**Agrupamiento**

* **Introducción**

El agrupamiento también conocido como *clustering*, es una técnica de aprendizaje no supervisado en la ciencia de datos. Busca organizar y clasificar grandes volúmenes de información, con base en similitudes, patrones en los datos, en grupos o *clusters* con características similares.[[1]](https://paperpile.com/c/xD7Rnv/fbz1)

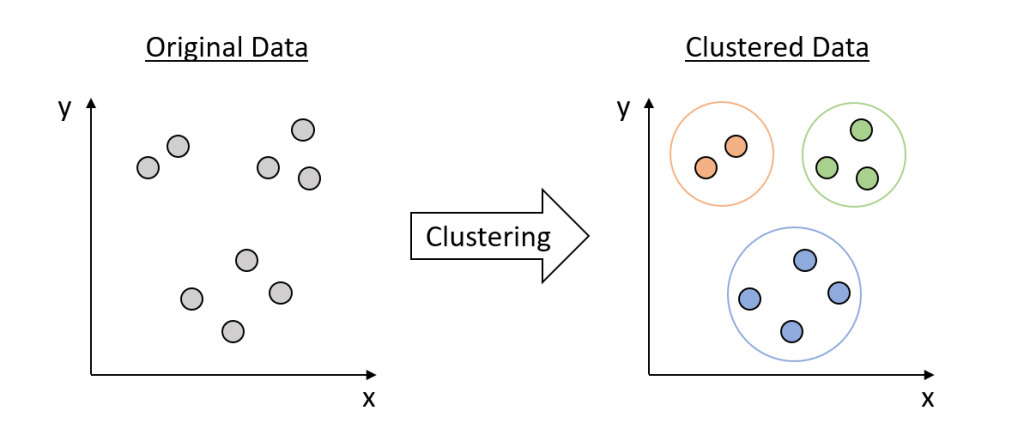


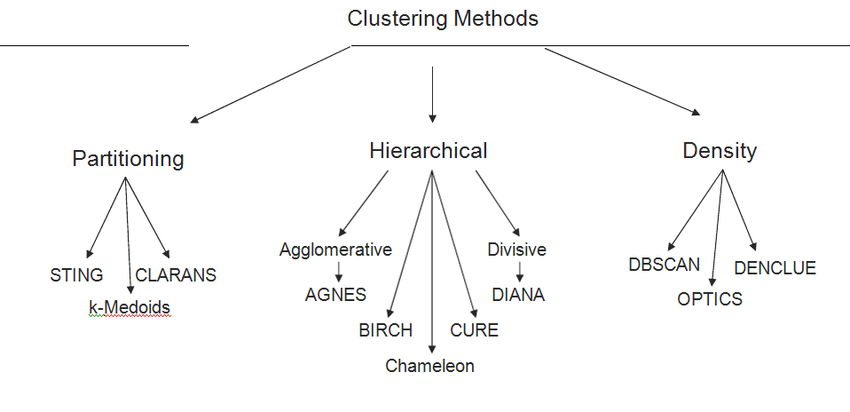
Figura 1. Representación de como el *clustering* puede ser empleado para identificar patrones en un conjunto de datos

Esta técnica a través de algoritmos que calculan la similitud entre datos, los agrupan en función de su proximidad. Dependiendo del tipo de datos que se posean y el objetivo del análisis pueden emplearse una gran variedad de algoritmos de agrupamiento. Algunos de los algoritmos más conocidos son k-means, DBSCAN y *hierarchical clustering*

Agrupar los datos en grupos o *clusters* nos ayudan a entender las relaciones no marcadas entre los datos, al ver características o rasgos distintivos que comparten una o varias categorías. También elagrupamiento nos permite detectar anomalías al observar la baja asociación de uno o varios datos con algún cluster y poder descartar estos datos aberrantes (*outliers*). Adicionalmente, puede ser empleado en la reducción de la complejidad de grandes conjuntos de datos, al reducir el número de variables[2]. [[2]](https://paperpile.com/c/xD7Rnv/HRHP)

Las aplicaciones del agrupamiento en el análisis de datos, es diversa, al poder ser empleado en numerosos campos, desde el marketing de una empresa, hasta en la clasificación de genes en biología.

* Algoritmos de *Clusterings*

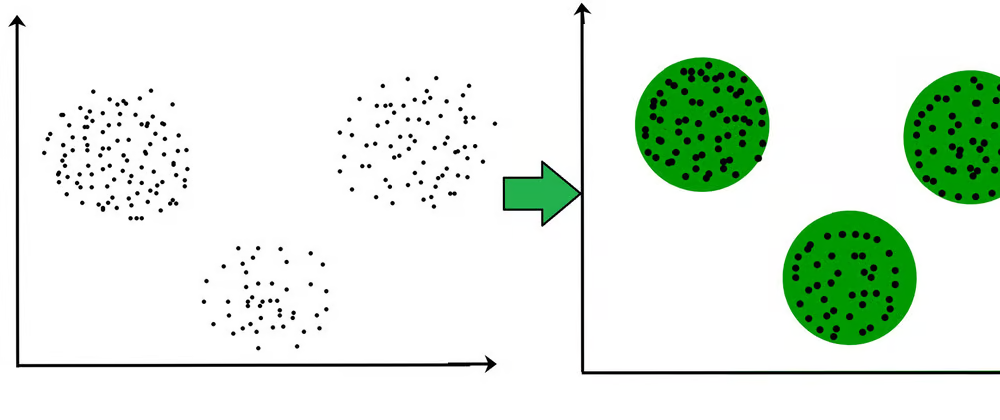
**

Así como existen diferentes tipos de algoritmos para los clusters, también existen diversas formas de definir a un cluster. Los diferentes enfoques de los modelos funcionan de manera distinta, dependiendo del tamaño de la información de entrada, la rigidez en las categorías, y el número de clusters para el conjunto de datos. Es muy posible que un algoritmo que funciona de manera adecuada para un conjunto de datos, trabaje de manera poco efectiva ante otro grupo de datos.[2] [[2]](https://paperpile.com/c/xD7Rnv/HRHP). Para definir un cluster, existen diferentes métodos de los cuales se parten, los que abordaremos son:

* *Clustering* basado en centroides
* *Clustering* jerarquico
* *Clustering* basado en distribución
* *Clustering* basado en densidad
* *Clustering* basado en rejillado

***Clustering* basado en centroides *(Centroid-based)***

Este método se centra en la partición o división del conjunto de datos en grupos de datos similares, conocidos como *clusters*. Se basa en la distancia que tienen los datos respecto al centro del grupo. La distancia a cada centro (*cluster´s centroid*), se refiere a la media o mediana de todos los puntos del grupo, según sean los datos[3].[[3]](https://paperpile.com/c/xD7Rnv/4Rxw) Ver figura 1

  
Imagen 2. Cluster basado en centroide (*Centroid-based clustering)*

Este tipo de algoritmo es ideal para conjuntos de datos que sean fácilmente separables, con *clusters* bien definidos. Es apropiado cuando los clusters son conocidos o pueden ser rápidamente estimados. Sin embargo, no es la mejor opción cuando hay solapamiento de los clusters o estos no poseen formas uniformes.

**Tipos de algoritmosbasado en centroides**

1. ***K-Means*:** Es el más empleado para este método, ya que minimiza la suma de las distancias entre los datos y sus correspondientes centroides. Trabaja de manera adecuada con *clusters* previamente definidos, de tamaño relativamente similar, y que no posean *outliers*. Es especialmente sensible a los *outliers*, debido que estos provocan un sobreajuste sobre los centroides al tratar de considerarlos.
2. ***K-Medoids*:** Variación de K-means, que utiliza medoides. Los medoides son datos representativos del conjunto de datos, donde en lugar de poseer un centroide de manera arbitraria como el centro del cluster, el algoritmo crea clusters empleando datos individuales como el centro del cluster (medoide). Debido al uso de medoides en lugar de centroides, es menos sensible a *outliers.*
3. **Fuzzy c-Means Clustering:** Variación de K-means que permite a los datos pertenecer a más de un cluster, con distintos grados de pertenencia.
4. ***Expectation Maximization (EM):*** Modelo que emplea modelos estadísticos para definir las relaciones entre los datos y los clusters.

Las aplicaciones para este tipo de algoritmo van desde la segmentación de imágenes, empleando el algoritmo *K-means.* Estas imágenes se pueden dividir en múltiples segmentos basándose en color, texto o alguna otra característica, hasta la detección de anomalías al identificar datos diferentes por medio de *K-medoids.*

***Clustering jerárquico (Hierarchical clustering)***

También conocido como agrupación de clusters basado en conectividad, agrupa los puntos de datos en función de la proximidad y la conectividad de sus características. Este método determina los clusters basándose en que tan cerca están datos unos de otros a través de todas las dimensiones. La idea es que los datos que estén más próximos unos de otros se encuentren más relacionados que los que se encuentran más alejados, asignándoles así una jerarquía . A diferencia de *k-means,* no se requiere de pre-especificar el número de clusters. En su lugar, el algoritmo crea una red gráfica de los clusters en cada nivel jerárquico. La forma de presentación más común es un dendrograma[[4](https://doi.org/10.1016/j.asoc.2023.110303) ] [[4]](https://paperpile.com/c/xD7Rnv/CDre) (Ver Figura 2).

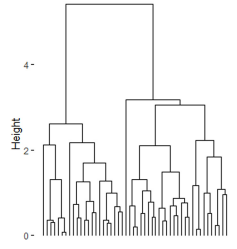
****

Imagen 3. Dendrograma de *Cluster*

Hay dos aproximaciones para *clustering* jerárquico:

1. **Aglomerativo:** Se comienza definiendo cada punto de datos como un *subcluster.* Definimos una métrica para medir la distancia entre todos los pares de subclusters en cada paso y seguimos fusionando los dos subclusters más cercanos en cada paso. Se repite el procedimiento hasta que solo quede un cluster en el sistema. Tiene como desventaja que es más lento que otros métodos de jerarquía
2. **División:** El algoritmo empleado para realizar este clustering divisivo es *k-means.* Donde se realiza en cada cluster el procedimiento de k-means, hasta encontrar todas las muestras de datos en el sistema o el número de clusters a obtener. Se debe conocer cuántos clusters se desean obtener a continuación.

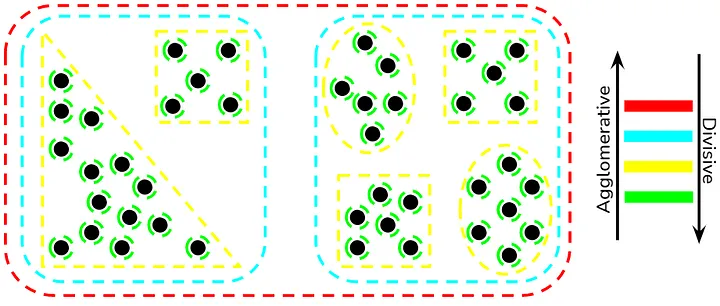


Imagen 4. Formación de cluster de división o aglomerativos. Donde cada punto representa un punto de datos. []5 [[5]](https://paperpile.com/c/xD7Rnv/aypO)

Las ventajas del método *Clustering* jerárquico *son:*

* Facilita el trabajar con formas más complejas, ya que se posee una forma gráfica y visualmente atractiva para la visualización de los clusters.
* Es fácil definir el número de clusters, al agrupar los datos en diversos niveles jerárquicos.

Las desventajas del método *Clustering* jerárquico *son:*

* Es un método fuertemente influenciado por heurísticas.
* El algoritmo matemático es pesado, debido a que se deben calcular las distancias entre los subclusters cada vez.
* Entré más aumente el tamaño de nuestro conjunto de datos, el analizar de manera visual se volverá más complejo.

**Clustering basado en distribución[6]** [[6]](https://paperpile.com/c/xD7Rnv/V7wK)

Llamados también clustering probabilístico, agrupa los datos basándose en su distribución probabilística. Este algoritmo supone que todos los datos son parte de un cluster, basándose en la probabilidad de que pertenezcan a un cluster dado. Funciona de manera que, del punto central, al aumentar la distancia del centro la probabilidad de que pertenezca al cluster disminuye. Los clusters se definen como grupos de datos que son más afines a tener la misma distribución antes que otra. El cálculo de ajuste de distribución, puede ser costoso computacionalmente al trabajar con grandes conjuntos de datos.

**Método Gaussian Mixture Model (GMM)**, supone que los datos de un conjunto de datos son generados de una mezcla de diversas distribuciones gaussianas. Cada una de estas distribuciones Gaussianas representa un cluster diferente. Es fundamental la elección del número de distribuciones para obtener resultados óptimos

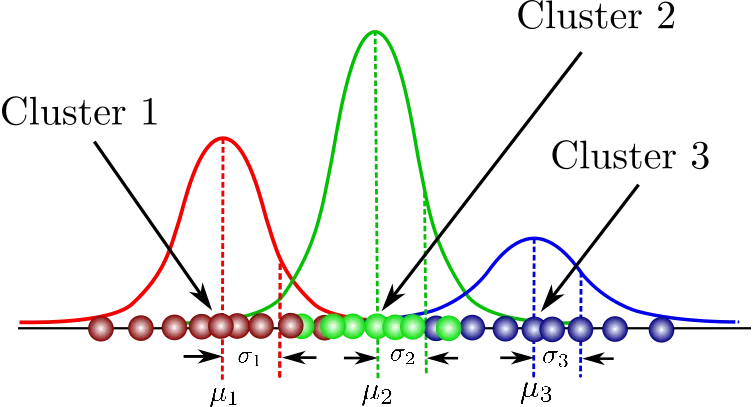


Imagen 5. Formación de distribuciones Gaussianas en conjunto de datos[6] [[6]](https://paperpile.com/c/xD7Rnv/V7wK)

***Clustering* basado en densidad**

Este algoritmo trabaja al detectar áreas donde los puntos de datos se encuentran concentrados, y también identificando las áreas por las cuales están separadas o hay un vacío. A diferencia de otros métodos basados en centroides, es capaz de detectar clusters de manera arbitraria. Esto puede ser de mucha utilidad cuando los clusters no están definidos de manera específica o por distribución También es capaz distinguir entre puntos que son parte de un cluster y los que son deberían ser considerados como ruido. Es decir que los clusters basados en densidad, son especialmente efectivos al trabajar con conjuntos de datos que posean ruido o *outliers,* y no se tenga como prioridad establecer un número de clusters en el conjunto de datos

**Método DBSCAN**[7][[7]](https://paperpile.com/c/xD7Rnv/Zwvw) **,** es un método centrado en la densidad. Utiliza un enfoque de agrupamiento espacial basado en la densidad para crear grupos con una densidad previamente establecida por el usuario que se coloca alrededor de un centroide espacial. El área inmediata alrededor de los centroides se llama vecindario, DBSCAN trata de definir vecindarios de cluster con una densidad especificada. Dentro de este algoritmo se tiene 3 tipos de puntos de datos (Imagen 5)

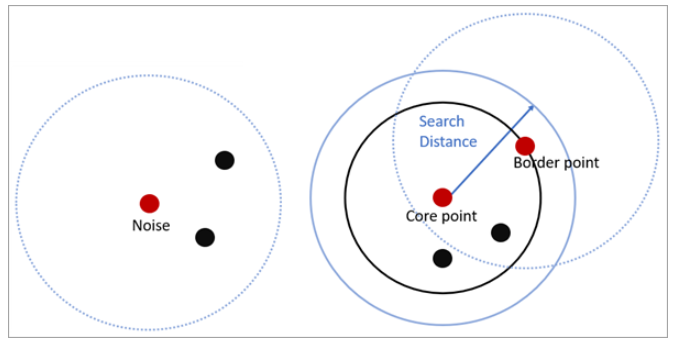


Imagen 6. Clustering de cuatro entidades mínimas por cluster. Donde el núcleo tiene cuatro vecinos dentro de la distancia de búsqueda incluyéndose a sí mismo. El punto de borde solo posee tres entidades dentro de la distancia establecida, pero al ser vecino del núcleo, se incluye dentro del clúster. El punto de ruido, no posee cuatro entidades de búsqueda y no es vecino del núcleo, por lo que no se incluye dentro de un cluster.

**Método HDBSCAN,** es una variante de DBSCAN pero no requiere de algún parámetro previo para ser establecida, ofreciendo una mayor flexibilidad que su original. Es menos sensible al ruido o los *outliers.* Un problema que presenta DBSCAN, es que puede llegar a presentar problemas identificando clusters que no tengan una densidad uniforme, esto es algo que HDBSCAN maneja de una manera mucho más efectiva.

**Clustering basado en rejillado (*Grid-based clustering*)**

Este método se centra en dividir el espacio del conjunto de datos en una estructura de rejilla. El agrupamiento se lleva a cabo hacia las celdas de la rejilla en lugar de los puntos de datos. Gracias a esto presenta como principal ventaja ante otros métodos, un mejor desempeño en el tiempo de procesamiento[8]. [[8]](https://paperpile.com/c/xD7Rnv/rCM2)

**Método STING,** del inglés “*STatistical INformation Grid*”, es un algoritmo que divide el espacio en celdas rectangulares y en varios niveles de celda con diferentes niveles de resolución. Es decir usa una estructura de rejilla multidimensional, cuantificando el espacio en un número finito de celdas.

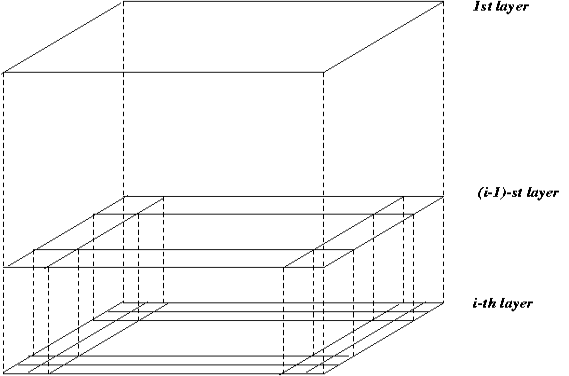


Imagen 6. Método de agrupación STING. Donde para cada celda, el nivel superior se divide en varias celdas más pequeñas para el siguiente nivel inferior[9] [[9]](https://paperpile.com/c/xD7Rnv/gbrm).

Las ventajas del *clustering* basado en rejillado *son:*

* Al ser un método basado en cuadrícula, es *query-independent*, ya que la información estadística de cada celda representa un resumen de la información de los datos que se encuentran en esta, permitiendo hacer consultas independientes
* La estructura de cuadrícula facilita el procesamiento en paralelo y las actualizaciones incrementales

A pesar de esto, presenta como principal desventaja, que los límites de *cluster* son horizontales o verticales, no puede detectar límites diagonales

**Métricas de evaluación para agrupamiento**

Al momento de obtener nuestros clusters es posible evaluar la cohesión y/o la separación entre ellos, es decir que tan cerca o alejados estan los puntos de un cluster en sí mismo o de otro. Dentro de la biblioteca de sklearn, existen dos métricas que nos pueden ayudar al momento de evaluar nuestros clusters obtenidos:

1. Silhoutte Score: Esta métrica mide la calidad de los clusters al calcular que tan cerca están los puntos dentro de un mismo clusters (cohesión) y comparándolo con los puntos de otros clusters (separación)
2. Davies-Bouldin Index: Esta evalúa que tan cerca están los puntos dentro de un cluster y la separación entre clusters

**Aplicaciones en diseño de fármacos**

**Principalmente son usados en reposicionamiento de fármacos, cribado virtual para el análisis de conformaciones y en el tratamiento de imágenes.**

| Applicación | Algoritmos | Referencia |
| --- | --- | --- |
| Elaboración de perfiles basados ​​en imágenes para descubrimiento de medicamentos: identificación de fenotipos asociado a enfermedades, predicción de actividad o toxicidad de un fármaco | *Clustering* basado en centroide, por algoritmo de k-means | [[10]](https://paperpile.com/c/xD7Rnv/fQx4) |
| Caracterización de interacciones fármaco-fármaco. Drug-drug interactions (DDIs) | *Cluster* jerárquico, SFM | [[11]](https://paperpile.com/c/xD7Rnv/5s7b) |
| Reposicionamiento de fármacos para tratamiento de COVID-19 | *Cluster* jerárquico, k-means, *spectral clustering* | [[12]](https://paperpile.com/c/xD7Rnv/xO9V) |
| Descubrimiento de candidatos a fármacos para la prevención de COVID-19 | *Clustering* aglomerativo jerárquico *(AHC)* | [[13]](https://paperpile.com/c/xD7Rnv/EfKi) |
| Descubrimientos de ligandos selectivos para MT1, en *ultra-large data bases* | *Clustering* por similaridad topológica | [[14]](https://paperpile.com/c/xD7Rnv/JeN3) |
| Drug Repurposing Using Modularity Clustering in Drug-Drug Similarity Networks Based on Drug-Gene Interactions |  |  |
| Drug Repurposing Based on Drug-Drug Interaction  DOI10.1111/cbdd.12378 |  |  |
| A Machine Learning-Enabled Pipeline for Large-Scale Virtual Drug Screening |  |  |

Focus en mi proyecto de tesis que sería en large-ultra large db

**Referencias:**

1. [MisApuntes P. Qué es el clustering y cómo funciona: guía completa. In: MisApuntes [Internet]. 9 Oct 2023 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/fbz1) <https://misapuntesdedatascience.es/que-se-entiende-por-clustering/>

2. [What is clustering? 21 Feb 2024 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/HRHP) <https://www.ibm.com/topics/clustering>

3. [Verma A. Centroid-based clustering: A powerful machine learning technique for partitioning datasets. In: DEV Community [Internet]. 7 Feb 2023 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/4Rxw) <https://dev.to/anurag629/centroid-based-clustering-a-powerful-machine-learning-technique-for-partitioning-datasets-41im>

4. [Hierarchical clustering: Visualization, feature importance and model selection. Appl Soft Comput. 2023;141: 110303. doi:](http://paperpile.com/b/xD7Rnv/CDre)[10.1016/j.asoc.2023.110303](http://dx.doi.org/10.1016/j.asoc.2023.110303)

5. [Pai P. Hierarchical clustering explained. In: Towards Data Science [Internet]. 7 May 2021 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/aypO) <https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8>

6. [Cota S. Fundamental tasks of AI — part 3 — clustering (2 of 2). In: Medium [Internet]. 7 Dec 2023 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/V7wK) <https://medium.com/@sasirekharameshkumar/deep-learning-basics-part-9-using-ai-for-clustering-2-of-2-d687b952a265>

7. [Cómo funciona el clustering basado en densidad. [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/Zwvw) <https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm>

8. [Improve IF. STING - Statistical Information Grid in Data Mining. In: GeeksforGeeks [Internet]. 2 Apr 2022 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/rCM2) <https://www.geeksforgeeks.org/sting-statistical-information-grid-in-data-mining/>

9. [Grid-Based Clustering - STING, WaveCluster & CLIQUE. Blogger; 6 Apr 2020 [cited 4 Jun 2024]. Available:](http://paperpile.com/b/xD7Rnv/gbrm) <https://www.datamining365.com/2020/04/grid-based-clustering.html>

10. [Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. 2021;20: 145–159. doi:](http://paperpile.com/b/xD7Rnv/fQx4)[10.1038/s41573-020-00117-w](http://dx.doi.org/10.1038/s41573-020-00117-w)

11. [Lv J, Liu G, Ju Y, Sun B, Huang H, Sun Y. Integrating multi-source drug information to cluster drug-drug interaction network. Comput Biol Med. 2023;162: 107088. doi:](http://paperpile.com/b/xD7Rnv/5s7b)[10.1016/j.compbiomed.2023.107088](http://dx.doi.org/10.1016/j.compbiomed.2023.107088)

12. [Bansal C, Deepa PR, Agarwal V, Chandra R. A clustering and graph deep learning-based framework for COVID-19 drug repurposing. Expert Syst Appl. 2024;249: 123560. doi:](http://paperpile.com/b/xD7Rnv/xO9V)[10.1016/j.eswa.2024.123560](http://dx.doi.org/10.1016/j.eswa.2024.123560)

13. [Supianto AA, Nurdiansyah R, Weng C-W, Zilvan V, Yuwana RS, Arisal A, et al. Cluster-based text mining for extracting drug candidates for the prevention of COVID-19 from the biomedical literature. J Taibah Univ Med Sci. 2023;18: 787–801. doi:](http://paperpile.com/b/xD7Rnv/EfKi)[10.1016/j.jtumed.2022.12.015](http://dx.doi.org/10.1016/j.jtumed.2022.12.015)

14. [Stein RM, Kang HJ, McCorvy JD, Glatfelter GC, Jones AJ, Che T, et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. Nature. 2020;579: 609–614. doi:](http://paperpile.com/b/xD7Rnv/JeN3)[10.1038/s41586-020-2027-0](http://dx.doi.org/10.1038/s41586-020-2027-0)