We thank the reviewers for their thoughtful feedback during this difficult time. We are glad they found our writing clear (R2), and our direct likelihood approach to generating knockoffs novel (R1, 3) yet simple (R4). It was particularly encouraging to see the reviewers acknowledge the usefulness of the Gumbel-Softmax trick in ensuring the validity of sampled knockoffs over a large number of potential swaps (R1, 2, 4). We are also glad the application of our method to a timely COVID-19 task stood out (R1, 3). The two types of questions brought up in the reviews involved either (a) possible additional experiments, or (b) clarifications of experimental details or notation. We will answer each question, and have incorporate all feedback into the final version. At the end, we discuss our COVID-19 data source.

Additional results/experimentation. Below, we discuss several thoughtful suggestions brought up by reviewers.

Comparison with second-order knockoffs (R1): We compare to Deep Knockoffs, which have been shown to outperform second-order knockoffs (Romano et al. 2018) in a variety of settings. Regardless, we have added a comparison with second-order knockoffs for completeness. **FDR and power as a function of the signal strength** (R1, 3): Both the KnockoffGAN and Deep Knockoffs papers explore this relationship and show that their methods are robust to the choice of signal strength in gaussian data. We observed this trend as well for DDLK, and figured this result would not sufficiently contrast each method. **Measures of knockoff quality and goodness-of-fit tests** (R1, 2): This is a great point and is also realistic for practitioners. We considered using the MMD-based two-sample test from Romano et al. (2018), but found issues as dimensionality increased (even just > 10) making it a poor test for knockoff quality. Ramdas et al. (2015) discuss this issue in detail and suggest that KL is a better measure of discrepancy in high dimensions. However, to evaluate KL, a likelihood is required. At least in the case of DDLK, we can inspect the KL term (L114) on a held-out set. We don't report this statistic in the paper simply because it is not available for likelihood-free models. An extension of our work could include a likelihood-free test for goodness of fit that avoids using MMDs. To check if knockoffs from DDLK will have high power, the conditional entropy term (L141) can be used: higher the conditional entropy, higher the power. We have added a line to the paper discussing empirical checks of knockoff quality. **Add feature selection frequencies in Table 1** (R1): This is a good suggestion, we have added this to Table 1. **Global null** (R2): We felt these experiments would not differ significantly from our existing ones. Valid knockoffs will yield FDR control regardless of how many important covariates there are. Under the global null, the family-wise error rate will also be controlled since in this case FDR = family-wise error rate. **Hardest swaps in two-sample test** (R2): This is a very interesting idea on its own with other potential applications. It definitely seems feasible to create a goodness-of-fit test using the hardest swaps.

Clarifications of experimental details and notation. The clarifications requested are concentrated in one of two categories: details about the Gumbel-Softmax sampling of swaps (R2), and understanding our mathematical notation (R4). We will clarify these below and incorporate all feedback in the final version.

Sampling swaps with β (R2): For clarity, we quickly recap the full procedure here. The parameter for the jth Gumbel-Softmax (GS) is β_j: the probability of sampling a 1. To sample a swap H, we first sample all d GS distributions. If and only if a 1 was sampled from the jth GS, then $j \in H$. In practice, the GS is a continuous relaxation of a discrete distribution and will only yield discrete samples if the temperature $\tau \to \infty$. For our implementation of DDLK, we use the straight-through GS estimator (line 174), which allows binary values to be sampled, but uses the continuous approximation in gradient computations. Further, we followed the annealing schedule approach for choosing τ as prescribed by Jang et al. (2017). We have clarified this process in section 3. **Gaussian experiment response fully dense** (R2): Thank you for pointing this out. The uploaded pdf was missing a sentence that states that 20 out of 100 coordinates were randomly chosen as important. Only these important coefficients get the value α_j. The rest get 0. This sentence has been added to the paper. **Second "q" should be"qw"** (R2): Thanks, we corrected this to $q_{\mathbf{w}}$. **Figure 2 visualizes only two dimensions** (R1): Figure 2 is intended to be a diagnostic to help understand differences between each method, rather than an exact measure of goodness-of-fit. **Mode collapse of baseline in L221** (R4): We agree that mode collapse is a sign of unsuccessful training. However, for Deep Knockoffs and KnockoffGAN, we used publicly available implementations. For Auto-Encoding Knockoffs, there was none, so we implemented the method as prescribed in the original paper. Further, we ran an extensive grid search over hyperparameters for each baseline method as discussed in the original papers. We have added a section discussing these in the appendix. **Flexibility of likelihood-free models** (R4): L39 suggests that likelihood-free generative models (e.g. GANs) can be easier to specify for arbitrary covariate distributions because they don't require a specific likelihood. We added a sentence to make this more clear. **Section 2** (R4): We have renamed section 2 to "Background" to remove this confusion. **Correlated gene features** (R4): We chose a dataset where the covariates are highly correlated to make knockoff generation more difficult. We added a sentence to reiterate this fact. **Mathematical notation** (R4): We have added an appendix section explicitly defining each object introduced in section 3 (e.g. what is $\mathbf{y}$, the set $[d]$, etc.).

COVID-19 data. (R3): We are glad the reviewer found the paper interesting and well written. Our COVID-19 EHR data was collected for quality improvement (QI) purposes rather than research, so no IRB approval was required. Through the QI process, the data was de-identified by someone else for another project making our use no longer for human subjects. This data has been made available through the HIPAA Business Associate Agreement to third parties.