Supplementary Information A Derivation of the ELBO Loss for VAE

$p_{\phi_d}(\mathbf{x} \mid \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{y_{cf}})$ and $p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{y_cf} \mid \mathbf{x})$ are the true likelihood and true posterior respectively. The posterior is hard to evaluate, so we have to approximate the true posterior to the product of the factorized known distributions $q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})$, $q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})$, $q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})$ and $q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x})$ by minimising the KL divergence as follows,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \mid\mid p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf} \mid \mathbf{x})\big) \\
&= \int\int\int\int q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \\
&\quad \Big[\log \frac{q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid x)}{p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf} \mid \mathbf{x})}\Big]\mathbf{dz}_x\mathbf{dz}_t\mathbf{dz}_{yf}\mathbf{dz}_{ycf} \\
&= \int\int\int\int q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \\
&\quad \Big[\log q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x}) + \log q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) + \log q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) + \log q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \\
&\quad - \log p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf} \mid \mathbf{x})\Big]\mathbf{dz}_x\mathbf{dz}_t\mathbf{dz}_{yf}\mathbf{dz}_{ycf} \\
&= \int\int\int\int q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \\
&\quad \Big[\log q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x}) + \log q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) + \log q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) + \log q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \\
&\quad - \log p_{\phi_d}(x \mid \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf}) - \log p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf}) \\
&\quad + \log p_{\phi_d}(x)\Big]\mathbf{dz}_x\mathbf{dz}_t\mathbf{dz}_{yf}\mathbf{dz}_{ycf} \\
&= \int q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x}) \log \frac{q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})}{p_{\phi_d}(\mathbf{z}_x)}\mathbf{dz}_x + \int q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) \log \frac{q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})}{p_{\phi_d}(\mathbf{z}_t)}\mathbf{dz}_t \\
&\quad + \int q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) \log \frac{q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x})}{p_{\phi_d}(\mathbf{z}_{yf})}\mathbf{dz}_{yf} + \int q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \log \frac{q_{\phi_x}(\mathbf{z}_{ycf} \mid \mathbf{x})}{p_{\phi_d}(\mathbf{z}_x)}\mathbf{dz}_{ycf} \\
&\quad - \int\int\int\int \Big[q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x})q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x})q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) \\
&\quad q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \log p_{\phi_d}(\mathbf{x} \mid \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf})\Big]\mathbf{dz}_x\mathbf{dz}_t\mathbf{dz}_{yf}\mathbf{dz}_{ycf} + \log p_{\phi_d}(\mathbf{x}) \\
&= KL\big(q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x}) \mid\mid p_{\phi_d}(\mathbf{z}_x))\big) + KL\big(q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) \mid\mid p_{\phi_d}(\mathbf{z}_t))\big) \\
&\quad + KL\big(q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) \mid\mid p_{\phi_d}(\mathbf{z}_{yf}))\big) + KL\big(q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \mid\mid p_{\phi_d}(\mathbf{z}_{ycf}))\big) \\
&\quad - \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(\mathbf{x} \mid \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf})] \\
&\quad + \log p_{\phi_d}(\mathbf{x})
\end{aligned}
$$

where, the distributions $q_{\phi_x}(\mathbf{z}_x|x), q_{\phi_t}(\mathbf{z}_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ and $p_{\phi_d}(x|\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf})$ are parameterized by the parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d$.

The KL divergence of two distributions is always greater than or equal to zero. So,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(\mathbf{z}_x \mid x) q_{\phi_t}(\mathbf{z}_t \mid x) q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid x) \,||\, p_{\phi_d}(\mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf} \mid x)\big) \geq 0, \\
&\log p_{\phi_d}(\mathbf{z}) \geq \mathcal{L}_{ELBO} \quad \text{where,} \\
&\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf};\ \mathbf{x}, \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf}) \\
&\quad = \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(\mathbf{x} \mid \mathbf{z}_x, \mathbf{z}_t, \mathbf{z}_{yf}, \mathbf{z}_{ycf})] \\
&\qquad - KL\big(q_{\phi_x}(\mathbf{z}_x \mid \mathbf{x}) \,||\, p_{\phi_d}(\mathbf{z}_x))\big) - KL\big(q_{\phi_t}(\mathbf{z}_t \mid \mathbf{x}) \,||\, p_{\phi_d}(\mathbf{z}_t))\big) \\
&\qquad - KL\big(q_{\phi_{yf}}(\mathbf{z}_{yf} \mid \mathbf{x}) \,||\, p_{\phi_d}(\mathbf{z}_{yf}))\big) - KL\big(q_{\phi_{ycf}}(\mathbf{z}_{ycf} \mid \mathbf{x}) \,||\, p_{\phi_d}(\mathbf{z}_{ycf}))\big)
\end{aligned}
$$

Supplementary Information B Derivation of the ELBO Loss for VAE

In DR-VIDAL we aimed to maximize the mutual information of the latent factors $\mathbf{z}_c$ and the output of the generator $G(\mathbf{z}_G, \mathbf{z}_c)$ as follows:

$$
\begin{aligned}
&I(\mathbf{z}_c;\ G(\mathbf{z}_G, \mathbf{z}_c)) = H(\mathbf{z}_c) - H(\mathbf{z}_c \mid G(\mathbf{z}_G, \mathbf{z}_c)) \\
&= H(\mathbf{z}_c) + \int\int p(\mathbf{Z}_c = \mathbf{z}'_c, \mathbf{X} = G(\mathbf{z}_G, \mathbf{z}_c)) \\
&\qquad \log p(\mathbf{Z}_c = \mathbf{z}'_c \mid \mathbf{X} = G(\mathbf{z}_G, \mathbf{z}_c))\, \mathbf{dz}_c \mathbf{dx} \\
&= H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}_G, \mathbf{z}_c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log(p(\mathbf{z}'_c \mid \mathbf{x})) \\
&= H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}_G, \mathbf{z}_c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[\frac{p(\mathbf{z}'_c \mid \mathbf{x})}{Q(\mathbf{z}'_c \mid \mathbf{x})} Q(\mathbf{z}'_c \mid \mathbf{x})\right] \\
&= H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}_G, \mathbf{z}_c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[\frac{p(\mathbf{z}'_c \mid \mathbf{x})}{Q(\mathbf{z}'_c \mid \mathbf{x})}\right] \\
&\qquad + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[Q(\mathbf{z}'_c \mid \mathbf{x})\right] \\
&= H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \int p(\mathbf{z}'_c \mid \mathbf{x}) \log \frac{p(\mathbf{z}'_c \mid \mathbf{x})}{Q(\mathbf{z}'_c \mid \mathbf{x})} dc' \\
&\qquad + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \mathbb{E}_{c' \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[Q(\mathbf{z}'_c \mid \mathbf{x})\right] \\
&= H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \left[KL\big(p(\mathbf{z}'_c \mid \mathbf{x} \,||\, Q(\mathbf{z}'_c \mid \mathbf{x}))\big)\right] \\
&\qquad + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[Q(\mathbf{z}'_c \mid \mathbf{x})\right] \\
&\geq H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c \mid \mathbf{x})} \log\left[Q(\mathbf{z}'_c \mid \mathbf{x})\right]
\end{aligned}
$$

$$\geq H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{z}_c \sim p(\mathbf{z}_c)} \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z},c)} \mathbb{E}_{\mathbf{z}'_c \sim p(\mathbf{z}_c|\mathbf{x})} \log \left[Q(\mathbf{z}'_c \mid \mathbf{x}) \right]$$

$$\geq H(\mathbf{z}_c) + \mathbb{E}_{\mathbf{z}_c \sim p(\mathbf{z}_c)} \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z},c)} \log \left[Q(\mathbf{z}_c \mid \mathbf{x}) \right]$$

(by Lemma 5.1 of [1])

$$= L_I(G, Q)$$

Supplementary Information C Supplementary Figures and Tables

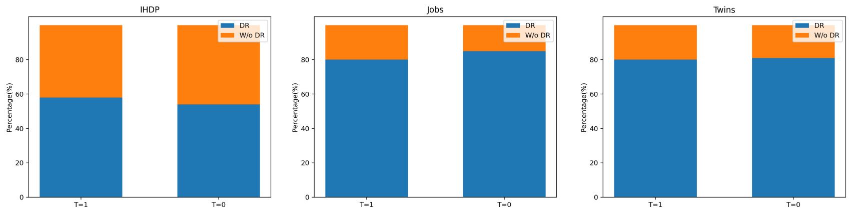

Fig. C1 Performance comparison of doubly robust vs. non-doubly robust version of DR-VIDAL. Panels, from left to right, show results on IHDP, Jobs and Twins datasets (100, 10, 100 iterations), respectively.

Table C1 Performance of the all the different DR-VIDAL configurations on the IHDP, Jobs and Twins datasets (1000, 10, and 100 realizations, respectively). Results show the out-of-sample (mean ± st.dev) error (PEHE) and policy risk (R_{Pol}).

	IHDP $\sqrt{\epsilon^{out-of-s}_{PEHE}}$	**Jobs** $R^{out-of-s}_{Pol}$	**Twins** $\sqrt{\epsilon^{out-of-s}_{PEHE}}$
DR-VIDAL	**0.62 ± 0.06**	**0.102 ± 0.01**	**0.318 ± 0.008**
DR-VIDAL (w/o DR loss)	0.85 ± 0.06	0.110 ± 0.01	0.324 ± 0.007
DR-VIDAL (w/o Info loss)	0.67 ± 0.04	0.109 ± 0.01	0.318 ± 0.012
DR-VIDAL (w/o DR + Info loss)	0.81 ± 0.05	0.113 ± 0.01	0.326 ± 0.008

References

[1] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems **29**, 2172–2180 (2016)

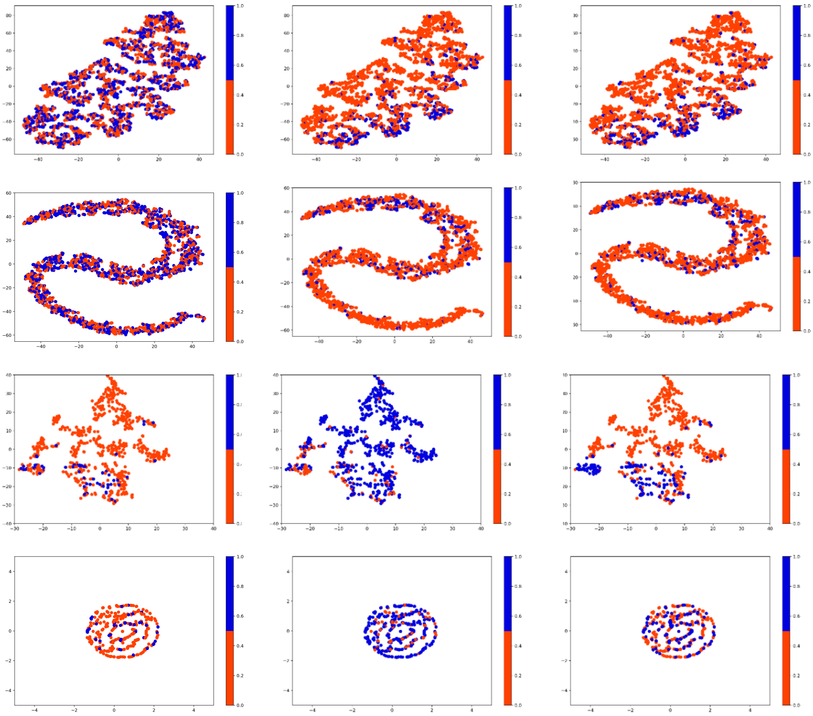

Fig. C2 Visualization of the latent representation learned by the VAE module of DR-VIDAL for the Twins and Jobs dataset using t-SNE. The 1^{st} and 2^{nd} panels show the t-SNE before and after training the network for Twins dataset. The 3^{rd} and 4^{th} panels show the same for Jobs dataset. From left to right, the plots show the t-SNE of treatment, factual and counterfactual outcomes.