Background Causal inference – an experiment with random or pseudo-random partitioning of units between a treatment and control group – has come to be understood as the gold standard for scientific settings where the end goal is intervening in some process to achieve a desired end, as in genetic engineering, clinical trials or public policy design. This is for good reason: causal inference is a technique for identifying the precise impact of a given intervention on the target outcome, which is essentially impossible otherwise due to issues of confounding. Nonetheless a number of issues still plague causal inference as a method of inquiry. Perhaps the most important is failure to generalize. We see this everywhere from MPRA estimated gene expression not predicting measured expression in cells [1] to the numerous nudges that work well in laboratories and fail when scaled [2]. Often these issues of generalization are related to a shift in the underlying population tested in the lab and the actual population intervened on. **This means there is a deep connection between understanding when we should expect treatments to translate to results and out-of-distribution prediction problems in the machine learning literature.** The second major problem with causal inference it is not integrative: information from one experiment rarely if ever informs our understanding of another experiment on a similar population. A clear consequence is that causal inference has produced a fractured landscape of treatment effects without real theoretical connections, especially in the social sciences. **Using latent representations of experimental units would allow for multitask learning which would effectively share information on treatment responses across treated units.**

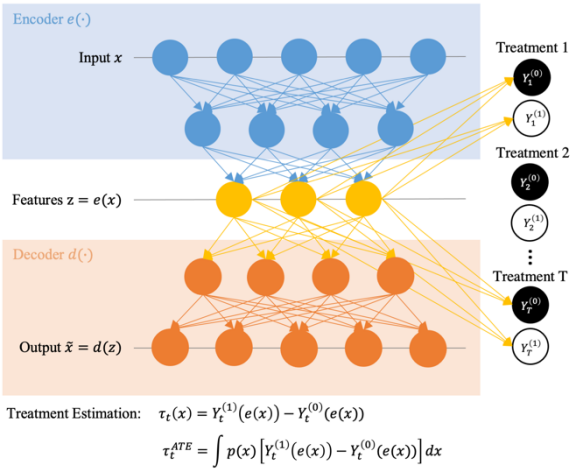
Related Work Machine Learning and causal inference is an emerging intersection with tremendous promise. Most work in the literature is focused on understanding treatment heterogeneity within a given study. Usually this is done by building a model to predict the outcome for treated and control units, then using that model to predict counterfactual treatment or control outcomes for each unit and taking the difference. The distribution of these differences captures the degree of treatment heterogeneity, which is often of interest especially in medical contexts where it is important to know if treatment effects are driven by broad effects or much higher than average effectiveness in some sub-group. Within this literature the closest work to my proposal is [3], which attempts to learn representations to improve the quality of these counterfactual predictions but does not focus either on out-of-distribution predictions for understanding generalization or learning representations for multiple experimental treatments.

Proposal Toward extending the literature on machine learning and causal inference to address the generalizability of treatments and allow sharing of information across treatment effects I propose to use learned representations of experimental units to allow for out-of-distribution prediction with calibrated uncertainty estimates and multi-task learning. Calibrated uncertainty in individual predictions should allow extrapolating from the experimental setting to the population of interest and looking at the confidence intervals to understand the expected range of outcomes. Multi-task learning, and in particular using a shared latent representation across experiments should enforce information sharing across different experimental settings, formally allowing the results of treatment in one experiment to inform the analysis of other experiments.

Research Plan Much of my work up to this point has been on representation learning of regulatory DNA and of political beliefs. In the political context I have found that even without tuning, representations can provide more robust out-of-distribution prediction. Along similar lines I have found that multi-task learning of latent representations also improves the out-of-distribution predictions. In the genetics context my work has shown that Gaussian processes offer well calibrated uncertainty estimation on samples far from the training distribution.

Aim I: Before digging deeper into method development I want to confirm these insights hold across other contexts. Does multi-task learning improve the quality of latent representations for out-of-distribution prediction in genomics as well as politics? Do Gaussian processes still provide well calibrated uncertainty if treated units are companies instead of basepairs? More generally I plan to build simulations to explore the dimensions of when and why these ideas hold up and hopefully to develop supporting theory.

Aim II: After confirming the results from my past work I want to extend these insights to develop a framework for embedding causal inference in deep learning. To build the learned representations I plan to explore Variational Auto-encoders, Auto-Encoding Generative Adversarial Networks, and comparing to a baseline using multitask learning directly and taking the last shared layer as the latent representation. The core idea is to use these latent representations to predict the outcomes for each unit in the control and treated conditions. Because of their high-quality uncertainty estimation, I plan to use Gaussian processes to make these predictions. Of course, if the learning of the latent representations is completely independent of outcome prediction there will be no information sharing across tasks. So, I am going to leverage another property of Gaussian processes: their differentiability. I plan to split training into two phases. The first unsupervised phase will just focus on training the auto-encoder for learning representations. The second phase will optimize the latent space for the predicting the outputs for all experimental settings simultaneously, updating the auto-encoder by back-propagating through the Gaussian processes (in the figure these are the yellow arrows pointing to outcomes).



Aim III: This framework is only useful if it actually works in practice. I want to conduct replications of several randomized trials that were first tested in the lab and then scaled. In particular I want to examine deworming studies from development economics [4], fixed/growth mindset work from the education literature [5], and α-1 adrenergic receptor antagonists for COVID-19 treatments [6].

The first of these failed to scale, and the second succeeded with limited effectiveness, and the third is an example where the experiment only involves older men but the target population for intervention is the general public. If my method correctly recovers the average treatment for these experiments, it would confirm the value in robustly extrapolating before taking the costly step of scaling treatments.

Intellectual Merit Should my approach to out-of-distribution confidence intervals prove successful it would have significant implications for the machine learning literature. Similarly, if integrating information across experiments proves useful for estimating treatment effects that will be very significant for work in causal inference, transforming the way we think about randomized trials. Instead of one-off experiments we could engineer large models that integrate as many effects as possible to mutually improve our understanding. Even if my main approach does not work as expected, in the process of completing this research, I will certainly be able to contribute to our understanding of when out-of-distribution prediction is easy and when it is hard, and to the literature on learning representations.

Broader Impacts Understanding when and how treatments effects will generalize when scaled up significantly is a crucial question in clinical settings and in public policy. If I am able to establish a framework that allows for more precise estimation of treatments when scaled it could greatly improve our understanding of who drugs are effective at treating, allowing greater patient understanding of expected outcomes and uncertainty. It would also improve the design of government programs, and the cost of designing government programs if extrapolation could substitute for running full scale experiments.

References [1] de Boer, Carl G., et al. "Deciphering eukaryotic gene-regulatory logic with 100 million random promoters." *Nature biotechnology* 38.1 (2020): 56-65. [2] Rai, Tage S. "Honesty “nudge” fails to replicate." *Science* 368.6488 (2020): 279-280. [3] Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." *International conference on machine learning*. 2016. [4] Miguel, Edward, and Michael Kremer. "Worms: identifying impacts on education and health in the presence of treatment externalities." *Econometrica* 72.1 (2004): 159-217. [5] Yeager, David S., et al. "A national experiment reveals where a growth mindset improves achievement." *Nature* 573.7774 (2019): 364-369. [6] Konig, Maximilian F., et al. "Preventing cytokine storm syndrome in COVID-19 using α-1 adrenergic receptor antagonists." *The Journal of Clinical Investigation* 130.7 (2020).