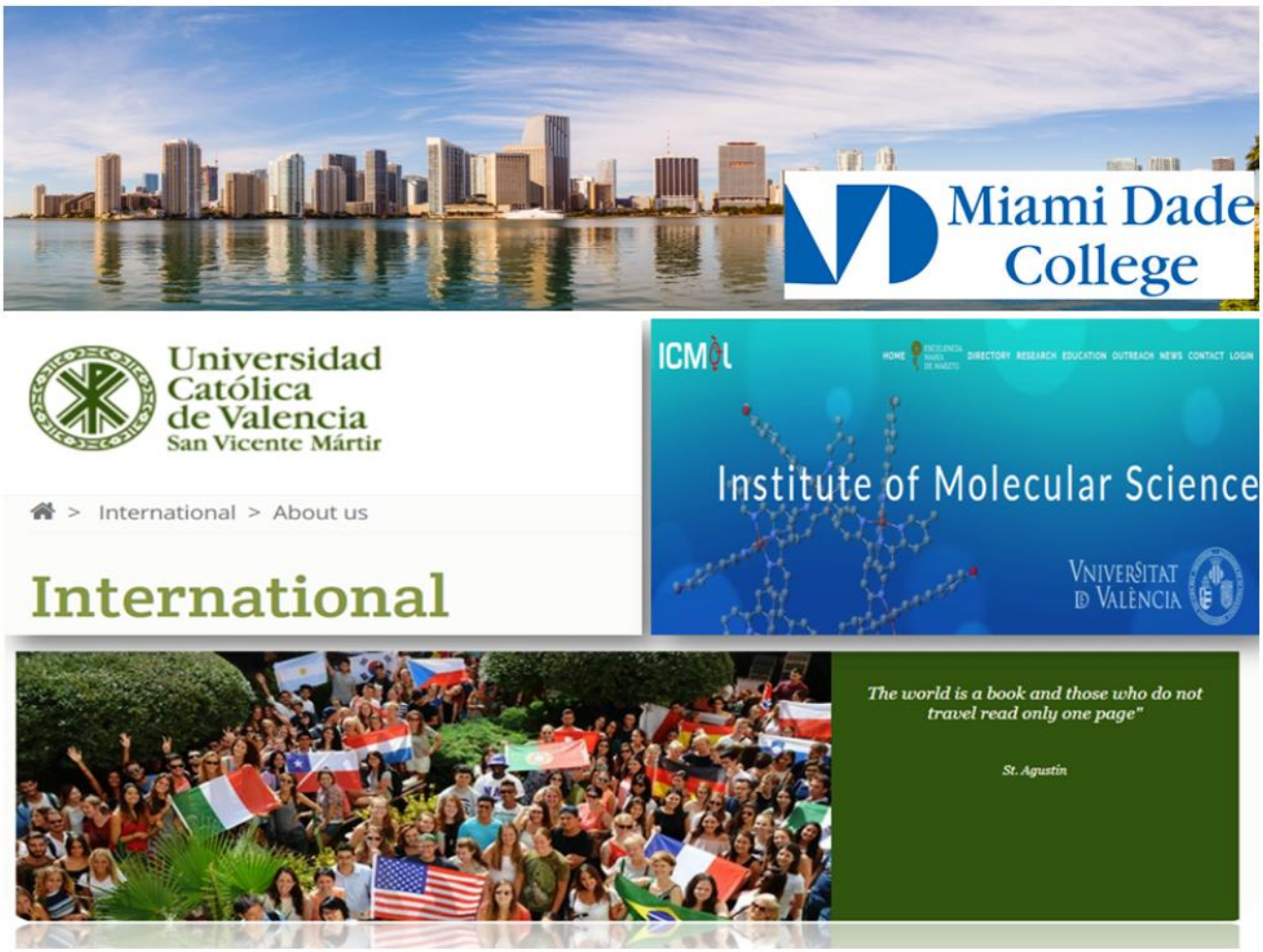
NICEXSM-09: North-Ibero-America Congress on Exp. & Simul. Methods, Valencia, Spain-Miami, USA, 2023.



Recent Topic in Computer-aided Drug Design and Discovery in Biomedical Research

Shan He

Department of Organic and Inorganic Chemistry, Faculty of Science and Technology, University of the Basque Country UPV/EHU, P.O.Box 644, 48080 Bilbao, Spain.
IKERDATA S.L INVESTIGO Next-generation Researcher.

Abstract. Drug design and discovery is a complex, expensive and arduous procedure taking into account the multiple existing diseases and their variants. This long process includes the identification of potential targets and the development of therapeutically safe and effective drugs.1 Computer-aided drug design (CADD) can make it less time- and resource-consuming. In recent research, computational and statistical techniques are used in an effective way to study biomedical compounds for target identification and hit hunting. The arrival of ML in this field of study offers important enhancement in the efficacy of drug

design and discovery process. The success drug design, discovery and development are in concordance with the computational methods and tools. They need to be accurate and use a reliable pre-processed data. Henceforward, Artificial Intelligence/Machine Leaning (AI) approaches to data pre-processing, modeling and representative applications in drug design and discovery will be introduced.

Drug design and discovery is a complex, expensive and arduous procedure taking into account the multiple existing diseases and their variants. This long process includes the identification of potential targets and the development of therapeutically safe and effective drugs.[1] Computer-aided drug design (CADD) can make it less time- and resource-consuming. In recent research, computational and statistical techniques are used in an effective way to study biomedical compounds for target identification and hit hunting. [2] Furthermore, CADD approach can applied the combined biochemical space to assure safety, efficacy and elude toxicity for the achievement of drug development. Recently, the advance and non-stop growth of big data in biological, chemical and pharmaceutical medicine, multiple machine learning (ML) algorithms have been optimized and used in the CADD strategy. The arrival of ML in this field of study offers important enhancement in the efficacy of drug design and discovery process. The success drug design, discovery and development are in concordance with the computational methods and tools. They need to be accurate and use a reliable pre-processed data.[3] Henceforward, Artificial Intelligence/Machine Leaning (AI) approaches to data pre-processing, modeling and representative applications in drug design and discovery will be introduced.

Clustering analysis is a hot research topic in the process of discovering cancer. This approach uses diverse profiles of gene expression, which is the most essential feature to successfully diagnose and treat the cancer disease. Several ensemble clustering systems have been established to achieve clustering utilizing tumor data. However, only few tumor data are incorporated a significant number of input clusterings. Hedjam *et al.* introduced two innovative phases in the ordinary fuzzy k-means algorithm to study the optimal number of input clusterings. In addition, they determined the optimal number of clusters in each clustering for ensemble clustering. The first step was to include a disadvantage for making the algorithm insensitive to the initialization of cluster centroids. The second step was to mechanize a clustering process for iteratively updating the feature weights. It is worth mentioning, the second step addressed the noise values in the dataset. Therefore, they proposed a collaborative clustering method, which combined a set of input clusterings into a final clustering, which have better overall quality. The experiments on real cancer gene expression profiles results illustrated that the proposed algorithm outperformed the well-known clustering algorithms.[4]

Following to this topic, Wang *et al.* proposed the unsupervised Linear Discriminate Analysis (Un-LDA) and first formulated as a flawlessly combined objective optimization which assures convergence in the iteratively unconventional resolving process. The objective optimization was in the ratio trace and the trace ratio forms. They formed a wide-ranging framework of a new method mutually clustering and unsupervised subspace learning. The use of Un-LDA allows to complete unsupervised subspace learning through the clearly obtainable subspace estimate matrix. This approaches can at the same time finish clustering and even clustering out-of-sample data through the explicitly presented transformation matrix. To dealt with the problem in solving the non-convex objective optimization. They statistically demonstrated that the Un-LDA optimization in both forms can be transformed into the simple K-means clustering optimization when the subspace was strongminded. The Un-LDA optimization was ultimately accomplished by alternatively optimizing the clusters using K-means. Moreover, the subspace using the supervised LDA methods and reiterating this entire process until

convergence or stopping criterion. The results showed that the proposed Un-LDA algorithms was comparable or even much superior to the counterparts.[5]

As last example, Ballester *et al.* presented a study where the accuracy of AutoDock Vina (commonly-used docking software) was optimisized by following a ML technique. In addition, they studied the factors that was responsible for this improvement and their generality.It is worth mentioning, with a proposed benchmark support, the athors showed that this improvement will be larger as more data becomes available for training Random Forest (RF) models. This can be explained by the regression models implying additive functional forms do not improve with more training data.Therefore, they discussed how the latter opened the door to new opportunities in scoring function development. They wanted to facilitate the translation of this advance to enhance structure-based molecular design. To do so, in this paper, they provided software to directly re-score Vina-generated posed and thus strongly improve their predicted binding affinity. The software is available at http://istar.cse.cuhk.edu.hk/rf-score-3.tgz and http://crcm. marseille.inserm.fr/fileadmin/rf-score-3.tgz.[6]

In conclusion, drug discovery is a very laborious, long and costly process. ML approaches can significantly decrease development time and costs. However, the majority of methods imply prior knowledge of their physicochemical characteristics or the structure of the drug. All research work mentioned above are beneficial for the early stage of drug discovery and design. Furthermore, they used big data and AI techniques to build computational and statistical models to solve various problems in drug discovery requires high-quality data as essential parts of research. In spite of the current success, there is still a big room for enhancement in terms of method accuracy.[7]

REFERENCE

1. (a) Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific reports* **2021,** *11* (1), 1-13; (b) Wang, H.; Zheng, X.; Jin, J.; Zheng, L.; Guan, T.; Huo, Y.; Xie, S.; Wu, Y.; Chen, W. LncRNA MALAT1 silencing protects against cerebral ischemia-reperfusion injury through miR-145 to regulate AQP4. *Journal of biomedical science* **2020,** *27* (1), 1-12.
2. Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity* **2021,** *25*, 1315-1360.
3. Katsila, T.; Spyroulias, G. A.; Patrinos, G. P.; Matsoukas, M.-T. Computational approaches in target identification and drug discovery. *Computational and structural biotechnology journal* **2016,** *14*, 177-184.
4. Khan, I.; Luo, Z.; Shaikh, A. K.; Hedjam, R. Ensemble clustering using extended fuzzy k-means for cancer data analysis. *Expert Systems with Applications* **2021,** *172*, 114622.
5. Wang, F.; Wang, Q.; Nie, F.; Li, Z.; Yu, W.; Wang, R. Unsupervised linear discriminant analysis for jointly clustering and subspace learning. *IEEE Transactions on Knowledge and Data Engineering* **2019,** *33* (3), 1276-1290.
6. Li, H.; Leung, K. S.; Wong, M. H.; Ballester, P. J. Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular informatics* **2015,** *34* (2-3), 115-126.
7. Lee, J. W.; Maria-Solano, M. A.; Vu, T. N. L.; Yoon, S.; Choi, S. Big data and artificial intelligence (AI) methodologies for computer-aided drug design (CADD). *Biochemical Society Transactions* **2022,** *50* (1), 241-252.