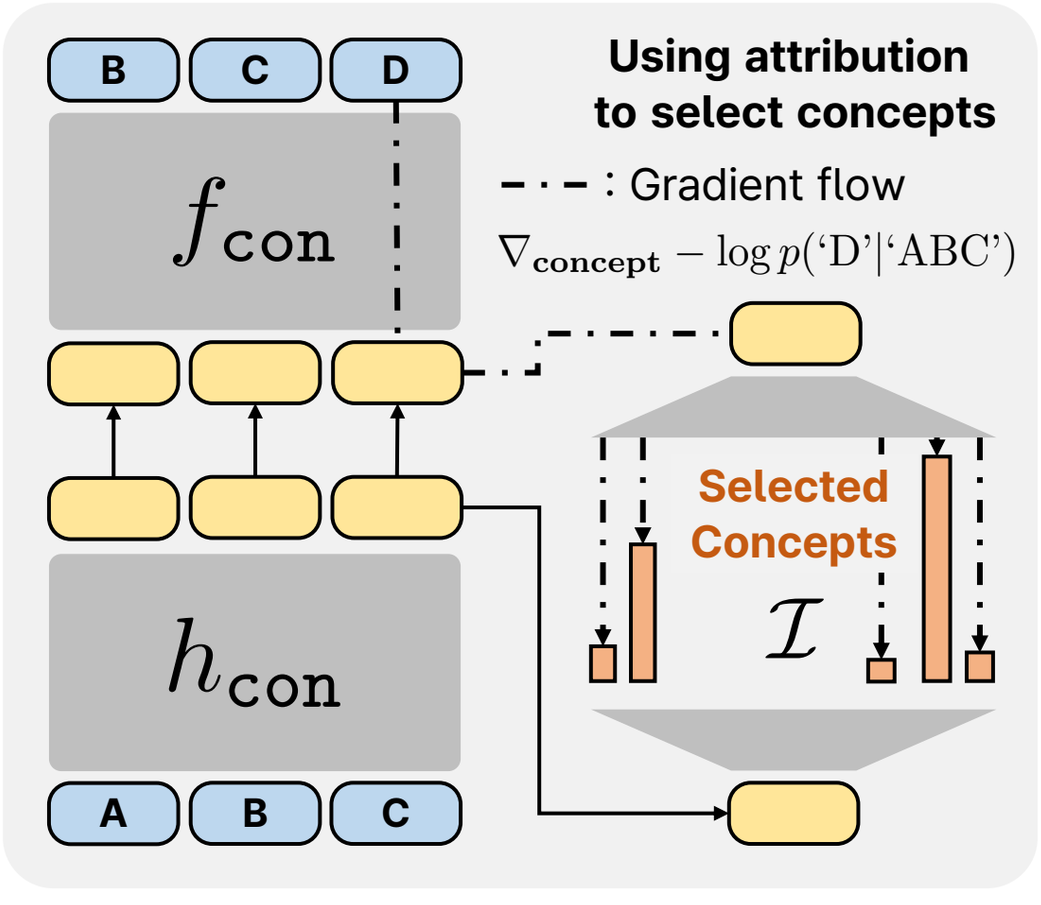
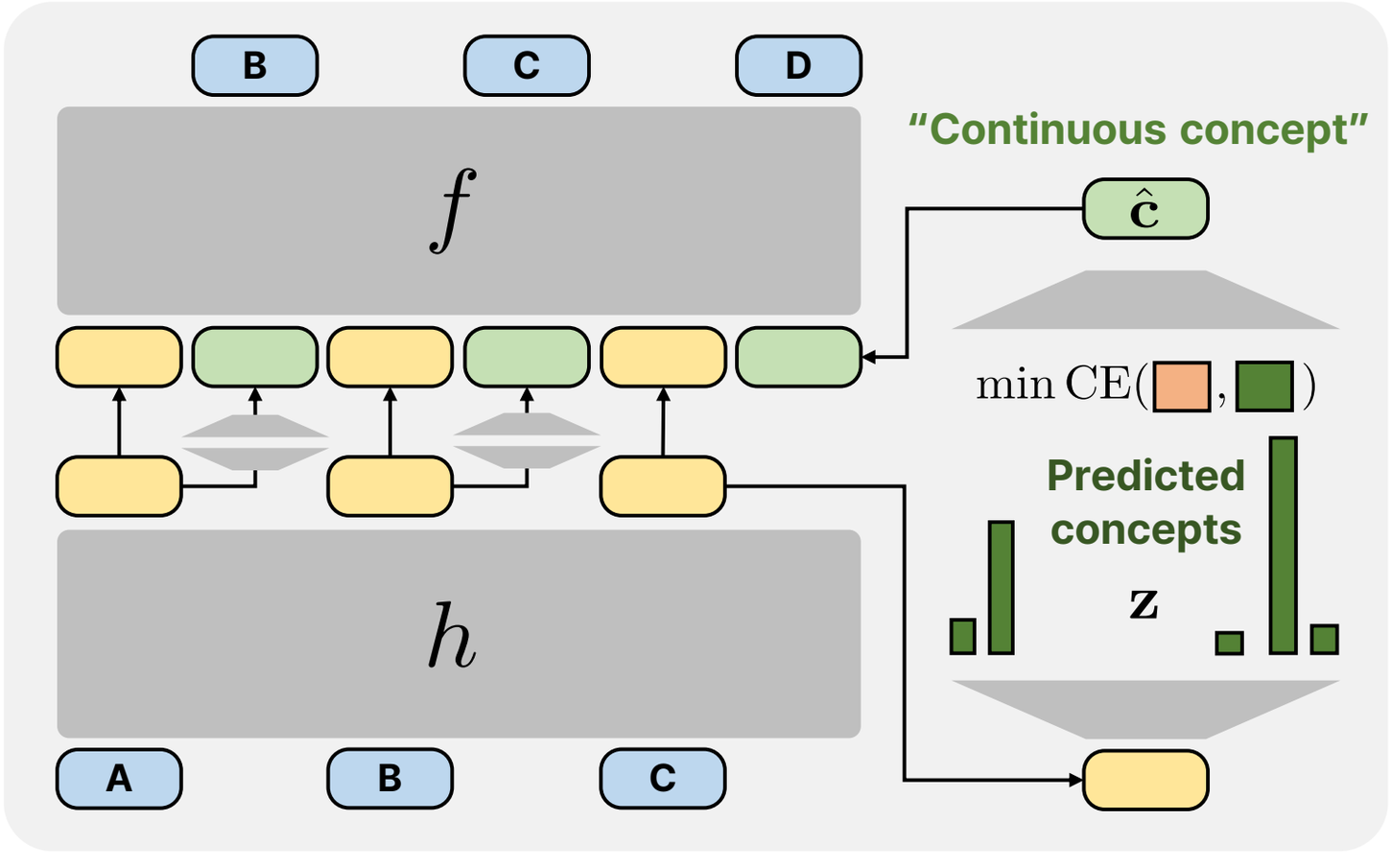


Extracting concepts from a pretrained SAE model's hidden state



Learning to predict concepts & Mixing/Interleaving continuous concepts into the hidden state