We would like to thank all the reviewers for their valuable reviews and constructive feedback which will help us improve the paper. We appreciate the reviewers' very positive comments, noting e.g. that the work is *"technically sound", "very well written and easy to understand", and "an important contribution"*, the idea of LLR *"opens room for developing novel techniques in future papers", "the experiments are strong"*, and it introduces *"a novel dataset that will be of utility to the field"*, etc. Due to space constraints in the rebuttal, we focus on only the major questions and comments below, but we will address the minor comments too in the camera ready.

R1: *"Quantitative metrics rather than simply say they 'largely overlap'."* The AUROC of 0.630 for likelihood (Table 3) also suggests that the two histograms (Figure 1a) are barely separable. We will add it to caption of Figure 1a in revision.

"The authors assume that (i) each dimension of the input space can be described as either background or semantic, and (ii) these are independent (Eq 1). How true is this in practice?" Thanks for the question. We made these two assumptions in lines 118-145 to intuitively describe the high level idea of our method, but note that we later relaxed (ii) in lines 146-154. Assumption (i) is relatively mild as many input modalities consist of background plus semantic components (given appropriate resolution) e.g. background and semantic pixels in images; background noise in audio signals; key words and stop words in text; genomics studies have modelled DNA sequences with motif and background models (Reinert et al., 2009; Wan et al., 2010; Zhai et al., 2010) with successful applications (Song et al., 2013; Chan et al., 2014; Ahlgren et al., 2016). *The independence assumption in (ii) was relaxed in line 146, so our method does not rely on (ii).* Specifically, we decompose joint as product of conditional distributions using the equality $p(\boldsymbol{x}) = \prod_{d=1}^{D} p(x_d|\boldsymbol{x}_{<d})$, and use auto-regressive models, which can model any $p(\boldsymbol{x})$ given sufficient capacity to represent true conditionals $p(x_d|\boldsymbol{x}_{<d})$; x_d is conditioned on all past $\boldsymbol{x}_{<d}$, so it does not introduce additional independence assumption. In Eq (4), we simply grouped likelihood terms into two groups depending on if x_d belongs to background or semantic group.

"LLR as defined in Eq (5) depends only on the semantic features–how are these identified in practice on the test set? We do not assume knowledge of z in our experiments (we will make it explicit in the text to avoid any confusion). If the modeling assumptions are correct, we expect LLR to be non-zero in the semantic part and be close to 0 on the background part. Since auto-regressive models give per-pixel likelihoods, we can validate if our hypothesis is correct by visualizing the heat map of per-pixel likelihoods and ratios. Figure 3 validates our hypothesis experimentally.

"Do you have an explanation for AUROC being significantly worse than random on Fashion MNIST?" Yes, we show that the proportion of zeros, i.e. the number of background pixels in an image, is highly positively correlated with the likelihood. MNIST (ood) images on average have a larger proportion of background pixels and so they get assigned higher likelihood than FashionMNIST (train). See Figures 2a-b and Figure S1,S2 in supp mat for more details.

"...the requirement to have knowledge of the data perturbation process." Our Bernoulli perturbations (Alg 1) make very few assumptions about data and are really simple to implement. We show that the same procedure works well for two different modalities (image and genomic data), demonstrating the generality of the methodology. We also found that training background model using just L_2 regularization (without any data perturbation) works effectively in some cases (cf. μ=0 in Table S3). We believe other ways of training the background model can further improve performance.

R2: *"Additional baselines for CIFAR-10 versus SVHN."* As mentioned in line 228, our goal for image experiments was to show that LLR effectively corrects for background statistics and significantly outperforms the likelihood, and not to claim that we achieve state-of-the-art performance on these datasets. That said, we agree that having other baselines will make the table more complete. Due to limited time for rebuttal, we implemented a few classifier-based methods (using ResNet) and added the corresponding evaluation metrics to Table 1. We will provide the full table in the camera ready version. Note that our LLR is completely unsupervised whereas classifier methods require labels. Using LLR on class-conditional generative models might further improve performance.

Table 1: Additional CIFAR-10 vs SVHN results.

	AUROC↑	AUPRC↑	FPR80↓
$p(\hat{y}\|\boldsymbol{x})$	0.914	0.877	0.084
Entropy of $p(\hat{y}\|\boldsymbol{x})$	0.924	0.894	0.132
Mahalanobis distance	0.901	0.920	0.127
Ensemble 5	0.940	0.911	0.033
Ensemble 10	0.943	0.913	0.024
Ensemble 20	0.946	0.916	0.019
$p(\hat{y}\|\boldsymbol{x})$ with noise class	0.922	0.889	0.063

"Report errors on metrics." Thank you for the suggestion. We will update those in the camera ready version. *"Runtimes."* The runtime of our LLR method is two times of the standard PixelCNN++ runtime.

"Releasing code for baselines and the benchmark." We agree that it is important to release the code and data and have been working on this since submission. The dataset and code for the genomic experiments have already been shared publicly online (for anonymity we will not share links here). All code will be publicly available by camera ready.

R3: Thanks for your positive comments and pointing us to the idea of learning a statistical potential in protein folding.

"For baseline method 9 (Choi et al.), is the generative model used the same as LLR (this paper's method)?" Yes, we do compare the two methods (LLR, WAIC) using the exact same generative model (PixelCNN++), as it allows us to compare the effectiveness of the two methods, while controlling for the generative model class.