**Insights and Perspectives on Machine Learning Applications in Chemoinformatics: A Review-Based Commentary**

**Ahmed Ibrahim**

This commentary is on the review:

[**Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review**](https://www.mdpi.com/1422-0067/24/14/11488)

[Sarfaraz K. Niazi](https://scholar.google.com/citations?hl=en&user=O2dgTYIAAAAJ&view_op=list_works&sortby=pubdate) 1,\* and [Zamara Mariam](https://scholar.google.com/citations?hl=en&user=x8MxPbMAAAAJ) 2

1 College of Pharmacy, University of Illinois, Chicago, IL 61820, USA

2 Zamara Mariam, School of Interdisciplinary Engineering & Sciences (SINES), National University of Sciences & Technology (NUST), Islamabad 24090, Pakistan; [zmariam.msbi21rcms@student.nust.edu.pk](mailto:zmariam.msbi21rcms@student.nust.edu.pk)  
\* Correspondence: niazi@niazi.com; Tel.: +1-312-297-0000

**Insights and Perspectives on Machine Learning Applications in Chemoinformatics: A Review-Based Commentary**

**This article is a summary for the review:** [Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review](https://www.mdpi.com/1422-0067/24/14/11488) **by** [Sarfaraz K. Niazi](https://scholar.google.com/citations?hl=en&user=O2dgTYIAAAAJ&view_op=list_works&sortby=pubdate), and [Zamara Mariam](https://scholar.google.com/citations?hl=en&user=x8MxPbMAAAAJ).

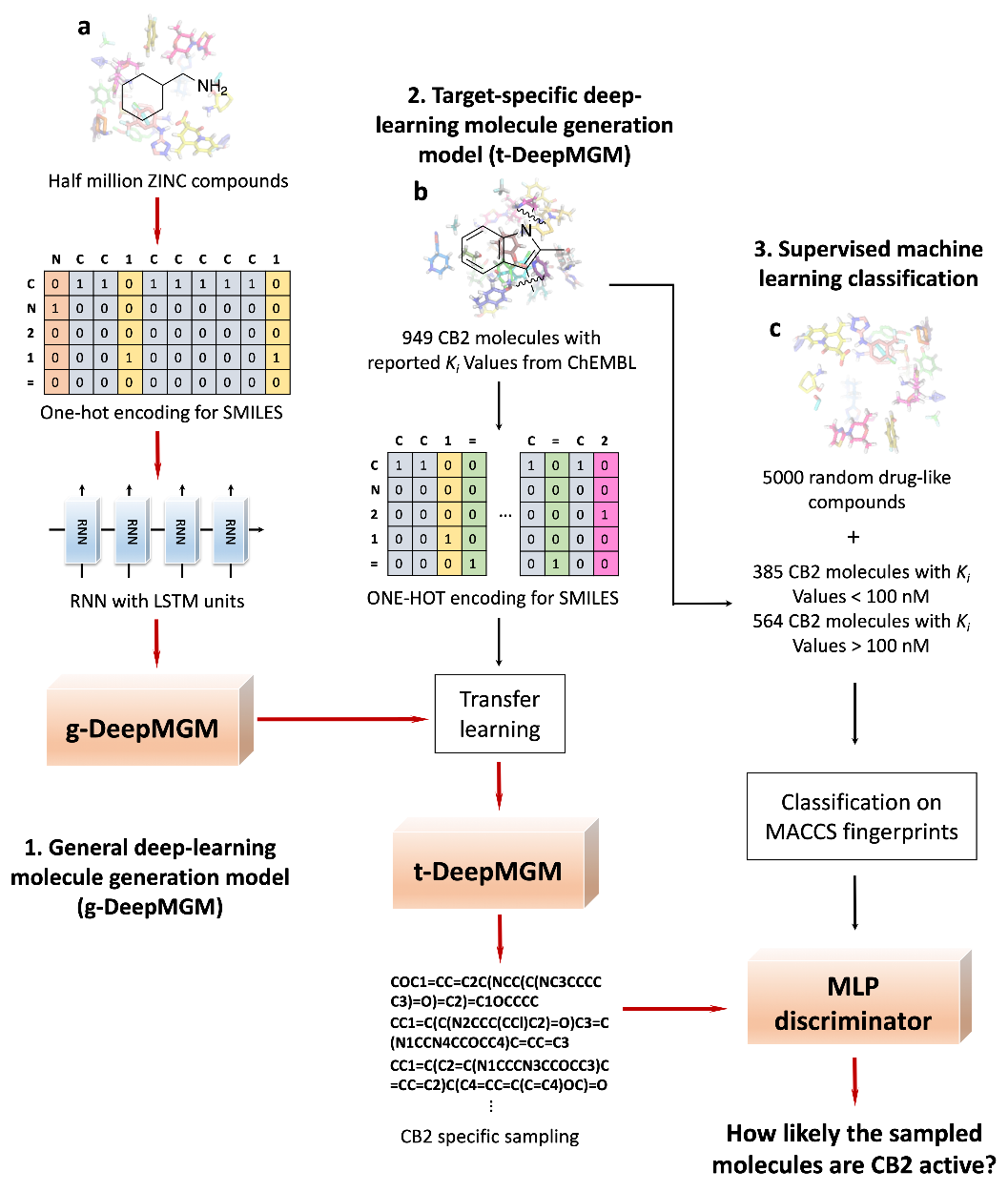
# **Introduction**

The classical methodology of placement of new drug in market is lengthy and demanding. It takes more than 6 years to design a new molecule, and clinical trials and approval processes would take additional 7 years. Furthermore, it is a risky industry with a failure rate of about 82%. The new techniques which integrate chemoinformatics, machine learning, and QSAR (Quantitative Structure Activity Relationship) modeling can now revolutionize the drug discovery processes through expediting its steps significantly. These novel approaches can accelerate mining literature to search for effective molecules, empower molecular modeling, mechanics, and dynamics, predict toxicological properties of candidates, and even generate molecules with desired and tailored properties. Such benefits have positive impacts on patient care and drug industry economics as well. To realize the significance of these effects, it is important to note that global small molecule market size reached $88.5 billion in 2024, and is projected to surpass $190 billion by 2034. In this article I aim to comment on and articulate insights and main topics discussed in the review, discerning the revolutionary effects of the alliance of machine learning and chemoinformatics.

# **Chemoinfromatics and Machine Learning Integration**

Chemoinformatics, is a discipline that encompasses the utilization of chemistry knowledge and computational techniques to help effectively in mining chemical libraries to retrieve chemical information and suggest molecules for further studying, predicting molecule activity and toxicity, and molecular docking. Application of machine learning and chemoinformatics to screen and process data from drug databases like ChEMBL, BindingDB, DrugBank, Inxight, and Protein Data Bank can effectively convey essential and profound knowledge of molecular structures, physiochemical properties, and descriptors, help create focused libraries, and enhance the discovery and design of novel molecules with required bioactivity.

***Example of General/Target-Specific Molecule Generation Models (g-DeepMGM and t-DeepMGM)***  
The **g-DeepMGM** (Deep Learning Molecular Generative Model) was trained on 500,000 molecule randomly collected from ZINC database. The molecules were represented as SMILES strings. Then 949 molecules with known ligands for specific targets (CB2 receptors) were grasped from ChEMBL and used to train **t-DeepMGM** which was built with transfer learning from the more generic g-DeepMGM model. The resultant t-DeepMGM has more focused properties and output five seed scaffolds for molecular sampling. This output should be specifically related to CB2 targets. A multi-Layer Perceptron **(MLP)** was constructed and trained against labelled dataset and was employed as discriminator to evaluate compounds’ activity for CB2 receptors.



# **Molecular Descriptors**

In order for molecular data to be readily utilized by QSAR models and machine learning approaches, multiple molecular descriptors are employed to quantitatively represent compounds' structures, biological, and physiochemical properties. They are categorized as in the following table;

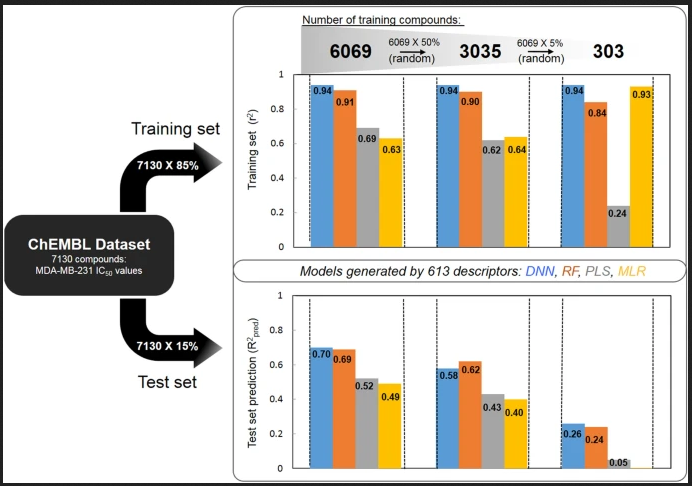
|  |  |  |
| --- | --- | --- |
| **Descriptor Dimension** | **Descriptor Type** | **Example** |
| 0D | The molecule’s atoms, bonds, and functional groups count | Molecular weight, LogP (partition coefficient) |
| 1D | Molecular properties in a linear manner | Molecular Formula, SMILES & SELFIES |
| 2D | Topological polar surface area (TPSA) | Molecular fingerprint (e.g., Morgan fingerprint),  Constitutional descriptors (e.g., atoms, bonds, and rings count) |
| 3D | Special properties of a molecule | Molecular shape descriptors (e.g., volume, surface area), Pharmacophore features |
| 4D | Electrostatic potential descriptors with spatiotemporal aspects | Molecular dynamics descriptors, solvent accessible surface area (SASA), radius of gyration (Rg), Time-dependent properties (e.g.,  dynamic polar surface area (dPSA), time-dependent dipole moment |

# **QSAR and Machine Learning Integration**

QSAR (Quantitative Structure Activity Relationship) is a computational approach to help establish and understand the relationship between chemical structures of compounds and their biological and physiochemical properties. This enables the prediction of how compounds’ activities and properties can vary with variations in their chemical structures. As a result of this provided predictions and knowledge, QSAR helps reducing the number of synthesized compounds for trials by facilitating selection of the most promising candidates. One of the examples of classical techniques to develop QSAR models is Hansch Analysis which used Multiple Linear Regression (MLP) to establish the required correlations. Another example is Free-Wilson Analysis, which focused mainly on the study of the effects of adding substituents to the molecule on biological activity, and used also linear models. A major challenge was to develop an effective QSAR model for huge and diverse datasets.

Currently, QSAR models can be constructed with supervised and unsupervised machine learning. In supervised learning, model is trained against labeled datasets with the input being the chemical structure with suitable representation and the output labels represent the corresponding properties like biological activity or toxicity. Unsupervised learning is used to explore unknown patterns or relationships within the chemical data through, for example, enabling clustering molecules based on structural similarities.

Comparative study between Deep Neural Networks (DNN) and conventional QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery showed that DNNs are remarkably superior in integrating immense chemical and biological data to make successful predictions ranging from in-vitro and in-vivo activities to ADMET properties (absorption, distribution, metabolism, excretion, and toxicity). The figure shows comparison of training and test sets’ prediction efficiencies between different models, DNN, RF (Random Forest), PLS (Partial Least Squares), and MLR (Multiple Linear Regression), compared with decreasing number of training compounds.



In another study, a DNN was used together with KNN (K-Nearest Neighbor) approach to build a QSAR model for a group of 1000 molecules having anticancer activity. This procedure succeeded to allow understanding the relationship between certain structural characteristics and the anticancer activity. SVM (Support Vector Machine) is a supervised machine learning algorithm that effectively classifies data by finding the hyperplane which maximize distance between classes. SVM is a reliable approach in QSAR modeling because of its efficiency in handling highly dimensional data and non-linear relationships.

CNNs (Convolutional Neural Networks) with their great capabilities in image analysis enabled protein structures visualization as 3D-images facilitating the prediction of effects of mutations on protein structures. CNNs can grasp molecular features from 2D representations of compounds. RNNs (Recurrent Neural Networks), also known as long short-term memory (LSTM) networks, can capture structural patterns from SMILES strings. DeepSnap, a deep-learning-based model, has been built to be able to predict toxicity of many compound utilizing 3D photographs of chemical structures without using descriptors. Overall, DNNs clearly enhanced QSAR models’ performance.

# **Interpretability and Explainability Challenges**

Being incapable to clearly explain why a machine learning model has given certain predictions or outputs renders some researchers unable to utilize and trust the results. Interpretability of ML-QSAR models enhances transparency and reproducibility, giving the space for stakeholders to take trusted decisions. Heat maps, feature importance plots, LIME (Local Interpretable Model-Agnostic Explanations), and SHAP (Shapley Additive Explanations) are techniques which can explain model’s predictions by relating them to specific features. STONED (Structure–Topology Optimization for Novel Explanatory Discoveries) allows skeletal, molecular structure-based explanations by producing legitimate molecular counterfactuals for any model without the need for training a counterfactual generator.

# **Conclusion**

The utilization of machine learning capabilities in chemoinformatics has expedited drug discovery processes through providing efficient hit-predictions, facilitating the design and development of new molecules with desired biological activities, and toxicological and physiochemical properties. It allowed efficient and rapid screening of datasets with high dimensionality, previously arduous to interpret. The current alliance between machine learning, chemoinformatics, and QSAR principles has revolutionized drug discovery field, and delivering great positive impacts in favor of patients and pharmaceutical industries economics.

(<https://www.mdpi.com/2073-4409/11/5/915>)

<https://www.sciencedirect.com/science/article/abs/pii/B9780128200452000064>

<https://www.nature.com/articles/s41598-020-73681-1>

(73) <https://pubs.acs.org/doi/10.1021/ci060149f>

<https://www.precedenceresearch.com/small-molecule-drug-discovery-market>