processes

MDPI

Review

Deep Learning Based Methods for Molecular Similarity Searching: A Systematic Review

Maged Nasser [1,*], Umi Kalsom Yusof [1,*] and Naomie Salim [2]

[1] School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia
[2] UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research, Universiti Teknologi Malaysia, Johor Bahru 81310, Johor, Malaysia
* Correspondence: maged.m.nasser@usm.my (M.N.); umiyusof@usm.my (U.K.Y.)

Abstract: In rational drug design, the concept of molecular similarity searching is frequently used to identify molecules with similar functionalities by looking up structurally related molecules in chemical databases. Different methods have been developed to measure the similarity of molecules to a target query. Although the approaches perform effectively, particularly when dealing with molecules with homogenous active structures, they fall short when dealing with compounds that have heterogeneous structural compounds. In recent times, deep learning methods have been exploited for improving the performance of molecule searching due to their feature extraction power and generalization capabilities. However, despite numerous research studies on deep-learning-based molecular similarity searches, relatively few secondary research was carried out in the area. This research aims to provide a systematic literature review (SLR) on deep-learning-based molecular similarity searches to enable researchers and practitioners to better understand the current trends and issues in the field. The study accesses 875 distinctive papers from the selected journals and conferences, which were published over the last thirteen years (2010–2023). After the full-text eligibility analysis and careful screening of the abstract, 65 studies were selected for our SLR. The review's findings showed that the multilayer perceptrons (MLPs) and autoencoders (AEs) are the most frequently used deep learning models for molecular similarity searching; next are the models based on convolutional neural networks (CNNs) techniques. The ChEMBL dataset and DrugBank standard dataset are the two datasets that are most frequently used for the evaluation of deep learning methods for molecular similarity searching based on the results. In addition, the results show that the most popular methods for optimizing the performance of molecular similarity searching are new representation approaches and reweighing features techniques, and, for evaluating the efficiency of deep-learning-based molecular similarity searching, the most widely used metrics are the area under the curve (AUC) and precision measures.

Keywords: molecular similarity searching; drug design; drug discovery; virtual screening; deep learning

Citation: Nasser, M.; Yusof, U.K.; Salim, N. Deep Learning Based Methods for Molecular Similarity Searching: A Systematic Review. *Processes* **2023**, *11*, 1340. https://doi.org/10.3390/pr11051340

Academic Editors: Adel Ali Ahmed, AbdulRahman Alsewari, Yousef Fazea and Waleed Ali

Received: 22 March 2023
Revised: 14 April 2023
Accepted: 20 April 2023
Published: 26 April 2023

1. Introduction

Over the past decade, searching databases for molecules that exhibit similarity to a given structure has become popular. The concept of similarity searching was being used in a wide range of applications and fields and comprises cheminformatics, chemistry, and pattern recognition. In recent times, molecular searching has become one of the key topics of cheminformatics study due to the increased demand for drug discovery research [1]. As discovered in the literature, the initial research on molecular searching was completed by Raymond et al. [2] in the middle of the 1980s, and they provided an algorithm for estimating the basic kind of substructure similarities known as atom pairs [2].

To perform a similarity search, three main components of similarity searching needed to be taken into consideration [1,2]. The structure representations are first considered to

describe molecules. The relative relevance representation value is then calculated using the weighting scheme. The similarity coefficient constitutes the final factor.

The similarity between the two structural representations is measured using the similarity coefficient, which is the critical part of the similarity search [3].

Considering the recent successes of deep-learning-based models in various application domains comprised of natural language processing (NLP) [4], image processing [5], and machine translations [6], deep learning models have been used for molecular similarity searching, which improves the capabilities by resolving several issues with the current models [7–9]. Particularly, the main idea of deep-learning-based approaches is based on how these models can be best utilized to select only the important features and how these important features can be used to improve the efficiency of molecular similarity searching.

Using deep learning methods for molecular similarity searching has become more prevalent as a result of more remarkable achievements in yielding high-quality performance [10]. Compared to the traditional similarity searching architectures, deep-learning-based models are recently used to identify the important molecular features in chemical datasets according to the molecular features representations [11,12]. Therefore, developing deep-learning-based models on how these important features are explored to improve the effectiveness of the similarity measure performance becomes a promising solution. This has been verified by recent research that used deep learning techniques to increase the performance and diversity of the proposed approaches [11,13,14]. For example, to improve the similarity searching performance, particularly for structurally heterogeneous molecules, a Siamese multi-layer perceptron design was introduced as a result of their capability to cope with complex data samples in various disciplines [11]. Similarly, Huber et al. [15] proposed a deep-learning-based MS2DeepScore method for structural similarity score prediction for spectrum pairs with improved accuracy. Using RNNs and GNNs, Yingkai Gao et al. [16] learned drug and protein embeddings on drug atomic graphs and protein sequences, respectively. These models all outperformed the traditional machine learning techniques and demonstrated exceptional success.

Since their recent introduction, there has been an explosion in the number and variety of deep learning models that have been applied to molecular searching. These models vary in their molecular representations, the model architectures, and the type of problems related to molecular design they address. To facilitate comparisons between the growing number of benchmarks that have been recently proposed to evaluate deep learning models, several excellent reviews have been written to summarize the development of this field. In other words, relatively limited secondary studies were explicitly available to analyze previous works and identify the challenges in the research area despite the several publications on deep learning methods for molecular similarity searching [17–19].

To be specific, ref. [20] focused on reviewing recent advances for the generation of novel molecules with desired properties considering the applications of generative adversarial networks (GANs), reinforcement learning (RL), and related techniques. The authors of [21] has also conducted a review to examine how the current deep generative models address the inverse chemical discovery paradigm. They begin introducing generative models for molecular designs and classify them based on their architecture and molecular representations. Then, the evolution and performance of the main molecular generation schemes were reviewed to help researchers extract important lessons for automatic chemical design. In addition, a general overview of the applications of deep learning in molecule generation and molecular property prediction has been detailed by [22]. The overview focused on the two key areas where deep learning has impacted molecular design, which include the prediction of molecular properties and the de novo generation of new molecules.

As can be observed, each of the above studies has addressed a specific aspect of optimizing molecules for computational metrics or quantitative estimates of drug design. Unlike the aforementioned works, this paper performs an integral approach of the complete SLR process on molecular similarity search based on deep learning models. By so doing,

it aims to fill the gap and provide both novice and new researchers with the required background knowledge within the field.

This study proposes an SLR approach for better reviewing, examining, and summarizing the trends and progress of the research works in this field as a synthesis of the best-quality works based on deep learning methods. This SLR was carried out in accordance with the guiding principle designed by [23]. It has been established that SLR, which employs the approach of a systematic collection of research articles using exclusion/inclusion methods, is the best way to conduct an objective review of recent studies. SLR is a method of evidence-based software engineering that compiles evidence to provide to practitioners and researchers [20]. The following highlights the important contributions of this SLR:

- We offer an SLR of the current deep-learning-based molecular similarity search methods. The researchers as well as the practitioners in the field will use this as guidance to better grasp the current trends and propose methods for addressing the existing challenges.
- We provide a quantitative analysis of the existing problems and as well identify a future direction for deep-learning-based methods for molecular similarity searching.
- We present descriptions of the state-of-the-art deep-learning-based molecular searching methods, including the datasets and their categories, and various performance measures employed to assess the performances of the deep learning methods for molecular similarity searching.

The remainder of this paper is structured as follows: the theoretical backgrounds of molecular similarity searches are presented in Section 2. The method adopted for extracting the relevant articles for our SLR study is provided in Section 3. The review's findings, based on the selected papers, are presented in Section 4. Finally, the study's limitations for future research areas and its findings are presented in Sections 5 and 6, respectively.

2. Theoretical Background

This section provides a theoretical overview of molecular similarity searches to enable researchers to better understand the concepts that use deep learning models to improve the performance of molecular similarity searching. It begins with brief overviews of the various methods for representing chemical structures in computer programs and databases. Then, different searching mechanisms used by chemical information systems are described as well as the calculation of the molecular descriptors that form the basis for the construction of various virtual screening methods.

2.1. Chemical Structure Representations

The systematic nomenclature is the main representation of the chemical compounds for a manual information retrieval system that uses a set of rules to generate systematic names for each compound. Due to the fact that there is not much flexibility, nomenclatures are often required to be represented using other methods for computerized systems. Three forms of chemical notation have been developed, which are connection tables, linear notations, and adjacency matrix (AM). The adjacency matrix is used mainly as a representation in the structure processing, the connection tables, and the linear notation is used for communication on computers and between humans and computers [24,25].

2.1.1. Connection Tables

The chemical structure is characterized in connection tables as a list of all non-hydrogen atoms in the molecule, along with details on how they are related to one another. Basically, a connection table is comprised of two parts: a list of the molecule atoms' atomic numbers and a list of the bonds represented as pairs based on the connected atoms [24–26]. Hydrogen atoms are not usually represented in the connection tables because they can be deduced from the bond orders and atom types, and more detailed tables consist of the bonding angles information for plotting the bonds. The 2D and 3D coordinates of the atoms could be incorporated as well to provide standard chemical drawings for molecules. The

3D chemical structure representation can be represented using the connection table by reporting the distance between all atoms and atoms' coordinates. However, a complete 3D connection table includes 3D features, and possible conformations of a molecule are most likely to consume a high amount of storage, so, usually, only the relative atomic coordinates are stored.

Figure 1 shows a simple connection table for Aspirin structure [26]. It consists of 13 atoms that are connected to each other by 13 bonds. The first and second atoms in the table are carbon (C), and they are connected using a bond of order 2. The coordinates of the first and second atoms are (−3.4639, −1.5375) and (−3.4651, −2.3648), respectively.

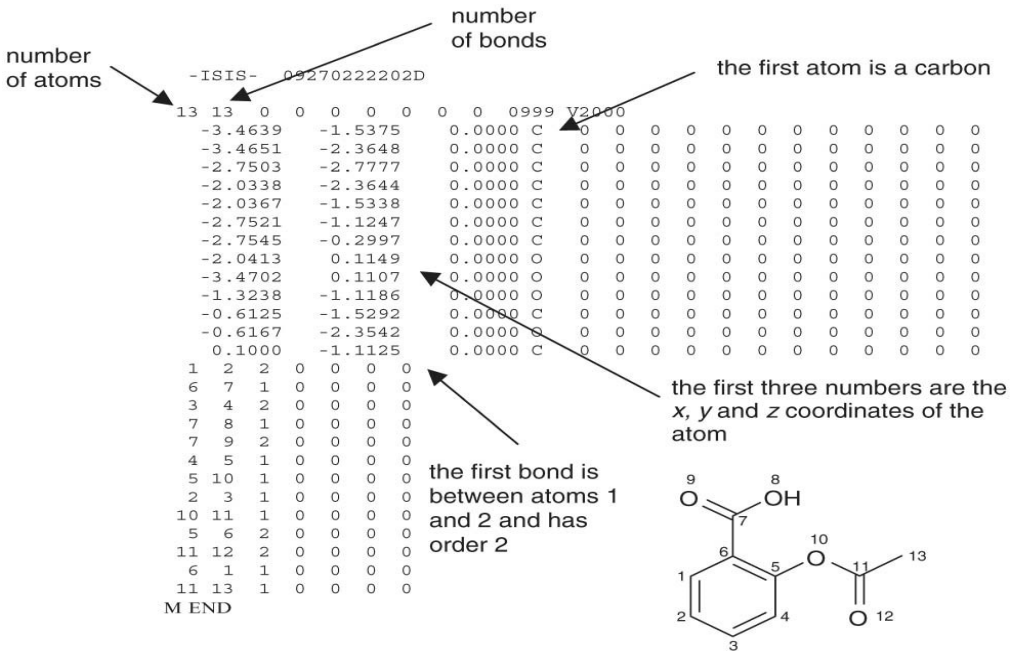


Figure 1. The Aspirin connection table.

2.1.2. Line Notations

Linear notation is one of the widely used chemical structure representations. It uses alphanumeric characters to encode the molecules in a more compact way than connection tables. Because of that, the use of linear notations is effective for storing vast numbers of molecules [24,26]. Therefore, the simplified molecular input line entry specification (SMILES) notation is the most acceptable and common linear notation because it has fewer rules and is easier to use than other notations.

The example of the SMILES strings for Aspirin is shown in Figure 2. However, there are many different ways to write SMILES strings (and to construct the connection table) for a certain molecule. This also implies that there are several approaches to number the atoms using the connection tables, and different sequences of SMILES notation can be obtained by starting on a different atom. For example, if we have N atoms, the different ways to number the connection tables are N!, which is considered a computationally unfeasible process. Therefore, the canonical representation of the chemical structures is employed to address the issue, with the canonical representation being defined as a unique ordering of molecule atoms. The Morgan algorithm is the most extensively used to determine the canonical order of molecule atoms; it is applied in conjunction with SMILES to offer a distinctive representation for each chemical structure.

Similar to the SMILES, the chemoinformatics research group at Universiti Teknologi Malaysia has developed a new method known as language for writing descriptors of outline shape of molecules (LWDOSM) [27], which is used to obtain a textual representation of 2D molecular structure based on its outline shape.

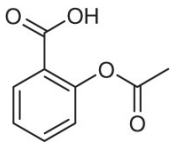
The chemical structure	
SMILES	OC(=O)c1ccccc1OC(=O)C
LWDOSM	"C–C–O–C–C–C–O–C–O–C–C–C–/C–C–C–C–O–C–O–C–C–C–C–/C–C"

Figure 2. The use of SMILES and LWDOSM for Aspirin.

2.1.3. Adjacency Matrix

The adjacency matrix (AM) is among the first methods introduced for molecular representation. The AM is a square matrix with dimensions equal to the number of atoms in the specific molecule [24,25]. Figure 3 shows the AM representation of Aspirin. The dimensions of the AM are presented as (13 × 13), which means that the Aspirin molecule contains 13 atoms and each value in each AM *Rowi* represents the level of bond connection between *Atomi* with all molecule atoms.

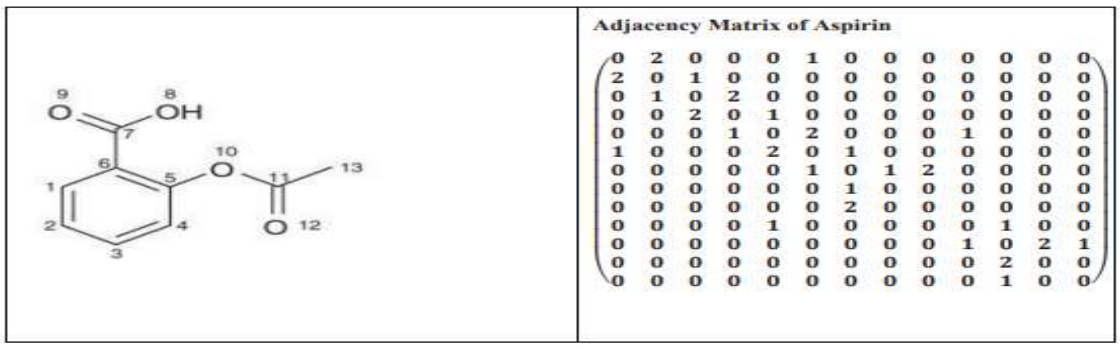


Figure 3. The representation of Aspirin using adjacency matrix.

The diagonal elements of an AM are always zero, indicating no connection because the molecular graph lacks recursive edges and self-connected nodes. The single, double, and triple bonds are represented by the numbers 1, 2, and 3 in the AM, respectively. Moreover, the extended bond types are often depicted in different systems with the extension to 4 and 5, for amide and aromatic bonds, respectively [27,28].

2.2. Molecular Descriptors

Molecular descriptors are numbers that describe the properties of molecules and are utilized to manipulate chemical structural information. For example, molecular descriptors can be used to represent the physicochemical properties of chemical structures, or they can be used to generate chemical structures using algorithmic techniques. There are numerous molecular descriptors used for various reasons in the literature, all of which differ in terms of the complexity of encoded information and the time needed to generate them. Furthermore, some descriptors include an experimental component, while others are generated solely by computational algorithms (e.g., 2D fingerprints) [26].

In chemoinformatics, researchers have developed several descriptors that can be used for similarity calculations and various purposes. Thus, an appropriate description is considered an essential part of the molecular similarity measure [27]. The descriptors consist of numerical values, which are presented in vectors of numbers that hold the

characters and properties of compounds. The descriptors are generated by encoding the molecule structure, which can be easily read by computers, similar to the previously mentioned representation, such as a 2D connection table. Moreover, similarity calculations are the most significant in molecular descriptors. The main three most prevalent types of descriptors used in chemoinformatics are 1D, 2D, and 3D descriptors. These categories of descriptors are further discussed in the following subsection.

2.2.1. 1D Descriptors

First, 1D descriptors are known as whole-molecule descriptors that characterize the chemical structure using simple properties, such as molecular weight, volume, hydrophobicity, or others. However, these types of descriptors are considered insufficient to determine the similarity between molecules. Normally, several types of descriptors can be used together for molecular searching in chemical datasets [29].

2.2.2. 2D Descriptors

Further, 2D descriptors are extensively explored in chemoinformatics and are generated from 2D molecular representations. There are various types of 2D descriptors, including simple count descriptors, physicochemical properties descriptors, topological indices, 2D fingerprints, and extended connectivity fingerprints.

Simple Counts Descriptors: The simplest 2D descriptors are dependent on simple counts of features, such as hydrogen bond acceptors, hydrogen donors, rotatable bonds, molecular weight, and ring systems (such as aromatic rings). Many of these structural properties can be classified as substructures, and their number of occurrences in a chemical compound can be easily estimated using substructure searching techniques based on the 2D connection table [30].

Physicochemical Properties Descriptors: The descriptors use the physicochemical properties of molecular compounds, which can be generated based on the hydrophobicity of molecules. One of the important properties of molecular structures that are used to determine the activity and transport of drugs is hydrophobicity, which is frequently approximated by the logarithms (log P) of the partitioning coefficients between n-octanol and water. The determination of log P values can be predicted either by experiments, which can often be difficult, or by developing various methods to predict the values of hydrophobicity, such as atom-based methods or property-based methods [31].

Topological Indices: Topological indices are single-valued (integer or real) descriptors that can be derived from a 2D graph representation of molecules [32]. Structures are described using this type of representation based on their sizes, degree of branching, and overall shape. In a molecule's structural diagram, each atom is represented by a vertex (node), and each bond is represented by an edge in the graph. Ref. [33] developed one of the most prominent topological indices, known as the molecular connectivity indices. Generally, there exist three generations of topological indices, which are the first generation of indices, which are integer numbers calculated from integer graph properties, the second generation of indices are real numbers calculated from integer graph properties, and the third generation includes real numbers calculated from real-valued graph properties. Subsequent work of [33] includes an extended version of topological descriptors that included electronic and valence state information, dubbed as the chi molecular connectivity indices. The topological state indices encode information (using numerical values) regarding the topological environment of an atom based on the encoding information of the atom in all paths emanating from that atom. The most prevalent forms of topological metrics are molecular connectivity indices.

2D Fingerprints: 2D fingerprints are frequently used in searching methods, such as substructure and similarity searching, to offer a fast-screening step. The dictionary-based and hashed-based are the main two types of 2D fingerprints. The fingerprints are generated by converting a chemical structure based on the kind of connection table into a string of zeros or ones, which identifies the presence or absence of molecule features [34].

Each bit position belongs to a particular substructural fragment or functional group. The fingerprint length is restricted to the count of fragments in the dictionary, in which each bit position in the binary string typically corresponds to a single distinct sub-structural fragment in a dictionary, so the bits either individually or collectively signify the existence or absence of fragments. For the 2D fingerprints, the dictionary contains hundreds to thousands of structural fragments, while it contains millions of structural fragments for 3D pharmacophore fingerprints. Unfortunately, the optimum fragments dictionary is often dataset-dependent [35]. The second type of 2D fingerprint is the hashed fingerprint, which is independent of the pre-defined fragment dictionary and applies to any type of chemical structure. The properties and biological activities of molecules often depend on features that are encoded by 2D fingerprints, which are important for searching algorithms. The hashed fingerprint of a molecule is a bit string (binary form) that contains information on the chemical structure. In this type of fingerprint, all the unique fragments present in a molecule are hashed based on a hashing function to conform to the length of the bit string [36].

2.2.3. 3D Descriptors

Further, 3D descriptors are generated based on geometrical representation (i.e., modeling the 3D environment of molecules). Since the geometrical representation requires knowledge of the relative positions of the atoms in 3D conformation, this type of descriptor provides more chemical information and discrimination power than topological descriptors for identical molecular structures and molecule conformations. Regardless of higher chemical information obtained by the 3D descriptors, there are some weaknesses; for example, calculating 3D descriptors is more computationally expensive compared to the processing of 2D descriptors, and most 3D descriptors (e.g., grid-based) require alignment rules to attain molecule comparability, and the complexity (for deriving 3D descriptors) can be increased significantly, particularly when several molecules' conformations instead of a single molecule conformation are considered. Hence, it is recommended to use 2D descriptors and other simple descriptors for large database screening [37–39].

However, the 3D descriptors can exploit the large information content to effectively find correlations between molecular structures and complex properties, such as biological activity [39]. Several examples of 3D descriptors include the 3D fragment screens, affinity fingerprints, potential-pharmacophore-point descriptors, the application of 3D atom environment for atom mapping similarity searching, and 3D molecular fields for field-based similarity searching.

2.3. The Advent of Deep Learning Architecture in Molecular Searching

Advances in deep learning, especially in computer vision and natural language processing, triggered the recent concern in drug discovery applications, particularly researchers in molecular similarity searching. Merck is credited with popularizing deep learning for computer-aided drug discovery in the Kaggle competition on Molecular Activity Challenge in 2012 [40]. Deep neural networks were trained using a multitask learning strategy in Dahl's winning solution [41]. As a result, several research studies employed these approaches for related drug discovery issues. These include the introduction of a Siamese multi-layer perceptron architecture [11] to improve the performance of similarity search, particularly for the structurally heterogeneous molecules, the prediction of structural similarity scores for spectrum pairs with a better accuracy using deep-learning-based MS2DeepScore method [15], the learning of drug and protein embeddings using GNNs and RNNs on drug atomic graphs and protein sequences [16], the analysis of therapeutic drug pharmacokinetic behavior predictions and their side effects [42], predicting small molecule–protein bindings [43], determining the chemotherapeutic responses of carcinogenic cells [44], the quantitative estimation of drug sensitivity [45], and modeling of the quantitative structure–activity relationship (QSAR) [46], among others. Significant drug-discovery-enabled discoveries of clinical drug candidates have been made possible as

a result of GPU-enabled deep learning architectures and the explosion of chemical genomics data. Moreover, companies focused on artificial intelligence (such as Benevolent AI, Insilco Medical, and Exscientia, among others) are recording achievements in advanced drug discoveries. These recent achievements suggest that further development and implementation of AI-driven methods enabled by GPU computing could significantly speed up the identification of new and enhanced treatments [47,48].

3. Methodology

This section outlines the process of the proposed review approach, which is strictly based on the guiding principle provided by [23]. It entails the clearly defined stages to evaluate and analyze the selected articles to find any gaps in the body of knowledge and to review their contributions to the research questions (RQs) in order to reach a conclusion. Three steps make up the SLR process: review planning, review execution, and review documentation. Figure 4 depicts the various elements of each step as well as the results of each phase. The next subsection describes each stage.

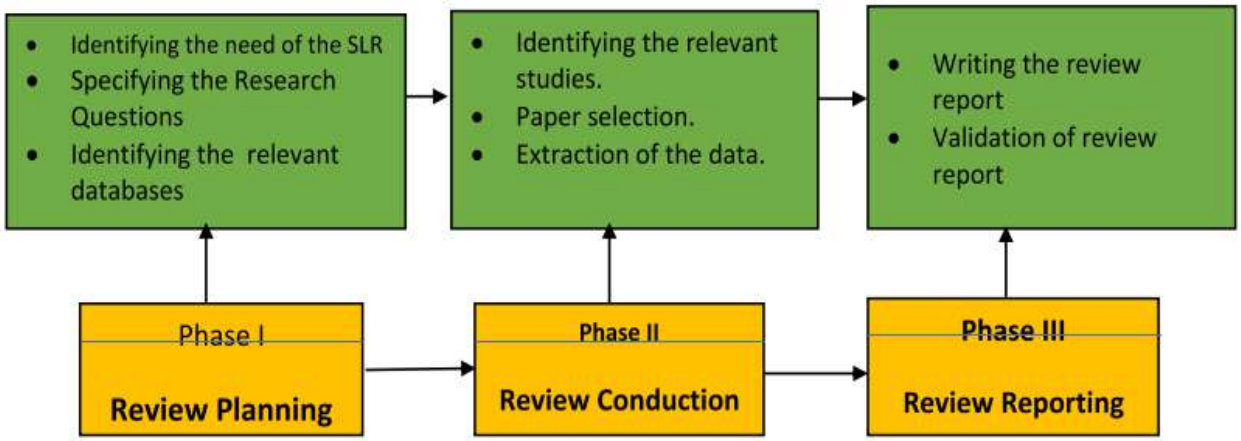


Figure 4. Three phases of the systematic literature review.

3.1. Review Planning

This phase involves preparing the review study, which comprises specifying the importance of the study, formulating the research questions, and choosing the appropriate online bibliographic databases.

3.1.1. The Significance of the Study

As previously noted, very few secondary studies on deep-learning-based molecular searching were undertaken in the area, and the current research was only focused on the traditional ML methods and the survey on state-of-the-methods. Therefore, it is necessary to explore the systematic review approach, which is considered the best way to provide an inclusive and objective analysis of research articles. We introduce this SLR to address the research questions (RQs) designed for this study. These are as follows:

- *RQ1:* What are the most common deep-learning-based approaches applied for molecular similarity searching?
- *RQ2:* What are the most effective ways to improve the performance of molecular similarity searching using deep learning methods?
- *RQ3:* What are the commonly used performance evaluation metrics for deep-learning-based molecular similarity searching?
- RQ4: What are the common datasets used for evaluating the deep-learning-based molecular similarity searching methods?
- *RQ5:* What are the research gaps, challenges, and future directions of deep-learning-based methods for molecular similarity searching?

3.1.2. Selection of the Online Databases

In this study, the major digital libraries comprising the IEEE explores, Web of Science, ScienceDirect, ACM (Association of Computing Machinery) digital library, and Springer were automatically searched in order to obtain the relevant publications for this review. Other comparable sources were not taken into consideration because they primarily indexed data from primary sources. The selected databases were considered on the basis that they are widely used and for the abundance of published papers they provide that are more relevant to our RQs. This investigation relates to works released between 2010 and 2023.

Search process: The search terms were carefully chosen given the designed RQs of this SLR. While searching for the relevant articles, various search strings containing various phrase combinations were used. The research employed the following keywords and synonyms: *MolecularSimilaritySearching, MolecularSearching, MolecularSimilarity, Deeplearning, and NeuralNetwork*. The search terms are to be used on the aforementioned digital libraries once the keywords and synonyms have been identified.

3.2. Conducting the Review

The selection of the primary studies is completed in the second phase of the SLR by using query strings to conduct searching taking into account the inclusion/exclusion criteria shown in Table 1. The selected articles will then be verified using the quality evaluation criteria.

Table 1. Inclusion and exclusion criteria.

Inclusion Approach	Exclusion Approach
The published and peer-reviewed articles that are written in English language only	Duplicate reports of the same studies are excluded
Studies that are directly related to the deep-learning-based molecular searching	Non-related books, theses, notes, tutorials, and studies are excluded from the review
Only publications from conferences as well as journals are considered.	This review excludes the articles that do not adequately describe an experimental study.
Articles published from 2010 to 2023 are only considered.	This review excludes the articles that authors are unable to access.

3.2.1. Paper Selection

Once the appropriate online databases were selected and the search terms were defined, the specified terms used in the search engines of the databases were considered to retrieve a number of 875 articles, as shown in Figure 5. As a result of the various techniques each database employed in its search engine, all the databases return a varied amount of publications.

By carefully analyzing the titles and examining the abstracts of the studies, we first eliminate the ones that are not relevant to our research focus. We then proceed to the conclusion section when the abstract did not include the relevant information, and 131 papers were obtained as a result. After applying the inclusion/exclusion criteria to the remaining publications, a list of 65 papers was retained. Figure 5 summarizes and illustrates all the steps.

3.2.2. Quality Assessment

We use the standard quality checklist questions listed in Table 2 by Kitchenham et al. [23] to verify the accepted publication's quality. To achieve this, we considered only the research that answered "yes" to at least seven questions [49]. To ensure that the findings significantly contribute to the review, the quality assessment will be taken into account together with the data extraction [23].

Table 2. Quality checklist.

No.	Quality Questions
1	Are the objectives of the research clearly stated?
2	Is reporting logical and precise?
3	Has the diversity context been studied?
4	Does the evidence relate to the interpretation and the conclusion?
5	Are the study's conclusions reliable?
6	Are they important if credible?
7	Is the research methodology properly described?
8	Could the investigation be repeated?
9	Are the details of the data collection processes well documented?
10	Is the report comprehensible and clearly written?

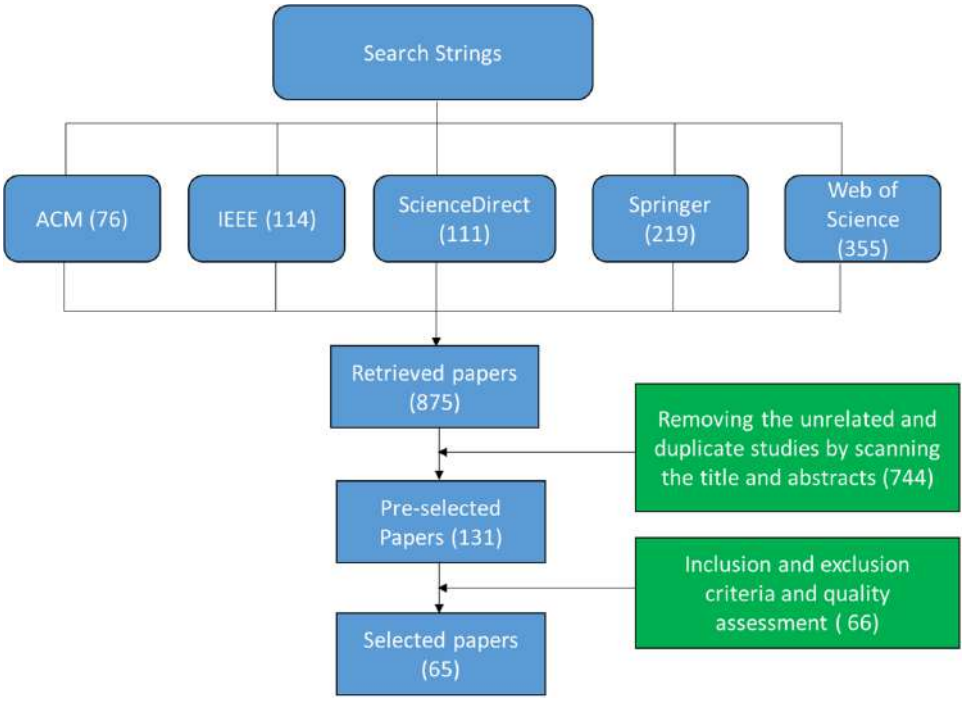


Figure 5. The paper selection process.

4. Results

The results of this study aimed to address the specified research questions as detailed in Section 3. This section is subdivided into five different subsections, which include the analysis of the searched articles for the purpose of this study as shown in the first subsection; the discussion of the various molecular similarity search deep-learning-based methods is in the second subsection; the third subsection lists various datasets and the areas where molecular similarity searching is completed using deep learning models; the fourth report is the various evaluation metrics used to measure the performance of the proposed methods; and the fifth subsection reports the challenges and potential future directions for the deep-learning-based molecular similarity searching models.

4.1. Selected Studies

The distribution of the various publications that were selected for this research is presented in this section based on the selection criteria outlined in Section 3. The 65 studies

were chosen to be used for further processing after the quality assessment technique was carefully followed. This study considered papers published written in English between the years 2010 and 2023 in journals and conference proceedings. This time frame was considered to accumulate many studies for this SLR. Figure 6 illustrates the distribution of the studies based on their publication years.

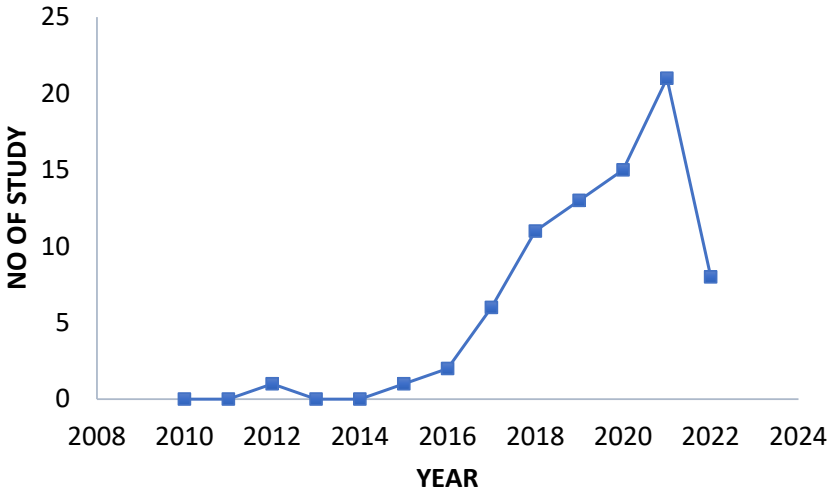


Figure 6. The selected studies according to publication years.

As shown in Figure 6, there has been a constant rise in the number of papers that have employed deep learning for molecular similarity searching since the first notable work in the field was published. Moreover, the majority of the articles on molecular similarity searches using deep learning models that were discovered in this study were mostly published within the last six years, as can be observed in the figure.

4.2. QR1: What Are the Most Common Deep-Learning-Based Approaches Applied for Molecular Similarity Searching?

This section discusses the first question, RQ1, which attempts to identify the studies based on deep learning methods for molecular similarity searching for this study. Table 3 presents the distribution of the deep learning techniques that were identified from the selected studies. Moreover, we provide brief descriptions of the key benefits and the number of research works found for each of the deep learning approaches in the table.

Autoencoder methods: AEs are nonlinear, unsupervised models for dimensionality reduction [50]. In general, a molecular structure candidate can be found using the AE model based on a desired physical property value, or a relationship between a molecular structure and a physical property can be embedded into a low-dimensional vector space (chemical space). By effectively modeling the chemical space's structure with more accuracy and lower dimensions than the original input space, the structure search is simplified. One of the basic methods that used AE for molecular searching was provided by [51]. The authors describe a novel approach for learning molecular embedding that integrates variational autoencoders (VAEs) and metric learning with some physical features. By enabling the local and continuous integration of molecular structures and physical features into the latent space of VAEs, this approach preserves the consistency of the relationship between the physical properties and the structural features of molecules to generate better predictions. The SMILES generated a smooth chemical latent space that enables continuous searching from one compound to another, as demonstrated by [52].

Adversarial AE (AAE): AAE is a deep learning method that enhances the VAE with the GAN structure to achieve feature generation and compression [17]. The VAE performs well by compressing the characteristics of compounds, but it shows inadequate performance to produce reliable results. On the other hand, the GAN generates binding compounds and produced reliable results, but it has a low diversity score and can be biased toward a single

mode. Insilico Medicine initially released the AAE [20,53] for the discovery and creation of novel compounds in 2016 [53], which was later extended by an improved model called druGAN model [54]. The AAE is a technique that can produce new compounds effectively and compress the data into the latent space. By incorporating a function to change the condition into AAE, Polykovskiy et al. created novel compounds by altering the input compound's synthetic accessibility (logP) and lipophilicity (logP) [55].

Recurrent Neural Networks: In contrast to MLPs and CNNs, recurrent neural networks (RNNs) [56] have the capacity to reuse internal information, which can be viewed as loops in the network. Recurrent neural units have the ability to produce molecules from known active ones [57]. The term "recurrent" refers to an input that uses the input from a previous time by performing a series of computations on it. RNNs were successfully used in the QSAR environment, and the SMILES encoding of molecules can be understood as sequential data for molecular prediction [58]. Similarly, [59] trained a recurrent neural network to create molecules with no restrictions on their properties. Their intention was to enrich current molecular pools for virtual screening using newly created compounds. Moreover, [60] provided a neural-network-embedding-based approximate nearest neighbor search model to find the compounds as well as solid-state catalyst systems in huge chemical datasets. The model embedding and approximate nearest neighbor search were both demonstrated. The first attempts to accurately reflect local atomic configurations, while the second offers a simple and cost-effective technique to search for near-real vectors in a large database. Given that ANN search is implemented together with the neural network model, GemNet and FAISS were employed to achieve these.

Long short-term memory (LSTM): The LSTM model was first used in the 1990s and became popular in the late 2000s [61]. The fast-vanishing issue with naive RNNs was addressed by the introduction of LSTM. When compared to the RNN, LSTM still performs better with longer sequence data. The LSTM has undergone numerous revisions since it was first introduced [62]. Moreover, gated recurrent units (GRU) with a more simple internal structure have been increasingly common in recent years [63]. In drug development, the LSTM and GRU have outperformed the RNN and are frequently employed in place of it [64,65]. The vanishing issue still occurs, making it difficult to use very long sequential data.

Multi-Layer Perceptron (MLP): The MLP, as a distinct variant of a feed-forward neural network, is considered to be the most basic deep learning model [11]. MLP can be used to convert similarity-based virtual screening's linear approach into nonlinear models for neural performance. As a result, MLPs have been incorporated into numerous molecular similarity searching methods currently in use. In this review, various research that uses MLP for molecular similarity searching has been discovered, as shown in Table 3. One of the basic MLP ways to exploit the MLP technique was proposed by Altalib and Salim [11]. The MLPs were among the first artificial neural networks to be employed in a drug discovery application successfully due to their high learning capacity and very few parameters [66]. Hence, current GPU processors make the MLP a cheaper model that is appropriate for the large cheminformatics datasets that are growing rapidly in computer-aided drug discovery [66]. Due to their broad adaptability, compound structures can be used in conjunction with a variety of data comprising fingerprints, transcriptomes, and molecular characteristics [67,68]. For instance, Chen et al. [31] integrated the target protein information with the PseAAC, PsePSSM, NMBroto, and structural features of the MLP with four hidden layers for drug–target interaction predictions, while the fingerprints were employed for the compound [69].

Convolutional Neural Networks (CNNs): The CNNs are a unique class of networks where convolutions are used in the computations of the hidden layers [70]. Their primary use is in extracting features from images, including edge detection, which makes this type of application very frequent [71–75]. Depending on the input data, 1D, 2D, and 3D convolutions could be considered. The data may form a time series in the one-dimensional case. Images and other planar grid-like structures are employed in 2D convolutions.

Three-dimensional tensors, such as 3D pictures, can be convoluted in three dimensions. Despite remarkable performance, convolutional neural networks have the drawback of rapidly increasing the parameter sizes as the network grows deeper, specifically in the case of the 3D, which makes the training process slow [19]. The 3D model of the protein–ligand binding site has also been used as an input for successful predictions in the field of binding affinity [76]. A deep convolutional neural network technique is presented by Berrhail et al. [77] to increase the efficiency of the ligand-based virtual screening method (DCNNLB). Their study featured two main contributions; the first was the design of a deep convolutional neural network (DCNN)-based model for LBVS. The best performance in terms of accuracy and recall was determined by comparing several topological network models. The second contribution was the creation of new learning representations for a more accurate representation of chemical compounds. This representation is based on the automatic feature learning that was extracted from the proposed model's weights. As a result, it is particularly effective in determining molecular similarity and LBVS process performances. Based on this strategy, the authors showed how the DCNN method can enhance model performance.

Graph Convolutional Neural Networks (GCNs): The basic concept of GCN is that a chemical characteristic can be recognized independently of its location by using trainable filters (a set of weights) in various layers of a GCN and modifying overlapping partial representations of an input graph. As recently demonstrated on a virtual exercise for selective CDK1 inhibitors, the neural embedding approach, which previously relied on low dimensional data vectors to effectively represent data in neural networks, has been extended to include various chemical fingerprints [78]. A typical strategy for overcoming the short-range nature of fingerprint representation is to combine several fingerprints and additional molecular descriptors. This was recently shown by Zhao et al. [79] when they searched for SARS-CoV-2 3CLpro inhibitors and discovered four naturally occurring compounds with antiviral properties [79]. According to Duvenaud et al. [80], adjacency in the vector representation indicates fragment similarity relevant to an interest assay endpoint, and the graph convolutional DNN (GCNN) technique may be utilized to dynamically train a fingerprint built for the most relevant chemical information [81,82].

Graph Attention Neural Networks (GANNs): The enhanced attention mechanism in GANNs makes them a special type of graph neural network [83]. In a graph context, this can be considered as ranking and assigning varying degrees of relevance to each node in a given vertex's neighborhood. For a certain task, some atoms and corresponding interactions may be more important. This is represented by including atom distances in the adjacency matrix, as in [84]. Subsequently, a feature node is generated by linearly combining its neighbors while considering the attention coefficient.

Generative Adversarial Networks (GANs): GANs have gained recognition recently as strong and versatile deep generative models. An adversarial game between a discriminator module and a generator forms the basis of GANs. Identifying genuine from fake data points produced by the generator network is the objective of the discriminator network. Using novel data points, a concurrently trained generator network tries to trick the discriminator into believing the generated results are genuine. Many enhancements and adjustments were proposed following the experiential success of GANs [85]. Researchers in the drug discovery field promptly applied these techniques to artificially synthesize data across different subproblems [86]. To generate and bias the generation toward preferred metrics, a method combining generative adversarial networks (GANs) and reinforcement learning was presented by Guimaraes et al. [87]. The GAN component of the reward function ensures that the model preserves the information gained from the data, whereas reinforcement learning biases data generation toward arbitrary metrics. This model has been evaluated in a variety of contexts, including the context of generating molecules encoded as text sequences (SMILES) as well as the context of generating the music, demonstrating in each case how to efficiently bias the generating processes toward the desired measures. The

model was developed based on the previous finding that produced sequence data using GANs and reinforcement learning.

A GAN-based generative modeling strategy was investigated by Méndez-Lucioet al. [88] at the nexus of molecular drug design and systems biology. Their effort to integrate chemistry and biology was verified by the creation of active-like molecules in response to the unique target of the gene expression profile. To achieve this, the conditional GANs with Wasserstein GAN were combined using a gradient penalty. Genetic algorithms and GANs have also been investigated to prevent mode collapses and subsequently explored a broader chemical space incrementally [89].

Table 3. Distribution of state-of-the-art deep-learning-based models.

Model	Description	Advantage	No. of Studies	References
AEs	The model can be used to identify molecules having similar properties either by looking for a prospective molecular structure in that space based on a desired physical property value or by embedding a relationship between a molecular structure and a physical property into a low-dimensional vector space (chemical space).	Ideal for feature dimensionality reduction and extracting hierarchical features.	10	[17,50–55,90,91]
RNNs	RNNs can be regarded as an extension of Markov chains with memory, which allows them to simulate autoregression in molecular sequences by learning long-range dependencies through their internal states.	Identify the time dependencies and textual sequence information.	05	[57–59,78,92]
MLPs	The basis of NNs is composed of these fully connected networks, which have input, hidden, and output layer(s) and nonlinear activation functions (such as sigmoid, ReLU, tanh, rectified linear unit, among others).	Transforms the neural performance models from linear to nonlinear	12	[11,34,66–69,93–97]
CNNs	Possibly the most popular NNs, CNNs process local subsections of the input using small receptive fields and hierarchical principles.	Allows feature extraction with contextual information	20	[16,19,65,70–77,98–106]
GCNs	GCNS consider graphs as relational structures	Captures the graphical structure of molecules, making them potentially of great use in several drug discovery applications.	12	[78–84,107–111]
GANs	Generative semi-supervised deep learning method	Enables information retrieval that is both generative and discriminative.	05	[85–89]

4.3. QR2: What Are the Most Effective Ways to Improve the Performance of Molecular Similarity Searching Using Deep Learning Methods?

Recently, several approaches have been presented to increase the performance of similarity searching; each method has a unique set of techniques and tools. We are able

to differentiate between certain works that use coefficients-based similarity and others that employ other techniques, such as new representation, scheme weighting, transform learning, standardization, and data fusion.

4.3.1. Representation Methods

Making representations that are syntactically and semantically important to the dataset and issue at hand is made possible by the ability to learn the latent representation of input molecules by eliminating the necessity of manually created descriptors [66]. The latent representation typically consists of a set of binary elements that indicate whether certain molecular characteristics are present or absent. Deep learning models have been utilized to generate numerical representations of molecules' structures in order to evaluate the differences between two molecules, as was discovered in recent studies [78,80,112]. Here, several deep-learning-based methods were recognized for creating numerical representations of molecules' structure from the selected studies considered for this review. In light of this, we summarize the main advantages of each deep learning technique used in the selected studies as follows:

Autoencoder (AE): AE is a model that reduces the input data into a representation of lower dimension as a code and utilizes a decoder module to rebuild this compact representation in a way that strongly resembles the original input [113]. To make use of its strong capability to learn a feature representation of molecules from low-level encodings of a vast corpus of chemical structures, Nasser et al. [7] presented a deep autoencoder. It uses the principles of neural machine translation to translate between two representations of chemical structures that are semantically related but syntactically distinct by condensing the relevant data pooled by both representations into a lower dimension of the representation vector. After the model has been trained, it is possible to obtain this chemical representation and utilize it as a new descriptor for molecular similarity searching.

SMILES2vec: Smiles2vec is an RNN that uses SMILES strings to automatically learn features to predict a variety of chemical attributes, such as toxicity, activity, solubility, and solvation energy [114,115]. To learn continuous embeddings from SMILES representations and make accurate predictions for a variety of datasets and tasks, SMILES2vec was explored by several authors [114,116–118]. The learned representations attained high performance and seemed to be more suitable to regression tasks than Morgan fingerprints by employing unsupervised pre-training of word2vec on the ChEMBL dataset. A Smiles2Vec deep learning model was proposed by Goh et al. [116] to learn molecular properties from the molecular structures of organic material. Furthermore, Phillips et al. [119] deployed a SMILES2vec model as a mixed CNN–GRU model that predicts chemical solubility from chemical compounds encoded as strings using SMILES.

GCNN: A GCNN is enhanced with attention and gate mechanisms [66,120] to systematically extract features from molecular graphs that are relevant to the target chemical quality, such as polarity, solubility, photovoltaic performance, and synthetic accessibility. The interaction between each atom and its neighbors is considered by the attention mechanism in order to distinguish between atoms in various chemical environments. For instance, the augmented GCNN can distinguish between polar and nonpolar functional groups, which are the essential structural elements for molecule polarity and solubility. Therefore, the model is able to precisely predict chemical attributes and cluster molecules with related properties together in a trained latent space.

Transformer networks: The remarkable success of transformer nets in language processing inspired several researchers on deep-learning-based drug discovery to explore the potential of the technology for learning long-term dependencies for sequences [121]. End-to-end neural regressions were performed by Shin et al. [122] to predict the scores of affinity between drug compounds and their target proteins. They achieved this by combining molecular token embedding with position embedding to learn molecular representations for the drug compounds, and they also used a CNN to learn new molecular representations for proteins. To predict drug–target interactions, Huang et al. [123] introduced MolTrans.

Target-specific molecular production was described by Grechishnikova [124] as a problem of transforming the SMILES representations of amino acid chains by utilizing a transformer encoder and decoder [124].

Mol2vec: Mol2vec is a "distributed representation" that calculates lower-dimensional latent vectors using environment data from molecular graph fragments [125]. Mol2vec was developed by Jaeger et al. [126] in response to the success of the commonly used word-embedding technique word2vec. The word2vec method addresses molecular substructures as "words" in the context of neighboring fragments based on the assumption that each word has a variety of meanings depending on the context. As similar fragments are encoded into similar latent vectors, latent vectors serve as distance-preserving reconstructions of the original molecules. The closeness of the corresponding latent vectors also increases with the similarity of the fragments. Studies relating to deep-learning-based molecular searches have demonstrated the efficiency of mol2vec molecular descriptors [125].

4.3.2. Weighting Scheme

Another component of importance is the weighting scheme, which is focused on assigning different degrees of weight to the different components of these representations. The effect of different weighting schemes on the utility of molecular similarity measures has been the subject of interesting studies [7,127]. In addition to the further extended literature related to the structure descriptors [81,128] and the similarity coefficients [1,3,24], several research articles have employed weighting schemes to improve the recall and accuracy performances [7,8,77,127,129,130]. It is based on this concept that non-related molecular fragments weigh equally to the relevant fragments in terms of biological activity. For instance, Ahmed et al. [129] introduced new approaches for similarity searching utilizing Bayesian inference networks (BIN), and their results outperformed all traditional methods. Additionally, they applied fragment reweighting approaches to the selection of attributes in order to enhance the Bayesian network.

It is believed that more studies may achieve better results by using different weighting functions. To this end, Abdo and Pupin [127] propose a LINGO approach to compare the LINGOs that are present in each molecule to determine how similar two molecules are to one another. By employing varying weights based on the length of the LINGOs, the similarity is determined using this approach, with longer LINGOs receiving greater weight and shorter LINGOs receiving less weight.

4.3.3. Data Fusion

By making minor adjustments to the existing data or altering the expression rule—a process known as data fusion—the researchers have begun investigating the benefits of combining the data. A common approach to image data fusion, known as geometric transformation, can be employed for data, such as the voxel [131]. A different approach introduces a small bit of background noise without affecting the performance of the data. The incorporation of Gaussian noise into the bioactivities and compound descriptors by Cortes-Ciriano and Bender [132] enhanced the model's predictive performance. Randomized SMILES is another common data augmentation technique in the field of drug development [133,134].

Depending on the starting point and direction, a compound can be written using a number of SMILES. A canonical SMILES was employed in the early stages of drug discovery using deep learning for consistent expression; however, randomized SMILES are used more frequently in the de novo drug design research field [135]. According to Kotsias et al. [134], utilizing randomized SMILES rather than canonical SMILES improved the quality of the generative model. Arús-Pous et al. [133] noted that randomized SMILES are mostly employed for de novo drug design; nevertheless, [58] demonstrated that randomized SMILES prove to be trained more reliably and outperformed the conventional form even when predicting IC50.

4.3.4. Transfer Learning

As previously discussed in the last section, lack of data is one of the main issues with AI-assisted drug discovery. It is challenging to train when attempting to identify a specific disease or recently found target because the data size is so limited. Moreover, applying augmentation to all the data is challenging. Alternatively, transfer learning is considered the best solution in a such situation [131,136].

Transfer learning is a component of lifelong learning that draws its inspiration from how rapidly humans derive new information from prior analogous experiences. This approach can be used to address a wide range of issues with insufficient data by fine-tuning a pre-trained model using a large dataset in another or a general field [137]. In order to improve performance, Shin et al. [120] integrated a chemical representation model that was learned through the PubChem database into the proposed DTI model. According to Panagiotis et al., in the de novo study using conditional RNN, the transfer learning technique demonstrated better performance in both DRD2 and CHEMBL25 [134].

4.3.5. Multi-Task Learning

Drug discovery studies commonly employ multi-task learning techniques [136]. Multi-task learning learns several tasks with several shared components. It allows for the training of the features that are challenging to train on the limited data. As demonstrated by Kearnes et al. [138], adopting multi-task learning models improved the AUC performance in comparison to the widely used logistic regression or random forest approach [138]. A few datasets exhibited a little decrease in AUC when utilizing multi-task learning; however, for most of the datasets, AUC increased significantly. Especially interesting is the considerable improvement in the performance of the datasets with a relatively smaller amount of data. From an industrial and practical standpoint, using a pre-trained model has the benefit of greatly reducing the training time and computational resources while also improving performance [139,140]. Hence, for representation learning, it is recommended to employ transfer learning.

4.4. QR3: What Are the Commonly Used Performance Evaluation Metrics for Deep-Learning-Based Molecular Similarity Searching?

What are the common datasets used for evaluating the deep-learning-based molecular similarity searching methods?

The ultimate objectives of any molecular similarity search are quality and high performance. In order to assess the effectiveness of molecular similarity searching, many metrics, including classification metrics and regression metrics, have been developed and used in different approaches. The several criteria employed to assess the deep-learning-based molecular similarity searching discovered in this review are presented in this section. Table 4 displays how the reviewed studies were distributed in relation to the prediction metrics.

4.4.1. Classification Metrics

Studies on molecular similarity searches have implemented different standard evaluation metrics, including precision, recall, specificity, sensitivity, and accuracy. These measures are computed using the confusion matrix. The simplest is accuracy, which is used to measure classifier performance. The accuracy metric, however, does not perform well when there are issues with skewness or class imbalance [149]. Due to the fact that the accurate prediction of the activity in practice is consistently categorized as positive, the precisions and recalls are frequently considered by several deep learning studies. F-score and precision–recall area under the curve (AUPR) are two metrics that allow for simultaneous measurement of precision and recall. Precision and recall are balanced by the F-score, often referred to as the F-measure. In light of this, an F1 score measures the weighted average of precision and recall. The PR-AUC signifies the trade-off between recall and precision while minimizing the effect of false positives. When the classes are unbalanced,

there are additional relevant metrics. For calculating accuracy, sensitivity and specificity are averaged [150]. The Matthews correlation coefficient (MCC) calculates the relationship between actual classes and expected labels. According to Chicco and Jurman [151], MCC provides more information about binary classifications compared to the F1 and accuracy metrics.

Table 4. The distribution of the reviewed studies of the prediction metrics.

Category	Metrics	No. of Study	References
Classification	precision	11	[7,8,11,65,92,97,100,103,104,130,141]
	Recall	04	[92,100,104,141]
	F-score	04	[67,99,141,142]
	Sensitivity	03	[65,92,103]
	Specificity	02	[65,92]
	MCC	04	[92,99,100,106]
	AUC	19	[16,50,65,82,90–92,96,97,99,100,103–107,111,141,143,144]
Regression	Pearson correlation coefficient (R)	02	[34,95]
	Squared correlation coefficient ($R2$)	06	[76,93,108,110,122,145]
	MSE	02	[145,146]
	RMSE	08	[34,76,82,95,111,122,146,147]
	MAE	02	[82,147]
Ranking	Concordance index (CI)	04	[122,145,146]
	Spearman's correlation coefficient (ρ)	04	[90,134,148]

The area under the curve is yet another often-employed evaluation metric for deep learning techniques (AUC). The term "AUC" refers to the region under a receiver–operator characteristic (ROC) curve that differentiates between the effectiveness of the classifier based on two different error types: the false positive and the false negative. The best classifier to achieve perfection is the top-left of the plot, where the ROC curve is the plot of the true positive rate against the false positive rate. The AUC value specifies how well the positive predictions are ranked. Moreover, the AUC is considered more sensitive to imbalanced datasets, which can amount to more false positives [152].

4.4.2. Regression Evaluation Metrics

Several evaluation metrics, including mean square error (MSE), root mean square error (RMSE), Pearson's correlation coefficient (R), and squared correlation coefficient, can be employed to evaluate binding affinity scores comprising the IC50 and pKd predicted by drug–target-interaction prediction models (R2). Different criteria can be used to measure the effectiveness of predictive QSAR models. In terms of binding affinity scores, the MSE is the average squared difference between the predictable and real scores. The RMSE is the squared root of MSE, as its name implies. R2 determines how closely the predicted value and the actual value match up (i.e., the goodness of fit). Some studies applied the modified $R2$ ($r2$) to the test set prediction [122,146].

By comparing the order of predictions with the order of ground truths, other measures, such as the concordance index (CI or C-index) and Spearman's correlation coefficient (ρ), can be used to measure the accuracy of rankings. CI is one of the frequently used ranking metrics in the prediction of drug–target-affinity [137,153]. The CI measures whether the predicted binding affinity values of two random drug–target combinations were in accordance with those actual values. Meanwhile, the degree and direction of the relationship between two ranking variables are measured by Spearman's correlation

coefficient metric. In recent studies, Spearman's correlation coefficient was combined with additional measures [90,134,137,148].

4.5. QR4: What Are the Common Datasets Used for Evaluating the Deep-Learning-Based Molecular Similarity Searching Methods?

About 15 datasets and the application categories identified in the selected articles are presented in Table 5. As could be observed from the table, the category for each dataset and the references where they were applied are indicated. The selected articles used at least one of the datasets. In most cases, a study uses two or more datasets, and some of the datasets are not frequently used.

Table 5. The 15 presented datasets and the domains.

Category	Dataset	Data Type	No. of Studies	References
Physical Chemistry	FressSolv	SMILES	01	[111]
	Lipophilcity	SMILES	01	[111]
	DrugBank	SMILES, 3D coordinates	15	[1,34,44,49,65,67,69,90–92,96,103–105,109,143]
	ChEMBL	SMILES	06	[34,82,97,100,110,141]
	DUD	SMILES, 3D coordinates	05	[16,100,102,107,154]
Biophysics	PubChem Bioassay	SMILES	02	[65,82]
	MUV	SMILES	03	[82,100,104]
	HIV	SMILES	01	[111]
	PDBlind	SMILES, 3D coordinates	05	[34,76,82,95,154]
	KEGG	SMILES	06	[65,90,92,97,105,109]
	Davis		05	[108,122,145,146]
	KIBA		05	[108,122,145,146]
Physiology	Tox21	SMILES	04	[82,98,99,111]
	ToxCast	SMILES	02	[82,108]
	SIDER	SMILES	04	[49,90,108,144]

Table 5 shows that, based on the studies that were considered, the DrugBank dataset seems to be the most widely used dataset, with ChEMBL and KEGG being identified as the second and third, respectively. The table also shows that a variety of publicly available datasets are mostly considered for the evaluation of deep learning techniques. In a nutshell, the selected studies covered about 15 datasets, while the majority of the researchers mainly focused on well-known datasets.

4.6. RQ5: What Are the Research Gaps, Challenges, and Future Directions of Deep-Learning-Based Methods for Molecular Similarity Searching?

This section aims to respond to RQ5 by highlighting the research gaps and the potential future of deep-learning-based methods for molecular similarity searching. There are certain gaps and potential future directions for research in this area despite the fact that the previous works have significantly contributed to the application of deep-learning-based algorithms for molecular similarity searching. Some of the significant issues that still need to be examined in future research include the following:

4.6.1. Data Imbalance

In data mining, deep learning techniques have proven to be efficient and provide promising solutions. For the deep-learning-based approach in the field of molecular searching, the evaluated datasets are almost imbalanced. First, since it requires expensive

research and a great deal of time to generate drug–target interaction data, the volume of data from drug discovery studies is on a small scale [22]. Moreover, the labeled data used in drug discovery are also extremely imbalanced. The high-throughput screening technique does not require a high frequency of active responses; hence, there are substantially fewer active responses in the high-throughput screening data than there are inactive responses. As a result, there are frequently only a few validated drugs available for drug–target interactions that are positive. The majority of the test findings in the PubChem Bioassay dataset had an active to inactive ratio of 1:40.92 (a hit rate of 2.385% of the total labeled activity values) [155].

4.6.2. Data Fusion Method

Data fusion reduces model overfitting and improves the general performance. An alternative is to supplement the data with a little quantity of noise that has no impact on the data performance. One popular data fusion method in drug discovery is randomized SMILES. Kotsias et al. [134] found that utilizing randomized SMILES rather than canonical SMILES improved the quality of the generative model. Despite the fact that Esben [58] showed that randomized SMILES trained more consistently and outperformed the canonical form even when predicting IC50, randomized SMILES is mostly utilized for de novo drug creation [135]. However, in many applications where a number of possible representations of a molecule are needed, such as the drug–target interaction, this requires information on the relationship between the ligand and target, which is not widely used.

4.6.3. Result Interpretation

Non-transparency of result interpretation is another issue in several applications of deep learning methods. The growing importance of the field of deep learning applications into specialized areas as compared to the basic tasks posed several challenges. While the model is not technically a black box, it is viewed as a black box because it can be difficult for a human to understand how the final result was attained [156]. This interpretation often lacks openness, which makes it challenging to accurately comprehend the process of reasoning or making informed decisions about the results.

4.6.4. More Reliable Features

To enhance the effectiveness of molecular similarity searching, there is a need to consider the use of more reliable feature representations and develop deeper architectures [7,8]. Therefore, developing deep-learning-based models on how these important features are explored to improve the effectiveness of the similarity measure performance becomes a promising solution. This has been verified by recent research that used deep learning techniques to increase the performance and diversity of the proposed approaches.

5. Limitations

In this study, the main studies on deep-learning-based molecular similarity searching have been reviewed and analyzed in a systematic approach. Many factors could have impacted the validity of the study. As a result, some of the limitations of this study were highlighted as follows:

- The data extraction method used for this SLR poses several significant limitations. The data used were strictly on the viewpoints of the predefined RQs despite the fact that they were considered to be reasonably sufficient. Therefore, there are chances that the readers will learn about some aspects that are not discussed in this study, and this could greatly improve research trends.
- Despite the fact that five search libraries (as listed in Section 3) were taken into consideration to identify relevant research articles, they are not all-inclusive, which can limit the validity of the study.
- This SLR is only limited to journal and conference papers that cover deep-learning-based molecular similarity searching. Using our search technique, some of the irrele-

vant research publications have been identified and excluded from our review in the first stages of the study. This ensures that the selected research papers conformed to the investigation's requirements. However, it has been suggested that using additional sources, such as additional sourcebooks, could have enhanced the review.

- We restricted our search to only English-language articles. This results in linguistic bias because it is possible that related articles in this field of study exist in other languages. Thankfully, every piece of writing we obtained for this study was written in English. As a result, we are not language-biased.

6. Conclusions

This study provided a systematic review that, strictly based on the publications published from 2010 to 2023, reviewed and analyzed the state-of-the-art deep-learning-based molecular similarity searching approaches. The goal of the study was to provide scholars and practitioners in the area a thorough understanding of molecular similarity searching based on deep learning models. The five main search libraries, which comprise the ACM Digital Library, Science Direct, IEEE Explore, Web of Science, and Springer, were used as the major data source of the study. The key findings of the study included a variety of deep learning techniques for molecular similarity searches, several datasets, and metrics that are widely used to assess the effectiveness of deep-learning-based molecular similarity searches. The most popular domains utilized for deep-learning-based molecular similarity searching were also identified in the study, alongside potential future research areas and existing challenges. In addition, we outlined several potential future directions for deep-learning-based molecular similarity searching. To enhance the effectiveness of molecular similarity searching, there is a need to consider the use of more reliable feature representations and develop deeper architectures. These were strictly based on the suggestions made for further research by the authors of the studies included in the study. Another potential research direction found in the related studies involved using more datasets.

Author Contributions: Writing—original draft, conceptualization, methodology, formal analysis: M.N.; writing—review and editing, supervision: U.K.Y. and N.S., project administration, funding acquisition: U.K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Creativity and Management Office (RCMO) and the School of Computer Sciences at the Universiti Sains Malaysia (USM).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Fatima Alsharjabi for proofreading and editing the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bero, S.A.; Muda, A.K.; Choo, Y.H.; Muda, N.A.; Pratama, S.F. Similarity Measure for Molecular Structure: A Brief Review. In Proceedings of the 6th International Conference on Computer Science and Computational Mathematics (ICCSCM), Langkawi, Malaysia, 4–5 May 2017.
2. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73. [CrossRef]
3. Willett, P. The Calculation of Molecular Structural Similarity: Principles and Practice. *Mol. Inform.* **2014**, *33*, 403–413. [CrossRef] [PubMed]
4. Schomacker, T.; Tropmann-Frick, M. Language Representation Models: An Overview. *Entropy* **2021**, *23*, 1422. [CrossRef] [PubMed]
5. Wang, Z.; Chen, J.; Hoi, S.C.H. Deep Learning for Image Super-Resolution: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3365–3387. [CrossRef]
6. Yang, S.; Wang, Y.; Chu, X. A survey of deep learning techniques for neural machine translation. *arXiv* **2020**, arXiv:2002.07526.
7. Nasser, M.; Salim, N.; Saeed, F.; Basurra, S.; Rabiu, I.; Hamza, H.; Alsoufi, M.A. Feature Reduction for Molecular Similarity Searching Based on Autoencoder Deep Learning. *Biomolecules* **2022**, *12*, 508. [CrossRef]

8. Altalib, M.K.; Salim, N. Similarity-Based Virtual Screen Using Enhanced Siamese Deep Learning Methods. *ACS Omega* **2022**, *7*, 4769–4786. [CrossRef]
9. Muegge, I.; Hu, Y. How do we further enhance 2D fingerprint similarity searching for novel drug discovery? *Expert Opin. Drug Discov.* **2022**, *17*, 1173–1176. [CrossRef]
10. Walters, W.P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263–270. [CrossRef]
11. Altalib, M.K.; Salim, N. Similarity-Based Virtual Screen Using Enhanced Siamese Multi-Layer Perceptron. *Molecules* **2021**, *26*, 6669. [CrossRef]
12. Bee, C.; Chen, Y.-J.; Queen, M.; Ward, D.; Liu, X.; Organick, L.; Seelig, G.; Strauss, K.; Ceze, L. Molecular-level similarity search brings computing to DNA data storage. *Nat. Commun.* **2021**, *12*, 4764. [CrossRef] [PubMed]
13. Devi, K.R.; Pradhan, J.; Bhutia, R.; Dadul, P.; Sarkar, A.; Gohain, N.; Narain, K. Molecular diversity of Mycobacterium tuberculosis complex in Sikkim, India and prediction of dominant spoligotypes using artificial intelligence. *Sci. Rep.* **2021**, *11*, 7365. [CrossRef] [PubMed]
14. Qi, S.; Gao, B.; Zhu, S. Molecular Diversity and Evolution of Antimicrobial Peptides in Musca domestica. *Diversity* **2021**, *13*, 107. [CrossRef]
15. Huber, F.; van der Burg, S.; van der Hooft, J.J.J.; Ridder, L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminform.* **2021**, *13*, 84. [CrossRef]
16. Gao, K.Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. Interpretable Drug Target Prediction Using Deep Neural Representation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 3371–3377.
17. Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K.F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2022**, *12*, e1608. [CrossRef]
18. Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.* **2020**, *37*, 1–12. [CrossRef] [PubMed]
19. Kimber, T.B.; Chen, Y.; Volkamer, A. Deep Learning in Virtual Screening: Recent Applications and Developments. *Int. J. Mol. Sci.* **2021**, *22*, 4435. [CrossRef] [PubMed]
20. Vanhaelen, Q.; Lin, Y.-C.; Zhavoronkov, A. The Advent of Generative Chemistry. *ACS Med. Chem. Lett.* **2020**, *11*, 1496–1505. [CrossRef] [PubMed]
21. Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics*; Springer: Cham, Switzerland, 2020; pp. 445–467.
22. Kim, J.; Park, S.; Min, D.; Kim, W. Comprehensive survey of recent drug discovery using deep learning. *Int. J. Mol. Sci.* **2021**, *22*, 9983. [CrossRef] [PubMed]
23. Kitchenham, B.; Pretorius, R.; Budgen, D.; Brereton, O.P.; Turner, M.; Niazi, M.; Linkman, S. Systematic literature reviews in software engineering—A tertiary study. *Inf. Softw. Technol.* **2010**, *52*, 792–805. [CrossRef]
24. Mathews, J.P.; Chaffee, A.L. The molecular representations of coal—A review. *Fuel* **2012**, *96*, 1–14. [CrossRef]
25. Li, Z.; Ni, G.; Wang, H.; Sun, Q.; Wang, G.; Jiang, B.; Zhang, C. Molecular structure characterization of lignite treated with ionic liquid via FTIR and XRD spectroscopy. *Fuel* **2020**, *272*, 117705.
26. Alsenan, S.A.; Al-Turaiki, I.; Hafez, A. Chemoinformatics for Data Scientists: An Overview. In Proceedings of the 22nd Annual International Conference on Information Integration and Web-Based Applications and Services (IIWAS), Chiang Mai, Thailand, 30 November–2 December 2020; pp. 456–461.
27. Berrhail, F.; Belhadef, H.; Hentabli, H.; Saeed, F. Molecular Similarity Searching with Different Similarity Coefficients and Different Molecular Descriptors. In Proceedings of the 2nd International Conference of Reliable Information and Communication Technology (IRICT), Johor, Malaysia, 23–24 April 2017; pp. 39–47.
28. Nasser, M.; Salim, N.; Hamza, H.; Saeed, F.; Rabiu, I. Improved Deep Learning Based Method for Molecular Similarity Searching Using Stack of Deep Belief Networks. *Molecules* **2021**, *26*, 128. [CrossRef] [PubMed]
29. Bagherian, M.; Sabeti, E.; Wang, K.; Sartor, M.A.; Nikolovska-Coleska, Z.; Najarian, K. Machine learning approaches and databases for prediction of drug–target interaction: A survey paper. *Brief. Bioinform.* **2021**, *22*, 247–269. [CrossRef] [PubMed]
30. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: A review and practical guide. *J. Cheminform.* **2020**, *12*, 56. [CrossRef] [PubMed]
31. Chen, D.; Huang, X.; Fan, Y. Thermodynamics-Based Model Construction for the Accurate Prediction of Molecular Properties From Partition Coefficients. *Front. Chem.* **2021**, *9*, 737579. [CrossRef] [PubMed]
32. Rada, J.; Rodriguez, J.M.; Sigarreta, J.M. General properties on Sombor indices. *Discret. Appl. Math.* **2021**, *299*, 87–97. [CrossRef]
33. Liu, W.; Wang, X.; Zhou, X.; Duan, H.; Zhao, P.; Liu, W. Quantitative structure-activity relationship between the toxicity of amine surfactant and its molecular structure. *Sci. Total Environ.* **2020**, *702*, 134593. [CrossRef]
34. Xie, L.; Xu, L.; Kong, R.; Chang, S.; Xu, X. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Front. Pharmacol.* **2020**, *11*, 606668. [CrossRef]
35. Willett, P. The Literature of Chemoinformatics: 1978–2018. *Int. J. Mol. Sci.* **2020**, *21*, 5576. [CrossRef]
36. Green, H.; Koes, D.R.; Durrant, J.D. DeepFrag: A deep convolutional neural network for fragment-based lead optimization. *Chem. Sci.* **2021**, *12*, 8036–8047. [CrossRef] [PubMed]

37. Arif, S.M.; Holliday, J.D.; Willett, P. The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening. In *Advances in Mathematical Chemistry and Applications*; Bentham Science: Sharjah, United Arab Emirates, 2015; pp. 92–112.
38. Polanski, J.; Gasteiger, J. Computer representation of chemical compounds. In *Handbook of Computational Chemistry*; Springer: Cham, Switzerland, 2017; pp. 1997–2039.
39. Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **2016**, *11*, 137–148. [CrossRef] [PubMed]
40. Markoff, J. Scientists See Advances in Deep Learning a Part of Artificial Intelligence. *New York Times*, 23 November 2012.
41. Dahl, G.E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv* **2014**, arXiv:1406.1231.
42. Yang, M.; Simm, J.; Lam, C.C.; Zakeri, P.; van Westen, G.J.P.; Moreau, Y.; Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. *Sci. Rep.* **2018**, *8*, 8322. [CrossRef]
43. Lee, K.; Kim, D. In-Silico Molecular Binding Prediction for Human Drug Targets Using Deep Neural Multi-Task Learning. *Genes* **2019**, *10*, 906. [CrossRef]
44. Tan, M. Prediction of anti-cancer drug response by kernelized multi-task learning. *Artif. Intell. Med.* **2016**, *73*, 70–77. [CrossRef]
45. Yuan, H.; Paskov, I.; Paskov, H.; González, A.J.; Leslie, C.S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **2016**, *6*, 31619. [CrossRef]
46. Simões, R.S.; Maltarollo, V.G.; Oliveira, P.R.; Honorio, K.M. Transfer and Multi-task Learning in QSAR Modeling: Advances and Challenges. *Front. Pharmacol.* **2018**, *9*, 74. [CrossRef]
47. Burki, T. A new paradigm for drug development. *Lancet Digit. Health* **2020**, *2*, E226–E227. [CrossRef]
48. Richardson, P.; Griffin, I.; Tucker, C.; Smith, D.; Oechsle, O.; Phelan, A.; Rawling, M.; Savory, E.; Stebbing, J. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *Lancet* **2020**, *395*, E30–E31. [CrossRef] [PubMed]
49. Genc-Nayebi, N.; Abran, A. A systematic literature review: Opinion mining studies from mobile app store user reviews. *J. Syst. Softw.* **2017**, *125*, 207–219. [CrossRef]
50. Peng, J.; Li, J.; Shang, X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinform.* **2020**, *21*, 394. [CrossRef] [PubMed]
51. Koge, D.; Ono, N.; Huang, M.; Altaf-Ul-Amin, M.; Kanaya, S. Embedding of Molecular Structure Using Molecular Hypergraph Variational Autoencoder with Metric Learning. *Mol. Inform.* **2021**, *40*, 2000203. [CrossRef]
52. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernandez-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef]
53. Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* **2017**, *8*, 10883–10890. [CrossRef]
54. Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* **2017**, *14*, 3098–3104. [CrossRef]
55. Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharm.* **2018**, *15*, 4398–4405. [CrossRef]
56. Rusk, N. Deep learning. *Nat. Methods* **2016**, *13*, 35. [CrossRef]
57. Nowak, D.; Bachorz, R.A.; Hoffmann, M. Neural Networks in the Design of Molecules with Affinity to Selected Protein Domains. *Int. J. Mol. Sci.* **2023**, *24*, 1762. [CrossRef]
58. Bjerrum, E.J. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv* **2017**, arXiv:1703.07076.
59. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131. [CrossRef] [PubMed]
60. Yang, Y.; Liu, M.; Kitchin, J.R. Neural network embeddings based similarity search method for atomistic systems. *Digit. Discov.* **2022**, *1*, 636–644. [CrossRef]
61. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
62. Mouchlis, V.D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A.G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22*, 1676. [CrossRef] [PubMed]
63. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
64. De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* **2018**, arXiv:1805.11973.
65. Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **2019**, *15*, e1007129. [CrossRef]
66. Pandey, M.; Fernandez, M.; Gentile, F.; Isayev, O.; Tropsha, A.; Stern, A.C.; Cherkasov, A. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **2022**, *4*, 211–221. [CrossRef]
67. Xie, L.; He, S.; Song, X.; Bo, X.; Zhang, Z. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC Genom.* **2018**, *19*, 667. [CrossRef]

68. Korkmaz, S. Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 4180–4190. [CrossRef]
69. Chen, C.; Shi, H.; Jiang, Z.; Salhi, A.; Chen, R.; Cui, X.; Yu, B. DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network. *Comput. Biol. Med.* **2021**, *136*, 104676. [CrossRef]
70. Hayes, T.L.; Krishnan, G.P.; Bazhenov, M.; Siegelmann, H.T.; Sejnowski, T.J.; Kanan, C. Replay in Deep Learning: Current Approaches and Missing Biological Elements. *Neural Comput.* **2021**, *33*, 2908–2950. [CrossRef] [PubMed]
71. Sun, Y.A.; Xue, B.; Zhang, M.; Yen, G.G. Evolving Deep Convolutional Neural Networks for Image Classification. *IEEE Trans. Evol. Comput.* **2020**, *24*, 394–407. [CrossRef]
72. Wang, B.; Xue, B.; Zhang, M. Particle swarm optimization for evolving deep convolutional neural networks for image classification: Single-and multi-objective approaches. In *Deep Neural Evolution: Deep Learning with Evolutionary Computation*; Springer: Singapore, 2020; pp. 155–184.
73. Khan, S.; Sajjad, M.; Hussain, T.; Ullah, A.; Imran, A.S. A Review on Traditional Machine Learning and Deep Learning Models for WBCs Classification in Blood Smear Images. *IEEE Access* **2021**, *9*, 10657–10673. [CrossRef]
74. Morales, D.; Talavera, E.; Remeseiro, B. Playing to distraction: Towards a robust training of CNN classifiers through visual explanation techniques. *Neural Comput. Appl.* **2021**, *33*, 16937–16949. [CrossRef]
75. Asokan, A.; Anitha, J.; Patrut, B.; Danciulescu, D.; Hemanth, D.J. Deep Feature Extraction and Feature Fusion for Bi-Temporal Satellite Image Classification. *CMC-Comput. Mater. Contin.* **2021**, *66*, 373–388. [CrossRef]
76. Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296. [CrossRef] [PubMed]
77. Berrhail, F.; Belhadef, H.; Haddad, M. Deep Convolutional Neural Network to improve the performances of screening process in LBVS. *Expert Syst. Appl.* **2022**, *203*, 117287. [CrossRef]
78. Mendolia, I.; Contino, S.; De Simone, G.; Perricone, U.; Pirrone, R. EMBER—Embedding Multiple Molecular Fingerprints for Virtual Screening. *Int. J. Mol. Sci.* **2022**, *23*, 2156. [CrossRef]
79. Zhao, J.; Ma, Q.; Zhang, B.; Guo, P.; Wang, Z.; Liu, Y.; Meng, M.; Liu, A.; Yang, Z.; Du, G. Exploration of SARS-CoV-2 3CLpro inhibitors by virtual screening methods, FRET detection, and CPE assay. *J. Chem. Inf. Model.* **2021**, *61*, 5763–5773. [CrossRef]
80. Duvenaudt, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gomez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015.
81. Chuang, K.V.; Gunsalus, L.M.; Keiser, M.J. Learning Molecular Representations for Medicinal Chemistry Miniperspective. *J. Med. Chem.* **2020**, *63*, 8705–8722. [CrossRef]
82. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388. [CrossRef]
83. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
84. Lim, J.; Ryu, S.; Park, K.; Choe, Y.J.; Ham, J.; Kim, W.Y. Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J. Chem. Inf. Model.* **2019**, *59*, 3981–3988. [CrossRef]
85. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3313–3332. [CrossRef]
86. Lin, E.; Lin, C.-H.; Lane, H.-Y. Relevant applications of generative adversarial networks in drug design and discovery: Molecular de novo design, dimensionality reduction, and de novo peptide and protein design. *Molecules* **2020**, *25*, 3250. [CrossRef]
87. 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* **2017**, arXiv:1705.10843.
88. Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **2020**, *11*, 10. [CrossRef] [PubMed]
89. Blanchard, A.E.; Stanley, C.; Bhowmik, D. Using GANs with adaptive training data to search for new molecules. *J. Cheminform.* **2021**, *13*, 14. [CrossRef] [PubMed]
90. Wang, H.; Wang, J.; Dong, C.; Lian, Y.; Liu, D.; Yan, Z. A Novel Approach for Drug-Target Interactions Prediction Based on Multimodal Deep Autoencoder. *Front. Pharmacol.* **2020**, *10*, 1592. [CrossRef] [PubMed]
91. Zhao, Y.; Zheng, K.; Guan, B.; Guo, M.; Song, L.; Gao, J.; Qu, H.; Wang, Y.; Shi, D.; Zhang, Y. DLDTI: A learning-based framework for drug-target interaction identification using neural networks and network representation. *J. Transl. Med.* **2020**, *18*, 434. [CrossRef] [PubMed]
92. Wang, Y.-B.; You, Z.-H.; Yang, S.; Yi, H.-C.; Chen, Z.-H.; Zheng, K. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 49. [CrossRef]
93. Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253–1268. [CrossRef] [PubMed]
94. Pandey, M.; Xu, Z.; Sholle, E.; Maliakal, G.; Singh, G.; Fatima, Z.; Larine, D.; Lee, B.C.; Wang, J.; van Rosendael, A.R.; et al. Extraction of radiographic findings from unstructured thoracoabdominal computed tomography reports using convolutional neural network based natural language processing. *PLoS ONE* **2020**, *15*, e0236827. [CrossRef]
95. Gao, K.; Nguyen, D.D.; Sresht, V.; Mathiowetz, A.M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373–8390. [CrossRef] [PubMed]

96. Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-Learning-Based Drug-Target Interaction Prediction. *J. Proteome Res.* **2017**, *16*, 1401–1409. [CrossRef] [PubMed]
97. Lee, H.; Kim, W. Comparison of target features for predicting drug-target interactions by deep neural network based on large-scale drug-induced transcriptome data. *Pharmaceutics* **2019**, *11*, 377. [CrossRef]
98. Hirohara, M.; Saito, Y.; Koda, Y.; Sato, K.; Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinform.* **2018**, *19*, 526. [CrossRef]
99. Matsuzaka, Y.; Uesawa, Y. Prediction Model with High-Performance Constitutive Androstane Receptor (CAR) Using DeepSnap-Deep Learning Approach from the Tox21 10K Compound Library. *Int. J. Mol. Sci.* **2019**, *20*, 4855. [CrossRef]
100. Rifaioglu, A.S.; Nalbat, E.; Atalay, V.; Martin, M.J.; Cetin-Atalay, R.; Doğan, T. DEEPScreen: High performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem. Sci.* **2020**, *11*, 2531–2557. [CrossRef] [PubMed]
101. Gonczarek, A.; Tomczak, J.M.; Zareba, S.; Kaczmar, J.; Dabrowski, P.; Walczak, M.J. Interaction prediction in structure-based virtual screening using deep learning. *Comput. Biol. Med.* **2018**, *100*, 253–258. [CrossRef] [PubMed]
102. Koes, D.; Ragoza, M.; Idrobo, E.; Hochuli, J.; Sunseri, J. Protein-ligand scoring with convolutional neural networks. *Abstr. Pap. Am. Chem. Soc.* **2017**, *57*, 942–957.
103. Mahmud, S.M.H.; Chen, W.; Jahan, H.; Dai, B.; Din, S.U.; Dzisoo, A.M. DeepACTION: A deep learning-based method for predicting novel drug-target interactions. *Anal. Biochem.* **2020**, *610*, 113978. [CrossRef]
104. Tsubaki, M.; Tomii, K.; Sese, J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **2019**, *35*, 309–318. [CrossRef] [PubMed]
105. Shao, K.; Zhang, Z.; He, S.; Bo, X. DTIGCCN: Prediction of drug-target interactions based on GCN and CNN. In Proceedings of the 32nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 9–11 November 2020; pp. 337–342.
106. Pu, L.; Govindaraj, R.G.; Lemoine, J.M.; Wu, H.-C.; Brylinski, M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.* **2019**, *15*, e1006718. [CrossRef] [PubMed]
107. Torng, W.; Altman, R.B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* **2019**, *59*, 4131–4149. [CrossRef]
108. Feng, Q.; Dueva, E.; Cherkasov, A.; Ester, M. Padme: A deep learning-based framework for drug-target interaction prediction. *arXiv* **2018**, arXiv:1807.09741.
109. Mongia, A.; Majumdar, A. Drug-target interaction prediction using multi-graph regularized deep matrix factorization. *BioRxiv* **2019**. [CrossRef]
110. Liu, K.; Sun, X.; Jia, L.; Ma, J.; Xing, H.; Wu, J.; Gao, H.; Sun, Y.; Boulnois, F.; Fan, J. Chemi-Net: A Molecular Graph Convolutional Network for Accurate Drug Property Prediction. *Int. J. Mol. Sci.* **2019**, *20*, 3389. [CrossRef]
111. Shang, C.; Liu, Q.; Chen, K.-S.; Sun, J.; Lu, J.; Yi, J.; Bi, J. Edge attention-based multi-relational graph convolutional networks. *arXiv* **2018**, arXiv:1802.04944.
112. Jeon, W.; Kim, D. FP2VEC: A new molecular featurizer for learning molecular properties. *Bioinformatics* **2019**, *35*, 4979–4985. [CrossRef]
113. Ponti, M.A.; Ribeiro, L.S.F.; Nazare, T.S.; Bui, T.; Collomosse, J. Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask. In Proceedings of the 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Niteroi, Brazil, 17–20 October 2017; pp. 17–41.
114. Hou, Y.; Wang, S.; Bai, B.; Chan, H.C.S.; Yuan, S. Accurate Physical Property Predictions via Deep Learning. *Molecules* **2022**, *27*, 1668. [CrossRef]
115. Goh, G.B.; Siegel, C.; Vishnu, A.; Hodas, N. Using Rule-Based Labels for Weak Supervised Learning A ChemNet for Transferable Chemical Property Prediction. In Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), London, UK, 19–23 August 2018; pp. 302–310.
116. Goh, G.B.; Hodas, N.; Siegel, C.; Vishnu, A. Smiles2vec: Predicting chemical properties from text representations. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
117. Hodas, N.; Siegel, C.; Vishnu, A.; Goh, G. SMILES2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv* **2018**, arXiv:1712.02034.
118. Gini, G.; Zanoli, F.; Gamba, A.; Raitano, G.; Benfenati, E. Could deep learning in neural networks improve the QSAR models? *SAR QSAR Environ. Res.* **2019**, *30*, 617–642. [CrossRef]
119. Phillips, L.; Goh, G.; Hodas, N. Explanatory masks for neural network interpretability. *arXiv* **2019**, arXiv:1911.06876.
120. Cáceres, E.L.; Tudor, M.; Cheng, A.C. Deep learning approaches in predicting ADMET properties. *Future Med. Chem.* **2020**, *12*, 1995–1999. [CrossRef]
121. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
122. Shin, B.; Park, S.; Kang, K.; Ho, J.C. Self-attention based molecule representation for predicting drug-target interaction. In Proceedings of the Machine Learning for Healthcare Conference, Ann Arbor, MI, USA, 8–10 August 2019; pp. 230–248.

123. Huang, K.X.; Xiao, C.; Glass, L.M.; Sun, J.M. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. *Bioinformatics* **2021**, *37*, 830–836. [CrossRef]
124. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci. Rep.* **2021**, *11*, 321. [CrossRef]
125. Shibayama, S.; Marcou, G.; Horvath, D.; Baskin, I.I.; Funatsu, K.; Varnek, A. Application of the mol2vec Technology to Large-size Data Visualization and Analysis. *Mol. Inform.* **2020**, *39*, 1900170. [CrossRef]
126. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35. [CrossRef]
127. Abdo, A.; Pupin, M. LINGO-DL: A text-based approach for molecular similarity searching. *J. Comput. Aided Mol. Des.* **2021**, *35*, 657–665. [CrossRef]
128. Das, N.R.; Mishra, S.P.; Achary, P.G.R. Evaluation of molecular structure based descriptors for the prediction of pEC50(M) for the selective adenosine A_{2A} Receptor. *J. Mol. Struct.* **2021**, *1232*, 130080. [CrossRef]
129. Ahmed, A.; Abdo, A.; Salim, N. Ligand-Based Virtual Screening Using Bayesian Inference Network and Reweighted Fragments. *Sci. World J.* **2012**, *2012*, 410914. [CrossRef]
130. Altalib, M.K.; Salim, N. Hybrid-Enhanced Siamese Similarity Models in Ligand-Based Virtual Screen. *Biomolecules* **2022**, *12*, 1719. [CrossRef]
131. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
132. Cortes-Ciriano, I.; Bender, A. Improved Chemical Structure–Activity Modeling Through Data Augmentation. *J. Chem. Inf. Model.* **2015**, *55*, 2682–2692. [CrossRef]
133. Arús-Pous, J.; Patronov, A.; Bjerrum, E.J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J. Cheminform.* **2020**, *12*, 38. [CrossRef]
134. Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E.J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2*, 254–265. [CrossRef]
135. Arús-Pous, J.; Awale, M.; Probst, D.; Reymond, J.-L. Exploring Chemical Space with Machine Learning. *Chimia* **2019**, *73*, 1018–1023. [CrossRef]
136. Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683–8694. [CrossRef]
137. Playe, B.; Stoven, V. Evaluation of network architecture and data augmentation methods for deep learning in chemogenomics. *bioRxiv* **2019**. [CrossRef]
138. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595–608. [CrossRef]
139. Zhang, X.Y.; Wang, S.; Zhu, F.Y.; Xu, Z.; Wang, Y.H.; Huang, J.Z. Seq3seq Fingerprint: Towards End-to-end Semi-supervised Deep Drug Discovery. In Proceedings of the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB), Washington, DC, USA, 29 August–1 September 2018; pp. 404–413.
140. Li, P.; Wang, J.; Qiao, Y.; Chen, H.; Yu, Y.; Yao, X.; Gao, P.; Xie, G.; Song, S. Learn molecular representations from large-scale unlabeled molecules for drug discovery. *arXiv* **2020**, arXiv:2012.11175.
141. Zhong, F.; Wu, X.; Li, X.; Wang, D.; Fu, Z.; Liu, X.; Wan, X.; Yang, T.; Luo, X.; Chen, K.; et al. Computational target fishing by mining transcriptional data using a novel Siamese spectral-based graph convolutional network. *BioRxiv* **2020**. [CrossRef]
142. Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, D1075–D1079. [CrossRef]
143. Zong, N.; Kim, H.; Ngo, V.; Harismendy, O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* **2017**, *33*, 2337–2344. [CrossRef]
144. Thafar, M.A.; Olayan, R.S.; Ashoor, H.; Albaradei, S.; Bajic, V.B.; Gao, X.; Gojobori, T.; Essack, M. DTiGEMS+: Drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminform.* **2020**, *12*, 44. [CrossRef]
145. Jiang, M.; Li, Z.; Zhang, S.; Wang, S.; Wang, X.; Yuan, Q.; Wei, Z. Drug–target affinity prediction using graph neural network and contact maps. *RSC Adv.* **2020**, *10*, 20701–20712. [CrossRef]
146. Zhao, Q.C.; Xiao, F.; Yang, M.Y.; Li, Y.H.; Wang, J.X. AttentionDTA: Prediction of drug-target binding affinity using attention model. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 64–69.
147. Kwon, Y.; Shin, W.-H.; Ko, J.; Lee, J. AK-Score: Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks. *Int. J. Mol. Sci.* **2020**, *21*, 8424. [CrossRef]
148. Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screen.* **2000**, *3*, 363–372. [CrossRef]
149. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y.A.; Gomaa, M.M.; Hassanien, A.E. Deep learning in drug discovery: An integrative review and future challenges. *Artif. Intell. Rev.* **2022**, 1–63. [CrossRef]
150. Bekkar, M.; Djemaa, H.K.; Alitouche, T.A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3*, 27–38.

151. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
152. Carrington, A.M.; Fieguth, P.W.; Qazi, H.; Holzinger, A.; Chen, H.H.; Mayr, F.; Manuel, D.G. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 4. [CrossRef]
153. Song, X.-Y.; Liu, T.; Qiu, Z.-Y.; You, Z.-H.; Sun, Y.; Jin, L.-T.; Feng, X.-B.; Zhu, L. Prediction of lncRNA-Disease Associations from Heterogeneous Information Network Based on DeepWalk Embedding Model. In *Intelligent Computing Methodologies, Proceedings of the 16th International Conference, ICIC 2020, Bari, Italy, 2–5 October 2020*; Springer Nature: Cham, Switzerland, 2020; Part III; pp. 291–300.
154. Gonczarek, A.; Tomczak, J.M.; Zaręba, S.; Kaczmar, J.; Dąbrowski, P.; Walczak, M.J. Learning deep architectures for interaction prediction in structure-based virtual screening. *arXiv* **2016**, arXiv:1610.07187.
155. Tran-Nguyen, V.-K.; Rognan, D. Benchmarking Data Sets from PubChem BioAssay Data: Current Scenario and Room for Improvement. *Int. J. Mol. Sci.* **2020**, *21*, 4380. [CrossRef]
156. Cho, Y.-R.; Kang, M. Interpretable machine learning in bioinformatics Introduction. *Methods* **2020**, *179*, 1–2. [CrossRef]