Supplementary Figures

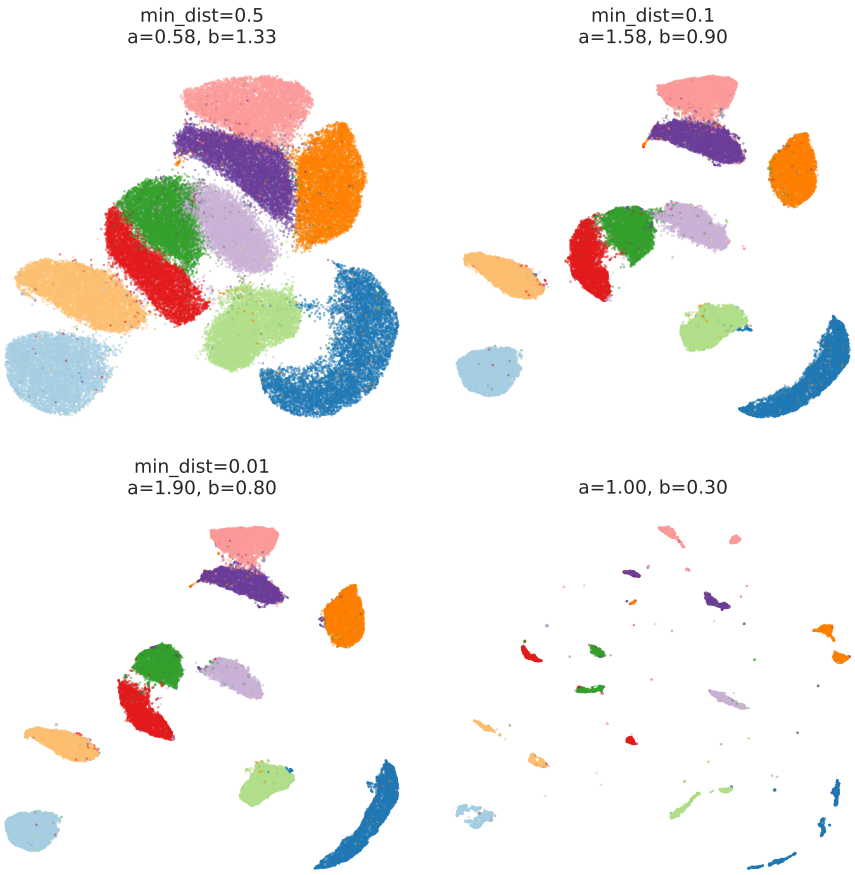

Fig. S1. UMAP visualisations of the MNIST data set with default parameters and `min_dist` set to 0.5, 0.1 (default), 0.01, and (a, b) manually set to $(1, 0.3)$. Each subplot shows the a and b values of the UMAP kernel $k(d) = (1 + ad^{2b})^{-1}$. With `min_dist` below 0.01 the b as a function of `min_dist` hardly changes.

As a side note, one can obtain an MNIST embedding very similar to the one that UMAP gives with default settings using t-SNE with late exaggeration of about 4, i.e. multiplying all attractive forces by 4 after the early exaggeration period (first 250 iterations) is over. This makes sense, given that the main difference between t-SNE and UMAP is in the repulsive term.

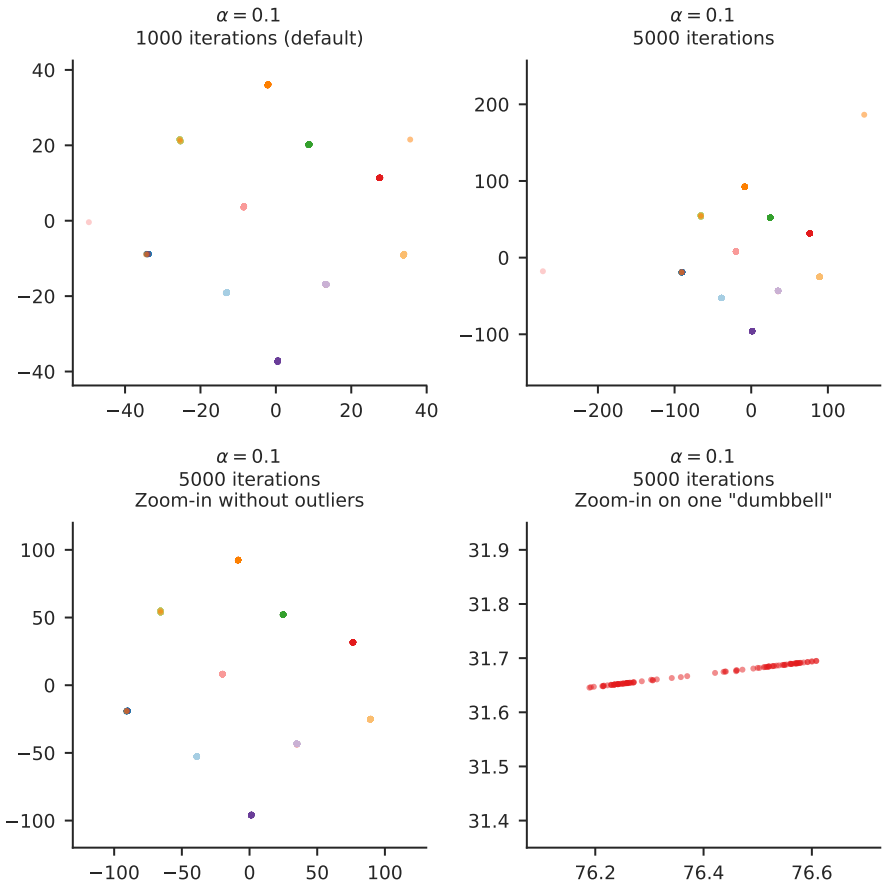

Fig. S2. Toy example with ten "dumbbell"-shaped clusters from Figure 2, here embedded with $\alpha = 0.1$. Top-left plot shows the result after 1000 gradient descent iterations (default). Note that the dumbbell shape is lost: whereas the number of visible clusters increased as α was lowered from 100 to 0.5 (Figure 2), it decreased when it was further lowered to 0.1. We believe the reason for this is that the strong repulsion between dumbbells "squashes" them in the beginning of optimisation into very compact blobs. It is likely that longer optimisation would resolve the dumbbell shapes. This is difficult to test because the kernel with $\alpha = 0.1$ is extremely wide and flat, leading to slow convergence. Top-right plot shows the result after 5000 iterations. Here a few outlying points get pushed to the periphery. Zooming-in to the main 10 clusters (bottom-left) still does not resolve the dumbbell shapes. Further zooming in on one of the dumbbells (bottom-right) shows that the points are squashed into 1D which may be a sign of poor convergence.

In a separate sets of experiments, we observed the similar phenomenon with MNIST: $\alpha = 0.2$ after 1000 iterations yielded fewer clusters than $\alpha = 0.5$. Our conclusion is that smaller values of α should be used with caution.

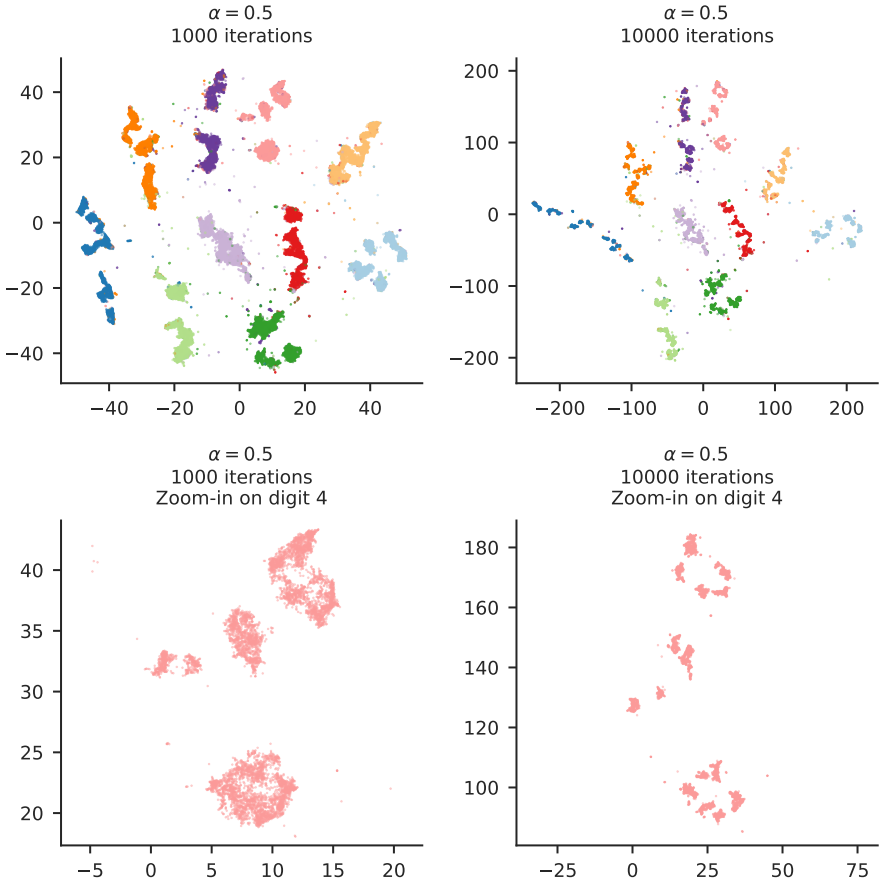

Fig. S3. t-SNE visualisation of the MNIST dataset with $\alpha = 0.5$. The top-left panel is identical to Figure 4C; it was obtained with 1000 gradient descent iterations (the default value). The top-right panel corresponds to 10 000 iterations and has many more isolated sub-clusters. This can also be seen in the bottom row showing the respective zoom-ins into the digit "4". At the same time, the embedding after 1000 iterations is not misleading and is simply a coarser-grained version of the embedding after 10 000 iterations.

Using 10 000 iterations is impractical: whereas 1000 iterations were finished in 1.5 minutes, 10 000 iterations took 4 hours 30 minutes. This is because FIt-SNE interpolation scheme uses regular interpolation grid with the number of nodes growing quadratically with the embedding size. While the left embedding is contained within $[-50, 50]^2$, the right one expands to $[-200, 200]^2$. In principle, an implementation based on the fast multipole method (FMM) could be developed to dramatically accelerate the gradient computation in this setting where most of the embedding space is "empty space", but current FIt-SNE implementation does not support it.

Note that the standard t-SNE embedding with $\alpha = 1$ also expands much further after 10 000 iterations, compared to the 1000 iterations. However, with $\alpha = 1$ it does not resolve additional sub-clusters, at least in MNIST.

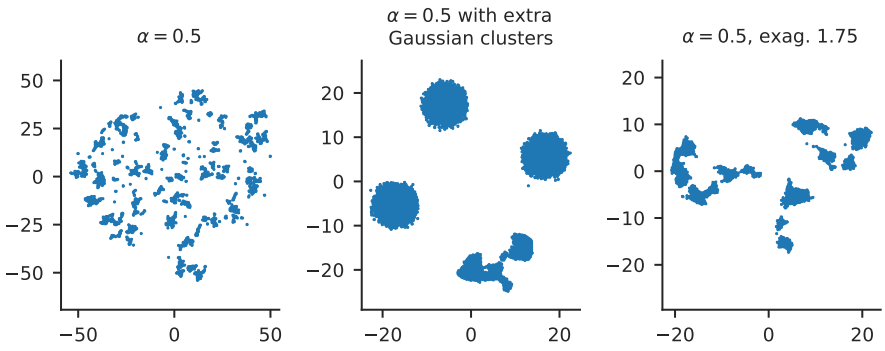

Fig. S4. t-SNE visualisation of a MNIST subset consisting of all images of digit "4" ($n = 6\,824$) (perplexity 50). Left: $\alpha = 0.5$, the same as in Figure 4C. Note the large number of isolated clusters. We believe this happens because the embedding rapidly expands to a larger area, compared to Figure S3 (bottom-left). One evidence for that is that re-running t-SNE after adding several random Gaussian clusters with $n = 7\,000$ each, roughly recovers the shape of the digit "4" archipelago from the full MNIST embedding (middle). Right: $\alpha = 0.5$ and exaggeration factor 1.5 [7], i.e. all attractive forces are multiplied by 1.5 after the end of the early exaggeration phase (during the early exaggeration they are multiplied by 12). This roughly recovers the sub-clusters from the full MNIST embedding (Figure S3). The relationship between α and exaggeration remains for future work.