We analyse a “dark-matter” SAE latent whose most activating tokens seem to have some interesting patterns.

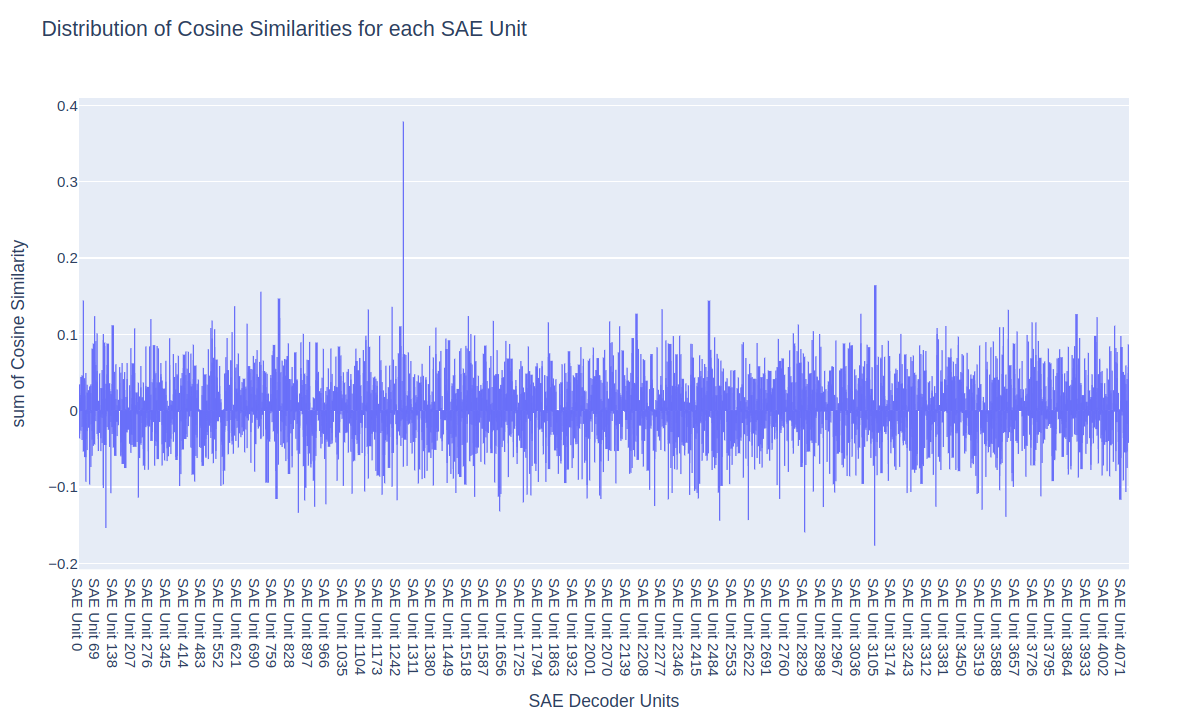
**Observation:** Latent 1264 does not selectively activate on a single functional annotation. Instead, it mostly activates on a group of annotations: RRE, env, ['cPPT/CTS', 'gag-pol'] or 'HIV-1 gag' - which are mostly associated with HIV.

(This holds across all three validation sets)

**Claim 1:** Latent 1264 is monosemantic for certain functional DNA components in HIV.

**Evidence** that Latent 1264 is HIV related:

1. BLAST confirms that the 15 most activating tokens + context are all strongly matching HIV sequences (the top-100 search results are +80% HIV related).
2. We train a linear probe to detect HIV related tokens, by labeling 1 iff the annotation includes the string ‘HIV’, we find the weights are extraordinarily aligned with the decoder row for latent 1264 (see image)
   1. Probe actually has low F1 (~20%)



1. ChatGPT, upon prompted to explain the most activating 100 tokens + context + annotation immediately explains it using HIV

**Claim 2:** This is an instance where an SAE latent has revealed a concept that we would have (unlikely) learned about through probing.

**Argument for 2:**

1. We would have been highly unlikely to probe for the combination of these 3-5 functional annotations
   1. We might well have for clearly HIV-related annotations
2. Given 1, we would have been highly unlikely to learn about this feature if we had just relied on probes

**Claim 3:** the decoder direction of this latent has special causal influence on the model outputs, i.e. if we ablate it in layer 12 from the residual stream, the change to the logits will be larger than for a random vector (for sequences that instantiate the concept?)  
  
**Evidence:** inconclusive. The differences in KL divergence tend to be greater when ablating the SAE decoder vector than random, but they are small (see fig). I’m not confident I’m calculating KL div correctly though   
  
