Is the complexity of representations good or bad? Should we care about it? What are the limitations of your findings? In VAEs low-complexity aggregate posteriors are clearly bad. This is illustrated by standard β-VAEs with $\beta \geq 4$. Fig. 3 (paper) and D.3, D.4 (Appendix) show that reduced complexity translates to reduced fidelity and diversity of samples. Thus, our contribution has direct applications in development of latent variable generative models, by: 1) allowing to compare models w.r.t capacity of the latent space and independence of dimensions, 2) improving sample fidelity by fitting the aggregate posterior density. Regarding CNNs, we show that standard networks converge to a different set of solutions than memorizing nets. This adds to the evidence that neural nets exploit patterns in data and speak to the debate around Zhang et al. [ICLR 2017] work. Possible immediate applications of complexity analysis can be in interpretability research. An overview by Gilpin et al. [DSAA 2018, pp. 80–89] cites a number of works that attempt explanation by capturing semantics of network units. We uncover cases (e.g. dropout nets) where converged representations are sensitive to network initialization, making usefulness of such explanations questionable in these settings. The main limitation of our results is that we cannot claim that complexity analysis explains performance of CNNs. That said, performance-oriented engineering of deep nets vastly exceeded understanding of the proposed algorithms. Questions as basic as "are converged solutions similar?" are being answered only recently. In this context our research does contribute to the knowledge of what's happening in deep nets, even if it's not an outright explanation of generalization in deep learning.

Why DP-GMM? Why not GMM with a fixed number of components? Main limitation of DP-GMM. DP-GMM posterior is consistent in total variation for distributions that are in the KL support of the prior. Under smoothness assumption for the approximated density, DP-GMM yields near minimax contraction rate. See e.g. [Ghosal & van der Vaart, 2017, "Fundamentals of Nonparametric Bayesian Inference", sections 7.2 and 9.4] for details. Due to this flexibility DP-GMMs are fairly conservative choice for density estimation. A mixture with fixed number of components requires guessing the "correct" number of components. There are heuristics for this, but a more principled approach would be to use a prior on a finite number of components, i.e. a mixture of finite mixtures (MFM) model [Miller & Harrison, J. Am. Stat. Assoc, 113(2018), pp. 340-356]. While there could be some merit in doing so, these models are more restrictive than DP-GMMs (see below). The main limitation of Gaussian mixtures is computational cost in very high-dimensional spaces – this must be taken into account when constructing neural representations.

Number of components as a measure of complexity. Number of components can be seen as a measure of complexity – sample complexity of learning a Gaussian mixture is linear in the number of components [Ashtiani et al., NeurIPS 2018, pp. 3416–3425]. But there are caveats. DP is a prior on infinite mixtures and will not concentrate on a finite number of components in the infinite data limit [Miller & Harrison, JMLR, 15(2014), pp. 3333-3370]. We can get consistency for the number of components with an MFM model. A more fundamental issue is that finite mixture models (including GMM with a fixed number of components) make sense only if the true data generating distribution is a finite mixture. To reason about the number of components we also need to know component distributions. Unless one already has a good understanding of the data generating process, these are fairly strong assumptions. The appeal of density analysis via infinite (in the limit) mixtures is that we can avoid making such assumptions.

Why is the KL with the Gaussian a good complexity measure? Why not MMD or Student's t-distribution? Basically, we choose reference distribution following maximum entropy principle. We pick distribution that encodes mean and variance of the data, but otherwise minimizes additional assumptions. Under maximum entropy principle this will be a Gaussian. It may seem that a reasonable alternative could be a maximum entropy distribution that exactly fits the support of the data (uniform distribution). However, the reference distribution and the posterior predictive would then have different supports, leading to problems with divergences (its unreasonable to restrict the posterior to the support of known data). T-distribution has less obvious justification – we could use it to measure divergence between prior and posterior predictive in DP-GMM. Reviewer #3 points out that the base distribution may coincide with the prior in VAEs with strong regularization. We actually leverage this to claim that under strong regularization posteriors in β-VAEs collapse to the prior (we will make this more explicit in the text). These results can also be seen as a sanity check for our model (see also third paragraph in Appendix D). KL divergence has intuitive interpretations in information theory. Further, it is much more common to reason about e.g. total correlation than distances between kernel mean embeddings.

Why is posterior predictive a mixture of t-distributions? Eq. 10 and total correlation. To be precise, it's the posterior predictive given component assignments that is a mixture of t-distributions – detailed derivation is in Appendix B. Eq. 10 is an estimate of the total correlation between dimensions in the posterior predictive – we will make that explicit in the text.

Lack of convergence to the prior in MMD-VAE. Is the MMD-VAE supposed to be the better model? In addition to IMQ kernel (used in the original MMD-VAE paper) we also experimented with an RBF kernel and did not observe substantial differences in results. Evaluating other kernels is quite interesting, but may fall outside the scope of this paper. Zhao et al. [AAAI-19, pp. 5885–5892] reports higher likelihoods for MMD-VAE.

Robustness of results w.r.t. changing the random labels. We use two datasets in CNN experiments. They have different labels and were permuted independently. Results for both datasets are compatible. Also, in initial experiments we did not fix random seeds and did not observe different outcomes due to specific label permutations.