Flores-Hernandez and Martinez-Ledesma
Journal of Cheminformatics (2024) 16:129
https://doi.org/10.1186/s13321-024-00916-y

Journal of Cheminformatics

REVIEW

Open Access

A systematic review of deep learning chemical language models in recent era

Hector Flores-Hernandez[1] and Emmanuel Martinez-Ledesma[2,3*]

Abstract

Discovering new chemical compounds with specific properties can provide advantages for fields that rely on materials for their development, although this task comes at a high cost in terms of complexity and resources. Since the beginning of the data age, deep learning techniques have revolutionized the process of designing molecules by analyzing and learning from representations of molecular data, greatly reducing the resources and time involved. Various deep learning approaches have been developed to date, using a variety of architectures and strategies, in order to explore the extensive and discontinuous chemical space, providing benefits for generating compounds with specific properties. In this study, we present a systematic review that offers a statistical description and comparison of the strategies utilized to generate molecules through deep learning techniques, utilizing the metrics proposed in Molecular Sets (MOSES) or Guacamol. The study included 48 articles retrieved from a query-based search of Scopus and Web of Science and 25 articles retrieved from citation search, yielding a total of 72 retrieved articles, of which 62 correspond to chemical language models approaches to molecule generation and other 10 retrieved articles correspond to molecular graph representations. Transformers, recurrent neural networks (RNNs), generative adversarial networks (GANs), Structured Space State Sequence (S4) models, and variational autoencoders (VAEs) are considered the main deep learning architectures used for molecule generation in the set of retrieved articles. In addition, transfer learning, reinforcement learning, and conditional learning are the most employed techniques for biased model generation and exploration of specific chemical space regions. Finally, this analysis focuses on the central themes of molecular representation, databases, training dataset size, validity-novelty trade-off, and performance of unbiased and biased chemical language models. These themes were selected to conduct a statistical analysis utilizing graphical representation and statistical tests. The resulting analysis reveals the main challenges, advantages, and opportunities in the field of chemical language models over the past four years.

Keywords Chemical language models (CLMs), Recurrent neural networks (RNNs), Transformers, Variational autoencoders (VAEs), Generative adversarial networks (GANs), Transfer learning, Reinforcement learning and conditional learning

*Correspondence:
Emmanuel Martinez-Ledesma
juanemmanuel@tec.mx
[1] Tecnológico de Monterrey, School of Engineering and Sciences, Monterrey 64710, Nuevo León, México
[2] Tecnológico de Monterrey, School of Medicine and Health Sciences, Monterrey 64710, Nuevo León, México
[3] Institute for Obesity Research, Tecnológico de Monterrey, Monterrey 64710, Nuevo León, México

Introduction

Chemical space and de novo molecule design

Molecule design aims to discover chemical entities that are distributed within a vast, intricate, and discontinuous space known as chemical space, which encompasses all possible atomic configurations that can produce molecules [1]. The search for molecules with specific properties in chemical space is a time-consuming

and expensive task due to the irregular distribution of molecules, where even slight changes in a molecule can result in significant changes in physicochemical properties [2]. For years, researchers considered molecule design as a process based on trial and error to explore various arrangements of functional groups or atoms, yielding molecules with diverse structures that can map regions of chemical space. This approach relied heavily on human knowledge and was limited to the human ability to identify complex chemical patterns from structures [3]. However, since the introduction of computational methods in the field of chemical sciences and supported by the exponential growth of available molecular data (e.g. chemical structure, physicochemical properties, bioactivity, toxicity, and others), from the last decade, molecular design experienced one of the most important advances in its history, driven mainly by methods capable of learning from data to generate novel chemical entities with specific properties, such as deep learning [4].

Deep learning is a subset of machine learning that performs predictive or generative tasks, specifically involving learning representation methods that enable computers to understand how to represent data from its raw form by performing multiple nonlinear matrix multiplications to learn multiple levels of abstraction [5]. As such, deep learning can learn chemistry rules for prediction or generative tasks by using different molecular representations that can be understood in their natural form by humans such as Simplified Molecular Input Line Entry System (SMILES), IUPAC nomenclature or molecular graphs, molecular fingerprints, among others [6–9]. The implementation of learning-based method has proven to be successful in multiple fields of chemistry, such as Quantitative Structure – Activity Relationships (QSAR), Quantitative Structure–Property Relationships (QSPR), and molecular generative models.[10, 11].

Furthermore, deep generative models are gaining rapid attention in the design of molecules. They possess the ability to learn implicit chemical knowledge from data by identifying structural patterns such as valency rules, reactive groups, molecular conformations, hydrogen bond donors and acceptors, among others to produce molecules with specific properties. Unlike hand-encoded rules-based or enumeration methods, which require human intervention to define chemical rules based on human knowledge to generate molecules, these models are independent and less prone to generating molecules that are unavailable for chemical synthesis due to unstable groups [12, 13].

Several deep learning architectures like recurrent neural networks (RNNs), transformers, variational autoencoders (VAEs), and generative adversarial networks (GANs) offer an efficient way to investigate chemical space using statistical techniques. These models can generate targeted molecules and investigate regions of the chemical domain through biased learning methods. Such methods can manipulate the molecular generator to yield molecules that meet specific conditions, showing analogous structures and therefore chemical properties [14–17]. On the other hand, such models could map large areas of chemical space by solely acquiring chemical regulations to reconstruct molecular structures from encoded molecular spaces [18, 19]. Deep generative models have become a useful tool for designing molecules due to their cost-effectiveness and time efficiency [20–22]. Due to the significance of deep learning models in contemporary molecule generation, it is imperative to use metrics for evaluating statistical methods that allow chemists and data scientists to compare the efficiency of different molecular generators. To address this requirement, benchmarking platforms have been introduced to quantify the quality and diversity of the distribution of generated molecules. Molecular Sets (MOSES) and GuacaMol are widely accepted benchmarks for measuring the quality, diversity, and fidelity of outputs generated by deep generative models, as well as their ability to explore chemical space [23, 24].

Since the start of the deep learning era in molecular design, other works have summarized the architectures mentioned above in terms of their theoretical background and applications for drug development or statistical approaches to explore the chemical space [12, 25–28]. Only a small portion of these reviews has systematically evaluated the implementation of deep learning architectures for molecular generation tasks. The primary challenges faced by deep learning architectures in molecule generation, as well as the most used deep learning architectures for de novo molecule generation and molecular representations, were examined through systematic evaluations of generative models [29]. In addition, Koutroumpa et al. conducted a systematic analysis of deep generative models to relate the validation of target molecules produced by these models in biological models. Deep generative models demonstrated their relevance in drug design and their capability for generating bioactive compounds, as evidenced in both in vitro and in vivo models [30]. Although these analyses are beneficial, it is essential to conduct a statistical evaluation of deep learning models using established standards like MOSES benchmark. This benchmarking platform can accurately reflect the robustness of models to generate novel, valid, and unique chemical entities.

The present study aims to comprehensively investigate the quality of deep learning architectures over the last

three years to generate chemical entities and to evaluate what are the most important features that affect the quality metrics of deep generative models through a systematic review. This work focuses on answering the following research questions:

1. How does the configuration of deep learning architectures and training size affect the quality metrics of the generated molecules?
2. Which deep learning architectures can achieve higher quality metrics of generated molecules?
3. Which hyperparameters have a higher impact on each deep learning architecture for better molecule generation performance?
4. What are the most common molecular representations for deep learning models?
5. What type of biased deep generative methods are most effective in generating novel molecules that show activity for a given target?

Therefore, a comprehensive systematic review is presented herein, encompassing a set of articles that assessed MOSES or Guacamol metrics from 2020- June 2024. This work presents a theoretical background on deep generative models and metrics found in the set of retrieves articles, followed by a detailed explanation of methodology using Prisma for article selection is provided, and a discussion section is presented highlighting the most significant finding from the systematic review. Finally, a statistical analysis is performed to analyze the robustness of deep learning models on molecular generation tasks.

Theoretical background

Since we aim to review, measure, and analyze deep learning techniques used in chemical language models, first we present some theoretical background of the concepts used in this area.

Molecule representations

Learning is a data-driven process. From the beginning, when computer science was introduced to chemistry, chemists have attempted to represent chemical entities through different methods that enable computer algorithms to acquire knowledge on how to build molecules. The implementation of deep generative models in chemistry requires the use of an appropriate and precise molecular representation that provides enough information for computers to process and learn through matrix operations. Molecular representations must meet specific requirements to ensure an accurate representation of a real molecule. These requirements include permutation invariance to ensure no alteration by changes in the specified order of atoms, translational invariance to prevent changes from translations in space, and rotational invariance to avoid changes from rotation operations [25] . On this basis, various representations have been developed through the years for deep learning applications depending on the challenges that deep generative models face and the requirements of the generation tasks. These molecular representations are illustrated in Fig. 1A.

Molecular graphs

Introduced more than 30 years ago, molecular graphs are one of the most revolutionary concepts for representing the chemical identities of covalently bonded molecules [32]. This approach involves mapping the atoms and bonds of a molecule onto a set of nodes (V) and edges (E) in a square matrix (G = (V, E)), where the matrix size reflects the total number of non-hydrogen atoms, usually called adjacency matrix. The adjacency matrix first enumerates each atom in the structure and then presents information about the types of atoms in its main diagonal, identified by their atomic numbers. Connectivity is determined by assigning a value of 0 to non-adjacent atoms in the structure or 1 to adjacent pairs of atoms, even this one for adjacency can be replaced by another number (between 1 and 4) to indicate the type of bond (single, double, triple, or aromatic), increasing the amount of information encoded in the matrix [33, 34]. Although molecular graphs are not memory efficient due to the large amount of information required to represent a single molecule, one-dimensional representations have been developed to overcome this limitation. These representations are based on strings, which require much less information and provide easily interpretable, human-friendly representations [35]. Additionally, one-dimensional representations have driven the rise of chemical language models in the recent era of deep learning in cheminformatics, changing the perspective of molecule generation.

Molecular representation for chemical language models

Nowadays, language models play a pivotal role in designing molecules through deep learning models [36]. Chemical language models, which use one-dimensional molecule representations as inputs, can generate molecules and learn the syntax, coherence, and grammar rules necessary to build them through training. These models train architectures using one-dimensional string representations of molecules [37, 38]. This section offers a concise overview of the most frequently used molecular representations for chemical language models.

Simplified molecular input line entry specification (SMILES)

The SMILES notation is a prevalent method employed for molecular representation in the field of deep learning.

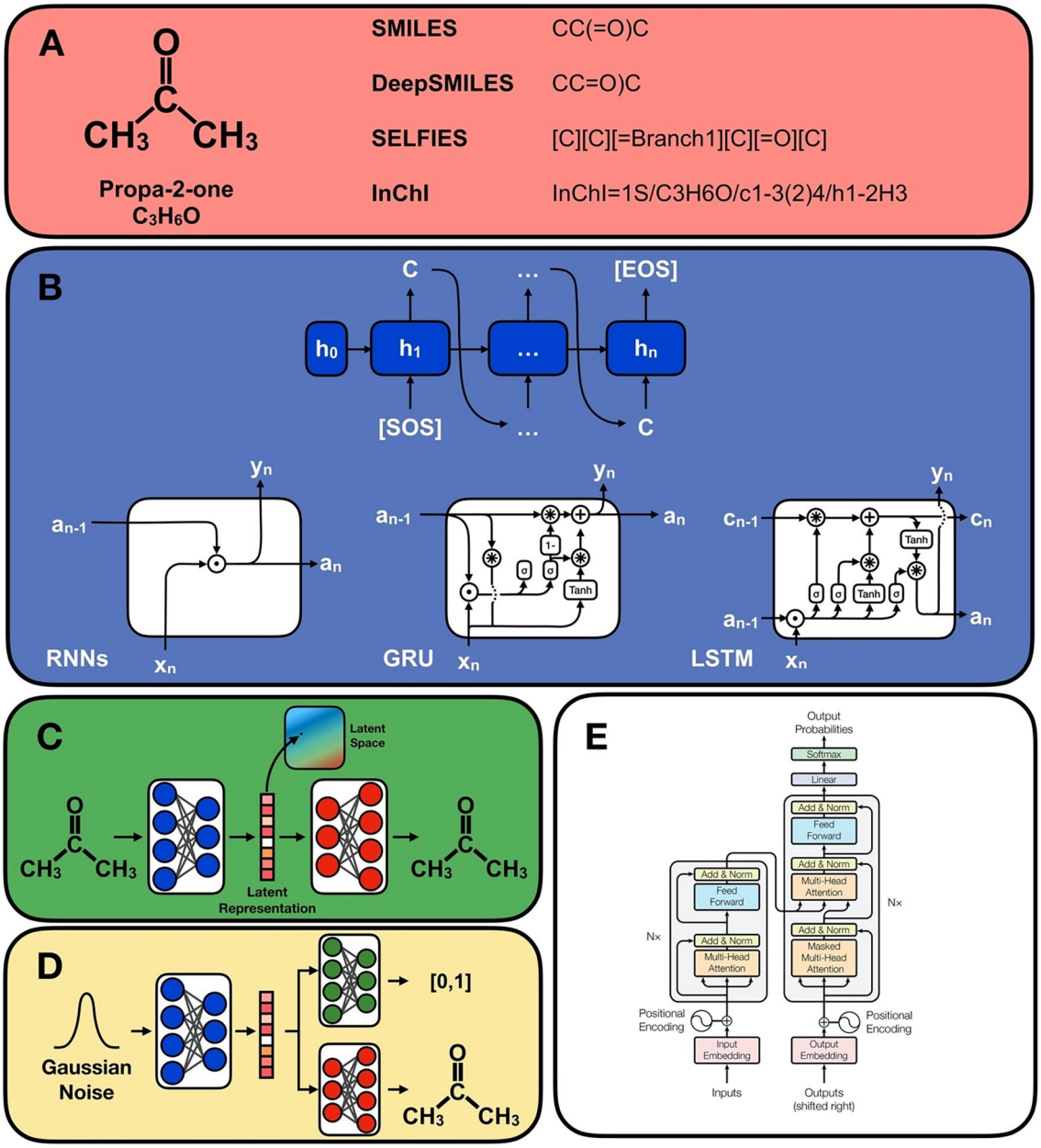

Fig. 1 Illustration of molecular representations and chemical language models. **A** displays various molecular representations of propa-2-one (acetone). **B** showcases RNNs as chemical language models and their autoregressive approach for generating chemical entities where [SOS] and [EOS] stands for start of sentence and end of sentence tokens, respectively. RNNs cells are also shown where the symbols •, *, and + denote dot-product, elementwise matrix multiplication, and addition, respectively. Each arrow corresponds to matrix multiplication utilizing a learnable matrix. Finally, in the context of RNNs, 'x', 'a', and 'c' correspond to the input, information matrix, and memory term, respectively. **C** illustrates the schematic representation of VAEs, while **D** presents the schematic representation of GANs. **E** displays the transformer model proposed by Vaswani and colleagues.31

This approach has been in practice for over three decades and it is comprised of character strings in ASCII (American Standard Code for Information Interchange) format. SMILES enables molecules to be represented by a series of tokens based on their chemical structure, where each element in the periodic table is assigned to a corresponding token using its atomic symbol. In the absence of a specified bond type, the bond type can be inferred as a single bond type. Conversely, when tokens appear in lowercase form, the bond type is identified as

aromatic. Otherwise, the bond type is explicitly indicated using non-alphanumeric tokens. Additionally, SMILES employs a special token (brackets) to specify branches or cycles within chemical structures. These notations and rules for molecular representation can be considered as an analogous form of language (chemical language) where words are present in the form of chemical tokens and sentences are molecules. However, similar to spoken languages, SMILES carries the potential for syntactic and grammar errors when dealing with branches and cycles, presenting a challenge for deep learning architectures aiming to accurately reconstruct syntactically valid molecules from latent space [39]. In addition, SMILES representation is a non-unique molecular representation, but this can be transformed into a unique molecular representation by implementing SMILES canonicalization algorithms, yet multiple SMILES can exist for a single molecule [40].

IUPAC international chemical identifier (InChI)

Introduced by IUPAC in 2013 as open-source software for molecule encoding, InChI is a string-based molecular representation that employs six layers and multiple sublayers to convey information about molecules. Each layer contains specific details regarding the chemical formula, atom connectivity, atomic charges, and stereochemistry, among others. It also provides information on the reactivity of atoms that may undergo chemical equilibrium, leading to the formation of constitutional isomers and resonance structures. The incorporation of multiple layers of information and its ability to provide information about structural and stereo isomers makes InChI the first canonical representation of molecules. However, unlike SMILES, due to the complexity level of InChI syntaxis and grammar, this representation method is not a user-friendly method and is also prone to valency and branching issues [41].

DeepSMILES

DeepSMILES is a SMILES-like molecule representation that encodes molecules using a syntax that avoids grammatical errors during molecule generation by utilizing one symbol for specifying branches and cycle closures. To indicate the length of a branch, *n* number of closing parentheses are used, where '*n*' represents the length of the branch. Similarly, cycles are represented by a number indicating the length of the cycle. This simplification of notation regarding SMILES allows DeepSMILES to solve grammar mistakes that arise when deep learning architectures learn molecular representations. However, this notation still leaves room for implementation that addresses valency constraints and the development of unique molecule representations [42].

Self-referencing embedded strings (SELFIES)

In 2020 Krenn M, et al., introduced SELFIES which represent molecules using string-based methods [43]. They use derivation rules to produce valid molecules by avoiding the use of brackets for branches and cycles and employing special symbols to indicate the start of the cycle or branch, ensuring the production of only valid molecules. Instead of utilizing an end marker, the length of the branch or ring is defined by the subsequent token in the string. This method additionally addresses valency constraints that do not yield valency penalties in molecules. Nonetheless, like SMILES notations, a single molecule corresponds to multiple SELFIES representations [43, 44].

Figure 1a illustrates an example of a molecular representation, highlighting the primary distinctions between one-dimensional molecular representation and the SMILES notation. In the latter, explicit atom and bond characters are employed, following the aforementioned rules. Similarly, DeepSMILES CC=OC streamlines the representation of branching and ring closure, while maintaining a similar visual representation for linear molecules such as prop-2-one. In contrast, SELFIES employs brackets to delineate each token in a sequence and utilizes an explicit token to describe the branching in the sequence.

Chemical language models (CLMs) as deep generative models

The quality of learning data is contingent upon both data representation and the manner in which information is processed between layers in deep learning models until a continuous vectorial representation is achieved that can reliably represent a molecule. CLMs customize natural language processing (NLP) algorithms to learn chemical grammar and syntax from one-dimensional molecular data [45–47]. Several successful deep learning architectures have been introduced into molecular generative models to process sequential text data. Thus far, these models have demonstrated notable progress in text generation tasks.

Recurrent neural networks (RNNs)

Introduced more than 40 years ago by Hopfield, RNNs are neural networks capable of processing information in a sequential form. RNNs are used to generate and manipulate sequentially structured data, such as one-dimensional molecular representations [48]. RNNs receive a sequence input and use a set of hidden layers connected between them in a recurrent manner to transform the discrete token representation, generated from one-dimensional data, into a continuous representation. The continuous representation is then fed

into a feedforward network to predict the adjacent token in the sequence. Information from the hidden layers is subsequently distributed to adjacent RNNs units, enhancing predictions with shared context. Assuming a benefit in predicting tokens via context-sharing among recurrent units, RNNs face a challenge wherein gradients may vanish or explode during the backpropagation process if the sequences become lengthy enough. This results in a nearly impossible task for RNNs to learn the long-term dependencies within the sequence. To address this limitation of RNNs, gated RNNs-units such as Long-Short Term Memory (LSTM) units and Gated Recurrent Units (GRU) are designed to implement short-term memory by adding trainable parameters that control the flow information of sequence dependencies, avoiding gradient vanishing or explosion at the cost of increasing the trainable parameters in the model (Fig. 1B) [49, 50].

Transformers

Since the successful performance of dynamic models for processing sequential data as RNNs, the search for novel architectures capable of capturing context from sequences began until meeting Transformers in 2017 by Vaswani and coworkers [31]. Transformers rank among the top deep learning architectures, surpassing RNNs in their ability to learn one-dimensional molecule representations, such as SMILES or SELFIES. This success results from their capacity to capture the relationship between tokens in sequences, independent of sequence length, which is attributed to the incorporation of attention mechanisms [51]. This architecture consists of an encoder-decoder model, where the encoder learns how to map molecules (from string-based methods) into a continuous representation, and the decoder learns how to reconstruct these models from continuous representation into a string-based representation. Positional embeddings and attention mechanisms are crucial for acquiring chemical language cognition abilities. By establishing connections between token positions and computing the attention coefficient using scaled dot product and softmax operations, these mechanisms facilitate language acquisition. To accomplish this, tokens are first embedded using word and positional embeddings methods which converts discrete token representation into a continuous representation. Next, attention mechanism is applied several times in parallel through a series of trainable matrix that allows to linear transform the vectorial representation of chemical sequences. Finally, a feedforward layer is used to generate a fixed-length representation of molecules to allow matrix operations in further layers (decoder model or softmax layer for chemical token prediction).

Variational autoencoders (VAEs)

Sequential models have surpassed generative models in the realm of chemical language applications. Nonetheless, there have been successful introductions of alternative methods for generating molecules. These methods are founded upon the compression of discrete data, into a continuous value vector, which is later reconstructed into discrete data [52]. First introduced in 2018 by Gómez-Bombarelli for molecular generative models, VAEs have proven to be a powerful tool for generating novel molecules. In principle, VAEs are generative models designed to model an unknown data distribution using a finite sample from the distribution. This model comprises an encoder and decoder components. The encoder maps the discrete representation of molecules into a continuous latent space using a low-dimensional vector [53]. This latent vector can be utilized for further classification or regression tasks to organize the space based on specific properties [53, 54]. On the other hand, the encoder is utilized to reconstruct molecules from latent space into its discrete representation. This process guarantees the capacity to learn how to generate molecules while complying with syntax and grammar rules. The VAE encoder and decoder may be different types of deep learning neural networks, including transformers, RNNs, and multilayer perceptron, among others.

Generative adversarial networks (GANs)

Since the successful application of autoencoders in generative chemistry and supported by their success in other fields such as image and audio generation, GANs have been introduced to chemistry. These models can increase the diversity of molecules generated while maintaining the probability of the data distribution [55]. GANs comprise two components: a generator and a discriminator. The generator component of GANs may employ various deep learning architectures like RNNs, transformers, or VAEs, among others. Once trained to build molecules, this component generates random molecules by inputting random noise into the model. In contrast, generating molecules with specific structural patterns depends on a discriminator component, a neural network that identifies if the created data represents an actual molecule or not. This element is essential for retraining the complete model until it can no longer differentiate between genuine and artificial data and updating the generator's parameters [56, 57].

Biased generative models

Until this point, a brief theoretical framework of the deep generative architectures involved in this work has been illustrated. Even though these models can generate large virtual libraries of molecules that are grammatically and

syntactically correct, producing molecules that are accessible for chemical synthesis and possess specific chemical properties or bioactivities remains a significant challenge for deep learning, particularly when working with biological targets or rare molecules that lack sufficient data for model training. To address these challenges, biased strategies have been incorporated into deep learning models. This allows for the exploration of chemical space in specific directions and the generation of molecules that are synthesis feasible or possess desired properties.

Reinforcement learning (RL)

Reinforcement Learning refers to a collection of techniques applied to solve decision problems in artificial intelligence (AI) models, such as deep learning. The RL methodology includes evaluating possible actions and their respective outcomes and subsequently devising a treatment plan that strives to achieve the optimal outcome [58]. RL techniques are implemented in generative deep learning models and can predict whether a generated molecule meets specific conditions. If a generated molecule is desired, the model is rewarded and updates to the model parameters allow for specific direction exploration of the chemical space, resulting in the generation of molecules with specific properties. Several reward functions have been developed to obtain molecules that are synthesizable, accessible, non-toxic, bioactive, permeable to biological membranes, and possess specific physicochemical properties [17, 58–62]. In addition, multi-objective optimization properties of molecules can be performed [14, 63, 64].

Transfer learning (TL)

On the other hand, the concept of Transfer Learning involves transferring knowledge learned by a model from a particular task to another model that can utilize this information to improve its performance in a comparable task. This approach can be utilized with molecular generative models for transferring expertise on synthesizing a particular group of molecules, thereby facilitating the production of novel molecules that exhibit desired properties [65]. TL enables generative models to apply their acquired knowledge in producing molecules with particular bioactivities or affinities to biological targets and transfer this knowledge to a model capable of generating grammatically and syntactically correct molecules [66]. This fine-tuning technique can overcome limitations in specialized datasets for biological targets or molecules with unique properties that lack the necessary information to enable models to learn syntax and grammar rules of chemical language [15, 67–70].

Conditional learning (CL)

Similarly, conditional learning (CL) can be utilized to direct deep learning architectures in the synthesis of molecules with specified, desired characteristics, obviating the necessity for incorporating a reinforcement learning agent to assess the outputs of the model or depending on fine-tuning data for retraining [71]. In contrast, CL enables models to integrate domain-specific data about chemical structures, such as chemical properties, biological activities, or functional groups, directly into the training process. This is typically achieved by embedding the relevant information in a vector format or through other encoding techniques, which are then used as input conditions for the model during training [72]. By conditioning the model on these learned representations, it is possible to sample novel chemical structures that possess similar or improved properties to those in the training dataset. This process enables the model to internalize both chemical sequences and their corresponding properties, thereby facilitating the generation of compounds that fulfill specific chemical or biological criteria. For instance, CL can facilitate the generation of drug-like molecules with optimized solubility, binding affinity, or metabolic stability, thereby significantly reducing the computational resources and time typically required for such tasks [73, 74].

Evaluation metrics for deep generative models

So far, we have presented a brief overview of deep learning architectures and bias techniques. However, it is crucial to evaluate the performance of these models fairly and objectively. To tackle this issue, Polykovskiy, et al. have introduced a set of evaluation metrics to identify common problems in generative models, such as overfitting, imbalanced frequent structures, and model collapse [24].

One of the key requirements for generative models is their ability to learn the syntax and grammar of chemical language models. To determine whether a model can generate valid chemical entities, a validity metric (Eq. 1) is introduced using RDkit software to calculate the percentage of chemical entities that do not violate basic chemical rules [75], It is recommended that this metric be calculated for at least 30,000 molecules.

$$Validity(V_m) = \frac{Valid\ molecules}{Molecules\ produced} \tag{1}$$

Equation 1. Validity model equation.

In addition, to evaluate the ability of the model to generate molecules while preserving uniqueness, we employ a uniqueness metric. This metric represents the

first 1,000 or 10,000 unique valid molecules produced by the model. The higher the uniqueness value, the better the model's performance, calculated using Eq. 2.

$$Uniqueness = \frac{set(V_m)}{V_m} \tag{2}$$

Equation 2. Uniqueness model's equation.

Additionally, to determine if the model is experiencing overfitting, novelty metrics measure the ratio of valid molecules (V_m) that do not appear in the training dataset (T_d). This indicates the model's proficiency in learning the data distribution and generating unique molecules. Larger values of novelty indicate a lower level of overfitting. The calculation of novelty is illustrated in Eq. 3.

$$Novelty = 1 - \frac{V_m}{T_d} \tag{3}$$

Equation 3. Novelty model's equation.

Methodology.

Literature search and screening

Following the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) methodology for a systematic review and meta-analysis, we conducted a thorough review to complete this manuscript. The PRISMA guidelines stress transparent and complete research practices, which we uphold in this review. The subsequent section discusses all relevant items in detail.

The sources for this study are derived from a peer-reviewed online database, as illustrated in Fig. 2. This research only considered articles indexed in Scopus, Web of Science, and Google Scholar. Advanced search filters were implemented in each scientific search engine to restrict findings to articles issued from January 2020 to June 2024. Boolean statements were implemented to retrieve articles, and the queries used are presented below:

1. "Molecule Generation" AND ("Deep Learning" OR "Artificial Intelligence").
2. ("Chemical Language Models" OR "Molecular Generative Models") AND ("Deep Learning" OR "Artificial Intelligence").
3. ("Chemical Language Models" OR "Molecular Generative Models" OR "Molecule Generation") AND ("Deep Learning" OR "Artificial Intelligence").
4. "Molecule Generation" AND ("RNNs" OR "Transformer" OR "VAEs" OR "Variational Autoencoders" OR "VAE" OR "GAN" OR "Generative Adversarial Networks" OR "GANs").
5. ("Molecular generation") AND ("recurrent neural networks" OR "transformers" OR "GPT").
6. ("Recurrent neural network" OR "transformer" OR "GPT"OR "VAE" OR "GAN") AND ("drug design" OR "de novo drug").

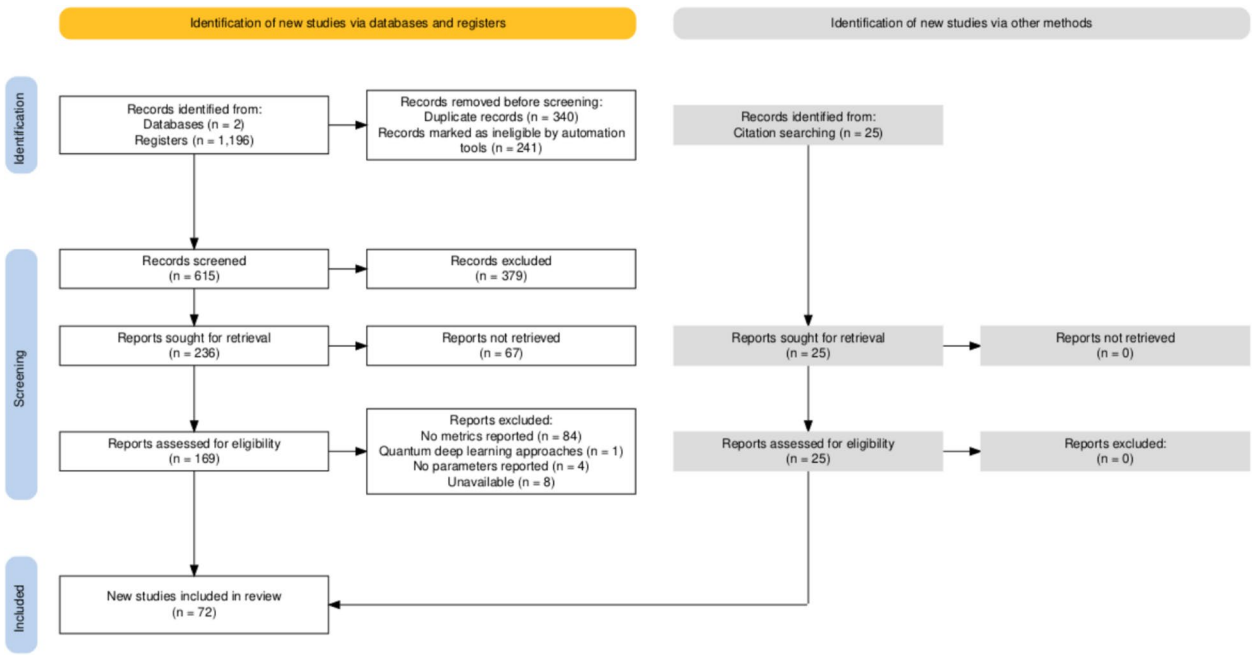

Fig. 2 PRISMA flow diagram showing the study selection process. PRISMA flow diagram was generated using [76].

Eligibility and criteria

As mentioned in the previous section, deep learning methods have transformed the molecular virtual library generation and molecular leads identification processes, leading to a considerable improvement in drug discovery's efficiency and accuracy. Numerous deep learning architectures have been introduced, and various approaches have been developed for designing molecules. This has made it necessary to use impartial metrics to compare and evaluate the pros and cons of various deep learning methods. To address this need for impartial metrics to compare deep learning models, MOSES and Guacamol benchmarking platforms were introduced in 2019 and 2020, respectively [23, 24]. This study was restricted to articles that fulfilled the specified criteria, as there was insufficient content reporting these metrics before their introduction in 2020. The metrics are the following:

1. The manuscript must be written in English.
2. The manuscript explicitly presents at least two metrics of uniqueness, validity, or novelty for the generated compounds (either in the article or in supplementary material).
3. The manuscript presents in detail the concept of uniqueness, validity, or novelty metrics, and these concepts fit the MOSES or GUACAMOL metric concepts.
4. The manuscript focuses on using deep learning generative models to generate de novo molecules.
5. The implemented model uses conventional deep learning generative methods without the use of quantum computer methods.
6. The manuscript was published between January 2020 and June 2024 in a peer-reviewed journal or pre-print services.

Data collection

For the analysis, we extracted article and journal details such as title, publication year, journal name, and the Scimagojr quartile category for each selected item. Additionally, we identified the database name, training dataset size, and physicochemical criteria used for selecting molecules, including K_i, IC_{50}, EC_{50}, LogP, and molecular weight. The data on the molecular representation used for training and the characteristics of molecules in the training dataset were also obtained. The extracted information of molecular representations encompasses the type of molecular representation, the upper and lower limits for the length of representation character and vocabulary for chemical language models, and a binary variable indicating canonization usage for molecular representation. Other aspects considered include the incorporation of stereochemistry in molecular representation, as well as the implementation of a salt-removing procedure. Furthermore, our study collected data on the architecture type, embedding length, number of layers and units in hidden layers, number of trainable parameters, use of dropout, activation temperature for the softmax function, batch size, epochs, learning rate, and optimizer type applied in deep generative architecture. Moreover, various columns that describe the model features were added depending on the features of each kind of deep learning architecture and can be consulted in supplementary material. Additionally, a binary variable was used to determine if the analyzed work employed biased techniques for generating molecules, such as RL, TL, or conditional learning. In cases where biased methods were used, details about the optimization objective of the biased model are provided. Finally, we collected information on the output format of the molecule representation, the number of molecules generated, their uniqueness, validity, novelty, scaffold diversity, scaffold novelty, fragment similarity, similarity to nearest neighbor (SNN), internal diversity, and Fréchet Chemical Distance (FCD) for both biased and unbiased models.

Finally, the selected set of research articles comprised 24 RNNs, 23 Transformers, 16 VAEs, 8 GANs and only 1 article for Structurated State Space Sequences (S4) for molecule generation. Since this work focuses on CLMs, only 10 out of 72 articles relate to graph approaches, which were solely incorporated for comparing CLMs and Graph Neural Network approaches in a general sense. A meta-analysis was performed for RNNs study case using the following outcomes: uniqueness, novelty and validity of the generated subset of molecules.

Results and discussion

Deep generative models

Since their successful introduction as generative models by Gómez-Bombarelli et al. and Segler et al. in 2018, deep learning has emerged as a fundamental tool for de novo molecule design, these have gained the attention of researchers to implement novel deep generative architectures and approaches that have shown to be useful for performing different tasks in other fields such as text generation, image and audio generation and among others [16, 53, 77].

CLMs have emerged as valuable tools for exploring chemical space through one-dimensional molecule string representation. CLMs have demonstrated advantages in molecule generation compared to other approaches. This is illustrated in Fig. 3, which depicts the growing disparity between research employing CLMs and other

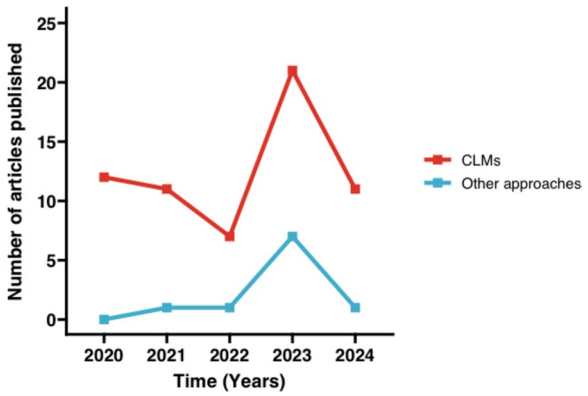

Fig. 3 Comparison of the number of deep generative models article publications from 2020 to June 2024

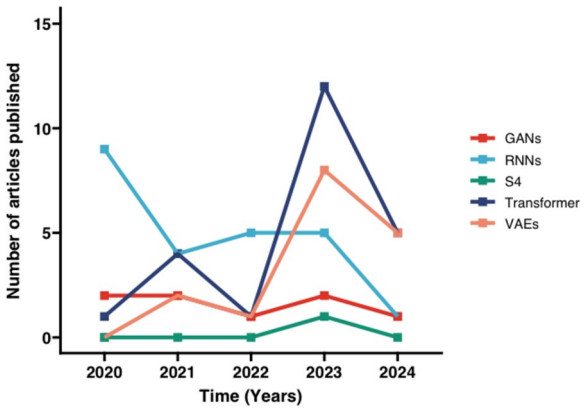

Fig. 4 Comparison of deep learning architecture models published from 2020 to June 2024

methodologies. The rise in publications on molecular generative models employing CLMs can be attributed, in part, to advancements in natural language processing (NLP), which has continued to flourish since the first semester of 2024. This is corroborated by the increasing number of publications pertaining to CLMs for molecule generation, which substantiates the efficacy of CLMs in identifying novel and efficacious molecules for the advancement of novel therapies or the treatment of new diseases [78]. Consequently, these methodologies have been effectively implemented in the domain of cheminformatics, employing diverse instruments to facilitate the identification of novel compounds [79, 80]. In contrast, alternative approaches essentially entail the use of graph neural networks for the generation of molecules. The graph models demonstrated the potential to serve as a robust tool for capturing spatial information about molecules, including atomic geometries and molecule topology [81–83]. This approach has been shown to enhance the validity rates and facilitate the incorporation of diverse spatial relations, such as pharmacophore groups and conformation energies [84–86]. However, graph-based models require a significant amount of computational time for training and generation of molecules, which can be up to one order of magnitude higher than that required by traditional CLMs. Additionally, the cost of training in terms of resources can be more expensive for graph-based models, and the ability to explore the chemical space may vary depending on the task due to the nature of the models, which process data and learn patterns in chemical distributions, such as the presence of large rings or branches [87, 87]. In terms of chemical space exploration, CLMs provides a range of architectural options that can efficiently handle one-dimensional molecular representation. These models are relatively straightforward to train and yield high-quality results during the inference stage [36]. Furthermore, they can be readily deconstructed to reveal the underlying generative models, which may eventually become explainable [88].

RNNs and their variations have gained widespread popularity for molecule generation since the inception of CLMs. In the early days of CLMs, RNNs, and their variants were utilized to understand the distribution of sequential data, particularly for SMILES. Objective evaluation was prioritized throughout this research. RNNs have demonstrated their ability to efficiently learn the grammar and syntax of generating or completing SMILES sequences, resulting in molecules with similar property distributions as those in the training dataset [89]. These results are shown in Fig. 4, where RNNs were previously a significant deep learning architecture extensively applied for generative tasks. However, RNNs have seen a reduced usage as molecular generative models due to the long training time and risk of degraded performance associated with long-term dependencies. The surge in this area is largely attributed to the implementation of advanced deep learning designs, such as transformers. These models enable parallel computing and surpass the restrictions of earlier methods by capturing more information about molecular representations more effectively [90].

Figure 4 shows a major increase in the implementation of transformers as generative models in recent years. This can be attributed to their ability to learn data distributions and produce precise predictions by the implementation of self-attention mechanisms and parallel computing [91]. On the other hand, Fig. 4 indicates that autoencoder approaches based on GANs and VAEs, which could be implemented as CLMs or graph models, remain as alternatives for molecule generation but with a low rate of use as generative models compared to sequential models. The observed low usage rate of GANs

and VAEs could be attributed to the high complexity regarding the time and memory space of these methods, as well as their reduced ability to generate large molecules [37, 92]. Despite, the low rate of implementation of autoencoder approaches, VAEs presented a constant rate of implementation as generative models since 2020. In 2023, S4 were introduced as chemical CLMs for drug discovery, demonstrating promising results at generation sequences based on SMILES strings. However, further details regarding the implementation of S4 can be found in the metrics evaluation of CLMs.

Finally, left panel of Fig. 5 presents the percentage of deep learning architectures utilized in the analyzed papers. The data indicates that RNNs and transformer models are the most frequently utilized, while GANs, VAEs, and the recently introduced S4 models are employed to a lesser extent in molecule generation. Notably, the slight difference in utilization percentages between RNNs and transformer models indicates a significant surge in the implementation of transformers since their introduction to the deep learning field in 2017. Contrastingly, RNNs were introduced more than 40 years ago. It is of note that despite the LSTM-RNNs having a greater number of trainable parameters compared to the GRU-RNNs, it remains the most frequently used model within RNN models. This finding could be attributed to the LSTM-RNNs longer exposure time, unlike the more recent introduction of the GRU-RNNs.

Databases and molecular representations

As previously stated in this article, the presence of data is crucial for learning. The efficacy of deep learning models is significantly influenced by the quality of data input. In generating molecular models, structural information about molecules is gathered in different

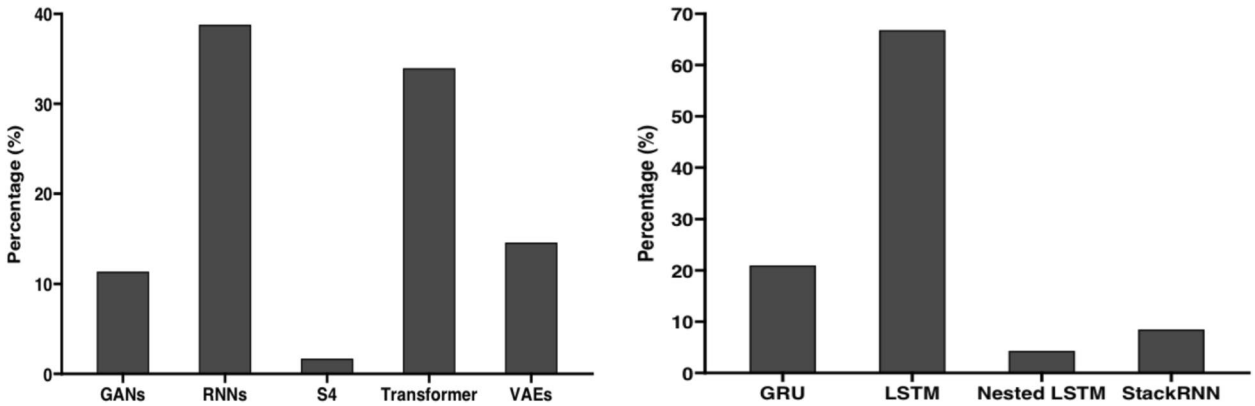

Fig. 5 Relative comparison of architectures used as generative CLMs in systematic review. Left panel represents the overall fraction of the deep learning architectures used in 49 retrieved CLMs. Right panel shows the fraction of RNN variations used in articles that used RNNs as generative CLMs

Table 1 Description of databases used in retrieved articles for analysis.

Database	Description	Number of molecules (millions)	Molecule representation	Articles	Ref
PubChem	Structural information of mostly small molecules	115.3	SMILES and InChI	4	[93]
ChEMBL	Bioactive molecules with drug-like properties and Bioactivity records of data	2.4	SMILES and InChI	27	[94]
Zinc	Structural information of drug-like molecules	750	SMILES	27	[95]
US patent database	Reactions extracted by text-mining from United States patents published between 1976 and September 2016	< 1.8	SMILES	1	[96]
DNA-Encoded Library[a]	Structural molecular, combinatorial screening, and DNA-encoded information	1040	SMILES	1	[97]
COCONUT	Natural products structural and biological information	0.695	SMILES and InChI	1	[98]
LINCS1000	A comprehensive resource of gene expression in human cells perturbated by small molecules	> 1	Not applicable	1	[99]

[a] Indicates databases created by authors and not publicly available, for this case reference indicates the article reference. Number of reported molecules up to September 2024

formats, with SMILES and InChI being the two main formats. Additionally, data about their physicochemical and structural characteristics is collected. Furthermore, molecular databases contain relevant information on bioactivities, biological targets, and other crucial biological data. Table 1 presents the databases used for training models in this study.

Drug-like molecular databases, such as Zinc and ChEMBL, provide information on molecular structures, bioactivity data, validated bioassays, and physicochemical properties. The extensive use of these drug-like databases is linked to current trends in drug design, which focus on deep learning-based methods for drug discovery and design [100, 101]. Approximately 71% of the articles discovered through this systematic review focus on designing and discovering targeted drugs. Of those 71% retrieved articles focused on drug discovery, roughly 9 out of 10 utilized the ChEMBL or Zinc database to train their deep generative models, which are primarily fueled by the curated structural data of synthesizable and validated drug-like molecules. This data serves as an extremely valuable source of information for the deep generative models.

Linked to the significant role that databases play in deep learning, molecular representations also play a crucial part in training deep generative models. Although SMILES has limitations, it is still a widely used molecular format for CLMs. In this study, 77.27% of the models used SMILES exclusively for training. This format allows for data augmentation through randomized SMILES and offers a format that is available in almost all databases in a compact memory format that can be learned easily for deep learning models [36, 37, 102, 103]. On the other hand, the other CLMs articles utilize NLP translation methods to generate targeted compounds using input formats such as target receptor sequences, gene expression signatures, IUPAC names, or physicochemical properties [104–110].

Training dataset size

In practice, data quality alone does not suffice for model training, the quantity of data also plays a crucial role. Therefore, high-quality data are essential for effective training. To learn chemical distributions and patterns, deep generative model training implements probabilistic estimators of data distributions to fit the original data distributions; this process involves the use of large amounts of chemical entities to estimate the unknown parameters and learn the data distribution. In CLMs, the Negative Log Likelihood (NLL) function is implemented for model training to minimize it. This minimization iterative process results in learning unknown parameters to model the chemical data distribution of sequences [18]. Since different models have varying levels of complexity, they require different amounts of data to learn unknown parameters. Figure 6 illustrates the differences in the training dataset used for each CLM in this analysis. Despite the different complexity of the architectures, no statistical difference was observed between the size of the training datasets in different deep generative models. However, a large dispersion for training dataset size in VAEs is observed. This is mainly driven by a value that corresponds to 30 million chemical entities extracted from PubChem to train TransAntivirus, a VAE architecture that uses IUPAC names for SMILES sequence prediction through transformers-based encoder-decoder. The encoders use a transformer that feeds the output prediction of the IUPAC name to the decoder, which is also a transformer that uses the latent representation of IUPAC names to predict the tokens in the SMILES sequences. This process involves learning two different chemical languages with different constraints and syntax, requiring many parameters to map each language system and large amounts of data to train all these parameters [110]. Similarly, an outlier is observed in the size of the Transformers training dataset, encompassing 50 million unique chemical structures retrieved from PubChem. These structures are employed to train a transformer-based structure–property multi-modal foundation model (SPMM), a proposal put forth by Chang and Ch. The aforementioned approach is analogous to that of a sequence-to-sequence model, which is capable of generating molecules from a property vector to a SMILES sequence. This enables the learning of a more nuanced representation of molecules, which is additionally adept at performing a multitude of tasks beyond mere molecule generation. These include property prediction and forward/ retro-reaction prediction without any loss of generation metric values. However, the learning process is exhaustive in terms of the data required, necessitating the learning of millions of molecules [111].

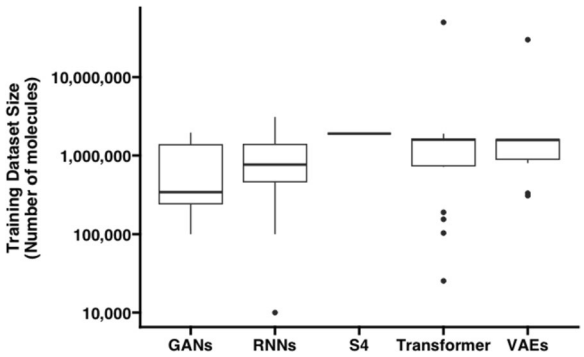

Fig. 6 Boxplot of training dataset size used in different deep learning models for molecular generation

Unbiased models

Validity

Molecule generation can lead to the production of invalid sequences beyond the chemical space. The assessment of molecular generative models' capacity to generate valid chemical entities is supported by a metric that measures the ability of CLMs to understand chemical language. To accurately compare the proportion of valid molecules generated by various architectures using CLMs, Fig. 7 presents the valid fraction of unbiased models as reported in the retrieved articles. No statistically significant differences were observed in the validity medians among the deep generative models, for either group or pairwise comparisons. Overall, Transformer architectures display high validity rates for molecule generation. This is accomplished using self-attention mechanisms, which adjust the size of the latent vector depending on the length of the sequence to retain uncompressed sequence information instead of compressing it into a fixed-length vector that may not effectively represent the interaction between chemical tokens. However, the analysis of Transformer architectures uncovered a molecular generative transformer model that achieves a validity rate as low as 6.9%. This study, conducted by Zhumagambetov et al., focuses on generating virtual libraries of compounds. The researchers discovered that increasing the variability of token sampling and adding Gaussian noise to the transformer decoder is a powerful technique for stochastically sampling molecules in chemical space. This technique can generate molecules with unique chemical structures that generalize well in unseen regions of chemical space. However, it is also prone to generating chemical entities that do not belong to the chemical space, thus reducing the validity ratio [112].

Although there are no outliers for GANs in Fig. 7, a significant data dispersion can be attributed to the constraints of learning chemical language from other biochemical sequences associated with the molecule target reaching values of validity as low as 8.5%. This strategy presented by Méndez-Lucio et al. is an important demonstration of the potential of GANs to generate targeted molecules resembling active compounds using only gene expression and SMILES as training data. This approach reduces the need for extensive information on bioactive molecules [104]. The model learns to generate both bioactive and realistic chemical entities in SMILES format from gene expression profiles. This approach improves the probability of generating molecules that induce a desired transcriptomic profile but comes at the cost of generating a high proportion of molecules that violate valence chemistry rules due to the challenging tasks of learning how to map valid and real chemical structures in chemical space and simultaneously learning SMILES syntax and grammar rules from gene expression signatures. Subsequently, other approaches have been developed to generate phenotype-tailored compounds using stained cell images as inputs, driven by the high cost of obtaining gene expression signatures and poor validity achieved in previous works. This proof-of-concept utilizes generative models and cellular morphology information to design compounds that have the potential to induce a desired biological response with a high validity rate (56.6%) by implementing SELFIES as the molecular representation [113]. Moreover, subsequent to the introduction of Mendez-Lucio, which generated targeted molecules from gene expression data, Liu and colleagues proposed a TransGEM. This methodology employs cell line and gene expression embeddings to create molecules using SELFIES sequences. The method is based on an encoder-decoder Transformer architecture, which has significantly enhanced the distribution generation metrics. These values now reach 85% for uniqueness and 100% for validity and novelty. However, the internal diversity is relatively low (79%), indicating that the model is unable to generate a diverse range of groups within the generated set. Nevertheless, this approach has demonstrated that the use of SELFIES and a tenfold binary encoding representation for the gene expression values, which are subsequently embedded into a dense vector, contains more information than other representations presented in the article (gene expression (GE) values and GE one-hot vectors). This representation serves as an input for the transformer decoder, which learns the interaction information between gene expression data and molecule embeddings, resulting in high distribution metric values [114].

In addition, other GAN-based CLMs utilizing discrete molecule representations have emerged as an active research area aimed at mitigating challenges, like model collapse, associated with generative models. For this

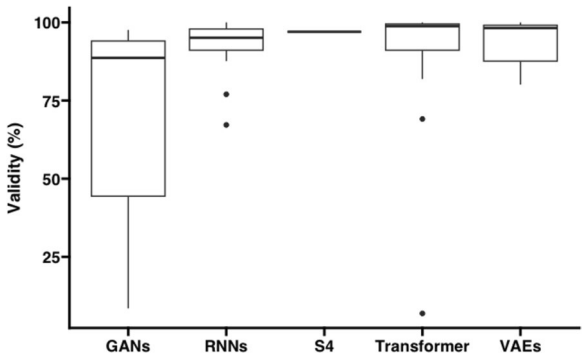

Fig. 7 Validity boxplot of unbiased models evaluated using at least 1000 chemical compounds generated by deep learning models

purpose, NLP methods have been adapted for molecular generative tasks, including MaskGAN, which generates molecules using SMILES as a molecular representation and a text fill-in-the-blank strategy [115]. Nonetheless, human expertise or other computational approaches are required to determine which scaffolds must be filled to generate novel compounds with potential therapeutic applications, limiting this approach to a proof-of-concept for CLMs that could achieve high molecular validity when the masking ratio is around ten percent of the sequence [116]. On the other hand, three different strategies implementing the idea proposed by Zhao and colleagues have used their Adversarially Regularized Autoencoders (ARAE) to train molecular GANs [117]. This approach is based on combining discrete autoencoders with GANs. Specifically, VAEs map discrete molecule representations onto a latent space. Generated latent vectors are then utilized to estimate the discrete distribution of molecules with the assistance of a GAN through adversarial training. ARAE sidesteps typical issues that arise when GANs attempt to learn discrete representations of molecules, which can result in a model collapse problem. The Conditional Adversarially Regularized Autoencoder (CARAE) approach was the first to introduce ARAEs to molecular representation using SMILES as discrete molecular representations. This approach includes a conditional module that can sample molecules with similar properties by tuning the latent vector using a property vector. Ultimately, the tuned latent vector is manipulated by the VAE decoder to reconstruct the original molecule [118]. Similarly, the cross-adversarial learning method for molecular generation (CRAG) approach utilizes ARAE with Projected Gradient Descent (PGD) to generate adversarial samples. The use of PGD leads to data augmentation without changing the actual molecule distribution, effectively addressing the challenge of precisely estimating representation distribution [119]. Without a doubt, both models can produce a high percentage of valid molecules, reaching values of 90.3% and 97.6%. In contrast, previous ARAE methodologies exhibited inadequate decoding proficiency for valid SMILES sequences, achieving values as low as 30.7% [120]. This shortcoming was primarily due to the lack of a smoothed latent landscape, resulting in empty areas that were later sampled by GANs, leading to the creation of invalid molecules by the decoder. This limitation is readily addressed by CARAE and CRAG through the incorporation of a property prediction neural network which is jointly trained with VAEs. This leads to an organized latent space that has a soft transition between encoded molecules with different properties, which significantly reduces the risk of GANs sampling empty spaces [53]. Furthermore, alternative methodologies for the generation of realistic chemical structures have been proposed, employing Generative Adversarial Imitation Learning (GAIL) [121]. This approach utilises a discriminator to direct the actor in emulating expert behaviour, through the training of a transformer to generate SMILES. In contrast, the contrastive discriminator is trained with a set of chemical structures exhibiting specific properties, with the objective of retraining the transformer and generating molecules that are analogous to the target structures. Although this model is susceptible to generating invalid molecules when working with SMILES, which can be attributed to the limited data used for training (~ 350 k chemical sequences), when SELFIES are employed as a molecular representation, the model is capable of generating a perfect validity score while maintaining internal diversity metrics, indicating that the model is capable of generating diverse chemical sequences that conserve the chemical properties of the training dataset in contrast to the expected behavior of generative CLMs presented by Skinnider [122]. However, VAEs have not shown any significant differences compared to other deep learning architectures, and their median validity results are similar to the median found for GANs.

In the final consideration, the multitude of versions of RNNs, these architectures were compared generally as a family with other architectures. The results revealed that their ability to generate valid molecules is close to transformer-based methods. To further investigate the behavior of individual RNN variants, the validity fraction of unbiased molecules generated by each variant was evaluated and presented in Fig. 8. Statistical analysis indicates no significant difference between the various recurrent deep architectures using group and pairwise median comparisons. It is important to note that only one work that implemented a Nested LSTM (NLSTM)

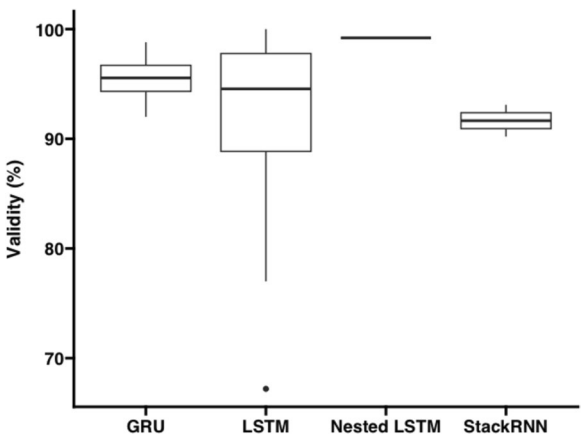

Fig. 8 Validity boxplot of unbiased models evaluated using at least 1000 chemical compounds generated by RNNs variants

for molecule generation was included in the analysis. This differs from LSTM or GRU in that it has a multi-level memory, which allows for a higher degree of freedom than other recurrent neural network (RNN) methods. This enables the model to handle internal memory over a longer range, but this feature comes at the cost of increasing the number of trainable parameters in memory cells and therefore the time needed to train in comparison to conventional RNN architectures. The NLSTM employs a pre-trained model to represent the tokens of molecules in a dense vector derived from the pre-trained Mol2Vec model. This provides more detailed information about chemical properties to each token. Subsequently, the NLSTM is capable of effectively handling the relation between tokens, thereby enabling the generation of molecules that are 97.6% valid even when the sample temperature of softmax is reduced to 0.75 [123].

To complete the analysis of validity in CLMs, validity was compared to the ability of models to generalize the learning of sequences in CLMs by producing novel molecules. In this context, while maintaining the ability to generate valid molecules, ideal generative models would allow the exploration of a larger chemical space by generating novel entities not present in the training dataset. To this end, we propose the metric Valid/Sample (Validity x Novelty) to compute the ratio of valid to novel molecules generated by deep learning structures, similar to Hong, et al. Novel/Sample metric. [118] After reviewing literature that reported the validity and novelty of unbiased models, we calculated the Valid/Sample ratio and used a box plot to identify low-value outliers. These outliers were mainly attributed to discontinuous sampling in latent molecular space, low rates of generalization learning, or inadequate methods for pre-processing chemical sequences [112, 124, 125].

To analyze the relationship between valid and novelty values, we eliminated outliers from the analysis. Figure 9 demonstrates that there is not a significant relationship between the validity and novelty values of generative models (p-value = 0.0618). However, a trend of the inverse relationship between the values of validity and novelty is noticeable (spearman coefficient, $\rho = -0.3575$), wherein 82.1% of models fail to attain values equal to or exceeding the median of both novelty and validity (95.6% and 96.5% respectively). This is attributed to the inherent balance of exploring the chemical space and generating valid chemical sequences. [126] Only a small fraction (17.9%) of the analyzed deep generative models can achieve high novelty values while retaining a high validity ratio of chemical entities. Thus, articles use different approaches to maximize validity and novelty such as reducing the gap between molecular representation in latent space by using ARAEs or VAEs approaches [110, 119]. In addition, fill-in-the-blank based strategies to generate novel molecules reduce the degrees of freedom of molecular complexity for molecule generation, resulting in a high validity rate that can be implemented either for SMILES or SELFIES, wherein due to the intrinsic properties of SELFIES models can generate 100% valid and novel chemical sequences [36, 127]. In a similar fashion to fill-in-the-blank methods, decorative approaches can reduce the complexity of molecule generation by decorating a carbon backbone with other functional groups [128]. It is noteworthy that a recent development of transformer-based variational autoencoders (VAEs) has been implemented by Zhu and colleagues. This approach focuses on the target generation of molecules using pharmacophore information. In this method, the information of pharmacophore groups is embedded using a gated graph neural network. Then, SMILES are embedded using a masking language model. Subsequently, all of the aforementioned information is utilized to generate a latent vector through an encoder transformer, which is then employed to train the encoder-decoder transformer in addition to the pharmacophore information embedding. In conclusion, novel molecules with a given pharmacophore group are generated by using a latent vector N(0,I) during the inference stage. This strategy is capable of designing drug molecules based on the superposition of known active molecules when the target is unknown, or the binding site is unclear. It achieves high novelty (97.6%), validity (98.2%), and uniqueness (97.9%) rates even when SMILES strings, which are prone to syntax errors, are used for molecule generation. This is due to the implementation of a large number of attention mechanisms during the training stage, which capture all the necessary relationships between sequences to ensure [129].

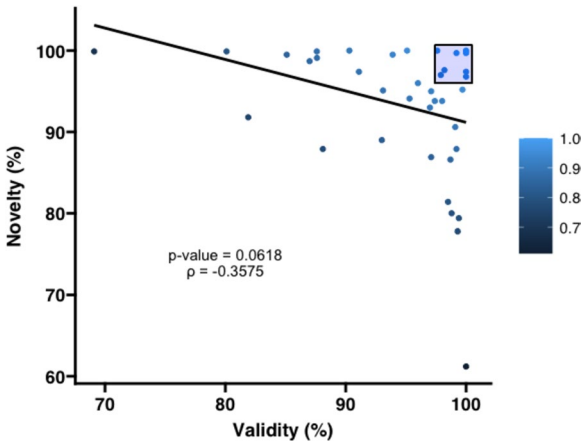

Fig. 9 Novelty trend respect to validity of unbiased models. Color scale indicates Validity x Novelty metric

One of the most promising deep learning architectures recently introduced is DRAGONFLY, which employs a LSTM as CLMs to generate molecules and graph neural network to encode the information of interaction between ligands and targets. This approach integrates information about receptors from graph neural networks, which are used to train the LSTM to learn the distribution of SELFIES that corresponds to each ligand based on its spatial 2D and 3D information about the receptor. This approach optimizes the acquisition of information regarding interaction networks between drug targets and their ligands while maintaining high-quality distribution metrics [130].

Uniqueness

Since the uniqueness of models represents a fundamental metric to compare the ability of models to generate diverse sets of chemical sequences, we evaluated and compared the uniqueness of unbiased models reported in CLMs with different architectures. Figure 10 shows boxplots of each general architecture used for CLMs and molecule generation, indicating no significant differences in medians between generative models. In general, deep generative models can reach high uniqueness values for molecular generated sets.

In addition, Fig. 10 shows the evaluation of a common obstacle faced by generative CLMs: the ability to generate unique compounds during the inference stage. Despite the advantages in terms of enhancing validity and novelty through scaffold decoration and fragment-linking approaches, the generation of unique molecules comes at a cost. These approaches reduce the chances of producing invalid molecules while maintaining an acceptable level of novelty. Nevertheless, the ability to produce unique chemical sequences is compromised due to redundancy problems associated with the training process. These records provide clear evidence that

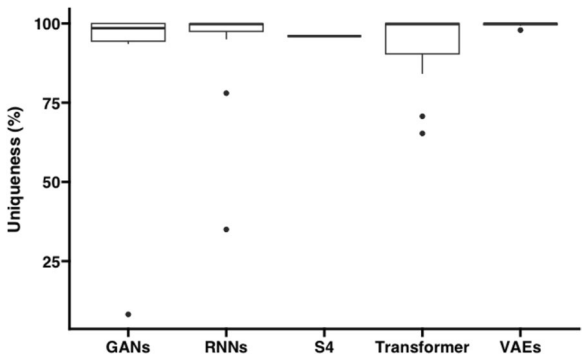

Fig. 10 Uniqueness box plot of unbiased models evaluated using at least 1000 chemical compounds generated by deep learning models

these models are susceptible to experiencing overfitting [74, 106, 124, 131, 132]. In contrast, incorporating the dynamic addition of blank positions to be filled or generating a token matrix instead of sequentially generating tokens to decorate scaffolds could improve the uniqueness ratio of generated molecules for CLMs [127, 128]. Additionally, transformer-based models utilizing sequential text generation for de novo molecule generation have demonstrated near-perfect or perfect uniqueness scores. This is achieved through the implementation of a masked self-attention mechanism, which prevents the model from attending further tokens in sequences and therefore prevents overfitting [61, 133–136]. While other architectures have concentrated on implementing RNNs or pooling algorithms to compute the attention mechanism and enhance the generation metrics of transformers, resulting in marginal improvements in generation metrics such as uniqueness, they have also led to a significant increase [138] In the number of trainable parameters and training time [137]. This suggests that GPT approaches are sufficient for chemical space exploration when the training data is sufficient to train a robust model. Nevertheless, this type of model may prove beneficial in instances where conditional generation is required or when particular properties must be present in chemical structures.

For this purpose, 4 out 9 CLMs VAE-based architecture met the analysis requirements and proved to be a potent tool that employs transformers and mask self-attention in their encoder-decoder structures, enabling accurate learning of information from IUPAC and SMILES sequences. This method, previously discussed for evaluating novelty and validity, has significant potential for designing and optimizing molecules. However, it requires extensive data and time for training, which may be a disadvantage in terms of requirements[110]. In addition, a novel approach called Generative Chemical Transformer (GCT) was introduced, which embeds transformer architecture in VAEs; this strategy takes advantage of the attention mechanism and its ability to pay sparse attention in chemical sequences to deeply understand the geometric structure of SMILES beyond the limitations of semantic discontinuity of chemical language, resulting in good performance in generalizing learning from SMILES sequences and thus drastically reducing the number of repetitive chemical entities produced by the model [138]. Nevertheless, alternative approaches that are based on the utilization of transformers as the foundation of VAE have been proposed by Yoshaki and colleagues. The authors put forth a proposal wherein the normalized latent vector, which has been learned from transformer encoding, is employed to feed a transformer decoder in conjunction with the SMILES embedding, thereby

enabling the generation of novel structures. This implies that the transformer decoder is capable of learning the distribution of SMILES from an embedding vector that contains both the latent vector and the SMILES embedding. This approach has the potential to yield results that are both unique and novel, with a value approaching that of a perfect score. However, this approach also has limitations, as it results in a loss of validity, with a score of 87%, due to the inability to smooth the decoding process in structures that do not adhere to the established chemical rules [139]. In contrast, Inukai and colleagues The authors proposed a transformer-based VAE that employs fragment tokenization to extract conformational molecules, as opposed to character-level tokenization and tree positional encoding. The former is used to generate embeddings that enrich the information about the position of fragments and simplify the handling of large and complex molecules. These are then fed into a decoder, while the latent vector generated from the encoder transformer is used to feed the cross-attention in the decoder transformer. This approach yields high distribution metrics (exceeding 97% for each metric) while reducing the time needed to train and generate large libraries in comparison to existing VAE architectures [140]. While other non-transformer-based VAE models have demonstrated the potential for high uniqueness and validity values, the implementation of NRC-VABS is a noteworthy example. NRC-VABS is a normalized reparametrized conditional LSTM-based VAE that employs a beam search to decode the latent vector. The model introduces Hx SMILES as a novel molecular representation, enabling the probabilistic model to learn long-term dependencies in SMILES strings and reducing complexity through the addition of two characters (^_), which indicate rings, and only closing brackets, which indicate branches. All of these are followed by a number that indicates the length of the ring or branch. The NRC-VABS approach allows for the exploration of chemical space in the surrounding area of targeted molecules. This is accomplished by creating a smoothed latent space that can be interpolated to generate novel molecules through a beam search. This approach permits the introduction of a variety of functional groups while preserving structural similarities with the targeted molecule, thereby maintaining the desired properties [141]. In conclusion, VAEs approaches achieve high uniqueness values with low dispersion; however, they are associated with model collapse or the generation of invalid molecules due to sparse latent space generation. This last limitation can be addressed by incorporating a prediction model that can accommodate the intricate latent space, as demonstrated by Liu et al. in their GRU-based VAEs for the generation of Alzheimer's drug molecules [142]. Thus far, the discussion has focused exclusively on traditional deep learning architectures. However, the introduction of S4 models as CLMs has been observed to exhibit high uniqueness values. These architectures are neural networks designed to handle long-range dependencies in sequential data, such as time series or natural language. The objective is to effectively capture both short-term and long-term patterns, which are often challenging for traditional models, such as recurrent neural networks (RNNs) or even transformers, to handle efficiently, particularly for very long sequences. The distinctive dual nature of S4s, encompassing convolution during training and recurrent generation, renders them especially fascinating for de novo design, commencing from SMILES and attaining high-quality metrics. Furthermore, they can be utilized in conditional generation to generate products that closely resemble natural ones [143].

Finally, the RNNs attained comparable levels of uniqueness to other architectures. However, each class of RNNs was evaluated in terms of uniqueness to fully analyze the data. Stack RNNs were excluded from statistical analysis because only one article met the requirements. However, this class of recurrent networks has demonstrated an ability to reach values close to 100% through parallel processing of information in each cell when sampling tokens in the sequence. However, this approach compromises space and time complexity. [144]. The analysis showed no significant difference between the performance of LSTM and GRU. However, LSTM exhibited an anomaly that has been previously studied for RNNs as groups and belongs to scaffold decorations approaches [124].

Biased models

Until now, we have reviewed the characteristics of unbiased models for molecule generation, but still left room to review deep learning strategies for exploring specific regions of chemical space. CLMs have utilized a range of techniques based on biased model weights to acquire knowledge on the general grammar and syntax of molecular representation, as well as the specific configuration and patterns of molecules that exhibit specific physicochemical properties or bioactivities.

Figure 11 illustrates the implementation of biased methods among deep learning architectures. The analysis examined three distinct biased methods, which were utilized with similar frequencies in various deep learning architectures. These methods are mainly used to generate targeted molecules when the available data alone is not sufficient to train deep generative models. This is especially important for rare or newly discovered targets that do not have enough data to train deep learning models.

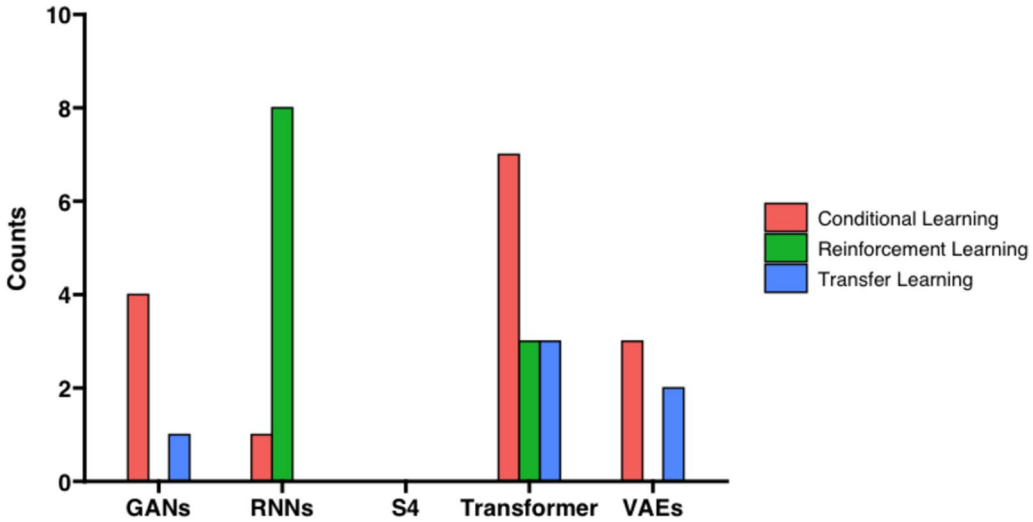

Fig. 11 Distribution of biased methods among different deep generative models

Table 2 Descriptive statistics and P-values are presented for the performance and training datasets of generative models that implemented TL

	Unbiased model	Target model	Samples	P-value
Training dataset size	1,128,920	2507	17	< 0.0001
Validity	98.05	95.5	10	0.1602
Uniqueness	97.9	90.2	11	0.0144
Novelty	91.6	96.0	8	0.8438

Median values are reported for both the Unbiased and target models, and P-values were calculated through the Mann–Whitney U test for paired samples. The number of articles meeting the required metric reporting criteria for analysis is indicated by the sample's column

Transfer learning

Among biased strategies, TL is the most prevalent method due to its user-friendly implementation strategies that are suitable for almost all architectures included in this analysis. For this specific case, fine-tuning is the commonly used transfer learning method for generating targeted molecules. This technique transfers previously learned parameters to the target model, which is responsible for recognizing specific patterns like common scaffolds, functional groups, and atomic configurations without having to re-learn the syntax and grammar of the chemical language. Fine-tuning strategies have proven to be effective tools for exploring chemical space in sequential-based models, as detailed in Table 2.

Such approaches do not compromise the validity and novelty of generated molecules compared to unbiased models and require significantly fewer chemical sequences to train the model. However, the cost of this reduction in training data is a decrease in the variety of generated molecules probably due to overfitting.

Fine-tuning has enabled the exploration of chemical space for areas containing molecules with bioactivities linked to proteins that are commonly expressed at high levels in various solid tumors, such as the Epidermal Growth Factor Receptor (EGFR). The resulting chemical sequences are accessible for chemical synthesis, and they have produced promising molecular docking results. Additionally, QSAR models suggest activities for EGFR [82, 107, 134]. Additionally, recent reports have highlighted the emergence of virtual libraries consisting of targeted chemical entities generated using deep generative models and transfer learning techniques. These models focus on generating molecules with the potential to target specific Alzheimer's-related proteins such as BACE-1 and ADAM10, schizophrenia-related targets including D2R, 5-HT1AR, and HT2AR, the Cannabinoid CB2 target, SARS-CoV-2 proteases or compounds exhibiting specific chemical properties [97, 103, 109, 124, 125, 145–149]. Finally, transfer learning has demonstrated the potential to generate target molecules for designing functional compounds that can be synthesized and experimentally tested against the Phosphatidylinositide 3-kinase receptor. This leads to the discovery of therapeutical leads with sub-micromolar activity, inhibiting the growth of cancerous tumor cells significantly in in vitro models [36].

Reinforcement learning

Unlike biased models used in transfer learning, RL is used for planning or decision making in sequential processes such as generation of de novo molecules. The

implementation of discriminative or predictive neural networks enables generative models to explore specific regions of the chemical space where bioactive or specific structures may exist. RL has been implemented in various deep learning models. However, this analysis has only been found to be applied to RNNs and Transformers (see Fig. 11). Table 3 presents the statistical summary of biased models that utilize RL for targeted molecule generation, revealing that the process of reinforcement learning does not affect molecular metrics.

RL utilizes models to guide generation models in learning parameters that enable molecular decoding in specific regions of chemical space, often utilizing scoring functions and predictive/classification deep learning methods. A wide variety of models and scoring functions have been introduced as agents for RL. First, 90.1% of the articles retrieved that employed reinforcement learning have opted to use Policy gradient methods. Policy gradient uses a score function that returns a value given the chemical properties of the generated molecules, this approach aims to find parameters that maximize the score function that induces the molecular generator to explore in a certain direction of the chemical space, since molecules that can achieve, higher rewards have tended to group in the chemical space. These methods have implemented many predictive modules to calculate the properties of generated molecules using machine learning or deep learning approaches. For example, linear regression-based models enable the computation of a range of physiochemical characteristics, molecular similarity, synthetic viability, and binding affinity metrics for generated chemical entities. These modules have proven to be a powerful tool in generating targeted chemical libraries for potential inhibitors of SARS-CoV-2, acetylcholinesterase, neuraminidase, and κ-opioid receptor [64, 144]. Additionally, other predictive models that employ deep learning methods, including multilayer perceptron, convolutional neural networks (CNNs), and LSTMs, among other deep learning architectures, enable the use of generated sequences. These models are not limited to using only chemical descriptors, outperforming chemical descriptors-based methods, and thus avoiding the addition of bias to specific regions of chemical space [17]. These methods enabled optimizing the properties of molecules and generating novel chemical entities with potential activity to permeate the blood–brain barrier [61, 62, 150]. Additionally, other strategies that implements policy gradient, such as Hill-Climbing, augmented Hill-Climb, and REINVENT algorithms, have demonstrated the ability to achieve state-of-the-art results for molecular generation metrics in terms of validity, novelty and uniqueness [131, 151, 152].

Finally, Monteiro et al. have introduced a novel approach based on evolutionary algorithms, demonstrating the potential of combining reinforcement learning, transfer learning, and nondominated sorting algorithms to generate unique and valid chemical sequences with desirable physicochemical and pharmacological properties for targeting biologically relevant molecules, such as the Adenosine A2A receptor (AA2AR), for therapeutic applications. This approach effectively handles the complex trade-off between validity and novelty, leading to unprecedented levels of uniqueness, novelty, and validity for the targeted molecules [136].

Table 3 Descriptive statistics and P-values are presented for the performance of generative models that implemented RL

	Unbiased model	Target Model	Samples	P-values
Validity	91.1	96.5	9	0.1289
Uniqueness	99.9	89.7	7	0.0935
Novelty	91.5	93.5	4	0.2500

Median values are reported for both the Unbiased and target models, and P-values were calculated through the Mann–Whitney U test for paired samples. The number of articles meeting the required metric reporting criteria for analysis is indicated by the sample's column

Conditional generation-based methods

Conditional Generation-Based Methods have emerged as a tool for generating chemical entities that satisfy desired properties. Conditional Generation is a technique that involves adding constraint tokens to chemical sequences that are not part of the molecule representation vocabulary. This enables the labeling of chemical entities, which in turn trains the model to recognize conditional chemical sequences. Consequently, the model learns to map similar constrained molecules on a latent space, thereby introducing a natural bias that facilitates the generation of molecules with similar properties. This method of conditioning structure generation eliminates the requirement for optimization loops and retraining epochs for fine-tuning. Since its ability to map molecules and generate latent spaces, encoder-decoder architectures like ARAEs, VAEs, and Transformers have been widely used to implement conditional generation strategies, as illustrated in Fig. 11. Furthermore, to assess the impact of conditional generation based methods on the generation metrics proposed in MOSES, despite the low number of samples for comparing metrics of generated chemical sequences, Table 4 demonstrates that conditional learning strategies can produce molecules as effectively as unbiased models that solely focus on learning chemical languages. The effectiveness of conditional learning in generating natural-like products has been demonstrated

Table 4 Descriptive statistics and P-values are presented for the performance of generative models that implemented conditional generation-based methods

	Unbiased model	Target model	Samples	P-value
Validity	98.5	96.8	11	0.4648
Uniqueness	99.9	97.5	10	0.0753
Novelty	89.3	99.6	8	0.2945

Median values are reported for both the Unbiased and target models, and P-values were calculated through the Mann–Whitney U test for paired

using Transformer architectures such as NIMO, which learns to reproduce molecules similar to terpenoids with high validity, novelty, and uniqueness values. This approach does not require exhaustive data sets or extensive computational resources, offering the potential to generate drug-like molecules with natural product scaffolds. These scaffolds may possess favorable ADME properties and minimal environmental impact [153].

Multi-objective compound optimization has been motivated by conditional generation methods as it possesses the ability to utilize a specific set of chemical properties to constrain the model producing virtual chemical libraries that not only can potentially be used as therapeutic agents by its structural similarity, but they may also share physicochemical properties. This achievement is demonstrated by the creation of virtual chemical libraries that can bind to EGFR, HTR1A, and S1PR1 receptors. The libraries were generated using conditional generation to produce compounds with physicochemical properties that follow empirical drug-like rules [135]. Multi-objective optimization is an effective strategy for exploring chemical space and obtaining molecules that follow various physicochemical constraints [133]. Furthermore, it is imperative to convert discrete molecular representations conditioned by the environment into a latent space. The use of VAE, ARAE, and Transformers lead to a viable approach for both multi-objective and single-objective molecular optimization, resulting in validity and uniqueness values reaching up to 85% and being as versatile as working for text-filling approaches and sequential generation, as well as, working with different molecular representations [74, 104, 109, 110, 113, 117, 119, 138].

Finally, other approaches have been introduced that combine more than one biasing strategy to generative models, such as combining conditional generation-based methods with transfer learning or reinforcement learning. These strategies demonstrate significant progress in developing deep generative models, providing a highly effective method for exploring unknown regions of chemical space in search of molecules with specific activities suitable for material or drug discovery [90]. In particular, the implementation of transfer learning and conditional generation-based methods has enabled the generation of molecules using multi-objective optimization, which shows structural features that promote their bioactivity against specific targets while using small datasets for training [134].

Conclusion

Molecule design is fundamental to the discovery of drugs and development of materials, which currently relies mostly on the exploration of chemical space using computational approaches. Within computational approaches, CLMs have been shown to be a versatile tool for exploring chemical space. They enable mapping of chemical space and exploration in specific directions, providing access to regions where bioactive molecules exist by using biased CLMs models. During the period of 2020 to June 2024, this systematic review evaluated deep learning molecular generative models utilizing MOSES metrics. The aim was to assess the model's metrics associated with the overfitting and learning process of chemical language.

Since the dawn of the data age, SMILES has remained the most widely used molecular representation format, owing to its readily accessible format. In contrast, SELFIES presents a promising representation format that can enhance model performance while reducing the trade-off between validity and novelty values. This work has shown that CLMs often use transformers and gated RNN variants as generative models, which is consistent with the trend observed in NLP for text generation that has evolved along with generative CLMs [154]. Nonetheless, in recent years, there has been an increasing tendency to apply the Transformers architecture more often, due to their self-attention mechanism. This allows models to efficiently learn long-term dependencies and thereby efficiently learn chemical language.

Finally, the performance of generative CLMs in terms of validity, uniqueness, and novelty is not statistically affected by targeted generative CLMs. Among biased models, TL has been the most used technique among TL, RL, and conditional learning for deep learning architectures when about one thousand molecular entities are available for optimization. However, incorporating multiple biased methods into deep generative models has proven to be a promising technique for targeted molecular generation. These models could produce molecules with improved chemical properties or bioactivities and can travel longer distances within chemical space to reach areas where specific molecules exist.

We hope this review provides a comprehensive understanding of deep generative models, ranging from their theoretical background to the practical implementation of generative CLMs, and offers a clear perspective on the evolution, progress, and opportunity areas of generative CLMs in recent years.

ABBREVIATIONS

ARAE	Adversarially regularized autoencoders
CL	Conditional learning
CLMs	Chemical language models
GANs	Generative adversarial networks
GRU	Gated recurrent unit
InChI	International chemical identifier
LSTM	Long short-term memory
MOSES	Molecular sets
RNNs	Recurrent neural networks
RL	Reinforcement learning
SELFIES	Self-referencing embedded strings
SMILES	Simplified molecular-input line-entry system
TL	Transfer learning
VAEs	Variational autoencoders

Author contributions

Conceptualization, H.F.-H. and E.M.-L.; methodology, H.F.-H.; formal analysis, H.F.-H. and E.M.-L.; data curation, H.F.-H.; writing and editing, H.F.-H. and E.M.-L.; supervision, E.M.-L. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partially supported by the Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT) (H.F.-H. scholarship) and by the Tecnológico de Monterrey.

Aavailability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 26 July 2024 Accepted: 17 October 2024
Published online: 18 November 2024

References

1. Polishchuk PG, Madzhidov TI, Varnek A (2013) Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Design 27(8):675–679. https://doi.org/10.1007/s10822-013-9672-4
2. Reymond J-L, Awale M (2012) Exploring chemical space for drug discovery using the chemical universe database. ACS Chem Neurosci 3(9):649–657. https://doi.org/10.1021/cn3000422
3. Lu C, Liu S, Shi W, Yu J, Zhou Z, Zhang X, Lu X, Cai F, Xia N, Wang Y (2022) Systemic evolutionary chemical space exploration for drug discovery. J Cheminform 14(1):19. https://doi.org/10.1186/s13321-022-00598-4
4. Maragakis P, Nisonoff H, Cole B, Shaw DE (2020) A deep-learning view of chemical space designed to facilitate drug discovery. J Chem Inf Model 60(10):4487–4496. https://doi.org/10.1021/acs.jcim.0c00321
5. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
6. Krasnov L, Khokhlov I, Fedorov MV, Sosnin S (2021) Transformer-based artificial neural networks for the conversion between chemical notations. Sci Rep 11(1):14798. https://doi.org/10.1038/s41598-021-94082-y
7. Karpov P, Godin G, Tetko IV (2020) Transformer-CNN: swiss knife for QSAR modeling and interpretation. J Cheminform 12(1):17. https://doi.org/10.1186/s13321-020-00423-w
8. Chuang KV, Gunsalus LM, Keiser MJ (2020) Learning molecular representations for medicinal chemistry. J Med Chem 63(16):8705–8722. https://doi.org/10.1021/acs.jmedchem.0c00385
9. Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H, Wang H (2022) Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell 4(2):127–134. https://doi.org/10.1038/s42256-021-00438-4
10. Li X, Fourches D (2020) Inductive transfer learning for molecular activity prediction: next-gen QSAR models with molpmofit. J Cheminform 12(1):27. https://doi.org/10.1186/s13321-020-00430-x
11. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V (2015) Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model 55(2):263–274. https://doi.org/10.1021/ci500747n
12. Vanhaelen Q, Lin Y-C, Zhavoronkov A (2020) The advent of generative chemistry. ACS Med Chem Lett 11(8):1496–1505. https://doi.org/10.1021/acsmedchemlett.0c00088
13. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 52(11):2864–2875. https://doi.org/10.1021/ci300415d
14. Goel M, Raghunathan S, Laghuvarapu S, Priyakumar UD (2021) MoleGuLAR: molecule generation using reinforcement learning with alternating rewards. J Chem Inf Model 61(12):5815–5826. https://doi.org/10.1021/acs.jcim.1c01341
15. Queiroz LP, Rebello CM, Costa EA, Santana VV, Rodrigues BCL, Rodrigues AE, Ribeiro AM, Nogueira IBR (2023) Transfer learning approach to develop natural molecules with specific flavor requirements. Ind Eng Chem Res 62(23):9062–9076. https://doi.org/10.1021/acs.iecr.3c00722
16. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci 4(1):120–131. https://doi.org/10.1021/acscentsci.7b00512
17. Pereira T, Abbasi M, Ribeiro B, Arrais JP (2021) Diversity oriented deep reinforcement learning for targeted molecule generation. J Cheminform 13(1):21. https://doi.org/10.1186/s13321-021-00498-z
18. Arús-Pous J, Blaschke T, Ulander S, Reymond J-L, Chen H, Engkvist O (2019) Exploring the GDB-13 chemical space using deep generative models. J Cheminform 11(1):20. https://doi.org/10.1186/s13321-019-0341-z
19. Li X, Xu Y, Yao H, Lin K (2020) Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors. J Cheminform 12(1):42. https://doi.org/10.1186/s13321-020-00446-3
20. Li L, Gupta E, Spaeth J, Shing L, Jaimes R, Engelhart E, Lopez R, Caceres RS, Bepler T, Walsh ME (2023) Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. Nat Commun 14(1):3454. https://doi.org/10.1038/s41467-023-39022-2
21. Li Y, Zhang L, Wang Y, Zou J, Yang R, Luo X, Wu C, Yang W, Tian C, Xu H, Wang F, Yang X, Li L, Yang S (2022) Generative deep learning enables the discovery of a potent and selective RIPK1 inhibitor. Nat Commun 13(1):6891. https://doi.org/10.1038/s41467-022-34692-w
22. Saka K, Kakuzaki T, Metsugi S, Kashiwagi D, Yoshida K, Wada M, Tsunoda H, Teramoto R (2021) Antibody design using LSTM based deep generative model from phage display library for affinity maturation. Sci Rep 11(1):5852. https://doi.org/10.1038/s41598-021-85274-7
23. Brown N, Fiscato M, Segler MHS, Vaucher AC (2019) GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model 59(3):1096–1108. https://doi.org/10.1021/acs.jcim.8b00839
24. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M, Kadurin A, Johansson S, Chen H, Nikolenko S, Aspuru-Guzik A, Zhavoronkov A (2020) Molecular sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol 11:565644
25. Mater AC, Coote ML (2019) Deep learning in chemistry. J Chem Inf Model 59(6):2545–2559. https://doi.org/10.1021/acs.jcim.9b00266
26. Zeng X, Wang F, Luo Y, Kang S, Tang J, Lightstone FC, Fang EF, Cornell W, Nussinov R, Cheng F (2022) Deep generative molecular design reshapes drug discovery. Cell Rep Med 3(12):100794. https://doi.org/10.1016/j.xcrm.2022.100794

27. Sousa T, Correia J, Pereira V, Rocha M (2021) Generative deep learning for targeted compound design. J Chem Inf Model 61(11):5343–5361. https://doi.org/10.1021/acs.jcim.0c01496
28. Ivanenkov Y, Zagribelnyy B, Malyshev A, Evteev S, Terentiev V, Kamya P, Bezrukov D, Aliper A, Ren F, Zhavoronkov A (2023) The hitchhiker's guide to deep learning driven generative chemistry. ACS Med Chem Lett 14(7):901–915. https://doi.org/10.1021/acsmedchemlett.3c00041
29. Martinelli DD (2022) Generative machine learning for de novo drug discovery: a systematic review. Comput Biol Med 145:105403. https://doi.org/10.1016/j.compbiomed.2022.105403
30. Koutroumpa N, Papavasileiou K, Papadiamantis A, Melagraki G, Afantitis A (2023) A systematic review of deep learning methodologies used in the drug discovery process with emphasis on in vivo validation. Int J Mol Sci 24:6573. https://doi.org/10.3390/ijms24076573
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:1
32. Balaban AT (1985) Applications of graph theory in chemistry. J Chem Inf Comput Sci 25(3):334–343. https://doi.org/10.1021/ci00047a033
33. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. 2017. Neural message passing for quantum chemistry. in international conference on machine learning. 1263–1272.
34. Raghunathan S, Priyakumar UD (2022) Molecular representations for machine learning applications in chemistry. Int J Quantum Chem 122(7):e26870. https://doi.org/10.1002/qua.26870
35. David L, Thakkar A, Mercado R, Engkvist O (2020) Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform 12(1):56. https://doi.org/10.1186/s13321-020-00460-5
36. Moret M, Pachon Angona I, Cotos L, Yan S, Atz K, Brunner C, Baumgartner M, Grisoni F, Schneider G (2023) Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nat Commun 14(1):114. https://doi.org/10.1038/s41467-022-35692-6
37. Flam-Shepherd D, Zhu K, Aspuru-Guzik A (2022) Language models can learn complex molecular distributions. Nat Commun 13(1):3293. https://doi.org/10.1038/s41467-022-30839-x
38. Skinnider MA, Stacey RG, Wishart DS, Foster LJ (2021) Chemical language models enable navigation in sparsely populated chemical space. Nat Mach Intell 3(9):759–770. https://doi.org/10.1038/s42256-021-00368-1
39. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
40. Weininger D, Weininger A, Weininger J (1989) SMILES 2 algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 29(2):97–101. https://doi.org/10.1021/ci00062a008
41. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC international chemical identifier. J Cheminform 7(1):23. https://doi.org/10.1186/s13321-015-0068-4
42. O'Boyle, N.; Dalke, A. Deep SMILES: An adaptation of smiles for use in machine-learning of chemical structures; 2018
43. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A (2020) Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn Sci Technol 1(4):045024. https://doi.org/10.1088/2632-2153/aba947
44. Krenn M, Ai Q, Barthel S, Carson N, Frei A, Frey NC, Friederich P, Gaudin T, Gayle AA, Jablonka KM, Lameiro RF, Lemm D, Lo A, Moosavi SM, Nápoles-Duarte JM, Nigam A, Pollice R, Rajan K, Schatzschneider U, Schwaller P, Skreta M, Smit B, Strieth-Kalthoff F, Sun C, Tom G, Falk von Rudorff G, Wang A, White AD, Young A, Yu R, Aspuru-Guzik A (2022) SELFIES and the future of molecular string representations. Patterns 3(10):100588. https://doi.org/10.1016/j.patter.2022.100588
45. Zheng S, Yan X, Yang Y, Xu J (2019) Identifying structure-property relationships through SMILES syntax analysis with self-attention mechanism. J Chem Inf Model 59(2):914–923. https://doi.org/10.1021/acs.jcim.8b00803
46. Ucak UV, Ashyrmamatov I, Ko J, Lee J (2022) Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. Nat Commun 13(1):1186. https://doi.org/10.1038/s41467-022-28857-w
47. Jaeger S, Fulle S, Turk S (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model 58(1):27–35. https://doi.org/10.1021/acs.jcim.7b00616
48. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558
49. Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of neural machine translation: encoder-decoder approaches. *arXiv preprint* arXiv:1409.1259*; 2014*.
50. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
51. Chen Y, Wang Z, Zeng X, Li Y, Li P, Ye X, Sakurai T (2023) Molecular language models: RNNs or transformer? Brief Funct Genomics 22(4):392–400. https://doi.org/10.1093/bfgp/elad012
52. Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint* arXiv:1312.6114*; 2013*.
53. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4(2):268–276. https://doi.org/10.1021/acscentsci.7b00572
54. Tevosyan A, Khondkaryan L, Khachatrian H, Tadevosyan G, Apresyan L, Babayan N, Stopper H, Navoyan Z (2022) Improving VAE based molecular representations for compound property prediction. J Cheminform 14(1):69. https://doi.org/10.1186/s13321-022-00648-x
55. Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (Organ) for Sequence Generation Models. *arXiv preprint* arXiv:1705.10843 **2017**.
56. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst 1:27
57. Blanchard AE, Stanley C, Bhowmik D (2021) Using GANs with adaptive training data to search for new molecules. J Cheminform 13(1):14. https://doi.org/10.1186/s13321-021-00494-3
58. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci Adv 4:7
59. Zhou Z, Kearnes S, Li L, Zare RN, Riley P (2019) Optimization of molecules via deep reinforcement learning. Sci Rep 9(1):10752. https://doi.org/10.1038/s41598-019-47148-x
60. Atance SR, Diez JV, Engkvist O, Olsson S, De MR (2022) Novo drug design using reinforcement learning with graph-based deep generative models. J Chem Inf Model 62(20):4863–4872. https://doi.org/10.1021/acs.jcim.2c00838
61. Mazuz E, Shtar G, Shapira B, Rokach L (2023) Molecule generation using transformers and policy gradient reinforcement learning. Sci Rep 13(1):8799. https://doi.org/10.1038/s41598-023-35648-w
62. Pereira T, Abbasi M, Oliveira JL, Ribeiro B, Arrais J (2021) Optimizing blood-brain barrier permeation through deep reinforcement learning for de novo drug design. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab301
63. Fang Y, Pan X, Shen H-B (2023) De novo drug design by iterative multiobjective deep reinforcement learning with graph-based molecular quality assessment. Bioinformatics. https://doi.org/10.1093/bioinformatics/btad157
64. Domenico A, Nicola G, Daniela T, Fulvio C, Nicola A, De ON (2020) Novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization. J Chem Inf Model 60(10):4582–4593. https://doi.org/10.1021/acs.jcim.0c00517
65. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J (2020) Transfer learning for drug discovery. J Med Chem 63(16):8683–8694. https://doi.org/10.1021/acs.jmedchem.9b02147
66. Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J (2020) Transfer learning for drug discovery. J Med Chem 63(16):8683–8694
67. 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. Commun Chem 1(1):68
68. Amabilino S, Pogány P, Pickett SD, Green DVS (2020) Guidelines for recurrent neural network transfer learning-based molecular generation of focused libraries. J Chem Inf Model 60(12):5699–5713. https://doi.org/10.1021/acs.jcim.0c00343

69. Pesciullesi G, Schwaller P, Laino T, Reymond J-L (2020) Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nat Commun 11(1):4874. https://doi.org/10.1038/s41467-020-18671-7
70. Singh S, Sunoj RBA (2022) Transfer learning approach for reaction discovery in small data situations using generative model. iscience 25(7):104661. https://doi.org/10.1016/j.isci.2022.104661
71. Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, Bozdaganyan M, Aliper A, Zhavoronkov A, Kadurin A (2018) Entangled conditional adversarial autoencoder for de novo drug discovery. Mol Pharm 15(10):4398–4405. https://doi.org/10.1021/acs.molpharmaceut.8b00839
72. Kang S, Cho K (2019) Conditional molecular design with deep generative models. J Chem Inf Model 59(1):43–52. https://doi.org/10.1021/acs.jcim.8b00263
73. Gebauer NWA, Gastegger M, Hessmann SSP, Müller K-R, Schütt KT (2022) Inverse design of 3d molecular structures with conditional generative neural networks. Nat Commun 13(1):973. https://doi.org/10.1038/s41467-022-28526-y
74. Yang Y, Zheng S, Su S, Zhao C, Xu J, Chen H (2020) SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. Chem Sci 11(31):8312–8322. https://doi.org/10.1039/D0SC03126G
75. Greg Landrum. RDKit: Open-Source Cheminformatics; http://Www.Rdkit.Org. http://www.rdkit.org Accessed 19 Oct 2023
76. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA (2022) PRISMA2020: an r package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 18(2):e1230. https://doi.org/10.1002/cl2.1230
77. Moret M, Grisoni F, Katzberger P, Schneider G (2022) Perplexity-based molecule ranking and bias estimation of chemical language models. J Chem Inf Model 62(5):1199–1206. https://doi.org/10.1021/acs.jcim.2c00079
78. Bajorath J (2024) Chemical language models for molecular design. Mol Inform 43(1):e202300288. https://doi.org/10.1002/minf.202300288
79. Ballarotto M, Willems S, Stiller T, Nawa F, Marschner JA, Grisoni F, De MD (2023) Novo design of Nurr1 agonists via fragment-augmented generative deep learning in low-data regime. J Med Chem 66(12):8170–8177. https://doi.org/10.1021/acs.jmedchem.3c00485
80. Grisoni F (2023) Chemical language models for de novo drug design: challenges and opportunities. Curr Opin Struct Biol 79:102527. https://doi.org/10.1016/j.sbi.2023.102527
81. Iwata H, Nakai T, Koyama T, Matsumoto S, Kojima R, Okuno Y (2023) VGAE-MCTS: a new molecular generative model combining the variational graph auto-encoder and monte carlo tree search. J Chem Inf Model 63(23):7392–7400. https://doi.org/10.1021/acs.jcim.3c01220
82. Hu C, Li S, Yang C, Chen J, Xiong Y, Fan G, Liu H, Hong L (2023) ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks. J Cheminform 15(1):91. https://doi.org/10.1186/s13321-023-00766-0
83. Zhang G, Zhang Y, Li L, Zhou J, Chen H, Ji J, Li Y, Cao Y, Xu Z, Pian C (2024) Exploring novel fentanyl analogues using a graph-based transformer model. Interdiscip Sci 16(3):712–726. https://doi.org/10.1007/s12539-024-00623-0
84. Kong Y, Zhao X, Liu R, Yang Z, Yin H, Zhao B, Wang J, Qin B, Yan A (2022) Integrating concept of pharmacophore with graph neural networks for chemical property prediction and interpretation. J Cheminform 14(1):52. https://doi.org/10.1186/s13321-022-00634-3
85. J Gilmer, SS Schoenholz, PF Riley, O Vinyals, GE Dahl. Neural message passing for quantum chemistry. In international conference on machine learning; PMLR, 2017; pp 1263–1272.
86. Chen B, Pan Z, Mou M, Zhou Y, Fu W (2024) Is fragment-based graph a better graph-based molecular representation for drug design? a comparison study of graph-based models. Comput Biol Med 169:107811. https://doi.org/10.1016/j.compbiomed.2023.107811
87. Zhang J, Mercado R, Engkvist O, Chen H (2021) Comparative study of deep generative models on chemical space coverage. J Chem Inf Model 61(6):2572–2581. https://doi.org/10.1021/acs.jcim.0c01328
88. Wu Z, Chen J, Li Y, Deng Y, Zhao H, Hsieh C-Y, Hou T (2023) From black boxes to actionable insights: a perspective on explainable artificial intelligence for scientific discovery. J Chem Inf Model 63(24):7617–7627. https://doi.org/10.1021/acs.jcim.3c01642
89. van Deursen R, Ertl P, Tetko IV, Godin G (2020) GEN: highly efficient smiles explorer using autodidactic generative examination networks. J Cheminform 12(1):22. https://doi.org/10.1186/s13321-020-00425-8
90. Wang J, Hsieh C-Y, Wang M, Wang X, Wu Z, Jiang D, Liao B, Zhang X, Yang B, He Q, Cao D, Chen X, Hou T (2021) Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. Nat Mach Intell 3(10):914–922. https://doi.org/10.1038/s42256-021-00403-1
91. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 5(9):1572–1583. https://doi.org/10.1021/acscentsci.9b00576
92. Kwon Y, Yoo J, Choi Y-S, Son W-J, Lee D, Kang S (2019) Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. J Cheminform 11(1):70. https://doi.org/10.1186/s13321-019-0396-x
93. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2023) PubChem 2023 update. Nucleic Acids Res 51(D1):D1373–D1380. https://doi.org/10.1093/nar/gkac956
94. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940. https://doi.org/10.1093/nar/gky1075
95. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA (2020) ZINC20—A free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model 60(12):6065–6073. https://doi.org/10.1021/acs.jcim.0c00675
96. D Lowe. Chemical reactions from US patents (1976 - Sep 2016). Figshare. *Dataset;* 2017.
97. Xiong F, Xu H, Yu M, Chen X, Zhong Z, Guo Y, Chen M, Ou H, Wu J, Xie A, Xiong J, Xu L, Zhang L, Zhong Q, Huang L, Li Z, Zhang T, Jin F, He X (2022) 3CLpro inhibitors: DEL-based molecular generation. Front Pharmacol 1:13
98. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: collection of open natural products database. J Cheminform 13(1):2. https://doi.org/10.1186/s13321-020-00478-9
99. NIH LINCS. LINCS L1000. NIH July 2023.
100. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6):1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039
101. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, Fisher J, Jansen JM, Duca JS, Rush TS, Zentgraf M, Hill JE, Krutoholow E, Kohler M, Blaney J, Funatsu K, Luebkemann C, Schneider G (2020) Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 19(5):353–364. https://doi.org/10.1038/s41573-019-0050-3
102. Schoenmaker L, Béquignon OJM, Jespers W, van Westen GJP (2023) UnCorrupt SMILES: a novel approach to de novo design. J Cheminform 15(1):22. https://doi.org/10.1186/s13321-023-00696-x
103. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G (2020) Generative molecular design in low data regimes. Nat Mach Intell 2(3):171–180. https://doi.org/10.1038/s42256-020-0160-y
104. Méndez-Lucio O, Baillif B, Clevert D-A, Rouquié D, De WJ (2020) Novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nat Commun 11(1):10. https://doi.org/10.1038/s41467-019-13807-w
105. Chen Y, Wang Z, Wang L, Wang J, Li P, Cao D, Zeng X, Ye X, Sakurai T (2023) Deep generative model for drug design from protein target sequence. J Cheminform 15(1):38. https://doi.org/10.1186/s13321-023-00702-2
106. Zheng S, Lei Z, Ai H, Chen H, Deng D, Yang Y (2021) Deep scaffold hopping with multimodal transformer neural networks. J Cheminform 13(1):87. https://doi.org/10.1186/s13321-021-00565-5
107. Wang X, Gao C, Han P, Li X, Chen W, Rodríguez Patón A, Wang S, Zheng P (2023) PETrans: de novo drug design with protein-specific encoding based on transfer learning. Int J Mol Sci 24(2):1146

108. Grechishnikova D (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Sci Rep 11(1):321. https://doi.org/10.1038/s41598-020-79682-4
109. Kotsias P-C, Arús-Pous J, Chen H, Engkvist O, Tyrchan C, Bjerrum EJ (2020) Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. Nat Mach Intell 2(5):254–265. https://doi.org/10.1038/s42256-020-0174-5
110. Mao J, Wang J, Zeb A, Cho K-H, Jin H, Kim J, Lee O, Wang Y, No KT (2023) Transformer-based molecular generative model for antiviral drug design. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.3c00536
111. Chang J, Ye JC (2024) Bidirectional generation of structure and properties through a single molecular foundation model. Nat Commun 15(1):2323. https://doi.org/10.1038/s41467-024-46440-3
112. Zhumagambetov R, Molnár F, Peshkov VA, Fazli S (2021) Transmol: repurposing a language model for molecular generation. RSC Adv 11(42):25921–25932. https://doi.org/10.1039/D1RA03086H
113. Marin Zapata PA, Méndez-Lucio O, Le T, Beese CJ, Wichard J, Rouquié D, Clevert D-A (2023) Cell morphology-guided de novo hit design by conditioning GANs on phenotypic image features. Digital Discov 2(1):91–102. https://doi.org/10.1039/D2DD00081D
114. Liu Y, Yu H, Duan X, Zhang X, Cheng T, Jiang F, Tang H, Ruan Y, Zhang M, Zhang H, Zhang Q (2024) TransGEM a molecule generation model based on transformer with gene expression data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btae189
115. Fedus, W.; Goodfellow, I.; Dai, A. M. Maskgan: Better Text Generation via Filling in The_. *arXiv preprint* arXiv:1801.07736 *2018*.
116. Lee YJ, Kahng H, Kim SB (2021) Generative adversarial networks for de novo molecular design. Mol Inform 40(10):2100045. https://doi.org/10.1002/minf.202100045
117. Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; LeCun, Y. Adversarially Regularized Autoencoders. In international conference on machine learning; PMLR, 2018; 5902–5911.
118. Hong SH, Ryu S, Lim J, Kim WY (2020) Molecular generative model based on an adversarially regularized autoencoder. J Chem Inf Model 60(1):29–36. https://doi.org/10.1021/acs.jcim.9b00694
119. Wu B, Li L, Cui Y, Zheng K (2022) Cross-adversarial learning for molecular generation in drug design. Front Pharmacol 12:1
120. Abbasi M, Santos BP, Pereira TC, Sofia R, Monteiro NRC, Simões CJV, Brito RMM, Ribeiro B, Oliveira JL, Arrais JP (2022) Designing optimized drug candidates with generative adversarial network. J Cheminform 14(1):40. https://doi.org/10.1186/s13321-022-00623-6
121. Ai C, Yang H, Liu X, Dong R, Ding Y, Guo F (2024) MTMol-GPT: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. PLoS Comput Biol 20(6):e1012229
122. Skinnider MA (2024) Invalid SMILES are beneficial rather than detrimental to chemical language models. Nat Mach Intell 6(4):437–448. https://doi.org/10.1038/s42256-024-00821-x
123. Zou J, Zhao L, Shi S (2023) Generation of focused drug molecule library using recurrent neural network. J Mol Model 29(12):361. https://doi.org/10.1007/s00894-023-05772-5
124. Bian Y, Xie X-Q (2022) Artificial intelligent deep learning molecular generative modeling of scaffold-focused and cannabinoid CB2 target-specific small-molecule sublibraries. Cells 11(5):915
125. Yasonik J (2020) Multiobjective de Novo drug design with recurrent neural networks and nondominated sorting. J Cheminform 12(1):14. https://doi.org/10.1186/s13321-020-00419-6
126. Harel S, Radinsky K (2018) Prototype-based compound discovery using deep generative models. Mol Pharm 15(10):4406–4416. https://doi.org/10.1021/acs.molpharmaceut.8b00474
127. Wei L, Fu N, Song Y, Wang Q, Hu J (2023) Probabilistic generative transformer language models for generative design of molecules. J Cheminform 15(1):88. https://doi.org/10.1186/s13321-023-00759-z
128. Liao Z, Xie L, Mamitsuka H, Zhu S (2023) Sc2Mol: A scaffold-based two-step molecule generator with variational autoencoder and transformer. Bioinformatics. https://doi.org/10.1093/bioinformatics/btac814
129. Zhu H, Zhou R, Cao D, Tang J, Li M (2023) A pharmacophore-guided deep learning approach for bioactive molecular generation. Nat Commun 14(1):6234. https://doi.org/10.1038/s41467-023-41454-9
130. Atz K, Cotos L, Isert C, Håkansson M, Focht D, Hilleke M, Nippa DF, Iff M, Ledergerber J, Schiebroek CCG, Romeo V, Hiss JA, Merk D, Schneider P, Kuhn B, Grether U, Schneider G (2024) Prospective de Novo drug design with deep interactome learning. Nat Commun 15(1):3408. https://doi.org/10.1038/s41467-024-47613-w
131. Langevin M, Minoux H, Levesque M, Bianciotto M (2020) Scaffold-constrained molecular generation. J Chem Inf Model 60(12):5637–5646. https://doi.org/10.1021/acs.jcim.0c01015
132. Diao Y, Liu D, Ge H, Zhang R, Jiang K, Bao R, Zhu X, Bi H, Liao W, Chen Z, Zhang K, Wang R, Zhu L, Zhao Z, Hu Q, Li H (2023) Macrocyclization of linear molecules by deep learning to facilitate macrocyclic drug candidates discovery. Nat Commun 14(1):4552. https://doi.org/10.1038/s41467-023-40219-8
133. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD (2022) MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model 62(9):2064–2076. https://doi.org/10.1021/acs.jcim.1c00600
134. Haroon S (2023) Generative pre-trained transformer (GPT) based model with relative attention for de novo drug design. Comput Biol Chem. https://doi.org/10.1016/j.compbiolchem.2023.107911
135. Wang Y, Zhao H, Sciabola S, Wang W (2023) CMolGPT: a conditional generative pre-trained transformer for target-specific de novo molecular generation. Molecules 28(11):4430
136. Monteiro NRC, Pereira TO, Machado ACD, Oliveira JL, Abbasi M, Arrais JP (2023) FSM-DDTR: end-to-end feedback strategy for multi-objective de novo drug design using transformers. Comput Biol Med 164:107285. https://doi.org/10.1016/j.compbiomed.2023.107285
137. Fan W, He Y, Zhu F (2024) RM-GPT: enhance the comprehensive generative ability of molecular GPT model via localRNN and realformer. Artif Intell Med 150:102827. https://doi.org/10.1016/j.artmed.2024.102827
138. Kim H, Na J, Lee WB (2021) Generative chemical transformer: neural machine learning of molecular geometric structures from chemical language via attention. J Chem Inf Model 61(12):5804–5814. https://doi.org/10.1021/acs.jcim.1c01289
139. Yoshikai, Y.; Mizuno, T.; Nemoto, S.; Kusuhara, H. A Novel Molecule Generative Model of VAE Combined with Transformer. *arXiv preprint* arXiv:2402.11950 *2024*.
140. Inukai, T.; Yamato, A.; Akiyama, M.; Sakakibara, Y. A Tree-transformer based vae with fragment tokenization for large chemical models; 2024.
141. Bhadwal AS, Kumar K, Kumar N (2024) NRC-VABS: normalized reparameterized conditional variational autoencoder with applied beam search in latent space for drug molecule design. Expert Syst Appl 240:122396. https://doi.org/10.1016/j.eswa.2023.122396
142. Liu D, Song T, Na K, Wang S (2024) PED: a novel predictor-encoder-decoder model for alzheimer drug molecular generation. Front Artif Intell 7:137418
143. Özçelik R, de Ruiter S, Criscuolo E, Grisoni F (2024) Chemical language modeling with structured state space sequence models. Nat Commun 15(1):6176
144. Hu P, Zou J, Yu J, De SS (2023) Novo drug design based on stack-RNN with multi-objective reward-weighted sum and reinforcement learning. J Mol Model 29(4):121. https://doi.org/10.1007/s00894-023-05523-6
145. Tan X, Jiang X, He Y, Zhong F, Li X, Xiong Z, Li Z, Liu X, Cui C, Zhao Q, Xie Y, Yang F, Wu C, Shen J, Zheng M, Wang Z, Jiang H (2020) Automated design and optimization of multitarget schizophrenia drug candidates by deep learning. Eur J Med Chem 204:112572. https://doi.org/10.1016/j.ejmech.2020.112572
146. Shi T, Huang S, Chen L, Heng Y, Kuang Z, Xu L, Mei H (2020) A molecular generative model of ADAM10 inhibitors by using GRU-based deep neural network and transfer learning. Chemom Intelligent Lab Syst. https://doi.org/10.1016/j.chemolab.2020.104122
147. Lee J, Myeong I-S, Kim Y (2023) The Drug-like molecule pre-training strategy for drug discovery. IEEE Access 11:61680–61687. https://doi.org/10.1109/ACCESS.2023.3285811
148. Li S, Wang L, Meng J, Zhao Q, Zhang L, Liu H (2022) De novo design of potential inhibitors against SARS-CoV-2 Mpro. Comput Biol Med 147:105728. https://doi.org/10.1016/j.compbiomed.2022.105728
149. Santana MVS, De S-J (2021) Novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. BMC Chem 15(1):8. https://doi.org/10.1186/s13065-021-00737-2

150. Suresh N, Kumar NCA, Subramanian S, Srinivasa G (2022) Memory augmented recurrent neural networks for De-novo drug design. PLoS ONE 17:6
151. Thomas M, O'Boyle NM, Bender A, de Graaf C (2022) Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. J Cheminform 14(1):68. https://doi.org/10.1186/s13321-022-00646-z
152. Thomas M, Smith RT, O'Boyle NM, de Graaf C, Bender A (2021) Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J Cheminform 13(1):39. https://doi.org/10.1186/s13321-021-00516-0
153. Shen X, Zeng T, Chen N, Li J, Wu R (2024) NIMO: a natural product-inspired molecular generative model based on conditional transformer. Molecules 29(8):1867
154. Fatima N, Imran AS, Kastrati Z, Daudpota SM, Soomro A (2022) A systematic literature review on text generation using deep neural network models. IEEE Access 10:53490–53503

Publisher's Note