JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Progress in Chemoinformatics through the Application of Machine Learning

[1]Koyel Misra*, [2]Subhajit Mukherjee, [3]Bipasha Mridha Ghosh, [4]Sanchita Sarkar
[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor, [4]Professor
[1]Department of Basic Science,
[1] NSHM Institute of Engineering & Technology, NSHM Knowledge Campus Durgapur, West Bengal-713212, India

Abstract: The understanding of molecular interactions, drug design, and chemical synthesis has undergone a paradigm shift due to the multidisciplinary integration of chemistry and informatics in modern scientific discovery. This study provides a comprehensive examination of the significant advancements in chemoinformatics enabled by the transformative potential of machine learning techniques.

The study begins by explaining the fundamental concepts and principles of chemoinformatics to provide a solid understanding of its crucial role in exploring the chemical universe. It then delves into the emerging field of machine learning, highlighting its applications and adaptations for chemical studies. Machine learning enhances chemoinformatics by enabling predictive modeling, pattern detection, and knowledge extraction from vast chemical datasets. Additionally, the study meticulously lists and evaluates various machine learning techniques used in chemoinformatics, such as molecular property prediction, drug discovery, virtual screening, and molecular generative models.

Highlights:

- Enhanced Predictive Accuracy: Machine learning techniques enable the development of highly accurate quantitative structure-activity relationship (QSAR) models, facilitating the prediction of chemical properties and biological activities with improved precision.
- High-Throughput Screening: Automation and AI-driven algorithms allow for the rapid screening of vast chemical libraries, accelerating drug discovery by identifying promising compounds and reducing the need for costly and time-consuming experimental assays.
- Molecule Generation: Generative models and deep learning algorithms can design novel molecules with desired properties, aiding in the creation of innovative drug candidates and materials.
- Interpretable Insights: Advanced machine learning methods provide interpretable insights into chemical structure-activity relationships, enabling chemists and researchers to gather a concrete knowledge of molecular interactions.
- Personalized Medicine: Machine learning in chemoinformatics supports the development of personalized medicine by tailoring drug treatments to individual genetic and physiological profiles, potentially revolutionizing healthcare.

Keywords: Chemoinformatics, Machine Learning, Molecular Sciences, Drug Discovery, Data-driven Chemistry, Big Data in Chemistry.

I. INTRODUCTION:

Frank K. Brown first used the word "chemoinformatics" in 1998 with the intention of accelerating invention of medicines and their development. Today, chemoinformatics is essential to all applied sciences. The general process of drug invention required around 15 years and huge amount of funding in 1998. Chemoinformatics and drug discovery have undergone a significant revolution thanks to recent advances in machine learning (ML) and artificial intelligence (AI). The market for small-molecule drug development is expected to generate revenues of $75.96 billion in 2024 and $163.76 billion by 2032 [1,2].

Machine learning has been utilized in chemoinformatics and drug discovery for more than two decades, utilization of artificial intelligence in various sectors, including chemistry, has raised the importance of Machine learning and broadened its field of application. Machine learning is now the most common artificial intelligence technique in robotics and chemistry, while other Artificial intelligence methodologies such as machine vision and its allied systems are only now being investigated. Importantly, growing curiosity in artificial intelligence has primarily prompted the use of deep learning using deep neural network (DNN) architectures for chemical applications such as compound activity prediction or procedure design [3-5] (Figure 1).

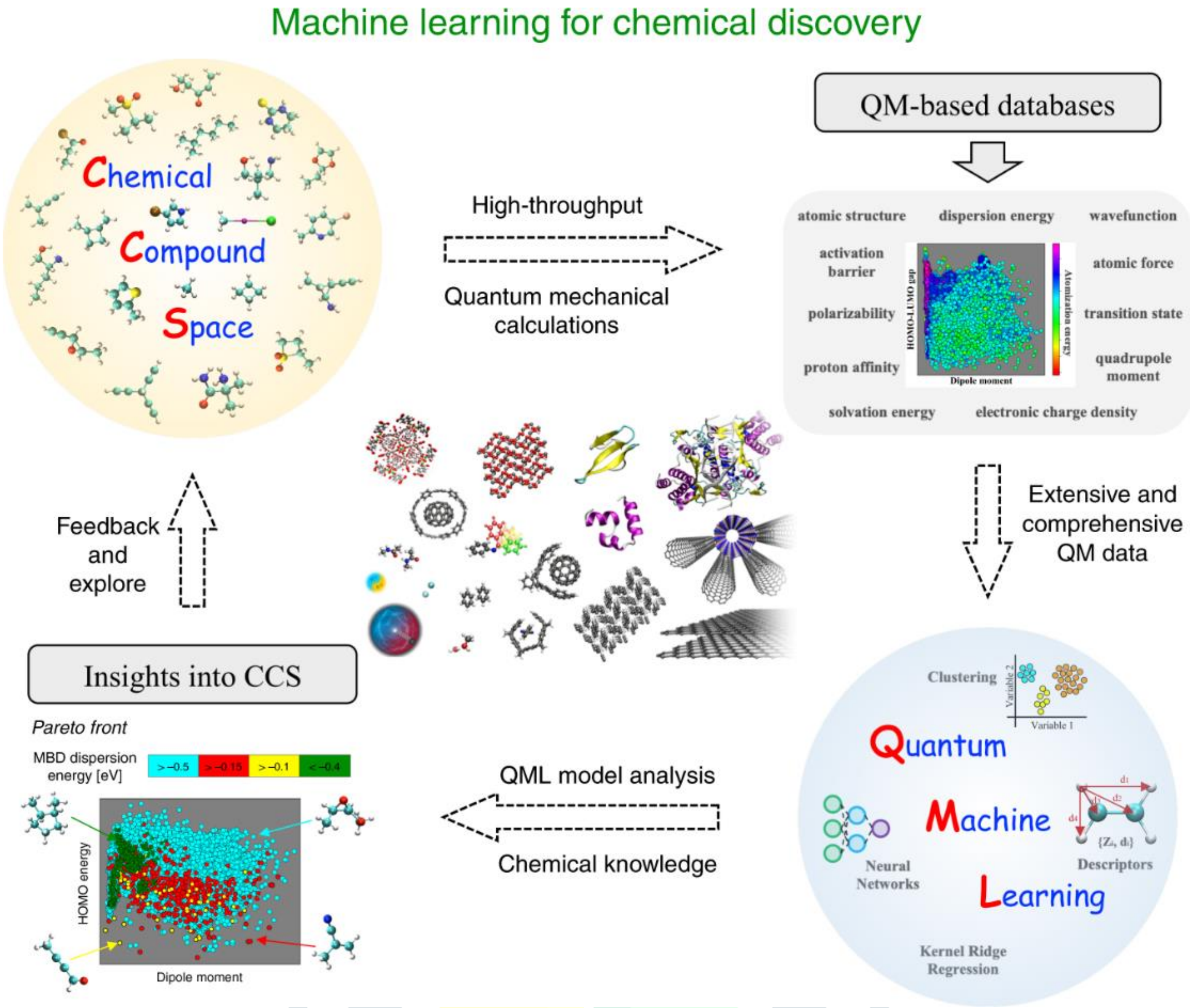

Figure 1: Application of machine learning in chemical discovery [5].

II. Chemoinformatics Exploration:

Chemoinformatics, which uses inductive learning to anticipate chemical processes, has developed as a powerful field in drug development at the junction of chemistry and informatics [3,4]. The utility of machine learning in cheminformatics has transformed the research pathway where the researchers discover, analyze, and forecast the properties and actions of molecules. It is primarily concerned with compound searching in available databases, molecular engineering and manipulation, library design, chemical space exploration, molecular network mining, pharmacophore and scaffold analysis [6-10].

III. Basic Fundamentals of Chemoinformatics:

Machine learning models predict chemical training data that is presented as mathematical equations. A complicated, layered computational technique is used to convert complex structures into chemical data generated by machine learning. Descriptor creation, , molecular dynamic simulations, graphs, chemical space searching, fingerprint building, similarity analysis, and so on are all part of the process. Each layer is intertwined with the previous layers, considerably altering the machine learning models' perception of the chemical data and improving their prediction skills.

(A) Data Mining and Chemical Databases

Chemical data is required for training ML models, and chemoinformatics is the use of chemical databases for storing and retrieving chemical information. These databases allow for the search of single molecules as well as the analysis of vast chemical datasets. Model training is strongly reliant on managing and utilizing vast volumes of molecular information, such as structures, bio-activities and other related properties. Natural compound databases such as LOTUS [11], COCONUT [12], Super Natural-II [13], NPASS [14], Sym Map [15], TCMSP [16], and TCMID [17] are useful sources. These databases contain detailed information about molecular structures and their physical character along with molecular descriptors.

Abductive approaches based on similarities in structure can be used to transmit knowledge about the mechanism using the known structures of these molecules. As previously stated, several similarity scores can be computed, taking into account the similarity of one dimensional structures (such as similarity based on SMILES or SELFIES [18]), two dimensional structures (such as two dimensional fingerprints and topological similarities), and three dimensional structures (such as similarity based on geometry). Several metrics useful for molecular similarity computations have been established in previous studies, such as Dice index, Tanimoto index, Overlap coefficient, Cosine coefficient, Soergel distance and Manhattan distance [19-21]. Moreover, drug databases such as ChEMBL [22], BindingDB [23], DrugBank [24], Inxight [25], and Protein Data Bank [26] can be used to obtain chemical bioactivity and structural data. Recurrent neural networks (RNN) and other generative models have been used to build novel chemical compounds with desirable features such as better activity with reduced toxicity.

(B) CHEMICAL DATA REPRESENTATION

Molecular data may be represented empirically, molecularly, and structurally in graphs, descriptors, fingerprints and so on [27,28]. In one study [29], a multivariate random forest model for gene characterization was trained using numerical data on gene sequence. Another study used numeric-based activity data to build a Nave Bayesian (NB) model that represented antagonist binding on estrogen receptors [30]. To forecast the characteristics of tiny molecules on the basis of ADMET (absorption, distribution, metabolism, excretion [31], a Machine learning-based model was trained using around thirty chemical numerical datasets from Merck.

(C) MOLECULAR DESCRIPTORS

Molecular descriptors are quantitative representations of the structural, physicochemical, and biological aspects of chemical molecules. These are quantitative metrics that are used for similarity analysis, virtual screening, and predictive modelling. Zero,one two, three, and four dimensional chemical molecular descriptors are classified [32-35].

1. Machine learning

Machine learning among artificial intelligence is one of the preferred method for producing practical software for various domains [10,11]. Machine learning is learnt from data using statistical approaches. Even hidden patterns and complicated data can be extracted from supplied data sets and expressed as mathematical objects using these techniques. Many artificial intelligence software developers now believe that in various aspects, desired input-output behavior is significantly easier than manual programming.

2. Deep Learning

Machine learning constrains by its input information. While evaluating photographs of standard size, the system will receive thousands of pixels. This implies that information must be received and grouped in order to be selected as essential to the activity. Deep learning is capable of dealing with such issues. To create accurate predictions, it employs “multi-layered neural networks”, massive volumes of data, and computation period. Unlike machine learning, deep learning does not require hand-engineering features from raw data.

IV. MACHINE LEARNING APPROACHES WIDELY USED IN MOLECULAR SCIENCES:

In comparison to traditional approaches (such as quantum mechanical calculations, density functional theory or molecular machine based methods, etc.), machine learning algorithms have vast applications in numerous areas of science to provide quick and more precise solutions. The relationship between molecular structure and its properties is largely being determined [36]. Machine learning models are used in universal approximation theorem for ANNs and QSPRs [37].

Machine learning techniques are classified using a variety of standards. One classification approach is dependent on whether human guidance is required for the machine learning system or not. These techniques are classified into two groups based on this: supervised and unsupervised learning.

(a) Supervised Learning

Among the machine learning approaches supervised learnings are most common [38]. Most predictions of molecular properties fall within this group. “Supervised learning is the process of learning a function that maps an input to an output based on human-labelled input-output pairs. The algorithms try to reduce the errors discovered during the learning process”. It is capable of extracting complex nonlinear trends and outperforms manually programmed standard models.

The supervised learning is classified as follows-

(i)Traditional machine learning methods

Traditional machine learning approaches is defined as the important algorithms which are often used as the basis for more advanced machine learning. Traditional algorithms include kernel-based approaches (such as SVMs), decision tree methods (such as Random Forests and XGBoost), Bayesian methods, and others. These algorithms are useful for classifying and predicting data. Regression problem such as molecular property prediction has been solved using methods like Kernel Ridge Regression (KRR) [39-41], Random Forests [42,43] and Elastic Net [44]. Despite their effectiveness in a variety of domains, usually these models rely on manually operated molecular descriptors derived by the symbolic molecular representation. Few machine learning systems use practical measurements as descriptors, such as physicochemical properties.

However, if the dataset size is limited, classic machine learning approaches are preferable over DNNs since DNNs tend to over fit.

(ii) Artificial Neural Networks (ANNs):

Artificial Neural Networks is analogous to biological neural networks [45,46] are one of the most commonly used computer models. The input x can be regarded of as being transformed into a new era, it gets associated with the output y. DNNs are formed when ANNs change features progressively over various layers. These are fantastic tools for discovering patterns and correlations that are very complicated for a person to solve and program physically.

DNNs incrementally learn high-level features from data, with each new concealed layer capturing higher level information than the prior layer. Domain expertise and the extraction of features manually are no longer required. As a result, DNNs may be learned to find meaningful chemical descriptors that are best matched to the provided data. However, because features must be trained from the ground up for each new dataset whose procedures can result in appropriate minimal data.

A feedforward neural network, or AN, is the most basic sort of ANN, in which information goes in only one way from input to output. more include recurrent neural networks (RNNs) and more.

(iii) Recurrent neural networks (RNNs):

Each iteration of vanilla ANN training forgets what it learned in the preceding iteration. When it recognizes correlations and patterns in periodic data such as the sequence of amino acids in proteins is a disadvantage. RNNs are ANN designs for their recurrent memory cells are capable of recalling input and modelling short-term dependencies. They are widely employed in the modelling of sequences and generation.

When simple RNNs are trained to anticipate dependencies that persist, the gradient decreases or bursts as it propagates back through time - this is known as the disappearing and bursting gradient difficulties [47,48]. This makes it impossible for RNNs to learn these features from extended sequences. “The long short term memory (LSTM) unit”, or its derivative, the gated “recurrent unit (GRU)”, incorporates "gates" that alleviate gradient difficulties. These gates determine how much of the past to remember, what to incorporate in the current state, and what to pass on as output to the next gate. Longer sequences can be achieved by Gradients. Because molecular representations such as SMILES involve long-term dependencies such as closing parenthesis and rings, LSTMs and GRUs are commonly utilized for inverse molecular design.

(b) Unsupervised:

Unsupervised learning, as opposed to supervised learning, is a method of learning without the use of labelled data. Instead of looking for predetermined classes of data, it merely seeks for data that can be categorized based on their commonalities. This is why it's often referred to as “clustering or grouping”. The system is taught using vast amounts of database and gather knowledge on its own. This part provides some instances of autonomous knowing for various processes.

(i) Auto encoders (AEs):

Character variational auto encoder (VAE), the initial machine learning-based generative model for molecules, was developed by Gomez-Bombarelli et al. [49] in 2016. In addition, the model provided a data-driven technique for molecular descriptors.

When VAEs are trained to replicate compounds as well as their characters, the latent space reorganizes molecules with more or less same qualities which are close to one another [50,51].

(ii) Generative adversarial networks (GANs):

GANs [52] are a quickly developing field of study. They are an ingenious method of learning this model composed of two sub-models: the generator model Gθ and the discriminator model Dϕ. These two models are ANNs that are often trained in tandem using stochastic gradient descent (SGD).

V. Machine-Learning-Based QSAR Modelling:

“Machine Learning-Based Quantitative Structure-Activity Relationship (QSAR)” Modelling is an effective chemoinformatics methodology that uses computational methods to predict the biological activity, physicochemical qualities, or toxicity of chemical compounds based on structural factors. In the pharmaceutical and chemical sectors, QSAR models are widely used to speed drug discovery, design novel molecules, and choose compounds for experimental testing.

(a) An Overview of QSAR Modelling

Quantitative Structure-Activity Relationship (QSAR) modelling is a chemoinformatics discipline that focuses on developing mathematical correlations between chemical structures and their biological or chemical functions. QSAR models use descriptors collected from the chemical structure of molecules to predict attributes or activities such as pharmacological potency or toxicity.

(b) Machine Learning's Role

Machine learning approaches have transformed QSAR modelling by allowing for the creation of more accurate and robust predictive models. Traditional QSAR models depended on linear regression, while newer machine learning methods such as random forests, support vector machines, and neural networks can capture complicated, nonlinear data correlations.

(c) Preparation of Data and Feature Engineering

The identification and extraction of important molecular descriptors or characteristics from chemical structures is a vital stage in QSAR modelling. In this step, chemical information is converted into numerical data that machine learning algorithms can process.

The calculation of properties such as molecular weight, LogP, chemical fingerprints, and topological indices may be part of feature engineering.

(d) Validation and data splitting

To evaluate the performance of QSAR models, proper data partitioning into training, validation, and test sets is critical. Cross-validation techniques, for example, help ensure that models generalize effectively to new data and avoid overfitting.

(e) Model Development and Optimization

To discover the correlations between chemical characteristics and the target activity, machine learning algorithms are trained using the training dataset. Model hyperparameters are optimized for performance, and several approaches can be compared to determine the best-performing model.

(f) Metrics of Performance

To measure the prediction capacity and accuracy of QSAR models, evaluation metrics such as RMSE (Root Mean Squared Error), R2 (Coefficient of Determination), and ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) are utilized.

(g) Application in Drug Development

QSAR models are critical in virtual screening, which involves computationally screening vast chemical libraries to uncover possible drug candidates.

They help in lead compound discovery, molecular structure optimization, and ADMET (Absorption, Distribution, Metabolism, Excretion, and Transport) prediction.

VI. Interpretability and Explainability of ML-QSAR Models:

The accurate prediction of bioactivities is the main target of QSAR analysis based on the validated models. As a part of learning to assess different QSAR model interpretation methodologies, six simulated datasets of varying complexity were created. These databases were employed for investigating of a vast range of descriptor and algorithm pairings as well as the Structure-Property Correlation Index (SPCI) approach of worldwide used interpretation. According to the findings, productivity may drop faster than interpretation performance and in some cases models with high predictability may have poor interpretation performance [53]. Several strategies can be used to improve the explanation and understanding of ML-QSAR models. The most important chemical descriptors or features contributing to the model's predictions can be identified using feature importance analysis. Heat maps and feature significance plots, for example, can help you comprehend the links among features and its results. Furthermore, model-independent techniques such as LIME (Local Interpretable Model-Agnostic Explanations) [54] or SHAP (Shapley Additive Explanations) [55] can provide insights into individual predictions by emphasizing the contributions of each feature. A new method for visualizing QSAR models is presented in a journal, which simplifies examination by introducing a new measure of structure similarity. The approach works by explaining models into a two-dimensional plane, where the distance between two models is proportional to the difference in their expected activities [56]. Another study combines direct kernel-based PLS with Canvas two dimensional fingerprints to generate anticipated QSAR models that may be projected onto the atoms of a molecule. The paper provides a model representation that may be utilized to detect the most significant atoms [57]. Various interpretation methodologies were developed; however, there are no relevant standards for assessing how well they relate to the interpretation of QSAR models. A paper suggests the STONED (Structure-Topology Optimization for Novel Explanatory Discoveries) approach, which generates molecular counterfactuals for individual model [53].

VII. Conclusion:

The use of machine learning techniques in chemoinformatics has greatly aided in the discovery and design of extremely effective medications. The importance of chemoinformatics and machine learning based QSAR in medicinal field in this review. Integrating computer methodologies has transformed the area, allowing for more appropriate exploration of chemical field and prediction of bio-activity. Different QSAR modelling approaches emphasize features required for subsequent small molecule creation. They have shown promise in forecasting molecular characteristics and properties, assisting in complex selection and optimization.

In upcoming fields of research, QSAR modelling provides exciting possibilities for advancement. Molecular docking Combined with QSAR models help in understanding the binding affinity and give useful information about ligand-target protein interactions [58] (Figure 2).

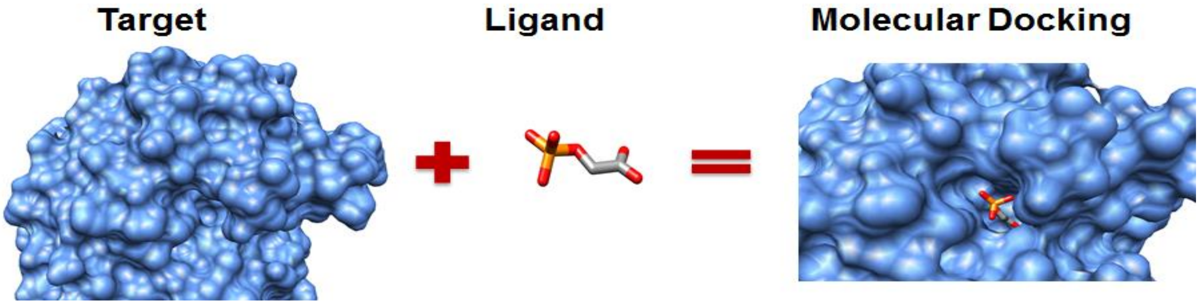

Figure 2: Information about ligand-target protein interactions [58].

QSAR models can help guide the decision-making and optimization of fragments in the development of innovative drug candidates in fragment-based design techniques.

The advantages of molecular docking tagged with QSAR models, fragment based design are accelerating the process of drug invention with economic benefits and increased success rate in developing new therapeutic agents.

VIII. Acknowledgment

The authors are thankful to NSHM Knowledge Campus Durgapur for providing the infrastructure and necessary support to carry out the study.

References

[1] Simm G., Pinsler R. and Hernàndez-Lobato J. M. 2020, Reinforcement learning for molecular design guided by quantum mechanics. In International Conference on Machine Learning, Nov 21, (PMLR) 8959.

[2] Brown F.K. 1998, Chapter 35—Chemoinformatics: What is it and How does it Impact Drug Discovery. In Annual Reports in Medicinal Chemistry; Bristol, J.A., Ed.; Academic Press: New York, NY, USA, Volume 33, pp. 375–384.

[3] Gawehn E, Hiss JA, Schneider G., Deep learning in drug discovery. Mol. Inform., 35(1), 2016, 3-14.

[4] Chen H., Engkvist O., Wang Y., Olivecrona M., Blaschke T. 2018, The rise of deep learning in drug discovery, Drug Discov. Today, 23, 1241.

[5] Tkatchenko A. 2020, Machine learning for chemical discovery, NATURE COMMUN., 11, 4125.

[6] Gasteiger, J. 2003, Handbook of Chemoinformatics; Wiley: New York, NY, USA,

[7] Varnek, A., Baskin, I.I. 2011, Chemoinformatics as a Theoretical Chemistry Discipline., Mol. Inform., 30, 20–32.

[8] Bajorath, J., Bajorath, J. (Eds.) 2011, Chemoinformatics and Computational Chemical Biology. In Methods in Molecular Biology; Springer Science+Business Media: Humana Totowa, NJ, USA.

[9] Kapetanovic, I.M. 2008, Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. Chem.-Biol. Interact. 171, 165–176.

[10] Dehmer, M., Varmuza, K., Bonchev, D. 2012, Statistical Modelling of Molecular Descriptors in QSAR/QSPR; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany.

[11] Lo, Y., Rensi, S.E., Torng, W., Altman, R.B. 2018, Machine learning in chemoinformatics and drug discovery. Drug Discov. Today,23, 1538–1546.

[12] Chandrasekaran, B., Abed, S.N., Al-Attraqchi, O., Kuche, K., Tekade, R.K. 2018, Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties; Elsevier: Amsterdam, The Netherlands, 731–755.

[13] Engel T. 2006, Basic Overview of Chemoinformatics., J. Chem. Inf. Model. 46, 2267–2277.

[14] Jordan M. I., Mitchell T. M. 2015, Machine learning: Trends, perspectives, and prospects Science, 349, 255.

[15] Hong Y., Hou B., Jiang H., Zhang J. 2020, Machine learning and artificial neural network accelerated computational discoveries in materials science Wiley Interdiscipl. Rev. Comput. Mol. Sci. 10, e1450.

[16] Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., Graham, J.G., Stephan, R., Page, R., Vondrášek, J. et al. 2021, The LOTUS initiative for open natural products research: Knowledge management through Wikidata. bioRxiv.

[17] Sorokina, M., Steinbeck, C. 2020, Review on natural products databases: Where to find data in 2020. J. Cheminform. 12, 20.

[18] Banerjee, P., Erehman, J., Gohlke, B.O., Wilhelm, T. Preissner, R.; Dunkel, M. 2015, Super Natural II—A database of natural products. Nucleic Acids Res. 43, D935–D939.

[19] Zeng, X., Zhang, P., He, W., Qin, C., Chen, S., Tao, L., Wang, Y., Tan, Y., Gao, D., Wang, B. et al. 2018, NPASS: Natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 46, D1217–D1222.

[20] Wu, Y., Zhang, F., Yang, K., Fang, S., Bu, D., Li, H., Sun, L., Hu, H., Gao, K., Wang, W. et al. 2019, SymMap: An integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res. 47, D1110–D1117.

[21] Ru, J., Li, P., Wang, J., Zhou, W., Li, B., Huang, C., Li, P., Guo, Z., Tao, W., Yang, Y. et al. 2014, TCMSP: A database of systems pharmacology for drug discovery from herbal medicines. J. Cheminform. 6, 13.

[22] Xue, R., Fang, Z., Zhang, M., Yi, Z., Wen, C., Shi, T. 2012, TCMID: Traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res. 41, D1089–D1095.

[23] Krenn, M., Aspuru-Guzik, A., Nigam, A., Friederich, P. 2020, Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. arXiv 1, 045024.

[24] Engel, T., Gasteiger, J. (Eds.) Chemoinformatics: Basic Concepts and Methods; Wiley: New York, NY, USA, 2018; Available online: https://www.wiley.com/en-dk/Chemoinformatics:+Basic+Concepts+and+Methods-p-9783527331093.

[25] Xue, H., Stanley-Baker, M., Kong, A.W.K., Li, H., Goh, W.W.B. 2022, Data considerations for predictive modelling applied to the discovery of bioactive natural products. Drug Discov. Today, 27, 2235–2243.

[26] Nikolova, N., Jaworska, J., 2003, Approaches to Measure Chemical Similarity—A Review. Qsar Comb. Sci. 221006–1026.

[27] Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. 2018, DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–D1082.

[28] Siramshetty, V.B.; Grishagin, I.; Nguyễn, Đ.T.; Peryea, T.; Skovpen, Y.; Stroganov, O.; Katzel, D.; Sheils, T.; Jadhav, A.; Mathé, E.A.; et al. 2022, NCATS Inxight Drugs: A comprehensive and curated portal for translational research. Nucleic Acids Res. 50, D1307–D1316.

[29] Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., Abola, E.E. 1998, Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Crystallogr. Sect. D Biol. Crystallogr. 54, 1078–1084.

[30] McDonagh, J. L, Silva, A. F, Vincent, M. A., Popelier, P. L. 2017, Machine learning of dynamic electron correlation energies from topological atoms J. Chem. Theory Comput. 14, 216.

[31] Meyer J. G., Liu S., Miller I. J., Coon J. J. and Gitter A. 2019, Learning drug functions from chemical structures with convolutional neural networks and random forests, J. Chem. Inf. Model. 59, 4438.

[32] Agar J. C., Naul B., Pandya S., van Der Walt S. 2019, Revealing ferroelectric switching character using deep recurrent neural networks Nat. Commun. 10, 1.

[33] Gómez-Bombarelli R., Wei J. N., Duvenaud D., Hernàndez-Lobato J.M., Sànchez-Lengeling B., Sheberla D., Aguilera-Iparraguirre J., Hirzel T. D., Adams R. P. and Aspuru-Guzik A. 2018, Automatic chemical design using a data-driven continuous representation of molecules ACS Central Sci. 4, 268.

[34] Sanchez-Lengeling B. and Aspuru-Guzik A. 2018, Inverse molecular design using machine learning: Generative models for matter engineering Science 361, 360.

[35] Pathak Y., Juneja K. S., Varma G., Ehara M., Priyakumar U. D. 2020, Deep learning enabled inorganic material generator Phys. Chem. Chem. Phys. 22, 26935.

[36] Simm G., Pinsler R., Hernàndez-Lobato J. M. 2020, Reinforcement learning for molecular design guided by quantum mechanics. In International Conference on Machine Learning, Nov 21 (PMLR), 8959.

[37] Olivecrona M., Blaschke T., Engkvist O. and Chen H. 2017, Molecular de-novo design through deep reinforcement learning J. Cheminform. 9, 1.

[38] Zhou Z., Kearnes S., Li L., Zare R. N. and Riley P. 2019, Optimization of molecules via deep reinforcement learning Sci. Rep. 9, 1.

[39] Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., Chong, J. 2016, BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 44, D1045–D1053.

[40] Ahuja K., Green W. H., Li Y. P., 2021, Learning to optimize molecular geometries using reinforcement learning J. Chem. Theory Comput. 17, 818.

[41] Murugan N. A., Poongavanam V., and Priyakumar U.D. 2019, Recent advancements in computing reliable binding free energies in drug discovery projects In Structural Bioinformatics: Applications in Preclinical Drug Discovery Process, SpringerChem., 221.

[42] Wu Z., Ramsundar B., Feinberg E. N., Gomes J., Geniesse C., Pappu A. S., Leswing K., and Pande V., 2018, MoleculeNet: A benchmark for molecular machine learning Chem. Sci, 9, 513.

[43] Ramakrishnan R., Dral P. O., Rupp M. and von Lilienfeld O. A., 2015, Big data meets quantum chemistry approximations: The machine learning approach J. Chem. Theory Comput. 11, 2087.

[44] Krogh A. 2008, What are artificial neural networks? Nat. Biotechnol. 26, 195.

[45] Kolen J. F., Kremer S. C. 2001, "Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies," in A Field Guide to Dynamical Recurrent Networks, IEEE, 237-243,

[46] Agar J. C., Naul B., Pandya S. and van Der Walt S. 2019, Revealing ferroelectric switching character using deep recurrent neural networks Nat. Commun. 10, 1.

[47] Gómez-Bombarelli R., Wei J. N., Duvenaud D., Hernàndez-Lobato J. M., Sànchez-Lengeling B., Sheberla D., Aguilera-Iparraguirre J., Hirzel T. D., Adams R. P. and Aspuru-Guzik A. 2018, Automatic chemical design using a data-driven continuous representation of molecules ACS Central Sci. 4, 268.

[48] Sanchez-Lengeling B., and Aspuru-Guzik A. 2018, Inverse molecular design using machine learning: Generative models for matter engineering Science 361, 360.

[49] Pathak Y., Juneja K. S., Varma G., Ehara M. and Priyakumar U. D. 2020, Deep learning enabled inorganic material generator Phys. Chem. Chem. Phys. 22, 26935.

[50] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. and Bengio Y., 2020, Generative adversarial networks Commun. ACM, 63, 139.

[51] Sutton R. S. and Barto A. G., 2018, Reinforcement Learning: An Introduction (MIT Press).

[52] Krakovsky M., Reinforcement Renaissance Commun. 2016, ACM, 59, 12.

[53] C3.ai. LIME: Local Interpretable Model-Agnostic Explanations. 2022. Available online: https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/#:~:text=What%20is%20Local%20Interpretable%20Model,to%20explain%20each%20individual%20prediction .

[54] Molnar, C. 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning. Available online: https://christophm.github.io/interpretable-ml-book/shap.html .

[55] Izrailev, S., Agrafiotis, D. 2004, A method for quantifying and visualizing the diversity of QSAR models. J. Mol. Graph. Model. 22, 275–284.

[56] An, Y., Sherman, W., Dixon, S.L. 2013, Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. J. Chem. Inf. Model. 53, 2312–2321.

[57] Wellawatte, G.P., Seshadri, A., White, A.J.P. 2022, Model agnostic generation of counterfactual explanations for molecules. Chem. Sci. 13, 3697–3705.

[58] Ogawa T., 2013, Protein Engineering - Technology and Application, ISBN: 978-953-51-1138-2.