AUCTORES
Globalize your Research

Biomedical Research and Clinical Reviews

Julian Lloyd Bruce *

Open Access

Research Article

An Overview of Unsupervised Machine Learning in Drug Discovery, Pharmacogenomics, and Pharmacovigilance

Julian Lloyd Bruce

1101 30th Street NW Suite #500 (Fifth Floor), Washington, D.C. 20007, United States.

***Corresponding Author:** Julian Lloyd Bruce, 1101 30th Street NW Suite #500 (Fifth Floor), Washington, D.C. 20007, United States.

Received Date: June 27, 2025; Accepted Date: July 31, 2025; Published Date: September 05, 2025

Citation: Julian L. Bruce, (2025), An Overview of Unsupervised Machine Learning in Drug Discovery, Pharmacogenomics, and Pharmacovigilance, *J. Biomedical Research and Clinical Reviews*, 11(2); **DOI:10.31579/2690-4861/220**

Abstract

As biomedical data continue to grow in complexity and often lack annotation, unsupervised machine learning (UML) has emerged as a powerful approach for uncovering hidden patterns and enabling data-driven discovery in pharmacology. This review examines the expanding role of UML in drug discovery, pharmacogenomics, and pharmacovigilance. Core techniques such as clustering, dimensionality reduction, and generative modeling support hypothesis-free analysis and contribute to translational insight. In drug discovery, UML enhances molecular representation, facilitates lead optimization, and guides scaffold development. Graph neural networks (GNNs) are emphasized for their ability to capture complex structural and chemical features that improve drug–target interaction prediction and de novo molecular generation. In clinical contexts, UML enables stratification of pharmacogenomic profiles and supports early detection of adverse drug events through anomaly detection and natural language processing applied to real-world data. By integrating insights from both algorithmic development and real-world applications, this review underscores the growing value of UML as a pivotal tool in advancing contemporary pharmacological research and data-driven decision-making.

Key words: artificial intelligence; drug discovery; pharmacogenomics; pharmacovigilance; computational biology

Introduction

As biomedical datasets continue to expand in size, complexity, and diversity, unsupervised machine learning (UML) has become an increasingly important tool in pharmacological research. These datasets originate from a range of sources including chemical compound libraries, multi-omics platforms, electronic health records (EHRs), and adverse event databases. Traditional supervised approaches often require extensive annotation and are limited by labeling biases or incomplete clinical outcomes. In contrast, UML algorithms operate without labeled outcomes, enabling autonomous identification of latent structures, statistical regularities, and hidden relationships that support hypothesis generation and translational insight.

This review examines the role of UML across three major domains of pharmaceutical science: drug discovery, pharmacogenomics, and pharmacovigilance. In drug discovery, clustering, dimensionality reduction, and generative modeling techniques improve molecular representation, facilitate scaffold identification, and accelerate lead optimization [1]. Special attention is given to graph neural networks (GNNs), which capture both local and global chemical structure features and enhance predictions of drug–target interactions as well as de novo compound generation. In pharmacogenomics, UML supports the stratification of patients by identifying molecular subgroups and integrating genomic, transcriptomic, and epigenomic signals to inform individualized dosing and therapeutic selection. In pharmacovigilance, UML enables the detection of latent safety signals using techniques such as anomaly detection and natural language processing, applied to large-scale structured and unstructured data from clinical and post-market surveillance settings [2].

By analyzing recent methodological developments and applied case studies, this review underscores UML's emergence as both a foundational methodology and a practical framework for uncovering biologically meaningful patterns in high-dimensional, unlabeled datasets. Its flexibility and scalability position UML as a central component of next-generation pharmacological research and data-driven clinical innovation.

Methods

A targeted literature review was performed using Google Scholar, PubMed, and IEEE Xplore to identify peer-reviewed articles relevant to unsupervised machine learning in drug discovery, pharmacogenomics, and pharmacovigilance. The search was restricted to publications from the past five years to maintain contemporary relevance. Keyword combinations such as "unsupervised learning," "graph neural networks," "pharmacogenomics," and "pharmacovigilance" were applied to capture both methodological advancements and translational applications. Articles were screened based on their abstracts, methodological integrity, and relevance to data-driven pharmaceutical research.

Methodology of Unsupervised Machine Learning

At its core, UML seeks to uncover latent structures, statistical regularities, or manifold representations within datasets that lack explicit labels or target variables. Unlike supervised machine learning (SML), which learns a function $f: X \rightarrow Y$ by minimizing error between predicted outputs and known ground truth labels Y, UML operates solely on the input space X [3]. This distinction is critical: whereas SML is constrained by the availability and quality of labeled data, UML autonomously infers patterns, enabling discovery in settings where annotation is scarce, noisy, or infeasible—such as high-dimensional omics data, unlabeled chemical libraries, or unstructured clinical narratives. UML algorithms identify clusters, reduce dimensionality, or learn generative representations that expose underlying biological heterogeneity, facilitate hypothesis generation, and augment downstream tasks like drug repurposing or patient stratification without relying on prior assumptions or labeling biases [4].

Key Components of UML

To operationalize these capabilities, UML relies on a structured pipeline composed of interdependent stages. Each component contributes to the model's ability to uncover meaningful representations in complex biomedical data. The key elements include:

1. Data Preparation & Feature Extraction: As with any machine learning pipeline, UML begins with rigorous preprocessing that includes cleaning, normalization, and encoding of datasets to ensure compatibility and reduce noise. In the absence of labeled outcomes, domain-informed feature engineering becomes especially important for preserving the biological, chemical, or clinical relevance of the data, particularly in high-dimensional omics or chemical descriptor spaces.

2. Algorithm Selection: The choice of algorithm depends on the data modality and the intended outcome. For example, clustering algorithms such as K-means, DBSCAN, and hierarchical clustering are used to uncover latent groupings in molecular datasets. Dimensionality reduction techniques like PCA, t-SNE, and UMAP help simplify complex feature spaces, making them more interpretable for visualization and structure discovery. Generative models, including variational autoencoders (VAEs), GANs, and transformer-based architectures, learn underlying data distributions to generate novel molecular or clinical representations.

3. Learning Process: UML models optimize internal parameters without supervision by maximizing intra-cluster similarity (e.g., in clustering) or minimizing reconstruction loss (e.g., in autoencoders). Since there is no explicit target variable, performance depends on how well the model captures structure, reduces entropy, or reconstructs meaningful latent spaces. In chemical modeling, this might manifest as embedding compounds into latent vectors that preserve pharmacophoric features.

4. Validation & Interpretability: Evaluating UML outcomes often requires indirect metrics such as silhouette score (clustering cohesion), reconstruction error (autoencoders), or visualization of latent spaces. Cross-validation is often complemented by expert annotation or alignment with known biological pathways. Increasingly, tools like SHAP or attention mechanisms are used to render UML models more transparent and biologically interpretable.

5. Deployment & Integration: Once trained, UML models are embedded into larger pipelines for downstream tasks—virtual screening, biomarker discovery, or pharmacovigilance signal detection. These models can be iteratively refined using updated datasets or combined with supervised components in hybrid architectures, enabling continuous learning and scalability in dynamic biomedical environments [3][4][5].

Summary of Existing Literature

Recent advances underscore the growing utility of UML in modeling molecular interactions, optimizing lead compounds, and enhancing post-market drug safety surveillance. For example, Mena-Yedra et al. introduced ALMERIA, a decision-support tool that estimates compound similarity and predicts molecular activity by accounting for conformational variability across large chemical libraries. Implemented with scalable infrastructure, ALMERIA demonstrated exceptional performance on the DUD-E benchmark dataset, achieving ROC AUC scores of 0.99, 0.96, and 0.87 across various partitions. Notably, the study emphasized the model's generalization capacity and interpretability, using SHAP analysis to elucidate feature contributions—an essential step toward transparent AI in drug discovery [6].

Complementing this, Yin et al. developed DeepDrug, a unified deep learning framework that integrates residual graph convolutional networks (Res-GCNs) and convolutional neural networks (CNNs) to learn both structural and sequential representations of drugs and proteins. DeepDrug outperformed state-of-the-art models across multiple tasks, including binary and multi-label classification of drug–drug and drug–target interactions. Beyond predictive accuracy, the authors applied DeepDrug to the DrugBank database for drug repurposing, identifying top-ranked candidates against SARS-CoV-2—seven of which had independent support for potential efficacy [7]. This highlights the model's translational potential in real-world therapeutic contexts.

Polanski's review offers a broader conceptual lens, tracing the evolution of UML in cheminformatics from early self-organizing maps (SOMs) to modern deep chemistry paradigms. He argues that UML excels not only in identifying chemically intuitive features but also in uncovering latent molecular patterns beyond human perception. The review underscores the promise of deep unsupervised architectures in scaffold hopping, feature learning, and molecular representation, while also noting current limitations—particularly the scarcity of high-quality, labeled chemical property data. Polanski concludes that while deep chemistry is still maturing, UML is poised to play a pivotal role in bridging data-driven discovery with rational drug design [8].

Further illustrating the breadth of UML applications in drug discovery, Zhang et al. introduced CASTELO, a hybrid framework that integrates machine learning with molecular modeling to streamline lead optimization workflows. By leveraging contact matrices derived from molecular dynamics simulations and encoding temporal dynamics through convolutional variational autoencoders (CVAEs), CASTELO identifies submolecular "hot spots" for chemical modification without requiring extensive structure–activity relationship data. The study demonstrated that CVAE-based clustering outperformed conventional methods in ranking atom subtypes for optimization, offering medicinal chemists a data-driven strategy to enhance potency while reducing development time [9].

In the realm of post-translational modification prediction, Luo et al. developed DeepPhos, a deep learning model tailored to identify protein phosphorylation sites with high accuracy. Unlike traditional predictors reliant on handcrafted features, DeepPhos employs densely connected convolutional blocks to capture hierarchical sequence representations. The model supports both general and kinase-specific predictions, outperforming existing tools across multiple benchmarks. Its architecture enables nuanced detection of phosphorylation motifs, which are critical for understanding signaling cascades and for designing kinase-targeted therapeutics. This underscores the value of deep unsupervised representations in proteomics [10].

Expanding into clinical informatics, Miotto et al. proposed Deep Patient, an unsupervised deep feature learning framework that generates patient-level embeddings from EHRs using stacked denoising autoencoders. Trained on

data from over 700,000 patients, Deep Patient captured latent health representations that significantly improved disease prediction across 78 conditions, including schizophrenia, diabetes, and various cancers. The model's ability to generalize across diverse clinical domains highlights the translational power of UML in precision medicine, enabling early risk stratification and personalized care strategies [11].

Offering a broader lens on machine learning's role in pharmaceutical R&D, Ibáñez Antolín emphasized the importance of unsupervised techniques—particularly clustering and dimensionality reduction—in early-stage applications such as target identification, biomarker discovery, and digital pathology analysis. The study highlighted that machine learning workflows often begin with extensive data preprocessing, where UML plays a pivotal role in revealing latent structure within high-dimensional omics and imaging datasets. This exploratory capability is especially valuable when labeled data are limited or incomplete, reinforcing the utility of unsupervised methods in hypothesis generation and mechanistic insight across translational research domains [1].

Finally, Vamathevan et al. provided a comprehensive review of ML applications across the drug development pipeline, highlighting UML's contributions to target validation, compound screening, and clinical trial optimization. The authors acknowledged challenges such as interpretability and reproducibility but emphasized that, when paired with high-quality data, UML can reduce attrition rates and accelerate decision-making. Notably, the review underscored the synergy between UML and supervised learning in hybrid models, advocating for integrative approaches that combine predictive power with biological plausibility [2].

Innovative Approaches: UML for Lead Optimization

The rapid advancement and adoption of UML in lead optimization reflects its growing utility in navigating high-dimensional chemical space without reliance on labeled potency data. By uncovering latent structural and physicochemical patterns through clustering, dimensionality reduction, and representation learning, UML enables the identification of scaffold relationships, substituent effects, and bioisosteric transformations that might otherwise remain obscured. This is particularly advantageous in early-stage optimization, where empirical activity data are often sparse or noisy. When integrated with graph-based encodings or generative frameworks, UML serves as a foundational layer for downstream predictive modeling, offering a scalable, hypothesis-free strategy for rational compound refinement [12].

In her chapter on AI-driven drug development, Ashenden frames lead optimization as a multidimensional balancing act—one that extends beyond potency to include solubility, metabolic stability, and toxicity [13]. She outlines how AI, including unsupervised methods, can integrate heterogeneous datasets such as high-throughput screening results, physicochemical descriptors, and ADMET profiles. This integration enables medicinal chemists to prioritize chemical modifications that enhance drug-likeness while minimizing downstream attrition. Ashenden emphasizes that UML is particularly well-suited for identifying liabilities early in the pipeline, where labeled outcomes are often unavailable or incomplete [13].

A broader methodological perspective is offered by Li et al., who review the landscape of machine learning-based scoring functions (ML-SFs) for structure-based lead optimization. While their primary focus is on supervised models, they highlight key limitations of classical scoring functions and emphasize the need for approaches that generalize across diverse protein–ligand complexes. The authors argue that unsupervised pretraining, including methods such as autoencoders, graph embeddings, and contrastive learning, can significantly improve the quality of molecular representations used in predictive tasks. Their analysis supports a hybrid modeling framework in which UML-derived features form the foundation for more accurate and transferable ML-SFs [14].

A more targeted application of unsupervised spatial learning can be seen in the development of DeltaDelta, a deep 3D convolutional neural network introduced by Jiménez-Luna et al. to rank congeneric ligands based on predicted potency differences [15]. The model architecture incorporates an unsupervised pretraining phase that learns spatial features from protein–ligand complexes, which are then fine-tuned using potency labels. To evaluate its performance, the researchers conducted one of the largest blind assessments to date, involving over 3,000 ligands and 13 targets sourced from Janssen, Pfizer, and Biogen. Across these datasets, DeltaDelta consistently outperformed traditional docking-based methods in ranking accuracy. The study highlights how embedding unsupervised spatial representation into lead optimization pipelines can produce models that are both predictive and broadly generalizable across diverse chemical series [15].

In the domain of fragment-based design, Green and Durrant introduced DeepFrag, a deep convolutional neural network trained on protein–ligand complexes with systematically removed fragments. The model learns to predict fragment vectors that complement the receptor environment, effectively identifying fragment–receptor compatibility patterns without explicit supervision. Benchmarking revealed that DeepFrag could recover the correct fragment from a library of over 6,500 candidates approximately 58% of the time. Even when the exact fragment was not retrieved, top-ranked alternatives were often chemically similar and synthetically tractable. The authors also released an open-source browser-based implementation, democratizing access to AI-assisted fragment elaboration and accelerating hypothesis generation in structure-guided design [16].

Visualizing Drug Data: The Role of Graph Neural Networks

Graph neural networks (GNNs) have become key technology in molecular representation learning, offering a flexible and expressive architecture for encoding chemical structures as graphs. In this framework, atoms are treated as nodes and covalent bonds as edges, allowing GNNs to capture both local atomic environments and long-range topological dependencies. This capability supports the modeling of stereoelectronic effects, conformational dynamics, and functional group connectivity, all of which are essential for accurate prediction of pharmacological properties. Unlike traditional descriptor-based models that depend on handcrafted molecular fingerprints, GNNs learn task-specific representations directly from graph topology through iterative message-passing mechanisms. This approach has demonstrated strong performance across a range of applications, including drug–target interaction (DTI) prediction, binding affinity estimation, molecular property regression, and de novo molecular generation [12].

A comparative study by Jiang et al. evaluated the performance of four descriptor-based models (SVM, XGBoost, RF, DNN) and four graph-based models (GCN, GAT, MPNN, Attentive FP) across 11 public datasets covering endpoints such as solubility, toxicity, and ADME properties. While descriptor-based models generally outperformed GNNs on smaller datasets, Attentive FP and MPNN demonstrated superior performance in multi-task and large-scale settings. The authors concluded that GNNs offer complementary advantages in capturing structural nuances and recommended their integration into hybrid modeling pipelines [17].

To enhance predictive robustness, Bongini et al. developed FP-GNN, a hybrid architecture that fuses molecular fingerprints with graph-based embeddings. Evaluated on 13 public datasets and 14 phenotypic screening datasets, FP-GNN consistently outperformed both traditional machine learning and deep learning baselines. The model also demonstrated resilience to noise and improved interpretability, suggesting that combining topological and fingerprint-derived features enhances generalizability in real-world drug discovery scenarios [18].

In the generative modeling space, Xiong et al. explored the use of GNNs for de novo molecular generation by integrating them with reinforcement learning and generative frameworks. Their approach enabled the generation of chemically valid, synthetically accessible molecules optimized for properties such as binding affinity and logP. Compared to SMILES-based models, GNN-based architectures achieved higher validity, novelty, and property alignment—highlighting their suitability for scaffold-constrained and multi-objective optimization in early-stage design [19].

For DTI prediction, Zhang et al. reviewed recent GNN-based models that incorporate attention mechanisms, multi-modal embeddings, and 3D structural information. These architectures outperform traditional sequence-based models by capturing spatial and topological dependencies critical for accurate interaction prediction. The authors emphasized that GNNs, particularly when combined with protein structure data, offer scalable and interpretable solutions for virtual screening and drug repurposing [20].

A more targeted application was proposed by Wang et al., who introduced DataDTA—a dual-graph GNN framework for predicting drug–target binding affinities. The model integrates molecular graphs with predicted protein pocket descriptors and sequence embeddings, using a dual-interaction aggregation strategy to capture both intra- and inter-molecular interactions. On benchmark datasets, DataDTA achieved a concordance index of 0.806 and a Pearson correlation of 0.814, outperforming several state-of-the-art baselines. These results underscore the value of combining structural and sequence-level features in affinity prediction [21].

Looking ahead, Abate et al. offered a forward-looking perspective on conditional de novo drug design using GNNs. Their review emphasized the importance of conditioning mechanisms, such as scaffold constraints or pharmacological profiles, to guide molecular generation toward specific objectives. Advances in graph-based generative models now support multi-objective optimization across drug-likeness, synthetic accessibility, and target specificity. These developments position GNNs as a scalable and customizable platform for rational drug design [22].

Stratifying Complexity: UML in Pharmacogenomics

The variability in drug response across individuals, shaped by genomic variation, epigenetic regulation, and environmental exposures, remains a central challenge in precision medicine. UML offers a data-driven framework for addressing this complexity by revealing latent structure in high-dimensional pharmacogenomic datasets without relying on predefined phenotypic labels. Unlike supervised models that require annotated outcomes such as therapeutic efficacy or adverse events, UML algorithms autonomously identify patterns in genomic variants, transcriptomic signatures, and pharmacokinetic trajectories. This enables the stratification of patients into molecularly coherent subgroups, the discovery of novel biomarkers, and the elucidation of genotype–phenotype relationships that may otherwise remain obscured. Techniques such as clustering, dimensionality reduction, and autoencoder-based representation learning are particularly well suited for integrating multi-omics data and generating biologically grounded hypotheses to inform individualized treatment strategies [12].

A widely accepted perspective on this approach is offered by Kalinin et al., who explored the use of deep unsupervised architectures such as autoencoders and deep belief networks to learn hierarchical representations from heterogeneous data sources, including gene expression, epigenetic marks, and electronic health records (EHRs). Their work demonstrated that these representations enhance downstream tasks such as drug response prediction and adverse event forecasting, particularly through the identification of regulatory variants in noncoding regions. The authors argue that unsupervised machine learning will play a central role in the development of scalable and interpretable frameworks that enable personalized medication selection and dosing strategies [23].

In a complementary study, Lautier et al. applied clustering algorithms to pharmacokinetic time-series data, specifically plasma concentration–time curves, to identify latent subgroups of drug metabolism. Using methods such as k-means and hierarchical clustering, they showed that UML could recover clinically meaningful pharmacokinetic phenotypes, including fast and slow metabolizers, without prior knowledge of covariates or outcomes. Their case study involving 250 PK curves demonstrated that unsupervised clustering could independently validate pharmacogenomic findings. This suggests its utility in guiding individualized dosing regimens for drugs with narrow therapeutic indices [24].

Expanding the scope to healthcare operations, Lopez et al. developed a hybrid framework that integrates UML with process mining and discrete-event simulation to model patient flow and treatment trajectories across clinical settings. By analyzing EHR-derived event logs, they identified common care pathways and deviations that may influence drug response and safety. Their approach enables simulation of treatment outcomes under varying protocols, offering a systems-level perspective on how molecular subtypes interact with real-world clinical workflows. This integration of UML with operational modeling highlights its potential to bridge molecular stratification with actionable clinical decision-making [25].

Uncovering Latent Safety Signals: UML in Pharmacovigilance

The growing complexity of pharmacotherapy, driven by demographic aging, multimorbidity, and widespread polypharmacy, has revealed critical limitations in conventional pharmacovigilance systems. These systems often depend on static alert thresholds, spontaneous reporting, and predefined rule sets, which may fail to capture emerging or context-specific safety concerns. UML introduces a data-centric alternative by enabling the detection of latent safety signals within large-scale, unlabeled clinical datasets. By modeling the probability distributions of high-dimensional sources such as EHRs, prescription logs, and adverse event registries, UML algorithms can autonomously identify anomalous prescribing behaviors, rare adverse drug reactions (ADRs), and systemic deviations in medication use [12]. Techniques such as one-class support vector machines (OCSVMs), isolation forests, and density-based clustering are particularly effective for detecting these outliers, especially when configured to incorporate patient-level covariates like renal function, age, and comorbidity burden. When combined with natural language processing (NLP), UML further extends its utility to unstructured data sources including clinical notes, discharge summaries, and social media content [4][5]. This integration supports a multi-modal, scalable pharmacovigilance framework that is better equipped to respond to evolving safety risks in real-world settings.

A compelling demonstration of this approach comes from Nagata et al., who applied OCSVMs to detect overdose and underdose prescriptions across 21 commonly used drugs using EHR data from Kyushu University Hospital. Each model was trained on three patient-specific features—age, weight, and prescribed dose—and evaluated against both real-world and synthetic dosing errors. The OCSVMs successfully identified 87.1% of clinically confirmed dosing anomalies and achieved high F1-scores on synthetic test sets (0.973 for overdose, 0.839 for underdose). Comparative analysis with other anomaly detection methods confirmed the superior precision and recall of the OCSVM approach, underscoring its potential as a real-time safeguard in high-risk prescribing environments [26].

Expanding the methodological scope, Yap advocates for integrating UML across diverse pharmacovigilance data streams, including EHRs, spontaneous reporting systems, and social media platforms. His work emphasizes the synergy between UML and NLP, particularly in mining unstructured text for early detection of rare ADRs and drug–drug interactions. By combining clustering, anomaly detection, and sentiment analysis, Yap proposes a hybrid framework that enhances signal detection while preserving clinical interpretability. He also stresses the importance of regulatory alignment and domain expertise to ensure that UML-generated insights are both actionable and trustworthy [27].

From an ethical and operational standpoint, Zou highlights the challenges of deploying UML in clinical pharmacovigilance. His analysis calls for transparent model design, clinician engagement, and robust data governance to mitigate risks such as alert fatigue, algorithmic bias, and overfitting. Zou also notes that UML can improve pharmacovigilance efficiency by filtering noise and prioritizing high-risk signals in large-scale surveillance systems—an essential capability in resource-constrained healthcare settings [28].

Finally, a systems-level lens was applied by Basile et al., who examined how UML can be used to link multi-omics datasets with clinical phenotypes to uncover mechanistic insights into ADRs. Their approach employed dimensionality reduction and clustering to delineate patient subpopulations

with differential drug responses, providing a foundation for hypothesis generation, biomarker discovery, and targeted safety assessments. Basile's work highlights UML not only as a detection engine but as a strategic tool for shaping future pharmacovigilance and regulatory science efforts [29].

Conclusions

Unsupervised machine learning (UML) is transforming pharmacological research by uncovering latent structure within complex, unlabeled datasets. This review has illustrated how UML contributes to drug discovery through scaffold identification and lead optimization with graph neural networks, supports patient stratification in pharmacogenomics using deep representations, and facilitates the detection of safety signals in pharmacovigilance by leveraging anomaly detection and natural language processing. These methods enhance molecular design, personalized medicine, and real-world drug safety monitoring. Although interpretability and regulatory challenges remain, UML offers a foundational approach for data-driven discovery in environments where annotation is limited or unavailable.

References

1. Antolín, I. A. (n.d.). Applications of machine learning in drug discovery and development. MDPI.
2. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463–477.
3. Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. International Journal of Computing and Digital Systems, 13(1), Article 172
4. Eckhardt, C. M., Madjarova, S. J., Williams, R. J., Ollivier, M., Karlsson, J., Pareek, A., & Nwachukwu, B. U. (2023). Unsupervised machine learning methods and emerging applications in healthcare. Knee Surgery, Sports Traumatology, Arthroscopy, 31(2), 376–381.
5. Pantanowitz, L., Pearce, T., Abukhiran, I., Hanna, M., Wheeler, S., Soong, T. R., Tafti, A. P., Pantanowitz, J., Lu, M. Y., Mahmood, F., & Gu, Q. (2024). Non-generative artificial intelligence (AI) in medicine: Advancements and applications in supervised and unsupervised machine learning. Modern Pathology.
6. Mena-Yedra, R., López Redondo, J., Pérez-Sánchez, H., & Martinez Ortigosa, P. (2024). ALMERIA: Boosting pairwise molecular contrasts with scalable methods. Informatica, 35(3), 617–648.
7. Yin, Q., Fan, R., Cao, X., Liu, Q., Jiang, R., & Zeng, W. (2023). DeepDrug: A general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction. Quantitative Biology, 11(3), 260–274.
8. Polanski, J. (2022). Unsupervised learning in drug design: From self-organization to deep chemistry. International Journal of Molecular Sciences, 23(5), 2797.
9. Zhang, L., Domeniconi, G., Yang, C. C., Kang, S. G., Zhou, R., & Cong, G. (2021). CASTELO: Clustered atom subtypes aided lead optimization—A combined machine learning and molecular modeling method. BMC Bioinformatics, 22, Article 1.
10. Luo, F., Wang, M., Liu, Y., Zhao, X. M., & Li, A. (2019). DeepPhos: Prediction of protein phosphorylation sites with deep learning. Bioinformatics, 35(16), 2766–2773.
11. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094.
12. Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. Artificial Intelligence Review, 55(3), 1947–1999
13. Ashenden, S. K. (2021). Lead optimization. In The era of artificial intelligence, machine learning, and data science in the pharmaceutical industry (pp. 103–117). Academic Press
14. Li, H., Sze, K. H., Lu, G., & Ballester, P. J. (2020). Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdisciplinary Reviews: Computational Molecular Science, 10(5), e1465.
15. Jiménez-Luna, J., Pérez-Benito, L., Martínez-Rosell, G., Sciabola, S., Torella, R., Tresadern, G., & De Fabritiis, G. (2019). DeltaDelta neural networks for lead optimization of small molecule potency. Chemical Science, 10(47), 10911–10918.
16. Green, H., & Durrant, J. D. (2021). DeepFrag: An open-source browser app for deep-learning lead optimization. Journal of Chemical Information and Modeling, 61(6), 2523–2529.
17. Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. Journal of Cheminformatics, 13, Article 1.
18. Bongini, P., Bianchini, M., & Scarselli, F. (2021). Molecular generative graph neural networks for drug discovery. Neurocomputing, 450, 242–252.
19. Xiong, J., Xiong, Z., Chen, K., Jiang, H., & Zheng, M. (2021). Graph neural networks for automated de novo drug design. Drug Discovery Today, 26(6), 1382–1393.
20. Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., & Li, X. (2022). Graph neural network approaches for drug-target interactions. Current Opinion in Structural Biology, 73, 102327.
21. Wang, C., Kumar, G. A., & Rajapakse, J. C. (2025). Drug discovery and mechanism prediction with explainable graph neural networks. Scientific Reports, 15(1), 179.
22. Abate, C., Decherchi, S., & Cavalli, A. (2023). Graph neural networks for conditional de novo drug design. Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(4), e1651
23. Kalinin, A. A., Higgins, G. A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I. D., Najarian, K., & Athey, B. D. (2018). Deep learning in pharmacogenomics: From gene regulation to patient stratification. Pharmacogenomics, 19(7), 629–650.
24. Lautier, J. P., Grosser, S., Kim, J., Kim, H., & Kim, J. (2024). Clustering plasma concentration-time curves: Applications of unsupervised learning in pharmacogenomics. Journal of Biopharmaceutical Statistics, Advance online publication.
25. Lopez, C., Tucker, S., Salameh, T., & Tucker, C. (2018). An unsupervised machine learning method for discovering patient clusters based on genetic signatures. Journal of Biomedical Informatics, 85, 30–39.
26. Nagata, K., Tsuji, T., Suetsugu, K., Muraoka, K., Watanabe, H., Kanaya, A., Egashira, N., & Ieiri, I. (2021). Detection of overdose and underdose prescriptions—An unsupervised machine learning approach. PLOS ONE, 16(11), e0260315.
27. Yap, K. Y. (2017). A pharmaco-cybernetics approach to patient safety: Identifying adverse drug reactions through unsupervised machine learning. In Healthcare ethics and training: Concepts, methodologies, tools, and applications (pp. 1291–1310). IGI Global.
28. Zou, C. (2018). Analyzing research trends on drug safety using topic modeling. Expert Opinion on Drug Safety, 17(6), 629–636.

29. Basile, A. O., Yahi, A., & Tatonetti, N. P. (2019). Artificial intelligence for drug toxicity and safety. Trends in Pharmacological Sciences, 40(9), 624–635.

DOI:10.31579/2692-9406/220