British Journal of Pharmacy and Pharmaceutical Sciences
DOI: 10.32996/bjpps
Journal Homepage: www.al-kindipublisher.com/index.php/bjpps

BJPPS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Machine Learning Approaches to Identify and Optimize Plant-Based Bioactive Compounds for Targeted Cancer Treatments

Fahmida Binte Khair[1], Mohammad Muzahidur Rahman Bhuiyan[2], Mia Md Tofayel Gonee Manik[2], Shafaete Hossain[1], Md Shafiqul Islam[3], Mohammad Moniruzzaman[3], Abu Saleh Muhammad Saimon[4]✉

[1]School of Business, International American University, Los Angeles, CA 90010, USA
[2]College of Business, Westcliff University, Irvine, CA 92614, USA
[3]Department of Computer Science, Maharishi International University, Iowa 52557, USA
[4]Department of Computer Science, Washington University of Science and Technology, Alexandria VA 22314, USA
Corresponding Author: Abu Saleh Muhammad Saimon, **E-mail**: bus.student@wust.edu

| ABSTRACT

Machine learning (ML) represents a breakthrough in drug discovery, markedly increasing efficiency in the search for plant-derived bioactive compounds with anticancer activity. While compounds derived from plants like vincristine and taxol are historical pillars of oncology, the emerging novel therapeutic agents aim to overcome limitations associated with classical therapies, such as toxicity and resistance. Some of the important ML algorithms in this context include decision trees, support vector machines, neural networks, and ensemble learning which allow predictions about bioactivity by managing complicated biological data and determining the effectiveness of different compounds while also optimizing therapeutic profiles. For anticancer compound discovery, supervised as well as unsupervised learning is used whereby activity can be predicted from known properties or compounds just clustered in huge phytochemical databases. Moreover, deep learning models are particularly adept at processing high-dimensional data like multi-omics data and discovering non-linear relationships which furthers our understanding of bioactive compounds at a systems level. While optimizing bioactive compounds, QSAR modeling alongside generative models helps in fine-tuning the molecular design for improved activity and reduced toxicity. ADMET profiling also ensures that the molecules are within the limits of pharmacokinetic and safety standards, thus smoothing out the passage from in silico predictions to experimental validation. The discussion closes with the consideration of some challenges, such as data integration, interpretability of models, computationally intensive tasks, and regulatory demands to be followed versus the promise of the future through cooperative platforms, accessible ML tools together with personalized medicine. Further emphasis is given on the need for continued research interdisciplinary collaborations as well as investments that will help harness the full potential of ML in plant-based anticancer drug discovery to improve treatment outcomes while minimizing adverse effects.

| KEYWORDS

Bioactive Compounds, Cancer, Drug Discovery, Machine Learning

| ARTICLE INFORMATION

ACCEPTED: 15 April 2024 **PUBLISHED:** 28 May 2024 **DOI:** 10.32996/bjpps.2024.1.1.7

1. Introduction

Cancer continues to be one of the top killer diseases around the world, thus prompting an ongoing search for available treatments that are more effective and safer. Historically, plant-based compounds have played a vital role in the treatment of cancer due to their wide range of bioactive properties. The anticancer drugs that have been derived from plants show a promise of therapeutic

potential in natural products. Vincristine and vinblastine, for example, which are used to treat leukemia and lymphoma, respectively, are obtained from Catharanthus roseus (Cragg & Newman, 2005). Another such compound is paclitaxel or taxol, which derives from the Pacific yew tree (*Taxus brevifolia*) and has spectacular efficacy against ovarian and breast cancers (Wani et al., 1971). All these have led to further investigation into plant-based sources as potential sources for new cancer therapies. Despite the success of plant-derived drugs, conventional therapies are often associated with limitations, such as toxicity and drug resistance, that can hinder long-term treatment efficacy (Newman & Cragg, 2016). Many chemotherapy agents are not only aimed at cancer cells but also at normal healthy cells resulting in very uncomfortable side effects that spoil the patients' quality of life under treatment. Besides, drug resistance is a major problem since cancer cells can change and develop mechanisms to counteract the drugs used against them; this makes previously effective medications lose their potency after a short period (Vasan et al., 2019). Thus, owing to these challenges, the search and development for novel phytobioactive compounds continue with high enthusiasm that could provide selective cytotoxicity along with target-specific action and fewer side effects compared to conventional therapies for a more sustainable and effective approach towards cancer management.

However, ML as part of artificial intelligence has dramatically changed contemporary drug discovery by facilitating the speedy accessing and interpreting of enormous datasets, a process that would otherwise be very lengthy and cumbersome with classical approaches. ML comprises numerous algorithms and statistical models that use data to learn, discover unseen patterns, and make predictions without being explicitly programmed for each task (LeCun et al., 2015; Rahaman et al., 2023; Islam et al., 2023). In the context of drug discovery, ML has been applied to predict biological activity, refine lead compounds, and detect possible adverse effects through the mining of complex datasets that include chemical structures, genomic information, and pharmacokinetic profiles (Chen et al., 2018). An explicit advantage of using ML in the drug discovery process is that it can undertake virtual screening. This is a computational approach that quickly evaluates thousands of compounds for biological activity. Virtual screening allows researchers to simulate interactions between natural product compounds and particular cancer targets, thereby predicting binding affinities and ranking the most promising candidates for further experimental evaluation (Stokes et al., 2020). The combination of molecular docking with machine learning algorithms enhances the predictive capability of virtual screening by enabling the identification of compounds with high therapeutic potential and specificity for cancer cells. In addition, machine learning plays an important role in the optimization phase of drug discovery when QSAR models are used to refine structures toward higher activity and lower toxicity as well as improved pharmacokinetics (Cherkasov et al., 2014).

This review applies machine learning in search, optimization, and validation of bioactive compounds from plant origins for targeted therapies against cancer. The overview touches upon some of the crucial ML algorithms used in the process of drug discovery, which include data collection, virtual screening, optimization of compounds and experimental validation. In addition, the review discusses challenges and possible directions in applying ML to plant-based drug discovery focused on the potential of ML to expedite the search process for safer and more effective natural remedies for cancer.

2.0 Machine Learning Techniques in Drug Discovery

2.1 Overview of ML Algorithms Used in Drug Discovery

The advent of machine learning (ML) has become feasible as it provided us with predictive models capable of handling and analyzing complicated biological data specifically for the identification of bioactive compounds. Some of the ML algorithms used are decision trees, support vector machines (SVM), neural networks, and ensemble learning models which play quite an important role in the drug discovery process. In decision trees, trees are widely used as modeling tools, mostly because of their simplicity and interpretability. Such an architecture enables researchers to visualize pathways of decisions that lead to a particular classification or prediction. In the context of modeling bioactivity, decision trees precisely indicate how the characteristics of compounds relate to biological responses (Chen et al., 2018). For the Support Vector Machines (SVM), SVMs are especially powerful in dealing with classification problems and have been extensively used to classify bioactive versus inactive compounds. SVMs determine an optimal hyperplane to differentiate between the compounds based on their molecular descriptors, which makes these classifiers highly Relevant for binary classification problems involved in the drug discovery process (Cortes and Vapnik,1995). In the case of Neural Networks, Deep neural networks are especially appropriate for capturing complex representations, many of which are intrinsically nonlinear. Their capability to learn from vast amounts of data makes them particularly useful for multi-task predictions and also exemplifies the application in toxicity prediction, as deep learning does (LeCun et al., 2015). Also, in Ensemble Learning, Techniques of ensemble learning, like random forests and gradient boosting, use multiple models to make more accurate predictions. This is a particularly robust approach because it minimizes overfitting, a characteristic that is especially sought after in high-dimensional biological data (Breiman 2001). Such algorithms are crucial in highlighting the intrinsic complexity of biological data, predicting the activity of compounds, and optimizing therapeutic potential; all three thus permit the development of plant-derived bioactive compounds with anticancer properties (Vamathevan et al., 2019).

2.2 Supervised and Unsupervised Learning in Bioactive Compound Discovery

Machine Learning models for drug discovery can be broadly categorized into supervised and unsupervised models. In Supervised Learning, the primary approach used to predict the activity of compounds is supervised learning. It relies on an already established database that contains known molecular descriptors and bioactivity outcomes. In supervised learning, random forests, neural networks, and support vector machines play a crucial role in determining anticancer activity by uncovering patterns in the given data set (Hughes et al., 2016). A model purely based on phytochemical information with bioactivities can screen unknown compounds for potential anticancer properties effectively, thus presenting a nice alternative to the screening done in laboratories through traditional methods (Lavecchia, 2015). For the Unsupervised Learning, Applications of unsupervised learning are evident in clustering and pattern recognition, where the method reveals inherent structures in data without any labels. K-means clustering and hierarchical clustering facilitate the grouping of compounds based on structural similarity or bioactivity characteristics, which may assist further in the discovery of new bioactive compounds (Xie et al., 2020). Very efficient unsupervised methods seek to mine huge databases of phytochemicals for clusters of newly identified compounds with potentially therapeutic uses (Zhang et al., 2018).

2.3 Deep Learning and Its Advantages for Complex Data Analysis

One of the important areas of machine learning, deep learning, has recently proven to be highly effective in drug discovery just due to its capability to handle high-dimensional and complex data. Specifically, convolutional neural networks and recurrent neural networks are types of deep learning architectures used in the drug discovery process. Such as Convolutional Neural Networks (CNNs), a crucial virtual screening when dealing with image-based data. In this sense, virtual screening allows obtaining visual representations of the structures of different compounds. CNNs are trained on thousands of images of chemicals, learning patterns in molecular structures that are indicative of bioactivity and thus speeding up and making the screening process more accurate. One of the most widely taught classes is Recurrent Neural Networks (RNNs). This specialized class addresses sequential data, like time series data relevant to pharmacokinetics. RNNs embody dependencies in sequences, which is especially handy for analyzing drug interactions over time or predicting gradual changes in the effect of drugs on cancer cells (Chen et al., 2018). Deep learning models are especially useful in drug discovery, precisely because of their ability to capture non-linear relationships that exist in multi-omics data. This data comprises genomic, transcriptomic, and metabolomic information all integrated with chemical structure information. It hence allows a truly holistic approach to understanding bioactivity by considering how compounds from plant sources interact within biological pathways (Gawehn et al., 2016). Indeed, deep learning models added to the integration of multi-omics for the profiling of compounds interacting with specific targets and therefore relevant toward the pursuit of multi-target anticancer therapies (Zhao & So, 2019).

3.0 Data Collection and Preprocessing for ML Applications

Identifying bioactive phytochemical compounds with anticancer activity requires a comprehensive database that contains information about phytochemicals, genomic profiles, and traditional medicine. In the case of PubChem, PubChem is a widely used repository with vast information regarding the structures, properties, and biological activities of molecules. It allows researchers to find substances that have potential anticancer effects and explore already known interactions with targets related to cancer, especially in the context of bioactivity assays (Kim et al., 2016). ChEMBL is a manually curated bioactivity database for small molecules. Such a database is fundamental in the drug discovery process since it provides information about the activity of compounds against biological targets, pharmacokinetics, and toxicity-all these factors are crucial determinants for ML-driven predictions (Gaulton et al., 2017). For the Phytochemical and Ethnobotanical Databases, these types of databases hold the information about substances used in traditional medicines, hence providing details on plant species that have historical medicinal value. Ethnobotanical data aid in the selection of potential species; it may exhibit anticancer activity and help the researcher to prioritize plants for further investigation (Daniyal & Ahmad, 2015). Such resources are invaluable in ML applications for drug discovery, offering precious datasets on phytochemicals, their chemical properties, and potential therapeutic effects.

However, Data regarding quality and consistency is what supports effective machine learning modeling, and thus data cleaning and preprocessing are needed steps. Missing Data and Normalization: Feature normalization will adjust the scales of the features so that variables, for instance, molecular weight and hydrophobicity become comparable with one another, thus enhancing the model's performance. Inherent in biological datasets like these are missing data that can also be handled through imputation methods such as K-nearest neighbors or mean substitution, ensuring robust ML modeling (Kotsiantis et al., 2006). Overfitting and increased computational expense accompany high-dimensional data. Dimensionality reduction methods, such as PCA and t-SNE, retain the most relevant features of the data, thus making models more interpretable and efficient. While PCA changes the data into principal components that hold most of the variance, t-SNE clusters akin data points—very vital in visualizing complex groupings within massive datasets. These steps of preprocessing form the backbone upon whose accuracy and efficiency the final ML models rely, meaning that algorithms would be able to find interesting patterns in aggregated data. In Molecular Descriptors, the three most popular molecular descriptors used to characterize any chemical compound are molecular weight, the number of hydrogen bond donors, and hydrophobicity. This determination is crucial for assessing a compound's bioactivity, permeability, and drug-like characteristics, all of which relate to anticancer activity (Todeschini & Consonni, 2009). For instance, molecular weight

influences absorption and bioavailability; the number of hydrogen bond donors is pertinent to binding affinity for protein targets, both of which are critically important in anticancer applications (Lipinski, 2004). For the Feature Selection Techniques, a feature selection technique, such as RFE, determines the most relevant features. Such techniques enhance not only the accuracy of the model but also the efficiency of processing models. RFE iteratively removes less informative features, hence refines the dataset which in turn helps the model to concentrate more on significant variables (Guyon et al., 2002). This becomes particularly important in high-dimensional biological datasets where redundant features can mask the performance of a model and also lead to overfitting.

4.0 Machine Learning Approaches for Identifying Bioactive Compounds

4.1 Virtual Screening and Molecular Docking

Virtual screening is a high-throughput screening method used in drug discovery to screen large libraries of compounds for potential biological activity (Figure 1). This was applied by machine learning (ML) algorithms to facilitate the detection of bioactive compounds. Machine learning quickly analyzes chemical libraries and predicts which compounds are likely to have the desired biological effects, thus reducing considerably the need for extensive experimental testing (Stokes et al., 2020). Molecular docking is an integral part of the virtual screening process, especially in the anticancer drug discovery process since it provides an insight into the binding affinity of a given compound against a target protein associated with cancer. In such an approach that imitates molecular interactions, docking scores are able to predict the chances of a compound successfully binding to a protein, which is tremendously important in inhibiting processes collapsed within cancer cells. The AutoDock and PyMOL combination, along with machine learning models, permit scientists to zero in on lead compounds that interact with proteins critical to cancer progression, such as kinases and growth factor receptors (Trott & Olson, 2010). Virtual screening along with molecular docking eliminates potential candidates in the discovery process based on anticipated binding affinities and emphasizes exertional effort on those identified as highly promising candidate's anticancer agents. Therefore, the integration of machine learning with virtual screening enhances the drug discovery pipeline and proves effective in identifying plant-based bioactive compounds exhibiting anticancer properties (Sliwoski et al., 2014).

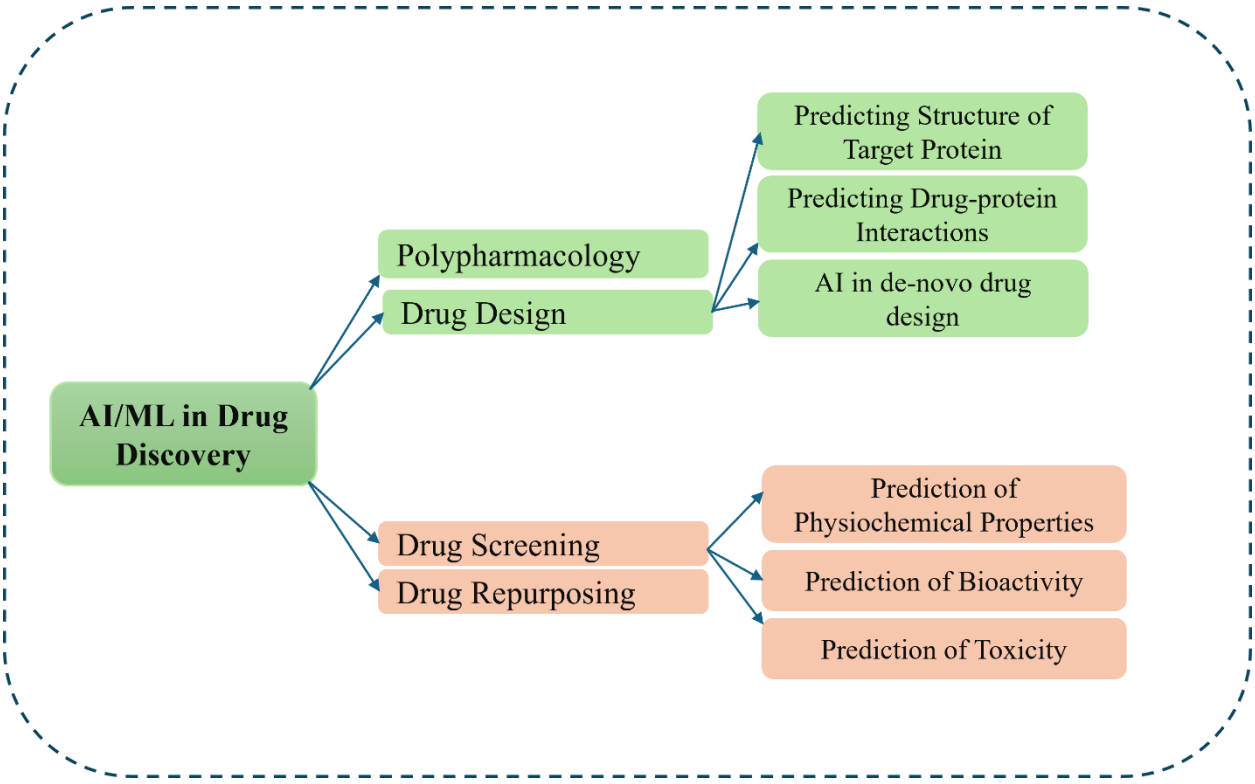

Figure 1. Applications of AI/ML in drug discovery.

4.2 Integration with Omics Data for Multi-Target Drug Discovery

Integration of multi-omics data, including genomics, transcriptomics, and metabolomics, along with phytochemical data is a powerful strategy to find multi-target anticancer agents. The integration of multi-omics enables researchers to study the impact of compounds on cancer cells at several levels and pinpoint targets in diverse biological pathways. Researchers can devise integrative models by combining omics data with ML that will predict the interactions of plant compounds within the biological

systems of cancer cells. Genomics tells about the gene expressions that respond to drug interactions. Transcriptomics show what these drugs do to RNA transcription levels, whereas metabolomics follows up with metabolic changes. Researchers can be supported by multi-omics in finding bioactive compounds that affect signaling pathways in cancer cells and metabolism, which would imply the identification of compounds capable of simultaneously modulating several cancer-related processes (Hasin et al., 2017). The holistic identification and optimization of bioactive compounds with multi-omics and ML in drug discovery can thus be targeted. Such an approach particularly highlights the insight into the synergistic action of compounds at multiple targets, which is invaluable for cancer therapy since intervention through multiple pathways minimizes the chances of drug resistance development and thereby increases therapeutic efficacy (Zhao & So, 2019).

5.0 Challenges and Limitations of ML in Plant-Based Drug Discovery

The performance of machine learning (ML) in drug discovery is critically dependent on data quality. Some of the most prevalent problems that affect models in this setting include data inconsistency, incomplete information, and integration obstacles with heterogeneous datasets. Inconsistency in data among multiple sources arises due to different formats, standards, and measurement methods; this ultimately leads to erroneous predictions (Sadiq & Indulska, 2000). The issue of missing values is a major problem for all biological datasets; gaps in key attributes for instances (like efficacy or toxicity of compounds) introduce bias into model training and predictive accuracy. Besides that, datasets from diverse fields themselves are very heterogeneous regarding their structures and scales of measurements; this calls for complex data harmonization techniques (Zhu et al., 2019). These issues need advanced data preprocessing, like imputation for missing values and standardization to harmonize datasets. Data integration tools, data lakes, or data warehouses also help consolidate and organize large volumes of data to make big data more usable and of higher quality in drug discovery (Kambatla et al., 2014). The more complex many ML models are, and especially deep learning algorithms, the more they challenge to interpret; interpretability is vital for establishing the reliability and trustworthiness of a prediction. Complex models like deep neural networks often operate as "black boxes," hiding their internal decision-making process from researchers. The opacity poses a grave problem in drug discovery: knowing how a model predicts (for instance, that a compound has bioactivity) is necessary for scientific validation of the result (Doshi-Velez & Kim, 2017).

Furthermore, Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been proposed as tools for enhancing interpretability through the explanation of model outputs. However, in particular for deep models, much of their actual decision process is opaque, and furthering the understanding gained from these methods; a central postulate is that interpretability is obtained at the cost of decreasing accuracy in the model being built. This remains a great challenge (Ribeiro et al., 2016). Machine learning, especially deep learning, is very resource-intensive and thus a form of adoption in drug discovery would be greatly impeded by the availability of resources. The training of deep learning models typically requires large datasets and high-performance computing resources like GPUs and TPUs, which are quite expensive and energy-consuming to accumulate (Strubell et al., 2019). Inference in ML models is computationally cheap compared to the training phase. In resource-limited environments, these overheads with computational infrastructure greatly limit the ability to develop and validate ML models productively. Cloud computing and distributed computing architectures are indeed the prevailing solutions for resource scalability. Google Cloud and AWS offer on-demand, flexible computing resources so that researchers can conduct large-scale ML experiments without the upfront investment in physical infrastructure (Cresswell et al., 2018). However, cloud-based resources introduce additional challenges, including data security and long-term costs, thus emphasizing the need for effective and scalable solutions in ML-driven drug discovery. ML in drug discovery raises ethical and regulatory concerns, especially about data privacy, ethical transparency, and adherence to regulatory standards. Data privacy is a critical issue since sensitive health information is often used by ML models to predict drug efficacy and patient responses. Regulations such as the GDPR (General Data Protection Regulation) in the European Union must be complied with to guarantee that personal data is utilized in an accountable and safe manner (Voigt & Von dem Bussche, 2017). The ethical considerations also extend to the question of transparency in algorithmic decision-making and the discrimination that may be inherited by ML models. Regulatory bodies like the FDA are paying closer attention to claims regarding the ethical use of AI and ML in drug discovery, as they put out guidelines stressing data integrity, reproducibility, and validation of models, (FDA, 2018). These regulatory requirements must be met so that drugs predicted by ML can enter clinical trials and obtain market approval. In summary, therefore of these challenges data quality, model interpretability, computational costs and also the ethical and regulatory standards must be in accordance for the successful integration of ML in drug discovery.

6.0 Future Directions and Applications

A significant step forward in exploring the anticancer activity of compounds is the application of integrated multi-omics data, including genomics, proteomics, and metabolomics (Figure 2). The use of multi-omics strategies permits researchers to investigate the intricate interplay between genes, proteins, and metabolites in neoplastic cells, thereby giving an almost complete picture of the mechanisms through which plant-derived substances exert their therapeutic effects (Hasin et al., 2017). Researchers can then zero in on molecular pathways that bioactive compounds influence more precisely to facilitate targeted drug development. For instance, proteomic data reveal protein targets involved in signaling within cancer cells while metabolomic data explain how compounds perturb cellular metabolism; both are critical for developing therapies that target multiple pathways simultaneously

(Zhang et al., 2018). An integrative multi-omics approach allows ML algorithms to capture patterns that would not be possible with a single omics distinction, thus providing a better insight into the efficacy of compounds and possibly uncovering synergistic effects. This can make the anticancer drug discovery much more accurate and could also towards the development of personalized, multi-targeted therapies against cancer (Zhou et al., 2019). It is suggested by collaborative platforms and open-source data sharing that they are important to speed up the research in herbal medicine for drugs. Open-access databases, for example, ChEMBL and PubChem databases, offer abundance of information on compound structures, bioactivities and pharmacological properties that help in establishing collaboration at the international level (Kim et al., 2016). Furthermore, initiatives like Open-Source Drug Discovery (OSDD) provide an opportunity for the researchers from various backgrounds to have equal access to contribution and data so that drug discovery could be democratized and data analysis conducted more thoroughly (Kaur et al., 2014). Such collaborations enhance the availability of information and make it reproducible because open data allow other researchers to confirm results and build upon what has already been done. With the culture of sharing and openness with data, collaborative platforms contribute significantly to plant-based therapeutics, which in turn accelerates the finding and confirmation of bioactive compounds.

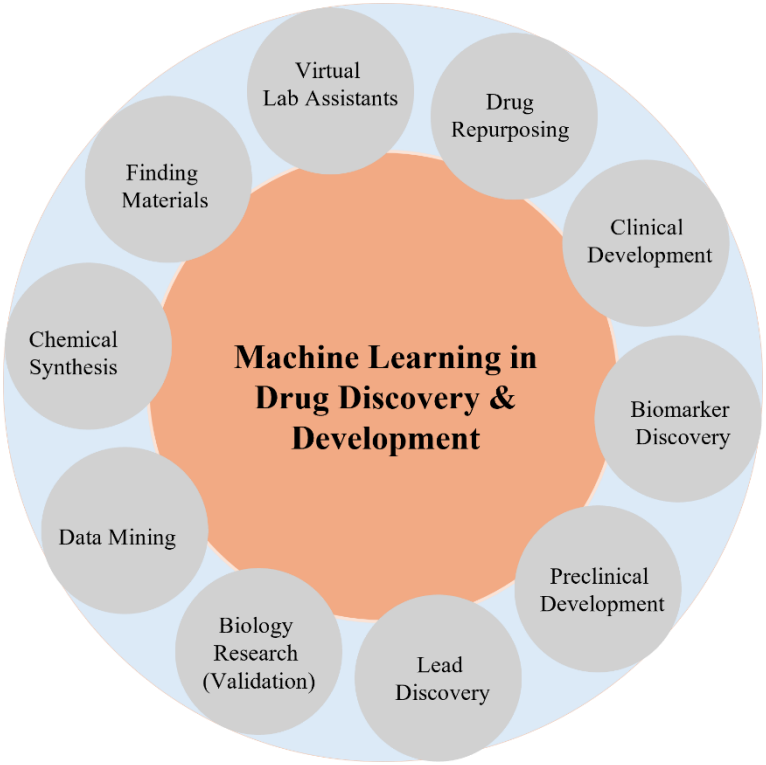

Figure 2. Various fields in drug discovery by using machine learning.

Moreover, the increased information available in the context of drug discovery has created a need for machine learning tools that do not require specialized knowledge. Accessible, user-friendly platforms are needed so that computational analyses become available to those researchers who might be less technically skilled but still have an important role to play in the discovering and analyzing of bioactive compounds. Such platforms as KNIME and Orange, which provide drag-and-drop usability, have proven potentially useful in enabling researchers with no programming knowledge to conduct sophisticated data analysis techniques like clustering and regression (Berthold et al., 2009). More inclusive availability of ML platforms would democratize the computational approaches in drug discovery, allowing botanists, pharmacologists, and clinicians to efficiently handle large data sets. By lowering barriers to entry, these tools facilitate interdisciplinary collaboration, which is crucial for the comprehensive nature of plant-based drug discovery. The promise of an ML-enabled personalized medicine lies in the ability to process enormous volumes of patient data and identify treatments that will be most effective for individuals based on their molecular profiles. This could revolutionize the management of cancer by allowing therapies to be used that are targeted, more effective, and less toxic than the conventional approaches (Ashley, 2016). Among these, plant-based compounds possess unique biological activities and relatively low toxicity; thus, they are especially appropriate for personalized treatment regimens which may lead to more effective and patient-oriented anticancer therapies.

7.0 Conclusion

Bioactive compounds- Specifically, machine learning (ML) has brought tremendous progress in the preliminary discovery of plant-based anticancer agents by facilitating identification, optimization, and validation of bioactive compounds. During the identification phase, ML methods allow compound libraries of millions to be screened quickly instead of traditionally painstakingly screening with laboratories-compound practically predicting the bioactivity of a compound from its structural features and history. Hence, it speeds up the discovery process for plant-derived antimicrobial agents with significant anticancer activity against reducing dependence on traditional time-consuming laboratory screening methods. Modern algorithms such as random forests support vector machines and neural networks enable ML models to learn about bioactivity patterns among plant constituents and thereby flag those with promising therapeutic properties potential. Once promising molecules are identified, machine learning helps in the optimization phase by predicting changes to the structure of the molecule that would increase efficacy and reduce toxicity. In this regard, QSAR modeling and generative models allow scientists to investigate changes in the structure of bioactive molecules, thereby fine-tuning plant-derived compounds for better binding affinity as well as pharmacokinetic profiles. This computational optimization is crucial in the drug development process because it creates a corridor toward developing compounds that will be therapeutic and safe for use. The impact of ML-guided plant-based discoveries on future cancer treatments is profound. As ML algorithms develop and become capable of integrating multi-omics data, the dream of personalized, targeted therapies for cancer just gets more realistic: plant-derived compounds will always be naturally varied in chemistry and can be tailored to target pathways in cancer, thus providing a more specific treatment possibly with fewer side effects than classical chemotherapy. Continuous research and interdisciplinary collaboration are required to fully exploit the potential of ML in plant-based anticancer drug discovery. Investments in the quality of data, computational infrastructure, and accessible ML platforms will further enhance teamwork among biologists, chemists, and data scientists. Thus, it will broaden the scope as well as the depth of machine learning applications in the drug discovery process. This integrated approach will hopefully expedite the emergence of new effective cancer therapies from plant-based compounds as viable options for targeted and personalized treatments against cancer.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Acknowledgement: We would like to express our gratitude to all the co-authors for their contribution and critical reviews from the anonymous reviewers.

ORCID ID:
Fahmida Binte Khair: https://orcid.org/0009-0001-5315-6258
Mohammad Muzahidur Rahman Bhuiyan: https://orcid.org/0009-0001-1774-9726
Mia Md Tofayel Gonee Manik: https://orcid.org/0009-0005-6098-5213
Shafaete Hossain: https://orcid.org/0009-0008-1622-9447
Md Shafiqul Islam: https://orcid.org/0009-0008-9067-498
Mohammad Moniruzzaman: https://orcid.org/0009-0006-5981-4473
Abu Saleh Muhammad Saimon: https://orcid.org/0009-0006-3147-1755

References

[1] Ashley, E. A. (2016). Towards precision medicine. Nature Reviews Genetics, 17(9), 507–522.
[2] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... & Thiel, K. (2009). KNIME: The Konstanz Information Miner. ACM SIGKDD Explorations Newsletter, 11(1), 26–31.
[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[4] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241–1250.
[5] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry, 57(12), 4977–5010.
[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
[7] Cragg, G. M., & Newman, D. J. (2005). Plants as a source of anti-cancer agents. Journal of Ethnopharmacology, 100(1-2), 72-79.
[8] Cresswell, K. M., Cunningham-Burley, S., & Sheikh, A. (2018). Health care robotics: Qualitative exploration of key challenges and future directions. Journal of Medical Internet Research, 20(12), e11017.
[9] Daniyal, M., & Ahmad, S. (2015). Herbal medicines for the treatment of cancer in the light of molecular docking. Pharmacognosy Reviews, 9(17), 89–97.
[10] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[11] FDA. (2018). Oncology drug products used with in vitro companion diagnostic devices: Regulatory requirements. Food and Drug Administration. Available from: [FDA Website](https://www.fda.gov)
[12] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Davies, M., & Overington, J. P. (2017). The ChEMBL database in 2017. Nucleic Acids Research, 45(D1), D945–D954.
[13] Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. Nature Reviews Drug Discovery, 15(3), 131–147.

[14] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1), 389–422.
[15] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Cell, 169(6), 1177–1190.
[16] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2016). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239–1249.
[17] Islam, M. R., Rahaman, M. M., Bhuiyan, M. M. R., & Aziz, M. M. (2023). Machine learning with health information technology: Transforming data-driven healthcare systems. Journal of Medical and Health Studies, 4(1), 89–96.
[18] Kaur, H., Kumar, V., & Singh, R. (2014). Open Source Drug Discovery: Collaborative platform for affordable and accessible drug discovery. Drug Discovery Today, 19(7), 892–899.
[19] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., … & Bolton, E. E. (2016). PubChem in 2017: New data content and improved web interfaces. Nucleic Acids Research, 45(D1), D955–D963.
[20] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Handling imbalanced datasets: A review. GMDH Shell, 20(1), 25-32.
[21] Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. Drug Discovery Today, 20(3), 318–331.
[22] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[23] Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies, 1(4), 337–341.
[24] Rahaman, M. M., Rani, S., Islam, M. R., & Bhuiyan, M. M. R. (2023). Machine Learning in Business Analytics: Advancing Statistical Methods for Data-Driven Innovation. Journal of Computer Science and Technology Studies, 5(3), 104-111.
[25] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135–1144).
[26] Sadiq, S., & Indulska, M. (2000). Open and Distance Learning: Trends, Policy and Strategy Considerations. Higher Education Policy, 13(1), 49-70.
[27] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacology & Therapeutics, 145, 1–13.
[28] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Brown, E. D. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688–702.e13.
[29] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.
[30] Todeschini, R., & Consonni, V. (2009). Molecular descriptors for chemoinformatics (Vol. 41). Wiley-VCH.
[31] Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 31(2), 455–461.
[32] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., & Zhao, J. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477.
[33] Vasan, N., Baselga, J., & Hyman, D. M. (2019). A view on drug resistance in cancer. Nature, 575(7782), 299–309.
[34] Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P., & McPhail, A. T. (1971). Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. Journal of the American Chemical Society, 93(9), 2325–2327.
[35] Zhang, B., Horvath, S., & Chen, Y. (2018). Genomic and proteomic insights into signaling and metabolic pathways in liver cancer. Cell Metabolism, 27(1), 1–13.
[36] Zhao, Q., & So, H. C. (2019). Drug repositioning for schizophrenia and depression/anxiety disorders: A machine learning approach leveraging expression data. Scientific Reports, 9(1), 1–12.
[37] Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2019). Human symptoms–disease network. Bioinformatics, 35(8), 1128–1135.
[38] Zhu, Y., Stephens, R., & Meltzer, P. S. (2019). CervixScan: radiomic analysis to predict patient response to radiation therapy in locally advanced cervical cancer. Nature Biotechnology, 37(12), 1458–1470.