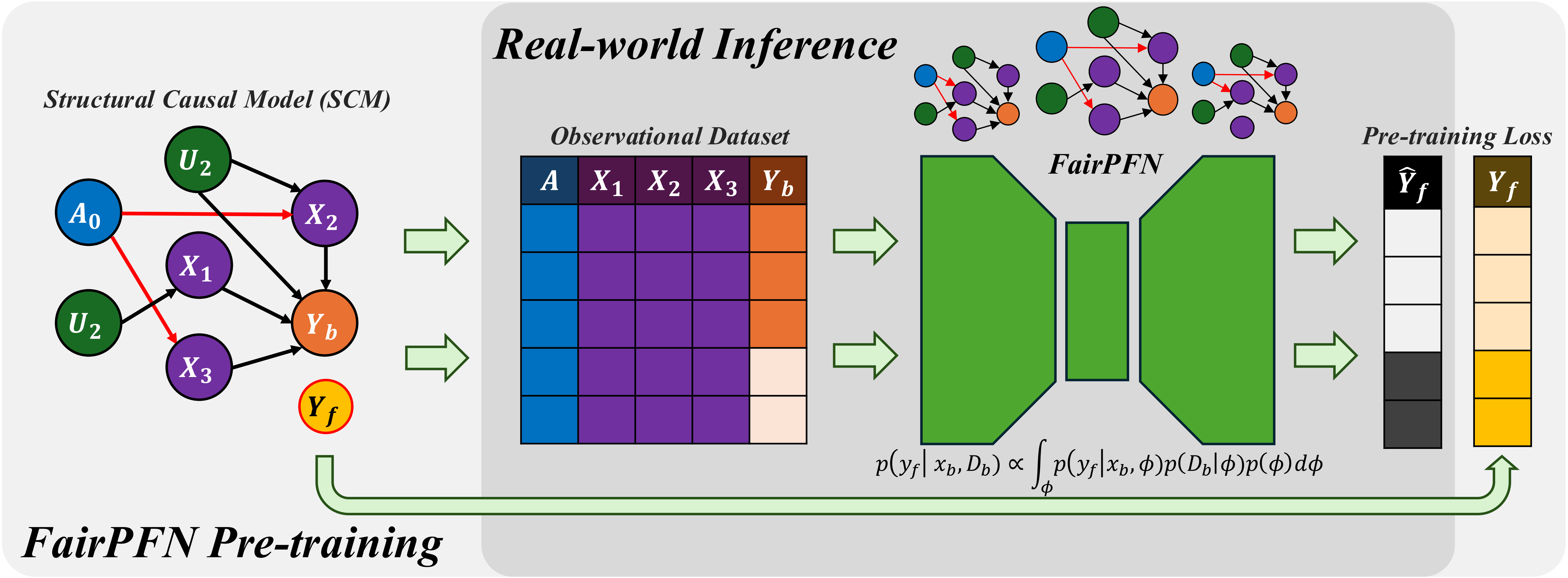
a) Data generation: For each pre-training dataset, we generate an SCM and sample a dataset $\boldsymbol{D}$ comprised of a protected attribute $\boldsymbol{A}$, potentially biased observables $\boldsymbol{X_b}$, and biased outcome $\boldsymbol{Y_b}$. We also sample a fair outcome $\boldsymbol{Y_f}$ by removing the outgoing edges of $\boldsymbol{A}$.

b) Transformer input: The observational dataset $\boldsymbol{D}$ is partitioned into training and validation splits. Given in-context examples $\boldsymbol{D_{train}}$ the transformer makes predictions on the inference set $\boldsymbol{D_{val}} = (\boldsymbol{A_{val}}, \boldsymbol{X_{val}})$

c) Fair prediction: The transformer makes predictions $\boldsymbol{\hat{Y}_f}$ on the validation set, and the pre-training loss is calculated with respect to the fair outcomes in the validation set. The transformer thus learns the mapping $\boldsymbol{X_b} \rightarrow \boldsymbol{Y_f}$


Real-world Inference
Structural Causal Model (SCM)
U_2
A_0
X_2
X_1
U_2
Y_b
X_3
Y_f
Observational Dataset
A
X_1
X_2
X_3
Y_b
FairPFN
$p(y_f| x_b, D_b) \propto \int_\phi p(y_f|x_b, \phi)p(D_b|\phi)p(\phi)d\phi$
Pre-training Loss
$\hat{Y}_f$
Y_f
FairPFN Pre-training