Geometric Deep Learning on Molecular Representations

Kenneth Atz[1,†], Francesca Grisoni[2,1,†*], Gisbert Schneider[1,3*]

[1]ETH Zurich, Dept. Chemistry and Applied Biosciences, RETHINK, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
[2]Eindhoven University of Technology, Dept. Biomedical Engineering, Groene Loper 7, 5612AZ Eindhoven, Netherlands.
[3]ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, Singapore.
† these authors contributed equally to this work
*f.grisoni@tue.nl, gisbert@ethz.ch

Abstract

Geometric deep learning (GDL), which is based on neural network architectures that incorporate and process symmetry information, has emerged as a recent paradigm in artificial intelligence. GDL bears particular promise in molecular modeling applications, in which various molecular representations with different symmetry properties and levels of abstraction exist. This review provides a structured and harmonized overview of molecular GDL, highlighting its applications in drug discovery, chemical synthesis prediction, and quantum chemistry. Emphasis is placed on the relevance of the learned molecular features and their complementarity to well-established molecular descriptors. This review provides an overview of current challenges and opportunities, and presents a forecast of the future of GDL for molecular sciences.

1 Introduction

Recent advances in deep learning, which is an instance of artificial intelligence (AI) based on neural networks [1, 2], have led to numerous applications in the molecular sciences, *e.g.*, in drug discovery [3, 4], quantum chemistry [5], and structural biology [6, 7]. Two characteristics of deep learning render it particularly promising when applied to molecules. First, deep learning methods can cope with "unstructured" data representations, such as text sequences [8, 9], speech signals [10, 11], images [12–14], and graphs [15, 16]. This ability is particularly useful for molecular systems, for which chemists have developed many models (*i.e.*, "molecular representations") that capture molecular properties at varying levels of abstraction (Figure 1). The second key characteristic is that deep learning can perform feature extraction (or feature learning) from the input data, that is, produce data-driven features from the input data without the need for manual intervention. These two characteristics are promising for deep learning as a complement to "classical" machine learning applications (*e.g.*, Quantitative Structure-Activity Relationship [QSAR]), in which molecular features (*i.e.*, "molecular descriptors" [17]) are encoded *a priori* with rule-based algorithms. The capability to learn from unstructured data and obtain data-driven molecular features has led to unprecedented applications of AI in the molecular sciences.

One of the most promising advances in deep learning is geometric deep learning (GDL). *Geometric deep learning* is an umbrella term encompassing emerging techniques which generalize neural networks to Euclidean and non-Euclidean domains, such as graphs, manifolds, meshes, or string representations [15]. In general, GDL encompasses approaches that incorporate a geometric prior, *i.e.*, information on the structure space and symmetry properties of the input variables. Such a geometric prior is leveraged to improve the quality of the information captured by the model. Although GDL has been increasingly applied to molecular modeling [5, 18, 19], its full potential in the field is still untapped.

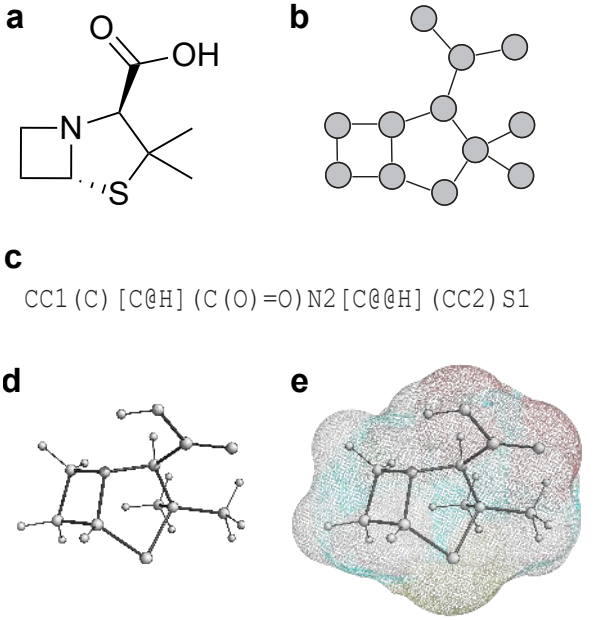


Figure 1: *Exemplary molecular representations for a selected molecule (i.e., the penam substructure of penicillin)*
a. Two-dimensional (2D) depiction (Kekulé structure).
b. Molecular graph (2D), composed of vertices (atoms) and edges (bonds).
c. SMILES string [20], in which atom type, bond type and connectivity are specified by alphanumerical characters.
d. Three-dimensional (3D) graph, composed of vertices (atoms), their position (x, y, z coordinates) in 3D space, and edges (bonds).
e. Molecular surface represented as a mesh colored according to the respective atom types.

The aim of this review is to (i) provide a structured and harmonized overview of the applications of GDL on molecular systems, (ii) delineate the main research directions in the field, and (iii) provide a forecast of the future impact of GDL. Three fields of application are highlighted, namely drug discovery, quantum chemistry, and computer-aided synthesis planning (CASP),

with particular attention to the data-driven molecular features learned by GDL methods. A glossary of selected terms can be found in Box 1.

2 Principles of geometric deep learning

The term *geometric deep learning* was coined in 2017 [15]. Although GDL was originally used for methods applied to non-Euclidean data [15], it now extends to all deep learning methods that incorporate geometric priors [21], that is, information about the structure and symmetry of the system of interest. Symmetry is a crucial concept in GDL, as it encompasses the properties of the system with respect to manipulations (transformations), such as translation, reflection, rotation, scaling, or permutation (Box 2).

Symmetry is often recast in terms of *invariance* and *equivariance* to express the behavior of any mathematical function with respect to a transformation $\mathcal{T}$ (*e.g.* rotation, translation, reflection or permutation) of an acting symmetry group [22]. Here, the mathematical function is a neural network $\mathcal{F}$ applied to a given molecular input $\mathcal{X}$. $\mathcal{F}(\mathcal{X})$ can therein transform equivariantly, invariantly or neither with respect to $\mathcal{T}$, as described below:

- *Equivariance.* A neural network $\mathcal{F}$ applied to an input $\mathcal{X}$ is *equivariant* to a transformation $\mathcal{T}$ if the transformation of the input $\mathcal{X}$ commutes with the transformation of $\mathcal{F}(\mathcal{X})$, *via* a transformation $\mathcal{T}'$ of the same symmetry group, such that: $\mathcal{F}(\mathcal{T}(\mathcal{X})) = \mathcal{T}'\mathcal{F}(\mathcal{X})$. Neural networks are therefore equivariant to the actions of a symmetry group on their inputs if and only if each layer of the network "equivalently" transforms under any transformation of that group.

- *Invariance.* Invariance is a special case of equivariance, where $\mathcal{F}(\mathcal{X})$ is invariant to $\mathcal{T}$ if $\mathcal{T}'$ is the trivial group action (*i.e.*, identity): $\mathcal{F}(\mathcal{T}(\mathcal{X})) = \mathcal{T}'\mathcal{F}(\mathcal{X}) = \mathcal{F}(\mathcal{X})$.

- $\mathcal{F}(\mathcal{X})$ is neither equivariant nor invariant to $\mathcal{T}$ when the transformation of the input $\mathcal{X}$ does not commute with the transformation of $\mathcal{F}(\mathcal{X})$: $\mathcal{F}(\mathcal{T}(\mathcal{X})) \neq \mathcal{T}'\mathcal{F}(\mathcal{X})$.

The symmetry properties of a neural network architecture vary depending on the network type and the symmetry group of interest and are individually discussed in the following sections. Readers can find an in-depth treatment of equivariance and group equivariant layers in neural networks elsewhere [23–26].

The concept of equivariance and invariance can also be used in reference to the molecular features obtained from a given molecular representation ($\mathcal{X}$), depending on their behaviour when a transformation is applied to $\mathcal{X}$. For instance, many molecular descriptors are invariant to the rotation and translation of the molecular representation by design [17], *e.g.,* the Moriguchi octanol-water partitioning coefficient [27], which relies only on the occurrence of specific molecular substructures for calculation. The symmetry properties of molecular features extracted by a neural network depend on both the symmetry properties of the input molecular representation and of the utilized neural network.

Many relevant molecular properties (*e.g.*, equilibrium energies, atomic charges, or physicochemical properties such as permeability, lipophilicity or solubility) are invariant to certain symmetry operations (Box 2). In many tasks in chemistry, it is thus desirable to design neural networks that transform equivariantly under the actions of pre-defined symmetry groups. Exceptions occur if the targeted property changes upon a symmetry transformation of the molecules (*e.g.*, chiral properties which change under inversion of the molecule, or vector properties which change under rotation of the molecule). In such cases, the inductive bias (learning bias) of equivariant neural networks would not allow for the differentiation of symmetry-transformed molecules.

While neural networks can be considered as universal function approximators [28], incorporating prior knowledge such as reasonable geometric information (geometric priors) has evolved as a core design principle of neural network modeling [21]. By incorporating geometric priors, GDL allows to increase the quality of the model and bypasses several bottlenecks related to the need to force the data into Euclidean geometries (*e.g.*, by feature engineering). Moreover, GDL provides novel modeling opportunities, such as data augmentation in low data regimes [29, 30].

3 Molecular GDL

The application of GDL to molecular systems is challenging, in part because there are multiple valid ways of representing the same molecular entity. Molecular representations[1] can be categorized based on their different levels of abstraction and the physicochemical and geometrical aspects they capture. Importantly, all of these representations are models of the same reality and are thus *"suitable for some purposes, not for others"* [60]. GDL provides the opportunity to experiment with different representations of the same molecule and leverages their intrinsic geometrical features to increase the quality of the model. Moreover, GDL has repeatedly proven useful in providing insights into relevant molecular properties for the task at hand, thanks to its feature extraction (feature learning) capabilities. In the following sections, we delineate the most prevalent molecular GDL approaches and their applications in chemistry, grouped according to the respective molecular representations used for deep learning: molecular graphs, grids, strings, and surfaces.

[1]Note that in this review the term "representation" is used solely to denote human-made models of molecules (*e.g.*, molecular graphs, 3D conformers, SMILES strings). To avoid confusion with other usages of the word "representation" in deep learning, we will use the term "feature" whenever referring to any numerical description of molecules, obtained either with rule-based algorithms (molecular descriptors) or learned (extracted) by neural networks.

Box 1: Glossary of selected terms

CoMFA and CoMSIA. Comparative Molecular Field Analysis (CoMFA) [31] and Comparative Molecular Similarity Indices Analysis (CoMSIA) [32] are popular 3D QSAR methods developed in the 1980s and 1990s, in which three-dimensional grids are used to capture the distributions of molecular features (*e.g.*, steric, hydrophobic, and electrostatic properties). The obtained molecular descriptors serve as inputs to a regression model for quantitative bioactivity prediction.
Convolution. Operation within a neural network that transforms a feature space into a new feature space and thereby captures the local information found in the data. Convolutions were first introduced for pixels in images [33, 34] but the term "convolution" is now used for neural network architectures covering a variety of data structures such as graphs, point clouds, spheres, grids, or manifolds.
Density Functional Theory (DFT). A quantum mechanical modeling approach used to investigate the electronic structure of molecules.
Data augmentation. Artificial increase of the data volume available for model training, often achieved by leveraging symmetrical properties of the input data which are not captured by the model (*e.g.*, rotation or permutation).
Feature. An individually measurable or computationally obtainable characteristic of a given sample (*e.g.*, molecule), in the form of a scalar. In this review, the term refers to a numeric value characterizing a molecule. Such molecular features can be computed with rule-based algorithms ("molecular descriptors") or generated automatically by deep learning from a molecular representation ("hidden" or "learned" features).
Geometric prior. An inductive bias incorporating information on the symmetric nature of the system of interest into the neural network architecture. Also known as *symmetry prior*.
Inductive bias. Set of assumptions that a learning algorithm (*e.g.*, a neural network) uses to learn the target function and to make predictions on previously unseen data points.
One-hot encoding. Method for representing categorical variables as numerical arrays by obtaining a binary variable (0, 1) for each category. It is often used to convert sequences (*e.g.*, SMILES strings) into numerical matrices, suitable as inputs and/or outputs of deep learning models (*e.g.*, chemical language models).
Quantitative Structure-Activity Relationship (QSAR). Machine learning techniques aimed at finding an empirical relationship between the molecular structure (usually encoded as molecular descriptors) and experimentally determined molecular properties, such as pharmacological activity or toxicity.
Reinforcement learning. A technique used to steer the output of a machine learning algorithm toward user-defined regions of optimality *via* a predefined reward function [35].
Transfer learning. Transfer of knowledge from an existing deep learning model to a related task for which fewer training samples are available [36].
Unstructured data. Data that are not arranged as vectors of (typically handcrafted) features. Examples of unstructured data include graphs, images, and meshes. Molecular representations are typically unstructured, whereas numerical molecular descriptors (*e.g.*, molecular properties, molecular "fingerprints") are examples of structured data.
Voxel. Element of a regularly spaced, 3D grid (equivalent to a pixel in 2D space).

3.1 Learning on molecular graphs

3.1.1 Molecular graphs

Graphs are among the most intuitive ways to represent molecular structures [62]. Any molecule can be thought of as a mathematical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, whose vertices ($\mathbf{v}_i \in \mathcal{V}$) represent atoms, and whose edges ($\mathbf{e}_{i,j} \in \mathcal{E}$) constitute their connection (Figure 3.1). In many deep learning applications, molecular graphs can be further characterized by a set of vertex and edge features.

3.1.2 Graph neural networks

Deep learning methods devoted to handling graphs as input are commonly referred to as graph neural networks (GNNs). When applied to molecules, GNNs allow for feature extraction by progressively aggregating information from atoms and their molecular environments (Figure 2a, [63, 64]). Different architectures of GNNs have been introduced [65], the most popular of which fall under the umbrella term of *message passing neural networks* [5, 66, 67]. Such networks iteratively update the vertex features of the l-th network layer ($\mathbf{v}_i^l \rightarrow \mathbf{v}_i^{l+1}$) *via* graph convolutional operations, employing at least two learnable functions ψ and ϕ, and a local permutation-invariant aggregation operator (*e.g.,* sum): $\mathbf{v}_i^{l+1} = \phi\left(\mathbf{v}_i^l, \bigoplus_{j \in \mathcal{N}(i)} \psi\left(\mathbf{v}_i^l, \mathbf{v}_j^l\right)\right)$.

Since their introduction as a means to predict quantum chemical properties of small molecules at the density functional theory (DFT) level [5], GNNs have found many applications in quantum chemistry [68–72], drug discovery [37, 73, 74], CASP [75], and molecular property prediction [76, 77]. When applied to quantum chemistry tasks, GNNs often use E(3)-invariant 3D information by including radial and angular information

Table 1: *Summary of selected geometric deep learning (GDL) approaches for molecular modeling.* For each approach, the utilized molecular representation(s) and selected applications are reported. 1D, one-dimensional; 2D, two-dimensional; 3D, three-dimensional.

GDL approach	**Molecular representation(s)**	**Applications**
Graph neural networks (GNNs)	2D and 3D molecular graph, and 3D point cloud.	Molecular property prediction in drug discovery [37, 38] and in quantum chemistry for energies [39–41], forces [41–43] and wave-functions [44], CASP [45, 46], and generative molecular design [47, 48].
3D convolutional neural networks (3D CNNs)	3D grid.	Structure-based drug design and property prediction [49, 50].
Mesh convolutional neural networks (geodesic CNNs or 3D GNNs)	Surface (mesh) encoded as a 2D grid or 3D graph.	Protein-protein interaction prediction and ligand-pocket fingerprinting [18].
Recurrent neural networks (RNNs)	String notation (1D grid).	Generative molecular design [19, 51], synthesis planning [52], protein structure prediction [53] and prediction of properties in drug discovery [54, 55].
Transformers	String notation encoded as a graph.	Synthesis planning [56], prediction of reaction yields [57], generative molecular design [58], prediction of properties in drug discovery [59], and protein structure prediction [6, 7].

into the edge features of the graph [43, 68, 69, 72, 78], thereby improving the prediction accuracy of quantum chemical forces and energies for equilibrium and non-equilibrium molecular conformations, as in the case of SchNet [79, 80] and PaiNN [43]. SchNet-like architectures were used to predict quantum mechanical wave-functions in the form of Hartree-Fock and DFT density matrices [81], and differences in quantum properties obtained by DFT and coupled cluster level-of-theory calculations [82].

GNNs for molecular property prediction have been shown to outperform human-engineered molecular descriptors for several biologically relevant properties [83]. Although including 3D information into molecular graphs generally improved the prediction of drug-relevant properties, no marked difference was observed between using a single or multiple molecular conformers for network training [84]. Because of their natural connection with molecular representations, GNNs seem particularly suitable in the context of explainable AI (XAI) [85], where they have been used to interpret models predicting molecular properties of preclinical relevance [38] and quantum chemical properties [86].

GNNs have been used for *de novo* molecule generation [47, 87–89], for example by performing vertex and edge addition from an initial vertex [87] (Figure 2b). GNNs have also been combined with variational autoencoders [48, 88–90] and reinforcement learning [47, 91, 92]. Finally, GNNs have been applied to CASP [45, 75, 93]; however, the current approaches are limited to reactions in which one bond is removed between the products and the reactants.

3.1.3 Equivariant message passing

A recent area of development of graph-based methods are SE(3)- and E(3)-equivariant GNNs (*equivariant message passing networks*) which deal with the absolute coordinate systems of 3D graphs [94, 95] (Figure 2b). Thus, these networks may be particularly well-suited to be applied to 3D molecular representations. Such networks exploit Euclidean symmetries of the system (Box 2).

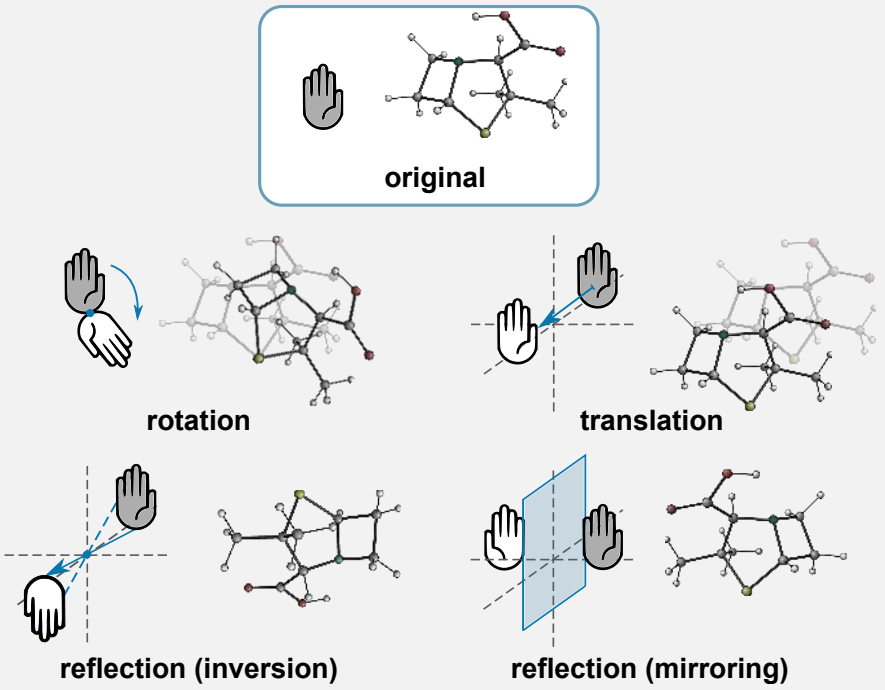
3D molecular graphs $\mathcal{G}_{3D} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, in addition to their vertex and edge features ($\mathbf{v}_i \in \mathcal{V}$ and $\mathbf{e}_{ij} \in \mathcal{E}$, respectively), also encode information on the vertex position in a 3D coordinate system ($\mathbf{r}_i \in \mathcal{R}$). By employing E(3)- [41] and SE(3)-equivariant [94] convolutions, such networks have shown high accuracy for predicting several quantum chemical properties such as energies [39, 40, 42, 43, 96–98], interatomic potentials for molecular dynamics simulations [41, 42, 99], and wave-functions [44]. SE(3) equivariant neural networks do not commute with reflections of the input (*i.e.* non-equivariant to reflections), and thereby enable SE(3) equivariant models to distinguish between stereoisomers of chiral molecules including enantiomers [94]. E(3) equivariant neural networks on the other side transform equivariantly with refelctions, which allows E(3) equivariant models only to distinguish between diastereomers and not eneantiomers. SE(3) neural networks are computationally expensive due to their use of *spherical harmonics* [100] and *Wigner D-functions* [101] to compute learnable weight kernels. E(3)-equivariant neural networks are computationally more efficient and have shown to perform equal to, or better than, SE(3)-equivariant networks, *e.g.*, for the modeling of quantum

Box 2: Euclidean symmetries in molecular systems

Molecular systems (and three-dimensional representations thereof) can be considered as objects in Euclidean space. In such a space, one can apply several symmetry operations (transformations) that are (i) performed with respect to three symmetry elements (*i.e.*, line, plane, point), and (ii) rigid, that is, they preserve the Euclidean distance between all pairs of atoms (*i.e.*, isometry). The Euclidean transformations are as follows:

- *Rotation.* Movement of an object with respect to the radial orientation to a given point.
- *Translation.* Movement of every point of an object by the same distance in a given direction.
- *Reflection.* Mapping of an object to itself through a point (inversion), a line or a plane (mirroring).

All three transformations and their arbitrary finite combinations are included in the *Euclidean group* [E(3)]. The *special Euclidean group* [SE(3)] comprises only translations and rotations.
Molecules are always symmetric in the SE(3) group, *i.e.*, their intrinsic properties (e.g., biological and physicochemical properties, and equilibrium energy) are invariant to coordinate rotation and translation, and combinations thereof. Several molecules are chiral, that is, some of their (chiral) properties depend on the absolute configuration of their stereogenic centers, and are thus non-invariant to molecule reflection. Chirality plays a key role in chemical biology; relevant examples of chiral molecules are DNA, and several drugs whose enantiomers exhibit markedly different pharmacological and toxicological properties [61].



chemical properties and dynamic systems [41]. Equivariant message passing networks have been applied to predict the quantum mechanical wave-function of nuclei and electron-based representations in an end-to-end fashion [102–104]. However, such networks are currently limited to small molecular systems because of the large size of the learned matrices, which scale quadratically with the number of electrons in the system.

3.2 Learning on grids

Grids capture the properties of a system at regularly spaced intervals. Based on the number of dimensions included in the system, grids can be 1D (*e.g.*, sequences), 2D (*e.g.*, RGB images), 3D (*e.g.*, cubic lattices), or higher-dimensional. Grids are defined by a Euclidean geometry and can be considered as a graph with a special adjacency, where (i) the vertices have a fixed ordering that is defined by the spatial dimensions of the grid, and (ii) each vertex has an identical number of adjacent edges and is therefore indistinguishable from all other vertices structure-wise [21]. These two properties render local convolutions applied to a grid inherently permutation invariant, and provide a strong geometric prior for translation invariance (*e.g.* by weight sharing in convolutions). These grid properties have critically determined the success of convolutional neural networks (CNNs), *e.g.*, in computer vision [33, 34], natural language processing [9, 105], and speech recognition [10, 11].

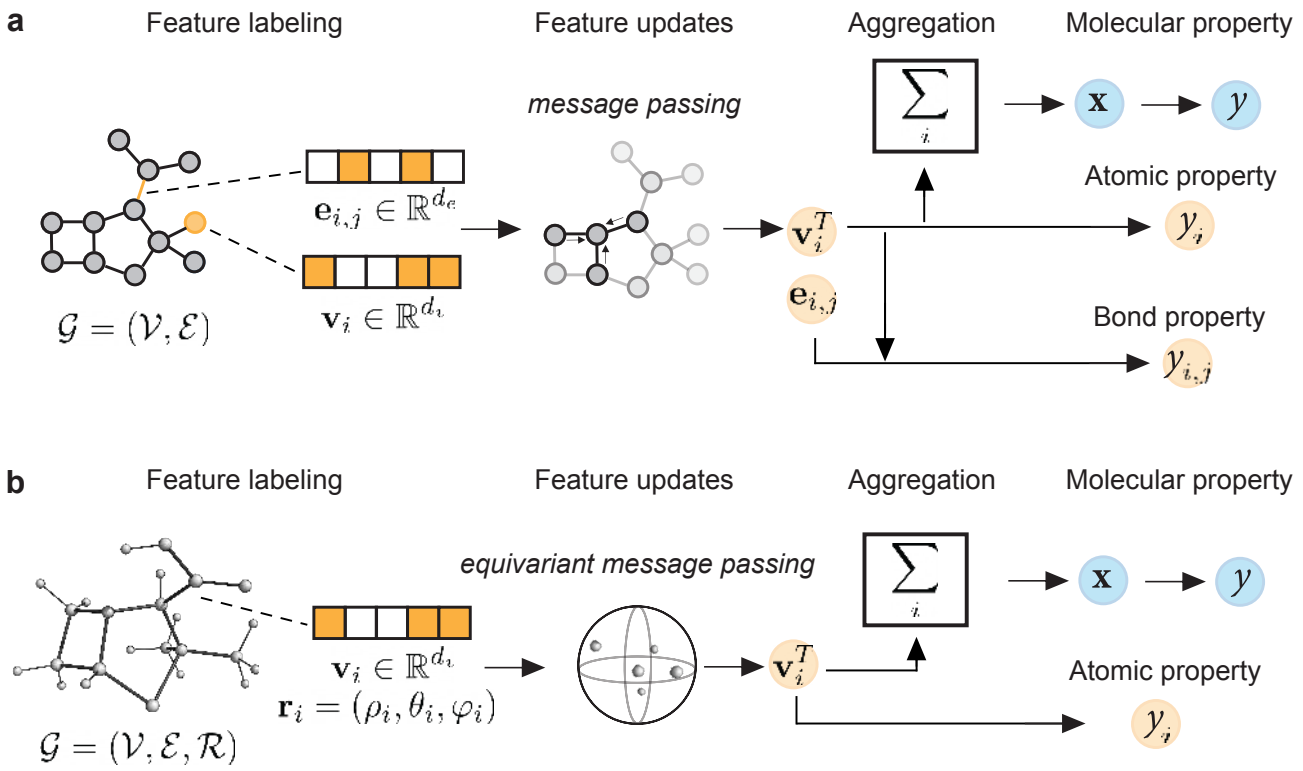


Figure 2: *Deep learning on molecular graphs.*
a. Message passing graph neural networks applied to two-dimensional (2D) molecular graphs: 2D molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with its labeled vertex (atom) features ($\mathbf{v}_i \in \mathbb{R}^{d_v}$), and edge (bond) features ($\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$). Vertex features are updated by iterative message passing for a defined number of time steps T across each pair of vertices v_i and v_j, connected *via* an edge $e_{j,i}$. After the last message passing convolution, the final vertex $\mathbf{v}_i^t$ can be (i) mapped to a bond (y_{ij}) or atom (y_i) property, or (ii) aggregated to form molecular features (that can be mapped to a molecular property y).
b. E(3)-equivariant message passing graph neural networks applied to three-dimensional (3D) molecular graphs: 3D graphs $\mathcal{G}_3 = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ that are labeled with atom features ($\mathbf{v}_i \in \mathbb{R}^{d_v}$), their absolute coordinates in 3D space ($\mathbf{r}_i \in \mathbb{R}^3$) and their edge features ($\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$). Iterative spherical convolutions are used to obtain data-driven atomic features ($\mathbf{v}_i^t$), which can be mapped to atomic properties or aggregated, and mapped to molecular properties (y_i and y, respectively).

3.2.1 Molecular grids

Molecules can be represented as grids in different ways. 2D grids (*e.g.*, molecular structure drawings) are generally more useful for visualization rather than prediction, with few exceptions [106]. Analogous with some popular pre-deep learning approaches, for example Comparative molecular field analysis (CoMFA) [31], and comparative molecular similarity indices analysis (CoMSIA) [32], 3D grids are often used to capture the spatial distribution of the properties within one (or more) molecular conformer. Such representations are then used as inputs to the 3D CNNs. 3D CNNs are characterized by a greater resource efficiency than equivariant GNNs, which until now have mainly been applied to molecules with fewer than approximately 1000 atoms. Thus, 3D CNNs have often been the method of choice when the protein structure has to be considered, *e.g.*, for protein-ligand binding affinity prediction [49, 50, 107–109], or active site recognition [110].

3.3 Learning on molecular surfaces

Molecular surfaces can be defined by the surface enclosing the 3D structure of a molecule at a certain distance from each atom center. Each point on such a continuous surface can be further characterized by its chemical (*e.g.*, hydrophobic, electrostatic) and geometric features (*e.g.*, local shape, curvature). From a geometrical perspective, molecular surfaces are considered as 3D meshes, *i.e.*, a set of polygons (faces) that describe how the mesh coordinates exist in the 3D space [111]. Their vertices can be represented by a 2D grid structure (where four vertices on the mesh define a pixel) or by a 3D graph structure. The grid- and graph-based structures of meshes enable applications of 2D CNNs, geodesic CNNs and GNNs to learn on mesh-based molecular surfaces. Recently, geodesic (2D) CNNs have been applied to learn on mesh-based representations of protein surfaces to predict protein-protein interactions and recognize corresponding binding sites [18]. This approach generated data-driven fingerprints that are relevant for specific biomolecular interactions. Approaches like 2D CNNs applied to meshes come with

certain limitations, such as the need for rotational data augmentation (due to their non-equivariance to rotations) and for enforcing a homogeneous mesh resolution (*i.e.*, uniform spacing of all the points in the mesh). Recently introduced GNNs for mesh-based representations have been shown to incorporate rotational equivariance into their network architecture and allow for heterogeneous mesh resolution [112]. Such GNNs are computationally efficient and have potential for modeling macromolecular structures; however, they have not yet found applications to molecular systems. Other studies have used 3D voxel-based surface representations of (macro)molecules as inputs to 3D CNNs, *e.g.,* for protein-ligand affinity [113] and protein binding-site [114] prediction.

3.4 Learning on string representations

3.4.1 Molecular strings

Molecules can be represented as molecular strings, *i.e.*, linear sequences of alphanumeric symbols. Molecular strings were originally developed as manual ciphering tools to complement systematic chemical nomenclature [115, 116] and later became suitable for data storage and retrieval. Some of the most popular string-based representations are the Wiswesser Line Notation [117], the Sybyl line notation [118], the International Chemical Identifier (InChI) [119], Hierarchical Editing Language for Macromolecules [120], and the Simplified Molecular Input Line Entry System (SMILES) [20].

Each type of linear representation can be considered as a "chemical language." In fact, such notations possess a defined syntax, *i.e.*, not all possible combinations of alphanumerical characters will lead to a "chemically valid" molecule. Furthermore, these notations possess semantic properties: depending on how the elements of the string are combined, the corresponding molecule will have different physicochemical and biological properties. These characteristics make it possible to extend the deep learning methods developed for language and sequence modeling to the analysis of molecular strings for "chemical language modeling" [121, 122].

SMILES strings – in which letters are used to represent atoms, and symbols and numbers are used to encode bond types, connectivity, branching, and stereochemistry (Figure 3a) – have become the most frequently employed data representation method for sequence-based deep learning [19, 52]. Whereas several other string representations have been tested in combination with deep learning, *e.g.,* InChI [123], DeepSMILES [124], and self-referencing embedded strings (SELFIES) [125], SMILES remains the *de facto* representation of choice for chemical language modeling [30]. The following text introduces the most prominent chemical language modeling methods, along with selected examples of their application to chemistry.

3.4.2 Chemical language models

Chemical language models are machine learning methods that can handle molecular sequences as inputs and/or outputs. The most common algorithms for chemical language modeling are *Recurrent neural networks* (RNNs) and *Transformers*:

- *RNNs* (Figure 3b) [126] are neural networks that process sequence data as *Euclidean* structures, usually *via* one-hot-encoding. RNNs model a dynamic system in which the hidden state (h_t) of the network at any t-th time point (*i.e.*, at any t-th position in the sequence) depends on both the current observation (s_t) and the previous hidden state (h_{t-1}). RNNs can process sequence inputs of arbitrary lengths and provide outputs of arbitrary lengths. RNNs are often used in an "autoregressive" fashion, *i.e.*, to predict the probability distribution over the next possible elements (tokens) at the time step $t+1$, given the current hidden state (h_t) and the preceding portions of the sequence. Several RNN architectures have been proposed to solve the gradient vanishing or exploding problems of "vanilla" RNNs [127, 128], such as long short-term memory [105] and gated recurrent units [129].

- *Transformers* (Figure 3c) process sequence data as *non-Euclidean* structures, by encoding sequences as either (i) a fully connected graph, or (ii) a sequentially connected graph, where each token is only connected to the previous tokens in the sequence. The former approach is often used for feature extraction in general (*e.g.,* in a Transformer-encoder), whereas the latter is employed for next-token prediction *e.g.* in a Transformer-decoder). The positional information of tokens is usually encoded by positional embedding or sinusoidal positional encoding [8]. Transformers combine graph-like processing with the so-called attention layers. Attention layers allow Transformers to focus on ("pay attention to") the perceived relevant tokens for each prediction. Transformers have been particularly successful in sequence-to-sequence tasks, such as language translation.

Extending early studies [19, 132, 133], RNNs for next-token prediction have been routinely applied to the *de novo* generation of molecules with desired biological or physicochemical properties, in combination with transfer [19, 134–136] or reinforcement learning [137, 138]. In this context, RNNs have shown remarkable capability to learn the SMILES syntax [19, 134], and capture high-level molecular features ("semantics"), such as physicochemical [19, 134] and biological properties [132, 135, 136, 139]. In this context, data augmentation based on SMILES randomization [133, 140] or bidirectional learning [141] have proven to be efficient for improving the quality of the chemical language learned by RNNs. Most published studies have used SMILES strings or derivative representations. In a few studies, one-letter amino acid sequences were employed for peptide design [51, 142–145]. RNNs have also been applied to predict ligand–protein interactions and the pharmacokinetic properties of drugs [54, 55],

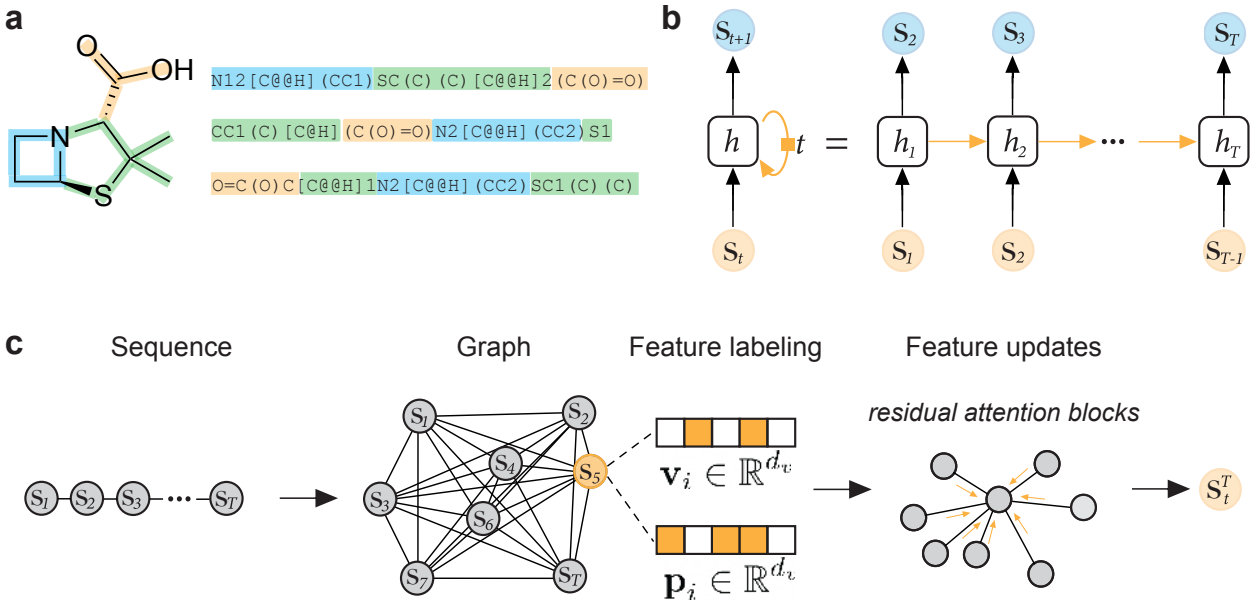


Figure 3: *Chemical language modeling.*
a. SMILES strings, in which atom types are represented by their element symbols, and bond types and branching are indicated by other predefined alphanumeric symbols. For each molecule, *via* the SMILES algorithm a string of T symbols ("tokens") is obtained ($\mathbf{s} = \{s_1, s_2, \ldots, s_T\}$), which encodes the molecular connectivity, herein illustrated *via* the color that indicates the corresponding atomic position in the graph (*left*) and string (*right*). A molecule can be encoded *via* different SMILES strings depending on the chosen starting atom. Three random permutations incorporating identical molecular information are presented.
b. Recurrent neural networks, at any sequence position t, learn to predict the next token s_{t+1} of a sequence $\mathbf{s}$ given the current sequence ($\{s_1, s_2, \ldots, s_t\}$) and hidden state h_t.
c. Transformer-based language models, in which the input sequence is structured as a graph. Vertices are featurized according to their token identity (*e.g.,* *via* token embedding, $\mathbf{v}_i \in \mathbb{R}^{d_v}$) and their position in the sequence (*e.g., via* sinusoidal positional encoding, $\mathbf{p}_i \in \mathbb{R}^{d_v}$). During transformer learning, the vertices are updated *via* residual attention blocks. After passing T attention layers, an individual feature representation $\mathbf{s}_t^T$ for each token is obtained.

protein secondary structure [53, 146], and the temporal evolution of molecular trajectories [147]. RNNs have been applied for molecular feature extraction [148, 149], showing that the learned features outperformed both traditional molecular descriptors and graph-convolution methods for virtual screening and property prediction [148]. The Fréchet ChemNet distance [150], which is based on the physicochemical and biological features learned by an RNN model, has become the *de facto* reference method to capture molecular similarity in this context.

Molecular Transformers have been applied to CASP, which can be cast as a sequence-to-sequence translation task, in which the string representations of the reactants are mapped to those of the corresponding product, or *vice versa*. Since their initial applications [56], Transformers have been employed to predict multi-step syntheses [151], regio- and stereoselective reactions [152], enzymatic reaction outcomes [153], and reaction yields and classes [57, 154]. Recently, Transformers have been applied to molecular property prediction [59, 155] and optimization [156]. Transformers have also been used for *de novo* molecule design by learning to translate the target protein sequence into SMILES strings of the corresponding ligands [58]. Representations learned from SMILES strings by Transformers have shown promise for property prediction in low-data regimes [157]. Furthermore, Transformers have recently been combined with E(3) and SE(3) equivariant layers to learn the 3D structures of proteins from their amino-acid sequence [6, 7]. These equivariant Transformers achieve state-of-the-art performance in protein structure prediction.

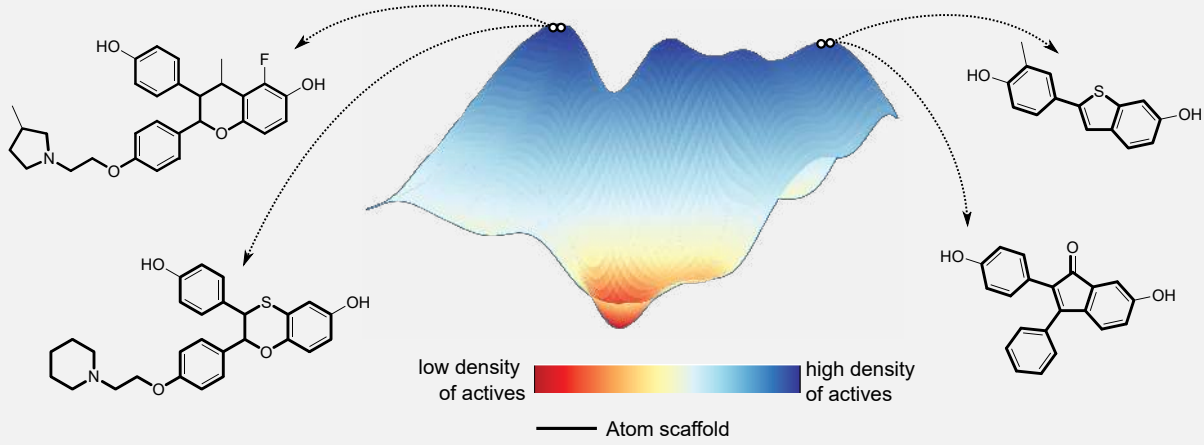
Other deep learning approaches have relied on string-based representations for *de novo* design, *e.g.*, conditional generative adversarial networks [158–160] and variational autoencoders [161, 162]. Most of these models, however, have limited or equivalent ability to automatically learn SMILES syntax, as compared to RNNs. 1D CNNs [163, 164] and self-attention networks [165–167] have been used with SMILES for property prediction. Recently, deep learning on amino acid sequences for property prediction was shown to perform *on par* with approaches based on human-engineered features [168].

4 Conclusions and outlook

Geometric deep learning in chemistry has allowed researchers to leverage the symmetries of different unstructured molecular representations, resulting in a greater flexibility and versatility of the available computational models for molecular structure generation

Box 3: Structure-activity landscape modeling with geometric deep learning

This worked example shows how geometric deep learning (GDL) can be used to interpret the structure-activity landscape learned by a trained model. Starting from a publicly available molecular dataset containing estrogen receptor binding information [130], we trained an E(3)-equivariant graph neural network (six hidden layers, 128 hidden neurons per layer) and analyzed the learned features and their relationship to ligand binding to the estrogen receptor. The figure shows an analysis of the learned molecular features (third hidden layer, analyzed *via* principal component analysis; the first two principal components are shown), and how these features relate to the density of active and inactive molecules in the chemical space. The network successfully separated the molecules based on both their experimental bioactivity and their structural features (*e.g.*, atom scaffolds [131]) and might offer novel opportunities for explainable AI with GDL.



and property prediction. Such approaches represent a valid alternative to classical chemoinformatics approaches that are based on molecular descriptors or other human-engineered features. For modeling tasks that are usually characterized by the need for highly engineered rules (*e.g.*, chemical transformations for *de novo* design, and reactive site specification for CASP), the benefits of GDL have been consistently shown. In published applications of GDL, each molecular representation has shown characteristic strengths and weaknesses.

Molecular strings, like SMILES, have proven particularly suited for generative deep learning tasks, such as *de novo* design and CASP. This success may be due to the relatively easy syntax of such a chemical language, which facilitates next-token and sequence-to-sequence prediction. For molecular property prediction, SMILES strings could be limited due to their non-univocity.

Molecular graphs have shown particular usefulness for property prediction, partly because of their human interpretability and ease of inclusion of desired edge and node features. The incorporation of 3D information (*e.g.*, with equivariant message passing) is useful for quantum chemistry related modeling, whereas in drug discovery applications, this approach has often failed to clearly outbalance the increased complexity of the model. E(3)-equivariant graph neural networks have also been applied for conformation-aware *de novo* design [169], but prospective experimental validation studies have not yet been published.

Molecular grids have become the *de facto* standard for 3D representations of large molecular systems, due to (i) their ability to capture information at a user-defined resolution (voxel density) and (ii) the Euclidean structure of the input grid.

Finally, *molecular surfaces* are currently at the forefront of GDL. We expect many interesting applications of GDL on molecular surfaces in the near future.

To further the application and impact of GDL in chemistry, an evaluation of the optimal trade-off between algorithmic complexity, performance, and model interpretability will be required. These aspects are crucial for reconciling the "two QSARs" [170] and connect computer science and chemistry communities. We encourage GDL practitioners to include aspects of interpretability in their models (*e.g.*, *via* XAI [85]) whenever possible and transparently communicate with domain experts. The feedback from domain experts will also be crucial to develop new "chemistry-aware" architectures, and further the potential of molecular GDL for concrete prospective applications.

The potential of GDL for molecular feature extraction has not yet been fully explored. Several studies have shown the benefits of learned representations compared to classical molecular descriptors, but in other cases, GDL failed to live up to its promise in terms of superior learned features. Although there are several benchmarks for evaluating machine learning models for property prediction [171, 172] and molecule generation [173, 174], at present, there is no such framework to enable the systematic evaluation of the usefulness of data-driven features learned by AI. Such benchmarks and

systematic studies are key to obtaining an unvarnished assessment of deep representation learning. Moreover, investigating the relationships between the learned features and the physicochemical and biological properties of the input molecules will augment the interpretability and applicability of GDL, *e.g.*, to modeling structure-function relationships like structure-activity landscapes (Box 3).

Compared to conventional QSAR approaches, in which the assessment of the applicability domain (*i.e.*, the region of the chemical space where model predictions are considered reliable) has been routinely performed, contemporary GDL studies lack such an assessment. This systematic gap might constitute one of the limiting factors to the more widespread use of GDL approaches for prospective studies, as it could lead to unreliable predictions, *e.g.*, for molecules with different mechanisms of action, functional groups, or physicochemical properties than the training data. In the future, it will be necessary to devise "geometry-aware" approaches for applicability domain assessment.

Another opportunity will be to leverage less explored molecular representations for GDL. For instance, the electronic structure of molecules has vast potential for tasks such as CASP, molecular property prediction, and prediction of macromolecular interactions (*e.g.* protein-protein interactions). Although accurate statistical and quantum mechanical simulations are computationally expensive, modern quantum machine learning models [175, 176] trained on large quantum data collections [177–179] allow quantum information to be accessed much faster with high accuracy. This aspect could enable quantum and electronic featurization of extensive molecular datasets, to be used as input molecular representations for the task of interest.

Deep learning can be applied to a multitude of biological and chemical representations. The corresponding deep neural network models have the potential to augment human creativity, paving the way for new scientific studies that were previously unfeasible. However, research has only explored the tip of the iceberg. One of the most significant catalysts for the integration of deep learning in molecular sciences may be the responsibility of academic institutions to foster interdisciplinary collaboration, communication, and education. Picking the "high hanging fruits" will only be possible with a deep understanding of both chemistry and computer science, along with out-of-the-box thinking and collaborative creativity. In such a setting, we expect molecular GDL to increase the understanding of molecular systems and biological phenomena.

5 Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF, grant no. 205321_182176) and the ETH RETHINK initiative.

6 Competing interest

G.S. declares a potential financial conflict of interest as co-founder of inSili.com LLC, Zurich, and in his role as scientific consultant to the pharmaceutical industry.

7 List of abbreviations

AI: Artificial Intelligence
CASP: Computer-aided Synthesis Planning
CNN: Convolutional Neural Network
DFT: Density Functional Theory
E(3): Euclidean Symmetry Group
GDL: Geometric Deep Learning
GNN: Graph Neural Network
QSAR: Quantitative Structure-Activity Relationship
RNN: Recurrent Neural Network
SE(3): Special Euclidean Symmetry Group
SMILES: Simplified Molecular Input Line Entry Systems
XAI: Explainable Artificial Intelligence
1D: One-dimensional
2D: Two-dimensional
3D: Three-dimensional

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
2. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61,** 85–117 (2015).
3. Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Molecular Informatics* **35,** 3–14 (2016).
4. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opinion on Drug Discovery,* 1–11 (2021).
5. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. *Neural message passing for quantum chemistry* in *International Conference on Machine Learning* (2017), 1263–1272.
6. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature,* 1–11 (2021).
7. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* (2021).
8. Vaswani, A. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems,* 5998–6008 (2017).
9. Brown, T. B. *et al.* Language models are few-shot learners. *arXiv:2005.14165* (2020).
10. Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* **29,** 82–97 (2012).

11. Mikolov, T., Deoras, A., Povey, D., Burget, L. & Černocky, J. Strategies for training large scale neural network language models. *2011 IEEE Workshop on Automatic Speech Recognition & Understanding,* 196–201 (2011).
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90 (2017).
13. Farabet, C., Couprie, C., Najman, L. & LeCun, Y. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence* **35,** 1915–1929 (2012).
14. Tompson, J. J., Jain, A., LeCun, Y. & Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in Neural Information Processing Systems,* 1799–1807 (2014).
15. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* **34,** 18–42 (2017).
16. Monti, F., Frasca, F., Eynard, D., Mannion, D. & Bronstein, M. M. Fake news detection on social media using geometric deep learning. *arXiv:1902.06673* (2019).
17. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references* (John Wiley & Sons, 2009).
18. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17,** 184–192 (2020).
19. Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* **4,** 120–131 (2018).
20. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36 (1988).
21. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478* (2021).
22. Marsden, J. & Weinstein, A. Reduction of symplectic manifolds with symmetry. *Reports on mathematical physics* **5,** 121–130 (1974).
23. Cohen, T. S. & Welling, M. *Group equivariant convolutional networks* in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48* (2016), 2990–2999.
24. Cohen, T. S. & Welling, M. Steerable cnns. *arXiv:1612.08498* (2016).
25. Cohen, T. S., Geiger, M., Köhler, J. & Welling, M. *Spherical CNNs* in *International Conference on Learning Representations* (2018).
26. Kondor, R. & Trivedi, S. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *International Conference on Machine Learning,* 2747–2755 (2018).
27. Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I. & Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chemical and pharmaceutical bulletin* **40,** 127–130 (1992).
28. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* **2,** 303–314 (1989).
29. Tetko, I. V., Karpov, P., Van Deursen, R. & Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature communications* **11,** 1–11 (2020).
30. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence,* 1–12 (2021).
31. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **110,** 5959–5967 (1988).
32. Klebe, G. Comparative molecular similarity indices analysis: CoMSIA. *3D QSAR in drug design,* 87–104 (1998).
33. LeCun, Y., Bengio, Y., *et al.* Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361,** 1995 (1995).
34. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).
35. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
36. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22,** 1345–1359 (2009).
37. Feinberg, E. N. *et al.* PotentialNet for molecular property prediction. *ACS Central Science* **4,** 1520–1530 (2018).
38. Jiménez-Luna, J., Skalic, M., Weskamp, N. & Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *Journal of Chemical Information and Modeling* **61,** 1083–1094 (2021).
39. Miller, B. K., Geiger, M., Smidt, T. E. & Noé, F. Relevance of rotationally equivariant convolutions for predicting molecular properties. *arXiv:2008.08461* (2020).

40. Anderson, B., Hy, T. S. & Kondor, R. Cormorant: Covariant Molecular Neural Networks. *Advances in Neural Information Processing Systems* **32,** 14537–14546 (2019).

41. Satorras, V. G., Hoogeboom, E. & Welling, M. E (n) Equivariant Graph Neural Networks. *arXiv:2102.09844* (2021).

42. Fuchs, F., Worrall, D., Fischer, V. & Welling, M. SE (3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Advances in Neural Information Processing Systems* **33** (2020).

43. Schütt, K. T., Unke, O. T. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *arXiv:2102.03150* (2021).

44. Unke, O. T. *et al.* SE (3)-equivariant prediction of molecular wavefunctions and electronic densities. *arXiv:2106.02347* (2021).

45. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science* **10,** 370–377 (2019).

46. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in Neural Information Processing Systems,* 2607–2616 (2017).

47. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Scientific Reports* **9,** 1–10 (2019).

48. Jin, W., Barzilay, R. & Jaakkola, T. *Junction tree variational autoencoder for molecular graph generation* in *International Conference on Machine Learning* (2018), 2323–2332.

49. Jiménez, J., Skalic, M., Martinez-Rosell, G. & De Fabritiis, G. K deep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling* **58,** 287–296 (2018).

50. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling* **57,** 942–957 (2017).

51. Grisoni, F. *et al.* Designing anticancer peptides by constructive machine learning. *ChemMedChem* **13,** 1300–1302 (2018).

52. Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. Found in Translation: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* **9,** 6091–6098 (2018).

53. Senior, A. W. *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and Bioinformatics* **87,** 1141–1148 (2019).

54. Wang, X. *et al.* Optimizing Pharmacokinetic Property Prediction Based on Integrated Datasets and a Deep Learning Approach. *Journal of Chemical Information and Modeling* **60,** 4603–4613 (2020).

55. Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence* **2,** 134–140 (2020).

56. Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science* **5,** 1572–1583 (2019).

57. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2,** 015016 (2021).

58. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific Reports* **11,** 1–13 (2021).

59. Morris, P., St. Clair, R., Hahn, W. E. & Barenholtz, E. Predicting Binding from Screening Assays with Transformer Network Embeddings. *Journal of Chemical Information and Modeling* **60,** 4191–4199 (2020).

60. Hoffmann, R. & Laszlo, P. Representation in chemistry. *Angewandte Chemie International Edition in English* **30,** 1–16 (1991).

61. Nguyen, L. A., He, H. & Pham-Huy, C. Chiral drugs: an overview. *International Journal of Biomedical Science: IJBS* **2,** 85 (2006).

62. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907* (2016).

63. Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., *et al. Interaction Networks for Learning about Objects, Relations and Physics* in *Advances in Neural Information Processing Systems* (2016), 4502–4510.

64. Battaglia, P. W. *et al.* Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261* (2018).

65. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. *AI Open* **1,** 57–81 (2020).

66. Geerts, F., Mazowiecki, F. & Pérez, G. A. Let's Agree to Degree: Comparing Graph Convolutional Networks in the Message-Passing Framework. *arXiv:2004.02593* (2020).

67. Duvenaud, D. *et al. Convolutional networks on graphs for learning molecular fingerprints* in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2* (2015), 2224–2232.

68. Klicpera, J., Groß, J. & Günnemann, S. *Directional Message Passing for Molecular Graphs* in *International Conference on Learning Representations* (2019).

69. Zhang, S., Liu, Y. & Xie, L. Molecular Mechanics-Driven Graph Neural Network with Multiplex Graph for Molecular Structures. *arXiv:2011.07457* (2020).

70. Withnall, M., Lindelöf, E., Engkvist, O. & Chen, H. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *Journal of Cheminformatics* **12,** 1 (2020).

71. Tang, B. *et al.* A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics* **12,** 1–9 (2020).

72. Liu, Y. *et al.* Spherical message passing for 3d graph networks. *arXiv:2102.05013* (2021).

73. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180,** 688–702 (2020).

74. Torng, W. & Altman, R. B. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling* **59,** 4131–4149 (2019).

75. Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning Graph Models for Retrosynthesis Prediction. *arXiv:2006.07038* (2020).

76. Li, J., Cai, D. & He, X. Learning graph-level representation for drug discovery. *arXiv:1709.03741* (2017).

77. Liu, K. *et al.* Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. *International Journal of Molecular Sciences* **20,** 3389 (2019).

78. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation* **15,** 3678–3693 (2019).

79. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148,** 241722 (2018).

80. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Communications* **8,** 1–8 (2017).

81. Schütt, K., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications* **10,** 1–10 (2019).

82. Bogojeski, M., Vogt-Maranto, L., Tuckerman, M. E., Müller, K.-R. & Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications* **11,** 1–11 (2020).

83. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* **59,** 3370–3388 (2019).

84. Axelrod, S. & Gomez-Bombarelli, R. Molecular machine learning with conformer ensembles. *arXiv:2012.08452* (2020).

85. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2,** 573–584 (2020).

86. Schnake, T. *et al.* XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv:2006.03589* (2020).

87. Li, Y., Vinyals, O., Dyer, C., Pascanu, R. & Battaglia, P. Learning deep generative models of graphs. *arXiv:1803.03324* (2018).

88. Simonovsky, M. & Komodakis, N. Graphvae: Towards generation of small graphs using variational autoencoders. *International Conference on Artificial Neural Networks,* 412–422 (2018).

89. De Cao, N. & Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv:1805.11973* (2018).

90. Flam-Shepherd, D., Wu, T. C. & Aspuru-Guzik, A. MPGVAE: Improved Generation of Small Organic Molecules using Message Passing Neural Nets. *Machine Learning: Science and Technology* (2021).

91. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in Neural Information Processing Systems,* 6410–6421 (2018).

92. Jin, W., Barzilay, R. & Jaakkola, T. Multi-objective molecule generation using interpretable substructures. *International Conference on Machine Learning,* 4849–4859 (2020).

93. Lei, T., Jin, W., Barzilay, R. & Jaakkola, T. Deriving neural architectures from sequence and graph kernels. *arXiv:1705.09037* (2017).

94. Thomas, N. *et al.* Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv:1802.08219* (2018).

95. Smidt, T. E., Geiger, M. & Miller, B. K. Finding symmetry breaking order parameters with Euclidean neural networks. *Physical Review Research* **3,** L012002 (2021).

96. Smidt, T. E. Euclidean symmetry and equivariance in machine learning. *Trends in Chemistry* (2020).

97. Hutchinson, M. *et al.* LieTransformer: Equivariant self-attention for Lie Groups. *arXiv:2012.10885* (2020).

98. Unke, O. T. *et al.* Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *arXiv:2105.00304* (2021).

99. Batzner, S. *et al.* SE (3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *arXiv:2101.03164* (2021).

100. Müller, C. *Spherical harmonics* (Springer, 2006).

101. Dray, T. A unified treatment of Wigner D functions, spin-weighted spherical harmonics, and monopole harmonics. *Journal of mathematical physics* **27,** 781–792 (1986).

102. Hermann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry* **12,** 891–897 (2020).

103. Pfau, D., Spencer, J. S., Matthews, A. G. & Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Physical Review Research* **2,** 033429 (2020).

104. Choo, K., Mezzacapo, A. & Carleo, G. Fermionic neural-network states for ab-initio electronic structure. *Nature Communications* **11,** 1–7 (2020).

105. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation* **9,** 1735–1780 (1997).

106. Rajan, K., Zielesny, A. & Steinbeck, C. DECIMER: towards deep learning for chemical image recognition. *Journal of Cheminformatics* **12,** 1–9 (2020).

107. Li, Y., Rezaei, M. A., Li, C. & Li, X. Deepatom: A framework for protein-ligand binding affinity prediction. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* 303–310 (2019).

108. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35,** 3329–3338 (2019).

109. Jiménez, J. *et al.* DeltaDelta neural networks for lead optimization of small molecule potency. *Chemical Science* **10,** 10911–10918 (2019).

110. Jiménez, J., Doerr, S., Martinez-Rosell, G., Rose, A. S. & De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33,** 3036–3042 (2017).

111. Ahmed, E. *et al.* A survey on deep learning advances on different 3D data representations. *arXiv:1808.01462* (2018).

112. Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. *Learning Mesh-Based Simulation with Graph Networks* in *International Conference on Learning Representations* (2020).

113. Liu, Q. *et al.* OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein-ligand affinity prediction. *Journal of Molecular Graphics and Modelling* **105,** 107865 (2021).

114. Mylonas, S. K., Axenopoulos, A. & Daras, P. DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins. *arXiv:2002.05643* (2020).

115. Barnard, J. M. Representation of Molecular Structures-Overview. *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes,* 27–50 (2003).

116. Wiswesser, W. J. Historic development of chemical notations. *Journal of Chemical Information and Computer Sciences* **25,** 258–263 (1985).

117. Wiswesser, W. J. The Wiswesser Line Formula Notation. *Chemical & Engineering News Archive* **30,** 3523–3526 (1952).

118. Ash, S., Cline, M. A., Homer, R. W., Hurst, T. & Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *Journal of Chemical Information and Computer Sciences* **37,** 71–79 (1997).

119. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI the worldwide chemical structure identifier standard. *Journal of Cheminformatics* **5,** 1–9 (2013).

120. Zhang, T., Li, H., Xi, H., Stanton, R. V. & Rotstein, S. H. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *Journal of Chemical Information and Modeling* **52,** 2796–2806 (2012).

121. Öztürk, H., Özgür, A., Schwaller, P., Laino, T. & Ozkirimli, E. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today* **25,** 689–705 (2020).

122. 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. *Angewandte Chemie International Edition* **53,** 8108–8112 (2014).

123. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **4,** 268–276 (2018).

124. O'Boyle, N. & Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *chemrxiv.7097960.v1.*

125. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1,** 045024 (2020).

126. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Learning internal representations by error propagation* tech. rep. (California Univ San Diego La Jolla Inst for Cognitive Science, 1985).

127. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **6,** 107–116 (1998).

128. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. *International Conference on Machine Learning,* 1310–1318 (2013).

129. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555* (2014).

130. Valsecchi, C., Grisoni, F., Motta, S., Bonati, L. & Ballabio, D. NURA: A curated dataset of nuclear receptor modulators. *Toxicology and Applied Pharmacology* **407,** 115244 (2020).

131. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **39,** 2887–2893 (1996).

132. Yuan, W. *et al.* Chemical Space Mimicry for Drug Discovery. *Journal of Chemical Information and Modeling* **57.** PMID: 28257191, 875–882 (2017).

133. Bjerrum, E. J. & Threlfall, R. Molecular generation with recurrent neural networks (RNNs). *arXiv:1705.04612* (2017).

134. Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Molecular Informatics* **37,** 1700111 (2018).

135. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Molecular Informatics* **37,** 1700153 (2018).

136. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Communications Chemistry* **1,** 1–9 (2018).

137. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics* **9,** 1–14 (2017).

138. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **4,** eaap7885 (2018).

139. Grisoni, F. *et al.* Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Science Advances* **7** (2021).

140. Arús-Pous, J. *et al.* Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics* **11,** 1–13 (2019).

141. Grisoni, F., Moret, M., Lingwood, R. & Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *Journal of Chemical Information and Modeling* (2020).

142. Müller, A. T., Hiss, J. A. & Schneider, G. Recurrent neural network model for constructive peptide design. *Journal of Chemical Information and Modeling* **58,** 472–479 (2018).

143. Nagarajan, D. *et al.* Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *Journal of Biological Chemistry* **293,** 3492–3509 (2018).

144. Hamid, M.-N. & Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* **35,** 2009–2016. ISSN: 1367-4803 (Nov. 2018).

145. Das, P. *et al.* Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering* **5,** 613–623 (2021).

146. Zhou, S., Zou, H., Liu, C., Zang, M. & Liu, T. Combining Deep Neural Networks for Protein Secondary Structure Prediction. *IEEE Access* **8,** 84362–84370 (2020).

147. Tsai, S.-T., Kuo, E.-J. & Tiwary, P. Learning molecular dynamics with simple language model built upon long short-term memory neural network. *Nature Communications* **11,** 1–11 (2020).

148. Gomez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **4,** 268–276 (2018).

149. Lin, X., Quan, Z., Wang, Z.-J., Huang, H. & Zeng, X. A novel molecular representation with BiGRU neural networks for learning atom. *Briefings in Bioinformatics* **21,** 2099–2111 (2019).

150. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling* **58,** 1736–1741 (2018).

151. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **11,** 3316–3325 (2020).

152. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature Communications* **11,** 1–8 (2020).

153. Kreutter, D., Schwaller, P. & Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chemical Science* (2021).

154. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence,* 1–9 (2021).

155. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv:2010.09885* (2020).

156. He, J. *et al.* Molecular Optimization by Capturing Chemist's Intuition Using Deep Neural Networks (2020).

157. Honda, S., Shi, S. & Ueda, H. R. SMILES transformer: pre-trained molecular fingerprint for low data drug discovery. *arXiv:1911.04738* (2019).

158. Mirza, M. & Osindero, S. Conditional generative adversarial nets. *arXiv:1411.1784* (2014).

159. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein generative adversarial networks. *International Conference on Machine Learning,* 214–223 (2017).

160. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Communications* **11,** 1–10 (2020).

161. Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science* **11,** 577–586 (2020).

162. Alperstein, Z., Cherkasov, A. & Rolfe, J. T. All smiles variational autoencoder. *arXiv:1905.13343* (2019).

163. Hirohara, M., Saito, Y., Koda, Y., Sato, K. & Sakakibara, Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC bioinformatics* **19,** 83–94 (2018).

164. Kimber, T. B., Engelke, S., Tetko, I. V., Bruno, E. & Godin, G. Synergy effect between convolutional neural networks and the multiplicity of SMILES for improvement of molecular prediction. *arXiv:1812.04439* (2018).

165. Zheng, S., Yan, X., Yang, Y. & Xu, J. Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of Chemical Information and Modeling* **59,** 914–923 (2019).

166. Lim, S. & Lee, Y. O. Predicting Chemical Properties using Self-Attention Multi-task Learning based on SMILES Representation. *arXiv:2010.11272* (2020).

167. Shin, B., Park, S., Kang, K. & Ho, J. C. Self-attention based molecule representation for predicting drug-target interaction. *Machine Learning for Healthcare Conference,* 230–248 (2019).

168. ElAbd, H. *et al.* Amino acid encoding for deep learning applications. *BMC bioinformatics* **21,** 1–14 (2020).

169. Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I. & Welling, M. E (n) Equivariant Normalizing Flows for Molecule Generation in 3D. *arXiv:2105.09016* (2021).

170. Fujita, T. & Winkler, D. A. Understanding the roles of the "two QSARs". *Journal of Chemical Information and Modeling* **56,** 269–274 (2016).

171. Hu, W. *et al.* Open graph benchmark: Datasets for machine learning on graphs. *arXiv:2005.00687* (2020).

172. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **9,** 513–530 (2018).

173. Polykovskiy, D. *et al.* Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in Pharmacology* **11** (2020).

174. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling* **59,** 1096–1108 (2019).

175. Von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry,* 1–12 (2020).

176. Unke, O. T. *et al.* Machine learning force fields. *Chemical Reviews* **121,** 10142–10186 (2021).

177. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **1,** 1–7 (2014).

178. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs: Quantum Mechanical Properties of Drug-like Molecules. *arXiv:2107.00367* (2021).

179. Von Rudorff, G. F., Heinen, S. N., Bragato, M. & von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S2 reactions. *Machine Learning: Science and Technology* **1,** 045026 (2020).