Toward AI/ML-assisted Discovery of Transition Metal Complexes

Hongni Jin[a,b*] and Kenneth M. Merz, Jr.[a,b]*

[a]Department of Chemistry, Michigan State University,

East Lansing, Michigan 48824, United States

[b]Center for Computational Life Sciences, Lerner Research Institute, The Cleveland Clinic,

Cleveland, Ohio 44106, United States

*Email: jinhongn@msu.edu merz@chemistry.msu.edu

Abstract

Traditional computational methods for molecule design are based on first principles calculation, which places a high demand on computing power. The increasingly powerful machine learning (ML) models have fundamentally transformed this landscape. Statistically, by learning the joint probability distribution between molecular or material structure and targeted properties, generative models can autonomously design numerous novel structures with satisfactory properties. This inverse design strategy clearly outperforms the traditional physics-based methods which requires human expertise and intuition, along with serendipity. To validate the generated molecules or materials for specific properties, classical discriminative models allow for fast large-scale screening of the quantitative structure-activity relationships. Generally, the completely ML-based workflow from generation to validation for the exploration of chemical space is accessible and provides outstanding benefits which traditional computational approaches struggle to achieve. In this review, we summarize recent advances in ML-assisted discovery for transition metal complexes and conclude with several existing challenges which impede the widespread practical applications of this technology to the class of problems.

1.Introduction

The flexible electron configurations in *d* or *f* orbitals in transition metals define new bonding patterns which differentiate them from covalent bonds in organic molecules.[1,2] Metal-ligand bonding is very complex, with a wide variety of coordination nubmers.[3] Moreover, ligand-ligand and metal-metal interactions also pose challenges in the design of organometallic compounds. On the other hand, the intricate structural features of organometallic complexes imparts important and novel properties to transition metal complexes (TMCs). For example, the spin states of transition metals remain of great interest, with increasing interest in the determination of the preferred ground spin state and on the factors that influence the preferred spin state and the overall energetics of the various possible spin states.[4] The existence of multiple energetically accessible spin states in transition metal complexes enriches the diversity of TMCs, leading to numerous technological applications. Taken together, these features of TMCs motivates the active exploration of unknown complexes with optimal properties.

The discovery of novel structures in transition metal chemistry involves several steps, including computational design/validation, experimental synthesis/testing with a focus on practical applications.[5] In this review, we focus on the ongoing development of computational strategies with particular emphasis on artificial intelligence(AI)/machine learning(ML) for the design of novel TMCs. Traditional computational approaches leverage explicit chemical principles with human intuition and expertise for rational *in silico* complex design. The detailed analysis obtained from these computational methods are valuable resources to guide the inverse design of new molecules. The rapid development of DFT methods facilitates the extensive investigation of electronic structure data derived from TMCs.[6] However, for the large-scale screening of promising

candidates, the efficiency and effectiveness of DFT calculations are far from satisfactory, regardless of the advancements in computing power.

To expediate scientific research in materials science, ML methods have been broadly used to assist in the search of targeted candidates.[7-10] Given tens of thousands of samples, the ML model is capable of autonomously looking for shared patterns and capturing minimal structural differences, *e.g.* conformers, with regard to predefined properties. The learned relations between structures and properties are then used to make predictions for unknown samples. With the unprecedent advancements in ML algorithms as well as the availability of extensive data sets in the past decade, the ML applications in chemical space has been transformed in several aspects: 1) the increasing complexity of the investigated systems from simple, single mocules[11] to challenging, composite systems such as crystal structures,[12] polymers.[13] 2) The development of ML-assisted functionals that provide a good understanding of electronic structure data rather than the use of neural network models without any physical knowledge.[14] 3) Improving the versatility of ML models for multitask prediction.[15] As a data-driven method, the limitation of a ML-based approach is generally recognized, to be the lack of interpretability. Unlike traditional computational methods, where the underlying physical and chemical principles are explicitly defined and understood, ML models often function as "black boxes", *i.e.*, they can provide accurate predictions, but the reasoning behind these predictions are opaque and without explanation. As a result, the insights into the chemical mechanisms obtained from ML models are limited. But the advantage of ML models is its ability to significantly decrease the computational cost while maintaining a high level accuracy. This efficiency allows for the sampling of larger chemical spaces and the fast simulation of

complex systems. And this capability is particularly beneficial in tasks such as drug discovery, material design, and process optimization, where speed and precision are crucial.

In Section 2, we first discuss various physics-knowledge based computational approaches for the design of TMCs, along with outstanding challenges. The general idea of these methods is the reassembly of available ligands with or without replacing functional substituents in the ligands. Next, we introduce generative ML models to explore the inorganic materials space. Unlike classical methods which require predefined ligands, generative models design novel structures from scratch without any prior knowledge. By learning from realistic data distributions in chemical spaces, generative models can sample numerous, appropriate structures within seconds, which notably accelerates the discovery process. A necessary step for these designed molecules, using either method, is the computational validation of the targeted properties before experimental synthesis. Currently, the main strategy is to use time-consuming DFT calculations. To optimize this process, in Section 3, we explore various ML methods to predict the energy-based properties of TMCs. We also discuss how these classical ML methods can combine with generative models to optimize the design workflow and maximize the efficiency of AI/ML-assisted discovery of new materials. The success of ML relies heavily on reference data. Although the chemical space of TMCs is less explored than that of organic molecules, significant progress has been made to cover TMC chemical space. In Section 4, we list some publicly available data sets in terms of complexes and ligands. These data sets are valuable resources for ML studies in transition metal chemistry. They can be either directly used to train new ML models for various properties or combined with newly generated complexes to expand the transition metal space. Finally, we conclude with

opportunities and challenges for the future of ML modeling of transition metal chemistry. The schematic illustration of this review is given in Figure 1.

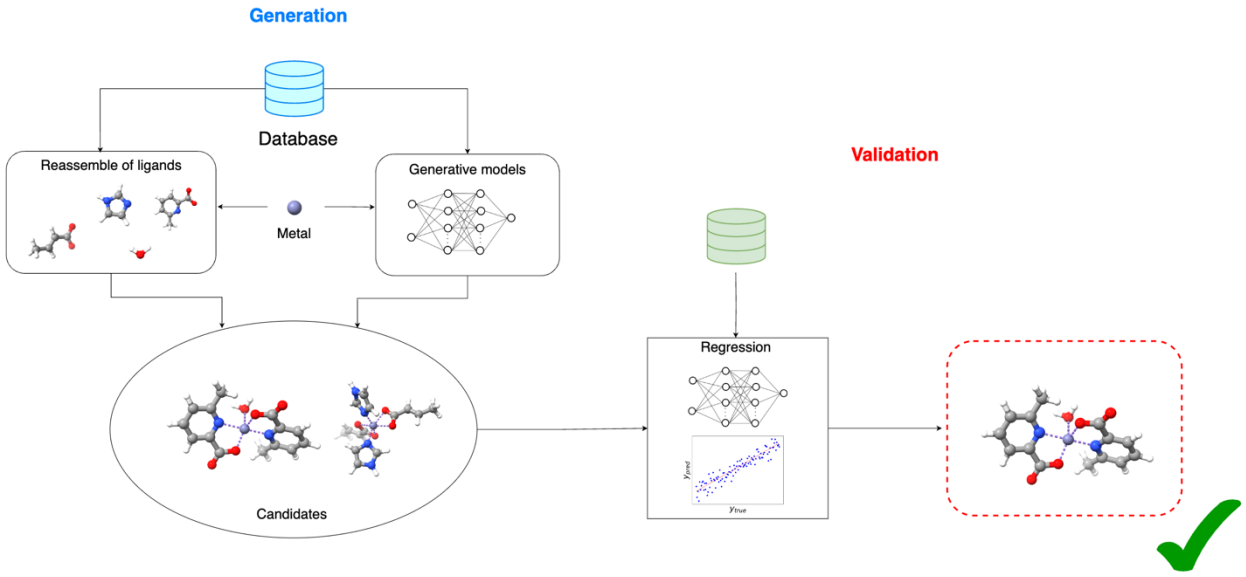

Figure 1. The design of transition metal complexes.

2. The Design of 3D Novel Structures

The study of 3D structure design is essential because well-defined 3D structures are a prerequisite to the exploration of chemical space in order to understand, for example, conformational space or to investigate complex chemical reaction mechanisms.[16-19] Extensive research effort has been expended toward understanding the representation of chemical structures, along with the development of automatic 3D structure generation.[20-24] Whereas most methods for *de novo* design are developed for drug-like organic molecules, successful strategies, to a lesser extent, have been explored in organometallic and transition metal chemistry.[25-27] The potential applications of TMCs, ranging from catalysis and materials science to medicine and environmental technology, underscore the importance for this area of research. As a broad class, organometallic molecules

are much more complex and structurally diverse than organic molecules, which poses new challenges to the discovery of novel structures in transition metal chemistry, requiring an in-depth understanding of both theoretical and experimental principles.[28-31] Herein, we provide a concise overview of various methods for the discovery of inorganic materials, with a focus on TMCs.

In Section 2.1 we discuss the traditional methods of designing novel TMCs. These methods leverage the existing TMCs as templates and then deconstructs the complexes to extract available ligands. By combining different ligands together, new structures of TMCs are then designed. One obvious limitation of this complex-level design is that it highly relies on already synthesized TMCs, and it is simply a selection of all possible recombination of available ligands. As a result, ligands are not designed from scratch, which limits the discovery of novel TMCs because the chemical space of known TMCs is smaller than that of organic molecules and accordingly the variety of ligands is also limited. Section 2.2 discusses how ML can overcome this drawback and open new pathways for the *de novo* design of TMCs. The schematic timeline of representative open-source tools for TMCs design is shown in Figure 2.

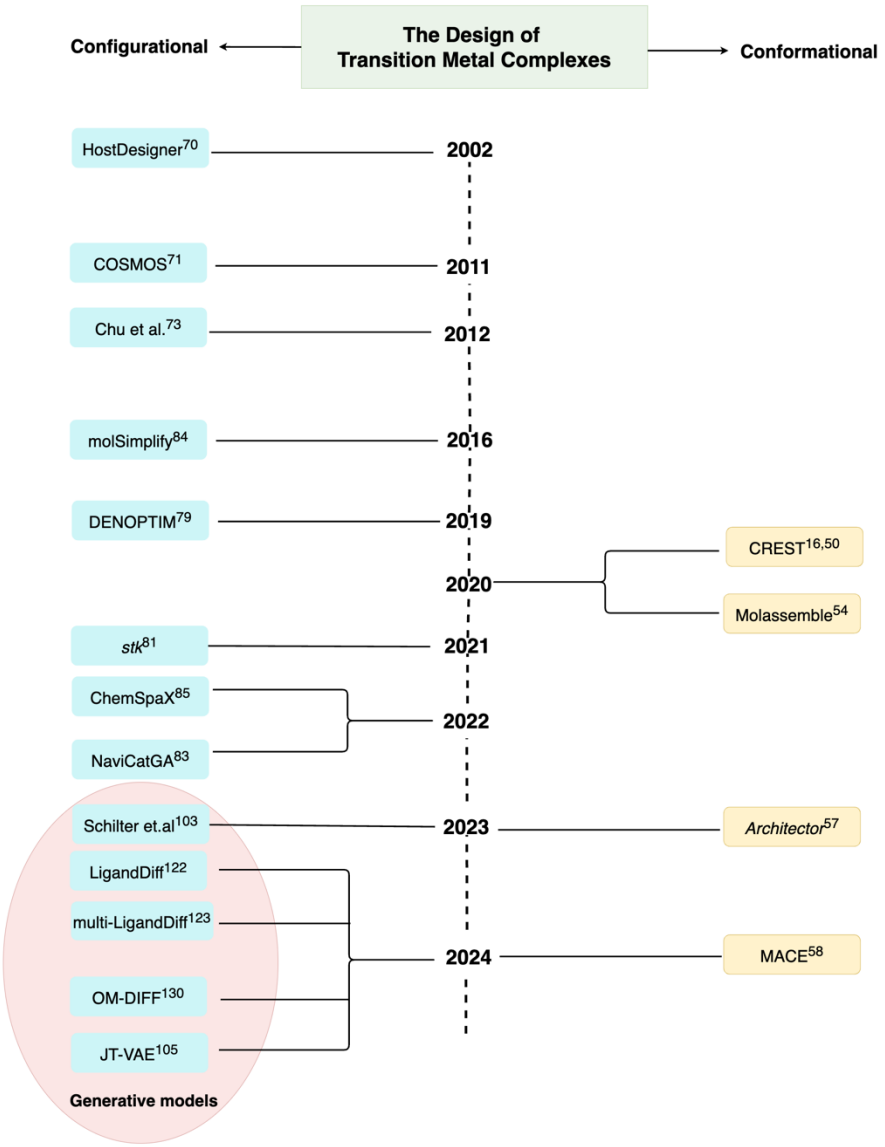

Figure 2. The timeline of strategic development for TMCs design. Representative approaches for configurational design are listed on the left, and available tools for conformation ensemble generation are shown on the right.

2.1 Traditional Strategies for the Design of Transition Metal Complexes

The first type of techniques we discuss focuses on conformer sampling in transition metal chemistry. The exploration of low-energy chemical space is essential for the investigation of

quantitative structure-activity relationships (QSAR) because the functional characteristics of molecules are largely determined by the most stable configuration.[32,33] In addition, conformer sampling also plays a crucial role in several different types of computational analyses, such as docking,[34,35] pharmacophore searching,[36] and receptor-based virtual screening.[37] Moreover, ligand conformations have profound effects on catalytic activities.[38,39] However, capturing reasonable conformer ensemble is significantly challenging because of the huge conformational space of flexible molecules. Therefore, software which can accurately and efficiently locate the local minimum on the potential energy surface (PES) is highly relevant.[40-42]

While most packages explore the low-energy chemical space of drug-like molecules, such as Balloon,[43,44] RDKit,[45] OMEGA,[46,47] and Forg2,[48,49] conformer sampling in transition metal chemistry is less well investigated. For example, CREST[16,50] develops a meta-dynamics driven search algorithm at a semiempirical quantum mechanics GFNn-xTB[51,52] level to explore the conformational ensembles for almost any chemical species in either the gas-phase or explicit solvation. In terms of the sampling of TMCs, bond breaking may occur due to the excess bias potential exerted to drive the sampling and bond constraints need to be applied to remedy this issue.[53] Simultaneously, Molassembler[54] proposes a graph-based method to construct the 3D structures of (in)organic molecules. Each molecule model is embedded as an undirected graph where atoms are represented as nodes and chemical bonds are encoded as edges. The spatial orientations of bonded atoms are traced in an index of permutation, including atom-centered stereopermutators which capture the orientation differences of neighbors around a central atom and bond-centered stereopermutators which capture bond rotation differences. Both types of representations are used for spatial modeling to generate a distance bounds matrix, which is finally

converted to spatial atom coordinates via distance geometry (DG)[55,56] to generate the conformer ensemble. However, the limited accessibility to stereopermutators may lead to failures to conformer generation and since no strict algorithm is introduced to control the generation of conformer at local minima, generated structures are not recommended for downstream tasks without further geometry optimization.

More recently, *Architector*[57] leverages the known experimental chemical space to design new complexes and energetically stable conformers. This python package highlights the conformer sampling of f-block systems which usually have more coordination sites than d-block complexes. Given a complex input, *Architector* first identifies the metal core, coordination number and ligand SMILIES. The metal symmetry is determined by referring to a predefined geometry table derived from CSD molecular symmetries. Ligand type, denticity and the corresponding ligand bite angle are assigned from a built-in ligand library of 27 ligand geometries. The mapping between ligands and an identified geometry is then automatically performed and redundant binding site mappings due to identical ligands are removed. For ligand generation, *Architector* uses DG to constrain the distance of atoms in the pre-optimized ligand. The final output is a list of conformers energetically ranked by GFN2-xTB. Results indicate that *Architector* generates structures lower in energy than the CSD complexes.

In a recent study, MACE[58] proposes a workflow to convert the metal center and ligand SMILES/MOL to 3D structures covering all possible stereoisomers of octahedral and square

planar complexes. Stereoisomers are a special class in complexes due to the unique spatial geometry of complexes where ligands around a metal center have different spatial orientations, leading to distinctive properties. For example, cisplatin shows great effectiveness as an anticancer agent and has been clinically applied, however, its trans-isomer, transplatin is found to be clinically ineffective.[59] The inclusion of stereoisomers can further enrich the diversity of TMCs. Similar to *Architector*, MACE first identifies the central metal, coordination geometry and ligand SMILES and lists all possible stereoisomers: octahedral (30) and square planar (3). After removing identical and "exotic" configurations due to non-localization, 3D atom coordinates are generated via RDKit-supported DG.[45] Post-optimization is implemented but it currently only supports the more qualitative Universal Force Field (UFF).

As we discuss above, most programs for conformer generation employ DG to explore the conformational space with the assumption that by constraining lower and upper distance bounds between all pairs of atoms, all possible conformers can be covered. Some empirical information, such as bond length, bond angles and torsion angles may also be added to the distance bounds matrix for better sampling. Nonetheless, the generated structures via this algorithm still can have distorted geometry and torsional angle values beyond the experimental range because within the distance constraints, atomic coordinates are randomly generated. Therefore, error checking or geometry filtration is essential to ensure high-quality conformations. In contrast, CREST utilizes a RMSD based bias potential to run meta-dynamics for its conformer search. The extra potential drives the structure far away from previously visited geometries, allowing for a wide screening of the PES. To ensure the lowest energy conformer is searched, all generated conformations are optimized and energetically sorted, and the meta-dynamics simulations are run multiple times until

no lower conformer is found. While CREST can provide reasonable chemical structures, its computational cost is far more expensive than the DG-based methods.

Another type of molecular design technique aims at generating configurationally diverse transition metal compounds. Different from conformer sampling where the covalent topology is maintained while changing the 3D shape of structures by rotating one or more bonds, the configurational exploration in transition metal domain attempts to attach configurationally new ligands to the metal center in order to design structurally diverse TMCs for the high-throughput screening of identification of complexes with desirable properties. This configuration-based design is equally as important as conformer sampling because the availability of experimental TMCs is quite limited, with only 242,829 mononuclear TMCs reported in the CSD.[60] The relative scarcity of TMCs reveals the enormous potential for the *in silico* design of new complexes which can guide experimental synthesis.[61,62] In addition, the configurationally diverse complexes are a necessity for conformer sampling as conformation generation is based on a complete complex structure. With more complexes available, conformation sampling has more sources to investigate QSAR models.

The complexity of TMCs poses both challenges and opportunities to the discovery of new complexes. For example, the complicated ligand-ligand interactions makes it difficult to design reasonable ligands that can spatially match the coordination site well in terms of steric hinderance.[63-67] Steric strain affects the connectivity, ligand arrangement, and the magnetic and electrochemical properties of complexes.[63] On the other hand, each ligand in a complex is structurally independent, which brings much flexibility to the ligand design. Ligands can be

partially modified by inserting or removing a functional group or the entire ligand can be replaced by an entirely different ligand, or even more than one ligand can be substituted. The flexible manipulation on ligands enriches the diversity of TMCs.[68,69]

To facilitate the investigation of the binding interactions between metal ion guests and host structures, HostDesigner[70] introduces scoring algorithms to identify promising host component from predefined fragments and the proposed host-guest systems are built via the LINKER and OVERLAY modules. LINKER connects fragments together from the library while OVERLAY superimposes linking fragments onto a predefined complex structure. COSMOS[71] uses a data-driven method to predict 3D structures of small molecules including both organic and organometallic systems. A fragment library which consists of rigid fragments and cyclic fragments was built by decomposing molecules in the Cambridge Structural Database (CSD).[72] For a given query molecule, COSMOS first deconstructs it into fragments, and the query fragments are then matched against the built-in library. Finally, the matched fragments are connected to construct 3D structures, along with adjusting the torsion angles to minimize steric interactions. COSMOS is a biased method because the distribution of fragments in the built-in library is highly imbalanced, indicating that fragments with high frequency are more likely to be matched than fragments with lower occurrence.

Chu et al.[73] proposed a fragment-based evolutionary algorithm (EA)[74-78] to *de novo* design functional TMCs via a search space of ligand scaffolds. As a generic and population-based optimization strategy, EA provides approximate solutions to problems that are difficult to exhaustively sample. The quality of candidate solutions is evaluated by a fitness function and the

evolutionary process is iterated until the predefined property criteria is satisfied. One prerequisite of EA is the presence of fitness functions to guide the optimization process. In the case study, the authors used a QSAR model as the only fitness function to generate ruthenium olefin metathesis catalysts.[73] Although this evolutionary method successfully optimizes multiple highly active catalysts, these promising structures are estimated to be synthetically inaccessible, which limits the practical applications of this method. To overcome this drawback, DENOPTIM[79] introduces a set of predefined connection rules to control the synthetic accessibility.[80] Significant effort to leverage EA to design organometallics has been reported recently, with such approaches as *stk*,[81] PoreMatMod.ji[82] and NaviCatGA.[83]

The Kulik lab has developed molSimplify[84] for the screening of TMCs. This toolkit predefines some common coordination geometries as templates to construct the geometry of a 3D-complex. Ligands are predefined as well. They can be either chosen from around 160 common built-in ligands or customed by users. The selected ligands are aligned to the coordination sites sequentially following the order of ligand denticity to minimize the global steric repulsion. Simple and fast force field optimizations are recommended after the entire structure is built for the sake of downstream tasks. To design functional complexes, ChemSpaX[85] targets exploring the local chemical space of molecular scaffolds. It allows users to place functional groups on a given complex to generate a series of derivatives from an initial skeleton. Similar to molSimplify, the functional substituents are specified from a predefined ligand database which includes 80 common ligands. And it also supports post-functionalization for either the newly placed substituent or the entire functionalized structure. Another feature of ChemSpaX is that it allows iterative functionalization on a skeleton, *i.e.*, the output of one functionalization can be the input for the

next functionalization. As a result, ChemSpaX can build large molecules with more than 200 atoms. However, it should be noted that a manual error check is necessary during serial functionalization because no chemical rules are considered to control the functionalization in ChemSpaX and exotic geometries may be generated, leading to poor synthetic feasibility.

The recent advances described propose powerful practical strategies to enrich the diversity of TMCs, but the *in silico* design of TMCs still faces challenges for large-scale generation of complexes. One common feature of these toolkits for configuration-based design of TMCs is the need of predefined ligands which can be either assigned from a built-in library or specified as input. Usually only up to hundreds of the most common ligands are included in the ligand library to simplify the coordination environment and thus improve the success rate. And this design strategy generates new complexes by identifying the possible combination of different ligands around a metal center. The new structure is defined at the complex level, while no ligands are newly designed from scratch, which makes the exploration of potential ligand space incomplete. Moreover, an important application of these methods is to provide tens of thousands of candidates for the high-throughput screening of TMCs to enumerate new structures with tailored properties. Although the fast development of computing power advances the computational analyses in chemistry, the computational cost of these methods for large-scale screening is still an issue, especially for EA-based methods, where tens of evolutionary cycles may be applied.

2.2 Generative Models for Inorganic Material Design

In the rapid developing AI area, thanks to both the ever-increasing computational resources and ever growing data sets, AI/ML-assisted strategies for automated molecular design show great

promise to outperform traditional methods.[5] Unlike some common ML models which make predictions, generative AI creates new data in various formats, such as text, image, video, audio, 3D models, *etc*. by learning the data distributions from known data sets.[86] This unique feature satisfies the goal of molecular design to intentionally search for tailored molecules or materials for practical applications in unknown chemical space.[87,88] Moreover, the self-learning characteristics of generative AI makes molecule generation autonomous without human supervision, which can avoid biases, leading to more objective and better decision-making. Another advantage of generative models is the large number of possible rational outputs, which provides large molecular samples for downstream tasks, *e.g*. wet-lab synthesis, computational analyses of structure-property relationships, or even building new ML models using the generated structures.

Traditional methods for molecule discovery greatly rely on serendipity, where the discovery of new molecules can be a matter of luck. These approaches lack the ability to support efficient design, leading to a trial-and-error process that can be time-consuming and resource-intensive. Researchers frequently depend on unexpected results, which can delay progress and make it challenging to achieve specific design goals. In contrast, generative models implement systematic strategies for molecule design by implicitly recognizing and leveraging complex patterns within existing data. Through deep learning, the complex feature representations related to targeted properties can be identified and extracted, which, however, are impossible to accomplish simply via chemical intuition and knowledge using traditional strategies. These implicit features can be parametrized via neural networks for recurrently directed search in chemical space.

Herein we mainly discuss several of the most common generative models used in molecular design: variational autoencoders (VAEs),[89] generative adversarial networks (GANs)[90,91] and diffusion models.[92-95] A general overview of generative models has been provided elsewhere[96-98], so we only focus on molecular design related to metals, like organometallics, inorganic materials, crystal structures, *etc*. Reviews on generative models for drug-like organic molecules have also appeared recently.[99,100] The overarching goal of generative models is to implement advanced algorithms embedded by the architecture, in order to model the latent representation of the high-dimensional probability distribution allowing for the generation of novel data within the existing distribution. The difference between various ML models depends on the architecture, *i.e*., the internal networking of each layer and the arrangement of layers.

VAEs are based on autoencoders[101,102], a class of unsupervised ML models which include an encoder to encode input data into a compressed representation in the latent space and a decoder to reconstruct the low-dimension representation to the representation of the original data as accurately as possible. En/decoder are typically composed of neural networks. Conceptually, the ultimate goal of autoencoders is to minimize the reconstruction loss so that the decoder can give output that closely resembles the original input. Autoencoders, while powerful for extracting latent representations, are trained to simply replicate the input rather than to learn how to generate truly new data from scratch. The emphasis of the model is reconstruction accuracy, and an extremely simplified situation is the model memorizes all training data so that the decoder exactly reproduces all original data without any reconstruction error. In this instance, no meaningful interpolation between the input and latent space is garnered and the model learns nothing. To really generate new data, several variants of autoencoders have be proposed.[102] As the most popular variant of

autoencoders, VAEs introduce variations into autoencoders. Plain autoencoders map the input into a fixed vector in the latent space, while VAEs map it into a distribution by sampling a prior distribution between the encoder and decoder, which allows for more flexible representation of the data. In addition, a new latent representation can be directly sampled from the latent space distribution and fed into the decoder and thus new data can be generated. Moreover, instead of consisting of distinct datapoints as in traditional autoencoders, the structured latent space distribution in VAEs allows for smooth interpolation between data points so that intermediate data points can be generated by interpolating between latent vectors. For example, given benzene and anthracene, a VAEs-based model is likely to generate the structure of naphthalene by interpolating the latent representation of both structures because the model is able to capture the overlap between two latent representations and learn the gradual change between them and thus generate realistic intermediate states. The overview of VAEs is shown in Figure 3.

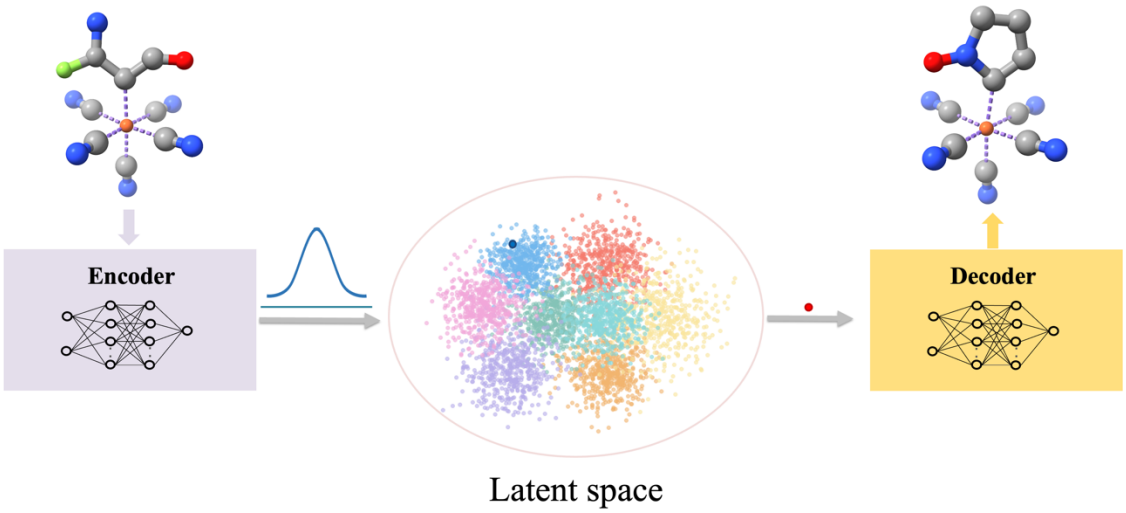

Figure 3. The architecture of a VAE model. By introducing a prior distribution, each input is projected to a sample (highlighted in blue) in the latent distribution. To generate new data, a latent representation (highlighted in red) is sampled and fed into the decoder.

Generative models based on VAEs have been extensively implemented to advance material discovery for various systems. Schilter et al.[103] proposed a hybrid model of VAE and a recurrent neural network (RNN) to design catalysts for Suzuki cross coupling reactions. The model was trained by taking both SMILES and SELFIES[104] of 7054 theoretically validated catalysts with data augmentation to build the latent space. Encoder and decoder of the VAE model are embedded within RNN, and a third neural network is implemented to interpolate the relationships between the latent representation and the reaction energy for conditional generation, *i.e*., the model is constrained to generate TMCs satisfying the catalysis of Suzuki cross coupling reactions. Results indicate that 84% of the generated samples are novel and valid catalysts.

Recently, Strandgaard *et.al* designed JT-VAE[105] for the dual-object inverse design of homoleptic metal complexes. JT-VAE models the latent representation of SMILES-embedded ligands with explicit identification of the metal coordination site. The graph-based ligands are further categorized into rings, bond, and atom clusters by using the junction tree method.[106] The model first unconditionally generated thousands of homoleptic TMCs which were then optimized at the DFT level with a focus on targeted properties. With these generated TMCs, the model was re-optimized by incorporating a third neural network to predict the targeted properties for conditional generation.

For inorganic crystal structures, iMatGen[107] constructs the latent space of solid-state materials by mapping the targeted compositions into image representation. To achieve this, iMatGen utilizes a two-step VAE for image reduction and material generation, respectively. A binary classifier is introduced to enhance the latent vector for the targeted formation energy related to the stability of

materials. The model capably reproduced some experimental vanadium oxide materials as well as designing novel stable structures. Later, Court *et al*.[108] extended iMatGen to design a VAE for not only generating new structures but also predicting eight properties of these generated structures simultaneously. In their work, the unit cells are embedded as electron-density maps and used to construct the latent space. To generate new structures, a UNet model is used to convert new electron-density maps sampled from the latent space to atomic structures. Finally, a graph convolutional neural network is used to predict the properties of generated structures. This conditional VAE model generates reasonable crystals with desired properties. However, the requirement of less than 40 atoms per unit cell limits the generation of complex structures. In addition, the model was trained on limited crystal-structure types, as a result, the model has limited generality and may not generate various crystal structures.

GANs are generative models which consist of two components, a generator and a discriminator, mainly implemented by neural networks. 'Adversarial' in GANs signifies the competitive interaction between the generator and discriminator, where the former creates new data instances that resemble the true data, *i.e*., the training data and shares them with the latter, then the discriminator evaluates the generated instances and distinguishes them from real data as accurately as possible. Both networks evolve together: the generator tries to produce more realistic data to fool the discriminator, while the discriminator strives to become better at identifying fake data received from the generator. The training dynamics is a min-max optimization problem, where the generator minimizes the likelihood of the discriminator correctly classifying its outputs as fake, and simultaneously, the discriminator maximizes its accuracy in identifying real from fake. Over time, this adversarial process leads to Nash equilibrium[109] where the generator creates highly

realistic data, meanwhile, the discriminator becomes highly adept at discriminating subtle differences between true and generated data. This dynamic optimization is shown in Figure 4.

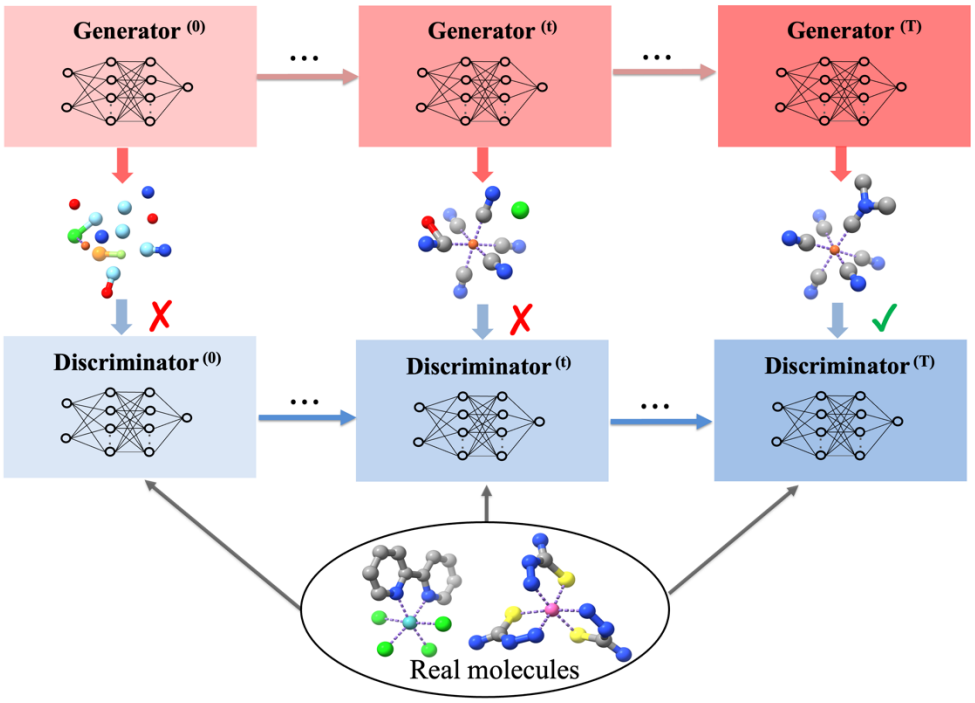

Figure 4. The dynamic optimization process of a GAN model. At the initial step s = 0, the generator only generates isolated atoms or bonds based on the sampled data from a prior distribution and at this step, it is easy for the discriminator to identify the fake data since the generated data is fundamentally different from real molecules. Over time s = t, the generator keeps improving its capability of generating realistic molecules and the discriminator also improve itself to better distinguish fake from real. At the Nash equilibrium s = T, the generator successfully generates real-like molecules.

CrystalGAN[110] is the first GAN model designed to explore novel crystal structures with increased structural complexity. It was trained with stable binary hydrogen storage materials MH, the model generates stable ternary crystal structures, AHB, where A or B=M. The model took the typical

crystallographic representation of the three lattice vectors of a unit cell and the atomic coordinates, *e.g*. H or M as inputs. The generator takes the representation of both AH and BH but generates the representation of AHB, while the discriminator identifies the combined representations. The model successfully generated novel crystal structures with the defined geometric constraints.

Kim *et al*.[111] also used the unit cell parameters and atomic coordinates of the ternary Mg-Mn-O system to design a GAN model for the generation of various novel crystal compositions. The GAN model includes three parts, where the generator takes the representation of the composition, and the critic evaluates the distance between the true data distribution and the training data distribution, inspired by WGANs,[112] and finally a classifier network is used for conditional generation of novel structures by evaluating the representation differences between the generated composition and the real compositions. Theoretical calculations indicated this proposed model generated 14 entirely novel structures with desirable photoanode properties.

Although the two previous models can generate new structures, the design space they cover is very limited because of the limited training data set used in the work, and thus the diversity of generated structures is a concern. To expand the design space, Dan et al.[113] proposed MatGAN for three separated data sets, each of which includes more than 63,000 inorganic compounds. MatGAN uses a sparse matrix of one-hot embedding for atom types and atom numbers in each material. A convolutional neural network with normalization layers were used in both the generator and discriminator. For 2 million generated materials, MatGAN gave valid results 84.5% of the time in terms of both charge neutrality and balanced electronegativity. The t-SNE analysis indicated that MatGAN explored new design space in inorganic materials.

To design complex-architecture materials without prior knowledge, Mao et al.[114] used GANs to analyze millions of various crystallographic architectures sampled from simulation. Both configurations and their properties were used to map the design space so that the discriminator interpolates the implicit relationships between configurations and properties, which guides the generators to promptly generate new configurations with expected properties. For example, the proposed GAN model designed more 400 2D architectures with the HS upper bounds of stiffness at various porosities.

GANs have been extensively applied to facilitate the exploration of inorganic material space, but they also have some shortcomings. First, difficulties in convergence, the adversarial loop between the generator and the discriminator makes the model difficult to train and converge. Second, mode collapse, previous studies observed that the generator sometimes repeatedly generates new data with a certain type of modes, although a wider variety of modes are included in the data distribution.[115,116] To tackle this issue, various variants of GANs have been proposed. The first solution is to avoid a fake local Nash equilibrium. Studies show that model collapse happens along with sharp gradients of the discriminator. To avoid this, DRAGAN proposed to regularize the discriminator by adding extra penalization terms so that its gradients are realistically constrained.[117] Similarly, LSGANs[118] used least squares loss to replace the original cross-entropy loss in the discriminator to overcome the vanishing gradient issue. Finally, several controversial topics, such as the generalization ability[119] and memorization issues,[120] are still being discussed with respect to GANs.

Diffusion models, as the newly emerging generative models, have outperformed GANs in a variety of application domains.[121] Inspired by nonequilibrium statistical mechanics, diffusion models consist of two Markov chains: a forward diffusion chain and a reverse diffusion chain. By iteratively adding noise sampled from a prior distribution to the input data, the forward process gradually destroys the complex data distribution and finally converts it to a simple, known and tractable distribution. Next, the reverse process learns to recover the real data distribution from the noised distribution. Central to the success of a diffusion model is its ability to parametrize the reverse process with a neural network. New datapoints in a targeted distribution are generated by first sampling random unstructured datapoints from a prior distribution and then sequentially removing noise via the learnable reverse process.

We recently proposed a diffusion model, LigandDiff, to design novel ligands for octahedral TMCs.[122] LigandDiff introduces the scaffold technique to the diffusion model and allows for the generation of 3D monodentate or polydentate ligands at a given coordination site. Specifically, given a metal center with several connecting ligands (context *u*), LigandDiff generates one appropriate ligand for the vacant coordination site to form a complete complex. But since the initialization of the reverse process is random, even with the same *u*, numerous ligands with diverse configurations can be designed. In LigandDiff, each complex is denoted as $x = [r, h_a, h_L]$, where *r* is the coordinate of each atom, h_a is the one-hot embedding of atom type, and h_L indicates the ligand group information, *i.e.*, which atoms are from the same ligand. In the forward process, an assigned well-structured ligand x_0^L in each complex is deconstructed to a group of isolated atoms x_T^L. The unstructured data is then used to parametrize a neural network ϕ to estimate the noise at a given step *t,* i.e., the distance between x_t^L and x_0^L. To generate a new ligand, LigandDiff starts

from x_T^L, sampled from a prior distribution, and removes the noise predicted by parametrized ϕ to reverse x_T^L to x_{T-1}^L. By iterating this denoising process from x_t^L to x_{t-1}^L, finally a new ligand x_0^L is generated. The overview of LigandDiff is given in Figure 3.

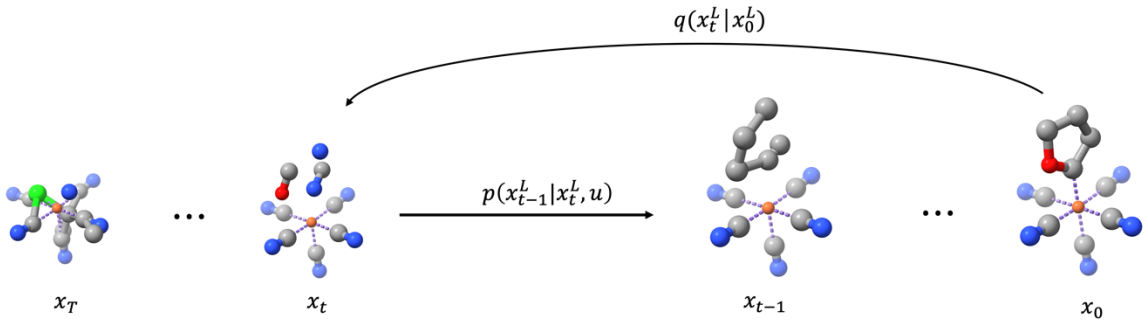

Figure 5. The dynamics of LigandDiff. The diffusion process q gradually adds noise to initial ligand x_0^L while the denoising process removes noise from x_T^L. Reproduced with permission from ref 122. Copyright 2024 American Chemical Society.

LigandDiff is capable of generating novel, unique and valid ligands with high synthetic accessibility and has great transferability to design ligands for any transition metal. Recently, we extended it to multi-LigandDiff,[123] which has two improvements over LigandDiff. First, multi-LigandDiff allows for partial to total generation of ligands for TMCs. Instead of generating only one ligand in LigandDiff, multi-LigandDiff can generate one or more ligands with or without ligands in context u. In addition, it allows users to predefine the ligand denticity of the generated ligands. To achieve this, the coordination site of a complex is embedded into multi-LigandDiff. With this extra embedding, users can customize the ligand denticity of each generated ligand. Ligand denticity influences organometallics in various ways, such as the stability of complexes,[124,125] the geometry of complexes,[126,127] and the reactivity of complexes.[128,129] This conditional design with regards of ligand denticity can facilitate the investigation of metal-ligand

interactions. As the successor of LigandDiff, multi-LigandDiff clearly outperforms LigandDiff in generating valid, novel and unique ligands for any transition metal.[123] Moreover, multi-LigandDiff is capable of generating ligands for non-octahedral complexes. As a universal tool for the discovery of TMCs, multi-LigandDiff can be used to design new complexes with any targeted properties based on scaffolds. That is to say, given a complex with known targeted properties, multi-LigandDiff can generate a series of derivatives of this complex by replacing existing ligands with new ligands from partially to totally. These newly generated complexes maintain the core structure of the reference structure as well as the main properties. Meanwhile, the newly generated ligands bring potentials to complexes for other properties. In our case study, multi-LigandDiff generated 338 theoretically validated Fe(II) spin-crossover(SCO) complexes with reference to only 47 experimentally validated Fe(II) SCO complexes.[123]

OM-DIFF incorporates a neural network predictor to a diffusion model for the inverse-design of novel catalysts for cross-coupling reactions.[130] By inducing the model to generate complexes with appropriate binding energy calculated at the DFT level, several optimized catalysts including Pd, Pt and Cu complexes were generated and theoretically identified to be potential catalysts for the Suzuki reaction. Crystal diffusion variational autoencoder (CDVAE)[131] combines VAE and diffusion models to design novel, stable periodic materials, like perovskites. Three neural networks are concurrently optimized: an encoder which encodes the crystal structures into the latent space; a property predictor which predicts the composition, lattice and number of atoms of a material; a diffusion model based decoder that denoises 3D crystal structures conditioned on the latent representation. Recently, Han et al.[132] improved CDVAE for the sake of inorganic crystal generation with desired compositions. The search in latent space is conditioned on targeted

compositions using gradient descent optimization. In addition, Alverson *et.al*.[133] compared GANs and diffusion models for unstructured crystal generation. The authors first proposed the CrysTens representation which efficiently captures both the pairwise distance matrix and the distance graphs of crystal structures. Three models, including Vanilla GANs, WGANs and diffusion models, were trained with the CrysTens representation of 53,856 CIFs and were evaluated on their ability to understand the symmetrical features of crystals. Results indicate that diffusion models clearly outperformed GANs in capturing the structural crystal information and covering the structural distribution of the real dataset.

Although generative models have paved the way for the design of molecules with tailored properties, AI-assisted methods still face challenges. Currently, almost all candidates designed by generative models are computationally validated for targeted properties and the evaluation of properties is performed by checking the energetics of candidates, such as the bonding affinities, the interaction energies and the free energies, *etc*. Such computation heavily relies on quantum mechanical methods, which however, are time-consuming and resource intensive. To further accelerate the discovery of chemical space, efficient strategies for property evaluations should be proposed. One possible approach is to combine generative models with machine learning potentials (MLPs). The latter is a family of ML methods which model the potential energy surface of chemical systems of interest. A well-trained model is able to predict the energetics with several orders of magnitude faster than traditional DFT methods.

3. Machine Learning Potentials for Transition Metal Complexes

3.1 Overview

Over the past 30 years, various ML algorithms, such as Random Forest (RF), Support Vector Machine (SVM), Extra-Trees (ET), GBoost (XGB), Neural Network (NN), *etc*. have been implemented to investigate the statistical relation between chemical structure and potential energy. By learning from tens of thousands of reference data from accurate yet computationally expensive electronic structures calculations, the parametrized ML models capably perceive the atomic interactions without a significant loss in accuracy and thus can be widely used in large-scale molecular simulations,[134] QSAR investigations,[135] and reaction rate determination.[136] Traditional MLPs based on fixed architectures are constrained to thousands of data points and have limited flexibility and transferability. In contrast, neural network potentials (NNPs), as the most widespread MLPs, are implemented by deep neural networks, composed of generally hundreds of neurons in several layers, which allows the model itself to freely determine the interconnections between neurons in consecutive layers. Dating back to 1995, NNPs have a long history and continue to experience rapid development. NNPs utilize molecular representations to learn the interatomic interactions in molecules. The atomic spatial configurations and interatomic relations are defined by descriptors to model the contribution of single atom to the entire system and the sum of individual contributions is counted as the total energy. Descriptors are either predefined based on intuition and experience from experts or implicitly learned by the model itself.

For example, Behler and Parrinello proposed the atom-centered symmetry functions (ACSFs) to capture radial and angular features of single atom with its neighbors.[137,138] These predefined descriptors transform atomic coordinates into chemically relevant representations for energies while maintaining translational, rotational, and permutational invariances. On the other hand, representative models, such as DTNN,[139] SchNet,[140] PAINN,[141] directly encode the geometric

structures into neural networks and the descriptors related to energies are automatically captured by the model. In NNPs, descriptors are proposed to define the local environments of atoms, which are constrained to a certain range from the central atom by a cutoff. To simplify the complexity and computation, the cutoff radius is usually predefined at 6~8 Å, and the interactions with pairwise distances beyond this cutoff are ignored. As a result, this type of NNPs only covers the short-range interactions and errors can occur by neglecting long range interactions.

To overcome this limitation, long-range interactions, such as electrostatics, van der Waals, can be explicitly added to the short-range component. The long-range interactions are calculated based on empirical formulas, like DFT-D3,[142] DFT-D4.[143] However, the generally parametrized formulas are not applicable to specific systems if high accuracy is required. To better model the long-range interactions, recently, implicit neural networks for long-range physics are proposed. For example, EwaldMP[144] embeds the Ewald summation into a neural network to capture the long-range interactions. The Ewald summation transforms the long-range part which decays slowly with distance into a Fourier space where the interactions decay quickly with frequency and thus can be summed up efficiently. Instead of using empirical formulas as general Ewald summation methods do,[145] EwaldMP achieves Ewald summation for nonlocal interactions by parametrizing neural networks. This parametrized Ewald summation for specific systems based on the training data set definitely outperforms the traditional long-range methods which are generally parametrized for all systems. In addition, So3krates proposes spherical harmonic coordinates (SPHCs) to capture nonlocal electronic effects on molecular systems with arbitrary length scales.[146] Instead of using only fixed atomic coordinates to learn the molecular representations, So3krates also defines SPHCs based on spherical harmonics during the message passing. The distance matrix for all

possible pair-wise atoms in SPHC space are introduced to model the nonlocal interactions. One advantage of SPHC space is that the spherical neighborhoods can include atoms that are far away in Euclidean space so that nonlocal interactions can be covered within short range part in SPHC space.

Although various NNPs have been proposed in the past 30 years, most models aim at delineating the PES of organic systems. The parametrization of TMCs is much more challenging and requires extra effort. First, TMCs have unique electronic properties, like charges, spin states, which effect the potentials of complexes. However, most NNPs only consider embeddings related to atomic coordinates and are only applied on neutral organic molecules. Without explicitly incorporating these electronic properties into neural networks, the proposed architectures are inappropriate for TMCs. Second, unlike organic molecules which have fixed bonding patterns, transition metals usually have flexible coordination types with different coordination numbers, and even ligands with the same atom coordinated to the metal center can have different ligand-field strengths. The increasing geometric complexity makes it difficult to model typical metal-ligand interactions. Moreover, NNPs designed for TMCs should be able to model manifold interactions, including the metal-ligand interactions, the ligand-ligand interactions, or even metal-metal interactions, which requires the architectures to be highly versatile. In addition, since almost any molecular species which can donate electrons can be a ligand, the diverse ligand types also increase the complexity of TMCs, and thus poses outstanding challenges to the generality of NNPs designed for TMCs. Overall, the optimal strategies for NNPs of TMCs are still challenging due to the the high-dimensional structural complexity of TMCs.

3.2 Available Approaches and Applications

Although the development of MLPs for TMCs faces multiple obstacles, significant progress has been achieved with the continuous advancement of cheminformatics. One aspect of the contribution of cheminformatics is attributed to the development of descriptors. Descriptors play a crucial role in data-driven research in chemistry. First, they can reduce the complex dimensionality of data by compressing chemical structures into a tractable string, vector or matrix. For example, 1D representations, e.g. SMILES, or SELFIES[104] are extensively used in cheminformatics when 3D structure isn't essential. Other descriptors, like bond order, the number of aromatic rings, the existence of functional groups are also widely used to decode structural information. More importantly, by means of these chemically intuitive characteristics, it is possible to understand a derived QSAR model. The mapping of descriptors in chemical space provides valuable insights into structural features for targeted properties. By leveraging derived structural information, inverse design of novel molecules is achievable. In terms of NNPs, descriptors play a pivotal role in extracting useful structure features related to energies. Such highly condensed descriptors provide a refined understanding of geometric structures to ML models, particularly those classical models whose simple architectures limit them in terms of interpreting complex 3D structures. As a result, ML models only need to focus on these essential descriptors rather than crude structures, which certainly improves the capability of ML models.

In 2017, Janet *et al*. trained neural networks to predict electronic structure properties of the first-row TMCs in multiple oxidation states as well as spin states.[147] 15-demensional predefined descriptors including complex-focused and ligand-focused features were used to encode the geometric information of TMCs. The trained model predicted the spin state ordering with an

error of around 3kcal/mol and metal-ligand bond length within 0.03 Å accuracy. However, since this model did not consider the 3D structure of TMCs, it is unable to capture the minimal structural differences in conformer ensembles of TMCs. Subsequently, Meyer *et al*. trained ML models based on the reaction energy of oxidative addition reactions to explore new catalysts for cross-coupling reactions.[148] The 3D structure of each catalyst was obtained by converting SMILES to atomic coordinates, and was used to calculate some commonly used descriptors, including the sorted Coulomb Matrix (CM), the Bag of Bonds (BoB) as well as the Spectrum of London and Axilrod–Teller–Muto potential (SLATM). Among them, SLATM representations were found be the most associated with the reaction energy, with an error of 2.61 kcal/mol. Similarly, Friederich *et al*. leveraged ML methods to explore optimal complexes with respect to catalysis.[149] Descriptors derived from autocorrelation and deltametric functions were fed into neural networks with Bayesian hyperparameter optimization to predict the H_2 activation energy with a MAE of around 2 kcal/mol. The authors then extended the feature space with molecular fingerprints (MFs) and utilized gradient boosting to select the most relevant features. Finally, Gaussian process regression with selected features yielded the lowest error of 0.59 kcal/mol. Simultaneously, Cordova *et al*. created a data set of more than 143,000 homogeneous nickel catalysts by replacing ligands at different coordination sites to estimate the catalytic activity for aryl ether cleavage via ML methods.[150]

Similar to Meyer *et al*., three types of descriptors, including CM, BoB and SLATM were individually computed and tested with a kernel ridge regression model. The parametrized model with BoB representations reported the lowest error of 4.28 kcal/mol. To facilitate the discovery of desirable transition metal catalysts for asymmetric hydrogenation of olefins (AHO), Hong's group

curated over 1,2000 AHO transformations and developed ML methods to identify promising candidates.[151] The combination of 2D MFs descriptor and the 3D many-body tensor representation (MBTR) yielded the lowest MAE of 0.317 kcal/mol in terms of ΔΔG. Xu et al. leveraged transition state (TS) knowledge in ML to predict enantioselectivity of pallada-electrocatalysed C–H activation.[152] From the distorted TS geometries generated with constrained optimization, a variety of descriptors at the atom-level and bond-level, such as the Hirshfeld charge, the condensed local softnesses, the bond order, *etc*. were computed to represent the targeted reaction. Their ML model trained with these knowledge-based descriptors reported a MAE of 0.236 kcal/mol with a R^2 of 0.909. In addition to these classical descriptors, Kneiding *et al*. proposed the natural quantum graph (NatQG) to encode the electronic structure data of TMCs based on natural bond orbital (NBO) analysis.[153] The NatQG representations derived from DFT computation, such as geometry optimization and NBO calculation, include topology and electronic structure information. The topology is encoded by either undirected molecular graph (u-NatQG) or directed graph (d-NatQG) with manifested donor-acceptor interactions. NBO data such as atomic charge, valence index and bond order are shared by nodes and edges in NatQG. The embedded NatQG is dynamically updated in a Graph Neural Network (GNN) to predict the quantum properties of TMCs, like the HOMO–LUMO gap.

All ML methods discussed above require predefined descriptors to predict energy-related properties. The exploration of suitable descriptors for ML models is an active area of research, with a variety of ligand descriptors derived from DFT calculations.[154] Recent advances also highlight the impact of descriptors in terms of ligand conformations on TMCs.[155] Although these already proposed descriptors have helped advance ML-based material discovery, the design of

new descriptors for targeted properties is never straightforward. It relies on computational expertise with a bit of serendipity, which introduces inadvertent bias that hinders further development. In addition, the intricate structures of TMCs pose extra requirements on descriptors because classical descriptors, like SMILES, MFs are inadequate for differentiating the structural complexity of configurational isomers, which are commonly observed in transition metal chemistry, *e.g.* *cis*-$[Co(NH_3)_4Cl_2]^+$ vs. *trans*-$[Co(NH_3)_4Cl_2]^+$. Although high-level descriptors based on 3D structures are available, they often involve quantum mechanical computation which hampers the efficiency of the entire ML workflow.

An alternative, and perhaps more straightforward ML approach, is to skip the computation of descriptors and directly employ 3D coordinates as the input of the ML model. By means of molecular architectures presented the model, the model is expected to automatically capture the complexity of structural differences via its mathematical architecture and infer the relations between structural features and properties. Generally, neural networks are widely used in this research area since they provide the flexibility to easily customize the architecture. For example, Garrison *et al.*[156] reported the tmQM_wB97MV data set by filtering out 155 anomalous complexes in the tmQM data set[157] and computing all remaining TMCs at the ωB97M-V/def2-SVPD level. For this refined data set, the authors implemented four GNNs models to predict the energies of TMCs. All tested GNNs models, including SchNet,[140] PaiNN,[141] SpinConv,[158] and GemNet-T,[159] only took atomic numbers and atomic coordinates as inputs. Compared with the original tmQM data set, the model trained on the tmQM-wB97MV data set clearly yield lower error which indicates that the original tmQM data set compiled at the TPSSh-D3BJ/def2-SVP level may be insufficient to describe the electronic structure of TMCs.

Despite the observed improvement at the cost of resource-intensive reference data, energy prediction can be further enhanced by optimizing the architecture of the ML model. All tested models only consider short-range interactions while neglecting long-range interactions, like dispersion which has been validated to be important in TMCs.[160,161] The pivotal role of dispersion in TMCs also explained why models trained on the tmQM-wB97MV data set outperform those on tmQM data set because the reference data of the former was calculated from ωB97M-V method which includes a better nonlocal correction than D3BJ. These findings also indicate that the tested models implicitly capture nonlocal interactions to a limited extent and the incorporation of embeddings related to long-range interactions will further improve the capability of ML models. In addition, electronic properties, such as charge, and spin states are also ignored, which makes it challenging for ML models to meticulously analyze electronic structure data.

Another limitation of this work is the inherent sparsity of the data. The configurationally diverse complexes in the data set ensure the generality of the trained models. However, minimal structural differences among conformers are difficult to capture because the models were never trained to differentiate conformer ensembles. In a conformational PES map the energy landscape of a molecule is a function of the positions of its nuclei. The detailed energy landscape is crucial to the understanding of molecular stability, reactivity, and dynamic behavior. NNPs designed for conformational PESs enable efficient sampling at a negligible computational overhead and allow for extensive conformational searches and molecular dynamics simulations, leading to a comprehensive view of the energy landscape.

Based on the previous discussion, several key factors to build satisfactory NNPs for TMCs can be summarized as follows: 1) the architecture of the NNPs should consider electronic properties; 2) long-range interactions should be explicitly involved; 3) the data set for NNPs should include both configurationally and conformationally diverse complexes.

To achieve this goal, recently, we developed NNPs for zinc and iron(II) complexes, namely Zn_NNPs[162] and Fe(II)_NNPs,[163] respectively. Both types of NNPs aim at accurately modeling the PES of conformer ensembles of complexes by covering the long-range interactions. To model the PES of zinc complexes with the inclusion of conformers, we first curated a subset of the tmQM data set[157] which includes 771 zinc complexes with various bonding patterns and ligand types to ensure a diverse data set for the generality of the trained NNPs. We then generated conformer ensembles for each zinc complex, leading to 39,599 conformations obtained via CREST and the RMSD of each pair of conformations is over 0.1 Å to avoid data redundancy. The single-point energy of each conformation was computed using r^2SCAN-3c method[164] with the D4 dispersion correction. To better cover long-range interactions, we improved the representations of EwaldMP.[144] Specifically, we incorporated partial charges into the model. Although all complexes are neutral in our data set, the partial charge distribution in each complex is significantly different, leading to varied long-range interaction contributions to the total energy. As an independent architecture, EwaldMP can be combined with any existing baseline model so both short-range and long-range interactions can be interpreted. In the original work,[144] EwaldMP and a baseline model share the same atomic embeddings in terms of atomic numbers and atomic coordinates. To highlight the significant role of partial charges in long-range interactions, we differentiated the representations for both components, *i.e.,* the embeddings of partial charges were fed into

EwaldMP, while the baseline model remained unchanged. The resulting Zn_NNPs model thus meets all three requirements mentioned above. We then tested both SchNet[140] and PaiNN[141] as baseline models and compared them with the original EwaldMP and our proposed approach.

Results indicate that Zn_NNPs outperforms both baseline models and the original EwaldMP, yielding the lowest MAE of 0.92 kcal/mol with respect to r^{2}SCAN-3c. Moreover, with reference to PWPB95/CBS with D4 correction, Zn_NNPs surpasses several semiempirical methods, including to GFN1-xtb, GFN2-xtb, PM6-D3H4X, and PM7 for relative conformational energies and capably locates the lowest energy conformation. Overall, this GPU-supported Zn_NNPs can predict the energies of zinc complexes at the DFT level in accuracy but it is several orders of magnitude faster.

We also tested our strategy for Fe(II) complexes in high-spin(HS) and low-spin(LS) states. The goal of our Fe(II) NNPs was to model the conformational PES of Fe(II) complexes and predict the spin-splitting energies ($\Delta E_{\mathrm{HS-LS}}$) simultaneously. Following the procedures in Zn_NNPs,[162] we selected 383 well-defined Fe(II) complexes with the charges {0,+1,+2} from a reported TMC database.[60] Both HS and LS states were assigned to each complex and 28834 spin-state-specific conformers were generated via CREST and optimized at the B97-3c level. The single-point energy was calculated using TPSSh/ def2-TZVP with the D4 correction. Complexes in this curated data set have varied charge and spin state properties.

To encode these global electronic properties, each property of each complex is shared among atoms in proportion to their atomic numbers. We tested several combinations of embeddings on

different models: 1) a baseline model with only classical atomic embeddings; 2) a baseline model with both classical atomic embeddings and electronic embeddings; 3) a baseline model with classical atomic embeddings + EwaldMP with electronic embeddings; 4) a baseline model and EwaldMP shared the same classical atomic embeddings; 5) a baseline model and EwaldMP shared both classical atomic embeddings and electronic embeddings. Results indicate that these electronic embeddings make contributions to modeling long-range interactions and SchNet with both classical atomic embeddings and electronic embeddings yields the lowest MAE for the prediction of the total energy and the spin-splitting energy, which are 0.037eV and 0.030eV, respectively. The combination of SchNet with classical atomic embeddings and Ewald with electronic embeddings performs a bit worse but clearly outperforms other combinations, indicating the significant importance of electronic embeddings to EwaldMP. To evaluate the ability of Fe_NNPs to identify the ground spin state of Fe(II) complexes, from a total number of 23446 pair of complexes in the HS state and LS state, we computed all possible complexes and recorded the prediction of the model. The trained model incorrectly predicted only 8 ground spin states, which surpassed semiempirical methods by several orders of magnitude.

This remarkable Fe_NNPs model which allows for efficient conformation sampling as well as ground spin state identification has a variety of promising applications for Fe(II) complexes. For example, in our multi-LigandDiff work,[123] we proposed a workflow to design Fe(II) SCO complexes (Figure 6). To quickly look for promising Fe(II) SCO complexes from thousands of candidates, we used Fe_NNPs to do the preliminary screening. The spin-splitting energies of 2231 pairs of Fe(II) complexes in both the HS and LS states were predicted by the Fe_NNPs model within seconds thanks to GPU parallel computing. 560 pairs of complexes identified by the model,

which were predicted to be promising SCO complexes, were further calculated using the TPSSh method with the D4 dispersion correction. Finally, 338 out of 560 complexes were labeled as Fe(II) SCO complexes. Evidently, the application of Fe_NNPs greatly expedites the entire design workflow via decreasing the computational load by 75% but without significant loss in accuracy in this qualitative analysis.

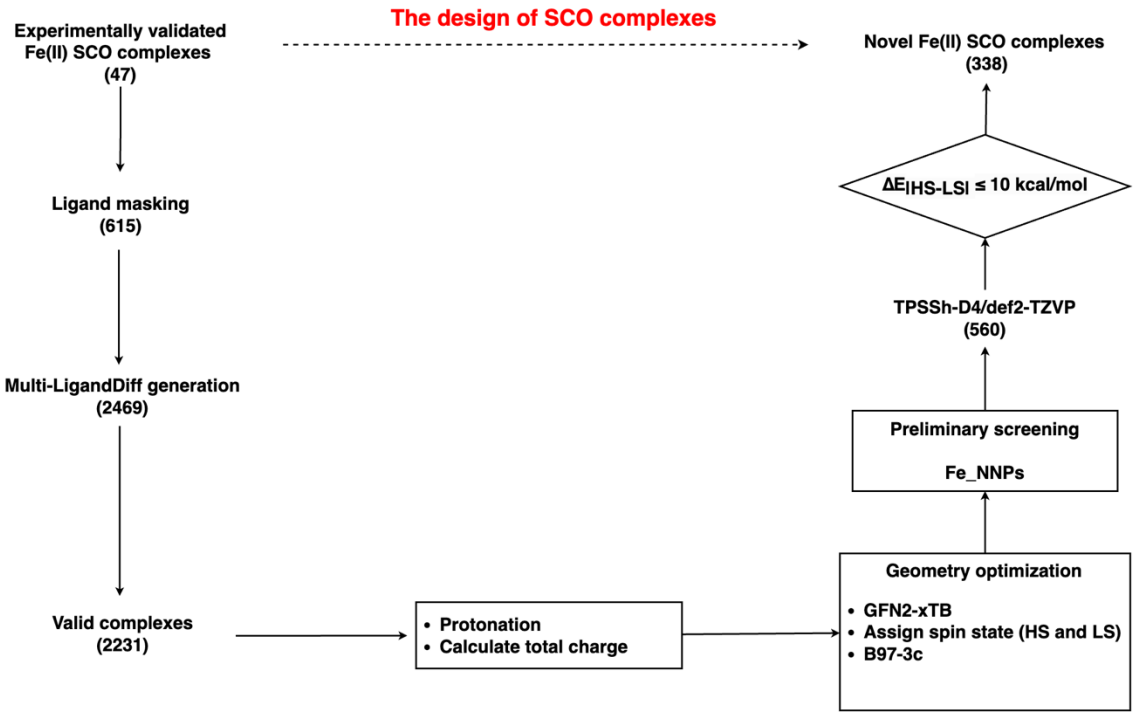

Figure 6. The design of novel Fe(II) SCO complexes with Fe_NNPs model. Reproduced with permission from ref 123. Copyright 2024 American Chemical Society.

Overall, the development of NNPs for TMCs can be concisely summarized in Figure 7. For a data set with less than 10k datapoints, traditional ML models and simple neural networks are a good choice to prevent overfitting. But considering the simplicity of these ML architectures, suitable descriptors, such as bond order, SLATM should be pre-computed and used as inputs of the ML models. In contrast, for high-level neural networks, easily accessible representations, including

atomic coordinates and atomic numbers are directly fed into the model with the expectation that the model can automatically capture the hidden structural features for targeted properties. In addition, charge and spin state information are essential attributes for TMCs, and the incorporation of them into neural networks can further improve the capability of NNPs for TMCs.

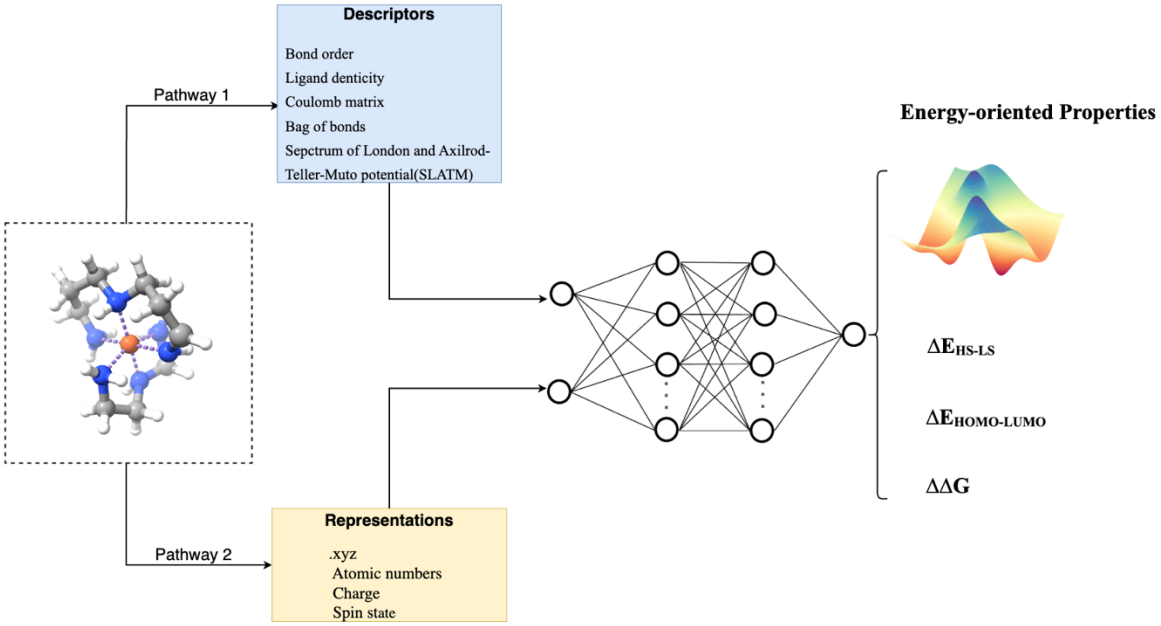

Figure 7. Two types of NNPs for TMCs. 1) Complex-level and ligand-level descriptors are used to facilitate neural networks to capture important structural information. 2) Various atomic representations derived from 3D structures without any refinement are directly passed into neural networks.

4. Data Availability

Despite these recent advancements discussed in Section 2 and Section 3, the discovery of novel TMCs is still experiencing difficulties due to two limiting factors identified in the existing reference data: poor diversity and inconsistent quality. As discussed in Section 2.1, traditional strategies for the design of novel TMCs require available ligands as templates to generate new complexes. And although generative models do not need any templates for the generation of

complexes, the training of generative models is based on available TMCs. That is to say, the diversity of already available TMCs is a limiting factor for the current strategies for the exploration of TMCs. Without massive reference data to cover a variety of configurationally diverse complexes, the generation of new complexes tends toward homogenization, leading to data redundancy. Unfortunately, compared with the explored chemical space for organic molecules, the exploration in transition metal chemistry is somewhat limited, leading to finite diversity for the TMC database. On the other hand, theoretical validation of generated complexes for targeted properties and the curation of reference data for NNPs are highly dependent on DFT computation, however, due to the complex structures of TMCs, the best protocol for DFT computation of transition metal complexes is still being explored.[165] To search for promising complexes from generated candidates, property evaluation via DFT computation is an easily accessible approach. However, it is challenging to pick appropriate DFT methods to assess TMCs for targeted properties due to a lack of widespread agreement on the "best" DFT functional. For example, for only 95 experimentally validated Fe(II) SCO complexes, 30 DFT functionals was unsuccessful in predicting SCO behavior for all of them.[166] Inaccurate assessment from DFT methods certainly leads to a loss of promising complexes, which underestimates the efficiency and effectiveness of these molecular design tools. Moreover, inappropriate DFT methods report poor reference data which misleads ML models to interpret the statistical relations between chemical structure and properties of interest, leading to poor performance of NNPs, *e.g.*, the tmQM-wB97MV data set [156] vs. the tmQM data set.[157] Therefore, the quality of reference data plays a crucial role in using ML methods to predict energy-oriented properties.

Herein, we provides an overview of several publicly available data sets used in the realm of transition metal chemistry (Table 1). All these easily accessible data sets are either directly curated or derived from the CSD. The tmQM data set[157] consists of 86,665 mononuclear, molecular, 3d~5d TMCs extracted from the 2020 release of the CSD, with a limited range of nonmetal elements, {H, B, C, N, O, F, Si, P, As, S, Se, Cl, Br, and I} and a fixed range of charge, {-1, 0, +1}. All reported TMCs were optimized at the GFN2-xTB level, along with some quantum properties computed at the TPSSh-D3BJ/def2-SVP level, such as the electronic and dispersion energies, HOMO/LUMO energies, HOMO/LUMO gap, dipole moment, and charge of the metal center. Recently, Garrison *et al*. revisited this tmQM data set, removed some erroneous complexes and recomputed the energies at the wB97MV level, leading to the reported tmQM-wB97MV data set.[156] From the same version of the CSD database as the tmQM data set, Kulik's group enumerated all 242,829 mononuclear 3d~5d TMCs, with a subset of 85,575 octahedral complexes.[60] From this comprehensive data set, the distribution of TMCs in chemical space were analyzed and visualized. For example, the majority of complexes in the CSD are 3d TMCs of {Fe, Co, Ni, Cu} with coordinating atoms of {C, N, O}. This full set includes 10 coordination geometries based on the coordination number from 4 to 7, along with some nonstandard sandwich complexes. Since octahedral geometries are most common in this reported set, the distribution of the symmetry classes for all octahedral complexes were analyzed and more than 20 types of symmetry classes exist in this subset where the most common type is the geometry with two identical equatorial bidentate ligands and two identical axial monodentate ligands. Because this full data set was curated without any constraint, some erroneous complexes with missing hydrogens or disorder are present in this set, thus data preprocessing, such as re-examination and filtration is recommended to ensure a set of high-quality complexes for downstream tasks.

Vela *et al*.[167] curated a database of 31,019 TMCs from the CSD updated to May 2021 to evaluate the ability of their *cell2mol* software to interpret the crystallographic data, such as the oxidation state of the metal center, the formal charge of each ligand and the molecular connectivity. These 31k complexes include eight transition metals in {Cr, Mn, Fe, Co, Ni, Cu, Ru, Re}, and anomalous complexes with missing H atoms or disorder were discarded. The curated set covers more than 13k distinctive ligands with total charge ranging from -6 to +2. However, all data sets mentioned above ignored the diversity of conformational space in transition metal chemistry. The conformations of ligands in complexes play a pivotal role in determining the physical-chemical properties of TMCs.[155] To alleviate the scarcity of data set by inclusion conformational ensembles, we reported the zinc_60 data set[162] with 39,599 zinc conformers and the Fe(II)_80 dataset[163] which includes over 28,000 conformers of Fe(II) complexes. The former includes neutral Zn conformations with less than 60 atoms with ligands including {H, C, N, O, Zn} atoms, while the latter consists of Fe(II) charged conformers in both the HS and LS states with less than 80 atoms with ligands including {H, C, N, O, P, S, Cl, Fe} atoms.

Recently, several ligand libraries have become publicly available which are valuable resources for the design of novel complexes with tailored properties. For instance, *kraren*[168] provides 331,776 virtual monodentate organophosphorus(III) ligands with their associated conformational space, along with 190 physicochemical descriptor properties. The authors first curated the conformer ensembles of 1558 commercially available ligands with a set of nonmetal elements including {H, B, C, N, O, F, Si, P, S} via CREST. By attaching different substituents to P atom in these 21,437 complexess, over 330,000 virtual ligands were generated. In addition, Chen *et al*. reported their

ReaLigands library which includes more 30,000 monodentate or polydentate ligands by deconstructing experimentally synthesized complexes reported in the CSD.[169] To assign charge for each ligand, the authors leveraged Random Forest models with several quantum-chemical descriptors, such as SCF iteration number, the HOMO-LUMO gap and internal forces calculated at the GFN2-xTB level. The most recently reported tmQMg-L set consists of 30k diverse and synthesizable TMC ligands with defined charges and metal coordination sites based on graph and NBO analyses.[6] The ligands were extracted from a subset of the tmQM data set and were highly diverse with 12 different ligand classes, such as phosphines, chelating amines, olefins, carbenes, etc. Although comparing with the publicly available data sets for organic systems, like QM9,[170] ANI-1,[171] these presented data sets cover limited chemical spaces of transition metals. But they have enumerated all realistic crystal structures reported in the CSD, along with extensive hypothetical ligands, and thus are good resources to expand the TMCs domain.

5. Discussion and Outlook

Despite the structural complexity of transition metal complexes, ranging from varied oxidation, spin states to the interplay between metal and ligands, notable advancements have been achieved in exploring this domain. Traditional physics-based knowledge approaches capably enumerate novel complexes by re-assembling discrete ligands extracted from available databases. However, the lack of widespread agreement on enumeration protocols results in a poor understanding of practical solutions under various conditions. The emergence of ML drives a paradigmatic transformation in this field. The ML-assisted strategies, *e.g*. generative models enrich the diversity of generated complexes with the aim of generating novel ligands from scratch without any prior knowledge. In addition, the ongoing development of MLPs allows for efficient screening of generated candidates for tailored properties. More importantly, the remarkable performance of ML

models makes it accessible to run simulations for large systems, like zinc-metalloproteins,[172] for which a force field or a DFT method could not achieve both efficiency and effectiveness simultaneously. It is acknowledged that ML provides numerous opportunities to explore the unknown domain in chemistry and is becoming increasingly important for the computational discovery of new materials. Despite the unprecedented progress of ML, challenges remain in maximizing the potentials of ML for the automated workflow for transition metal materials design.

ML can circumvent human biases and preconceived notions about molecules and materials, thus reanimating the understanding of researchers about chemistry and motivating them to expand the search space for new molecules and materials. However, the inspiration obtained from these ML models is very limited due to the lack of interpretability which exists in all ML models. Although ML methods provide good solutions to various chemical problems, the inference behind these predictions is not transparent, making it difficult to understand how specific molecular features influence the predicted properties or activities. In addition, in some cases such as reaction pathways, the dynamic transition between different states is crucial, however, ML can only give the final state without any intermediate information provided. This hampers a clear understanding of reaction mechanisms and quantum mechanical methods, to some extent, alleviate this dilemma, but these electronic structure methods do not have the same level of accuracy for transition metals as they do for organic molecules. The insights on TMCs are very limited due to the complexity of structures and large chemical space for TM complexes also impedes the utilization of DFT methods in transition metal chemistry because of the computational cost. The lack of fully validated electronic structure methods also influences the quality of reference data, thus misleading the fitting of ML models. Finally, for practical

applications, only computational validation on designed molecules is not enough because there still exists a fundamental gap between theory and experiments.[166] Whether these computationally validated molecules really have targeted properties is still an open question and need further experimental analysis.[173] The computational design is certainly not the endpoint. Rather, these design tools should be used by experimentalists to explore TMC chemical space to help in precisely locating targets for synthesis and evaluation.

In conclusion, ML can facilitate the automated design of molecules and materials, despite some outstanding challenges. Its ability to quickly analyze massive datasets, identify patterns and make predictions surpasses that of physics-based methods. Automation without any prior knowledge makes it a handy tool to solve a variety of scientific problems. Ongoing research and advancements in ML techniques, combined with efforts to improve data availability and model interpretability, are likely to further enhance the impact of ML in this field.

Acknowledgments

The authors are grateful for the financial support from the National Institutes of Health (Grant Numbers GM130641).

Table 1. Publicly available data sets in transition metal chemistry.

Dataset	Source	Type	Complex or Ligand	Number	Include conformer	Metal	Nonmetal	Charge
tmQM[132]	CSD 2020	experimental	complex	86,665	No	3d~5d	H, B, C, N, O, P, S, Si, As, Se, F, Cl, Br, I	{-1,0,+1}
tmQM-wB97MV[127]	tmQM	experimental	complex	86,507	No	3d~5d	H, B, C, N, O, P, S, Si, As, Se, F, Cl, Br, I	{-1,0,+1}
Nandy et al.[43]	CSD 2020	experimental	complex	242,829	No	3d~5d	H, B, C, N, O, P, S, Si, As, Se, F, Cl, Br, I	-
Cell2mol[167]	CSD 2021	experimental	complex ligand	31019 13819	No	Cr, Mn, Fe, Co, Ni, Cu, Ru, Re	-	[-6, +2]
Zinc_60[162]	tmQM	hypothetic	complex	39599	Yes	Zn	H, C, N, O	0
Fe(II)_80[163]	Nandy et al.[43]	hypothetic	complex	28834	Yes	Fe	H, C, N, O, P, S, Cl	{0,1,2}
Kraren[168]	Literature	hypothetic	ligand	331776	yes	-	H, B, C, N, O, F, Si, P, S	-
ReaLigands[169]	tmQM	experimental	ligand	>30000	no	-	H, B, C, N, O, P, S, Si, As, Se, F, Cl, Br, I	[-3, +1]
tmQMg-L[6]	tmQMg[153]	experimental	ligand	29764	no	-	H, B, C, N, O, P, S, Si, As, Se, F, Cl, Br, I	-

References

1. Gerloch, M.; Constable, E. C. *Transition Metal Chemistry: The Valence Shell in d-Block Chemistry*; Verlagsgesellschaft mbH:Weinheim, 1994.
2. Chen, D.; Xie, Q.; Zhu, J. Unconventional Aromaticity in Organometallics: The Power of Transition Metals. *Acc. Chem. Res.* **2019**, *52*, 1449–1460.
3. Klamm, B. E.; Windorff, C. J.; Celis-Barros, C.; Marsh, M. L.; Meeker, D. S.; Albrecht-Schmitt, T. E. Experimental and Theoretical Comparison of Transition-Metal and Actinide Tetravalent Schiff Base Coordination Complexes. *Inorg. Chem.* **2018**, *57*, 15389–15398.
4. Swart, M.; Gruden, M. Spinning around in Transition-Metal Chemistry. *Acc. Chem. Res.* **2016**, *49*, 2690–2697.
5. Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825–24836.
6. Kneiding, H.; Nova, A.; Balcells, D. Directional Multiobjective Optimization of Metal Complexes at the Billion-System Scale. *Nat. Comput. Sci.* **2024**, *4*, 263.
7. Zunger, A. Inverse Design in Search of Materials with Target Functionalities. *Nat. Rev. Chem.* **2018**, *2* (4).
8. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360.
9. Guo, K.; Yang, Z.; Yu, C.-H.; Buehler, M. J. Artificial Intelligence and Machine Learning in Design of Mechanical Materials. *Mater. Horiz.* **2021**, *8*, 1153.
10. Li, J.; Lim, K.; Yang, H.; Ren, Z.; Raghavan, S.; Chen, P.-Y.; Buonassisi, T.; Wang, X. AI Applications through the Whole Life Cycle of Material Discovery. *Matter* **2020**, *3*, 393.
11. Balabin, R. M.; Lomakina, E. I. Support Vector Machine Regression (LS-SVM)—an Alternative to Artificial Neural Networks (ANNs) for the Analysis of Quantum Chemistry Data? *Phys. Chem. Chem. Phys.* **2011**, *13*, 11710.
12. Graser, J.; Kauwe, S. K.; Sparks, T. D. Machine Learning and Energy Minimization Approaches for Crystal Structure Predictions: A Review and New Horizons. *Chem. Mater.* **2018**, *30*, 3601.
13. Martin, T. B.; Audus, D. J. Emerging Trends in Machine Learning: A Polymer Perspective. *ACS Polym. Au* **2023**, *3*, 239.
14. Dick, S.; Fernandez-Serra, M. Machine Learning Accurate Exchange and Correlation Functionals of the Electronic Density. *Nat. Commun.* **2020**, *11*.
15. Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and Transferable Multitask Prediction of Chemical Properties with an Atoms-in-Molecules Neural Network. *Sci. Adv.* **2019**, *5*.
16. Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; Spicher, S.; Steinbach, P.; Wesołowski, P. A.; Zeller, F. Crest—a Program for the Exploration of Low-Energy Molecular Chemical Space. *J. Chem. Phys.* **2024**, *160*,114110.
17. Duran-Frigola, M.; Mosca, R.; Aloy, P. Structural Systems Pharmacology: The Role of 3D Structures in Next-Generation Drug Development. *Chemistry & Biology* **2013**, *20*, 674–684.

18. Li, X.; Lu, S.; Zhang, G. Three-Dimensional Structured Electrode for Electrocatalytic Organic Wastewater Purification: Design, Mechanism and Role. *Journal of hazardous materials* **2023**, *445*, 130524–130524.Bruice, T. C. Computational Approaches: Reaction Trajectories, Structures, and Atomic Motions. Enzyme Reactions and Proficiency. *Chem. Rev.* **2006**, *106*, 3119–3139.
19. Nad−a Došlić; Goran Kovačević; Ljubić, I. Signature of the Conformational Preferences of Small Peptides: A Theoretical Investigation. *J. Phys. Chem. A* **2007**, *111*, 8650–8658.
20. Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
21. Ishikawa, Y. A Script for Automated 3-Dimentional Structure Generation and Conformer Search from 2-Dimentional Chemical Drawing. *Bioinformation* **2013**, *9*, 988–992.
22. Bochkov, A. Y.; Toukach, P. V. CSDB/SNFG Structure Editor: An Online Glycan Builder with 2D and 3D Structure Visualization. *J. Chem. Inf. Model.* **2021**, *61*, 4940–4948.
23. Dey, F.; Caflisch, A. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.
24. Takeda, S.; Kaneko, H.; Funatsu, K. Chemical-Space-Based de Novo Design Method to Generate Drug-like Molecules. *J. Chem. Inf. Model.* **2016**, *56*, 1885–1893.
25. Burello, E.; Rothenberg, G. In Silico Design in Homogeneous Catalysis Using Descriptor Modelling. *Int. J. Mol. Sci.* **2006**, *7*, 375.
26. Comba, P.; Kerscher, M. Computation of Structures and Properties of Transition Metal Compounds. *Coord. Chem. Rev.* **2009**, *253*, 564.
27. Drummond, M. L.; Sumpter, B. G. Use of Drug Discovery Tools in Rational Organometallic Catalyst Design. *Inorg. Chem.* **2007**, *46*, 8613.
28. Klamm, B. E.; Windorff, C. J.; Celis-Barros, C.; Marsh, M. L.; Meeker, D. S.; Albrecht-Schmitt, T. E. Experimental and Theoretical Comparison of Transition-Metal and Actinide Tetravalent Schiff Base Coordination Complexes. *Inorg. Chem.* **2018**, *57*, 15389–15398.
29. Gruden, M.; Browne, W. R.; Swart, M.; Duboc, C. Computational versus Experimental Spectroscopy for Transition Metals. *Transition Metals in Coordination Environments* **2019**, 161–183.
30. Lin, Z. Interplay between Theory and Experiment: Computational Organometallic and Transition Metal Chemistry. *Acc. Chem. Res.* **2010**, *43*, 602–611.
31. Bonney, K. J.; Schoenebeck, F. Experiment and Computation: A Combined Approach to Study the Reactivity of Palladium Complexes in Oxidation States 0 To IV. *Chem. Soc. Rev.* **2014**, *43*, 6609-6638.
32. Li, J.; Maravelias, C. T.; Van Lehn, R. C. Adaptive Conformer Sampling for Property Prediction Using the Conductor-like Screening Model for Real Solvents. *Ind. Eng. Chem. Res.* **2022**, *61*, 9025.
33. Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95.
34. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578.
35. Das, S.; Shimshi, M.; Raz, K.; Nitoker Eliaz, N.; Mhashal, A. R.; Ansbacher, T.; Major, D. T. EnzyDock: Protein–Ligand Docking of Multiple Reactive States along a Reaction Coordinate in Enzymes. *J. Chem. Theory Comput.* **2019**, *15*, 5116.

36. Schwab, C. H. Conformations and 3D Pharmacophore Searching. *Drug Discov. Today Technol.* **2010**, *7*, e245.
37. Lyne, P. D. Structure-Based Virtual Screening: An Overview. *Drug Discov. Today* **2002**, *7*, 1047.
38. Crawford, J.; Sigman, M. Conformational Dynamics in Asymmetric Catalysis: Is Catalyst Flexibility a Design Element? *Synthesis (Mass.)* **2019**, *51*, 1021.
39. Baber, R. A.; Haddow, M. F.; Middleton, A. J.; Orpen, A. G.; Pringle, P. G.; Haynes, A.; Williams, G. L.; Papp, R. Ligand Stereoelectronic Effects in Complexes of Phospholanes, Phosphinanes, and Phosphepanes and Their Implications for Hydroformylation Catalysis. *Organometallics* **2007**, *26*, 713.
40. Das, S.; Merz, K. M., Jr. Molecular Gas-Phase Conformational Ensembles. *J. Chem. Inf. Model.* **2024**, *64*, 749-760.
41. Hatfield, M.; Lovas, S. Conformational Sampling Techniques. *Curr. Pharm. Des.* **2014**, *20*, 3303-3313.
42. McNutt, A. T.; Bisiriyu, F.; Song, S.; Vyas, A.; Hutchison, G. R.; Koes, D. R. Conformer Generation for Structure-Based Drug Design: How Many and How Good? *J. Chem. Inf. Model.* **2023**, *63*, 6598-6607.
43. Puranen, J. S.; Vainio, M. J.; Johnson, M. S. Accurate Conformation-dependent Molecular Electrostatic Potentials for High-throughput *in Silico* Drug Discovery. *J. Comput. Chem.* **2010**, *31*, 1722–1732.
44. Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.
45. Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know to Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.
46. Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572–584.
47. Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA with Respect to Retrieving Bioactive Conformations. *J. Mol. Graph. Model.* **2003**, *21*, 449–462.
48. Miteva, M. A.; Guyon, F.; Tuffery, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38*, W622–W627.
49. Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: A FRee Online druG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35*, W568–W572.
50. Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.
51. Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All Spd-Block Elements (Z = 1–86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.
52. Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—an Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole

Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
53. Bursch, M.; Hansen, A.; Pracht, P.; Kohn, J. T.; Grimme, S. Theoretical Study on Conformational Energies of Transition Metal Complexes. *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.
54. Sobez, J.-G.; Reiher, M. Molassembler: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 3884–3900.
55. Blaney, J. M.; Dixon, J. S. Distance Geometry in Molecular Modeling. *Reviews in Computational Chemistry*. Wiley January 1994, pp 299–335.
56. Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; John Wiley & Sons, 1988.
57. Taylor, M. G.; Burrill, D. J.; Janssen, J.; Batista, E. R.; Perez, D.; Yang, P. Architector for High-Throughput Cross-Periodic Table 3D Complex Building. *Nat. Commun.* **2023**, *14*.
58. Chernyshov, I. Y.; Pidko, E. A. MACE: Automated Assessment of Stereochemistry of Transition Metal Complexes and Its Applications in Computational Catalysis. *J. Chem. Theory Comput.* **2024**, *20*, 2313–2320.
59. Kishimoto, T.; Yoshikawa, Y.; Yoshikawa, K.; Komeda, S. Different Effects of Cisplatin and Transplatin on the Higher-Order Structure of DNA and Gene Expression. *Int. J. Mol. Sci.* **2019**, *21*, 34.
60. Nandy, A.; Taylor, M. G.; Kulik, H. J. Identifying Underexplored and Untapped Regions in the Chemical Space of Transition Metal Complexes. *J. Phys. Chem. Lett.* **2023**, *14*, 5798–5804.
61. Fedorova, E. V.; Buryakina, A. V.; Zakharov, A. V.; Filimonov, D. A.; Lagunin, A. A.; Poroikov, V. V. Design, Synthesis and Pharmacological Evaluation of Novel Vanadium-Containing Complexes as Antidiabetic Agents. *PLoS One* **2014**, *9*, e100386.
62. Mondal, B.; Neese, F.; Ye, S. Toward Rational Design of 3d Transition Metal Catalysts for CO_2 Hydrogenation Based on Insights into Hydricity-Controlled Rate-Determining Steps. *Inorg. Chem.* **2016**, *55*, 5438–5444.
63. Medlycott, E. A.; Hanan, G. S.; Abedin, T. S. M.; Thompson, L. K. The Effect of Steric Hindrance on the Fe(II) Complexes of Triazine-Containing Ligands. *Polyhedron* **2008**, *27*, 493–501.
64. Gothard, N. A.; Mara, M. W.; Huang, J.; Szarko, J. M.; Rolczynski, B.; Lockard, J. V.; Chen, L. X. Strong Steric Hindrance Effect on Excited State Structural Dynamics of Cu(I) Diimine Complexes. *J. Phys. Chem. A* **2012**, *116*, 1984–1992.
65. Wang, Y.; Han, K. Steric Hindrance Effect of the Equatorial Ligand on Fe(IV)O and Ru(IV)O Complexes: A Density Functional Study. *J. Biol. Inorg. Chem.* **2010**, *15*, 351–359.
66. Fujisawa, K.; Kanda, R.; Miyashita, Y.; Okamoto, K.-I. Copper(II) Complexes with Neutral Bis(Pyrazolyl)Methane Ligands: The Influence of Steric Hindrance on Their Structures and Properties. *Polyhedron* **2008**, *27*, 1432–1446.
67. Kuppuraj, G.; Dudev, M.; Lim, C. Factors Governing Metal−Ligand Distances and Coordination Geometries of Metal Complexes. *J. Phys. Chem. B* **2009**, *113*, 2952–2960.
68. Younus, H. A.; Ahmad, N.; Su, W.; Verpoort, F. Ruthenium Pincer Complexes: Ligand Design and Complex Synthesis. *Coord. Chem. Rev.* **2014**, *276*, 112–152.

69. Matsuoka, W.; Harabuchi, Y.; Maeda, S. Virtual Ligand Strategy in Transition Metal Catalysis toward Highly Efficient Elucidation of Reaction Mechanisms and Computational Catalyst Design. *ACS Catal.* **2023**, *13*, 5697–5711.
70. Hay, B. P.; Firman, T. K. HostDesigner: A Program for the de Novo Structure-Based Design of Molecular Receptors with Binding Sites That Complement Metal Ion Guests. *Inorg. Chem.* **2002**, *41*, 5502–5512.
71. Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress. *J. Chem. Inf. Model.* **2011**, *51*, 760–776.
72. Taylor, R. Life-Science Applications of the Cambridge Structural Database. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58*, 879–888.
73. Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for *de Novo* Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134*, 8885–8895.
74. Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107–6132.
75. Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087.
76. Clark, D. E.; Westhead, D. R. Evolutionary Algorithms in Computer-Aided Molecular Design. *J. Comput. Aided Mol. Des.* **1996**, *10*, 337–358.
77. Chakraborti, N. Genetic Algorithms in Materials Design and Processing. *Int. Mater. Rev.* **2004**, *49*, 246–260.
78. Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-like Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545–552.
79. Foscato, M.; Venkatraman, V.; Jensen, V. R. DENOPTIM: Software for Computational *de Novo* Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59*, 4077–4082.
80. Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 767–780.
81. Turcani, L.; Tarzia, A.; Szczypiński, F. T.; Jelfs, K. E. *Stk*: An Extendable Python Framework for Automated Molecular and Supramolecular Structure Assembly and Discovery. *J. Chem. Phys.* **2021**, *154*.
82. Henle, E. A.; Gantzler, N.; Thallapally, P. K.; Fern, X. Z.; Simon, C. M. PoreMatMod.Jl: Julia Package for *in Silico* Postsynthetic Modification of Crystal Structure Models. *J. Chem. Inf. Model.* **2022**, *62*, 423–432.
83. Laplaza, R.; Gallarati, S.; Corminboeuf, C. Genetic Optimization of Homogeneous Catalysts. *Chemistry Methods* **2022**, *2*.
84. Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. MolSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37*, 2106– 2117.
85. Kalikadien, A. V.; Pidko, E. A.; Sinha, V. *ChemSpaX*: Exploration of Chemical Space by Automated Functionalization of Molecular Scaffold. *Digit. Discov.* **2022**, *1*, 8–25.

86. van der Zant, T.; Kouw, M.; Schomaker, L. Generative Artificial Intelligence. In *Studies in Applied Philosophy, Epistemology and Rational Ethics*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2013; pp 107–120.
87. Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative Models for Automatic Chemical Design. In *Machine Learning Meets Quantum Physics*; Springer International Publishing: Cham, 2020; pp 445–467.
88. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360.
89. Kingma, D.P. and Welling, M. Auto-Encoding Variational Bayes. arXiv Preprint arXiv:1312.6114. https://arxiv.org/abs/1312.6114, 2013.
90. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139.
91. Gui, J.; Sun, Z.; Wen, Y.; Tao, D.; Ye, J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3313.
92. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. arXiv Preprint arXiv:2006.11239, https://arxiv.org/abs/2006.11239, 2020.
93. Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. arXiv Preprint arXiv:1503.03585, https://arxiv.org/abs/1503.03585, 2015.
94. Song, Y.; Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. arXiv Preprint arXiv:1907.05600, https://arxiv.org/abs/1907.05600, 2019.
95. Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. arXiv Preprint arXiv:2011.13456, https://arxiv.org/abs/2011.13456, 2020.
96. Gm, H.; Gourisaria, M. K.; Pandey, M.; Rautaray, S. S. A Comprehensive Survey and Analysis of Generative Models in Machine Learning. *Comput. Sci. Rev.* **2020**, *38*, 100285.
97. Kumar, S.; Musharaf, D.; Musharaf, S.; Sagar, A. K. A Comprehensive Review of the Latest Advancements in Large Generative AI Models. In *Communications in Computer and Information Science*; Springer Nature Switzerland: Cham, 2023; pp 90–103.
98. Sengar, S.S., Hasan, A.B., Kumar, S. and Carroll, F., 2024. Generative Artificial Intelligence: A Systematic Review and Applications. arXiv Preprint arXiv:2405.11029, https://arxiv.org/abs/2405.11029, 2024.
99. Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; Wei, L. Deep Generative Models in *DE Novo* Drug Molecule Generation. *J. Chem. Inf. Model.* **2024**, *64*, 2174.
100. Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative Models for DE Novo Drug Design. *J. Med. Chem.* **2021**, *64*, 14011.
101. Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. In *Readings in Cognitive Science*; Elsevier, 1988; pp 399–421.
102. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders. In *Machine Learning for Data Science Handbook*; Springer International Publishing: Cham, 2023; pp 353–374.
103. Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. Designing Catalysts with Deep Generative Models and Computational Data. A Case Study for Suzuki Cross Coupling Reactions. *Digit. Discov.* **2023**, *2*, 728.

104.Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024.
105.Strandgaard, M.; Linjordet, T.; Kneiding, H.; Burnage, A.; Nova, A.; Jensen, J. H.; Balcells, D. Deep Generative Model for the Dual-Objective Inverse Design of Metal Complexes. *ChemRxiv*, 2024.
106.Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. arXiv Preprint arXiv:1802.04364, https://arxiv.org/abs/1802.04364, 2018.
107.Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370.
108.Court, C. J.; Yildirim, B.; Jain, A.; Cole, J. M. 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4518.
109.Ratliff, L. J.; Burden, S. A.; Sastry, S. S. Characterization and Computation of Local Nash Equilibria in Continuous Games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*; IEEE, pp.917-924, 2013.
110.Nouira, A., Sokolovska, N. and Crivello, J.C. Crystalgan: Learning to Discover Crystallographic Structures with Generative Adversarial Networks. arXiv Preprint arXiv:1810.11203. https://arxiv.org/abs/1810.11203, 2018.
111.Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* **2020**, *6*, 1412.
112.Arjovsky, M., Chintala, S. and Bottou, L. Wasserstein GAN. arXiv Preprint arXiv:1701.07875. https://arxiv.org/abs/1701.07875, 2017.
113.Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative Adversarial Networks (GAN) Based Efficient Sampling of Chemical Composition Space for Inverse Design of Inorganic Materials. *Npj Comput. Mater.* **2020**, *6*.
114.Mao, Y.; He, Q.; Zhao, X. Designing Complex Architectured Materials with Generative Adversarial Networks. *Sci. Adv.* **2020**, *6*, 4169.
115.Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U. and Sutton, C. VEEGAN: Reducing Mode Collapse in Gans Using Implicit Variational Learning. arXiv Preprint arXiv:1705.07761, https://arxiv.org/abs/1705.07761, 2017.
116.Bau, D.; Zhu, J.-Y.; Wulff, J.; Peebles, W.; Zhou, B.; Strobelt, H.; Torralba, A. Seeing What a GAN Cannot Generate. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE, 2019.
117.Kodali, N., Abernethy, J., Hays, J. and Kira, Z., 2017. On Convergence and Stability of Gans. arXiv Preprint arXiv:1705.07215, https://arxiv.org/abs/1705.07215, 2017.
118.Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z. and Paul Smolley, S. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, pp.2794-2802, 2017.
119.Ma, T. Generalization and Equilibrium in Generative Adversarial Nets (GANs) (Invited Talk). In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*; ACM: New York, NY, USA, 2018.
120.Nagarajan, V., Raffel, C. and Goodfellow, I.J. Theoretical Insights into Memorization in GANs. *In Neural Information Processing Systems Workshop (Vol. 1, p. 3)*, 2018.

121. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.-H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* **2024**, *56*, 1.
122. Jin, H.; Merz, K. M., Jr. LigandDiff: De Novo Ligand Design for 3D Transition Metal Complexes with Diffusion Models. *J. Chem. Theory Comput.* **2024**, *20* (10), 4377-4384.
123. Jin, H.; Merz, K. M. Partial to Total Generation of 3D Transition Metal Complexes. *J. Chem. Theory Comput.* **2024**, *20,*8367-8377.
124. Clough, T. J.; Jiang, L.; Wong, K.-L.; Long, N. J. Ligand Design Strategies to Increase Stability of Gadolinium-Based Magnetic Resonance Imaging Contrast Agents. *Nat. Commun.* **2019**, *10*, 1420.
125. Toporivska, Y.; Mular, A.; Piasta, K.; Ostrowska, M.; Illuminati, D.; Baldi, A.; Albanese, V.; Pacifico, S.; Fritsky, I. O.; Remelli, M.; et al. Thermodynamic Stability and Speciation of Ga(III) and Zr(IV) Complexes with High-Denticity Hydroxamate Chelators. *Inorg. Chem.* **2021**, *60*, 13332-13347.
126. Preston, D.; Kruger, P. E. Using Complementary Ligand Denticity to Direct Metallosupramolecular Structure about Metal Ions with Square-planar Geometry. *ChemPlusChem* **2020**, *85*, 454-465.
127. Meagley, K. L.; Garcia, S. P. Chemical Control of Crystal Growth with Multidentate Carboxylate Ligands: Effect of Ligand Denticity on Zinc Oxide Crystal Shape. *Cryst. Growth Des.* **2012**, *12*, 707-713.
128. Deka, H.; Ghosh, S.; Saha, S.; Gogoi, K.; Mondal, B. Effect of Ligand Denticity on the Nitric Oxide Reactivity of Cobalt(II) Complexes. *Dalton Trans.* **2016**, *45*, 10979-10988.
129. Smits, N. W. G.; van Dijk, B.; de Bruin, I.; Groeneveld, S. L. T.; Siegler, M. A.; Hetterscheid, D. G. H. Influence of Ligand Denticity and Flexibility on the Molecular Copper Mediated Oxygen Reduction Reaction. *Inorg. Chem.* **2020**, *59*, 16398-16409.
130. Cornet, F.; Benediktsson, B.; Hastrup, B.; Schmidt, M. N.; Bhowmik, A. Om-Diff: Inverse-Design of Organometallic Catalysts with Guided Equivariant Denoising Diffusion. *ChemRxiv*, 2024.
131. Xie, T.; Fu, X.; Ganea, O.-E.; Barzilay, R.; Jaakkola, T. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. arXiv Preprint arXiv:2110.06197, https://arxiv.org/abs/2110.06197, 2021.
132. Han, S.; Lee, J.; Han, S.; Moosavi, S. M.; Kim, J.; Park, C. Design of New Inorganic Crystals with the Desired Composition Using Deep Learning. *J. Chem. Inf. Model.* **2023**, *63*, 5755-5763.
133. Alverson, M.; Baird, S. G.; Murdock, R.; Ho, (Enoch) Sin-Hang; Johnson, J.; Sparks, T. D. Generative Adversarial Networks and Diffusion Models in Material Discovery. *Digit. Discov.* **2024**, *3*, 62-80.
134. Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning. *Science* **2019**, *365*.
135. Iribarren, I.; Garcia, M. R.; Trujillo, C. Catalyst Design within Asymmetric Organocatalysis. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*.
136. Koner, D.; Unke, O. T.; Boe, K.; Bemish, R. J.; Meuwly, M. Exhaustive State-to-State Cross Sections for Reactive Molecular Collisions from Importance Sampling Simulation and a Neural Network Representation. *J. Chem. Phys.* **2019**, *150*.

137.Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*.
138.Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*.
139.Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*.
140.Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions arXiv Preprint arXiv:1706.08566, https://arxiv.org/abs/1706.08566, 2017.
141.Schütt, K. T.; Unke, O. T.; Gastegger, M. Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra. arXiv Preprint arXiv:2102.03150, https://arxiv.org/abs/2102.03150, 2021.
142.Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate*Ab Initio*Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*.
143.Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A Generally Applicable Atomic-Charge Dependent London Dispersion Correction. *J. Chem. Phys.* **2019**, *150*.
144.Kosmala, A., Gasteiger, J., Gao, N. and Günnemann, S. Ewald-Based Long-Range Message Passing for Molecular Graphs. arXiv Preprint arXiv:2304.04791, https://arxiv.org/abs/2303.04791, 2023.
145.Wells, B. A.; Chaffee, A. L. Ewald Summation for Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 3684.
146.Frank, J. T.; Unke, O. T.; Müller, K.-R. So3krates: Equivariant Attention for Interactions on Arbitrary Length-Scales in Molecular Systems. arXiv Preprint arXiv:2205.14276, https://arxiv.org/abs/2205.14276, 2022.
147.Janet, J. P.; Kulik, H. J. Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8* (7), 5137-5152.
148.Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069-7077.
149.Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11* (18), 4584-4601.
150.Cordova, M.; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C. Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* **2020**, *10* (13), 7021-7031.
151.Xu, L.-C.; Zhang, S.-Q.; Li, X.; Tang, M.-J.; Xie, P.-P.; Hong, X. Towards Data-driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed Engl.* **2021**, *60* (42), 22804-22811.
152.Xu, L.-C.; Frey, J.; Hou, X.; Zhang, S.-Q.; Li, Y.-Y.; Oliveira, J. C. A.; Li, S.-W.; Ackermann, L.; Hong, X. Enantioselectivity Prediction of Pallada-Electrocatalysed C–H Activation Using Transition State Knowledge in Machine Learning. *Nat. Synth.* **2023**, *2*, 321-330.

153. Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; De Bin, R.; Balcells, D. Deep Learning Metal Complex Properties with Natural Quantum Graphs. *Digit. Discov.* **2023**, *2*, 618-633.
154. Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561-6594.
155. Baidun, M. S.; Kalikadien, A. V.; Lefort, L.; Pidko, E. A. Impact of Model Selection and Conformational Effects on the Descriptors for in Silico Screening Campaigns: A Case Study of Rh-Catalyzed Acrylate Hydrogenation. *J. Phys. Chem. C Nanomater. Interfaces* **2024**, *128*, 7987-7998.
156. Garrison, A. G.; Heras-Domingo, J.; Kitchin, J. R.; dos Passos Gomes, G.; Ulissi, Z. W.; Blau, S. M. Applying Large Graph Neural Networks to Predict Transition Metal Complex Energies Using the tmQM_wB97MV Data Set. *J. Chem. Inf. Model.* **2023**, *63*, 7642-7654.
157. Balcells, D.; Skjelstad, B. B. TmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135-6146.
158. Shuaibi, M.; Kolluru, A.; Das, A.; Grover, A.; Sriram, A.; Ulissi, Z.; Zitnick, C. L. Rotation Invariant Graph Neural Networks Using Spin Convolutions. arXiv Preprint arXiv:2106.09575, https://arxiv.org/abs/2106.09575, 2021.
159. Gasteiger, J.; Becker, F.; Günnemann, S. GemNet: Universal Directional Graph Neural Networks for Molecules. arXiv Preprint arXiv:2106.08903, https://arxiv.org/abs/2106.08903, 2021.
160. Roy Chowdhury, S.; Nguyen, N.; Vlaisavljevich, B. Importance of Dispersion in the Molecular Geometries of Mn(III) Spin-Crossover Complexes. *J. Phys. Chem. A* **2023**, *127*, 3072-3081.
161. Dixon, I. M.; Khan, S.; Alary, F.; Boggio-Pasqua, M.; Heully, J.-L. Probing the Photophysical Capability of Mono and Bis(Cyclometallated) Fe(Ii) Polypyridine Complexes Using Inexpensive Ground State DFT. *Dalton Trans.* **2014**, *43*, 15898-15905.
162. Jin, H.; Merz, K. M., Jr. Modeling Zinc Complexes Using Neural Networks. *J. Chem. Inf. Model.* **2024**, *64*, 3140-3148.
163. Jin, H.; Merz, K. M., Jr. Modeling Fe(II) Complexes Using Neural Networks. *J. Chem. Theory Comput.* **2024**, *20*, 2551-2558.
164. Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss Army Knife" Composite Electronic-Structure Method. *J. Chem. Phys.* **2021**, *154*.
165. Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem. Weinheim Bergstr. Ger.* **2022**, *134*.
166. Vennelakanti, V.; Taylor, M. G.; Nandy, A.; Duan, C.; Kulik, H. J. Assessing the Performance of Approximate Density Functional Theory on 95 Experimentally Characterized Fe(II) Spin Crossover Complexes. *J. Chem. Phys.* **2023**, *159*.
167. Vela, S.; Laplaza, R.; Cho, Y.; Corminboeuf, C. Cell2mol: Encoding Chemistry to Interpret Crystallographic Data. *Npj Comput. Mater.* **2022**, *8*.
168. Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzi, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

169. Chen, S.-S.; Meyer, Z.; Jensen, B.; Kraus, A.; Lambert, A.; Ess, D. H. ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design. *J. Chem. Inf. Model.* **2023**, *63*, 7412–7422.
170. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*.
171. Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A Data Set of 20 Million Calculated off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*.
172. Ding, Y.; Huang, J. DP/MM: A Hybrid Model for Zinc–Protein Interactions in Molecular Dynamics. *J. Phys. Chem. Lett.* **2024**, *15*, 616–627.
173. Karl, T. M.; Bouayad-Gervais, S.; Hueffel, J. A.; Sperger, T.; Wellig, S.; Kaldas, S. J.; Dabranskaya, U.; Ward, J. S.; Rissanen, K.; Tizzard, G. J.; Schoenebeck, F. Machine Learning-Guided Development of Trialkylphosphine $Ni^{(I)}$ Dimers and Applications in Site-Selective Catalysis. *J. Am. Chem. Soc.* **2023**, *145*, 15414-15424.