**Results Structure**

1. **Do genomic LMs sometimes represent functional annotations?**
   1. Yes. Probe for CMV enhancer >> Probe for Control Task
2. **If yes, are these concepts learned through pretraining?**
   1. Yes. Probe for CMV enhancer on pretrained model >> Probe for CMV enhancer on random init model
3. **Do gLMs represent concepts that don’t align with our functional annotations?** 
   1. Yes. SAE finds HIV related concept that spans across 5+ functional annotations
4. **Are these concept-directions “causally relevant” in the models forward passes?** 
   1. Yes. They can be used to predictably steer the model
   2. TO-DO: experiments:
      1. Does ablating them from the residual stream cause greater change to logits (on relevant inputs) than random directions?
      2. Maybe: compare different steering methods?

**Questions:**

* How novel is any of this?
  + Not so familiar with the background literature
* Should I present the results in the format of case studies (one for 1-2, anther for 3-4) and then put remaining results for other annotations/concepts into the appendix?