Thanks to all reviewers for the insightful comments! We have already open-sourced most of our code (link hidden to preserve anonymity) and will provide the link at the beginning of Section 5 of the final paper. Detailed responses below:

R1: The overall method seems to be not end-to-end. Algorithm 1 is in fact an end-to-end algorithm. All the trainable parameters β and $\{W_l, b_l\}_{l \leq L-1}$ are updated simultaneously in the main SGD step. The only parameters that are not updated are the W_L and b_L, which are fixed random features that are frozen throughout the learning. The precision matrix update is only a side computation for summary statistics derived from the SGD-updated model. As a result, the optimization procedure is very similar to the determinisitic SGD training. We use the same initialization for the hidden weights as in the deterministic model, and use the default Glorot uniform initializer for the GP output layer weights β.

R1: Clarity regarding Eqn (5). Thanks for highlighting a point that merits more clarification. Briefly, different from a classic minimax problem, we derived (5) under the assumption that we have partial knowledge about p^* (i.e., we know the domain probability $p^*(x \in X)$), therefore it is possible for the known property of p^* to enter into the final expression. Please see Appendix B for a full statement of the motivation, the assumptions and the proof for Eqn. (5). In particular, please see line 619-626 for an explanation of why (5) is structured as such. Eqn (5) (corresponds to Appendix Proposition 2) is used to motivate Section 2.2. In the final paper, we will include additional explanations / pointers to Appendix B around line 100, and replace "Proposition 2" with "Equation (5)" on line 110 to improve clarity.

R2: It seems that the $\mathcal{X}$ is with L_2-norm as distance metric. We in fact allow the metric for $\mathcal{X}$ to be non-Euclidean so it reflects the semantically meaningful distance in the data space (please see statement on line 55-56, the discussion on line 141-149, and the proof for Proposition 1 in Appendix D.2. which does not impose restriction on $||.||_X$ and is based on the theoretical work of [3]). In addition, we'd like to point out that in the vision / language experiments (Section 5.2), the SNGP has superior performance in distinguishing in-domain / out-of-domain data, which is not likely if SNGP can only preserve a L_2 metric, which is not suitable for an image / language manifold.

R3: Ablation study. We have conducted such ablation study in Appendix C, where DNN-SN and DNN-GP are ablated versions of SNGP. Figure 2-3 shows that in the 2D example, the uncertainty surface of a DNN-SN behaves similarly to a deterministic DNN, while that of a DNN-GP is lacking in preserving input distance. Table 4-6 shows in the vision and language experiments, DNN-SN and DNN-GP tend to outperform the deterministic baseline, but underperform SNGP.

R3: Comparison to methods designed explicitly for OOD detection. Please see table for performance comparison to popular OOD methods evaluated using area under precision recall curve (AUPRC) (we will add it to Appendix C). We denoted CIFAR-10/-100 as C10/100. As shown, despite not designed explicitly for OOD, SNGP is competitive and sometimes outperforms other OOD approaches, especially on difficult near-OOD tasks (e.g, CIFAR 10 v.s. 100 and vice versa). Mahalanobis = Mahalanobis with feature ensemble and inputprocessing.

Method / AUPRC	C10 vs SVHN	C10 vs C100	C100 vs SVHN	C100 vs C10
MSP+OE	89.4	76.2	52.9	32.6
Mahalanobis	99.1	-	98.4	-
ODIN	92.5	-	93.9	-
SNGP	99.0	90.5	92.3	80.1

R4: Unclear why the distance to the training data should be used for the uncertainty measure / whether this distance awareness property is indeed advantageous. Intuitively, given a testing example that "looks different" from the training data (i.e., far from the training data manifold), a model's uncertainty metric is expected to return a high value (see, e.g., Fig 1a). Such definition of model uncertainty (or "epistemic" uncertainty) in terms of distance/dissimilarity from observed data has been widely adopted in both the UQ and the ML literature (c.f. Kiureghian and Ditlevsen (2009).Aleatory or epistemic? Does it matter? , Kendall and Gal (2017).What Uncertainties Do We Need in BDL for CV?, *NeurIPS* and the many papers citing them), and empirically can be measured by a model's OOD accuracy (see Table 1-2, and Table 4-6 in Appendix). Quoting other reviewers: "Empirical results on benchmark datasets show superior performance in OOD." (R1), "This work conducted convincing experiments on various datasets...showed the advantage of the proposed method." (R2), etc.

R4: Why the special setting in this paper is inevitable...Either the theoretical derivations or the empirical results do not support why the practitioners have to use the proposed modification. Contrary to local methods, vanilla DNN models tend to have difficulty in achieving the distance-preservation property shown in Eqn (6). For example, vanilla DNNs are found to be vulnerable to adversarial examples - they can be sensitive to tiny perturbation in the input space, yet sometimes insensitive to semantics-altering edits to the training data - i.e., not input distance aware [33,34]. To this end, the paper's theoretical result (Proposition 1) ensures SNGP's ability in guaranteeing distance preservation, and the empirical result (Table 1-2, and Appendix C.2) shows that such modification leads to concrete improvement in ECE/OOD performance when compared to an unmodified baseline.

R4: ...disentanglement...is contrary to the explanation that the distance-preservation matters. Disentanglement and distance-preservation (i.e. invariance) are both important properties for a representational learning algorithm, and they do not contradict each other (see, e.g., Achille and Soatto (2018).Emergence of Invariance and Disentanglement in Deep Representations, *JMLR*). For a DNN hidden mapping $h : \mathcal{X} \to \mathbb{R}^d$ which is a coordinate transform from the input space to a hidden space $h(x) \in \mathbb{R}^d$, *disentanglement* describes $h(x)$'s ability in separating salient latent features from the noise among its d coordinates, while *distance preservation* describes $h(x)$'s ability in translating a semantically meaningful (often non-Euclidean) measure in the data manifold into that in the Euclidean space [30]. With suitable model specification, disentanglement can happen jointly with distance preservation (e.g., see Figure 2 of [30]). There have been many work that try to achieve both for the purpose of generalization and adversarial robustness, notably via Lipschitz regularization or invertible (i.e. bi-Lipschitz) networks [35, 69] (also, e.g., Engstrom. (2019) Adversarial Robustness as a Prior for Learned Representations)