Thank you very much for all for the thorough and thoughtful reviews. We want to address some of the concerns about our work and hopefully clear up any potential confusion as well.

Reviewer 2 –

Comparison to other algorithms – One of the major complaints reviewer 2 had about our paper was the lack of comparison to other localization algorithms beyond center of mass (COM). To clarify, the task we were aiming to perform was *unsupervised*, *large-scale*, *pre-sorting* localization. Although we did cite other localization algorithms, these preexisting methods require spike sorting to be done before localization (to get average waveforms) and are also computationally expensive. For this reason, we wanted to compare to the only other localization method that is unsupervised, used consistently pre-sorting, and can scale to arbitrarily large datasets: COM. Our opinion is shared with reviewer 3 who also mentioned in his/her review that other localization algorithms would not scale to the large-scale datasets we were targeting. Hopefully this clarifies why we did not benchmark against other algorithms and alleviates reviewer 2's concern (we can include this reasoning about our benchmarks more clearly in the camera-ready version).

Non-Gaussian amplitudes – Reviewer 2 also mentioned he/she was concerned with our amplitudes being modelled as Gaussian random variables as spikes are non-Gaussian events. To clarify our modelling assumption, we chose a Gaussian amplitude model for convenience when working with a VAE and because the peak amplitude variations in the observed data seemed well-approximated by a Gaussian. It is a fair point that spikes are non-Gaussian, however, we were trying to model the peak amplitude variations rather than the actual spikes.

Reviewer 1-

Scalability – Reviewer 1 mentioned that our algorithm may not scale up to larger probes (with 10000+ channels) since we use all the channels as inputs to our model. However, our method does not actually use all the channels on the MEA. The observed data for our model is from a *subset* of channels that is centered on the channel with the largest negative amplitude. We construct this subset by taking a small radius (a hyperparameter that is typically set to 40-50 microns) of channels around the central channel. This subset consists of anywhere from 4 to 25 channels for the radii we chose in the paper. That means that our method would scale well to 10000+ channel MEAs.

Data Augmentation – Reviewer 1 mentioned that the data augmentation we implement seems unnecessary given our model. To clarify, the data augmentation only affects spikes detected near the edge of the array (spikes detected near the center of the array are unaffected by the augmentation) and provides two important benefits. First, since the prior location for each spike is at the center of the subset of channels used for the observed data, for spikes detected near the edge of the array, *the data augmentation actually puts the prior closer to the edge* and is, therefore, much more informative for localizing spikes near/off the edge of the array. Second, since spikes detected near the edge of the array typically have smaller amplitudes and are seen on fewer channels, our augmentation reduces the number of noisy channels with little signal in these cases. We experienced a performance increase with this data augmentation and would be happy to include an empirical analysis of this increase, if wanted, in the appendix of the camera-ready version.

Spike Sorting – Reviewer 1 is correct in saying that our "spike sorting" method of using a GMM and searching over the number of clusters would not be useable for real datasets where the number of clusters is unknown. However, since our work does not include the development of a new detection or classification algorithm (we solely focus on localization and location features) we just wanted to show that *our location features were more discriminable than those of COM*. This is why we show, for any number of mixtures in our GMMs, that our location features upper bound those of COM in both the true positive rate and accuracy of the clusterings. Therefore, any algorithm that uses centroid-based localization as features for their classification (Herding Spikes - Hilgen 2017, IronClust/JrClust - Jun 2017), could improve their accuracy by using our localization method instead. We actually use HerdingSpikes2 paired with our new localization method in our analysis of the real datasets to show how our method can be easily integrated into preexisting spike sorting pipelines. Since we did not develop new detection or clustering methods, we decided a GMM was the best and most direct way to show the advantages of improved localization, rather than comparing an integration into a full pipeline. For our claim that combining locations and waveforms was introduced in 2017, we were referring to how they were explicitly combined for clustering; reviewer 2 was correct about the implicit combination existing earlier.

Reviewer 3-

Data Augmentation – Reviewer 3 was interested in seeing the increase in performance over COM without the augmentation. As mentioned before, our augmentation only affects spikes detected near the edge and, despite that, we still see strong improvement over COM in the center of the MEA where the augmentation has no effect.

Writing - We can address the issues with experimental values and model discrepancies in the camera-ready version.

Thank you again to all the reviewers for their thoughtful comments and feedback. We believe our work provides a solid contribution to the electrophysiological community and can be both a tool and a benchmark for future localization work.