# Main Research Questions and Sub Questions

Initially:

1. **If I properly train SAEs on gLM activations, do I find monosemantic latents for biological concepts?**
   1. What does proper training consist of?
   2. Once trained, how do I check for monosemantic latents?
   3. How do measure monosemanticity?
2. **If I find monosemantic SAE-latents for biological concepts, are these biological concepts represented in the model (as SAE-latent directions)?**
3. **If these biological concepts are represented in the model in this way, can we use them to usefully change the models outputs?**

Now:

1. **Are there gLMs whose internal representations encode some information about biological function?**
   1. Can probes learn to predict some biological functions from these representations better than appropriate controls?
      1. What are appropriate controls?
   2. If probes can, does this mean internal representations *encode* this information in an important sense?
2. **If yes, does this extend to some genomic regions whose function biology hasn’t characterised yet?**
   1. How could we get evidence for that?
3. **If yes, how can we best extract this information to characterise these regions?**

# Evidential Status