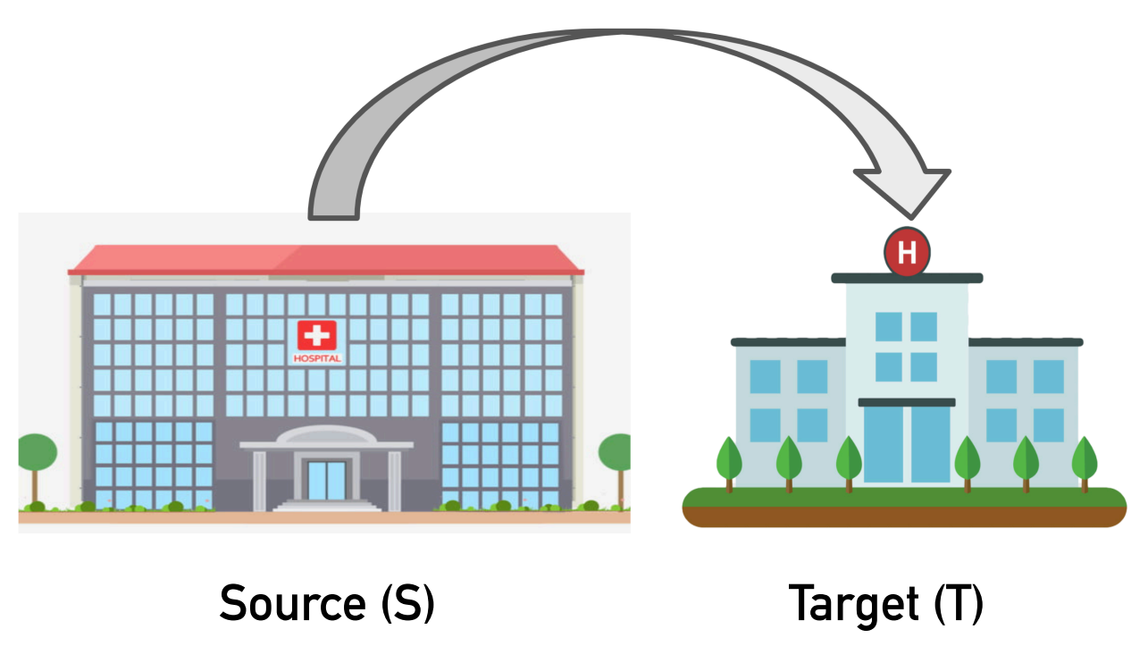
Counterfactually Guided Policy Transfer in Clinical Settings

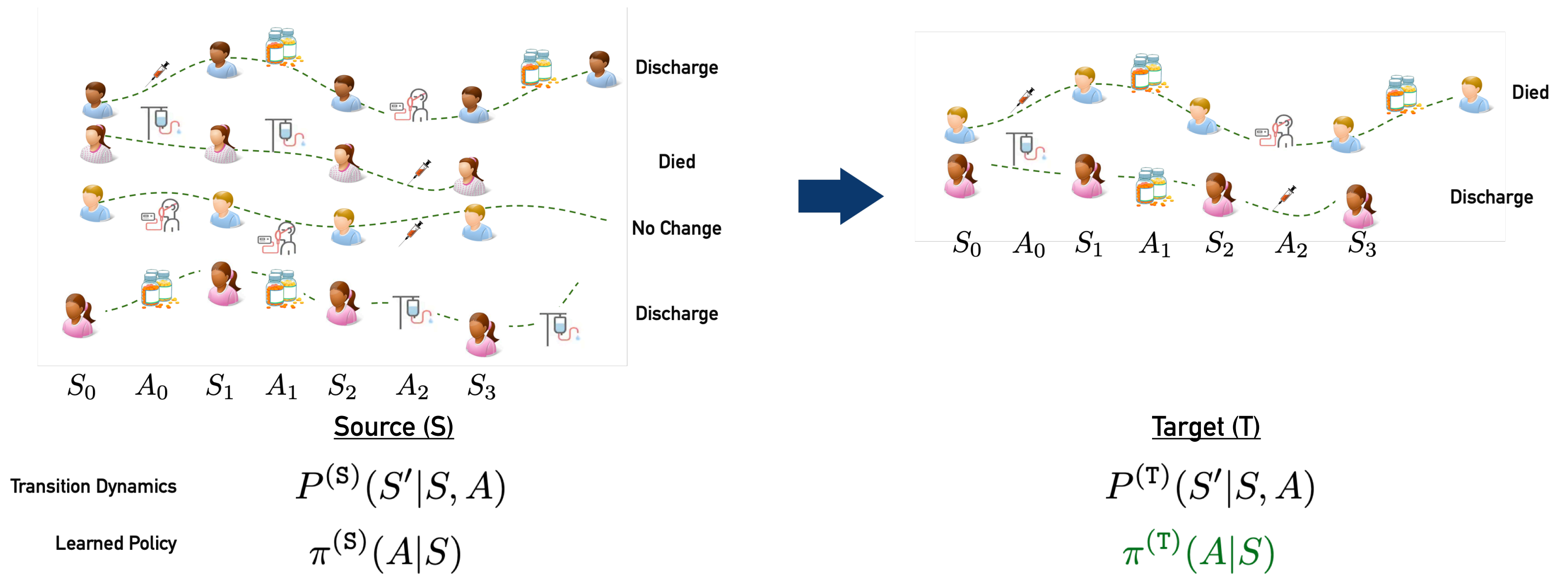
Taylor W. Killian[1,2] Marzyeh Ghassemi[3] Shalmali Joshi[4]

[1]University of Toronto, [2]Vector Institute
[3]Massachusetts Institute of Technology
[4]CRCS, Harvard University (SEAS)

Motivation

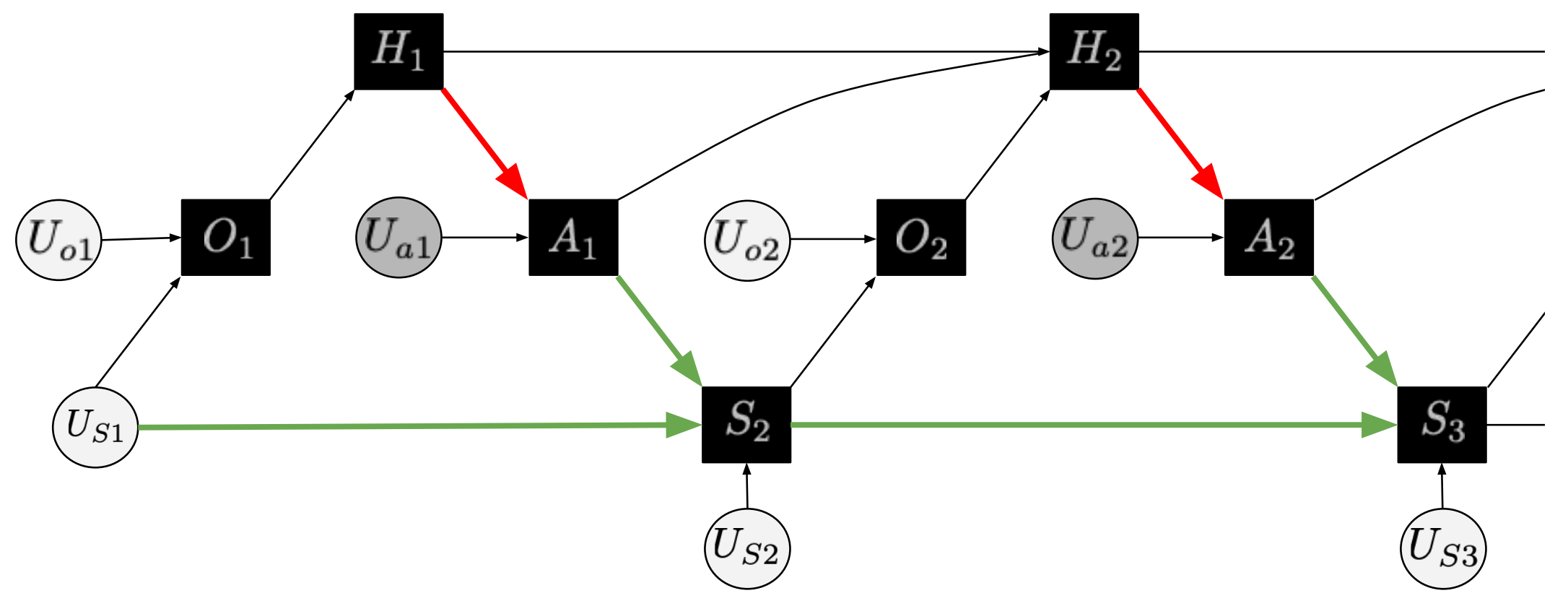
Domain shift between training and deployment clinical environments significantly limits the ability to transfer trained treatment models. These challenges are magnified when data limitations and unobserved confounding (e.g subpopulation composition) are present in the deployment environment.





Off-policy Transfer as Counterfactual Inference

Following Buesing, et al[1] and we assume observation-treatment interactions can be modeled by a POMDP structured as a Structural Causal Model



The edges of this graphical model captures relevant causal dependencies

$$\begin{aligned} S_i &:= f_s(S_{i-1}, A_{i-1}, U_{s,i}) && \approx T(S_i|S_{i-1}, A_{i-1}) \\ O_i &:= f_o(S_i, U_{o,i}) && \approx \Omega(O_i|S_i) \\ A_i &:= f_\pi(H_i, U_{a,i}) && \approx \pi(A_i|H_i) \end{aligned}$$

Gumbel-Max SCM

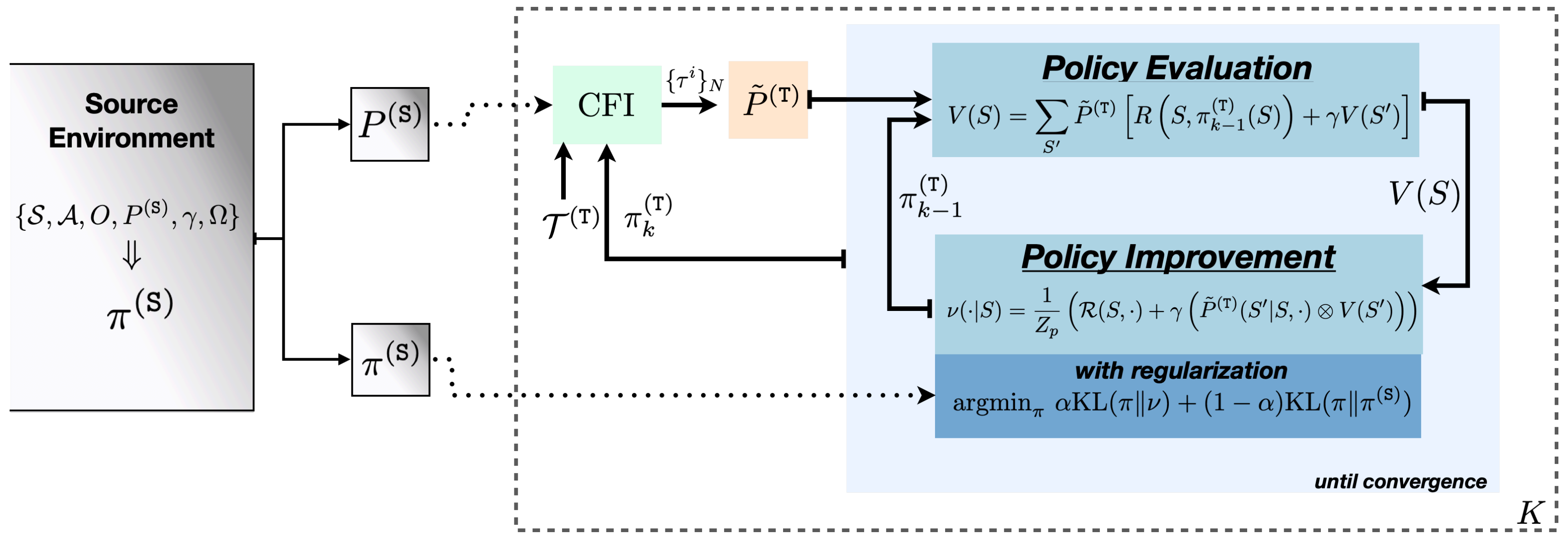
Oberst and Sontag[2] introduced the Gumbel-Max SCM, which ensures that counterfactual queries preserve observed outcomes when sampling from discrete, categorical distributions (embedding the Gumbel-Max Trick[3]).

All nodes in a Gumbel-Max SCM are modeled with the following functional form:

$$X_i := \arg\max_j \ \log p(X_i = j|\mathbf{PA}_i) + g_j$$

given independent Gumbel variables $\mathbf{g} \ = \ \{g_1, g_2, \ldots, g_k\}$.

Counterfactually Guided Policy Transfer


Source Environment
$\{\mathcal{S}, \mathcal{A}, O, P^{(S)}, \gamma, \Omega\}$
$\Downarrow$
$\pi^{(S)}$
$P^{(S)}$
$\pi^{(S)}$
CFI
$\{\tau^i\}_N$
$\tilde{P}^{(T)}$
$\mathcal{T}^{(T)}$
$\pi_k^{(T)}$
Policy Evaluation
$V(S) = \sum_{S'} \tilde{P}^{(T)} \left[R\left(S, \pi_{k-1}^{(T)}(S)\right) + \gamma V(S')\right]$
$\pi_{k-1}^{(T)}$
$V(S)$
Policy Improvement
$\nu(\cdot|S) = \frac{1}{Z_p}\left(\mathcal{R}(S,\cdot) + \gamma\left(\tilde{P}^{(T)}(S'|S,\cdot) \otimes V(S')\right)\right)$
with regularization
$\text{argmin}_\pi \ \alpha \text{KL}(\pi\|\nu) + (1-\alpha)\text{KL}(\pi\|\pi^{(S)})$
until convergence
K

To enable the transfer of trained treatment policies in offline settings (where we only have access to a collection of observed trajectories $\mathcal{T}$), we model the common generative process using a causal mechanism to guide policy development in the target environment. This builds on two phases of regularization:

1. Leverage structural similarities to facilitate the use of $P^{(S)}$ as an informative prior to improve counterfactual transition estimation in **T**

2. Constraining $\pi^{(T)}$ to remain close to $\pi^{(S)}$ in order to avoid unsafe behaviors in regions of poor data support in **T**

Counterfactual Regularization

To improve the counterfactual sampling of the observed transition statistics $P^{(T)}$ we use the more accurate $P^{(S)}$ as an informative prior when establishing posterior estimates of the exogenous variables $\mathbf{U}$ defining the causal mechanisms, used to infer the Gumbel parameters. That is we use:

$$p(\mathbf{g}^{(T)}) = p(\mathbf{g}^{(S)}|\tau^{(S)})$$

when estimating the posterior over these Gumbels:

$$\begin{aligned} p(\mathbf{g}^{(T)}|\tau^{(T)}, P^{(S)}) &\propto p(\tau^{(T)}|\mathbf{g}^{(T)})p(\mathbf{g}^{(T)}) \\ &= p(\tau^{(T)}|\mathbf{g}^{(T)})p(\mathbf{g}^{(S)}|P^{(S)}) \end{aligned}$$

We use a mixture parameterization of this posterior, conditioned on observation k′ in **T**

$$\begin{aligned} p(g_1^{(T)}, \ldots, g_n^{(T)}|k') =& w^{(T)} p(g_1^{(T)}, \ldots, g_n^{(T)}| \log P^{(T)}, k') \\ &+ w^{(S)} p(g_1^{(T)}, \ldots, g_n^{(T)}| \log P^{(S)}, k') \end{aligned}$$

Regularized Policy Iteration

To avoid dangerous overconfidence in regions of low data-support in **T** we regularize $\pi^{(T)}$ through minimizing the KL-divergence between a proposed policy ν (derived from the policy improvement step of Policy Iteration) and the learned optimal $\pi^{(S)}$:

$$\nu(\cdot|S) = \frac{1}{Z_p}\left(\mathcal{R}(S,\cdot) + \gamma\left(\tilde{P}^{(T)}(S'|S,\cdot) \otimes \mathbf{V}(S')\right)\right)$$

$$\Rightarrow \pi_{k-1}^{(T)} = \arg\min_\pi \ \alpha \ \text{KL}(\pi\|\nu) + (1-\alpha) \ \text{KL}(\pi\|\pi^{(S)})$$

Simulated Clinical Validation of CFPT

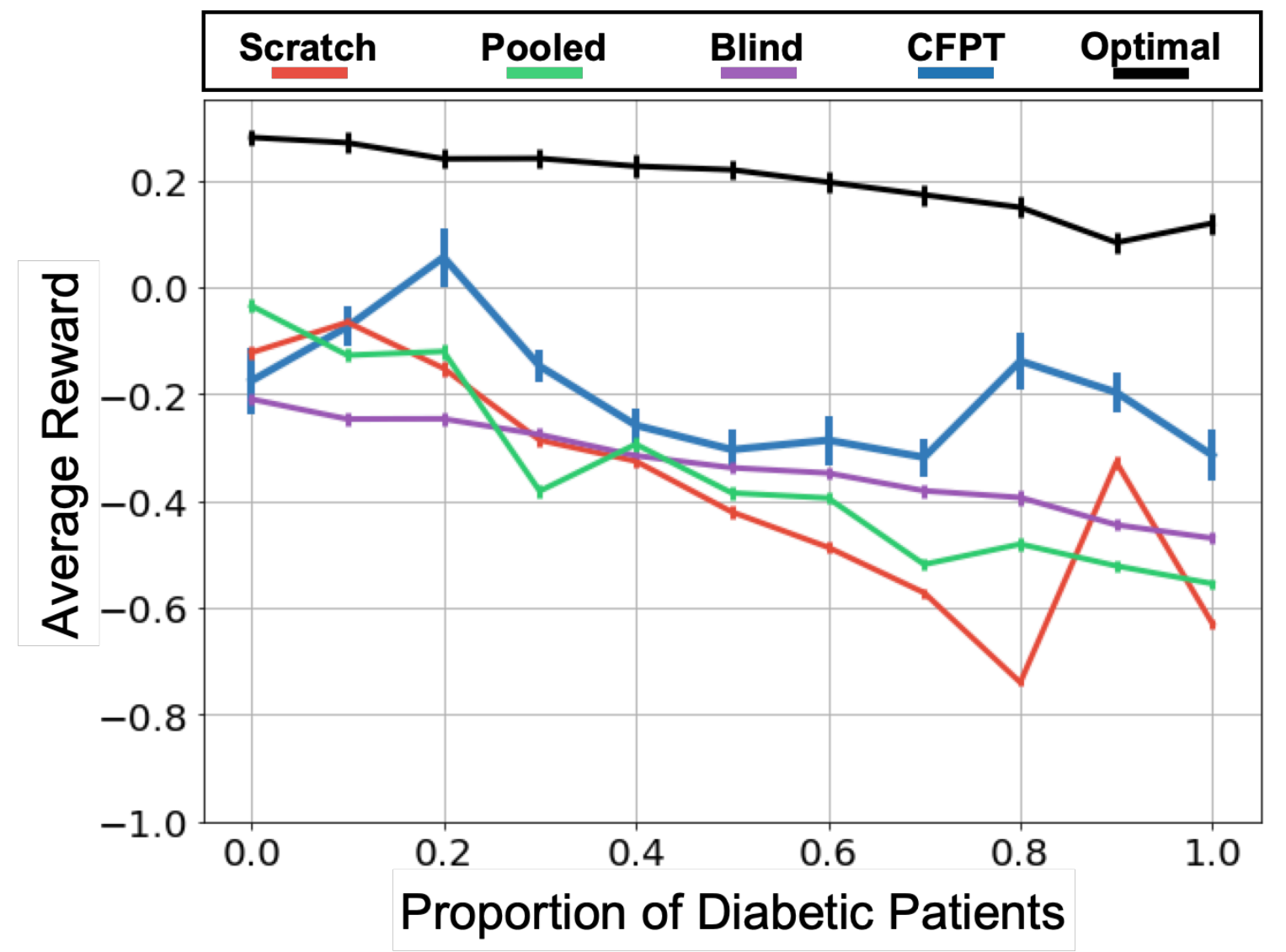
We demonstrate the benefits of the two-phase regularization of CFPT through a simulated clinical task of treating septic patients with the following features:

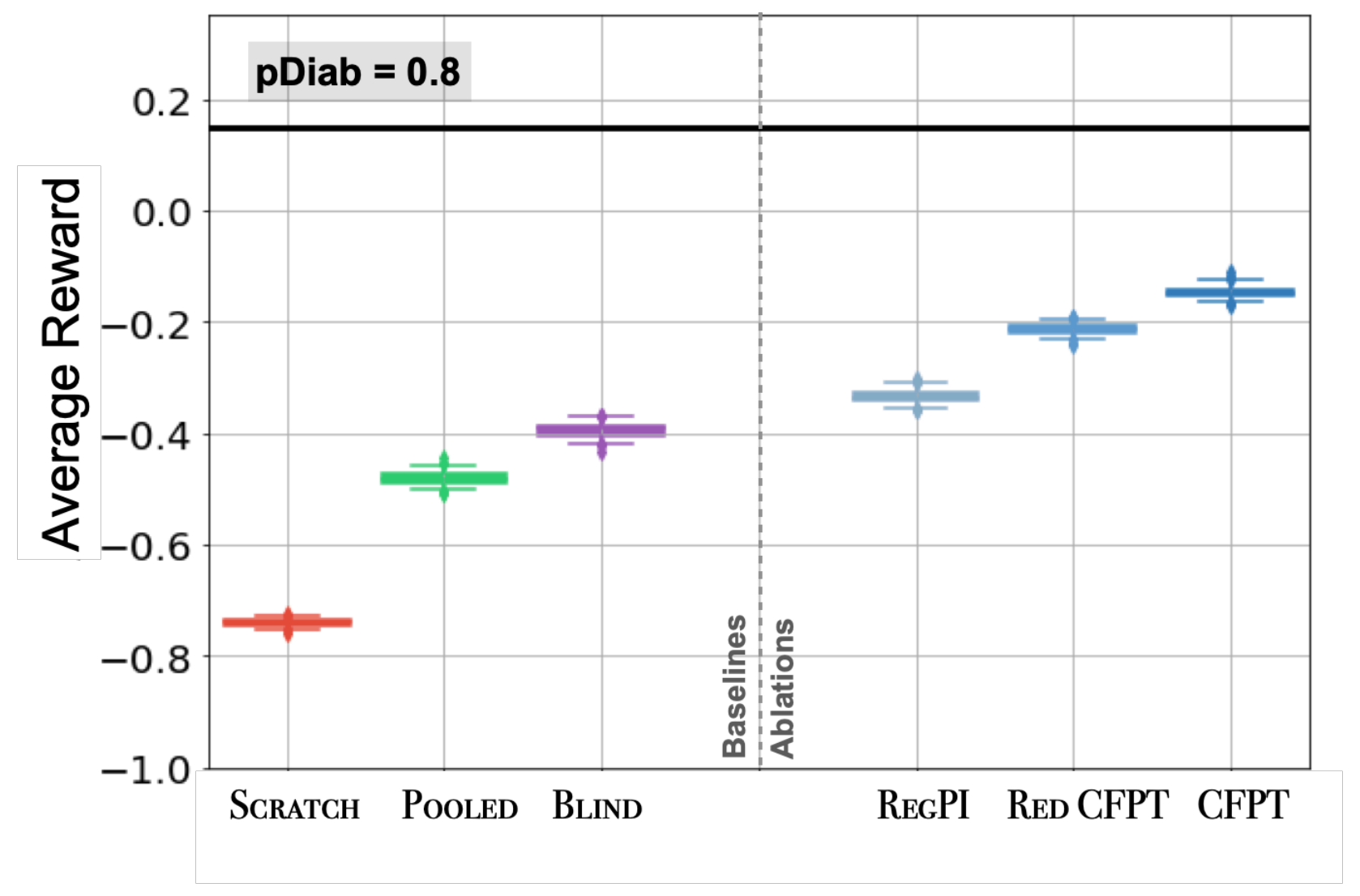
- **Heart Rate** - **Systolic Blood Pressure** - **Percent 0₂**
- **Glucose** - **Diabetic Status** (Unobserved)

We consider the task of transferring treatment policies from a data-rich environment where diabetic patients are in the minoirty (20%) to a data-scarce environment where the proportion of diabetic patients has shifted, and may be in the majority.

CFPT is Robust Under Domain Shift

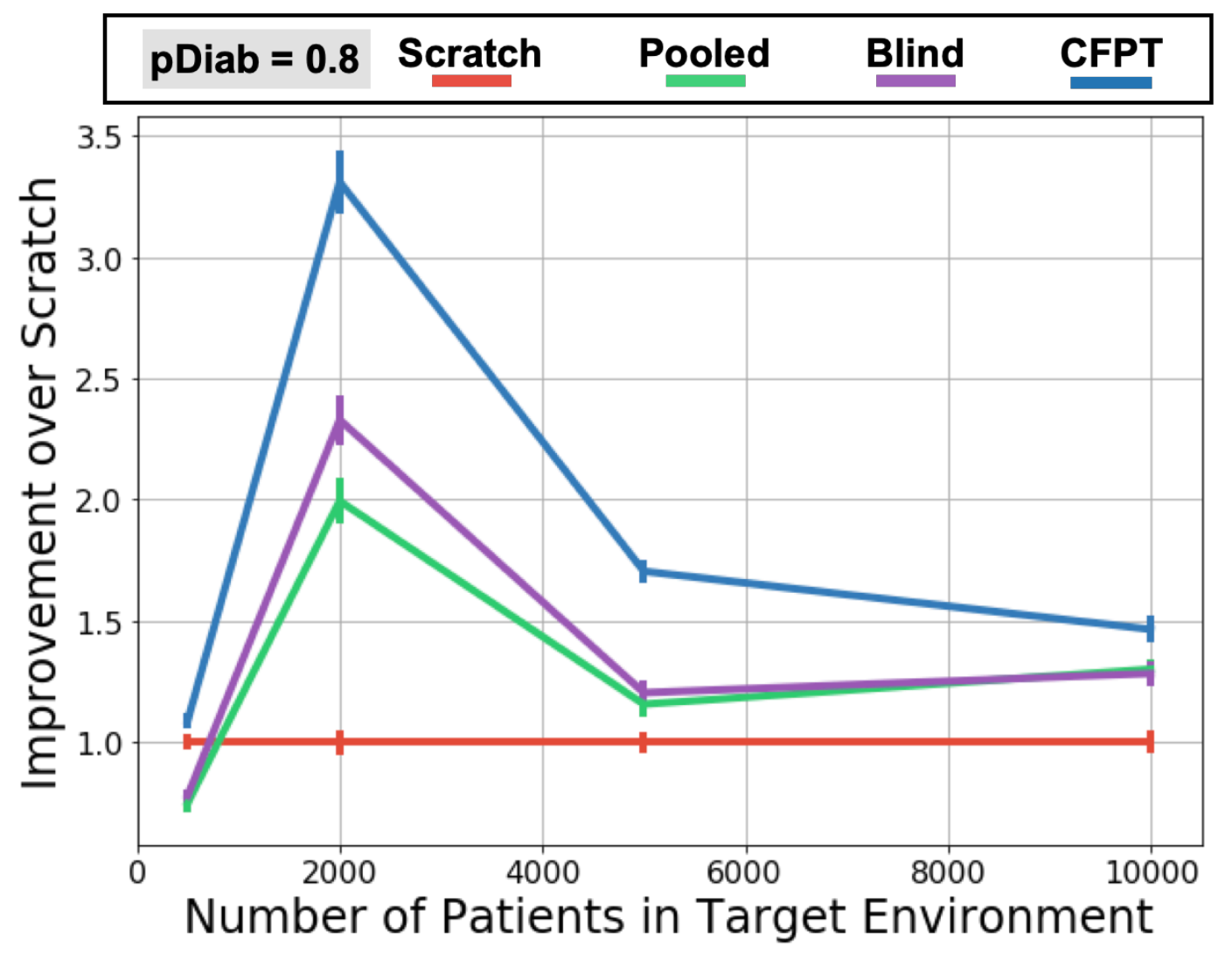
We compare CFPT to a set of standard transfer baselines and ablations across a set of target environments by varying proportions of diabetic patients



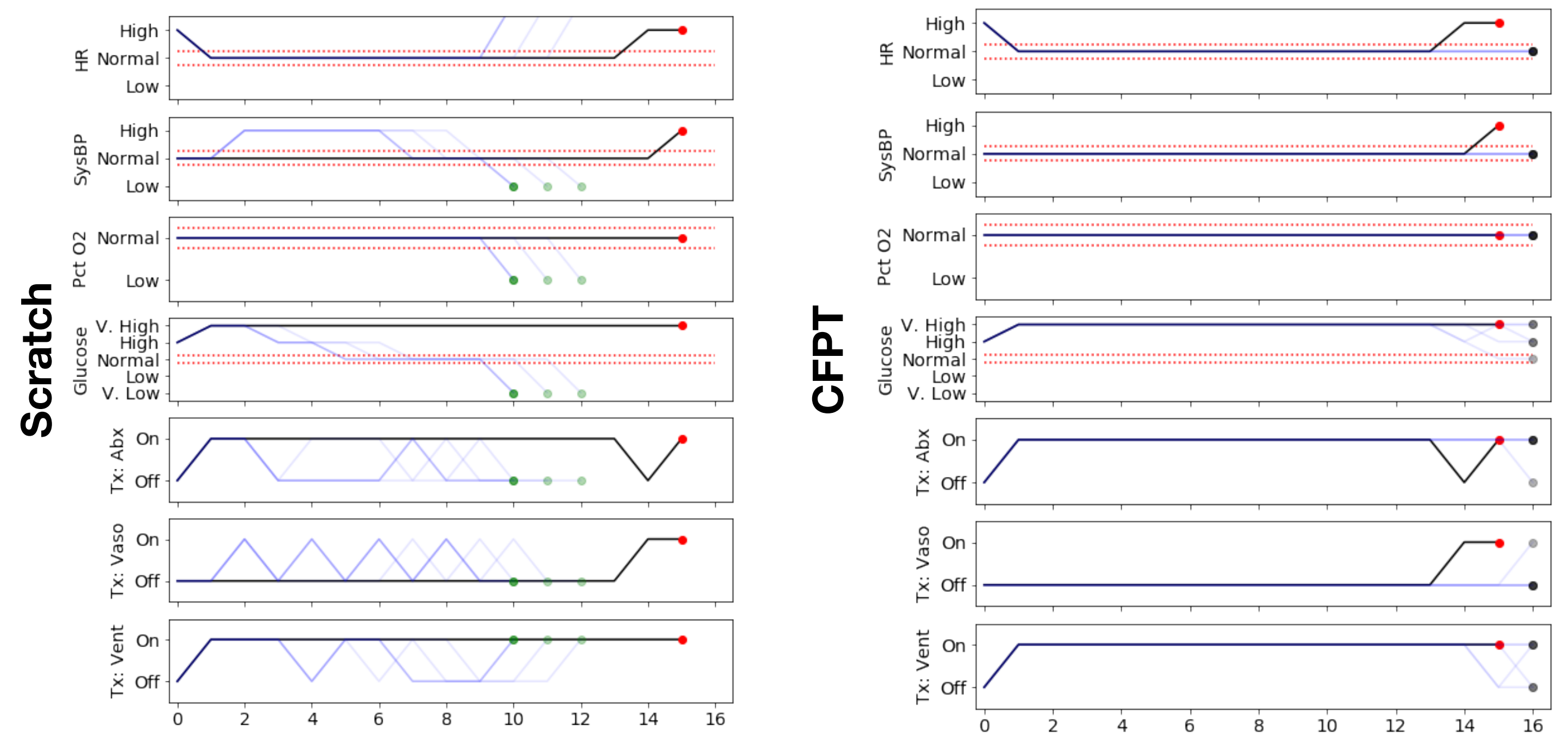


CFPT Improvement Across Levels of Data-Scarcity

We test the extent of the benefits of transfer via CFPT when the amount of data in the target environment increases. CFPT, by virtue of the regularization procedures maintains improvement over standard transfer baselines.



CFPT Develops Stable Treatment Policies



References and Acknowledgement

[1] Buesing, et al. "Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search." *ICLR*. 2018.
[2] Oberst and Sontag. "Counterfactual off-policy evaluation with gumbel-max structural causal models." *ICML*. 2019.
[3] Maddison, Tarlow and Minka. "A* sampling." *NeurIPS*. 2014

We gratefully acknowledge the fruitful discussions and advice gained from members of Vector Institute and the MIT Healthy ML group.