Thank you for your detailed reviews and comments. We hope our clarifications, which we will include in the final version of the paper, will strengthen your confidence in the novelty and significance of our results. We begin by addressing two crucial concerns that were raised.

Novelty. We believe that the following results are significantly novel:

(a) Theory: we prove generative models can recover the sub- and supermodularity of target distributions. These are fundamental properties for combinatorial optimization, and as such this result is an important step for the theoretical analysis of generative models over discrete spaces.

(b) Algorithms: DPPNet sampling is the only DPP sampling algorithm to generalize to new data without requiring updates to pre-computed information. Even very recent work [2, 3] (published after NeurIPS) requires pre-processing that relies on the immutability of the DPP kernel. In comparison, DPPNet can draw samples from new kernels as long as the feature representations of the new items are drawn from the same distribution as the training data. This significantly increases the scope of application for DPPs.

(c) Experiments: we show that current neural architectures are, under the right training conditions, able to represent DPPs to a degree of precision sufficient to replace DPPs in downstream applications.

Comparison to MAP. We will update our work to include NLL results for the MAP approximations for standard DPPs; we do not expect DPPNet to outperform the DPP mode. Although DPPNet mode has the same complexity as DPPNet sampling, the same does not hold for standard sampling (in particular, [4] and [1] grow as $\mathcal{O}(N^3)$ and hence will be slower than MCMC sampling [5] for which we provide a timing comparison). Since standard DPP sampling costs $\mathcal{O}(N^3)$, our timing results for standard sampling on the Nystrom experiments provide a lower bound on how much acceleration we can expect over previous MAP inference algorithms.

Reviewer 4. Thank you for your review and your comment about MAP algorithms. We hope our above clarification answers your question; we will update our paper to clarify this important point.

Objective function. We will write out the objective function explicitly and clarify the NLL notation.

Fast sampling related work ([2, 3]). These works (made available online after the NeurIPS submission deadline) are indeed highly relevant. Both speed up DPP sampling given a polynomial time pre-processing step. However, this pre-processing needs to be re-applied when the ground set is changed unless the change belongs to a specific family of transformations [3]. This is not the case of DPPNet. For this reason, [2, 3] are complementary to DPPNet; DPPNet will be more efficient when the true DPP changes overtime, but [2, 3] should be preferred for fixed kernels. We will gladly update our work to include this discussion.

To be more precise, the Nystrom approximation of [2] has to be computed every time the ground set changes. If the ground set changes frequently, this is prohibitively expensive as soon as $k \geq 5$, costing $\mathcal{O}(Nk^6 \log^2 \frac{N}{\delta} + k^9 \log^3 \frac{N}{\delta})$ [2, Thm 1 for DPPs, page 9]. The tree construction [3] can be pre-processed only if samples are drawn from DPPs whose kernels are of the form $L = B^\top WWB$ with fixed B and varying diagonal W. In comparison, as long as the new features are drawn from the same distribution as the training data, we show experimentally that DPPNet generates high quality samples without requiring re-training or additional pre-processing.

Reviewer 5. Thank you for the kind review — as recommended, we focus on concerns raised by other reviewers.

Applications. Applications of DPPs to problems in ML have been limited by the poly(N) cost of sampling when the ground set varies often (*e.g.*, certain recommender systems settings). DPPNet is a viable method to address such obstacles, and applying DPPNet to such problems is planned future work.

Reviewer 6. Thank you for the detailed comments; we have summarized the key novel contributions of our work at the beginning of the rebuttal; we will be certain to emphasize these in the final version of our paper.

Line 151. We will clarify this. We mean that if the kernel for training data has an expensive computational cost (*e.g.*, needs to be learned from data), this expensive computational cost will only be required during training, and not when generalizing to new or updated datasets.

Training objective/algorithm. We will clarify this. *DPP literature review in appendix.* We will add this.

[1] L. Chen, G. Zhang, and E. Zhou. Fast greedy map inference for DPP to improve recommendation diversity. In *NeurIPS*. 2018.

[2] M. Dereziński, D. Calandriello, and M. Valko. Exact sampling of DPPs with sublinear time preprocessing, 2019.

[3] J. Gillenwater, A. Kulesza, Z. Mariet, and S. Vassilvtiskii. A tree-based method for fast repeated sampling of DPPs. In *ICML*, 2019.

[4] I. Han, P. Kambadur, K. Park, and J. Shin. Faster greedy MAP inference for determinantal point processes. In *ICML*, 2017.

[5] C. Li, S. Jegelka, and S. Sra. Fast DPP sampling for Nystrom with application to kernel methods. In *ICML*, 2016.