**Garbage in, Einstein out**

A Mathematical Study of Einstein from Noise

I-Ping Tu

Institute of Statistical Science, Academia Sinica

A cryo-EM 3D structure is solved from many noisy 2D projections of individual molecules. Two keys that make this 3D reconstruction a challenging computational task is its high level of noise and the unknown pose parameters of each individual molecule. Often times, reference is used to initiate the search of orientation, which has incurred the risk of coalescing images with low or no signal to the reference, known as the ‘Einstein from noise’ problem. Here, we investigate this phenomenon from model-bias viewpoint in terms of image dimensionality and sample size. By using mathematical modeling, we derive a surprisingly simple form accurately predicting the correlation value between Einstein face and the spurious image arising from averaging the sorted top images of purely Gaussian noise images. This theoretical value increases with n (the number of images) and m (the number of images sorted for averaging) but decreased with p (the dimensionality of image). To avoid ‘Einstein from noise’ pitfall, we propose a denoising method as a data pre-processing tool to increase the SNR. We observe that this tool makes significant improvement in either computation time or clustering average quality in 2D clustering of various cryo-EM analysis packages.