# LITERATURE REVIEW

**The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data.**

**Abdullah *et al. ,*  [1](2022)**

**A)Problem Statement**

**The study addresses the critical problem of assessing and managing COVID-19 risk in Indonesian provinces during the early stages of the pandemic. The challenge is to determine the proximity or similarity between provinces based on the number of confirmed COVID-19 cases, recovered cases, and deaths. This assessment is pivotal for making informed decisions and formulating policies to mitigate the spread of the virus.**

**B)Method**

**The research method involved a series of steps:**

**Data collection from the Indonesian COVID-19 Acceleration Task Force website on April 19, 2020.**

**Division of the data into three categories: confirmed cases, recovered cases, and deaths, segregated by provinces.**

**Exclusion of predominant data to ensure a balanced analysis.**

**Utilization of R Software version 3.6.3, along with packages like "tidyverse," "cluster," and "factoextra."**

**Data preparation steps included structuring data with rows as observations and columns as variables, handling missing values through deletion or estimation, and standardizing data for comparability.**

Data were analyzed using the K-Means Clustering method as a technique for performing data groupings.

**C)Advantages/Disadvantages**

**Advantages of the K-Means Clustering method:**

**Efficiently identifies patterns and groups within the data.**

**Provides clear insights into provincial similarities.**

**Disadvantages:**

**Sensitive to initial cluster centroids, which may lead to different outcomes.**

**Requires prior knowledge of the desired number of clusters (k).**

**D)Performance**

**The study's primary findings revealed the formation of three distinct clusters of provinces based on COVID-19 data.**

**These clusters provide insights into the varying COVID-19 situations across the provinces, aiding policy formulation.**

**E)Conclusion**

**In conclusion, the study successfully applied the K-Means Clustering method to group Indonesian provinces based on COVID-19 data.**

**The results revealed three distinct clusters, indicating different levels of COVID-19 risk and impact.**

**These findings have significant implications for government policymaking, particularly in terms of implementing targeted restrictions and preventive measures.**

**Furthermore, the methodology employed demonstrates the potential of data-driven approaches for future predictions and effective decision-making based on provincial data.**

***References:******[1]*** *Susilo, Ahmar, Rusli, Hidayat,. The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Qual Quant 56, 1283–1291 (2022). https://doi.org/10.1007/s11135-021-01176-w*

**Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering.**

**Shutaywi, *et al. ,*  [2](2021)**

**A)Problem Statement**

**The problem of assessing the performance of clustering methods in machine learning is a critical concern in various domains. Clustering serves the purpose of grouping data points with similar characteristics, but selecting an appropriate clustering algorithm and its parameters can be challenging. The choice of clustering method often depends on the specific data and problem context. Additionally, traditional performance metrics for clustering often necessitate labelled data for evaluation, limiting their applicability in unsupervised scenarios.**

**B)Method**

**The proposed method addresses the inherent challenges in clustering performance evaluation by introducing an innovative approach that harnesses the power of the Silhouette index. The Silhouette index, a versatile metric, measures clustering quality by considering within-cluster cohesion and between-cluster separation. Importantly, it does not rely on labelled data, making it well-suited for unsupervised learning scenarios. The method employs multiple kernel functions, including Gaussian, polynomial, and hyperbolic tangent kernels, in the context of kernel k-means clustering. These kernels project data into higher-dimensional spaces, allowing for more flexible and complex cluster shapes.**

**C)Advantages/Disadvantages**

**Advantages: The method offers a solution to the longstanding challenge of evaluating clustering performance without the need for labelled data. This unsupervised approach broadens its applicability and relevance.**

**By incorporating multiple kernel functions, the method mitigates the sensitivity to kernel selection, leading to more robust and reliable clustering outcomes. Weighted clustering results based on the Silhouette index scores enable the extraction of the most informative clusters, thereby enhancing the overall quality of clustering assignments.**

**Disadvantages: While the Silhouette index represents a valuable unsupervised performance metric, it may not capture all facets of clustering quality, and its effectiveness can vary depending on dataset characteristics. The method assumes the availability of multiple kernel functions, which may not always be applicable or readily accessible for every dataset or problem.**

**D)Performance**

**The performance of the proposed method is rigorously assessed through simulations on benchmark datasets. To account for the randomness associated with the initialization of cluster centers in kernel k-means, a Monte Carlo approach is employed. This approach involves averaging results over multiple trials to provide a more reliable assessment of the method's clustering performance. The Silhouette index serves as the primary metric for evaluating clustering quality, and the proposed weighted clustering approach demonstrates its effectiveness in producing high-quality cluster assignments.**

**E)Conclusion**

**In conclusion, the proposed method represents a significant advancement in the field of performance evaluation for clustering in machine learning. By leveraging the Silhouette index and multiple kernel functions, it addresses common challenges associated with clustering evaluation, including the reliance on labelled data and the sensitivity to kernel choices. The method's unsupervised nature demonstrated performance on benchmark datasets, and robustness to random initialization of cluster centers make it a valuable tool for enhancing clustering outcomes in a diverse array of applications. Future research may explore further refinements, extensions, and real-world applications of this approach, contributing to the continued advancement of unsupervised clustering methodologies.**

***References:[2] M.; Kachouie, N.N, Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy 2021, 23, 759. https://doi.org/10.3390/e23060759***

**Research on K-Value Selection Method of K-Means Clustering Algorithm**

**Yuan, *et al. ,*  [3](2019)**

**A)Problem Statement**

**Cluster analysis is a fundamental task in data mining, facilitating the identification of patterns and structures within datasets without prior knowledge. One of the critical challenges in cluster analysis, specifically in K-means clustering, is determining the appropriate number of clusters, denoted as K. This literature review examines four widely used K-value selection algorithms: the Elbow Method, Gap Statistic, Silhouette Coefficient, and Canopy, exploring their methodologies, strengths, weaknesses, and performance in addressing this problem.**

**B)Method**

**Elbow Method Algorithm: The Elbow Method calculates the sum of squared errors (SSE) for various K values, searching for an "elbow point" where the SSE begins to level off. While its simplicity is an advantage, it struggles when the SSE versus K relationship lacks a clear elbow, making K-value selection challenging.**

**Gap Statistic Algorithm: Gap Statistic introduces reference measurements through Monte Carlo sampling to determine the optimal K value by comparing clustering results to those of a reference dataset. However, this method can be computationally expensive, particularly for large-scale datasets.**

**Silhouette Coefficient Algorithm: The Silhouette Coefficient algorithm combines cohesion and separation metrics, maximizing silhouette values across different K values. Yet, its computational complexity, especially the requirement for distance matrix calculations (O(n^2)), limits its applicability to extensive datasets.**

**Canopy Algorithm: Canopy divides datasets into overlapping subsets using predefined distance thresholds, enhancing fault tolerance and noise immunity. It excels in handling large and complex datasets.**

**C)Advantages and Disadvantages**

**Each algorithm exhibits distinct characteristics:**

**Elbow Method offers simplicity but relies on an identifiable elbow point, limiting its effectiveness when the elbow is unclear.**

**Gap Statistics provides reliable results but may become impractical for large datasets due to their computational demands.**

**Silhouette Coefficient delivers a balanced evaluation of cluster quality but struggles with the computational complexity inherent in distance matrix calculations.**

**Canopy excels in handling large and complex datasets, offering robust fault tolerance, and noise immunity.**

**D)Performance**

**The performance of these algorithms varies depending on the dataset size and complexity. For small datasets, all methods can yield acceptable results. However, as the dataset size and complexity increase, the Canopy algorithm emerges as the most suitable choice due to its efficient computational handling of extensive data.**

**E)Conclusion**

**In conclusion, selecting the appropriate K value in K-means clustering is a critical challenge with several viable solutions. Each of the four algorithms reviewed here has its strengths and limitations. For small datasets, all are viable options. However, when dealing with large and complex datasets, the Canopy algorithm offers significant advantages in terms of computational efficiency, fault tolerance, and noise immunity. Future research should explore these algorithms' performance with real-world multidimensional data to gain deeper insights into their applicability and potential for further improvements.**

***References: [3]C.,Yang, H Research on K-Value Selection Method of K-Means Clustering Algorithm. J 2019, 2, 226-235. https://doi.org/10.3390/j2020016***

**A new coreset framework for clustering**

**Vincent Cohen-Addad *et al. ,*  [4] (2021).**

**A)Problem statement**

Clustering is a fundamental problem in machine learning, where the goal is to group data points into clusters such that points within a cluster are more similar to each other than to points in other clusters. Clustering is used in a wide variety of applications, such as image segmentation, natural language processing, and anomaly detection.

One of the challenges of clustering is that it can be computationally expensive to cluster large datasets. This is because traditional clustering algorithms, such as k-means, require examining all pairs of data points.

**B)Method**

The paper proposes a new coreset framework for clustering. A coreset is a small subset of the original dataset that preserves the clustering properties of the original dataset. This means that clustering the coreset will produce a clustering that is very similar to the clustering of the original dataset.

The paper's coreset framework is based on the idea of using a carefully chosen subset of the data points to represent the original dataset. The coreset is constructed by iteratively adding data points to the coreset until the coreset is sufficiently representative of the original dataset.

**C)Advantages/disadvantages**

One of the main advantages of the paper's coreset framework is that it is very efficient. This is because the coreset is much smaller than the original dataset, so clustering the coreset is much faster than clustering the original dataset.

Another advantage of the paper's coreset framework is that it is very effective. The experiments in the paper show that the coreset framework is able to produce clusters that are very similar to the clusters produced by traditional clustering algorithms, even when the coreset is very small.

One of the main disadvantages of the paper's coreset framework is that it can be difficult to choose the right subset of data points to use as the coreset. The paper provides some heuristics for choosing the coreset, but there is no guarantee that these heuristics will always produce a good coreset.

**D)Performance**

The researchers evaluated the performance of the coreset framework on a variety of clustering datasets. The results showed that the coreset framework was able to produce clusters that were very similar to the clusters produced by traditional clustering algorithms, even when the coreset was very small.

**E)Conclusion**

The paper presents a new coreset framework for clustering. The coreset framework is very efficient and effective, and it is able to produce clusters that are very similar to the clusters produced by traditional clustering algorithms, even when the coreset is very small.

Overall, the paper presents a significant contribution to the field of machine learning. The coreset framework is a powerful new tool for clustering large datasets, and it is likely to be used in a wide range of applications in the future.

***References: [4]David Saulpic, Chris Schwiegelshohn. A new coreset framework for clustering. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2021). Association for Computing Machinery, New York, NY, USA, 169–182. https://doi.org/10.1145/3406325.3451022***

**A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots.**

**Ran, *et al.* , [5](2021)**

**A)Problem statement**

Urban hotspots are areas of high activity or concentration. They can be used to identify areas of interest for businesses, governments, and researchers. However, identifying urban hotspots can be challenging, as they may be difficult to define and may change over time.

**B)Method**

The research paper proposes a novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots. The proposed algorithm is based on the traditional K-means clustering algorithm, but it incorporates a noise algorithm to help identify and remove outliers from the data. This makes the proposed algorithm more robust to noise and more effective at capturing urban hotspots.

**C)Advantages/disadvantages**

One of the main advantages of the proposed algorithm is that it is simple and easy to implement. It is also computationally efficient, which makes it suitable for large datasets.

Another advantage of the proposed algorithm is that it is effective at capturing urban hotspots, even in the presence of noise. This is because the noise algorithm helps to identify and remove outliers from the data, which can improve the accuracy of the clustering results.

One of the disadvantages of the proposed algorithm is that it is sensitive to the choice of the number of clusters (K). It is important to choose a value of K that is appropriate for the data, and there is no general rule of thumb for doing this.

Another disadvantage of the proposed algorithm is that it is not able to identify overlapping hotspots. This is because the K-means clustering algorithm only allows each data point to belong to one cluster.

**D)Performance**

The researchers evaluated the performance of the proposed algorithm on a dataset of taxi GPS data from Beijing, China. The results showed that the proposed algorithm was able to identify urban hotspots more accurately than traditional K-means clustering algorithms.

**E)Conclusion**

The research paper presents a novel K-means clustering algorithm with a noise algorithm for capturing urban hotspots. The proposed algorithm is simple, efficient, and effective at capturing urban hotspots, even in the presence of noise. However, the proposed algorithm is sensitive to the choice of the number of clusters (K) and is not able to identify overlapping hotspots.

Overall, the research paper presents a significant contribution to the field of urban data analysis. The proposed algorithm is a powerful new tool for identifying urban hotspots, and it is likely to be used in a wide range of applications in the future.

***References: [5],X.; Zhou, X.; Lei, M.; Tepsan, W.; Deng, W.,A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. Appl. Sci. 2021, 11, 11202. https://doi.org/10.3390/app112311202***

**Clustering objectives in wireless sensor networks: A survey and research direction analysis**

**Amin Shahraki, *et al.*  [6](2020)**

**A)Problem statement**

Clustering is a fundamental problem in wireless sensor networks (WSNs), where the goal is to group sensor nodes into clusters such that nodes within a cluster are more similar to each other than to nodes in other clusters. Clustering is used in WSNs to improve energy efficiency, scalability, reliability, and other performance metrics.

**B)Method**

The paper "Clustering Objectives in Wireless Sensor Networks: A Survey and Research Direction Analysis" reviews a wide range of clustering techniques for WSNs, classifying them based on their clustering objectives. These objectives include:

* Energy efficiency
* Scalability
* Reliability
* Quality of service (QoS)
* Security

The paper also discusses the different network properties that are supported by the different clustering techniques. These properties include:

* Mobility
* Heterogeneity

**C)Advantages/disadvantages**

One of the main advantages of the paper is that it provides a comprehensive and up-to-date overview of clustering techniques in WSNs. The paper also provides statistics on the chosen metrics, such as the number of clustering techniques, the clustering objectives they address, and the network properties they support.

One of the disadvantages of the paper is that it does not go into much detail about the design and implementation of the different clustering techniques. This is because the paper is intended to be a survey, not a tutorial.

**D)Performance**

The paper reviews a wide range of clustering techniques for WSNs, and it provides statistics on the performance of these techniques. The statistics show that energy efficiency is the most common clustering objective, followed by scalability and reliability. The statistics also show that most clustering techniques are designed for static and homogeneous sensor networks.

**E)Conclusion**

The paper is a valuable resource for researchers and practitioners who are interested in designing or implementing clustering algorithms for WSNs. The paper provides a comprehensive and informative overview of the different clustering techniques, their clustering objectives, the network properties they support, and their performance.

***References: [6]****Amir Taherkordi, Øystein Haugen, Frank Eliassen, Clustering objectives in wireless sensor networks: A survey and research direction analysis, Computer Networks, Volume 180,2020,107376, ISSN 1389-1286,https://doi.org/10.1016/j.comnet.2020.107376*.

**Unsupervised K-Means Clustering Algorithm**

**K. P. Sinaga , et al. , [7](2020)**

**A)Problem statement**

Clustering is a common unsupervised learning task that involves grouping data points into clusters such that data points within a cluster are more similar to each other than to data points in other clusters. K-means clustering is a popular clustering algorithm that divides data points into k clusters, where k is a user-specified value.

However, existing k-means clustering algorithms have several limitations. First, these algorithms can be inefficient, as they require the calculation of the distance between all pairs of data points. Second, these algorithms can be sensitive to outliers, as they assign data points to the cluster with the closest centroid.

**B)Method**

The proposed algorithm addresses the limitations of existing k-means clustering algorithms by using a weighted distance metric to assign data points to clusters. The weighted distance metric gives more weight to data points that are closer to the centroid. This helps to improve the efficiency of the algorithm by reducing the number of distance calculations that need to be performed. The weighted distance metric also helps to improve the robustness of the algorithm to outliers by assigning outliers to clusters that are not too close to the centroid.

**C)Advantages/disadvantages**

The proposed algorithm has several advantages over existing k-means clustering algorithms. First, the algorithm is more efficient than existing algorithms, as it does not require the calculation of the distance between all pairs of data points. Second, the algorithm is more robust to outliers, as it uses a weighted distance metric to assign data points to clusters.

One of the disadvantages of the proposed algorithm is that it requires the user to specify the number of clusters, k. This can be a difficult task, as the optimal number of clusters is often unknown.

**D)Performance**

The proposed algorithm was evaluated on a variety of data sets, and it was shown to be effective in clustering data with different characteristics. The algorithm was also shown to be more efficient and robust than existing k-means clustering algorithms.

**E)Conclusion**

The proposed algorithm is a promising new approach to k-means clustering. The algorithm is more efficient and robust than existing algorithms, and it is a good choice for clustering data with different characteristics.

References:[7] **M. -S. Yang,"Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.**

**An Overview of Fairness in Clustering**

**A. Chhabra, *et al.*  , [8](2021**

**A)Problem statement**

Clustering is a common unsupervised learning task that involves grouping data points into clusters such that data points within a cluster are more similar to each other than to data points in other clusters. Clustering algorithms are often used in a variety of applications, such as customer segmentation, medical diagnosis, and fraud detection.

However, clustering algorithms can be unfair, as they can lead to the segregation of data points based on sensitive attributes, such as race, gender, or religion. This can have a negative impact on individuals and groups, as it can lead to discrimination and unequal treatment.

**B)Method**

The paper reviews a variety of methods for achieving fairness in clustering algorithms. These methods can be classified into two broad categories:

* Pre-processing methods: These methods modify the data before it is clustered. For example, these methods can be used to remove sensitive attributes from the data or to transform the data to make it more fair.
* Post-processing methods: These methods modify the clustering results after the clustering algorithm has been run. For example, these methods can be used to reassign data points to clusters to achieve fairness.

**C)Advantages/disadvantages**

Pre-processing methods are generally easier to implement than post-processing methods. However, pre-processing methods can also be more disruptive to the data, as they may remove important information.

Post-processing methods are generally more flexible than pre-processing methods. However, post-processing methods can be more difficult to implement, as they may require the design of a new clustering algorithm.

**D)Performance**

The performance of fairness-aware clustering algorithms is a subject of ongoing research. However, some studies have shown that fairness-aware clustering algorithms can improve the fairness of clustering results.

**E)Conclusion**

Fairness in clustering is an important issue that needs to be addressed. The paper by Chhabra, Masalkovaitė, and Mohapatra provides a comprehensive overview of the challenges and approaches to achieving fairness in clustering. The paper also discusses the challenges of evaluating the fairness of clustering algorithms. There is no single agreed-upon metric for evaluating fairness in clustering. However, some common metrics include the following:

* Balance: This metric measures the proportion of data points from each protected group in each cluster.
* Isolation: This metric measures the degree to which data points from different protected groups are separated into different clusters.
* Equalized odds: This metric measures the probability of being assigned to a particular cluster given a protected attribute.

The paper concludes by discussing the future of fairness in clustering. The authors argue that there is a need for more research on fairness-aware clustering algorithms and on the development of new metrics for evaluating fairness in clustering.

***References: [8]****K. Masalkovaitė and P. Mohapatra, "An Overview of Fairness in Clustering," in IEEE Access, vol. 9, pp. 130698-130720, 2021, doi: 10.1109/ACCESS.2021.3114099*

**Ultra-Scalable Spectral Clustering and Ensemble Clustering**

**D. Huang, *et al. ,*  [9](2020)**

**A)Problem statement**

Spectral clustering is a popular clustering algorithm that has been shown to be effective for clustering data with complex relationships. However, spectral clustering can be computationally expensive for large datasets.

Ensemble clustering is a technique that combines the results of multiple clustering algorithms to improve the overall accuracy and robustness of the clustering results. However, ensemble clustering can also be computationally expensive for large datasets.

**B)Method**

U-SPEC and U-SENC address the limitations of existing spectral clustering and ensemble clustering algorithms by being more scalable and robust.

U-SPEC uses a hybrid representative selection strategy and a fast approximation method for K-nearest representatives to reduce the computational complexity of spectral clustering. U-SENC integrates multiple U-SPEC clusters into a single framework to improve the robustness of U-SPEC while maintaining its high efficiency.

**C)Advantages/disadvantages**

The main advantages of U-SPEC and U-SENC are their scalability and robustness. U-SPEC and U-SENC are significantly faster than existing spectral clustering and ensemble clustering algorithms for large datasets. U-SENC is also more robust to noise and outliers than existing spectral clustering and ensemble clustering algorithms.

One of the disadvantages of U-SPEC and U-SENC is that they require more memory than existing spectral clustering and ensemble clustering algorithms. This is because U-SPEC and U-SENC need to store the entire affinity matrix of the data.

**D)Performance**

U-SPEC and U-SENC were evaluated on a variety of large-scale datasets, and they were shown to be more efficient and robust than existing spectral clustering and ensemble clustering algorithms.

For example, on a dataset with 10 million data points, U-SPEC was able to cluster the data in about 10 minutes, which is significantly faster than existing spectral clustering algorithms, which can take hours or even days to cluster such large datasets.

**E)Conclusion**

U-SPEC and U-SENC are two novel algorithms for large-scale spectral clustering and ensemble clustering. U-SPEC and U-SENC are more scalable and robust than existing spectral clustering and ensemble clustering algorithms. U-SPEC and U-SENC have the potential to be used in a wide variety of applications, such as image clustering, social network analysis, and customer segmentation.

***References:[9] C. -D. Wang, J. -S. Wu, J. -H. Lai and C. -K. Kwoh, ;"Ultra-Scalable Spectral Clustering and Ensemble Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1212-1226, 1 June 2020, doi: 10.1109/TKDE.2019.2903410***

**Syed, K. Comprehensive Analyses of Cytochrome P450 Monooxygenases and Secondary Metabolite Biosynthetic Gene Clusters in Cyanobacteria**

**Khumalo, et al. , [10](2020)**

**A)Problem statement**

Cytochrome P450 monooxygenases (CYPs) are a diverse group of enzymes that play a key role in the metabolism of xenobiotics and secondary metabolites in cyanobacteria. Secondary metabolites are natural products that are not essential for growth or reproduction, but they can provide a variety of benefits to cyanobacteria, such as protection from predators and competitors. However, there is a lack of comprehensive information on the diversity and role of CYPs in cyanobacteria.

**B)Method**

The authors of the paper "Comprehensive Analyses of Cytochrome P450 Monooxygenases and Secondary Metabolite Biosynthetic Gene Clusters in Cyanobacteria" by Khumalo et al. (2023) analyzed the genomes of 114 cyanobacterial species to identify CYPs and secondary metabolite biosynthetic gene clusters (BGCs). They then used this information to study the diversity and role of CYPs in cyanobacteria.

**C)Advantages/disadvantages**

One of the main advantages of this study is that it provides a comprehensive analysis of CYPs and secondary metabolite BGCs in cyanobacteria. This information is valuable for understanding the diversity and role of CYPs in cyanobacteria and for developing new methods for the production of valuable secondary metabolites from cyanobacteria.

One of the disadvantages of this study is that it is limited to the analysis of the genomes of 114 cyanobacterial species. It is possible that there are other cyanobacterial species that contain CYPs that are not represented in this study.

**D)Performance**

The authors of the study found that cyanobacterial CYPs are highly diverse and that they belong to a wide range of CYP families. They also found that only 8% of CYPs were part of BGCs, suggesting that most cyanobacterial CYPs are involved in primary metabolism. However, the authors also identified a number of CYPs that are likely involved in secondary metabolite biosynthesis, including CYPs that are part of BGCs for the production of terpenes, polyketides, and non-ribosomal peptides.

**E)Conclusion**

This study provides new insights into the diversity and role of CYPs in cyanobacteria. The authors suggest that cyanobacterial CYPs have the potential to be exploited for the production of valuable secondary metabolites, such as antibiotics, antitumor agents, and other pharmaceuticals. This study is a significant contribution to the field of cyanobacterial research. The authors have provided a comprehensive analysis of CYPs and secondary metabolite BGCs in cyanobacteria. This information could be used to develop new methods for the production of valuable secondary metabolites from cyanobacteria.

***References: [10];M.J.; Nzuza, N.; Padayachee, T.; Chen, W.; Yu, J.-H.; Nelson, D.R.; Syed, K. Comprehensive Analyses of Cytochrome P450 Monooxygenases and Secondary Metabolite Biosynthetic Gene Clusters in Cyanobacteria. Int. J. Mol. Sci. 2020, 21, 656.*** *https://doi.org/10.3390/ijms21020656*