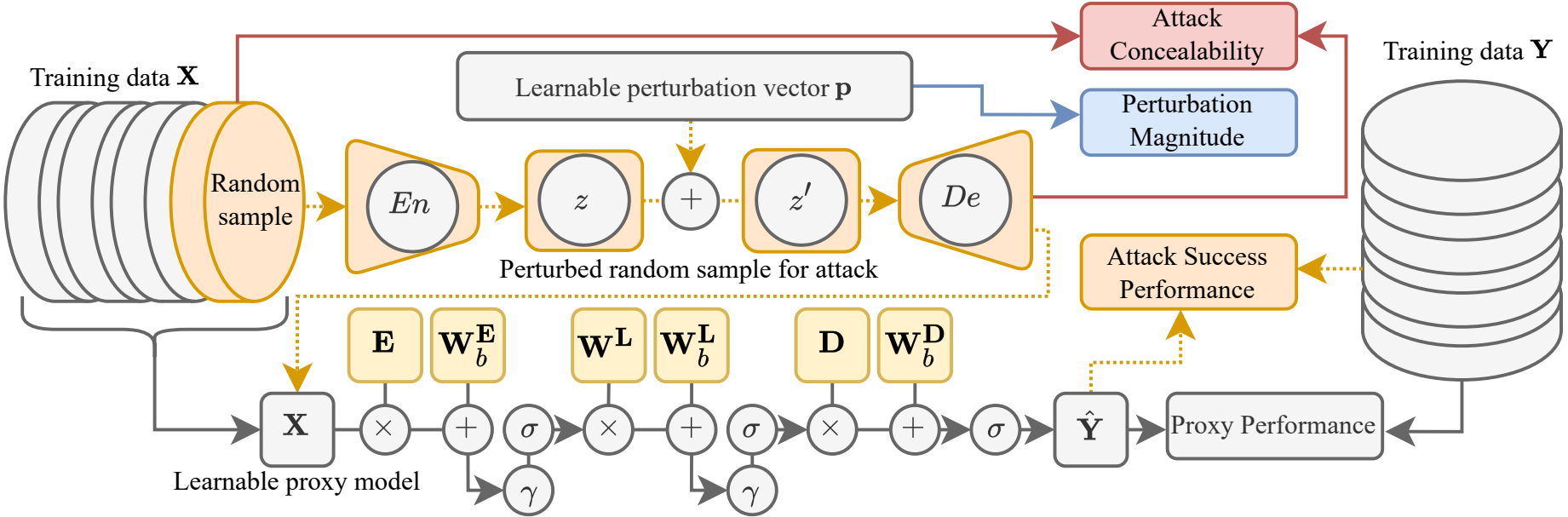
Attack Concealability
Training data $\mathbf{Y}$
Training data $\mathbf{X}$
Learnable perturbation vector $\mathbf{p}$
Perturbation Magnitude
Random sample
En
z
$+$
z'
De
Perturbed random sample for attack
Attack Success Performance
$\mathbf{E}$
$\mathbf{W}_b^{\mathbf{E}}$
$\mathbf{W}^{\mathbf{L}}$
$\mathbf{W}_b^{\mathbf{L}}$
$\mathbf{D}$
$\mathbf{W}_b^{\mathbf{D}}$
$\mathbf{X}$
$\times$
$+$
σ
γ
$\hat{\mathbf{Y}}$
Proxy Performance
Learnable proxy model