




Review

# Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges

Guang Chen <sup>1</sup>, Zhiqiang Shen <sup>1</sup>, Akshay Iyer <sup>2</sup>, Umar Farooq Ghumman <sup>2</sup>, Shan Tang <sup>3</sup>, Jinbo Bi <sup>4</sup>, Wei Chen <sup>2,\*</sup> and Ying Li <sup>1,5,\*</sup>

<sup>1</sup> Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269, USA; guang.chen@uconn.edu (G.C.); zhiqiang.shen@uconn.edu (Z.S.)

<sup>2</sup> Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA; akshayiyer2021@u.northwestern.edu (A.I.); UmarGhumman2018@u.northwestern.edu (U.F.G.)

<sup>3</sup> State Key Laboratory of Structural Analysis for Industrial Equipment, Department of Engineering Mechanics, and International Research Center for Computational Mechanics, Dalian University of Technology, Dalian 116023, China; shantang@dlut.edu.cn

<sup>4</sup> Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA; jinbo.bi@uconn.edu

<sup>5</sup> Polymer Program, Institute of Materials Science, University of Connecticut, Storrs, CT 06269, USA

\* Correspondence: weichen@northwestern.edu (W.C.); yingli@engr.uconn.edu (Y.L.)

Received: 1 December 2019; Accepted: 2 January 2020; Published: 8 January 2020



**Abstract:** Organic molecules and polymers have a broad range of applications in biomedical, chemical, and materials science fields. Traditional design approaches for organic molecules and polymers are mainly experimentally-driven, guided by experience, intuition, and conceptual insights. Though they have been successfully applied to discover many important materials, these methods are facing significant challenges due to the tremendous demand of new materials and vast design space of organic molecules and polymers. Accelerated and inverse materials design is an ideal solution to these challenges. With advancements in high-throughput computation, artificial intelligence (especially machine learning, ML), and the growth of materials databases, ML-assisted materials design is emerging as a promising tool to flourish breakthroughs in many areas of materials science and engineering. To date, using ML-assisted approaches, the quantitative structure property/activity relation for material property prediction can be established more accurately and efficiently. In addition, materials design can be revolutionized and accelerated much faster than ever, through ML-enabled molecular generation and inverse molecular design. In this perspective, we review the recent progresses in ML-guided design of organic molecules and polymers, highlight several successful examples, and examine future opportunities in biomedical, chemical, and materials science fields. We further discuss the relevant challenges to solve in order to fully realize the potential of ML-assisted materials design for organic molecules and polymers. In particular, this study summarizes publicly available materials databases, feature representations for organic molecules, open-source tools for feature generation, methods for molecular generation, and ML models for prediction of material properties, which serve as a tutorial for researchers who have little experience with ML before and want to apply ML for various applications. Last but not least, it draws insights into the current limitations of ML-guided design of organic molecules and polymers. We anticipate that ML-assisted materials design for organic molecules and polymers will be the driving force in the near future, to meet the tremendous demand of new materials with tailored properties in different fields.

**Keywords:** de novo materials design; machine learning; data-driven algorithm; organic molecules; polymers; materials database

## 1. Introduction

Polymeric materials are ubiquitously encountered in our daily life, ranging from familiar synthetic plastics, such as polystyrene, to natural biopolymers, such as DNA and proteins. Their exceptional chemical, physical, biological and mechanical properties [1–3] provide the broad range of applications in biomedical, chemical, and materials science fields. A polymer is usually a long chain molecule with many covalently bonded organic molecules, or repeating units. The chemical and molecular structures of these repeating units can determine the properties of polymeric materials. Thus, the choice of a specific repeating unit grants the potential of inverse materials design. This concept of molecular design for organic molecules and polymers has been widely adopted in many fields, such as organic photovoltaics [4–7], polymer dielectrics [8,9], metal-organic frameworks (MOFs) [10–13], organic light-emitting diodes [14,15], high energetic materials [16–18], and design of drug-like molecules [19–21].

When searching and designing a new material with target properties, it is of great significance to explore the complex quantitative structure property/activity relation (QSPR/QSAR). Specifically, establishing the relationship between the molecular structure and material properties is the major task for design of organic molecules and polymers. However, due to enormous combinations of the repeating units in a polymer chain or of atoms in an organic molecule, the chemical space of polymers and organic molecules is usually extremely large. For example, the nearly-infinite space of drug-like molecules is between  $10^{23}$  and  $10^{60}$  [22]. Another example is the GDB-17 database, which has 166 billion molecules generated by enumeration of up to 17 atoms for an organic molecule [23]. As a result, identifying promising molecular candidates exhaustively through the chemical space is an extreme challenge if a traditional trial-and-error approach is applied, which is similar to the search of the needle in a haystack. Therefore, novel concepts of materials design and methods with effective searching capability are the keys to overcome this challenge.

The development of materials design has experienced three stages, to date. The first stage is the conventional experimentally-driven and trial-and-error materials design, guided by experience, intuition, and conceptual insights (domain knowledge). This traditional method has enjoyed success in the invention of many important drug molecules, such as penicillin [24,25]. However, this approach has unavoidable limitations. For instance, only certain macroscopic properties are available, while others can hardly be measured. Moreover, this method suffers from by-chance discovery, loss of generality, and is extremely time-, labor-, and cost-consuming. For example, it takes 13 years, on average, for discovery of a new drug molecule [26]. In the second materials design stage, thanks to the progress made in computational technologies, modeling and simulation have dominated this field for materials design. Computational methods, such as density functional theory (DFT) [27,28] and molecular dynamics (MD) [29,30], have facilitated fast materials design through high-throughput virtual screening. It is especially powerful and useful in predicting material properties when an analytic formula does not exist. For instance, in the realm of de novo drug discovery, in silico modeling to develop QSAR has been established as a plausible approach [31,32]. Yet, computer modeling still suffers from several limitations. For example, the design process is usually computationally expensive in terms of time and resources. Further, it only gives a direct mapping from a structure to its properties, while an inverse mapping from material properties to molecular structures is usually difficult to find. In the past, the inverse QSAR has been used to map a favorable region in terms of predicted activity to the corresponding molecular structures [33–35]. However, this is not a trivial problem: first the solutions of molecular descriptors corresponding to the region need to be resolved using the QSAR model, and these then need to be mapped back to the corresponding molecular structures. The fact that the molecular descriptors chosen need to be suitable both for building a forward predictive QSAR model, as well as for translation back to molecular structure, is one of the major obstacles for this type of approach.

With the growth of materials database, as well as the development of data science and artificial intelligence (AI) in general, in particular, the invention of AlphaGo [36], we are facing a new age

that has been termed the “fourth paradigm of science” [37] or the “fourth industrial revolution” [38]. This advance brings materials design into its third stage, in which data-driven methods have emerged to tackle the inverse materials design problems. In addition to experimental approaches, theoretical means, and computer simulations, data-driven materials design approaches have been considered to be the “fourth pillar” in scientific research [3]. To date, many breakthroughs and research works flourish in de novo design of organic molecules and polymers by data-driven methods. Successful frameworks include material informatics [39,40], polymer informatics [41,42], and polymer genome [43,44], to name a few. An excellent example is the development of energetic materials, which has lagged behind other materials discovery, since many of them, such as TNT and TATB, came into the market after World War II [45,46]. Using traditional methods, the most recent successful design is the invention of CL-20. However, this design case takes 30 years from its initial synthesis to be embraced in industrial application [47]. Very recently, with state-of-the-art data science techniques, the discovery of high energetic materials has been dramatically accelerated. It has been reported that, with the help of materials genome approach [48–51], a new insensitive explosive with high-energy density has been designed with much less effort [52].

Automation of organic molecules and materials design is considerably less developed than that of inorganic materials due to challenges associated with searching the vast design space defined by the almost infinite combinations of molecular constituents, microstructures and synthesis conditions [24,53,54]. Machine learning (ML), a subset of AI, has been considered as a promising method to deal with inverse molecular design. An ML-based approach can explore the underlying pattern of QSPR/QSAR in an accelerated, while efficient, manner. Because of its superior capability, the large design space can be searched exhaustively. For example, the complexity of the game Go is to the order of  $10^{140}$  possible solutions [55]. Therefore, ML-assisted approaches have great potential to overcome aforementioned challenges and provide huge opportunities in materials design [56–59]. To perform an ML-assisted materials design task, acquisition of a database containing uniformly distributed data of interests is the first step, which remarkably affects the performance of the ML models to be built. The database can be obtained from public databases, published literature, or self-built by experiments or numerical simulations. In addition to the quality of the database, data representation also plays an important role in developing the ML model. The molecular structure is encoded into the feature representation as the input of ML models to establish QSAR [60]. Afterwards, the database is then divided into a training dataset and test dataset to build and validate the ML model, respectively. With the predictive ML model at hand, promising molecules can be generated through reinforcement learning (RL) or Bayesian optimization (BO).

Although there have been a few excellent review articles on this topic [58,60–64], this review is focused on a comprehensive and broad introduction of opportunities and challenges in the ML-guided design of organic molecules and polymers in a more friendly way to beginners. Particularly, succinct summaries of material databases, methods for feature generation, and suitable ML algorithms are provided, which can serve as a tutorial for beginners. We delineate a detailed procedure of how to use ML to predict properties of a material, and more significantly, the inverse problem of designing materials that show appropriate properties in a variety of case studies. The discussion is comprehensive, covering nearly every facet of the ML implementation and inverse design. In this review, the methods of ML-assisted property prediction and material design developed in the last decade are reviewed in a chronological order. We review ML methodologies ranging from the simplest method of linear regression to state-of-the-art deep learning (DL) methods. We believe that AI or ML will revolutionize diverse scientific fields, especially for the design of organic molecules and polymers. To embrace the potential provided by this relatively new design paradigm with less effort, this work is aimed to support material researchers who want to take advantage of ML in their own research. The paper is organized as follows. Nine typical design examples using ML-assisted approaches in various fields are presented in Section 2. In particular, to differentiate the molecular and microstructure designs, examples at either the molecular level or the microstructure level for organic photovoltaics are given.

In Section 3, we reveal the open problems and challenges facing in the ML-assisted materials design and discuss possible solutions. Section 4 completes this work with a few concluding remarks. We expect ML-assisted materials design methods to play an essential role in the near future and hope that this work inspires future directions in this field.

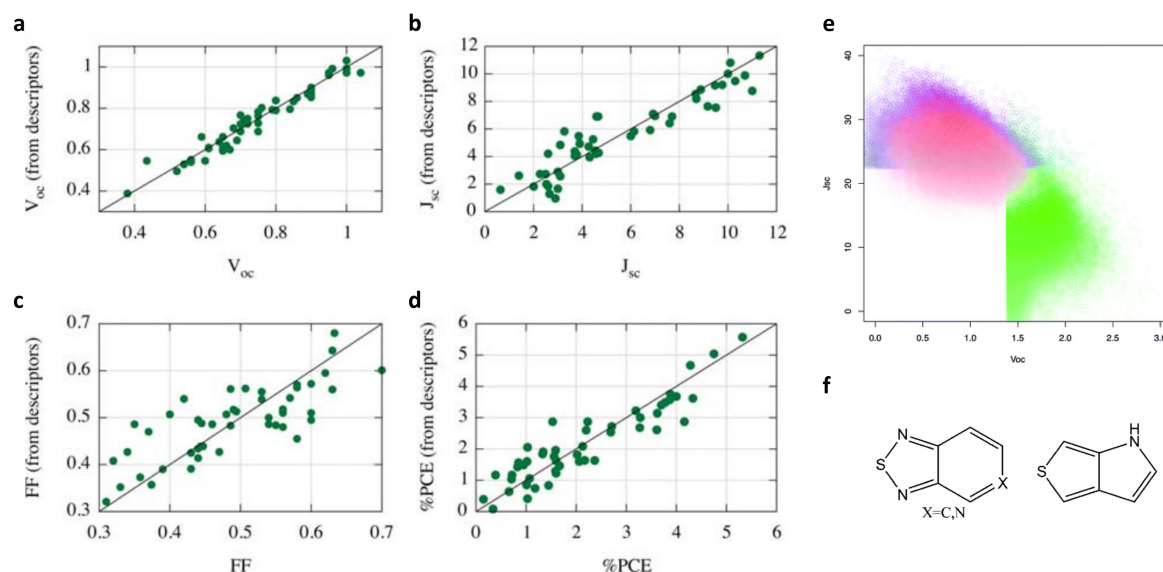
## 2. Case Studies of ML-Assisted Materials Design

### 2.1. Molecular Design of Organic Photovoltaics (OPV)

The need for clean energy is of global importance as the consumption of traditional energy especially crude oil keeps increasing [65]. Among many clean energy techniques, design of materials to utilize solar power has been recognized as one of the most promising solutions because the huge amount of solar energy available makes it an important source of electricity. Most commercially available solar materials are crystalline silicon based. However, they suffer from a few disadvantages, like high production costs [66] and low efficiency [67]. Therefore, design of alternative materials is of great importance to better utilize solar power. The organic photovoltaic (OPV) serves as a great potential candidate due to its low cost, abundance, and installation versatility [68]. A good design example of OPVs has been developed by Aspuru-Guzik's group. As a part of the Harvard Clean Energy Project (CEP) [69], in addition to the widely used quantum chemical computation, ML-guided high-throughput screening (HTS) was also proposed to accelerate the material discovery of a high efficiency bulk-heterojunction (BHJ) solar cell [70]. High power conversion efficiency (PCE) is the desired property that the authors tried to design for organic solar materials within the virtual chemical library.

The first step in an HTS approach is to get a pool of candidate molecules. The authors obtained a chemical library through combinatorial generation of 30 heterocyclic building blocks by either linking or fusing basic building blocks together. In this way, they finally got a library including 2.6 million conjugated molecules. To construct a prediction model to screen the generated molecular library, a dataset to build the ML model is needed. To this end, 50 molecules obtained from the literature were used as a training dataset. This dataset enclosed the molecular structures and the associated properties, which allowed them to develop a prediction model. The molecular structure information was represented by 33 physicochemical and topological descriptors calculated by the software ChemAxon [71]. Instead of mapping the molecular descriptors with the desired PCE property directly, intermediate properties, including the filler factor (FF), the short circuit current density ( $J_{sc}$ ), and the open circuit voltage ( $V_{oc}$ ), were adopted to develop the QSPR in an indirect way.

The multi-linear regression model was chosen to develop the QSPR. Figure 1a–d show the prediction performance of multiple linear regression for the selected properties. The R squared values are 0.96, 0.92, 0.66, and 0.89 for  $V_{oc}$ ,  $J_{sc}$ , FF, and %PCE, respectively. The correlations of  $V_{oc}$  and  $J_{sc}$  are great and acceptable for PCE. But the correlation of FF is poor. The authors tried other methods to mitigate this drawback by using different regression models on  $V_{oc}J_{sc}$  and different descriptors. However, the results showed that better fit of  $V_{oc}$  and  $J_{sc}$  than FF were observed. Therefore, the ML models developed for  $V_{oc}$ ,  $J_{sc}$ , and  $V_{oc}J_{sc}$  were chosen to screen the virtual molecular library.



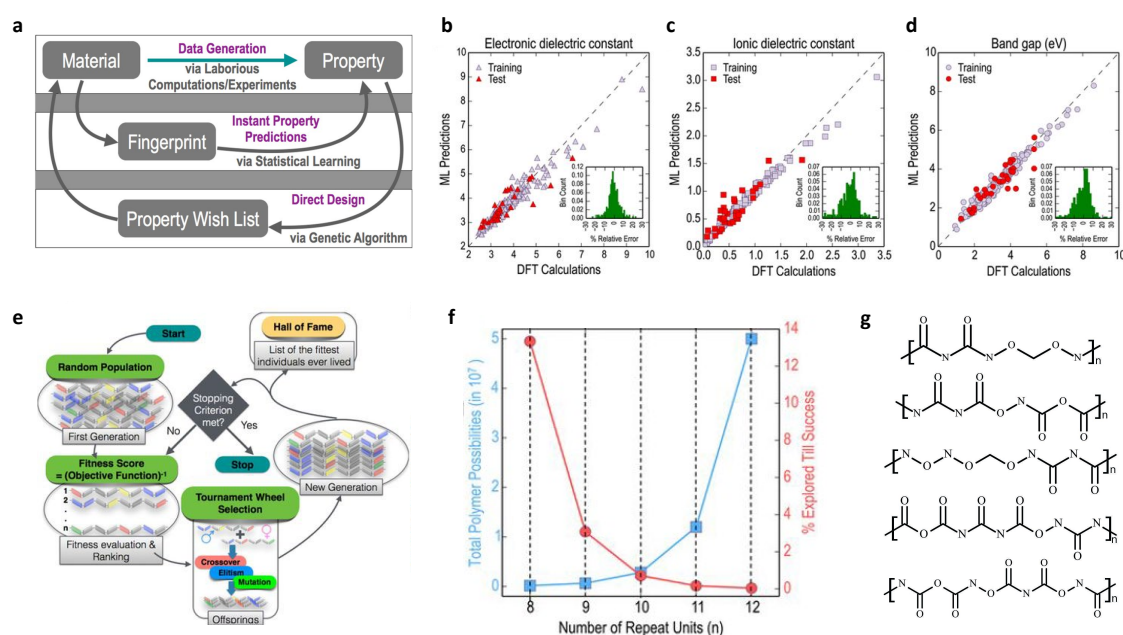
**Figure 1.** Machine learning (ML)-guided design of organic photovoltaics. (a–d) The performance of ML models on desired properties:  $V_{oc}$ ,  $J_{sc}$ , filler factor (FF), and power conversion efficiency (%PCE); (e) top 10% screened molecules with highest predicted  $V_{oc}$  (green),  $J_{sc}$  (blue), and  $V_{oc}J_{sc}$  (red); (f) the most promising building blocks screened by  $V_{oc}J_{sc}$  model. The figures are adapted from Reference [70] with permission from The Royal Society of Chemistry.

The top 10% molecules filtered out by  $V_{oc}$  ML model (green),  $J_{sc}$  ML model (blue), and  $V_{oc}J_{sc}$  ML model (red) are given in Figure 1e. The most promising candidates are located in the upper left corner of the contour plot. Based on the  $V_{oc}J_{sc}$  model, it is interesting to find that the building blocks presented in many desired molecules are benzothiadiazole or pyridinethiadiazole motif and thienopyrrole motif (see Figure 1f). The highest predicted efficiency in the library is 10.36%, which is larger than the highest one in the training dataset of 5.32%. It further verifies that the ML-guided methods can help to accelerate the discovery of materials with excellent properties.

As an early work published in 2011, the framework proposed was innovative. However, there are several aspects to improve. First, the training dataset with just 50 cases may not be large enough to accurately figure out the underlying relationships between molecular structures and properties, compared to the large space of the virtual library to screen. Second, the obtained ML models do not have a test dataset for validation, which might not guarantee the accuracy of the ML models. Last but not least, the ML algorithm employed may be too simple to capture the complicated relation. Other ML models that are good at capturing nonlinear mapping could be tested.

## 2.2. Design of Polymer Dielectrics

Polymer dielectrics have been used in various applications, such as organic field-effect transistors [72], insulators [73], energy storage [74], and capacitors [44]. To meet such a high diverse demand from industry, the design of polymer dielectrics is a big challenge. With the help of ML, as well as intellectual algorithms, unprecedented progress has been made. For example, excellent work has been done by Ramprasad's group [8,9,75]. In one of their works, state-of-the-art computational tools, ML algorithm, and genetic algorithm (GA) were combined to solve direct property prediction and inverse material design problems. Figure 2a illustrates the outline for the proposed method. Three phases are involved in this method. Data generation using first principle calculation comes in the first phase, which is followed by the construction of an ML model (second phase). After the structure–property relation is established, the third phase is the inverse design of polymers with target properties through genetic algorithm.



**Figure 2.** ML-guided design of dielectric polymers. (a) Three phases involved in this design approach; (b–d) the performance of ML model on the desired properties: electronic dielectric constant, ionic dielectric constant, and band gap; (e) the flow chart of genetic algorithm to identify promising candidates with desired properties; (f) the relation between number of building blocks and the number of possible polymers, as well as the percentage of the polymers needed to be considered; (g) the optimized molecular structures with 8~12 units (C and H are not displayed explicitly). The figures are adapted from Reference [9] with permission, copyright 2016 Springer Nature.

The training and test dataset came from first principle calculations. The candidates are chosen based on a particular chemical subspace which includes seven linearly repeated chemical building blocks (CH<sub>2</sub>, NH, CO, C<sub>6</sub>H<sub>4</sub>, C<sub>4</sub>H<sub>2</sub>S, CS, and O). These building blocks are commonly found for many polymer materials and can effectively represent electronic and dielectric properties of polymeric materials [76]. In order to obtain the dataset at a reasonable computational cost, four blocks forming a repeating unit was set for generating the candidate space. Furthermore, to avoid generating chemically invalid molecules, the authors pre-screened the data and 284 candidates were eventually determined.

The crystal structures of candidates were obtained by the minima hopping method [77], as well as density functional theory. The primary properties of dielectric polymers were calculated by DFT, including bandgap ( $E_{\text{gap}}$ ) and dielectric constant ( $\epsilon_{\text{elec}}$ ,  $\epsilon_{\text{ionic}}$ , and  $\epsilon_{\text{total}} = \epsilon_{\text{elec}} + \epsilon_{\text{ionic}}$ ). Molecular structures of these candidates were represented by fingerprints, which can be represented by a  $7 \times 1$  vector ( $M_I$ ),  $7 \times 7$  matrix ( $M_{II}$ ), and  $7 \times 7 \times 7$  matrix ( $M_{III}$ ) for individual block, block pair, and block triplet. The value in the vector or matrix denotes the occurrence frequency of corresponding building block, block pairs, or block triplet. With this representation, the chemical structure can be converted into a numerical form to build a relation between chemical structures and desired dielectric properties (high  $\epsilon_{\text{total}}$  and  $E_{\text{gap}}$ ) by a selected ML model.

The kernel ridge regression (KRR) with Gaussian kernel, an ML algorithm able to reproduce nonlinear relationships [78], was applied to develop a mapping from molecular structure to dielectric properties. The dataset was divided into 90% and 10% for training and test purposes, respectively. In the training step, the cross-validation scheme was also implemented to reduce over-fitting and ensure the generality of the model. Figure 2b–d show the performance of the prediction by the ML model on electric dielectric constant, ionic dielectric constant, and bandgap, compared by DFT calculations. We can tell that the prediction agrees well with the DFT calculation, which verifies the direct prediction model constructed. Moreover, the authors showed that the model developed for

4-block polymers could be applied to other repeated units with arbitrary number of building blocks. This impressive conclusion for extrapolation can guide inverse material design for polymers with longer repeating units.

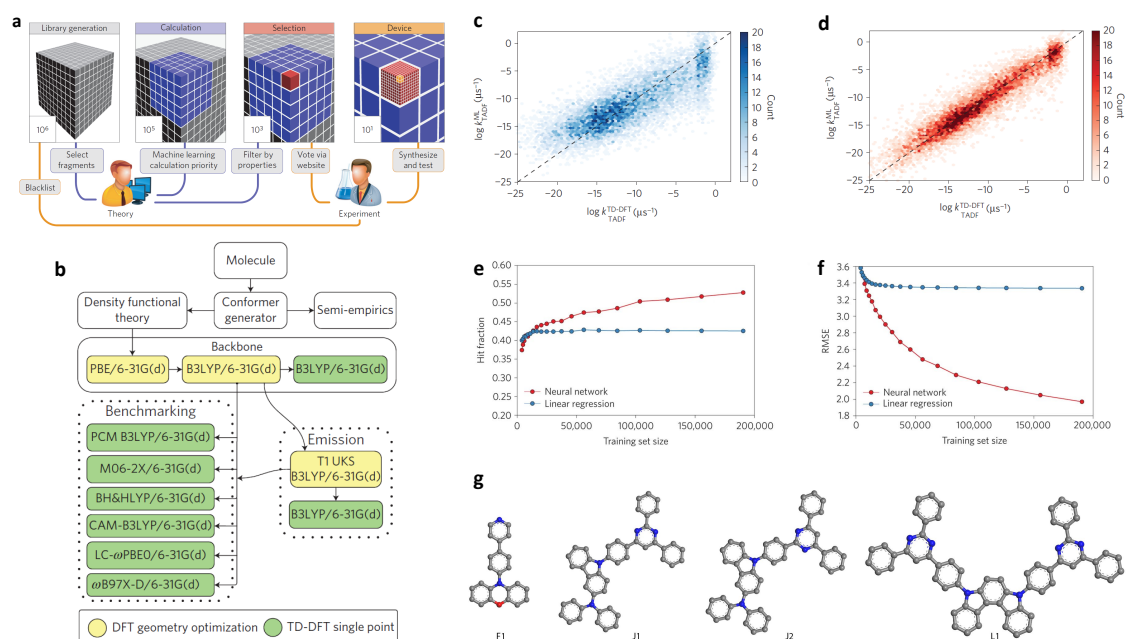
To tackle the inverse design problem, the first question to answer would be how to generate the molecules. Enumeration could be an answer. However, the possible candidates grow exponentially as the block number increases, which is intractable for a case by case approach. To this end, the genetic algorithm (GA) was applied closely cooperated with the ML model. GA is an effective algorithm to find optimum candidates using the strategy imitating biological evolution, which involves crossover, mutation, and elitism [79]. In this method, a pool of candidates (initial generation) is generated randomly; for example, 300 polymers with an 8-block unit were considered by the authors. In each generation, the population undergoes different evolution operations, and a fitness score is used to evaluate the properties for an individual polymer, which is performed by the constructed KRR model. Candidates with higher fitness scores survive to the next generation (offspring). After the prescribed number of generations is reached or optimum fitness scores are obtained, the process stops. The whole procedure is explained in Figure 2e.

Figure 2f depicts the relationship between the number of building blocks and the space of total possible candidates, as well as the percentage of favorable candidates. One can see that, as the number of repeating units increases, the number of candidates increases exponentially, while the percentage of the best candidates decreases rapidly, which means enumeration and one-by-one screening is a high cost yet ineffective approach. Figure 2g shows the best structure examples given the desired properties (target  $E_{\text{gap}} = 5\text{eV}$  and  $\epsilon_{\text{total}} = 5$ ), with the number of building blocks ranging from 8 to 12.

### 2.3. Molecular Design of Organic Light-Emitting Diodes

Organic light-emitting diodes (OLEDs) have great potential to be applied in the third-generation display devices. OLEDs have promising properties, including energy-saving and long-lasting working time [80], both of which outperform liquid crystal display (LCD) technology. Traditional fluorescent emitters allow only singlet-singlet transition, resulting in the limited harvest efficiency. Even though phosphorescent OLEDs have higher harvest efficiency, their application is restricted by high costs since a necessary ingredient (iridium) is very expensive. The thermally-activated delayed fluorescent (TADF) technique has been proposed to mitigate these limitations [81,82]. Using this method, excellent work has been done. For example, Aspuru-Guzik et al. reported that experimentally tested external quantum efficiency (EQE) were up to 22% [83]. In their work, the authors combined the quantum chemical computation, cheminformatics [84,85], machine learning, industrial expertise, synthesis, and testing to construct an effective design framework for the design of blue TADF OLEDs emitters. Figure 3a illustrates the schematic of the framework.

Utilizing their in-house code based on the RDKit package [86], a virtual chemical library of OLED molecular candidates was generated as a starting point. The generation of the library was directed by chemical intuition, quantum simulation, and experimental work. To increase the pool of the library, the authors began with a group of fragments and used combinatorial enumeration with constraints. The imposed constraints include the following considerations: the structure requirement of the TADF OLEDs molecules to be in the form of donor-(bridge)<sub>x</sub>-acceptor (where  $x$  can be 0, 1 or 2), the symmetry in molecular substitution, molecular size, vapor processing, optical properties, and synthesis accessibility. Furthermore, a blacklist of unwanted structures, such as chemically unstable molecules, was compiled to further constrain the growth of the library. Following these rules, 110 donors, 105 acceptors, and seven bridges came out. The library finally grew into a big space with 1.6 million candidate molecules.



**Figure 3.** Integrated design of organic light-emitting diodes (OLEDs). (a) Schematic of the integrated design method; (b) flow chart of quantum chemical computation; (c,d) the coefficient of determinant for linear regression (0.80) and neural network (0.94); (e,f) the relation between hit fraction and root mean square error (RMSE) with respect to the training set size; (a–f) are adapted from Reference [83] with permission, copyright 2016 Springer Nature; (g) the best candidate molecular structures (gray, blue, and red nodes) denote carbon, nitrogen, and oxygen atoms, respectively.

Direct screening of this library by quantum chemical simulation alone is intractable. Therefore, 40,000 candidates were randomly selected and evaluated by time-dependent DFT (TD-DFT). The flow chart of the calculation is shown in Figure 3b. Since the nonradiative decay rates are difficult to predict by theoretical or numerical approaches, the delayed fluorescent rate constant  $k_{\text{TADF}}$  is selected as the desired property, which was estimated through its relation to singlet-triplet gap  $\Delta E_{ST}$  and oscillator strength  $f$  [83]:

$$k_{\text{TADF}} = \alpha \frac{f}{1 + 3 \exp(\Delta E_{st}/kT)}, \quad (1)$$

where  $\alpha$  is material constant,  $k$  the Boltzmann constant, and  $T$  the absolute temperature. The two parameters  $\Delta E_{ST}$  and  $f$  were obtained through DFT quantum chemical calculation on the candidate molecules. In this fashion, the property ( $k_{\text{TADF}}$ ) of candidate molecules can be evaluated by DFT calculations. As a result, the training dataset enclosing molecular structures and respective properties is ready to build a mapping between them using ML.

To construct an ML model, the molecular structure was first represented in a simplified molecular-input line-entry system (SMILES) form [87], and then was converted to a vector with fixed length using extended connectivity fingerprints (ECFP) [88]. Note that the SMILES is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings. Then the neural network (NN) with two hidden layers was selected as the ML model, which was trained to map a molecular ECFP to its property  $k_{\text{TADF}}$ . The training objective was to minimize the root mean square error (RMSE) of predicted  $\log(k_{\text{TADF}})$ .

After this ML model was established, it was then employed to further screen the rest of the library. The candidates that gave the best predictions were moved up for TD-DFT calculation. In the meantime, the neural network model was retrained as new data were added, which further increased the predictive accuracy of the ML model. Prediction by linear regression was also carried out for comparison. It is easy to see from Figure 3c,d that the coefficient of determination ( $R^2$ ) of the test



dataset were 0.80 and 0.94 for linear regression and neural network algorithm, respectively. Compared to the linear regression, the neural-network-based ML model performs better with a large dataset. It is also found that the performance of NN model is more dependent on the training data size (Figure 3e,f); while with larger data size, the performance improvement of LR model is imperceptible.

When the library was screened to a human-tractable range, a collaborative decision-making procedure was conducted by two to six synthetic organic chemists to evaluate the molecules. Four optimum molecules (Figure 3g) were then synthesized and tested to validate the prediction by the ML model. They showed that the EQE is over 22%. It was proved that an ML-guided materials design approach for OLEDs can be really powerful.

#### 2.4. Design of Polymeric Solar Cell

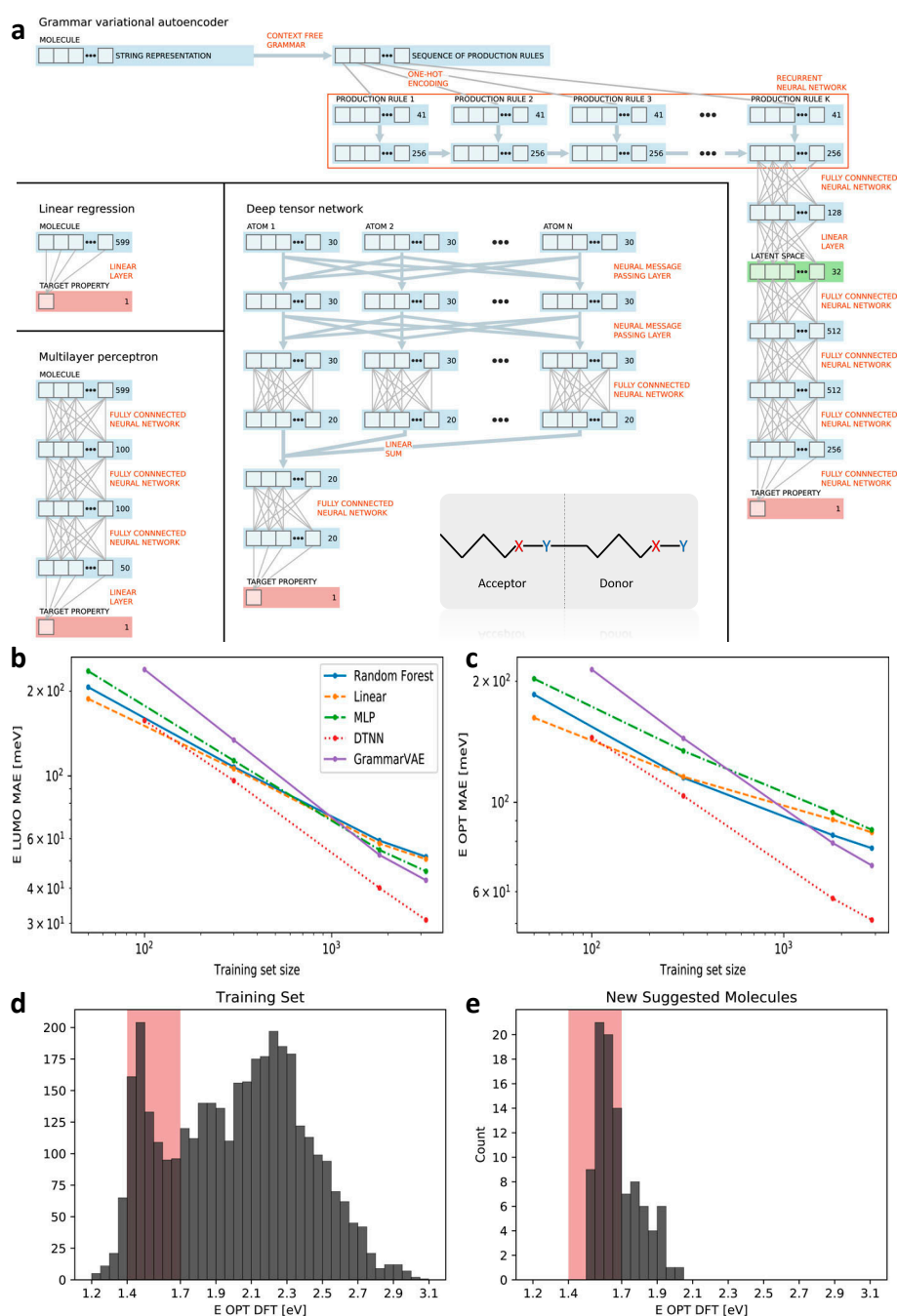
Polymeric materials are widely applied as active photoabsorbers in most organic solar cells. The donor-acceptor type of polymers is a prospective candidate for design of solar cells. By tuning the donor and acceptor units, the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) can be designed [89]. In order to capture sufficient solar energy, the ideal optical gap is to fall in the range of 1.1–1.7 eV [89]. Due to the large number of possible donor-acceptor combinations, it is suitable to use ML models to find promising solar cells. A recent work by Jorgensen et al. gives an excellent design example. In their work, they not only obtained direct property prediction models but also realized inverse materials design with desired properties by the so-called grammar variational autoencoder (GrammarVAE) [89]. Among many of the important properties of polymeric solar cells, LUMO ( $\epsilon_{\text{LUMO}}$ ) and the lowest optical transition energy ( $\epsilon_{\text{opt}}$ ) were adopted by the authors as the design targets.

The figure in the shaded box in Figure 4a illustrates a typical organic solar cell molecule considered by the authors. To build a virtual library of possible molecules, 13 acceptor, 10 donor moieties, 9 possible side-chains, and atomic substitutions were used. Combinatorial enumeration of the basic elements resulted in  $10^{14}$  monomer structures. The authors first pre-screened the library by ignoring the count number differences, ignoring relative positions of side groups, and pre-optimizing the structures. A virtual library containing 3938 monomers was then constructed. DFT calculations were conducted to get the corresponding properties to form a complete dataset.

Representation of molecules is one important element in the development of an ML-based materials design approach. The commonly utilized approaches for molecular descriptors include: Coulomb matrix [90,91], bag-of-bonds [91,92], and molecular graphs [61,63,93]. However, the authors argued that the spatial information of molecules utilized by these representations might not be available [89]. SMILES string is another commonly used molecular structure representation. Generative ML models, for example, the variational autoencoder (VAE), are usually applied to generate molecular SMILES to get a chemical library [24,54,63]. But this method still suffers from generating chemically invalid molecules. To alleviate the problems, the authors integrated a simple string featurization and GrammarVAE. By imposing chemical grammar at the decoding phase, they can prohibit generating syntactically invalid molecules.

Three different representation methods were employed for comparison, including fixed length vector representation, string representation, and XYZ-coordinate. The former two representations do not need the spatial configuration of the molecular structure, while the third representation requires spatial configurations obtained by DFT calculation. Though computationally costly, the authors intended to show how the spatial information could influence the accuracy of ML algorithms. Five types of ML models were selected to construct the structure–property relations, namely linear ridge regression (LRR), multi-layer perceptron (MLP, three hidden layers with 100, 100, and 50 units, respectively), random forest regression (RFR), the deep tensor neural network (DTNN), and the grammar variational autoencoder (GrammarVAE). To distinguish the performances of these different ML algorithms, various combinations of molecular representations and ML models were carried out. Specifically, LRR, MLP, and RFR used fixed length vector representation; DTNN adopted the

XYZ-coordinates; and GrammarVAE applied the simple string representation. The schematic of the data flow in each ML model is shown in Figure 4a.



**Figure 4.** Design framework of polymeric solar cell. (a) Data flow of four different ML models (the gray shaded box in the bottom of it is a representation of donor-acceptor structure with X and Y the side groups, the number of which are variable); (b,c) performance comparison of different ML models; (d,e) molecules distribution in training dataset and new suggested molecules for  $\epsilon_{opt}$  (shaded area stands for target property range). These figures are adapted from Reference [89] with permission, copyright 2018 AIP publishing. DFT = density functional theory.

The performances of these models are compared in Figure 4b,c. They show the performances of all the models increase as the data size increases. They also show that DTNN model behaves better than the other. The better performance of DTNNA is considered to be a result of utilizing spatial

information as input. In addition, we can see GrammarVAE gives the lowest mean average error when a large dataset is employed for model development.

GrammarVAE was then used to inversely design promising candidates with desired properties. All the data was used as a training dataset to generate SMILES strings. BO method [54] and approximated calculation of conditional probability were applied to assist generation of molecules. The top 100 candidates were selected and evaluated by DFT calculations of  $\epsilon_{opt}$ , which are shown in Figure 4d,e. The results show that the percentage of promising candidates increased from 11% (random distribution of the molecules in a wide property range) to 61% in the screened dataset (the molecules distribute close to the desired property range). It confirms that ML-based algorithms can accelerate the process of materials discovery with desired optical properties.

### 2.5. Design of High Energetic Materials

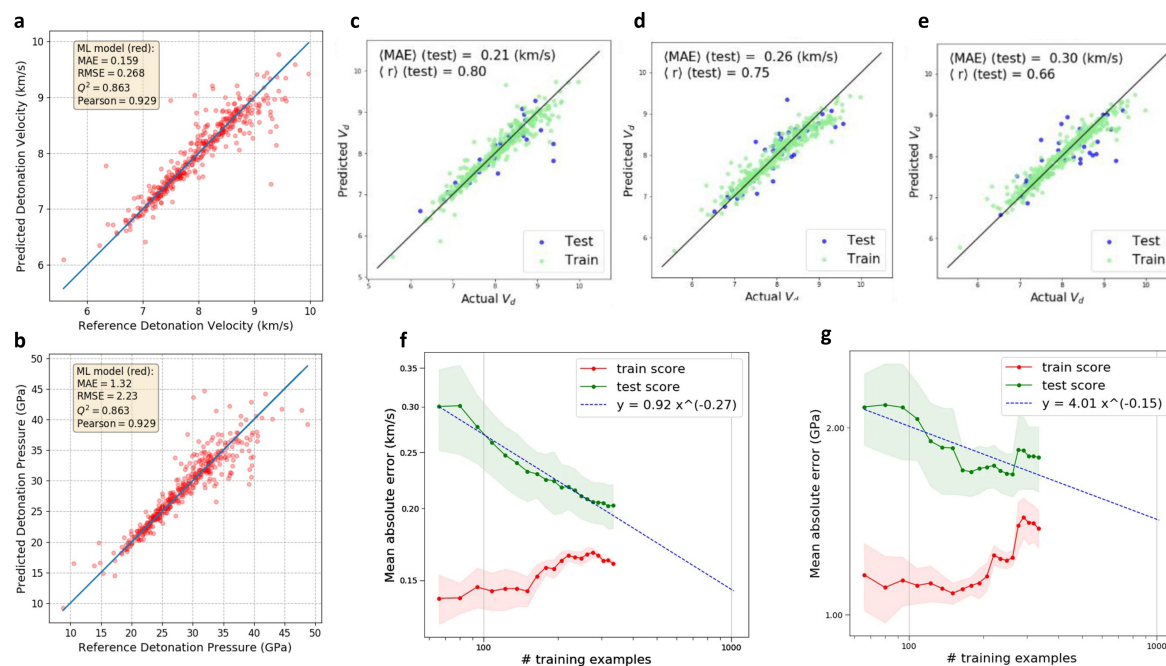
The development of energetic materials with high energy density yet low sensitivity to external stimulus has great challenges. The trade-off between high energy density and low sensitivity brings with many difficulties when designing energetic materials [52]. In addition, the limited number of available databases restricts the application of an ML-assisted design method. The data scarcity of energetic materials originates from two aspects: (1) experiments are inherently dangerous; and (2) quantum chemical calculations are prohibitively expensive. Therefore, the major challenge in ML-assisted design of energetic materials is how to develop a sound model with limited data. Recent works by Chung et al. demonstrated that good performance could still be achieved by using only 109 molecular data points [91,94]. In these works, several featurizations and various ML models were combined for dual study objectives: fast prediction of detonation properties (explosive energy, detonation pressure, and velocity) and comparison of the molecular representations and ML models.

The dataset of 109 candidates in their studies came from the literature [95], which is uniformly distributed in ten distinct compound classes and, more importantly, verified by experiments. DFT calculation and analytical approach were utilized to obtain the properties of these candidates. Nine different properties are considered, such as density, heat of formation of the solid, TNT equivalent per cubic centimeter, etc. The authors indicated that with a diverse training dataset, the model trained should be relatively generative.

To establish a reasonable ML model, it is required that the number of molecular features should be much less than the total number of molecules in the dataset. As a result, featurization is of great importance. Popular representation methods, such as SMILES, may not be suitable since these representation methods need a large dataset to train the ML model. The authors suggested that chemical intuition and domain knowledge can assist feature selection. In one earlier work [91], they chose five featurizations: custom descriptor set (CDS, a vector including 21 customized parameters, like raw counts of carbon and nitrogen), sum over bond (SoB, a vector that contains how many different bonds are presented), Coulomb matrix (CM, coordinates and nuclear charges were transformed to the Coulomb matrix eigenvalue spectra representation, which is invariant of the transformation and rotation of the molecular structure), bag of bonds (BoB, a bag containing the number of occurrence of different bonds), and fingerprinting (it transfers molecular graphs into vector form). The kernel ridge regression (KRR), ridge regression (RR), support vector regression (SVR), random forest (RF), and k-nearest neighbors (KNN) were selected as the ML models. In a later work [94], they also adopted LASSO regression, Gaussian process regression (GPR), and neural network (NN) to develop prediction models using the same dataset, plus additional 309 molecules from another reference [96].

They showed that the SoB featurization always performed the best compared to the rest in prediction of detonation properties. Among all of these ML models, KRR and RR outperformed SVR, RF, and KNN; GPR and NN outmatched LASSO regression. The performance of the neural network model for the predictions of detonation velocity and pressure are shown in Figure 5a,b. The accuracy of ML predictions by LASSO, GPR, and NN are demonstrated in Figure 5c–e, respectively. Good

performances are observed for the ML models applied. Furthermore, to study the influence of data diversity on the ML model, the authors adopted another small dataset including 25 molecules (narrow and in the same molecular class) [97] to construct a prediction model. They found that the ML model developed was not generative beyond this specific class of molecules since it did not capture the difference between various classes. Thus, when predicting properties of molecules from other classes, the model performance is poor. They emphasized the importance of the applicability domain of ML models. The learning curves of data dependent modeling have also been investigated, which are shown in Figure 5f,g. We can see that, as the data size increases, the gaps between the training and test curves decrease.



**Figure 5.** Detonation property prediction of energetic materials. (a,b) The performance of the neural network model for prediction of detonation velocity and pressure; (c–e) prediction accuracy of LASSO, Gaussian process regression (GPR), and neural network (NN), respectively; (f,g) left: learning curves of ML model for detonation energy; right: detonation pressure. (a–e) are adapted from Reference [94] with permission. (f,g) are adapted from Reference [91] with permission, copyright 2018 Springer Nature.

## 2.6. Design of Polyimides with High Refractive Index (RI)

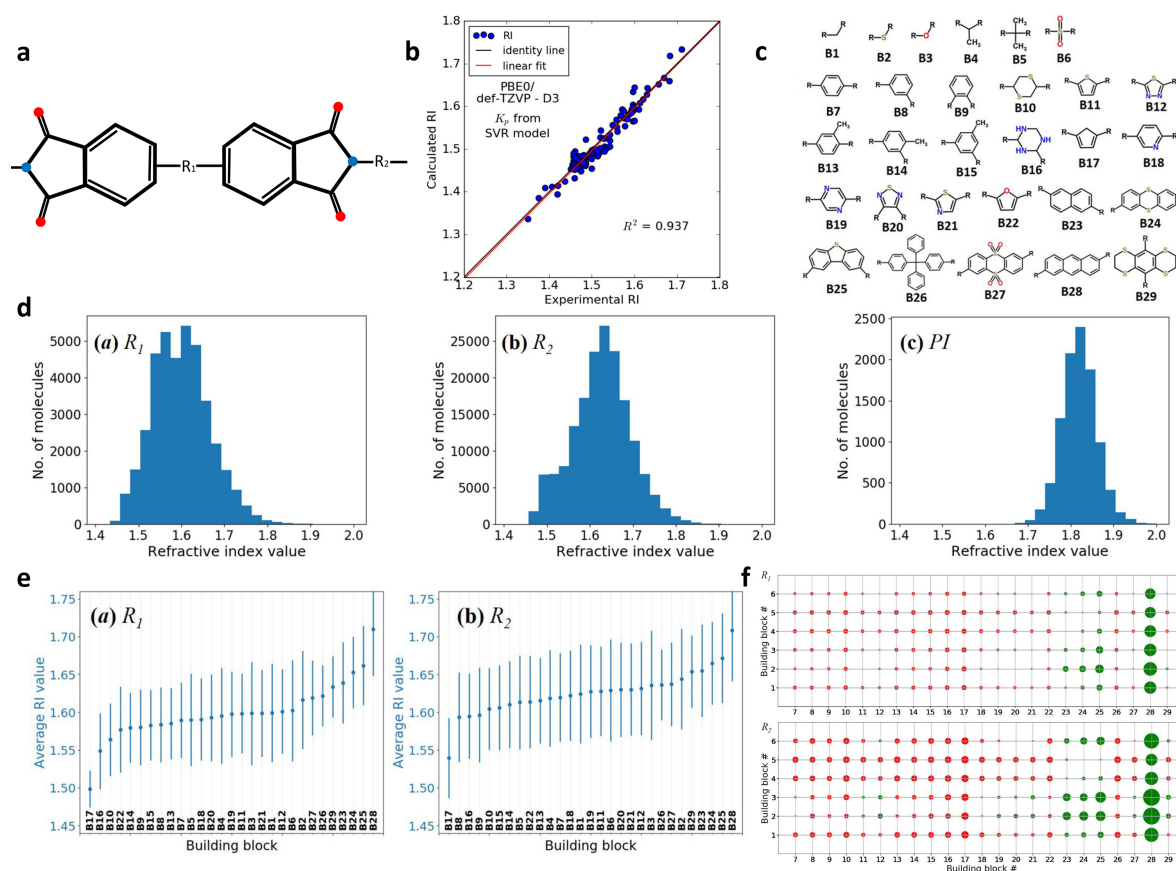
Polyimides (PIs) are potential materials for the next generation optic and optoelectronic applications. To achieve superior performances, the desired values of refractive index (RI) should be larger than 1.7 [98]. However, the typical RI values of many current PIs are between 1.3~1.5. Therefore, it is of great significance to design new PIs from the molecular or monomer level with desired RI values. A monomer structure of PIs is determined by two basic blocks called R1 and R2 [99], as shown in Figure 6a. The properties of PIs can be tuned by controlling these two basic blocks [100]. In the design of basic blocks, there are still problems to solve. First, the chemical space is huge because there is large amount of possible combinations of the basic blocks. Second, screening all the candidates by traditional methods is cumbersome and not cost-effective. Recent publications have showed that utilization of ML algorithm and first principle calculations can accelerate the discovery of promising PIs [100].

Different from the aforementioned examples, the work did not construct an ML model by directly mapping a PI's molecular structure to the RI values. Instead, they tried to build a relationship between the molecular structures of PIs and two intermediate properties (the polarizability  $\alpha$  and the number

density  $N$ ). These two intermediate properties are related with RI values through analytical analysis. An analytical equation between  $\alpha$ ,  $N$  and RI values are given as follows:

$$n_r = \sqrt{\frac{1 + 2\alpha N/3\epsilon_0}{1 - \alpha N/3\epsilon_0}}, \quad (2)$$

where  $n_r$  is the RI value, and  $\epsilon_0$  is the material permittivity in vacuum. The values of polarizability  $\alpha$  were obtained from first principle calculations. The number density  $N$  was obtained by applying data modeling via the van der Waals volume  $V_{dWV}$  and packing fraction  $K_p$  ( $N = K_p/V_{dWV}$ ). In addition,  $V_{dWV}$  was calculated by Slonimskii's method [101], and  $K_p$  was predicted by support vector regression (SVR). As a result, given a molecular structure, the RI value can be predicted by using this protocol with integrated first principle calculations and ML models.



**Figure 6.** Accelerated design of polyimides (PIs) with high refractive index. (a) The core structure of PI with R1 and R2 group (blue and red nodes denote nitrogen and oxygen atoms, respectively); (b) the performance of the support vector regression (SVR) model (adapted from Reference [102] with permission, copyright 2018 AIP publishing); (c) 29 building blocks; (d) the distribution of molecules versus the RI values for R1, R2, and PIs; (e) RI values in terms of each building block for R1 and R2; (f) Z-score of building pairs for R1 and R2 (c–f) are adapted from Reference [100] with permission, copyright 2019 American Chemical Society.

To develop an ML model, a dataset of 84 polymers from literature was selected, in which the  $K_p$  was obtained by experiments. The number of monomer units was used for the structure feature, and  $K_p$  value was the property to be related to. All the data was treated as a training dataset to build an SVR model. To validate the ML model, an external dataset including 112 non-conjugate polymers was selected as the test dataset. The performance of the SVR model is shown in Figure 6b. As we can see, a good correlation has been found with  $R^2 = 0.94$ . Thus, the ML model is acceptable for  $K_p$  prediction.

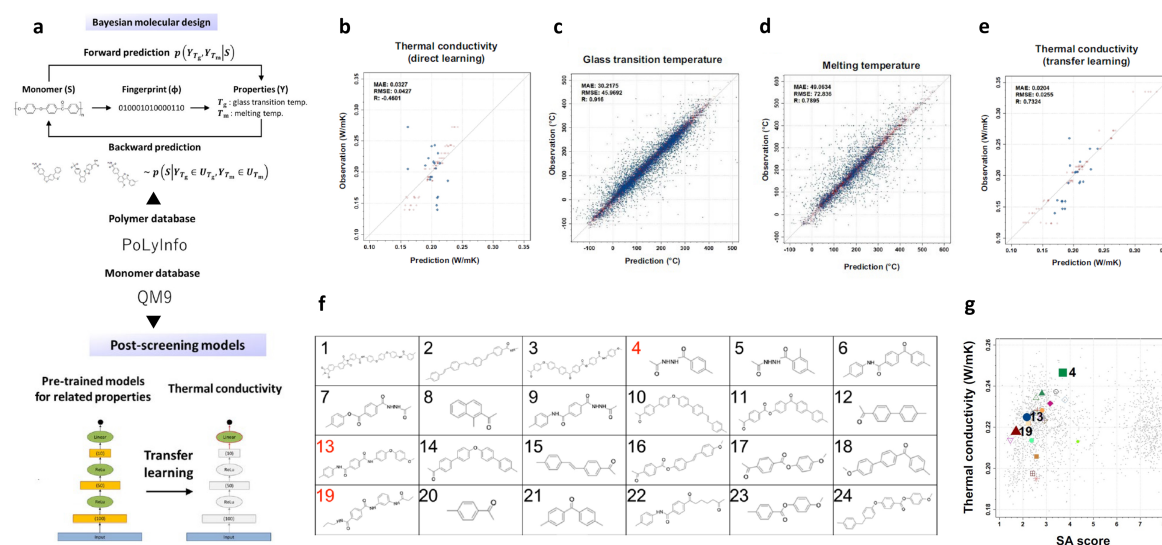
Once the prediction model is constructed, it is then employed for the high-throughput screening to find promising candidates. To this end, a molecular library was determined first. A group of 29 building blocks, including 6 linkers (B1–B6) and 23 aromatic moieties (B7–B29), serves as the base to generate the library (see Figure 6c). Following a combinatorial approach, the authors found 38,619 candidates for R1 and 171,172 candidates for R2, resulting in a chemical space including 6.6 billion candidates for PIs. Instead of searching the whole chemical space, they narrowed down the candidate pool by only selecting the top 100 R1 and 100 R2 candidates with highest RI values. In this way, the search space was reduced to just 10,000 candidates.

The distributions of RI values with respect to the R1 group, R2 group, and the combined PI structures are shown as Figure 6d. As we can see, the authors found promising candidates with RI values larger than 1.7. In the meantime, they also tested the influence of individual block and block pairs on the RI value through the Z-score analysis. With the help of Z-score analysis, the correlations of each building block and block pairs to the RI values are quantified. This sensitivity analysis can distinguish contributions from block or block pairs, so that the key group can be located. From Figure 6e, we can tell that R1 and R2 candidates containing B24, B25, and B28 have higher RI values than the rests. Figure 6f suggests that block pairs containing B28 give the best RI performance. In a nutshell, the building block B28 seems to be the most promising one. It inspires us that a data mining technique can be applied to find underlying patterns between structures and properties, which can help reduce dimensionality and select the main features among many.

### 2.7. Polymers with High Thermal Conductivity

With the rapid development of size and integration density in electronic devices, the massive heat generated poses challenges on the performance and lifespan of these devices. Thus, the discovery of heat dissipating polymeric materials with high thermal conductivity is of great importance to maintain reliable performance, as well as a long lifetime, for these electronic devices [103–105]. However, in the design of polymeric materials, an ML model to directly link a structure and thermal conductivity is difficult to construct. It is because common public databases including thermal conductivity are limited. Moreover, due to the massive chemical space, building a large database by molecular dynamics is currently restricted by the extremely long computation time. Nonetheless, a recent work gave a good idea for how to deal with this design problem. In this work, the authors incorporated ML algorithm, expertise from material synthesis, and advanced measurement technology [106]. The synthesized polymers were reported to have thermal conductivities of 0.18–0.41 W/mK, which match the cutting-edge ones in thermoplastics without any fillers. Figure 7a shows the schematic of the ML-guided design process, which consists of a forward prediction step, a Bayesian molecular design step, and a backward prediction step.

In the first and forward prediction step, the data was obtained from the public database PoLyInfo [107] and QM9 [90,108]. There were a hundred properties recorded in accordance with constitutional repeating units. However, the data for thermal conductivity had only 28 instances, which limited the performance of a predictive ML model. Additionally, as a second order tensor, thermal conductivity is sensitive to polymer processing. As a result, the direct structure-property relation is less convincing. As indicated in Figure 7b, the coefficient of determination is only  $-0.4601$ , which is very bad. Alternatively, an indirect approach was adopted by the authors for the QSPR development. Based on heat conduction theory and reported literature [109,110], they first mapped structures of molecules to proxy properties, including glass transition temperature  $T_g$ , melt temperature  $T_m$ , density  $\rho$ , and heat capacity  $C_V$ . A transfer learning process was then applied to link the structure to the desired thermal conductivity property. More importantly, the data of the surrogate properties was adequate to obtain a reasonable model. For instance, PoLyInfo recorded 5917 and 3234 structures for  $T_g$  and  $T_m$ , respectively.



**Figure 7.** Integrated design for polymers with high thermal conductivity. (a) The proposed ML approach for materials discovery; (b) the performance of a direct learning algorithm; (c,d) validations of the trained linear regression model for glass transition temperature and melt temperature, respectively; (e) validation of the transfer learning; (f) the screened molecular candidates, in which the number are synthesized in red color; (g) validation of the synthesized molecules. The figures are adapted from Reference [106] with permission, copyright 2019 Springer Nature.

Molecular fingerprints (ECFP) was selected as the molecular feature. A linear regression model was trained based on 80% of the dataset. As illustrated in Figure 7c,d, the models gave good predictions for glass transition temperature and melt temperature. By using different training data, 1000 pre-trained models were constructed, in which the weight parameters involved were refined using the limited data of thermal conductivity in the library. From this model pool, the best transferable model to predict thermal conductivity was identified. Following these approaches, a transfer learning model was constructed to correlate thermal conductivity to the molecular structure. However, the performance of the model is unreliable since the test dataset only has 28 data points. Therefore, in the molecular design step, the intermediate properties  $T_g$  and  $T_m$  were adopted as design targets, while the transferred model was employed as a post-screening tool for accelerated molecular generation. Figure 7e verified the transfer learning model with certain precision.

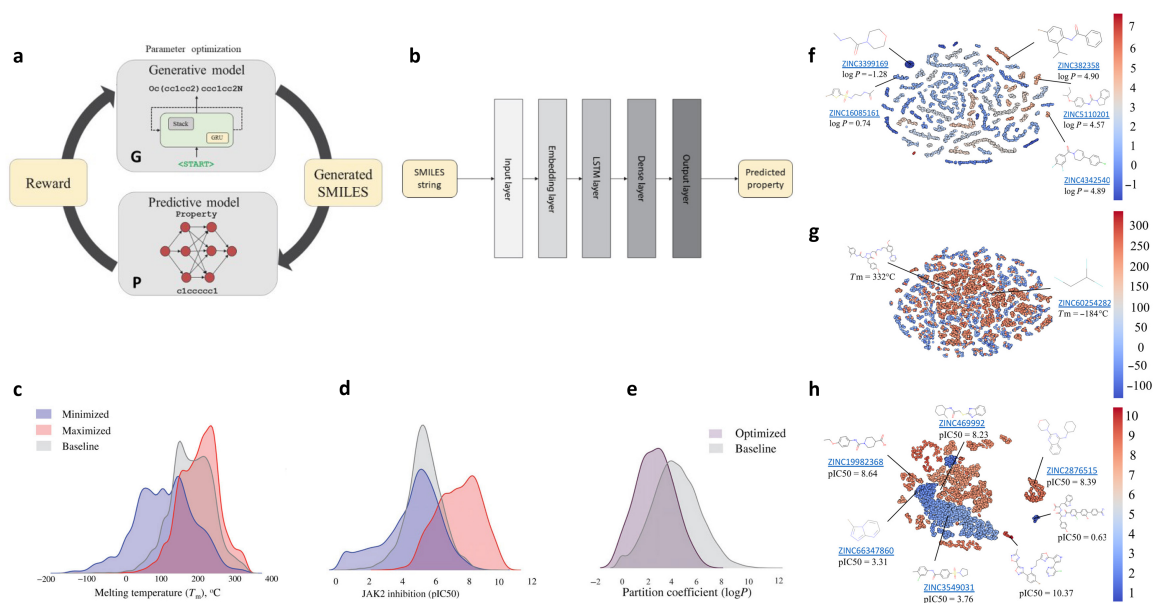
The next step was the generation of a molecular library. A successful dataset with high percentage of chemical valid structures is the key for molecular design. To avoid the generation of a large amount of invalid structures by using chemical fragments and their combinations, the so-called Bayesian molecular design method was carried out [111]. The Bayesian method requires significantly less data for model training compared to the recurrent neural network (RNN) [112] and variational autoencoder (VAE) [113]. Incorporating the SMILES representation and Bayesian ML model, the authors used the self-developed n-gram technique for molecular generation. In the molecular generation process, constraints were considered, including synthetic accessibility (SA), ease of processing, validity of chemical bond, chemical stability, and liquid-crystalline polymers (LCPs) likeness. Finally, 1000 candidates were generated.

The third step was to screen these 1000 candidates. The authors considered three key factors, namely LCP-like structures, higher SA score, and ease of processing (which required  $T_g \leq 300$  °C). As a consequence, 24 candidates were identified for further investigation (see Figure 7f). Three promising candidates among these 24 candidates were further synthesized and tested. The comparison between the prediction and experimental results of the screened structures is showed in Figure 7g. We can see that the screened candidates have desired properties, which verifies the proposed protocols.

## 2.8. De Novo Drug-Like Molecular Design

Thanks to the recent advances of computer science techniques, various ML models have been introduced into drug discovery, for instance, the recurrent neural networks (RNNs), the generative adversarial network (GANs), and variational autoencoder (VAE). Many previous works [54,112,114,115] focused these models on screening of chemical libraries, or they may not be able to design drug-like molecules by well controlling the properties in a specific range [116]. On the other hand, Popova et al. recently reported that, using deep reinforcement learning (RL) algorithm, they were able to bias the generation of molecular library towards specific range of chemical, physical, and/or biological properties [116].

In their work, two deep neural networks, along with generative and predictive models, were developed independently. These two models were then combined into a RL framework, as shown in Figure 8a. There are two phases in the proposed framework. In the first phase, the generative model and predictive model are trained separately. In the second phase, they are trained together with RL to generate new molecules with targeted properties. The generative model aims to learn the underlying pattern between a molecule and its SMILES string so that it can generate novel chemically feasible candidates. The predictive model (Figure 8b) is designed to evaluate the candidates generated by the generative model. In the RL step, the predictive model assigns rewards if the desired candidates are generated, otherwise assigns penalties to the given candidates. In this way, the generation of new molecules is biased towards the region with target properties.



**Figure 8.** De novo drug-like molecular design framework. (a) The proposed framework of reinforcement learning (RL) model; (b) flow chart of the predictive model; (c–e) properties distributions of RL model versus baseline model (no RL); (f,g,h) clustering of generated molecules. The figures are adapted from Reference [116] with permission, copyright 2018 AAAS.

In developing the generative neural network, the database adopted for training is obtained from the ChEMBL21 database (<https://www.ebi.ac.uk/chembl/>). The deep network of this model had a Gated Recurrent Unit (GRU) layer with 1500 units and a stack augmentation layer with 512 units. In order to generate chemically legitimate molecules (correct valence, balance of ring opening and closure, or bracket sequences with different brackets styles) in SMILES form (the output of the model), the authors used the stack memory augmented network to learn the underlying grammar towards chemical feasible strings. The model was then trained with about 1.5 million structures from the ChEMBL21 database to learn the SMILES grammar of real chemical structures. To validate the model, they generated 1 million molecules. The validation checker by using ChemAxon [71] showed that



95 percent of the generated compounds were valid and chemically sensible. It was also shown that less than 0.1% of the generated molecules was from the training dataset. It indicates that the model did learn the grammar instead of memorizing the training dataset. Moreover, the synthetic accessibility (SA) score calculation showed that most generated molecules were synthetically accessible (99.5% molecules have SA score below 6 above which molecules are considered not easy to be synthesized). By comparison, the same ML model without the stack memory showed different results in two aspects: (1) the chemically valid percentage was reduced to 86%; (2) the number of similar molecules to the training dataset was increased. All of these aspects justify the importance of using stack memory for learning.

When developing the predictive neural network,  $T_m$ ,  $\log P$ , and  $\text{pIC}_{50}$  for JAK2 were selected as the target properties. SMILES strings were selected as the only molecular representation. The deep network employed had three layers: a long short-term memory (LSTM) layer with 100 neurons and tanh activation function, a dense layer with 100 neurons and Rectified Linear Unit (ReLU) activation function, and the output layer with one neuron and identity activation function. 5-fold cross-validation technique was adopted to build the predictive ML model. The model was then applied to external prediction for model validation. Good predictive performance was observed. For example, the prediction gave a  $R^2 = 0.91$  for  $n$ -octanol/water partition coefficient ( $\log P$ ). In addition, the prediction for melting temperature ( $T_m$ ) was shown to be comparable to the prediction by state-of-the-art descriptor based random forest model [117].

Incorporated with the generative and predictive models, the RL was then formed to design molecules with controllable physical/chemical/biological properties ( $T_m$ ,  $\log P$ , and JAK2 inhibition). To design target properties with maximum or minimum range, the authors gave high rewards to molecules with more benzene rings, and more small group substitutes (like  $-\text{NH}_2$ ,  $-\text{OH}$ ). On the other hand, they penalized molecules with undesired structures like bromine or carboxyl group. Figure 8c–e demonstrated the property distributions of training dataset and generated library (10,000 molecules). One can see that the generated molecules shifted the properties from the baseline to maximum or minimum range, which verified the model constructed. To further visualize the generated molecules, the t-distributed stochastic neighbor embedding for dimensionality reduction was carried out as shown in Figure 8f–h. In these figures, a point refers to a molecule and is colored by its property value. We can see that there are clusters for  $\log P$  and JAK2 inhibition, while no cluster for  $T_m$ , which can provide useful information for those molecules.

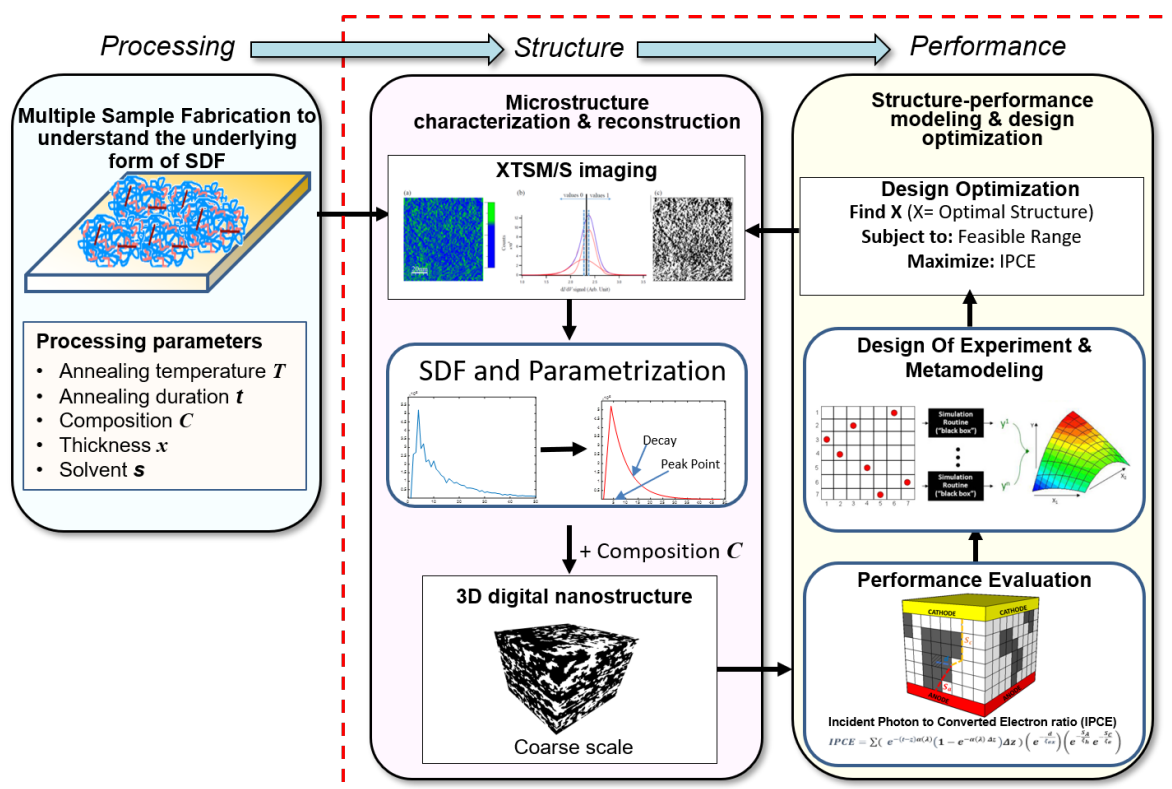
### 2.9. Microstructure Design of Organic Photovoltaic Solar Cells (OPVCs)

The materials design approaches presented in previous cases were attempted at only molecular levels, while the current example zooms out to the nanoscale level. In particular, these previous cases mainly focus on the chemistry–property relationship for organic molecules and polymers, while the processing–microstructure–morphology–performance relationship also plays an important role in the design of advanced functional materials and devices [118–122]. Different from the OPV example given in Section 2.1, where the focus was more on chemical composition and resulting molecular structure, another approach identified is the optimization of the microstructure for OPVCs to improve the performance (which in this case is called Incident Photon to Converted Electron-IPCE efficiency) [120]. It is because the ultimate performance of the devised devices is highly dependent on the material processing and the microstructure. The processing–structure–performance (PSP) relationship [123] is the key to design materials with targeted performances. As such, a case study on microstructure design of OPVCs is presented to make this study more comprehensive.

The quest for high performance metamaterials by cost-effective fabrication techniques has put design of nanostructures material systems (NMSs) [124] in the limelight. Compared to the costly top-down fabrication techniques for periodic NMSs, the quasi-random nanostructures can be fabricated by low cost, bottom-up, processes [125]. One such example of quasi-random nanostructures is OPVCs [6,126,127]. In order to get optimal performance, multiple structure and processing parameters,

such as thickness of active layer, donor-acceptor ratio, annealing temperature, etc., need to be optimized simultaneously. There have been multiple attempts to optimize the performance by changing one or two parameters at a time, but a more recent approach provides a framework which can take more parameters in an efficient manner to reach the optimal structure [128].

Although the authors propose to complete the PSP chain, they focus their work on the first stage, i.e., to establish the Structure-Performance (S-P) relationship (enclosed in red dotted box in Figure 9). Because of the high dimensionality of the underlying structure, it becomes necessary to reduce the dimensionality by extracting only the salient and potentially useful features of the microstructure. Microstructure characterization and reconstruction (MCR) [119,129] provides a quantitative approach to analyze microstructure, while reducing the dimensionality. Among the many MCR techniques [119,129–132], the authors chose spectral density function (SDF) [118,133–135] because of its proven efficacy to represent [118,133,134] and design [134] quasi-random NMSs, and also because of its physical association with processing conditions [118]. SDF is a one-dimensional function of spatial frequency, calculated as the radial average of the squared magnitude of Fourier spectrum of a quasi-random structure [135,136] and represents the structural correlation in Fourier space. To make the design problem more efficient, this one-dimensional function can be further reduced to a couple of variables by parameterizing the SDF curve. For this study, two parameters were selected, i.e., Decay and Peak Point, to represent SDF.



**Figure 9.** A framework for designing active layer of organic photovoltaic solar cells (OPVCs) via spectral density function. This figure is adapted with permission from Reference [128], copyright 2018 American Society of Mechanical Engineering. SDF = spectral density function.

After characterization, a statistically equivalent 3D microstructure was created using SDF. A novel, physics-based performance evaluation strategy was developed in this work to evaluate the efficiency of this reconstruction. In the next step, a performance optimization problem was setup to determine the optimal microstructure. Every 3D digital reconstruction required significant computational time. To save the computational cost of reconstruction and make the design more efficient, ML technique was brought in. Metamodel, which is a popular category in ML, allows the usage of a few intelligently

selected datapoints to estimate the original model. This metamodel can then be used to find the global/local minimum. The metamodel in this study was based on the kriging technique [137] and Optimal Latin Hyper Sampling [138] technique was used to find the data points. For three input variables (two SDF parameters, and Volume Fraction), 45 data points were reconstructed and evaluated to fill the design space to build the metamodel. The range for SDF parameters was based on a couple of SDFs from X/STM images of OPVC samples, and the range for Volume Fraction was selected based on the literature studies. In the final step, Sobol sensitivity analysis [139] was also carried out to extract most important variables.

The study showed a 36.75% increase in the IPCE value after structural optimization. Upon investigation, it was also concluded that the PCBM (Phenyl-C61-butyric acid methyl ester) volume fraction (or composition) was the most influential design variable followed by Decay (one of the SDF parameters). The current work lacks experimental validation, so the next logical step for this study could be to complete the PSP chain and verify the optimized result against physical experiments.

### 3. Discussion

The aforementioned nine examples about organic molecule and polymer design demonstrate the huge potential for applying ML to accelerate the materials design process in various fields, as summarized in Table 1. However, challenges still exist in application of ML-guided materials design approach. In what follows, details about the problems associated with the ML approach will be discussed. In particular, materials database, feature selection and extraction, and ML models for molecular generation and inverse materials design will be further discussed.

#### 3.1. Materials Database

The acquisition of a large materials database is the first step and the foundation to develop an ML model. Within the database, material structures and corresponding properties are enclosed, which could be obtained either from experimental works or numerical simulations. The characteristics of the database strongly influence the capability of the ultimate ML model since the model is trained and validated by the database. These characteristics include adequate size, diversity, and uniformity across the chemical space [140]. Diversity and uniformity are even more important to construct a predictive ML model for interpolation, extrapolation, and exploration. Even with a small database, effective ML models could still be built with reasonable accuracy, as shown in the energetic materials design case [91,94] and polymer thermal conductivity design case [106].

Considering the tremendous effort needed to establish a database from scratch, a better way is to use existing public databases, provided that the structures and properties of interests are available. Some popular materials databases are listed in Table 2 [57,60]. If public databases are not available in some cases, natural language processing (NLP) can be applied to extract data from published literature. For example, Cooper et al. [141] adopted the text-mining tool ChemDataExtractor [142] to construct a database of 9431 dye candidates in their polymeric solar cell study. Their database constructed through NLP contains the information of chemical structures, maximum absorption wavelengths, and molar extinction coefficients. Based on this database, the authors then identified a promising candidate from the molecular library by high-throughput screening, which was further verified by experiments. The experimental test showed that it demonstrated comparable PCE to the organometallic dye N719. The solar cell study successfully shows the effectivity of using NLP for data mining in materials science fields.

**Table 1.** Summary of the nine ML-guided materials design examples.

Materials	Design Feature	Design Scope	Data Size	Representation	ML Model
Organic photovoltaics (2011)	Self-built library and screening	Power conversion efficiency (molecular level)	2.6M	Molecular descriptors	MLR
Polymer dielectrics	Self-build library; building blocks for molecular generation; genetic algorithm	bandgap and dielectric constant (molecular level)	284	Fingerprints	KRR
Organic light-emitting diodes	Self-build library and screening; building blocks for molecular generation	delayed fluorescent rate constant (molecular level)	40,000	ECFPs	ANN
Polymer solar cell (2018)	Self-build library and screening; building blocks for molecular generation; various combinations of feature representations and ML models are compared	highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) (molecular level)	3938	Fixed length vector; string; spatial coordinate	LRR; MLP; RF; DTNN; GrammarVAE
High-energetic material	Material design with limited data; various combinations of feature representations and ML models are compared	high energy density and low sensitivity (molecular level)	109; 309	CDS; SoB; CM; BoB; fingerprints	KRR; RR; SVR; RF; kNN; LASSO; GPR; ANN
Polyimides with high refractive index	Self-build library and screening; building blocks for molecular generation; ML model construction with limited data	polarizability and number density (molecular level)	196	Number of monomer units	SVM
Polymer with high thermal conductivity	ML model construction with limited data; transfer learning	thermal conductivity (molecular level)	28; 5917; 3234	ECFPs	Bayesian model
de novo drug-like molecule	Material design with arbitrary target property range; SMILES strings as input for molecular generation	physical/chemical/biological properties (molecular level)	1.5M	SMILES	DNN; RL
Organic photovoltaic solar cells (2019)	Polymer composite design; bottom-up nanofabrication; microstructure characterization and reconstruction	IPCEfficiency (microstructure level)	45	Microstructure characterization	SDF

Note: ECFPs: extended-connectivity fingerprints; CDS: custom descriptor set; SoB: sum over bonds; CM: Coulomb matrix; BoB: bag of bonds; SMILES: simplified molecular-input line-entry system. MLR: multi-linear regression; KRR: kernel ridge regression; ANN: artificial neural network; LRR: linear ridge regression; MLP: multi-layer perceptron; RF: random forest; DTNN: deep tensor neural network; GrammarVAE: grammar variational autoencoder; KRR: kernel ridge regression; RR: ridge regression; SVR: support vector regression; kNN: k-nearest neighbors; GPR: Gaussian process regression; SVM: support vector machine; DNN: deep neural network; RL: reinforcement learning; SDF: spectral density function.

**Table 2.** Some public materials databases enclosing structures and properties.

Database	Type	Description	URL
AFLOWLIB	Computation	Database of 2,961,744 material compounds with over 527,190,432 calculated properties	<a href="http://afowlib.org">http://afowlib.org</a>
BNPAH	Computation	Structures and properties of 77 polycyclic aromatic hydrocarbons and 33,059 B, N substituted compounds	<a href="https://moldis.tifrh.res.in/datasets.html">https://moldis.tifrh.res.in/datasets.html</a>
ChemDiv	Comp./Exp.	Collection of over 1,500,000 individually crafted, lead-like, drug-like small molecules	<a href="http://www.chemdiv.com/complete-list/">http://www.chemdiv.com/complete-list/</a>
ChemSpider	Experiment	A free chemical structure database providing fast text and structure search access to over 67 million structures	<a href="https://chemspider.com">https://chemspider.com</a>
ChEMBL	Experiment	A manually-curated database of bioactive molecules with drug-like properties	<a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a>
Citration	Experiment	A premier open database and analytics platform for the world's material and chemical information	<a href="https://citration.com">https://citration.com</a>
CMR	Computation	A collection of molecules obtained from electron-structure codes	<a href="https://cmr.fysik.dtu.dk">https://cmr.fysik.dtu.dk</a>
COD	Experiment	A collection of crystal structures of organic, inorganic, metal-organics compounds, and minerals, excluding biopolymers	<a href="http://www.crystallography.net/cod/">http://www.crystallography.net/cod/</a>
CSD	Experiment	A database of over one million small-molecule organic and metal-organic crystal structures	<a href="https://www.ccdc.cam.ac.uk">https://www.ccdc.cam.ac.uk</a>
DrugBank	Experiments	Drug database with comprehensive drug target information	<a href="https://www.drugbank.ca/">https://www.drugbank.ca/</a>
eMolecules	N/A	Commercially available with over seven million compounds for drug discovery	<a href="https://reaxys.emolecules.com/index.php">https://reaxys.emolecules.com/index.php</a>
Energetics	Computation	A database of energetic molecules	<a href="https://git.io/energeticmols">https://git.io/energeticmols</a>
GDB	Computation	A database containing hypothetical small organic molecules	<a href="http://gdb.unibe.ch/downloads">http://gdb.unibe.ch/downloads</a>
HCEP	Computation	Harvard Clean Energy project for solar absorber materials	<a href="https://cepdb.molecularspace.org">https://cepdb.molecularspace.org</a>
HOPV15	Comp./Exp.	A collation of experimental photovoltaic data from the literature and calibrated by DFT calculation	<a href="https://figshare.com/articles/HOPV15_Dataset/1610063/4">https://figshare.com/articles/HOPV15_Dataset/1610063/4</a>
ICSD	Experiment	A database of inorganic crystal structure	<a href="https://icsd.fiz-karlsruhe.de">https://icsd.fiz-karlsruhe.de</a>
MatNavi	Experiment	A materials databases of polymer, ceramic, alloy, superconducting material, composite, and diffusion	<a href="http://mits.nims.go.jp">http://mits.nims.go.jp</a>
MatWeb	Experiment	A database of material properties of polymers, metals, ceramics, and semiconductor	<a href="http://matweb.com">http://matweb.com</a>
MP	Computation	Computed information on known and predicted materials	<a href="https://materialsproject.org">https://materialsproject.org</a>
NIST CW	Experiment	A database of thermochemical properties	<a href="https://webbook.nist.gov/chemistry">https://webbook.nist.gov/chemistry</a>
NIST MDR	Experiment	A repository of material data being updated	<a href="https://materialsdata.nist.gov">https://materialsdata.nist.gov</a>
NOMAD	Computation	A repository to host, organize, and share material data	<a href="https://nomad-repository.eu">https://nomad-repository.eu</a>
NREL MD	Computation	A computational materials database for renewable energy applications	<a href="https://materials.nrel.gov">https://materials.nrel.gov</a>
OQMD	Computation	A database of DFT-calculated thermodynamic and structural properties	<a href="http://oqmd.org">http://oqmd.org</a>
PubChem	Experiment	A chemical database of chemical and physical properties, biological activities, and safety and toxicity information	<a href="https://pubchem.ncbi.nlm.nih.gov">https://pubchem.ncbi.nlm.nih.gov</a>
QM	Computation	Small organic molecules calculated by DFT	<a href="http://quantum-machine.org/datasets/">http://quantum-machine.org/datasets/</a>
TEDesignLab	Comp./Exp.	Thermoelectric material design	<a href="http://tedesignlab.org">http://tedesignlab.org</a>
ZINC	Computation	Database of commercially-available compounds for virtual screening	<a href="https://zinc15.docking.org">https://zinc15.docking.org</a>

### 3.2. Machine Learning Model

#### 3.2.1. Feature Selection and Extraction

With the obtained database at hand, the next key step is to transform a material's structure into a digital coded format that can be used as the input for ML models. However, there are no fixed rules for feature selection. In general, representations of chemical structures are expected to meet several requirements. First of all, the representation should be based on chemical or physical knowledge. It is shown that representations embedding the physics and structural information help ML models perform better [60,61]. Additionally, the representation of a certain chemical structure should not change under transformations, such as spatial translation and rotation [143]. Moreover, it is better for the representation methods to be unique and invertible [60]. Generally, molecular representations can be classified into three types: discrete, continuous, and weighted graphs [24]. Several commonly used representations are categorized in Table 3. Among them, molecular descriptors, fingerprints, SMILES, and graphs are the most commonly used representations of chemical structures for organic molecules.

Molecular descriptors are either experimentally measured or theoretically derived properties of molecules [144], which have long been adopted for developing QSAR/QSPR [145–147]. A molecular descriptor transforms the information of a chemical structure into a mathematical form, which makes it easy to develop a quantitative relationship between structures and corresponding properties [148,149]. There are many types of molecular descriptors, such as 3D types [150,151] and topological types [152]. Some tools can be used to generate molecular descriptors, for example, the RDKit package [86], Dragon [153,154], etc.

Fingerprint, a special molecular descriptor, uses a vectored form to represent a chemical structure. Two- and three-dimensional fingerprints have been widely adopted to represent small molecules by bit strings for evaluation of compound similarity [155]. There are several classes of fingerprints, such as topological fingerprints, structural fingerprints, circular fingerprints (e.g., extended connectivity fingerprints, ECFP [88]), pharmacophore fingerprints, and hybrid fingerprints [155]. Fingerprints can be easily applied for inverse materials design. To illustrate, Huan et al. [143] adopted motif-based fingerprints to describe structures composed of C, H, O, N, and F and obtained the fingerprints-properties relationship by ML models. With this mapping at hand, they realized fast materials design by reconstructing fingerprints with target properties. However, there are limitations associated with this representation approach. For example, though it can express the presence of an element and how many of them are presented by fingerprints, it lacks the stacking information [9].

SMILES represents a molecular structure by text string, which is probably the most popular representation method. For demonstration, the SMILES form of benzene ring is C1=CC=CC=C1. However, different SMILES forms may represent the same molecule [60]. Additionally, it can not represent metallic, periodic materials, and properties that are three-dimensional geometry-dependent [156]. Furthermore, a high percentage of invalid candidates may be generated when implemented in molecular generation [60]. Nonetheless, successful applications can still be achieved based on this representation. For example, Ikebata et al. [111] developed a chemical language model following SMILES notation to generate chemically favorable molecules in a Bayesian molecular design method. The model obtained good performance of design of small organic molecules with desired internal energy and HOMO-LUMO gap.

Graph-based representation is deemed a promising method to take care of the shortcomings associated with SMILES representation [157]. It has been successfully applied to generate small molecular graphs [158,159]. In a molecular graph, atoms and bonds are represented by nodes and edges. Among many of the ML models, the VAE model is usually combined with molecular graph representation for molecular design through an encode-decode process of graphs [158,160]. Recently, Li et al. [157] proposed a conditional graph generative model which incorporated DNNs to resolve the sequential design of a molecular graph. They showed that the model built outperformed SMILES-based representation on the multi-objective (drug likeness, synthetic accessibility, etc.) drug design problem.

Feature representation can significantly affect the performance of an ML model [7]. It is reported that fingerprints are suitable for DNNs, while molecular graphs are more favorable for CNNs and RNNs [61]. It is also shown from the energetic materials design case that the sum over bond representation performs better than the custom descriptor set, Coulomb matrix, and bag of bonds representation [91,94].

**Table 3.** Common feature representations of organic molecules and tools for feature generation.

Representation	Description	References
SMILES	Line notation for describing a chemical structure using text strings	[87,112,115,116]
Fingerprints	A special descriptor using vector of fixed or variable length to represent a chemical structure	[58,143,155,161]
Molecular graphs	A representation of chemical structures by graph theory	[157–160]
Coulomb matrix	A matrix representation embedded nuclear coordinates and charges, similar representations include Ewald sum matrix, Sine matrix	[90,91,162–164]
Smooth overlap of atomic orbitals (SOAP)	A special descriptor encoding atomic structures using local expansion of atomic density	[165–167]
Atom-centered symmetry functions (ACSF)	A special descriptor representing the local environment near an atom using two- or three-body functions	[168–170]
Bag of bonds	A vector enclosing chemical bonds and corresponding numbers	[91,92]
Grids of molecules	A visual form of molecules generated by their coordinates	[61,171,172]
Tools	Description	References
CDK	Chemistry Development Kit: open-source Java libraries for cheminformatics to generate various descriptors, fingerprints, etc.	[173–176]
ChemDes	A free web-based tool for generation of molecular descriptors (3679 types) and fingerprints (59 types)	[144,177]
ChemMine	A free online tool for analyzing and clustering small molecules, including similarity search and properties calculations	[178,179]
OEChem	Programming library for chemistry and cheminformatics with small molecules	[180–182]
Open Babel/Pybel	Open-source chemical toolbox to search, convert, analyze, and store data	[183–185]
PaDEL	A software to generate molecular descriptors (1875 types) and fingerprints (12 types) using CDK	[186,187]
PubChemPy	An open-source python library to interact with PubChem	[188]
RDKit	A collection of cheminformatics and machine-learning tools	[86,189]

### 3.2.2. ML Methods and Model Validation

Since material properties are always assessed in the study of materials science, supervised learning methods are commonly used to learn the QSAR/QSPR from the samples that are labeled with assessed property values. Several popular ML models and respective features are briefly reviewed here. More details can be found in several recent review articles on ML methods for materials science [21,56–58,190].

*Linear regression* is a statistical method to regress a target or response variable on a set of explanatory variables or features. It associates a weight with each feature and sums them up to predict the response. This gives rise to a linear function in terms of the weights. Linear regression commonly minimizes the squared sum, i.e., sum of the squared error on each observed example, which leads to the least squares method. This method, although simple, is often the baseline choice for a problem.

*Gaussian process regression (GPR)* is a statistical model where observations occur in a continuous domain, such as time or space. In a Gaussian process, every point in a continuous input space is associated with a normally distributed random variable. Inference of continuous values with a Gaussian process prior is known as Gaussian process regression. Gaussian process regression is a non-parametric Bayesian approach towards regression problems. It can capture a wide variety of relations between inputs and outputs by utilizing a theoretically infinite number of parameters and determine the level of complexity from data through the means of Bayesian inference.

*Decision tree (DT)* consists of a tree structure of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It consists of nodes representing attributes, edges to branch by values of a selected attribute represented by the node from which the edge comes, and leaf nodes corresponding to the class labels or the response values. Constructing a decision tree is a top-down building procedure. It starts from a root node with the entire training data. Then in each decision node, it finds the best test attribute. By splitting the training data according to a value on this attribute, it diminishes the mixture of classes between the split sets as much as possible. To classify a test example, we start from the root, evaluate the relative test attribute, and then take the corresponding branch. This process is repeated until a leaf is encountered. The new example is then classified to the class that the leaf is labeled.

*k-nearest neighbors (kNNs)* are used for evaluation of continuous data labels and have been widely adopted for regression and classification problems. The property value is obtained by averaging over the number of  $k$  nearest neighbors, where  $k$  is an integer number specified by the users. The weight associated with each nearest neighbor can be equally contributed from its  $k$  nearest neighbors or assigned with different values considering their relative distances. For example, the nearest neighbor has higher weight than more distant neighbors. As such, performance of kNNs are greatly related to the local structures of the data.

*Support vector machines (SVMs)* are supervised learning methods for classification and regression analysis. SVM minimizes a loss function together with a regularizer for the best classifier (or regression model). The regularizer helps SVM find the classifier reaching the largest margin between the different classes of examples, or helps SVM penalize the models that use a lot of predictors or input variables to find sparse models. The reason that SVMs lead to superior performance is because they are able to construct a non-linear model via a linear mechanism by applying the so-called “kernel” mapping. In other words, using kernel calculation, an SVM maps the data from the input space to a high-dimensional feature space that has nonlinear terms of the input variables as coordinates first, then builds a linear model in this feature space.

*Random forest (RF)* is an ensemble learning method that predicts class labels or a response by constructing a multitude of decision trees during training. For classification, it determines the label of a new example by evaluating the mode of the classes of the individual trees. For regression, it predicts the output value of an example by averaging the values from individual trees. These decision trees are trained by a statistical method called “bagging” (which stands for bootstrap aggregating) that has been proved to reduce the model variance; hence, RF is usually not expected to overfit data. These trees differ from the standard decision trees because each node is divided using the best combination of input variables. As the trees grow, an extra randomness is brought into the process to re-shuffle a subset of features from which RF searches for the best combined feature. RF tends to give high accuracy performance in practice.

*Artificial neural networks (ANNs)* are networks connected by layers of artificial neurons which mimic the human brain. A single neuron outputs weighted inputs through a so-called activation function. A typical ANN is consisted of one input layer, an output layer, and one or more intermediate layers called hidden layers. Deep neural networks (DNNs) are special ANNs with more than one hidden layers which have superior learning power. Using nonlinear activation functions, ANNs or DNNs demonstrate excellent capability in solving highly nonlinear problems.

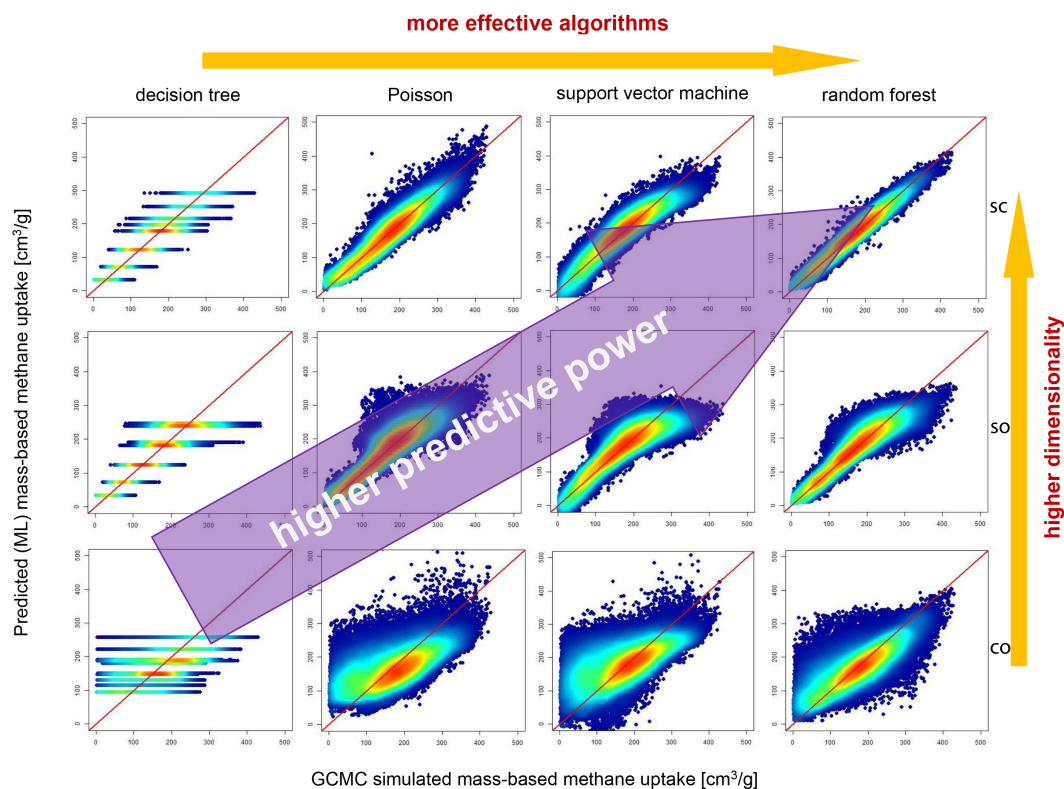


*Convolutional neural networks (CNNs)* are a typical deep learning method. Unlike the multi-layer perceptron (MLP, a type of feed forward neural network that consists of fully connected layers), CNNs take a multi-channelled volume as the input, such as an RGB image with three channels for the red, green, and blue color elements. In a convolutional layer, which is the main building block of CNNs, there are a set of independent filters which are also multi-channelled weights of small size. Each filter is independently convolved with the image in a manner that slides the filter over the complete image; along the way, it computes the dot product between the filter and the image patch in the sliding window. The convolved features are then passed through an activation function and a pooling layer. In a pooling layer, such as the max pooling, the maximum value from the convolved features is returned. The pooling layers help reduce the spatial size of the convolved images and extract features or representations of the image that are rotational and translational invariant. The last layers of CNNs are often the fully-connected layers to predict a response or class label based on image representation produced by the convolutional layers.

*Generative adversarial networks (GANs)* are a family of generative models that use deep neural networks. Unlike a discriminative model that is concerned with how to map from inputs to a response label, a generative model is concerned with how the input data can be generated given the label, so it often learns the distribution of the inputs. A GAN consists of two neural networks, a generator and a discriminator, that play adversarial game between each other. The generator attempts to generate new and synthetic data instances, e.g., images (after training based on a set of observed examples), whereas the discriminator evaluates them for authenticity. The goal of the generator is to generate fake images without being caught by the discriminator. The goal of the discriminator is to identify images created by the generator as fake. Hence, training a GAN requires to optimize a minimax objective function that comprises the two opposing losses. GANs have gained a lot attention lately because of their huge application potential in science, video games, fashion, art, and advertising.

A validated ML model is the foundation for accurate material property prediction and inverse materials design. However, a major issue in developing such an ML model is overfitting. In order to construct a sound ML model,  $n$ -fold cross-validation is usually adopted in which the dataset is split into  $n$  folds and  $n$  rounds of building ML models are carried out. In each experiment, a distinct fold is selected as the test dataset, while the left folds are the training dataset, and the overall predictive performance of the ML model is evaluated by combining the validation results from all  $n$  ML models.

In summary, feature representations and ML models are two key ingredients in developing predictive ML models. Not only is individual selection of feature representation and ML model important on the ultimate performance of the ML model, their combined effect is also worthy of notice. For example, Pardakhti et al. considered the methane uptake of metal-organic frameworks (MOFs) to test the predictive capability of different ML models [191]. The dataset is taken from the database of hypothetical MOFs (hMOFs) [192], from which 130,398 candidates were extracted. Both volumetric-based uptake and mass-based uptake of methane were available in the database. Four different ML algorithms were adopted, namely decision tree (DT), Poisson regression (PR), support vector machine (SVM), and random forest (RF). To evaluate the influence of different descriptors on predictive capabilities of these ML models, the structural only (SO), chemical only (CO), and structural and chemical (SC) variables were tested to make a comparison. The structural descriptor took into account of void fraction, surface area, density, dominant pore diameter, maximum pore diameter, interpenetration capacity, and the number of interpenetration framework, while the chemical descriptor considered the types and the number of atoms, degree of unsaturation, metallic percentage, oxygen to metal ratio, electronegative atoms to total atoms ratio, and weighted electronegativity per atom. The comparison of different combinations of ML models and descriptors is shown in Figure 10. We can see that the RF model gives the best performance among these ML models, with SC descriptors.

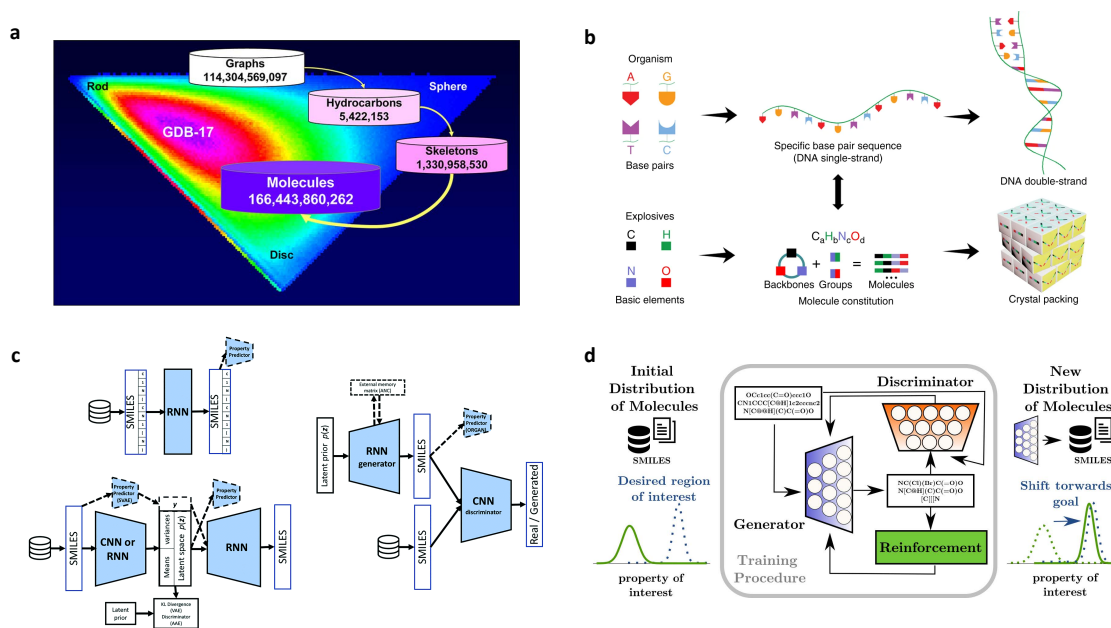


**Figure 10.** The choice of ML models and descriptors that leads to different performance of ML predictions on methane uptake of metal-organic frameworks (MOFs). Reprinted with permission from Reference [191], copyright 2017 American Chemical Society.

### 3.3. Molecular Generation

Molecular generation plays a key role in the inverse materials design process. For example, with a high-throughput screening method, molecules must be generated to form a virtual molecular library. There are different ways for molecular generation. The easiest and most direct way to generate organic molecules is perhaps by enumeration or combinatorial approaches. Taking enumeration as an example, by usage of up to 11 atoms of C, N, O, and F (saturated with H), GDB-11 database of 26.4 million molecules has been constructed [193]. Using the same fashion, GDB-13 database of 970 million molecules [108], and GDB-17 database of 166 billion molecules [23] have also been generated, as shown in Figure 11a.

The molecular generation process, in the case of GDB-17 database, starts from mathematical graphs generated by the GENG [194] program. Using up to 17 nodes by GENG and PLANARG checker of planarity of the graphs to avoid crossed bonds, 114 billion graphs is created initially. The graphs are then subjected to removal of ring strain and small rings. Consequently, 5.4 million hydrocarbons (nodes stand for C and edges denote single bond) are selected. Afterwards, the 5.4 million candidates are transferred into 1.3 billion “skeletons” by selectively substituting bond orders (single, double, and triple bonds) for the graph edges. Similarly, combinatorially substituting N and O for graph nodes, (C) is also employed. Following post-processing for oximes, nitro, S, CF<sub>3</sub>, and halogens, the GDB-17 database is finally generated. To verify the uniqueness of the generated molecules, the database is compared with the public archives of PubChem [195], ChEMBL [196], and DrugBank [197] in terms of molecules with up to 17 atoms. It has been found that the number of molecules in the GDB-17 database is much bigger than the total number of molecules from these three databases. Another finding is GDB-17 molecules include more rings, especially small rings and nonaromatic heterocycles. Last but not least, it contains enormous isomers of known drugs and represents various scaffold types [23].



**Figure 11.** Typical molecular generation methods. (a) GDB molecular database generated by direct enumeration (adapted from Reference [23] with permission, copyright 2012 American Chemical Society); (b) high-energetic molecules generated by a material genome approach (adapted from Reference [52] with permission, copyright 2018 Springer Nature); (c) molecular generation by CNNs or RNNs with SMILES representation (adapted from Reference [60] with permission, copyright 2019 Royal Society of Chemistry); (d) molecular generation by generative adversarial network (GANs) using SMILES representation (adapted from Reference [25] with permission).

In addition to combinations of single atoms, building fragments can also be used for combinatorial molecular generation. As mentioned in the case study of polymer dielectrics, seven building blocks were used for combinatorial generation of monomers [9]. Additionally, in the case study of polyimides design with high RIs, two types of building fragments were adopted to generate PIs [100,102]. The idea is that some building units or functional groups characterize main features of specific materials with desired properties. Thus, using these building blocks for materials design is a fast approach. Another example is the material genome approach for energetic materials design [52]. C, H, O, and N were selected as the basic elements to generate small molecules mimicking the A, G, C, and T base pairs that generate DNA molecules, which are illustrated in Figure 11b. In this work, 1 parent aromatic ring from 14 candidates and certain numbers of substituent groups (each group can have 0~3 counts) from three typical groups are chosen to combinatorially generate the organic molecules for energetic materials. With this approach, a promising candidate was successfully discovered with high energy density, as well as good thermal and mechanical stability to external stimulus, which are comparable to the most popular explosives.

Though combinatorial molecular generation methods have proven successful in real applications, they suffer from generating chemically unfavorable molecules. Therefore, much more effort is needed to virtually screen the database. Additionally, for a specific materials design task, only molecules with desired properties are needed. The molecules may only be a small portion of the whole chemical space, resulting in low efficiency of the generation effort. Moreover, combinatorial methods using known building blocks are experience biased, which significantly limits the novelty of the generated molecules [23]. To solve these problems, chemical constraints are usually applied during the generation process. These chemical constraints, such as valency rules, functional group stability criteria [23], balance of ring opening and closure, synthetic accessibility [116], and medicinal chemistry [198], can be imposed to generate chemically meaningful molecules. However, new issues still remain. On one

hand, if too many constraints are imposed during the molecular generation, it is subjected to loss of novelty [116]. On the other hand, if the constraints are not enough, loss of validity of the chemical library will weaken the capability of the method [89]. It is challenging to balance these two factors for molecular generation.

Recently, deep neural networks (DNNs)-based methods have emerged to handle these problems. They have great advantages in learning the molecular grammar for representing molecules, such as SMILES, from molecule databases and creating molecules based on the information learned. There are several popular methods, such as recurrent neural networks (RNNs), autoencoder based method, generative adversarial networks (GANs), and reinforcement learning (RL). Several recent review articles provide details of these molecular generation methods [24,60,93]. Here, we briefly discuss these methods and highlight their features.

RNNs are the most popular models for sequence modeling and generation, thus having emerged as powerful generative models in different domains, such as natural language processing, music generation, and speech. Segler et al. [112] treated the task of generating SMILES as a language model attempting to learn the statistical structure of SMILES syntax by training an RNN using a large corpus of SMILES. The model can then create large sets of novel molecules with similar physicochemical properties to the molecules in the training set. Gupta et al. [199] trained a long short-term memory (LSTM)-based RNN model to generate libraries of valid SMILES strings and used a common ML strategy—transfer learning—to fine-tune a model previously trained on other sequence data. The resultant model generated molecules that were structurally similar to drugs with known activities against specific targets. This approach was successful for “low data” situations in early stage molecular discovery.

Figure 11c demonstrates the usage of RNNs or CNNs for molecular generation by learning the underlying pattern of valid strings of real molecules. Although these deep learning models could not reach 100% accuracy of finding the molecules with desired properties, the generated molecules can reach a high percentage of chemically valid species [116].

An autoencoder is a neural network that is trained to attempt to encode an input variable into latent variables (in the so-called representation or code layer) and then decode the latent variables to reproduce the input information. Variational autoencoder (VAE) is a framework for training two neural networks—an encoder and a decoder—in a Bayesian way to learn a good representation of an input structure, such as SMILES, in the code layer. Gomez-Bombarelli et al. [54] proposed a generation model using VAE to generate chemical structures after training the VAE on a large number of SMILES. The resultant latent space became a generative model. Sampling on the latent vectors and running through the decoder yield new SMILES.

SMILES strings do not directly reflect the structural similarities between molecules. Moreover, a molecule may have multiple SMILES representations, though canonization algorithms exist [200]. As a consequence, the generated molecules lack diversity and validity. Simonovsky et al. [158] proposed the GraphVAE model, which was formulated in the framework of VAE to generate molecular graphs for small molecules by predicting adjacency matrices of molecular graphs.

Some methods only generate molecules that have the same number of atoms because they can be trained using such a dataset. For instance, the GraphVAE was trained on graph data, where each graph example represented a molecule with nodes corresponding to atoms and edges reflecting the bond between any two atoms. The VAE model learns an adjacency matrix of fixed size to characterize the connectivities among the atoms, so all the training molecules had the same number of atoms. In reality, molecules that have similar structural-activity properties can have a different number of atoms and bonds. Samanta et al. [201] developed NeVAE, another deep generative model for molecular graphs, and this model allowed for graphs with different numbers of nodes and edges.

GANs are deep generative models composed of two networks: a generator and a discriminator. The generative network generates candidates, whereas the discriminator network evaluates them against the truly observed molecules. Cao et al. [159] proposed the molecular GAN (MolGAN) model,

which was the first to address the generation of graph-structured data in the context of molecular synthesis using GANs. The model combined a reinforcement learning (RL) objective to encourage the generation of molecules with specific desired chemical properties by providing a reward to the created molecules in a real-time fashion during an iterative process. The limitation of MolGAN was the susceptibility to mode-collapse: both the GAN and the RL objective did not encourage the diversity of created molecules, thus the model's tendency to be pulled towards a solution that only involves little sample variability.

Adversarial autoencoders use the GAN framework as a variational inference algorithm for both discrete and continuous latent variables in probabilistic autoencoders. Kadurin et al. [114] proposed a deep adversarial autoencoder model for identification and generation of new compounds that made a use of available biological and chemical data. The model used the NCI-60 cell line assay data for 6252 compounds profiled on MCF-7 cell line. The model output was used to screen 72 million compounds in PubChem and selected candidate molecules with potential anticancer properties. A successful example using GANs for molecular generation is the ORGANIC framework, as shown in Figure 11d, which aims to utilize GANs for molecular generation and RL for biasing lead candidates [25].

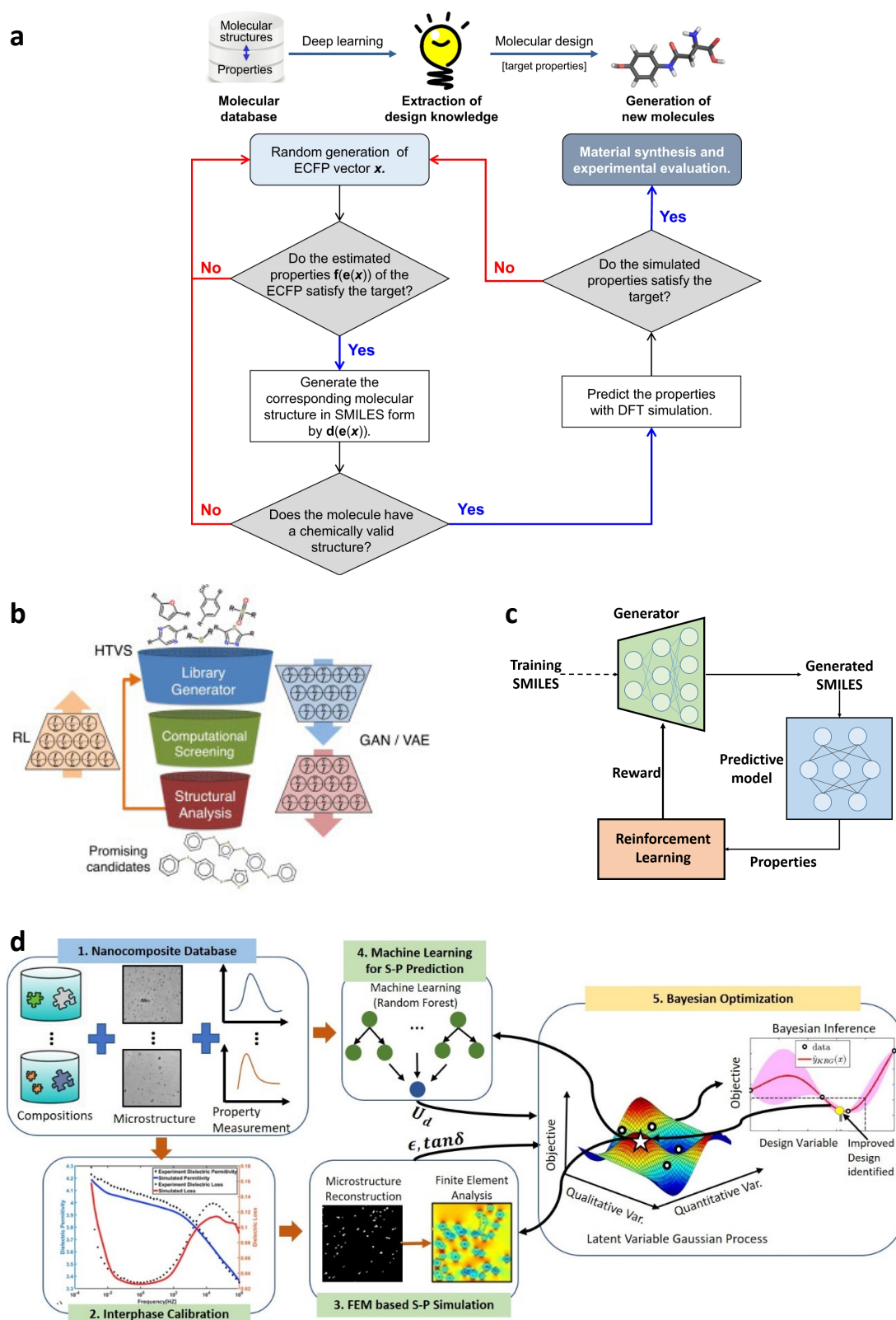
RL learns essential rules of a system by interacting with the environment, where the system resides. It learns how to take suitable actions for a given system state to maximize the expected reward in the particular situation. As is shown in the de novo drug design case, Popova et al. [116] applied RL techniques on top of a string generator to generate the SMILES strings of molecules. They successfully generated molecules with given desirable properties, though they struggled with chemical validity. You et al. [202] proposed a graph convolutional policy network (GCPN) for molecular graph generation through RL. The model was trained to optimize domain-specific rewards and adversarial loss through policy gradient, and act in an environment that incorporated domain-specific rules. The resultant model achieved 61% improvement on chemical property optimization over state-of-the-art baselines (such as GraphVAE and JT-VAE). Zhavoronkov et al. [203] developed a deep generative model, generative tensorial RL (GNETRL), for de novo small molecule design. They used GNETRL to discover potent inhibitors of discoidin domain receptor 1 (DDR1). Four compounds were active in biochemical assays, and two were validated in cell-based assays. A lead candidate was tested and demonstrated favorable pharmacokinetics in mice.

### 3.4. Inverse Materials Design

There are several levels of inverse materials design by ML models. The first level is to use trained ML model as a screening tool for high-throughput screening. In this design method, molecular generation and candidate screening are two major steps. In the second level, the two steps are combined into one cycle using RL, which biases the integrated model toward generation of valid molecules with desired properties directly. As an active learning strategy, Bayesian optimization (BO) [204] could enable an adaptive model when searching the whole chemical space with multiple objectives [24], which is an example of the third level.

#### 3.4.1. Materials Design by High-Throughput Screening

In materials design by high-throughput screening, an ML model mapping molecular structures to their properties is established as a screening tool to filter the molecular library. An example of this design process is demonstrated in Figure 12a. In this framework, a predictive ML model is constructed first, which is then used to screen molecules generated by ML model using ECFP featurization by discriminating whether the generated molecules have the desired properties. It is supplemented by other screening constraints until it passes validation by synthesis and experimental evaluation.



**Figure 12.** Illustration of materials design approaches. (a) ML-assisted materials screening (adapted from Reference [205] with permission, copyright 2018 Springer Nature); (b) high-throughput virtual screening integrated with ML models (adapted from Reference [206] with permission, copyright 2019 Elsevier) and (c) inverse molecular design by RL; (d) integration of various modules for design of insulating nanocomposites by Bayesian optimization (BO). ECFP = extended connectivity fingerprints.

However, this protocol suffers from several shortcomings. First, the candidate molecules are generated randomly without sufficiently learning the underlying grammar and rules of chemically valid structures, which could result in a high percentage of invalid molecules. Thus, it has to determine the legitimacy of generated molecules one by one, which would take enormous additional efforts. Additionally, the way of randomly generating molecules may not search through the whole chemical space uniformly and efficiently. Besides, virtual screening by DFT calculation is introduced which contradicts with the starting point of using ML models to avoid time-consuming modeling and simulation, and eventually to realize accelerated materials design. Last but not least, the development of predictive ML models for screening is separated from molecular generation, which is not an efficient approach since the percentage of desired molecules in the whole chemical space is low.

#### 3.4.2. Reinforcement Learning for Materials Design

In the second level of materials design, molecular generation is combined with molecular discrimination using RL which allows for accelerated materials design, as illustrated in Figure 12b,c. A generative model, by using either molecular graph or SMILES representation, is trained to learn the grammar of valid molecules, which can be realized by the GAN or VAE model. And then the generated molecules are judged by the predictive ML model. The RL model decides to give a reward or a penalty based on the properties predicted. Using RL, it tends to maximize the rewards, which finally biases the molecular generation process towards the space of molecules with target properties. By using such a data-driven method that attempts to learn the underlying probability distribution over a large set of chemical structures, the search over the chemical space can be reduced to only molecules seen as reasonable, without introducing the rigidity of rule based approaches.

However, this protocol suffers from a few drawbacks. First of all, when using SMILES representation, the rewards or penalties can only be given after the sequence of SMILES is generated [24], which does not allow timely feedback to generate valid SMILES. Furthermore, the use of SMILES sequence affects the capability of RL since a syntactically invalid SMILES sequence may represent a valid chemical structure [24]. Lastly, designing the reward functions is difficult when many objectives are presented, i.e., multiple properties are to be optimized. For multi-variable optimization problems, Bayesian optimization is a favorable method [207,208].

#### 3.4.3. Bayesian Optimization for Materials Design

Bayesian optimization (BO), an adaptive sampling approach driven by ML model and uncertainty quantification (UQ), has recently received significant interest in materials design. Among the many global optimization methods reported in literature, BO stands out due to its capability of locating the global optima for highly non-linear functions within tens of objective function (i.e., material properties) evaluations. BO accomplishes this by repeating these three steps [207,208]:

1. An ML model is trained on available data to predict material property of interest from the design variables and supply uncertainty quantification over the design space.
2. An acquisition function uses the prediction and UQ to determine the best design to evaluate next.
3. The design recommended by acquisition function is evaluated and added to the dataset.

This procedure is usually terminated after user-specified maximum iterations are completed. Gaussian Process modeling [209,210] is a popular choice of ML model for BO due to its inherent ability for UQ without incurring additional computational cost, although RF [211] and ensembles of SVM [212] have also been used in the past. The acquisition function essentially gauges the benefit of evaluating a design by interrogating the ML model predictions and associated uncertainties. The acquisition function must decide between exploration and exploitation of design space, which may be contradictory goals. The best performing acquisition functions generally strike a balance between the two. Commonly used acquisition functions are expected improvement [213], Probability of

Improvement [214], and knowledge gradient [215]. The review article by Shahriari et al. [216] provides detailed discussions about ML models and acquisitions functions used in BO.

BO enables significant acceleration of materials design and discovery, as demonstrated by its successful application across a wide variety of materials covering alloys [217], polymers [208, 218], inorganic compounds [219–221], and drug-like molecules [222–224]. A recent article by Lookman et al. [225] reviews application of BO in materials science and highlights existing challenges.

Here, we use the polymer nanocomposite as an example to demonstrate and discuss the BO for inverse materials design. Polymer nanocomposites are attractive candidates for electrical insulation [226,227]. The three key electrical properties of interest for this application—breakdown strength ( $U_d$ ), dielectric permittivity ( $\epsilon$ ) and loss ( $\tan \delta$ )—can be tuned by careful design of constituents, their composition and nanocomposite morphology. This was demonstrated by Iyer et al. [208] using the aforementioned BO approach, with a framework that integrates experimental data in different modules of design process.

Their search space consisted of two polymers (polystyrene and polymethylmethacrylate), silica nanoparticles with three distinct monofunctional silane coupling agents (octyldimethylmethoxysilane, chloropropylethoxysilane, aminopropylethoxysilane), and infinitely many possibilities of nanocomposite microstructure. The salient microstructural features are captured by nanoparticle volume fraction (composition descriptor) and the scale parameter (dispersion descriptor) obtained from Spectral Density Function (SDF) [134,228]. The ability of represent microstructures using only a few variables is a critical aspect in materials design and has received significant interests in the last few decades. Bostanabad et al. [119] reviewed the prevalent techniques in this area and discuss their applicability.

Figure 12d illustrates the flow of information between state-of-the-art computational and experimental techniques leveraged in this design process. It commences by preparing a database (Module 1) containing TEM images and measured dielectric properties for the aforementioned polymer–surface coupling agent combinations. All TEM images were binarized using Niblack algorithm [229] and analyzed to evaluate nanoparticle volume fraction and dispersion. This process helps determine the bounds for the two microstructural design variables. The measured dielectric permittivity and loss are used to calibrate interphase dielectric properties (Module 2) for each polymer–surface coupling agent combination, through the procedure described in Wang et al. [230]. The calibrated interphase parameters are then used in a finite element program [231] (Module 3) to evaluate  $\epsilon$ ,  $\tan \delta$  for microstructures arising in the iterative BO loop. Evaluating breakdown strength is computationally expensive and circumvented by training a Random Forrest model on experimental data to predict  $U_d$  from design variables.

Due to presence of qualitative (choice of polymer, surface coupling) and quantitative (nanoparticle volume fraction and scale parameter) design variables, latent variable Gaussian process [232]—a novel ML method well suited for mixed variable BO—is employed. The aim is to maximize  $U_d$  and minimize  $\epsilon$ ,  $\tan \delta$  for electrical insulation. When formulated as a single objective problem, results show that BO outperforms genetic algorithm [233] by consistently identifying global optimum within 100 iterations. When formulated as a multi-objective problem, multi-objective BO identifies several designs on the Pareto frontier. All Pareto designs comprise polystyrene, clearly indicating its favorability over polymethylmethacrylate for electrical insulation. Additionally, high (low) filler volume fraction and dispersion leads to high (low)  $\tan \delta$ ,  $\epsilon$  and  $U_d$ . The designer may choose any Pareto design as the solution based on his/her preferences.

#### 4. Conclusions

Machine learning, especially deep learning techniques, have emerged as a new and powerful method to accelerate materials design in various fields. The growth of materials databases enclosing chemical structures and corresponding properties provides huge potential to use ML-guided methods for design of organic molecules and polymers. The adequacy of data ensures that ML models



sufficiently learn the underlying mapping between structures and properties, which guarantees the accuracy of QSPR/QSAR for forward property prediction. Additionally, ML models can also be used for inverse materials design in an accelerated manner. Since ML models are able to learn the syntax of chemically meaningful representations of molecules, they can be used to generate novel molecules in a more efficient way than the classical methods. Thus, it has many advantages to deal with issues associated with inverse materials design for organic molecules and polymers.

In this work, we have reviewed recent progress using ML-guided materials design in chemical, biomedical, and materials science fields. Nine representative design examples are highlighted and examined. More importantly, as scale of materials design influences the ultimate performance of the materials, both molecular and microstructure designs of OPVs are presented and discussed. Challenges and issues associated with ML models to tackle materials design problems are discussed. Specifically, the challenges are as follows:

- (i) *Acquisition of a diverse database.* There are many public databases available for various materials, such as the ones summarized in Table 2. If no database of interest is available, we can build one by experiments or simulations. As a result, it is generally not challenging to acquire a database, rather it is challenging to obtain a “good” one. “Good” means that the database is diverse or uniform across the chemical space [140] since this feature of a database significantly affects the capabilities (interpolation, extrapolation, and exploration) of the ML model to be built. With a diverse or uniform database in the chemical space, the ML model guarantees the prediction by interpolation, while with a database in a limited region or class, the prediction is weakened by extrapolation. However, since the whole chemical space is nearly infinite and not clearly known, how can we determine if the database is uniform or not? To overcome this challenge, two areas of algorithmic approaches should be considered [140]: algorithms to perform searches, and more general machine learning and statistical modeling algorithms to predict the chemistry under investigation. The combination of these approaches should be capable of navigating and searching chemical space more efficiently, uniformly, quickly and, importantly, without bias [140].
- (ii) *Feature representation.* Most ML models need all inputs and outputs to be numeric, which requires the data to be represented in a digital form. Many types of representation methods are widely used, such as molecular descriptors, SMILES, and fingerprints, as summarized in Table 3. However, are they universal for all property predictions? Taking fingerprints as an example, it is known that different functional groups (substructures) of a complex structure may have distinct influences on the properties. Therefore, if one fingerprint method with certain bits does demonstrate predictive power in one property prediction, will it have the same capability in another property prediction? In addition, which representation is more suitable to work with specific ML models so that the model can have strong predictive capability? All of these questions require us to be cautious for the feature representation, selection, and extraction by applying the ML models for different materials and properties.
- (iii) *ML algorithms and training.* When conducting a materials design task, the choice of a suitable ML model should be carefully considered. There are many available ML models to choose as reviewed in the Discussion section, but it is not as easy as just to choose any one randomly. Choosing a suitable ML model depends on the database availability and the feature representation method. Which ML model is the best for a certain material property prediction? Does it depend on the type of materials? Can a model that is built with strong predictive power for one material be applicable to other similar but different materials? What about applying to a totally different material? Additionally, when training the selected ML model, there are usually some hyperparameters to be set. It is not trivial to set them without any knowledge of the ML algorithms. In order for the ML model to have better predictive power, the setting of these hyperparameters needs learning efforts, from the user’s point of view.
- (iv) *Interpretation of results.* ML models do show good prediction power in some cases. However, how to explain the constructed model, for example, the DNN model, is still an open question

even in the field of computer science. When applying ML models to materials design, is there any unified theory to physically or chemically interpret the relationship established between a chemical structure to its properties? Can the model built increase our understanding of materials? What role should we consider ML models to be in materials design?

- (v) *Molecular generation.* Molecular generation plays an important role in the design of de novo organic molecules and polymers. As we have discussed, there are several deep generative models, including generative adversarial networks, variational autoencoders, and autoregressive models, rapidly growing for the discovery of new organic molecules and materials [24,60,93]. It is very important to benchmark these different deep generative models for their efficiency and accuracy. Very recently, Zhavoronkov and co-workers have proposed MOlecular SEtS (MOSES) as a platform to benchmark different ML techniques for drug discovery [234]. Such a platform is extremely helpful and useful to standardize the research on the molecular generation and facilitate the sharing and comparison of new ML models. Therefore, more efforts are needed to further design and maintain these benchmark platforms for organic molecules and polymers.
- (vi) *Inverse molecular/materials design.* Currently, RL has been widely used for the inverse molecular/materials design, due to its ease of integration with deep generative ML models [25,36,116]. RL usually involves the analysis of possible actions and outcomes, as well as estimation of the statistical relationship between these actions and possible outcomes. By defining the policy or reward function, the RL can be used to bias the generation of organic molecules towards most desirable domain [24,25,116]. Nevertheless, the inverse design of new molecules and materials typically requires multi-objective optimization of several target properties concurrently. For instance, drug-like molecules should be optimized with respect to potency, selectivity, solubility, and drug-likeness properties for drug discovery [116]. Such a multi-objective optimization problem poses significant challenges for the RL technique [235–237], combined with the huge design space of organic molecules. Comparing with RL technique, BO is more suitable and effective for multi-objective optimization and multi-point search [238–240]. Yet, the design of new molecules and materials involve both continuous/discretized and qualitative/quantitative design variables, representing molecular constituents, material compositions, microstructure morphology, and processing conditions. For these mixed variable design optimization problems, the existing BO approaches are usually restrictive theoretically and fail to capture complex correlations between input variable and output properties [207,208,232]. Therefore, new RL or BO methods should be formulated and developed to resolve these issues.

Nevertheless, with these successful design examples summarized in Table 1, we are positive that the gap between the promise of ML-assisted materials design approaches and their broad applications in reality will be minimized. We anticipate that ML-assisted materials design for organic molecules and polymers will be the driving force in the near future, when accelerated materials design is fully realized to meet the tremendous demand of new materials with tailored properties in different fields.

**Author Contributions:** Y.L. and W.C. conceived and designed this study; G.C. and Z.S. performed the case studies on ML-assisted materials design and discussions; A.I. wrote the BO approach for inverse materials design; U.F.G. wrote the case study for microstructure design of OPVC; S.T. helped to analyze the data; J.B. wrote the molecular generation and ML models; Y.L. and W.C. guided the work of G.C. and G.C. wrote initial draft of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** W.C. would like to acknowledge the support from NSF grants (CMMI-1729743, EEC-1530734, and CMMI-1662435) and Center for Hierarchical Materials Design 70NANB19H005. Y.L. would like to thank the support from NSF grants (OAC-1755779, CMMI-1762661 and CMMI-1934829).

**Acknowledgments:** G.C., Z.S. and Y.L. are grateful for support from the Department of Mechanical Engineering at the University of Connecticut. This work was partially supported by a fellowship grant (to Z.S.) from GE's Industrial Solutions Business Unit under a GE-UConn partnership agreement. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Industrial Solutions or University of Connecticut. This research benefited from the computational resources and staff contributions provided by the Booth Engineering Center for

Advanced Technology (BECAT) at the University of Connecticut. Last but not least, the authors thank Alessandro Fisher and Jason Yang for proofreading of this work.

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

## References

1. Brazel, C.S.; Rosen, S.L. *Fundamental Principles of Polymeric Materials*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
2. Lei, T.; Wang, J.Y.; Pei, J. Roles of flexible chains in organic semiconducting materials. *Chem. Mater.* **2013**, *26*, 594–603. [[CrossRef](#)]
3. Afzal, M.A.F. From Virtual High-Throughput Screening and Machine Learning to the Discovery and Rational Design of Polymers for Optical Applications. Ph.D. Thesis, State University of New York at Buffalo, Buffalo, NY, USA, 2018.
4. Kippelen, B.; Brédas, J.L. Organic photovoltaics. *Energy Environ. Sci.* **2009**, *2*, 251–261. [[CrossRef](#)]
5. Schmidt-Mende, L.; Fechtenkötter, A.; Müllen, K.; Moons, E.; Friend, R.H.; MacKenzie, J.D. Self-organized discotic liquid crystals for high-efficiency organic photovoltaics. *Science* **2001**, *293*, 1119–1122. [[CrossRef](#)] [[PubMed](#)]
6. Brabec, C.J. Organic photovoltaics: Technology and market. *Sol. Energy Mater. Sol. Cells* **2004**, *83*, 273–292. [[CrossRef](#)]
7. Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A.A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; et al. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **2019**, *5*, eaay4275. [[CrossRef](#)]
8. Wang, C.; Pilania, G.; Boggs, S.; Kumar, S.; Breneman, C.; Ramprasad, R. Computational strategies for polymer dielectrics design. *Polymer* **2014**, *55*, 979–988. [[CrossRef](#)]
9. Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T.D.; Lookman, T.; Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **2016**, *6*, 20952. [[CrossRef](#)]
10. Zhou, H.C.; Long, J.R.; Yaghi, O.M. Introduction to metal–organic frameworks. *Chem. Rev.* **2012**, *112*, 673–674. [[CrossRef](#)]
11. James, S.L. Metal-organic frameworks. *Chem. Soc. Rev.* **2003**, *32*, 276–288. [[CrossRef](#)]
12. Furukawa, H.; Cordova, K.E.; O’Keeffe, M.; Yaghi, O.M. The chemistry and applications of metal-organic frameworks. *Science* **2013**, *341*, 1230444. [[CrossRef](#)]
13. Bucior, B.J.; Rosen, A.S.; Haranczyk, M.; Yao, Z.; Ziebel, M.E.; Farha, O.K.; Hupp, J.T.; Siepmann, J.I.; Aspuru-Guzik, A.; Snurr, R.Q. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19*, 6682–6697. [[CrossRef](#)]
14. Burroughes, J.H.; Bradley, D.D.; Brown, A.; Marks, R.; Mackay, K.; Friend, R.H.; Burns, P.; Holmes, A. Light-emitting diodes based on conjugated polymers. *Nature* **1990**, *347*, 539. [[CrossRef](#)]
15. Gross, M.; Müller, D.C.; Nothofer, H.G.; Scherf, U.; Neher, D.; Bräuchle, C.; Meerholz, K. Improving the performance of doped  $\pi$ -conjugated polymers for use in organic light-emitting diodes. *Nature* **2000**, *405*, 661. [[CrossRef](#)]
16. Agrawal, J.P. Recent trends in high-energy materials. *Prog. Energy Combust. Sci.* **1998**, *24*, 1–30. [[CrossRef](#)]
17. Talawar, M.; Sivabalan, R.; Mukundan, T.; Muthurajan, H.; Sikder, A.; Gandhe, B.; Rao, A.S. Environmentally compatible next generation green energetic materials (GEMs). *J. Hazard. Mater.* **2009**, *161*, 589–607. [[CrossRef](#)]
18. Bushuyev, O.S.; Brown, P.; Maiti, A.; Gee, R.H.; Peterson, G.R.; Weeks, B.L.; Hope-Weeks, L.J. Ionic polymers as a new structural motif for high-energy-density materials. *J. Am. Chem. Soc.* **2012**, *134*, 1422–1425. [[CrossRef](#)]
19. Kuntz, I.D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078–1082. [[CrossRef](#)]
20. Silverman, R.B.; Holladay, M.W. *The Organic Chemistry of Drug Design and Drug Action*; Academic Press: Cambridge, MA, USA, 2014.
21. Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520–10594. [[CrossRef](#)]
22. 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](#)]

23. 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](#)]
24. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365. [[CrossRef](#)]
25. 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**. [[CrossRef](#)]
26. 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.
27. Parr, R.G. Density functional theory of atoms and molecules. In *Horizons of Quantum Chemistry*; Springer: Berlin/Heidelberg, Germany, 1980; pp. 5–15.
28. Cohen, A.J.; Mori-Sánchez, P.; Yang, W. Insights into current limitations of density functional theory. *Science* **2008**, *321*, 792–794. [[CrossRef](#)] [[PubMed](#)]
29. Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: Cambridge, MA, USA, 2001.
30. Rapaport, D.C.; Rapaport, D.C.R. *The Art of Molecular Dynamics Simulation*; Cambridge University Press: Cambridge, UK, 2004.
31. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476–488. [[CrossRef](#)] [[PubMed](#)]
32. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010. [[CrossRef](#)] [[PubMed](#)]
33. Churchwell, C.J.; Rintoul, M.D.; Martin, S.; Visco, D.P., Jr.; Kotu, A.; Larson, R.S.; Sillerud, L.O.; Brown, D.C.; Faulon, J.L. The signature molecular descriptor: 3. Inverse-quantitative structure–activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graph. Modell.* **2004**, *22*, 263–273. [[CrossRef](#)] [[PubMed](#)]
34. Wong, W.W.; Burkowski, F.J. A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *J. Cheminform.* **2009**, *1*, 4. [[CrossRef](#)]
35. Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR analysis for chemical structure generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286–299. [[CrossRef](#)]
36. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484. [[CrossRef](#)]
37. Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **2016**, *4*, 053208. [[CrossRef](#)]
38. Gil, Y.; Greaves, M.; Hendler, J.; Hirsh, H. Amplify scientific discovery with artificial intelligence. *Science* **2014**, *346*, 171–172. [[CrossRef](#)]
39. Rajan, K. *Informatics for Materials Science and Engineering: Data-driven Discovery for Accelerated Experimentation and Application*; Butterworth-Heinemann: Oxford, UK, 2013.
40. Sarkisov, L.; Kim, J. Computational structure characterization tools for the era of material informatics. *Chem. Eng. Sci.* **2015**, *121*, 322–330. [[CrossRef](#)]
41. Adams, N. Polymer informatics. In *Polymer Libraries*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 107–149.
42. Audus, D.J.; de Pablo, J.J. Polymer Informatics: Opportunities and Challenges. *ACS Macro. Lett.* **2017**, *6*, 1078–1082. [[CrossRef](#)] [[PubMed](#)]
43. Kim, C.; Chandrasekaran, A.; Huan, T.D.; Das, D.; Ramprasad, R. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **2018**, *122*, 17575–17585. [[CrossRef](#)]
44. Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T.D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785–796. [[CrossRef](#)]
45. Council, N.R. *Advanced Energetic Materials*; National Academics Press: Washington, DC, USA, 2004.
46. Pagoria, P. A comparison of the structure, synthesis, and properties of insensitive energetic compounds. *Propellants Explos. Pyrotech.* **2016**, *41*, 452–469. [[CrossRef](#)]

47. Nielsen, A.T.; Chafin, A.P.; Christian, S.L.; Moore, D.W.; Nadler, M.P.; Nissan, R.A.; Vanderah, D.J.; Gilardi, R.D.; George, C.F.; Flippen-Anderson, J.L. Synthesis of polyazapolycyclic caged polynitramines. *Tetrahedron* **1998**, *54*, 11793–11812. [[CrossRef](#)]
48. White, A. The materials genome initiative: One year on. *MRS Bull.* **2012**, *37*, 715–716. [[CrossRef](#)]
49. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002. [[CrossRef](#)]
50. de Pablo, J.J.; Jones, B.; Kovacs, C.L.; Ozolins, V.; Ramirez, A.P. The materials genome initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **2014**, *18*, 99–117. [[CrossRef](#)]
51. Green, M.L.; Choi, C.; Hatrick-Simpers, J.; Joshi, A.; Takeuchi, I.; Barron, S.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J.; et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.* **2017**, *4*, 011105. [[CrossRef](#)]
52. Wang, Y.; Liu, Y.; Song, S.; Yang, Z.; Qi, X.; Wang, K.; Liu, Y.; Zhang, Q.; Tian, Y. Accelerating the discovery of insensitive high-energy-density materials by a materials genome approach. *Nat. Commun.* **2018**, *9*, 2444. [[CrossRef](#)] [[PubMed](#)]
53. Gubernatis, J.; Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2018**, *2*, 120301. [[CrossRef](#)]
54. 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](#)]
55. Van Den Herik, H.J.; Uiterwijk, J.W.; Van Rijswijk, J. Games solved: Now and in the future. *Artif. Intell.* **2002**, *134*, 277–311. [[CrossRef](#)]
56. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials discovery and design using machine learning. *J. Mater.* **2017**, *3*, 159–177. [[CrossRef](#)]
57. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [[CrossRef](#)]
58. Ramprasad, R.; Batra, R.; Pilia, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: Recent applications and prospects. *Npj Comput. Mater.* **2017**, *3*, 54. [[CrossRef](#)]
59. Kailkhura, B.; Gallagher, B.; Kim, S.; Hiszpanski, A.; Han, T.Y.J. Reliable and explainable machine-learning methods for accelerated material discovery. *Npj Comput. Mater.* **2019**, *5*, 1–9. [[CrossRef](#)]
60. 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](#)]
61. Xu, Y.; Yao, H.; Lin, K. An overview of neural networks for drug discovery and the inputs used. *Expert Opin. Drug Discov.* **2018**, *13*, 1091–1102. [[CrossRef](#)] [[PubMed](#)]
62. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250. [[CrossRef](#)] [[PubMed](#)]
63. Dimitrov, T.; Kreisbeck, C.; Becker, J.S.; Aspuru-Guzik, A.; Saikin, S.K. Autonomous molecular design: Then and now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825–24836. [[CrossRef](#)] [[PubMed](#)]
64. Kumar, J.N.; Li, Q.; Jun, Y. Challenges and opportunities of polymer design with machine learning and high throughput experimentation. *MRS Commun.* **2019**, 1–8. [[CrossRef](#)]
65. Outlook, A.E. Energy information administration. *Dep. Energy* **2010**, 92010, 1–15.
66. Gaudiana, R. Third-generation photovoltaic technology- the potential for low-cost solar energy conversion. *J. Phys. Chem. Lett.* **2010**, *1*, 1288–1289. [[CrossRef](#)]
67. Imamzai, M.; Aghaei, M.; Thayoob, Y.H.M.; Forouzanfar, M. A review on comparison between traditional silicon solar cells and thin-film CdTe solar cells. In Proceedings of the National Graduate Conference (Nat-Grad, 2012), Kajang, Malaya, 8–10 November 2012; pp. 1–5.
68. Heeger, A.J. Semiconducting polymers: The third generation. *Chem. Soc. Rev.* **2010**, *39*, 2354–2371. [[CrossRef](#)]
69. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R.S.; Gold-Parker, A.; Vogt, L.; Brockway, A.M.; Aspuru-Guzik, A. The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251. [[CrossRef](#)]

70. Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sanchez-Carrera, R.S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849–4861. [[CrossRef](#)]
71. ChemAxon. Available online: <https://www.chemaxon.com/> (accessed on 3 September 2019).
72. Kim, C.; Wang, Z.; Choi, H.J.; Ha, Y.G.; Facchetti, A.; Marks, T.J. Printable cross-linked polymer blend dielectrics. Design strategies, synthesis, microstructures, and electrical properties, with organic field-effect transistors as testbeds. *J. Am. Chem. Soc.* **2008**, *130*, 6867–6878. [[CrossRef](#)]
73. Müller, K.; Paloumpa, I.; Henkel, K.; Schmeisser, D. A polymer high-k dielectric insulator for organic field-effect transistors. *J. Appl. Phys.* **2005**, *98*, 056104. [[CrossRef](#)]
74. Mannodi-Kanakkithodi, A.; Treich, G.M.; Huan, T.D.; Ma, R.; Tefferi, M.; Cao, Y.; Sotzing, G.A.; Ramprasad, R. Rational Co-Design of Polymer Dielectrics for Energy Storage. *Adv. Mater.* **2016**, *28*, 6277–6291. [[CrossRef](#)] [[PubMed](#)]
75. Sharma, V.; Wang, C.; Lorenzini, R.G.; Ma, R.; Zhu, Q.; Sinkovits, D.W.; Pilania, G.; Oganov, A.R.; Kumar, S.; Sotzing, G.A.; et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **2014**, *5*, 4845. [[CrossRef](#)] [[PubMed](#)]
76. Miller, R.L. Crystallographic data and melting points for various polymers. *Wiley Database Polym. Prop.* **2003**. [[CrossRef](#)]
77. Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911–9917. [[CrossRef](#)]
78. Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810. [[CrossRef](#)]
79. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Dielectr. Electr. Insul.* **2002**, *6*, 182–197. [[CrossRef](#)]
80. Jou, J.H.; Kumar, S.; Agrawal, A.; Li, T.H.; Sahoo, S. Approaches for fabricating high efficiency organic light emitting diodes. *J. Mater. Chem. C* **2015**, *3*, 2974–3002. [[CrossRef](#)]
81. Tao, Y.; Yuan, K.; Chen, T.; Xu, P.; Li, H.; Chen, R.; Zheng, C.; Zhang, L.; Huang, W. Thermally activated delayed fluorescence materials towards the breakthrough of organoelectronics. *Adv. Mater.* **2014**, *26*, 7931–7958. [[CrossRef](#)]
82. Zhang, Q.; Li, B.; Huang, S.; Nomura, H.; Tanaka, H.; Adachi, C. Efficient blue organic light-emitting diodes employing thermally activated delayed fluorescence. *Nat. Photonics* **2014**, *8*, 326. [[CrossRef](#)]
83. Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M.A.; Chae, H.S.; Einzinger, M.; Ha, D.G.; Wu, T.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120. [[CrossRef](#)]
84. Olsson, T.; Oprea, T.I. Cheminformatics: A tool for decision-makers in drug discovery. *Curr. Opin. Drug Discov. Dev.* **2001**, *4*, 308–313.
85. Akella, L.B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325–330. [[CrossRef](#)]
86. Landrum, G. Rdkit documentation. *Release* **2013**, *1*, 1–79.
87. 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](#)]
88. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [[CrossRef](#)]
89. Jørgensen, P.B.; Mesta, M.; Shil, S.; García Lastra, J.M.; Jacobsen, K.W.; Thygesen, K.S.; Schmidt, M.N. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **2018**, *148*, 241735. [[CrossRef](#)]
90. Rupp, M.; Tkatchenko, A.; Müller, K.R.; Von Lilienfeld, O.A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301. [[CrossRef](#)]
91. 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](#)]
92. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O.A.; Müller, 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](#)]

93. Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future Med. Chem.* **2019**, *11*, 567–597. [[CrossRef](#)]
94. Barnes, B.C.; Elton, D.C.; Boukouvalas, Z.; Taylor, D.E.; Mattson, W.D.; Fuge, M.D.; Chung, P.W. Machine learning of energetic material properties. *arXiv* **2018**, arXiv:1807.06156.
95. Huang, L.; Massa, L. Applications of energetic materials by a theoretical method (discover energetic materials by a theoretical method). *Int. J. Energ. Mater. Chem. Propul.* **2013**, *12*, 197–262. [[CrossRef](#)]
96. Mathieu, D. Sensitivity of energetic materials: Theoretical relationships to detonation performance and molecular structure. *Ind. Eng. Chem. Res.* **2017**, *56*, 8191–8201. [[CrossRef](#)]
97. Ravi, P.; Gore, G.M.; Tewari, S.P.; Sikder, A.K. DFT study on the structure and explosive properties of nitropyrazoles. *Mol. Simul.* **2012**, *38*, 218–226. [[CrossRef](#)]
98. Liu, J.g.; Ueda, M. High refractive index polymers: Fundamental research and practical applications. *J. Mater. Chem.* **2009**, *19*, 8907–8919. [[CrossRef](#)]
99. Odian, G. *Principles of Polymerization*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
100. Afzal, M.A.F.; Haghghatlari, M.; Prasad Ganesh, S.; Cheng, C.; Hachmann, J. Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining. *J. Phys. Chem. C* **2019**, *123*, 14610–14618. [[CrossRef](#)]
101. Slonimskii, G.; Askadskii, A.; Kitaigorodskii, A. The packing of polymer molecules. *Polym. Sci. USSR* **1970**, *12*, 556–577. [[CrossRef](#)]
102. Afzal, M.A.F.; Cheng, C.; Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *J. Chem. Phys.* **2018**, *148*, 241712. [[CrossRef](#)]
103. Hansson, J.; Nilsson, T.M.; Ye, L.; Liu, J. Novel nanostructured thermal interface materials: A review. *Int. Mater. Rev.* **2018**, *63*, 22–45. [[CrossRef](#)]
104. Razeeb, K.M.; Dalton, E.; Cross, G.L.W.; Robinson, A.J. Present and future thermal interface materials for electronic devices. *Int. Mater. Rev.* **2018**, *63*, 1–21. [[CrossRef](#)]
105. Wan, X.; Feng, W.; Wang, Y.; Wang, H.; Zhang, X.; Deng, C.; Yang, N. Materials Discovery and Properties Prediction in Thermal Transport via Materials Informatics: A Mini Review. *Nano Lett.* **2019**, *19*, 3387–3395. [[CrossRef](#)]
106. Wu, S.; Kondo, Y.; Kakimoto, M.A.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Comput. Mater.* **2019**, *5*, 5. [[CrossRef](#)]
107. Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. In Proceedings of the 2011 International Conference on Emerging Intelligent Data and Web Technologies, Tirana, Albania, 7–9 September 2011; pp. 22–29.
108. Blum, L.C.; Raymond, J.L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733. [[CrossRef](#)]
109. Morikawa, J.; Tan, J.; Hashimoto, T. Study of change in thermal diffusivity of amorphous polymers during glass transition. *Polymer* **1995**, *36*, 4439–4443. [[CrossRef](#)]
110. Allen, P.B.; Feldman, J.L. Thermal conductivity of disordered harmonic solids. *Phys. Rev. B: Condens. Matter* **1993**, *48*, 12581. [[CrossRef](#)]
111. Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 379–391. [[CrossRef](#)]
112. Segler, M.H.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2017**, *4*, 120–131. [[CrossRef](#)]
113. Lim, J.; Ryu, S.; Kim, J.W.; Kim, W.Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform.* **2018**, *10*, 31. [[CrossRef](#)]
114. 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. [[CrossRef](#)]
115. 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](#)]
116. Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885. [[CrossRef](#)]

117. Tetko, I.V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A.E.; Charochkina, L.; Asiri, A.M. How accurately can we predict the melting points of drug-like compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320–3329. [[CrossRef](#)]
118. Lee, W.K.; Yu, S.; Engel, C.J.; Reese, T.; Rhee, D.; Chen, W.; Odom, T.W. Concurrent design of quasi-random photonic nanostructures. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 8734–8739. [[CrossRef](#)]
119. Bostanabad, R.; Zhang, Y.; Li, X.; Kearney, T.; Brinson, L.C.; Apley, D.W.; Liu, W.K.; Chen, W. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. *Prog. Mater. Sci.* **2018**, *95*, 1–41. [[CrossRef](#)]
120. Ghumman, U.F.; Iyer, A.; Dulal, R.; Wang, A.; Munshi, J.; Chien, T.; Balasubramanian, G.; Chen, W. A Spectral Density Function Approach for Design of Organic Photovoltaic Cells. In Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers Digital Collection, Quebec City, QC, Canada, 26–29 August 2018.
121. Munshi, J.; Ghumman, U.F.; Iyer, A.; Dulal, R.; Chen, W.; Chien, T.; Balasubramanian, G. Effect of polydispersity on the bulk-heterojunction morphology of P3HT: PCBM solar cells. *J. Polym. Sci. Part B Polym. Phys.* **2019**. [[CrossRef](#)]
122. Munshi, J.; Dulal, R.; Chien, T.; Chen, W.; Balasubramanian, G. Solution Processing Dependent Bulk Heterojunction Nanomorphology of P3HT/PCBM Thin Films. *ACS Appl. Mater. Interfaces* **2019**, *11*, 17056–17067. [[CrossRef](#)]
123. Olson, G.B. Computational design of hierarchically structured materials. *Science* **1997**, *277*, 1237–1242. [[CrossRef](#)]
124. Gleiter, H. Nanostructured materials: Basic concepts and microstructure. *Acta Mater.* **2000**, *48*, 1–29. [[CrossRef](#)]
125. Biswas, A.; Bayer, I.S.; Biris, A.S.; Wang, T.; Dervishi, E.; Faupel, F. Advances in top–down and bottom–up surface nanofabrication: Techniques, applications & future prospects. *Adv. Colloid Interface Sci.* **2012**, *170*, 2–27. [[PubMed](#)]
126. Brabec, C.; Scherf, U.; Dyakonov, V. *Organic Photovoltaics: Materials, Device Physics, and Manufacturing Technologies*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
127. Brabec, C.J.; Dyakonov, V.; Parisi, J.; Sariciftci, N.S. *Organic Photovoltaics: Concepts and Realization*; Springer: Berlin/Heidelberg, Germany, 2013.
128. Ghumman, U.F.; Iyer, A.; Dulal, R.; Munshi, J.; Wang, A.; Chien, T.; Balasubramanian, G.; Chen, W. A Spectral Density Function Approach for Active Layer Design of Organic Photovoltaic Cells. *J. Mech. Des.* **2018**, *140*, 111408. [[CrossRef](#)]
129. Liu, Y.; Greene, M.S.; Chen, W.; Dikin, D.A.; Liu, W.K. Computational microstructure characterization and reconstruction for stochastic multiscale material design. *Comput.-Aided Des.* **2013**, *45*, 65–76. [[CrossRef](#)]
130. Xu, H.; Li, Y.; Brinson, C.; Chen, W. A descriptor-based design methodology for developing heterogeneous microstructural materials system. *J. Mech. Des.* **2014**, *136*, 051007. [[CrossRef](#)]
131. Yeong, C.; Torquato, S. Reconstructing random media. II. Three-dimensional media from two-dimensional cuts. *Phys. Rev. E: Stat. Nonlinear Soft Matter Phys.* **1998**, *58*, 224. [[CrossRef](#)]
132. Xu, H.; Dikin, D.A.; Burkhart, C.; Chen, W. Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials. *Comput. Mater. Sci.* **2014**, *85*, 206–216. [[CrossRef](#)]
133. Yu, S.; Wang, C.; Zhang, Y.; Dong, B.; Jiang, Z.; Chen, X.; Chen, W.; Sun, C. Design of non-deterministic quasi-random nanophotonic structures using Fourier space representations. *Sci. Rep.* **2017**, *7*, 3752. [[CrossRef](#)]
134. Yu, S.; Zhang, Y.; Wang, C.; Lee, W.k.; Dong, B.; Odom, T.W.; Sun, C.; Chen, W. Characterization and design of functional quasi-random nanostructured materials using spectral density function. *J. Mech. Des.* **2017**, *139*, 071401. [[CrossRef](#)]
135. van Lare, M.C.; Polman, A. Optimized scattering power spectral density of photovoltaic light-trapping patterns. *ACS Photonics* **2015**, *2*, 822–831. [[CrossRef](#)]
136. Lee, W.K.; Jung, W.B.; Nagel, S.R.; Odom, T.W. Stretchable superhydrophobicity from monolithic, three-dimensional hierarchical wrinkles. *Nano Lett.* **2016**, *16*, 3774–3779. [[CrossRef](#)]
137. Kleijnen, J.P. Kriging metamodeling in simulation: A review. *Eur. J. Oper. Res.* **2009**, *192*, 707–716. [[CrossRef](#)]



138. Jin, R.; Chen, W.; Simpson, T.W. Comparative studies of metamodelling techniques under multiple modelling criteria. *Struct. Multidiscip. Optim.* **2001**, *23*, 1–13. [[CrossRef](#)]
139. Zhang, X.Y.; Trame, M.; Lesko, L.; Schmidt, S. Sobol sensitivity analysis: A tool to guide the development and evaluation of systems pharmacology models. *CPT: Pharmacomet. Syst. Pharmacol.* **2015**, *4*, 69–79. [[CrossRef](#)]
140. Gromski, P.S.; Henson, A.B.; Granda, J.M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128. [[CrossRef](#)]
141. Cooper, C.B.; Beard, E.J.; Vázquez-Mayagoitia, Á.; Stan, L.; Stenning, G.B.; Nye, D.W.; Vigil, J.A.; Tomar, T.; Jia, J.; Bodedla, G.B.; et al. Design-to-Device Approach Affords Panchromatic Co-Sensitized Solar Cells. *Adv. Energy Mater.* **2019**, *9*, 1802820. [[CrossRef](#)]
142. Swain, M.C.; Cole, J.M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904. [[CrossRef](#)]
143. Huan, T.D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B: Condens. Matter* **2015**, *92*, 014106. [[CrossRef](#)]
144. Dong, J.; Cao, D.S.; Miao, H.Y.; Liu, S.; Deng, B.C.; Yun, Y.H.; Wang, N.N.; Lu, A.P.; Zeng, W.B.; Chen, A.F. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* **2015**, *7*, 60. [[CrossRef](#)]
145. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, NY, USA, 2000.
146. Puzyn, T.; Leszczynski, J.; Cronin, M.T. *Recent Advances in QSAR Studies: Methods and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
147. Varmuza, K.; Dehmer, M.; Bonchev, D. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
148. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
149. Tauler, R.; Walczak, B.; Brown, S.D. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*; Elsevier: Amsterdam, The Netherlands, 2009.
150. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693–705. [[CrossRef](#)]
151. Todeschini, R.; Gramatica, P. New 3D molecular descriptors: The WHIM theory and QSAR applications. In *3D QSAR in Drug Design*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 355–380.
152. Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and QSPAR*; CRC Press: Boca Raton, FL, USA, 2000.
153. Dragon7. Available online: <https://chm.kode-solutions.net/index.php/> (accessed on 3 September 2019).
154. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **2006**, *56*, 237–248.
155. Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **2016**, *11*, 137–148. [[CrossRef](#)]
156. Tabor, D.P.; Roch, L.M.; Saikin, S.K.; Kreisbeck, C.; Sheberla, D.; Montoya, J.H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **2018**, *3*, 5. [[CrossRef](#)]
157. Li, Y.; Zhang, L.; Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *J. Cheminform.* **2018**, *10*, 33. [[CrossRef](#)]
158. Simonovsky, M.; Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 412–422.
159. De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* **2018**, arXiv:1805.11973.
160. Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained graph variational autoencoders for molecule design. In Proceedings of the Neural Information Processing Systems 2018, Montréal, QC, Canada, 3–8 December 2018; pp. 7795–7804.
161. Batra, R.; Tran, H.D.; Kim, C.; Chapman, J.; Chen, L.; Chandrasekaran, A.; Ramprasad, R. A General Atomic Neighborhood Fingerprint for Machine Learning Based Methods. *J. Phys. Chem. C* **2019**, *123*, 15859–15866. [[CrossRef](#)]

162. Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; Lilienfeld, A.V.; Müller, K.R. Learning invariant representations of molecules for atomization energy prediction. In Proceedings of the Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 440–448.
163. Faber, F.; Lindmaa, A.; von Lilienfeld, O.A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101. [[CrossRef](#)]
164. Himanen, L.; Jäger, M.O.; Morooka, E.V.; Canova, F.F.; Ranawat, Y.S.; Gao, D.Z.; Rinke, P.; Foster, A.S. DDescribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949. [[CrossRef](#)]
165. Bartók, A.P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115. [[CrossRef](#)]
166. De, S.; Bartók, A.P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769. [[CrossRef](#)] [[PubMed](#)]
167. Caro, M.A. Optimizing many-body atomic descriptors for enhanced computational performance of machine-learning-based interatomic potentials. *arXiv* **2019**, arXiv:1905.02142.
168. Behler, J.; Lorenz, S.; Reuter, K. Representing molecule-surface interactions with symmetry-adapted neural networks. *J. Chem. Phys.* **2007**, *127*, 07B603. [[CrossRef](#)] [[PubMed](#)]
169. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106. [[CrossRef](#)] [[PubMed](#)]
170. Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P. wACSF-Weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.* **2018**, *148*, 241709. [[CrossRef](#)]
171. 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](#)] [[PubMed](#)]
172. Sunseri, J.; King, J.E.; Francoeur, P.G.; Koes, D.R. Convolutional neural network scoring and minimization in the D3R 2017 community challenge. *J. Comput.-Aided Mol. Des.* **2019**, *33*, 19–34. [[CrossRef](#)] [[PubMed](#)]
173. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500. [[CrossRef](#)]
174. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E.L. Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120. [[CrossRef](#)]
175. May, J.W.; Steinbeck, C. Efficient ring perception for the Chemistry Development Kit. *J. Cheminform.* **2014**, *6*, 3. [[CrossRef](#)]
176. Willighagen, E.L.; Mayfield, J.W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliaskova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2. 0: Atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*, 33. [[CrossRef](#)]
177. ChemDes. Available online: <http://www.scbdd.com/chemdes/> (accessed on 12 December 2019).
178. Backman, T.W.; Cao, Y.; Girke, T. ChemMine tools: An online service for analyzing and clustering small molecules. *Nucleic Acids Res.* **2011**, *39*, W486–W491. [[CrossRef](#)]
179. ChemMine Tools. Available online: <https://chemminetools.ucr.edu/> (accessed on 12 December 2019).
180. OEChem Toolkit. Available online: <https://docs.eyesopen.com/toolkits/python/oechemtk/index.html> (accessed on 12 December 2019).
181. Stahl, M.; Mauser, H. Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model.* **2005**, *45*, 542–548. [[CrossRef](#)] [[PubMed](#)]
182. Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207. [[CrossRef](#)]
183. Open Babel. Available online: <http://openbabel.org/> (accessed on 12 December 2019).
184. O’Boyle, N.M.; Morley, C.; Hutchison, G.R. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*, 5. [[CrossRef](#)] [[PubMed](#)]
185. O’Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [[CrossRef](#)] [[PubMed](#)]

186. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [[CrossRef](#)] [[PubMed](#)]
187. PaDEL-Descriptor. Available online: <http://www.yapcsoft.com/dd/padeldescriptor/> (accessed on 12 December 2019).
188. PubChemPy. Available online: <https://pubchempy.readthedocs.io/en/latest/> (accessed on 12 December 2019).
189. RDKit. Available online: <https://www.rdkit.org/> (accessed on 12 December 2019).
190. Mueller, T.; Kusne, A.G.; Ramprasad, R. Machine learning in materials science: Recent progress and emerging applications. *Rev. Comput. Chem.* **2016**, *29*, 186–273.
191. Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S.L.; Srivastava, R. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Comb. Sci.* **2017**, *19*, 640–645. [[CrossRef](#)]
192. Wilmer, C.E.; Leaf, M.; Lee, C.Y.; Farha, O.K.; Hauser, B.G.; Hupp, J.T.; Snurr, R.Q. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **2012**, *4*, 83. [[CrossRef](#)]
193. Fink, T.; Reymond, J.L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.
194. McKay, B.D. *Practical Graph Isomorphism*; CRC Press: Boca Raton, FL, USA, 1981.
195. Wang, Y.; Xiao, J.; Suzek, T.O.; Zhang, J.; Wang, J.; Bryant, S.H. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633. [[CrossRef](#)]
196. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100–D1107. [[CrossRef](#)] [[PubMed](#)]
197. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; et al. DrugBank 3.0: A comprehensive resource for ‘omics’ research on drugs. *Nucleic Acids Res.* **2010**, *39*, D1035–D1041. [[CrossRef](#)] [[PubMed](#)]
198. Ruddigkeit, L.; Blum, L.C.; Reymond, J.L. Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2013**, *53*, 56–65. [[CrossRef](#)]
199. Gupta, A.; Müller, A.T.; Huisman, B.J.; Fuchs, J.A.; Schneider, P.; Schneider, G. Generative recurrent networks for de novo drug design. *Mol. Inf.* **2018**, *37*, 1700111. [[CrossRef](#)]
200. O’Boyle, N.M. Towards a Universal SMILES representation—A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **2012**, *4*, 22. [[CrossRef](#)]
201. Samanta, B.; Abir, D.; Jana, G.; Chattaraj, P.K.; Ganguly, N.; Rodriguez, M.G. Nevae: A deep generative model for molecular graphs. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 1110–1117.
202. You, J.; Liu, B.; Ying, Z.; Pande, V.; Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 6410–6421.
203. 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](#)]
204. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [[CrossRef](#)]
205. Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.S.; Jung, Y.; Kim, S.; et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *NPJ Comput. Mater.* **2018**, *4*, 67. [[CrossRef](#)]
206. Haghghatlari, M.; Hachmann, J. Advances of machine learning in molecular modeling and simulation. *Curr. Opin. Chem. Eng.* **2019**, *23*, 51–57. [[CrossRef](#)]
207. Zhang, Y.; Apley, D.; Chen, W. Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables. *arXiv* **2019**, arXiv:1910.01688.
208. Iyer, A.; Zhang, Y.; Prasad, A.; Tao, S.; Wang, Y.; Schadler, L.; Brinson, L.C.; Chen, W. Data-Centric Mixed-Variable Bayesian Optimization For Materials Design. *arXiv* **2019**, arXiv:1907.02577.

209. Rasmussen, C. *Gaussian Processes in Machine Learning, Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003.
210. Bostanabad, R.; Kearney, T.; Tao, S.; Apley, D.W.; Chen, W. Leveraging the nugget parameter for efficient Gaussian process modeling. *Int. J. Numer. Methods Eng.* **2018**, *114*, 501–516. [[CrossRef](#)]
211. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
212. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
213. Mockus, J.; Tiesis, V.; Zilinskas, A. The application of Bayesian methods for seeking the extremum. In *Towards Global Optimisation 2*; North-Holland: Amsterdam, The Netherlands, 1978.
214. Kushner, H.J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.* **1964**, *86*, 97–106. [[CrossRef](#)]
215. Scott, W.; Frazier, P.; Powell, W. The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM J. Optim.* **2011**, *21*, 996–1026. [[CrossRef](#)]
216. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2015**, *104*, 148–175. [[CrossRef](#)]
217. Wen, C.; Zhang, Y.; Wang, C.; Xue, D.; Bai, Y.; Antonov, S.; Dai, L.; Lookman, T.; Su, Y. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* **2019**, *170*, 109–117. [[CrossRef](#)]
218. Li, C.; de Celis Leal, D.R.; Rana, S.; Gupta, S.; Sutti, A.; Greenhill, S.; Slezak, T.; Height, M.; Venkatesh, S. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Sci. Rep.* **2017**, *7*, 5683. [[CrossRef](#)]
219. Xue, D.; Balachandran, P.V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **2016**, *7*, 11241. [[CrossRef](#)]
220. Balachandran, P.V.; Xue, D.; Theiler, J.; Hogden, J.; Lookman, T. Adaptive strategies for materials design using uncertainties. *Sci. Rep.* **2016**, *6*, 19660. [[CrossRef](#)] [[PubMed](#)]
221. Yuan, R.; Tian, Y.; Xue, D.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Lookman, T. Accelerated Search for BaTiO<sub>3</sub>-Based Ceramics with Large Energy Storage at Low Fields Using Machine Learning and Experimental Design. *Adv. Sci.* **2019**, *6*, 1901395. [[CrossRef](#)] [[PubMed](#)]
222. Winkler, D.A.; Burden, F.R. Bayesian neural nets for modeling in drug discovery. *Drug Discov. Today BIOSILICO* **2004**, *2*, 104–111. [[CrossRef](#)]
223. Lavecchia, A. Machine-learning approaches in drug discovery: Methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331. [[CrossRef](#)] [[PubMed](#)]
224. Madhukar, N.S.; Khade, P.K.; Huang, L.; Gayvert, K.; Galletti, G.; Stogniew, M.; Allen, J.E.; Giannakakou, P.; Elemento, O. A Bayesian machine learning approach for drug target identification using diverse data types. *Nat. Commun.* **2019**, *10*, 1–14. [[CrossRef](#)] [[PubMed](#)]
225. Lookman, T.; Balachandran, P.V.; Xue, D.; Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput. Mater.* **2019**, *5*, 21. [[CrossRef](#)]
226. Tanaka, T.; Montanari, G.; Mulhaupt, R. Polymer nanocomposites as dielectrics and electrical insulation—perspectives for processing technologies, material characterization and future applications. *IEEE Trans. Dielectr. Electr. Insul.* **2004**, *11*, 763–784. [[CrossRef](#)]
227. Weidner, J.R.; Pohlmann, F.; Gröppel, P.; Hildinger, T. Nanotechnology in high voltage insulation systems for turbine generators—First results. In Proceedings of the 17th ISH, Hannover, Germany, 22–26 August 2011.
228. Torquato, S.; Haslach, H. Random heterogeneous materials: Microstructure and macroscopic properties. *Appl. Mech. Rev.* **2002**, *55*, B62–B63. [[CrossRef](#)]
229. Niblack, W. *An Introduction to Digital Image Processing*; Strandberg Publishing Company: Birkerød, Denmark, 1985.
230. Wang, Y.; Zhang, Y.; Zhao, H.; Li, X.; Huang, Y.; Schadler, L.S.; Chen, W.; Brinson, L.C. Identifying interphase properties in polymer nanocomposites using adaptive optimization. *Compos. Sci. Technol.* **2018**, *162*, 146–155. [[CrossRef](#)]
231. Zhao, H.; Li, Y.; Brinson, L.C.; Huang, Y.; Krentz, T.M.; Schadler, L.S.; Bell, M.; Benicewicz, B. Dielectric spectroscopy analysis using viscoelasticity-inspired relaxation theory with finite element modeling. *IEEE Trans. Dielectr. Electr. Insul.* **2017**, *24*, 3776–3785. [[CrossRef](#)]
232. Zhang, Y.; Tao, S.; Chen, W.; Apley, D.W. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. *Technometrics* **2019**, 1–12. [[CrossRef](#)]
233. Goldberg, D.E. *Genetic Algorithms*; Pearson Education India: Delhi, India, 2006.

234. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. Molecular sets (moses): A benchmarking platform for molecular generation models. *arXiv* **2018**, arXiv:1811.12823.
235. Van Moffaert, K.; Drugan, M.M.; Nowé, A. Scalarized multi-objective reinforcement learning: Novel design techniques. In Proceedings of the 2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Singapore, 16–19 April 2013; pp. 191–199.
236. Van Moffaert, K.; Nowé, A. Multi-objective reinforcement learning using sets of pareto dominating policies. *J. Mach. Learn. Res.* **2014**, *15*, 3483–3512.
237. Mossalam, H.; Assael, Y.M.; Roijers, D.M.; Whiteson, S. Multi-objective deep reinforcement learning. *arXiv* **2016**, arXiv:1610.02707.
238. Khan, N.; Goldberg, D.E.; Pelikan, M. Multi-objective Bayesian Optimization Algorithm. In Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation (GECCO'02), New York, NY, USA, 9–13 July 2002; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2002; p. 684.
239. Laumanns, M.; Ocenasek, J. Bayesian optimization algorithms for multi-objective optimization. In Proceedings of the International Conference on Parallel Problem Solving from Nature, Granada, Spain, 7–11 September 2002; pp. 298–307.
240. Wada, T.; Hino, H. Bayesian Optimization for Multi-objective Optimization and Multi-point Search. *arXiv* **2019**, arXiv:1905.02370.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).