molecules

MDPI

Review

Artificial Intelligence for Autonomous Molecular Design: A Perspective

Rajendra P. Joshi and Neeraj Kumar *

Computational Biology Group, Biological Science Division, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99352, USA; rajendra.joshi@pnnl.gov
* Correspondence: neeraj.kumar@pnnl.gov; Tel.: +1-509-372-6422

Abstract: Domain-aware artificial intelligence has been increasingly adopted in recent years to expedite molecular design in various applications, including drug design and discovery. Recent advances in areas such as physics-informed machine learning and reasoning, software engineering, high-end hardware development, and computing infrastructures are providing opportunities to build scalable and explainable AI molecular discovery systems. This could improve a design hypothesis through feedback analysis, data integration that can provide a basis for the introduction of end-to-end automation for compound discovery and optimization, and enable more intelligent searches of chemical space. Several state-of-the-art ML architectures are predominantly and independently used for predicting the properties of small molecules, their high throughput synthesis, and screening, iteratively identifying and optimizing lead therapeutic candidates. However, such deep learning and ML approaches also raise considerable conceptual, technical, scalability, and end-to-end error quantification challenges, as well as skepticism about the current AI hype to build automated tools. To this end, synergistically and intelligently using these individual components along with robust quantum physics-based molecular representation and data generation tools in a closed-loop holds enormous promise for accelerated therapeutic design to critically analyze the opportunities and challenges for their more widespread application. This article aims to identify the most recent technology and breakthrough achieved by each of the components and discusses how such autonomous AI and ML workflows can be integrated to radically accelerate the protein target or disease model-based probe design that can be iteratively validated experimentally. Taken together, this could significantly reduce the timeline for end-to-end therapeutic discovery and optimization upon the arrival of any novel zoonotic transmission event. Our article serves as a guide for medicinal, computational chemistry and biology, analytical chemistry, and the ML community to practice autonomous molecular design in precision medicine and drug discovery.

Keywords: autonomous workflow; therapeutic design; computer aided drug discovery; computational modeling and simulations; quantum mechanics and quantum computing; artificial intelligence; machine learning; deep learning; machine reasoning and causal inference and causal reasoning

check for updates

Citation: Joshi, R.P.; Kumar, N. Artificial Intelligence for Autonomous Molecular Design: A Perspective. *Molecules* **2021**, *26*, 6761. https://doi.org/10.3390/molecules26226761

Academic Editor: Rita Prosmiti

Received: 16 August 2021
Accepted: 29 October 2021
Published: 9 November 2021

1. Introduction

Synthesizing and characterizing small molecules in a laboratory with desired properties is a time-consuming task [1]. Until recently, experimental laboratories have been mostly human operated; they relied completely on the experts of the field to design experiments, carry out characterization, analyze, validate, and conduct decision making for the final product. Moreover, the experimental process involves a series of steps, each requiring several correlated parameters that need to be tuned [2,3], which is a daunting task, as each parameter set conventionally demands individual experiments. This has slowed down the discovery of high-impact small molecules and/or materials, in some case by decades, with possible implications for diverse fields, such as in energy storage, electronics, catalysis, drug discovery, etc.

Moreover, the high-impact materials of today come from exploring only a fraction of the known chemical space. Larger portions of the chemical space are still uncovered, and it is expected to contain exotic materials with the potential to bring unprecedented advances to state-of-the-art technologies. Exploring such a large space with conventional experiments will take time and a lot of resources [4–7]. In this scenario, complete automation of laboratories is long overdue and has been used with limited success in the past [8–12]. The concept of laboratory automation is not new [13]. It was used with limited success for material discovery in the past. More recently, automation has re-emerged as the approach of potential interest due to the significant development in computing architecture, sophisticated material synthesis, and characterization techniques, increasing the successful adoption of deep learning based models in physical and biological science domains. Automating the computational design of small molecules that integrates physics-based simulations and optimization with ML approaches is a feasible and efficient alternative instead; it significantly contributes in expediting autonomous molecular design.

High throughput quantum mechanical calculations, such as density functional theory (DFT), based simulations are the first step towards this goal of providing insight into larger chemical space and have shown some promise in accelerating novel molecule discovery. However, the physics based modeling still requires human intelligence for different decision-making processes, and for instance, it cannot autonomously guide small-molecule therapeutic design steps, thus slowing down the entire process. In addition, the inverse design of molecules is equally difficult with quantum mechanical simulations alone. The amount of data produced by these high throughput methods is so large that it cannot be analyzed in real-time with conventional methods. Autonomous computational design and characterization of molecules is more important in the scenarios where existing experimental/computational approaches are inefficient [14,15].

One such particular example is the challenge associated with identifying new metabolites in a biological sample from mass spectrometry data, which requires mapping the fragmented spectra of novel molecules to the existing spectral library, making it slow and tedious. In many cases, such references libraries do not exist, and an ML-integrated, automated workflow could be an ideal choice to deploy for the rapid identification of metabolites and the expansion of the existing libraries for future reference. Such a workflow has shown the early ability to quickly screen molecules and accurately predict their properties for different applications. The synergistic use of high throughput methods in a closed loop with machine-learning-based methods capable of inverse design is considered vital for autonomous and accelerated discovery of molecules [11].

In this contribution, we discuss how computational workflows for autonomous molecular design can guide the bigger goal of laboratory automation through active learning approaches. At first, we assess the performance of current state-of-the-art artificial intelligence (AI)-guided molecular design tools, mainly focusing on small molecule for therapeutic design and discovery. We start with an extensive discussion of popular molecular representation with various formulation and data generation tools used in advanced ML and deep learning (DL) models. We also benchmark the physics informed predictive ML by comparing various property predictions, which is critical for small-molecule design. In the end, we highlighted the cutting edge AI tools to utilize these ML models for inverse design with desired properties.

2. Results and Highlights

2.1. Components of Computational Autonomous Molecular Design Workflow

The workflow for computational autonomous molecular design (CAMD) must be an integrated and closed-loop system (Figure 1) with: (i) efficient data generation and extraction tools, (ii) robust data representation techniques, (iii) physics-informed predictive machine learning models, and (iv) tools to generate new molecules using the knowledge learned from steps i–iii. Ideally, an autonomous computational workflow for molecule discovery would learn from its own experience and adjust its functionality as the chemical environment or the targeted functionality changes through active learning. This can be achieved when all the components work in collaboration with each other, providing feedback while improving model performance as we move from one step to other.

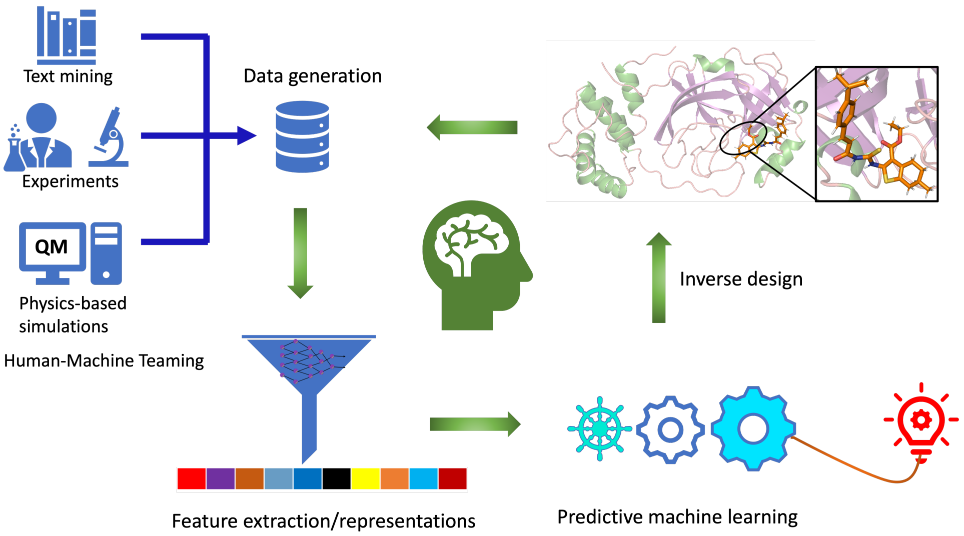
Text mining
Experiments
QM
Physics-based simulations
Human-Machine Teaming
Data generation
Feature extraction/representations
Predictive machine learning
Inverse design

Figure 1. Closed-loop workflow for computational autonomous molecular design (CAMD) for medical therapeutics. Individual components of the workflow are labeled. It consists of data generation, feature extraction, predictive machine learning and an inverse molecular design engine.

For data generation in CAMD, high-throughput density functional theory (DFT) [16,17] is a common choice mainly because of its reasonable accuracy and efficiency [18,19]. In DFT, we typically feed in 3D structures to predict the properties of interest. Data generated from DFT simulations is processed to extract the more relevant structural and properties data, which are then either used as input to learn the representation [20,21] or as a target required for the ML models [22–24]. Data generated can be used in two different ways: to predict the properties of new molecules using a direct supervised ML approach and to generate new molecules with the desired properties of interest using inverse design. CAMD can be tied with supplementary components, such as databases, to store the data and visualize it. The AI-assisted CAMD workflow presented here is the first step in developing automated workflows for molecular design. Such an automated pipeline will not only accelerate the hit identification and lead optimization for the desired therapeutic candidates but can actively be used for machine reasoning to develop transparent and interpretable ML models. These workflows, in principle, can be combined intelligently with experimental setups for computer-aided synthesis or screening planning that includes synthesis and characterization tools, which are expensive to explore in the desired chemical space. Instead, experimental measurements and characterization should be performed intelligently for only the AI-designed lead compounds obtained from CAMD.

The data generated from inverse design in principle should be validated by using an integrated DFT method for the desired properties or by high throughput docking with a target protein to find out its affinity in the closed-loop system, then accordingly update the rest of the CAMD. These steps are then repeated in a closed loop, thus improving and optimizing the data representation, property prediction, and new data generation component. Once we have confidence in our workflow to generate valid new molecules, the validation step with DFT can be bypassed or replaced with an ML predictive tool to make the workflow computationally more efficient. In the following, we briefly discuss the main component of the CAMD, while reviewing the recent breakthroughs achieved.

2.2. Data Generation and Molecular Representation

ML models are data-centric—the more data, the better the model performance. A lack of accurate, ethically sourced well-curated data is the major bottleneck limiting their use in many domains of physical and biological science. For some sub-domains, a limited amount of data exists that comes mainly from physics-based simulations in databases [25,26] or from experimental databases, such as NIST [27]. For other fields, such as for bio-chemical reactions [28], we have databases with the free energy of reactions, but they are obtained with empirical methods, which are not considered ideal as ground truth for machine learning models. For many domains, accurate and curated data does not exist. In these scenarios, slightly unconventional yet very effective approaches of creating data from published scientific literature and patents for ML have recently gained adoption [29–32]. These approaches are based on the natural language processing (NLP) to extract chemistry and biology data from open sources published literature. Developing a cutting edge NLP-based tool to extract, learn, and reason the extracted data would definitely reduce timeline for high throughput experimental design in the lab. This would significantly expedite the decision making based on the existing literature to set up future experiments in a semi-automated way. The resulting tools based on human–machine teaming is much needed for scientific discovery.

2.3. Molecular Representation in Automated Pipelines

Robust representation of molecules is required for accurate functioning of the ML models [33]. An ideal molecular representation should be unique, invariant with respect to different symmetry operations, invertible, efficient to obtain, and capture the physics, stereo chemistry, and structural motif. Some of these can be achieved by using the physical, chemical, and structural properties [34], which, all together, are rarely well documented so obtaining this information is considered cumbersome task. Over time, this has been tackled by using several alternative approaches that work well for specific problems [35–40] as shown in Figure 2. However, developing universal representations of molecules for diverse ML problems is still a challenging task, and any gold standard method that works consistently for all kind of problems is yet to be discovered. Molecular representations primarily used in the literature falls into two broad categories: (a) 1D and/or 2D representations designed by experts using domain specific knowledge, including properties from the simulation and experiments, and (b) iteratively learned molecular representations directly from the 3D nuclear coordinates/properties within ML frameworks.

Expert-engineered molecular representations have been extensively used for predictive modeling in the last decade, which includes properties of the molecules [41,42], structured text sequences [43–45] (SMILES, InChI), molecular fingerprints [46], among others. Such representations are carefully selected for each specific problem using domain expertise, a lot of resources, and time. The SMILES representation of molecules is the main workhorse as a starting point for both representation learning as well as for generating expert-engineered molecular descriptors. For the latter, SMILES strings can be used directly as one hot encoded vector to calculate fingerprints or to calculate the range of empirical properties using different open source platforms, such as RDkit [47] or chemaxon [48], thereby bypassing expensive features generation from quantum chemistry/experiments by providing a faster speed and diverse properties, including 3D coordinates, for molecular representations. Moreover, SMILES can be easily converted into 2D graphs, which is the preferred choice to date for generative modeling, where molecules are treated as graphs with nodes and edges. Although significant progress has been made in molecular generative modeling using mainly SMILES strings [43], they often lead to the generation of syntactically invalid molecules and are synthetically unexplored. In addition, SMILES are also known to violate fundamental physics and chemistry-based constraints [49,50]. Case-specific solutions to circumvent some of these problems exist, but a universal solution is still unknown. The extension of SMILES was attempted by more robustly encoding rings and branches of molecules to find more concrete representations with high semantical and syntactical validity using canonical SMILES [51,52], InChI [44,45], SMARTS [53], DeepSMILES [54], DESMILES [55], etc. More recently, Kren et al. proposed 100% syntactically correct and robust string-based representation of molecules known as SELFIES [49], which has been increasingly adopted for predictive and generative modeling [56].

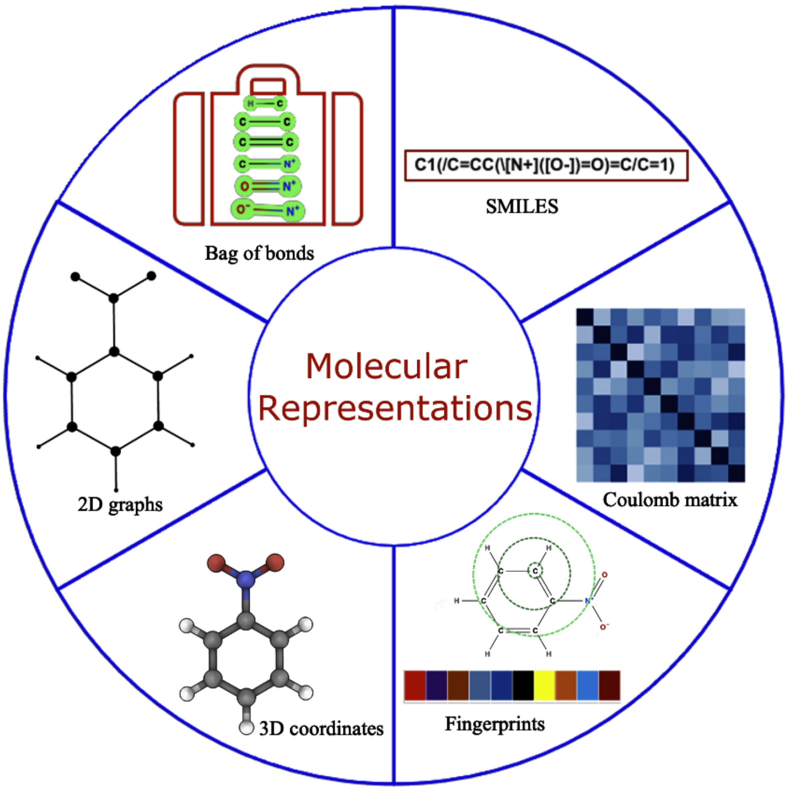

Figure 2. Molecular representation with all possible formulation used in the literature for predictive and generative modeling.

Recently, molecular representations that can be iteratively learned directly from molecules have been increasingly adopted, mainly for predictive molecular modeling, achieving chemical accuracy for a range of properties [34,57,58]. Such representations as shown in Figure 3 are more robust and outperform expert-designed representations in drug design and discovery [59]. For representation learning, different variants of graph neural networks are a popular choice [37,60]. It starts with generating the atom (node) and bond (edge) features for all the atoms and bonds within a molecule, which are iteratively updated using graph traversal algorithms, taking into account the chemical environment information to learn a robust molecular representation. The starting atom and bond features of the molecule may just be one hot encoded vector to only include atom-type, bond-type, or a list of properties of the atom and bonds derived from SMILES strings. Yang et al. achieved the chemical accuracy for predicting a number of properties with their ML models by combining the atom and bond features of molecules with global state features before being updated during the iterative process [61].

Molecules are 3D multiconformational entities, and hence, it is natural to assume that they can be well represented by the nuclear coordinates as is the case of physics-based molecular simulations [62]. However, with coordinates, the representation of molecules is non-invariant, non-invertible, and non-unique in nature [35] and hence not commonly used in conventional machine learning. In addition, the coordinates by itself do not carry information about the key attribute of molecules, such as bond types, symmetry, spin states, charge, etc., in a molecule. Approaches/architectures have been proposed to create robust, unique, and invariant representations from nuclear coordinates using atom-centered Gaussian functions, tensor field networks, and, more robustly, by using representation learning techniques [34,58,63–66], as shown in Figure 3.

Chen et al. [34] achieved chemical accuracy for predicting a number of properties with their ML models by combining the atom and bond features of molecules with global state features of the molecules and are updated during the iterative process. The robust representation of molecules can also only be learned from the nuclear charge and coordinates of molecules, as demonstrated by Schutt et al. [58,63,65]. Different variants (see Equation (1)) of message passing neural networks for representation learning have been proposed, with the main differences being how the messages are passed between the nodes and edges and how they are updated during the iterative process using hidden states h_v^t. Hidden states at each node during the message passing phase are updated using

$$m_v^{t+1} = \sum M_t(h_v^t, h_w^t, h_{vw}^t), \quad h_v^{t+1} = S_t(h_v^t, m_v^{t+1}) \tag{1}$$

where M_t and S_t are the message and vertex update functions, whereas h_v^t and h_{vw}^t are the node and edge features. The summation runs over all the neighbor of v in the entire molecular graph. This information is used by a readout phase to generate the feature vector for the molecule, which is then used for the property prediction.

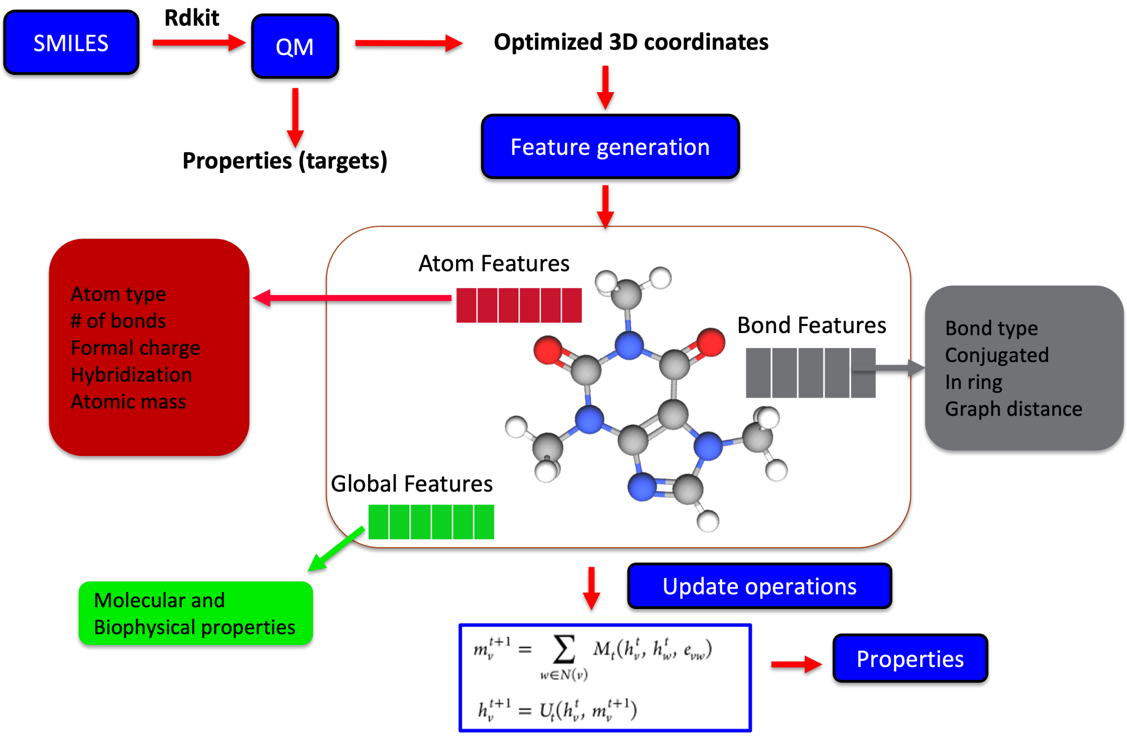

Figure 3. The iterative update process used for learning a robust molecular representation either based on 2D SMILES or 3D optimized geometrical coordinates from physics-based simulations. The molecular graph is usually represented by features at the atomic level, bond level, and global state, which represents the key properties. Each of these features are iteratively updated during the representation learning phase, which are subsequently used for the predictive part of model.

These approaches, however, require a relatively large amount of data and computationally intensive DFT optimized ground state coordinates for the desired accuracy, thus limiting their use for domains/datasets lacking them. Moreover, representations learned from a particular 3D coordinate of a molecule fail to capture the conformer flexibility on its potential energy surface [66], thus requiring expensive multiple QM-based calculations for each conformer of the molecule. Some work in this direction based on semi-empirical DFT calculations to produce a database of conformers with 3D geometry has been recently published [66]. This, however, does not provide any significant improvement in predictive power. These methods, in practice, can be used with empirical coordinates generated from SMILES using RDkit/chemaxon but still require the corresponding ground state target properties for building a robust predictive modeling engine as well as optimizing the properties of new molecules with generative modeling.

Moreover, in these physics-based models, the cutoff distance is used to restrict the interaction among the atoms to the local environments only, hence generating local representations. In many molecular systems and for several applications, explicit non-local interactions are equally important [67]. Long-range interactions have been implemented in convolutional neural networks; however, they are known to be inefficient in information propagation. Matlock et al. [68] proposed a novel architecture to encode non-local features of molecules in terms of efficient local features in aromatic and conjugated systems using gated recurrent units. In their models, information is propagated back and forth in the molecules in the form of waves, making it possible to pass the information locally while simultaneously traveling the entire molecule in a single pass. With the unprecedented success of learned molecular representations for predictive modeling, they are also adopted with success for generative models [57,69].

2.4. Physics-Informed Machine Learning

Physics-informed machine learning (PIML) is the most widely studied area of applied mathematics in molecular modeling, drug discovery, and medicine [58,63,65,70–76]. Depending upon whether the ML architecture requires the pre-defined input representations as input features or can learn their own input representation by itself, PIML can be broadly classified into two sub-categories. The former is well covered in several recent review articles [70–75]. We will focus only on the latter, which has been increasingly adopted in predictive machine learning recently with unprecedented accuracy for a range of properties and datasets. A number of related approaches for predictive feature/property learning have been proposed in recent years under the umbrella term graph-based models so-called graph neural networks (GNNs) [77–79] and extensively tested on different quantum chemistry benchmark datasets. GNN for predictive molecular modeling consists of two phases: representation learning and property prediction, integrated end-to-end in a way to learn the meaningful representation of the molecules while simultaneously learning how to use the learned feature for the accurate prediction of properties. In the feature-learning phase, atoms and bond connectivity information read from the nuclear coordinates or graph inputs are updated by passing through a sequence of layers for robust chemical encoding, which are then used in subsequent property prediction blocks. The learned features can than be processed using dimensionality reduction techniques before using them in a subsequent property prediction block, as shown in Figure 4.

In one of the first works on embedded feature learning, Schütt et al. [63] used the concept of many body Hamiltonians to devise the size extensive, rotational, translational, and permutationally invariant deep tensorial neural network (DTNN) architecture for molecular feature learning and property prediction. Starting with the embedded atomic number and nuclear coordinates as input, and after a series of refinement steps to encode the chemical environment, their approach learns the atom-centered Gaussian-basis function as a feature that can be used to predict the atomic contribution for a given molecular property. The total property of the molecule is the sum over the atomic contribution. They demonstrated chemical accuracy of 1 kcal mol^{-1} in the total energy prediction for relatively small molecules in the QM7/QM9 dataset that contains only H, C, N, O, and F atoms.

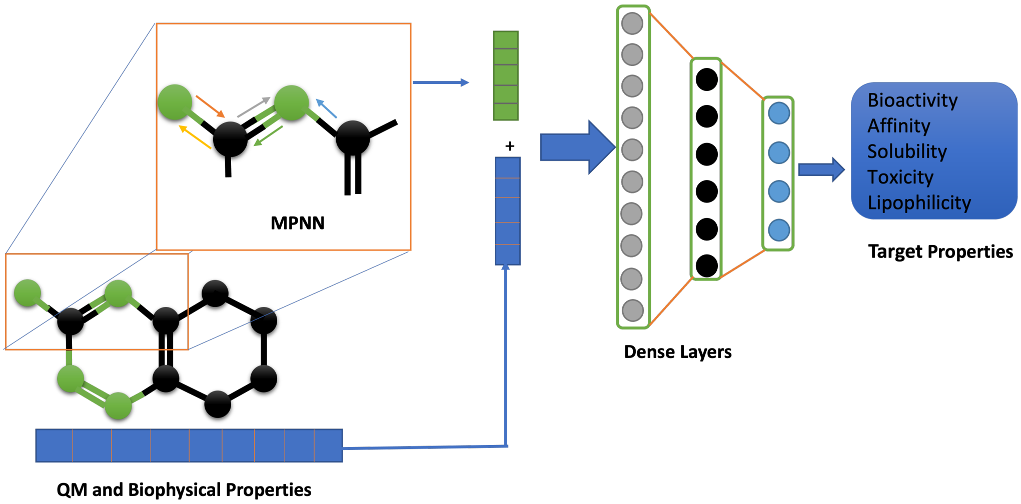

Figure 4. Physics-informed ML framework for predictive modeling. It takes into account the properties obtained from quantum mechanics-based simulation or from experimental data to ultimately generate features in addition to the standard process used in benchmark models (e.g., message passing neural network (MPNN).

Building on DTNN, Schütt et al. [58] also proposed a SchNet model, where the interactions between the atoms are encoded using a continuous filter convolution layer before being processed by filter generating neural networks. The predictive power of their model was further extended for electronic, optical, and thermodynamic properties of molecules in the QM9 dataset compared to only the total energy in DTNN, achieving state-of-the-art chemical accuracy in 8 out of 12 properties. The improved accuracy was observed over a related approach of Gilmer et al. [37], known as message passing neural network (MPNN), on a number of properties except polarizability and electronic spatial extent. In contrast to the SchNet/DTNN model, which learns atom-wise representation of the molecule, MPNN learns the global representation of molecules from the atomic number, nuclear coordinates, and other relevant bond-attributes and uses it for the molecular property prediction. It is critical to mention that MPNN is more accurate for the intensive properties (α, $\langle R^2 \rangle$) where the decomposition into individual atomic contributions is not required. The performance of SchNet is further improved by Jørgensen et al. [80] by making edge features inclusive of the atom receiving the message.

In another related model, Chen et al. [34] proposed an integrated framework with unique feature update steps that work equally well for molecules and solids. They used several atom attributes and bond attributes and then combined it with the global state attribute to learn the feature representation of molecules. It was claimed that their method is outperforming the SchNet model in 11 out of 13 properties, including U0, U, H, and G in the benchmark QM9 dataset. However, they trained their model for respective atomization energies (P $-$ $n_X X_p$, P = U0, U, H, and G) in contrast to the parent U0, U, H, and G trained model of Schnet. Based on our extensive assessment, a fair comparison of the model should be made between the similar quantities. These models also demonstrated that a model trained for predicting a single property of molecules with a graph-based model will always outperform the model optimized for predicting all the properties simultaneously. Other variants of MPNN are also published in the literature with slight improvements in accuracy for predicting some of the properties in the QM9 dataset over the parent MPNN [61,80]. The key features of a few benchmark models with their advantages and disadvantages are listed in Table 1. One particular approach is of Jorgenson et al. [80], where they extended the SchNet model in a way that the message exchanged between the atoms depends not only on the atom sending it but also on the atom receiving it. The comparison of mean absolute errors obtained from some of the benchmark models with their target chemical accuracy are reported in Table 2. This shows that the appropriate ML models, when used with the proper representation of molecules and a well-curated accurate dataset, a well-sought state-of-the-art chemical accuracy from machine learning can be achieved.

Table 1. Highlights and benchmark of predictive ML methods, their comparison, including their key features, advantages, and disadvantages.

Methods	Key Feature	Advantage	Drawbacks
MPNN [60]	• Message exchanged between the atoms depends only on the feature of the sending atom and the corresponding edge features and is independent of the representation of the atom receiving the message • Generate global representation of the molecule • Predicted property of the molecule is the function of global representations of the molecule • Generate messages centered on the atoms	• Achieved chemical accuracy in 11 out of 13 properties in QM9 data • Performs well for intensive properties	• Including the state of the message-receiving atom (dubbed as pair message) increases the property prediction error • The message passed from atom A to atom B can be transmitted back to atom B, resulting in noise
d-MPNN [61]	• Learns molecular representation centered on bonds instead of atoms • Update on MPNN that combines the learned representation with the prior known fixed atomic, bond, and global molecular descriptors	• Avoid noise resulting from the message being passed along any path by using directed messages • Use only SMILES string to generate input representation	• Does not use spatial information as a part of input features
SchNet [58]	• Learns the atomistic representations of the molecules • The total property of the molecule is the sum over the atomic contributions • Learns representations only by using the atomic number and geometry as atom and bond features, respectively	• Improves the performance on 8 out of 13 properties in QM9 data compared to MPNN • Performs relatively well compared to MPNN for extensive properties • Requires only the nuclear charge and nuclear coordinates for learning input representations	• Relatively poor performance for intensive properties compared to MPNN • Use optimized 3D coordinates
MEGNet [34]	• Learns the global representations of the molecules • Uses several atomic and bond properties of the atom and bond as atom and bond features • Adds the global state attribute of molecule in addition to atom and bond feature	• Improves the performance on all the extensive properties compared to MPNN and SchNet • Works equally well for molecules and solid • Provides good accuracy with RDkit-generated 3D coordinates	• Larger error for intensive properties compared to MPNN • It calculates MAE errors for atomization energies of U0, U, H, and G and compares with MAE on U0, U, H, and G of SchNet
SchNet-edge [80]	• Edge feature also depends on the features of the atom receiving the message	• Improves the accuracy of the model over SchNet/MPNN in all the properties in the QM9 dataset	• Requires optimized 3D coordinates

Table 2. Mean absolute errors obtained from several benchmark methods on 12 different properties using the QM9 molecular dataset. Bold represents the lowest mean absolute errors among the models. * represents the property trained for respective atomization energies. Target corresponds to the chemical accuracy for each property desired from the predictive ML models.

Property	Units	MPNN	SchNet-Edge	SchNet	MegNet	Target
HOMO	eV	0.043	**0.037**	0.041	0.038 ± 0.001	0.043
LUMO	eV	0.037	**0.031**	0.034	**0.031 ± 0.000**	0.043
band gap	eV	0.069	**0.058**	0.063	0.061 ± 0.001	0.043
ZPVE	meV	1.500	1.490	1.700	**1.400 ± 0.060**	1.200
dipole moment	Debye	0.030	**0.029**	0.033	0.040 ± 0.001	0.100
polarizability	Bohr^2	0.092	**0.077**	0.235	0.083 ± 0.001	0.100
R^2	Bohr^2	0.180	**0.072**	0.073	0.265 ± 0.001	1.200
U_0	eV	0.019	0.011 *	0.014	**0.009 ± 0.000 ***	0.043
U	eV	0.019	0.016 *	0.019	**0.010 ± 0.000 ***	0.043
H	eV	0.017	0.011 *	0.014	**0.010 ± 0.000 ***	0.043
G	eV	0.019	0.012 *	0.014	**0.010 ± 0.000 ***	0.043
C_v	cal $(\text{mol K})^{-1}$	0.040	0.032	0.033	**0.030 ± 0.000**	0.050

2.5. Inverse Molecular Design

To achieve the long overdue goal of exploring a large chemical space, accelerated molecular design, and generation of molecules with desired properties, inverse design is unavoidable. It is generally known that a molecule should have specific functionalities for it to be an effective therapeutic candidate against a particular disease, but in many cases, new molecules that host such functionalities are not easily known with a direct approach. Furthermore, the pool where such molecules may exist is astronomically large [81–83] (approx. 10^{60} molecules), making it impossible to explore each of them by quantum mechanics-based simulations or experiments.

In such scenarios, inverse design is of significant interest, where the focus is on quickly identifying novel molecules with desired properties in contrast to the conventional, so-called direct approach where known molecules are explored for different properties. In inverse design, we usually start with the initial dataset, for which we know the structure and properties, and map this to a probability distribution and then use it to generate new, previously unknown candidate molecules with desired properties very efficiently. Inverse design uses optimization and search algorithms [84,85] for the purpose and, by itself, can accelerate the lead molecule discovery process, which is the first step for any drug development. This paradigm holds even more promise when used in a closed loop with synthesis, characterization, and different test tools in such a way that each of these steps receives and transmits feedback concurrently, thus improving each other over time. This has shown some promise recently by substantially reducing the timeline for the commercialization of molecules from its discovery to days, which is otherwise known to span over a decade in most cases. In one recent work, Zhavoronkov et al. [1] designed, developed, and tested a workflow that integrates deep reinforcement learning with experimental synthesis, characterization, and test tools for the de novo design of drug molecules as potential inhibitors of the discoidin domain receptor-1 in 21 days. Such a paradigm shift in the design of drugs is possible only because of recently developed deep generative model architectures. Here, we briefly discuss some of the breakthrough architectures along with the recent applications in drug discovery.

Variational autoencoders [86] (VAEs) and its different variants have been extensively used for generating small molecules with optimal physio-chemical and biological properties. VAEs consist of an encoder and decoder network, where the encoder functions as a compression tool for compressing high-dimensional discrete molecular representations to a continuous vector in low-dimensional latent space, whereas the decoder recreates the original molecules from the compressed space. Within VAEs, recurrent neural networks (RNN) [87] and convolution neural networks (CNN) [88] are commonly used as encoding networks, whereas several RNN-based architectures, such as GRU and LSTM, are used as the decoder network. RNN independently has also been used to generate molecules. Bombarelli et al. [86] first used VAEs to generate molecules in the form of SMILES strings from latent space while simultaneously predicting their properties. For property prediction, they coupled the encoder–decoder network with the predictor network, which uses the vector from latent space as an input. SMILES strings generated from their VAEs do not always correspond to valid molecules. To improve on this, Kusner et al. [89] proposed a variant of VAEs known as the grammar VAE that imposes a constraint on SMILES generation by using context-free grammars rules. Both of these works employed string-based molecular representations. More recent works have focused on using molecular graphs as input and output for variational auto-encoders [90] using different variants of VAEs, among others [89–91], such as stacked auto-encoder, semi-supervised deep autoencoders, adversial autoencoder, and Junction Tree Variational Auto-Encoder (JT-VAE), for generating molecules for drug discovery. In JT-VAE [91], tree-like structures are generated from the valid sub-graph components of molecules and encoded along with a full graph to form two complementary latent spaces: one for the molecular graph and another for the corresponding junction tree. These two spaces are then used for hierarchical decoding, generating 100% valid small molecules. Further improvement on this includes using JT-VAE in combination with auto-regressive and graph-to-graph translation methods for valid large-molecule generation [92].

Generative adversarial networks (GANs) are another class of NN popular for generating molecules [93–95]. They consist of generative and discriminative models that work in coordination with each other where the generator is trained to generate a molecule and the discriminator is trained to check the accuracy of the generated molecules. Kadurin et al. [95] successfully first used the GAN architecture for de novo generation of molecules with anti-cancer properties, where they demonstrated higher flexibility, more efficient training, and processing of a larger dataset compared to VAEs. However, it uses unconventional binary chemical compound feature vectors and requires cumbersome validation of output fingerprints against the PubChem chemical library. Guimaraes et al. [96] and Sanchez-Lengeling et al. [97] used a sequence-based generative adversarial network in combination with reinforcement learning for molecule generation, where they bias the generator to produce molecules with desired properties. The works of Guimaraes et al. and Sanchez-Lengeling et al. suffer from several issues associated with a GAN, including mode collapse during training, among others. Some of these issues can be eliminated by using the reinforced adversarial neural computer method [98], which extends their work. Similar to VAEs, GANs have also been used for molecular graph generation, which is considered more robust compared to SMILES string generation. Cao et al. [94] non-sequentially and efficiently generated the molecular graph of small molecules with high validity and novelty from a jointly trained GAN and reinforcement learning architectures. Maziarka et al. [92] proposed a method for graph-to-graph translation, where they generated 100% valid molecules identical with the input molecules but with different desired properties. Their approach relies on the latent space trained for JT-VAE and a degree of similarity of the generated molecules to the starting ones can be tuned. Mendez-Lucio et al. [99] proposed conditional generative adversarial networks to generate molecules that produce a desired biological effect at a cellular level, thus bridging the system's biology and molecular design. A deep convolution NN-based GAN [93] was used for de novo drug design targeting types of cannabinoid receptors.

Generative models, such as GANs, RNNs, and VAEs, have been used together with reward-driven and dynamic decision making reinforcement learning (RL) techniques in many cases with unprecedented success in generating molecules. Popova et al. [100] recently used deep-RL for the de novo design of molecules with desired hydrophobicity or inhibitory activity against Janus protein kinase 2. They trained a generative and a predictive model separately first and then trained both together using an RL approach by biasing the model for generating molecules with desired properties. In RL, an agent, which is a neural network, takes actions to maximize the desired outcome by exploring the chemical space and taking actions based on the reward, penalties, and policies setup to maximize the desired outcome. Olivecrona et al. [101] trained a policy-based RL model for generating the bioactives against dopamine receptor type 2 and generated molecules with more than 95% active molecules. Furthermore, taking an example of the drug Celecoxib, they demonstrated that RL can generate a structure similar to Celecoxib even when no Celecoxib was included in the training set. De novo drug design has so far only focused on generating structures that satisfy one of the several required criteria when used as a drug. Stahl et al. [102] proposed a fragment-based RL approach employing an actor-critic model for generating more than 90% valid molecules while optimizing multiple properties. Genetic algorithms (GAs) have also been used for generating molecules while optimizing their properties [103–106]. GA-based models suffer from stagnation while being trapped in at the regions of local optima [107]. One notable work alleviating these problems is by Nigam et al. [56], where they hybridize a GA and a deep neural network to generate diverse molecules while outperforming related models in optimization.

All of the generative models discussed above generate molecules in the form of 2D graphs or SMILES strings. Models to generate molecules directly in the form of 3D coordinates have also recently gained attention [57,108,109]. Such generated 3D coordinates can be directly used for further simulation using quantum mechanics or by using docking methods. One of such first models is proposed by Niklas et al. [57], where they generate the 3D coordinates of small molecules with light atoms (H, C, N, O, F). They then use the 3D coordinates of the molecules to learn the representation to map it to a space, which is then used to generate 3D coordinates of the novel molecules. Building on this for a drug discovery application, we recently proposed a model [69] to generate 3D coordinates of molecules while always preserving the desired scaffolds, as depicted in Figure 5. This approach has generated synthesizable drug-like molecules that show a high docking score against the target protein. Other scaffold-based models to generate molecules in the form of 2D graphs/SMILES strings are also published in the literature [110–114].

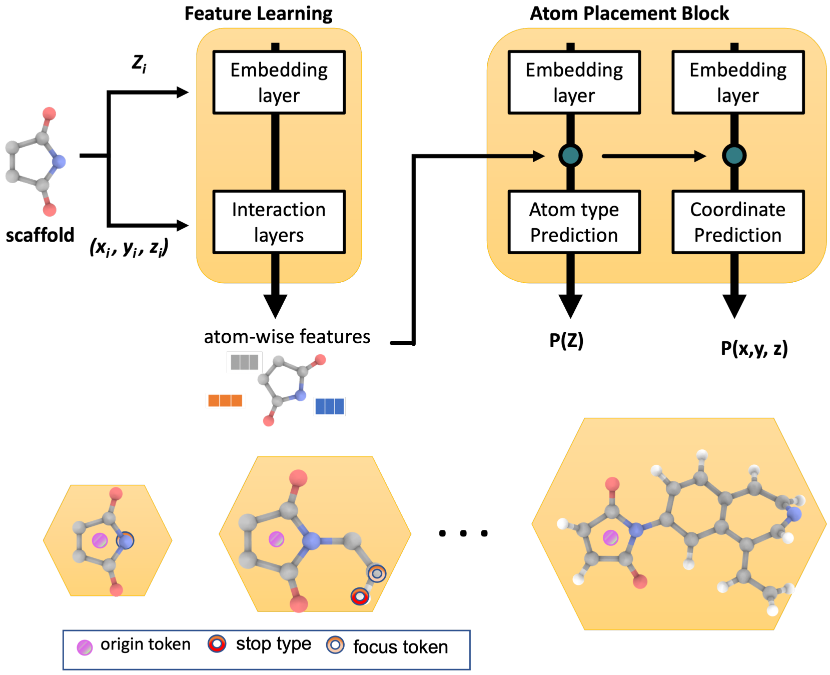

Figure 5. Generative model such as 3D-scaffold [69] can be used to inverse design novel candidates with desired target properties starting from core scaffold or functional group.

Recently, with the huge interest in the development of architecture and algorithms required for quantum computing, quantum version of generative models such as the quantum auto-encoder [115] and quantum GANs [116] have been proposed, which carry huge potential, among others, for drug discovery. The preliminary proof of concept work of Romero et al. [115,116] shows that it is possible to encode and decode molecular information using a quantum encoder, demonstrating generative modeling is possible with quantum VAEs, and more work, especially in the development of supporting hardware architecture, is required in this direction.

2.6. Protein Target Specific Molecular Design

The efficacy and potency of generated molecules against a target protein should be examined by predicting protein–ligand interactions (PLIs) and estimating key biophysical parameters. Figure 6 shows some of the computational methods frequently used in the literature (independently or together) for PLI prediction. Computationally, high throughput docking simulations [117–119] are most efficient and are used to numerically quantify and rank the interaction between the protein and ligand in terms of a docking score. These scores are based on the binding affinity of the ligand with the protein target and are used as the primary filter to narrow down high-impact candidates before performing more expensive simulations. Docking simulations are commonly used in combination with more accurate approaches to avoid false positives for pose prediction. Molecular mechanics (MM) simulations are another popular choice [120] but lack the accuracy that is generally required for making concrete decisions. Recently, all atoms molecular dynamics (MD) and hybrid QM/MM approach are increasingly adopted for studying protein–ligand interactions. It considers QM calculations for simulating the ligands and vicinity of protein where it docks while uses MM for simulating the rest of protein structure, providing improved accuracy over classical MM/docking simulations. Performing QM simulation even only for ligands and protein vicinity is computationally very expensive compared to relatively quick docking simulations. To expedite, QM simulations for ligands/protein vicinity can be replaced with state-of-art ML-based predictive model that has recently achieved chemical accuracy in predicting several properties of small molecules.

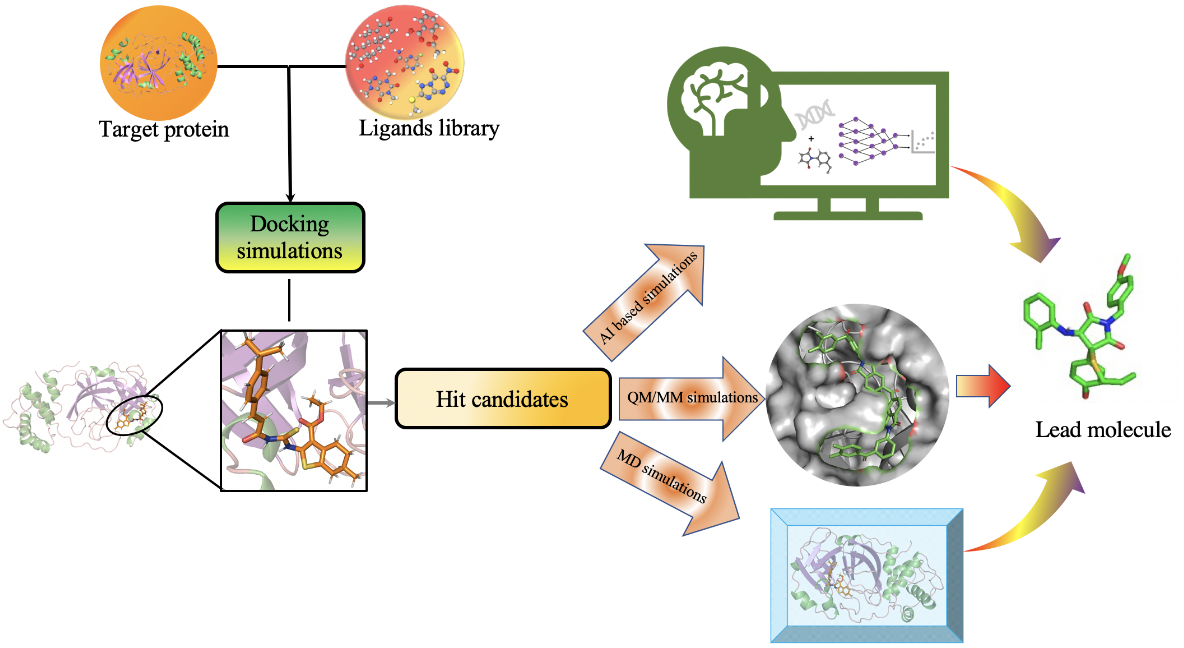

Figure 6. Molecular modeling methods used to study protein–ligand interactions including molecular docking simulations, molecular mechanics methods, hybrid Quantum Mechanics/Molecular Mechanics simulations, and deep learning models for the activity and affinity prediction.

In this regards, several deep learning architectures have been used for efficient and accurate predictions of PLI parameters. These models vary among each other depending upon how protein or ligands are represented within the model [121–124]. For instance, Karimi et al. [125] proposed a semi-supervised deep learning model for predicting binding affinity by integrating RNN and CNN, wherein proteins are represented by an amino acid sequence and ligands in the form of SMILES strings. Other studies have used graph representations of ligand molecules with a string-based sequence representation of proteins [126,127]. Recently, Lim et al. [128] used a distance-aware GNN that incorporates 3D coordinates of both ligands and protein structures to study PLI outperforming existing models for pose prediction. The development and deployment of robust and accurate PLI models within a closed loop should be conducted in a way that encodes 3D coordinates of both protein and generated ligand molecules while simultaneously including and differentiating each ligand–residue interaction. This is important for accurately predicting the desired PLI interactions and biophysical parameters while designing high throughput novel molecules. It will contribute to efficiently narrow down the candidates during lead optimization, which ultimately will be subjected to further experimental characterization before it can be used for pre-clinical studies

3. Conclusions and Future Perspectives

The success of current ML approaches depends on how accurately we can represent a chemical structure for a given model. Finding a robust, transferable, interpretable, and easy-to-obtain representation that obeys the physics and fundamental chemistry of the molecules that work for all different kinds of applications is a critical task. If such a spatial representation is available, it would save lot of resources while increasing the accuracy and flexibility of molecular representations. Efficiently using such representations with robust and reproducible ML architectures will provide a predictive modeling engine that would be ethically sourced with molecules metadata. Once a desired accuracy for diverse molecular systems for a given property prediction is achieved, it can routinely be used as an alternative to expensive QM-based simulations or experiments. In the chemical and biological sciences, a major bottleneck for deploying ML models is the lack of sufficiently curated data under similar conditions that is required for training the models. Finding architecture that works consistently well enough for a relatively small amount of data is equally important. Strategies such as active learning (AL) and transfer learning (TL) are ideal for such scenarios to tackle problems [129–133]. Graph-based methods for end-to-end feature learning and predictive modeling have been successfully used on small molecules consisting of lighter atoms. For larger molecules, robust representation learning and molecule generation parts must include non-local interactions, such as Van der Waals and H-bonding, while building predictive and generative models.

Equally important is developing and tying a robust, transferable, and scalable state-of-the-art platform for inverse molecular design in a closed loop with a predictive modeling engine to accelerate the therapeutic design, ultimately reducing the cost and time required for drug discovery. Many of the ML models used for inverse design use single biochemical activity as the criteria to measure the success of a generated candidate therapeutic, which is in contrast to a real clinical trial, where small-molecule therapeutics are optimized for several bio-activities simultaneously, leading to multi-objective optimization. Our contribution serves as inspiration to develop a CAMD workflow that should be engineered in a way to optimize multiple objective functions while generating and validating therapeutic molecules. Validation of all the newly generated lead molecules for a given target or disease-based models, if characterized by experiments or quantum mechanical simulations, is an very expensive task. We need to find ways to auto-validate molecules (using an inbuilt robust predictive model), which would be ideal to save resources and expedite molecular design. In addition, CAMD workflows should be able to quantify the uncertainty associated with it using statistical measures. For an ideal case, such uncertainty should decrease over the time as it learns from its own experience and reason in series of closed-loop experiments.

Currently, CAMD workflows are generally built and trained with a specific goal in mind. Such workflows need to be re-configured and re-trained to work for different objectives in therapeutic design and discovery. Designing and engineering a single automated CAMD setup for multiple experiments (multi-parameter optimization) through transfer learning is a challenging task, which can hopefully be improved based on the scalable computing infrastructure, algorithm, and more domain-specific knowledge. It would be particularly very helpful for the domains where a relatively small amount of data exist. Having such a CAMD infrastructure, algorithm and software stack would speedup end-to-end antiviral lead design and optimization for any future pandemics, such as COVID-19.

Author Contributions: Conceptualization, N.K.; methodology, N.K. and R.P.J.; software, N.K. and R.P.J.; validation, N.K. and R.P.J.; formal analysis, R.P.J.; investigation, N.K. and R.P.J.; resources, N.K. and R.P.J.; data curation, N.K. and R.P.J.; writing—original draft preparation, R.P.J.; writing—review and editing, N.K. and R.P.J.; visualization, N.K.; supervision, N.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on responses to COVID-19, with funding provided by the Coronavirus CARES Act. Additional support was provided by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on responses to COVID-19, with funding provided by the Coronavirus CARES Act. Additional support was provided by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830. Computing resources was supported by the Intramural program at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL; grid.436923.9), a DOE office of science user facility sponsored by the Office of Biological and Environmental Research and operated under Contract No. DE-AC05-76RL01830. We thank Andrew McNaughton and Carter Knutson at PNNL for discussion on artificial intelligence, physics informed modeling, and providing the feedback on topic areas to complete the review.

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

Sample Availability: Not available.

References

1. Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov, M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040. [CrossRef] [PubMed]
2. Kackar, R.N. Off-Line Quality Control, Parameter Design, and the Taguchi Method. *J. Qual. Technol.* **1985**, *17*, 176–188. [CrossRef]
3. Kim, E.; Huang, K.; Tomala, A.; Matthews, S.; Strubell, E.; Saunders, A.; McCallum, A.; Olivetti, E. Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **2017**, *4*, 1–9. [CrossRef]
4. Leelananda, S.P.; Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **2016**, *12*, 2694–2718. [CrossRef]
5. DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **2016**, *47*, 20–33.
6. Murcko, M.A. Envisioning the future: medicine in the year 2050. *Disruptive Sci. Technol.* **2012**, *1*, 89–99. [CrossRef]
7. Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; Persinger, C.C.; Munos, B.H.; Lindborg, S.R.; Schacht, A.L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214. [PubMed]
8. Nicolaou, C.A.; Humblet, C.; Hu, H.; Martin, E.M.; Dorsey, F.C.; Castle, T.M.; Burton, K.I.; Hu, H.; Hendle, J.; Hickey, M.J.; et al. Idea2Data: Toward a new paradigm for drug discovery. *ACS Med. Chem. Lett.* **2019**, *10*, 278–286. [CrossRef]
9. Vidler, L.R.; Baumgartner, M.P. Creating a virtual assistant for medicinal chemistry. *ACS Med. Chem. Lett.* **2019**, *10*, 1051–1055. [CrossRef]

10. Struble, T.J.; Alvarez, J.C.; Brown, S.P.; Chytil, M.; Cisar, J.; DesJarlais, R.L.; Engkvist, O.; Frank, S.A.; Greve, D.R.; Griffin, D.J.; et al. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63*, 8667–8682. [CrossRef]
11. Godfrey, A.G.; Masquelin, T.; Hemmerle, H. A remote-controlled adaptive medchem lab: An innovative approach to enable drug discovery in the 21st Century. *Drug Discov. Today* **2013**, *18*, 795–802. [CrossRef]
12. Farrant, E. Automation of Synthesis in Medicinal Chemistry: Progress and Challenges. *ACS Med. Chem. Lett.* **2020**, *11*, 1506–1513. [CrossRef]
13. Winicov, H.; Schainbaum, J.; Buckley, J.; Longino, G.; Hill, J.; Berkoff, C. Chemical process optimization by computer—A self-directed chemical synthesis system. *Anal. Chim. Acta* **1978**, *103*, 469–476. [CrossRef]
14. Marklund, E.; Degiacomi, M.; Robinson, C.; Baldwin, A.; Benesch, J. Collision Cross Sections for Structural Proteomics. *Structure* **2015**, *23*, 791–799. [CrossRef] [PubMed]
15. Li, Y.; Kuhn, M.; Gavin, A.C.; Bork, P. Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. *Bioinformatics* **2019**, *36*, 1213–1218. [CrossRef]
16. Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871. [CrossRef]
17. Kohn, W.; Sham, L.J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138. [CrossRef]
18. Jain, A.; Hautier, G.; Moore, C.J.; Ong, S.P.; Fischer, C.C.; Mueller, T.; Persson, K.A.; Ceder, G. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295–2310. [CrossRef]
19. Qu, X.; Jain, A.; Rajput, N.N.; Cheng, L.; Zhang, Y.; Ong, S.P.; Brafman, M.; Maginn, E.; Curtiss, L.A.; Persson, K.A. The Electrolyte Genome project: A big data approach in battery materials discovery. *Comput. Mater. Sci.* **2015**, *103*, 56–67. [CrossRef]
20. Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F.R.; Miller, T.F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, 124111. [CrossRef] [PubMed]
21. Lee, S.J.R.; Husch, T.; Ding, F.; Miller, T.F. Analytical Gradients for Molecular-Orbital-Based Machine Learning. *arXiv* **2020**, arXiv:2012.08899.
22. Dral, P.O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347. [CrossRef] [PubMed]
23. Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M.E.; Müller, K.R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223. [CrossRef]
24. Joshi, R.P.; McNaughton, A.; Thomas, D.G.; Henry, C.S.; Canon, S.R.; McCue, L.A.; Kumar, N. Quantum Mechanical Methods Predict Accurate Thermodynamics of Biochemical Reactions. *ACS Omega* **2021**, *6*, 9948–9959. [CrossRef]
25. Ramakrishnan, R.; Dral, P.O.; Rupp, M.; von Lilienfeld, O.A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022. [CrossRef]
26. Ruddigkeit, L.; van Deursen, R.; Blum, L.C.; Reymond, J.L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875. [CrossRef]
27. Shen, V.; Siderius, D.; Krekelberg, W.; Mountain, R.D.; Hatch, H.W. *NIST Standard Reference Simulation Website, NIST Standard Reference Database Number 173*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2017.
28. Seaver, S.M.; Liu, F.; Zhang, Q.; Jeffryes, J.; Faria, J.P.; Edirisinghe, J.N.; Mundy, M.; Chia, N.; Noor, E.; Beber, M.E.; et al. The ModelSEED Biochemistry Database for the Integration of Metabolic Annotations and the Reconstruction, Comparison and Analysis of Metabolic Models for Plants, Fungi and Microbes. *Nucleic Acids Res.* **2021**, *49*, D575–D588. [CrossRef] [PubMed]
29. Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **2019**, *6*, 1–11. [CrossRef]
30. Zheng, S.; Dharssi, S.; Wu, M.; Li, J.; Lu, Z. Text Mining for Drug Discovery. In *Bioinformatics and Drug Discovery*; Larson, R.S., Oprea, T.I., Eds.; Springer: New York, NY, USA, 2019; pp. 231–252.
31. Singhal, A.; Simmons, M.; Lu, Z. Text mining for precision medicine: Automating disease-mutation relationship extraction from biomedical literature. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 766–772. [CrossRef]
32. Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117*, 7673–7761. [CrossRef]
33. Huang, B.; von Lilienfeld, O.A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102. [CrossRef]
34. Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572. [CrossRef]
35. Elton, D.C.; Boukouvalas, Z.; Fuge, M.D.; Chung, P.W. Deep learning for molecular design—A review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849. [CrossRef]
36. Bjerrum, E.J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv* **2017**, arXiv:1703.07076.
37. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural Message Passing for Quantum Chemistry. *arXiv* **2017**, arXiv:1704.01212.
38. Hamilton, W.L.; Ying, R.; Leskovec, J. Representation Learning on Graphs: Methods and Applications. *arXiv* **2017**, arXiv:1709.05584.

39. 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]
40. Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A.S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530. [CrossRef]
41. Rupp, M.; Tkatchenko, A.; Muller, K.R.; Von Lilienfeld, O.A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301. [CrossRef] [PubMed]
42. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O.A.; Muller, K.R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331. [CrossRef] [PubMed]
43. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
44. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI—The worldwide chemical structure identifier standard. J. cheminform. *J. Cheminform.* **2013**, *5*, 1–9. [CrossRef] [PubMed]
45. Grethe, G.; Goodman, J.; Allen, C. International chemical identifier for chemical reactions. *J. Cheminform.* **2013**, *5*, O16. [CrossRef]
46. Elton, D.C.; Boukouvalas, Z.; Butrico, M.S.; Fuge, M.D.; Chung, P.W. Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* **2018**, *8*, 9059. [CrossRef]
47. Landrum, G. RDKit: Open-Source Cheminformatics Software. 2016. Available online: http://rdkit.org/ (accessed on 20 December 2020).
48. Cxcalc, ChemAxon. Available online: https://www.chemaxon.com (accessed on 20 December 2020).
49. Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024. [CrossRef]
50. Available online: https://aspuru.substack.com/p/molecular-graph-representations-and (accessed on 20 December 2020).
51. Koichi, S.; Iwata, S.; Uno, T.; Koshino, H.; Satoh, H. Algorithm for advanced canonical coding of planar chemical structures that considers stereochemical and symmetric information. *J. Chem. Inf. Model.* **2007**, *47*, 1734–1746. [CrossRef] [PubMed]
52. O'Boyle, N.M. Towards a Universal SMILES representation—A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **2012**, *4*, 1–14. [CrossRef]
53. Daylight Chemical Information Systems Inc. Available online: http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed on 20 December 2020).
54. O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *Chemrxiv* **2018**, 1–9. [CrossRef]
55. Maragakis, P.; Nisonoff, H.; Cole, B.; Shaw, D.E. A Deep-Learning View of Chemical Space Designed to Facilitate Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 4487–4496. [CrossRef]
56. Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. *arXiv* **2020**, arXiv:1909.11655.
57. Gebauer, N.; Gastegger, M.; Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; Volume 32, pp. 7566–7578.
58. Schütt, K.T.; Kessel, P.; Gastegger, M.; Nicoli, K.A.; Tkatchenko, A.; Müller, K.R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448–455. [CrossRef]
59. Minnich, A.J.; McLoughlin, K.; Tse, M.; Deng, J.; Weber, A.; Murad, N.; Madej, B.D.; Ramsundar, B.; Rush, T.; Calad-Thomson, S.; et al. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 1955–1968. [CrossRef]
60. St. John, P.C.; Phillips, C.; Kemper, T.W.; Wilson, A.N.; Guan, Y.; Crowley, M.F.; Nimlos, M.R.; Larsen, R.E. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **2019**, *150*, 234111. [CrossRef] [PubMed]
61. 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]
62. Göller, A.H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's in silico ADMET platform: A journey of machine learning over the past two decades. *Drug Discov. Today* **2020**, *25*, 1702–1709. [CrossRef] [PubMed]
63. Schütt, K.T.; Arbabzadah, F.; Chmiela, S.; Müller, K.R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890. [CrossRef]
64. Schütt, K.T.; Sauceda, H.E.; Kindermans, P.J.; Tkatchenko, A.; Müller, K.R. SchNet—A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722. [CrossRef] [PubMed]
65. Schütt, K.; Kindermans, P.J.; Sauceda Felix, H.E.; Chmiela, S.; Tkatchenko, A.; Müller, K.R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2017; pp.991–1001.
66. Axelrod, S.; Gomez-Bombarelli, R. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. *arXiv* **2020**, arXiv:2006.05531.

67. Yue, S.; Muniz, M.C.; Calegari Andrade, M.F.; Zhang, L.; Car, R.; Panagiotopoulos, A.Z. When do short-range atomistic machine-learning models fall short? *J. Chem. Phys.* **2021**, *154*, 034111. [CrossRef]
68. Matlock, M.K.; Dang, N.L.; Swamidass, S.J. Learning a Local-Variable Model of Aromatic and Conjugated Systems. *ACS Cent. Sci.* **2018**, *4*, 52–62. [CrossRef]
69. Joshi, R.P.; Gebauer, N.W.A.; Bontha, M.; Khazaieli, M.; James, R.M.; Brown, J.B.; Kumar, N. 3D-Scaffold: A Deep Learning Framework to Generate 3D Coordinates of Drug-like Molecules with Desired Scaffolds. *J. Phys. Chem. B* **2021**. [CrossRef] [PubMed]
70. Gertrudes, J.; Maltarollo, V.; Silva, R.; Oliveira, P.; Honorio, K.; da Silva, A. Machine Learning Techniques and Drug Design. *Curr. Med. Chem.* **2012**, *19*, 4289–4297. [CrossRef] [PubMed]
71. Talevi, A.; Morales, J.F.; Hather, G.; Podichetty, J.T.; Kim, S.; Bloomingdale, P.C.; Kim, S.; Burton, J.; Brown, J.D.; Winterstein, A.G.; et al. Machine Learning in Drug Discovery and Development Part 1: A Primer. *CPT Pharmacomet. Syst. Pharmacol.* **2020**, *9*, 129–142. [CrossRef]
72. Lo, Y.C.; Rensi, S.E.; Torng, W.; Altman, R.B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546. [CrossRef]
73. Agarwal, S.; Dugar, D.; Sengupta, S. Ranking Chemical Structures for Drug Discovery: A New Machine Learning Approach. *J. Chem. Inf. Model.* **2010**, *50*, 716–731. [CrossRef]
74. Rodrigues, T.; Bernardes, G.J. Machine learning for target discovery in drug development. *Curr. Opin. Chem. Biol.* **2020**, *56*, 16–22. [CrossRef]
75. Gao, D.; Chen, Q.; Zeng, Y.; Jiang, M.; Zhang, Y. Applications of Machine Learning in Drug Target Discovery. *Curr. Drug Metab.* **2020**, *21*, 790–803. [CrossRef]
76. Dahal, K.; Gautam, Y. Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. *Open J. Stat.* **2020**, *10*, 694–705. [CrossRef]
77. Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv* **2015**, arXiv:1509.09292.
78. Faber, F.A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S.S.; Dahl, G.E.; Vinyals, O.; Kearnes, S.; Riley, P.F.; von Lilienfeld, O.A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264. [CrossRef]
79. Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B.G. Benchmarking graph neural networks for materials chemistry. *NPJ Comput. Mater.* **2021**, *7*, 84. [CrossRef]
80. Jørgensen, P.; Jacobsen, K.; Schmidt, M. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. In Proceedings of the 32nd Conference on Neural Information Processing Systems, NIPS 2018, Montréal, QC, Canada, 3–8 December 2018.
81. Polishchuk, P.G.; Madzhidov, T.I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput. Aided Mol. Des.* **2013**, *27*, 675–679. [CrossRef]
82. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; et al. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2015**, *44*, D1202–D1213. [CrossRef]
83. Coley, C.W. Defining and Exploring Chemical Spaces. *Trends Chem.* **2021**, *3*, 133–145. [CrossRef]
84. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2018**, *2*, 1–16. [CrossRef]
85. Kuhn, C.; Beratan, D.N. Inverse strategies for molecular design. *J. Phys. Chem.* **1996**, *100*, 10595–10599. [CrossRef]
86. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-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] [PubMed]
87. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2015**, arXiv:1409.2329.
88. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
89. Kusner, M.J.; Paige, B.; Hernández-Lobato, J.M. Grammar Variational Autoencoder. In Proceedings of the 34th International Conference on Machine Learning—Volume 70, ICML'17, Sydney, Australia, 6–11 August 2017; pp. 1945–1954.
90. Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A.L. Constrained Graph Variational Autoencoders for Molecule Design. *arXiv* **2018**, arXiv:1805.09076.
91. Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. Learning Multimodal Graph-to-Graph Translation for Molecular Optimization. *arXiv* **2018**, arXiv:1812.01070.
92. Jin, W.; Barzilay, R.; Jaakkola, T.S. Multi-Resolution Autoregressive Graph-to-Graph Translation for Molecules. *Chemrxiv* **2019**, 1–13. [CrossRef]
93. Bian, Y.; Wang, J.; Jun, J.J.; Xie, X.Q. Deep Convolutional Generative Adversarial Network (dcGAN) Models for Screening and Design of Small Molecules Targeting Cannabinoid Receptors. *Mol. Pharm.* **2019**, *16*, 4451–4460. [CrossRef]
94. Cao, N.D.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* **2018**, arXiv:1805.11973.
95. 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] [PubMed]

96. 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.
97. Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G.L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). *Chemrxiv* **2017**, 1–18. [CrossRef]
98. Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204. [CrossRef]
99. 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. Comm.* **2020**, *11*, 1–10. [CrossRef]
100. Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, 7885. [CrossRef]
101. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9*, 1758–2946. [CrossRef]
102. Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59*, 3166–3176. [CrossRef]
103. O'Boyle, N.M.; Campbell, C.M.; Hutchison, G.R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115*, 16200–16210. [CrossRef]
104. Virshup, A.M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D.N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303. [CrossRef]
105. Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D.N. Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe. *J. Chem. Inf. Model.* **2015**, *55*, 529–537. [CrossRef]
106. Jensen, J.H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572. [CrossRef]
107. Paszkowicz, W. Properties of a genetic algorithm equipped with a dynamic penalty function. *Comput. Mater. Sci.* **2009**, *45*, 77–83. [CrossRef]
108. Simm, G.N.C.; Pinsler, R.; Csányi, G.; Hernández-Lobato, J.M. Symmetry-Aware Actor-Critic for 3D Molecular Design. *arXiv* **2020**, arXiv:2011.12747.
109. Simm, G.N.C.; Pinsler, R.; Hernández-Lobato, J.M. Reinforcement Learning for Molecular Design Guided by Quantum Mechanics. *arXiv* **2020**, arXiv:2002.07717.
110. Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *J. Chem. Inf. Model.* **2020**, *60*, 77–91. [CrossRef]
111. Lim, J.; Hwang, S.Y.; Moon, S.; Kim, S.; Kim, W.Y. Scaffold-based molecular design with a graph generative model. *Chem. Sci.* **2020**, *11*, 1153–1164. [CrossRef]
112. 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]
113. Zhang, K.Y.J.; Milburn, M.V.; Artis, D.R., Scaffold-Based Drug Discovery. In *Structure-Based Drug Discovery*; Springer: Dordrecht, The Netherlands, 2007; pp. 129–153.
114. Scott, O.B.; Edith Chan, A.W. ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. *Bioinformatics* **2020**, *36*, 3930–3931. [CrossRef]
115. Romero, J.; Olson, J.P.; Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci. Technol.* **2017**, *2*, 045001. [CrossRef]
116. Allcock, J.; Zhang, S. Quantum machine learning. *Natl. Sci. Rev.* **2018**, *6*, 26–28. [CrossRef] [PubMed]
117. Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.T.; Ban, F.; Norinder, U.; Gleave, M.E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6*, 939–949. [CrossRef]
118. Stein, R.M.; Kang, H.J.; McCorvy, J.D.; Glatfelter, G.C.; Jones, A.J.; Che, T.; Slocum, S.; Huang, X.P.; Savych, O.; Moroz, Y.S.; et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **2020**, *579*, 609–614. [CrossRef]
119. Lyu, J.; Wang, S.; Balius, T.E.; Singh, I.; Levit, A.; Moroz, Y.S.; O'Meara, M.J.; Che, T.; Algaa, E.; Tolmachova, K.; et al. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566*, 224–229. [CrossRef] [PubMed]
120. Guterres, H.; Im, W. Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations. *J. Chem. Inform. Model.* **2020**, *60*, 2189–2198. [CrossRef]
121. Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv* **2015**, arXiv:1510.02855.
122. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D.R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957. [CrossRef]
123. 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]
124. Stepniewska-Dziubinska, M.M.; Zielenkiewicz, P.; Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **2018**, *34*, 3666–3674. [CrossRef]
125. Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329–3338. [CrossRef] [PubMed]

126. Li, S.; Wan, F.; Shu, H.; Jiang, T.; Zhao, D.; Zeng, J. MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities. *Cell Syst.* **2020**, *10*, 308–322.e11. [CrossRef]
127. 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 2018 International Joint Conference on Artificial Intelligence, IJCAI, Stockholm, Schweden, 13–18 July 2018; Volume 2018, pp. 3371–3377.
128. 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] [PubMed]
129. Li, B.; Rangarajan, S. Designing compact training sets for data-driven molecular property prediction through optimal exploitation and exploration. *Mol. Syst. Des. Eng.* **2019**, *4*, 1048–1057. [CrossRef]
130. Warmuth, M.K.; Rätsch, G.; Mathieson, M.; Liao, J.; Lemmen, C. Active Learning in the Drug Discovery Process. In *Advances in Neural Information Processing Systems 14*; Dietterich, T.G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; pp. 1449–1456.
131. Fusani, L.; Cabrera, A.C. Active learning strategies with COMBINE analysis: new tricks for an old dog. *J. Comput. Aided Mol. Des.* **2019**, *33*, 287–294. [CrossRef]
132. Green, D.V.S.; Pickett, S.; Luscombe, C.; Senger, S.; Marcus, D.; Meslamani, J.; Brett, D.; Powell, A.; Masson, J. BRADSHAW: A system for automated molecular design. *J. Comput. Aided Mol. Des.* **2019**, *34*, 747–765. [CrossRef]
133. Zhang, W.; Li, R.; Zeng, T.; Sun, Q.; Kumar, S.; Ye, J.; Ji, S. Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. *IEEE Trans. Big Data* **2016**, *6*, 322–333. [CrossRef]