I am treating this as practice for unsupervised workflow, i.e., I drop the labels until the very end and trying to find conclusions without using these labels.

Section A:

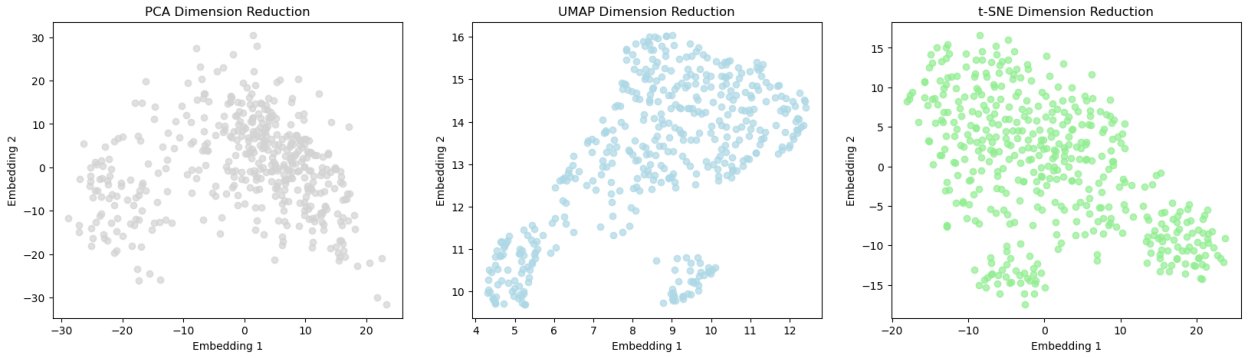
Using my analysis of this dataset and results on dimension reduction (here), the dimension reduction techniques can be decided between these UMAP and t-SNE (PCA included as baseline to demonstrate that nonlinear structure exists in data).

Note - Initial experiments used standardized (z-scored) gene expression features prior to dimensionality reduction and clustering. However, these transformations removed biologically meaningful variance, resulting in degraded cluster separation. Subsequent analysis on the unscaled data produced much clearer boundaries indicating that raw expression magnitudes carry essential discriminative signal. All results therefore use the unscaled feature space for dimensional reduction and clustering.

Grid-Search results (from previous work)

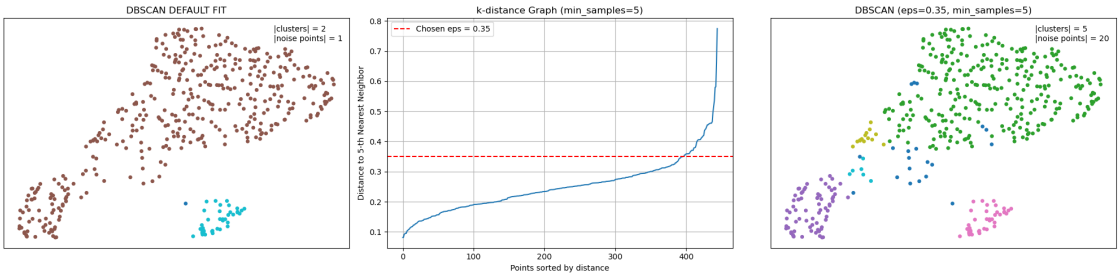
```
Tuning UMAP: 100%|██████████| 60/60 [00:35<00:00,  1.68combo/s]

Grid search completed in 35.74 seconds

Best UMAP configuration:
n_neighbors          10
min_dist             0.1
metric          euclidean
jaccard          0.224755
```

```
Tuning t-SNE: 100%|██████████| 30/30 [00:36<00:00,  1.23s/combo]

Grid search completed in 36.88 seconds

Best t-SNE configuration:
perplexity              30
learning_rate           50
metric           euclidean
jaccard           0.253707
```



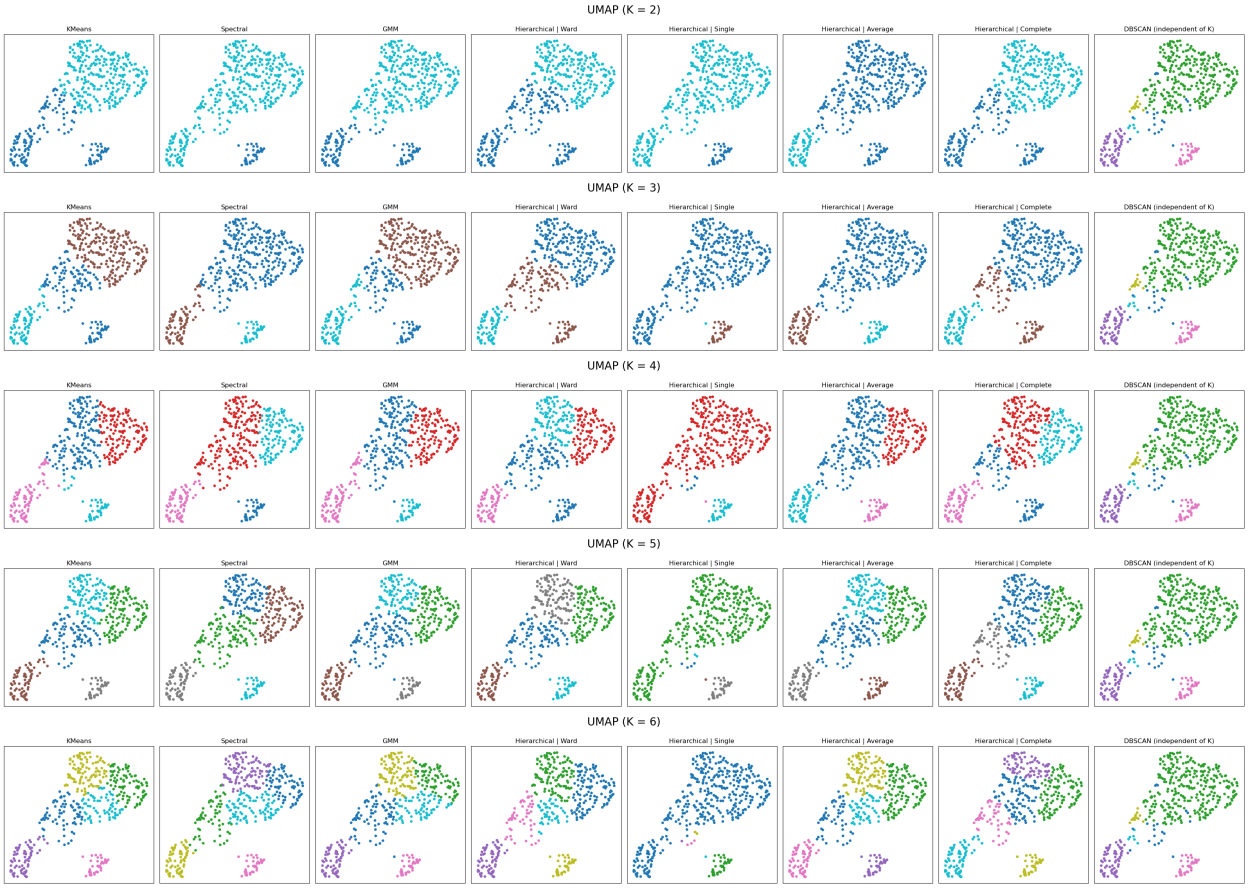
Ultimately – from my previous analysis UMAP provides better separation and preserves a balance of local and global structure more effectively than t-SNE so I will use UMAP for all downstream clustering. This decision was rooted in previous quantitative analysis.

Now fitting the clustering methods with default settings or slightly tuned through methods like elbow test – approximations I the EDA phase before exploring more quantitative unsupervised hyperparameter tuning. This is necessary because some clustering algorithms yield degenerate results – e.g. methods like DBSCAN which are extremely sensitive to hyperparameters (see DBSCAN example below).



As you can see the default fit returns 2 clusters which is not extremely informative – after applying the elbow test we start to see further seperation as the clustering appears to do a significantly better job at partitioning the structure in the data.

Now to fit all clustering methods across a few selections of K to see how the partitions drastically differ as not all these methods are nested – so they lack some favorable properties which make selection of K crucial. See plots below.



The progression from $K = 2$ to $K = 6$ demonstrates increasingly granular partitions of the BRCA samples. At $K = 2$, most algorithms identify a coarse separation likely corresponding to a biologically meaningful binary subtype (e.g., basal-like vs non-basal). As K increases, K-Means and GMM yield relatively balanced subdivisions, whereas hierarchical variants (especially single linkage) produce unstable or chained structures. DBSCAN remains agnostic to K and reveals density-defined groupings consistent across runs, though highly sensitive to ε. These observations motivate a quantitative evaluation in the next section to objectively select an optimal K based on internal validation metrics. In the next section we perform hyperparameter tuning by applying quantitative measures such as sillhouette scores, stability and generalizibility in order to further tune these models and yield a defendable consistent workflow – informing the selection of the "best" K.

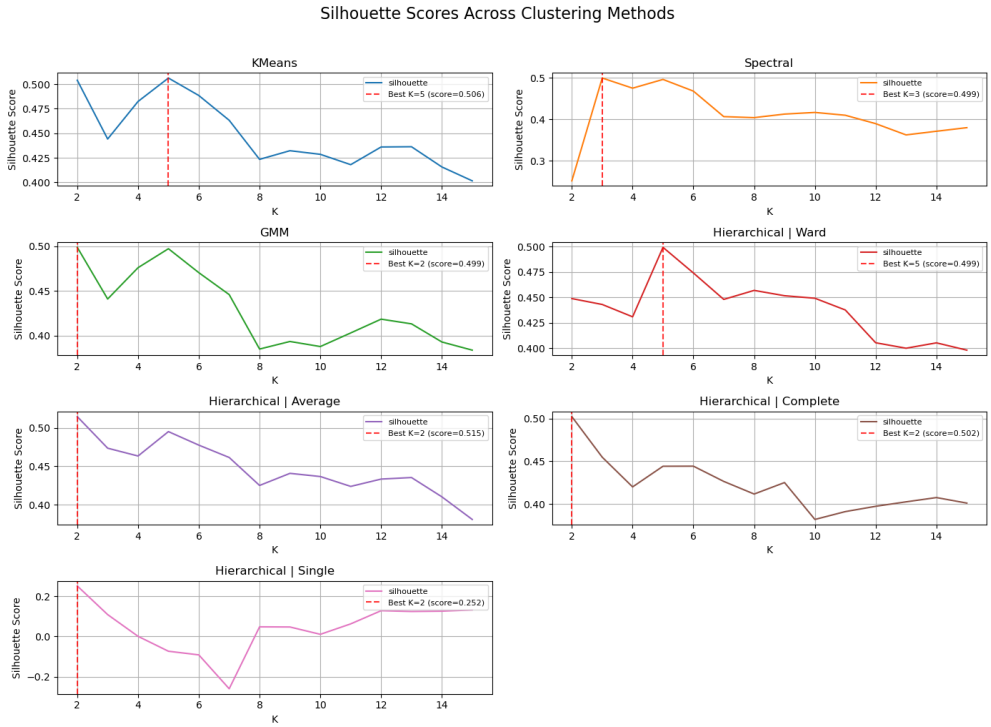
Section B:

In this section the main goal is to tune for the "best" K - other hyperparameters are considered but the focus is on determining K.

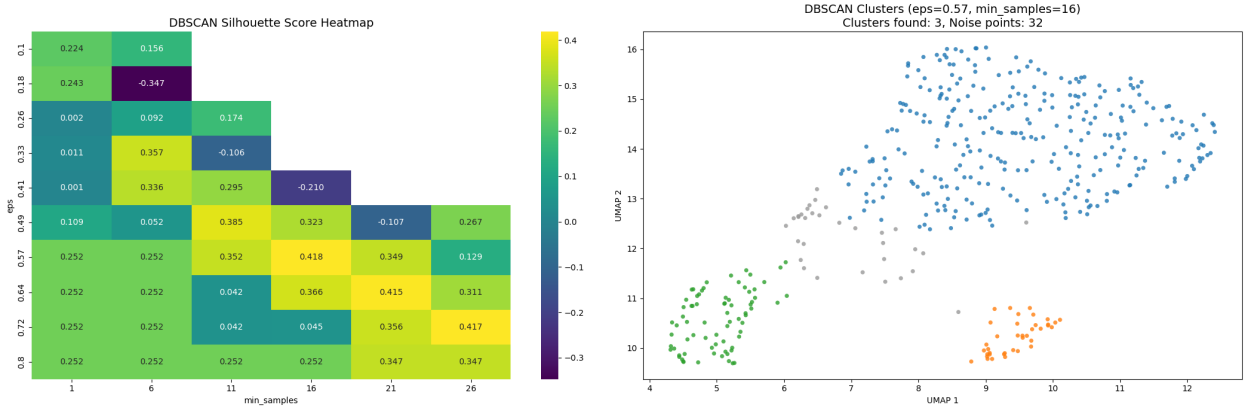
Silhouette Score

The silhouette score can be calculated for each data point i as $s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}}$ where b_i is the average distance between point i and all points in the nearest other cluster (inter-cluster seperation) and a_i is the average distance between point i and all other points in the same cluster (intra-cluster cohesion). The overall silhouette score is the average of s_i across all points $i \in \{\mathbb{Z} : [1, n]\}$ given by $S = \frac{1}{n}\sum_i s_i$.

For methods that take a hyperparameter K clusters, I have plotted the silhouette score across clusters (below).



Now calculating the stability score across values of K is not viable for methods like DBSCAN that does not take this as a hyperparameter so instead we seek $K^* = \underset{\theta}{\text{argmax}}\, S_\theta$ where $\theta = (\varepsilon, \text{min\_samples})$ – i.e., take the K^* number of clusters that occur when the silhouette score is maximized across a grid of parameters.



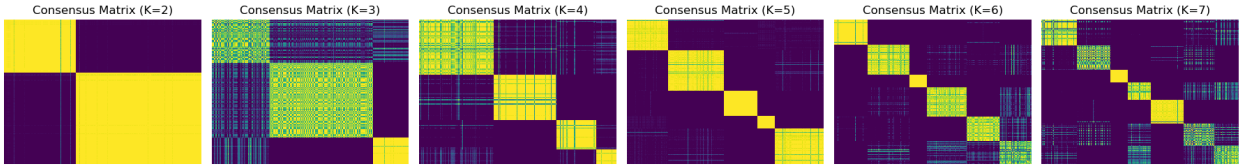
The results indicate that the silhouette score S_θ is maximized at 0.418 for hyperparameter $\theta = (0.57, 16)$. Visuallizing the results on the right plot indicate that DBSCAN finds $K = 3$ clusters – hence we see that DBSCAN agrees with Spectrals $K = 3$ decision.

The silhouette scores do not align perfectly across methods but there is definitely strong agreement between the values $K \in \{2, 3, 5\}$ - this makes sense intuitively upon inspecting the UMAP data projected into a 2d embeddings across different values of K.

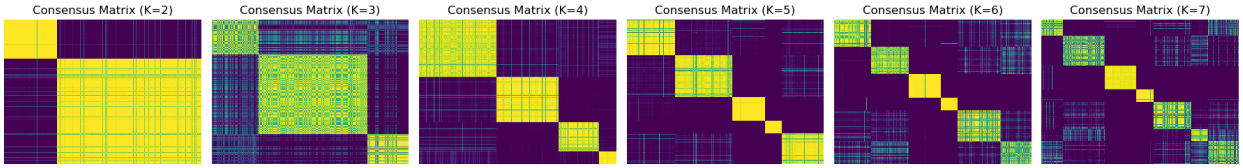
Stability

To assess clustering robustness across hyperparameters, we compute a consensus matrix using bootstrap resampling. In these heatmaps, a strong yellow diagonal pattern of K squares indicates high assignment stability - yellow cells represent samples that are consistently co-clustered across resamples. Across K-Means, Gaussian Mixture Models (GMM), and Spectral Clustering, we observe clear, well-defined block structures at lower values of K (although Kmeans and GMM are unstable at $K = 3$) suggesting these configurations yield the most reproducible partitions. The patterns become increasingly fragmented for larger K, i.e., less stable.

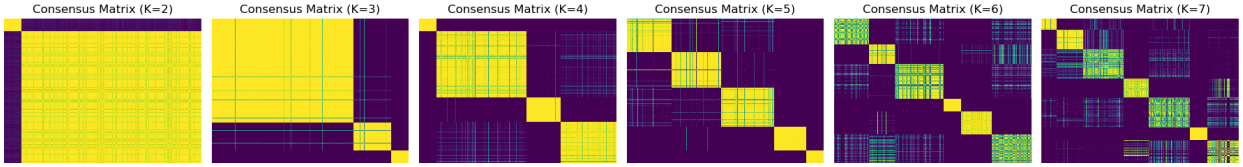
K-Means

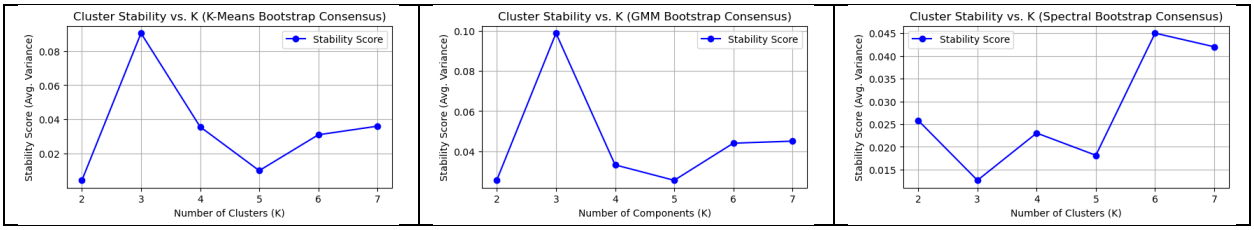


Gaussian Mixture Models (GMM)



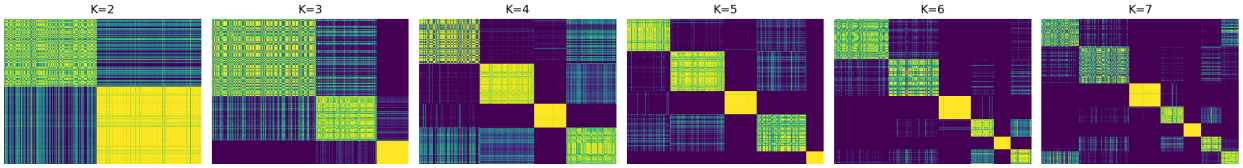
Spectral Clustering



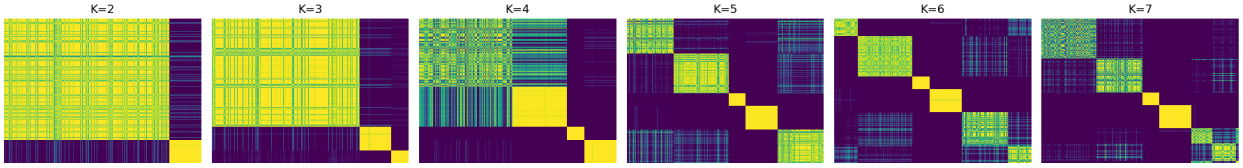


Now applying this analysis to hierarchical clustering, I was initially surprised by the results, which show lower stability compared to methods like K-Means and GMM. This outcome seems counterintuitive, as hierarchical clustering produces a nested family of clusterings that would intuitively suggest greater robustness. However, in practice, the greedy nature of the linkage criterion makes the method highly sensitive to small data perturbations, reducing its stability under bootstrap resampling.

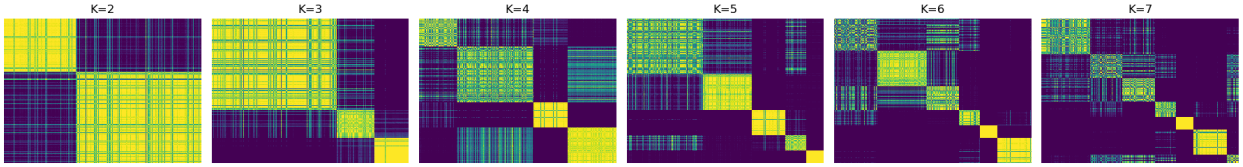
Agglomerative Clustering (linkage="ward")

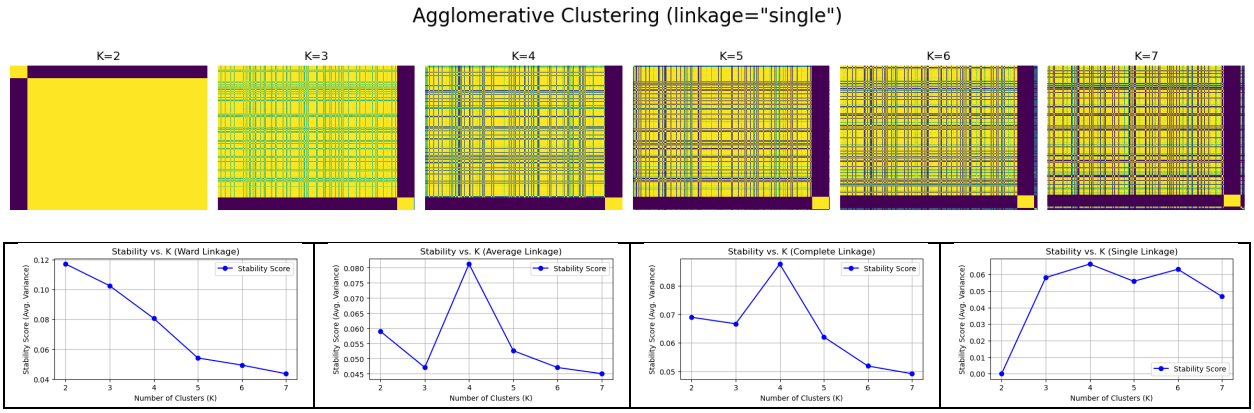


Agglomerative Clustering (linkage="average")



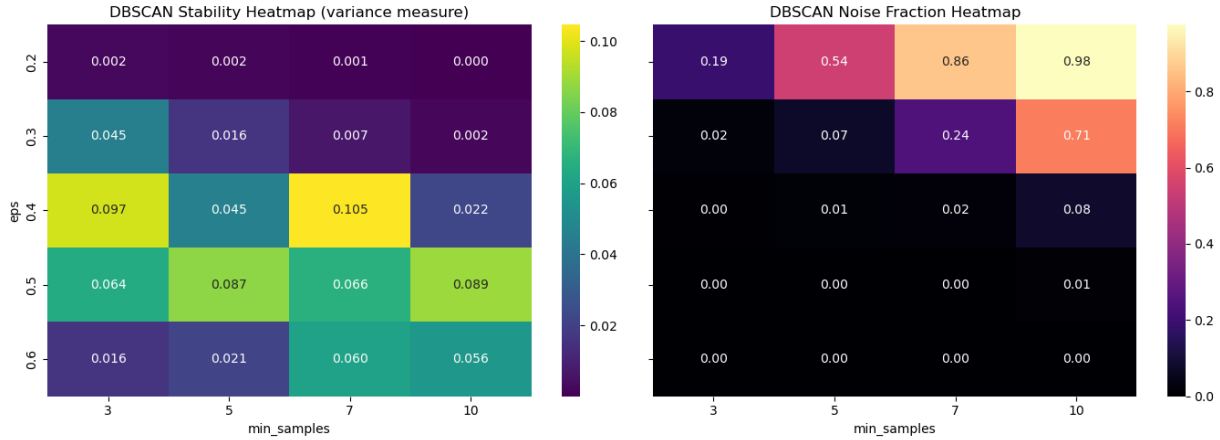
Agglomerative Clustering (linkage="complete")



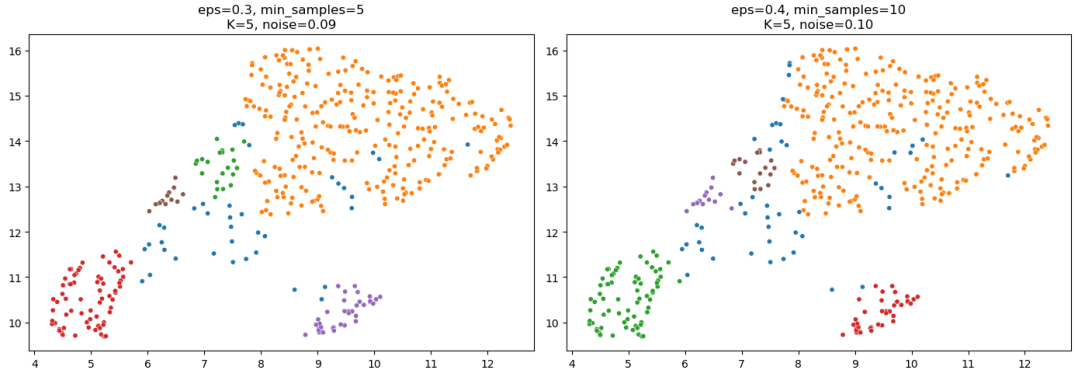


Note that linkages such as single linkage appear artificially stable because of their chaining behavior, which leads to a degenerate and nearly deterministic assignment—consistently linking points into one dominant cluster. Generally, there appears to be higher stability across linkages at higher instances of $K \in \{5,6,7\}$.

Lastly, to assess the stability of density-based methods such as DBSCAN, instead of using a consensus matrix of co-assignment across observations, I computed a heatmap of stability using the variance measure obtained from bootstrap resampling. Lower values on the left heatmap indicate higher stability (i.e., lower variance across resamples). The accompanying heatmap on the right shows the fraction of points identified as noise for each hyperparameter combination. High stability regions often coincide with high noise fractions, suggesting that these "stable" configurations are in fact degenerate — the model becomes overly conservative and labels most points as noise, leading to trivially stable but uninformative solutions. Therefore, interpreting DBSCAN stability requires balancing true clustering robustness against degeneracy due to excessive noise assignment.

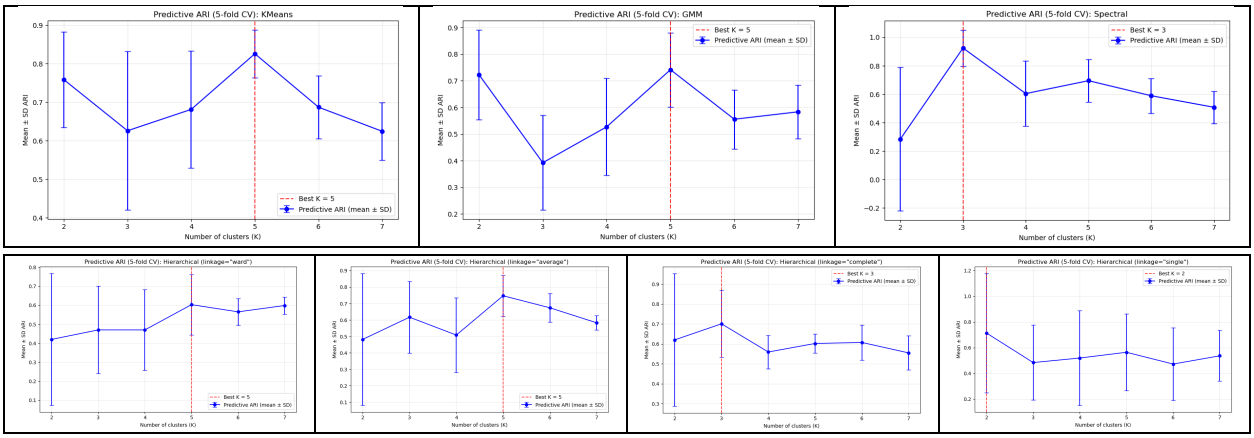


Selecting the most parameter combinations that have the minimum variance conditioned on the noise fraction being less than some threshold and plotting the results in 2d space indicates that DBSCAN yields higher stability at $K = 5$. See the visual below.



Generalizability

To evaluate generalizability, a cross-validation framework is applied where each fold is split into training and testing subsets. On each split, the selected clustering algorithm is fit independently on both subsets to obtain separate cluster assignments. A Random Forest classifier is trained on the training data using its cluster labels to learn the corresponding decision boundaries, which are then used to predict cluster membership on the test data. The Adjusted Rand Index (ARI) is computed between the predicted and test cluster labels to quantify generalizability, as ARI is invariant to label permutations and therefore correctly measures the agreement between independent clustering assignments. The mean results of this algorithm with error bars are plotted below.



The generalizability is optimal across methods for $K \in \{3,5\}$ except for Hierarchical Clustering using single linkage – which we can safely ignore as this method yields poor results as the chaining property is not desirable for this dataset (UMAP embeddings are imbalanced elliptical like clusters).

This technique cannot be applied to DBSCAN because upon split the number of clusters can change so while this algorithm for generalizability is invariant to permutation it requires the number of clusters to align. Methods to assess generalizability agnostic to K for DBSCAN are scoring how often each point remains a core across bootstraps (and with which cluster). Then aggregate as the average core-membership probability for each reference cluster; high core consistency implies strong generalizability.

Generalizability across methods strongly favors $K = 5$ and is consistent with $K \in \{2,3,5\}$.

Results

Values of K across various methods – top K (ordered)				Hierarchical				
Method	K-Means	GMM	Spectral	Ward	Average	Complete	Single	DBSCAN
Silhouette Score	{2,5}	{2,5}	{3,5}	{5}	{2,5}	{2}	{2}	{3}
Stability	{2,5}	{2,4,5}	{2,3,4,5}	–	–	–	–	{5}
Generalizibility	{2,5}	{2,5}	{3,5}	{5}	{5}	{3}	{2}	–

The stability results are unreliable across hierarchical clustering for all linkages (referenced in that section) due to greedy algorithm that is not robust when bootstrapping. The results are indeed relatively consistent across methods and validation techniques with clear themes of $K \in \{2,3,5\}$.

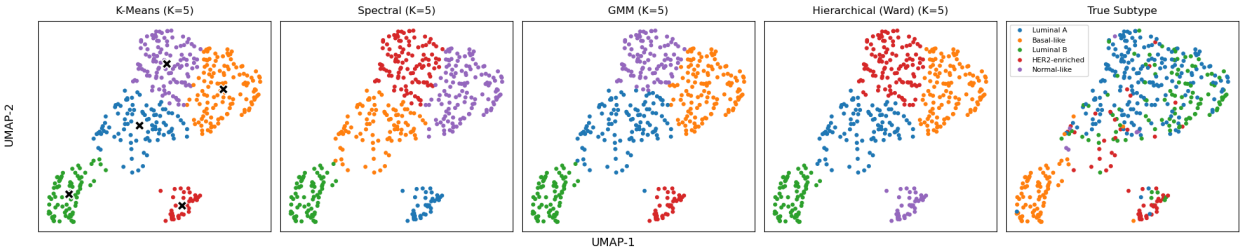
Across all methods, K-Means yields the most consistent results finding $K \in \{2,5\}$ as relevant. It produced stable clustering assignments – suggesting convergence to reliable local optima – along with strong silhouette scores, stability and generalizability relative to the other methods. Moreover, its computational efficiency with respect to the other more complex models further reinforces that K-Means is the most effective and practical method for this dataset. Lastly, when one considers a more rigorous workflow and tuning across all hyperparameters and dimension reductions and their combinations it is ideal to have a simplistic, faster fitting model.

Overall, the results across clustering methods suggest the presence of meaningful underlying structure in this dataset, as each approach identified consistent grouping patterns. With domain expertise, these clusters can likely be mapped to biologically interpretable distinctions — for example, the $K = 2$ configuration may correspond to a binary characteristic in breast cancer tumor profiles, such as hormone receptor activation status or another key molecular feature. Since the true labels are available, the next section will explore these discovered patterns in greater depth to evaluate their biological relevance.

***Section C*:**

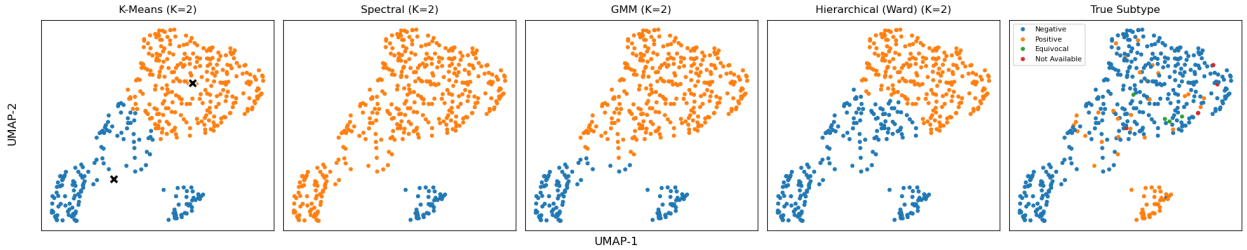
In this section, we use the clinical metadata – or labels – to identify if some of the underlying structure we found in the data can be explained by this known information. Below is a plot of the UMAP embeddings in 2d colored with the clinical metadata.

Subtype: There are 5 different subtypes present in the tumor samples [Luminal A, Luminal B , Basal-like, HER2-enriched, Normal-like]. These clustering algorithms all yield relatively similar results and strongly distinguish the Basal-like and HER2-enriched subtypes. There is some overlap between Luminal A and Luminal B and these algorithms seems to do okay but not great at distinguishing between the two. However, this could likely be the reason that $K = 5$ was such a common cluster value when performing hyperparameter tuning.



HER2-Status: For this clinical metadata there is essentially a binary decision Negative or Positive (with instances of Equivocal or Not Available being sparse). The only method that seems to capture this pattern well for $K = 2$ is Spectral Clustering – but this is mainly due to methods like K-Means seeking balanced gaussian clusters. Furthermore, the cluster on the bottom right shows strong separation (and while UMAP is an attraction repulsion method, so this distance is distorted it still preserves some global distances and clearly is dissimilar enough to distinguish its own tight cluster).

PR-Status: For this clinical metadata seems to align well with GMM where $K = 2$ was a persistent clustering (K-Means and Hierarchical also seem to do fine well). This is one potential interpretation for the structure of this pattern – where Spectral seems to not agree and may instead be capturing a pattern like HER2-Enriched Positive or Negative.



Lastly, if we go back to this tuned version of DBSCAN from before with $(\varepsilon,\text{min\_samples}) = (0.57,16)$, K ends up being 3 – we can see that one interpretation for this could simply be a distinguishing subtypes where the distinction between Luminal A and Luminal B fails due to similarity but HER2-enriched and Basal-like are dissimilar enough to be captured.

Overall, the clustering methods and unsupervised hyperparameter tuning was extremely successful at detecting that there were strong patterns for $K \in \{2,3,5\}$ where for larger K over partitioning was occurring and structure of the data was less evident which aligns with the clinical metadata.