INNF+ Workshop 2020

Phillip Urquijo, *Justin Tan*

Decorrelation via Disentanglement

October 26, 2020

Background Sculpting

Classifier output $f(X;\theta_f) \sim$ p(signal|data). Only accept events above a given probability.

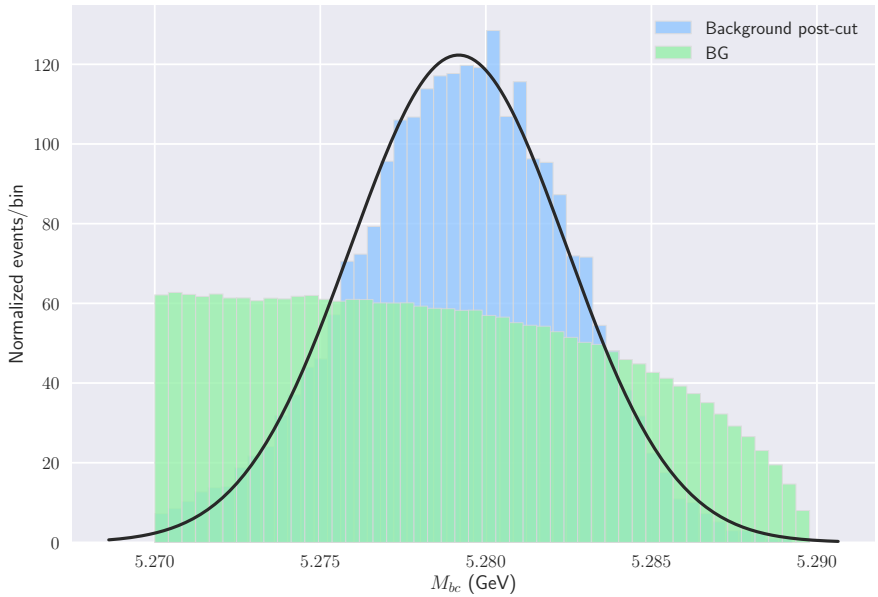


Figure 1: Continuum M_{bc} before (green) and after (blue) 0.995 suppression

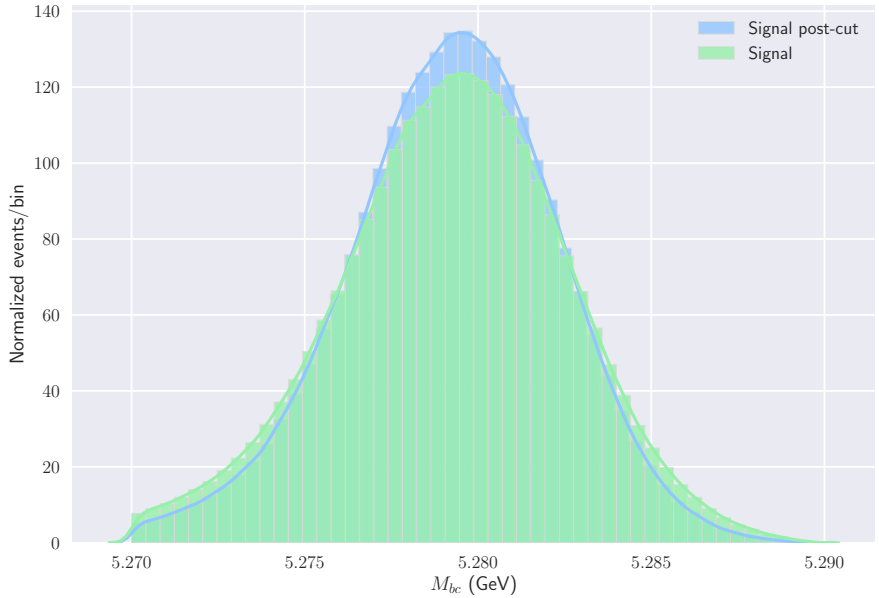


Figure 2: Signal M_{bc} before (green) and after (blue) 0.995 suppression

e.g. invariant mass. Background looks like signal post-selection - difficult to quantify background contribution via mass sidebands.

Decorrelation

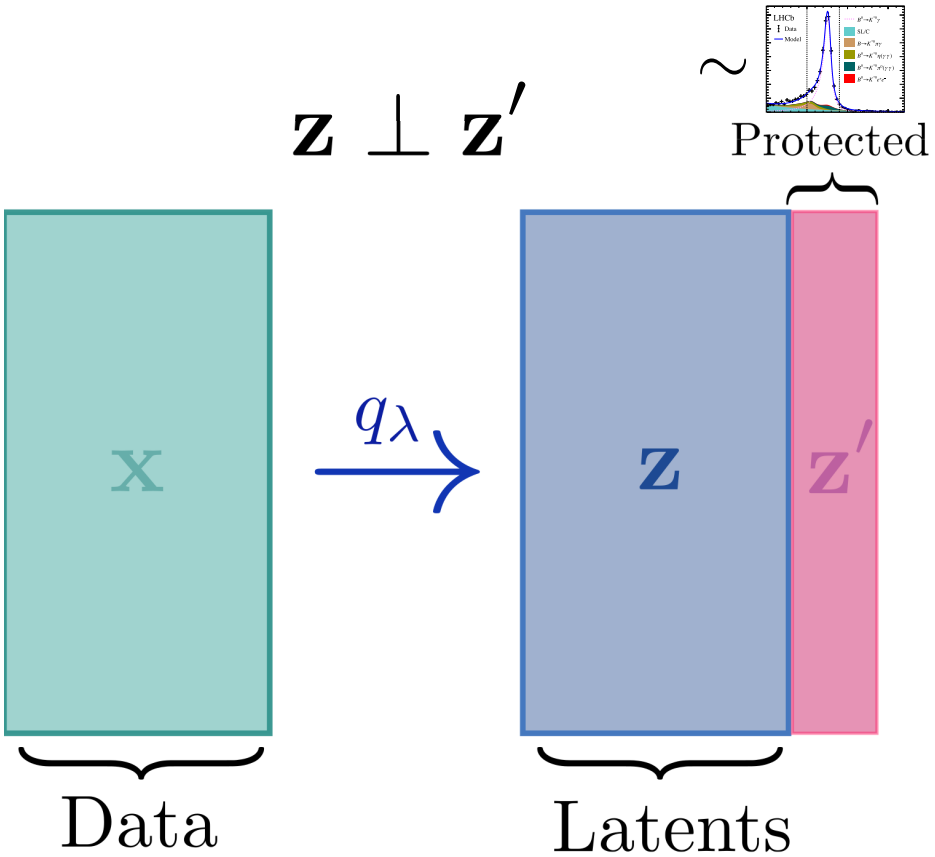
- Most decorrelation approaches:

$$\mathcal{L}_{Decorr.} = \mathcal{L}_{Clf.} + \lambda \cdot \mathcal{L}_{?} \tag{1}$$

- $\mathcal{L}_{?}$: summarizes dependency of classifier output on protected variables.
- We propose to decouple the classification task from the decorrelation process.
- Two-stage procedure:
  - First, learn a representation z non-informative of protected variables.
  - Use representation in lieu of original data x downstream.

Disentanglement

- Want to find representation of event data that is:
  - Non-informative of physical observables (e.g. M_{inv}).
  - Maximally informative of the original data.
- Propose to isolate influence of protected variables in a subspace of latent representation.



Latent Variable Models

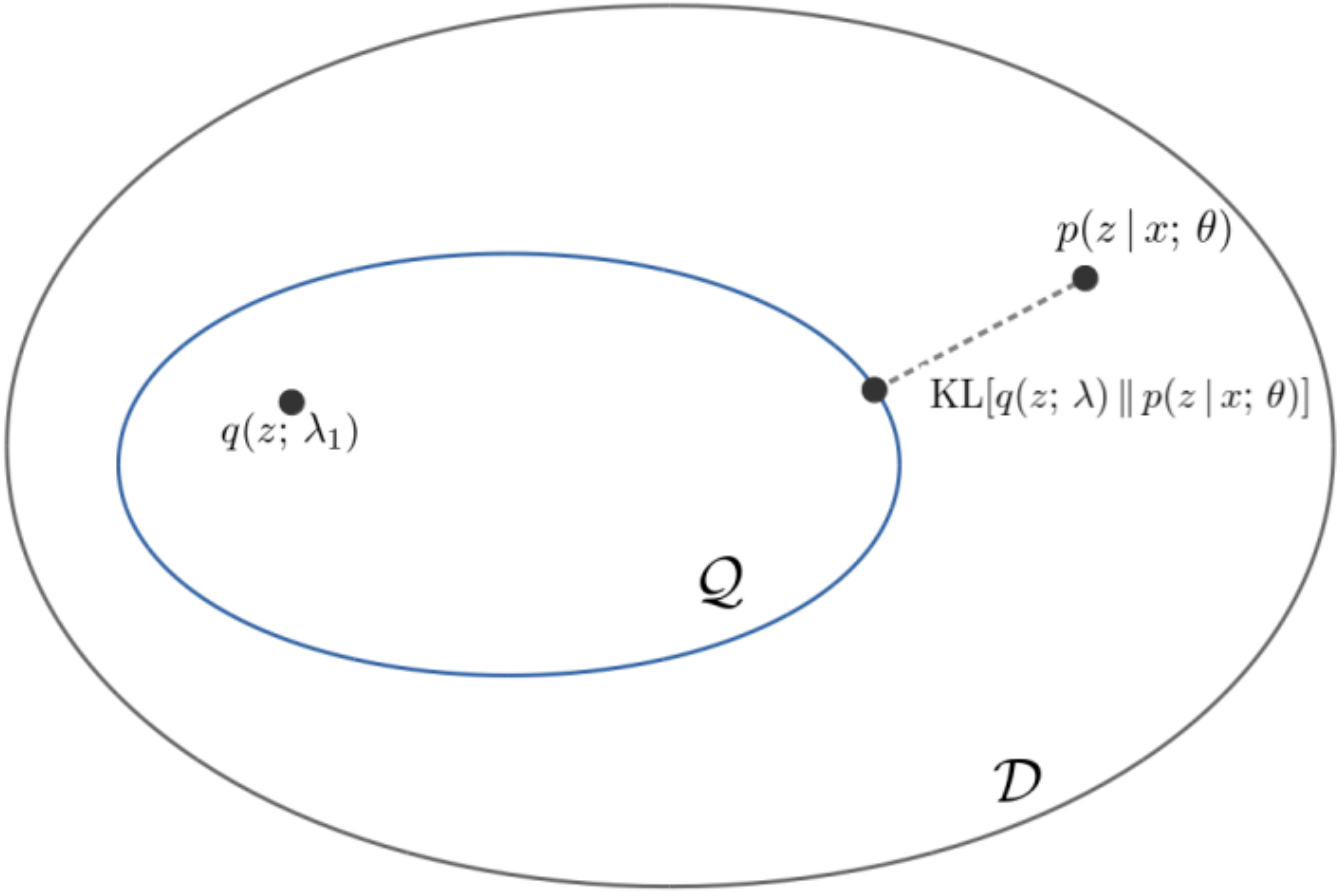
- Want compact description of observed data x.
- Introduce hidden/latent variables z, joint density $p_\theta(x, z) = p_\theta(x|z)p(z)$.
- z provides low-dimensional representation of x.
- Want to infer latents z responsible for generating x:

$$p_\theta(z|x) = \frac{p_\theta(x, z)}{\int_z dz\, p_\theta(x|z)p(z)} \tag{2}$$

- Problem: quadrature takes exponential time in dim(z), likely non-analytic.

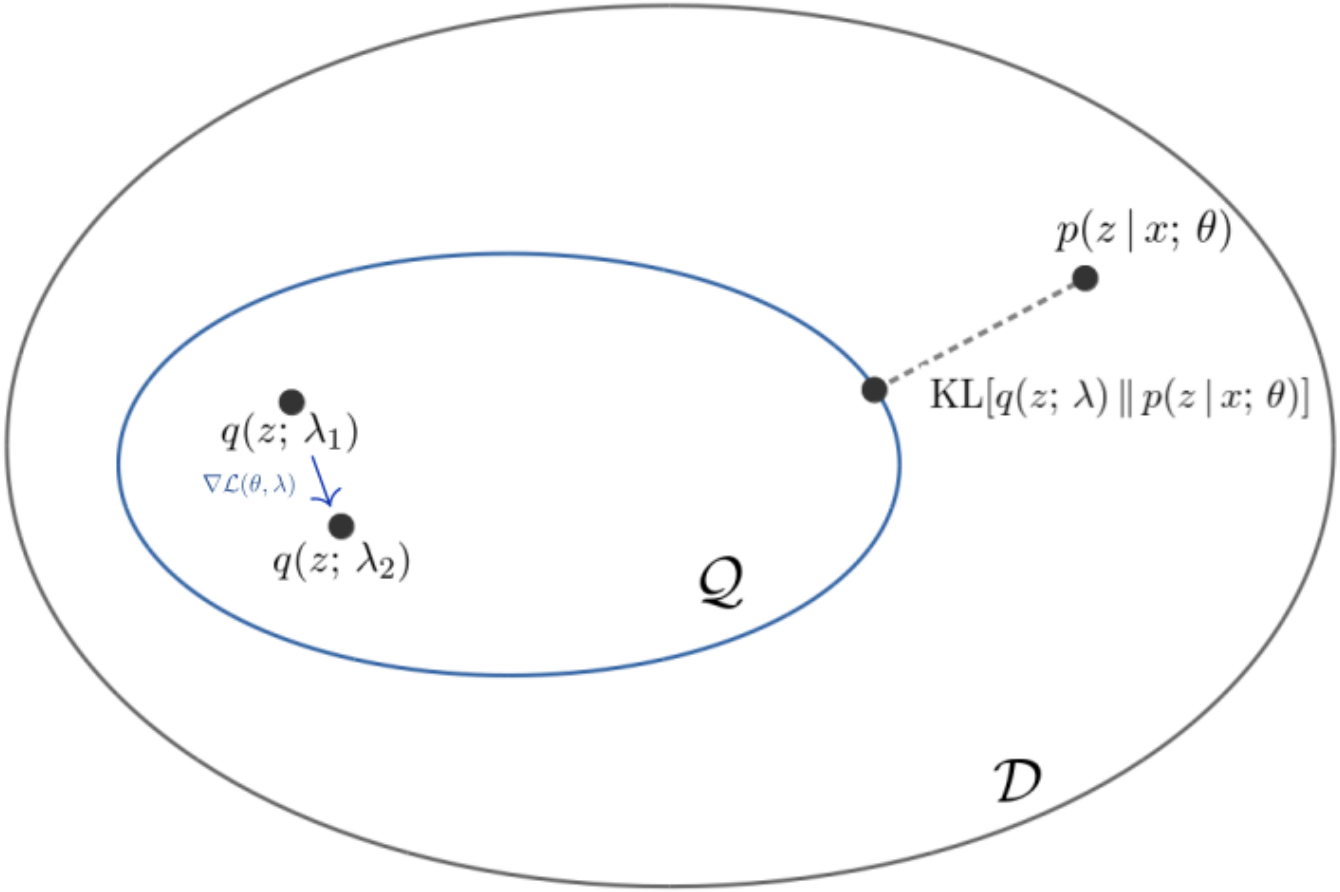
Inference as Optimization

- Approximate true $p_\theta(z|x)$ with $q_\lambda(z;x) \in \mathcal{Q}$.
  - $\mathcal{Q}$: 'Simple' family of parametric distributions.
- Minimize 'distance' $D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right)$.
- Inference (hard) $\approx$ optimization (easy).



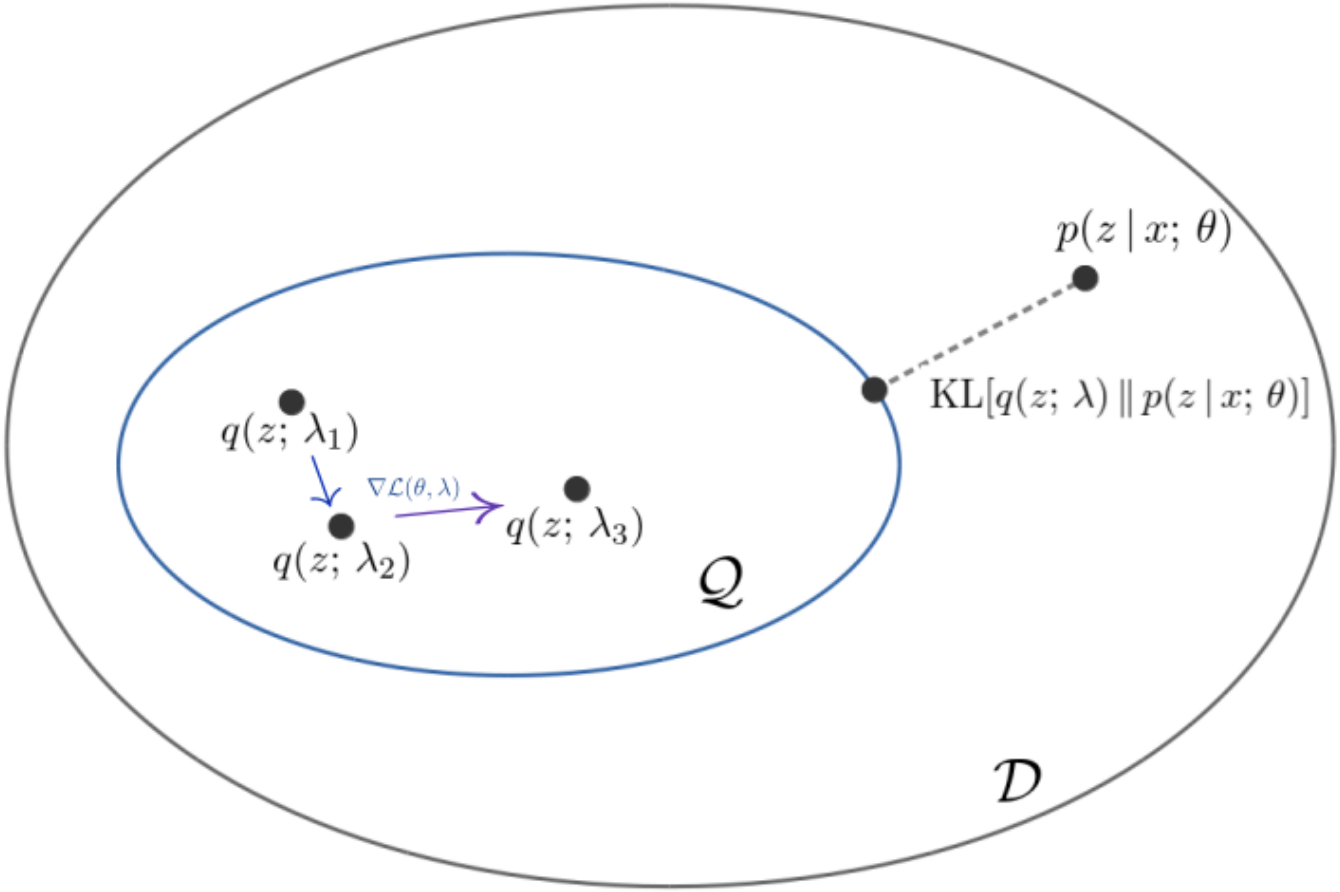
Inference as Optimization

- Approximate true $p_\theta(z|x)$ with $q_\lambda(z;x) \in \mathcal{Q}$.
  - $\mathcal{Q}$: 'Simple' family of parametric distributions.
- Minimize 'distance' $D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right)$.
- Inference (hard) $\approx$ optimization (easy).



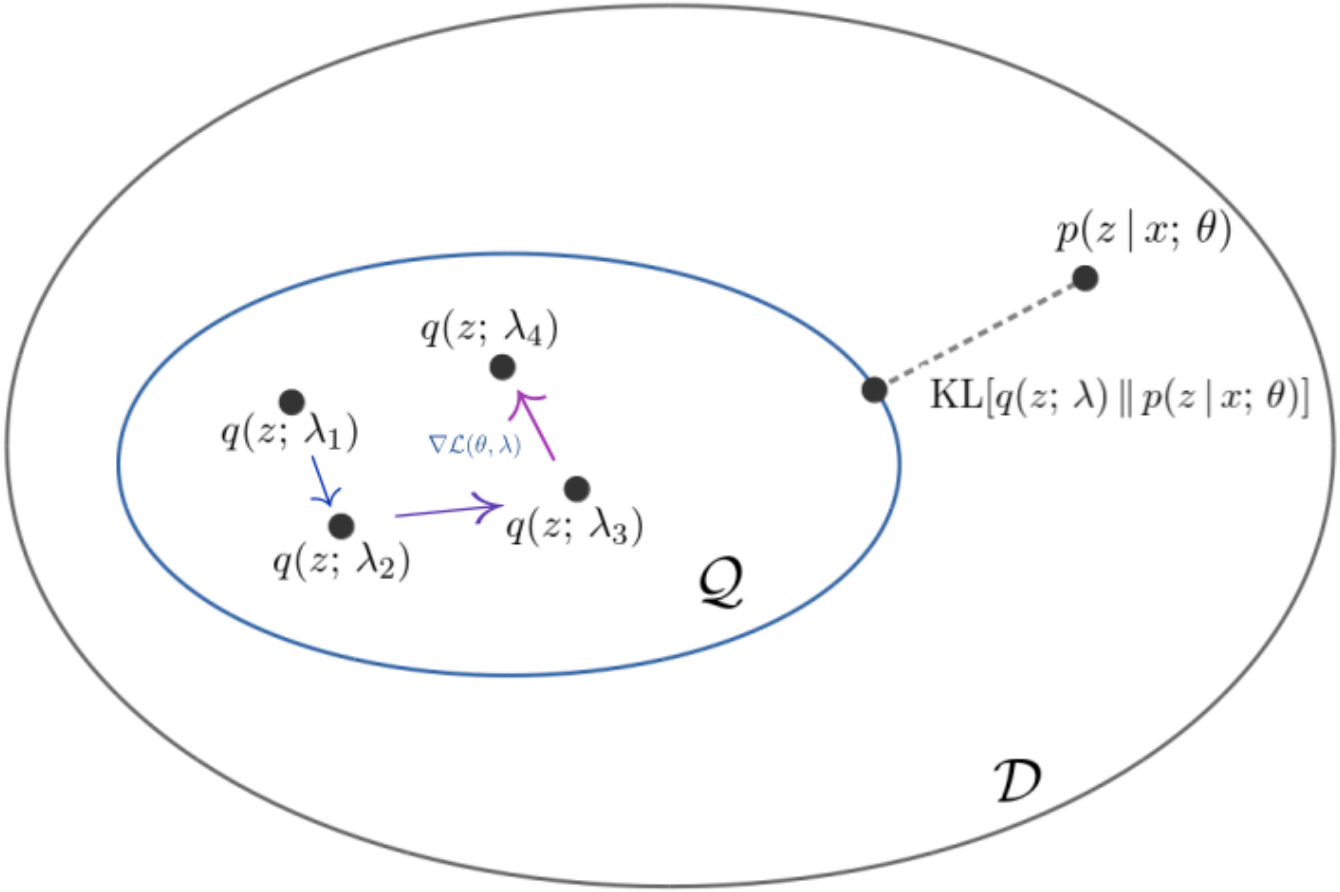
Inference as Optimization

- Approximate true $p_\theta(z|x)$ with $q_\lambda(z;x) \in \mathcal{Q}$.
  - $\mathcal{Q}$: 'Simple' family of parametric distributions.
- Minimize 'distance' $D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right)$.
- Inference (hard) $\approx$ optimization (easy).



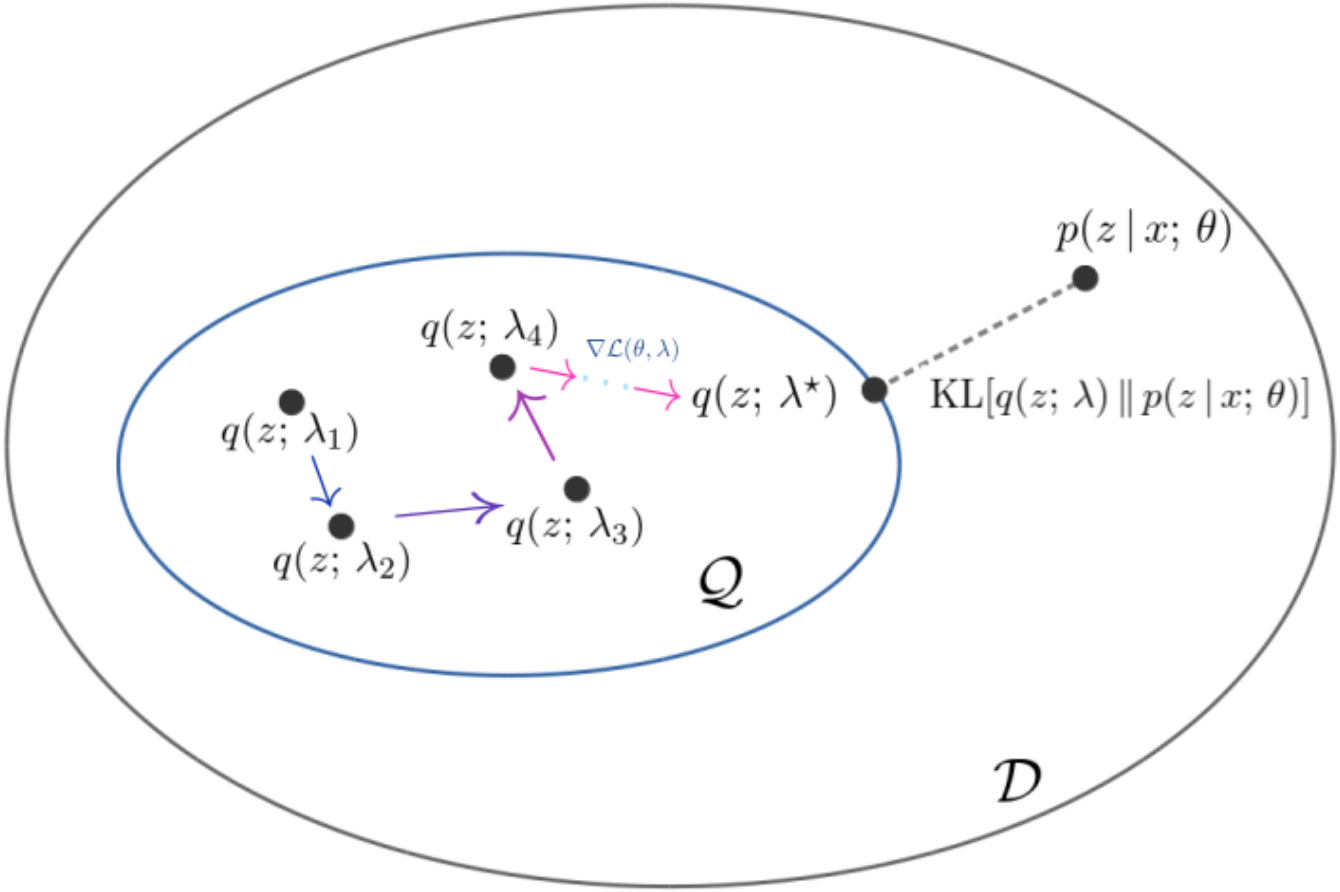
Inference as Optimization

- Approximate true $p_\theta(z|x)$ with $q_\lambda(z;x) \in \mathcal{Q}$.
  - $\mathcal{Q}$: 'Simple' family of parametric distributions.
- Minimize 'distance' $D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right)$.
- Inference (hard) $\approx$ optimization (easy).



Inference as Optimization

- Approximate true $p_\theta(z|x)$ with $q_\lambda(z;x) \in \mathcal{Q}$.
  - $\mathcal{Q}$: 'Simple' family of parametric distributions.
- Minimize 'distance' $D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right)$.
- Inference (hard) $\approx$ optimization (easy).



Inference as Optimization

- Equivalent to maximization of Evidence Lower BOund (ELBO):

$$\mathrm{ELBO}(x) \triangleq \mathbf{E}_{q_\lambda(z;x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\lambda(z;x)}\right] \leq \log p_\theta(x) \tag{3}$$

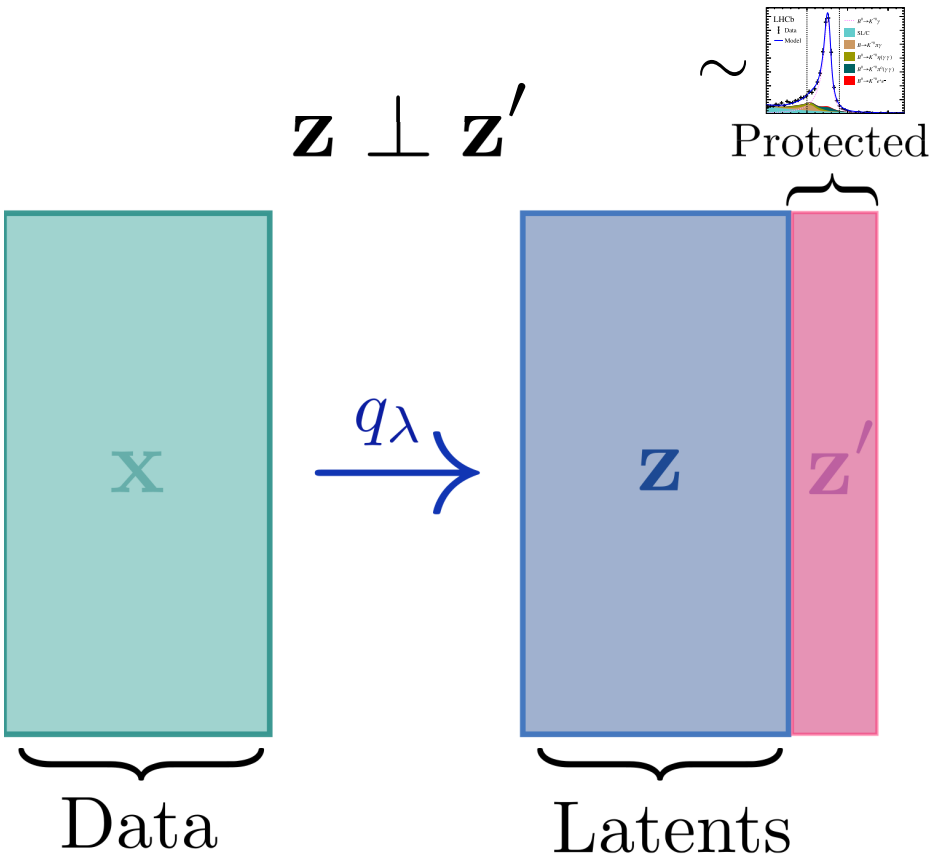
$$\min_{q_\lambda \in \mathcal{Q}} D_{\mathrm{KL}}\left(q_\lambda(z;x) \,\|\, p_\theta(z|x)\right) \iff \max_{\theta,\lambda} \mathrm{ELBO}(x) \tag{4}$$

- Variational Autoencoding (VAEs) framework: q_λ and conditional model p_θ parameterized by neural networks.
  - Instead of generation, interested in inference.

Disentanglement

- Different components of z should capture distinct semantic information about x.
- How to encourage independence between latent dimensions?
- Try to factorize average joint variational posterior over z, z':

$$\mathbf{E}_{p(x)}\left[q_\lambda(z, z'; x)\right] \approx \mathbf{E}_{p(x)}\left[q_\lambda(z; x)\right] \mathbf{E}_{p(x)}\left[q_\lambda(z'; x)\right]. \tag{5}$$



ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\mathbf{E}_{p(\mathrm{x})}\left[\mathrm{ELBO}(\mathrm{x})\right] = \mathbf{E}_{p(\mathrm{x})}\left[\mathbf{E}_{q_\lambda(\mathrm{z}|\mathrm{x})}\left[\log p_\theta(\mathrm{x}|\mathrm{z})\right] - D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right]$$

$$\begin{aligned}
\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right] &= \mathbb{I}_q(x;z) + D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) \\
&= \underbrace{\mathbb{I}_q(x;z)}_{(1)} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j)\right)}_{(2)} \\
&\quad + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i)\right)}_{(3)}
\end{aligned}$$

(1) Observed-latent mutual information.

ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\mathbf{E}_{p(\mathrm{x})}\left[\mathrm{ELBO}(\mathrm{x})\right] = \mathbf{E}_{p(\mathrm{x})}\left[\mathbf{E}_{q_\lambda(\mathrm{z}|\mathrm{x})}\left[\log p_\theta(\mathrm{x}|\mathrm{z})\right] - D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right]$$

$$\begin{aligned}
\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right] &= \mathbb{I}_q(x;z) + D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) \\
&= \underbrace{\mathbb{I}_q(x;z)}_{(1)} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j)\right)}_{(2)} \\
&\quad + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i)\right)}_{(3)}
\end{aligned}$$

(1) Observed-latent mutual information.

(2) Latent Total Correlation.

ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\mathbf{E}_{p(\mathrm{x})}\left[\mathrm{ELBO}(\mathrm{x})\right] = \mathbf{E}_{p(\mathrm{x})}\left[\mathbf{E}_{q_\lambda(\mathrm{z}|\mathrm{x})}\left[\log p_\theta(\mathrm{x}|\mathrm{z})\right] - D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right]$$

$$\begin{aligned}
\mathbf{E}_{p(x)}\left[D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)\right] &= \mathbb{I}_q(x;z) + D_{\mathrm{KL}}\left(q(z) \,\|\, p(z)\right) \\
&= \underbrace{\mathbb{I}_q(x;z)}_{①} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j)\right)}_{②} \\
&\quad + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i)\right)}_{③}
\end{aligned}$$

① Observed-latent mutual information.

② Latent Total Correlation.

③ Dimension-wise KL.

ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\min_{q_\lambda \in \mathcal{Q}} \left(\underbrace{\mathbb{I}_q(x; z)}_{(1)} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j) \right)}_{(2)} + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i) \right)}_{(3)} \right)$$

Aggregate Posterior: $q(z) = \frac{1}{N} \sum_n q_\lambda(z; x_n)$ empirical average of $q_\lambda(z; x_n)$.

(1) Observed-latent mutual information.

(1) Predictiveness of observed data from latents.

ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\min_{q_\lambda \in \mathcal{Q}} \left(\underbrace{\mathbb{I}_q(x; z)}_{(1)} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j) \right)}_{(2)} + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i) \right)}_{(3)} \right)$$

Aggregate Posterior: $q(z) = \frac{1}{N} \sum_n q_\lambda(z; x_n)$ empirical average of $q_\lambda(z; x_n)$.

(1) Observed-latent mutual information.

(2) Latent Total Correlation.

(2) Statistical dependency between dimensions of z.

ELBO Decomposition

[Hoffman, Johnson, ELBO Surgery, 2016]

$$\min_{q_\lambda \in \mathcal{Q}} \left(\underbrace{\mathbb{I}_q(x; z)}_{(1)} + \underbrace{D_{\mathrm{KL}}\left(q(z) \,\|\, \prod_j q(z_j) \right)}_{(2)} + \underbrace{\sum_i D_{\mathrm{KL}}\left(q(z_i) \,\|\, p(z_i) \right)}_{(3)} \right)$$

Aggregate Posterior: $q(z) = \frac{1}{N} \sum_n q_\lambda(z; x_n)$ empirical average of $q_\lambda(z; x_n)$.

1. Observed-latent mutual information.
2. Latent Total Correlation.
3. Dimension-wise KL.

(3) Complexity penalty on dimensions of $q(z)$.

Disentangling Objective

$$\mathbf{E}_{p(\mathrm{x})}\left[\underbrace{\mathbf{E}_{q_\lambda(\mathrm{z}|\mathrm{x})}\left[\log p_\theta(\mathrm{x}|\mathrm{z})\right] - D_{\mathrm{KL}}\left(q_\lambda(\mathrm{z};\mathrm{x}) \,\|\, p(\mathrm{z})\right)}_{\mathrm{ELBO}(\mathrm{x})}\right]$$
$$- (\beta - 1) D_{\mathrm{KL}}\left(q(\mathrm{z}) \,\|\, \prod_{i=1}^{d} q(z_i)\right)$$

Maximize w.r.t. variational parameters (λ, θ) for $\beta > 1$.

- Total correlation hard to estimate, but multiple (biased) estimators developed:
  - Chen, et. al. Isolating Sources of Disentanglement in VAEs, 2018. (β-TCVAE).
  - Kim, et. al. Disentangling by Factorizing, 2018. (Factor-VAE).

Identifying latent dimensions

Identify selected dimensions z' of the latent code with the protected variables v with cross-entropy-like objective:

$$R_s(\mathrm{x}, \mathrm{v}) = \sum_{i=1}^{|\mathrm{v}|} \left(v_i \log \sigma(r(\mathrm{x})_i) + (1 - v_i) \log \left(1 - \sigma(r(\mathrm{x})_i)\right) \right). \tag{6}$$

- $r(\mathrm{x})$ is mean of the variational posterior q_λ : $r(\mathrm{x}) = \langle q_\lambda(z; x) \rangle$.
- We protect the invariant mass: $v = M_{inv}$.

Final Objective

$$\mathbf{E}_{p(\mathrm{x})}\left[\mathrm{ELBO}(\mathrm{x}) + \lambda_s \mathbf{E}_{p(\mathrm{v}|\mathrm{x})}\left[R_s(\mathrm{x}, \mathrm{v})\right]\right] - (\beta - 1) D_{\mathrm{KL}}\left(q(\mathrm{z}) \,\|\, \prod_{i=1}^{d} q(z_i)\right) \tag{7}$$

- Maximize w.r.t. variational parameters θ, λ.
- Adopt mean of variational posterior as representation:

$$r(\mathrm{x}) = \int_z dz\, z\, q_\lambda(z; x)$$

.

- Excise dimensions z' to yield final representation for use in arbitrary downstream tasks without exploiting protected vs:

$$r(x)' = \langle q_\lambda(z \setminus z'; x) \rangle \tag{8}$$

Only disentangle protected factors

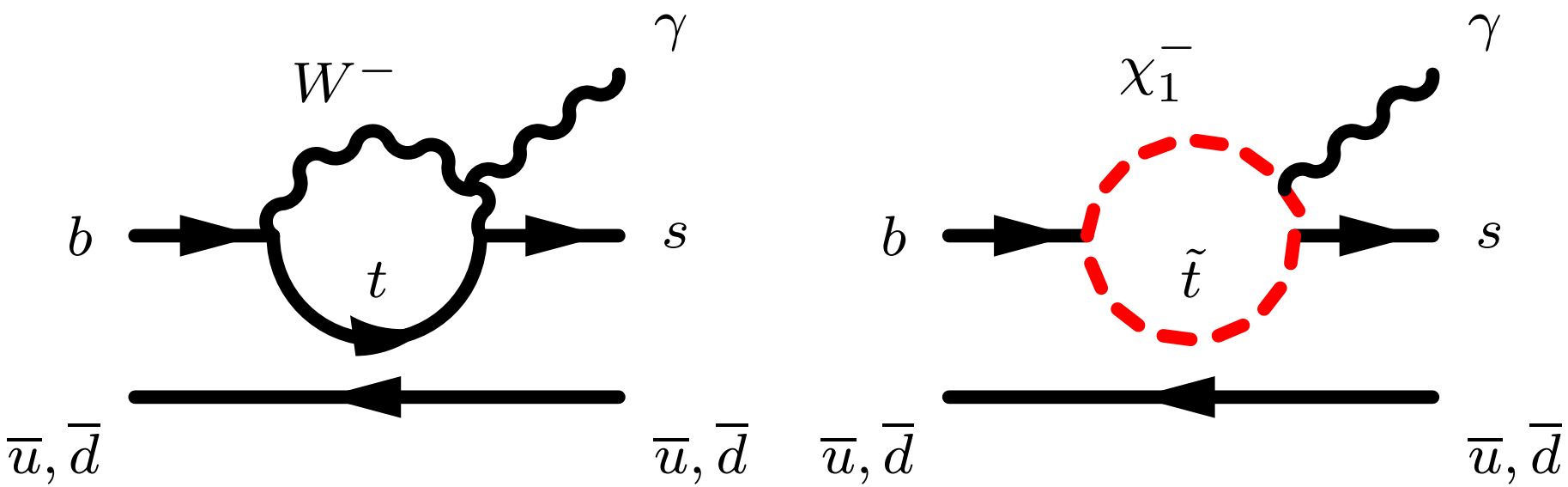
Only disentangle protected dimensions z' (with indices S) instead of full z:

$$\mathcal{L}(\theta, \phi) = \mathbf{E}_{p(x)} \left[\mathbf{E}_{q_\lambda(z;x)} \left[\log p_\theta(x|z) \right] + \lambda_s \mathbf{E}_{p(v|x)} \left[R_s \left(q_\lambda(z; x), v \right) \right] \right]$$
$$- \left(\alpha \mathbb{I}_q(x; z) + \beta \mathbf{E}_{q(z)} \left[\log \frac{q(z)}{q(z \setminus z') \prod_{j \in S} q(z'_j)} \right] \right.$$
$$\left. + \gamma D_{\mathrm{KL}} \left(q(z \setminus z') \,\|\, p(z \setminus z') \right) + \delta \sum_{j \in S} D_{\mathrm{KL}} \left(q(z'_j) \,\|\, p(z'_j) \right) \right)$$

- Each term estimated through stratified importance sampling.
- We call this approach β-TCVAE-Sensitive.

Experiments

- Apply to B-physics analysis of 'penguin' $b \to s\gamma$ transition (BF $\approx 10^{-5}$).
- Background estimation relies on beam-constrained mass spectrum: $M_{bc}^2 = E_{beam}^2 - p_B^2$.
  - Set protected $v = M_{bc}$.
- Use output representation $\langle q_\lambda(z \setminus z'; x) \rangle$ as input to 'downstream' classification task
  - 10^6 events, $1:10$ signal ratio, M_{bc}-correlated features removed.



Experiments

- Quantify results:
  - Disentanglement: Mutual Information Gap.

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

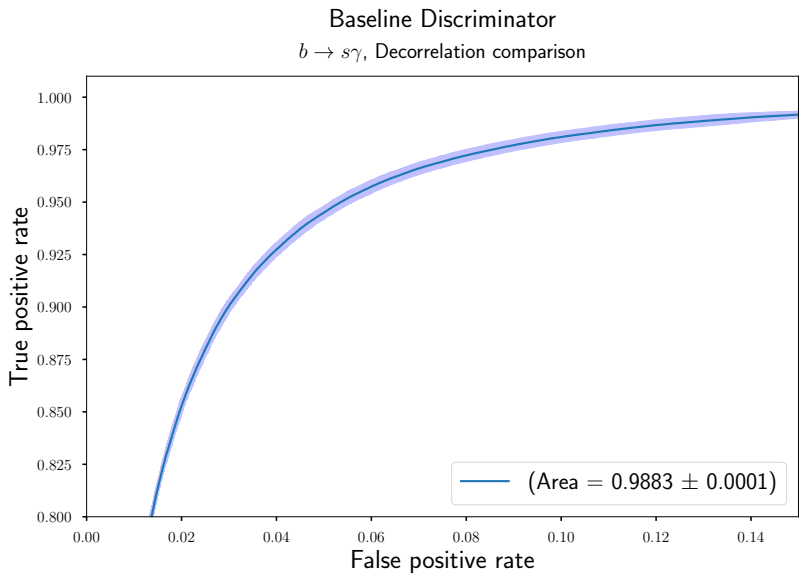
$$j^{(k)} = \operatorname*{argmax}_j I(z_j; v_k)$$

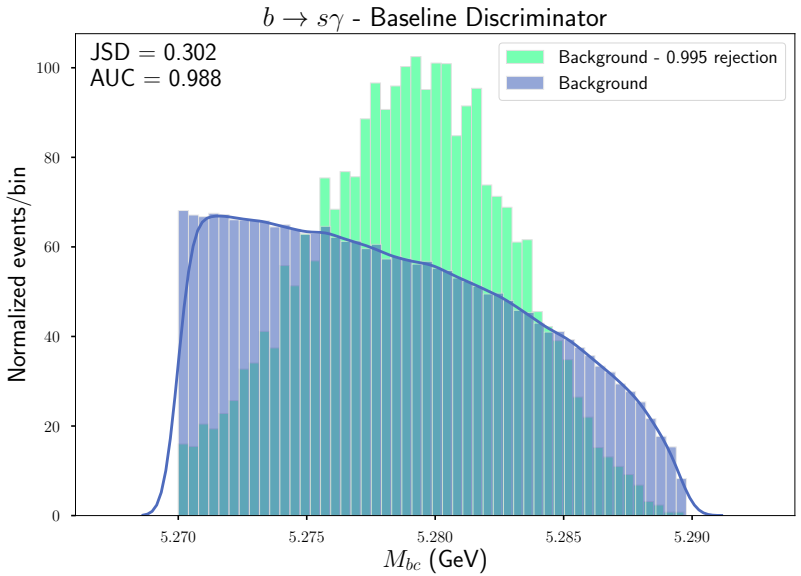
  - Sculpting extent: (Discretized) Jensen-Shannon divergence of background mass spectra before/after 0.995 background rejection.

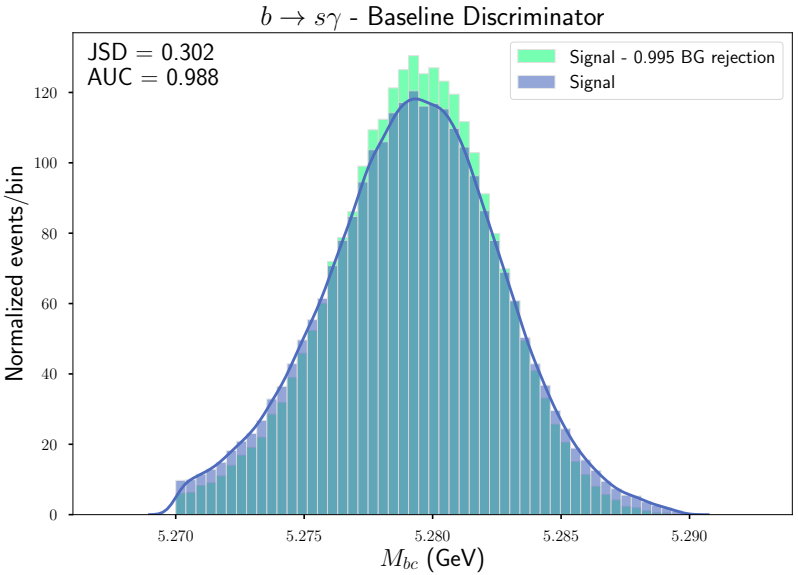
$$\mathrm{JSD} = D_{\mathrm{JS}} \left(N_{\mathrm{bkg}}^{\mathrm{pass}}(m) / \sum_i N_{\mathrm{bkg},i}^{\mathrm{pass},i}(m) \,\middle\|\, N_{\mathrm{bkg}}^{\mathrm{fail}}(m) / \sum_i N_{\mathrm{bkg},i}^{\mathrm{fail},i}(m) \right).$$

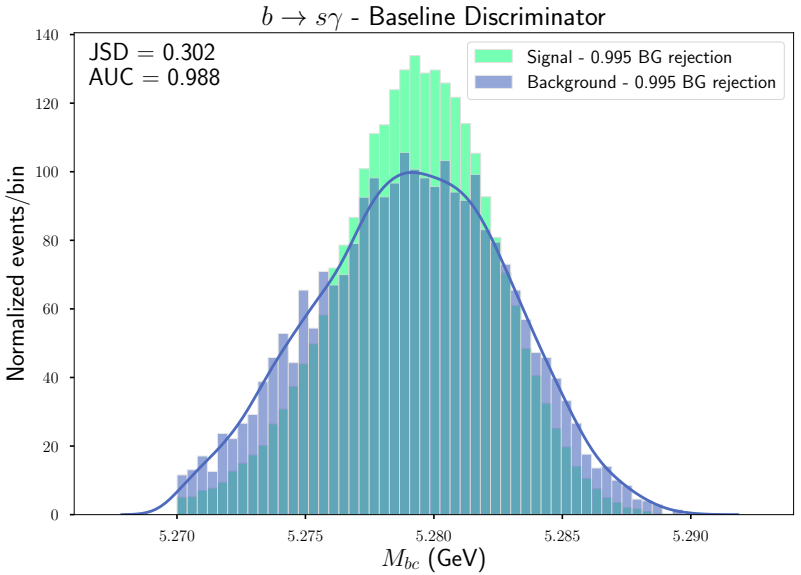
  - Classification: AUC-ROC.

Baseline Discriminator

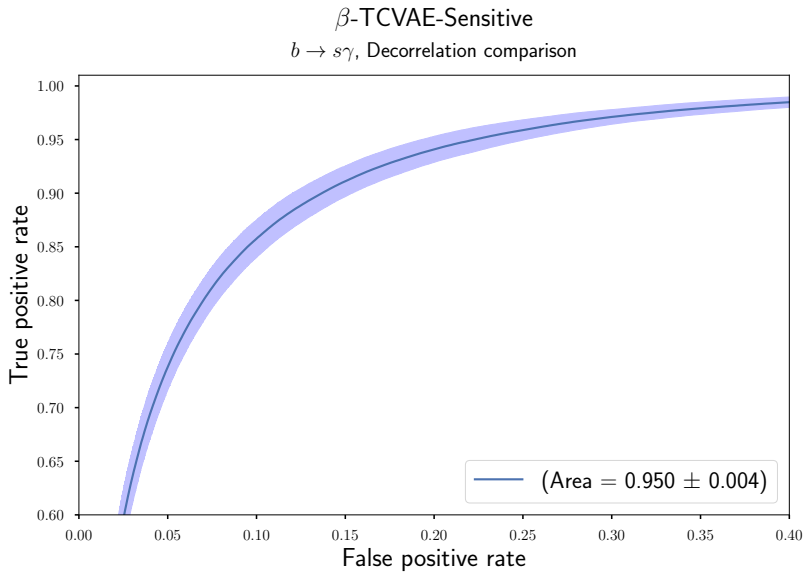


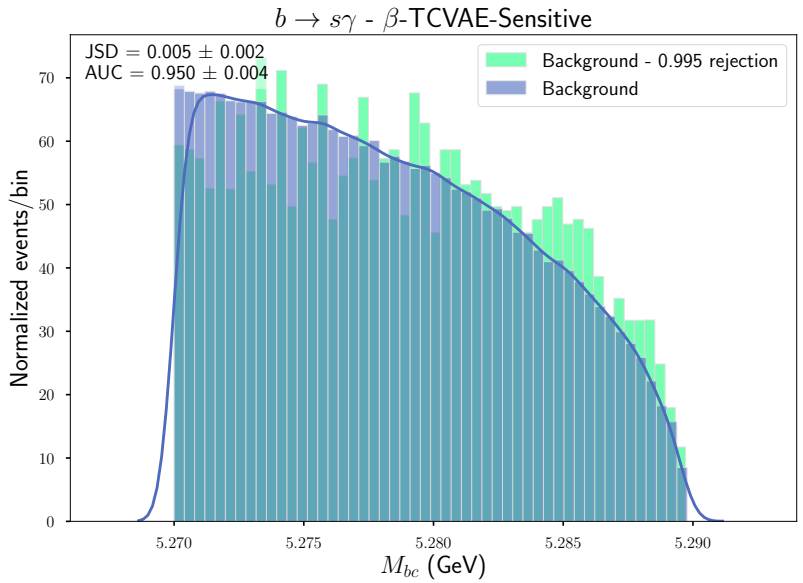


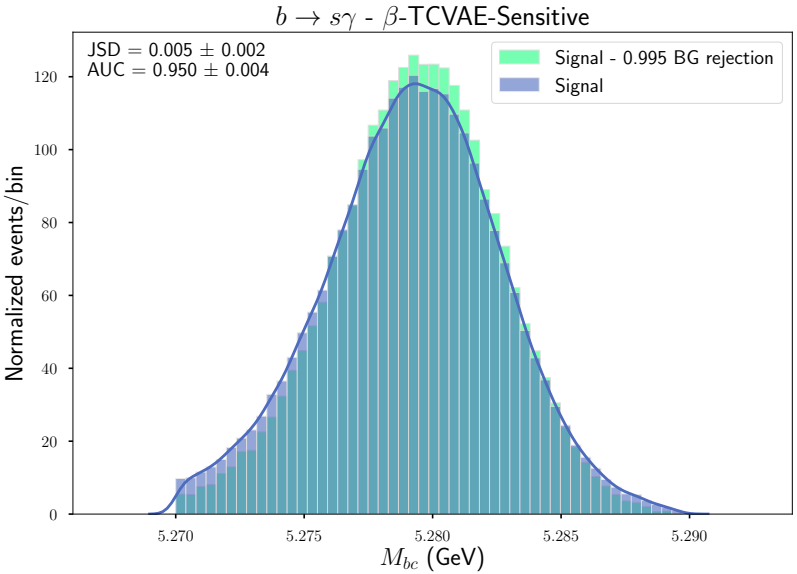


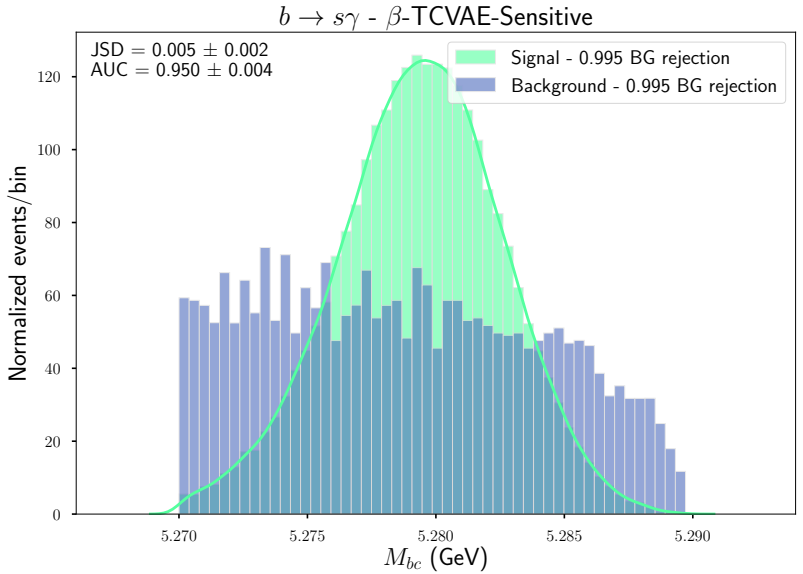


Decorrelated Representation

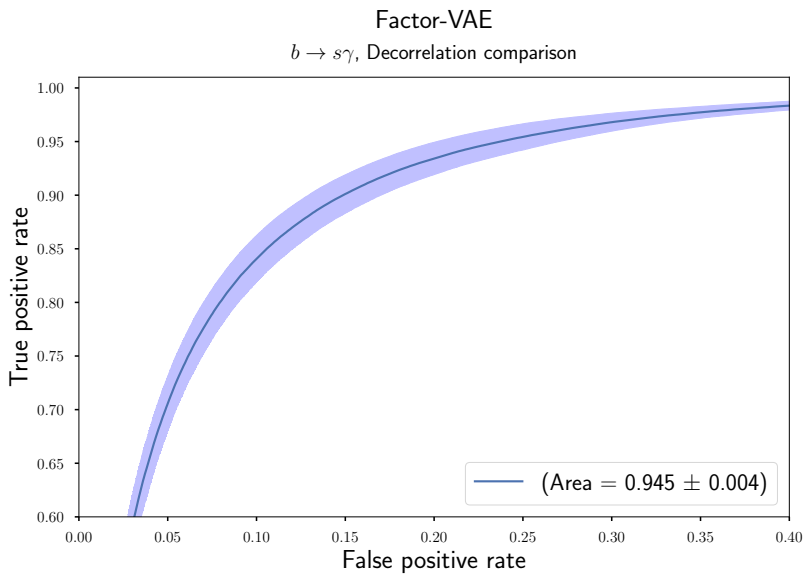


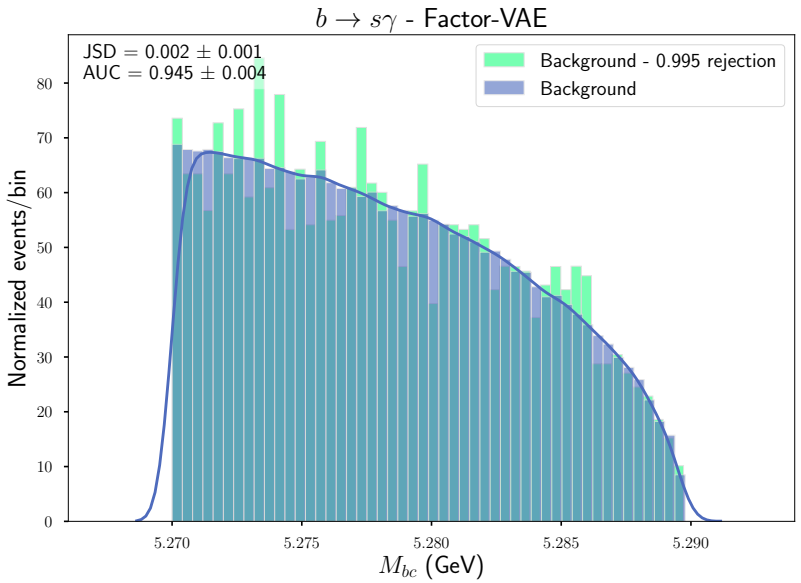


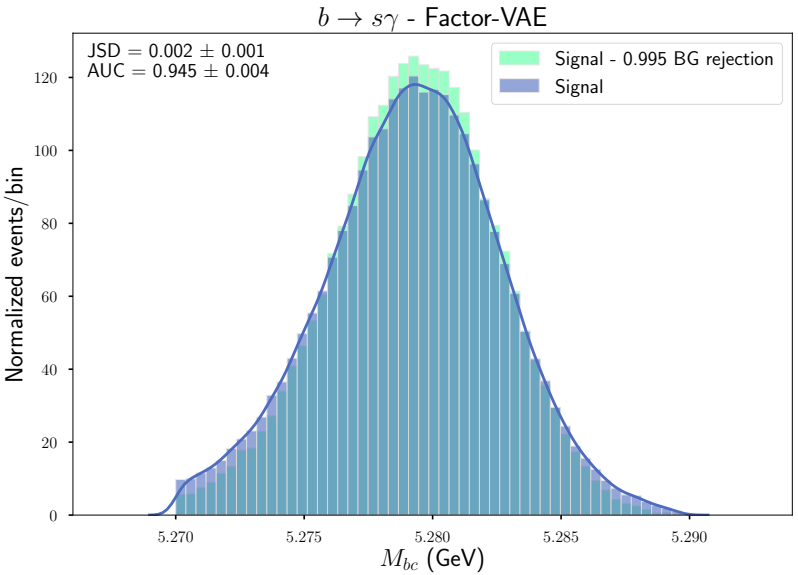


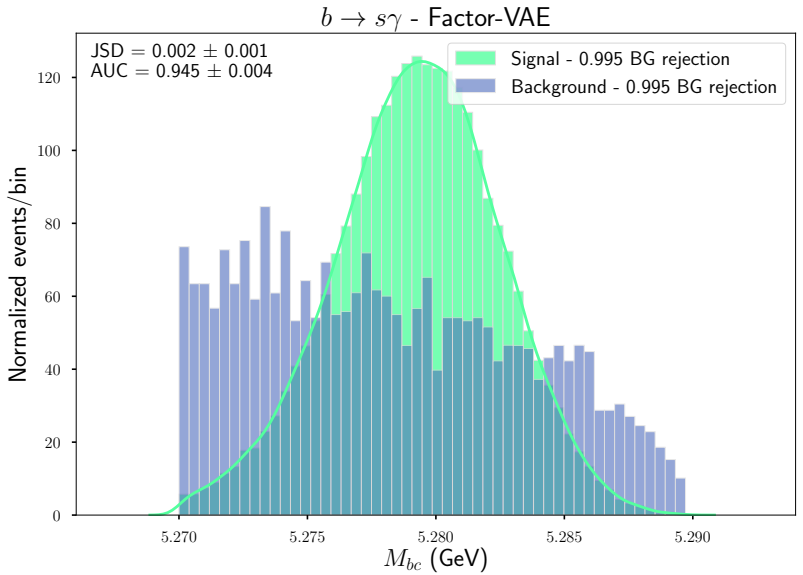


Decorrelated Representation (Factor-VAE)

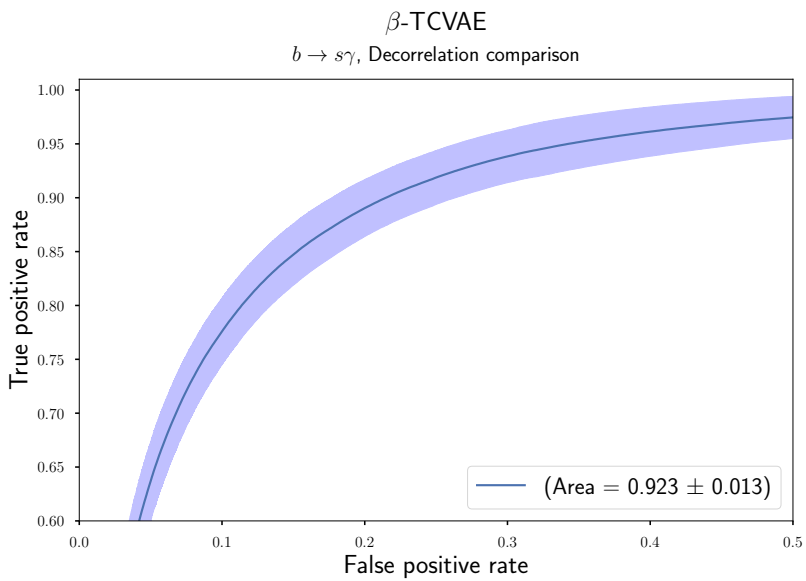


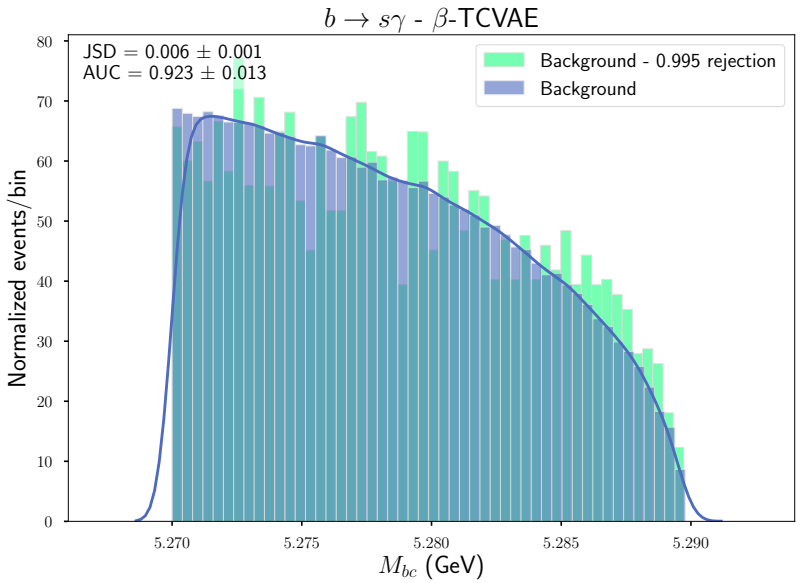


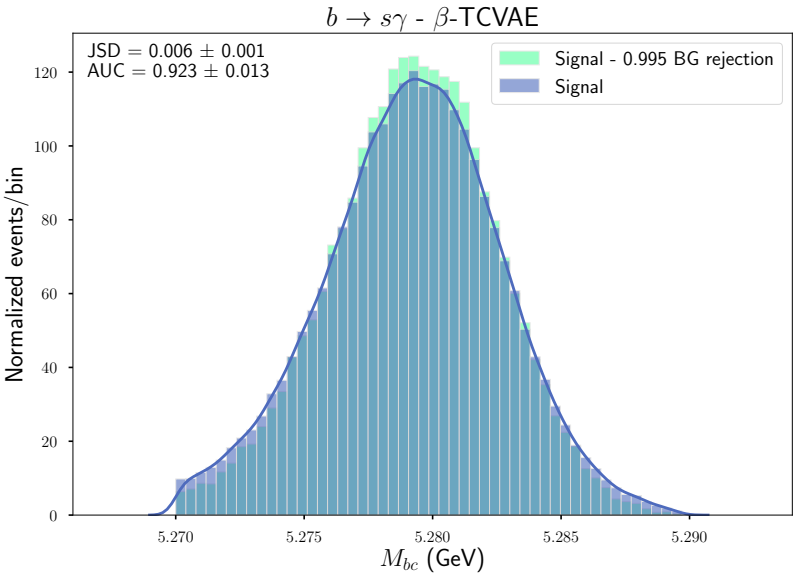


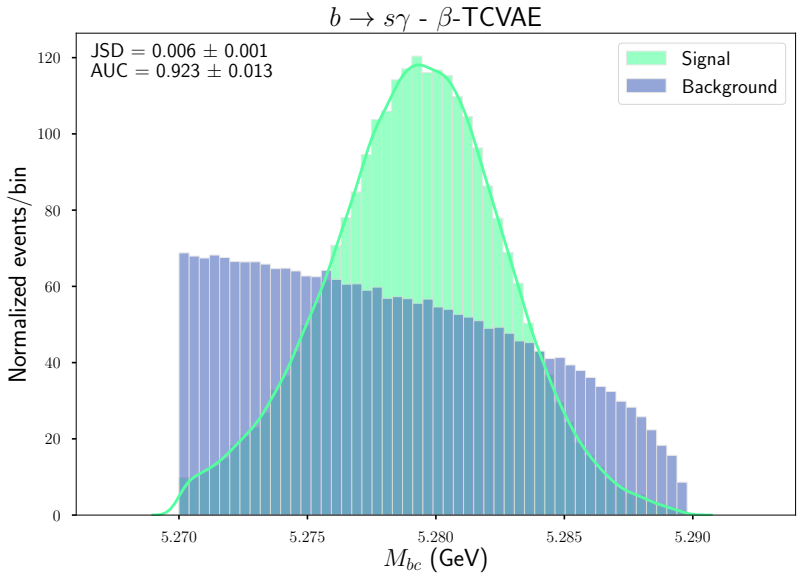


Decorrelated Representation (β-TCVAE)

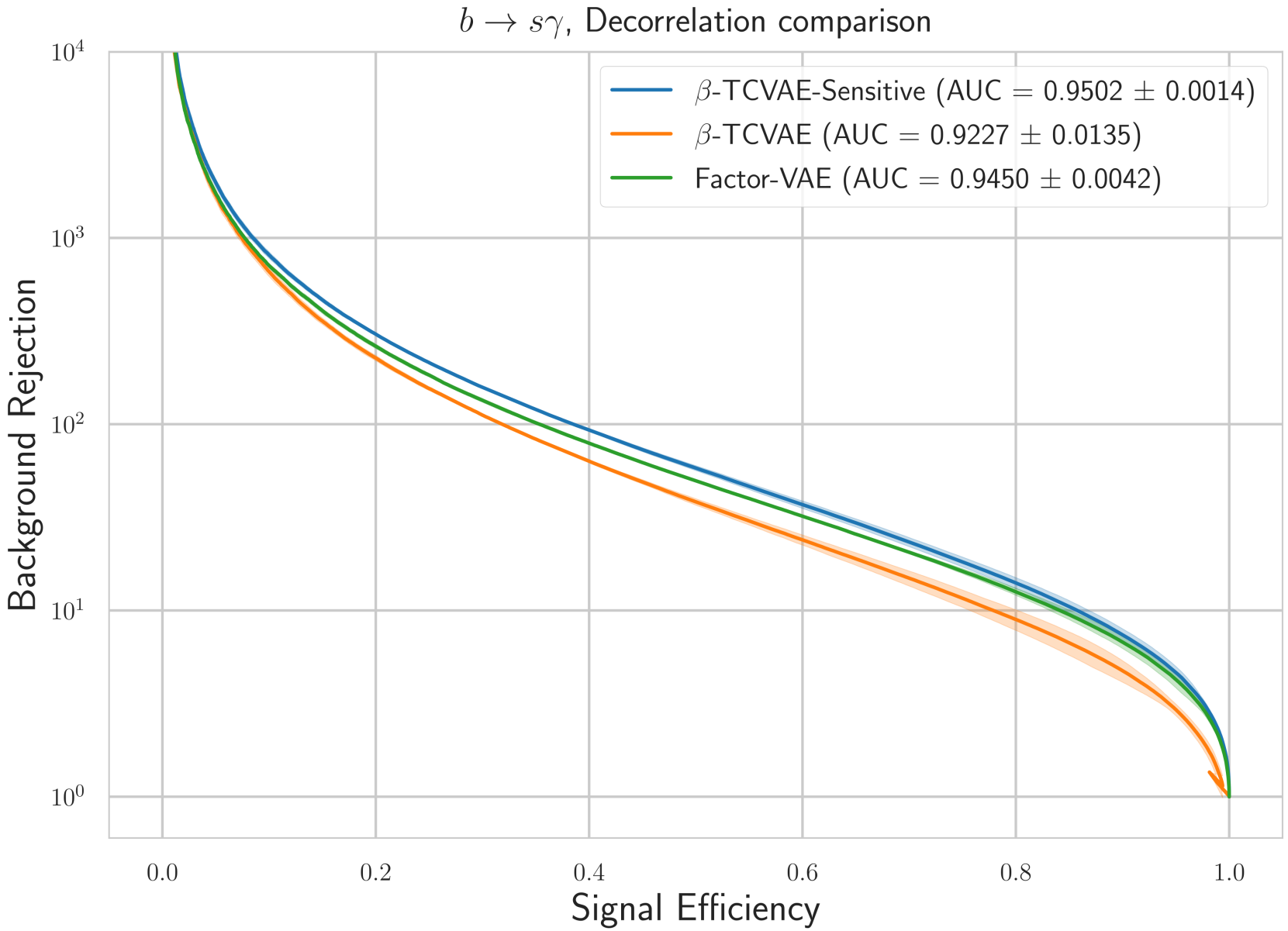








Efficiency Plots



Results

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 1: Metrics obtained when the invariant representation $z \setminus t$ is used as input for the downstream discriminator. Higher is better for AUC and MIG, while lower is better for JSD.

Hyperparameter Robustness

Scan over the weighting λ_s of the supervised regularization term R_s (Eq. 6).

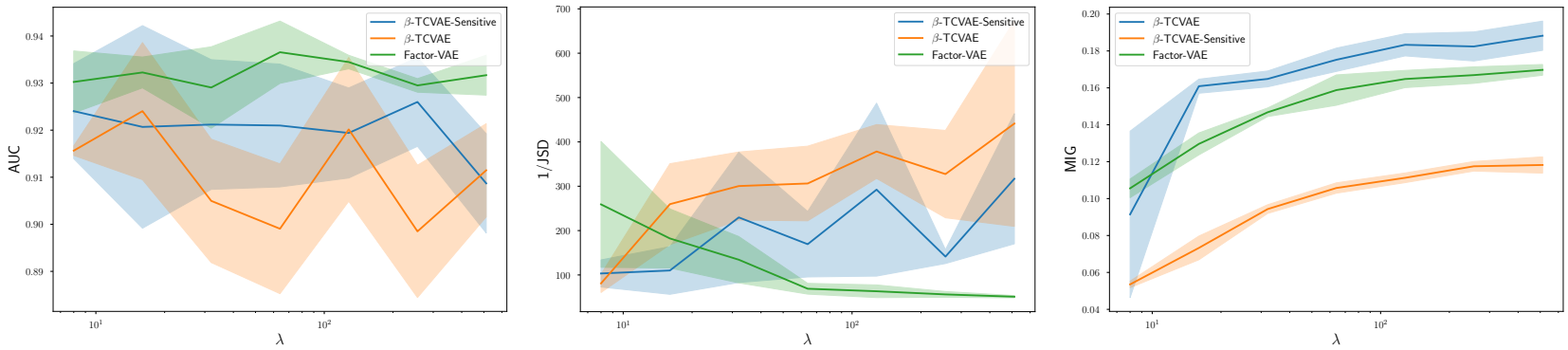


Figure 8: AUC (left), 1/JSD (middle), and MIG (right) versus β (γ) for β-TCVAE (orange), β-TCVAE-Sensitive (blue) and Factor-VAE (green). Higher is better for each metric.

Hyperparameter Robustness

Scan over the weighting β of the latent total correlation term $\mathbf{E}_{q(z)}\left[\log\left(q(z)/\prod_i q(z_i)\right)\right]$ (Eq. 7).

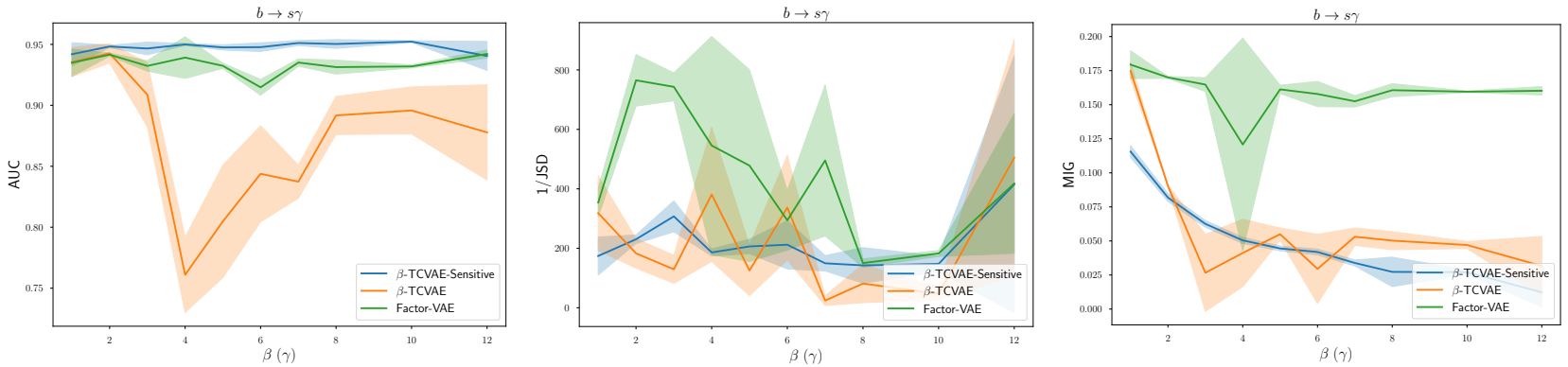


Figure 9: AUC (left), 1/JSD (middle), and MIG (right) versus β (γ) for β-TCVAE (orange), β-TCVAE-Sensitive (blue) and Factor-VAE (green). Higher is better for each metric.

Disentangling Future Work

- Generative modelling via *intervention* on disentangled factors.
  - Manipulate generated event characteristics in latent space.
- Anomaly detection via isolation of invariant mass dimension.
  - Associate latent dimension with invariant mass and perform approximate posterior inference.
- Discrete-variable decorrelation via discrete latent relaxations.

Conclusions

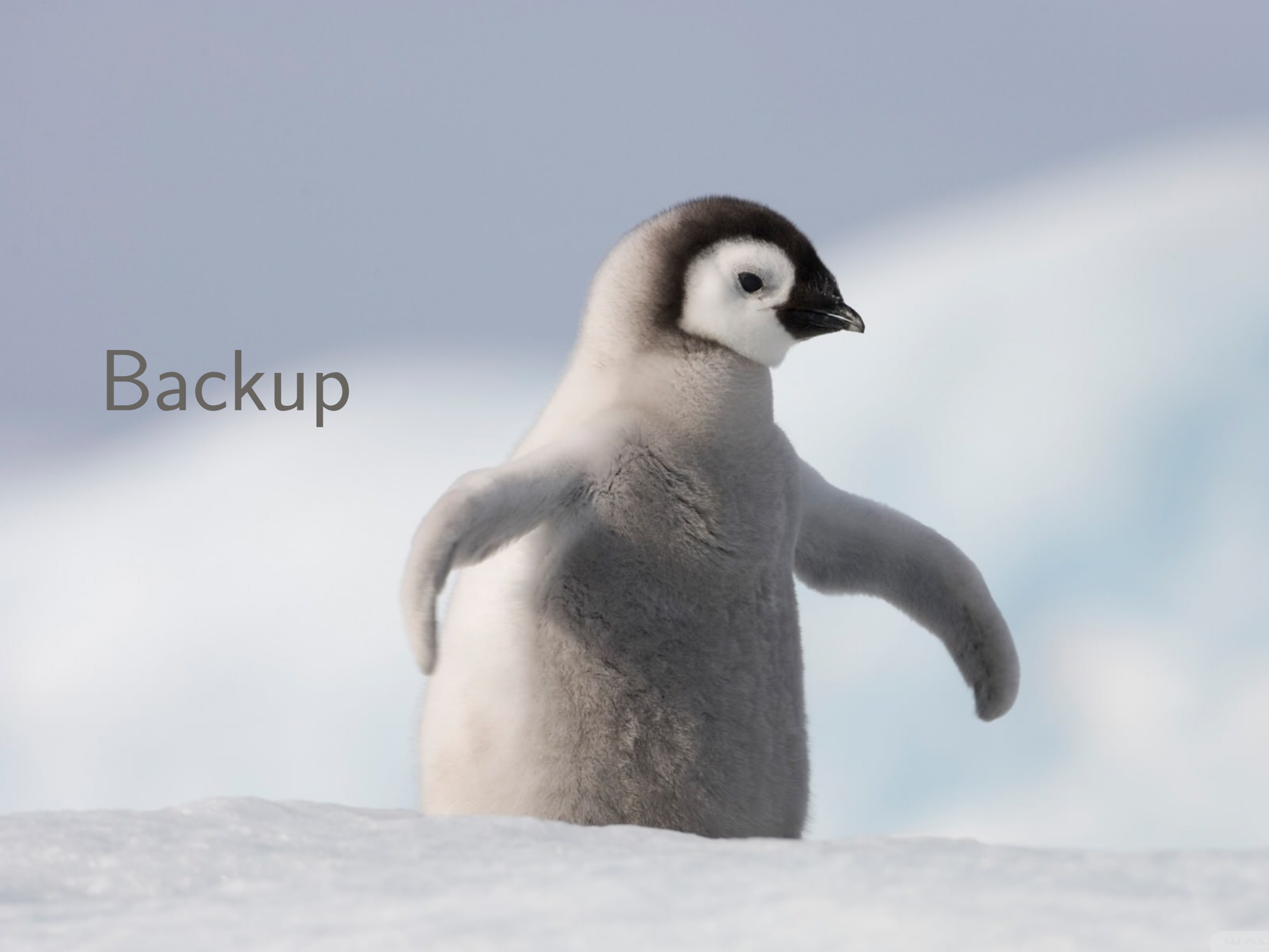
- Presented a method for decorrelation based on latent variable representations in VAEs.
- Robust, non-adversarial training procedure.
- Decorrelation process decoupled from and agnostic to the downstream task.

Thanks for listening!

justin.tan@unimelb.edu.au

github.com/Justin-Tan

Backup



Proofs

Proposition (*Factorization of $q(z) \implies$ factorization of $r(x) = \int dz\, z\, q_\lambda(z; x)$*)

Proof.

$q(z) \triangleq \mathbf{E}_{p(x)}[q_\lambda(z; x)]$. Take expectation of $r(x)$ over $p(x)$:

$$\int dx\, r(x)p(x) = \int dx\, dz\, z\, q_\lambda(z; x)p(x) \tag{9}$$

$$= \int dz\, z \int dx\, q_\lambda(z; x)p(x) \tag{10}$$

$$= \int dz\, z\, q(z) \tag{11}$$

$$= \int dz\, z \prod_{i=1}^{\dim(\mathcal{Z})} q(z_i) \tag{12}$$

$$= \prod_{i=1}^{\dim(\mathcal{Z})} \int dz_i\, z_i\, q(z_i). \tag{13}$$

□

Visualizations

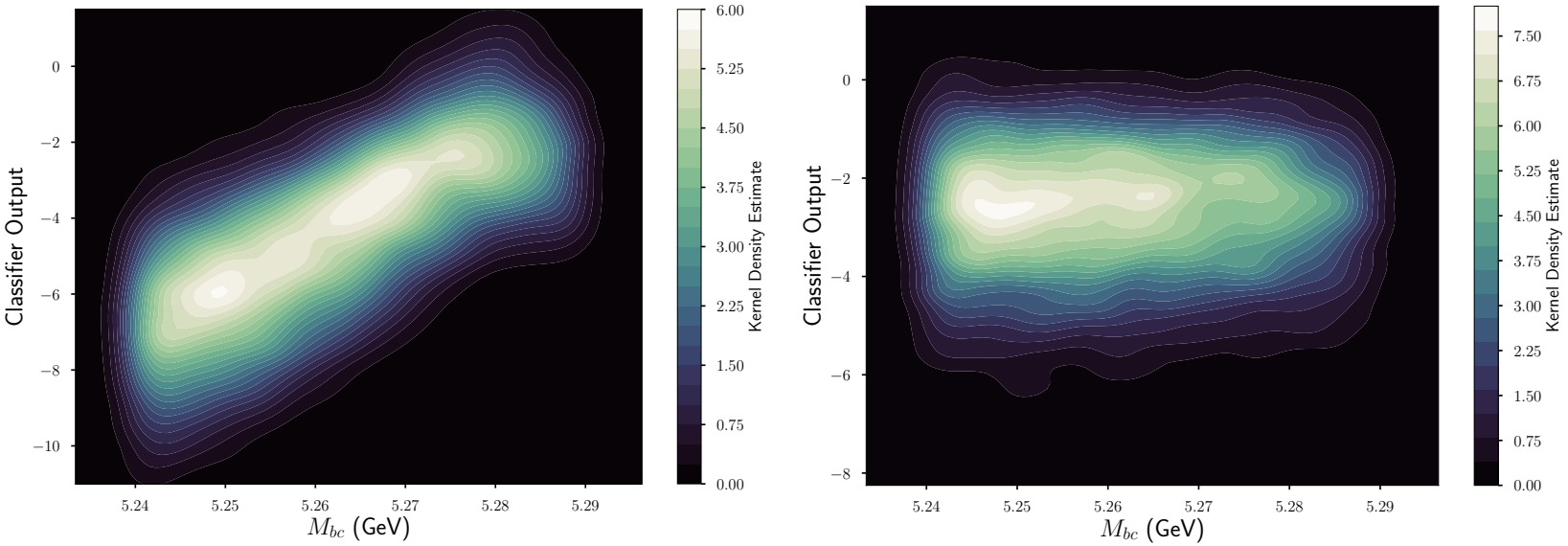


Figure 10: Visualization of invariance of the downstream classification task to the sensitive variable $s = M_{bc}$. Bivariate kernel density estimate between M_{bc} and the classifer output, $\log\left(\frac{p}{1-p}\right)$, where p is the probability that a given event belongs to the positive signal class given by the downstream classifier. 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 baseline classifier output (left) 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. The output of the invariant classifier (right) does not exhibit significant dependency on M_{bc}, with the classifier response flat throughout the full M_{bc} range.