We are grateful to all reviewers for their high-quality consideration of our manuscript. There are some key clarifications we want to make regarding the nature of the viral escape problem, and some additional analyses we did during the intervening review period. While our comments are organized by reviewer, we hope the totality of our response will be useful to all. We will also link to a public GitHub for reproducibility in the final manuscript version.

Reviewer 1: Thanks for noting our novelty and empirical performance. • "*Appropriateness of 'grammaticality'*": We wanted to clarify a key point about viral escape (and apologize for our lack of clarity), which we believe addresses your concern. To escape from human immunity, not only does a mutation need to preserve infectivity, but it *also* must be functionally/antigenically altered enough so that antibody recognition no longer works. So, the fact that our training data comes solely from infectious virus, which would be highly probable (or "grammatical") sequences under our language model (LM), is a key feature of our approach. We intuit that it is "grammaticality" that encodes infectivity, while it is "semantic change" that encodes the functional/antigenic alteration essential to immune escape (rather than, e.g., a mutation that preserves infectivity but is functionally neutral and therefore does not affect antibody recognition). • "*consider these comparisons*": Regarding points about comparison/benchmarking on viral finetuned embedding models and generative models of evolutionary data, during the review period, we have auspiciously performed new benchmarks that we can hopefully add to the paper. We newly benchmark with EVcouplings (a refined version of the method in Hopf et al., *NBT*, 2017) on the same viral training dataset as our model and we were also able to update Bepler/TAPE results as well. Among these, the highest AUC for H3 is EVcouplings independent with 0.691, for H1 is EVcouplings epistatic with 0.726, and for Env is TAPE with 0.574, which are all lower than the CSCS results with our LM. Importantly, however, we note that, fundamentally, CSCS is presented in generality here so these methods are not strictly "competitor methods" in the sense that, if one were to work better, it would still be incorporable within the CSCS framework. • "*slightly ad hoc nature of the CSCS objective*": One nice thing about the CSCS objective is that it has a straightforward interpretation and its simplicity (combined with our empirical results) illustrates a rather direct connection between our modeling intuition and nature, and we can certainly replace the appeal to BO with an appeal to Lagrange multipliers (thanks for this point). • "*Correctness*": We apologize for the lack of clarity and indeed note a random 80/20 CV split within the training dataset before application to the temporally held-out test dataset; we will clarify in the Appendix. • "ℓ_1 *rather than Euclidean*": We used ℓ_1 since it has nicer properties than, e.g., ℓ_2 in high-dimensional spaces (Aggarwal et al., *ICDT*, 2001) but other distance metrics could be empirically quantified. • "*Clarity*": Thanks for your detailed points; we will incorporate.

Reviewer 2: Thanks for recognizing our impact and your correctness points are really helpful. We will definitely discuss embedding closeness of antonyms in the NLP setting and add an explicit remark on how LM probabilities, our definition of (soft) grammaticality, can also encode NLP pragmatics. More generally, we will clarify our definitions of these terms, borrowed from NLP, in the viral protein setting. We will incorporate all your detailed specific suggestions.

Reviewer 3: Thanks for noting the strength of our unsupervised setup and broader impacts. • "*conditionally independent*": We apologize for a lack of explanation; conditional independence is by construction of the model architecture, since we use the entire hidden-layer output as the latent variable embedding. We noted CI just to show that conditioning on $\hat{\mathbf{z}}_i$ (or plugging in the embedding values before the final softmax layer) is sufficient to compute the final mutation probability, so $\hat{p}(x_i|\mathbf{x}_{[N]\setminus\{i\}}, \hat{\mathbf{z}}_i) = \hat{p}(x_i|\hat{\mathbf{z}}_i)$ for the learned distribution $\hat{p}$. • "*continuous measure*": In the intervening time, we performed additional analysis correlating the CSCS objective with continuous differential selection scores, with CSCS also performing the best (which we can include). We do note selection scores are consistently and clearly bimodal, indicating more or less binary escape. • "*practical implications*": Thanks for this, which is essential to discuss; typically, a physical constraint (e.g., mutational library size or sequencing cost) directly provides a top N to acquire, which is the best way (and already a practical one) to use CSCS predictions now. Useful further work could identify an absolute threshold beyond which to acquire mutations (perhaps by quantifying prediction uncertainty).

Reviewer 4: Thanks for noting your conceptual interest and the strength of our empirical results. • "*theoretical detail*"/"*how the method works*": We apologize for sparsity of detail. The fundamental reasons typically given for how/why LMs work use an appeal to the distributional hypothesis (our paper's refs [22, 25]). Our work builds off of recent extensions of LMs to protein sequences and is motivated by the broader impact of this approach for studying infectious viruses. • "*how the semantic change was computed*": Sorry for the confusion here and you're right; for a given protein sequence, we evaluate the BiLSTM at each position to obtain a sequence-length-by-embedding-dimension matrix, which is also what previous protein embedding approaches produce (refs [9, 44, 4]). This matrix can be flattened (as we do) or averaged across the sequence dimension to compare proteins/compute distances.• "*exhaustive search*": Since we focus on single-token mutations, an exhaustive search (scales with alphabet-size-by-sequence-length) suffices for our purposes, especially since current experimental validation data is also limited to single-residue mutations. We note for (viral) proteins, alphabet size (i.e., number of natural amino acids) is constant and sequence length is typically in the high 10^2 or low 10^3 range. Combinatorial search will require a different strategy, which suggests highly interesting future work. • "*choice of beta*": We find good robustness of β values reasonably close to 1 (e.g, 0.5-2). Results do start to change after more than a 2X increase or decrease (e.g., when $\beta = 0.25$).