 doi:10.20944/preprints202105.0567.v1

# Artificial intelligence and machine learning in medicinal chemistry and validation of emerging drug targets

Sameer Quazi[1], Rohit Jangi[1]

1. GenLab BioSolutions Private Limited, Bangalore, Karnataka, India.

**Abstract:**

Artificial learning and machine learning is playing a pivotal role is the society especially in the field of medicinal chemistry and drug discovery. Particularly its algorithms, neural networks or other recurrent networks drive this area. In this review, we have taken into account the diverse use of AI in a number of pharmaceutical industry including discovery of drugs, repurposing, development of pharmaceutical drug and its clinical trials. In addition the efficiency of these artificial or machine learning programs in achieving the target drugs in short time period, along with accurate dosage and cost effectively of the drug has also been discussed. Numerous applications of AI in property prediction such as ADMET have been used for prediction of strength of this technology in QSAR. In case of de-novo synthesis, it results in generation of novel drug molecules with unique design proving this a promising field fir drug design. Moreover its involvement in synthetic planning, ease of synthesis and much more contribute to automated drug discovery in near future.

**Keywords:**

**Artificial intelligence AI, machine learning, algorithms, QSAR, drug discovery etc.**

## 1) Introduction:

It has been estimated that an average cost of bringing drug to the marketplace is found to exceed about 3 billion dollars. This increment in cost is attributed to the two factors which include, the number of clinical trials resulting in attraction to about 80-85% as required for approving drugs for humans. Secondly, the complexity of the discovery phase of drugs demanding a considerable amount of investment for both time and resources (Díaz, Dalton, & Giraldo, 2019). A strong pipeline of candidate drug in preclinical trials will in turn have significant downstream effects in consideration of total approval. The advancement in both the computer software's and in-vitro approaches aim to employ as well as improve the various other aspects of drug discovery cycle and the test refered as quitensial design make test analyse DMTA. The area of increasing interest is utilization of data driven synthetic tools with the objective to reduce the number of failures with the subsequent increase in output of drugs during synthesis of novel molecular drug subunits. History from 1960 shows that computer aided synthesis planning CASP when the Corey group first disclosed LASHA which presents a rule based approach for retrosynthetic planning. This publication was the prime key for providing definition to heuristics involving chemical synthesis which could be a valuable tool for the software involving synthesis planning of drugs. From 1960s to 1990, many groups disclosed that the advancement in computer based planning of synthesis

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

were limited by computation resources and rely wholly on human based rules for reaction (Yang, Wang, Byrne, Schneider, & Yang, 2019).

The early progenitors provide the basis for many of the commercial software as for example Synthia formally termed as chematica. In addition to ICSynth where the rules coded by hand are utilized in addition to following guidelines for heuristics in relation to negative synthetic pathways.

From the past few decades, more automated methods have been found regarding synthesis process which use the subset of an artificial learning method being referred as machine learning for inferring the reactivity from previously available data which has provided a visible alternative to expert rule-based algorithm. Hence both the expert based and ML can come under the umbrella niche of AI approach. The first one using the information from crafted knowledge and second presents the example of using statistical learning method. Each of them has its own specific advantage for drug synthesis planning process. But the machine learning has extended to incorporate further new reactions because they are published based on extraction or training pipelines which in turn reduces burden on experts belonging to this category i.e. researchers etc. As much as reactions operate in the industry or company, the automated method provide prediction of the candidate molecule and results more robustly(Chan, Shan, Dahoun, Vogel, & Yuan, 2019).

Both the rule based and machine-based learning has provided valuable tool for planning of synthetic route being executed in laboratory and evaluated by chemist during research. As for example Synthia has developed route for the medically related compounds which is far better than the routes developed by experts. As Seiger et al. expressed that their researchers have not preferred the last or previous route developed by literature but has taken notes from their novel algorithm based route in double blind evolution process in double blind evolution process. Automated platforms are coupled to synthesis planning tools for a varying level of human intervention. However this field is in its early stages to use CASP or fully automated planning, this resulting in initial success which provides the better tool for the drug development process following DMTA cycle (Jing, Bian, Hu, Wang, & Xie, 2018).

2) **History of machine learning in molecular design of medicinal chemistry:**

In the computed based molecular design, artificial intelligence and machine learning is not a novel research. The previously done research by Hansh and Fujita in addition to Free and Wilson has come forward with the quantitative structure activity relationship modelling program. In their historic work they employed datasets as that of dozen of chemical derivatives to fit in the equation which in turn will anticipate phenotypic complex effects such as that of toxicity. Based on this research, a lot of scientists started working on the identification and analyzing approaches suitable for description of chemicals in more detail in order to catch the characteristics which give their properties such as the 3D structure and pharmacophores but also focused on autonomously learned representations (Zhang, Tan, Han, & Zhu, 2017). Based on the increased information on structure, and generation of data via combinatorial libraries and screening, the very first application of much complex machine learning becomes feasible. The up growing field of QSAR had a very hard lesson in 1990 based on the validation of model, controlled experiments and other up or downs. More specifically, the applications of computational model acting as a hard filter fir the number of data sets didn't get cover in training data which lead to an increase in disappointment in the field later (Stephenson et al., 2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

With the time an increase in understanding. Of Arithmetic principles and statistical data suggested the concept of domains of applicability being introduced in this field. This predictive confidence enabled the drug hunters for increasing the transparency of their tools in addition to their expectations. Hence it became the first step for the application of machine learning in drug discovery and design in 2000 which in turn rebuilt the trust of the researchers and allow them the usage of these tools (Zhavoronkov, Vanhaelen, & Oprea, 2020). In 2015, the advancement in computation such as involvement of GPUs in modern frameworks and additionally increase in capacity of RAM at the same time enabled the thinking of neural nets more feasibly. The most famous kaggle challenge used a deep Neural Net in order to win SAR challenges set by Merck. This competition proved to be a turning point where the learning of deep artificial intelligence methods had overcome the other machine based leaning approaches. Hence it proved to be a more useful tool for computational based molecular drug design. Moreover this deep learning can also trace its roots late in 1956 workshop run at Dartmouth College (Saikin, Kreisbeck, Sheberla, Becker, & Aspuru-Guzik, 2019). Instead on a long history of artificial learning, the field still have faced many pitfalls in the sense that expectations did not match the reality. Thus provided a huge number of setbacks to field, which later required time for recovery. At the current, multiple applications of AI are in working providing promising future to derive molecular descriptors and their relationship to biological properties. Hence the algorithms in this way is linked to requirement of big data sets for providing useful solutions because they give a large number of opportunities for navigation of huge data sets (Smith, Roitberg, & Isayev, 2018).

3) **Utilization of AI/ML in drug discovery:**

Artificial intelligence and machine learning has been used at different stages in number of early studies for the identification of target drug, generation or utilization of lead in addition to its optimization and pre-clinical developmental stage. For the target identification, the Al has used the data sets which are heterologous for the identification of pattern so that the underlying mechanism in both disease and target disease can be understood easily. In case of lead generation of lead and its optimization, the algorithms of Al and machine learning is involved in improving the score and quantitative structure analysis relationship model QSAR in the screening pipelines and supporting the de-Novo synthesis of drug designing process (Zhavoronkov, 2018). Lastly in the pre-clinical development and design, artificial intelligence is involved in generation of predictive models based on the physiochemical properties by processing of a huge amount of chemical data in an efficient way and for further absorption, metabolism, distribution and excretion ADME toxicity (Zhavoronkov, 2018).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

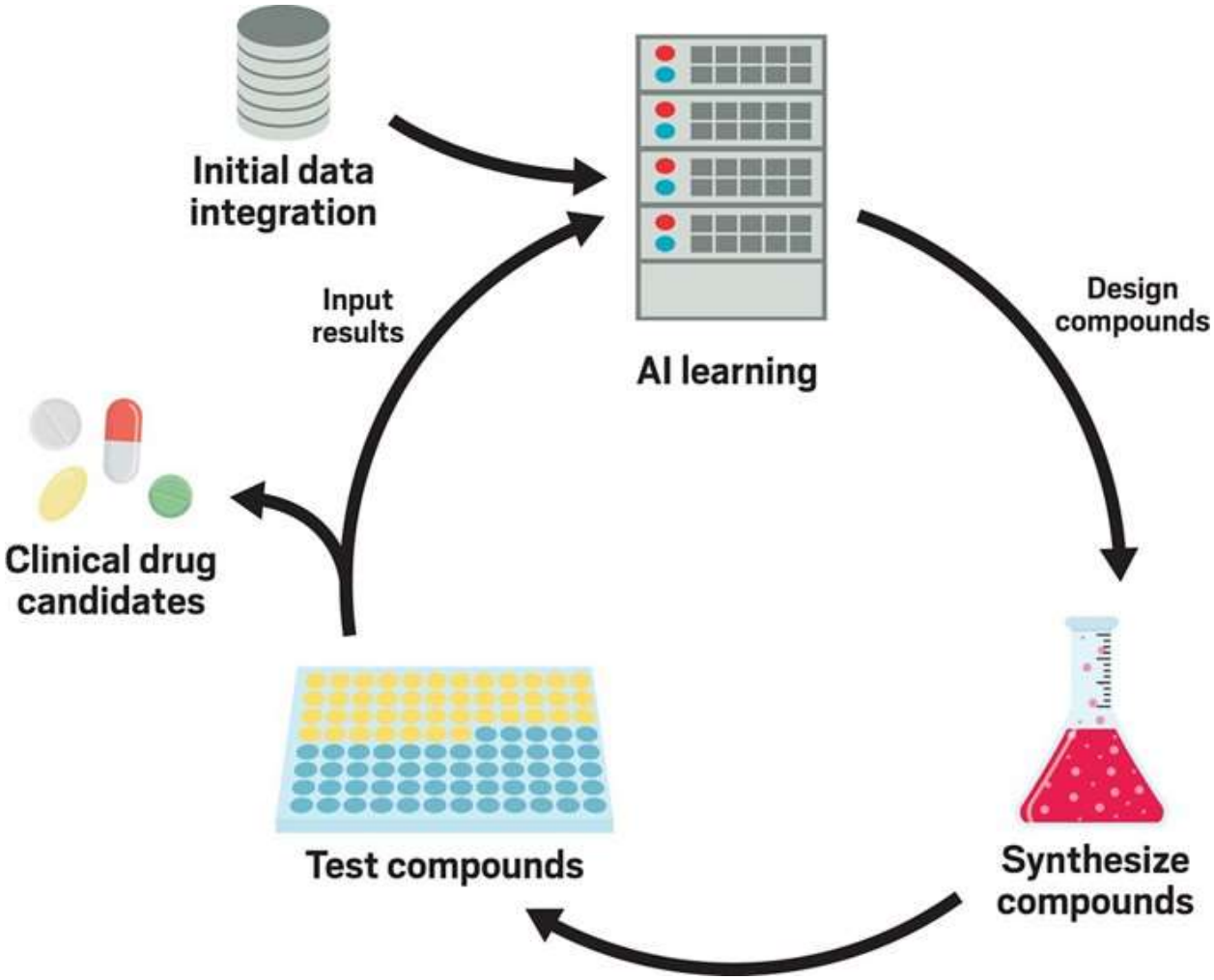


**Fig. 01. The cycle showing the involvement of AI learning in drug design involved in identification of target drug, optimization and lastly pre- clinical or clinical integration**

4) **Overview of AI/ML algorithms:**

Artificial learning utilities a large number of models for intelligence systems which are being classified further by learning procedures. Artificial intelligence algorithms can be used as the machine learning algorithm but still they are nor the same. So it is important to clarify both of the terms at first (Zhang et al., 2017). According to definition by FDA, artificial intelligence can be defined as the science which involves the engineering of artificial intelligence but on the other hand machine learning can be described as the technique of artificial intelligence for the design of algorithms being used for the analysis of huge amount of datasets. Hence ML techniques can be classified as AI techniques but at the same time not all the AI techniques are said to be involved in machine learning. **Table 1** presents the name and definition of most of the algorithms involved in drug discovery or medicinal chemistry. AI-related learning procedures are extensively arranged as administered, un-administered, semi supervised, dynamic, support, move, and multiple tasks learning. Various calculations are utilized in those learning structures to perform explicit functions like grouping or categorizing. Notwithstanding, accomplishment with AI requires more than

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

preparing an AI models only (Chen, Engkvist, Wang, Olivecrona, & Blaschke, 2018). A powerful AI work process includes

- Formulation of problem
- Preparation of data
- Extraction of features
- Selection of datasets for testing and training purposes
- Applying the model to all datasets and its refinement

| Name of algorithm | Functionalities |
|---|---|
| **Support vector machines** | Classification |
| **K-nearest-neighbors** | Classification/Regression |
| **Artificial neural networks** | Classification/Regression |
| **Deep neural network** | Classification/Regression |
| **Random Forest** | Classification/Regression |
| **Multiple regression** | Regression |
| **K-means clustering** | Clustering |
| **Fuzzy clustering** | Clustering |
| **Hierarchical clustering** | Clustering |
| **Principal component analysis** | Dimensionality reduction |
| **Generative adversarial networks** | Dimensionality reduction/Anomaly detection |
| **Active learning** | High-performance classifier |
| **Reinforcement learning** | Dynamic programming |
| **Transfer learning** | Pertained deep neural network |
| **End-to-end learning** | Decrease variance |
| **Ensemble learning** | Training of a deep neural network |

**Table no.01. Algorithms along with their functionality used in machine learning/ Artificial learning.**

5) **Prediction of the target protein in drug designing:**

During the development of a drug molecules, it is essential to identify a solid target of drug molecule for the successful treatment. Number of proteins are involved in the disease, some of which are over expressed in either case. Hence for targeting of disease, it is most essential to identify the targeted protein for designing a drug molecule (Ekins et al., 2019). Artificial intelligence in this way is helpful in the prediction of 3D structure of proteins as the design would be quite similar to the chemical environment of target protein sites. So it will be helpful in prediction of the effect of compound on target disease in addition to the safety considerations before the synthesis or production of drug molecule. The tools based on DNN such as AI and alpha tools was used for the estimation of distance between the amino acids and corresponding angles between the peptides for prediction of accurate 3D structure and surprisingly 26 structures out of 43 were found to be accurate (Stephenson et al., 2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

In another study, researchers used RNN for prediction of protein structures considering three stages such as computation, geometry and assessment. It involved the encoding of the primary structure of protein followed by torsion angles and partial backbone of amino acids by geometric stage, upstream of which is encoded as input and a new back one was refered as output. The final unit presented 3D structure as output. Assessment of the produced structure in this way was done using the distance based root mean square deviation matrix (Rifaioglu et al., 2019). The parameters were optimized to keep the deviation root mean square in between the predicted and experimental structures. The author predicted that artificial intelligence based merged took small time in prediction of protein structures as compared to Alpha fold. But alpha fold possess better accuracy as compared to AI in predicting protein structures having most of the sequences similar to that of experimental one. A study was also conducted using MATLAB in addition to NN tool box based on feed forwarded supervise learning and back propagation error algorithms. MATLAB was used for training if input and output datasets while NNs were proved to be learning algorithms and were used for evaluation. The accuracy in prediction of the 2D structure was found to be 63% in this scenario (Zhang et al., 2017).

6) **Prediction of drug-protein interactions:**

Understanding the interaction between drug and protein is much important for the purpose of therapy. The prediction of the drug with the receptor is important for understanding the efficacy and significance for designing drugs as it prevents polypharmacology consequences. Artificial intelligence methods are used for the efficacy of therapy by prediction of accurate drug-protein interaction. Wang et al. reported the use of SVM approach for training of almost 15000 dug-protein interactions which were developed based on the primary sequences and structure of small characteristic molecules for finding out 9 new drug molecules along with their interactions with 4 new targets at the same time (Mirzaei, 2020). Yu et al employed two RF models for possible prediction of drug interaction by using pharmacological and chemical data against already present platforms and then validation them using SVM for specifity and sensitivity. All of these models were used for production of drug target association which in turn predict drug disease and target associations. Hence this results in the speeding of drug discovery process. Xiao et al., used synthetic minority over sampling techniques for obtaining optimized data in development of drug target. It further involve four sub descriptors for identification of G protein coupled receptors, ion channels, nuclear receptors respectively. This was further competed with jackknife predictors where the results showed that former suppressed the later in both efficiency and productivity (Jia, Li, Hao, & Yang, 2020).

This ability of artificial intelligence in predicting drug –drug and drug-protein interaction was further for repurposing and reformulation of the existing drug with the preventing any further pharmacological failure at the same time. Repurposing of drug makes it qualified directly for the Phase 2 of the drug clinical trial. This in addition to fast production, also reduces the cost or expenditure as the reformulation of existing drug or medicine is much more cost effective then formulation of a totally new drug molecule. The guilt by association approach can be used to figure the inventive relationship of a medication and infection or disease, which is either an information based or computationally determined approach (Gupta et al., 2021). In a computationally determined approach, the ML approach is broadly utilized, which uses strategies like SVM, NN, regression, and DL. Logistic regression involve the platforms , like PREDICT, SPACE, and other ML considering drug–drug, drug- target comparability, the closeness in between target atoms,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

compound design, and quality articulation profiles while repurposing a medication (Maltarollo, Kronenberger, Espinoza, Oliveira, & Honorio, 2019; Warmuth et al., 2003).

Cellular based deep learning technology involve the prediction of therapeutic uses of topotecan which in the research are being employed as topoisomerase inhibitors. It can also be used for prevention of multiple sclerosis via inhibition if human retinoic acid related receptors gamma. The platform is still under study of US patent. The unadministerd or unsupervised category of the algorithms include self-organizing maps SAM for the repurposing of drug molecules. They used a ligand based approach for several off targets of a set of drug molecules having recognized biological activities which oater can be used for different compounds. Recently a drug was repurposed having activity against novel corona virus strain SARC-CoV, or flu like diseases using protease inhibitors and is done by DNN. Here in this case, artificial intelligence platform used for training include extend connectivity fingerprints, functional class fingerprints and octanal water partition algorithms were used (Zernov, Balakin, Ivaschenko, Savchuk, & Pletnev, 2003).

This drug protein interaction can also result in production of off target drugs side effects by interaction of one drug molecule with a large number for receptors at the same time. Based on rationale of polypharmacology, this will be helpful in the prediction and design of a new drug molecule and hence results in generation of safer drug molecules (Burbidge, Trotter, Buxton, & Holden, 2001). Other AI platforms such as SOM along with other databases can be used to link number of compounds linked with a number of target molecules at the same time. Bays and SEA profiles can be used for the pharmacological target of drug with the possible targets at the same time. Li et al at the same time used kinomex based on the use of artificial intelligence using DNN for detection of kinases based on their chemical structures. The platform uses DNN with the higher bioactivity of 14000 and the amount of kinase based on it is about 300 (Ivanciuc, 2007). This is helpful practically for the study of any drug towards the kinase molecule. Thus will be helpful in designing novel chemical modifiers. The study uses NVP-BHG712 compound as model for predicting its accuracy and also to predict off targets with great accuracy. One of them may instances is Cyclical cloud based staining platform which is used to find receptors interacting with small molecules in order to predict off target interaction. Thus it will be helpful in finding out the possible side effects at the same time (Byvatov, Fechner, Sadowski, & Schneider, 2003).

7) **Artificial intelligence in property prediction:**

During drug discovery, the clinical drug molecules must meet a totally different criterion. Next to identification of target, the compound should be specific to a small amount of targets and also possess the properties such as absorption, excretion, metabolism and toxicity ADMET. Hence the optimization of the procedure is the multi-dimensional process and challenge. Various in silico produces are applied for this purpose and are employed along with optimization protocols for designing of drug compound (Gui, Pan, Lin, Li, & Yuan, 2017). In addition to it, several machine learning protocols are being employed such as SVM, random forest etc. One of the most important aspect of machine learning is the easy access to large data sets which make it presiquite for applying artificial learning. In the pharmacological industry, a large datasets are collected during the optimization of drug molecules for a number of different properties. These data sets for targets and anti-targets are available for a large number of chemical series and are used for training of machine learning programs for optimization of artificial learning techniques (S. Y. Kim et al., 2011).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

Prediction of activities against many kinase molecules is a great example. As in the case of using random forest derived from at least 200 different data sets combining with the other housekeeping datasets. Random forest showed better activity and productivity as compared to any other machine learning software. Only DNN showed the performance with greater sensitivity but with the least or no specifity (Mohammad, Sulaiman, & Khalaf, 2011). Hence the researchers prefer random forest soft wares because they are easy to train and use. Several recent reviews show the novel activities of machine learning. DNN have been used at many places for the property prediction. In a recent research, the comparison of DNN was performed in relative to other machine learning or artificial intelligence based techniques and it was shown that the FNN show better efficiency and activity regarding biological properties, ADMET and others to physiochemical parameters (Byvatov et al., 2003). As for example in Kaggle competition, DNN showed better results as compared to random foresting predicting 2D structure in an effective way. The study revealed that properties and results of DNN was proved to use hyper parameters utilizing architecture having a hidden number of layers and in turn a number of neurons in between and activation functions. Definition of these hidden functions is important for the efficient performances (Li, Meng, Cai, Yoshino, & Mochida, 2009).

Lemselik et al. performed the similar study showing better performance by DNN as compared to other using data set from ChEMBL. During this validation was done by temporal analysis where training and data sets are separated on the basis of their publication date. This way of validation is more straightforward. And the validation performance measures are much smaller than that of random split which is found to be much closer to real life productivity (Mitchell, Michalski, & Carbonell, 2013).

Deep learning has also been used for prediction of toxicity. As result from rox 12 competition showed that DNN showed better activity at 12 points as compared to other corresponding techniques. In thus study some of the emphasis was provided to molecular descriptors. Absence of one or more toxicophore was included in descriptors in addition to other physiochemical descriptors (Charniak, 1985). The results showed that DNN has the ability to extract molecular descriptors with known toxic elements and hence predict more descriptors in hidden layers as compared to others. **Fig 2** illustrates these features demonstrated by toxicophore. It has been found that relevant structural elements can be derived from DNN making drug discovery process easier without the involvement of human experties in field of toxicology. In addition to it the composition of training datasets also influence the productivity and applicability of domains of model by network possessing a large number of barriers in these automated learning procedures. The deep tox uses a pipeline of such models but it is also dominated by DNN prediction methods. It outperformed about 9 approaches out of 12 in learning toxic endpoints (McCarthy, 2007).

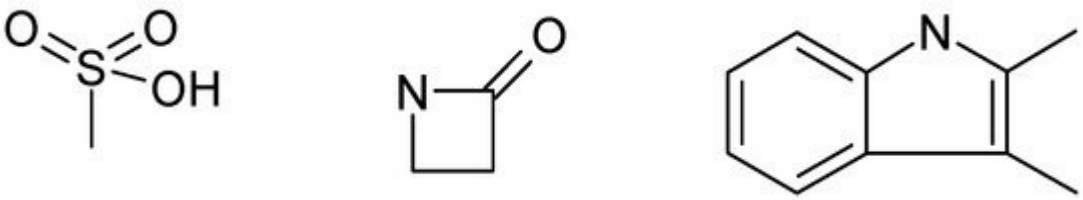


**Fig2 Features of toxicophore identified from Tox21 datasets**

In another example, the production of drug induced liver injury has been done for the prediction of toxic ends. Following this procedure, the network was trained for almost 476 compounds and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

the performance was done over 198 chemical compounds of medicinal chemistry. Several excellent parameters based in statistics can be achieved with having up to 89.6% accuracy, sensitivity of 83% and specifity of almost 94% (Thrall et al., 2018). The molecular descriptors from the two algorithms such as PaDEL and Mold were used utilizing the description from UG-RNN methodology for structural modeling of compounds in combination with the other methods such as bisection method. In this method, the descriptors are elicited from the chemical structures derived from graphs in the form of undirected structures. At the nodes, heavy atoms were presented but at the edges bond were relying (Stewart, Sprivulis, & Dwivedi, 2018). The graph thus obtained was fed to recursive neuronal network RNN as shown in **fig. 03**. The final output is summarized over several iterations. The descriptors obtained from it showed better progression as compared to the other two descriptor sets. Hence using neural network in the fields of cheminformatics is a novel trend. But the mist of the example shoes that chemical descriptor overcomes the results of classical descriptors by neural net (Das, Dey, Pal, & Roy, 2015).

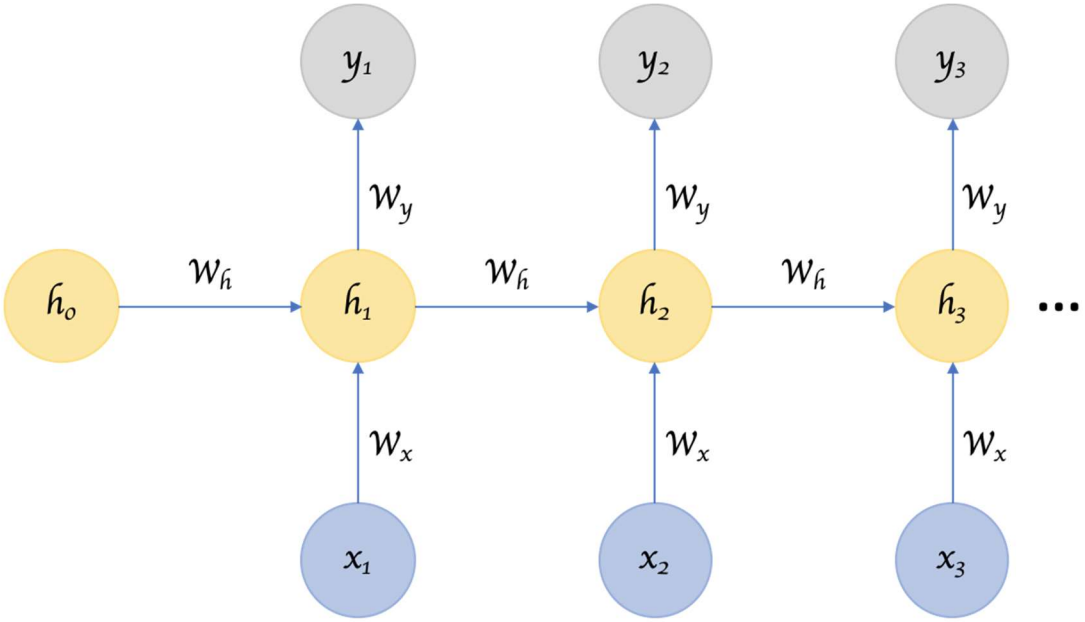


**Fig. 03. Recurrent neural network depiction the next one carrying data from previous one. The ho have hidden state and are updated on the basis of input entered.**

QSAR and machine learning are usually trained for one end point although a huge number of endpoints are available DNNS allows the combination of multitask learning based to combine the prediction of several endpoints. As multitask learning improves results and prediction to several orders, the fact elucidated by study which compared the multi task learning with that of single task at the same time (Ghahramani, 2015). Ramsundar et al. utilized the multitask learning for up to 2000 assays with the increase in performance of model has been observed in case of multitask learning procedure and in addition it appeared to be more stringer than the large number if tasks at the same time. The improvement in the datasets has also been observed when the analysis is done for several compounds at once and sharing of these active compounds with other sources. Multitask learning in turn is influenced by both amount of data and number if tasks. In another example industrial sized ADMET datasets benefits for multitask learning could be identified, although improvements proved to be dataset dependent (Shah et al., 2019).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

Conclusive remarks about the best performance software is based on the two types of validation including temporal and random split type. Only adding a large amount if data does not results in increase in the productivity. While multi task learning exerts positive effects on the number of data sets where the drop in the productivity has been observed easily (Ullah, Al-Turjman, Mostarda, & Gagliardi, 2020).Xou et al predicts that in multi task learning some of the information is borrowed from other sources or endpoints which make them improved method for increased efficiency and productivity. Deep learning has also been observed for explaining the potential energy of several compounds. Thus replacing the computational chemical method with that of fast learning methodology (Galbusera, Casaroli, & Bassani, 2019).

8) **De-Novo synthesis of Drug:**

As the name suggests, it involves the generation of a large number of new compounds which are active too without the previously found reference to other compounds and it has been discovered about 25 years ago (Méndez-Lucio, Baillif, Clevert, Rouquié, & Wichard, 2020). A large number of approaches and software solutions have been used for this purpose. But the de novo synthesis is not observed mostly in the drug discovery and is used for the compounds which are synthetically difficult to access in any way. Thus field has also been revolutionized due to involvement if artificial intelligence and machine learning (Bung, Krishnan, Bulusu, & Roy, 2021). One of the most important approach involving AI includes variation auto encoder being consist of artificial encoder, neural networks and in addition them an encoder network (Yu & Buehler, 2020). The encoder network is involved in the translation of the chemical structure produced by SMILES into the real value continuous vector. On the other hand decoder is involved in the translation of these vectors into chemical structures again. For most of the translational structures, any one if the molecules dominate having structural modifications with smallest probability. Researcher used the training model based on QED based drug likeness score and SAS. In this way, the path of molecules with improved target properties can be obtained (Merk, Grisoni, Friedrich, & Schneider, 2018) .

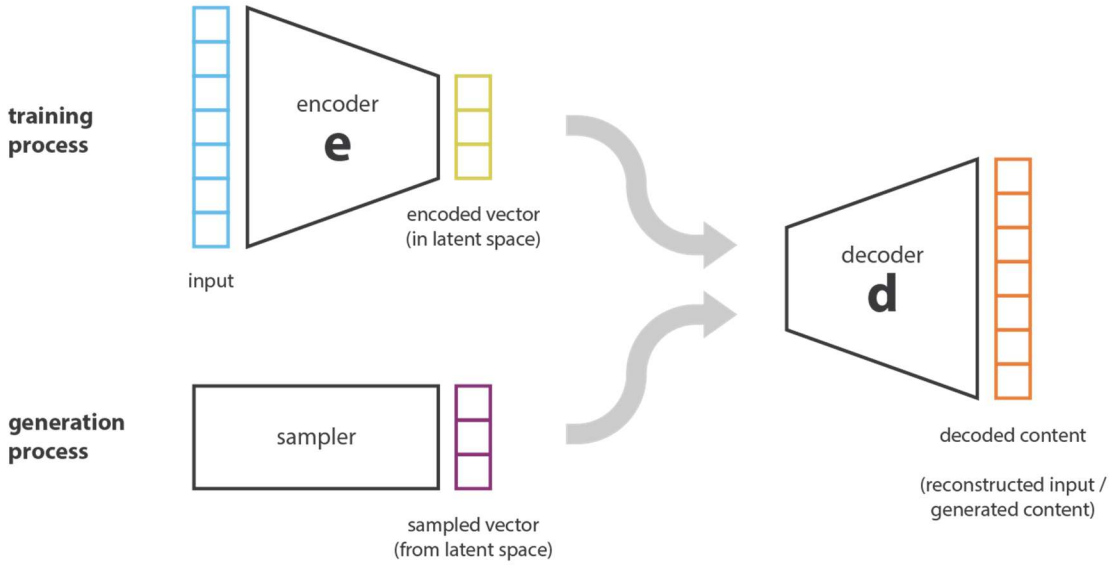


**Fig. 03. A vibrational auto encoder consists of two systems which transform input to encoder vector and sampled vector while decoder retranslate it into decoded content or output.**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

In another experiment the performance of vibrational auto encoder was compared to that of Adversarial auto encoders. The later comprised of generative model for production of novel chemical structures. The second discriminative model was trained to find out the differences in between real and that of synthetic molecules and it in turn produced more number of validated structures as compared to that of vibrational auto encoders (Button, Merk, Hiss, & Schneider, 2019). In addition to it, an in silico approach was used for prediction of active structure against dopamine type 2 receptors. Kadurin et al. employed the use of generative adversarial networks GAN for suggestion of the compounds with cancerous properties (Empel & Koenigs, 2019). Recursive neural network has been also used successfully for the de no o synthesis of drug molecules using artificial learning. Technically, they have been developed under the area of language learning in artificial intelligence. They take a small amount of information as an input, with the SMILES results in string of chemical structures in the form of latters (Struble et al., 2020). Hence in this way RNN can be used for the prediction of chemical structures. For acknowledging RNN with the grammar and language of SMILES, they are trained using chemical compounds being taken from existing chemical compounds which are either taken from CHEMBLE or commercially available. The similar approach was used for the generation of peptide structures. Bias generated compounds were treated with reinforcement structures for getting desired properties (Jiménez-Luna, Grisoni, Weskamp, & Schneider, 2021).

Another strategy involve transfer learning used for the generation of novel chemical compounds with the desired biological activities. In the first step, the network is allowed to learn the grammar of SMILES using a huge training set. In the second step training is done with the compound having desired property. Few additional trainings are also required for the generation of the compounds biased to be novel into the chemical space occupied by active molecules (Mak & Pichika, 2019). Based on such approach, five molecules were synthesized and design activity was confirmed for four of them using neural network harmone receptors. Hence several different architectures are used and trained which are capable of developing much meaningful and desired structures. Hence the chemical space for these novel compounds can be explored with the property distribution of generated molecules similar to that of training space. The first five applications of tis was performed with 5-6 molecules which later showed desired activity. But on conclusion more experience and training is needed according to size of sample and chemical feasibility of proposed molecules (de Almeida, Moreira, & Rodrigues, 2019).

9) **Use of artificial intelligence and machine learning in synthesis planning**

Organic molecules synthesis is the most critical part of drug discovery program. New molecules are synthesized continuously based on the optimization of previously existing compounds for the production of molecules with desired properties. But in certain situations, certain things result in the restriction of chemical space available for novel discovery of accessible compounds. Hence the synthesis planning is a key element in the drug discovery process (Coley, Green, & Jensen, 2018). According to it, new computational approaches have been developed for the assistance of synthetic planning. According to this three different aspects are to be distinguished which include,

- Prediction of outcomes of a reaction based on given set of inputs.
- Prediction of yield of chemical reactions as well as involvement of retrosynthetic planning.

The retrosynthetic planning is dominated by knowledge based system based on expert derived rules or from the rules directly derived from reaction datasets. Recent analysis proved that a large

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 May 2021 doi:10.20944/preprints202105.0567.v1

number of computational techniques have been used for forward synthesis prediction (Struble et al., 2020). They offer the ranking of routes from retro synthesis procedure. In one of the approach, quantum mechanics chemical descriptors are combined with manually encoded rules combined with machine learning for the prediction of reaction and its products at the end. This process or method is used for the prediction of multi-step analysis reaction. In another example of analytic approach, a much deeper neural reaction has been used for training of the reactions being present in Reaxys. This has dominated the previously present expert system for the prediction and has reported best accuracy of 79% for the reaction of 822 templates (E. Kim et al., 2020).

The publically obtained data sets have no concern with the chemical reactions which have gone failure and are not considered in prediction. So this becomes a limitation for machine learning approach. Therefore in another research the datasets for analysis were combined with the negative results of chemical reactions (Molga, Szymkuć, & Grzybowski, 2021). At first, the reactions were gone through and classified based upon neural network. On the basis of about 15000 training reactions were run according to US patent permission and only one was identified as the reaction giving product with accuracy of about 72%. In another research, scientist used an approach involving no template for the prediction of reaction and increasing coverage of chemical reactions being involved. Machine learning involved chemical reaction prediction showed improved learning as depicted in validated studies (Wipke, Ouchi, & Krishnan, 1978). But still this field further consideration for future development based on involvement of catalyst. The combination of three neural networks with that of Monte Carlo tree search has provided us with the excellent results at the end. For this purpose training and test datasets were taken from the Reaxys databases and then were splitted into the time. About 476 molecules were synthesized after 2015 and about 80% correctness was purposed (Fortunato, Coley, Barnes, & Jensen, 2020). Hence machine learning gas proved to be a valuable tool for handling of dataset which cannot he handled by the humans in an unbiased manner. For synthesis planning the combination of both artificial intelligence and machine learning based approaches in combination with knowledge based approach are proved to be useful for prediction of chemical reactions (Kishimoto, Buesser, Chen, & Botea, 2019). While using only machine learning approach for huge dataset is proved to be an excellent approach with great productivity. But still detailed analysis is necessary for the use of quantum chemical methods for use in future (Napoli, Laurenço, & Ducournau, 1994).

## 10) **Conclusion:**

The continuous involvement of artificial intelligence and machine leaning in the field of drug discovery only aims to reduce the challenges faced by pharmaceutical companies. Hence have a strong impact over the drug development process along with the life cycle of drug molecule which will results in new startups in this field. Currently the medical field is facing a large number of challenges which include increased cost of drugs or therapies in one other ways. With the involvement of artificial intelligence and machine learning in manufacturing of drugs, personalized medications with desired dosage and ADME features can be developed according to need of individual patient any time. This involvement not only enhances the quality of the products but also decreases the time needed by drug to be discovered. In addition it also ensures the safety of the production process but also provide the better utilization of resources along with making it cost effective and lastly increases the importance of automation of a reaction. This technique not only provide quick identification of drug molecule but in addition also contributes to further routes being suggested for synthesis of these compounds along with identification of target and drug – protein interactions. Furthermore artificial intelligence also contributes to optimization of

candidate drug molecule or reaction for predicting correct dosage. It also ensures batch to batch consistency along with the improved and quick descion, leading to prediction and production of better quality drug in small time. It also ensures the accuracy of these compounds in the clinical trials based on its algorithms as well as also ensure proper positioning and cost of drug in the market. Although there are no drugs available prepared on the basis of artificial learning in market, and still there are specific challenges faced by scientists to implement these procedures, it is most probably predicted that artificial intelligence and machine leaning will become a invaluable tool by pharmaceutical industry in the near future.

## 11. Conflict of Interest.

Authors declare no conflict of interest.

## References:


1. Bung, N., Krishnan, S. R., Bulusu, G., & Roy, A. (2021). De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future medicinal chemistry, 13*(06), 575-585.
2. Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & chemistry, 26*(1), 5-14.
3. Button, A., Merk, D., Hiss, J. A., & Schneider, G. (2019). Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature machine intelligence, 1*(7), 307-315.
4. Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences, 43*(6), 1882-1889.
5. Chan, H. S., Shan, H., Dahoun, T., Vogel, H., & Yuan, S. (2019). Advancing drug discovery via artificial intelligence. *Trends in pharmacological sciences, 40*(8), 592-604.
6. Charniak, E. (1985). *Introduction to artificial intelligence*: Pearson Education India.
7. 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.
8. Coley, C. W., Green, W. H., & Jensen, K. F. (2018). Machine learning in computer-aided synthesis planning. *Accounts of chemical research, 51*(5), 1281-1289.
9. Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications, 115*(9).
10. de Almeida, A. F., Moreira, R., & Rodrigues, T. (2019). Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry, 3*(10), 589-604.
11. Díaz, Ó., Dalton, J. A., & Giraldo, J. (2019). Artificial intelligence: a novel approach for drug discovery. *Trends in pharmacological sciences, 40*(8), 550-551.
12. Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., . . . Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature materials, 18*(5), 435-441.

13. Empel, C., & Koenigs, R. M. (2019). Artificial-Intelligence-Driven Organic Synthesis—En Route towards Autonomous Synthesis? *Angewandte Chemie International Edition, 58*(48), 17114-17116.
14. Fortunato, M. E., Coley, C. W., Barnes, B. C., & Jensen, K. F. (2020). Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. *Journal of chemical information and modeling, 60*(7), 3398-3407.
15. Galbusera, F., Casaroli, G., & Bassani, T. (2019). Artificial intelligence and machine learning in spine research. *JOR spine, 2*(1), e1044.
16. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452-459.
17. Gui, G., Pan, H., Lin, Z., Li, Y., & Yuan, Z. (2017). Data-driven support vector machine with optimization techniques for structural health monitoring and damage detection. *KSCE Journal of Civil Engineering, 21*(2), 523-534.
18. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, 1-46.
19. Ivanciuc, O. (2007). Applications of support vector machines in chemistry. *Reviews in computational chemistry, 23*, 291.
20. Jia, C.-Y., Li, J.-Y., Hao, G.-F., & Yang, G.-F. (2020). A drug-likeness toolbox facilitates ADMET study in drug discovery. *Drug discovery today, 25*(1), 248-258.
21. Jiménez-Luna, J., Grisoni, F., Weskamp, N., & Schneider, G. (2021). Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opinion on Drug Discovery*, 1-11.
22. Jing, Y., Bian, Y., Hu, Z., Wang, L., & Xie, X.-Q. S. (2018). Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The AAPS journal, 20*(3), 1-10.
23. Kim, E., Jensen, Z., van Grootel, A., Huang, K., Staib, M., Mysore, S., . . . Jegelka, S. (2020). Inorganic materials synthesis planning with literature-trained neural networks. *Journal of chemical information and modeling, 60*(3), 1194-1201.
24. Kim, S. Y., Moon, S. K., Jung, D. C., Hwang, S. I., Sung, C. K., Cho, J. Y., . . . Lee, H. J. (2011). Pre-operative prediction of advanced prostatic cancer using clinical decision support systems: accuracy comparison between support vector machine and artificial neural network. *Korean journal of radiology, 12*(5), 588.
25. Kishimoto, A., Buesser, B., Chen, B., & Botea, A. (2019). Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning.
26. Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. *Energy Conversion and Management, 50*(1), 90-96.
27. Mak, K.-K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. *Drug discovery today, 24*(3), 773-780.
28. Maltarollo, V. G., Kronenberger, T., Espinoza, G. Z., Oliveira, P. R., & Honorio, K. M. (2019). Advances with support vector machines for novel drug discovery. *Expert opinion on drug discovery, 14*(1), 23-33.
29. McCarthy, J. (2007). What is artificial intelligence.

30. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., & Wichard, J. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications, 11*(1), 1-10.
31. Merk, D., Grisoni, F., Friedrich, L., & Schneider, G. (2018). Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Communications Chemistry, 1*(1), 1-9.
32. Mirzaei, M. (2020). Drug discovery: a non-expiring process. *Advanced Journal of Chemistry-Section B, 2*(2), 46-47.
33. Mitchell, R., Michalski, J., & Carbonell, T. (2013). *An artificial intelligence approach*: Springer.
34. Mohammad, M. N., Sulaiman, N., & Khalaf, E. T. (2011). A novel local network intrusion detection system based on support vector machine. *Journal of Computer Science, 7*(10), 1560.
35. Molga, K., Szymkuć, S., & Grzybowski, B. A. (2021). Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Accounts of chemical research, 54*(5), 1094-1106.
36. Napoli, A., Laurenço, C., & Ducournau, R. (1994). An object-based representation system for organic synthesis planning. *International Journal of Human-Computer Studies, 41*(1-2), 5-32.
37. Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in bioinformatics, 20*(5), 1878-1912.
38. Saikin, S. K., Kreisbeck, C., Sheberla, D., Becker, J. S., & Aspuru-Guzik, A. (2019). Closed-loop discovery platform integration is needed for artificial intelligence to make an impact in drug discovery. *Expert opinion on drug discovery, 14*(1), 1-4.
39. Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., . . . Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ digital medicine, 2*(1), 1-5.
40. Smith, J. S., Roitberg, A. E., & Isayev, O. (2018). Transforming computational drug discovery with machine learning and AI. In: ACS Publications.
41. Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., . . . Cao, R. (2019). Survey of machine learning techniques in drug discovery. *Current drug metabolism, 20*(3), 185-193.
42. Stewart, J., Sprivulis, P., & Dwivedi, G. (2018). Artificial intelligence and machine learning in emergency medicine. *Emergency Medicine Australasia, 30*(6), 870-874.
43. Struble, T. J., Alvarez, J. C., Brown, S. P., Chytil, M., Cisar, J., DesJarlais, R. L., . . . Griffin, D. J. (2020). Current and future roles of artificial intelligence in medicinal chemistry synthesis. *Journal of medicinal chemistry, 63*(16), 8667-8682.
44. Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology, 15*(3), 504-508.
45. Ullah, Z., Al-Turjman, F., Mostarda, L., & Gagliardi, R. (2020). Applications of artificial intelligence and machine learning in smart cities. *Computer Communications, 154*, 313-323.
46. Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., & Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences, 43*(2), 667-673.

 doi:10.20944/preprints202105.0567.v1

47. Wipke, W. T., Ouchi, G. I., & Krishnan, S. (1978). Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artificial Intelligence, 11*(1-2), 173-193.
48. Yang, X., Wang, Y., Byrne, R., Schneider, G., & Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews, 119*(18), 10520-10594.
49. Yu, C.-H., & Buehler, M. J. (2020). Sonification based de novo protein design using artificial intelligence, structure prediction, and analysis using molecular modeling. *APL bioengineering, 4*(1), 016108.
50. Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., & Pletnev, I. V. (2003). Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of chemical information and computer sciences, 43*(6), 2048-2056.
51. Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today, 22*(11), 1680-1685.
52. Zhavoronkov, A. (2018). Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. In: ACS Publications.
53. Zhavoronkov, A., Vanhaelen, Q., & Oprea, T. I. (2020). Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? *Clinical Pharmacology & Therapeutics, 107*(4), 780-785.