Journal of Medicinal Chemistry

pubs.acs.org/jmc

Perspective

Learning Molecular Representations for Medicinal Chemistry

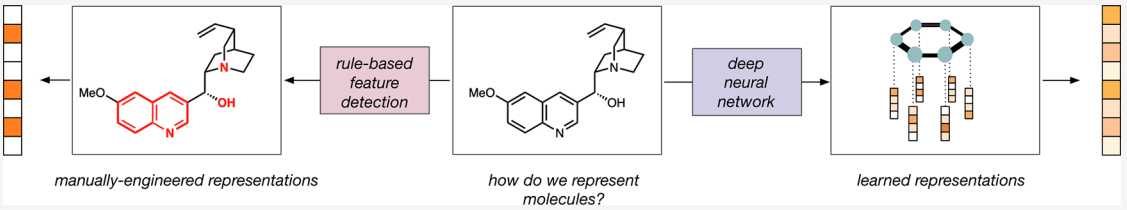
Miniperspective

Kangway V. Chuang, Laura M. Gunsalus, and Michael J. Keiser*

Cite This: J. Med. Chem. 2020, 63, 8705−8722

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: The accurate modeling and prediction of small molecule properties and bioactivities depend on the critical choice of molecular representation. Decades of informatics-driven research have relied on expert-designed molecular descriptors to establish quantitative structure−activity and structure−property relationships for drug discovery. Now, advances in deep learning make it possible to efficiently and compactly *learn* molecular representations directly from data. In this review, we discuss how active research in molecular deep learning can address limitations of current descriptors and fingerprints while creating new opportunities in cheminformatics and virtual screening. We provide a concise overview of the role of representations in cheminformatics, key concepts in deep learning, and argue that learning representations provides a way forward to improve the predictive modeling of small molecule bioactivities and properties.

1. INTRODUCTION

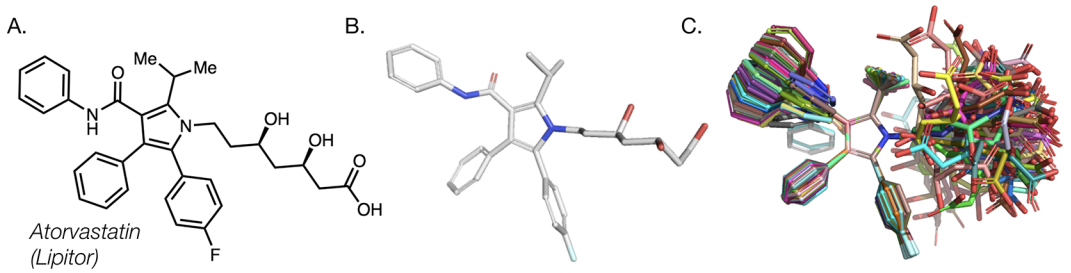
1.1. Why Does Representation Matter? The ability to learn key patterns from complex sets of observations is a central aspect of human intelligence.[1] Expert chemists leverage this ability to find small-molecule leads and optimize drug-like properties in therapeutic discovery,[2] where intricate chemical and biological processes govern the interactions of small molecules. Successfully navigating this discovery landscape is a testament to human ingenuity in complex pattern recognition, particularly in relating molecular structure to resulting biological function.[3] An often overlooked but inextricable aspect of structural pattern recognition lies in *how molecules are represented*. For example, Figure 1 shows the statin Lipitor drawn in a variety of human-interpretable ways. Pictorially, most organic chemistry textbooks teach the canonicalized bond-line notation (Figure 1, left), where a molecule is depicted as a chemical graph and each unlabeled vertex corresponds to a carbon atom.[4] This visual notation readily illustrates the topology of molecules; yet varying orientation and viewpoint of illustration can occlude or reveal salient patterns for expert chemists.[5] Furthermore, bond-line notation neglects important aspects of three-dimensional shape (Figure 1, middle) and relevant conformational dynamics for flexible molecules (Figure 1, right).

Just as humans depend on clear representations for pattern recognition, effective molecular representation is crucial for statistical and machine learning.[6,7] Decades of drug discovery research in quantitative structure−activity relationship (QSAR) studies[8] and molecular similarity analysis[2] have demonstrated that accurate predictions rely on the choice of molecular features, also known as *molecular descriptors*.[9−12] In turn, computational chemists often select or engineer these descriptors using domain-specific expertise and a strong working knowledge of the causal factors underlying the observed data. In machine learning, this process of manually discovering and creating relevant features is known as *feature engineering*.[7] The choice of features is essential to any machine learning pipeline and directly affects the type of learning algorithms that can be used. To illustrate this concept, consider the toy classification example in Figure 2. Most real-world processes produce data that are not linearly separable;[13] e.g., no linear decision boundary can be found that discriminates between the red and blue classes in Figure 2A. Domain expertise and an understanding of the physical processes that generate the observed data can inspire feature transformations that simplify learning. In this example, a switch from Cartesian to polar

Special Issue: Artificial Intelligence in Drug Discovery

Received: March 5, 2020
Published: May 4, 2020

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722



D. CC(C)C1=C(C(=C(N1CCC(CC(CC(=O)O)O)O)C2=CC=C(C=C2)F)C3=CC=CC=C3)C(=O)NC4=CC=CC=C4

E. logP: 6.98, Molecular Weight: 558.6 g/mol, Hydrogen Bond Donors: 4, Hydrogen Bond Acceptors: 6

Figure 1. Common visual, human-interpretable representations of molecules. (A) The commonly used bond-line notation represents molecules as graphs, with each vertex corresponding to an atom and each edge corresponding to a bond. (B) Three-dimensional visual representation. (C) Aligned three-dimensional overview of different generated conformations. (D) Canonical SMILES line notation. (D) Examples of computed molecular descriptors for atorvastatin.

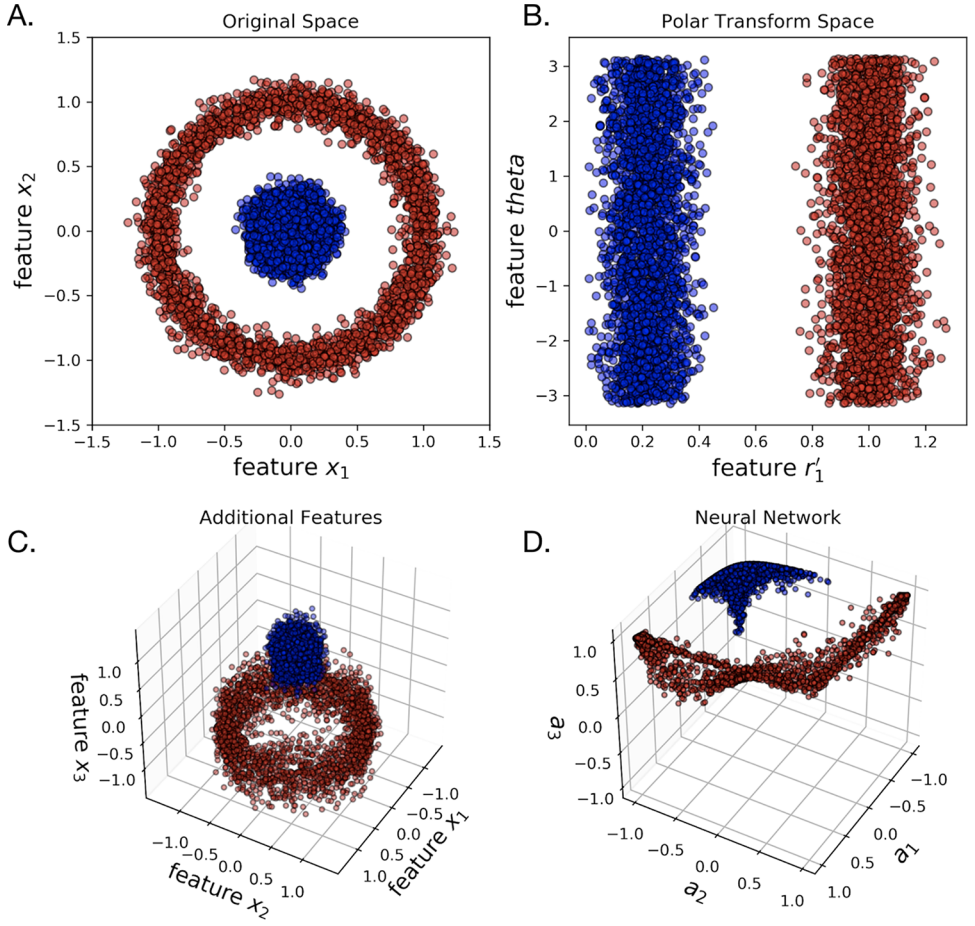


Figure 2. Choice of representation plays a key role in learning. (A) The red and blue classes as shown are not linearly separable. (B) Domain knowledge can allow for the application of new ways to represent data: in this case, the simple transformation of data into a polar coordinate system results makes the red and blue classes trivially separable with a linear decision boundary. (C) A goal of feature engineering is to discover new features that may aid in discriminating between classes. (D) In contrast, a neural network automatically *learns* a new representation space (here, the activations a_i of three hidden layer neurons) that results in finding a linear decision boundary.

coordinates allows a linear decision boundary to trivially separate the two classes (Figure 2B). Alternatively, adding new relevant features can help to distinguish classes in a new dimension (Figure 2C). Despite the continued success of this approach in machine learning, feature engineering can be difficult and time-consuming.[14] Furthermore, engineering good features *a priori* is inherently challenging in new and unexplored systems, as little is known about causal factors behind the observations and the distribution of data.

In contrast to feature engineering, deep learning algorithms perform a type of *feature learning*, also known as *representation learning*.[6,15] Rather than relying on expert-encoded features, the

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

deep learning paradigm *learns* compact and expressive representations directly from observed data.[16,17] Figure 2D highlights how a simple neural network can automatically learn a new internal representation that is linearly separable without the need for additional engineering.[18] This paradigm shift is at the core of the deep learning renaissance in computer vision and natural language processing, where deep neural networks now achieve superhuman performance in object recognition and sentiment analysis.[17] In this review, we discuss how this paradigm can improve performance in ligand-based virtual screening while creating new research directions in molecular machine learning.

1.2. Scope of This Review. This review focuses on representation learning for small molecule bioactivity and property prediction. Although we touch on key ideas and concepts along the way, an in-depth tutorial on machine and deep learning is beyond the scope of this review, and we direct interested readers toward dedicated resources.[15,19] Below, we provide a brief context on molecular representations (section 2) and deep learning (sections 3.1 and 3.2) and relate innovations in deep learning to concepts in cheminformatics and similarity analysis (section 3.3). We then discuss in greater technical detail how representations learned with deep neural networks present opportunities and challenges in this field (section 4).

As recent interest in machine and deep learning has led to a flood of innovative applications across chemistry and biology, we are unable to comprehensively cover all work relevant to this field. Applications of deep learning to protein structure-based methods such as docking[20−23] and methods for calculating electronic and materials properties[24] are likewise outside the scope of this review and not explicitly covered. Instead, we direct readers interested in comprehensive overviews on machine learning in cheminformatics,[8,25,26] medicinal chemistry,[27,28] and materials[24,29] research toward recent reviews.

2. DO WE NEED NEW REPRESENTATIONS FOR MOLECULES?

2.1. A Short History of Molecular Representations. From lipophilicity to three-dimensional geometry, thousands of experimental and theoretical descriptors have been developed for cheminformatic applications in drug discovery.[10] With so many options, does the field need new molecular representations? Existing descriptors each encode different information, and no single representation performs universally well across all tasks. For instance, real-valued descriptors such as molecular weight or polarity may correlate well to a property such as boiling point but may suffer in complex molecular recognition tasks such as binding, where aspects of structure or geometry provide crucial information. Choosing a specific representation depends heavily on its precise application,[30] and the researcher must often assess multiple sets of descriptors to obtain good predictive performance.[12]

Accordingly, most molecular representations encode information optimized for a specific use. As a historical example, early chemical information systems primarily focused on storage and retrieval of small-molecule information, and efforts focused on compact and unique molecular representations for disk-efficient storage and rapid indexing. The widely used simplified input line entry system[31] (SMILES) and subsequent international chemical identifier[32] (InChI) are just two examples of lexical representations for this purpose, as they compactly store molecular graph information in a standardized format to facilitate information search. Similarly, the need to efficiently query growing chemical databases for rapid substructure search drove the development of key-based bitstring "fingerprints", with each bit denoting the absence or presence of a molecular feature or substructure.[33,34] These fingerprints, akin to Bloom filters[35] used in data retrieval, resulted in a fast method for bitwise-comparison of molecular features and allowed for rapid filtering and search in chemical databases.

The development of new representations also reflects the shifting trends in research and the evolving technological landscape. For example, growing efforts in structure−activity modeling and molecular similarity analysis in the 1980s and 1990s prompted the creation of new bitstring representations and the reoptimization of old ones. Topological fingerprints based on atom pairs[36] as well as local circular neighborhoods[37] provided more general-purpose representations that were specifically designed for bioactivity prediction and similarity analysis.[2] Likewise, the growing push for bioactivity modeling prompted the reoptimization of the molecular access system (MACCS) keys fingerprints that were originally designed for substructure searching based on expert-encoded features.[38] As QSAR studies progressed, the increasing influence of X-ray crystallographic structures combined with additional computational power drove the design of representations and methods that capture aspects of three-dimensional structure and shape.[39] Fingerprints based on geometric distances[40,41] and methods such as the rapid overlay of chemical structures (ROCS)[42] formed new opportunities to leverage spatial information for 3D-QSAR and shape-similarity analyses that were previously inaccessible. Despite the initial promise of incorporating 3D information into molecular fingerprints, these methods have not made a significant impact on QSAR and machine learning approaches and are often outperformed by 2D fingerprinting methods.[2,43] Bioactive conformations of molecules are not generally known *a priori*, and only encoding a single, lowest energy conformer out of a vast conformational ensemble can introduce unwanted bias. To date, representations that effectively incorporate conformational ensembles have not yet been developed.

Cheminformatic applications continue to integrate tightly with machine learning at ever-increasing scales, with recent applications in chemical synthesis[44−50] and systems pharmacology.[51] These large-scale investigations increasingly necessitate general-purpose molecular representations that can capture the rich diversity of chemical space; yet recent studies indicate that existing molecular descriptors are insufficiently expressive for many applications.[52] To meet the demands of this evolving research area, we anticipate that new molecular representations must be developed. Specifically, active work in deep learning creates a promising way forward for the flexible representation learning of small molecules. Below, we delve into aspects of deep learning for small molecules and discuss advantages, opportunities, and limitations within the field.

2.2. What Makes a Good Molecular Representation? What qualities make for a good molecular representation? A general intuition in machine learning is that a good representation makes the learning task easier, a concept illustrated through the toy example in Figure 2.[6,15] The same intuition holds when applied to molecules, as identifying key structural features is paramount to revealing bioactivity and property relationships:[12] a good *molecular* representation makes the subsequent learning task easier. A number of practical and theoretical considerations are necessary for the design of molecular representations, including invariance to atom-

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

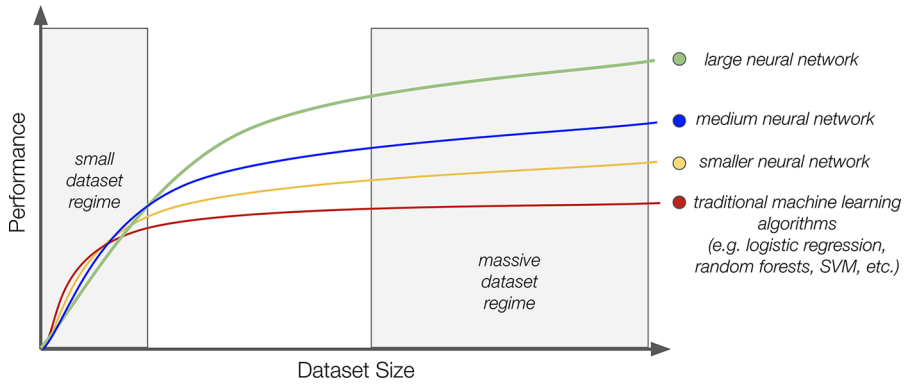


Figure 3. Conceptual illustration of machine learning model performance as a function of data set size. In the small data set regime, traditional machine learning methods often outperform neural networks. As data set size increases, neural networks with increasing capacity outperform other methods. Figure adapted from Ng.[55]

numbering and an unambiguous, computable definition.[10] Here, we focus on a few essential aspects of representations for use in molecular learning tasks. Molecular representations should be the following.

Expressive. Chemical space is vast; yet single-atom perturbations to molecular structure can lead to dramatic differences in both physicochemical properties and biological activities. For example, subtle differences in protonation state and tautomerization can lead to drastic shifts in molecular function and remain an ongoing challenge in cheminformatics.[53] Representations should both faithfully capture the richness and diversity of chemical space and distinguish subtle differences between molecules.

Parsimonious. The cost of experiments at scale limits the size and diversity of chemical data sets. To ensure that models learn important patterns over noise, it is critical to maintain parsimony in the input feature space for a machine learning task. As often attributed to Einstein, "Everything should be made as simple as possible but not simpler." In the same spirit, representations should be compact and retain expressiveness without eliminating critical information.

Invariant. Because the same molecular input should consistently generate the same output, molecular representations must be invariant to aspects such as atom-numbering. Additional types of invariance can facilitate learning by limiting the space of learnable functions to better suit a specific application domain. For example, circular fingerprints such as the extended connectivity fingerprint (ECFP) use an internal reference frame that only encodes structural topology and are therefore invariant to rototranslation and molecular conformation.[37] As such, these fingerprints have found widespread success in molecular similarity analysis as they provide sufficient information and are fast to compute and compare.[2]

Interpretable. For scientific applications of machine learning, it is critical to ensure that model performance derives from learning relevant patterns instead of by exploiting confounding variables, experimental noise, or other possible artifacts. Representations that can be traced to a structural interpretation allow human experts to assess the patterns learned by their models and provide a sanity check against domain knowledge.[54] Crucially, integrating these interpretability studies into chemical and biological discovery will be necessary for the continued advancement of machine learning in the sciences.

3. WHY DEEP LEARNING AND WHY NOW?

The past few years have witnessed a resurgence of interest in neural networks for drug discovery,[56] leading to excitement[57] and skepticism[58] from experts in the field. Given the limited success of neural networks for drug discovery in the past, an important question is whether deep learning is well suited for small-molecule drug discovery.[59] Below, we provide a short overview of neural network fundamentals to center the discussion, provide context for neural networks in drug discovery, and relate innovations in computer vision and natural language processing to modern small-molecule drug discovery. Although fundamentally different in application domain and field-specific challenges, recent innovations across fields foreshadow exciting new research directions for deep learning in medicinal chemistry.

3.1. Artificial Neural Networks and Deep Learning.

Artificial neural networks (ANNs) are a class of computing systems inspired by the biological networks of the human brain.[19] In the simplest case, a shallow, fully connected, or feed-forward network is a directed computational graph consisting of three layers: an input layer, a single hidden layer, and an output layer (Figure 4A). Each layer has a variable number of computational units, referred to as neurons, that perform a nonlinear transformation on their input data. Conceptually, information is propagated in a layer-wise fashion, as each layer receives the output of the previous layer. *Deep neural networks*, or *deep learning*, refer to neural networks that possess multiple hidden layers. Overall, these flexible models are universal mathematical function approximators that can learn arbitrarily complex functions given a sufficient number of neurons and training examples. As a type of machine learning, neural networks learn parameters of the model *directly from data* using the *backpropagation algorithm*, a gradient-based optimization method for computational graphs.[60] Backpropagation allows for the training of neural networks as long as they are end-to-end differentiable; i.e., they are composed of smooth, continuous functions. Colloquially, a trained model refers to a neural network architecture along with learned weights connecting all of its neurons.

Decades of research have explored a variety of architectures, each suited to a different purpose. In addition to the standard *feed-forward* network that consists of fully connected layers (equivalently known as *dense* or *affine* layers), recurrent, convolutional, and graph convolutional architectures have been developed for different domains and data types (Figure

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

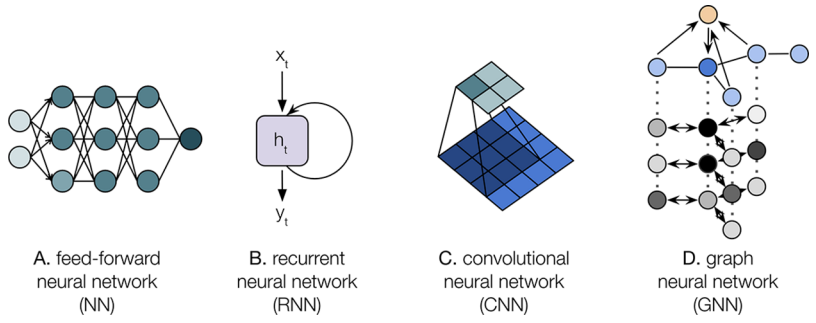


Figure 4. A schematic representation of flexible neural network architectures. (A) A feed-forward network is fully connected between layers. (B) Recurrent neural network architectures include self-loops and can handle variable-length data including text. (C) Convolutional neural networks operate on regular grid inputs and aggregate local spatial information that are particularly effective for image data. (D) Graph neural networks are an emerging area of research for learning on irregular and unstructured data such as social networks.

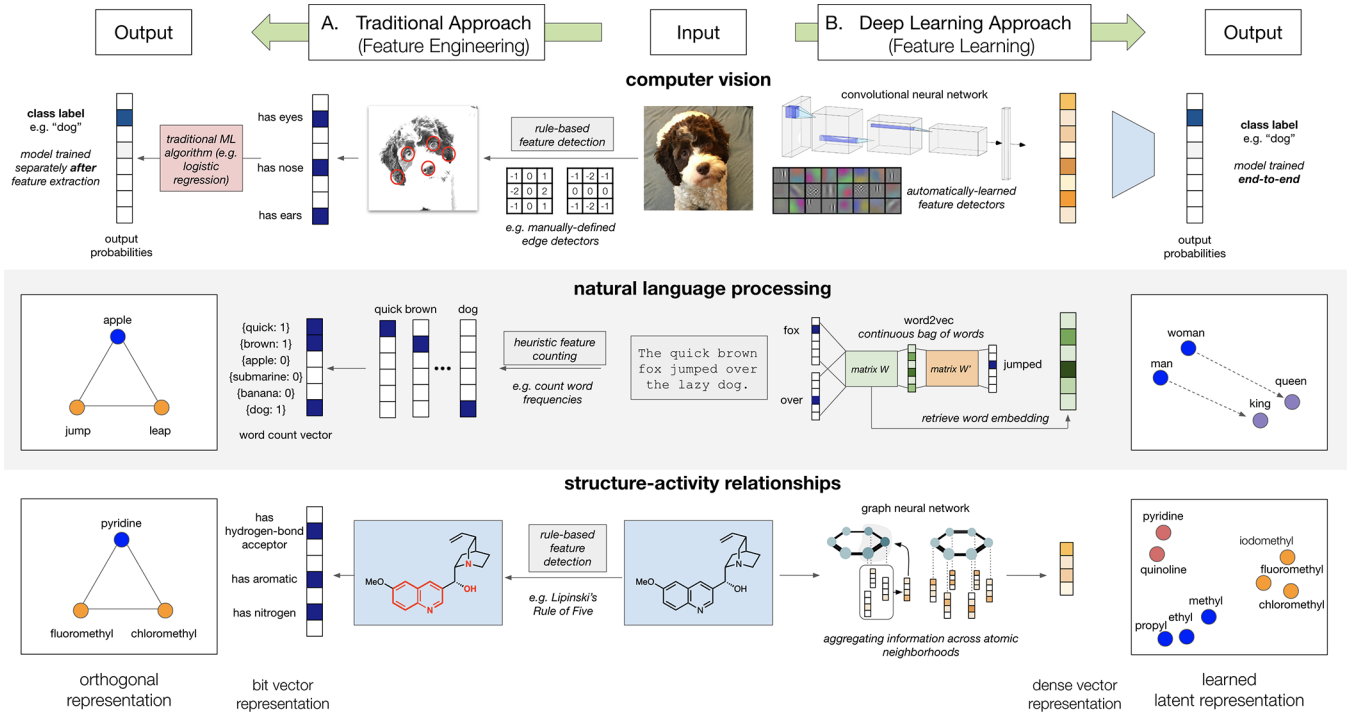


Figure 5. Feature engineering vs feature learning for capturing important similarity relationships. (A) Traditional approaches in computer vision, natural language processing, and cheminformatics use hard-coded feature extractors that do not capture similarity between inputs. Similar objects may thus be far apart in the encoding space. (B) Deep learning approaches automatically construct an intermediate *latent space* that can naturally capture meaningful relationships.

4). *Recurrent neural networks* (RNNs) such as gated recurrent units[61] and long short-term memory units (LSTM)[62] are widely used to capture temporal dependence in sequence-based data such as time series and text. *Convolutional neural networks* (CNNs) emerged as the standard for image processing by capturing local, spatial relationships with learnable filters.[15] *Graph neural networks* (GNNs) operate on unordered data structures such as social networks and are particularly well suited to molecules.[63] Together, these modular units allow deep networks to operate on a broad range of data and combined data types to provide flexible learning.

Increased data availability, innovations in algorithms, and advancements in computational hardware have driven the recent explosion in deep learning.[64] Although neural networks have been used for decades, deep networks were notoriously difficult to train and easily overfit small data sets.[65] Technological and scientific trends in the past 3 decades have rapidly increased the rate of data collection, and the push for "big data" and the availability of open-source data sets have helped to mitigate challenges in overfitting with highly parametrized networks (Figure 3).[66] From an algorithmic perspective, networks possessing more than a few layers often suffered from the problem of vanishing or exploding gradients that prevented models from learning effectively.[65] The development of new initialization schemes,[65] neuron activation functions,[67] and gradient-based optimization methods[68−70] has led to considerable improvements in model training and enabled researchers to efficiently train exceptionally deep networks.[71] Finally, hardware innovations and the widespread availability of graphics processing units (GPUs) powered rapid parallel matrix computations necessary for gradient-based training of neural networks.[64] Together, these advancements in data, algorithms, and hardware have led to dramatically reduced times to train and evaluate deep networks as well as increased scalability.

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

3.2. Neural Networks in QSAR and Drug Discovery. Neural networks have a checkered past in QSAR and drug discovery.[8] The first application of ANNs in medicinal chemistry dates back nearly 5 decades to the classification of dioxolane-containing small molecules using the perceptron algorithm.[72] Despite this early example, QSAR studies did not widely incorporate neural networks until the 1990s. The prior decade had seen a surge of interest in parallel distributed systems, leading to the rediscovery and popularization of the back-propagation algorithm by Rumelhart et al. for efficient training.[60] As research in this area found renewed interest in the form of *connectionism*, so did its applications to QSAR studies across industrial and academic settings.[73] Despite limited successes, neural networks fell out of favor due to their propensity to overfit, as well as their perceived black-box nature.[56] Small data sets tended to exacerbate overfitting with highly parametrized models, and strong predictive performance came at the cost of model interpretability. As machine learning research continued to progress, ANNs were supplanted by other algorithms such as random forests[74] and support vector machines[75] that were less prone to overfitting. These algorithms are still widely used and achieve strong performance especially in small-data regimes (Figure 3).

In 2012, the Merck Molecular Activity Challenge hosted by Kaggle reignited interest in deep learning for drug discovery.[76,77] In this online data science contest, competitors were challenged to predict bioactivities across 15 related tasks using precalculated molecular descriptors for nearly 50 000 molecules. A team led by Dahl et al. won the competition using multitask deep neural networks and outperformed state-of-the-art random forest models by nearly 15% to win the competition.[78] Although follow-up studies at Merck demonstrated a more modest performance gain over random forests models,[79,80] this competition repopularized neural networks for drug discovery. In the following years, significant attention shifted toward deep learning approaches for predicting small molecule activities, physicochemical properties, as well as adsorption, distribution, metabolism, and excretion (ADME).[27,28] Deep learning methods are now an active area of academic and pharmaceutical research[81] and have continued to evolve beyond the evaluation of existing neural networks. Innovative methods continue to be developed at a rapid pace to address challenges in molecular design and optimization. In section 4, we discuss the strengths and limitations of several of these key innovations from the point of view of representation learning.

3.3. From Feature Engineering to Feature Learning. Deep learning shifts computer vision and natural language processing paradigms: whereas prior approaches relied heavily on expert *feature engineering*, deep neural networks automatically perform a type of *feature learning* directly from data (Figure 5). For instance, traditional computer vision methods extensively use expert-defined image filters, including edge, color, and texture detectors.[82] By contrast, convolutional neural networks *automatically* generate hierarchical combinations by *learning* good feature extractors through gradient-based optimization. As a concrete example, a traditional computer vision approach for dog classification may use existing knowledge of animals to detect known patterns, such as the presence of eyes, nose, or a tail. These extracted features can then be used with any machine learning algorithm to train a dog classifier. Instead, a supervised deep learning approach does not require existing knowledge and learns discriminating features of dogs purely by example, using raw image data along with their labels (e.g., dog, cat, submarine).

The power of this approach was demonstrated in the ImageNet Large Scale Visual Recognition Challenge, an open computer vision competition to evaluate algorithms for image classification at scale.[83] In 2012, the CNN-based architecture AlexNet won with a top-five error rate of 15.3% and a margin of almost 11% compared to the runner-up.[84] The success of AlexNet ignited interest in deep convolutional networks for computer vision and contributed significantly to the current revolution in deep learning. In the years following AlexNet, very deep CNNs with new architectures rapidly surpassed human-level performance, reaching top-five error rates better than 5%.

Why is deep learning so effective for image recognition, and what do these models learn? A key aspect of deep architectures is the concept of *hierarchical learning* of representations. The lowest layers of a neural network learn relatively simple features that are nonlinearly combined into higher-order concepts as they propagate through the network (Figure 5). This hierarchical organization, with multiple intermediate layers of representation, is key to the predictive power of deep networks and provides improved computational complexity, sharing of statistical strength, and increased expressiveness.[85] Concretely, Zeiler and Fergus explored the hierarchical learning process of AlexNet,[86] and demonstrated that the initial layers learn features such as colors and edges, intermediate layers capture more complex textures, and deeper layers recognize complex features including noses and eyes. Finally, the deepest layers of a CNN, closest to the output, learn task-specific abstractions such as animal faces and poses. All together, these layers effectively transform raw input features into learned ones that are useful for visual recognition and classification.

The same shift toward feature learning led to dramatic improvements in natural language processing (Figure 5). Text analysis models based on word counts alone, such as *bag-of-words*, are limited in document classification and sentiment analysis. Because each word is encoded as an orthogonal vector, these representations explicitly disregard any *similarity* between words. For instance, in these models, "jump" and "leap" are equally dissimilar as "jump" and "apple". Alternatively, predictions models such as *word2vec* are built on the assumption that *similar words appear in similar contexts* and yield richer word representations (i.e., word embeddings) and improve language models.[87] For example, in a *continuous-bag-of-words* model, a shallow neural network is trained to predict a word from its surrounding context. As a result, the network learns word embeddings from their contextual usage, providing an expressive, *dense* representation that captures relationships between words (e.g., woman is to man as queen is to king).[88]

Although computer vision and natural language processing appear distinct from cheminformatics, we can draw parallels between these fields of research (Figure 5). As discussed in section 2, decades of cheminformatic research have resulted in expert-defined molecular representations that qualify as *feature engineering*. Yet the advances in computer vision and natural language processing indicate that *learning* molecular representations drives real opportunities across tasks such as small molecule affinity prediction. Just as visual representations of dogs can be built hierarchically from simple edges, shapes, and *local spatial* information, molecular representations can be built hierarchically from *local atomic* environments and substructures with deep learning.[89] Grzybowski and co-workers recognized the striking resemblance between modern cheminformatics and computational linguistics based on frequency-based approaches.[90,91] Indeed, the commonly used circular molecular

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

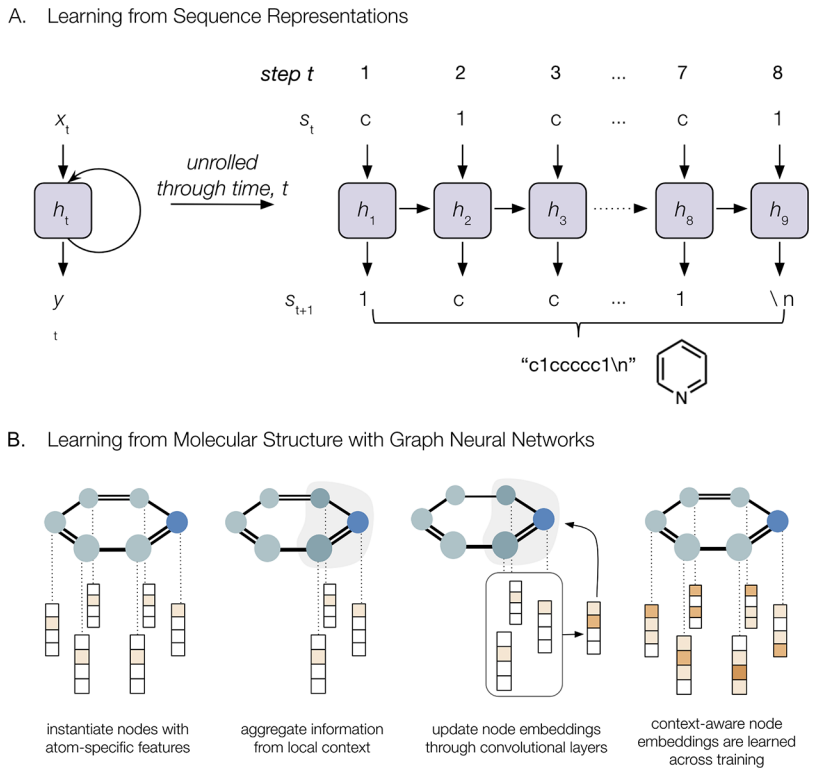


Figure 6. Deep learning enables flexible learning from diverse input types such as sequences and graphs. (A) Recurrent neural networks can be used to learn molecular representations of molecules from sequence-based data such as SMILES. (B) Graph neural networks aggregate information about local neighborhoods to provide expressive representations of small molecules.

fingerprint, the extended-connectivity fingerprint (ECFP), effectively encodes a "bag-of-fragments" representation for a molecule in direct parallel to the "bag-of-words" representation for text and uses identical techniques such as *feature hashing* to create a sparse molecular fingerprint.[37] Together, these parallels suggest that representation learning will improve the state of predictive modeling for small molecules.

Critically, small-molecule drug discovery breaks standard assumptions in many technological applications of machine learning. Most machine learning algorithms operate on the assumption that training and testing data are independently and identically distributed (the i.i.d. assumption).[19] For example, we would expect a standard image classifier trained to exclusively distinguish cats from dogs to generalize to new images of cats and dogs. This model will likely produce nonsensical classifications if asked to evaluate pictures of humans. In stark contrast, real-world drug-discovery breaks this standard i.i.d. assumption. The optimization and design of small molecules necessarily explore structural variations drawn from intentionally novel regions of chemical space. Large structural changes to small-molecule hits are typically required to become a lead. For a model to be useful to the practicing medicinal chemist, it must generalize to *out-of-distribution* examples. Despite this fundamental shift in distributions, the transition from expert-defined features to learned features will still benefit cheminformatics and medicinal chemistry. Below, we outline key advances and developments in learning representations for small molecules and discuss how these methods will continue to improve the state-of-the-art in predictive modeling for drug discovery.

4. OPPORTUNITIES IN LEARNING MOLECULAR REPRESENTATIONS

The sections above discuss how expert-designed molecular descriptors advanced the field of cheminformatics for the prediction of small molecule properties and how the shift from feature engineering to feature learning transformed computer vision and natural language processing. In this section, we explore several specific advantages and active areas of research for learning small molecule representations from chemical informatic and pharmacology data.

4.1. Learning from Flexible Input Representations.

The major advances in computer vision and natural language processing operate directly on *complete* and *raw* input structures, such as the native pixels of an image. Translating these developments to molecular deep learning is an exciting direction. Traditional machine learning requires fixed-length inputs that do not handle variable-length and unstructured data. However, broad variation in the structure and size of molecules makes it difficult to engineer fixed-length representations that are sufficiently expressive for learning. Developments in deep learning now operate on fundamentally unstructured and variable data types that create opportunities to explore new and meaningful molecular representations. Existing string- and graph-based formats, which are designed to encode the complete composition and bonding of molecules, are promising starting points for deep learning tasks.

Learning from string-based representations such as SMILES has attracted interest as they compactly encode molecular structure and are already widely used for storage in chemical databases. SMILES follow a human-interpretable syntax that constitutes a formal grammar system, allowing researchers to directly adapt methods and architectures from natural language

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

processing and neural machine translation to problems in cheminformatics.[92] For example, the SMILES representation for pyridine is "c1ccncc1," (Figure 6A). However, until recently, only a handful of methods have attempted to learn *directly* from SMILES due to challenges in reducing variable-length strings into an expressive, fixed-length representation.[93,94]

In an early report, Segler et al. applied an RNN-based model to generate focused chemical libraries with antimalarial and antibacterial activities using a two-step approach (Figure 6A): (1) they first train a deep recurrent neural network model on 1.4 million molecules extracted from the ChEMBL[95] database to learn the SMILES syntax, and then (2) they *fine-tune* the resulting model by further training against a smaller corpus of known active molecules.[96] The authors demonstrate that the final, trained model can generate focused libraries of new and valid molecules for antibacterial discovery. Together, this study and others have demonstrated the ability to learn meaningful internal representations of chemical space from a previously intractable input type, with useful applications to small molecule property and activity prediction.

Although convenient, the SMILES representation has several critical flaws for learning: (1) Two similar molecules can produce two dramatically different SMILES representations, since multiple valid but different SMILES can describe the same molecule. (2) SMILES representations are brittle; single character changes can produce invalid molecules. (3) Most molecules are inherently nonlinear, yet SMILES collapses complex structures into a single linear sequence. These flaws have made the SMILES syntax empirically difficult to learn using standard convolutional and recurrent architectures, and as a result, sophisticated model architectures and large amounts of data are required to effectively overcome the grammatical dependencies of these linear representations. Recent work has demonstrated that modifications to the SMILES syntax,[97] new sampling methods such as SMILES-augmentation,[98] and specialized architectures such as stack-RNNs[99] address many of these shortcomings. Furthermore, more robust lexical representations, such as self-referencIng embedded strings (SELFIES), enhance learning and consistently yield valid molecules.[100]

Alternatively, an exciting emerging direction is learning directly on molecular structures using graph neural networks. As discussed in section 1, graphs are the standard for molecular depiction, with nodes corresponding to atoms and edges corresponding to bonds (Figure 1A). Graph neural networks (GNNs) specifically learn from this simple representation. Just as CNNs aggregate local spatial information across a regular grid (Figure 4C), GNNs generalize this concept to non-Euclidean inputs such as networks (Figure 4D). Graph learning proceeds in several steps (Figure 6B): First, existing molecular features are directly encoded into each node representation, such as atom type and hybridization. Throughout the layers in a GNN, node representations are updated with information passed from their surrounding neighborhoods in a framework known as *message passing*.[101] This process of iterative message passing and updates allows information to flow across the graph to create a continuous and dense representation of each node.

This algorithmic procedure tightly corresponds with the ECFP fingerprinting algorithm, where information is collected from the local environment radially. In contrast to circular fingerprints, however, graph neural networks use fully differentiable neural network layers to automatically learn useful hierarchical features by propagating raw atomic information across locally bonded neighborhoods. Critically, although traditional fingerprints and graph neural networks both exploit expert-defined and *engineered* atomic and bond features (e.g., atom type, hybridization, partial charges, etc.), graph neural network layers progressively transform and aggregate arbitrarily sized molecular graphs into a relevant *learned* vector (i.e., an embedding). As such, the information aggregation stages of these networks evolve *specifically for the task* and are fundamentally distinct from traditional fingerprints and descriptors that rely on predefined means to aggregate chemical substructure patterns (such as hash functions).

In practice, early reports by Duvenaud et al. on the development of neural graph fingerprints[102] and Kearnes et al. on molecular graph convolutions[103] demonstrated strong performance relative to traditional fingerprints on aqueous solubility[104] and bioactivity tasks. Subsequently, a number of variations on graph neural network architectures have been developed for the prediction of physicochemical properties, bioactivities,[23,105] and small molecule energies.[101,106,107] These early studies demonstrated how end-to-end learning on native graph structures can improve on traditional representations in a variety of tasks.

Overall, the flexibility of deep neural networks and their ability to handle diverse input data types have led to innovative approaches for small molecule representation and have already demonstrated consistent, albeit modest, improvements over traditional methods. Although still limited in their practicality due to the need for substantial expertise, these early studies have opened up new avenues for continued research.

4.2. Learning Molecular Similarity and Chemical Space through Continuous Representations. The *similar property principle* of cheminformatics states that *similar compounds should have similar properties*.[108] Although this principle agrees with expert intuition, molecular similarity remains ill-defined. Computational measures such as the Tanimoto coefficient (Tc) primarily reflect the similarity of their chosen molecular representation, yet the similarity of these representations serve only as a correlative proxy to the ultimate goal of determining *function* from molecular structure.[2] Matched molecular pairs[109] corresponding to an activity cliff[110] illustrate this concept at one extreme: although the two molecules may be structurally similar, the stark difference in activity indicates they are not functionally similar. At the other extreme, two active molecules with different scaffolds illustrate the inverse concept: both molecules will have low structural and topological similarity based on their molecular fingerprints, yet they can be considered functionally similar as active molecules against the same protein target. Molecular representations that faithfully capture aspects of similarity in an expressive, yet parsimonious fashion, are therefore of considerable interest.

Popular bit- and count-based fingerprinting methods generate discrete representations that are not unique and frequently lose information through bit collisions. In contrast, deep neural networks naturally learn continuous representations that are unique, have more representational power, and can learn notions of *task-specific molecular similarity*. To illustrate this example, consider the circular fingerprint ECFP.[37] The ECFP algorithm effectively encodes each molecule as a "bag-of-fragments" based on local atomic environments, generating unique integer identifiers that are subsequently hashed into a fixed-length representation. As a result, each fragment is necessarily and completely distinct; e.g., a chloroethyl group is as different to a fluoromethyl group as it is to a benzene

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

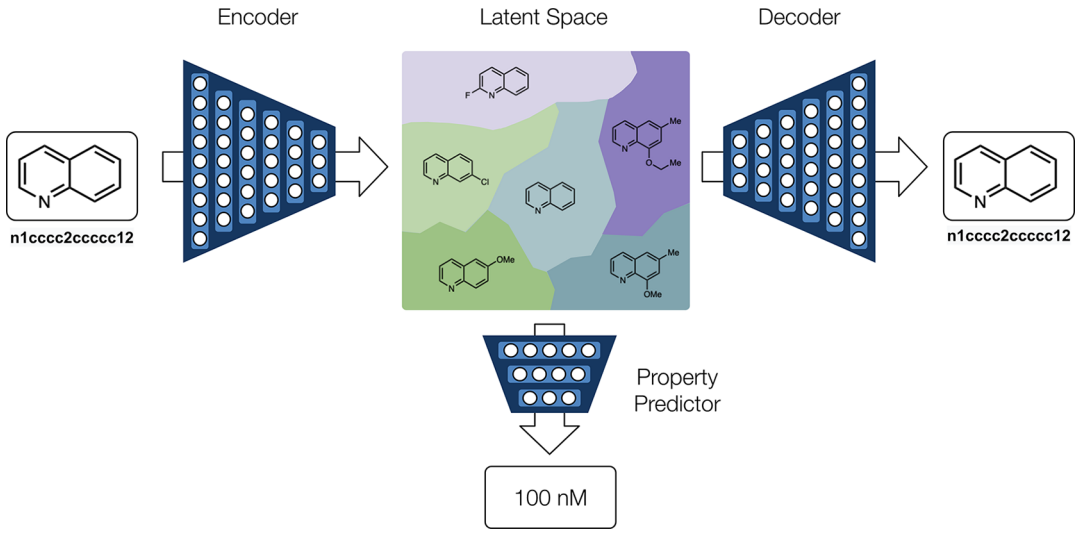


Figure 7. Continuous latent space optimization. Gomez-Bombarelli et al. developed a variational autoencoder to map discrete SMILES inputs to a continuous latent space. The generation of a smooth and continuous latent representation allows for gradient-based optimization methods, interpolation of chemical space, and inverse design of small molecules with targeted properties.

substructure; no notion of similarity is encoded between fragments. At the other extreme, modifications to ECFPs such as functional-class fingerprints[37] (FCFPs) use generic atom types to enforce that similar groups are encoded identically (e.g., chloroethyl and fluoroethyl map to the same bit identifier). Mapping similar but nonidentical fragments to the same bit necessarily reduces the expressivity of the fingerprint but can effectively increase performance in low-data regimes.

In contrast, neural graph fingerprints reported by Duvenaud et al. provide a continuous generalization of the ECFP algorithm, replacing a hash function with a single layer of a neural network. This approach allows each molecular fragment to be similarly encoded based on the predictive task. Learning the continuum of similarities increases the expressiveness of these representations, allowing for subtle differences in molecules to be faithfully captured. Furthermore, the parameters of these networks are *learned* directly from the training data, meaning that these notions of molecular similarity are defined by its desired function. Although graph neural networks are more flexible and expressive, a substantial amount of data can still be necessary to learn what is trivial to expert chemists. For example, a small neural network trained on a few hundred examples may struggle to learn common bioisosteric relationships (e.g., benzene and thiophene) or basic periodic trends. However, when these methods are applied to sufficiently large and high-quality data sets, they consistently outperform standard machine methods by considerable margins.[23,111]

In a broader context, learning a smooth and continuous representation provides advantages beyond improved similarity measures for improved predictive performance. Specifically, the chemical space learned by deep neural networks has several advantages: (1) A smooth and continuous chemical space can be generated from discrete molecules in an automatic and data-driven fashion; (2) the continuous and hierarchical representations learned by the network are unique and more expressive; (3) fast gradient-based methods can be used for optimization of chemical properties. Concretely, this learned chemical space allows for new visualizations of chemical space to understand the diversity and biases of chemical libraries,[112] can improve performance across a range of predictive tasks or provide enriched leads, and accelerate molecular optimization.

A seminal report by Gomez-Bombarelli et al. illustrated these advantages through the application of a *variational autoencoder* (Figure 7) network.[113] Specifically, the autoencoder architecture consists of two parts: an *encoder* network that transforms an input molecule (in SMILES representation) into a reduced-dimensionality, chemical *latent space*, and a *decoder* network that maps points from this latent space back to a molecular output. The entire autoencoder is trained through an *unsupervised learning* approach that takes on a simple objective: to reconstruct its input. This unsupervised reconstruction task allows for large corpuses of unlabeled data or even hypothetical, drug-like molecules to learn a smooth representation of chemical space. In this vein, the authors trained the network on 250 000 drug-like molecules, represented by their SMILES strings, from ZINC15[114] to provide a chemical latent space for small-molecule exploration. Furthermore, by coupling this latent space to a neural network on labeled examples, Gomez-Bombarelli et al. illustrate that powerful gradient-based optimization methods, when combined with Bayesian inference, can be used to guide molecular optimization in tasks such as water−octanol partition toward new libraries with desirable properties. Critically, the decoder network allows for points in this chemical space to be directly sampled, allowing humans to visualize and assess the different regions of this learned space. This *generative* aspect of this network further provides opportunities in *de novo* drug design, as detailed in the following section.

4.3. Learning New Molecules with Generative Models.

Inverse molecular design is a longstanding challenge in modern drug discovery. Whereas standard QSAR models map molecular structures to an activity or property, inverse QSAR models turn this notion on its head, instead seeking to generate new molecular structures that satisfy optimal properties or activities.[24] On paper, this approach is highly attractive as a hypothesis generator for new molecules with desirable properties, yet historically, inverse-design models have been unable to map a continuous activity or property back to discrete, viable molecules.[115,116] The intractability of this problem has instead led researchers to adopt virtual screening approaches to rapidly

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

evaluate pre-enumerated compound libraries of synthetically accessible molecules.[117] *Generative models* in deep learning now directly address the inverse design problem, leading to new opportunities for *de novo* drug design.

As discussed in section 4.2 above, the approach described by Gomez-Bombarelli et al. using a variational autoencoder constitutes a type of generative model (Figure 7). The key to this generative process is the same primary objective as any autoencoder: to reconstruct its input. In the classic scenario, the encoder learns to compress the original input into a high-dimensional, chemical latent space, and the decoder learns to generate molecules from this space. Critically, the decoding process learned through the reconstruction task is a *direct* solution to inverse molecular design. To exploit this chemical space for molecules with specific properties, the latent space can be tuned through joint training with a prediction network, allowing for specific regions of the space to be sampled and new molecules to be generated. In this example, a variety of tasks include the generation of molecules with specific octanol−water coefficients, synthetic accessibility score (SAS),[118] and quantitative estimation of drug-likeness (QED).[119] Although the predictive tasks in this study are relatively simple in the scope of drug discovery, this study demonstrates the feasibility of this new conceptual framework and highlighted future challenges in the area.

The ability to generate focused libraries for *de novo* drug design has inspired numerous approaches, including a variety of autoencoder[120] and recurrent neural network[96] architectures. Input representation, however, remains a critical aspect of these generative models that must ultimately learn the rules or syntax of any molecular input format. To date, most generative models have largely focused on SMILES representations for inputs and generated outputs, and early studies found that only a fraction of generated SMILES corresponded to valid molecules or that generated molecules were unrealistic.[113] To this end, the development of new architectures and increased training set sizes have led to vastly improved results. For example, a recent study by Popova et al. used stack-RNNs[121] to generate millions of molecules with 95% structural validity.[99] The authors compared 1 million generated molecules against 320 million drug-like molecules from ZINC15[114] and found that 3% existed in the ZINC15 database. Along with further evidence, these studies suggested that the generated molecules were both new and drug-like.

Whereas these models have largely operated on SMILES strings, models that *directly* produce molecular graphs remain attractive. Although encoding graphs is straightforward, the *generation* of graphs is significantly more challenging.[122] Active research in this area has recently provided methods to generate potential small molecules with excellent graph validity, including *junction-tree variational autoencoders* by Jin et al.[123] and *MolGAN* by DeCao and Kipf.[124] Together, these generative models can act as hypothesis generators for drug design and discovery.

Deep generative models are beginning to directly address longstanding challenges in small molecule drug design. However, these innovations do not come without practical challenges; specifically, both the novelty and accessibility of generated molecules must be considered. Early reports focused largely on generating valid graphs that correspond to physically valid molecules. As models have progressed, the focus has shifted toward evaluating novelty.[125,126] Critically, if generative models are to guide drug design, they cannot merely produce trivial extensions of the training data set. It remains unclear whether the latent spaces of generative models, which effectively interpolate across the chemical space of the training data, are capable of usefully *extrapolating* into new regions of chemical structure space. Furthermore, current generative models are torn between novelty and accessibility. Vendor catalog libraries enumerate molecules that are guaranteed (or highly likely) to be synthetically accessible. In contrast, even generated molecules that appear drug-like may lack a clear synthetic route. The difficulty in validating chemical matter from these hypothesis generators still drastically limits translation of generative models into practical use.

With these considerations, two key questions remain open in the field: (1) Can generative models be practically applied to prospective discovery? (2) How do we evaluate their success if testing of their predictions is difficult? Recent reports have begun to prospectively assess generative models through synthesis and experimentation.[127,128] Zhavoronkov et al. trained a deep generative model for the discovery of discoidin domain receptor family member 1 (DDR1) kinase inhibitors with training data from the literature.[128] The authors used a SMILES-based autoencoder to propose candidate DDR1 kinase inhibitors and synthesized six candidates for testing, in collaboration with a contract research organization experienced in the manufacture of drug-like small molecules. The top compound exhibited an impressive IC_{50} of 10 nM, and the integrated study was completed within weeks. In response, Walters and Murcko comment that the top generated molecule has striking similarity to a known DDR1 inhibitor,[129−131] which substantiates concerns of limited novelty in generative drug design.

We must consider generative model practicality against alternative approaches that rely instead on the rapid screening of large pre-enumerated chemical libraries. For example, Stokes et al. recently adopted a virtual screening approach for the discovery of novel antibiotics.[132,133] Prospective testing of predicted antibiotics from the Drug Repurposing Library identified a new broad-spectrum antibiotic, halicin. Furthermore, virtual screening of the ZINC15[114] database produced promising hits. This work illustrates the effectiveness of deep learning in a screening workflow that exclusively considers synthesizable and valid molecules. Prospective validation still cannot scale well against the number of predictions available, both for enumeration-based and generative approaches, although "active" or "online" learning frameworks may at least focus experimentation where it is most informative.[24] While generative models may conceptually be an inherently more satisfying approach to surveying vast chemical spaces than enumerative alternatives, pragmatic considerations in their use and evaluation nonetheless currently limit this potential.

4.4. Learning Shared Representations with Multitask and Transfer Learning. The multidimensional optimization of affinity and physicochemical properties is a central challenge in small-molecule therapeutic discovery. A small-molecule lead must be simultaneously optimized with multiple objectives: (1) to maintain high affinity to its intended target, (2) to improve desirable physicochemical properties that dictate its adsorption, distribution, metabolism, and excretion (ADME) properties, (3) to preserve selectivity against undesired off-targets. Heuristic guidelines, such as Lipinski's rule of five,[134,135] are general recommendations to lead optimization. However, this process remains experimentally laborious, as numerous analogs must be synthesized and tested. Predictive methods that simultaneously evaluate multiple targets and properties, and

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

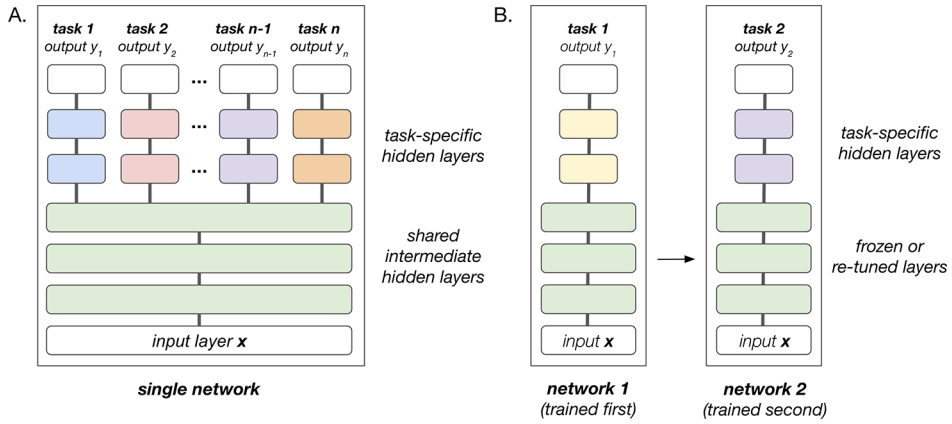


Figure 8. Approaches to learning shared representations across different tasks. (A) A multitask network simultaneously predicts multiple targets using shared hidden layers. These can correspond to multiple different protein targets or properties, and are jointly trained as a single network. (B) Transfer learning occurs sequentially: A network is first trained on one task, and the early layers of the network are reused as feature extractors in a subsequent, typically related task.

methods that capitalize on historical data across programs, can accelerate molecular optimization and discovery efforts. To this end, the machine learning concepts of *multitask learning*[136] and *transfer learning*[137] are of considerable interest in drug discovery. Just as humans intuitively apply information and knowledge gained from one problem to solve a new one, these approaches leverage knowledge gained from one predictive task to facilitate another, providing better predictive performance or requiring fewer examples for training.

In contrast to *single task learning*, where a model considers an individual prediction, *multitask learning* encompasses models that explicitly consider two or more tasks at once, such as predicting entire molecular target profiles simultaneously (Figure 8A), and naturally reflects the polypharmacological optimization of drug design. In general, multitask networks share internal hierarchical representations that can exploit similarities and nuanced differences across tasks, resulting in improved learning efficiency and model performance compared to single-task models.[136] In the context of medicinal chemistry, bioactivity data collected on one protein target often inform us about another; e.g., a model that learns important aspects of small molecule inhibition of one protein kinase should aid in learning important features for other protein kinases, as many inhibitors target a highly conserved ATP-binding site.[59]

Empirically, multitask learning strategies can improve model performance and robustness. For instance, the winning multitask model developed by Dahl et al. for the Merck Molecular activity challenge, discussed above, outperformed models trained on a single target at a time.[78] Ramsundar et al. demonstrated that multitask networks can be applied across hundreds of diverse protein targets simultaneously, with a modest performance increase.[138] Although these early reports sparked enthusiasm over the potential of multitask networks, systematic studies have delineated some limits of this approach and begun to decipher practical use cases. Investigations by Kearnes et al. on modeling real-world industrial data sets confirmed that modest performance benefits can be gleaned through a multitask approach; however, adding large numbers of auxiliary tasks was not guaranteed to improve model performance compared to simpler multitask models.[139] Xu et al. further studied the effectiveness of multitask learning for activity prediction, illustrating that multitask learning is often beneficial when task activities are correlated but also revealed cases of *negative transfer*, where the addition of another task is *detrimental*.[80] As an example, Rodríguez-Pérez and Bajorath recently investigated multitask neural networks on a data set of 19 030 kinase inhibitors across 103 human kinases.[140] As these tasks are largely related, the authors observed consistent improvements using multitask models over single-task models. Fare et al. also evaluated multitask networks and explicitly considered task selection from a suite of auxiliary tasks that include activities (e.g., anti-HIV and antimalarial activities) and physicochemical property properties (e.g., experimental thermochemical properties and solubilities).[141] In this study, the authors developed a pairwise score to prioritize auxiliary tasks for multitask training and again observed that selecting correlated tasks tend to improve multitask network performance. All together, these studies illustrate that although multitask learning can improve model performance, task selection is critical for building these networks.

Whereas multitask learning leverages shared underlying hierarchical representations, representations may also be *transferred* between tasks (Figure 8B). In the case of *transfer learning*, a fully trained network trained on one task may be subsequently applied to another, with the underlying assumption that information learned from one domain may be useful for another. Following the kinase example from above, a fully trained model to predict inhibition of one kinase inhibitor might be carefully retrained on a *different* kinase inhibitor. This strategy has proven exceptionally effective in computer vision, where a network trained on millions of images from ImageNet (e.g., cats, dogs, etc.) can subsequently achieve high classification performance when *fine-tuned* on a few hundred new training examples. Furthermore, the same ImageNet models can be subsequently fine-tuned across fields, such as classifying skin cancer lesions at expert dermatologist levels.[142]

Transferring knowledge from existing and historical chemical data to new data sets with limited examples would similarly enable drug discovery. In parallel to the successes of computer vision, the ability to build strong and generalizable models from only a few hundred examples could significantly impact the discovery pipeline. In an ideal case, pretraining a large neural network model on a sufficiently massive and diverse chemical data set would compensate for the small number of observations in a new medicinal chemistry campaign. However, in contrast, drug discovery spans a greater diversity of underlying chemical

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

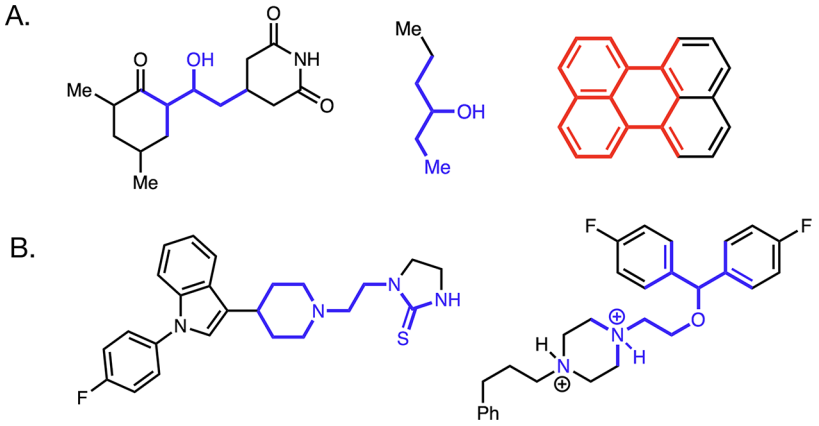


Figure 9. Feature interpretability in deep neural networks. (A) Duvenaud et al.'s analysis of neural graph fingerprints to identify the most (blue) and least (red) favorable substructural motifs in an aqueous solubility task. (B) Chen et al.'s report on a deep reinforcement learning model to general explicit rationales. These models aptly identify lipophilic, basic amine motifs (blue highlight) responsible for inhibition of hERG.

and physical processes, and data sets operate on a much smaller scale. These limitations currently hinder both the investigation and development of more universal models for chemical prediction, and the current scope and limitations of transfer learning are not well understood. Efforts to better understand transferability between different models are ongoing, and early reports illustrate that supervised pretraining between similar bioactivity and property prediction tasks can offer modest improvements.[141,143]

In a recent example, Hu and Liu et al. evaluated transfer learning strategies for graph neural networks in the context of biological and chemical prediction tasks.[143] Specifically, the authors investigated whether representations learned from training on a subset of ChEMBL bioactivity data, consisting of 456K molecules across 1310 protein targets, could be used to improve performance on a variety of smaller bioactivity tasks from MoleculeNet.[144] Empirically, the authors find that a careful pretraining procedure at both the atom and molecule levels (node- and graph-levels) can improve downstream model performance by an average of 7.2% in ROC-AUC (receiver operator characteristic, area under the curve) when compared to models trained from scratch. However, the authors find that the choice of training task and graph neural network architecture strongly dictated the performance. Their experiments demonstrate both examples of improved performance (*positive transfer*) and decreased performance (*negative transfer*) depending on the training setup. These observations highlight future opportunities for future investigation into transfer learning approaches.

4.5. Interpretability Matters for Molecular Deep Learning. Deep neural networks can effectively learn patterns automatically from data, yet automatic learning does not guarantee meaningful learning. These powerful algorithms can uncover complex relationships within data, but without careful data curation, this can lead to unintended pattern recognition or learning of hidden biases. Although machine learning is quantitative by nature, relying solely on performance metrics such as accuracy and the receiver operator characteristic often leads to overoptimistic models that do not generalize to new data.[145] To generate actionable new scientific knowledge instead of simply fitting data, deep learning models must learn truly salient patterns that reflect underlying physical processes.[54]

Neural networks have long been known to sacrifice interpretability for performance,[146,147] but preliminary methods for *model interpretability*[148] now shed light on the internal decision processes of neural networks, showing which features are salient to final predictions.[149] For example, systematic studies that quantify the impact of removing features on predictive performance, a process known as *feature ablation*,[150] have been used for decades to understand the importance of individual features to model predictions. More recently, gradient-based saliency maps[86,151] and attention-based[152,153] models highlight regions of images and keywords in text that are most important for performance in computer vision and natural language processing. Applying this toolkit to drug design can provide insight into which molecular features influence predictive performance, driving the development of more robust and generalizable models.

Several early reports illustrate the utility of these tools in understanding models trained on property prediction tasks (Figure 9). For example, Mayr et al. manually examined the hidden units of a trained network to reveal substructures that contribute to molecular toxicity.[154] Duvenaud et al. analyzed graph neural networks trained for aqueous solubility (Figure 9A, left). Their findings aligned with chemical intuition: models considered hydroxyl-containing motifs most important for solubility and extended polyaromatic systems most predictive for insolubility (Figure 9A, right).[102] Whereas these post hoc approaches provide insight into implicit, black-box processes, other models explicitly integrate interpretability into their design. For example, Chen et al. developed a deep reinforcement learning approach to generate explicit rationales of small-molecule bioactivity prediction.[155] Models trained to predict inhibitors of the human ether-a-go-go-related gene (hERG) aptly identified lipophilic, basic tertiary amines as key structural motifs, consistent with expert intuition (Figure 9B).[156,157] Importantly, in all cases the networks *learn* these structural motifs without expert-encoded knowledge, illustrating the applicability of interpretability methods to small molecules.

However, the studies above largely serve as sanity checks on simple tasks, where the examples shown are consistent with expert intuition. In these cases, a relatively small set of local features are known to drive molecular properties through experimental mechanistic confirmation (e.g., a hydroxyl group for solubility).[147] An important caveat to these studies is that the examples shown are confirmatory and nonexhaustive. In the absence of more rigorous and systematic testing, these interpretations may be subject to confirmation bias.[158] For example, a systematic study by Sheridan on interpreting

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

machine learning QSAR models demonstrates that feature ablation alone can produce misleading results and overlook even ideal cases where individual atom contribution is already known.[159] Furthermore, useful applications in drug discovery attempt to model complex physical processes. In new applications, little may be known about the underlying chemistry, and a combination of features may drive observed molecular properties. Interpretability studies do not yet provide a simple answer, and methods that enforce interpretability may additionally sacrifice model performance.[146,160]

No single test is guaranteed to provide a useful answer. Instead, interpretability tools allow us to interrogate models with thoughtful experimental design. A hypothesis-driven approach can test whether models learn what is relevant to the underlying causal process and flag instances of unintended pattern recognition, learning spurious correlations, and data set bias. For example, McCloskey et al. studied synthetic data sets of drug activity constructed such that the true underlying activities are predefined but hidden.[145] By applying the technique of integrated-gradients,[161] the authors illustrated that neural networks trained on these data sets readily capitalized on spurious correlations to achieve perfect accuracy on held-out test sets. This study demonstrates that performance metrics alone are not reliable evidence of generalizability and further shows how the thoughtful application of attribution methods can be effectively used for scientific inquiry.

Understanding what models learn not only improves model robustness but also opens avenues for hypothesis generation in drug development.[147] For example, integrating explainable models into a discovery setting can augment medicinal chemists' decision processes for hit-to-lead optimization: a model that explores binding activity may remind chemists to preserve core motifs most salient for activity and direct chemists to explore perturbations that improve ADME properties. Rather than viewing these deep learning models as a replacement for medicinal chemists, these tools provide an opportunity to direct optimization campaigns. Explainable models can distill implicit patterns encoded within an empirical data set into standalone structure−activity hypotheses that can be tested explicitly. Likewise, the shift from expert-defined features to feature learning does not obviate the need for expert knowledge or hypothesis-driven research; instead, domain expertise will be crucial to ensure that models learn meaningful patterns. We anticipate that interpretability provides a bridge between practicing chemists and deep learning models that will further enhance our scientific understanding in a virtuous cycle.

4.6. Limitations of Deep Learning. As with any promising new technology, the strengths and opportunities outlined above are accompanied by unique challenges and limitations. To readily integrate deep learning as a practical method in drug discovery pipelines, we must address the following.

Data and Data Set Considerations. In comparison to other machine learning algorithms, deep neural networks require much larger amounts of data to ensure model generalizability and prevent overfitting. Massive open data sets play an important role in the success of deep learning, yet chemical data sets frequently suffer from comparatively small sample sizes and limited chemical diversity. As illustrated conceptually in Figure 3, deep neural networks excel in massive data set regimes that are rarely accessible for many medicinal chemistry applications. The current practical advantage of deep learning, particularly for early stage discovery projects, is especially limited. In these settings, traditional fingerprints such as ECFP in combination with random forests remain an effective and practical choice.

Although open-source data sets, including ZINC15[114] and ChEMBL,[95] as well as public benchmarks[144] have enabled many of the early studies covered in this review, building more robust and generalizable models requires improved data and more rigorous benchmarks. Aggregated and literature-based databases continue to suffer from publication and analog bias, missing data, and an underrepresentation of negative data points. Additionally, high data variability and data quality are still significant challenges. In parallel, many existing benchmarks are small, not diverse, and can easily be overfit. Despite these limitations, a handful of studies using high-quality data at larger scale demonstrate that neural network approaches can provide practical benefit to drug discovery in real-world settings.[111,162] Improving accessibility of large amounts of high quality data and the careful development of challenging and realistic benchmarks will together drive innovation and improve the applicability of deep learning models.

Training Cost. Unlike established molecular featurization and fingerprinting methods, such as the MACCS keys and ECFP that are fast to compute on scale, most of the learning methods discussed above require time-intensive training and optimization steps. Parallel computation on GPUs has accelerated model development, yet these methods are still slower than existing fingerprinting algorithms that do not require learning parameters.

Deep Learning Experience and Expertise. The flexible nature of deep learning requires careful data set curation, model training, and evaluation procedures to ensure that models generalize. Despite open-source initiatives and the integration of deep learning packages into scientific computing software,[144] basic coding knowledge is required to implement most state-of-the-art models. As a result, deep learning does not currently offer a turnkey solution for predictive modeling of small molecules. At present, simpler machine learning models are more readily integrated into fully automated discovery pipelines.

Reproducibility. Deep learning models are trained via stochastic initialization and optimization and can be exquisitely sensitive to their settings ("hyperparameters"). The stochastic nature of these methods, along with the rapid development of the field, has led to a reproducibility crisis in AI.[163] Per Gundersen and Kjensmo, "Reproducibility in empirical AI research is the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team."[164] Reproducible procedures and access to original data are necessary to provide actionable scientific insight to the greater community.[165] In order to build useful and practical models, best practices must be established to encourage researchers to report detailed experimental procedures and open-source code repositories that include trained models and to release the precise training and test data used in their studies.[166] Adhering to these best practices and data sharing efforts will improve the odds that new research in deep learning effectively translates to real-world drug discovery. Active research in this area and open-source initiatives and data-sharing efforts are currently underway in the community to address these limitations, and we anticipate that continued work will allow for the democratization of deep learning in drug discovery.

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

5. FUTURE DIRECTIONS, OUTLOOK, AND CONCLUDING REMARKS

Recent years have seen an explosion in deep learning research and innovation. Despite high expectations for drug discovery, deep learning techniques alone are not a panacea. Rather, these approaches offer value in addressing concrete challenges within small molecule predictive modeling and require further development before integration into practical discovery pipelines. Many of the techniques overviewed in this review are currently being evaluated in prospective drug discovery. The interdisciplinary teams that closely integrate computation and experiment will soon see these methods guide practical innovation toward real experimental discovery.

Looking ahead, molecular representations that capture the dynamics of complex systems will be of increasing importance. Representations that faithfully encode aspects of three-dimensional spatial relationships, conformational dynamics, and kinetic pathways will build a strong foundation for future predictive tasks. These innovations, alongside concurrent deep learning work in structural biology and proteomics, will allow for integrative modeling of higher-order, complex systems. Deep learning approaches will drive new hypotheses and experimental procedures by considering complex physical systems from the atomic to protein resolution. In total, we anticipate that the future is bright for deep learning in small-molecule innovation.

AUTHOR INFORMATION

Corresponding Author

Michael J. Keiser − *Department of Pharmaceutical Chemistry, Department of Bioengineering & Therapeutic Sciences, Institute for Neurodegenerative Diseases, Kavli Institute for Fundamental Neuroscience, Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California 94143, United States;* orcid.org/0000-0002-1240-2192; Phone: 415-886-7651; Email: keiser@keiserlab.org

Authors

Kangway V. Chuang − *Department of Pharmaceutical Chemistry, Department of Bioengineering & Therapeutic Sciences, Institute for Neurodegenerative Diseases, Kavli Institute for Fundamental Neuroscience, Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California 94143, United States;* orcid.org/0000-0002-0652-8071

Laura M. Gunsalus − *Department of Pharmaceutical Chemistry, Department of Bioengineering & Therapeutic Sciences, Institute for Neurodegenerative Diseases, Kavli Institute for Fundamental Neuroscience, Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California 94143, United States;* orcid.org/0000-0003-3444-5617

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.0c00385

Notes

The authors declare no competing financial interest.

Biographies

Kangway V. Chuang is an Arnold O. Beckman Postdoctoral Fellow at the University of California, San Francisco. He obtained both his B.S. and Ph.D. in Chemistry from the California Institute of Technology. His research interests lie at the intersection of chemistry and machine learning.

Laura M. Gunsalus is a graduate student at the University of California, San Francisco. She obtained her B.S. in Neuroscience from Carnegie Mellon University. Her research explores learning meaningful representations of physical and biological systems.

Michael J. Keiser is a Chan Zuckerberg Initiative Ben Barres Investigator and an Allen Distinguished Investigator. He joined the UCSF faculty as an Assistant Professor in 2014, in the Department of Pharmaceutical Chemistry and the Institute for Neurodegenerative Diseases, and holds appointments in the Department of Bioengineering & Therapeutic Sciences, the Kavli Institute for Fundamental Neuroscience, and the Bakar Computational Health Sciences Institute. Before this, he cofounded a startup bringing systems pharmacology methods for drug−target prediction to pharma and the U.S. FDA. He holds degrees from Stanford, including a B.Sc. in Computer Science. His lab combines machine learning with chemical biology methods to investigate how drug-like small molecules perturb protein networks to achieve their therapeutic effects.

ACKNOWLEDGMENTS

We gratefully acknowledge Dr. Madeleine Kieffer, Dr. Ziyang Zhang, and members of the Keiser laboratory at UCSF for insightful discussions and helpful feedback. We thank the Arnold and Mabel Beckman Foundation for generous research support to K.V.C. through an Arnold O. Beckman Postdoctoral Fellowship, and the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation, for Grant 2018-191905 to M.J.K.

ABBREVIATIONS USED

ADME, absorption, distribution, metabolism, excretion; ANN, artificial neural network; CNN, convolutional neural network; RNN, recurrent neural network; DNN, deep neural network; DL, deep learning; ECFP, extended connectivity fingerprint; GNN, graph neural network; GPU, graphics processing unit; InChI, international chemical identifier; LSTM, long short-term memory unit; MACCS, molecular access system; ML, machine learning; NN, neural network; QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; ROCS, rapid overlay of chemical structures; SGD, stochastic gradient descent; VAE, variational autoencoder; SMILES, simplified molecular input line entry system

REFERENCES

(1) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLoS One* **2012**, *7* (11), No. e48476.

(2) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (8), 3186−3204.

(3) Gomez, L. Decision Making in Medicinal Chemistry: The Power of Our Intuition. *ACS Med. Chem. Lett.* **2018**, *9* (10), 956−958.

(4) Wade, L. G.; Simek, J. W. Introduction and Review. In *Organic Chemistry*; Pearson: Glenview, IL, 2016; pp 1−41.

(5) Ellerbrock, P.; Armanino, N.; Ilg, M. K.; Webster, R.; Trauner, D. An Eight-Step Synthesis of Epicolactone Reveals Its Biosynthetic Origin. *Nat. Chem.* **2015**, *7* (11), 879−882.

(6) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35* (8), 1798−1828.

(7) Zheng, A.; Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*; O'Reilly Media, Inc.: Sebastopol, CA, 2018.

(8) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4* (5), 468−481.
(9) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. *J. Chem. Inf. Model.* **2008**, *48* (4), 766−784.
(10) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; John Wiley & Sons: Weinheim, Germany, 2009.
(11) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29* (6−7), 476−488.
(12) Grisoni, F.; Consonni, V.; Todeschini, R. Impact of Molecular Descriptors on Computational Models. *Methods Mol. Biol.* **2018**, *1825*, 171−209.
(13) Jain, A. N. Chemoinformatics for Drug Discovery. In *The Challenge of Creativity in Drug Design*; Bajorath, J., Ed.; John Wiley & Sons: Hoboken, NJ, 2013; pp 33−50.
(14) Ng, A. Y. Machine Learning and AI via Brain Simulations. https://helper.ipam.ucla.edu/publications/gss2012/gss2012_10595.pdf (accessed Apr 15, 2020).
(15) Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, 2016.
(16) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313* (5786), 504−507.
(17) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436−444.
(18) Olah, C. Neural Networks, Manifolds, and Topology. http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/ (posted April 6, 2014, accessed Jan 27, 2019).
(19) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2016.
(20) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. *arXiv* **2015**, arXiv:1510.02855.
(21) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57* (4), 942−957.
(22) Hochuli, J.; Helbling, A.; Skaist, T.; Ragoza, M.; Koes, D. R. Visualizing Convolutional Neural Network Protein-Ligand Scoring. *J. Mol. Graphics Modell.* **2018**, *84*, 96−108.
(23) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4* (11), 1520−1530.
(24) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(25) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23* (8), 1538−1546.
(26) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59* (6), 2545−2559.
(27) Panteleev, J.; Gao, H.; Jia, L. Recent Applications of Machine Learning in Medicinal Chemistry. *Bioorg. Med. Chem. Lett.* **2018**, *28* (17), 2807−2815.
(28) Xu, Y.; Yao, H.; Lin, K. An Overview of Neural Networks for Drug Discovery and the Inputs Used. *Expert Opin. Drug Discovery* **2018**, *13*, 1091−1102.
(29) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.
(30) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discovery* **2010**, *9* (4), 273−276.
(31) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(32) McNaught, A. The IUPAC International Chemical Identifier. *Chem. Int.* **2006**, *28* (6), 12−15.
(33) Ahrens, E. K. F. Customisation for Chemical Database Applications. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Berlin, 1988; pp 97−111.
(34) Christie, B. D.; Leland, B. A.; Nourse, J. G. Structure Searching in Chemical Databases by Direct Lookup Methods. *J. Chem. Inf. Model.* **1993**, *33* (4), 545−547.
(35) Bloom, B. H. Space/time Trade-Offs in Hash Coding with Allowable Errors. *Commun. ACM* **1970**, *13* (7), 422−426.
(36) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Model.* **1985**, *25* (2), 64−73.
(37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(38) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.
(39) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A.; Kelley, B. *J. Med. Chem.* **2010**, *53* (10), 3862−3886.
(40) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 128−136.
(41) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504−1519.
(42) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48* (5), 1489−1495.
(43) Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7* (17), 903−911.
(44) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725−732.
(45) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, *23* (25), 6118−6128.
(46) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *Advances in Neural Information Processing Systems 30*, Long Beach, CA, U.S., Dec. 4−9, 2017; Curran Associates, Inc.; pp 2607−2616.
(47) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604−610.
(48) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434−443.
(49) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572−1583.
(50) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589−604.
(51) Zitnik, M.; Agrawal, M.; Leskovec, J. Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics* **2018**, *34* (13), i457−i466.
(52) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7* (1), 3582.
(53) Sayle, R. A. So You Think You Understand Tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 485−496.
(54) Chuang, K. V.; Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chem. Biol.* **2018**, *13* (10), 2819−2821.
(55) Ng, A. Introduction to Deep Learning. http://cs230.stanford.edu/files/C1M1.pdf (accessed Apr 10, 2020).
(56) Baskin, I. I.; Winkler, D.; Tetko, I. V. A Renaissance of Neural Networks in Drug Discovery. *Expert Opin. Drug Discovery* **2016**, *11* (8), 785−795.

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

(57) Smith, J. S.; Roitberg, A. E.; Isayev, O. Transforming Computational Drug Discovery with Machine Learning and AI. *ACS Med. Chem. Lett.* **2018**, *9* (11), 1065−1069.
(58) Jordan, A. M. Artificial Intelligence in Drug Design-The Storm Before the Calm? *ACS Med. Chem. Lett.* **2018**, *9* (12), 1150−1152.
(59) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57* (8), 2068−2076.
(60) Hinton, G. E.; Rumelhart, D. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323* (9), 533−536.
(61) Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct 25−29, 2014; Association for Computational Linguistics, 2014; DOI: 10.3115/v1/d14-1179.
(62) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735−1780.
(63) Zhang, Z.; Cui, P.; Zhu, W. Deep Learning on Graphs: A Survey. *arXiv* **2018**, arXiv:1812.04202.
(64) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85−117.
(65) Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, May 13−15, 2010; PMLR, 2010; pp 249−256.
(66) Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* **2009**, *24* (2), 8−12.
(67) Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, Apr 11−13, 2011; PMLR, 2011; pp 315−323.
(68) Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, U.S., Jun 17−19, 2013; JMLR.org, 2013; pp 1139−1147.
(69) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference for Learning Representations*, San Diego, CA, May 7−9, 2015; https://arxiv.org/abs/1412.6980.
(70) Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, Jul 6−11, 2015; JMLR.org, 2015; pp 448−456.
(71) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, Jun 27−30, 2016 ;IEEE, 2016; pp 770−778.
(72) Hiller, S. A.; Golender, V. E.; Rosenblit, A. B.; Rastrigin, L. A.; Glaz, A. B. Cybernetic Methods of Drug Design. I. Statement of the problem—The Perceptron Approach. *Comput. Biomed. Res.* **1973**, *6* (5), 411−421.
(73) Devillers, J. *Neural Networks in QSAR and Drug Design*; Academic Press: San Diego, CA, 1996.
(74) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.
(75) Czermiński, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Quant. Struct.-Act. Relat.* **2001**, *20* (3), 227−240.
(76) Merck Molecular Activity Challenge. https://www.kaggle.com/c/MerckActivity (accessed Mar 11, 2019).
(77) Markoff, J. Scientists See Advances in Deep Learning a Part of Artificial Intelligence. *N. Y. Times* **2012** November 24.
(78) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. *arXiv* **2014**, arXiv:1406.1231.
(79) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263−274.
(80) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57* (10), 2490−2504.
(81) Schneider, P.; Walters, W. P.; Plowright, A. T.; Sieroka, N.; Listgarten, J.; Goodnow, R. A.; Fisher, J.; Jansen, J. M.; Duca, J. S.; Rush, T. S.; Zentgraf, M.; Hill, John E.; Krutoholow, E.; Kohler, M.; Blaney, J.; Funatsu, K.; Luebkemann, C.; Schneider, G. Rethinking Drug Design in the Artificial Intelligence Era. *Nat. Rev. Drug Discovery* **2020**, *19*, 353−364.
(82) Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Science & Business Media: New York, NY, 2010.
(83) Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115* (3), 211−252.
(84) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60* (6), 84−90.
(85) Bengio, Y.; Courville, A. Deep Learning of Representations. In *Handbook on Neural Information Processing*; Bianchini, M., Maggini, M., Jain, L. C., Eds.; Springer: Berlin, 2013; pp 1−28.
(86) Zeiler, M. D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Computer Vision—ECCV 2014*, Zurich, Switzerland, Sep 6−12, 2014; Springer International Publishing, 2014; pp 818−833.
(87) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26*, Lake Tahoe, NV, U.S., Dec 5−10, 2013; Curran, 2013; pp 3111−3119.
(88) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the 1st International Conference on Learning Representations*, Scottsdale, AZ, May 2−4, 2013; https://arxiv.org/abs/1301.3781.
(89) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. *Workshop Paper at the 28th Conference on Neural Information and Processing Systems*, Montreal, CA, Dec 8−13, 2014; NIPS; Vol. *27*, pp 1−9.
(90) Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A. Organic Chemistry as a Language and the Implications of Chemical Linguistics for Structural and Retrosynthetic Analyses. *Angew. Chem., Int. Ed.* **2014**, *53* (31), 8108−8112.
(91) Woźniak, M.; Wołos, A.; Modrzyk, U.; Górski, R. L.; Winkowski, J.; Bajczyk, M.; Szymkuć, S.; Grzybowski, B. A.; Eder, M. Linguistic Measures of Chemical Diversity and the "keywords" of Molecular Collections. *Sci. Rep.* **2018**, *8*, 7598.
(92) Jastrzębski, S.; Leśniak, D.; Czarnecki, W. M. Learning to SMILE(S). *arXiv* **2016**, arXiv:1602.06289.
(93) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. SMILES-Based Optimal Descriptors: QSAR Analysis of Fullerene-Based HIV-1 PR Inhibitors by Means of Balance of Correlations. *J. Comput. Chem.* **2010**, *31* (2), 381−392.
(94) Worachartcheewan, A.; Mandi, P.; Prachayasittikul, V.; Toropova, A. P.; Toropov, A. A.; Nantasenamat, C. Large-Scale QSAR Study of Aromatase Inhibitors Using SMILES-Based Descriptors. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 120−126.
(95) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945−D954.
(96) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120−131.

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722

(97) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv*, **2018**, https://doi.org/10.26434/chemrxiv.7097960.v1.
(98) Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv* **2017**, arXiv:1703.07076.
(99) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4* (7), No. eaap7885.
(100) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. SELFIES: A Robust Representation of Semantically Constrained Graphs with an Example Application in Chemistry. *arXiv* **2019**, arXiv:1905.13741.
(101) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning—Volume 70*, Sydney, NSW, Australia, Aug 6−11, 2017;JMLR.org, 2017; pp 1263−1272.
(102) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*, Montreal, Quebec, Canada, Dec 7−12, 2015; Curran Associates, Inc.; pp 2224−2232.
(103) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30* (8), 595−608.
(104) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-like Molecules. *J. Chem. Inf. Model.* **2013**, *53* (7), 1563−1575.
(105) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3* (4), 283−293.
(106) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.
(107) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. In *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, Dec 4−9, 2017; Curran Associates, Inc.; pp 991−1001.
(108) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley-Interscience: New York, NY, 1990.
(109) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54* (22), 7739−7750.
(110) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932−2942.
(111) Feinberg, E. N.; Sheridan, R.; Joshi, E.; Pande, V. S.; Cheng, A. C. Step Change Improvement in ADMET Prediction with Potential Net Deep Featurization. *arXiv* **2019**, arXiv:1903.11789.
(112) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62* (3), 1116−1124.
(113) 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* (2), 268−276.
(114) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324−2337.
(115) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from Y to X). *J. Chem. Inf. Model.* **2016**, *56* (2), 286−299.
(116) Miyao, T.; Funatsu, K.; Bajorath, J. Exploring Differential Evolution for Inverse QSAR Analysis. *F1000Research* **2017**, *6*, 1285.
(117) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224−229.
(118) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1* (1), 8.
(119) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90−98.
(120) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15* (10), 4398−4405.
(121) Joulin, A.; Mikolov, T. Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. In *Advances in Neural Information Processing Systems 28*, Montreal, Quebec, Canada, Dec 2−8, 2015; Curran Associates, Inc.; pp 190−198.
(122) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *Proceedings of the 27th Conference on Artificial Neural Networks and Machine Learning*, Rhodes, Greece, Oct 4−7, 2018; Springer, 2018; pp 412−422.
(123) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, Jul 10−15, 2018; PMLR, 2018; pp 2323−2332.
(124) De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *arXiv* **2018**, arXiv:1805.11973.
(125) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58* (9), 1736−1741.
(126) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Vladimir Veselov, Mark; Kadurin, A.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv* **2018**, arXiv:1811.12823.
(127) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37* (1−2), 1700153.
(128) 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.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038−1040.
(129) Walters, W. P.; Murcko, M. Assessing the Impact of Generative AI on Medicinal Chemistry. *Nat. Biotechnol.* **2020**, *38* (2), 143−145.
(130) Zhavoronkov, A.; Aspuru-Guzik, A. Reply to "Assessing the Impact of Generative AI on Medicinal Chemistry.". *Nat. Biotechnol.* **2020**, *38* (2), 146−146.
(131) Gao, M.; Duan, L.; Luo, J.; Zhang, L.; Lu, X.; Zhang, Y.; Zhang, Z.; Tu, Z.; Xu, Y.; Ren, X.; Ding, K. Discovery and Optimization of 3-(2-(Pyrazolo[1,5-A]pyrimidin-6-Yl)ethynyl)benzamides as Novel Selective and Orally Bioavailable Discoidin Domain Receptor 1 (DDR1) Inhibitors. *J. Med. Chem.* **2013**, *56* (8), 3281−3295.
(132) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370−3388.
(133) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688−702.
(134) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1−3), 3−26.

(135) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technol.* **2004**, *1* (4), 337−341.
(136) Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28* (1), 41−75.
(137) Pratt, L.; Jennings, B. A Survey of Connectionist Network Reuse through Transfer. In *Learning to Learn*; Springer: Boston, MA, 1996; pp 19−43.
(138) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv* **2015**, arXiv:1502.02072.
(139) Kearnes, S.; Goldman, B.; Pande, V. Modeling Industrial ADMET Data with Multitask Networks. *arXiv* **2016**, arXiv:1606.08793.
(140) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4* (2), 4367−4375.
(141) Fare, C.; Turcani, L.; Pyzer-Knapp, E. O. Powerful, Transferable Representations for Molecules through Intelligent Task Selection in Deep Multitask Networks. *arXiv* **2018**, arXiv:1809.06334.
(142) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542* (7639), 115−118.
(143) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-Training Graph Neural Networks. *arXiv* **2019**, arXiv:1905.12265.
(144) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513−530.
(145) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (24), 11624−11629.
(146) Johansson, U.; Sönströd, C.; Norinder, U.; Boström, H. Trade-off between Accuracy and Interpretability for Predictive in Silico Modeling. *Future Med. Chem.* **2011**, *3* (6), 647−663.
(147) Polishchuk, P. Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57* (11), 2618−2639.
(148) Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608.
(149) Castelvecchi, D. Can We Open the Black Box of AI? *Nature* **2016**, *538* (7623), 20.
(150) Cohen, P. R.; Howe, A. E. How Evaluation Guides AI Research: The Message Still Counts More than the Medium. *AI Mag.* **1988**, *9* (4), 35.
(151) Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V. N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. *Proceedings*, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, March 12 −15, 2018; IEEE, 2018; DOI: 10.1109/wacv.2018.00097.
(152) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations*, San Diego, CA, U.S., May 7−9, 2015; http://arxiv.org/abs/1409.0473.
(153) Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, June 12−17, 2016; Association for Computational Lingustics, 2016; pp 1480−1489.
(154) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 24.
(155) Chen, B.; Coley, C.; Barzilay, R.; Jaakkola, T. Using Deep Reinforcement Learning to Generate Rationales for Molecules. https://openreview.net/pdf?id=HkJ1rgbCb, 2018.
(156) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440* (7083), 463−469.
(157) Cavalli, A.; Buonfiglio, R.; Ianni, C.; Masetti, M.; Ceccarini, L.; Caves, R.; Chang, M. W. Y.; Mitcheson, J. S.; Roberti, M.; Recanatini, M. Computational Design and Discovery of "Minimally Structured" hERG Blockers. *J. Med. Chem.* **2012**, *55* (8), 4010−4014.
(158) Nickerson, R. S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.* **1998**, *2* (2), 175−220.
(159) Sheridan, R. P. Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It? *J. Chem. Inf. Model.* **2019**, *59* (4), 1324−1337.
(160) Lipton, Z. The Mythos of Model Interpretability. *Commun. ACM* **2018**, *16* (3), 36.
(161) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70*, Sydney, NSW, Australia, Aug 6−11, 2017; PMLR, 2017; pp 3319−3328.
(162) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; Cuozzo, J. W.; Guié, M.-A.; Guilinger, J. P.; Huguet, C.; Hupp, C. D.; Keefe, A. D.; Mulhern, C. J.; Zhang, Y.; Riley, P. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit-Finding. *arXiv* **2020**, arXiv:2002.02530.
(163) Hutson, M. Artificial Intelligence Faces Reproducibility Crisis. *Science* **2018**, *359* (6377), 725−726.
(164) Gundersen, O. E.; Kjensmo, S. State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, Feb 2−7, 2018; AAAI, 2018; pp 1644−1651.
(165) Walters, W. P. Where's The Code? https://practicalcheminformatics.blogspot.com/2019/05/wheres-code.html (posted May 3, 2019, accessed Apr 10, 2020).
(166) Schaduangrat, N.; Lampa, S.; Simeon, S.; Gleeson, M. P.; Spjuth, O.; Nantasenamat, C. Towards Reproducible Computational Drug Discovery. *J. Cheminf.* **2020**, *12* (1), 9.

https://dx.doi.org/10.1021/acs.jmedchem.0c00385
J. Med. Chem. 2020, 63, 8705−8722