \neq y \\ 1, & \text{if } h(x) = y \end{cases}, \tag{2.5}$$

the squared loss for regression of continuous targets:

$$\ell_{\text{sq}}(h(x), y) = (h(x) - y)^2, \tag{2.6}$$

or the negative log-likelihood loss for density estimation:

$$\ell_{\text{nll}}(h(x)) = -\log p(h(x)|x), \tag{2.7}$$

which is equivalent to performing maximum likelihood estimation.

The risk is defined as the expected loss of the classifier with respect to the true probability distribution over $p(x, y) = p(y|x)p^*(x)$ (where the conditional $p(y|x)$ accounts for possible stochasticity in the labelling function f):

$$R(h) = \int dx\, dy\, p(x, y)\ell(h(x), y) \tag{2.8}$$

$$= \mathbf{E}_{p(x,y)}\left[\ell(h(x), y)\right]. \tag{2.9}$$

However, the true distribution $p(x, y)$ is unknown to the learner and instead in practice we compute the empirical risk over the training dataset S:

$$\hat{R}(h) = \frac{1}{|S|}\sum_{i\in S}\ell(h(x_i), y_i). \tag{2.10}$$

The problem of finding the optimal predictor now becomes the minimization of the risk function over the training set - this paradigm is referred to as *empirical risk minimization* (ERM).

$$\hat{h} = \operatorname*{argmin}_h \hat{R}(h). \tag{2.11}$$

However, naively minimizing the empirical risk on the training dataset leads to the possibility of overfitting - where the learned hypothesis achieves excellent performance on the training set but fails to generalize to unseen examples. For instance, a function that predicts the value $h(x) = y_i$ when $x = x_i$ but 0 otherwise will achieve the minimum possible loss on the training set S but does not capture any of the underlying structure of the dataset and will generalize poorly to new examples. To resolve this, we introduce an inductive bias to the learner - before observing the data, a class of hypotheses $\mathcal{H}$ is fixed to which the hypothesis function of the learner must belong to:

$$\hat{h} = \operatorname*{argmin}_{h\in\mathcal{H}} \hat{R}(h). \tag{2.12}$$

By restricting the learner to choose a hypothesis $h \in \mathcal{H}$, we bias the solution toward a particular set of predictors. The hypothesis class $\mathcal{H}$ is typically chosen based on prior domain knowledge about the problem at hand. For example, the hypotheses learnt may be restricted to be the set of linear functions, which is the case for support vector machines and perceptrons, or decision trees, which form the basis for tree boosting models. For the learner to be able to generalize well to unseen examples, the true risk $R(h)$ should be close to the empirical risk $\hat{R}(h)$ over the training set. A significant portion of statistical learning theory is devoted to investigating the discrepancy between these quantities (Vapnik, 1992). Practical bounds on the generalization error - the difference between the empirical risk and actual risk $|R(h) - \hat{R}(h)|$ for certain hypothesis classes $\mathcal{H}$, such as neural networks, remain an active field of research.

Regularized Risk Minimization

In regularized risk minimization (RRM), the learner seeks to find a hypothesis $h \in \mathcal{H}$ which minimizes the sum of the empirical risk and a regularization function $\Omega : h \to \mathbb{R}^+$. The primary objective of regularization is to reduce the generalization gap $|R(h) - \hat{R}(h)|$ without significantly increasing the empirical risk function. The objective of RRM, subject to inductive biases now becomes

$$\operatorname*{argmin}_{h\in\mathcal{H}}\left(\hat{R}(h) + \lambda\Omega(h)\right) \Leftrightarrow \operatorname*{argmin}_{h\in\mathcal{H}} \hat{R}(h) \text{ subject to: } \Omega(h) \leq C, \tag{2.13}$$

where $\lambda > 0$ is the Lagrange multiplier of the corresponding constrained optimization problem in Eq. 2.13. For each $\lambda > 0\ \exists\ C > 0$ such that the implication holds, and vice-versa. λ can be interpreted as controlling the relative

importance of the regularization and empirical risk terms. Intuitively, the regularization function $\Omega(h)$ measures the complexity of the hypothesis h, and therefore the RRM objective contains an additional term which penalizes the hypothesis complexity. During optimization, the learner must strike a compromise between achieving a low empirical risk and learning simpler hypotheses. Regularization was traditionally motivated as a means to prevent overfitting - more complex hypotheses tend to memorize the idiosyncrasies of the training set better - achieving a low empirical risk, but exhibit poor generalization performance because the same idiosyncrasies may not feature in new examples. In this sense regularization can be viewed as imposing a 'soft' inductive bias on the possible class of hypothesis $\mathcal{H}$. One of the simplest and most widely used forms of regularization is Tikhonov, or L_2 regularization. If the learner has internal degrees of freedom represented by the weights w, then the Tikhonov regularizer is given by the L_2 norm of the weights: $\Omega(w) = \|w\|^2$. Perhaps the most well-known example of Tikhonov regularization appears in linear regression with the squared loss. Here the weights w are the regression coefficients and the learning procedure is known as ridge regression:

$$\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(\langle w, x_i \rangle - y_i \right)^2 + \lambda \|w\|^2 \right). \tag{2.14}$$

Historically, the form of the regularizer was constrained by the need for computationally efficient optimization of the joint objective (Fan et al., 2008). The advent of machine learning frameworks with reverse-mode automatic differentiation support ((Abadi et al., 2015), (Paszke et al., 2019)) automates the process of calculating gradients of arbitrary differentiable functions, enabling efficient optimization of the regularization term in the objective in Eq. 2.13 via gradient-based methods (Arjovsky and Bottou, 2017) for a wide variety of functional forms of $\Omega(h)$. With increased liberty on the functional form of the regularization term, the role of the regularizer has grown beyond a means to penalize hypothesis complexity to a way of enforcing 'desirable structure' on the hypothesis. Examples of this include:

- Enforcing Lipschitz continuity of the hypothesis function - intuitively the output $h(x)$ should not change significantly when x changes only slightly. This can be achieved by penalizing the Frobenius norm of the Jacobian matrix associated with the hypothesis function:

$$\Omega(h) = \left\| \frac{\partial h(x)}{\partial x} \right\|_F^2. \tag{2.15}$$

- The information bottleneck principle in representation learning (Tishby et al., 1999). Here, we would like to learn some compressed representation of the data $h = h(x)$ that is informative of y but with non-essential information about x removed. This is achieved by minimizing the mutual information (Section 2.1.1) between x and h, $I(x; h)$ but maximizing the mutual information between y and h, $I(y; h)$. Here $I(x; h)$ can be interpreted as the risk term and $I(y; h)$ can be interpreted as the regularization term in the information bottleneck objective:

$$\min_h I(x; h) - \lambda I(y; h). \tag{2.16}$$

Appropriate regularization of the objective function to enforce desirable properties of the hypothesis returned by the learner forms a major component of the proposed approaches in this thesis in this chapter and the next.

Machine Learning in High Energy Physics

As discussed above, the increasing sophistication of machine learning methods and increased availability of high-quality data have seen rapid adoption in many scientific fields, among them high-energy physics (HEP) (Apostolakis et al., 2018). Machine learning has found useful application in many stages along the analysis pipeline, from approximate event simulation to theoretical parameter inference in models of new physics. However, as with any scientific field, there are characteristic challenges in the application of machine learning to high-energy physics (Keck, 2017).

- Quantum field theory is inherently stochastic - meaning that knowledge of the absolute ground truth properties for any event cannot be obtained. This leads to reliance on simulation of physical processes to approximate the ground truth.
- Physical processes of interest when testing models of new physics are typically either extremely rare or conjectured to exist but hitherto unobserved. This makes their experimental signatures hard to isolate from the overwhelmingly dominant set of background events - necessitating the analysis of large amounts of data to draw statistically significant conclusions.
- Statistical and systematic uncertainties and their relationship to supervised learning models must be treated carefully to ensure the veracity of the final measurement.

Each of these points must be considered when applying machine learning to experimental analyses. Behnke et al. (2013) provide a comprehensive overview of the idiosyncrasies of applying machine learning to high-energy physics. In this thesis we restrict our attention to the field of event classification in HEP. Event classification in HEP is concerned with separation of signal from non-interesting background events. Here, a model is trained on simulated data to distinguish signal from background events based on variables which are descriptive of the underlying physics of the signal process. The trained model is then expected to generalize to the domain of real data. Machine learning algorithms capable of reliable signal discrimination in collider environments have installed themselves as an integral part of modern high-energy physics. However, dangerous pitfalls surround the naive application of machine learning models to supervised learning in HEP, as discussed above. The approach outlined in this chapter of the thesis treats the problem of building a classifier that makes decisions invariant of certain observables as a machine learning problem, in the sense that the desired outcome can be achieved by autonomous optimization of some objective function on data without having to encode additional logic in the analysis procedure by hand.

Notation

In what follows, we focus on a supervised binary classification task for simplicity. $x \in \mathbb{R}^D$ denote the input vector to the supervised learner, let $y \in \{0, 1\}$ denote the associated ground truth labels. Depending on context, x and y may be overloaded to denote a pair of arbitrary random variables. Let the data $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$ denote the training dataset consisting of N independent samples from the distribution $p(x, y)$. The objective of the learning algorithm is to infer the random variable y from the instance x. Let the output of the learner be denoted by the encoding or representation $E_\theta : X \to \mathbb{R}^{D'}$, with parameters θ optimized to minimize some classification loss. We may interchangeably refer to this as the classifier response. Let z denote the target random variable which should remain invariant with respect to the encoding, with the total dataset $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ generated by the joint $p(x, y, z)$. z can take either continuous or categorical values. For simplicity, take $z \in \mathbb{R}$ and $E_\theta(x) \in \mathbb{R}$. All results straightforwardly generalize to the multidimensional case. Louppe et al. (2017) propose that invariance of the distribution of the encoding E_θ with respect to the target variable z requires:

$$p\left(E_\theta(x) = c, z = z'\right) = p\left(E_\theta(x) = c\right) p\left(z = z'\right), \tag{2.17}$$

for all $z' \in Z$ and all possible encoding outputs c, i.e. $E_\theta(x)$ and z are independent random variables[1].

Intuitively, information about the encoding should not reveal information about z and vice-versa. In practice, if the target variable z is strongly correlated with the truth labels y, the independence condition may be too restrictive; such an encoding E_θ may not exist. In this case, the learning procedure must reach a compromise between discrimination power and statistical invariance. Lastly, we assume $c = 1$ throughout this chapter.

Mutual Information

The mutual information $I(x; y)$ between the random variables x and y is a symmetric measure of the information shared between x and y. The mutual information can be expressed as the expected value of the log-ratio between the joint distribution and the product of the marginals or the difference between the marginal and conditional entropies:

$$I(x; y) \triangleq \mathbf{E}_{(x,y)\sim p(x,y)} \left[\log \frac{p(x, y)}{p(x)p(y)}\right] \tag{2.18}$$

$$= H(x) - H(x|y) \tag{2.19}$$

$$= H(y) - H(y|x). \tag{2.20}$$

The second form invites the intuitive interpretation of mutual information as the reduction in uncertainty in either variable due to knowledge of the other. By Jensen's Inequality, $I(x; y) \geq 0$ with equality if and only if x and y are independent random variables. In this case, information about one variable provides no information about the other. Note that if x is an invertible, deterministic function of y, knowledge of either variable fully determines the other - hence $H(x|y) = H(y|x) = 0$ and the mutual information reduces to the entropy $H(x) = H(y)$ - the inherent uncertainty in either variable.

[1] As the initialization/optimization procedure may be stochastic, we regard $E_\theta(x)$ as a random variable in its own right.

Invariant Representation Learning

The majority of work in the field of invariant representation learning is based on minimizing some measure of informativeness between the learnt representation and the sensitive or protected attribute the representation should be invariant to. A ubiquitous theme is to minimize the ability of an auxiliary network r to predict the value of the sensitive attribute given the current representation as input. The auxiliary network is trained jointly together with a 'parent' model f optimized for two purposes; firstly performance on the primary task, typically classification or regression of some target variable, and to secondly minimize the performance of the auxiliary network by reducing the information content in the representation about the sensitive attributes. Below the sensitive attributes are denoted by z, the observed input data is denoted by x and the intermediate representation h is taken to be the unnormalized output of the parent network $h = f(x)$. For simplicity, assume the parent model is tasked with classification of target values y, and that f and r are represented by probabilistic classifiers whose outputs $q_f(y|x)$ and $q_r(z|f(x))$ are estimates of the probability that instance x belongs to class y and the probability of observing sensitive value z given intermediate representation value $f(x)$, respectively. Then the parent and auxiliary network are trained jointly under the following prototypical objective:

$$\min_{f,r} \mathcal{L}, \quad \mathcal{L} = \mathcal{L}_f - \lambda \mathcal{L}_r \tag{2.21}$$

$$\mathcal{L}_f = \mathbf{E}_{p(x)} \left[\mathbf{E}_{p(y|x)} \left[-\log q_f(y|x) \right] \right] \tag{2.22}$$

$$\mathcal{L}_r = \mathbf{E}_{p(x)} \left[\mathbf{E}_{p(z|x)} \left[-\log q_r(z|f(x)) \right] \right]. \tag{2.23}$$

Where the Lagrange multiplier $\lambda > 0$ is a hyperparameter that controls the relative importance of invariance versus performance on the parent task. As optimization of the joint loss is a zero-sum game when $\lambda = 1$, and the networks f and r have competing objectives, the auxiliary network r is commonly referred to as the *adversary*. If the sensitive attribute is a discrete quantity then the auxiliary discriminator loss $\mathcal{L}_r$ takes the form of the cross-entropy. If the sensitive attribute is continuous one may discretize the distribution or attempt to model the density $q_r(z|f(x))$ through a mixture density network. This method does not scale well to the case of multiple sensitive attributes $(z_1, \dots z_m)$ as a separate adversary r_i must be introduced to model the attribute z_i, such that the joint loss becomes

$$\min_{f,r_1,\dots r_m} \mathcal{L}, \quad \mathcal{L} = \mathcal{L}_f - \sum_{i=1}^{m} \lambda_i \mathcal{L}_{r_i}, \tag{2.24}$$

where λ_i is the weighting associated with adversary r_i. Note that an inner optimization loop is required to optimize the parameters each network r_i. There are several disadvantages to this general scheme. Firstly, gradient-based adversarial training procedures are vulnerable to saturation of gradients and instability during optimization (Arjovsky and Bottou, 2017), and may require the introduction of heuristic measures to stabilize the nested optimization loop. Secondly, the performance of adversarial training methods is highly sensitive to the choice of hyperparameter configuration (Lucic et al., 2017). Thirdly, the additional hyperparameters introduced by the auxiliary discriminator results in an combinatorial increase in the size of the joint hyperparameter space, compounding the second difficulty.

This pattern has been applied to learning representations for algorithmic fairness ((Zemel et al., 2013), (Edwards and Storkey, 2015), (Song et al., 2019), (Zhao et al., 2019)), domain adaption ((Ganin and Lempitsky, 2015), (Tzeng et al., 2017), (Saito et al., 2018)) and experimental analyses in particle physics ((Louppe et al., 2017), (Shimmin et al., 2017)). Another notable approach to invariant representation learning involves enforcing an information-bottleneck between the intermediate representation and the protected values by penalizing an estimate of the mutual information between the representation and the protected values ((Moyer et al., 2018), (Song et al., 2019)), and the approach proposed in Chapter 2.2, each of which minimizes different bounds on the mutual information. Unlike the adversarial approach, it is easier to enforce a multidimensional correlation between the intermediate representation and multiple sensitive data attributes s_i under this method as it does not require the introduction of a distinct adversary for each s_i.

2.2 Mutual Information Regularization

2.2.1 Enforcing Invariance in Learning Algorithms

One way of promoting invariance of the target z with respect to the representation of the data x, $E_\theta(x)$ is to bound the information content of z stored in the encoding $E_\theta(x)$, so that observation of the encoder output does not significantly

reduce uncertainty in the z spectrum. To achieve this, we minimize the mutual information between z and $E_\theta(x)$. This leads to the following objective function, to be minimized with respect to θ:

$$\mathcal{L}_\theta(\mathbf{z}, \mathcal{D}) = \mathbf{E}_{(x,y)\sim p(x,y)} \left[\log q_\theta(y|x)\right] + \lambda I\left(E_\theta; z\right) \tag{2.25}$$

Here $\mathcal{D} = (\mathbf{x}, \mathbf{y})$ denotes the training dataset and $q_\theta(y|x)$ is the softmax of the encoding $E_\theta(x)$, which gives the probability of correct classification of data x as label y, calibration issues aside. This can be seen as the standard cross-entropy loss plus a regularization term promoting the desired invariance. Here $\lambda \geq 0$ is a Lagrange multiplier controlling the strength of the regularization. λ can be thought of as a hyperparameter determining the tradeoff between discrimination power and invariance. Note that instead of the cross-entropy loss we can use any other generic loss term that optimizes for some downstream task.

2.2.2 A divergence minimization perspective

Unfortunately, the expectation (2.18) is intractable for continuous non-Gaussian random variables, making the objective (2.25) impractical to optimize directly for non-discrete target variables z. To find a tractable objective, we note that the mutual information (2.18) can be written as the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the joint distribution $p(x, y)$ and the product of the marginal distributions $p(x)p(y)$:

$$I(x; y) = D_{\mathrm{KL}}\left[p(x, y) \,||\, p(x)p(y)\right]. \tag{2.26}$$

The KL divergence $D_{KL}(p\|q)$, can be thought of as a measure of the 'distance' between probability distributions p and q. It is strictly non-negative but is not symmetric with respect to the inputs p, q, i.e. $D_{\mathrm{KL}}(p|q) \neq D_{\mathrm{KL}}(q\|p)$ in general. Higher values of the KL divergence indicate a greater degree of similarity between the input distributions. The KL divergence belongs to the more general class of f-divergences (Ali and Silvey, 1966), a collection of disimilarity measures between two given probability distributions that take the form of an expectation of the density ratio. While the KL-divergence admits a dual representation (Donsker and Varadhan, 1975) that permits tractable estimation through a variational formulation, we are only concerned with driving the log-ratio between the joint distribution and product of the marginals to zero. Or equivalently, the ratio to unity. Hence we are free to minimize any f-divergence between the joint and the product of the marginals. Specifically, we are interested in divergence functionals of the form

$$d_f\left[p(x, y) \,||\, p(x)p(y)\right] = \int dx\, dy\; p(x)p(y) f\left(\frac{p(x, y)}{p(x)p(y)}\right), \tag{2.27}$$

Where the generator function $f : \mathbb{R}^+ \to \mathbb{R}$ is a convex, lower-semicontinuous function satisfying $f(1) = 0$. Different choices of f define different types of statistical divergences. For example, choosing $f(u) = u \log u$ recovers the KL divergence.

While the computation of f-divergences between two arbitrary continuous distributions is intractable in general, Nguyen et al. (2008) derive a variational representation of the f-divergence that can be used to obtain a lower bound on the divergence. This is extended by Nowozin et al. (2016) to parametric variational divergence minimization. The scheme hinges on the following variational lower bound on the f-divergence between absolutely continuous density functions p and q:

$$d_f(p \,||\, q) \geq \sup_{T \in \mathcal{T}} \left(\mathbf{E}_{p(x)}\left[T(x)\right] - \mathbf{E}_{q(x)}\left[f^\star \circ T(x)\right]\right). \tag{2.28}$$

Where $f^\star$ denotes the convex conjugate of f, and $\mathcal{T}$ is some arbitrary class of functions $T : \mathbb{R}^d \to \mathrm{dom}_{f^\star}$ which parameterize the estimation. We find the variational representation of the Jensen-Shannon divergence, a symmetrized and bounded form of the KL-divergence, to be more stable under optimization than the corresponding KL-based objective and hence optimize Equation 2.28 under the generator function corresponding to the Jensen-Shannon divergence (Nowozin et al., 2016). In what follows, let $T_\varphi : E \times Z \to \mathbb{R}$ denote the variational function. E_θ and T_φ are represented by neural networks with parameters θ and φ respectively, although in principle either can be represented with any sufficiently expressive differentiable function. Then the variational lower bound on the Jensen-Shannon divergence between the joint distribution $p(x, z)$ and product of marginals $p(x)p(z)$ is found to be (with some abuse of notation):

$$\begin{aligned}\hat{I}\left(\theta, \varphi\right) = \;& \mathbf{E}_{(x,z)\sim p(x,z)}\left[\log \sigma\left(T_\varphi\left(E_\theta(x), z\right)\right)\right] \\ & + \mathbf{E}_{x\sim p(x), z\sim p(z)}\left[\log\left(1 - \sigma\left(T_\varphi\left(E_\theta(x), z\right)\right)\right)\right] \\ & \leq I\left(E_\theta(x); z\right).\end{aligned} \tag{2.29}$$

Using this estimate of D_{JS} as a proxy for the idealized mutual information term in Equation 2.25, the regularization procedure requires us to find a saddle point of the following minimax objective function, which is minimized with respect to θ and maximized with respect to φ (Algorithm 1).

$$\mathcal{L}_{\theta,\varphi}(\mathbf{z}, \mathcal{D}) = \mathbf{E}_{(x,y)\sim p(x,y)} \left[\log q_\theta(y|x)\right] + \lambda \hat{I}\left(\theta, \varphi\right) \tag{2.30}$$

$$(\hat{\theta}, \hat{\varphi}) = \underset{\theta}{\operatorname{argmin}} \left(\mathbf{E}_{p(x,y)} \left[\log q_\theta(y|x)\right] + \lambda \underset{\varphi}{\operatorname{argmax}}\, \hat{I}\left(\theta, \varphi\right) \right). \tag{2.31}$$

2.2.3 A density ratio estimation perspective

The above objective (2.30) is slightly unsatisfying from a theoretical perspective as the 'outer loop' of the optimization procedure involves minimization of a *lower bound* on the mutual information $\hat{I}(\theta, \varphi)$ in θ. This term can be decreased by legitimate decrease in the mutual information between the representation and the protected variable or an artificial deterioration of the bound.

In what follows we abuse notation slightly such that:

$$(x, z) \sim p_\theta(x, z) \Leftrightarrow (x', z) \sim p(x', z), \quad x \sim E_\theta(x'), \tag{2.32}$$

and analogously for the marginal $p_\theta(x)$. To build an alternative optimization step for the encoder E_θ, we can view the objective through the lens of sample-based density ratio estimation, where a discriminator is built to approximate the ratio $\frac{p(x,y)}{p(x)p(y)}$ in the f-divergence in Equation 2.27. The sigmoid term in Equation 2.29 can be identified as the output of a discriminator D_φ with parameters φ trained to distinguish between samples from the joint $p_\theta(x, z)$ and factored marginals $p_\theta(x)p(z)$. That is, if we label samples $(x, z) \sim p_\theta(x, z)$ as $y = 1$ and samples $(x, z) \sim p_\theta(x)p(z)$ as $y = 0$, then $D_\varphi(x, z) = P(y = 1|x, z)$. Then the mutual information can be approximated as:

$$\begin{aligned} I\left(E_\theta(x); z\right) &= \mathbf{E}_{p_\theta(x,z)} \left[\log \frac{p_\theta(x, z)}{p_\theta(x)p(z)}\right] \\ &= \mathbf{E}_{p_\theta(x,z)} \left[\log \frac{D^*(x, z)}{1 - D^*(x, z)}\right] \\ &\approx \mathbf{E}_{p_\theta(x,z)} \left[\log \frac{D_\varphi(x, z)}{1 - D_\varphi(x, z)}\right]. \end{aligned} \tag{2.33}$$

Where in the second line we have used the fact that the Bayes-optimal discriminator D^* between densities p and q is given by[2] $D^*(\mathbf{x}) = p(\mathbf{x})/\left(p(\mathbf{x}) + q(\mathbf{x})\right)$ (Sønderby et al., 2016). We approximate the discriminator D_φ after a fixed number of gradient steps as being close to the optimal discriminator. Note there is no guarantee that the discriminator D_φ will be optimal at any stage of the optimization process, and there will in general be some approximation error in Eq. 2.33 that is difficult to quantify. In practice we find that taking a small number of gradient steps, around 4-8, for the discriminator for each gradient step of the encoder E_θ works well. To learn the parameters φ, we optimize the discriminator D_φ with respect to a proper scoring rule for binary classification (Gneiting and Raftery, 2007). Choosing the Bernoulli score, which we minimize over, corresponds to the well-known cross-entropy loss and recovers the discriminator objective in Equation 2.29.

As a side note, using the fact that the optimal discriminator minimizes the Bernoulli loss connects back to the divergence minimization framework. The Jensen-Shannon divergence between densities $p(x), q(x)$ is defined as:

$$D_{\text{JS}}\left[p(x)||q(x)\right] \triangleq \frac{1}{2}\left(D_{\text{KL}}\left[p(x)||m(x)\right] + D_{\text{KL}}\left[q(x)||m(x)\right]\right) \tag{2.34}$$

Where the mixture density is defined as $m(x) = \left(p(x) + q(x)\right)/2$. Let $m_\theta(x, z)$ denote the corresponding mixture between the joint and the marginals, then minimizing the Bernoulli loss is equivalent to approximation of the Jensen-Shannon divergence, up to a linear transformation:

[2] And further assumed equal prior probabilities for each class.

$$
\begin{aligned}
\sup_{\varphi} \hat{I}\left(\theta, \varphi\right) &= \mathbf{E}_{p_\theta(x,z)}\left[\log \frac{p_\theta(x,z)}{p_\theta(x,z)+p_\theta(x)p(z)}\right] \\
&+ \mathbf{E}_{p_\theta(x)p(z)}\left[\log \frac{p_\theta(x)p(z)}{p_\theta(x,z)+p_\theta(x)p(z)}\right] \\
&= D_{\mathrm{KL}}\left[p_\theta(x,z)||m_\theta(x,z)\right] \\
&+ D_{\mathrm{KL}}\left[p_\theta(x)p(z)||m_\theta(x,z)\right] - \log 4 \\
&= 2\, D_{\mathrm{JS}}\left[p_\theta(x,z)||p_\theta(x)p(z)\right] - \log 4. \qquad (2.35)
\end{aligned}
$$

Returning to the main thread of this section, the update schemes for the encoder E_θ and discriminator D_φ networks[3] may now employ distinct objective functions, which allows us to circumvent the saturating gradient issues empirically observed with minimax optimization procedures on deep networks. The discriminator is trained using the Bernoulli loss to distinguish samples from the joint $p_\theta(x,z)$ and marginals $p_\theta(x)p(z)$ in order to estimate the density ratio. This ratio is then used in the approximation of the mutual information (Equation 2.33), which we substitute into Equation 2.25. The procedure repeats in an alternating fashion (Algorithm 2). We perform a comprehensive empirical study on the proposed methods in the following sections.

Algorithm 1 Encoder Training with Mutual Information-based Regularization

The encoder E_θ and variational function T_ϕ are represented by deep neural networks with parameters θ and ϕ respectively, although in principle any differentiable parametric function with sufficient capacity may be used. The expectations present in the respective objective functions Eq. 2.29, Eq. 2.30 are approximated using Monte Carlo samples. The inner and outer learning rate schedule is held constant at $\eta_t = \eta'_t = 10^{-4}$. We find increasing the number of inner loop iterations for each outer step K tightens the lower bound and tends to increase the effectiveness of the invariance constraint at the expense of computation time. The empirical marginal distribution may be approximated by shuffling along the batch dimension for pairs of variables drawn from the joint distribution.

Require: Regularization coefficient $\lambda \geq 0$, inner learning rate schedule η_t, outer learning rate schedule η'_t.

1: Initialize the parameters of the encoder network E_θ and variational function T_ϕ.
2: **for** $n = 1$ to N **do**
3: **for** $k = 1$ to K **do**
4: Sample $\{(x_1, z_1), \ldots, (x_B, z_B)\} \sim \mathbb{P}_{XZ}$ from the joint distribution.
5: Sample $\{\tilde{z}_1, \ldots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.
6: Update T_ϕ by ascending the lower bound:

$$
\hat{I}(\theta, \phi) = \frac{1}{B}\sum_{i=1}^{B}\left[\log \sigma\left(T_\phi(E_\theta(x_i), z_i)\right) - \log\left(1 - \sigma\left(T_\phi(E_\theta(x_i), \tilde{z}_i)\right)\right)\right]
$$

$$
\phi \leftarrow \phi + \eta_t \nabla_\phi \hat{I}(\theta, \phi)
$$

7: **end for**
8: Sample $\{(x_1, y_1, z_1), \ldots, (x_B, y_B, z_B)\} \sim \mathbb{P}_{XYZ}$ from the joint distribution.
9: Sample $\{\tilde{z}_1, \ldots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.
10: Update E_θ by descending the objective:

$$
\mathcal{L}(\theta, \phi) = \frac{1}{B}\sum_{i=1}^{B}\left[-\log p_\theta\left(y_i|x_i\right) + \lambda \hat{I}\left(\theta, \phi\right)\right]
$$

$$
\theta \leftarrow \theta - \eta'_t \nabla_\theta \mathcal{L}(\theta, \phi)
$$

11: **end for**

2.2.4 Equivalence to Adversarial Training

Proposition 1 (*Minimizing the mutual information between the encoding and protected variable is formally equivalent to adversarial training*)**.** Let $h = f(x)$ denote the learnt encoding of the data x. Then minimization of the mutual

[3] We subsume the statistic T_φ network into the discriminator, $D_\varphi = \sigma \circ T_\varphi$.

Algorithm 2 Encoder Training with Mutual Information-based Regularization (alternative formulation).

Require: Regularization coefficient $\lambda \geq 0$, inner learning rate schedule η_t, outer learning rate schedule η'_t.

Initialize the parameters of the encoder network E_θ and discriminator $D_\varphi = \sigma \circ T_\phi$.
for $n = 1$ to N **do**
  for $k = 1$ to K **do**
    Sample $\{(x_1, z_1), \ldots, (x_B, z_B)\} \sim \mathbb{P}_{XZ}$ from the joint distribution.
    Sample $\{\tilde{z}_1, \ldots, \tilde{z}_B\} \sim \mathbb{P}_Z$ from the marginal distribution.
    Update D_φ by minimizing the cross-entropy:

$$\varphi \leftarrow \varphi - \eta_t \nabla_\varphi \left\{ -\frac{1}{B} \sum_{i=1}^{B} \left[\log D_\varphi \left(x_i, z_i \right) + \log \left(1 - D_\varphi \left(x_i, \tilde{z}_i \right) \right) \right] \right\}$$

  end for
  Sample $\{(x_1, y_1, z_1), \ldots, (x_B, y_B, z_B)\} \sim \mathbb{P}_{XYZ}$ from the joint distribution.
  Update E_θ by descending the objective:

$$\mathcal{L}(\theta, \varphi) = \frac{1}{B} \sum_{i=1}^{B} \left[-\log p_\theta \left(y_i | x_i \right) + \lambda \log \frac{D_\varphi \left(x_i, z_i \right)}{1 - D_\varphi \left(x_i, z_i \right)} \right]$$
$$\theta \leftarrow \theta - \eta'_t \nabla_\theta \mathcal{L}(\theta, \varphi)$$

end for

information $I(z; h)$ between the encoding h and the protected variable z is equivalent to maximization of the loss of an adversary trained to predict the value of z from h, as detailed in the adversarial training procedure (Eq. 2.21) proposed by Louppe et al. (2017). Furthermore, minimization of the loss of an adversary predicting the value z from h is equivalent to maximization of a lower bound on the mutual information $I(z; h)$.

The proof of this proposition can be found in Appendix A.1. To the best of our knowledge, this is the first such result connecting mutual information optimization to adversarial training in the context of learning invariant representations. While minimizing the mutual information between the representation is formally equivalent to the adversarial training procedure (2.21), they may not be practically equivalent. This is due to several reasons. Firstly, approximating the mutual information between continuous quantities is intractable in general, necessitating the use of variational bounds $\hat{I}(z; h)$ parameterized by neural networks (Poole et al., 2019), which offer notoriously poor guarantees on the approximation error $|I(z; h) - \hat{I}(z; h)|$. Secondly, the tightness of the bounds in Appendix A.1 are poorly understood in general. Finally, the final performance of either approach will be strongly dependent on the optimization procedure, which will be different across the two approaches.

2.3 Experiments

2.3.1 Bivariate Gaussian Classification

The proposed method is first applied to a toy example involving binary classification of bivariate data similar to the toy example proposed in Louppe et al. (2017).

In this first illustrative example, data is sampled from bivariate normal distributions as follows:

$$\begin{aligned} x_0 &\sim \mathcal{N}\left(\begin{pmatrix} z \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right), \\ x_1 &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right). \end{aligned} \tag{2.36}$$

Where $z \sim \mathcal{N}(0, 1)$ simulates uncertainty in the x-coordinate of samples drawn from the first distribution. The data is then passed through a shallow neural network consisting of two hidden layers with 32 units and ReLU activation, which we refer to as the classification network. The objective of this network is to distinguish between samples drawn from each distribution in Equation 2.36. The network is trained using the standard binary cross entropy loss and the resulting

(interpolated) decision surface for ~ 25000 examples drawn in equal proportion is shown in the left panel of Figure 2.2. The strong variation of the network output along the y-axis reflect that samples from the distributions can be most easily distinguished through their respective y-coordinates. However, the curvature of the contours away from $x = 0$ indicates that the network has learned to exploit the fact that samples with an x-coordinate sufficiently removed from zero are more likely to originate from the first distribution.

To reduce the dependency of the network on the potentially noisy x-variable, we penalize an estimate of the mutual information between the logits of the neural network and the x-coordinate of the input data. We optimize a lower bound on the desired log-ratio as described in Algorithm 1 through a shallow neural network of similar capacity to the classification network, which remains unchanged. The degree to which the output encoding of the classification network should be independent of the x-coordinate can be controlled through the hyperparameter λ in the objective function 2.30, we use $\lambda = 10$ for this example. The corresponding decision surface is shown in the right panel of Figure 2.2. We observe that the contours have been 'flattened' to be largely parallel to the x-axis, demonstrating a high degree of invariance to the x-coordinate.

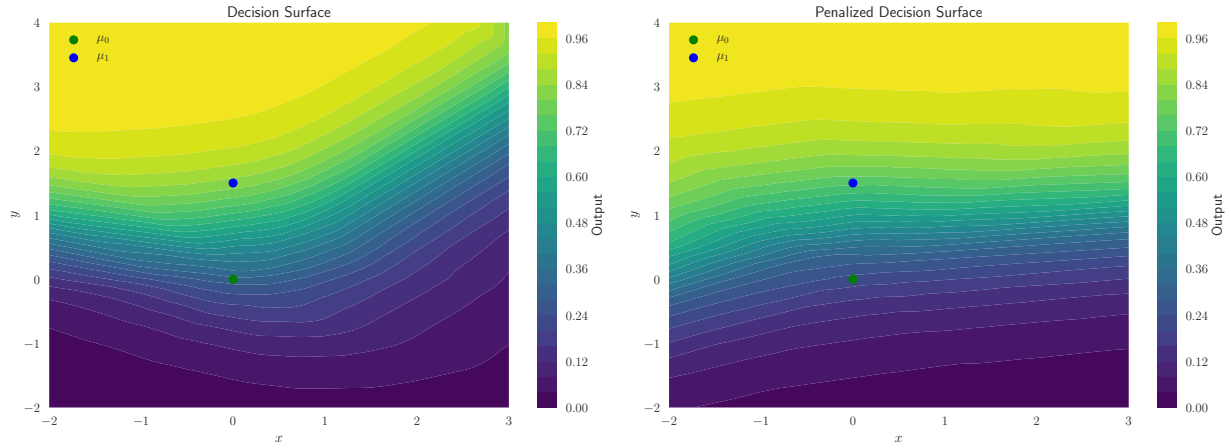


Figure 2.2: Decision surface in Cartesian space for a binary discriminator between two multivariate Gaussians (2.36) with respective means μ_0, μ_1 indicated on the plot. The colors indicate the degree of confidence $0 \geq p \geq 1$ of the discriminator that a sample with the given coordinates belongs to the positive class, arbitrarily chosen to be the second distribution in Eq. 2.36. **Left:** Decision surface resulting from training using the standard binary cross-entropy objective. Curvature of contours away from $x = 0$ demonstrates the network strong dependency on the x-coordinate within this region. **Right:** Decision surface after introduction of a penalty on a lower bound of the mutual information between the x-variable and the network output. Straightening of the contours parallel to the x-axis indicates reduced dependency on the x-variable.

2.3.2 Application to $b \to s\gamma$

We next consider an representative example from high-energy physics and demonstrate that the proposed method is able to overcome common difficulties associated with the use of machine learning in modern experimental analyses.

We train a supervised learning algorithm to distinguish true signal $b \to s\gamma$ events from background processes misidentified in reconstruction. The data X consists of kinematic quantities and variables describing event topology, the full variable list is shown in Appendix A.2. All variables with a Pearson correlation between the target variable z greater than 0.1 are removed from the input. After training, a threshold on $E_\theta(x)$ is selected as the decision boundary corresponding to rejection of 0.995 of background events (while the ideal fraction of rejected events will be analysis-dependent, this is within the realistic order-of-magnitude of background reduction factor for precision measurements of rare flavor decays).

Baseline

As a simple baseline, we consider removing a subset of variables with the greatest dependency on M_{bc} from the input feature set and investigate the effect on classification performance and the degree of background sculpting. To identify

Variable (x_k)	$I(x_k; z)/H(x_k)$	Description
$\max_{d\in D}(p_z)_d$	0.216	p_z summed over all B meson daughters.
$\max_{d\in D}(p_x)_d$	0.210	p_y summed over all B meson daughters.
$\max_{d\in D}(p_y)_d$	0.208	p_x summed over all B meson daughters.
γ-m^2 recoil	0.0407	Invariant mass squared of system recoiling against photon.
E_{CMS} (γ)	0.0406	Photon CMS energy.
$\cos(\theta_{T_{B-ROE}})$	0.0404	Cosine of the angle between the B and ROE thrust axes.
p_T (γ)	0.0389	Photon transverse momentum.
$\max_{d\in D}(p_T)_d$	0.0388	Maximum transverse momentum of B meson daughters.
$\max_{d\in D} E_d$	0.0367	Maximum energy of B meson daughters.
$\max_{d\in D} p_d$	0.0358	Maximum 3-momentum magnitude of B meson daughters.

Table 2.1: Variables x_k with greatest normalized mutual information $I(x_k; z)/H(x_k)$ with protected variable $z = M_{bc}$. Note that as the photon is typically the most energetic daughter in the $b \to s\gamma$ transition and $m^2 = E^2 - p^2 = 0$ for photons, some variables will be identical up to scaling and detector resolution effects. A more comprehensive definition of these features may be found in Appendix A.2.

these variables, first we estimate the mutual information between each feature $x_k \in X$ and the $z = M_{bc}$ variable over $100,000$ samples. As the mutual information between continuous non-Gaussian variables is intractable to compute, we discretize each distribution into 24 bins and apply the estimator proposed in Vinh et al. (2009).[4] For discrete variables, $H(x|z) \geq 0$, so $0 \leq I(x_k; z) \leq H(x_k)$. Therefore we consider the normalized mutual information $I(x_k; z)/H(x_k)$ to account for differences in the entropy of each variable. The 10 features with the highest normalized mutual information with M_{bc} are listed in Table 2.1.

We note that various kinematic variables associated with the daughter particles of the reconstructed B meson exhibit the greatest normalized mutual information with $M_{bc} = \sqrt{E^2_{\text{beam}} - |p_B|^2}$. This is because the beam-constrained mass can be expressed analytically (through a nonlinear function with one or more unknowns) in terms of these variables. The radiative photon will typically be the daughter particle with the highest energy and momentum, and the momentum of the B meson candidate can be related to the momentum of the radiative photon through conservation of momentum, from which M_{bc} can be determined. Therefore such kinematic quantities feature prominently in Table 2.1. Plots demonstrating the relationship between M_{bc} and the first 6 features from Table 2.1 can be found in Figure 2.3. We remove the top-10, top-20, top-30 and top-40 variables with the greatest normalized mutual information with M_{bc} from the feature set and investigate the effect on classification performance and the degree of background sculpting. The classifier is a simple fully connected neural network with 2 hidden layers with 256 units and ReLU activation, trained using the Adam optimizer (Kingma and Ba, 2015) with learning rate 10^{-4}.

Note that the correlation coefficient should not be relied upon as an absolute measure of dependency as this can only capture the linear relationship between two variables. The same applies for nonlinear measures of dependency such as the mutual information. We observe from Figure 2.5 that even aggressively removing variables related to M_{bc} does not prevent the classifier from using information about M_{bc} to distinguish between signal and background events. There still remains a residual dependency on M_{bc} in the remaining features that the classifier can combine to retrieve information about M_{bc}. Furthermore, from Figure 2.4, classification power deteriorates significantly with removal of these features, as variables with high separation power in this scenario are often also dependent on M_{bc}. Therefore eliminating variables based on their exhibited correlation with M_{bc} or using nonlinear measures of dependency such as the mutual information is not a reliable strategy to prevent the classifier from exploiting information about M_{bc}. Instead, more advanced methods are required.

Changing Regularition Strength λ

We briefly examine the effect of the Lagrange multiplier λ on our proposed approach, note $\lambda = 0$ corresponds to traditional cross-entropy minimization. Figure 2.6 shows the tradeoff inherent in the objective (2.29) - higher values of λ promote invariance of $E_\theta(x)$ with respect to z at the expense of discrimination power. Figures 2.7 and 2.8 demonstrate the effect on the background mass spectrum at 0.99 background rejection - approximately the order of magnitude of background rejection required to recover an acceptable signal fraction in searches for rare decays. At $\lambda = 100$, the background and signal M_{bc} spectra are largely invariant under selection - permitting measurement of the signal and

[4] We find the mutual information tends to increase slightly with the number of bins, but is relatively stable under the number of samples considered.

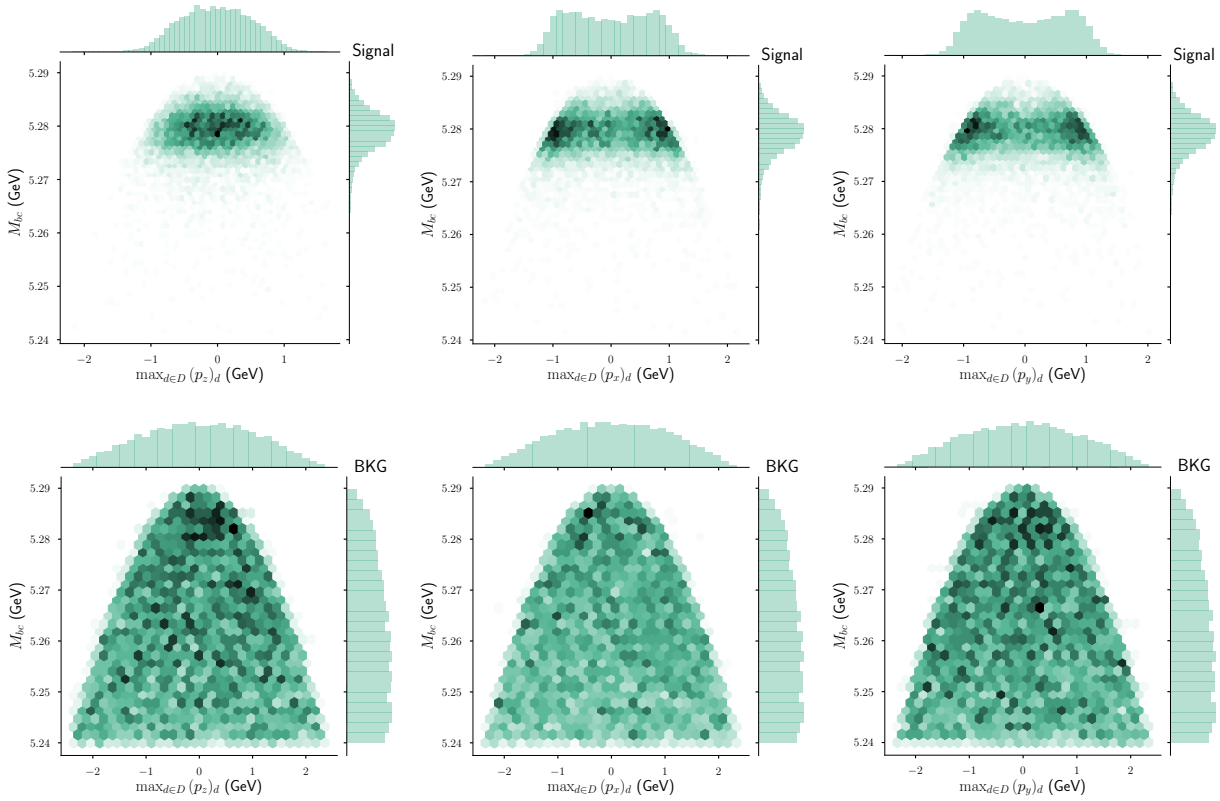


Figure 2.3: (1/2) 2D correlation plots between M_{bc} and features with high M_{bc} dependence. Continued in Figure 2.3

background yields under a maximum likelihood fit. As λ decreases, the background spectrum begins to mimic the signal structure until it is nearly degenerate with the resonant signal peak at $\lambda = 0$, making estimation of the background rate in the signal region difficult. Increasing the regularization strength λ comes with the caveat that the signal efficiency will in general be lower at the same level of background rejection for higher values of λ. One may adjust the λ parameter to more efficiently navigate the Pareto frontier between discrimination power and invariance to M_{bc}.

Model Evaluation

We compare our approach with training a standard feedforward network with no penalty and the adversarial training procedure proposed in (Louppe et al., 2017). To achieve a fair comparison, we fix the classifier architecture to a 4-layer densely connected ReLU network with 512 nodes at each layer. The classifier parameters are optimized using the Adam (Kingma and Ba, 2015) optimization algorithm. For the unpenalized network only the classifier learning rate and the parameter β_1 in Adam are tuned. For both the adversarial and MI-minimizing approach we fix the classifier learning rate at 10^{-4} and tune the ratio of learning rate between the classifier and the auxillary network, the relative penalty strength, relative number of training steps and auxillary network architecture. The search algorithm considers a total of 64 random samples over hyperparameter space, with trial scheduling controlled by the Hyperband early stopping algorithm (Li et al., 2016).

After rejection of a suitable number of background events, an unbinned extended maximum likelihood fit is then performed in the beam-constrained mass M_{bc} to determine the parametric form of the distributions used to model the respective signal and background spectra. The statistical significance obtained from the maximum likelihood fit is used as an evaluation metric to balance the dual considerations of discrimination and invariance. Technical details on the fitting procedure can be found in Appendix A.3.

Hyperparameters are optimized according to the reward signal $\hat{N}_S/\delta N_S$ reported on an independent test set, where $\hat{N}_S$ is the number of signal events returned through an extended maximum likelihood fit to the combined signal and background beam-constrained mass distribution. δN_S denotes the standard error of the maximum likelihood estimator $\hat{N}_S$ obtained through an approximation of the inverse of the expected value of the Hessian. Intuitively, this metric looks

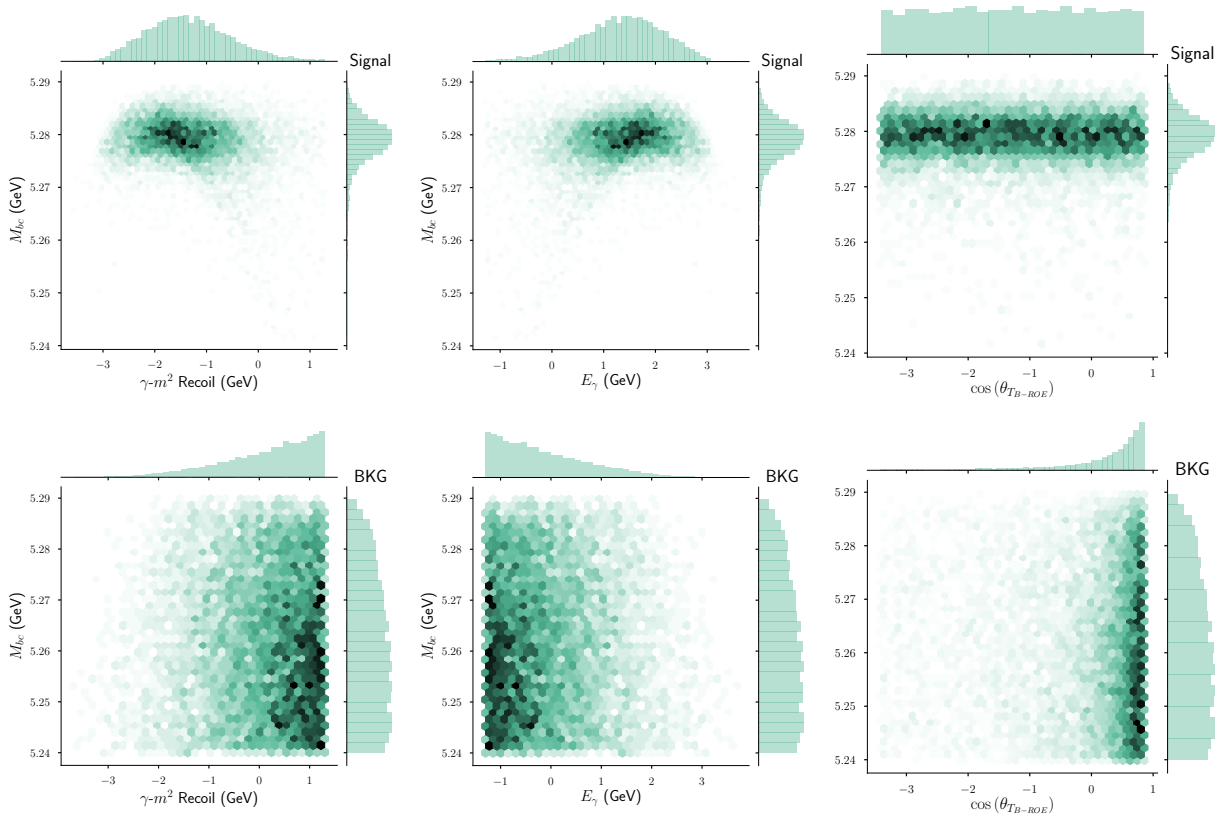


Figure 2.3: (2/2) 2D correlation plots between M_{bc} and features with high M_{bc} dependence. Darker hexes indicate a higher density of events. Note that features except M_{bc} have been normalized to zero mean and unit variance. Marginal distributions of each variable are shown on the top and right hand axes. Even-numbered rows: signal $b \to s\gamma$ events. Odd-numbered rows: background light quark continuum $e^+e^- \to q\bar{q}$ or faking B hadronic decay background. The signal plots exhibit a rich nonlinear substructure that can be exploited by a standard neural network to obtain information about M_{bc}. Note that E_γ is the same as the invariant mass of the system recoiling against the radiative photon up to a square root and detector resolution effects.

to balance discrimination power against distortion of the mass spectrum. Figure 2.9 demonstrates that this quantity demonstrates certain properties that mark it as a reasonable choice of reward signal, although one shortcoming of this reward is that it fails to penalize the inaccuracy of the signal yield reported by the fit, warranting further study into a more principled metric. We note that it is reasonably straightforward to find hyperparameters that yield superior results to the best-performing configuration found through this search procedure with some educated guesswork and time. Further details on the hyperparameter search ranges can be found in Appendix A.4.

To evaluate model sensitivity to hyperparameters, we present the distribution of the maximum (not final) reward on the test set achieved by each hyperparameter run (Figure 2.10). In addition to the approaches introduced in this Section, we also consider the adversarial approach to correlation (Louppe et al., 2017) as well as an normalized bound on the mutual information (Poole et al., 2019). An additional heuristic loss that aims to, maximize $\log \sigma \left(T_\phi(E_\theta(x_i), \tilde{z}_i)\right.$ instead of minimizing $\log \left(1 - \sigma \left(T_\phi(E_\theta(x_i), \tilde{z}_i)\right)\right)$ is also trialled. The motivation originates from the non-saturating loss proposed in (Goodfellow et al., 2016) to empirically avoid gradient saturation during the early stages of training. To investigate model stability, we select the best-performing hyperparameter configuration from each sweep, and rerun training 4 times with different random seeds for 4 epochs. Metrics are recorded periodically and we select the top-performing parameter setting within each run and report generalization performance on the validation set, which is independent from the training and test sets. (Table 2.2).

As a final note, it may be fruitful to consider alternative f-divergences/distances beyond the Kullback-Leiber and Jensen-Shannon variants considered here, especially in higher-dimensional settings. - We ran experiments using the Wasserstein-1 distance (Villani, 2009) in lieu of the Jensen-Shannon divergence but could not obtain useful gradients from the objective corresponding to Eq. 2.29, we speculate that convergence of the Wasserstein distance may be

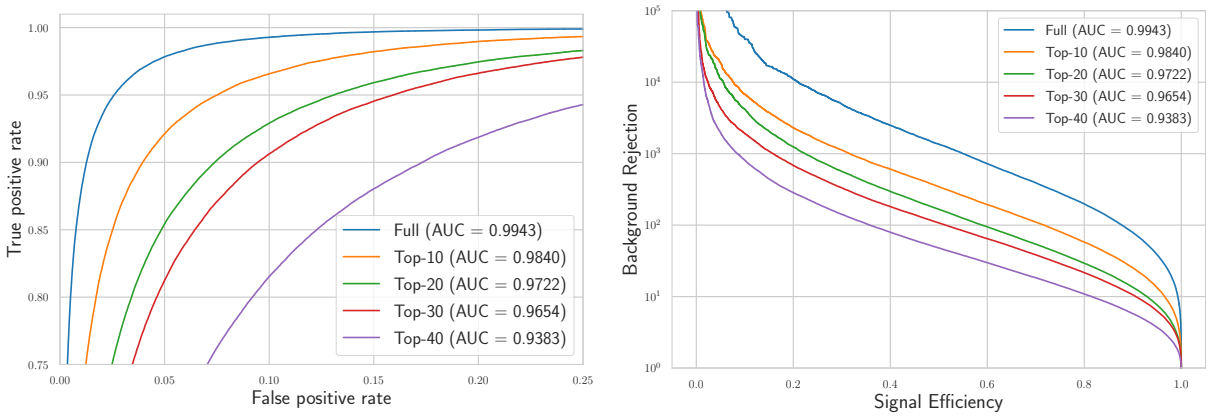


Figure 2.4: Classification performance after baseline feature removal. 'Full' corresponds to providing the full dataset as classifier input while 'Top-X' corresponds to removal of the X features with highest normalized mutual information with M_{bc} from the classifier input. **Left:** Receiver operating characteristic curves for each label. **Right:** Signal efficiency versus background rejection for each label. Removing features related to M_{bc} results in significant deterioration in performance as features with high discrimination power are typically related to M_{bc}.

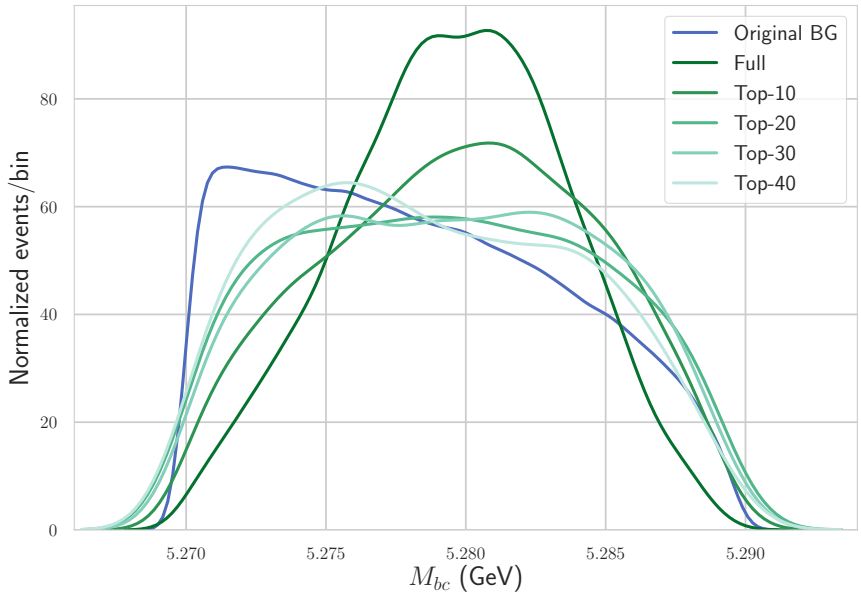


Figure 2.5: Illustration of mass sculpting after baseline feature removal. Shown are the kernel density estimates of the background M_{bc} distribution after selection by a classifier trained on the following input types. 'Full' corresponds to providing the full dataset as classifier input while 'Top-X' corresponds to removal of the X features with highest normalized mutual information with M_{bc} from the classifier input. 'Original-BG' denotes the shape of the distribution before applying any selection. We observe that even aggressively pruning features from the input dataset mitigates but does not prevent a non-uniform selection efficiency in M_{bc}.

too weak in the low-dimensional setting we consider. Similarly, attempting to match the moments between the joint and marginal distributions through the maximum-mean-discrepancy distance (Gretton et al., 2012) did not allow our model to convergence. Note that it is not necessary to regularize the mutual information between the final output and the sensitive variables and one may instead constrain the intermediate representations or 'hidden layers' of the base neural network. For multiple sensitive input variables it is possible to train multiple statistic networks, but a more principled solution may be to constrain a generalization of the mutual information to multiple random variables, the total correlation (Watanabe, 1960), between all variables. However, we leave this avenue to future work.

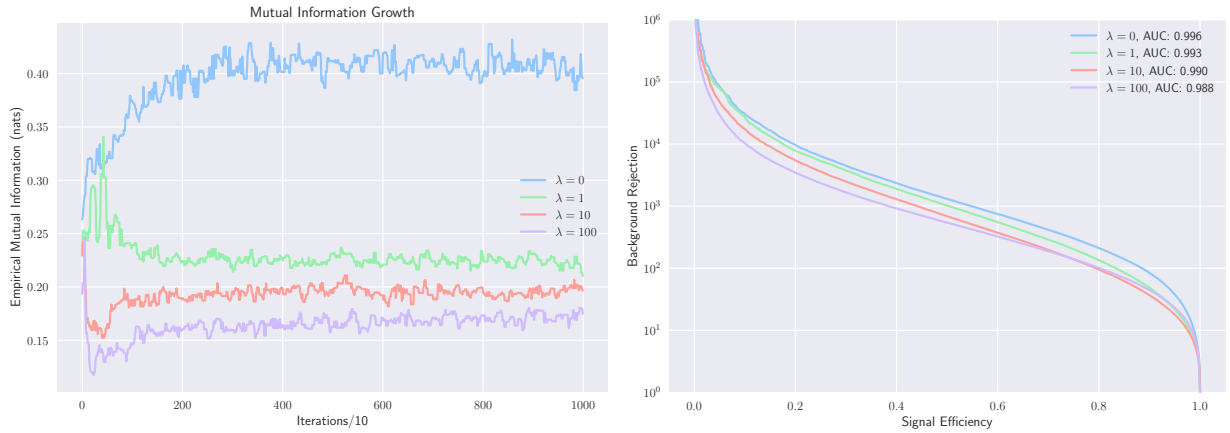


Figure 2.6: Left: Evolution of mutual information between encoding $E_\theta(x)$ and z over training for different values of the hyperparameter λ. As the objective is based on Jensen-Shannon divergence minimization, we use the estimator proposed in (Gao et al., 2014). Right: Signal efficiency versus background rejection level for different thresholds on the encoding $E_\theta(x)$. Increasing λ promotes invariance of the distribution of $E_\theta(x)$ with respect to z at the expense of discrimination power.

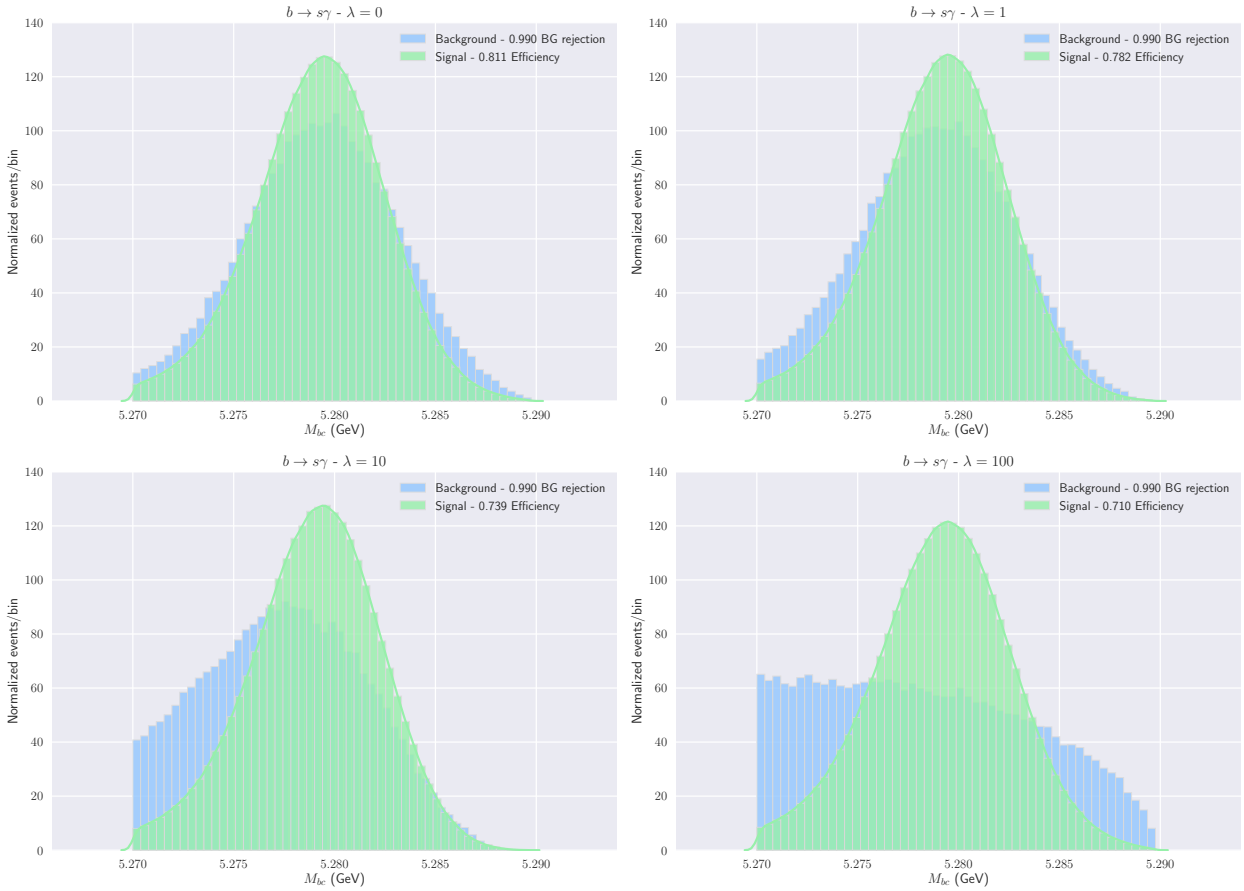


Figure 2.7: Sculpting of background spectrum toward resonant signal structure for different values of λ. The M_{bc} distribution of background events after rejection of 99.5% of background events are shown in blue. The M_{bc} distribution of signal events after rejection of 99.5% of The background spectrum is near degenerate with the signal spectrum at $\lambda = 0$ (top left). For a given background rejection level (fixed throughout at 0.99), increasing invariance comes at the expense of signal efficiency.

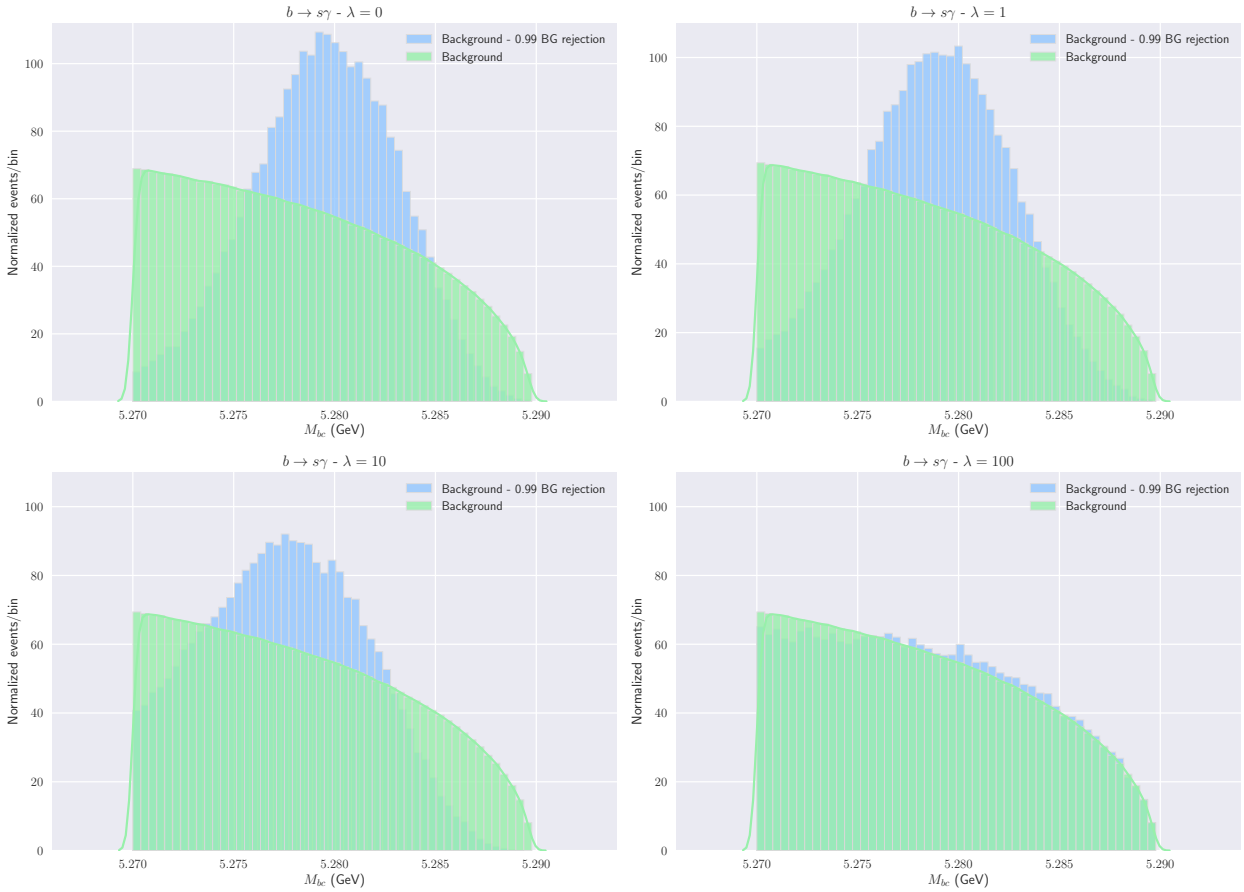


Figure 2.8: Post-selection distortion of background spectrum for different values of λ. This highlights the role of λ as guiding the network toward the desired behaviour of achieving a uniform selection efficiency in the sensitive variable M_{bc} for the background class.

Model	AUC	$N_S/\delta N_S$
MI Penalty	0.970 ± 0.024	38.7 ± 0.4
Adversarial Penalty	0.952 ± 0.020	35.1 ± 0.1
No Penalty	0.995 ± 0.013	29.9 ± 2.0

Table 2.2: We select the top-performing hyperparameter configuration from the search illustrated in Figure 2.10 over all the MI-penalty-based models, which encompass the (minimax, heuristic, MINE and Approx-KL) variants, as well as the Adversarial penalty and baseline (no penalty) approach. Shown is the mean of 4 experiments with different random seeds for the initial network weights.

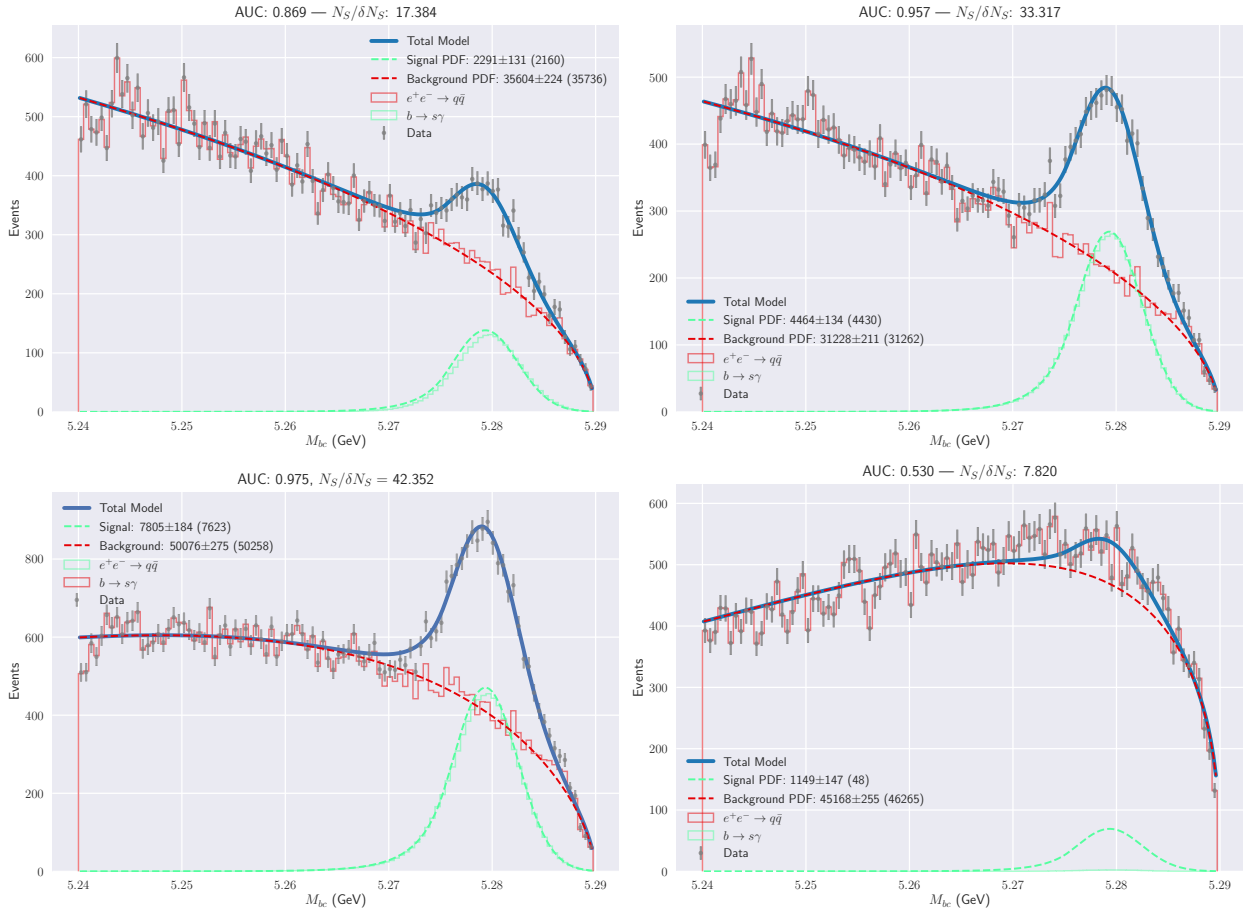


Figure 2.9: Visualization of distributions post-selection for different values of the reward $N_S/\delta N_S$. Signal (continuum) is shown in green (red), with combined data shown in blue. The dashed lines represent the pdf obtained from an unbinned extended maximum likelihood fit, while the solid stepped histograms represent the ground-truth signal/background distributions. The reward tends to increase with increased discrimination power, and is able to identify failure modes (bottom right) with a heavily distorted background spectrum.

2.4 Application to Lepton ID at Belle II

The separation of low-momentum lepton (e, μ) candidates from hadronic background is vital to many parts of the Belle II physics program. For example, tests of lepton flavor universality violation through semi-tauonic B decays such as $B \to D^*\tau\nu$ or measurements of CKM elements via the $b \to u(c)\ell\nu$ transition. This is a nontrivial task in the low momentum region (≤ 0.6 GeV/c). In this region, muons have insufficient energy to reach the KLM and must be detected by the electromagnetic calorimeter (ECL). Furthermore, at low momenta electrons lose significant energy via bremsstrahlung and can be faked by hadronic decays. To discriminate between muon and electron candidates detected by the ECL, information from the ECL measurements and tracking systems is combined to yield a collection of physical observables sensitive to the different physics governing interactions of hadrons and leptons with the ECL crystals. These observables are passed as input to event classification algorithms, which seek to distinguish electrons/muons from hadronic background. It is desirable that the classifier response in this situation be invariant to the magnitude of the 3-momentum p. Input parameters of theoretical models tested using lepton identification at Belle II may be strongly momentum-dependent. Hence introducing a bias into the momentum distributions of detected leptons may increase systematic uncertainties associated with experimental tests of these models.

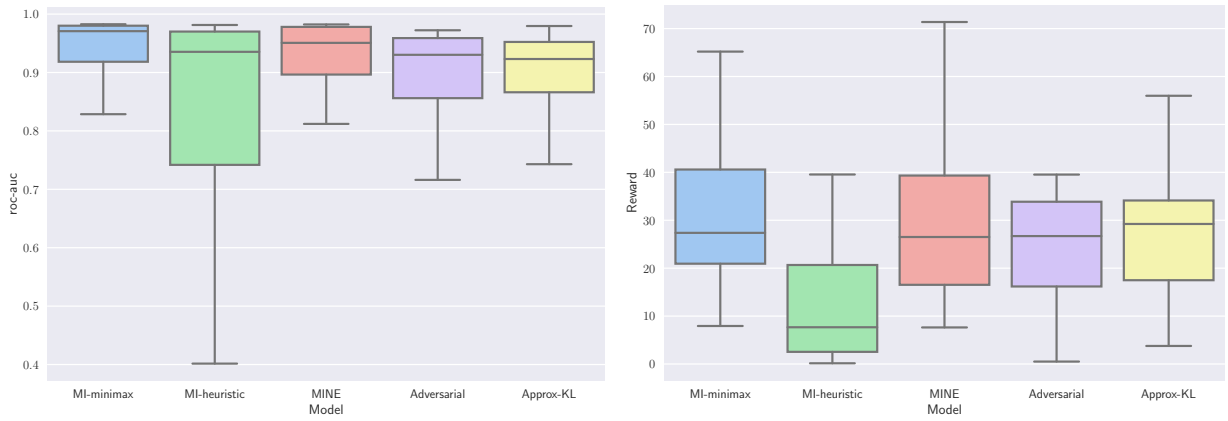


Figure 2.10: Variance of AUC (left) and reward (right) for different models. We take the maximum of each metric to simulate early stopping. 64 samples from hyperparameter space are permitted for each model, with an allotted computation budget of 48 hours on a 24-core machine with 4 P100 Nvidia GPUs. 'Minimax' refers to the approach outlined in Algorithm 1. 'Approx-KL' refers to the approach outlined in Algorithm 2. 'Heuristic' refers to the approach in Algorithm 2, where the denominator in Eq. 2.33 is maximized instead of minimizing the density ratio, conjectured to avoid gradient saturation issues during the early stages of training (Arjovsky and Bottou, 2017). 'Adversarial' refers to the minimax optimization approach defined in (Louppe et al., 2017). 'MINE' substitutes the estimate of the mutual information in Algorithm 1 using an unnormalized variational bound (Poole et al., 2019). We find no statistically significant difference in performance, measured by the reward metric defined above, between the proposed minimax and KL approximation methods, as well as the adversarial approach and MINE approaches.

2.4.1 Comparing Approaches for Particle Identification at Belle II

The current Belle II charged particle identification (PID) scheme is based on a likelihood ratio for each particle hypothesis $H \in \{e, \mu, \pi, K, p\}$. The likelihood $\mathcal{L}_H$ for each particle hypothesis is constructed as a function of measurements from several subdetectors, including:

- Particle energy loss by ionisation (dE/dx) in the silicon vertex detector (SVD) and central drift chamber (CDC).
- Measurements of the mass-dependent optical response of time-of-propagation detectors (TOP) and ring-imaging Cherenkov detectors (ARICH).
- Measurements of the energy deposition in the scintillation crystals of the electromagnetic calorimeter (ECL).
- Measurements in the K_L^0 -and-muon detector (KLM) that account for the different penetration range and scattering of muons and hadrons.

Each subdetector d is associated with a likelihood $\mathcal{L}_H^d = p(x|H)$ for each particle hypothesis H and measurements x. The likelihood is a function of the distributions associated with observables relevant to the detector. The distribution parameters are estimated via maximum likelihood on simulated Monte Carlo data. Under the Naive Bayes' assumption that the measurements are independent, the joint likelihood for hypothesis H reduces to:

$$\mathcal{L}_H = \prod_{d \in D} \mathcal{L}_H^d. \tag{2.37}$$

Where $D = \{\text{SVD, CDC, } \ldots\}$ is the set of all subdetectors. The posterior $p(H|x)$ is recovered from Bayes' Theorem and the assumption that the prior probabilities are identical for all hypotheses: $p(H = h) = p(H = h')$ for all $h, h' \in \{e, \mu, \pi, K, p\}$.

$$p(H|x) = \frac{p(x|H)p(H)}{\sum_h p(x|H = h)p(H = h)} \tag{2.38}$$

$$= \frac{\mathcal{L}_H}{\sum_h \mathcal{L}_h}. \tag{2.39}$$

Lepton identification in the electromagnetic calorimeter (ECL) is based on fitting a univariate probability density function to the distribution of the ratio of E/p - the ratio of the deposited energy measured by the ECL and the reconstructed 3-momentum of the topologically matched track - in distinct momentum/angular/charge bins. This variable provides excellent discrimination between electrons and particles such as muons and charged hadrons. For electrons, particle energy will be absorbed completely by the calorimeter, resulting in maximum energy deposition. Muons are minimum ionizing particles (MIPs), hence only a small part of the total energy is deposited within the ECL and the E/p distribution is longer-tailed due to energy loss from nuclear interactions within the ECL material. Separation of electrons from hadronic background at low momenta based solely on E/p is difficult due to the increased hadronic background and electron energy loss from bremsstrahlung in the detector material. Conversely, distinguishing muons from hadronic background is easier at low momenta as hadrons lose energy more easily than muons from ionization and hadronic interactions with the detector material - the longer the path length, the more likely hadronic interaction is to occur, resulting in in long E/p hadronic tail. However, using E/p solely as a discriminating variable presents challenges when attempting to separate muons from the MIP component of the hadronic shower - resulting in deteriorated signal efficiency in the low-momentum region (Figure 2.11).

Hadronic cascades are comprised primarily of secondary charged and/or neutral pions, so for the remainder of this section we restrict our attention to binary discrimination between ℓ/π events, although we expect the same methods can be applied to discrimination between leptonic and kaon or proton events without modification. Leptonic events $\ell = e,\ \mu$ will be referred to as signal and hadronic events will be referred to as background hereafter. Discrimination of leptons from the hadronic background using the E/p likelihood is efficient in the high-momentum region but sub-optimal at low-momenta ($< 0.6\,\mathrm{GeV}/c$), where the background from misidentified hadronic decays is largest. This especially true for electrons, as shown in Figure 2.11. We observe a deterioration in separation power for both electrons and muons in the low-momentum region - the signal efficiency observes a sharp turn at approximately 1 GeV/c - indicating that the E/p separation power is highly dependent on the magnitude of the 3-momentum in the lab frame, p_{lab} in that region. To investigate whether an improvement in separation power and removal of p-dependence is possible, we turn to supervised learning methods based on a set of features which are discriminative of lepton and hadronic events.

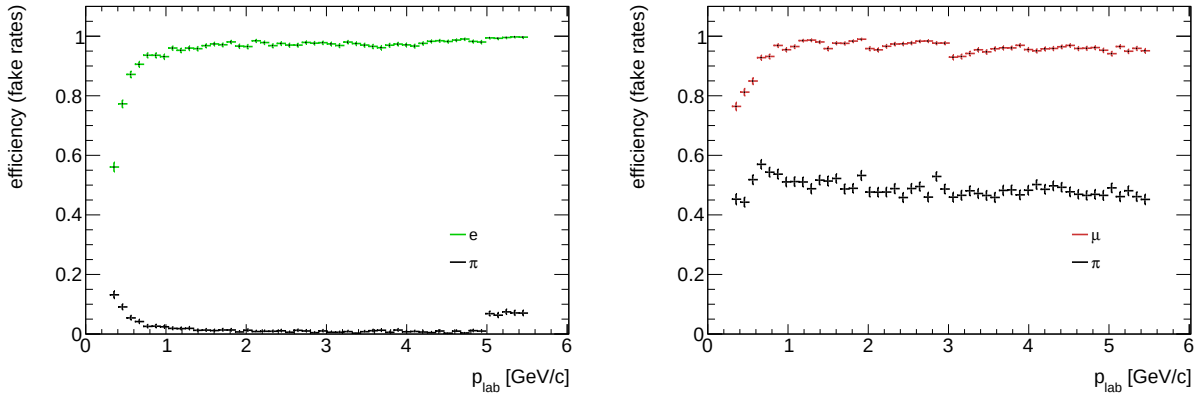


Figure 2.11: Binary PID performance at average signal efficiency $\langle \epsilon_S \rangle = 0.9$, measured in terms of efficiency (background fake rate) versus momentum in the lab frame in particle gun samples using the univariate E/p likelihood for electrons (left) and muons (right). Exhibits deteriorated performance in the low-momentum region. Figures produced by (Marco Milesi, internal Belle II note, 2019).

Dataset

We conduct this study using 10^6 single particle events for e, μ, π generated with the `ParticleGun` module of the official Belle II software framework `basf2`. Beam background events sampled from pseudo-random triggered early 2019 (Phase 3) data are overlaid. Particles were generated according to uniform distributions of in 3-momentum magnitude $(0.2 < p < 5.5)\ \mathrm{GeV}/c$, polar angle $(0 < \theta < 2\pi)$, and azimuthal angle $(0 < \phi < 2\pi)$, so all particles have an equal prior over momentum and angular (θ, ϕ) space.

We use features derived from measurements by the electromagnetic calorimeter (ECL) detector as input to the supervised learning algorithms described below. We briefly describe each input variable below and plot the corresponding distributions for signal and background events in Figures 2.12 and 2.13.

Electromagnetic calorimeter (ECL) response variables

- E_{cluster}: Cluster energy, given by the weighted sum of all ECL crystal energies within the ECL cluster, corrected for leakage and background contamination.
- The absolute values of the $(4,0)$ ($|Z_{40}|$) and $(5,1)$ ($|Z_{51}|$) Zernike moments, used as shape descriptors of the electromagnetic showers generated in the ECL crystals by incident particles.
- $E1E9$: Ratio of energy deposited in the central cluster cell seed to the total energy deposited in the inner 3×3 crystals around the central crystal.
- $E9E21$: Ratio of energies deposited in the central inner 3×3 and outer 5×5 (excluding corners) crystals around the central crystal. Hadronic showers tend to have a large lateral component due to large momentum transfers in nuclear interactions. Both ratios $E1E9$, $E9E21$ are close to 1 for true photon candidates but lower for asymmetric hadronic shower events.
- E_{max} Cluster: Energy of the most energetic crystal in the ECL.
- Cluster E/p: Ratio of the energy deposited by a charged particle in the ECL to the particle's momentum as measured from the curvature of the track geometrically matching the cluster.
- S_2: The second moment of an ECL shower, defined as the following:

$$S_2 = \frac{\sum_{i=0}^{N_c} w_i E_i r_i^2}{\sum_{i=0}^{N_c} w_i E_i}. \tag{2.40}$$

  Where the index i runs over ECL crystals, E_i denotes single crystal energies, w_i are the crystal weights - as each crystal can be related to multiple showers, w_i is the fraction of the energy of the i-th crystal associated with the shower in question, r_i is the distance of the i-th digit to the shower center projected to a plane perpendicular to the shower axis, and N_c denotes the total number of crystals.
- Cluster LAT: Lateral energy distribution of cluster, defined as the following:

$$S = \frac{\sum_{i=3}^{n} w_i E_i r_i^2}{\sum_{i=3}^{n} w_i E_i r_i^2 + w_0 E_0 r_0^2 + w_1 E_1 r_0^2}. \tag{2.41}$$

  Where E_i are the single crystal energies sorted by energy, w_i are the crystal weights as defined above, r_i is the distance of the i-th digit to the shower center projected to a plane perpendicular to the shower axis, and $r_0 \approx 5$ cm is the distance between two crystals. This quantity peaks around 0.3 for radially symmetrical electromagnetic showers and is larger for hadronic events.

To erase the dependency of separation on the lepton momentum, we follow the procedure outlined in Section 2.2. We apply the approximate-KL training procedure outlined in Algorithm 2 with a regularization coefficient of $\lambda = 10$ and learning rate schedules of $\eta_k = 4 \times 10^{-4}$ (inner loop) and $\eta_k' = 10^{-4}$ (outer loop), which appear to work well without requiring any further tuning of hyperparameters. The architecture for the classifier and KL-estimator is the same as the examples detailed in Section 2.2. We reuse the same dataset and data splits used to train the current BDT discriminator, consisting of approximately 750000 training events and 200000 testing events. The input data is normalized to zero mean and unit variance and outliers with a Z-score above 5 are truncated to avoid complications associated with training neural networks on input features with significantly different variances. Training was conducted for 128 epochs on a 24-core SLURM cluster with a single Tesla P100 GPU, with a total duration of approximately 1.5 hours. All results are reported on the independent test set. No hyperparameters were fine-tuned for the purposes of this study.

We compare the classification network trained with the mutual information regularization term as described above, which we refer to as the 'regularized NN', with a standard 4-layer feedforward neural network with 256 neurons per layer trained for the same task in Figure 2.14. For the former approach, we observe the signal efficiency still exhibits large kinematic dependency in the low-momentum region. In the electron sample in particular, the performance of the standard NN suffers from kinematic dependency and deteriorated performance in the low-momentum region under 0.6 GeV. For the latter approach, we observe that the kinematic dependency of the signal efficiency / background fake rate in the low-p region has been almost completely erased, at the expense of mildly reduced overall discrimination power (Figure 2.15). This will enable superior identification of leptons in the critical low-p region in future analyses. One can control the tradeoff between overall discrimination power and sensitivity to p by tuning the regularization coefficient λ. More importantly, in analyses of the $b \to s\ell\ell$ transition, a uniform classifier response across momentum space will prevent introducing systematic effects into the measurement of observables parameterized by the dilepton invariant mass squared, $m_{\ell\ell}^2 = q^2$, as explained in Section 1.1.6.

In Figure 2.15 we plot the ROC curves for the neural network and regularized neural network over the full momentum range and the low-momentum ($p < 1$ GeV) range. While both classifiers can discriminate between signal/background satisfactorily at high momenta, the regularized NN strongly outperforms the NN at low momenta as the training

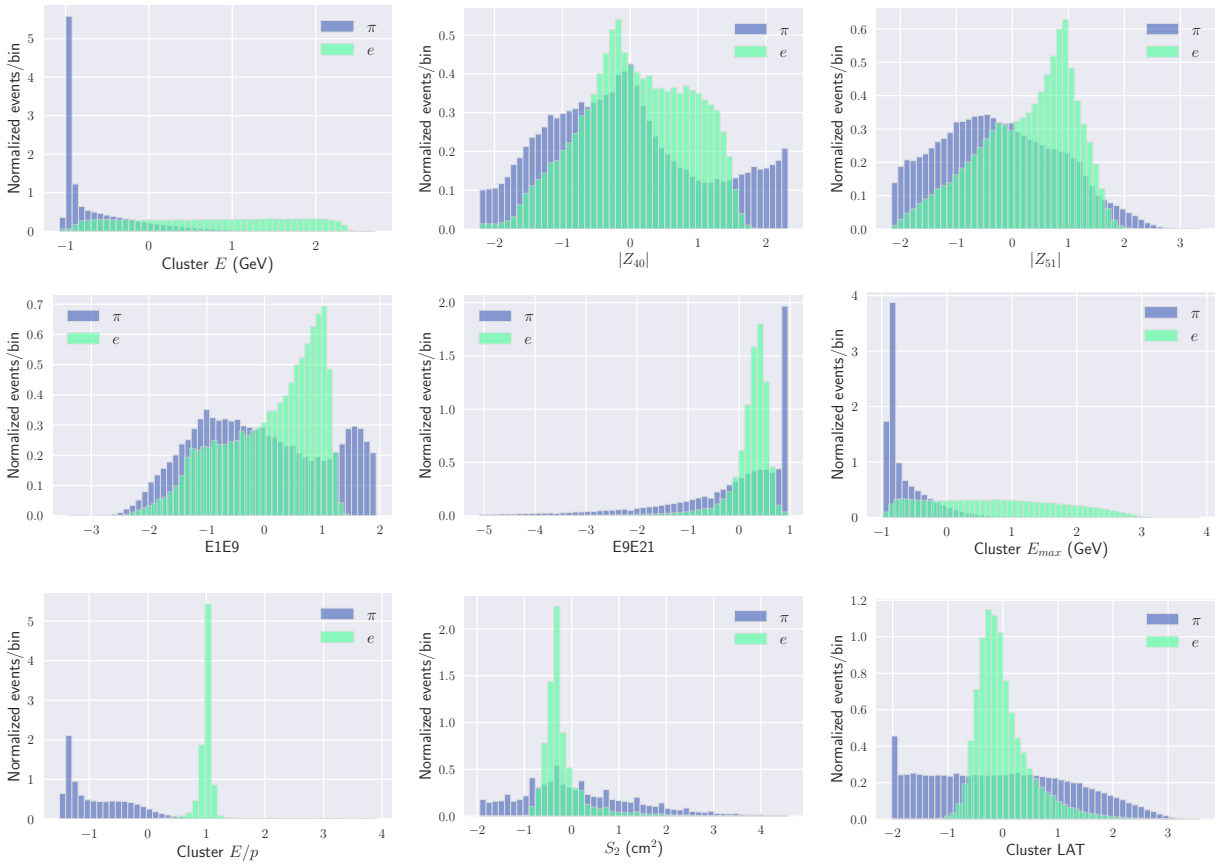


Figure 2.12: Distributions of input features to supervised learning algorithms for discrimination of electron events (green) from hadronic background (blue) in the ECL. Note that each variable has been normalized to zero mean and unit variance.

procedure forces it to be insensitive to the value of $|\mathbf{p}|$. The standard NN is free to condition on $|\mathbf{p}|$ - while this may be an excellent separating variable at high-momenta ($|\mathbf{p}| > 1$ Gev), at low-momenta this is not a reliable indicator of event class, as signal events are obscured by high-multiplicity low-momentum background candidates from hadronic decays. Therefore over-reliance on $|\mathbf{p}|$ can lead to decreased performance in the low-momentum region. For muon identification, separation of muons from hadrons in the low-momentum region is easier than throughout the whole momentum region due to the MIP property of muons and hadronic interactions with the detector material, but more difficult at high-momenta where muons must be separated from the MIP component of the hadronic shower. We observe the regularized NN outperforms the NN in both the low-momentum region and across the full momentum range.

2.5 Conclusion

In this chapter we introduced a method to encourage the hypothesis output by a supervised learning algorithm to be invariant to changes in given sensitive or protected variables based on regularization of an estimate of the mutual information between the learner output and the sensitive variable. We demonstrated application of the developed technique to $b \to s\gamma$ measurements and lepton identification systems at the Belle II experiment. Within the field of high-energy physics, this method may be applied to any experimental analysis where it is necessary to employ a supervised learning algorithm to increase the signal to noise ratio without introducing a bias into the distributions of sensitive attributes under this selection. Such situations commonly arise in high-energy physics. For example, as discussed in Section 2.2, signal and background yields for an experimental measurement are typically extracted through a maximum likelihood fit to the shape of the distribution of certain variables. Biasing the distribution of these variables under selection increases the systematic uncertainty associated with the results of the maximum likelihood fit. Another example comes from experimental analyses which aim to test proposed models of new physics. If the

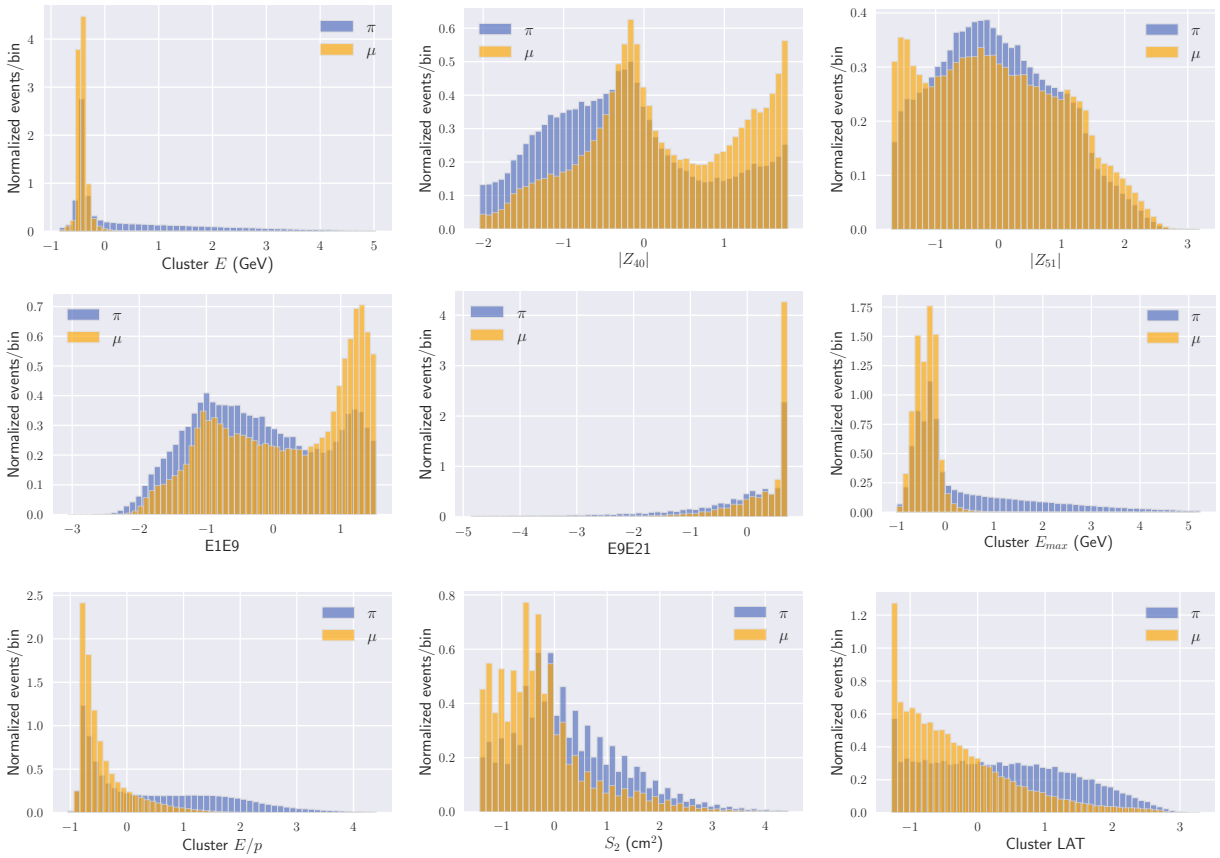


Figure 2.13: Distributions of input features to supervised learning algorithms for discrimination of muon events (orange) from hadronic background (blue) in the ECL.

parameters of the model are dependent on given observables, learning algorithms must avoid biasing these observables to prevent undesirable systematic effects on the measurement. In the continued absence of concrete signatures of new physics, the requisite sensitivity of experimental analyses must increase to probe larger regions of parameter space. Methods such as this will become increasingly important due to the dual demands of a high signal-to-noise ratio and low systematic uncertainty. We have implemented the proposed approach for learning invariant representations within the `basf2` framework, under the multivariate analysis software package used by the Belle II collaboration. We have made the code used for training and several examples available open-source under the Apache License 2.0 at https://github.com/Justin-Tan/invariant_reps.

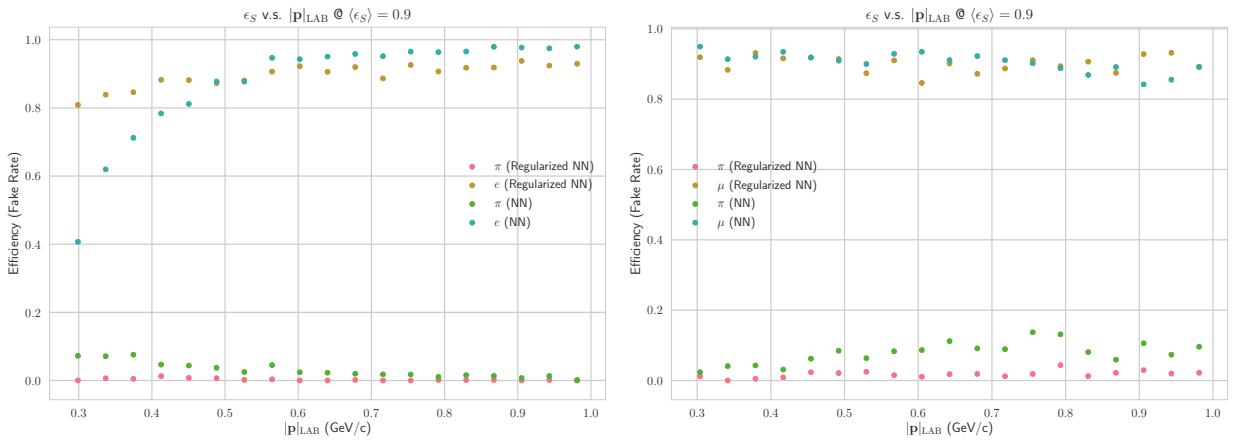


Figure 2.14: Binary PID performance at an average signal efficiency $\langle\epsilon_S\rangle = 0.9$, measured in terms of efficiency (background fake rate) versus momentum in the lab frame in particle gun samples. Here the standard neural network (NN) and regularized neural network (regularized NN) signal efficiencies are overlaid for electrons (left) and muons (right). For electron identification, the standard NN exhibits high kinematic dependency and deteriorated performance in this region, while the performance of the regularized NN shows no kinematic dependency at the expense of slightly reduced discrimination power across the whole $|\mathbf{p}|$ spectrum. For muon identification, the standard NN exhibits lower background rejection in the high momentum region, while background efficiency is uniformly lower throughout the whole momentum region for the regularized NN.

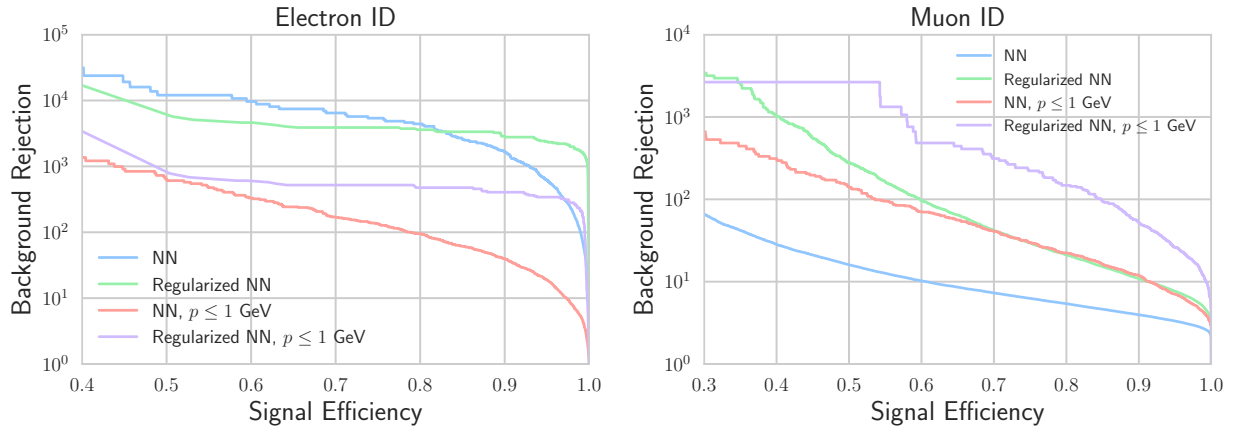


Figure 2.15: Signal efficiency versus background rejection plots for the neural network and regularized neural network in different momentum regions for electron ID (left) and muon ID (right). Blue and green curves show the classifier performance across the entire momentum space, while the violet and red curves show performance in the low momentum $p < 0.1$ GeV region.

CHAPTER 3

INVARIANCE THROUGH DISENTANGLEMENT

In this chapter we return to the problem first introduced in Chapter 2 of encouraging the hypothesis output by a supervised learner (Section 2.1.1) to be invariant to changes in given sensitive or protected variables. We adopt a representation learning perspective (Bengio et al., 2013) in this chapter, and formulate the problem as finding a intermediate representation of the original data that is invariant to changes in the sensitive attributes. The main contribution of this chapter is a means of achieving this by learning an intermediate representation which is *disentangled* with respect to the underlying factors of variation of the original data. Informally, we model the instances $\{x_i\}_{i=1}^N$ in our training set as a set of observations from some high dimensional random variable. The true data generative process of this random variable (Eq. 2.3) is assumed to be described by a finite set of (generally unknown) ground-truth factors of variation which provide a minimal but sufficient description of the original high dimensional data. A disentangled representation of the data is one in which each factor of variation is described by a distinct portion of the representation, such that the influence of each factor of variation is isolated in a particular subspace of the representation. Further discussion of this property may be found in Section 3.4. By excising the subspace of the representation corresponding to protected attributes of the input data, the resulting reduced representation will be invariant to changes in these attributes. Consequently, learning algorithms may use this reduced representation in lieu of the original data to learn hypotheses which do not depend on the protected attributes.

This chapter is structured as follows: in Section 3.1 we briefly review latent variable models and variational inference, which underpin our proposed approach. In Section 3.2 we define disentanglement more formally and outline previously proposed methods for achieving disentanglement in latent variable models. In Section 3.3 we review related work and place our work into context. Section 3.4 contains the main contribution of this chapter; a description of our proposed procedure for achieving invariance through disentanglement. Section 3.5 contains empirical results demonstrating the validity of the proposed method as well as comparisons to previously proposed approaches using the same high-energy physics application from Chapter 2.

3.1 MOTIVATION

3.1.1 Generative Models

As discussed in Section 2.1, machine learning in experimental high-energy physics has traditionally been dominated by the discriminative learning paradigm, where we attempt to infer the category of process responsible for generating some observed datapoint by modelling the likelihood $p(\mathsf{category}|\mathsf{datapoint})$. The primary application of this is to distinguish between events of interest and background processes on the basis of detector measurements at particle colliders.[1]

A more general task is to directly model the underlying process responsible for generating the observed data. Such models are known as *generative models* (Goodfellow et al. (2016), Ch. 20). The resulting model can be used for a variety of downstream tasks such as approximate simulation, representation learning and density estimation. Formally, let x represent the set of all variables describing our observed data. x may describe the pixel intensities of a natural images, term frequencies in text corpora, or closer to home, measurements describing the kinematic and geometric

[1] One may refer to Apostolakis et al. (2018) for a comprehensive overview of applications of machine learning to high-energy physics.

properties of a recorded event in the detector of a collider experiment. We assume x is a random sample from some underlying process associated with a ground truth probability distribution we denote as $p^*(x)$:

$$x \sim p^*(x). \tag{3.1}$$

Typically we are able to freely sample from the distribution but unable to directly evaluate the density. To approximate the ground-truth distribution, we introduce a model with parameters θ: $p_\theta(x)$, and learn the parameters θ through by minimization of some divergence between the model and the ground-truth distributions (Stummer and Vajda, 2012), such that $p_\theta(x) \approx p^*(x)$ for all possible x. Recently generative models have received considerable attention from the high-energy physics community, primarily in the context of fast simulation of calorimeter radiation patterns ((Paganini et al., 2018), (de Oliveira et al., 2017)).

Latent Variable Models

Latent variable models augment the set of observed variables x with unobserved, or latent random variables z which explain hidden causes of the data. Every observed data point x_i is modelled as independently generated by a different latent variable z_i. Generative models extended to the latent variable case attempt to model the complete joint distribution $p_\theta(x, z)$ over both latent and observed variables. This is achieved by defining a prior distribution $p(z)$ and a conditional likelihood $p_\theta(x|z) = p(x|f_\theta(z))$ that models the relationship between the latent variables z and the observed data x, where a function f_θ, parameterized by θ is introduced to explicitly model this relation. The functional form of the prior $p(z)$ and likelihood term $p_\theta(x|z)$ will depend on the choice of model and the type of the observed data. The marginal distribution over the observed variables may be obtained through marginalization of the latents in the joint distribution:

$$p_\theta(x) = \int dz\, p_\theta(x, z) \tag{3.2}$$

$$= \int dz\, p_\theta(x|z)p(z). \tag{3.3}$$

The advantage of this formulation is that we can approximate the complex ground truth distribution $p^*(x)$ over the original space as the combination of simpler distributions over the complete variable space. Despite the simplicity of the individual constituent distributions in Eq. 3.3, the final marginal distribution $p_\theta(x)$ can be thought of as a weighted mixture of these distributions, permitting $p_\theta(x)$ to be a very flexible model in the same way that mixtures of Gaussians are significantly more expressive than individual Gaussians. In particular, if the parameters θ are the output of nonlinear transformations such as multilayer feedforward dense networks (here after referred to as neural networks), then the marginal distribution can be transformed to an arbitrarily complex distribution, potentially making it capable of modelling the original high-dimensional data x. Generative latent variable models invite a natural interpretation in which the observed data x is generated by an unobserved underlying set of latent variables z that provides a more fundamental representation of the data. First the latent variable z is sampled from some prior distribution $p(z)$, and the observation x is subsequently sampled from the conditional likelihood term $p_\theta(x|z)$.

$$z \sim p(z), \tag{3.4}$$

$$x|z \sim p(x|f_\theta(z)) = p_\theta(x|z). \tag{3.5}$$

From a representation learning perspective, the objective of a latent variable model is to infer the values of the latent variables z responsible for generating the observed data by evaluating the posterior distribution $p(z|x)$ over the latent variables. The advantage of this is that the structure of the original high dimensional data x may be hard to interpret or even define. By modelling x as a transformation of a latent variable z, we can avoid dealing with complicated high-dimensional distributions over x directly and deal instead with simpler distributions over the (typically significantly lower-dimensional) latent space. This permits us to more easily enforce desirable structure in the lower-dimensional latent space and use this learned latent representation of the data x in downstream tasks.

3.1.2 Variational Inference

In general there is no closed-form expression for the integral expression of the marginal likelihood $p_\theta(x)$ (Eq. 3.3)[2]. Furthermore, if the latent space dimension is sufficiently large, as is the case in most practical applications, numerical integration techniques require a prohibitive number of samples to approximate (Eq. 3.3) due to the curse of dimensionality.[3] Therefore, we are not able to directly optimize $p_\theta(x)$ through gradient-based maximum likelihood estimation. Note that intractability of the marginal $p_\theta(x)$ implies intractability of the posterior $p_\theta(z|x)$ and vice-versa by Bayes' Theorem:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)}. \tag{3.6}$$

The likelihood $p(x|z)$ and prior $p(z)$ are efficiently evaluated by design, but computation of the posterior $p(z|x)$ necessitates evaluation of the marginal likelihood. However, introduction of a tractable approximation to the posterior will enable us to approximate the marginal likelihood. *Variational inference* (Jordan et al., 1999) phrases the problem of approximating the posterior $p_\theta(z|x)$ in latent variable models as an optimization problem. First some family of tractable distributions $\mathcal{Q}$ over the latent variables is introduced, parameterized by variational parameters ϕ. The goal is to find the member of this family $q_\phi \in \mathcal{Q}$ that is the closest approximation to the model posterior $p_\theta(z|x)$, $q_\phi(z) \approx p_\theta(x|z)$ for all observed x. q_ϕ will be referred to as the *variational posterior*. The resulting variational posterior $q_\phi(z)$ may then be used as a stand-in for the true posterior $p_\theta(z|x)$ in downstream tasks. Note that the true posterior will typically not lie in $\mathcal{Q}$, and the quality of the variational approximation will be dependent on the class of distributions considered in $\mathcal{Q}$.

Learning Objective

To uncover the objective function we should use during variational inference, note that the marginal log-likelihood can be expressed as:

$$\begin{aligned}
\log p_\theta(x) &= \int dz\, q_\phi(z) \log p_\theta(x) && (3.7)\\
&= \int dz\, q_\phi(z) \left(\log p_\theta(z,x) - \log p_\theta(z|x) + \log q_\phi(z) - \log q_\phi(z)\right) && (3.8)\\
&= \int dz\; q_\phi(z) \left(\log \frac{p_\theta(z,x)}{q_\phi(z)} + \log \frac{q_\phi(z)}{p_\theta(z|x)}\right) && (3.9)\\
&= \mathbf{E}_{q_\phi}\left[\log \frac{p_\theta(z,x)}{q_\phi(z)}\right] + D_{\mathrm{KL}}\left(q_\phi(z) \| p_\theta(z|x)\right) && (3.10)\\
&\geq \mathbf{E}_{q_\phi}\left[\log \frac{p_\theta(z,x)}{q_\phi(z)}\right] && (3.11)\\
&\triangleq -\mathcal{F}(\theta,\phi). && (3.12)
\end{aligned}$$

The second last inequality (Eq. 3.11) arises as the KL divergence is non-negative. Eq. 3.12 is referred to as the negative *variational free energy* or (log-) *evidence lower bound* (ELBO), because it lower bounds the marginal log-likelihood $\log p_\theta(x)$, also referred to as the log-evidence. We observe the log evidence is equal to the sum of the ELBO and the KL divergence between the variational distribution $q_\phi(z)$ and posterior $p_\theta(z|x)$. The closer the variational posterior $q_\phi(z)$ is to the posterior $p_\theta(z|x)$ as measured by the KL divergence, the tighter this lower bound becomes. Therefore, to learn the parameters of our generative model, we should optimize the set of model parameters (θ, ϕ) to maximize $-\mathcal{F}(\theta,\phi)$, approximately maximizing the log-evidence with gap given by the KL divergence $D_{\mathrm{KL}}\left(q_\phi(z)\|p_\theta(z|x)\right)$.

[2] Certain notable exceptions exist, such as the case where both $p(z)$ and $p_\theta(x|z)$ are normal, and z is a linear function of x. Even if a closed-form solution is possible, e.g. through conjugacy of the prior and likelihood, evaluation of the integral still requires exponential time even in relatively simple cases (e.g. (Blei et al., 2016))

[3] Integrals where evaluation is blocked by these collective obstacles will be referred to as *intractable* hereafter.

As the logarithm is monotonic, this will approximately maximize the log-evidence $p_\theta(x)$. Note that the ELBO is the negative of the KL divergence in Eq. 3.11 plus the log-evidence.

$$\mathbf{E}_{q_\phi}\left[\log \frac{p_\theta(z,x)}{q_\phi(z)}\right] = -D_{\mathrm{KL}}\left(q_\phi(z)\|p_\theta(z|x)\right) + \log p_\theta(x). \tag{3.13}$$

As the log-evidence is a constant with respect to $q_\phi(z)$, maximization of the ELBO is equivalent to minimization of $D_{\mathrm{KL}}\left(q_\phi(z)\|p_\theta(z|x)\right)$. Thus the objective of variational inference can be formulated as finding the distribution $q_\phi \in \mathcal{Q}$ that is closest to the true posterior under the KL divergence:

$$\operatorname*{argmin}_{q\in\mathcal{Q}} D_{\mathrm{KL}}\left(q_\phi(z)\|p_\theta(z|x)\right). \tag{3.14}$$

To understand the form of the ELBO further, we observe it may be expressed as the sum of the log-likelihood under the variational posterior q_ϕ and the negative of the KL divergence between q_ϕ and the latent prior $p(z)$:

$$\mathbf{E}_{q_\phi}\left[\log \frac{p_\theta(z,x)}{q_\phi(z)}\right] = \mathbf{E}_{q_\phi}\left[\log p(x|z) - \log \frac{q_\phi}{p(z)}\right] \tag{3.15}$$

$$= \underbrace{\mathbf{E}_{q_\phi}\left[\log p_\theta(x|z)\right]}_{\text{Expected log-likelihood}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(z)\|p(z)\right)}_{\text{Complexity Penalty}}. \tag{3.16}$$

The first term in Eq. 3.16, the expected log-likelihood under $q_\phi(z|x)$, has an interpretation as a 'reconstruction term' which encourages informativeness of latent variables with the observed data. The second term is the negative KL divergence between the variational posterior q_ϕ and the latent prior, which is typically chosen to be an isotropic Gaussian for continuous z. This acts as a regularizer for the learnt posterior by penalizing the complexity of the learned q_ϕ distribution. Intuitively, this can be thought of as imposing the principle of Ockham's Razor on the variational posterior - favouring the simplest possible distribution over the latent variables z explaining the observed data x. Note that in the case where we condition the parameters ϕ of the learnt posterior on the observed data x one should replace $q_\phi(z) \to q_\phi(z|x)$ in the previous discussion.

3.1.3 Variational Autoencoders

While there are many approaches to optimization of the objective in Eq. 3.14 ((Jordan et al., 1999), (Wainwright and Jordan, 2008)), we focus here on the *Auto-Encoding Variational Bayes* (AEVB) algorithm (Kingma and Welling, 2014). Latent variable models trained via AEVB are referred to as variational autoencoders (VAEs). Unlike traditional approaches to variational inference where the variational parameters are learnt per datapoint, under the VAE framework, the variational posterior q_ϕ is represented by a parametric distribution with the same form across the whole dataset. For the observed data instance x_i, the distribution parameters of the variational posterior are a function of the corresponding data point z_i, but the parameters of this function - the variational parameters ϕ - are learned using the entire dataset through optimization of the ELBO. Therefore the AEVB framework optimizes a set of variational parameters which are common to all datapoints. This amortizes the computation cost across the entire observed dataset, permitting the AEVB procedure to scale effectively to large datasets. In the VAE framework, the distributions $p_\theta(x|z)$ and $q_\phi(z|x)$ are parameterized by neural networks. The former is commonly referred to as the generative/decoder model and the latter as the inference/encoder model, after the autoencoding literature.[4] For simplicity of notation, we take the parameters of the neural networks to be subsumed into the model parameters (θ, ϕ). In general, if the latent and data distributions are represented as distributions f, g with parameters $\{\mu_i\}_{i=1}^m$ and $\{\nu_i\}_{i=1}^n$, respectively, then the distributions are modelled as follows:

$$(\mu_1, \ldots, \mu_{D_z}) = E_\phi(x), \tag{3.17}$$

$$q_\phi(z|x) = f\left(z; \mu_1, \ldots, \mu_m\right), \tag{3.18}$$

$$(\nu_1, \ldots, \nu_{D_x}) = D_\theta(z), \tag{3.19}$$

$$p_\theta(x|z) = g\left(x; \nu_1, \ldots, \nu_n\right). \tag{3.20}$$

[4] Although it should not be forgotten that a variational autoencoder is fundamentally a model for performing variational inference and not strictly an autoencoder, despite the strong resemblance of their respective objectives.

Where the encoder E_ϕ and decoder D_θ functions are represented by neural networks with parameters ϕ and θ, respectively. The former accepts a sample from the true data distribution and returns the parameters of the latent distribution, while the latter accepts a sample form the learned latent distribution and outputs the parameters of the model data distribution. A convenient modelling choice in VAEs is to assume the following factorized form for the family of distributions containing the variational posterior:

$$\mathcal{Q}_f = \left\{ q_\phi(z|x) = \prod_{i=1}^{D_z} q_\phi(z_i|x_i) \;\middle|\; \phi \in \Phi \right\}. \tag{3.21}$$

Where $\phi \in \Phi$ represents the parameters of the function mapping the observed data x to the distribution parameters of $q_\phi(z|x)$, and D_z is the dimension of the latent space. It is common to use a factorized (diagonal-covariance) Gaussian to represent the variational distribution q_ϕ, and take the prior to be the isotropic standard Gaussian, $p(z) = \mathcal{N}(\mathbf{0}, \mathbb{1})$ In this case the scale and location parameters for the distribution $q_\phi(z|x)$ are given as:

$$(\mu_\phi, \log \operatorname{diag}(\Sigma_\phi)) = E_\theta(x, \phi), \tag{3.22}$$

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi, \Sigma_\phi) \tag{3.23}$$

$$= \prod_{i=1}^{D_z} \mathcal{N}(z_i; (\mu_\phi)_i, (\Sigma_\phi)_{ii}). \tag{3.24}$$

Where the covariance matrix $\Sigma_\phi = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{D_z}^2\right)$. The choice of distribution to represent the generative model p_θ depends on the type of data being considered. For continuous data a common choice is a factorized Gaussian, while for natural images $p_\theta(x|z)$ can be taken to be a Bernoulli distribution over each pixel. If factorized distributions are not sufficiently expressive to represent the variational posterior or generative model, it is possible to remove this restriction and represent the inference and generative models as unfactorized distributions. For example, full-covariance Gaussians, or the end product of a sequence of simple transformations from a base distribution ((Rezende and Mohamed, 2015), (Dinh et al., 2017)). This increases the range of distributions our model is able to represent at the expense of additional computation.

Optimization

As neural networks are differentiable with respect to their parameters, the ELBO can be efficiently optimized with respect to the generative and variational parameters through standard gradient-based optimization techniques. Both sets of parameters (θ, ϕ) are optimized simultaneously. Optimization over ϕ helps the variational posterior $q_\phi(z|x)$ to approach the true posterior $p_\theta(z|x)$, keeping the ELBO tight around the marginal log-likelihood. Optimization over θ improves the quality of generated samples, and pushes the lower bound on $\log p_\theta(x)$ up. Optimization over While it is straightforward to compute unbiased Monte Carlo estimators of the gradient $\nabla_\theta \mathcal{F}(\theta, \phi)$ as the gradient operator ∇_θ and expectation $\mathbf{E}_{q_\phi}$ commute, this is not true of ∇_ϕ and $\mathbf{E}_{q_\phi}$. Computing unbiased Monte Carlo estimates of $\nabla_\phi \mathcal{F}(\theta, \phi)$ requires an additional step referred to as the reparameterization trick ((Kingma and Welling, 2014), (Rezende and Mohamed, 2015)). Briefly, the goal is to express the expectation $\mathbf{E}_{q_\phi}$ as an expectation over an ancillary random variable ϵ which does not depend on ϕ, allowing the expectation and gradient operator ∇_ϕ to commute. This can be achieved by letting z be a diffeomorphism (invertible, differentiable transformation) of ϵ, $z = T(\epsilon)$. Then the expectation of an arbitrary function of z with respect to q_ϕ becomes $\mathbf{E}_{q_\phi}[f(z)] = \mathbf{E}_{p(\epsilon)}[f(T(\epsilon))]$, allowing the expectation and gradient with respect to ϕ to commute. This allows both the generative model and variational posterior to be jointly trained in an efficient manner using modern libraries implementing gradient-based optimization through reverse-mode automatic differentiation.

The final objective for the base VAE framework, taken in expectation over the data distribution is shown below. In practice we average over the empirical data distribution (Eq. 3.26).

$$\max_{\theta,\phi} \mathbf{E}_{p(x)} \left[\mathbf{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - D_{\mathrm{KL}} (q_\phi(z|x) \| p(z)) \right] \tag{3.25}$$

$$\approx \max_{\theta,\phi} \sum_{n=1}^{N} \mathbf{E}_{q_\phi(z|x_n)} [\log p_\theta(x_n|z)] - D_{\mathrm{KL}} (q_\phi(z|x_n) \| p(z)). \tag{3.26}$$

3.2 Latent Space Disentanglement

As mentioned in Section 3.1, we may derive some representation of the latent space from the variational posterior $q_\phi(z|x)$ and use this learned representation in subsequent downstream tasks. Explicitly modelling the distributions $q_\phi(z|x)$ and $p_\theta(x|z)$ is useful as this allows us to apply information-theoretic arguments to constrain the structure of the latent space, where this may be more difficult to achieve with representations without a probabilistic interpretation. For example, to constrain the complexity of the learnt latent space, we may penalize the KL divergence between the learnt distribution and some simple base distribution. To encourage independence between the dimensions of a learnt representation, we can enforce a factorial structure in the distribution, etc. The caveat is that the expectations involved in these information-theoretic constructs may be intractable and require some approximation procedure to estimate, for example importance-sampling or variational approximation.

Examples of desirable properties of this learned representation that may be useful on subsequent downstream tasks include interpretability, low description length/compactness. and disentanglement of the learned latent space ((Shu et al., 2019), (Higgins et al., 2018)). While the exact definition of a disentangled representation remains a point of contention, informally, we wish to for the latent variables in our model to encoder information about the true generative factors of the observed data. Therefore, the latent space should be partitioned into distinct regions - each of which corresponds to a distinct underlying *factor of variation* (Bengio et al., 2013) of the original data. Traversing each region of latent space while holding the others fixed should only lead to a change in a single factor of the learned representation $r(x)$ of the original data x. By factors of variation we refer to explanatory factors of the data. For example, when describing images of geometric objects, we may distinguish them by scale, shape, color and orientation, none of which depend on another *a priori*. When describing portraits of human faces, example factors may be related to the facial pose (azimuth, evaluation), as well as facial attributes such as facial width, gender etc. Visual examples illustrating disentanglement of the latent space on simple geometric image datasets are shown in Figures 3.2 and 3.3.

One important assumption for disentangled representation learning is that different factors of variation change approximately independently of one another throughout the input distribution, allowing learning algorithms to identify and isolate each explanatory factor. It is posited that disentangled representations may be a useful property when learning real-world downstream tasks, but such investigations remain in their infancy. Notably, van Steenkiste et al. (2019) conduct a large-scale experimental study that demonstrates disentangled representations exhibit increased sample efficiency on challenging abstract visual reasoning tasks. Locatello et al. (2019a) demonstrate that disentangled representations are correlated with increased fairness on downstream tasks requiring the prediction of some variable without bias against some sensitive attribute of the input data.[5]

3.2.1 Informal Definition

Higgins et al. (2018) attempt to define a disentangled vector representation in terms of the symmetry properties of the representation under group actions corresponding to symmetry transformations in the data space. While this is a useful definition for synthetic datasets used in the study of disentangled representations ((Matthey et al., 2017), (Burgess and Kim, 2018)) with known and exact symmetries, it remains to be seen how to generalize this to real-world data where the true factors of variation are generally unknown. Furthermore, while it is intuitively clear how to express physical transformations such as reflection and rotation as group actions, it is unclear how to extend this to arbitrary factors of variation. Earlier work (Eastwood and Williams, 2018) provide a set of less comprehensive but intuitive criteria that disentangled representations of the latent space should satisfy:

- **Modularity:** A modular representation is one in which each individual latent dimension encodes no more than one factor of variation. This is a universally agreed-upon criterion for disentanglement. If multiple generative factors are projected onto a single dimension, they will be related in a nonlinear way by definition, potentially making it difficult to isolate each factor.
- **Compactness:** A compact representation is one in which each generative factor is encoded by an individual latent dimension. i.e. there is a one-to-one mapping between latents and factors of variation. This is a more controversial criterion than modularity. Most of the current VAE-based methods for disentanglement discussed below ((Higgins et al., 2017), (Burgess et al., 2018), (Kim and Mnih, 2018), (Chen et al., 2018), (Shu et al., 2019)) are developed to enforce a one-to-one correspondence between the latent and factors of variation as opposed to a many-to-one mapping, and assume this requirement to hold. We will assume compactness is necessary for disentanglement throughout this work. However, note that there are simple cases in which it is

[5] Note they work in the setting where the sensitive attribute s is independent of the target variable y, which may be unrealistic in real-world applications such as algorithmic fairness.

natural for compactness to be violated. For example, to fully specify a rotation in 3D, one must specify the axis of rotation, the angle of rotation about the axis, and the orientation of the rotation. It is not possible to encode all this information in a single dimension and so a compact representation cannot encode 3D rotations.

- **Explicitness:** An explicit representation is one in which the values of all the factors of variation are related linearly to the learned representation. In practice this is too strong a requirement for VAE-based disentanglement methods, as this requires the learned representation to be:
  - Informative of all factors of variation - unrealistic in real-world scenarios where the true factors of variation are typically unknown.
  - Able to recover each factor of variation under a linear transformation. This is not realistic as a complexity penalty (e.g. through the KL term $D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)$in the ELBO) is enforced on the variational posterior, and we offload some of the representational capacity to the generative model $p_\theta(x|z)$ by parameterizing the latter as a neural network with nonlinear activations.

For the remainder of this work we will assume that a disentanglement representation should satisfy modularity and compactness, adopting the following somewhat informal definition that will be sufficient for our purposes:

Definition 3.2.1. (Latent Space Disentanglement) Let the factors of variation of the observed data x be denoted by $v = (v_1, v_2, \ldots, v_K)$. Assume we obtain a latent representation of x by training a latent-variable model with latent variables $z \in \mathcal{Z}$. A disentangled representation $z = (z_1, z_2, \ldots z_{D_z})$ of the observed data is one in which:

a) The dimensions $\{z_i\}$ of the latent representation are statistically independent.

b) The mutual information $I(z_i; v_j)$ for each latent dimension z_i should be nonzero only for a single factor v_j.

If both of these conditions are satisfied only for a subset of z we call the representation partially disentangled.

Intuitively, each dimension z_i should exclusively store information about a single factor of variation v_j. Note that the above definition satisfies modularity but only satisfies compactness if z_i has nonzero mutual information with unique factor v_j. As the mutual information is an intractable quantity we are not able to enforce both conditions exactly for continuous (z, v) and must resort to a relaxation of the conditions, as discussed in Section 3.2.2. No constraints are imposed on the exact nature of the relation between z_i and v_j. Ideally we would like to learn a representation in which z_i is a deterministic function of factor v_j, such that the conditional entropy $H(v_j|z_i) = 0$ and the mutual information $I(z_i; v_j) = H(v_j) - H(v_j|z_i) = H(v_j)$. However, this may not be possible and we may only be able to approximate this condition through a learnable nonlinear mapping represented by e.g., a neural network, as discussed in Section 3.2.2. Suter et al. (2019) provide a more formal definition that allows for the possibility that the generative factors are statistically dependent when influenced by confounding factors, but we will not consider this (important) added complication for now.

3.2.2 Unsupervised Learning of Disentangled Representations

In practice most approaches to learning disentangled representations of the latent space under the VAE framework augment the evidence lower bound objective (Eq. 3.25) with terms motivated by information theory. These terms are typically chosen to achieve constraints on the variational posterior $q_\phi(z|x)$ or its marginal or aggregate posterior $q(z) = \int dx\ q_\phi(x, z)$. The main approaches are briefly outlined below, where it is implicit that any of the below proposed objective functions are to be maximized.

Information Bottleneck

Early approaches to consider disentanglement in VAEs (Higgins et al., 2017) propose the β-VAE, introducing a titular hyperparameter $\beta > 1$ which reweights the KL divergence between the variational posterior and prior in Eq. 3.25.

$$\mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right]. \tag{3.27}$$

Where the variational posterior $q_\phi(z|x)$ is generally chosen to be diagonal-covariance Gaussian and the prior $p(z)$ is chosen to be the isotropic standard Gaussian. Burgess et al. (2018) argue that the variational posterior acts as an information bottleneck (Tishby et al., 1999), for the log-likelihood component of the ELBO, $\mathbf{E}_{q_\phi}\left[\log p_\theta(x|z)\right]$. Informally, the KL divergence term measures the amount of information encoded in the latent space about the data

distribution. Penalizing this term with weighting $\beta > 1$ decreases the capacity of the latent distribution. Recalling that the log-likelihood component rewards learnt latent densities which explain the observed data, the authors further argue this complexity penalty on q_ϕ forces each dimension of the learnt latent representation to specialize onto distinct and individual factors of variation. While this argument is intuitively compelling, there is no guarantee that an efficient representation of the data should be axis-aligned with the true factors of variation of the data. In the same vein, the authors propose the annealed-VAE, which progressively increases the capacity of the posterior over training to allow the distribution to learn one factor of variation at a time:

$$\mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \gamma|D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right) - C|\right]. \tag{3.28}$$

Where $\gamma > 0$ and C is annealed from zero to a positive value over the course of training.

Factorial Representations

Another proposed method of learning disentangled latent representations is to enforce a factorial structure in the variational posterior over z, to encourage the dimensions of the learnt latent space to be statistically independent - although this is evidently not a sufficient condition - if the variational posterior is an isotropic Gaussian then the latent dimensions are independent but it cannot convey any information about the data distribution. The KL divergence component in the ELBO can be decomposed as:

$$\begin{aligned}
\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] &= \int dx dz\, p(x) q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z)} && (3.29)\\
&= \int dx dz\, p(x) q_\phi(z|x) \log \frac{q_\phi(z|x) q(z)}{p(z) q(z)} && (3.30)\\
&= \int dx dz\, p(x) q_\phi(z|x) \log \frac{q_\phi(z|x)}{q(z)} + \int dx dz\, q(x,z) \log \frac{q(z)}{p(z)} && (3.31)\\
&= \mathbf{E}_{q(x,z)}\left[\log \frac{q(x,z)}{p(x)q(z)}\right] + \int dz\, q(z) \log \frac{q(z)}{p(z)} && (3.32)\\
&= I_q(x;z) + D_{\mathrm{KL}}\left(q(z)\|p(z)\right). && (3.33)
\end{aligned}$$

Where $I_q(x;z)$ is the mutual information between x and z under the empirical joint distribution $q(x,z) = q_\phi(z|x)p(x)$, where $p(x)$ is the true data distribution. As β-VAE and associated variants penalize this term, $I_q(x;z)$ is also penalized, restricting the information content in z about x. Kim and Mnih (2018) argue that penalizing $I_q(x;z)$ more than the standard VAE is neither necessary or desirable for disentangling, as this can lead to poor reconstruction of x from the latent codes. Recall that $q(z) \approx \sum_n q_\phi(z|x)p(x)$ reflects the average structure of the latent representation over the real data distribution. Thus, assuming a factorial prior, penalizing $D_{\mathrm{KL}}\left(q(z)\|p(z)\right)$ directly encourages independence between the latent dimensions. Inspired by this, Kim and Mnih (2018) propose the following objective:

$$\mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] - \gamma D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right). \tag{3.34}$$

The term $D_{\mathrm{KL}}\left(q(z)\|\prod_i^d q(z_i)\right)$ is the KL divergence between the joint distribution $q(z_1, \ldots z_d)$ and the product of marginals $q(z_1)q(z_2)\ldots q(z_d)$, referred to as the total correlation (TC) (Watanabe, 1960). The total correlation is a particular generalization of the mutual information to the case of more than two random variables. Like the mutual information, it is intractable to evaluate or optimize over for the case of continuous non-Gaussian densities. A low TC corresponds to a low degree of dependency between the latent dimensions - hence penalization of this term is expected to promote factorial structure of the latent representation. Kim and Mnih (2018) make use of the density-ratio trick described in the preceding section to yield a differentiable estimate of the density ratio appearing in the TC term. This requires training an auxiliary discriminator D_φ, typically taken to be a fully-connected neural network to output an estimate $D_\varphi(z)$ of the probability the sample z is drawn from the joint $q(z)$ versus the product of marginals (see Section 2.2.3 for further details). The discriminator is optimized with respect to a proper scoring rule for binary classification (Gneiting and Raftery, 2007). Choosing the Bernoulli score, which is minimized over, corresponds to the well-known cross-entropy loss. Thus, factor-VAE requires an inner optimization loop over a non-convex objective to estimate the TC term used to compute the objective.

Independently, Chen et al. (2018) propose a similar extension of the ELBO based on the following alternative decomposition of the KL divergence component:

$$\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] = I_q(x;z) + D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right) + \sum_i D_{\mathrm{KL}}\left(q(z_i)\|p(z_i)\right). \tag{3.35}$$

The first term in Eq. 3.35 is the mutual information $I_q(x;z)$ between x and z under the empirical joint distribution $q(x,z) = q_\phi(z|x)p(x)$, as described previously. The second term is the total correlation term discuss above, and the third term is the dimension-wise KL divergence between the aggregate posterior and factorial prior, preventing individual latent dimensions from deviating significantly from the corresponding prior. Chen et al. (2018) provide empirical evidence to demonstrate that the TC term is the most important term in the decomposition for learning disentangled representations, and only weight the TC term in Eq. 3.35 in their proposed objective, referred to as β-TCVAE:

$$\mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]\right] - \left(I_q(x;z) + \beta D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right) + \sum_i D_{\mathrm{KL}}\left(q(z_i)\|p(z_i)\right)\right). \tag{3.36}$$

The terms $\log q(z)$ and $\log \prod_i q(z_i)$ are estimated using Monte Carlo estimates obtained through importance sampling, which avoids the inner optimization loop for the discriminator parameters necessary to estimate the TC term using factor-VAE. Note that both Factor-VAE and β-TCVAE can be interpreted as the ELBO (Eq. 3.25) plus a TC term with prefactor $\lambda - 1$. As it may not be possible to enforce independence between the latent dimensions z_i exactly, this can be considered a relaxation of the first condition in Definition 3.2.1. In this relaxation, we consider disentangled representation learning as an constrained optimization problem where we maximize the ELBO subject to the constraint $D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right) < \epsilon$ with Lagrange multiplier $\lambda - 1$ and corresponding $\epsilon > 0$.

$$\min_{\theta,\phi} \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] - (\lambda - 1)\left(D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right) - \epsilon\right), \tag{3.37}$$

$$\implies \min_{\theta,\phi} \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] \text{ subject to: } D_{\mathrm{KL}}\left(q(z)\|\prod_{i=1}^{d} q(z_i)\right) < \epsilon. \tag{3.38}$$

Caveats

We should emphasize that while the constraints imposed on the variational posterior $q_\phi(z|x)$ through the different proposed augmentations to the ELBO intuitively appear to facilitate a disentangled representation, there is no reason why the dimensions of the learnt representation should correspond or be axis-aligned with the true factors of variation of the dataset. Locatello et al. (2018) demonstrate that in the unsupervised case where the true factors of variation are not provided or unknown, learning disentangled representations is fundamentally impossible without inductive biases on the model or data, requiring supervision-based hyperparameter tuning to be reliably achieved. In practice unsupervised disentanglement learning is still possible in practice as we can restrict our model class to exploit structure in real-world data (e.g. restricting the neural architecture to convolution-based networks to exploit locality in natural images).

3.2.3 Introducing Supervision

The previous discussion has focused on the unsupervised learning of disentangled representations without access to the ground truth factors of variation v. Assuming we have access to v, we may augment any of the proposed objectives in Section 3.2.2 with some supervised penalty to encourage individual latent dimensions to correspond to the labelled factors. Separating out the objective into unsupervised $R_u(q_\phi(z|x))$ and supervised terms $R_s\left(q_\phi(z|x), v\right)$, with respective weights $\lambda_u > 0$, $\lambda_s > 0$:

$$\max_{\theta,\phi} \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]\right] + \lambda_u \mathbf{E}_{p(x)}\left[R_u\left(q_\phi(z|x)\right)\right] + \lambda_s \mathbf{E}_{p(x)p(v|x)}\left[R_s\left(q_\phi(z|x), v\right)\right]. \tag{3.39}$$

Where R_u is intended to represent any of the above unsupervised objective functions, modulo the log-likelihood term. Many different choices for the supervised penalty R_s are possible, however in this work we consider a function reminiscent of the binary cross-entropy objective, proposed in Locatello et al. (2019b):

$$R_s(q_\phi(z|x), v) = \sum_{i=1}^{\min(D_z, |v|)} \left(v_i \log \sigma(r(x)_i) + (1 - v_i) \log \left(1 - \sigma(r(x)_i)\right)\right). \tag{3.40}$$

Where we assume no natural ordering of the latent dimensions. If a natural ordering is desirable, the sum indices in Eq. 3.40 should be set accordingly. Here $r(x)$ is chosen to be the mean of the variational posterior $q_\phi(z|x)$, $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function, the factor v is normalized to the range $[0, 1]$ and the sum runs over the dimensions of the latent representation. This has no natural interpretation when the factor v is not a discrete binary random variable, and thus the cross-entropy loss does not correspond to maximum likelihood estimation, but performs well empirically. Introducing supervision may not be possible in most real-world datasets where the true factors of variation are generally unknown. However, for applications such as learning fair representations with respect to known sensitive attributes of the input data (Dwork et al., 2012), or generative modelling for fast simulation, ground truth factors are available, allowing supervision to be applied.

3.2.4 Quantifying Disentanglement

Several different metrics have been proposed ((Higgins et al., 2017), (Kim and Mnih, 2018), (Chen et al., 2018), (Locatello et al., 2018)) to quantify the disentanglement of latent representations. In this work we adopt the Mutual Information Gap (MIG) proposed by Chen et al. (2018) based on estimating the empirical mutual information between the latent representation z and ground truth factors v. The empirical mutual information between each dimension of the latent representation z_j and each ground truth factor is estimated, then the MIG is taken to be the average, normalized difference in mutual information between the latent variables with the highest and second highest mutual information with each ground truth factor v_k:

$$\text{MIG}(z, v) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{H(v_k)} \left(I(z_{j^{(k)}}; v_k) - \max_{j \neq j^{(k)}} I(z_j; v_k)\right). \tag{3.41}$$

Where $j^{(k)} = \text{argmax}_j \, I(z_j; v_k)$, the index of the latent dimension with the highest mutual information with ground truth factor v_k. As the mutual information can be expressed as the difference between the marginal and conditional entropies, $I(z_j; v_k) = H(v_k) - H(v_k|z_j) \leq H(v_k)$ (where the last inequality only holds if z, v are quantized), the mutual information between each dimension z_j and each ground truth factor is normalized by the entropy of the ground truth factor $H(v_k)$. Chen et al. (2018) estimate the MIG using a Monte Carlo estimate obtained through importance sampling. However, their procedure requires knowledge of the true generative process $p(x|v)$, which is not possible to obtain exactly in general even if the ground truth factors v are known. It is possible to use their proposed procedure to calculate MIG on synthetic datasets ((Matthey et al., 2017), (Burgess and Kim, 2018)) with uniformly distributed factors of variation and with known $p(x|v)$. For cases where this is not possible, we quantize continuous factors of variation and each dimension of the latent representation obtained from 10000 samples from the variational posterior $q_\phi(z|x)$ into 24 bins[6] and estimate the discrete mutual information to calculate the MIG.

The MIG metric is capable of detecting axis-alignment of the latent representation with ground truth factors of variation by construction - if the dimensions of the latent representation are rotated with respect to the true factors, each latent dimension may contain information about multiple factors, which manifests itself in a low gap. By similar logic, the MIG is able to detect representations which violate modularity or compactness (Section 3.2.1). Furthermore, it does not rely on an auxiliary classifier to compute, and is unbiased as it does not require setting of auxiliary hyperparameters, unlike the metrics proposed in ((Higgins et al., 2017), (Kim and Mnih, 2018)). The disadvantage is that it requires approximation of an intractable quantity through Monte Carlo sampling, which is computationally expensive to evaluate for large datasets, and that it penalizes representations which may be disentangled but violate compactness.

3.3 Background / Related Work

[6]The discrete mutual information computed this way is fairly insensitive to the number of bins for the datasets we consider, or increasing the number of samples from $q_\phi(z|x)$ above 10000.

3.3.1 Disentangled Representation Learning

Disentangled representation learning is closely related to the field of dimensionality reduction/data compression. Methods for dimensionality reduction are concerned with learning a lower-dimensional representation of the observed data $x \in \mathcal{X}$ that is able to capture the most important factors of variation in x, modulo some additional criteria. Formally, for N samples from the observed data $\{x_n\}_{n=1}^N$, each of dimension D, the observed data can be represented as a data matrix $\boldsymbol{X} \in \mathbb{R}^{D \times N}$, where the ith column of $\boldsymbol{X}$ corresponds to the instance x_i. Methods for dimensionality reduction aim to find a low-rank approximation $\boldsymbol{Z} \in \mathbb{R}^{k \times D}$ to the original data matrix X with $k \ll N$ that preserves certain important properties of the data - where the exact properties depend on the method. We briefly review a popular subset of dimensionality reduction approaches below and discuss their relation to VAEs and disentanglement.

Autoencoders Autoencoder models attempt to reconstruct observed data from some lower-dimensional learned encoding. Autoencoder models consist of an encoder $E : \mathcal{X} \to \mathcal{Z}$ and an decoder $D : \mathcal{Z} \to \mathcal{X}$, typically implemented as shallow neural networks. The encoder learns a mapping from the original data space $\mathcal{X}$ to the lower dimensional latent space $\mathcal{Z}$, and the decoder learns a mapping in the opposite direction such that the reconstructed output is produced as $\hat{x} = D(E(x))$. Both encoder and decoder are jointly trained to minimize an objective function

$$\min_{D,E} R(x, D(E(x))) + \Omega(z). \tag{3.42}$$

consisting of the reconstruction error $R(x, \hat{x})$, which rewards fidelity of the output plus some regularization term on the learnt latent code $z = E(x)$. It is not useful for the decoder to exactly implement the inverse mapping of the decoder - the regularization term $\Omega(z)$ encourages desirable structure in the latent space, such as sparsity, robustness against noise corruption and the Lipschitz property (Goodfellow et al., 2016), Ch. 14. The ELBO objective used in VAEs (Eq. 3.25) has the same superficial form, with the expected log-likelihood playing the role of the reconstruction term and the KL divergence between the variational posterior and prior as the regularization term. However, it should be reiterated that variational autoencoders are fundamentally a scalable method to perform variational inference and should not be considered a particular class of autoencoder, despite the similarity of both objectives. For autoencoders the lower dimensional representation corresponds to the encoding $\mathbf{Z} = E(\mathbf{X})$.

Principal Component Analysis (PCA) PCA projects the D-dimensional input data onto the k-dimensional linear space minimizing the reconstruction error - taken to be the Euclidean norm between the original (mean-centered) data matrix $\boldsymbol{X}$ and the projected data matrix. That is, the objective is to find the orthogonal projection matrix $\boldsymbol{P}^*$ among the set of N-dimensional rank-k projection matrices $\mathcal{P}_k$ minimizing:

$$\min_{\boldsymbol{P} \in \mathcal{P}_k} \|\boldsymbol{P}\boldsymbol{X} - \boldsymbol{X}\|^2. \tag{3.43}$$

The optimal solution of this objective corresponds to projection of each data point onto the k eigenvectors u_i with largest associated eigenvalues λ_i of the sample covariance matrix $\boldsymbol{C} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T$. By definition of the covariance matrix, the eigenvectors of $\boldsymbol{C}$ with greatest eigenvalue correspond to directions of highest variance in the data. Hence PCA also corresponds to linear projection of the data into the subspace that maximizes the variance in the projected space. The lower dimensional representation obtained from PCA are the principal components $\mathbf{Z} = \mathbf{U}_k^T\mathbf{X}$, where $\mathbf{U}_k$ is the matrix formed by the top k eigenvectors of $\mathbf{C}$. By the Spectral Theorem for real, symmetric matrices, the eigenvectors of the covariance matrix u_i form an orthonormal basis of the space spanned by the eigenvectors. Hence the principal components of the data will be uncorrelated and in some sense form a minimal representation of the data that captures the maximum variance possible. However, there is no guarantee that the axes of maximal variance will correspond to the true generative factors of the data - the covariance between two random variables only measures the linear dependency between the two. If the latent factors have a nonlinear relationship with the observed data, PCA will fail to recover the true underlying factors of variation of the dataset. Note that the optimal solution for a linear autoencoder - i.e. an autoencoder in which the encoder and decoder are implemented as projection matrices, recovers the same subspace as the one spanned by the principal components (Bourlard and Kamp, 1987).

Factor Analysis Factor analysis is a latent variable model that assumes a standard Gaussian prior over the latent variables $p(z) = \mathcal{N}(z; \mathbf{0}, \mathbb{1})$. The generative model used in factor analysis assumes the observed data x is conditionally independent given the values of the latent variables z, and that there is a linear relationship between the latent factors and observed data:

$$p(x|z) = \mathcal{N}(x; \mathbf{W}z + \mu, \Psi). \tag{3.44}$$

Where the covariance Ψ is a diagonal matrix. As Gaussian distributions are self-conjugate, marginalization of latent variables can be done in closed form:

$$p(x) = \int dz\, p(x|z)p(z) \tag{3.45}$$

$$= \mathcal{N}\left(x; \mu, \mathbf{W}\mathbf{W}^T + \Psi\right). \tag{3.46}$$

The latent factors z explain the dependencies between the observed data variables through the $\mathbf{W}\mathbf{W}^T$ term in the covariance of the marginal $p(x)$. while the additive diagonal-covariance Gaussian noise represents the individual variation of each latent factor. The generative model $p_\theta(x|z)$ used in variational autoencoders can be considered a generalization of factor analysis where the distribution parameters of $p_\theta(x|z)$ are a nonlinear function of the latent code z (so marginalization of latents cannot be done exactly), implemented through a neural network. Furthermore, the variational posterior learnt under VAEs allows us to sample more likely latent variables conditional on the observed data rather than sampling from a simple Gaussian prior. Note that the choice of distribution for the prior, inference and generative models in VAEs are not restricted to be Gaussian.

Probabilistic Principal Component Analysis (pPCA) pPCA (Tipping and Bishop, 1999) is a special case of factor analysis where the covariance of the generative model is restricted to be isotropic, $\Psi = \sigma^2\mathbb{1}$, hence $p(x) = \mathcal{N}\left(x; \mu, \mathbf{W}\mathbf{W}^T + \sigma\mathbb{1}\right)$. They show that the principal axes of the data $\mathbf{X}$ can be recovered from the columns of the weight matrix $\mathbf{W}$ under the maximum likelihood solution. As $\sigma \to 0$ the expected value of the posterior $p(z|x)$ is the orthogonal projection of $x - \mu$ onto the space spanned by the top eigenvectors of the sample covariance matrix, reducing to the PCA representation. Lucas et al. (2019) show that the global optimum for a linear VAE (where the encoder and decoder are modelled as projection matrices) corresponds to the maximum likelihood solution for pPCA, a parallel to the analogous result for linear autoencoders and PCA.

3.3.2 Invariant Representations

Real-world datasets may contain 'sensitive' or 'nuisance' attributes that downstream tasks should not condition upon. For instance, when determining identity from facial portraits, we would like our model to make decisions which are independent of the intensity of scene lighting, facial elevation, facial azimuth, etc. Therefore, we would like to learn intermediate representations of the original data that are invariant to such sensitive factors which can be subsequently used in downstream tasks. We briefly outline two major applications of invariant representations below.

Applications to High-Energy Physics

Learning invariant representations is an important concept in experimental high-energy physics, as described in the previous chapter. In measurements of physical observables associated with some signal process of interest, it is necessary to measure the signal yield. This is usually obtained through a parametric maximum likelihood fit to the spectrum of one or more physically meaningful 'fit-variables' capable of distinguishing between events originating from signal processes of interest and background processes of no interest. By design, the spectrum of fit-variables exhibit strong dichotomy between signal and background events - encouraging supervised learning algorithms for classification to condition their output on these fit-variables. After classification and selection of signal events conditional on the output of the supervised learner, this tends to result in significantly higher acceptance of events in signal-enriched regions of the spectrum. In many cases, this results in distortion of the false-positive background spectrum to mimic structure in the signal spectrum - an example of this is shown in Figure B.6. The resulting bias in the spectrum of fit-variables can lose the physical interpretation of the original structure and introduce systematic uncertainties to experimental analyses that are not easily characterized. Several approaches to learning invariant representations have been proposed by the experimental high-energy physics community ((Louppe et al., 2017), (Shimmin et al., 2017)), in addition to the approach proposed in the previous chapter. These approaches are based on an adversarial-type minimax training strategy in which an auxiliary discriminator is trained in an inner optimization loop to minimize (maximize) some term in the objective which the base classifier attempts to maximize (minimize) and are described in more detail in Section 2.1.1.

Algorithmic Fairness

The field of algorithmic fairness ((Dwork et al., 2012), (Zemel et al., 2013)) is dependent upon the concept of invariant representations. In algorithmic fairness, the objective is to prevent automated decision systems from discriminating against certain protected demographics. For instance, it is undesirable for systems for insurance rate determination or

police allocation to exhibit bias in their output against demographic subgroups of the population. This is challenging as systems trained on past data are likely to inherit historical bias. As Zemel et al. (2013) state, to overcome this, "the automated decision-maker [should be] blind to some attributes."

Invariant Representations and VAEs

Using a factorized family of distributions $\mathcal{Q}_f$ (Eq. 3.21) to model the variational posterior $q_\phi(z|x)$ in VAEs ostensibly leads naturally to learning invariant representations. However, factorization of the per-datapoint posterior does not guarantee factorization of the aggregate structure of latent variables under the data distribution, $q(z) = \sum_n q_\phi(z|x_n)p(x_n)$, as discussed in Section 3.2.2. Methods for learning invariant representations using VAEs are centered around enforcing statistical independence between the sensitive factors s and the latent representation z. Louizos et al. (2015) use a regularizer that performs moment-matching by penalizing the maximum mean discrepancy (Gretton et al., 2012) between all possible marginal posterior distributions over latent variables $\{q_\phi(z|s = s_1), q_\phi(z|s = s_2), \ldots, q_\phi(z|s = s_K)\}$ where $(s_1, s_2, \ldots s_K)$ denote the different possible values of the sensitive factors. However, the number of terms in the loss function increases combinatorially with the number K of possible values that s can assume - making it difficult to scale beyond the binary case. As a consequence, the method is unable to handle continuous sensitive factors. (Wang et al., 2018) use the aforementioned adversarial training procedure (Eq. 2.21) to penalize informativeness between the latent representation and the sensitive factors, where $f(x) \sim q_\phi(z|x)$ are taken to be samples from the variational posterior. For disentangled representations, where $q(z)$ is constrained to be approximately factorial; information about each generative factor is isolated in a unique latent subspace. From this, it is natural to produce a disentangled representation by removing the latent subspace corresponding with the sensitive factors. The difficulty lies in enforcing this latent-sensitive correspondence without sacrificing disentanglement of the latent representation. Creager et al. (2019) independently propose to use disentangling VAEs to learn invariant representations but can only handle the case of discrete factors of variation and discrete sensitive factors, and do not model the sensitive factors using latent random variables.

Placing this work into context, we observe that VAEs may be viewed as a probabilistic, nonlinear dimensionality reduction technique distinct from previous latent variable models. Unlike factor analysis and related techniques, which assume a linear relationship between latent factors and the observed data, latent representations learnt under VAEs can model more complex relationships between the dimensions of the observed data. Unlike autoencoders, the probabilistic formulation of variational autoencoders allows us to build domain knowledge into our model through the prior on the latent space, provides an inherent measure of uncertainty when making decisions based on the latent factors, and allows us to use effective information-theoretic arguments to constrain the structure of the learned latent space. Disentanglement in the learned representation arises from enforcing a factorial structure in the variational posterior $q_\phi(z|x)$, as discussed in Section 3.2.2. By encouraging independence between the latent factors corresponding to sensitive attributes of the input data and the remainder of the latent representation, we can derive representations which are invariant to the sensitive attributes from the variational posterior learnt through the VAE framework, as we discuss in the following section.

3.4 Invariance through Disentanglement

In both the above applications, the challenge lies in learning an encoding of data conducive to success of the downstream task while being robust against changes in the sensitive attributes. Most downstream tasks in the real-world will fall under the category of classification/regression, so the objective is to learn a representation that is informative of the target variables without being dependent on the sensitive variables - i.e. a representation in which the influence of the target variables is disentangled from the influence of the sensitive variables. We propose to learn an intermediate representation derived from the latent code learned under the variational autoencoder framework - i.e. derived from the variational posterior $q_\phi(z|x)$. The basic premise is as follows: to achieve invariance to the sensitive data attributes, we explicitly identify the sensitive factors with the latent variables learned under a latent variable model through supervision (Section 3.2.3). In the case of the variational autoencoder, the latent variables correspond to distinct dimensions of the latent code z. If the latent representation is disentangled with respect to the sensitive factors, the remaining latent variables should not contain any information about the sensitive factors. This enables us to quarantine the influence of the sensitive data attributes within a subspace of the latent representation, which we remove to obtain the intermediate representation for subsequent use in downstream tasks. Let us use the following notation to be consistent with the literature on disentanglement:

- Let $x \in \mathbb{R}^{D_x}$ denote all input variables.

- Let $z \in \mathbb{R}^{D_z}$ denote all latent variables.
- Let $s \in \mathbb{R}^{D_s}$ denote the sensitive factors which we would like our learned representation to be invariant to.
- Let $t \in \mathbb{R}^{D_t}$ denote the latent variables corresponding to s, with $t \subseteq z$. Let the set S denote the indices of the dimensions of t, the latent variables corresponding the the sensitive inputs s. Without loss of generality, we can assume they run from $\{1, \ldots, D_t\}$.

3.4.1 Representation Definition

Louppe et al. (2017) propose that the representation $r(x)$ and the sensitive, possibly multidimensional variable s should be independent random variables for $r(x)$ to be considered an invariant representation against s. In this context, it is natural to let $r(x)$ be a function of the variational posterior $q_\phi(z|x)$. If $q_\phi(z|x)$ is modelled as a factorized distribution (commonly taken to be a diagonal-covariance Gaussian), each latent dimension is independent by construction for each individual datapoint. If we are able to identify s with distinct latent dimensions t and take the final representation to be sampled from the variational posterior with t omitted:

$$r(x) \sim q_\phi(z \setminus t|x). \tag{3.47}$$

Then the independence criterion between s and $r(x)$ can be satisfied. In practice it is difficult to enforce an exact correspondence between dimensions of the latent representation and s if the true distribution over s lies outside the family of variational distributions considered, and because of the complexity penalty $D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)$ in the ELBO. Furthermore, we find the additional stochasticity from sampling hurts the performance of downstream classification task unacceptably on the datasets considered. As an alternative, we note factor-VAE and β-TCVAE explicitly enforce factorization of the posterior averaged over the data distribution, or *aggregate posterior*:

$$q(z) = \int dx\, q_\phi(z|x)p(x) \tag{3.48}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} q_\phi(z|x_n). \tag{3.49}$$

Through penalization of a differentiable estimate of the total correlation $D_{\mathrm{KL}}\left(q(z)\|\prod_i q(z_i)\right)$ (Section 3.2.2), while β-VAE and annealed-VAE are hypothesized to do so (Burgess et al., 2018) through restricting the capacity of the variational posterior $q_\phi(z|x)$. We make the following observation:

Proposition 2 (*Factorization of the aggregate posterior implies factorization of the mean representation* $r(x) = \int dz\, z q_\phi(z|x)$ *averaged over the data distribution*)**.**

Proof.

$$\int dx\, r(x)p(x) = \int dx dz\, z\, q_\phi(z|x)p(x) \tag{3.50}$$

$$= \int dz\, z \int dx\, q_\phi(z|x)p(x) \tag{3.51}$$

$$= \int dz\, z\, q(z) \tag{3.52}$$

$$= \int dz\, z \prod_{i=1}^{D_z} q(z_i) \tag{3.53}$$

$$= \prod_{i=1}^{D_z} \int dz_i\, z_i\, q(z_i). \tag{3.54}$$

□

This leads us to take the initial base representation $r(x)$ to be the mean of the variational posterior:

$$r(x) = \int dz\, z\, q_\phi(z|x) \tag{3.55}$$

$$\approx \sum_{m=1}^{M} z_m\, q_\phi(z_m|x). \tag{3.56}$$

However, it may not be possible to exactly enforce the restriction $q(z) = \prod_i q(z_i)$, in which case the dimensions of the mean representation $r(x)$ will not be independent in general. This is in agreement with the empirical evidence presented in Section 5 of Locatello et al. (2018) that demonstrate that regularization of the total correlation of the aggregate posterior $q(z)$ does not lead to zero correlation between the dimensions of the mean representation. As previously, we omit the sensitive latent dimensions t from the mean representation, to yield the final invariant representation for use in subsequent downstream tasks.

3.4.2 Latent-sensitive Correspondence

To match latent dimensions t with sensitive data attributes s we use the supervised penalty discussed in Section 3.2.3.

$$R_s(q_\phi(z|x), s) = -\sum_{i=1}^{\min(D_z, |s|)} \left(s_i \log \sigma(t_i) + (1 - s_i) \log(1 - \sigma(t_i))\right). \tag{3.57}$$

With the sensitive factors s normalized to the range $[0, 1]$. This is a suboptimal way of enforcing correspondence of sensitive latents t with sensitive factors s as it ignores the probabilistic interpretation of the VAE. Future work may explore more principled forms of distribution matching between the marginals $q(t)$ and $p(s)$. One possible approach is to model $q(t)$ using normalizing flows (Rezende and Mohamed, 2015) and minimize the KL divergence between $q(t)$ and target density $p(s)$.

3.4.3 General Objective

To learn the representation $r(x)$, we optimize the following general objective:

$$\max_{\theta,\phi} \mathbf{E}_{p(x)} \left[\mathbf{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] + R_u\left(q_\phi(z|x)\right) + \lambda_s \mathbf{E}_{p(s|x)} \left[R_s\left(q_\phi(z|x), s\right)\right]\right]. \tag{3.58}$$

The first term, the expected log-likelihood, encourages informativeness of the latent representation with the input x. The second term denotes any of the unsupervised extensions to the ELBO objective discussed in Section 3.2.2 used to achieve disentanglement of the latent representation. The third enforces correspondence between sensitive factors s and selected, distinct dimensions of the latent representation t. The intermediate representation is taken to be the mean of the variational posterior learned through optimization of this objective, with t omitted. Note that approaches to disentanglement which reward factorization of the aggregate posterior, $q(z) \approx \prod_i q(z_i)$ do not discriminate between sensitive latents t and non-sensitive latents $z \setminus t$. For downstream tasks which involving prediction of a target variable from the intermediate representation that is invariant to the sensitive factors s, it is not neceessary for all latent dimensions to be disentangled, only the latent dimensions corresponding to s. Below we propose an objective based on an extension of β-TCVAE where the learning procedure specializes in the disentanglement of a subset of the latent dimensions t and allows the remaining latent dimensions $z \setminus t$ to be maximally informative of the observed data.

3.4.4 Optimizing for Disentanglement of Sensitive Factors

To begin, we wish to decompose the KL divergence $D_{\mathrm{KL}}\left(q(z|x)\|p(z)\right)$ in the ELBO into tractable/interpretable terms. This proof follows similar logic as the proof in Appendix C.1 of Chen et al. (2018).

$$
\begin{aligned}
&\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)\right] = \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log p(z)\right]\right] \\
&= \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log q_\phi(z|x) - \log p(z) + \log q(z) - \log q(z) + \log q(z\setminus t)\prod_{j\in S} q(t_j) - \log q(z\setminus t)\prod_{j\in S} q(t_j)\right]\right] \\
&= \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z|x)}{q(z)}\right] + \mathbf{E}_{q(z)}\left[\log \frac{q(z)}{q(z\setminus t)\prod_{j\in S} q(t_j)}\right] + \mathbf{E}_{q_\phi(z|x)}\left[\log q(z\setminus t) + \sum_{j\in S}\log q(t_j) - \log p(z)\right]\right] \\
&= \underbrace{\mathbf{E}_{q(x,z)}\left[\log \frac{q_\phi(z|x)p(x)}{q(z)p(x)}\right]}_{\text{I}} + \underbrace{\mathbf{E}_{q(z)}\left[\log \frac{q(z)}{q(z\setminus t)\prod_{j\in S} q(t_j)}\right]}_{\text{II}} + \underbrace{D_{\mathrm{KL}}\left(q(z\setminus t)\|p(z\setminus t)\right)}_{\text{III}} + \underbrace{\sum_{j\in S} D_{\mathrm{KL}}\left(q(t_j)\|p(t_j)\right)}_{\text{IV}}.
\end{aligned}
$$

To see the origin of the last equality, note that, as $z \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$, we can exploit the isotropy of the prior:

$$
\begin{aligned}
&\mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log q(z\setminus t) + \sum_{j\in S}\log q(t_j) - \log p(z)\right]\right] && (3.59)\\
&= \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log q(z\setminus t) - \log p(z\setminus t) + \sum_{j\in S}\left(\log q(t_j) - \log p(t_j)\right)\right]\right] && (3.60)\\
&= \mathbf{E}_{q(z)}\left[\log \frac{q(z\setminus t)}{p(z\setminus t)} + \sum_{j\in S}\log \frac{q(t_j)}{p(t_j)}\right] && (3.61)\\
&= \mathbf{E}_{q(z\setminus t)q(t|z\setminus t)}\left[\log \frac{q(z\setminus t)}{p(z\setminus t)}\right] + \sum_{j\in S}\mathbf{E}_{q(z)}\left[\log \frac{q(t_j)}{p(t_j)}\right] && (3.62)\\
&= \mathbf{E}_{q(z\setminus t)}\left[\log \frac{q(z\setminus t)}{p(z\setminus t)}\right] + \sum_{j\in S}\mathbf{E}_{q(t_j)q(z\setminus t_j|t_j)}\left[\log \frac{q(t_j)}{p(t_j)}\right] && (3.63)\\
&= \mathbf{E}_{q(z\setminus t)}\left[\log \frac{q(z\setminus t)}{p(z\setminus t)}\right] + \sum_{j\in S}\mathbf{E}_{q(t_j)}\left[\log \frac{q(t_j)}{p(t_j)}\right] && (3.64)\\
&= D_{\mathrm{KL}}\left(q(z\setminus t)\|p(z\setminus t)\right) + \sum_{j\in S} D_{\mathrm{KL}}\left(q(t_j)\|p(t_j)\right). && (3.65)
\end{aligned}
$$

Interpreting each term in the decomposed KL divergence, $D_{\mathrm{KL}}\left(q_\phi(z|x)\right)\|p(z))$:

1. $I_q(x;z)$, the mutual information between x and z under the joint distribution $q_\phi(z|x)p(x)$. Ostensibly this enforces informativeness between the observed data and the latent variables and is required for z to be a useful representation of the original data x. However compression of x through this information bottleneck-style term is thought to lead to disentangled representations (Achille and Soatto, 2018). We note that if there is no requirement for the latents $z \setminus t$ to be disentangled, we can restrict this term to only include the latent dimensions t in lieu of the mutual information $I_q(x;z)$ over the full set of latents:
$$
I_q(x;t) = \mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\phi(t|x)\|q(t)\right)\right]. \tag{3.66}
$$
2. The KL divergence between $q(z)$ and the product of the marginals $q(z\setminus t)\prod_{j\in S} q(t_j)$. Penalizing this term encourages independence between the sensitive latent dimensions t as well as independence of t from the remaining latent dimensions $z \setminus t$. This term does not exert any pressure on the non-sensitive latent dimensions $z \setminus t$ to be disentangled between themselves.
3. The KL divergence between the aggregate posterior and factorized prior, excluding sensitive latent dimensions t. This term encourages independence between dimensions of $z \setminus t$, discouraging deviation of $q(z \setminus t)$ from an isotropic Gaussian.

4. The dimension-wise KL divergence between each dimension j of the aggregate posterior over the sensitive latents t and the standard normal prior over $p(z_j)$. This both penalizes complexity in the aggregate posterior over the individual latent dimensions t_j, and encourages the factorization $q(t) \rightarrow \prod_j q(t_j)$.

The terms in this decomposition are computed using the same importance sampling strategy proposed in Chen et al. (2018). We refer to this approach as the β-TCVAE-Sensitive loss in the following sections.

3.4.5 Final Objective

Putting the individual terms together and assigning each a respective weighting, we arrive at the proposed objective,

$$\begin{aligned}\mathscr{L}(\theta, \phi) = \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] + \lambda_s \mathbf{E}_{p(s|x)}\left[R_s\left(q_\phi(z|x), s\right)\right]\right] \\ - \left(\alpha I_q(x; z) + \beta \mathbf{E}_{q(z)}\left[\log \frac{q(z)}{q(z \setminus t)\prod_{j \in S} q(t_j)}\right] + \gamma D_{\mathrm{KL}}\left(q(z \setminus t) \| p(z \setminus t)\right) + \delta \sum_{j \in S} D_{\mathrm{KL}}\left(q(t_j) \| p(t_j)\right)\right),\end{aligned}$$

$$\max_{\theta, \phi} \mathscr{L}(\theta, \phi). \tag{3.67}$$

The first line can be interpreted as the 'reconstruction' terms for both non-sensitive and sensitive observations discussed in Section 3.4.3 - learned latent variables z should be informative of the observed data x. The second line consists of the previously discussed breakdown of $D_{\mathrm{KL}}\left(q_\phi(z|x) \| p(z)\right)$ with respective weightings. We retain the mean of the learned posterior over the non-sensitive latent factors $q_\phi(z \setminus t|x)$ as the learned intermediate representation. Ideally the number of latent factors z should correspond to the number of generative factors v, although this is not generally possible to enforce except in the case of toy datasets. Empirically and in agreement with the experiments conducted in ((Kim and Mnih, 2018), (Chen et al., 2018)), we find that setting the weights of the decomposed KL term identically to one with the exception of the β parameter, works well out-of-the-box and avoids a combinatorial explosion of possible configurations when tuning hyperparameters.

Using the language of Definition 3.2.1, minimizing the β-TCVAE-Sensitive loss results in a partial disentanglement of the latent space. This is useful for the case where only some of the factors of variation of the data are known and supervision can only be applied for this subset. A summary of the procedure for end-to-end learning of invariant representations is given in Algorithm 3.

3.5 Experiments

3.5.1 DSprites Variants

We make use of the DSprites (Matthey et al., 2017) toy dataset as a sanity check of our proposed method. The DSprites dataset consists of 64×64 binary images of 2D polygons with 5 discretized independent generative factors:

- Shape (3)
- Scale (6)
- Rotation (40)
- Cartesian coordinates (32×32)

The dataset includes all possible combinations of these latent factors, for a total of $3 \times 6 \times 40 \times 32 \times 32 = 737280$ images. We also consider the more challenging DSprites-Scream dataset (Locatello et al., 2018). This has the same generative factors as the DSprites dataset, except additional noise is dynamically overlaid during training. At each iteration, a random 64×64 section of The Scream (Munch, 1893) is sampled as the background, randomly colorized, and the sprite is embedded into the image by inverting the pixel colors of the sampled patch at the original sprite location. Example images from both datasets are shown in 3.1.

Recall that the goal of disentangled representation learning is to encode information about each generative factor in independent latent variables. To visualize this, we plot the reconstructed images obtained by traversal of the latent space

Algorithm 3 Invariant Representation Learning through Disentanglement

Procedure to learn a latent-variable representation of the observed data x which is invariant to changes in the sensitive attributes s. The mapping from x to distribution parameters of $q_\phi(z|x)$ and the mapping from z to distribution parameters of $p_\theta(x|z)$ are represented by deep neural networks E_ϕ and D_θ with parameters ϕ and θ respectively. For simplicity, assume that $q_\phi(z|x)$ and $p_\theta(x|z)$ take the form of diagonal-covariance Gaussians, but generalizes to the case of arbitrary (continuous) distributions. Note i indexes latent dimensions and n indexes data examples in the following.

Require: Inference model $q_\phi(z|x)$, generative model $p_\theta(x|z)$, dataset $\mathcal{D} = \{x, s\}$ containing observed data x and sensitive attributes s. Indices S of latent dimensions marked to be associated with s.

1: Initialize the parameters of the encoder network E_ϕ and decoder network D_ϕ.

2: **for** $n = 1$ to N **do**

3: Sample $\{(x_1, s_1), \ldots, (x_B, s_B)\} \sim \mathbb{P}_{XS}$ from the data distribution.

4: Compute distribution parameters of variational posterior $q_\phi(z|x)$:

$$(\mu_\phi, \log \Sigma_\phi) = E_\phi(x_n)$$

5: Sample from variational posterior $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi, \Sigma_\phi)$ through reparameterization:

$$z_n = \mu_\phi + \Sigma_\phi^{1/2} \odot \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$$

6: Compute distribution parameters of generative model $p_\theta(x|z)$:

$$(\mu_\theta, \log \Sigma_\theta) = D_\theta(x_n)$$

7: Compute log-density of data x_n under generative model $p_\theta(z|x)$:

$$\log p_\theta(x_n|z_n) = -\sum_{i=1}^{D_z} \left(\log\left(2\pi \left(\Sigma_\theta\right)_{ii}\right) + \|\Sigma_\theta^{-1}(x_n - \mu_\theta)\|^2 \right)$$

8: Compute supervised regularization term (omitting data index n for readability):

$$R_s(q_\phi(z|x), s) = \sum_{i=1}^{\min(D_z, |s|)} \left(s_i \log \sigma(t_i) + (1 - s_i) \log(1 - \sigma(t_i))\right), \quad t = (\mu_\theta)_S$$

9: Compute unsupervised disentanglement term $R_u(q_\phi(z|x))$ (see Section 3.2.2).

10: Update (ϕ, θ) by ascending the objective:

$$\mathscr{L} = \frac{1}{B} \sum_{n=1}^{B} \log p_\theta(x_n|z_n) + \lambda_s R_s(q_\phi(z_n|x_n), s_n) + R_u(q_\phi(z_n|x_n))$$

$$\theta \leftarrow \theta, +\eta_t \nabla_\theta \mathscr{L}$$
$$\phi \leftarrow \phi, +\eta_t \nabla_\phi \mathscr{L}$$

11: **end for**

12: **return** Trained inference model $q_\phi(z|x)$.

13: To obtain a representation of x approximately invariant to s, compute the mean $z = \langle q_\phi(z|x) \rangle$, and return $z' = z \setminus t$.



Figure 3.1: Example images randomly sampled from the DSprites (left) and DSprites-Scream datasets (right).

Encoder	**Decoder**
Input: $\mathbb{R}^{64\times64\times C}$	Input: $\mathbb{R}^{10}$
4×4 conv., 32 ReLU, stride 2	Dense, 256 ReLU
4×4 conv., 32 ReLU, stride 2	$4 \times 4 \times 64$ ReLU
4×4 conv., 64 ReLU, stride 2	4×4 upconv., 64 ReLU, stride 2
4×4 conv., 64 ReLU, stride 2	4×4 upconv., 32 ReLU, stride 2
Dense, 256	4×4 upconv., 32 ReLU, stride 2
Dense, 10 (μ), Dense, 10 ($\log \Sigma$)	4×4 upconv., C, stride 2

Table 3.1: Convolutional network architectures used to parameterize the variational posterior (encoder) and generative model (decoder) distributions for DSprites experiments. Above C is the number of channels in the input images, conv. denotes the convolution operation, upconv. denotes the transposed convolution operation, and $(k \times k)$ prior to conv. or upconv. denotes the size of the convolutional kernel.

for DSprites (Figure 3.2 and DSprites-Scream (Figure 3.3). We use a factorized (diagonal-covariance) 10-dimensional Gaussian distribution for the variational posterior $q_\phi(z|x)$ and a factorized Bernoulli distribution for the generative model $p_\theta(x|z)$[7]. The architectures of the networks used to parameterize the approximate posterior and generative model distributions are shown in Table 3.1. We train the model for 42 epochs using the Adam optimizer (Kingma and Ba, 2015) with learning rate 5×10^{-4}. To obtain the below latent traversals, we fix all latent dimensions but one in each row. The unfixed latent dimension is then traversed from the 5% quantile to the 95% quantile in evenly spaced steps[8], with the remaining latent dimensions fixed to the mean of the prior. The latent values obtained are then passed through the generative network $p_\theta(x|z)$, which outputs the mean parameter of the Bernoulli distribution over each pixel in the output image. This mean parameter can be interpreted as the pixel intensity, producing a reconstructed image. The rows in each latent traversal are arranged in descending order of the KL divergence between the standard normal and the corresponding dimension of the variational distribution averaged over the dataset.

We observe from Figure 3.2 that the generative factors corresponding to (x, y) position as well as scale are able to be easily modelled by the top three latent dimensions for the DSprites dataset. Encoding information about rotation and shape is more difficult as changes in these factors result in a lower pixel-wise difference compared to position and scale, but the fourth and fifth latent traversals appear to indicate that the corresponding dimension captures some information about rotation and/or shape. Turning to Figure 3.3, when noise is introduced, it is harder for the latent variables to model the true generative factors of the data, with the latent dimensions wasting capacity modelling noise factors such as color and background. However, both β-TCVAE-Sensitive and Factor-VAE are able to model the (x, y) position and scale factors well (dimensions $(3, 4, 9)$ for β-TCVAE-Sensitive and dimensions $(1, 2, 6)$ for Factor-VAE). Traversals of latent dimensions with low average KL divergence from the prior $p(z)$ (shown in the bottom rows of Figures 3.2 and 3.3) do not affect the reconstructed image. In this sense they are inactive - the remaining active latent dimensions encode all necessary information about the data.

Unbiased Classification

We consider a toy example derived from the DSprites-Scream dataset. Here, we attempt to differentiate square sprites from elliptical and heart-shaped sprites without using information about the (x, y) coordinates of the sprites. Note that this classification task is fairly challenging as the image color and background texture are overlaid randomly dynamically during training. Example instances of training data are shown in Figure 3.4. Each factor of variation of the DSprites-Scream dataset is uniformly distributed across the corresponding domain. In particular, sprite coordinates are distributed uniformly in the unit square, so first we introduce an artificial sampling bias in which the probability of

[7] Here, an image is modelled as the product of independent Bernoulli distributions over each pixel.

[8] For the standard Gaussian prior, this corresponds to a latent traversal range from $\Phi^{-1}(0.05)$ to $\Phi^{-1}(0.95)$, where Φ^{-1} denotes the inverse of the standard Gaussian CDF.

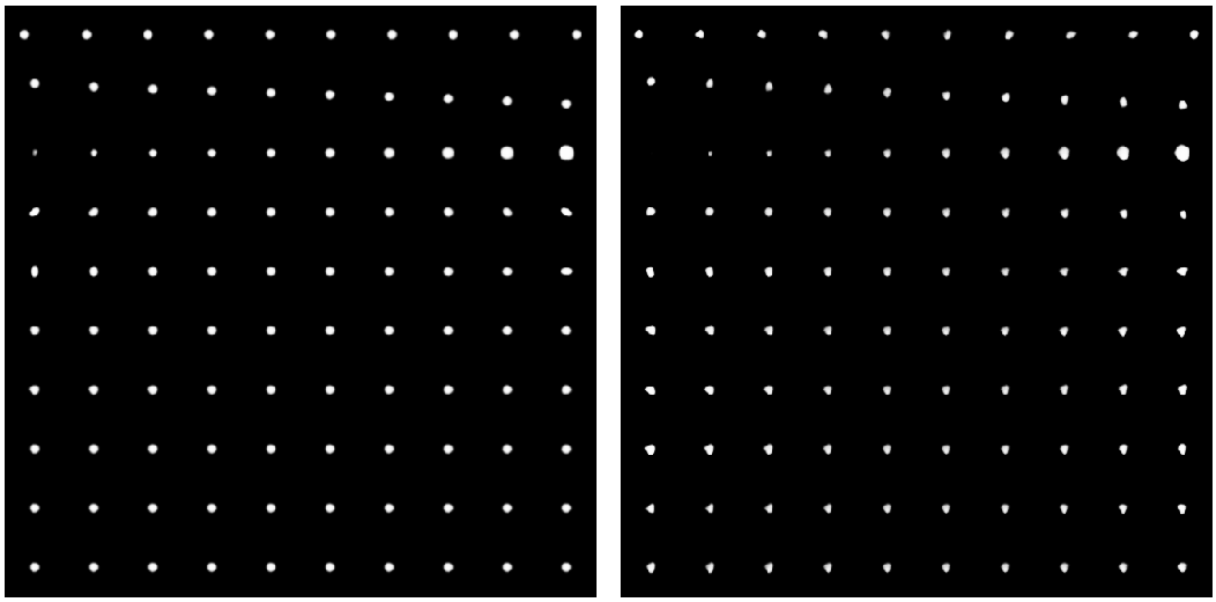


Figure 3.2: Latent traversals for DSprites-Scream obtained via β-TCVAE-Sensitive (left) and Factor-VAE (right). Each row corresponds to linear traversal of a latent dimension from $\Phi^{-1}(0.05)$ to $\Phi^{-1}(0.95)$ with the remaining latent dimensions held constant to the mean of the prior, where Φ is the inverse of the standard Gaussian CDF. Rows are ordered by the magnitude of the KL divergence between the standard normal and corresponding dimension of the variational distribution averaged over the dataset. We observe that the latent dimensions are able to capture information about position and scale well, while modelling rotation and shape is more difficult.

sampling an image containing a square sprite at position (x, y) is given by a Gaussian distribution with mean located at the center of the unit square with isotropic covariance.

$$(x, y) = \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix} \middle| \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \frac{1}{25}\mathbb{1}\right). \tag{3.68}$$

The corresponding probability density function is shown in Figure 3.4. Now the distribution of images with square sprites over Cartesian space is very sharply peaked at $(0.5, 0.5)$, and it is *a priori* more probable that square sprites have position close to the center of the unit square. The (x, y) distribution of elliptical and heart-shaped sprites remain uniform across the unit square. We would like the classifier responsible for separating square from non-square sprites to be insensitive to the sprite coordinates. To achieve this, we optimize the β-TCVAE-Sensitive objective (Eq. 3.4.5), summarized in Algorithm 3, with a 10-dimensional latent space, using the same encoder and decoder architectures and optimization procedure as detailed above. The (x, y) factors of variation are matched with the first and second latent dimensions using the supervised penalty (Eq. 3.40). The mean of the variational posterior, $\langle q_\phi(z|x)\rangle \in \mathbb{R}^{10}$, is taken to be the representation of the image x.

For the downstream classification task, we use a shallow neural network with two hidden layers with 64 units each and ReLU activation, trained using the binary cross-entropy loss and the Adam optimizer (Kingma and Ba, 2015) for 8 epochs. The classifier output is a number between 0 and 1 that can be interpreted as the probability that a given image contains a square sprite. We visualize the decision surface of this classifier in the Cartesian plane in Figure 3.5 for two cases: for the first case, we use the full latent representation $\langle q_\phi(z|x)\rangle \in \mathbb{R}^{10}$ as input to the classifier. For the second, we excise the $[1, 2]$ latent dimensions corresponding to the positional factors of variation and use the resulting representation $\langle q_\phi(z \setminus t|x)\rangle \in \mathbb{R}^8$ as input to the classifier. Averaged over 3 independent runs (using 3 independently trained inference models q_ϕ), we obtain an accuracy of 0.777 ± 0.005 when trained on the full latent representation and an accuracy of 0.756 ± 0.008 when trained on the reduced latent representation, showing that removal of bias from downstream classification algorithms arrives at the expense of reduced discrimination power.

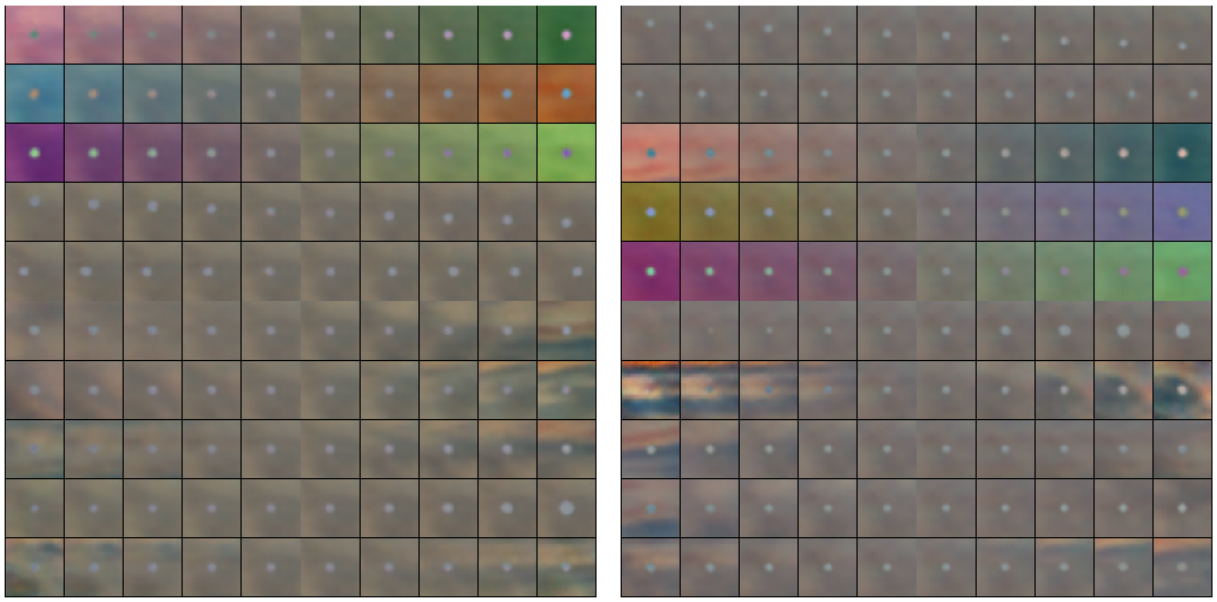


Figure 3.3: Latent traversals for DSprites-Scream obtained via β-TCVAE-Sensitive (left) and Factor-VAE (right). Each row correpsonds to linear traversal of a latent dimension from $\Phi^{-1}(0.05)$ to $\Phi^{-1}(0.95)$ with the remaining dimensions held constant to the mean of the prior, where Φ is the inverse of the standard Gaussian CDF. Rows are ordered by the magnitude of the KL divergence between the standard normal and the corresponding dimension of the variational distribution averaged over the dataset. Compared to the DSprites dataset, when noise is overlaid, it is harder for the latent variables to model the true generative factors of the data, with the latent dimensions wasting capacity modelling noise factors such as color and background. However, both β-TCVAE-Sensitive and Factor-VAE are able to model the (x, y) position and scale factors well (dimensions $(3, 4, 9)$ for β-TCVAE-Sensitive and dimensions $(1, 2, 6)$ for Factor-VAE).

3.5.2 Application to High Energy Physics

We also consider the same high-energy physics example introduced in Chapter 2, where the challenge is to learn an intermediate representation that is capable of:

- **a**) Reliable discrimination between penguin $b \to s\gamma$ (signal) processes from non-interesting continuum and $B\bar{B}$ events (background).
- **b**) Preservation of the distribution of the beam-constrained mass (M_{bc}) spectrum of both classes under selection of a sample of signal-enriched events with signal-background purity ratio of up to $1000 : 1$. Recall that the beam-constrained mass is defined as the invariant mass of the reconstruct B meson candidate, with the energy constrained to the beam energy: $M_{bc}^2 = E_{\text{beam}}^2 - |p|^2$.

To fulfill the second criterion, we would like the representation to be invariant to M_{bc}, which assumes the role of the sensitive variable s in the general objective (Eq. 3.58). The dataset considered here is a subset of the one discussed previously, as using a factorized Gaussian distribution for both the inference $q_\phi(z|x)$ and generative $p_\theta(x|z)$ models means we are unable to model discrete sensitive factors or latent variables. We discuss potential extensions to the model to overcome this restriction in Section 3.5.3. The observed data x is restricted to 34 features from the original dataset with smooth continuous marginal distributions, which are described in Appendix B.3. No variables with Pearson correlation coefficient with M_{bc} above 0.1 is included in the dataset. As previously, the features are scaled to zero mean and unit variance. The size of the dataset considered is restricted to approximately 3 million events for training, and approximately 2 million events in the test and validation sets each as the training process is more computationally demanding than the approach outlined in Chapter 2.2. We train our models for 42 epochs using the Adam optimizer (Kingma and Ba, 2015) with a base learning rate of 5×10^{-4}. If the objective function has not improved for 3 epochs, we reduce the learning rate by a factor of $1/2$. We note that the results appear relatively insensitive to the choice of starting learning rate in the range $[10^{-4}, 10^{-3}]$. We use a batch size of 4096 to reduce the variance of the Monte Carlo estimator of $\log q(z)$ and $\log \prod_i q(z_i)$ proposed by Chen et al. (2018).

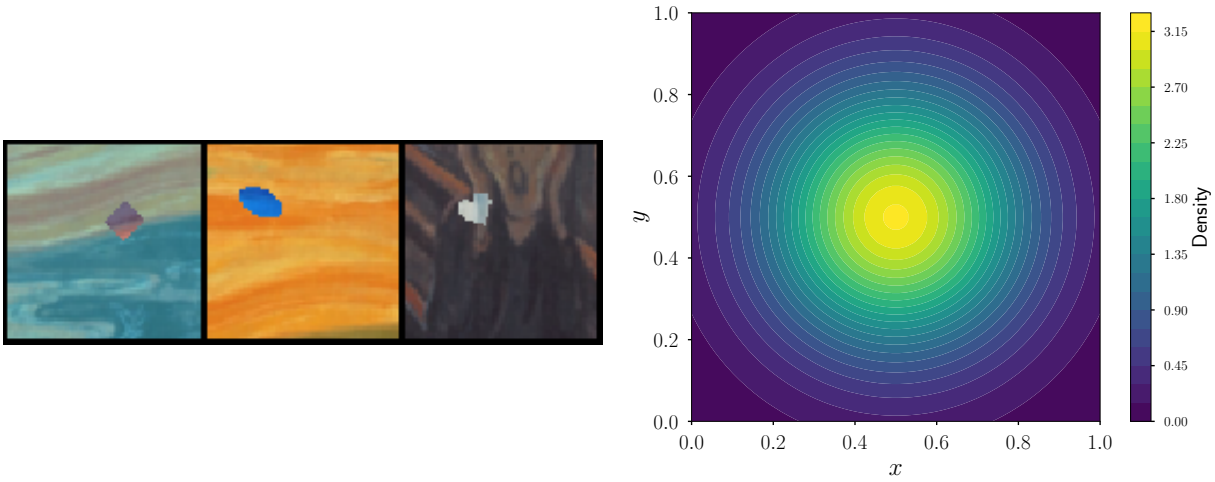


Figure 3.4: **Left:** Example instances from DSprites-Scream dataset for square, elliptical and heart-shaped sprites, from right to left. **Right:** Contour plot of the sampling distribution over Cartesian coordinates (Eq. 3.68) used for square sprites.

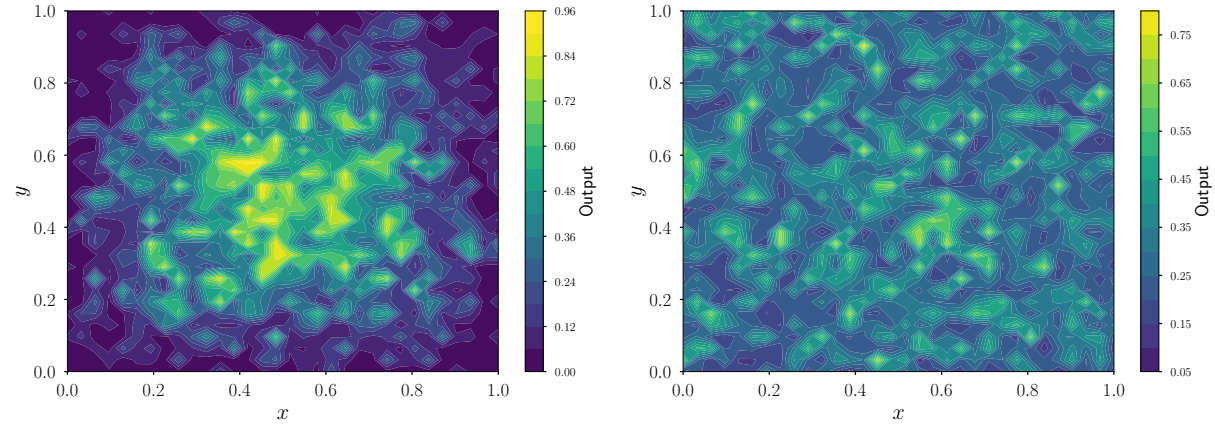


Figure 3.5: Decision surfaces of downstream classifier trained on the DSprites-Scream dataset with sampling bias, in the Cartesian plane. **Left:** Decision surface of classifier trained on the full latent representation $\langle q_\phi(z|x)\rangle$, exhibiting the same sampling bias present in the dataset - the classifier is more likely to predict that images with sprite coordinates close to the center of the unit square contain a square sprite, due to imbalance in the original dataet. **Right:** Decision surface of classifier trained on the latent representation $\langle q_\phi(z \setminus t|x)\rangle$ with information about the coordinates removed. The classifier response is uniform across the Cartesian plane and does not exhibit any bias in coordinate space.

Evaluation

We optimize the parameters of the inference and generative models on the training set. For evaluation we freeze the parameters of the inference model and take the mean of the variational posterior $\langle q_\phi(z|x)\rangle$ on new data as the intermediate representation for the subsequent classification task, with the latent dimensions t corresponding to the sensitive factors removed. We use a simple densely connected neural network for the downstream binary classification task with four hidden layers with 256 units each and ReLU activations at each layer, trained using the binary cross entropy loss to distinguish between signal and background events:

$$\mathcal{L}_C = \frac{1}{B}\sum_{n=1}^{B} y_n \log p_n + (1-y_n)\log(1-p_n)\,. \tag{3.69}$$

Where B is the batch size, $y_n \in \{0, 1\}$ denote the ground truth labels (0 for background and 1 for signal) for instance x_n, and $p \in [0, 1]$ denotes the normalized response of the classifier. The downstream network is trained using the Adam optimizer (Kingma and Ba, 2015) for 16 epochs. We will refer to as the downstream classifier/discriminator for brevity. We consider three metrics for this task:

- The MIG metric (Chen et al., 2018) discussed in section 3.2.4. If we only have a single sensitive factor s, which we take to be the beam-constrained mass (M_{bc}), then the MIG reduces to the difference in mutual information between s and the latent dimension with the highest and second highest mutual information with s, normalized by the entropy of s. A high MIG indicates that the latent subspace t is disentangled from the remaining latent dimensions and the representation is compact (Section 3.2.1).
- To directly quantify the extent of sculpting of the mass distribution we use an estimate of the Jensen-Shannon divergence between the mass distribution of background events passing and failing signal event selection for a given level of background rejection (ATLAS Collaboration, 2018). We use a working point of rejection of 0.995 of background events. While this rejection requirement appears arbitrary, the rejection level lies within the realistic order of magnitude required by analyses of rare decays, given the high beam background at modern colliders. Recall the Jensen-Shannon divergence (JSD) between densities $p(x), q(x)$ can be considered a symmetrized form of the KL divergence.

$$D_{\text{JS}}\left[p(x)||q(x)\right] \triangleq \frac{1}{2}\left(D_{\text{KL}}\left[p(x)||M(x)\right] + D_{\text{KL}}\left[q(x)||M(x)\right]\right). \tag{3.70}$$

  Where the mixture density is defined as $M(x) = (p(x) + q(x))/2$. As the expectations involved are intractable, we quantize the mass distribution m into 32 bins between 5.24 and 5.29 GeV and estimate the JSD between the normalized discretized variables:

$$\text{JSD} = D_{\text{JS}}\left(N^{\text{pass}}_{\text{bkg}}(m)/\sum_i N^{\text{pass},i}_{\text{bkg},i}(m) \,\middle|\middle|\, N^{\text{fail}}_{\text{bkg}}(m)/\sum_i N^{\text{fail},i}_{\text{bkg},i}(m)\right). \tag{3.71}$$

  A smaller JSD (larger 1/JSD) indicates a higher degree of invariance of decisions made by the classifier to the mass variable. Intuitively, the larger the difference between the distributions of background events mistakenly accepted by the selection and background events rejected by the selection, the larger this metric will be and the worse the extent of background sculpting. Any efficiently computable distance between probability distributions, such as the Cramér–von Mises distance (Cramér, 1928), will exhibit similar behaviour.
- To quantify classification performance, we consider the area under the receiver operating characteristic curve (AUC) obtained when the discriminator is trained on the intermediate representation. This is equal to the probability that a random signal example will be ranked higher than a random background example.

We report results using each of the VAE-based methods for learning disentangled representations reviewed in Section 3.2.2. Unless explicitly stated, we use the general disentanglement objective (Eq. 3.58) with supervision. If a model is trained without supervision, we denote this by 'unsupervised'. Without labels for the factors v, we are unable to directly enforce a latent-sensitive correspondence - the latent dimensions t are unobserved, so we remove the latent dimension(s) with the highest mutual information with the sensitive factor(s) s from the mean of the variational posterior to obtain the intermediate representation. Unless otherwise stated, the hyperparameters used for each method are detailed in Appendix B.1. As a baseline we consider the scenario where the classification network is trained over the full 34 features of the observed data. We tune hyperparameters on the validation set, and report all results on the independent test set. Results of a non-comprehensive scan over important hyperparameters are given in Appendix B.1.

Results

As a baseline, we report the performance obtained when the downstream discriminator is given as input the full set of 34 input variables. The performance obtained should be an upper bound on the maximum achievable performance by methods which use the latent representation in lieu of the original observed data. As observed in Chapter 2.2, the background M_{bc} distribution after rejection of 99.5% of background events is sculpted from the original distribution to mimic the signal distribution (Figure B.6). As the signal and background distributions appear degenerate, the relative contributions of signal and background events to the total measured yield can no longer be characterized easily, precluding measurement of the signal yield through a maximum likelihood fit to the M_{bc} distribution.

We show the corresponding plots in Figure B.7 when the M_{bc}-invariant representation obtained through our proposed approach using the β-TCVAE-Sensitive loss (Eq. 3.58) is used in lieu of the original set of features. The background

M_{bc} distribution retains its shape under event selection, and the background and signal distributions remain distinct from one another after applying event selection. This allows the relative contributions from both sources to be properly characterized via a maximum likelihood fit to some template probability distribution function. Corresponding plots for the remainder of the loss functions considered in Section 3.2.2 can be found in Appendix 3.3.2.

We perform a basic scan over a subset of the hyperparameters for each disentanglement loss, the details of which may be found in Appendix B.1. We take the hyperparameters from the median AUC configuration for each loss function and perform three independent runs with the given hyperparameter configuration, with different random initialization seeds. We report the mean and standard deviation of the AUC, JSD and MIG metrics obtained in Table 3.2. We note that the β-TCVAE-Sensitive loss achieves the best classification performance, while the Factor-VAE loss achieves the lowest disentanglement and exhibits the least amount of background sculpting. We note that the β-VAE and Annealed-VAE losses, which focus on penalizing the mutual information $I_q(x;z)$ (Section 3.2.2) do not score highly on any of the metrics. losses which penalize the total correlation $D_{\mathrm{KL}}\left(q(z)\|\prod_i q(z_i)\right)$ (β-TCVAE-Sensitive, β-TCVAE and Factor-VAE) score highly on all three metrics. From the results in Table 3.2, we observe that β-TCVAE-Sensitive and Factor-VAE are generally the best-performing models and fairly interchangeable for the purposes of learning invariant representations on this dataset, followed by β-TCVAE. We find that the β-VAE and Annealed-VAE losses in general do not find appropriate representations on this dataset.

Model	AUC	JSD	MIG
β-TCVAE-Sensitive	$\mathbf{0.950 \pm 0.002}$	0.0046 ± 0.0016	0.071 ± 0.009
β-TCVAE	0.923 ± 0.014	0.0059 ± 0.0011	0.16 ± 0.03
Factor-VAE	0.945 ± 0.004	$\mathbf{0.0021 \pm 0.0007}$	$\mathbf{0.152 \pm 0.005}$
Annealed-VAE	0.915 ± 0.006	0.04 ± 0.02	0.062 ± 0.003
β-VAE	0.903 ± 0.021	0.03 ± 0.02	0.064 ± 0.017
Baseline	0.988 ± 0.000	0.324 ± 0.0074	-

Table 3.2: Metrics obtained when the invariant representation $z \setminus t$ is used as input for the downstream discriminator. Shown is the mean and standard deviation over 3 trials obtained by using the hyperparameters of the median AUC configuration for each loss in the scan detailed in Appendix B.1. Higher is better for AUC and MIG, while lower is better for JSD.

Multidimensional Decorrelation

Many analyses in high-energy physics extract physical observables by fitting models of the signal and background distribution over two or more fit variables. One example is the extraction of signal and background yields through a combined fit to the M_{bc} and ΔE[9] spectra at B-factory experiments (e.g. (Saito et al., 2015)). Another example is extractiion of the fraction, mass and width of intermediate resonances in a given decay through a two-dimensional fit to the Dalitz plane of the decay (e.g. (Aaij et al., 2015)). In this case all fit variables should be protected against exploitation by any system separating signal and background. Our method naturally handles the case of multiple protected variables provided the true labels for each variable are known. This is more difficult to achieve for the adversarial method proposed (Louppe et al., 2017) as a separate adversary network must be trained to predict each protected variable, significantly increasing the complexity of the aggregate model with each protected variable added. In Figures 3.8 and 3.9 we show results of training the downstream classifier on a representation which is invariant to M_{bc} and ΔE. The distribution of both variables remains fairly stable under selection. This is to be contrasted with Figure 3.10, where the signal and background M_{bc} and ΔE distributions peak in the same region after selection of a signal-rich subsample, although this effect is less pronounced in ΔE. However, restricting the downstream classifier from exploiting information in ΔE further decreases classification power. Future work should investigate the relationship between invariance and downstream performance to more efficiently navigate this Pareto frontier.

3.5.3 Future Work

- The distribution of many variables in the input data is multimodal, so the joint distribution over all input variables x will be non-Gaussian. Thus the generative model $p_\theta(x|z)$ cannot represented accurately using a

[9] $\Delta E = E_B - E_{\text{beam}}$ is defined as the difference between the reconstructed energy of the B meson candidate and the beam energy.

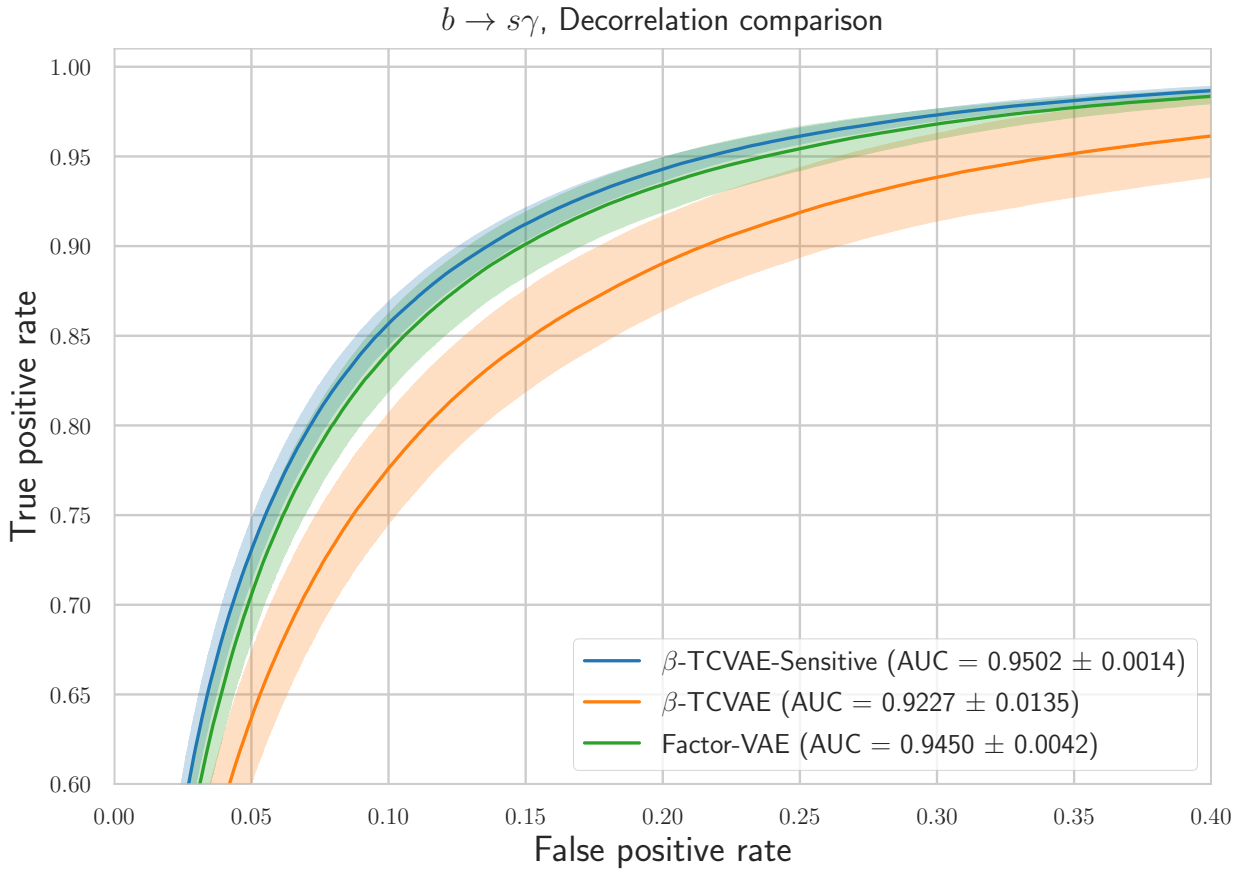


Figure 3.6: False positive rate versus true positive rate when trained on an M_{bc}-invariant representation obtained via optimizing the β-TCVAE-Sensitive, β-TCVAE and Factor-VAE losses. Shown is the mean and standard deviation over 3 trials obtained by using the hyperparameters of the median AUC configuration for each loss in the scan detailed in Appendix B.1. We observe that β-TCVAE-Sensitive appears to yield the best AUC, followed by Factor-VAE and β-TCVAE.

factorized Gaussian. Similarly, there is not guarantee that the latent space can be modelled using a factorized Gaussian. In order to make the generative model more expressive, we attempted to represent $p_\theta(x|z)$ as a normalizing flow model (Rezende and Mohamed, 2015). The essential idea is to express samples from a complex distribution, in this case the target data distribution $p^*(x)$, as a differentiable, invertible transformation T_ψ of samples from a simple base distribution $p(u)$, such that

$$x = T_\psi(u), \quad u \sim p(u). \tag{3.72}$$

Here the base distribution $p(u)$ is taken to be a factorized Gaussian as in the standard VAE model. The transformed distribution is obtained through change of variables (Rudin, 1987):

$$p^*(x) = p(u) \left|\det J_T(u)\right|^{-1}. \tag{3.73}$$

Where J_T denotes the Jacobian of the transformation. We can attempt to force to the generative model to approach the true data density through optimizing the transformation parameters ψ to minimize the KL divergence between the generative model and the target data distribution $p^*(x)$:

$$\begin{aligned} D_{\mathrm{KL}}\left(p^*(x)\|p_\theta(x|z)\right) &= -\mathbf{E}_{p^*(x)}\left[\log p_\theta(x|z)\right] - H(p^*) && (3.74)\\ &= -\mathbf{E}_{p^*(x)}\left[\log p_\theta\left(T_\psi^{-1}(x)|z\right)\right] + \log\left|\det J_{T^{-1}}(x)\right| - H(p^*). && (3.75) \end{aligned}$$

Where $H(p^*)$ denotes the entropy of the data distribution, which is a constant with respect to ψ, and the last line is due to the inverse function theorem. In order for this optimization to be tractable, the time complexity of

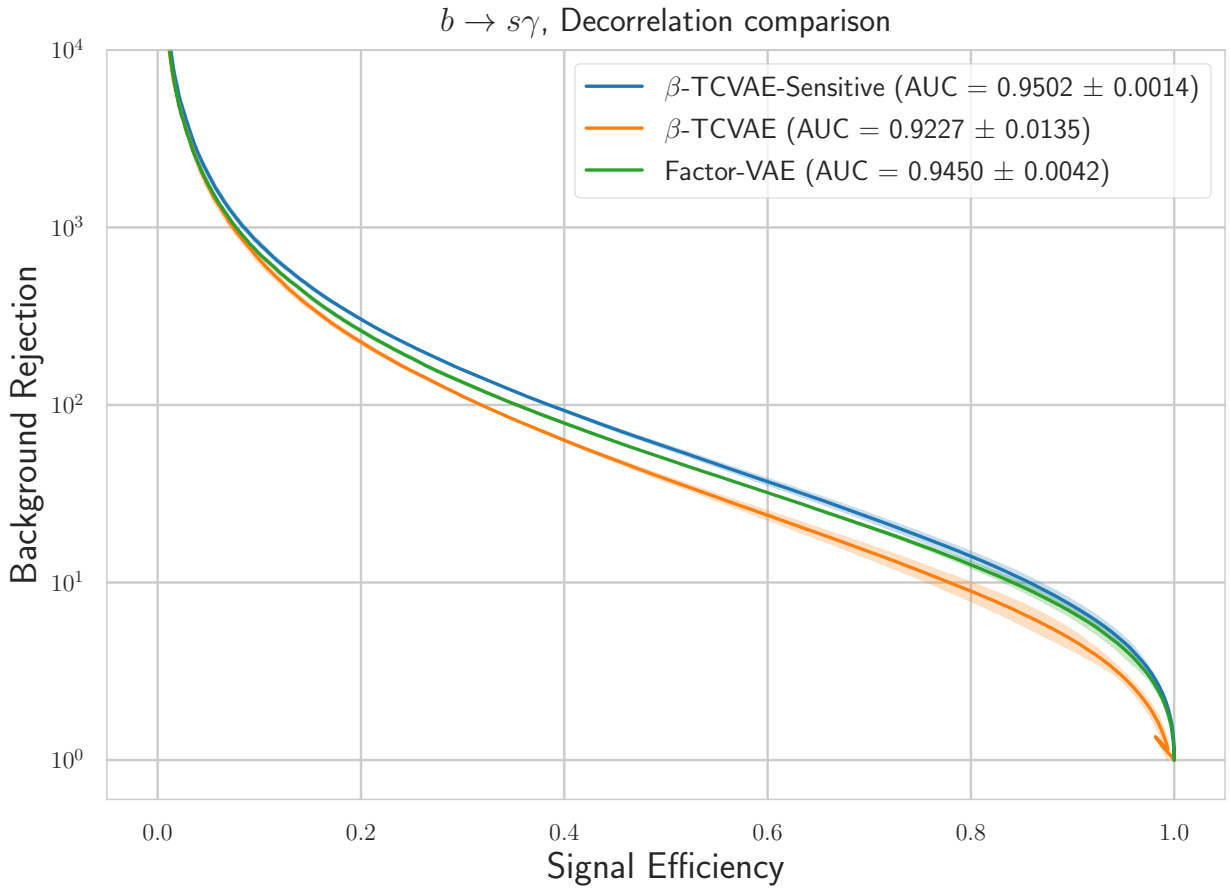


Figure 3.7: Signal efficiency versus background rejection when trained on an M_{bc}-invariant representation obtained via optimizing the β-TCVAE-Sensitive, β-TCVAE and Factor-VAE losses. The signal efficiency ϵ_S is defined as the ratio of surviving signal events to original signal events under some selection procedure, while the background rejection $1/\epsilon_B$ is defined as the ratio of original background events to surviving background events under some selection procedure. Shown is the mean and standard deviation over 3 trials obtained by using the hyperparameters of the median AUC configuration for each loss in the scan detailed in Appendix B.1.

computing the Jacobian should be $O(D_x)$ (linear in the dimension of x). We implemented the transformation T_ψ using the neural network based parameterization proposed in Real-NVP (Dinh et al., 2017). We find that this increases the expected log-likelihood $\mathbf{E}_{q_\phi}[\log p_\theta(x|z)]$, thus improving the ELBO. This suggests that normalizing flows improve the generative capability of our model. However, we do not observe significantly improved performance on the downstream task or disentanglement score that merit the increased computation associated with computing T_ψ, so we do not consider this in our final results. There are several applications of the proposed approaches in this section where sampling from the learned distribution is important and hence the output distributions must be accurately modelled - for example, fast generative modelling. For such applications, using normalizing flows to increase the expressivity of the modelling distributions should be the subject of future work.

- As the inference model $q_\phi(z|x)$ is modelled as a factorized Gaussian, our approach is not equipped to handle discrete variables in the input data or enforce invariance against discrete sensitive factors. Part of the difficulty in handling discrete factors of variation is that the inference $q_\phi(z|x)$ and generative models $p_\theta(x|z)$ must be differentiable with respect to their respective distribution parameters. One possibility for future work is to parameterize the distribution over discrete variables with a continuous relaxation of the categorical distribution (Jang et al., 2016) and combine this with the standard parameterization of distributions over continuous variables in a hybrid framework (Dupont, 2018).
- There are many possible candidates for the supervised penalty to be used in Eq. 3.58. Using the cross-entropy-like term in (Eq. 3.40) may be suboptimal as it does not exploit the probabilistic formulation of VAEs. While

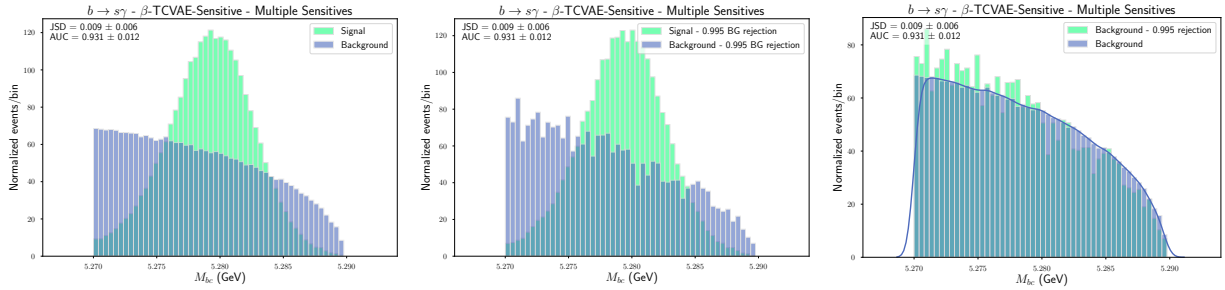


Figure 3.8: M_{bc} distributions after enforcing a multidimensional decorrelation. 'Post-selection' refers to the resultant sample after rejection of 99.5% of background events. **Left:** Original signal (green) and background (blue) distributions. **Middle:** Post-selection signal (green) and background (blue) distributions. **Right:** Original background (blue) and post-selection background (green). We note that restricting the downstream classifier from exploiting information in both M_{bc} and ΔE further reduces discrimination power.

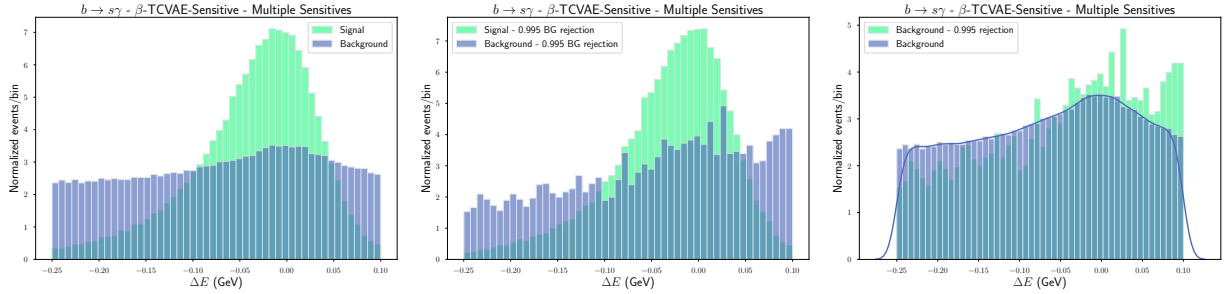


Figure 3.9: ΔE distributions after enforcing a multidimensional decorrelation. 'Post-selection' refers to the resultant sample after rejection of 99.5% of background events. **Left:** Original signal (green) and background (blue) distributions. **Middle:** Post-selection signal (green) and background (blue) distributions. **Right:** Original background (blue) and post-selection background (green). We note that restricting the downstream classifier from exploiting information in both M_{bc} and ΔE further reduces discrimination power.

this works well in practice, it may more principled to use some estimator of the mutual information between the latent dimensions and the generative factors instead, for example, the approach proposed in Chapter 2.2 or the non-parametric relative entropy estimator proposed in Kolchinsky and Tracey (2017).

- One drawback of this method is an inability to control the tradeoff between downstream performance and the degree of invariance through an explicit hyperparameter, as was possible under the approaches discussed in Chapter 2.2, which take this hyperparameter to be the weighting of the function of the auxiliary discriminator output in the joint loss function. This is desirable as this can simplify finding the Pareto frontier between downstream performance and invariance by sweeping this hyperparameter. Note that the Lagrange multiplier of the supervised loss does not fulfill this role as the representation must be disentangled on top of informativeness between sensitive factors and corresponding latent dimensions. We leave investigation of this point to future work.

3.6 Conclusion

Machine learning methods capable of efficiently separating signal from background events have installed themselves as an important part of the analysis procedure in high-energy physics. This is particularly true for indirect searches for new physics in loop-suppressed decays at the intensity frontier, where signal events are rare and the collected data is dominated by background processes. However, methods which solely aim to maximize the signal-to-noise ratio may be detrimental to the significance of any potential discovery or exclusion by introducing a sampling bias into the distribution of physically important variables under selection. This introduces systematic effects into the analysis which are difficult to characterize, as we have argued in this thesis. To resolve this complication, machine learning methods should distinguish signal from background events without using information about such protected variables.

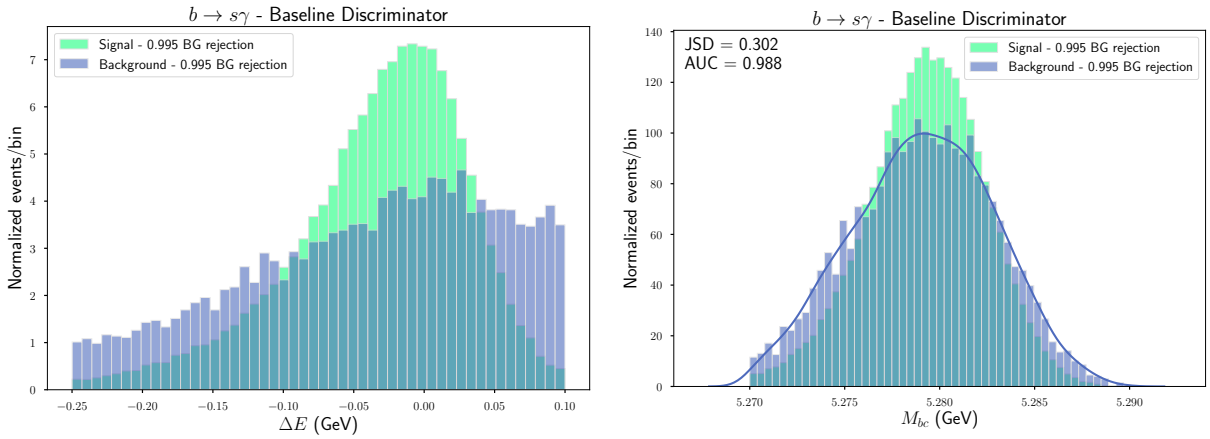


Figure 3.10: M_{bc} and ΔE distributions obtained after selection by a baseline discriminator trained using maximum likelihood without constraints. **Left:** Post-selection signal (green) and background (blue) ΔE distributions. **Right:** Post-selection signal (green) and background (blue) distributions. In both cases the surviving background suffers selection bias, sculpting the shape of the distribution to resemble the structure in the signal spectrum. This introduces unwanted systematic effects when attempting to model the background and signal-background hypotheses using the joint M_{bc} and ΔE distribution.

The challenge in developing such methods lies in striking a balance between increasing the signal-to-noise ratio and avoiding the introduction of excessive systematic uncertainty.

In this thesis we approached this task as an optimization problem of finding a representation of the observed data that is invariant to the given protected quantities. Two novel approaches to this problem were outlined in this thesis. Chapter 2 proposed minimizing the mutual information between the learnt representation and the protected quantities jointly with the original classification loss. Chapter 3 proposed using the variational autoencoder framework to learn structured latent representations to make it possible to isolate the influence of these protected quantities in a distinct subspace of the representation. This subspace may be removed, excising information about the protected variables from the representation, and the resulting reduced representation used for the original classification task. We further proposed a novel objective function that can be used to learn disentangled representations which achieves favourable results when compared to previous objective functions on the datasets considered in the experimental section of this chapter. Although our representation-learning approaches were developed with the expectation that the learnt representation would be used for classification, our approaches are task-agnostic and may be applied to any downstream tasks requiring the invariance condition. We demonstrate the applicability of our proposed methods on two important classification tasks at the Belle II experiment, discrimination of leptons from hadronic background, and identification of signal events from the rare $b \to s\gamma$ transition.

In the continued absence of an unambiguous signal of new physics, demands on the sensitivity of experimental analyses testing extensions of the Standard Model will grow. Methods such as those we have proposed, capable of reducing statistical uncertainties without undue increase in systematic uncertainty, are expected to become an increasingly important part of experimental high-energy physics.

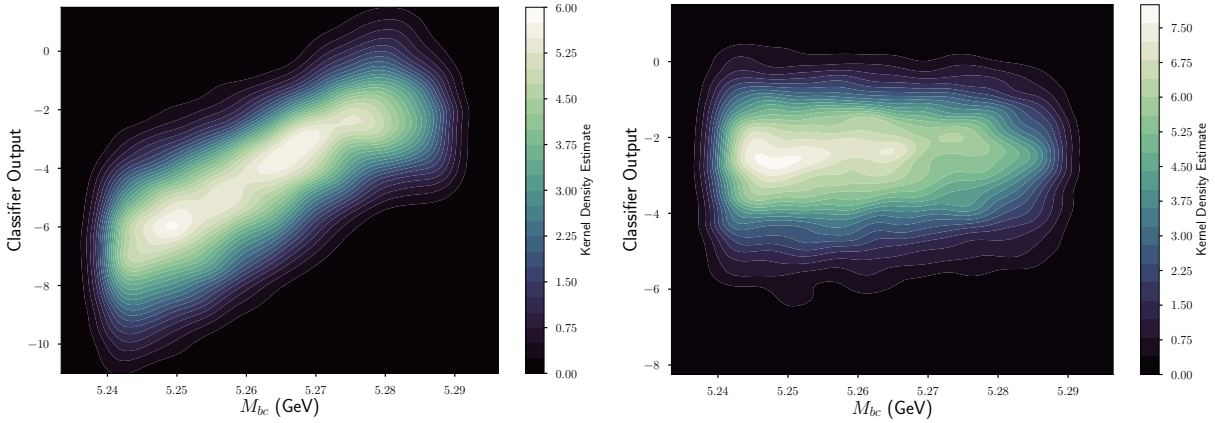


Figure 3.11: Visualization of invariance of the downstream classification task to the sensitive variable $s = M_{bc}$. The plots show the bivariate kernel density estimate between M_{bc} and the classifer output, $\log\left(\frac{p}{1-p}\right)$, where p is the probability that a given event belongs to the positive signal class given by the downstream classifier. Note that the output is a monotonic function of p - larger values of the classifier output indicate higher confidence by the classifier that an example belongs to the positive class. The left hand figure shows the output of the classifier when trained on the baseline feature set (AUC 0.989) while the right hand figure shows the output of the classifier when trained on the invariant representation obtained using β-TCVAE-Sensitive (AUC 0.950). We observe that the output of the baseline classifier is strongly correlated with M_{bc}, and that it outputs higher confidence that an event belongs to the signal class around the signal region $M_{bc} \in (5.27, 5.29)$ GeV, where the signal M_{bc} distribution is sharply peaked (Figure B.6). In contrast, the output of the invariant classifier does not exhibit significant dependency on M_{bc}, with the classifier response flat throughout the full M_{bc} range. This enables downstream tasks to make decisions largely independent of the sensitive variables, at the expense of reduced discrimination power.

REFERENCES

Aaij, R. et al. (2015). Dalitz plot analysis of $B^0 \to \overline{D}^0 \pi^+ \pi^-$ decays. *Phys. Rev.*, D92(3):032002.

Aaij, R. et al. (2017). Test of lepton universality with $B^0 \to K^{*0} \ell^+ \ell^-$ decays. *JHEP*, 08:055.

Abadi, M. et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.

Achille, A. and Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. In *J. Mach. Learn. Res.*, volume 19.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142.

Altmannshofer, W. et al. (2018). The Belle II Physics Book.

Amhis, Y. et al. (2016). Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2016.

Apostolakis, J. et al. (2018). HEP Software Foundation Community White Paper Working Group - Detector Simulation.

Arjovsky, M. and Bottou, L. (2017). Towards Principled Methods for Training Generative Adversarial Networks. *arXiv preprint*, page arXiv:1701.04862.

ATLAS Collaboration (2018). Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS. Technical Report ATL-PHYS-PUB-2018-014, CERN, Geneva.

Behnke, O., Kröninger, K., Schörner-Sadenius, T., and Schott, G., editors (2013). *Data analysis in high energy physics*. Wiley-VCH, Weinheim, Germany.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Bennett, J. (October 2017). Kaon/Pion efficiency report, B2GM.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blake, T., Lanfranchi, G., and Straub, D. M. (2017). Rare B Decays as Tests of the Standard Model. *Prog. Part. Nucl. Phys.*, 92:50–91.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *arXiv preprint arXiv:1601.00670*.

Bourlard, H. and Kamp, Y. (1987). Auto-association by multilayer perceptrons and singular value decomposition. Manuscript, Philips Research Laboratory, Brussels, Belgium.

Burgess, C. and Kim, H. (2018). 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*.

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, Long Beach, California, USA. PMLR.

de Oliveira, L., Paganini, M., and Nachman, B. (2017). Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis. *Comput. Softw. Big Sci.*, 1(1):4.

del Amo Sanchez, P. e. a. (2010). Study of $B \to X\gamma$ Decays and Determination of $|V_{td}/V_{ts}|$. *Phys. Rev.*, D82:051101.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using Real NVP. *arXiv preprint*, page arXiv:1605.08803.

Donoghue, J. F., Golowich, E., and Holstein, B. R. (2014). *Dynamics of the Standard Model*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 2 edition.

Donsker, M. and Varadhan, S. (1975). Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Dupont, E. (2018). Learning Disentangled Joint Continuous and Discrete Representations. *arXiv preprint*, page arXiv:1804.00104.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Eastwood, C. and Williams, C. K. I. (2018). A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*.

Edwards, H. and Storkey, A. (2015). Censoring Representations with an Adversary. *arXiv preprint*, page arXiv:1511.05897.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Ferber, T. (October 2017). MC9 Photon resolution study, B2GM.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Gao, S., Steeg, G. V., and Galstyan, A. (2014). Efficient estimation of mutual information for strongly dependent variables. *CoRR*, abs/1411.2003.

Glashow, S. L., Iliopoulos, J., and Maiani, L. (1970). Weak interactions with lepton-hadron symmetry. *Phys. Rev. D*, 2:1285–1292.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Hassabis, D. e. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a Definition of Disentangled Representations. *arXiv preprint*, page arXiv:1812.02230.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Hu, Q.-Y., Li, X.-Q., and Yang, Y.-D. (2017). $b \to k^* \mu\mu$ decay in the aligned two-higgs-doublet model. *The European Physical Journal C*, 77(3):190.

Hurth, T. and Nakao, M. (2010). Radiative and Electroweak Penguin Decays of B Mesons. *Ann. Rev. Nucl. Part. Sci.*, 60:645–677.

James, F. and Roos, M. (1975). Minuit – a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10:343–367.

Jang, E., Gu, S., and Poole, B. (2016). Categorical Reparameterization with Gumbel-Softmax. *arXiv preprint*, page arXiv:1611.01144.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Keck, T. (2017). *Machine learning algorithms for the Belle II experiment and their validation on Belle data*. PhD thesis, Karlsruher Institut für Technologie (KIT).

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kolchinsky, A. and Tracey, B. D. (2017). Estimating mixture entropy with pairwise distances. *Entropy*, 19(7).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Lees, J. P. e. a. (2012). Exclusive Measurements of $b \to s\gamma$ Transition Rate and Photon Energy Spectrum. *Phys. Rev.*, D86:052012.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2016). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *ArXiv preprint*.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. (2019a). On the Fairness of Disentangled Representations. *arXiv preprint*, page arXiv:1905.13662.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. (2018). Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. (2019b). Disentangling Factors of Variation Using Few Labels. *arXiv preprint*, page arXiv:1905.01258.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The Variational Fair Autoencoder. *arXiv preprint*, page arXiv:1511.00830.

Louppe, G., Kagan, M., and Cranmer, K. (2017). Learning to pivot with adversarial networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 981–990. Curran Associates, Inc.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. (2019). Dont blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems 32*, pages 9403–9413. Curran Associates, Inc.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2017). Are GANs Created Equal? A Large-Scale Study. *arXiv preprint*, page arXiv:1711.10337.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Misiak, M. and Steinhauser, M. (2017). Weak radiative decays of the b meson and bounds on $m_{H^\pm}$ in the two-higgs-doublet model. *The European Physical Journal C*, 77(3):201.

Misiak, M. e. a. (2015). Updated NNLO QCD predictions for the weak radiative B-meson decays. *Phys. Rev. Lett.*, 114(22):221801.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. The MIT Press.

Moyer, D., Gao, S., Brekelmans, R., Steeg, G. V., and Galstyan, A. (2018). Invariant representations without adversarial training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9102–9111, Red Hook, NY, USA. Curran Associates Inc.

Munch, E. (1893). The scream.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2008). Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Advances in neural information processing systems*, pages 1089–1096.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc.

Paganini, M., de Oliveira, L., and Nachman, B. (2018). CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev.*, D97(1):014021.

Paszke, A. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035.

Paul, A. and Straub, D. M. (2017). Constraints on new physics from radiative B decays. *Journal of High Energy Physics*, 4:27.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180, Long Beach, California, USA. PMLR.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML 2015, pages 1530–1538.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA.

Saito, K., Yamamoto, S., Ushiku, Y., and Harada, T. (2018). Open Set Domain Adaptation by Backpropagation. *arXiv preprint*, page arXiv:1804.10427.

Saito, T., Ishikawa, A., Yamamoto, H., Abdesselam, A., Adachi, I., Aihara, H., Al Said, S., Arinstein, K., Asner, D. M., Aushev, T., et al. (2015). Measurement of the b$^-$ x s γ branching fraction with a sum of exclusive decays. *Physical Review D*, 91(5):052004.

Saito, T. e. (2015). Measurement of the $\bar{B} \to X_s\gamma$ Branching Fraction with a Sum of Exclusive Decays. *Phys. Rev.*, D91(5):052004.

Shimmin, C., Sadowski, P., Baldi, P., Weik, E., Whiteson, D., Goul, E., and Søgaard, A. (2017). Decorrelated Jet Substructure Tagging using Adversarial Neural Networks. *Phys. Rev.*, D96(7):074034.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. (2019). Weakly Supervised Disentanglement with Guarantees. *arXiv preprint*, page arXiv:1910.09772.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. (2016). Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. (2019). Learning controllable fair representations. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR.

Stummer, W. and Vajda, I. (2012). On bregman distances and divergences of probability measures. *IEEE Transactions on Information Theory*, 58(3):1277–1288.

Suter, R., Miladinovic, D., Schölkopf, B., and Bauer, S. (2019). Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6056–6065, Long Beach, California, USA. PMLR.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):611–622.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial Discriminative Domain Adaptation. *arXiv preprint*, page arXiv:1702.05464.

van Steenkiste, S., Locatello, F., Schmidhuber, J., and Bachem, O. (2019). Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv 1905.12506*, abs/1905.12506.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In Moody, J. E., Hanson, S. J., and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. Morgan-Kaufmann.

Villani, C. (2009). *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg.

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1073–1080, New York, NY, USA. Association for Computing Machinery.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Wang, Y., Koike-Akino, T., and Erdogmus, D. (2018). Invariant Representations from Adversarially Censored Autoencoders. *arXiv preprint*, page arXiv:1805.08097.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4(1):66–82.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA. PMLR.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2019). Conditional Learning of Fair Representations. *arXiv preprint*, page arXiv:1910.07162.

APPENDIX A

APPENDIX: INVARIANCE THROUGH REGULARIZATION

A.1 PROOF OF PROPOSITION 1

Let $h = f(x)$ with encoder f denote the intermediate representation of the data x. We allow for the possibility that h is a stochastic function of x, with corresponding conditional distribution $q_f(h|x)$. Then minimization of the mutual information $I(z;h)$ between the encoding h and the protected variable z is equivalent to maximization of the loss of an adversary trained to predict the value of z from h, as detailed in the adversarial training procedure (Eq. 2.21) proposed by Louppe et al. (2017). Furthermore, minimization of the loss of an adversary predicting the value z from h is equivalent to maximization of a lower bound on the mutual information $I(z;h)$.

As in Section 2.2, assume the dataset $\{\mathbf{x}, \mathbf{z}\}$ consists of N independent samples from the joint $p(x,z) = p(z|x)p(x)$. The mapping $h = f(x)$ induces a conditional distribution over z given the representation h, defined as:

$$q_f(z|h) = \int dx\, p(x) \frac{p(z|x)q_f(h|x)}{q(h)}. \tag{A.1}$$

Where the marginal $q(h) = \int dx\, p(x) q_f(h|x)$. Note that Bayes' Theorem implies $q_f(h|z) = \frac{q_f(z|h)q(h)}{p(z)}$. Recall that the adversary network r outputs an estimate of the probability of observing the sensitive value z given input $h = f(x)$, $q_r(z|h)$. The adversary attempts to match this to the true probability of observing z given h, $q_f(z|h)$, by minimizing the adversary loss $\mathcal{L}_r = \mathbf{E}_{p(x)}\left[\mathbf{E}_{q_f(z|x)}\left[-\log q_r(z|h = f(x))\right]\right]$. Recall that the cross-entropy between two probability distributions is defined as $H(p(x), q(x)) = -\mathbf{E}_{p(x)}\left[\log q(x)\right]$. Hence the adversary loss is equivalent to minimizing the cross entropy loss between q_f and q_r in expectation over the dataset, $\mathcal{L}_r = \mathbf{E}_{q_f(h|x)p(x)}\left[H(q_f(z|h), q_r(z|h))\right]$.

Proof. We begin by using the definition of the mutual information between two random variables (Eq. 2.18).

$$\begin{aligned}
I(z;h) &= \mathbf{E}_{q_f(z|h)q(h)}\left[\log \frac{q_f(z|h)}{p(z)}\right] && \text{(A.2)}\\
&= \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_f(z|h)\right] - \mathbf{E}_{q_f(h|z)p(z))}\left[\log p(z)\right] && \text{(A.3)}\\
&= \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_f(z|h)\right] + H(z) && \text{(A.4)}\\
&= H(z) - H(z|h). && \text{(A.5)}
\end{aligned}$$

In Equation A.5 we use the definition of the conditional entropy $H(y|x) = -\mathbf{E}_{p(x,y)}\left[\log p(x|y)\right]$, which measures the uncertainty remaining in y given observation of x. The larger the conditional entropy, the less information observation of x conveys about y. Intuitively, the mutual information $I(h;z)$ is minimized by optimizing the encoding network f to maximise the conditional entropy $H(z|h) = -\mathbf{E}_{q_f(z|h)q(h)}\left[\log q_f(z|h)\right]$, making it difficult for the adversary to predict z based on h.

To complete the proof, we start from Eq. A.4:

$$
\begin{aligned}
I(z;h) &= \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_f(z|h)\right] + H(z) && \text{(A.6)}\\
&= \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_f(z|h)\right] + H(z) - \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_r(z|h)\right] + \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_r(z|h)\right] && \text{(A.7)}\\
&= \mathbf{E}_{q(h)}\left[D_{\mathrm{KL}}\left(q_f(z|h)\|q_r(z|h)\right)\right] + H(z) + \mathbf{E}_{q_f(z|h)q(h)}\left[\log q_r(z|h)\right] && \text{(A.8)}\\
&\geq H(z) + \mathbf{E}_{q_f(h|x)p(x)}\left[\mathbf{E}_{q_f(z|h)}\left[\log q_r(z|h)\right]\right] && \text{(A.9)}\\
&= H(z) - \mathbf{E}_{q_f(h|x)p(x)}\left[H(q_f(z|h), q_r(z|h))\right]. && \text{(A.10)}\\
&= H(z) - \mathcal{L}_r. && \text{(A.11)}
\end{aligned}
$$

In Eq. A.8 we used the definition of the KL divergence. In Eq. A.9 we used the non-negativity of the KL divergence. In Eq. A.10 we used the definition of the cross-entropy. Note that $H(z)$ is a constant with respect to the parameters to be optimized over. Hence minimizing the mutual information $I(z;h)$ through the encoder f is equivalent to maximizing a lower bound on the adversary loss $\mathcal{L}_r$. Similarly, minimizing the adversarial loss $\mathcal{L}_r$ maximizes a lower bound on the mutual information $I(z;h)$.

Thus the adversarial training procedure (Eq. 2.21) is formally equivalent to the mutual information regularization approach for invariant representation learning outlined in Algorithms 1 and 2, where the loss function of the adversary serves as a proxy for the mutual information $I(z;h)$. □

A.2 TRAINING FEATURES

A total of 101 variables are used as input to the classification network. They can be roughly divided into variables describing event topology, and kinematic variables of the B meson and radiative photon. All variables are scaled to zero mean and unit variance, with the necessary transformation calculated on the training set and applied to the test and validation sets. No variable with Pearson correlation above 0.1 with the beam-constrained mass M_{bc} was included. All kinematic variables are evaluated in the $\Upsilon(4S)$ center of mass (CMS) frame unless stated otherwise. A brief description of each input variable is listed below, grouped into their corresponding categories. For a comprehensive overview of event shape or continuum suppression variables one may refer to Section 5 of the Belle II Theory Interface Platform report (Altmannshofer et al., 2018).

Kinematic / Angular variables

- The angular distribution $(\cos(\theta), \phi)$ of the reconstructed B meson and radiative photon.
- The cosine of the helicity angle between the candidate B and the radiative photon. Defined as the angle between the line defined by the momentum difference of the hadronic system and photon in the rest frame of the B and the momentum of the γ in the lab frame.
- The spatial coordinates of the vertex of the B meson relative to the interaction point, and associated uncertainties.
- The impact parameter of the B meson.
- The maximum value recorded between the daughters of the signal B for each four-momentum component $(E, \vec{p})$, and the transverse momentum p_t.
- The energy and transverse momentum of the radiative photon.

Vertexing-based variables

- ΔZ: The distance along the beamline axis of the decay vertices of the signal and tag B mesons, recorded in the lab frame. Asymmetric beam energy results in a Lorentz boost to the coherent $B\bar{B}$ pair produced in a collison event, enabling non-negligible separation before their respective decays. In contrast, the produced light quark pair in continuum events fragments in a very short timescale, placing the spurious reconstructed B decay vertices closer to the interaction point.
- χ^2: The χ^2 probability of the vertex fit.
- ΔB: The boost direction between the reconstructed and tag B.
- $\cos\left(\theta_{Vtx-Z}\right)$: The cosine of the angle between the momentum vector of the B candidate and the vector between the interaction point and fitted vertex point. This should peak at 1 for signal events.

Rest-of-Event (ROE) variables The ROE contains detected tracks and calorimeter clusters in a collision event that do not originate from the reconstructed B meson, and are attributed to the companion $\bar{B}$ along with extra tracks and clusters from other background processes and additional tracks arising from the in-flight decay of charged particles. In a typical analysis the rest of event is typically cleared of spurious track and cluster events by the application of cuts, however in this study the total energy, total momentum, (p_x, p_y, p_z) 3-momentum, invariant mass and beam constrained mass of the ROE are included as inputs.

Electromagnetic calorimeter (ECL) response variables These are variables associated with energy deposition in the ECL by the radiative photon.

- Cluster E1E9 (E9E21): Ratio of energies deposited in the central inner 3×3 and outer 5×5 (excluding corners) crystals around the central crystal. This ratio is close to 1 for true photon candidates but lower for asymmetric hadronic shower events.
- Cluster second moment: Average distance of the incident particle from the central cluster, weighted by cluster energy.
- The $(4, 0)$ and $(5, 1)$ Zernike moments, used as shape descriptors of the electromagnetic showers generated in the ECL crystals by incident particles.

Event shape variables These variables characterize events by the geometry of the final state distribution and are derived from lower-level kinematic quantities. The beam energies at B-factories are tuned to be slightly above the $\Upsilon(4S)$ threshold, hence the entangled $B\bar{B}$ pair is produced at relatively low momentum in the CMS frame, resulting in an approximate isotropic decay with no correlation between the decay products of each B. In contrast, light quark continuum events $e^+e^- \to qq$ exhibit a two-jet structure, with particles concentrated in two narrow back-to-back cones.

- Fox-Wolfram moments, which contain information about the angular correlation and energy distribution of the final state. These are parameterized in terms of the opening angle between the momenta of final state particles θ_{ij}, and can be expressed as a superposition of spherical harmonics weighted by particle momenta:

$$H_l = \sum_{ij} \frac{|\vec{p}_i||\vec{p}_j|}{\sum_i |\vec{p}_i|} P_l\left(\cos\theta_{ij}\right) \tag{A.12}$$

  The degree l can range from $(0, 8)$. We also include the corresponding normalized ratio $R_n = H_n/H_0$ and the 16 Super Fox-Wolfram moments, which constrain the sum to run over the B primary daughters and the B final state daughters, respectively.
- Cleo cone variables, which calculate the momentum flow into concentric cones at angular intervals of 10° around the signal B thrust axis up to right angles.
- The square of the missing mass of the system, which sums over all particles in the system and peaks at 0 for correctly reconstructed B candidates.

$$m^2_{miss} = \left(\sqrt{s} - \sum_n E_n\right)^2 - \sum_n |\vec{p}_n|^2 \tag{A.13}$$

- Thrust-based quantities. The thrust axis is given by the axis maximizing the longitudinal projections of the momenta of final state particles:

$$T = \max_{|\vec{v}_T|=1} \frac{\sum_i |\vec{v}_T \cdot \vec{p}_i|}{\sum_i |\vec{p}_i|} \tag{A.14}$$

  The sum can be restricted to subsets of particles to compute different thrust axes. We consider the following thrust-related quantities:
  - T_B: The magnitude of the signal B meson thrust axis.
  - T_{ROE}: The magnitude of the rest-of-event (ROE) - all unreconstructed particles) thrust axis.
  - $\cos\left(\theta_{T_{B-ROE}}\right)$: Cosine of the angle between the thrust axis of the signal B and ROE thrust axis.
  - $\cos\left(\theta_{T_{B-z}}\right)$: Cosine of the angle between the thrust axis of the signal B and the z axis.

  The isotropy of B decays in the CMS frame relative to the collimation of continuum decay products enable angular quantities derived from the thrust axes to differentiate between signal/background.

A.3 Maximum Likelihood Fit

The signal yield is extracted through an unbinned extended maximum likelihood fit to the M_{bc} distribution of the remaining data after rejecting 0.995 of background events, chosen to be within the realistic order of magnitude of rejection required for precision decay studies. The signal probability distribution function (PDF) is modelled using the empirically motivated crystal ball distribution, a Gaussian core with a power-law tail used to account for energy deposition in crystal calorimeters. The background PDF is modelled using the generalized ARGUS distribution, which is nominally a smoothly falling function over the considered fit range (5.24, 5.29) GeV. The signal PDF shape parameters are fixed to the original signal sample before selection, while the background shape parameters are allowed to float, along with the signal and background yields. The χ and p parameters of the generalized ARGUS function

	MI	Adv	No penalty
LR ratio	$L(-2, 2)$	$L(-2, 2)$	-
Classifier LR	10^{-4}	10^{-4}	$L(-5, -3)$
Iterations	$U(1, 32)$	$U(1, 32)$	-
λ	$U(10, 100)$	$U(10, 100)$	-
Activation	[ELU, ReLU]	[ELU, ReLU]	-

Table A.1: Ranges labelled $U(x, y)$ indicate values sampled from a uniform distribution with range $[x, y]$. Ranges labelled $L(x, y)$ indicate values sampled from a log-uniform distribution. That is, the exponent is sampled uniformly from the range $[x, y]$ with a base of 10.

are restricted to the ranges $(0, 12)$ and $(0.2, 2)$ respectively to prevent egregious departure of the background spectrum from its original distribution. If the optimization procedure fails to converge, the reward for the episode is recorded as -1. The maximum likelihood fit is accomplished using the `MINUIT` (James and Roos, 1975) `C++` package, through the iminuit Python interface. The signal yield N_S is given by the MLE point estimate, with standard error δN_S given by the expected value of the inverse of the Hessian.

A.4 Hyperparameter Optimization

In the experiments described in Section 2.1, we compare the proposed model against the adversarial training setup proposed in Louppe et al. (2017) and a model trained using the standard cross-entropy classification loss. In the adversarial training approach, the authors introduce a network that attempts to predict the value of the sensitive variable z from the encoding $E_\theta(x)$. The training procedure is adversarial as the loss function of the secondary network contributes negatively to the objective of the encoder network, hence both networks are optimized toward a saddle point in parameter space. Instead of directly attempting to regress the value of z, we partition z into 10 bins and attempt to predict the bin label from the encoder output $E_\theta(x)$, as proposed in Shimmin et al. (2017).

Hyperparameter configurations are evaluated using the asynchronous Hyperband algorithm (Li et al., 2016), an early stopping method that allocates more resources to promising configurations. The reduction rate η and the number of brackets are both set to the default recommended value of 3. Sequential optimization methods such as Bayesian optimization were found to require prohibitively large wall clock times to reach an acceptable level of performance across all models. In our experiments, we set the batch size to 1024 and the classifier learning rate to 10^{-4}. We use batch normalization in the classifier and layer normalization in the statistic network. The inputs to the statistic network are standardized to have zero mean and unit variance. The hyperparameters optimized are briefly described below. The ranges for each hyperparameter are shown in Table A.1.

- Learning rate ratio: The ratio of the learning rate between the classifier and the statistic/adversary network. The classifier learning rate is fixed at 10^{-4} when tuning the mutual information penalty and adversarial approaches, and floated when tuning the unpenalized mode.
- Iterations: The number of gradient steps for the statistic/adversary network for each gradient step of the classifier.
- λ: Multiplier of the mutual information penalty/adversarial loss term.
- Activation: The nonlinearity used in the statistic/adversary network.

All experiments presented in this chapter were run on a node with 24 Intel Xeon E5-2650 v4 processors and 4 NVIDIA Tesla P100 GPUs unless indicated otherwise. The described algorithms in Chapter 2.2 are implemented using the Tensorflow framework (Abadi et al., 2015). For ease of reproducibility, the accompanying source code for Chapter 2.2 is available at `https://github.com/Justin-Tan/invariant_reps` under the Apache License 2.0.

APPENDIX B

APPENDIX: INVARIANCE THROUGH DISENTANGLEMENT

B.1 DISENTANGLEMENT HYPERPARAMETERS

Here we list the baseline hyperparameters considered for each approach to learning disentangled representations reviewed in Section 3.2.2. Each hyperparameter is set to the value recommended by their respective authors. For hyperparameter coordinate scans we sweep over different values of the given hyperparameter while holding the remainder fixed to the baseline and show the mean with $\pm 1\sigma$ confidence regions over 3 trials with different random seeds unless otherwise specified. Note that the objective here is not to find the optimum hyperparameter configuration - this will be empirically dataset-independent and require many samples from hyperparameter space to understand the interactions between hyperparameters. The goal here is to evaluate sensitivity to hyperparameters of each possible implementation of our proposed approach and find a general 'Goldilocks zone' for each hyperparameter.

Model	Hyperparameter	Baseline	Scan
β-VAE	β	4	$[2, 4, 6, 8, 12, 16, 32]$
Annealed-VAE	γ	30	$[10, 20, 30, 50, 75, 100]$
	C	5	None
Factor-VAE	γ	12	$[1, 2, 3, 4, 5, 6, 7, 8, 10, 12]$
β-TCVAE	α	1	$[-1, 4, 6, 8, 16]$
	β	1	$[1, 2, 3, 4, 5, 6, 7, 8, 10, 12]$
	γ	1	None
β-TCVAE-Sensitive	α	1	$[-1, 4, 6, 8, 16]$
	β	6	$[1, 2, 3, 4, 5, 6, 7, 8, 10, 12]$
	γ	1	None
All	λ_s	64	$[8, 16, 32, 64, 128, 256, 512]$

Table B.1: Range of hyperparameters for each model considered in hyperparameter scans described in Appendix B.1. The 'baseline' column denotes the default value for each hyperparameter used in our experiments.

For the discriminator network necessary to estimate the TC term used in factor-VAE, we use a 5-layer MLP with 512 hidden units with ReLU activations, trained using the Adam optimizer (Kingma and Ba (2015)) with learning rate 5×10^{-5}. One gradient step is taken for each gradient step of the VAE model.

B.1.1 Latent Dimension

We scan over the latent dimension of the intermediate representation for β-TCVAE-Sensitive with all relevant hyperparameters set to the baseline in Figure B.1. We observe that increasing the number of latent dimensions increases the channel capacity and makes the $D_{\mathrm{KL}}\left(q_\phi(z|x)\|p(z)\right)$ term in the ELBO less harsh - with more dimensions, each latent dimension can encode information about the observed data through small variations of the learned posterior

$q_\phi(z|x)$ from the prior $p(z)$. With fewer dimensions, the KL term is more restrictive as the posterior distribution over each dimension must deviate from $p(z)$ more significantly to encode the same amount of information. This allows downstream classification performance to improve as the number of dimensions are increased, at the expense of a higher background distortion - indicated by decreasing 1/JSD. We choose a latent dimension of 8 as a compromise between invariance and downstream performance.

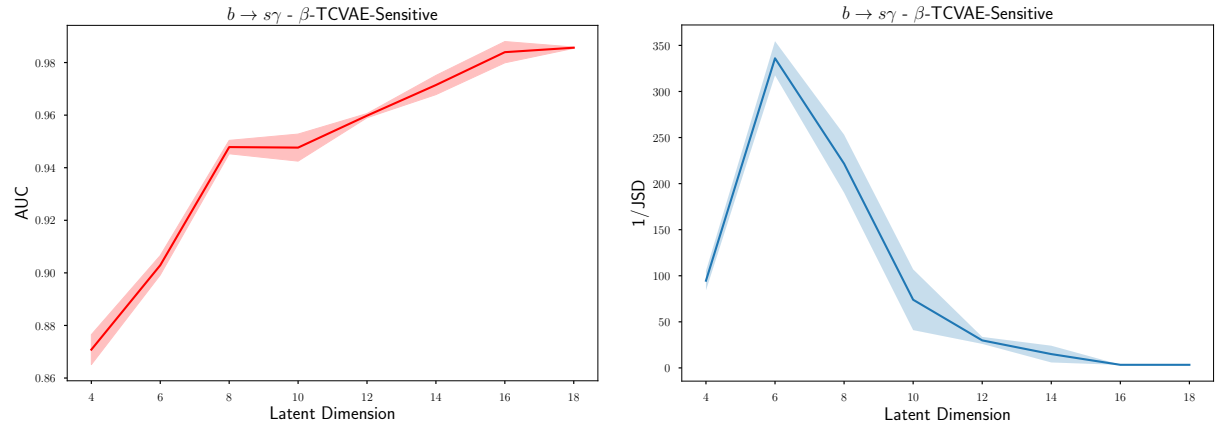


Figure B.1: AUC (left) and 1/JSD (right) of downstream classifier versus dimension of the latent representation for β-TCVAE-Sensitive. Higher is better for each metric. The shaded area denotes the $\pm 1\sigma$ confidence region.

B.1.2 Mutual Information Weighting

We scan over the weighting α of the mutual information between the observed data and latent code $I_q(x; z)$ for β-TCVAE-Sensitive and β-TCVAE in Figure B.2. Empirically, the 1/JSD metric peaks strongly at $\alpha = 1$ for β-TCVAE. Downstream classification performance and invariance as measured by the 1/JSD metric are fairly insensitive to values of $\alpha \geq 1$ for β-TCVAE-Sensitive, so we fix $\alpha = 1$ in all experiments.

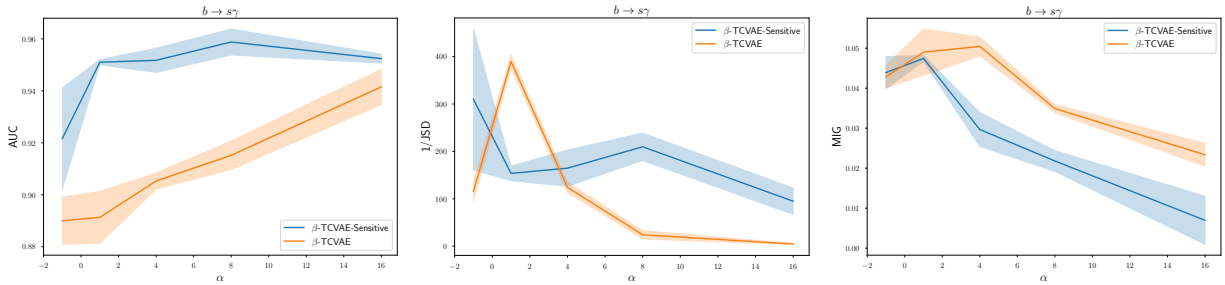


Figure B.2: AUC (left), 1/JSD (middle), and MIG (right) versus α for β-TCVAE (orange) and β-TCVAE-Sensitive (blue). Higher is better for each metric. The shaded area denotes the $\pm 1\sigma$ confidence region.

B.1.3 Total Correlation Weighting

We scan over the weighting β of the total correlation term $\mathbf{E}_{q(z)}\left[\log\left(q(z)/\prod_i q(z_i)\right)\right]$ used in β-TCVAE and Factor-VAE in Figure B.3. To ensure the objective remains a proper lower bound on the ELBO we only consider β (γ) ≥ 1. Note in Factor-VAE this term is referred to as γ. We observe β-TCVAE exhibits some instability for values of $\beta > 2$. Classification performance for both β-TCVAE-Sensitive and Factor-VAE remain consistent over the range of β considered. Disentanglement, as measured by MIG appears to degrade with increasing β for all three approaches, somewhat counter-intuitively, meriting further attention. We speculate that excessive pressure on the TC term may lead to regression of the distribution of each dimension to the unit Gaussian prior as the easiest pathway to independence, and may be solved by annealing the dimension-wise KL term from zero to a nonzero value over the course of training. We set $\beta = \gamma = 1$ for all further experiments.

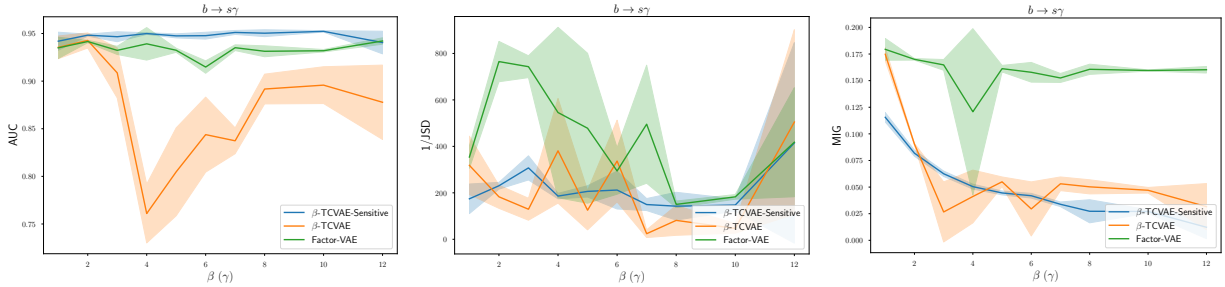


Figure B.3: AUC (left), 1/JSD (middle), and MIG (right) versus β (γ) for β-TCVAE (orange), β-TCVAE-Sensitive (blue) and Factor-VAE (green). Higher is better for each metric. The shaded area denotes the $\pm 1\sigma$ confidence region.

B.1.4 Supervised Regularization Weighting

We scan over the weighting λ_s of the supervised regularization term (Eq. 3.40) intended to match latent dimensions with sensitive factors s in Figure B.4 Due to the number of values considered, each model is only trained for 32 epochs for this particular scan. We observe that the discrimination power, as measured by the AUC appears to be relatively constant across the values of λ_s considered. The extent of mass sculpting as measured by the reciprocal of the JSD (higher is better) increases with λ_s, as does the extent of disentanglement, quantified by the MIG. This makes intuitive sense as increasing the supervised regularization strength should force the sensitive latent dimensions t to be more strongly related to s, while it is not beneficial for other dimensions to be correlated with s due to the pressure placed upon the latent dimensions to be independent. We choose $\lambda = 128$ for all subsequent results as this appears to result in the least amount of distribution shift in M_{bc} after selection without significantly compromising discrimination power.

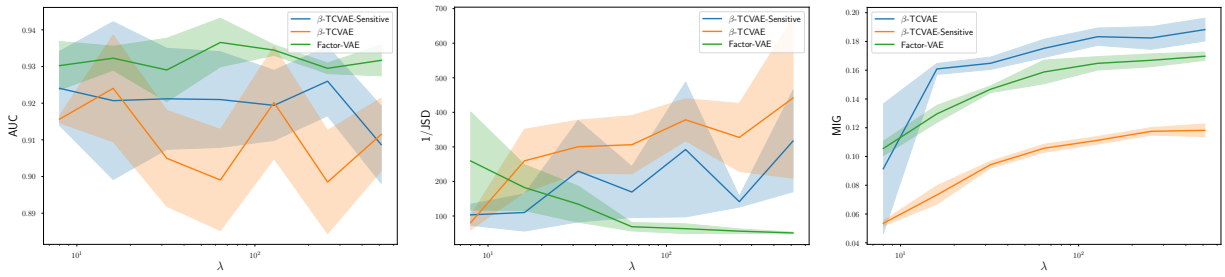


Figure B.4: AUC (left), 1/JSD (middle), and MIG (right) versus β (γ) for β-TCVAE (orange), β-TCVAE-Sensitive (blue) and Factor-VAE (green). Higher is better for each metric.

B.1.5 Hyperparameter Sensitivity

To illustrate hyperparameter sensitivity of each of the proposed methods, we plot the distributions of each metric collected throughout all hyperparameter runs listed in Table B.1, discarding runs with a final AUC below 0.9 (insufficient discrimination) or with a JSD above 0.1 (excessive distribution distortion). We observe that there does not appear to be a monotonic relationship between JSD and MIG, somewhat counter-intuitively. We observe that β-TCVAE-Sensitive and Factor-VAE exhibit the least variance against changes in hyperparameters. Representations learnt with β-TCVAE-Sensitive result in the best discrimination power, as measured with AUC, while representations learnt with Factor-VAE are the most disentangled, as measured by MIG, and result in the least amount of distribution shift, as measured by JSD. Representations obtained under Annealed-VAE and β-VAE exhibit substandard discrimination power and invariance to changes in M_{bc}, unlike methods trained to explicitly minimize the total correlation term $\log\left(q(z)/\prod_i q(z_i)\right)$.

B.2 Signal/Background M_{bc} Plots

Here we plot the receiver operating characteristic curve and visualize the background and signal distributions before and after selection for each of the considered methods from Section 3.5.2.

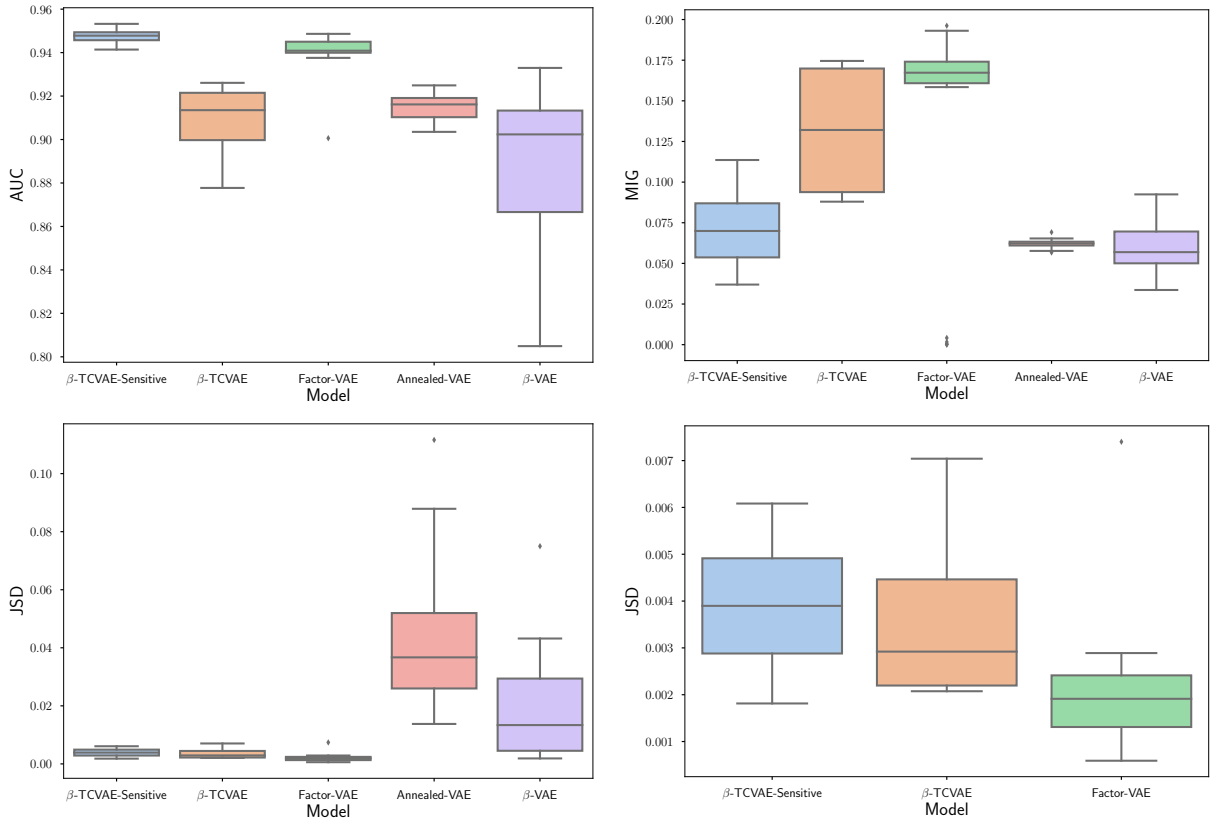


Figure B.5: Distribution of metrics obtained for each model over all hyperparameter runs. Runs with AUC below 0.9 and JSD above 0.1 are treated as failed runs and discarded. Clockwise from top left: AUC score, MIG score, JSD score restricted to β-TCVAE-Sensitive, β-TCVAE and Factor-VAE, JSD Score. Higher is better for AUC and MIG, lower is better for JSD.

B.2.1 Using the Full Latent Representation

We attempted to train a disentangling model using the supervised objective (Eq. 3.58) and use the full latent representation z for the downstream instead of the representation with sensitive latents omitted, $z \setminus t$. We observe an unacceptable level of backround distortion after selection, similar to the baseline discriminator. We report the JSD and AUC metrics for this case in Table B.2.

Model	AUC	JSD
β-TCVAE-Sensitive	0.981 ± 0.004	0.373 ± 0.007
β-TCVAE	0.961 ± 0.001	0.38 ± 0.04
Factor-VAE	0.977 ± 0.007	0.3 ± 0.1
Annealed-VAE	0.92 ± 0.03	0.17 ± 0.02
β-VAE	0.94 ± 0.03	0.3 ± 0.2

Table B.2: AUC and JSD metrics for the case when the full latent representation z is used as input for the downstream discriminator, as opposed to the invariant representation $z \setminus t$ obtained by removing sensitive latents t. Shown is the mean and standard deviation over 3 trials with different random seeds.

All experiments presented in this chapter were run on a node with 24 Intel Xeon E5-2650 v4 processors and 4 NVIDIA Tesla P100 GPUs unless indicated otherwise. The described algorithms in Chapter 3 are implemented using the Pytorch framework (Paszke et al., 2019). The source code for Chapter 3 will be made available at `https://github.com/Justin-Tan/disentangled_invariant_reps` under the Apache License 2.0 after thesis publication.

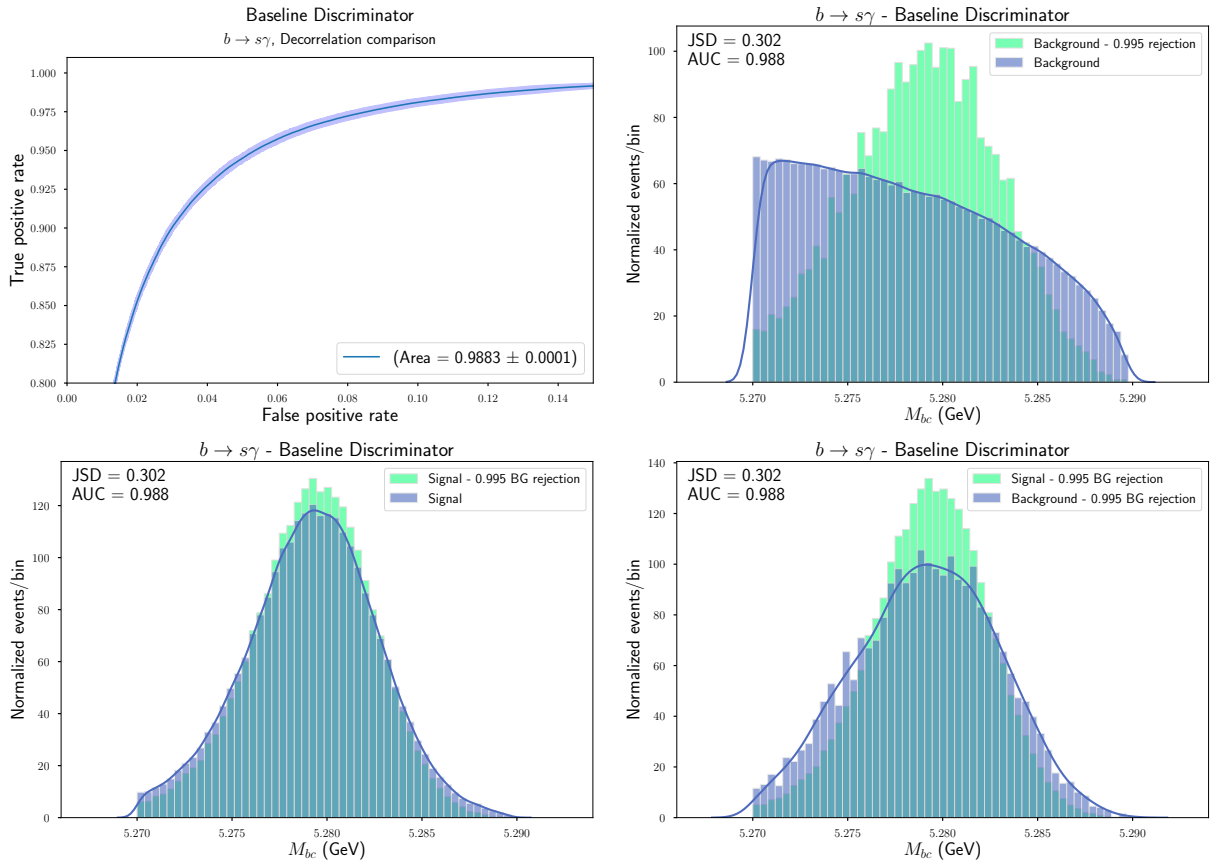


Figure B.6: Results of signal selection by the baseline discriminator at 99.5% background rejection level. Top left: Receiver operating characteristic (ROC) curve, Top right: Background distribution before and after rejection of 99.5% of background events, exhibiting severe sculpting. Bottom left: Signal distribution before and after rejection of 99.5% of background events. Bottom right: Signal and background distributions after rejection of 99.5% of background events,

B.3 Observed Dataset

The observed high-energy physics data x used when learning disentangled representations consists of 34 continuous quantities, which form a subset of the full dataset described in detail in Appendix A.2. We briefly describe each input variable and plot the corresponding distributions for both signal $b \to s\gamma$ events as well as misreconstructed background events below. As previously, all kinematic quantities are evaluated in the $\Upsilon(4S)$ center of mass (CMS) frame and each variable is normalized to zero mean and unit variance.

Kinematic / Angular Variables

- H: The cosine of the helicity angle between the candidate B and the radiative photon. Defined as the angle between the line defined by the momentum difference of the hadronic system and photon in the rest frame of the B and the momentum of the γ in the lab frame.
- R_2, The ratio of the second and zeroth order Fox-Wolfram moments.
- T_B: The magnitude of the signal B meson thrust axis.
- T_{ROE}: The magnitude of the rest-of-event (ROE) - all unreconstructed particles) thrust axis.
- The angular distributions $(\cos(\theta), \phi)$ of the reconstructed B meson and radiative photon candidate, respectively.
- $p_T(\gamma)$: Tranverse momentum of photon candidate.
- $\cos\left(\theta_{T_{B-ROE}}\right)$: Cosine of the angle between the thrust axis of the signal B and rest of event (ROE) thrust axis.
- $\cos\left(\theta_{T_{B-z}}\right)$: Cosine of the angle between the thrust axis of the signal B and the z axis.
- γ-E recoil: Energy recoil against candidate photon.

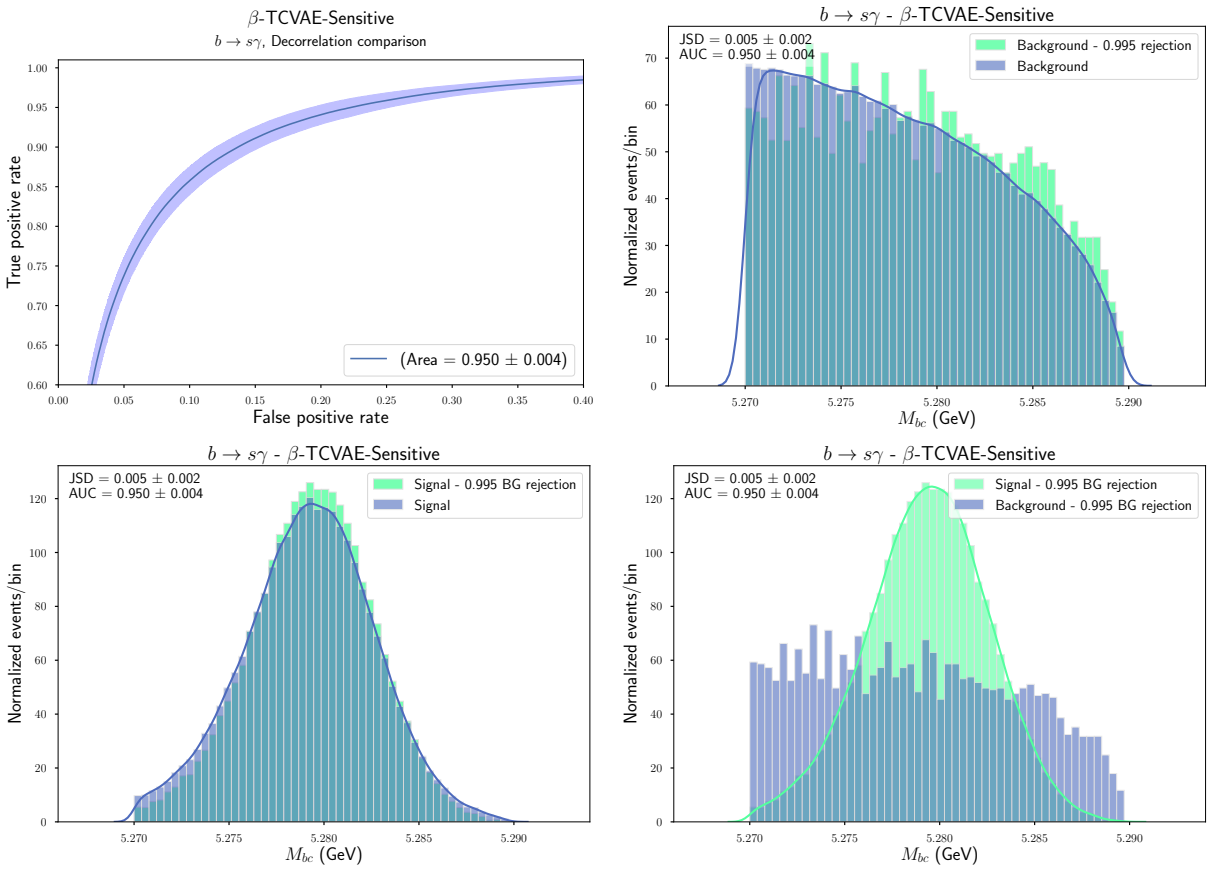


Figure B.7: Results of signal selection with representation obtained by β-TCVAE-Sensitive at 99.5% background rejection level. Top left: Receiver operating characteristic (ROC) curve, Top right: Background distribution before and after rejection of 99.5% of background events. Bottom left: Signal distribution before and after rejection of 99.5% of background events. Bottom right: Signal and background distributions after rejection of 99.5% of background events,

- γ-m^2 recoil: Invariant mass squared of system recoiling against candidate photon.
- γ-p recoil: Magnitude of 3-momentum recoiling against candidate photon.
- $\max_{d \in D}\ E_d$: Maximum energy of B meson daughters.
- $\max_{d \in D}\ p_d$: Maximum 3-momentum magnitude of B meson daughters.
- $\max_{d \in D}\ (p_T)_d$: Maximum traverse momentum magnitude of B meson daughters.
- $\max_{d \in D}\ (p_x)_d$: Maximum p_x magnitude of B meson daughters.
- $\max_{d \in D}\ (p_y)_d$: Maximum p_y magnitude of B meson daughters.
- $\max_{d \in D}\ (p_z)_d$: Maximum p_z magnitude of B meson daughters.
- $\max_{d \in D}\ p_{xd}$: Maximum p_x magnitude of B meson daughters.
- $\sum_{d \in D}\ (p_T)_d$: The sum of the traverse momentum over all daughters of the B meson.
- $\sum_{d \in D}\ (p_x)_d$: The sum of p_x over all daughters of the B meson.
- $\sum_{d \in D}\ (p_y)_d$: The sum of p_y over all daughters of the B meson.
- $\sum_{d \in D}\ (p_z)_d$: The sum of p_z over all daughters of the B meson.

Electromagnetic calorimeter (ECL) response variables

- δE_γ: Error in measured energy of photon candidate detected by the ECL.
- $E_{CMS}\ (\gamma)$: Photon CMS energy.
- E_{cluster}: Cluster energy, given by the weighted sum of all ECL crystal energies within the ECL cluster, corrected for leakage and background contamination.

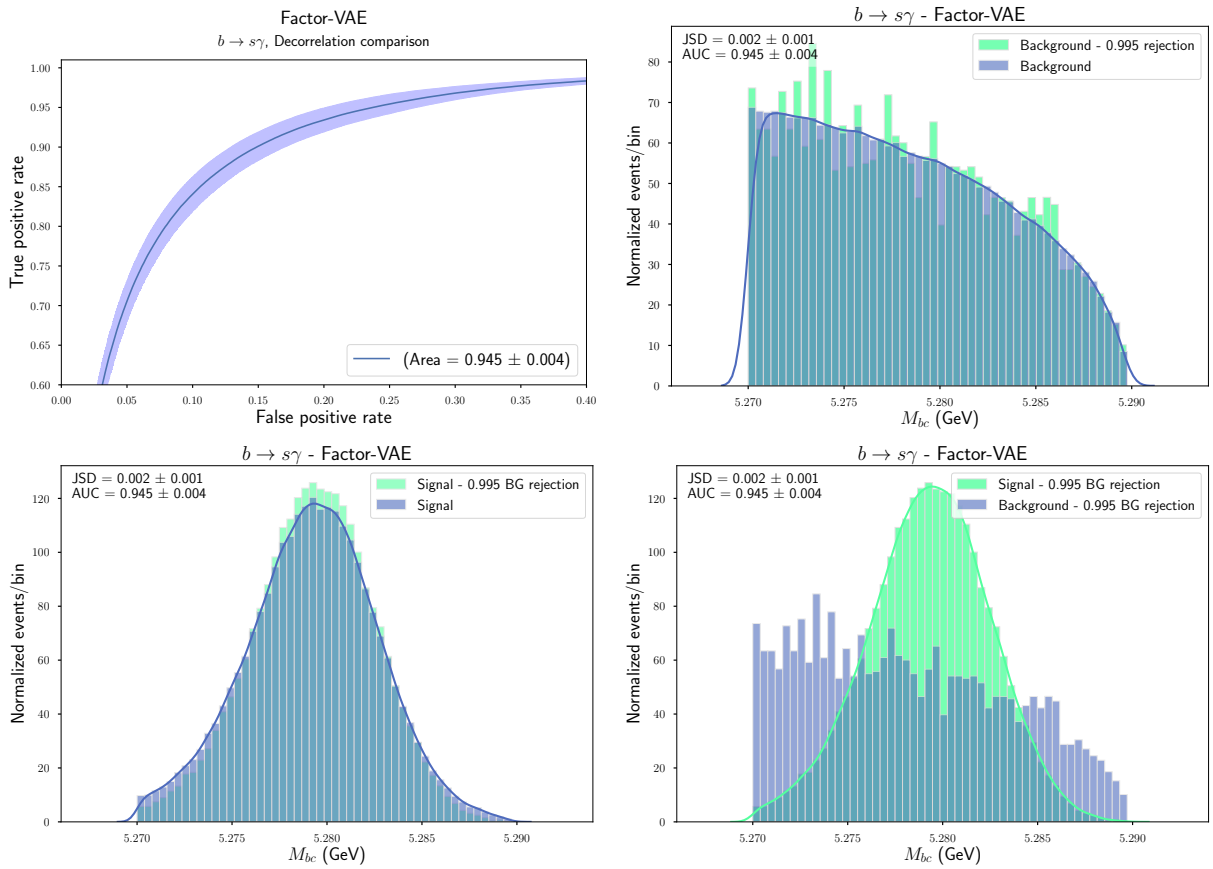


Figure B.8: Results of signal selection with representation obtained by Factor-VAE at 99.5% background rejection level. Top left: Receiver operating characteristic (ROC) curve, Top right: Background distribution before and after rejection of 99.5% of background events. Bottom left: Signal distribution before and after rejection of 99.5% of background events. Bottom right: Signal and background distributions after rejection of 99.5% of background events,

- The absolute values of the $(4,0)$ ($|Z_{40}|$) and $(5,1)$ ($|Z_{51}|$) Zernike moments, used as shape descriptors of the electromagnetic showers generated in the ECL crystals by incident particles.
- $E1E9$: Ratio of energy deposited in the central cluster cell seed to the total energy deposited in the inner 3×3 crystals around the central crystal.
- $E9E21$: Ratio of energies deposited in the central inner 3×3 and outer 5×5 (excluding corners) crystals around the central crystal. Both ratios $E1E9$, $E9E21$ are close to 1 for true photon candidates but lower for asymmetric hadronic shower events.
- E_{max} Cluster: Energy of the most energetic crystal in the ECL.
- Cluster $\cos\theta$: Cosine of polar angle of the centroid of the ECL cluster.
- Cluster ϕ: Azimuthal angle of the centroid of the ECL cluster.
- δE Cluster: Uncertainty in ECL cluster energy.
- Cluster E_{uncal}: Uncalibrated energy of ECL cluster.
- Cluster LAT: Lateral energy distribution of cluster. It is defined as the following:

$$S = \frac{\sum_{i=3}^{n} w_i E_i r_i^2}{\sum_{i=3}^{n} w_i E_i r_i^2 + w_0 E_0 r_0^2 + w_1 E_1 r_0^2}. \tag{B.1}$$

  Where E_i are the single crystal energies sorted by energy, w_i are the crystal weights, r_i is the distance of the i-th digit to the shower center projected to a plane perpendicular to the shower axis, and $r_0 \approx 5$ cm is the distance between two crystals. This quantity peaks around 0.3 for radially symmetrical electromagnetic showers and is larger for hadronic events.

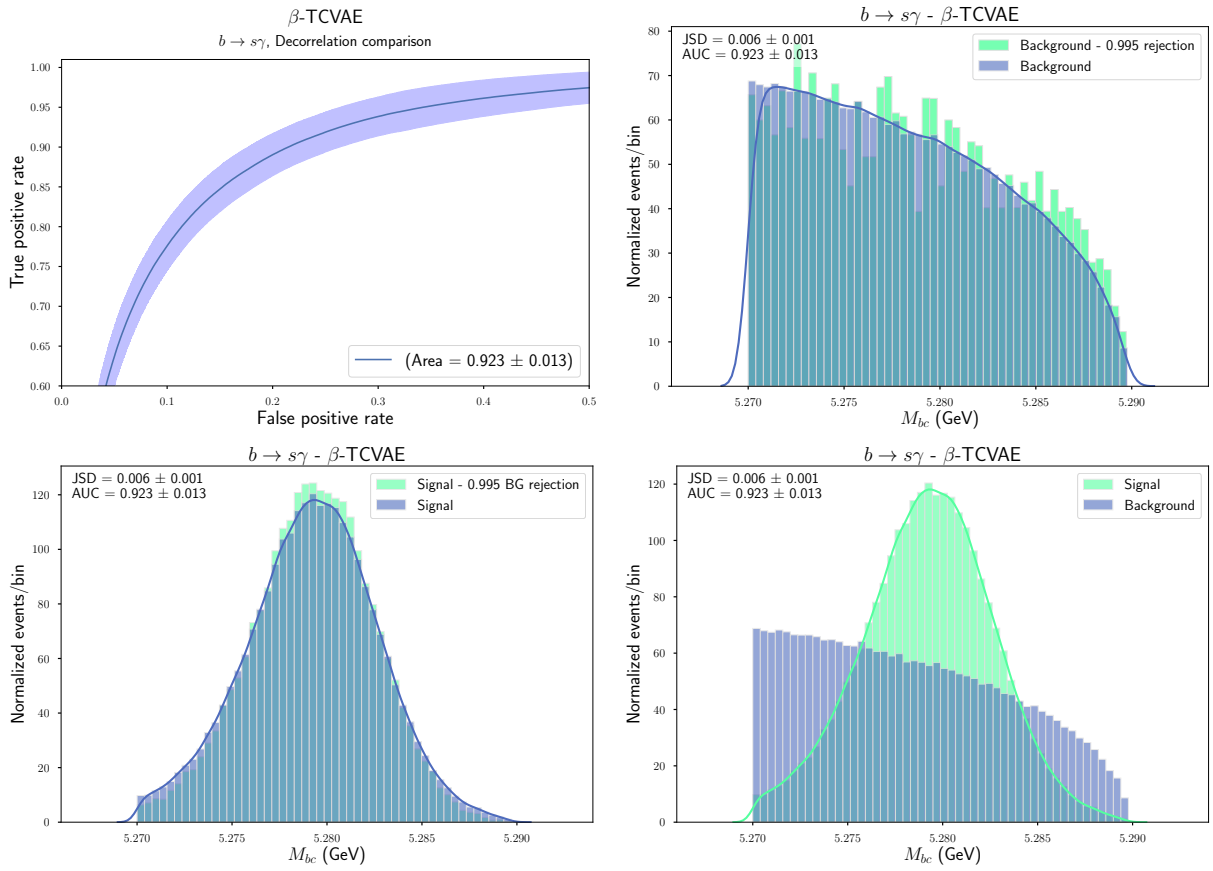
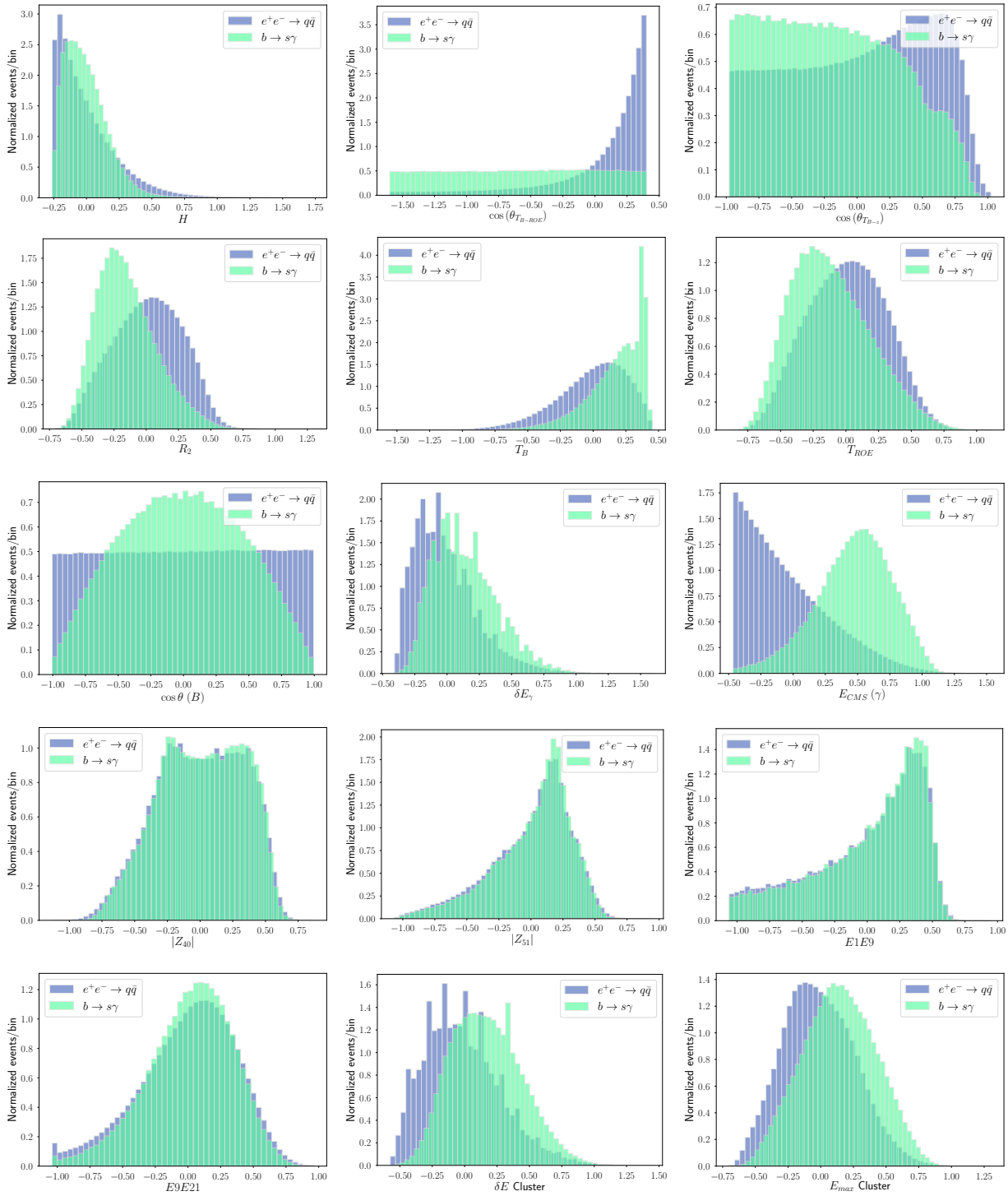
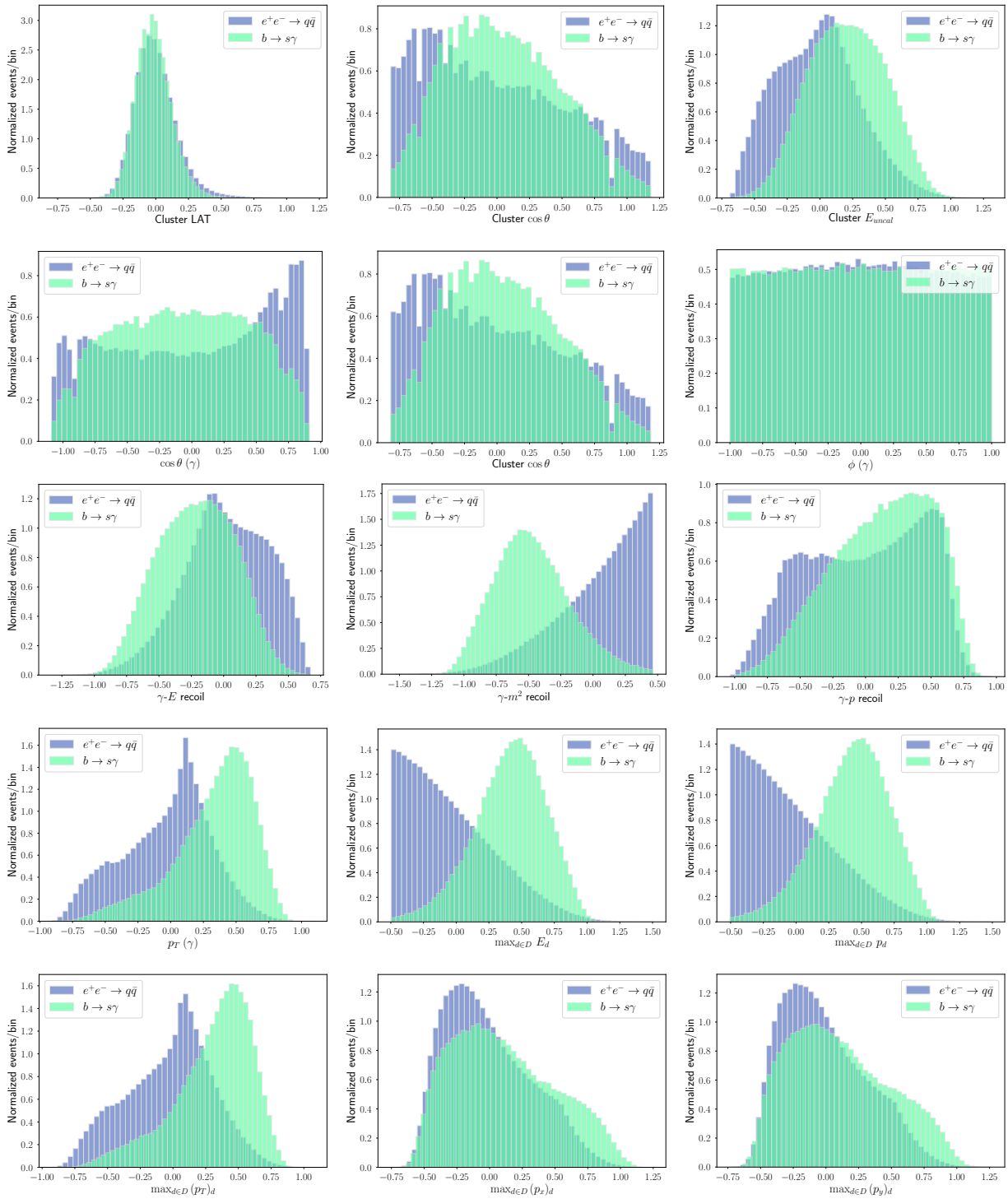


Figure B.9: Results of signal selection with representation obtained by β-TCVAE at 99.5% background rejection level. Top left: Receiver operating characteristic (ROC) curve, Top right: Background distribution before and after rejection of 99.5% of background events. Bottom left: Signal distribution before and after rejection of 99.5% of background events. Bottom right: Signal and background distributions after rejection of 99.5% of background events,


$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Normalized events/bin
H
$\cos(\theta_{T_{B-ROE}})$
$\cos(\theta_{T_{B-z}})$
R_2
T_B
T_{ROE}
$\cos\theta\ (B)$
δE_γ
$E_{CMS}\ (\gamma)$
$|Z_{40}|$
$|Z_{51}|$
$E1E9$
$E9E21$
δE Cluster
E_{max} Cluster


$e^+e^- \to q\bar{q}$
$b \to s\gamma$
Normalized events/bin
Cluster LAT
Cluster $\cos\theta$
Cluster E_{uncal}
$\cos\theta$ (γ)
ϕ (γ)
γ-E recoil
γ-m^2 recoil
γ-p recoil
p_T (γ)
$\max_{d\in D} E_d$
$\max_{d\in D} p_d$
$\max_{d\in D} (p_T)_d$
$\max_{d\in D} (p_x)_d$
$\max_{d\in D} (p_y)_d$

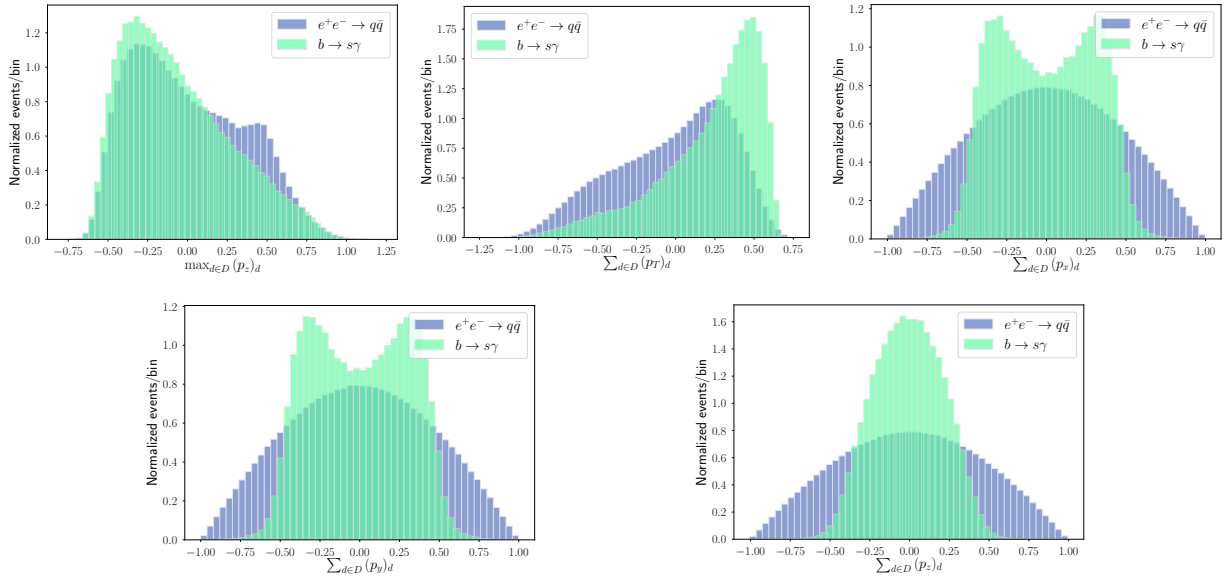


Figure B.10: Marginal distributions of observed data x