In my experience, simulated data is kinda useless for training, especially when the data is so noisy. Fun blue sky project but I strongly doubt we can build a convincing generative model for the kind of data where a spearman of 3-4% is considered good, clearly our data is barely understood by our models