A. Proposition 1

Proposition Let P be the distribution under the null hypothesis H_0. Let μ be the measure associated with the distribution of test statistic $\phi_p(\mathbf{x})$ under the null. Then, assuming conditional $\mathbf{x} \mid \phi_p(\mathbf{x})$ is not degenerate on μ-non-measure zero set, there exists a set of alternative distributions $Q \in \mathcal{Q}$ where $Q \neq_d P$ and the test has power equal to the false positive rate. In other words, the test does no better than random guessing.

Proof. We first construct a distribution $q(\mathbf{x}) \neq_d p(\mathbf{x})$ but where $q(\phi_p(\mathbf{x})) = p(\phi_p(\mathbf{x}))$.

The roadmap for this part of the proof is as follows: for some function f, we write

$$\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) = \mathbb{E}_{p(\phi_p(\mathbf{x}))}\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] \tag{3}$$

We then identify $q(\mathbf{x}|\phi_p(\mathbf{x}))$ and f_p such that the inner difference of expectations is non-zero, which implies inequality in distribution via $\mathbb{E}_{q(\mathbf{x})}(f_p) \neq \mathbb{E}_{p(\mathbf{x})}(f_p)$. We do not change the distribution in the outer expectation $p(\phi_p(\mathbf{x}))$. We finally define $q(\mathbf{x}) = p(\phi_p(\mathbf{x}))q(\mathbf{x}|\phi_p(\mathbf{x}))$.

We now show how to construct f_p, q. Let $(\Omega_{\phi_p(\mathbf{x})}, \mathcal{F}_{\phi_p(\mathbf{x})})$ be the probability space associated with $\phi_p(\mathbf{x})$, with probability measure $\mu = \mathbb{P}_{p(\phi_p(\mathbf{x}))}$. By assumption, $p(\mathbf{x} \mid \phi_p(\mathbf{x}))$ is non-degenerate on some μ non-measure zero set. This means there exists a set $\Phi \in \mathcal{F}_{\phi_p(\mathbf{x})}$ with $\mu(\Phi) > 0$ such that $\forall \phi_p(\mathbf{x}) \in \Phi$, $\exists A_{\phi_p(\mathbf{x})} \subset \text{supp}(p(\mathbf{x} \mid \phi_p(\mathbf{x})))$ such that $0 < \mathbb{P}_{p(\mathbf{x} \mid \phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) < 1$.

Let g be any function for which $\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(g) < \infty \; \forall \phi_p(\mathbf{x}) \notin \Phi$. Then define

$$f_p(\mathbf{x}) \triangleq \mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right] \mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \Phi\right] g(\mathbf{x}) \tag{4}$$

Define the conditional $q(\mathbf{x}|\phi_p(\mathbf{x}))$ with normalization constant $C_{\phi_p(\mathbf{x})}$ and $0 < \lambda < 1$:

$$\begin{aligned} q(\mathbf{x}|\phi_p(\mathbf{x})) \triangleq\; & \mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right] \Big[\frac{1}{C_{\phi_p(\mathbf{x})}}\Big(\lambda p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right]\Big)\Big] \\ & + \mathbb{1}\left[\phi_p(\mathbf{x}) \notin \Phi\right] p(\mathbf{x}|\phi_p(\mathbf{x})) \end{aligned} \tag{5}$$

For $\phi_p(\mathbf{x}) \notin \Phi$, $q(\mathbf{x}|\phi_p(\mathbf{x})) = p(\mathbf{x}|\phi_p(\mathbf{x}))$. Therefore, $\mathbb{1}\left[\phi_p(\mathbf{x}) \notin \Phi\right][\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f)] = 0$. For $\phi_p(\mathbf{x}) \in \Phi$, $q(\mathbf{x}|\phi_p(\mathbf{x}))$ moves density away from points in $A_{\phi_p(\mathbf{x})}$ relative to $p(\mathbf{x}|\phi_p(\mathbf{x}))$, given that $0 < \lambda < 1$.

For simplicity, we construct q such that $\text{supp}(q(\mathbf{x} \mid \phi_p(\mathbf{x}))) = \text{supp}(p(\mathbf{x} \mid \phi_p(\mathbf{x})))$. This is to avoid any issues with an invalid joint distribution $q(\mathbf{x}, \phi_p(\mathbf{x})) \neq q(\mathbf{x})$ if $q(\mathbf{x}|\phi_p(\mathbf{x})) = 0$ (the left-hand side would be 0 while the right-hand side would be greater than 0 $\forall \mathbf{x} \in \text{supp}(p(\mathbf{x}))$.

We now show that this construction leads to $\mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) > 0$, implying inequality in distribution.

$$\begin{aligned}
& \mathbb{E}_{p(\mathbf{x})}(f_p) - \mathbb{E}_{q(\mathbf{x})}(f_p) && (6)\\
=& \mathbb{E}_{p(\phi_p(\mathbf{x}))}\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] && (7)\\
=& \mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(f_p)\Big] && (8)\\
=& \mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\mathbb{E}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right]) - \mathbb{E}_{q(\mathbf{x}|\phi_p(\mathbf{x}))}(\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right])\Big] && (9)\\
=& \mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} q(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x})\Big] && (10)\\
=& \mathbb{E}_{p(\phi_p(\mathbf{x}))}\mathbb{1}\left[\phi_p(\mathbf{x}) \in \Phi\right]\Big[\int_{A_{\phi_p(\mathbf{x})}} p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x} - \int_{A_{\phi_p(\mathbf{x})}} \frac{1}{C_{\phi_p(\mathbf{x})}}\lambda p(\mathbf{x}|\phi_p(\mathbf{x}))d\mathbf{x})\Big] && (11)\\
>& 0 && (12)
\end{aligned}$$

Line 11 follows from the substitution of $q(\mathbf{x}|\phi_p(\mathbf{x}))$ defined in Equation (5). Line 12 follows from the fact that $\frac{\lambda}{C_{\phi_p(\mathbf{x})}} < 1$, shown below:

$$C_{\phi_p(\mathbf{x})} = \int_{\mathcal{X}} \lambda p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \in A_{\phi_p(\mathbf{x})}\right] + p(\mathbf{x}|\phi_p(\mathbf{x}))\mathbb{1}\left[\mathbf{x} \notin A_{\phi_p(\mathbf{x})}\right] d\mathbf{x} \tag{13}$$

$$= \lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^C_{\phi_p(\mathbf{x})}) \tag{14}$$

$$= \lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + 1 - \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) \tag{15}$$

$$\frac{\lambda}{C_{\phi_p(\mathbf{x})}} = \frac{\lambda}{\lambda \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + 1 - \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})})} \tag{16}$$

$$= \frac{1}{\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \frac{1}{\lambda}\left[\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^c_{\phi_p(\mathbf{x})})\right]} \tag{17}$$

$$< 1 \tag{18}$$

Line 18 holds since the denominator in the previous line is greater than 1: Since $0 < \lambda < 1$, $\frac{1}{\lambda} > 1$. Then, $\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \frac{1}{\lambda}\left[\mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^c_{\phi_p(\mathbf{x})})\right] > \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A_{\phi_p(\mathbf{x})}) + \mathbb{P}_{p(\mathbf{x}|\phi_p(\mathbf{x}))}(A^c_{\phi_p(\mathbf{x})}) = 1$.

Having constructed the density q, we now proceed with the second part of the proposition: for any specified false positive rate, any test based on ϕ_p has power equal to the false positive rate when the OOD samples come from q.

Recall that $q(\phi_p(\mathbf{x})) = p(\phi_p(\mathbf{x}))$. Then, for any rejection rule $\phi_p(\mathbf{x}) \notin \Phi_{\text{Accept}}$, the probability of rejection is the same regardless of whether the sample $\mathbf{x}$ is drawn from P or q:

$$\forall \Phi_{\text{Accept}}, \quad \mathbb{P}_{\mathbf{x}\sim q}(\phi_p(\mathbf{x}) \notin \Phi_{\text{Accept}}) = \mathbb{P}_{\mathbf{x}\sim p}(\phi_p(\mathbf{x}) \notin \Phi_{\text{Accept}}). \tag{19}$$

Therefore, the power of the test (i.e. rejecting under the $H_A : \mathbf{x} \sim q$) is equal to the false positive rate (i.e. rejecting under $H_0 : \mathbf{x} \sim p$). When power and false positive rate are equal for all possible values of the false positive rate, then the result is an ROC curve $y = x$ with AUC 0.5. This is equivalent to random guessing with rejection rate based on the false positive rate chosen for the test. □

B. Rejection Rules Can Be Written in the Form $\phi_p(\mathbf{x}) < k$

Lemma 1 *Any rejection rule involving intervals, i.e. $\phi(\boldsymbol{x}) \notin \Phi$ can be recast as a rule of the form $\phi'(\boldsymbol{x}) < k$.*

Proof If we have a one-sided rule, i.e. an interval Φ where one of the endpoints is $-\infty$ or ∞, we simply reverse the sign if necessary, and for two-sided rules, i.e. a bounded interval, we can find the midpoint of the interval m, where $\Phi = [m - k, m + k]$, and recast the rule to $|\phi(\mathbf{x}) - m| < k$.

Rejection rules of this form match the same "rejection" rules used for binary classification more broadly. For added clarity, we define some OOD detection methods based on their rejection rules in this form. For instance, the likelihood-based test rejects when the negative log likelihood is above a certain threshold k, whereas the typicality test rejects when the distance to the training set entropy is above k.

C. Details for Experiment 5.2

In this experiment, we compare a partially trained GLOW model p_θ with a pretrained GLOW model (Kingma & Dhariwal, 2018) which we use as our data distribution P. First, we generate samples from P by sampling from the pretrained GLOW model [4]. We use temperature 1 for sampling to ensure our samples come from the distribution specified by the model. We generate 40,000 samples for training and 10,000 samples for evaluation. These are the analogous sizes to the train and test sets of the CIFAR-10 dataset, which is the dataset the model P was pretrained on.

The glow (GLOW) model p_θ is made of 3 blocks, each with 8 affine coupling layers with 400 hidden units per layer. The network is trained with Adamax at learning rate 0.001, which stays constant after 10 epochs of warmup. We use batch size 64 during training. We intentionally limit the training (50 epochs with 10 epochs of warmup) to make the model

[4] https://openaipublic.azureedge.net/glow-demo/logs/abl-1x1-aff.tar

mis-estimation clear. Our model achieves an average bits per dimension of 3.67 on the test samples, versus 3.45 for the true model (lower is better).

The true model is a larger model than p_θ, consisting of 3-blocks each with 32 affine coupling layers with 400 units each.

We evaluate OOD performance on the test set of the model samples and the test set of CelebA.

D. Existing Model Architectures Can Yield Good OOD Detectors

We directly optimize a PIXELCNN++ to distinguish between FashionMNIST and MNIST by replacing the maximum likelihood training objective with one which simultaneously maximizes likelihood on FashionMNIST images while minimizing likelihood of MNIST images. Our objective is similar to that of Kirichenko et al. (2020), who show that flows can distinguish problematic OOD dataset pairs when optimized directly to do so.

$$\frac{1}{N_{in}} \sum_{x \in \mathcal{D}_{in}} \log p_\theta(x) - \frac{1}{N_{ood}} \sum_{x' \in \mathcal{D}_{ood}} \min(\log p_\theta(x'), c) \tag{20}$$

Replicating the architecture of Ren et al. (2019), we train our model across five random seeds using the MLE objective and five seeds with the above objective. We use the same training hyperparameters as Ren et al. (2019): 50,000 steps at a learning rate of 0.0001 with exponential decay rate of 0.999995 per step, batch size of 32, and Adam optimizer with momentum parameters 0.95 and 0.9995. Our results, shown in Table 2, demonstrate that the PIXELCNN++ architecture has the capacity to push down probabilities on problematic OOD samples while maintaining high in-distribution likelihoods.

Table 2. Existing models can be optimized to distinguish datasets. PIXELCNN++ trained via the negative training (NT) objective in Equation (20) can achieve near-perfect OOD detection while maintaining comparable held-out log likelihoods (LL) with models trained via maximum likelihood estimation (MLE). Results averaged over 5 random seeds.

	Fashion LL	OOD AUC
MLE	-1550 ± 6	0.097 ± 0.004
NT	-1562 ± 7	1.000 ± 0.000