EAS Journal of Humanities and Cultural Studies
Abbreviated Key Title: EAS J Humanit Cult Stud
ISSN: 2663-0958 (Print) & ISSN: 2663-6743 (Online)
Published By East African Scholars Publisher, Kenya

OPEN ACCESS

Volume-7 | Issue-3 | May-Jun-2025 | DOI: https://doi.org/10.36349/easjhcs.2025.v07i03.005

Review Article

AI & Machine Learning in Lead Discovery: Deep-Learning Architectures for *de novo* Design, Property Prediction and Inverse QSAR

Arnav Kumar[1], Subhas S Karki[2*]

[1]Research Scholar, KLE College of Pharmacy, P.B. No 1062, II block Rajajinagar, Bengaluru-560010, India
[2]Professor & Head, Department of Pharmaceutical Chemistry, KLE College of Pharmacy, Bangalore, India

Article History
Received: 22.04.2025
Accepted: 29.05.2025
Published: 31.05.2025

Journal homepage:
https://www.easpublisher.com

Quick Response Code

Abstract: Traditional lead discovery has relied on quantitative structure–activity relationships (QSAR) and physics-based screening, but exhaustively searching chemical space is infeasible. Modern workflows therefore employ deep learning to learn predictive structure–property mappings. Graph neural networks (GNNs) and transformer models have become widely adopted for molecular property prediction and design, as they natively operate on graph-structured or sequential chemical representations. Variational autoencoders, generative adversarial networks and related architectures embed molecules in continuous latent spaces, enabling inverse QSAR: one can sample or optimize structures to match target bioactivity and physicochemical criteria. These generative models can propose entirely new scaffolds with desired attributes, effectively 'designing' candidate leads beyond known libraries. Despite these advances, significant challenges remain. Data sparsity and bias limit model robustness, and many molecular properties (e.g. ADMET endpoints) are measured on limited datasets. Interpretability is limited – deep models often act as black boxes, motivating development of explainable AI techniques. Ensuring scalability to ultra-large libraries and embedding chemical constraints (synthetic feasibility, drug-likeness) is nontrivial. Moreover, lead optimization is inherently multi-objective: models must balance potency, selectivity, and pharmacokinetics, requiring complex trade-offs during design. Looking ahead, emerging strategies promise to address these gaps. Self-supervised pretraining on massive unlabelled chemical corpora is improving feature learning, while explainable AI methods aim to highlight key substructures driving predictions. Early quantum-enhanced machine learning frameworks show promise for accelerating optimization and generation of candidates. Multimodal models that integrate chemical structure with biological assays and omics data may yield richer lead profiles. Federated learning is beginning to enable collaborative QSAR without sharing proprietary data – recent work shows federated QSAR matches the performance of pooled-data models under privacy constraints. In sum, these technical advances in deep architectures and data paradigms are poised to transform AI-driven lead discovery, making de novo design and property prediction more predictive and efficient than ever.

Keywords: Lead discovery, Deep learning, Graph neural networks, Transformer models, Inverse QSAR, Generative modelling.

INTRODUCTION

The pharmaceutical development process is known for being lengthy, costly, and prone to failure. Advancing a single drug candidate to market can require more than ten years and a price tag in the billions, with success rates in clinical trials generally falling below 10 percent [1]. The conventional drug discovery process faces significant challenges due to very high attrition rates—approximately only 10% of compounds that enter preclinical studies make it to clinical use—and it is burdened by substantial costs and prolonged timelines. The typical expenditure to develop a new molecular entity from the initial target identification to FDA approval exceeds $2 billion, and the development period frequently goes beyond 10 to 15 years [2,3]. These inefficiencies arise from dependence on extensive high-

***Corresponding Author:** Subhas S Karki
Professor & Head, Department of Pharmaceutical Chemistry, KLE College of Pharmacy, Bangalore, India

throughput screening of large chemical libraries, repeated cycles of medicinal chemistry, and tedious in vitro/in vivo assays, all of which result in low success rates, increasing costs, and postponed availability of new therapies for patients [4]. Artificial intelligence (AI) and machine learning (ML) aim to alleviate these bottlenecks by rapidly sifting through chemical space and predicting biological activity with higher accuracy. Early computational approaches in medicinal chemistry (quantitative structure–activity relationships, QSAR) relied on engineered descriptors and statistical models. In recent years, deep learning (DL) – where models learn features directly from raw representations – has revolutionized this field[5]. Contemporary AI technologies have the capability to automatically identify intricate patterns within chemical data, facilitating quicker identification and enhancement of leads.

Drug discovery challenges: Identifying new drug candidates presents significant challenges, as it requires the evaluation of extensive chemical libraries and the refinement of promising compounds, both of which are expensive and time-consuming undertakings. Loss of potential candidates often occurs due to inadequate pharmacokinetic properties or toxicity issues encountered in later development phases. The implementation of AI/ML is viewed as a potential means to mitigate risks in this process by analyzing historical data. In fact, recent literature highlights that advancements in AI/ML have the potential to speed up development, reduce expenses, and enhance success rates[1]. For instance, artificial intelligence can rank compounds that meet specific criteria prior to their synthesis. Nonetheless, leveraging AI necessitates a comprehension of both its capabilities and constraints regarding pharmaceutical research and development.

From QSAR to deep learning: Conventional QSAR models connect predefined molecular descriptors or fingerprints to biological activity or characteristics through regression or classification techniques. While these approaches have proven effective (for instance, in forecasting solubility, lipophilicity, or toxicity), they depend on descriptors created by experts and typically only identify linear associations. The emergence of deep neural networks has allowed models to automatically extract features directly from unprocessed inputs. Recurrent neural networks (RNNs) are utilized to process SMILES strings[5] and Graph convolutional networks applied to molecular graphs have surpassed previous methods in numerous tasks. The domain has quickly evolved from basic feed-forward networks to intricate architectures (such as graph neural networks, transformers, variational autoencoders, and generative adversarial networks). These advancements not only enhance property prediction but also facilitate the creation of completely novel molecules.

This review examines the transition from traditional QSAR to cutting-edge deep learning architectures and explores significant AI approaches for de novo molecular design, predicting properties, and generating "inverse QSAR." We will also address hybrid AI workflows, existing challenges (such as data sparsity, model interpretability, scalability, and synthetic feasibility), and potential future directions including self-supervised pretraining, explainable AI, and quantum-enhanced machine learning.

In the following sections, we initially outline important machine learning techniques applied in chemistry, then consider their uses in (1) de novo molecular design, (2) property prediction, and (3) inverse QSAR. Finally, we wrap up with integrated design pipelines, key challenges, and future outlooks.

Rise of AI/ML in Lead Discovery

To tackle these issues, computational methods have progressed from the initial quantitative structure–activity relationships (QSAR) developed in the 1960s to contemporary machine-learning (ML) and deep-learning (DL) techniques.

Era	Methodology	Key Milestones
1960s–1980s	Classical QSAR	Hansch & Fujita (1964); Free–Wilson (1964)
1990s–2010	Machine Learning (SVM, RF, k-NN, etc.)	Emergence of SVM for activity prediction; Random Forest QSAR[4]
2012–present	Deep Learning (CNNs, RNNs, GNNs)	AlphaFold (2018); GENTRL for de novo design[6]

QSAR established a foundation by linking molecular descriptors to biological activity. Later advancements in machine learning techniques, such as support vector machines and random forests, enhanced predictive accuracy and generalizability. In recent times, deep learning models—including graph neural networks for molecular graphs and variational autoencoders for generative modeling—have shown remarkable efficacy in both property prediction and the generation of novel molecules [6,8].

Fundamentals of Machine Learning Techniques

Machine learning (ML) has emerged as an essential resource in the field of chemistry, driving progress in numerous applications like de novo molecular design, property forecasting, and quantitative structure-activity relationship (QSAR) modeling. The combination of ML techniques with chemically relevant descriptors greatly improves the predictive power of these models, rendering them crucial for the fields of chemistry, biology, and materials science [9].

Supervised Learning

Supervised learning is a key machine learning technique in which models are educated using labeled datasets, where each input is associated with a particular output. This method is often used for predictive tasks, aiming to uncover relationships from past data to predict future results. Prominent techniques in supervised learning consist of regression analysis, decision trees, and support vector machines, which are widely applied for different predictive tasks in the field of chemistry [10,11]. The effectiveness of supervised learning models is heavily dependent on the quality of the data utilized, making it essential to conduct thorough data preprocessing and feature selection to enhance results [12,13]. This issue is made worse by the "Curse of Dimensionality," which indicates that adding more features can result in reduced model effectiveness because of sparse data distribution [14].

Unsupervised Learning

In contrast, unsupervised learning consists of training models using data that lacks labeled results, enabling the identification of concealed patterns or clusters within the dataset. Methods like clustering and dimensionality reduction are essential for investigating intricate chemical datasets [13]. Unsupervised approaches have been effective in examining simulation data and conducting exploratory data analysis, providing insights into the chemical landscape that may not be easily discerned through supervised methods [12,13].

Deep Learning

Deep learning, which is a branch of machine learning that utilizes neural networks with several layers, has become increasingly popular in predicting property values and designing molecules. Recent progress in geometric deep learning frameworks allows for the modeling of intricate molecular structures and characteristics, leading to precise predictions for a diverse array of chemical compounds [15]. Deep learning models offer significant benefits due to their capability to understand complex relationships within extensive datasets, which improves predictive accuracy for thermochemical properties as well as molecular interactions (Fig.1).

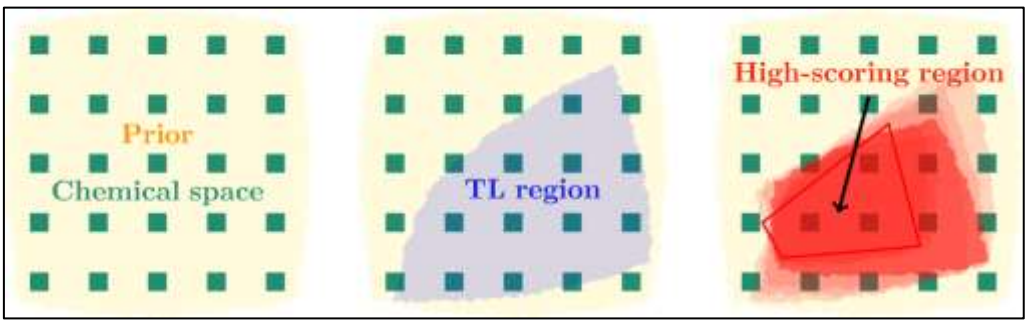

Fig.1: Illustration of idealized behaviour of priors, transfer learning agents and deep learning/staged learning agents. In all cases, the models describe the probability of sampling a given token sequence corresponding to a specific molecule (green squares), represented by a coloured fill. The prior model is trained to increase probability over all drug–like molecules. A transfer learning agent built from this prior increases the likelihood on a specific region (blue, middle). In staged learning (red, right), starting from the transfer learning agent, likelihood of sampling high-scoring sequences is iteratively increased, resulting in concentration on high-scoring regions (red polygon)

Applications in Molecular Design and Property Prediction

The incorporation of these machine learning methods into de novo molecular design and property prediction has yielded encouraging outcomes. Sophisticated algorithms are being utilized to enhance drug formulation processes and to forecast the physicochemical characteristics of new compounds more accurately than conventional approaches [16]. By leveraging extensive datasets, which include experimental information from trustworthy sources, machine learning models can aid in discovering and defining new materials [17].

Fundamental ML Techniques in Chemistry
Descriptor-based Methods

Classical cheminformatics starts with the use of molecular descriptors and fingerprints. Descriptors encompass properties such as molecular weight, logP, and polar surface area, as well as the counts of various fragments, while fingerprints represent substructures in the form of bit strings (e.g., Morgan/ECFP fingerprints). These elements act as inputs for machine learning algorithms (such as random forests and support vector machines) utilized for either regression or classification within the QSAR/QSPR framework. Descriptor-based QSAR continues to hold significance, particularly in areas with limited datasets. A recent investigation into ADME-Tox models revealed that traditional 1D/2D/3D descriptors combined with gradient-boosted trees frequently produced optimal results on medium-sized datasets. In fact, the selection and preprocessing of descriptors (including feature selection and dimensionality reduction) can significantly influence the accuracy of the model [18. Nevertheless, descriptor models rely on human knowledge and might overlook intricate nonlinear relationships between structure and properties.

Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as the leading approach for representing molecules. In a GNN, a molecule is represented as a graph where the atoms are the nodes and the bonds make up the edges. Through multiple "message-passing" iterations, the embeddings of the nodes are progressively refined by gathering information from their neighbouring nodes. This process enables the network to acquire high-level features from the entire molecular graph without relying on predefined descriptors. GNNs have demonstrated remarkable effectiveness in predicting molecular properties; when provided with sufficient training data, they frequently surpass traditional models that depend on descriptors [19]. For instance, GNNs are particularly effective at forecasting ADMET characteristics (absorption, distribution, metabolism, excretion, toxicity) that are important in drug development [20]. They have also been utilized for activities such as predicting binding affinity and forecasting reaction outcomes. One limitation is that GNNs typically need quite extensive datasets to realize their full capabilities – training a deep graph model often necessitates thousands of labelled instances [20]. However, after they are trained, GNNs can swiftly assess numerous potential compounds during virtual screening.

Transformer Models

Transformer models, initially created for natural language tasks, have been modified for use with chemical information. In models that use SMILES notation, a transformer analyses the sequence of characters in a SMILES string through self-attention mechanisms. Extensive transformer models (such as ChemBERTa [21], Chemformer) models are frequently pretrained using a self-supervised approach on vast quantities of unlabeled SMILES. For example, ChemBERTa-2 [22] was trained on a carefully selected collection of 77 million PubChem SMILES and attained competitive accuracy in property prediction on MoleculeNet benchmarks [23]. These models successfully acquire a chemical language: the attention mechanism detects relationships between substructures and context, and can recognize functional motifs. Similar to graph neural networks, transformer-based models require a large amount of data for training, yet they gain from transfer learning: an individual pretrained model can be adapted for various downstream tasks.

Autoencoders and Variational Autoencoders (VAEs)

Autoencoders reduce molecular representations into a continuous latent vector space and subsequently decode them to recreate the molecule. A significant variation, the variational autoencoder (VAE), captures a probabilistic distribution in the latent space. In the field of chemistry, VAEs have been developed using either SMILES strings or graph representations. For graph-based VAEs, models such as the Junction-Tree VAE (JT-VAE, Jin *et al.,* 2018) have been implemented [24]. Initially, molecules are represented as a hierarchical structure of chemical subcomponents, which are subsequently refined into a valid molecular form. This leads to a continuous latent space that facilitates both interpolation and gradient-based optimization of molecular configurations. Importantly, Variational Autoencoders (VAEs) can be tailored based on specific properties: it is possible to sample within the latent space close to vectors associated with favorable activities, or to incorporate predictors in the decoding procedure [24,25]. Inverse QSAR approaches frequently utilize conditional VAEs to create molecules that possess specific descriptors or predicted targets [22]. The decoder network within a VAE acts as a trained molecular generator, generating valid chemical structures from latent points.

Generative Adversarial Networks (GANs)

Generative adversarial networks operate through a two-player framework: a generator network creates molecules while a discriminator network is trained to tell apart the generated molecules from genuine ones. In the field of chemistry, GANs have been utilized for generating SMILES or molecular graphs (such as MolGAN, 2018) [26]. The adversarial framework has the potential to yield crisp (high-quality) samples. Nonetheless, training GANs is recognized for its instability, and there is a possibility of mode collapse [27]. In application, molecular generators that are based on GANs frequently integrate reinforcement learning to uphold validity and property limitations. Unlike VAEs, GANs do not explicitly capture a distribution of latent space, resulting in a more complex approach for inverse design [26,28]. Nevertheless, GANs are still one category of deep generative models utilized for molecules [29].

1. *de novo* Molecular Design

De novo design involves creating completely new chemical structures that possess specific properties, without relying on pre-existing molecules. Deep generative models have emerged as effective tools for accomplishing this objective [30]. Initial methods utilized RNNs (LSTM networks) to create SMILES strings, which were trained on chemical libraries; these models are capable of generating new chemically valid strings and have established a robust baseline [31,32]. Lately, generative models that use graphs (such as VAEs, normalizing flows, and GANs) along with transformer-based decoders have been developed.

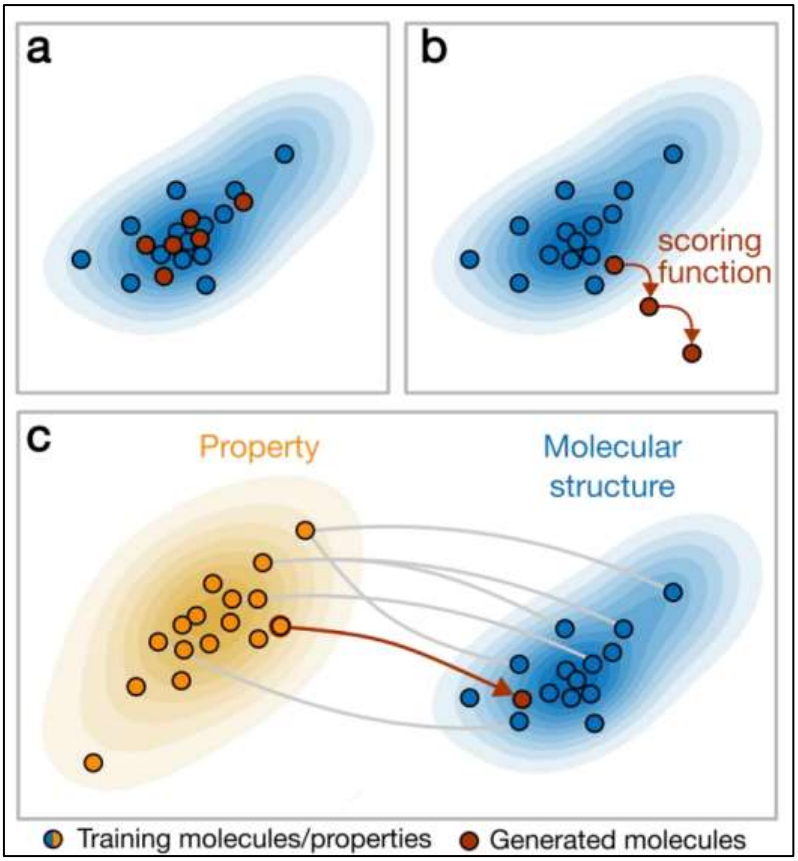

Fig. 2: Approaches for model training and molecular generation. (a) Distribution learning, where models generate molecules that statistically resemble those in the training set in terms of physico-chemical and biological properties. (b) Goal-directed generation, which optimizes molecules towards a pre-defined objective, often using reinforcement learning, guided by a scoring function. (c) Conditional generation, where models are explicitly trained to design molecules with specified properties, by incorporating property constraints into the generation process

A significant challenge in generative design is assessment. Standardized tasks and metrics (such as validity, uniqueness, novelty, and Fréchet ChemNet Distance) are provided by benchmarks like MOSES (Molecular Sets) and GuacaMol (2019) to facilitate the comparison of models [32]. For instance, Polykovskiy *et al.*, demonstrated that a character-level RNN trained on SMILES (CharRNN) attained the highest Fréchet ChemNet Distance compared to various baselines, suggesting it effectively mirrored the training distribution [33]. The research indicated that VAE models frequently suffer from overfitting, which results in low novelty, as many of the generated molecules closely resemble those in the training set [22]. Notably, CharRNN not only achieved similar property distributions but also identified several new scaffolds (approximately 11% scaffold novelty). Overall, contemporary generative models can closely replicate the property distributions found in the training set, including molecular weight, logP, druglikeness, and synthetic-accessibility. A common approach involves generating a large dataset (for instance, 30,000 molecules) and evaluating validity (which should be close to 100% for models that are well-trained), novelty (the proportion of samples not present in the training set), internal diversity, and resemblance to established reference sets [35].

Generative architectures:

Table 1 provides an overview of different model categories for de novo design. RNNs handle SMILES sequences in an autoregressive manner, processing one character at a time, whereas Transformers leverage self-attention with positional encodings for SMILES. Graph-based autoencoders (like JT-VAE) and normalizing flows (such as GraphAF) are also included [36]. Molecular structures can be created incrementally, atom by atom. Recently, diffusion models, originally developed for image generation, have been modified for molecular applications (learning to revert a noise-influenced process). Certain models produce 3D coordinates directly; for example, Li *et al.*, presented L-Net and DeepLigBuilder, a graph-generating VAE designed for the complete design of 3D drug-like molecules that fit into protein binding pockets [37]. In a research study regarding the main protease of SARS-CoV-2, DeepLigBuilder suggested new compounds that exhibit high predicted affinity and binding characteristics akin to established inhibitors [37,38].

Active design and case studies: Reinforcement learning can guide generators to achieve specific goals. For instance, policy gradients can be applied to influence a SMILES-RNN or a latent-vector generator to optimize a

reward (such as predicted affinity or QED) [39]. Haddad *et al.*, (2025) introduced a proximal policy optimization technique that functions within the latent space of a pretrained Variational Autoencoder (VAE). By exploring the latent space for areas linked to high-value properties, they achieved performance that matched or surpassed previous methods on standard optimization tasks [40]. Interestingly, their reinforcement learning approach was able to impose a substructure constraint while enhancing properties, representing a sophisticated "scaffold-constrained" generation pertinent to lead optimization [41].

The body of literature includes various case studies that demonstrate de novo generation followed by experimental confirmation. For example, Ivanenkov and colleagues employed deep generative RNNs to suggest DDR1 kinase inhibitors, which progressed to lead optimization in collaboration with in silico Medicine (2019) [42]. In recent times, design influenced by AI was utilized for the COVID-19 Mpro target [43,44], using hybrid generative-docking approaches (such as DeepLigBuilder) showcases the process: (i) create extensive virtual libraries of innovative molecules, (ii) assess them using predictive models or docking techniques, (iii) rank the top candidates for synthesis and testing. New AI-enhanced platforms are designed to automate this process in a continuous cycle, combining generative models, predictive assessment, and iterative refinement. Within an active-learning framework, the model is regularly updated with the latest experimental results, thereby enhancing its recommendations [41].

Table 1: Representative generative model architectures for *de novo* design

Model Category	Representation	Example Model(s)	Key Feature / Notes
Recurrent (SMILES)	SMILES sequence	CharRNN (SMILES LSTM)	Autoregressive SMILES generation; simple baseline with high validity and novelty [33].
Transformer (SMILES)	SMILES sequence	Chemformer, ChemBERTa	Self-attention on SMILES; benefits from large-scale pretraining [22].
VAE (Graph)	Molecular graph	JT-VAE	Encodes molecule to latent vector; decodes via substructure tree to generate valid graphs [45].
GAN (Graph)	Molecular graph	MolGAN	Adversarial training; can incorporate property predictors or rewards [46].
Flow (Graph)	Molecular graph	GraphAF	Autoregressive normalizing flow; exact likelihood and tractable inference [36].
Diffusion (Graph/3D)	Graph or 3D	GraphDF	Generative diffusion processes to produce molecular graphs or conformers (emerging) [36].
3D Molecular (Graph)	3D atomic coords	DeepLigBuilder (L-Net)	Generates 3D ligand structures within target sites; combined with docking/MCTS for design [37].
Reinforcement Learning	SMILES/Graph	REINVENT	Optimizes objectives via policy gradients; can bias generation toward specific properties [44].

2. Property Prediction and Benchmarks

Deep learning is extensively applied to forecast molecular characteristics (such as logP, solubility, binding affinity, toxicity, etc.) based on their structure. In real-world applications, researchers typically evaluate new techniques using standardized datasets. A notable compilation in this area is MoleculeNet [48]- - A collection of public benchmarks (ESOL, FreeSolv, Lipophilicity, HIV, Tox21, BACE, ClinTox, etc.) is available for assessing models on regression or classification tasks. Additional significant datasets consist of PCQM4Mv2 [49] (a substantial quantum chemistry dataset for predicting HOMO/LUMO), Open Graph Benchmark - Drug Discovery (OGB-GDD) [50]. Assignments and exclusive internal datasets. Table 2 enumerates typical sources and activities. Molecular property datasets vary from containing thousands to millions of compounds. ChEMBL [51], which includes bioactivity measurements, and PubChem [52] supply millions of annotated instances for different objectives. Virtual databases such as ZINC house numerous drug-like substances (primarily for synthesis). More focused, smaller collections emphasize particular endpoints. For example, ESOL [53] contains approximately 1,100 substances with determined water solubility, FreeSolv [54] Approximately 600 hydration free energies, QM955 [55] There are 134,000 small molecules for which properties have been computed using DFT. The Tox21 initiative and similar challenges include approximately 7,000 compounds that have been labeled for toxicity across various assays.

Predictive models encompass GNNs, transformers, and traditional approaches. Typically, GNNs that utilize extensive features tend to exceed the performance of fingerprint + machine learning baselines when there is ample data [19]. Chithrananda *et al.*, demonstrated that pretraining transformers (ChemBERTa) using 77 million SMILES (Fig.3) results in representations that are comparable to GNNs on tasks within MoleculeNet [22]. Nevertheless, traditional techniques continue to excel with small datasets: the combinations of descriptors and XGBoost models referenced earlier frequently outperform deep learning models on smaller ADME datasets [18]. Consequently, choosing a model relies on the data at hand and the intricacy of the task.

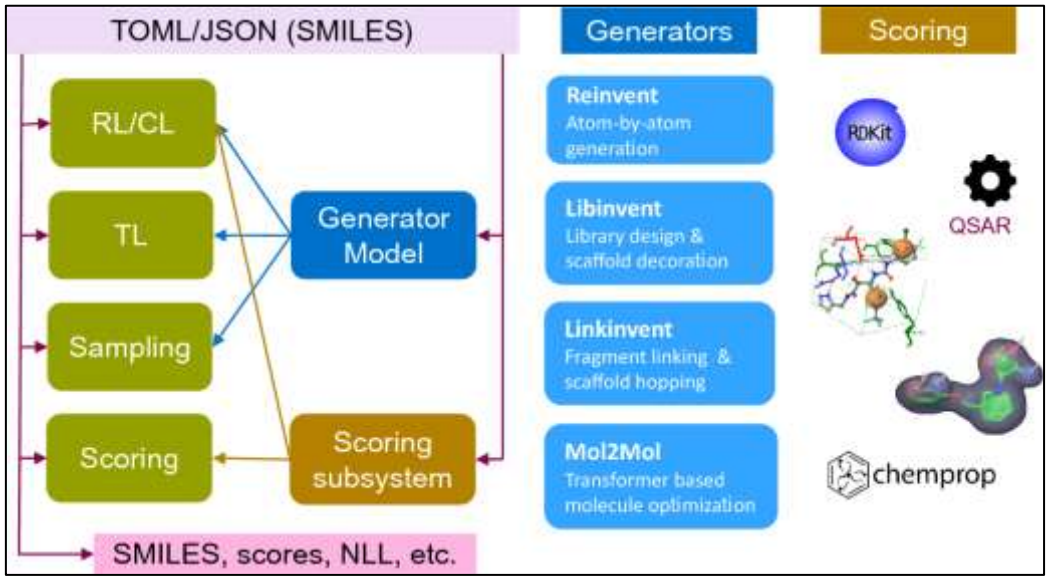

Fig.3: Information flow in REINVENT 4 for all run modes (green boxes) depicted in the left row. Also shown are the supported generators and the scoring subsystem. A input configuration file in TOML or JSON format controls all aspects of the software. The configuration file may contain "seed" SMILES for the Lib/Linkinvent and Mol2Mol2 generators. Input SMILES strings are needed for staged learning, TL and scoring.

Assessment through benchmarking is essential. In the context of de novo design, MOSES and GuacaMol offer tasks focused on distribution learning and goal orientation [56]. In property prediction, scientists utilize metrics such as ROC-AUC [57]. There is now a greater focus on robustness to out-of-distribution (OOD) scenarios: models must be able to generalize to unfamiliar scaffolds rather than merely interpolate based on known chemical structures. Recent research, such as "BOOM," has introduced OOD benchmarks specific to chemistry [58,59]. To sum up, deep learning provides effective methods for predicting properties, but achieving success necessitates thorough assessment on appropriate benchmarks and consideration of dataset biases.

Table 2: Common molecular datasets and benchmarks

Dataset/Benchmark	Scope / Size	Task / Focus
ChEMBL (PubChem)	~2-3 million curated compounds	Bioactivity data (IC50, Ki) for many targets
ZINC	~20+ million commercial compounds	Virtual screening library of drug-like molecules
QM9	134,000 small molecules	Quantum properties (HOMO/LUMO energies, dipole, etc)
Tox21	~8,000 (12 assays)	Toxicity classification
ESOL (MoleculeNet)	~1,128	Aqueous solubility (regression)
FreeSolv (MoleculeNet)	~643	Hydration free energy
HIV, BACE, etc.	10,000–50,000	Bioactivity classification (MoleculeNet)
MOSES (Polykovskiy *et al.*,)	1.94 million (subset of ZINC)	*de novo* generation benchmark (validity, novelty)
GuacaMol	1.6 million (ExCAPE-DB)	*de novo* benchmark tasks and metrics
PDBBind	~16,000 protein–ligand complexes	Binding affinity prediction (regression)

3. Inverse QSAR Approaches

Inverse QSAR (or inverse QSPR) involves creating new molecules that are expected to exhibit specific properties. Unlike traditional QSAR, which predicts a property based on a structure, inverse QSAR aims to identify a structure that corresponds to a predetermined property value. Deep generative models enable this process by conditioning on property information. For instance, one could develop a conditional VAE that accepts a target descriptor or label as input and samples latent points that decode into molecules likely to display that property [5]. Remadna *et al.,* created a conditional VAE that uses an attention mechanism (AcVAE) [60]. which translates targeted descriptor vectors into latent space to produce matching SMILES. Likewise, generative adversarial networks and flow models can be directed by additional inputs or goals.

Reinforcement learning effectively addresses inverse QSAR by refining objectives (Fig.4). In this method, a foundational generative model suggests molecules, a property predictor assigns rewards in accordance with target objectives, and the generator is adjusted using policy gradients. This process involves learning within latent space to enhance a property score. The PPO method [40] the previously mentioned example is one illustration. Additionally, another approach is Bayesian optimization [61] in the latent space, train a surrogate model (such as a Gaussian process) on the latent representations of a Variational Autoencoder (VAE), and then employ acquisition functions to suggest potentially valuable latent points, which can be decoded into structures.

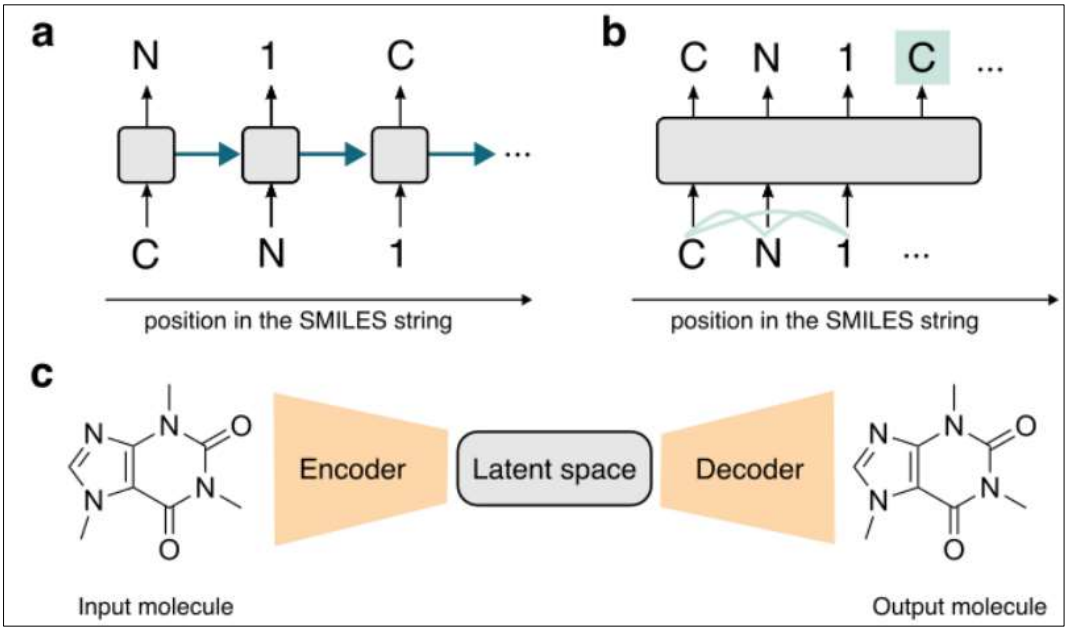

Fig.4: (a) Recurrent Neural Networks, which learn to predict the next token in a SMILES string, using information on all the previous tokens. The network hidden state is updated in a recurrent way, to perform a prediction at any steps while keeping track of the preceding portions of the string. (b) Transformers, which learn all pair relationships between sequence tokens to perform a prediction. (c) Variational Autoencoder, where an encoder is trained to transform an input molecule (e.g., a graph or a string) into a fixed-dimension latent vector, and a decoder is trained to convert such vectors back into molecular representations

Inverse QSAR techniques have achieved success in design that is driven by properties [62]. For example, by representing enzyme binding affinity or solubility as objectives, scientists have produced candidate inhibitors or lead-like compounds. A key issue is validity: the molecule produced needs to be chemically plausible. Modern techniques implement syntactic or chemical constraints (such as valence rules in JT-VAE decoding) to guarantee validity. The primary aim is to create an in silico "designer" that processes high-level targets and generates candidate structures that fulfill those requirements [61-63].

Challenges and Current Limitations

Although there has been remarkable advancement, drug design driven by AI encounters substantial obstacles.

- **Data limitations**: High-quality labeled datasets (such as bioactivities and ADMET results) are frequently limited, inconsistent, or proprietary. A lot of deep learning models require substantial amounts of data. Public databases often contain biases (such as an overabundance of specific chemotypes), which can cause models to overfit or overlook new chemical structures. Data curation and augmentation techniques (like label augmentation and transfer learning) are ongoing areas of research aimed at addressing these challenges. Even when large datasets of unlabeled data are available, the absence of negative examples (inactive molecules) can hinder the reliability of models [64].
- **Interpretability:** Deep networks frequently function as "black boxes," creating challenges for chemists to have confidence in predictions without explanation [66]. Techniques for Explainable AI (XAI) – including attention visualization, feature attribution, and counterfactual reasoning – are starting to clarify how models make decisions. For example, modern frameworks integrate XAI with language models to produce understandable explanations of structure–property relationships [67]. Despite advancements, understanding model predictions is still restricted; chemists typically require explicit explanations for the predicted activity or toxicity of a compound, a need that explainable AI aims to address [68].
- **Scalability:** The molecular landscape is unimaginably vast. Despite rapid machine learning inference, thoroughly exploring this space is

unfeasible. Generative models focus on specific areas and might overlook distant chemotypes [69]. The computational expense can be significant, particularly for 3D generative models or simulations. Additionally, models need to be capable of scaling to accommodate extremely large training datasets when employed. Finding a balance between model complexity, size, and training efficiency remains a persistent issue [70].

- **Synthetic feasibility:** Artificial intelligence models can suggest new molecules that are not feasible to synthesize. Even if a molecule is chemically valid, it might include unstable functional groups or necessitate intricate multi-step synthesis processes [64]. Metrics such as the synthetic accessibility (SA) score and criteria for pan-assay interference (PAINS) are utilized for the generated candidates, though they are not flawless [33]. Sophisticated techniques integrate retrosynthetic planning into the design process, guaranteeing that suggested structures have feasible synthetic pathways. The objective is to create "AI-suggested" compounds that medicinal chemists can realistically synthesize [71].
- **Multi-objective trade-offs:** Pharmaceutical compounds need to meet various criteria at the same time (effectiveness, specificity, ADMET properties, originality). Focusing on improving one characteristic can negatively impact others. Designing with multiple objectives is still a complex task: models have to manage trade-offs and identify Pareto-optimal solutions. One method to tackle this is using reinforcement learning with combined rewards, but consistently achieving a balance among objectives can be quite challenging [72].

These difficulties emphasize that AI tools are not simple solutions. They need to be incorporated carefully, tested through experiments, and regularly improved. Confidence estimates from models and constraints based on domain knowledge are frequently necessary to guarantee reliability.

Future Perspectives

Looking ahead, several trends are poised to shape AI in lead discovery:

- **Self-Supervised and Transfer Learning:** Unlabeled chemical information is plenteous, so self-supervised pretraining will develop. Expansive atomic dialect models (analogs of BERT/GPT) will be pretrained on colossal Grins or chart datasets [73]. Early work (ChemBERTa) [22] illustrates that scaling up pretraining makes strides downstream property expectation. Within the future, we anticipate "foundation models" for chemistry: enormous pretrained systems that chemists fine-tune for particular targets, assignments, or security endpoints.
- **Explainable and Causal AI:** Past black-box expectation, AI models will ought to give noteworthy knowledge. Joining XAI strategies will permit models to highlight key substructures or intuitive driving a forecast. Systems like XpertAI [67] appear that combining ML with expansive dialect models can create natural-language clarifications of SAR discoveries. Eventually, interpretable models may propose mechanism-based alterations to atoms. There's too intrigued in causal models that can foresee the impact of chemical changes, moving toward model-guided theory era [74].
- **Quantum Machine Learning:** Quantum computing and quantum-inspired ML offer a tantalizing wilderness. Later work by Gircha *et al.*, (2023) presented a quantum–classical generative demonstrate for KRAS inhibitors, which utilized quantum Boltzmann machines to show likelihood disseminations [76]. They produced 15 candidate particles and tentatively approved two promising hits, illustrating that quantum-enhanced calculations can outflank simply classical ones in certain plan errands. As quantum equipment develops, half breed quantum-classical ML calculations may speed up assignments like computing atomic Hamiltonians or investigating complex disseminations. Activities just like the Quantum Computing for Sedate Disclosure challenge (2023) reflect developing cooperative energy between quantum computing and AI in medicate plan. Whereas still early, quantum-ML integration may in the long run empower investigating locales of chemical space recalcitrant to classical computing [77].
- **Integrated Biology and Multi-Modal Data:** Future AI models may together reason over different information sorts: quality expression profiles, omics information, quiet records, and atomic structures. For case, a show might connect medicate atoms to cellular phenotypes or illness pathways to superior anticipate adequacy. This multi-modal integration, conceivably intervened by chart representations combining compounds, proteins, and pathways, might make a more all-encompassing sedate disclosure AI. Essentially, coordination basic science (protein 3D structures, docking reenactments) with generative chemistry models will move forward target-specific plan [78].
- **Federated and Collaborative Learning:** Privacy-preserving methods (unified learning) may permit sharing experiences from restrictive pharma information without uncovering the information itself. This might quicken demonstrate change over teach. Stages that standardize chemical and natural datasets will too offer assistance decrease duplication of exertion and make strides reproducibility [79].

In outline, AI in lead disclosure is moving toward bigger, more common models that learn from differing information, give logical comes about, and indeed use rising equipment like quantum processors.

These progresses guarantee to form medicate plan speedier, cheaper, and more inventive, but they will require near collaboration between computational researchers and chemists.

CONCLUSIONS

Profound learning and AI have on a very basic level changed how we approach lead revelation. From upgraded QSAR to generative de novo plan, neural systems presently handle numerous angles of sedate plan. In this survey, we overviewed the key designs (RNNs, GNNs, transformers, VAEs, GANs) and outlined their applications to particle era and property forecast. We talked about how these strategies can be combined in iterative workflows with dynamic learning, whereas moreover sketching out their current confinements (information shortage, interpretability, possibility). Looking forward, progresses in self-supervised learning, reasonable AI, and quantum computing are likely to advance quicken advance. By coordination these innovations shrewdly, the medicate revelation handle may ended up much speedier and more proficient, bringing modern treatments to patients more rapidly.

REFERENCES

1. McNair D. Artificial intelligence and machine learning for lead-to-candidate decision-making and beyond. Annual review of pharmacology and toxicology. 2023 Jan 20;63(1):77-97.
2. Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. *Principles of early drug discovery*. Br. J. Pharmacol. 162, 1239–1249 (2011).
3. Mullard, A. *2019 FDA drug approvals*. Nat. Rev. Drug Discov. 19, 77–81 (2020).
4. Cherkasov, A. *et al., QSAR modeling: where have you been? Where are you going to?* J. Med. Chem. 57, 4977–5010 (2014).
5. Bort W, Mazitov D, Horvath D, Bonachera F, Lin A, Marcou G, Baskin I, Madzhidov T, Varnek A. Inverse qsar: reversing descriptor-driven prediction pipeline using attention-based conditional variational autoencoder. Journal of Chemical Information and Modeling. 2022 Nov 4;62(22):5471-84.
6. Zhavoronkov, A. *et al., Deep learning enables rapid identification of potent DDR1 kinase inhibitors*. Nat. Biotechnol. 37, 1038–1040 (2019).
7. Gómez-Bombarelli, R. *et al., Automatic chemical design using a data-driven continuous representation of molecules*. ACS Cent. Sci. 4, 268–276 (2018).
8. Stokes, J. M. *et al., A Deep Learning Approach to Antibiotic Discovery*. Cell 180, 688–702.e13 (2020).
9. Haghighatlari M, Li J, Heidar-Zadeh F, Liu Y, Guan X, Head-Gordon T. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. Chem. 2020 Jul 9;6(7):1527-42.
10. Nivas Marimuthu A, McGuire BA. A Machine Learning Pipeline for Molecular Property Prediction using ChemXploreML. arXiv e-prints. 2025 May:arXiv-2505.
11. Meyers J, Fabian B, Brown N. De novo molecular design and generative models. Drug discovery today. 2021 Nov 1;26(11):2707-15.
12. Brown N, Fiscato M, Segler MH, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. Journal of chemical information and modeling. 2019 Mar 19;59(3):1096-108.
13. Nguyen, Hoang (2024). De novo Molecular Design using Deep Learning. Open Access Te Herenga Waka-Victoria University of Wellington. Thesis. https://doi.org/10.26686/wgtn.25044362
14. Keith JA, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Muller KR, Tkatchenko A. Combining machine learning and computational chemistry for predictive insights into chemical systems. Chemical reviews. 2021 Jul 7;121(16):9816-72.
15. Husnain A, Rasool S, Saeed A, Hussain HK. Revolutionizing pharmaceutical research: Harnessing machine learning for a paradigm shift in drug discovery. International Journal of Multidisciplinary Sciences and Arts. 2023 Sep 27;2(4):149-57.
16. Jiang J, Zhang C, Ke L, Hayes N, Zhu Y, Qiu H, Zhang B, Zhou T, Wei GW. A review of machine learning methods for imbalanced data challenges in chemistry. Chemical Science. 2025.
17. Valizadeh A, Amirhosseini MH, Ghorbani Y. Predictive precision in battery recycling: unveiling lithium battery recycling potential through machine learning. Computers & Chemical Engineering. 2024 Apr 1;183:108623.
18. Orosz Á, Héberger K, Rácz A. Comparison of descriptor-and fingerprint sets in machine learning models for ADME-Tox targets. Frontiers in Chemistry. 2022 Jun 8;10:852893.
19. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, Metni H, van Hoesel C, Schopmans H, Sommer T, Friederich P. Graph neural networks for materials science and chemistry. Communications Materials. 2022 Nov 26;3(1):93.
20. Atz K, Cotos L, Isert C, Håkansson M, Focht D, Hilleke M, Nippa DF, Iff M, Ledergerber J, Schiebroek CC, Romeo V. Prospective de novo drug design with deep interactome learning. Nature Communications. 2024 Apr 22;15(1):3408.
21. Sadaghiyanfam S, Kamberaj H, Isler Y. Leveraging ChemBERTa and machine learning for accurate toxicity prediction of ionic liquids. Journal of the Taiwan Institute of Chemical Engineers. 2025 Jun 1;171:106030.
22. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885. 2020 Oct 19.

23. Zhang R, Wang X, Liu K, Zhou Y, Wang P. M2Mol: Multi-view Multi-granularity Molecular Representation Learning for Property Prediction. InInternational Conference on Database Systems for Advanced Applications 2024 Jul 2 (pp. 264-274). Singapore: Springer Nature Singapore.
24. Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. InInternational conference on machine learning 2018 Jul 3 (pp. 2323-2332). PMLR.
25. Kondratyev V, Dryzhakov M, Gimadiev T, Slutskiy D. Generative model based on junction tree variational autoencoder for HOMO value prediction and molecular optimization. Journal of Cheminformatics. 2023 Feb 2;15(1):11.
26. Alqahtani H, Kavakli-Thorne M, Kumar G. Applications of generative adversarial networks (gans): An updated review. Archives of Computational Methods in Engineering. 2021 Mar;28:525-52.
27. Saxena D, Cao J. Generative adversarial networks (GANs) challenges, solutions, and future directions. ACM Computing Surveys (CSUR). 2021 May 8;54(3):1-42.
28. Akkem Y, Biswas SK, Varanasi A. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Engineering Applications of Artificial Intelligence. 2024 May 1;131:107881.
29. Mehmood R, Bashir R, Giri KJ. Deep generative models: a review. Indian Journal of Science and Technology. 2023 Feb 17;16(7):460-7.
30. Zeng X, Wang F, Luo Y, Kang SG, Tang J, Lightstone FC, Fang EF, Cornell W, Nussinov R, Cheng F. Deep generative molecular design reshapes drug discovery. Cell Reports Medicine. 2022 Dec 20;3(12).
31. Ross J, Belgodere B, Hoffman SC, Chenthamarakshan V, Navratil J, Mroueh Y, Das P. Gp-molformer: A foundation model for molecular generation. arXiv preprint arXiv:2405.04912. 2024 Apr 4.
32. Thomas M, O'Boyle NM, Bender A, De Graaf C. MolScore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. Journal of Cheminformatics. 2024 May 30;16(1):64.
33. Kuznetsov M, Polykovskiy D. MolGrow: A graph normalizing flow for hierarchical molecular generation. InProceedings of the AAAI Conference on Artificial Intelligence 2021 May 18 (Vol. 35, No. 9, pp. 8226-8234).
34. Anishchenko I, Kipnis Y, Kalvet I, Zhou G, Krishna R, Pellock SJ, Lauko A, Lee GR, An L, Dauparas J, DiMaio F. Modeling protein-small molecule conformational ensembles with ChemNet. bioRxiv. 2024 Sep 25.
35. Dupont E, Teh YW, Doucet A. Generative models as distributions of functions. arXiv preprint arXiv:2102.04776. 2021 Feb 9.
36. Luo Y, Yan K, Ji S. Graphdf: A discrete flow model for molecular graph generation. In International conference on machine learning 2021 Jul 1 (pp. 7192-7203). PMLR.
37. Li Y, Pei J, Lai L. Structure-based de novo drug design using 3D deep generative models. Chemical science. 2021;12(41):13664-75.
38. Wang S, Lin T, Peng T, Xing E, Chen S, Kara LB, Cheng X. TopMT-GAN: a 3D topology-driven generative model for efficient and diverse structure-based ligand design. Chemical Science. 2025;16(6):2796-809.
39. Lu Z, Wei J, Wang Z. Active steering controller for driven independently rotating wheelset vehicles based on deep reinforcement learning. Processes. 2023 Sep 6;11(9):2677.
40. Haddad R, Litsa EE, Liu Z, Yu X, Burkhardt D, Bhisetti G. Targeted molecular generation with latent reinforcement learning. Scientific Reports. 2025 Apr 30;15(1):15202.
41. Li C, Yamanishi Y. SpotGAN: a reverse-transformer GAN generates scaffold-constrained molecules with property optimization. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2023 Sep 17 (pp. 323-338). Cham: Springer Nature Switzerland.
42. Ivanenkov Y, Zagribelnyy B, Malyshev A, Evteev S, Terentiev V, Kamya P, Bezrukov D, Aliper A, Ren F, Zhavoronkov A. The hitchhiker's guide to deep learning driven generative chemistry. ACS Medicinal Chemistry Letters. 2023 Jun 30;14(7):901-15.
43. Hu Q, Xiong Y, Zhu GH, Zhang YN, Zhang YW, Huang P, Ge GB. The SARS-CoV-2 main protease (Mpro): structure, function, and emerging therapies for COVID-19. MedComm. 2022 Sep;3(3):e151.
44. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, Engkvist O. Reinvent 4: Modern AI–driven generative molecule design. Journal of Cheminformatics. 2024 Feb 21;16(1):20.
45. Alghamdi A, Abouzied AS, Alamri A, Anwar S, Ansari M, Khadra I, Zaki YH, Gomha SM. Synthesis, molecular docking, and dynamic simulation targeting main protease (Mpro) of new, thiazole clubbed pyridine scaffolds as potential COVID-19 inhibitors. Current Issues in Molecular Biology. 2023 Feb 7;45(2):1422-42.
46. Truong GB, Pham TA, To VT, Le HS, Van Nguyen PC, Trinh TC, Phan TL, Truong TN. Discovery of Vascular Endothelial Growth Factor Receptor 2 Inhibitors Employing Junction Tree Variational Autoencoder with Bayesian Optimization and Gradient Ascent. ACS omega. 2024 Nov 12;9(47):47180-93.
47. Fan J, Hong SK, Lee Y. Validity Improvement in MolGAN-Based Molecular Generation. IEEE Access. 2023 Jun 2;11:58359-66.

48. Wu Y, Ni X, Wang Z, Feng W. Enhancing drug property prediction with dual-channel transfer learning based on molecular fragment. BMC bioinformatics. 2023 Jul 21;24(1):293.
49. Lu S, Gao Z, He D, Zhang L, Ke G. Highly accurate quantum chemical property prediction with uni-mol+. arXiv preprint arXiv:2303.16982. 2023 Mar 16.
50. Liu R, Krishnan A. Open biomedical network benchmark: a Python toolkit for benchmarking datasets with biomedical networks. InMachine Learning in Computational Biology 2024 Mar 15 (pp. 23-59). PMLR.
51. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic acids research. 2024 Jan 5;52(D1):D1180-92.
52. Kim S. Exploring chemical information in PubChem. Current protocols. 2021 Aug;1(8):e217.
53. Karwounopoulos J, Kaupang Å, Wieder M, Boresch S. Calculations of absolute solvation free energies with Transformato─ Application to the FreeSolv database using the CGenFF force field. Journal of Chemical Theory and Computation. 2023 Aug 24;19(17):5988-98.
54. Low K, Coote ML, Izgorodina EI. Explainable solvation free energy prediction combining graph neural networks with chemical intuition. Journal of Chemical Information and Modeling. 2022 Nov 1;62(22):5457-70.
55. Williams NJ, Kabalan L, Stojanovic L, Zólyomi V, Pyzer-Knapp EO. Hessian QM9: A quantum chemistry database of molecular Hessians in implicit solvents. Scientific data. 2025 Jan 3;12(1):9.
56. Flores-Hernandez H, Martinez-Ledesma E. A systematic review of deep learning chemical language models in recent era. Journal of Cheminformatics. 2024 Nov 18;16(1):129.
57. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023 Feb 17;16(1):4.
58. Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T. Wilds: A benchmark of in-the-wild distribution shifts. InInternational conference on machine learning 2021 Jul 1 (pp. 5637-5664). PMLR.
59. Antoniuk ER, Zaman S, Ben-Nun T, Li P, Diffenderfer J, Demirci B, Smolenski O, Hsu T, Hiszpanski AM, Chiu K, Kailkhura B. BOOM: Benchmarking Out-Of-distribution Molecular Property Predictions of Machine Learning Models. arXiv preprint arXiv:2505.01912. 2025 May 3.
60. Remadna I, Terrissa LS, Al Masry Z, Zerhouni N. RUL prediction using a fusion of attention-based convolutional variational autoencoder and ensemble learning classifier. IEEE Transactions on Reliability. 2022 Jul 27;72(1):106-24.
61. Guo J, Ranković B, Schwaller P. Bayesian optimization for chemical reactions. Chimia. 2023 Feb 22;77(1/2):31-8.
62. Born J, Manica M. Trends in deep learning for property-driven drug design. Current medicinal chemistry. 2021 Nov 1;28(38):7862-86.
63. Born J, Manica M. Trends in deep learning for property-driven drug design. Current medicinal chemistry. 2021 Nov 1;28(38):7862-86.
64. Tiwari PC, Pal R, Chaudhary MJ, Nath R. Artificial intelligence revolutionizing drug development: Exploring opportunities and challenges. Drug Development Research. 2023 Dec;84(8):1652-63.
65. Rahate KP, Mondal R. Applications of AI in drug discovery: Its challenges, opportunities, and strategies. Approaches to Human-Centered AI in Healthcare. 2024:86-120.
66. Benítez JM, Castro JL, Requena I. Are artificial neural networks black boxes?. IEEE Transactions on neural networks. 1997 Sep;8(5):1156-64.
67. Wellawatte GP, Schwaller P. Human interpretable structure-property relationships in chemistry using explainable machine learning and large language models. Communications Chemistry. 2025 Jan 14;8(1):11.
68. Rahate KP, Mondal R. Applications of AI in drug discovery: Its challenges, opportunities, and strategies. Approaches to Human-Centered AI in Healthcare. 2024:86-120.
69. Bettanti A, Beccari AR, Biccarino M. Exploring the future of biopharmaceutical drug discovery: can advanced AI platforms overcome current challenges? Discover Artificial Intelligence. 2024 Dec;4(1):1-6.
70. Xu J, Stevenson J. Drug-like index: a new approach to measure drug-like compounds and their diversity. Journal of Chemical Information and Computer Sciences. 2000 Sep 25;40(5):1177-87.
71. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert opinion on drug discovery. 2021 Sep 2;16(9):949-59.
72. Lambrinidis G, Tsantili-Kakoulidou A. Multi-objective optimization methods in novel drug design. Expert opinion on drug discovery. 2021 Jun 3;16(6):647-58.
73. Liu J, Yang M, Yu Y, Xu H, Li K, Zhou X. Large language models in bioinformatics: applications and perspectives. arXiv preprint arXiv:2401.04155. 2024 Jan 8.
74. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh CY, Hou T. From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. Journal of Chemical

Information and Modeling. 2023 Dec 11;63(24):7617-27.
75. Gircha AI, Boev AS, Avchaciov K, Fedichev PO, Fedorov AK. Hybrid quantum-classical machine learning for generative chemistry and drug design. Scientific Reports. 2023 May 22;13(1):8250.
76. Kwon T, Kim H. Quantum biological convergence: quantum computing accelerates KRAS inhibitor design. Signal Transduction and Targeted Therapy. 2025 May 14;10(1):1-2.
77. Hassan SS, Khan LU, Park YM, Guizani M, Han Z, Ratnarajah T, Hong CS. Quantum Machine Learning for 6G Space-Air-Ground Integrated Networks: A Comprehensive Tutorial and Survey. Authorea Preprints. 2025 May 15.
78. Ali H. Artificial intelligence in multi-omics data integration: Advancing precision medicine, biomarker discovery and genomic-driven disease interventions. Int J Sci Res Arch. 2023;8(1):1012-30.
79. Huang D, Ye X, Sakurai T. Multi-party collaborative drug discovery via federated learning. Computers in Biology and Medicine. 2024 Mar 1;171:108181.

Cite This Article: Arnav Kumar & Subhas S Karki (2025). AI & Machine Learning in Lead Discovery: Deep-Learning Architectures for *de novo* Design, Property Prediction and Inverse QSAR. *EAS J Humanit Cult Stud, 7*(3), 88-100.