MACHINE LEARNING FOR TRANSITION METAL COMPLEXES

By

Hongni Jin

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry – Doctor of Philosophy

2024

ABSTRACT

Transition metal complexes, dubbed ‘Lego molecules’, are composed of small molecules, ions, or atoms arranged around a central metal. The diversified research field of organometallic compounds includes but is not limited to the study of metal-ligand interactions, structure-property relationships, and practical applications. This dissertation leverages machine learning techniques to expedite the research in this domain. The first part focuses on neural network potentials (NNPs). A Zn_NNPs model was built to depict the potential energy surface of zinc complexes. In this work, a simple but useful embedding of partial charges was proposed, which could model the long-range interactions accurately. Furthermore, an Fe_NNPs model was designed to identify the lowest energy spin state of Fe (II) complexes. The model integrates electronic characteristics such as total charge and spin state to account for long-range interactions effectively. For each model, a high-quality data set including tens of thousands of distinctive conformations was well curated using metadynamics. The third model is a scaffold-based diffusion model, called LigandDiff which can generate valid, novel, and unique ligands for organometallic compounds. Users only need to specify the desired size of the ligand, LigandDiff then generates a diverse and potentially infinite number of ligands of that size from scratch. Collectively, these models surpass traditional computational methods on both accuracy and efficiency, demonstrating substantial potential to accelerate transition metal complexes research.

This dissertation is dedicated to my elder brother, Jay Jin, for his unconditional support, encouragement, and belief in me.

ACKNOWLEDGEMENTS

Thank my families for their support and love. As the youngest child of my generation in our large family, I have always received extra care and attention from my family members. Thank my parents for their continuous encouragement and trust which have given me the confidence to explore the unlimited possibilities of my life. And I also want to express my special gratitude to my elder brother, who sets an excellent model for me to follow. He is truly my personalized ChatGPT, always providing me valuable and timely advice about my life.

I am grateful to my advisor, Professor Kenneth M. Merz Jr. for his dedication and support. As an educator, he has given me trust and the freedom to pursue my interests. As a researcher, his extensive knowledge has answered many of my questions. As an editor, he has taught me to express my ideas clearly and succinctly in scientific writing. And I thank my committee members (Professor Angela K. Wilson, Professor Gary J. Blanchard, and Professor Guo-Wei Wei) for their insightful feedback on my research, especially their valuable experience on scientific writing. I also appreciate all members of the Merz group for their useful advice on my work.

Last but not least, thank all my friends throughout my life for their companionship. Their presence brightens my life.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

TMCs	Transition metal complexes
CN	Coordination number
AI	Artificial intelligence
ML	Machine learning
MD	Molecular dynamics
PES	Potential energy surface
MLPs	Machine learning potentials
NNPs	Neural network potentials
RMSE	Root mean square error
GNNs	Graph neural networks
CSD	Cambridge Structural Database
MAE	Mean absolute error
HS	High spin
LS	Low spin
SCO	Spin crossover
SE	Splitting energy
DDPM	Denoising Diffusion Probabilistic Model

CHAPTER 1 INTRODUCTION

1.1 General Introduction of Transition Metal Complexes

Transition metals are chemical elements which mainly lie in the *d* block of the periodic table. They include 3d elements from Sc to Zn, 4d elements from Y to Cd, and 5d elements from Hf to Au. The existence of transition metals on earth varies a lot, from ubiquitous to rare; 3d Fe ranks as the fourth most abundant element in the Earth's crust while 4d Tc is only artificially produced.[1] Most transition metals display slivery color, but copper and gold are usually in slightly red, and mercury is the only one which is in liquid at ambient temperatures.

The research of these transition metals started in the nineteenth century and quickly drew people's attention since the compounds of these transition metals exhibit different properties than typical covalent organic compounds. First, the partially filled *d* orbitals can accommodate electrons from outside to form dative covalent interactions. This bonding pattern is usually called coordination, which is a representation of Lewis acid-base interaction. Such coordination compounds are interchangeably called complexes, where the transition metal is the electron acceptor, while the electron donor is known as a ligand. The coordination sphere typically contains the central transition metal atom or ion along with the ligands that are bonded to it. The type of ligand is very flexible, and the sole criterion for ligands is that they must possess at least one pair of electrons to donate. Therefore, the ligands cover a diverse range of chemicals, either atoms, molecules, or ions. Some common ligands are water, ammonia, and chloride. The coordination number (CN) is the count of ligand attachment sites surrounding a transition metal center in a complex, ranging from 1 to 16[2,3] which depends on both the central metal and the ligands. Some metals prefer certain coordination numbers since these coordination numbers may stabilize electron energies. Moreover, the nature of the ligands also plays a role. The shape and the size of ligands greatly determine the coordination number. Larger ligands typically give rise to a reduced coordination number. Overall, this diverse coordination range further enriches the chemical space of TMCs.

Second, a characteristic feature of transition metals is their ability to exist in various oxidation states, because they can lose electrons easily in contrast to the alkali metals and alkaline earth metals. Alkali metals, with a single electron in their *s* orbitals, typically exhibit a +1oxidation state. Similarly, alkaline earth metals, which have two electrons in the *s* orbitals, almost invariably show +2 oxidation state. In contrast, transition metals are more complex because their oxidation states are determined by both the charge of ligands as well as the overall charge of the complex. For instance, most transition metals in *3d* block have oxidation states of +2 or +3, like Co^{2+}/ Co^{3+}, Ni^{2+}/ Ni^{3+}, Fe^{2+}/Fe^{3+}, but the element Mn has more choices, ranging from +2 to +7. The flexible oxidation states give TMCs intriguing redox properties, resulting in a myriad of potential applications that will be discussed later.

In addition, the spin state of transition metals can vary. The spin state describes the spin configurations of the central metal's *d* electrons. It manifests as the unpaired electrons of the metal. Usually, a complex can be categorized as either high spin or low spin, sometimes an intermediate spin state is also possible. The spin state is determined by the energy gap between the crystal field splitting energy (Δ) and the pairing energy (P). When Δ exceeds P, electrons occupy all the lower energy orbitals and pair up within them prior to climbing to the higher energy orbitals. In this case, it is energetically more advantageous for electrons to pair and occupy all of the low energy orbitals. Conversely, when the pairing energy exceeds the crystal field energy, electrons will first fill all available orbitals singly before any pairing occurs, regardless of the orbitals' energy levels. Many properties of TMCs are highly related to the spin state, such as the magnetic properties and the spin-crossover properties. The overall unique properties of TMCs can be briefly summarized below:

1. Various charge states. The transition metals are usually cationic in aqueous solution. But the charge is manipulated by the coordination environment so that the whole complex can be either cationic, neutral, or anionic.
2. Unique interactions with ligands. The dative covalent bonding between ligands and transition metal center is selective and specific. The ligands can tune the electron configurations of the *d* orbitals of the metal, thereby forming unique overall property that the individual metal or ligands do not own.
3. Diverse coordination geometries. The flexible choice of CN provides a wide variety of coordination shapes different from the organic molecules.

> Furthermore, isomerism exists extensively in coordination compounds which further enriches the diversity of geometries. For instance, one type of isomerism is attributed to the relative arrangement of the ligands. A planar complex $M(L_1)_2(L_2)_2$, where M is the metal, L_1 and L_2 are two different ligands, can be in either *cis* or *trans* form, depending on the relative position of the same type of ligand.

TMCs have extensive applications in many fields.[4-20] In biology, optical imaging is an important tool in life sciences for disease diagnostics. Many luminescent TMCs have shown promising applications in bioimaging and biosensing due to their remarkable photophysical properties.[21,22] For instance, oxygen deprivation in biological systems can cause various diseases such as fatty liver, cerebral infarction, diabetic retinopathy, and cancer.[23-25] Investigating the mechanism of diseases related to oxygen deprivation requires O_2 imaging technology that is capable of detecting O_2 in real time with high selectivity and high stability. Studies have shown that a wide range of TMCs including Pd (II) and Pt (II) porphyrins, Ir (III) complexes, Ru (II) complexes have strong phosphorescence within the visible to near-infrared spectrum, with notably extended lifetimes exceeding 1.0 microseconds and they have been successfully used to monitor O_2 level in cell nucleus,[26] PC12 cells,[27] bone marrow,[28] *etc*. Transition metals also play important catalytic and structural roles in biology. For instance, iron is an essential ion for the process of the respiration and electron transport in biological systems.[29] Fe is bound to a porphyrin, and the whole complex, called a heme, is the major oxygen carrier in blood. Small ligand like CO can easily bind with the heme so that the cellular uptake of oxygen is blocked which deactivates oxygen transport, leading to the death of the organisms. Such process is called carbon monoxide poisoning.[30] In addition to respiration, Fe actively interacts with many enzymes to support the redox processes,[31] electron transfer reactions,[32] and even nitrogen fixation in plants.[33] Another important application of TMCs is metallodrugs. Barnett Rosenberg and colleagues at Michigan State University fortuitously unveiled that cisplatin has anticancer effects, which initiated metallodrug research.[34] Currently, the platinum-based metallodrug family including carboplatin and oxaliplatin has been extensively used to treat ovarian, breast, colon, testicular and prostate cancer.[35] Some other TMCs have also shown intriguing properties as promising metallodrug candidates. The Ru-based complexes exhibit high affinity for DNA and can reversely bind to the double helix.[36] They are expected to be potential candidates as antineoplastic drugs since they can tune the tumor cell cycle and cause apoptosis.[37] A couple of complexes such as KP1019, KP1339 and NAMI-A are being evaluated in clinical trials.[38] Copper(II) complexes have emerged as an promising chemotype against cancer since they are capable of generating reactive nitrogen species and reactive oxygen species, resulting in cellular death.[39] A typical family example of Copper metallodrugs is Casiopeínas, which have phenanthroline or bipyridine bidentate ligands in the coordination sphere, and the second charged ligand is either O-O or N-O coordinated, such as acetylacetonate, salicylaldehyde and aminoacidate. These compounds are able to increase endonuclease G and activate caspase 3, thereby leading to apoptosis.[40] Recently other new copper complexes similar to Casiopeínas have also been successfully synthesized. And the results indicate that these compounds have remarkable cytotoxic and antiproliferative bioactivities in multiple cancer cells, like breast cancer, melanoma cancer, osteosarcoma cancer, cervical cancer, colon cancer and ovarian carcinoma.[41-43] Overall, the mechanisms of metallodrugs against diseases include inhibition of angiogenesis, induction of cell apoptosis, alteration of the cytoskeleton, *etc*.[44] And they open novel pathways for treating diseases that traditional organic medications cannot address due to issues with drug resistance.[45] But currently, the comprehensive functions of TMCs in disease diagnostics and treatments are still unclear and more research is greatly needed.

TMCs have also been widely used in material science. One important application is their great contributions as organometallic catalysts to chemical synthesis. The properties of transition metal catalysts are determined, to large degree, by the ligands coordinated to the metal. The electronic properties, the size of the ligand and the ligand bite angle are a couple of important parameters to consider for the design of new TMCs as catalysts.[46,47] For instance, the electrocatalytic CO_2 conversion offers tremendous potential to use carbon-free renewable energy to meet green chemistry.[48] The high energy barrier for electrochemical CO_2 activation is related to the high energy requirement for breaking the linear neural molecule into the bent radical anion, leading to a high overpotential demand for the one electron reduction from CO_2 to $CO_2^{\cdot -}$. But once the transition metal catalyst is introduced to the system, this critical potential can be met by the reduction potential of the catalyst, thus making the conversion.[49] The whole process starts from the coordination of CO_2 to the central transition metal so that the electron can transfer to CO_2. The selection of ligands in TMCs is of great importance in this step. The suitable ligands should be flexible electron carriers that can accept or donate electrons via redox reactions. Examples include pyridines and imines. Currently, TMCs including elements Mo, W, Mn, Rh, Re, Cr, Fe, Ru, Ni, Pd, Pt, Cu, Zn, Os, Ir, have been successfully synthesized as catalysts for electrochemical conversion of CO_2.[50]

One more example of TMCs in material science is the utilization of TMCs as photosensitizers. Photosensitization is a process where the energy transfers from the photosensitizer to a substrate so that the substrate can be activated to undergo further chemical transformations.[51] The generation of singlet oxygen is a typical example of photosensitization reaction.[52,53] The reaction is initiated by the excitation of the photosensitizer via the absorption of a single photon, resulting in a high energy singlet state. The photosensitizer then proceeds to convey its energy to the ground state molecular oxygen ($^3O_2$), undergoing an internal intersystem crossing to reach an excited triplet state. And the transferred energy allows the production of metastable excited state singlet oxygen ($^1O_2$). TMCs are ideal candidates for photochemical applications due to their intense absorption in the visible light spectrum. The effectiveness of a photosensitizer depends on the presence of readily available, low-energy valence exited states. For TMCs, it usually corresponds to the charge transfer states between metal and ligand, either from metal to ligand or ligand to metal. For instance, Ru (II) complexes are most widely used photosensitizers because their low-lying valence excited states are predominantly long-lived metal-to-ligand charge transfer states.[54] And considering that Ru (II) is extremely rare on earth, recently other transition metals, especially the first-row TMCs have been largely explored. The relative inaccessibility of MLCT excited states may limit the viability of these complexes to activate photosensitization, but well-designed novel complexes show a promising balance between cost, abundance and efficiency.[55]

1.2 Machine Learning

Artificial intelligence (AI) refers to a computer system’s capacity to emulate human cognitive functions, including learning and problem-solving with machine learning (ML) representing a subset of AI applications. It is a process where mathematical models are well designed to make predictions via learning implicitly from available data. The beauty of ML is that the learning process is driven by the model itself without direct human intervention. The learning process is evaluated by the loss function of a ML model and the overall goal is to reduce the error between the true values and the predicted values as much as possible.

Depending on the type of used data, machine learning mainly includes supervised learning and unsupervised learning. The former uses labeled training data, thereby having a baseline understanding of what the correct output should be. By contrast, the latter uses unlabeled data, which means the model learns independently to understand the inherent structure of the given data without any specific guidance. While the type of data is a distinctive difference between both models, they also have different goals and applications which set them apart from each other. Supervised learning is used to investigate the underlying relation between input and output while unsupervised learning is more focused on discovering new patterns and relationships in raw, unlabeled data. For instance, a supervised model might be designed to predict the flight times under the conditions of weather, airport traffic, peak flight hours, *etc.* But an unsupervised model might be used to automatically categorize some unlabeled images. In this case, the data is not labeled so the model does not know the object in the image, but the model can discern the common characteristics of the same type of images, thus classifying the images correctly. Traditional machine learning requires feature engineering with human intervention. Features related to the target property need to be carefully determined and extracted and then fed into the model. They usually have simple framework and thus are easy to interpret. Examples of traditional machine learning model include linear regressions, logistic regressions, support vector machines, decision trees, random forests, *etc*. However, feature engineering is a meticulous and time-intensive task, requiring experts' knowledge to pinpoint the pertinent features for the model. In addition, due to their simple and fixed structures, they are usually unable to model complex and high-dimensional data, thus limiting their application domains. To circumvent the limitations of traditional machine models, deep learning is used. Deep learning uses neural networks to process raw data without any feature engineering which eliminates the human intervention, thus allowing the use of large amounts of data. Artificial neural networks (ANNs), or neural networks are proposed to imitate the mechanism which human brain operates. The human brain consists of millions of neurons, all interconnected and they are sending electric signals back and forth to each other to help human process information and make decisions. And neural networks were first designed by the inspiration of these biological neurons dating back to 1943. A typical neural network architecture consists of one input layer, a minimum of one hidden layer as well as one output layer. And every layer has multiple nodes, i.e. the neurons in the biological systems. Usually, nodes are fully connected between two consecutive layers. And the connection strength is determined by the weights. The input layer processes the data input from outside, analyzes it and then passes the data to the hidden layers. The hidden layers further transform the data via nonlinear functions and pass the transformed data to the output layer. Subsequently, the output gives the final result of all the data processed by the neural network. The transfer of information from one layer to the subsequent one in a neural network is termed as feedforward propagation, which is usually used for training. Once the error between the predictions and the true values is determined, backpropagation is activated to minimize the error by fine-tuning the neural network’s weights and biases. The backpropagation uses gradient descent algorithm to propagate the error from the output back to the input layer. The gradient descent algorithm computes the gradient of the error function in relation to the model’s parameters, such as weights and biases, and then updates the weights and biases in the direction of the negative gradient to reduce the error. In essence, feedforward and backpropagation work together in a neural network's training phase: feedforward makes predictions, backpropagation assesses and corrects them, and this cycle repeats until the network optimally learns to map inputs to the correct output.

Machine learning in chemistry is not new. The earliest application of data science techniques to chemistry research is the determination of molecular formula from low resolution mass spectrometry in 1969.[56] Machine learning has wide-ranging applications in chemistry: For instance, traditional ML methods have made contributions to quantitative structure activity relationship (QSAR) applications.[57,58] Usually, a set of molecular descriptors which are precomputed molecular physicochemical properties are fed into a traditional ML model, such as linear regression model, random forests, support vector machines, *etc*.. With such a well-trained model, the target property of new, unseen molecules can be quickly predicted. In addition, using ML to accelerate traditional quantum mechanical (QM) calculations has been emerging in the last few years. Having a deep understanding of the electronic structure of chemical systems is crucial for the design of molecules and materials. The most accurate way to decode the chemical structures is to solve the Schrödinger equation used to calculate the wave function of a given system. However, this equation can be solved exactly only for the single electron system, *e.g.*, the hydrogen atom but not for multi-electron systems, such as the Helium atom. For larger molecules, carefully chosen approximations are needed at the cost of losing accuracy as little as possible.[59] Machine-learning potentials (MLPs) with reference to high-level quantum chemistry methods have gained increasing attention in computational chemistry.[60-63] Well-built MLPs can reach as high accuracy as their reference method but at much lower computational cost. Since the

MLPs are built based on non-linear functions which do not require any physical knowledge, such methods are very flexible and can be used for almost any system, such as a molecule,[64,65] nanoporous materials,[66] oxides,[67] and metals.[68,69] And one more example of ML for chemistry is computational material design. The design of novel structures with desirable properties is a core part of chemistry. In the early years, materials discovery was primarily driven by serendipity.[70] Compared to traditional high-throughput screening methods, deep generative models can generate rational molecules at a reduced cost since little human intervention is required and much time can be saved. Once a generative model has been well designed, unlimited and diverse new molecules can be automatically generated within seconds. Currently, several generative models have been widely used for molecular generation, including variational autoencoder (VAE), convolutional neural network (CNN), Transformer-based models, recurrent neural network (RNN), flow-based models, generative adversarial network (GAN), and diffusion models.[71]

The powerful learning ability of ML has revolutionized many fields in modern society, and we have already witnessed great progress in chemistry that interacts with ML. Currently, AI or specifically ML, is still quickly developing, and powerful large language models (LLMs), such as ChatGPT, demonstrate potential for advancing human development. However, how these tools can benefit chemists thus accelerating chemistry research, has been underexplored to date.

CHAPTER 2 MODELING ZINC COMPLEXES USING NEURAL NETWORKS

2.1 Zinc Complexes

Zinc is a crucial trace element responsible for all livings on our planet.[72-75] Zinc deficiency can result in a variety of health issues related to skin, bones, and the reproductive, digestive, and immune systems.[76] With an $[Ar]3d^{10}4s^2$ electronic configuration, zinc has only one oxidation state, but exists in different isotopic forms of mass ranging from 66 to 70. The completely filled d-shell enables Zn^{2+} to coordinate various ligands in highly flexible geometry and run fast ligand-exchange reactions.[77] The coordinating atom of zinc is usually N, O, S and these electron-donor ligands attach to the central meal zinc in tetrahedral, trigonal bipyramidal, or octahedral geometries, among which tetrahedral is the most common geometry with a coordination number of four.

As the second most essential and abundant element in human bodies, transition metal zinc plays important structural and catalytic roles in various biological activaties.[78-82] Zinc is a good electron acceptor, thus serving as a Lewis acid in catalysis. Its structural functions are validated by the fact that zinc is found in various protein structures and superstructures. The importance of zinc to biological systems can be briefly summarized below: first, the importance of zinc to the gene is indispensable based on the fact that approximately 25% of the zinc compound of rat liver is identified in the nucleus.[83] Specifically, zinc actively participates in the process of genetic stability and gene expression in various ways, such as the structure of chromatin, DNA repair, RNA transcription, DNA and RNA polymerases as well as programmed cell death.[84] Second, the d^{10} configuration makes Zn^{2+} redox inactive, an ideal antioxidant which inhibits any possibilities of free radical reactions. And this property is crucial for antioxidant protection in biological systems. The antioxidant effect of zinc can be mediated via multiple ways including the direct activity of zinc ions, the regulatory effect on metallothionein induction as well as its structural functions in antioxidant enzymes. Specifically, zinc ions can directly bind to thiol functional groups to prevent oxidation.[85] Meanwhile, zinc is also identified as a crucial component of the antioxidant enzyme known as Cu, Zn-superoxide dismutase (SOD$_1$), which plays a vital role in defending the body against oxidative stress. When there is a deficiency of zinc in the body, the activity of SOD1 can be suppressed, potentially leading to increased oxidative damage in cells.[86] Furthermore, it has been indicated that zinc can indirectly influence the functionality of various other antioxidant enzymes.[87] Third, zinc complexes are reported to have great medicinal effects for a variety of diseases. For instance, zinc complexes have shown appealing properties as photosensitizers in photodynamic therapy against cancers.[88] Compared to platinum derivatives which are the most widely used anticancer agents, zinc complexes have lower toxicity, which make them ideal alternatives. Besides, Zn complexes have been widely used as anti-Alzheimer agents,[89] anticonvulsant[90] and for antidiabetic treatment.[91-93]

2.2 Neural Network Potentials

Electronic structure theory methods enable the understanding of molecules, materials on the quantum level and thus complement the experimental studies. The rapid development of computational resources allows large-scale electronic structure simulations for simple systems. However, the ever-increasing demand for accurate computational calculations of large and complex systems is currently infeasible to achieve. Density functional theory (DFT), currently the computational backbone within the electronic structure theory, provides a practical compromise between chemical accuracy and computational cost, but the explicit form of Kohn-Sham DFT is still unclear and the functionals are formulated in increasingly complicated analytical forms as they climb the Jacob's ladder.[94] In addition, the functionals are crafted to satisfy certain physical conditions, including asymptotic behaviors and scaling characteristics, however these designs highly depend on human heuristics, particularly in the intermediate regime where the asymptotic principles are not applicable.[95]

Alternatively, force fields methods have been proposed to model the chemical reactions for large system by summing over the bonded and nonbonded interactions.[96] The bonded term is used to describe the simple interactions at close distances between the directly bonded atoms, as well as angles and dihedrals among atoms connected through shared bonding partners. Conversely, nonbonded terms model the pairwise interactions between atoms, mainly for electrostatics with

Coulomb's law and dispersion with Lennard-Jones parameters[97]. The simple format of classical force fields (FF) methods greatly improves the computational efficiency comparing to DFT methods, thus making it possible to model the system which includes thousands of atoms via molecular dynamics (MD) simulations. Even though a fundamentally sound analysis of chemical interactions is ensured, deep insights derived from MD simulations are usually limited due to the low accuracy of FF methods.[98] Especially for systems where the polarization and many-body interactions have to consider, large error may appear using FF methods since both types of interactions are completely ignored in classical FF methods.

Considering the limitations of both DFT and classical FF methods, new pathways for accurately and efficiently modeling the electronic structure of molecules, clusters and materials are highly required. And ML should be a good choice. ML methods strive to discern the implicit relationships between inputs, either predefined chemical descriptors or just xyz coordinates and outputs, i.e., the chemical properties. The learning process relies on the provided data set. And a well-trained model can capture the fundamental physical laws of quantum mechanics embedded in the data. Practically, ML methods do not need to deal with any mathematical formulas that obey the physical principles to represent the structure-property relation, which greatly simplifies the calculations. This unique ability allows to investigate the realm of chemistry and forecast the attributes of new molecules and materials with unparalleled efficiency and remarkable accuracy.[99-101]

A potential energy surface (PES) delineates the relationship between a system's energy and its geometrical parameters, such as the atom coordinates, with the assumption of the Born-Oppenheimer approximation. Each point on the PES represents a unique conformation of the given configuration at different energy levels. As a result, the PES is utilized to locate stable conformers, i.e. local minima, to investigate the minimum energy pathways among numerous possible conformers, and to find transition states which include all information about the chemical reactions. Furthermore, the PES is also utilized to run MD simulations to gain understanding of the reaction dynamics.[104] The PES aims at providing an understanding of systems at the atomic level, such as small organic molecules, liquids, solids, and polymers because all stable and metastable structures, atomic vibrations, transition states and activation barriers between various structures can be tracked on the PES.

However, MD simulations assisted by the PES is challenging. On one hand, for large systems the PES can only be generated using quantum mechanics/molecular mechanics (QM/MM) techniques since it is practically impossible to determine the complete PES for over thousands of atoms due to the complexity of the system. For instance, the PES was used to investigate the catalytic mechanism of enzymes.[105] The insights into electrons movement in a chemical reaction and the mechanistic role of active site in residues can all be obtained by analyzing the PES. However, the semiempirical methods used in QM/MM are often unreliable for providing accurately microscopic chemical properties. On the other hand, to develop the full PES, the energy of a nonlinear molecule with N atoms is associate with a function of $3N$-6 degrees of freedom. And the computational expense for calculating the PES escalates quickly with an increase in the number of atoms because the electron structure theory or DFT methods are utilized, and the computational scaling of DFT is typically cubic with respect to the number of atoms ($O(N^3)$), whereas the coupled-cluster singles, doubles, and perturbative triples [CCSD(T)] method scales as the seventh power ($O(N^7)$) with the number of atoms.

To circumvent the limitations mentioned above, machine learning potentials (MLPs) were proposed to represent the multi-dimensional PES two decades ago.[106] MLPs learn the contours of the PES using reference data curated from first-principles calculations. These MLPs implicitly encode the atomic interactions in regard to nuclear charges and atomic positions without a significant loss of accuracy compared to the electronic structure calculations but they are much faster than these traditional quantum calculations.

To build a ML model to investigate the potential energy surface (PES), suitable reference data is of great importance. The reference data is usually obtained from *ab initio* calculations. But the *ab initio* calculations are only required for data preparation, since once the model is trained with the reference data, any predictions can be directly obtained from the model with *ab initio* level accuracy. The reference data determine the applicability domain of the trained model as well as the reliability.

Any deficiencies in the data will unavoidably result in imperfections of the trained model. Therefore, the reference data stands as a crucial element of any model. But in computational chemistry, compiling the data sets is not easy. This is primarily because each reference point derives from calculations that are both computationally intensive and complex, restricting the volume of data that can be gathered. Besides, the chemical space of molecules, clusters and materials is extremely large, and it is not easy to locate the representative geometries efficiently. The most common strategies for sampling and generating the reference data sets can be summarized below:

(1) *ab initio* molecular dynamics (AIMD) Sampling. In this method, dynamical trajectories at finite-temperature are produced by using forces derived from electronic structure computations. Although this method is expensive, it is able to accurately describe the chemical process, such as chemical bond breaking and forming. In most cases, the system is simplified so that only N nuclei and N_e electrons are considered to meet the Born-Oppenheimer approximation and the dynamics of the nuclei are assumed on the ground-state electronic surface. But due to the impossibility of precisely solving the differential equations of the ground-state electronic structure, approximate electronic structure methods are widely used, among which DFT is the most common one due to its good performance at a relatively acceptable

computational cost.

(2) Normal Mode Sampling. This method does not require to run any MD simulations. To generate a set of data with the energy range around the minima energy structure, it starts from an equilibrium point on the PES, where atoms at the geometry's energy minima are randomly displaced with the normal modes. To achieve this, the normal mode coordinates at the minimum position are first calculated and the displacement coordinates are then obtained based on the setting of harmonic potential which is derived from the normal mode coordinates. The single point energy of the newly generated geometry is calculated as the reference data while the displaced coordinates are as input. A typical example of this sampling method is ANI-1.[102] Although this method is efficient since no MD simulations are involved, only samples close to minima can be generated which limits the energy space, thus resulting redundant geometries in the data set.

(3) Metadynamics Sampling. Metadynamics[103] is a dynamic sampling method which is capable of biasing configurations away from the positions that have already been visited. This enhanced sampling method uses collective variables to define a biased potential, compelling the system to explore less probable states, thus all states on the PES can be sampled. The collective variables are a predefined number of degrees

of freedom. In metadynamics, the externally applied bias potential, a function of the collective variables, is iteratively added to the Hamiltonian of the system, which induces the system to sample the high-energy area. One benefit of this approach is that the high-energy landscape on the PES which is usually ignored in classical MD simulations can be frequently visited. Therefore, the data collected from this method is evenly distributed, instead of being limited to a narrow energy window.

One drawback of ML methods is the limited extrapolation abilities since the trained models can only make reliable predictions in the training data domains. For data curation in ML, the sufficient sampling of the PES is therefore crucial. In the three sampling methods discussed above, metadynamics sampling is highly recommended.

In the past 20 years, various MLPs have been proposed. Models based on traditional ML include support vector machine potentials,[107] atomic cluster expansion potentials,[108] spectral neighbor analysis potentials,[109,110] gaussian approximation potentials,[111,112] moment tensor potentials,[113] gradient domain machine learning,[114] *etc.* For example, Roman and coworkers developed a least-squares support vector machines (LS-SVM) to investigate the interatomic potentials of around 200 small organic molecules which only contain H, C, N, O, F five elements.[107] To accurately model the pairwise interactions in the system, constitutional descriptors, such as the mole fractions of different atoms, the size of molecule, and quantum-chemical descriptors including average polarizability, dipole moment, quadrupole moment, HOMO-LUMO gap were used to decode the structure of molecules. With these molecular descriptors, it is practically accessible to have a clear intuition of what the model learns from the provided reference data, thus allowing to understand the structure-property relations. The LS-SVM is able to deal with multivariate calibration problems and solve both linear and nonlinear multivariate calibrations. And regularization technique can be introduced to the LS-SVM model to better balance between overfitting and underfitting. The findings indicated that the LS-SVM demonstrates greater efficiency compared to ANNs for training. Furthermore, for two extra test sets, the LS-SVM model showed better interpolation and extrapolation abilities than ANNs.

Although in this specific example of LS-SVM, traditional ML methods outperform ANNs, this conclusion is not universal. For large amount of data, which is necessary for real-world applications, ANNs definitely surpasses traditional ML methods, due to its structural flexibility. The potentials modelled using neural networks are call neural network potentials (NNPs). Starting from 1995 NNPs have been developed a lot and can be classified into four generations.[115]

Doren and coworkers first proposed a single feedforward neural network model to estimate the adsorption energy of H_2 molecule on the Si (100) cluster surface.[116] This first-generation neural network is a two-dimensional system since the input layer has only three input coordinates and the single neuron in the output layer gives the predicted energy. At least one intermediate hidden layer resides between the initial input and the final output layer. But no physical information is embedded in the nodes of these hidden layers since they are used just to increase the neural network's flexibility and allow for the processing of complex features and patterns in the input data before reaching the output. Adding more hidden layers and nodes increases the network's capacity for flexibility, thus having better generalizability. The relation between any two nodes in the two consecutive layers is defined as

$$y = wx + b \tag{1}$$

where w is weight, b is bias, and both are learnable parameters. However, this simple linear representation is not enough. A nonlinear function, or activation function is applied so that any arbitrary function can be well represented. The output of a node is then defined as

$$y = \sigma(wx + b) \tag{2}$$

where σ is the activation function. Various activation functions are frequently used. Some examples are available from **Figure 35** to **Figure 38** (See APPENDIX B: FIGURES). The advantages of this first-generation NNPs are obvious: they can accommodate numerous training parameters so that any physical principles can be modelled. And no preliminary equations are required to represent these physical rules which make this method simple to implement. Furthermore, the high flexibility also enables the model to accurately represent the energy of the system in regard to atomic coordinates. Finally, this method allows the calculation of analytic derivatives. Hence, both energy and forces can be calculated at speeds that are much faster than first-principles methods. However, the limitations of this single feed-forward neural network are also obvious. First, this method is not applicable to large system, since the extra input neurons of the model is required whenever a new atom is introduced to the system. For a system with hundreds of atoms, the number of input nodes is excessively large, and it can impede efficient training of the model due to increased computational demands and potential overfitting. Second, in this simple feed-forward ANN, the symmetry is not guaranteed. Changing the input order of each atom results in different energy output, which is contradictory to the fundamental principle that the PES should inherently be invariant under translation, rotation, and permutation of identical particles. For example, for a water monomer, the input can be the pairwise distances of these atoms, since the two O-H bonds should be chemically equivalent, exchanging the input order of two H atoms should not change the energy of this water molecule. However, each weight and bias in the neural network are usually numerically different, different input orders possibly give rise to different outputs. Finally, once the NNP is determined, it can be only applied to a specific class of system because the framework of this neural network limits it to predict the potential energy for a system with different size of atoms. Any new system containing more or less atoms needs to be retrained.

The restrictions of the low-dimensional neural networks prevent its application to complex system. To solve these issues, Behler and Parrinello introduced the second-generation NNPs in 2007.[117] The first improvement is to abandon the single feedforward neural network. Instead, in the second-generation NNPs, each atom is given a neural network to predict the atomic energy. By aggregating these atomic energies, the cumulative energy for the entire system can be determined by

$$E_T = \sum_{i}^{N} E_i \tag{3}$$

where E_i is the atomic energy and E_T is the total energy across all N atoms. The unique design of this method is that each element shares the same neural network with the same weights and biases which reduces the computational cost. In addition, this atomic neural network allows to model the PES for any arbitrary system since if a new atom is introduced to the existing system, a corresponding neural network of that element can be easily added to the whole framework of neural network. On the other hand, an atomic neural network can also be easily deleted if one atom is removed. The second improvement is that the locality approximation is introduced. Instead of including all pairwise interatomic interactions, interactions are only taken into account between atoms that fall within a predefined cutoff radius. This type of short-range interactions can cover the main interactions among atoms without a significant loss of accuracy. Finally, the so-called atom-centered symmetry functions (ACSFs) are introduced to encode the local structural fingerprints of the atomic environments. Meanwhile, this type of descriptors can also keep the translational, rotational and permutation invariance. One important component of ACSFs is the cutoff function which is used to define the local atomic environments. The cutoff function should meet some criteria: (1) it should be differentiable to ensure there is no discontinuity in the descriptor numbers as well as the corresponding derivatives; (2) It should decay smoothly to zero at the cutoff radius to make sure the interactions decrease at larger distances and become zero outside the cutoff. A common cutoff function[117] adapted from the cosine function is defined as

$$f_c(R_{ij}) = \begin{cases} 0.5\left[cos\left(\pi\frac{R_{ij}}{R_c}\right) + 1\right], & R_{ij} \leq R_c \\ 0 & , R_{ij} > R_c \end{cases} \tag{4}$$

where R_{ij} denotes the relative position between atom i and atom *j*. R_c denotes a predefined cutoff radius. The shape of this cutoff function is given in **Figure 1**. The local atomic environments are then described based on this cutoff function. The ACSFs include two types of functions, i.e., the radial and angular symmetry functions. Both symmetry functions depend on the distance between the center atom i and its neighbor j. They complement each other to fully describe the local atomic environments of each atom. And each type of symmetry functions has a range of different functional form.

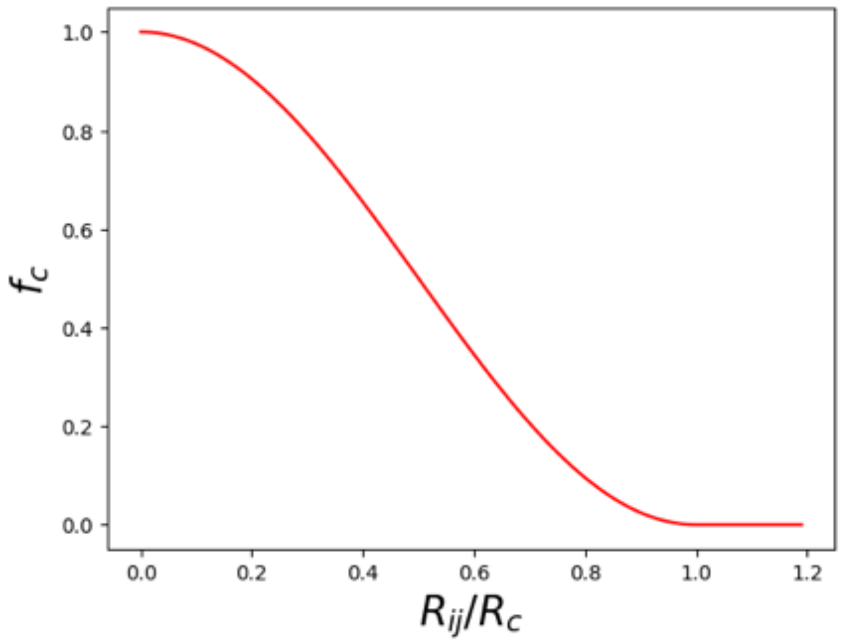


Figure 1. The cosine adapted cutoff function.

But the symmetry functions have to meet some requirements. First, like the cutoff function, they should also decay in value as the neighbor gets away from the center atom and become zero beyond the cutoff range to reflect the real physical interactions along with distance. Moreover, they should be able to capture the minimal differences of similar structures, such as conformers. Each unique structure should have unique representations decoded by these symmetry functions. Finally, they should not depend on the number of neighbors since the number of atoms within the threshold value varies in different molecules, but the dimensionality of the input layer of the neural network keeps fixed once it is determined.

The radial symmetry functions are two-body terms, based on the pairwise distances among atoms. They are designed to describe the radial environment of atoms. The frequently used radial function is a sum of Gaussians of all neighbors,

$$G_i^{rad} = \sum_{j\neq i}^{N} e^{-\eta\left(R_{ij}-R_c\right)^2} f_c(R_{ij}) \tag{5}$$

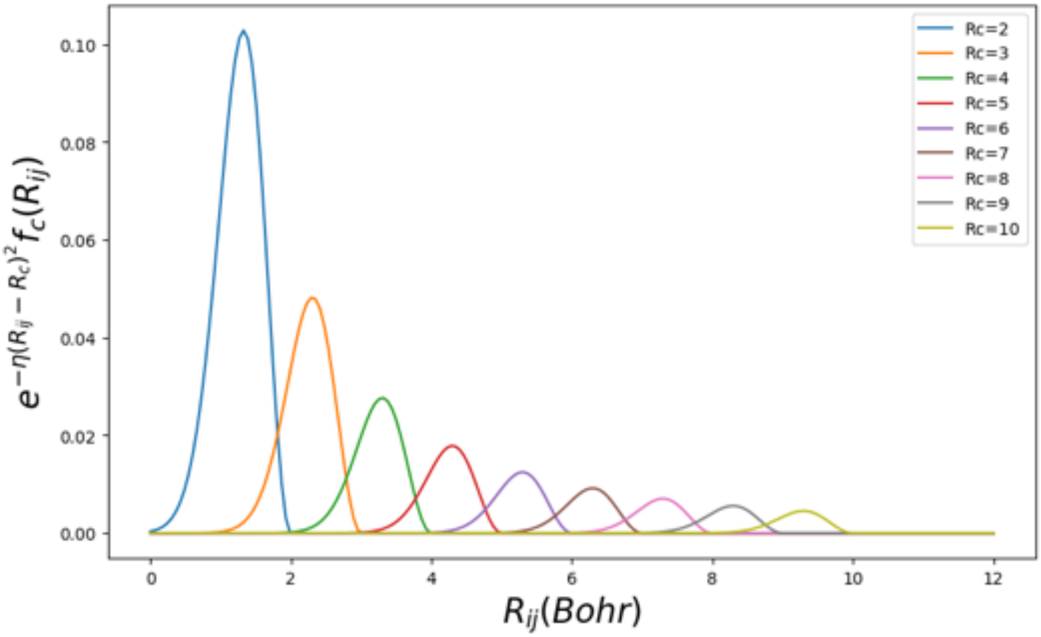


Figure 2. Radial symmetry functions(η=2).

Figure 2 shows the radial functions with a set of different cutoff radius. Summing over these Gaussian functions ensures the representation of local atomic environment is not affected by the number of neighbors so that molecules with various system size can all be fed into the neural network. The radial function's spatial range is controlled by the parameter η to provide a smooth transition of both potentials and forces at the cutoff.[118] However, only using the radial symmetry functions is not enough to exactly describe the local atomic environments because the radial functions only cover the radial environment, but for some systems, such as square planar coordination and tetrahedral coordination, if all the bond lengths are the same, the radial function is not able to distinguish both geometries. The angular symmetry function is then designed to solve this issue. Angular terms are constructed to incorporate the angles for any three atoms. A typical form is

$$G_i^{ang} = 2^{1-\zeta}\sum_{j,k\neq i}^{all}\left(1+\lambda\, cos\,\theta_{ijk}\right)^{\zeta} e^{-\eta\left(r_{ij}^2+r_{ik}^2+r_{jk}^2\right)}\, f_c\left(r_{ij}\right)f_c\left(r_{ik}\right)f_c\left(r_{jk}\right) \qquad (6)$$

where θ_{ijk} is the angle spanned by the atoms i, j, and k. The parameter $\lambda = \pm\, 1$, inverts the shape of the cosine function to capture an accurate depiction across different values of θ_{ijk}, while ζ controls its width. As shown in **Figure 3**, the exponent ζ can achieve good angular resolution.

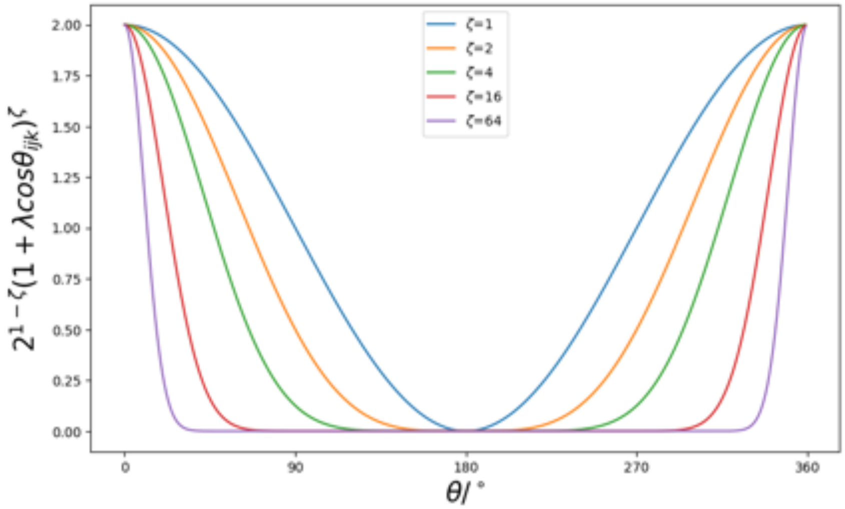


Figure 3. The angular term $2^{1-\zeta}\left(1+\lambda\, cos\,\theta_{ijk}\right)^{\zeta}$ with a set of ζ and $\lambda = 1$.

The term $2^{1-\zeta}$ is a normalizing factor to control the range of the angular symmetry functions. And since θ_{ijk} and $2\pi - \theta_{ijk}$ should give the same angular function value, cosine function is used to keep the symmetry. Both radial and angular symmetry functions use a set of different parameters to fully describe the atomic environments.

The ACSFs are useful descriptors for NNPs and have shown successful applications in a variety of systems. For example, ANI-1[65] is a well-trained NNPs for small organic molecules that only have less than 8 heavy atoms of carbon, nitrogen, oxygen. From ~58k neural molecules, the authors used normal mode sampling method to generate ~17.2 million conformers. They then redesigned the symmetry functions. First, they introduced a predefined parameter to shift the angular functions. Second, the cutoff radius was added to the distance exponent part of the angular function. Both modifications allow the descriptors to recognize different molecular features more accurately, such as bonding patterns, functional groups. The ANI-1 model uses 32 evenly spaced hyperparameters for radial functions and 16 shifting hyperparameters for modified angular functions, resulting in total 768 predefined parameters. It includes 3 hidden layers, each of which has 128, 128, 64 neurons. The results indicate that ANI-1 model can yield a RMSE of 0.6 kcal/mol with regard to DFT reference method and also has good transferability of predicting the energetics of larger systems with 10-24 heavy atoms.

Although the second-generation NNPs with ACSFs greatly improve the performances compared to the first-generation single feed-forward neural network, one limitation is that these predefined descriptors require much human expertise since each hyperparameter needs to be manually selected or adapted through numerous tests which is laborious and time-consuming. In addition, an increasing number of input dimensions can rapidly lead to high computational cost for both descriptor calculation and model evaluation in multicomponent system. To overcome the bottleneck in ACSFs, learnable descriptors are introduced. The base framework is 'message passing neural networks' (MPNNs), in which every molecule is considered as a three-dimensional graph.[119] MPNNs are adapted from Graph Neural Networks (GNNs)[120], in which every sample is considered as a graph. A graph is an ensemble of objects with specific connectivity. And it includes node, edge, and global three parts of attributes. Generally, the graph represents the relations (edges) between a collection of nodes. The simplest GNN applies an individual multilayer perceptron (MLP) to each element in a graph, i.e., for each node, a MLP is applied and returns a learned node, and a learning embedding can also be applied to single edge and the whole graph. Finally, an updated graph is obtained. One drawback of this simplest GNN is that it does not consider the connectivity of the graph, since each node, edge and global context is processed independently. But the connectivity contains very important information, for example, the connectivity of an atom with its neighbors reflects its local environment which greatly affects its own atomic contribution to the total energy. MPNNs were then proposed to overcome this limitation. A MPNN aims at updating all attributes of a graph while maintaining the graph symmetry, i.e. the connectivity is unchanged during the course of transformation. In other words, MPNNs follow a 'graph-in, graph-out' mechanism meaning that only the information of a graph loaded into its nodes, edges and global context are progressively transformed, but the connectivity index of any at least two nodes are still kept. Molecules in chemistry are good examples of graphs since all atoms in the molecule are interacting with each other via electrons. And the interactions can be reflected by bonds connecting two atoms. Therefore, for a given molecule, each atom can be considered as a node and each bond is regarded as an edge in GNNs. The task on the graph can be either node-level, edge-level or global-level. For example, on the global level, we can predict the property of the whole graph, like whether a ligand binds to a receptor. And for the edge-level property, we can predict the type of a bond, either single, double, or even triple. Finally, for node-wise property, GNNs can be designed to estimate the atomic partial charges of a given molecule. In terms of NNPs based on MPNNs, the node-level prediction is of great concern. The goal is to predict the atomic energy of each node and sum them up to get the total energy of the molecule. MPNNs leverage the connectivity of graphs which makes GNN models more sophisticated. Further, multiple GNN layers can be stacked together, through which, a node can eventually incorporate information across the entire graph. The MPNNs include in three steps:

1. Each node in the graph collects all the neighbors' embeddings.
2. All these neighboring messages are aggregated through pooling technique.
3. The pooled messages are passed through nonlinear layers to update the node information.

Specifically, each node's embedding is first randomly initialized, $h^0 \in \mathbb{R}$ and each node exchanges its own information with its neighbors via message passing block M_t. The central atom first collects its neighbors' message,

$$m_i^{t+1} = \sum_j M_t\left(h_i^t, h_j^t, e_{ij}\right) \tag{7}$$

where m_i^{t+1} is the total message passed from the neighbors *j* to the central node *i* at step *t*, and the bonding information of pairwise atoms e_{ij} are included to model the interactions between *i* and *j*. The central node is then updated based on both the gathered message m_i^{t+1} and its own representation h_i^t via update block,

$$h_i^{t+1} = U_t(h_i^t, m_i^{t+1}) \tag{8}$$

The message passing and update block together is called one interaction. And each node goes through the interaction block multiple times, allowing the message to be disseminated throughput the molecule. At the final step *T*, each node transforms its updated embedding into atomic energy,

$$E_i = R(h_i^T) \tag{9}$$

Finally, all atomic energies are summed up as the total energy via eq 3. Compared to neural networks based on ACSFs, MPNNs do not require any predefined descriptors because they are replaced by neural networks to learn all pairwise interactions implicitly. In MPNNs M_t, U_t, R are all nonlinear functions designed by neural networks. And they are much more expressive than manually selected descriptors due to their flexible framework.

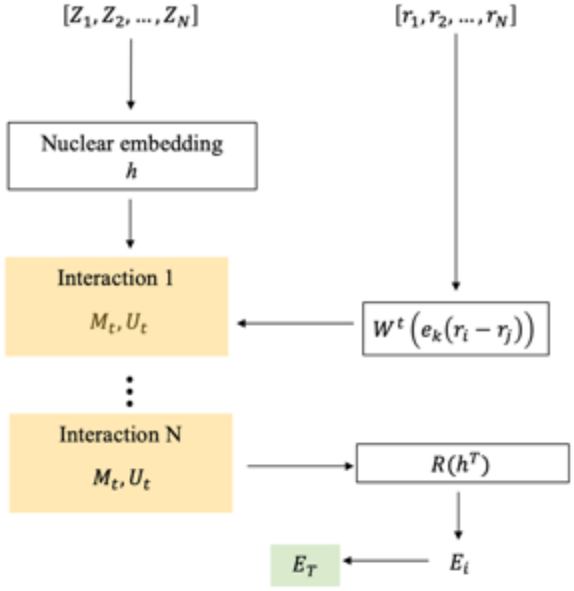


Figure 4. The schematic process of SchNet.

A well-designed MPNNs for NNPs is SchNet.[121] The inputs of SchNet are the nuclear charges Z= $\{Z_1, Z_2, \dots,\ Z_N\}$ and the atom positions R= $\{r_1, r_2, \dots,\ r_N\}$. First, these nuclear charges are transformed to high-dimensional features to represent the atom type meaning that the same atom type gets the same initial nuclear embedding.

This is a common step for initialization in MPNNs. The first feature of SchNet is the expansion of distances by a set of Gaussian basis,

$$e_k(r_j - r_i) = \exp\left(-\gamma\left(\|r_j - r_i\| - \mu_k\right)^2\right) \tag{10}$$

eq. 10 is similar to the radial symmetry functions in ACSFs to extend the distance in space by adding some nonlinear transformations, $\mathbb{R}^3 \rightarrow \mathbb{R}^D$ where D is the dimension of the expanded distance. Both γ and μ_k are hyperparameters to determine the degree of expansion. Another contribution of SchNet is to decode the edge information. Instead of using the direct relative positions of pairs of atoms, the authors used convolutional filters to further transform the bonding information. These convolutional filters are designed by neural network,

$$m_i^{t+1} = \sum_j M_t\left(h_i^t, h_j^t, e_{ij}\right) = \sum_j h_j^t \circ W^t\left(e_k\left(r_j - r_i\right)\right) \tag{11}$$

where $\circ$ represents the element-wise multiplication and W^t are convolutional filters at step *t*. The advantage of W^t is that the atomic positions can be mapped from $\mathbb{R}^D \rightarrow \mathbb{R}^F$, where F is the hidden features. The framework of SchNet is given in **Figure 4**. As the results shows in the original work, SchNet models the interatomic potentials very well. The authors first tested a well-known benchmark data set QM9[122] which includes ~131k organic molecules with less than nine heavy atoms of carbon, nitrogen, oxygen, fluoride. SchNet evaluated 12 molecular properties which are minimal energy, enthalpy, the energy of HOMO and LUMO along with the corresponding energy gap, polarizability, harmonic frequency, dipole moments, *etc*. SchNet shows good performance on all these tested properties. For example, the mean absolute error (MAE) of minimal energy predictions is 0.014 eV.

In the past five years, a lot of variants of SchNet have been proposed to further improve the performances of NNPs, like PAINN,[123] tensor field networks (TFNs),[124] NequIP, [125]SE (3)-transformers[126] to name a few. The common feature of these NNPs is that they are all equivariant GNNs. For example, in order to overcome the limitation of only invariant representations in SchNet, a new type of rotationally equivariant representations were proposed in PAINN. As shown in **Figure 5** (a), only distances and angles are considered in SchNet and since both representations are rotationally invariant, they are unable to differentiate both structures. However, the directional vectors in **Figure 5** (b) can easily recognize the differences in both structures. Therefore, geometric vectors and tensors should be introduced to NNPs to make the model more expressive. Another advantage of these equivariant GNNs is that with the equivariant atom-wise representations, the model can predict tensorial properties accurately, such as forces, molecule dipole moment, polarizability, etc.

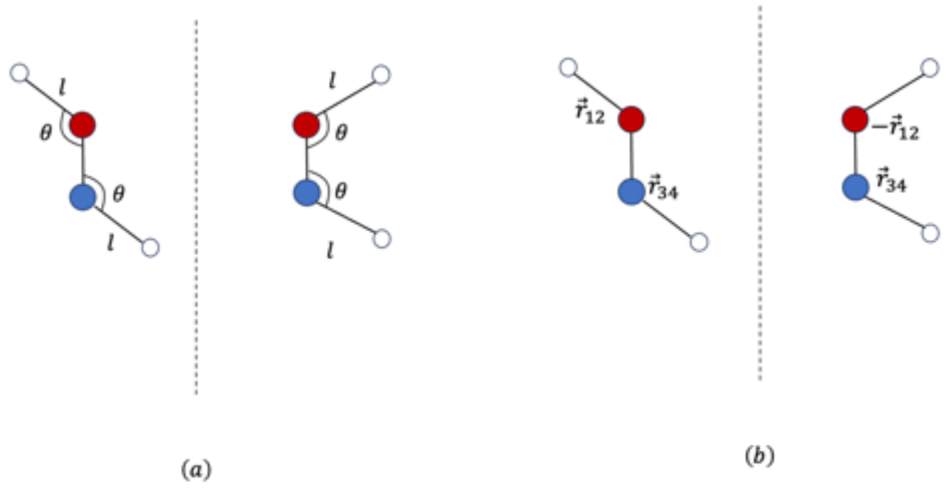


Figure 5. Invariant representations and equivariant representations.

The second-generation NNPs with either predefined or learnable descriptors have greatly advanced MLPs. They are the mainstream methods in machine learning for MD simulations. However, their limitations are also very obvious that only short-range interactions are considered. This design can dramatically decrease the computational complexity but for systems where long-range interactions play an important role, large errors may appear. Electrostatics is a major component of long-range interactions, and the relatively weak dispersion interactions can also contribute a lot to large systems.[127] It is necessary to incorporate long-range interactions in MLPs since it not only covers the interactions beyond cutoff radius, following Coulomb's law, which adds physically meaningful energy term to MLPs but also reduces the radius to smaller threshold, which facilitates a reduced sampling configurational space. A few NNPs which contain electrostatic interactions explicitly are called third-generation NNPs. The first example is an extension of HDDPs proposed in 2011.[128,129] In this method, a second neural network is designed to predict the atomic partial charges. Like the neural network for atomic energy predictions, the partial charges are dependent on a set of ACSFs representing the local atomic environments. The electrostatic energy E_{elec} is subsequently computed based on the predicted partial charges q with Coulomb's law. Finally, the system's total energy is calculated as

$$E_{total} = E_{short} + E_{elec} = \sum_{i}^{N} E_i + \sum_{i>j}^{N} \frac{q_i q_j}{R_{ij}} \tag{12}$$

Practically, the short-range interactions can also cover electrostatic interactions up to the cutoff range, in order to avoid redundant calculation in both parts, the atomic charge neural networks are developed based on reference atomic charges calculated from *ab initio* methods. And the electrostatic energy is calculated with the predicted partial charges. By subtracting the electrostatic energy from the total reference energy, the reference short-range is obtained. Another variant of third-generation NNPs is to include both the electrostatic interactions and dispersion interactions. One typical example is PhysNet.[130] This method uses Grimme's D3 method to explicitly include dispersion corrections in NNPs. Another feature of this model is that the atomical charges and energies are predicted simultaneously with the shared neural network, further improving the NNP' s efficiency.

The partial charges in the third-generation NNPs are considered to be localized as they are determined solely by only the atomic positions within the cutoff sphere. However, in systems where partial charges are influenced by molecular characteristics beyond the cutoff range, this local charge approximation may not hold true. A representative example is the change of global charge state in the system by chemical reactions, such as protonation, deprotonation, or ionization. All three generations of NNPs are unable to describe the PES well because they cannot recognize the electronic differences of these configurationally identical structures. Potentials which can capture the nonlocal or global attributes are defined as fourth-generation NNPs. It should be noted that currently there is no consistent terminology to define third/fourth generation NNPs since the distinction between nonlocal interactions and long-range interactions is still unclear. In this work, any NNPs that involve some interactions which only depend on local property is regarded as third-generation NNPs, otherwise, if the electronic structure is considered or global property is involved, the NNPs are classified as fourth-generation NNPs. The charge equilibration neural network (CENT) is the first fourth-generation NNP.[131] The core idea is to use charge equilibration method[132] to redistribute the electrons throughout the entire system, giving rise to the electrostatic interaction minimization. The total energy is calculated as

$$E_{total} = \sum_{i}^{N}(E_i^0 + \chi_i q_i + \frac{1}{2}J_i q_i^2) + \frac{1}{2}\iint \frac{\rho(r)\rho(r')}{|r-r'|}drdr' \qquad (13)$$

where E_i^0 are the potentials of free atoms, χ_i are the atomic electronegativities, q_i are the atomic charges and J_i are the element-wise hardness values. The charge density ρ is calculated by Gaussian distributions. The inputs of CENT are still ACSFs to represent the atomic environments, but the predicted partial charges are redistributed before calculating the electrostatic interactions. The charge redistribution strategy implemented in NNPs shows remarkable performances for systems with predominantly ionic bonding.[133-135] Recently, some other fourth-generation NNPs have also been proposed, such as BpopNN,[136] 4D-HDNNP.[137,138] Since the introduction of second-generation NNPs, MLPs have gained considerable attention. And the swift advancement in this area is still going on without reaching full maturity. Currently, most work focus on organic molecules due to its simple electronic structures and already available public data set, such as ISO17,[121] QM9,[122] MD17,[139] ANI-1,[140] etc. However, the exploration of TMCs in MLPs is still being ignored, possibly because of the unclear interactions between ligand and metals as well as the scarcity of large data set. Although a few available data sets for TMCs are already available,[141-143] none of them considers multiple conformations. The process of conformational sampling is pivotal in defining the macroscopic and physicochemical properties of molecules because the three-dimensional arrangement of atoms greatly influences a variety of properties.[144] Furthermore, molecular conformations mediate biological activities as well. For example, DNA only binds to a zinc transcriptional regulator once zinc coordinates in a specific manner and tunes a conformational change in the transcription regulator.[145]

In this work, a NNP to model the PES of zinc complexes was built by following steps below:

1. A dataset including zinc complexes with both configurational and conformational diversity was curated.
2. A neural network which covers both short-range and long-range interactions was well designed.
3. The results indicated this proposed model is orders of magnitude faster than DFT methods without losing significant accuracy.

2.3 Zinc Data Set Curation

The transitional metal quantum mechanics (tmQM) data set, including tens of thousands of GFN2-xTB[146] optimized TMCs extracted from the Cambridge Structural Database (CSD),[147] was used to get the configurational ensemble of zinc complexes. From the tmQM data set, 771 complexes were extracted and each of them has 60 atoms or less. And only H, C, N, O elements are included in these complexes. All complexes are neutral, closed-shell, and mononuclear. Some examples are given in **Figure 6**. The size distribution and the element distribution of the 771 complexes are given in **Figure 7** and **Figure 8**, respectively. A variety of ligand types and bonding patterns are present in this data set. As shown in **Table 1**, the coordination number of these complexes is various, ranging from 2 to 8. And totally 38 denticity types were identified in this data set. In 2435 ligands extracted from these 771 complexes, 829 ligands are unique (evaluated by SMILES). The denticity distribution of the ligands is given in **Figure 9**. Several polydentate ligands are given in **Figure 10**.

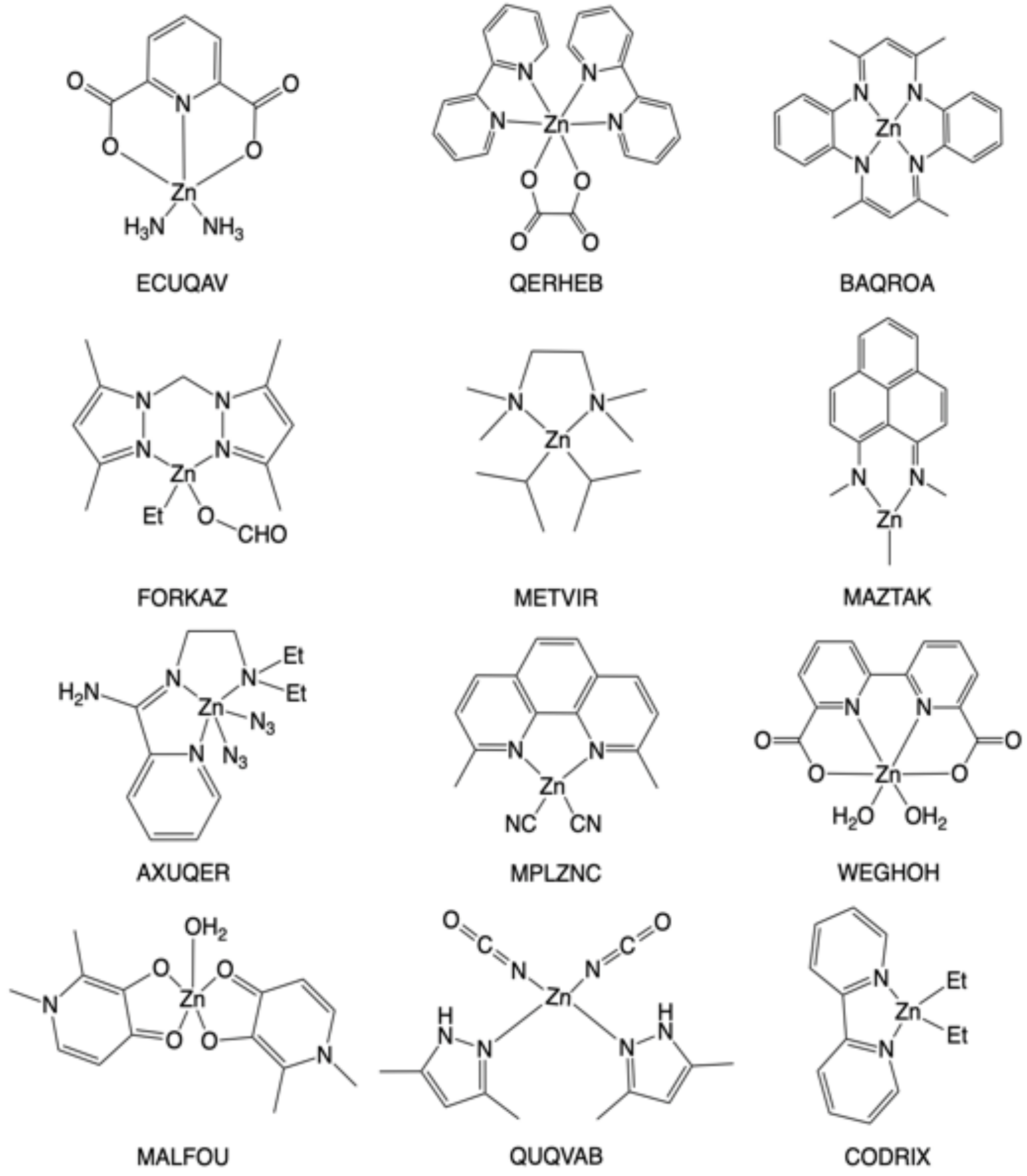


Figure 6. Some structures of zinc complexes with the refcode taken from the CSD.

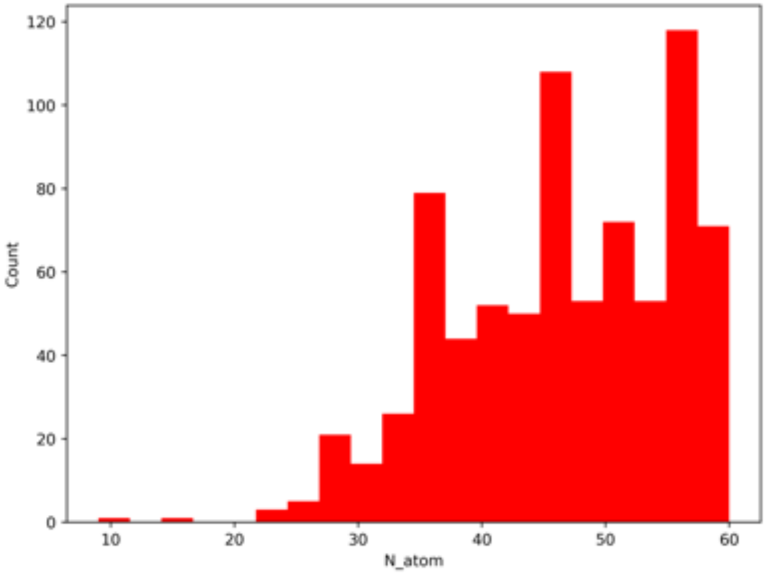


Figure 7. This size distribution in 771 zinc complexes.

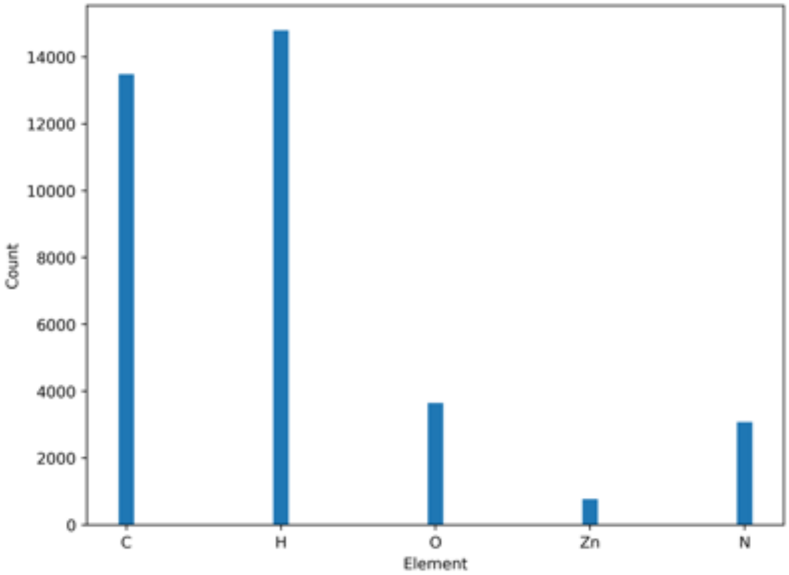


Figure 8. The element distribution in 771 zinc complexes.

Table 1. The coordination types in the 771 complexes.

CN[a]	Denticity type	Count	CN	Denticity type	Count
2	1,1	25	6	1,1,1,3	10
3	1,1,1	3		1,1,4	4
	1,2	18		1,1,2,2	134
	3	1		1,2,3	14
4	1,1,1,1	27		1,5	3
	1,1,2	23		2,2,2	27
	1,3	21		2,4	3
	2,2	66		3,3	41
	4	17		6	10
5	1,1,1,1,1	1	7	1,1,1,1,3	4
	1,1,1,2	8		1,1,2,3	14
	1,1,3	28		1,3,3	10
	1,4	27		2,2,3	20
	1,2,2	42		4,3	2
	2,3	8	8	1,1,3,3	52
	5	2		1,3,4	1
6	1,1,1,1,1,1	40		2,2,2,2	1
	1,1,1,1,2	10		2,3,3	54

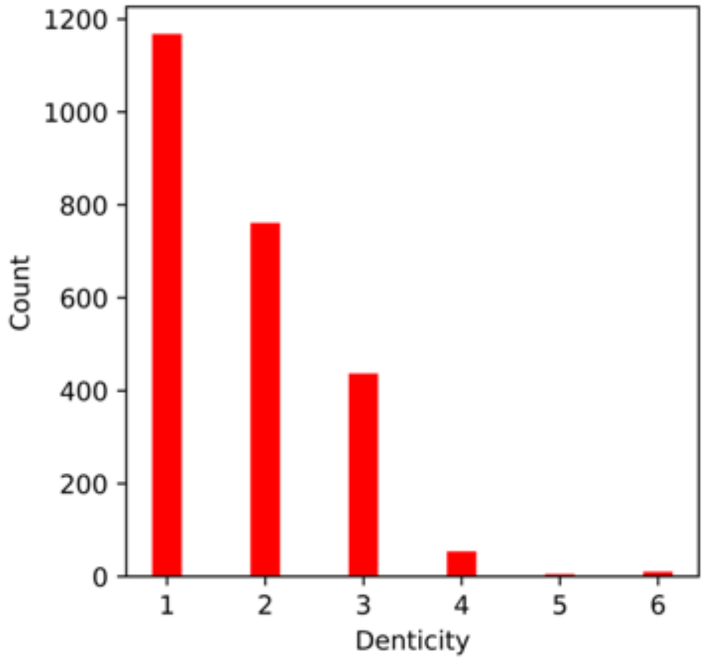


Figure 9.The denticity distribution in 771 complexes.

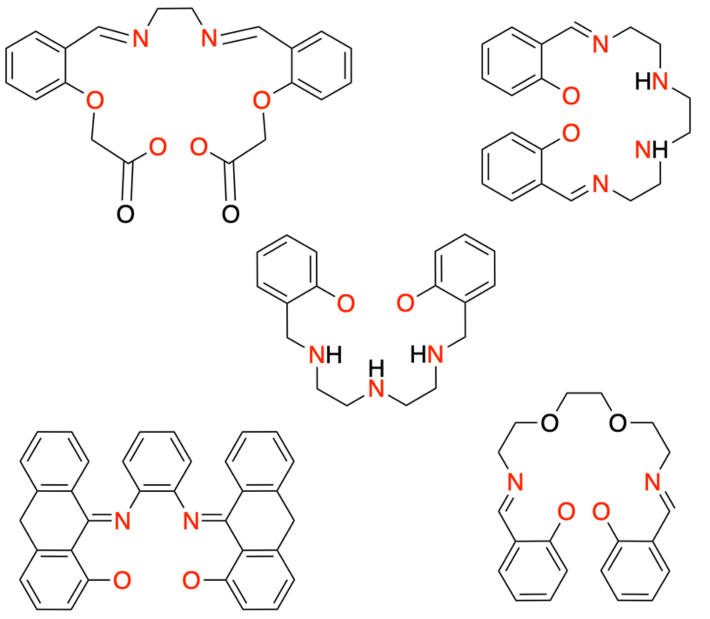


Figure 10.Representative examples of polydentate ligands. The coordinating atoms are in red.

To build the PES of zinc complexes, multiple conformations of each complex were generated using metadynamics. As illustrated in section 2.2, metadynamics outperforms other sampling methods. The canonical MD simulation is a commonly used method to prepare a data set for NNPs. However, it is an energetically uneven sampling method, *i.e.,* geometries in low energy regions are more frequent to generate than those in high energy area. As a result, a lot of redundant geometries with low energy are present in the data set while geometries with high energy are few. The neural network trained on such data sets is biased to energy predictions meaning that it can reach high accuracy in dense configuration regions, while large errors may appear for high-energy geometries. Such uneven sampling can be overcome by introducing additional potentials against previously generated geometries to induce frequent sampling in high-energy space. In this work, the automatic conformation search engine, CREST[148] was used to generate conformations for each zinc complex. In CREST the biased potential is applied as the sum of multiple Gaussian functions in relation to the root-mean-square deviation (RMSD),

$$V_{bias} = \sum_{i}^{n} k_i exp(-\alpha_i \Delta_i^2) \tag{14}$$

where n is the number of all reference structures, the RMSD related to the *i*th reference structure is applied as the collective variable Δ_i, k_i is the pushing strength which is scaled by the size of the structure, and the parameter α_i determines the width of the potential. A dozen metadynamics simulations with different combinations of k_i and α_i were performed in parallel to get a complete conformation ensemble. With the 771 zinc complexes, 53247 conformations were gained via CREST. To avoid redundant sampling, RMSD filtration was applied. People usually calculate the RMSD in a straightforward manner without any structural adjustment, so the value is likely to be too big. But in this work any two conformations were recentered and then rotated onto each other to get the real or "minimum" structural differences. Any conformation within 0.1 Å was removed. This step ensures only highly distinctive conformations were present in the data set. An example of RMSD distribution with 9 conformations is given in **Figure 11**.

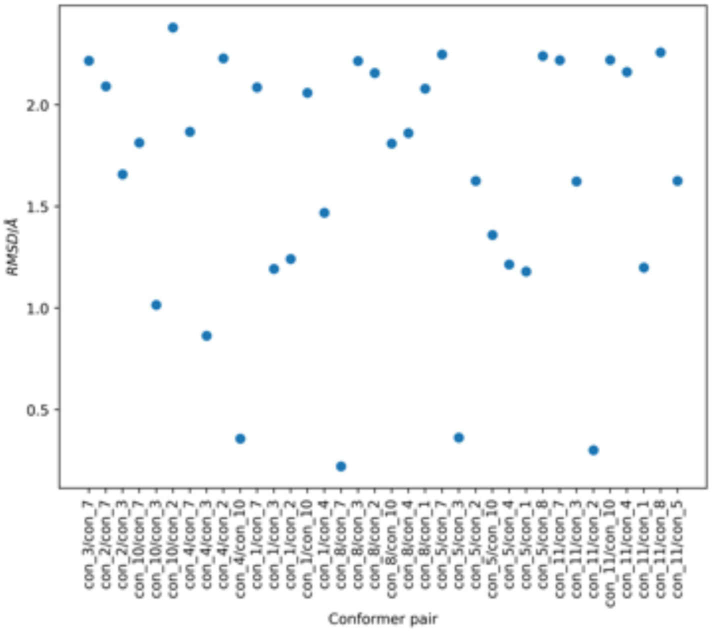


Figure 11. A RMSD distribution of 9 conformations. (CSD code: UGIBAP).

Finally, 39599 conformations were present in the final Zinc_60 dataset. The chemical diversity of this conformation ensemble is given in **Figure 12~14**.

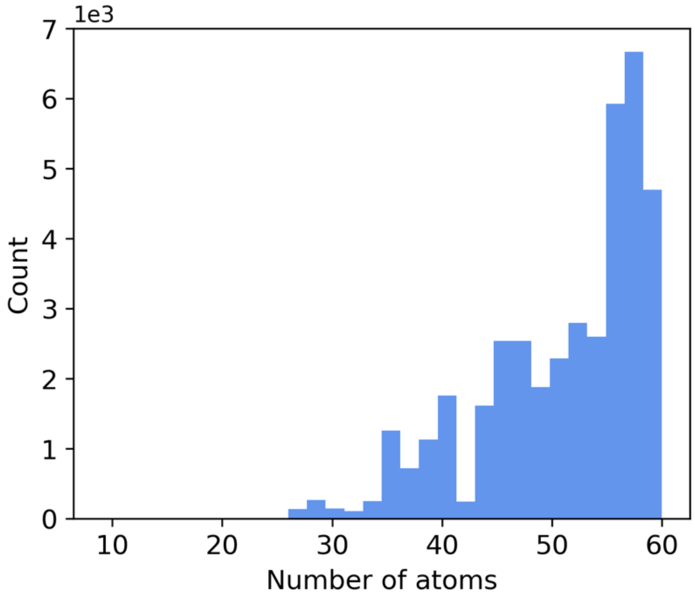


Figure 12. The molecular size distribution in Zinc_60 dataset.

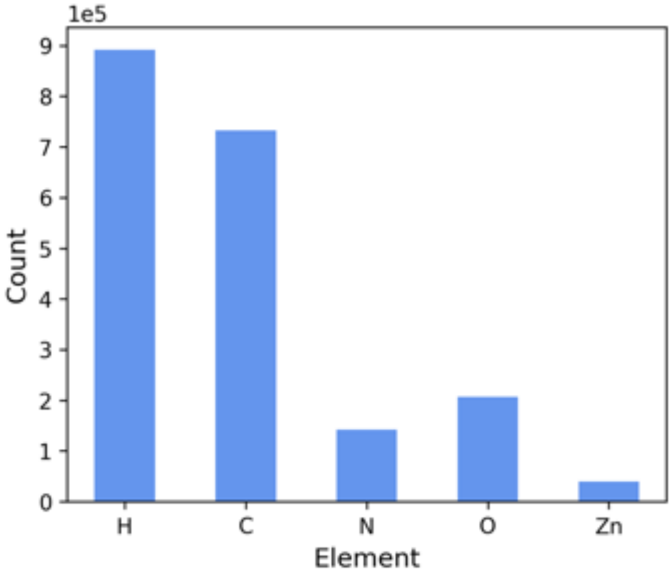


Figure 13. The element distribution of Zinc_60 dataset.

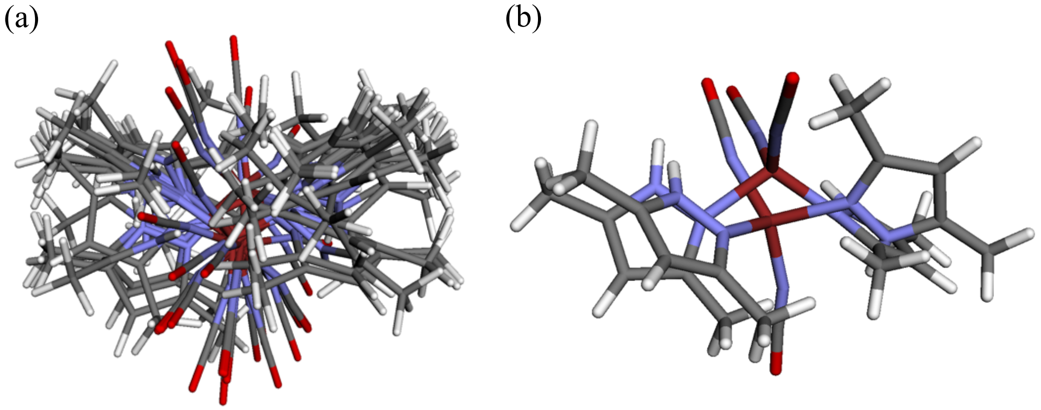


Figure 14. (a) An ensemble of 19 conformations (CSD code: QUQVAB) ;(b) The geometries of the lowest energy and the highest energy in the ensemble, $\Delta E = 5.76\ \mathrm{kcal/mol}, \mathrm{RMSD} = 2.47\text{Å}$.

The unique CSD code of 771 complexes and the number of conformations in Zinc_60 data set is given in **Table 12** (See APPENDIX A: TABLES). The single-point energies of all the 39599 conformations were calculated using the meta-GGA r^{2}SCAN-3c[149] method. All calculations were conducted using ORCA 5.0.4,[150] with *TightSCF* and all other parameters set to the default. Finally, the Zinc_60 dataset was divided into training, validation, and test sets with a distribution ratio of 8:1:1.

2.4 Zn_NNPs Framework

One limitation of the second-generation NNPs is its local approximation that only short-range interactions within the cutoff range are considered. To overcome this drawback, some models explicitly include the long-range interactions, for example, the electrostatic interactions can be included using the Coulomb's law and the dispersion interactions can be contained using Grimme's D3[151] or D4[152] method. However, either the D3 correction used in the ML models, like PhysNet[130] or even the D4 correction is a generally good method to calculate the dispersion interaction in most systems, but neither is 100% accurate since currently people don't have a full understanding of the long-range interactions in complex chemical systems. And it is possible that there are some factors missed in physical laws affect the dispersion interactions. Moreover, one advantage of both dispersion correction methods is their simple form because the parameters in these corrections are predefined by training some datasets. But it is hard to guarantee that these universal parameters are perfectly applicable to all systems. It is possible that large errors may appear for a specific class of systems, like TMCs in this work.

To model the long-range interactions more accurately, in this work we referred to EwaldMP[153] which is a neural network to implicitly model the long-range interactions of our zinc complexes. The basic idea is to embed Ewald summation[154,155] into neural networks. Usually, the long-range or nonlocal interactions are divided into electrostatic, repulsion, and dispersion three contributions. The total term of these potential is given as

$$E(r) = \frac{q_1 q_2}{4\pi\epsilon_0 r} + \frac{A_{rep}}{r^{12}} - \frac{B_{dis}}{r^6} \tag{15}$$

One challenge of calculating the long-range interactions is the slow convergence of potentials with distance. The potentials are defined in non-negative range and the closed forms of the sum are unknown. To practically evaluate them, they are usually truncated at predefined cutoff radius. Potentials beyond this cutoff distance are ignored, which raises systematic errors to the calculation. The Ewald summation was proposed to solve this problem. The main contribution of this method is to transform a single conditionally or slowly convergent sum into two quickly convergent sums by using Fourier transform. The rapid convergence ensures a high accuracy of the calculations, thereby avoiding the systematic artifacts of cutoff-based approximations. The method is mainly designed for electrostatic interactions,[156] but later it has also been used for dispersion interactions,[157] as well as higher order electrostatic multipoles.[158] EwaldMP refers to this method but implements it in neural network. And as a standalone block, EwaldMP can be added to any short-range model. Moreover, in principle, it can account for both long-range electrostatics and long-range dispersion interactions together. The advantages of this method are (i) a neural network is used which is much more powerful than the simple equations in D3 or D4 corrections because the ML model can deal with complicated physical laws via its flexible framework. It should be able to be more comprehensive in evaluating the long-range interactions; (ii) each conformation in this work is a zinc complex, a very specific class of system. The long-range interactions in these conformations should be similar. Hence, the model can learn specifically to just model the long-range interactions of zinc complexes, which should be more accurate than the D3 or D4 method which use the general parameters for all systems. But EwaldMP was not directly copied in our work. One contribution of our work is to introduce the partial charges as the input of EwaldMP. In the original EwaldMP, partial charges are never considered. The long-range interactions are captured with only the Cartesian coordinates and the embedding of the atomic types. Our hypothesis is that such embeddings are not good enough to capture the long-range interactions of zinc complexes. In our opinion, the partial charges should be also considered since they are an essential part of the long-range interaction. We first proposed a simple but efficient way to embed the partial charges. The inputs of the model are the atom types, represented by the nuclear charges, $Z_i \in N$, the Cartesian coordinates $r_i \in R^3$, as well as the total charge of each complex, $Q \in Z$. Initially, the single nuclear charge number of each atom is transformed to high-dimensional features to get the nuclear embedding $x^0 \in R^F$, where F is the number of features. Instead of feeding the nuclear embedding into EwaldMP, in this work, the nuclear embedding was further transformed to partial charges via well designed neural networks,

$$q = \sigma(Wx^0 + b) \tag{16}$$

where σ is the activation function, W and b are trainable parameters. To better model the electronic structures, these partial charge embeddings are then scaled to make sure the sum of these partial charges are equal to the total charge Q via

$$q_s = q + (Q - \sum_{i}^{N} q_i)/N \tag{17}$$

$$\hat{q} = residual(q_s) \tag{18}$$

where N is the number of atoms. A two-layer residual block is used to avoid gradients vanishing or exploding.[159] These predicted partial charges are then input into EwaldMP for modelling the long-range interactions. To test performances of the proposed partial charge embedding on predicting the long-range interactions, three types of models were run on the Zinc_60 data set: i) Baseline. Only short-range interactions are covered in the baseline models. In this work, we used SchNet and PAINN as baseline model. Please see Chapter 2.2 for more details about both models;(ii) Baseline+EwaldMP. Both the baseline and EwaldMP share the same nuclear embedding as the inputs of the model. This is the original design in EwaldMP in which baseline model covers short-range interactions while EwaldMP cover long-range interactions;(iii) Baseline+EwaldMP_Q. In this case, the baseline model still accepts the nuclear embedding as input, while EwaldMP now accepts the partial charge embedding. The adapted EwaldMP was named as EwaldMP_Q for clarity in this work. The whole process is illustrated in **Figure 15**.

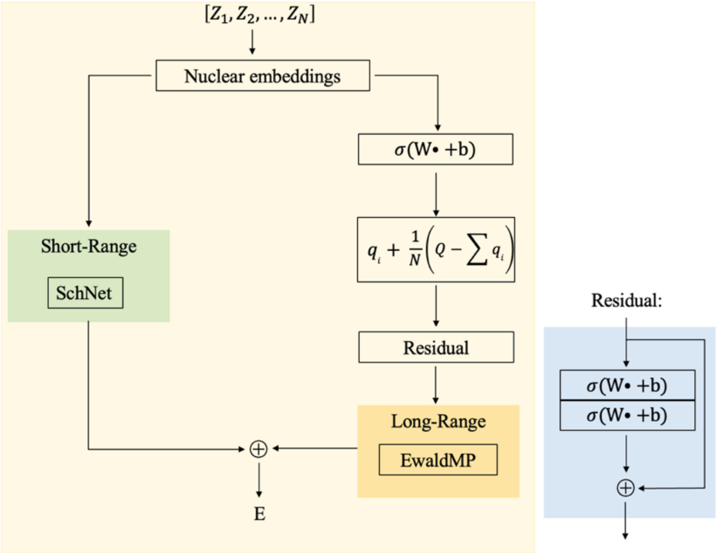


Figure 15. The schematic model used in this work.

The initial learning rate of all models is 5×10^{-4}. The batch size of SchNet-related models is 64. The warmup technique is used with a warmup factor of 0.2 up to the first 30000 steps and decays at steps of 60000, 90000, 120000 with a decay factor of 0.1. For the PAINN-related models, the batch size is 32 and the Adam optimizer is used with the plateau scheduler. All data and code are freely available at https://github.com/Neon8988/Zinc_NNPs.

2.5 Results and Discussion

We first compared the performance of the three types of models. The total potential energy of each conformation in the test set were predicted by each model and the results are given in **Table 2**.

Table 2. Mean absolute errors (MAE) for energy predictions in kcal/mol.

Model	MAE
SchNet	1.20
SchNet+EwaldMP	2.42
SchNet+ EwaldMP_Q	0.92
PAINN	1.36
PAINN+EwaldMP	1.50
PAINN+ EwaldMP_Q	1.02

Our proposed model, i.e., baseline+EwaldMP_Q yield the lowest errors which are 0.92 kcal/mol and 1.02 kcal/mol, respectively. Compared to the baseline models, our proposed models decrease the error by 23.33% and 25.00% with regard to SchNet and PAINN. The results also show that the importance of long-range interactions in modeling zinc complexes. And as the partial charges play an important role in long-range interactions,[152,160] accurately predicting the partial charges is necessary in NNPs for zinc complexes. The original EwaldMP takes the nuclear embedding as inputs to make sure atoms of the same type are initialized identically. But the global attribute is completely ignored. In contrast, we further transform the nuclear embedding into partial charges via neural network and scale them to ensure the sum of partial charges match the global charge of the complex.

This implementation gives some physical meanings to the NNPs to better model the potentials. The results indicate the original EwaldMP raises the largest errors in both cases, which increases the error by 101.67% and 10.29% from the baseline model, respectively.

To further evaluate our proposed EwaldMP_Q scheme, we compared this NNP with some widely used semiempirical methods, *i.e.*, GFNn-xTB(N=1, 2),[161,162] PM6-D3H4X and PM7.[163-165] To achieve this, we extracted 3838 conformations from the test set. Not all conformations were used for this evaluation since some configurations has only one conformation in the test set which is not applicable to compare the relative energies in terms of different tested methods. GFNn-xtb calculations were conducted in the *xtb*[166] package (version 6.6.1). The PM6-D3H4 and PM7 calculations were performed using version 22.0.6 of the MOPAC program[167]. We then calculated the conformational energies at the double hybrid PWPB95[168]/CBS (def2-TZVPP/def2-QZVPP) level with the D4 correction based off its known accuracy for TMCs.[169] These calculations were performed in ORCA 5.0.4.[150] with *TightSCF* and the RIJK approximation.[170] The PWPB95 method is used because it is very accurate in predicting the conformational energy of TMCs. Hence, it can be used to validate whether the reference method (r^2SCAN-3c) we used to curate our data set is good enough. And it can also indicate how good our

ML method is compared to this double-hybrid method which is accurate but very expensive. The results are given in **Table 3**.

Table 3. Performance of all tested methods on 3838 conformations with respect to PWPB95-D4/CBS method.[a]

Method	MAE	Count
r^2SCAN-3c	0.65	150
SchNet+ EwaldMP_Q	1.32	529
GFN1-xTB	2.15	764
GFN2-xTB	2.35	970
PM6-D3H4X	2.39	796
PM7	2.41	782

[a]The mean absolute errors (MAE) are given in kcal/mol. The count is the number of conformations which have lower energy than the actual most stable conformation, i.e. the lowest energy conformation at PWPB95-D4 level.

The proposed NNP clearly outperforms these semiempirical methods with a MAE value of 1.32 kcal/mol. And it is expected that it is worse than the r^2SCAN-3c method because the latter is the reference method of our NNP. The overall trends of these relative conformational energies are given in **Figure 16**. The reference PWPB95-D4 method is along the x-axis while the various tested methods are along the y-axis. The Pearson correlation coefficient (r_p) is also reported. The first subfigure shows the correlation between PWPB95-D4 and r^2SCAN-3c. And expectedly, it is generally good with a MAE of 0.65kcal/mol which indicates this method's reliability for curating the reference data of NNPs. And our ML method follows a similar trend. A detailed example of the relative conformational energy is given in **Figure 17**. The r^{2}SCAN-3c method and the proposed NNP follow a generally similar trend as PWPB95-D4, while the relative energies evaluated by these semiempirical methods fluctuate dramatically.

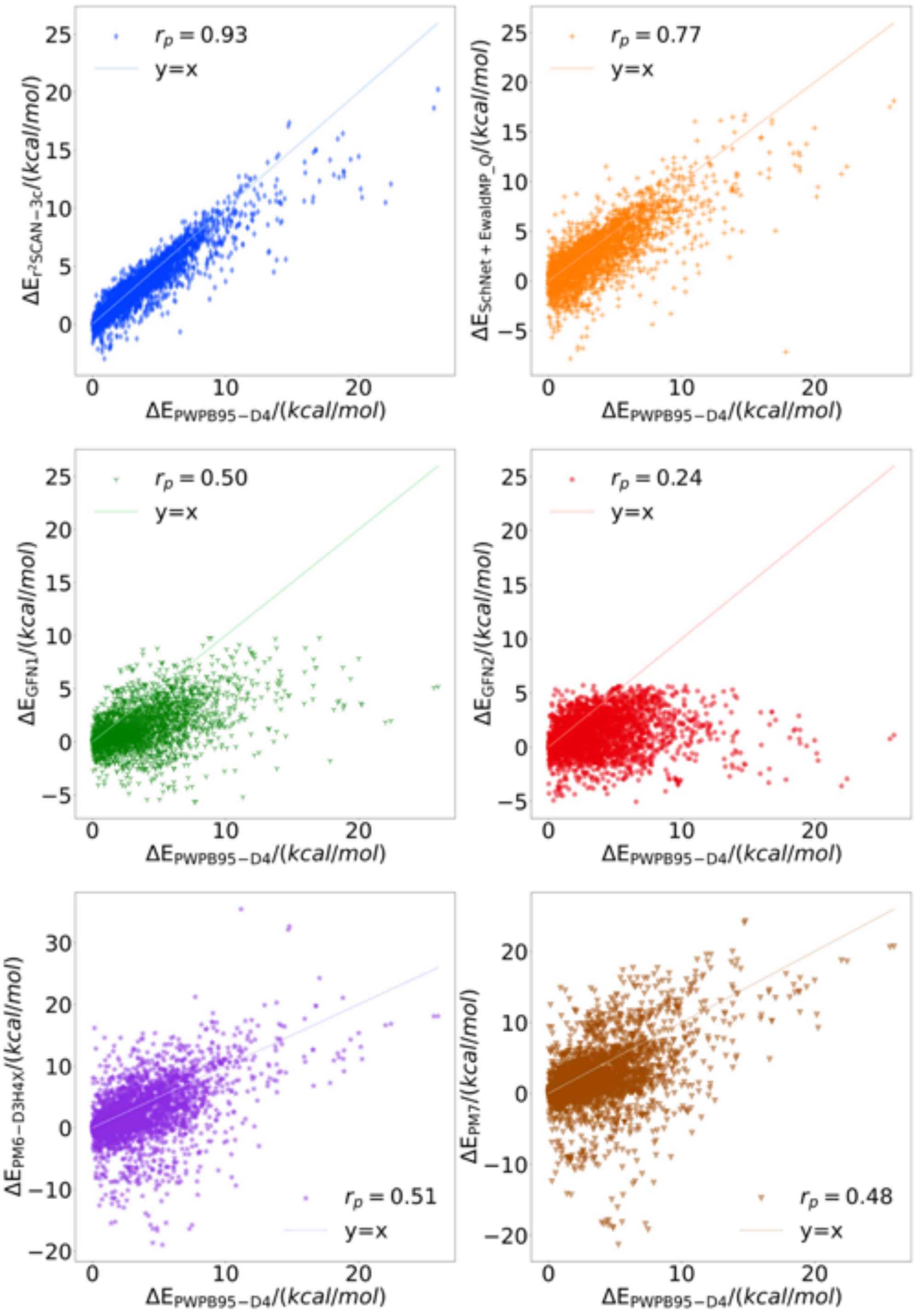


Figure 16. The relative conformational energies in all tested methods with reference to the PWPB95-D4 method.

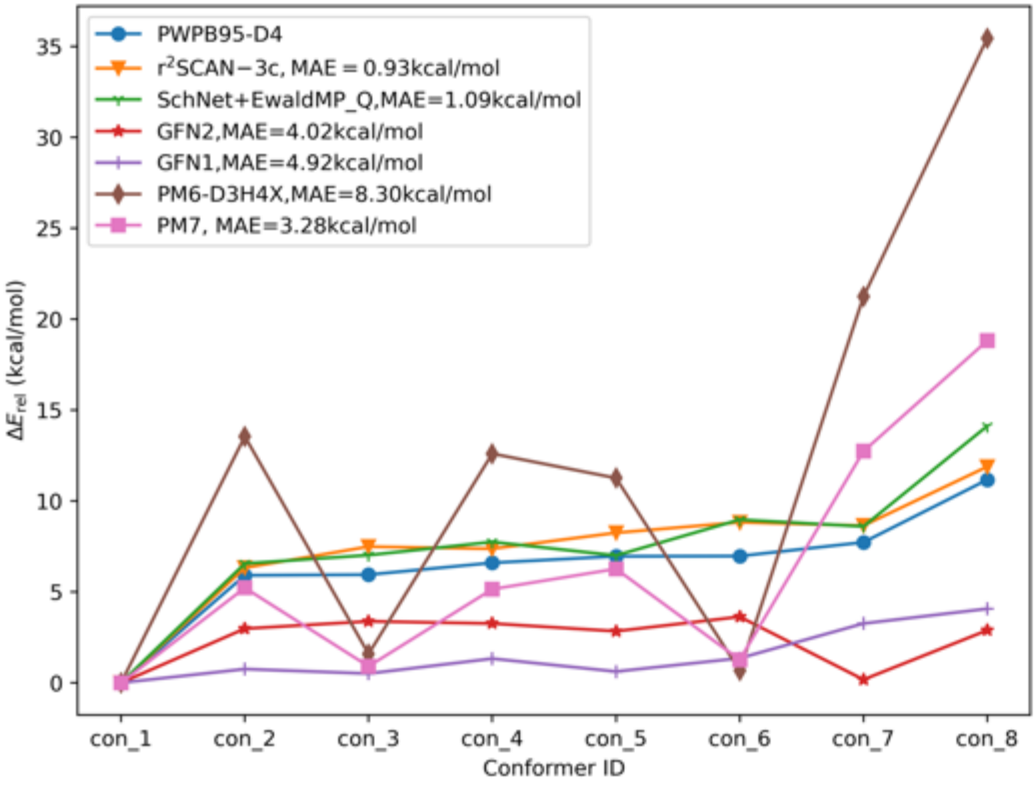


Figure 17. The relative energies of an ensemble with 8 conformations (CSD code: YUMWOT).

We then did some qualitative analysis. The 3838 conformations were grouped into 418 ensembles, each of which has more than one conformation with the same configurational structure. Each ensemble was individually ranked by energy. The lowest energy conformation of each ensemble at the PWPB95-D4 level was regarded as the most stable conformation. We then counted the number of conformations which has lower energy than the reference geometry ranked by each method. This analysis indicates the possibility of each method to incorrectly pinpoint the most stable conformation. As indicated in **Table 3**, for the r^{2}SCAN-3c method, only 150 conformations, *i.e.*, the number of y-axis values below 0 in **Figure 16**, were incorrectly predicted to have lower energy than the actual lowest energy conformation at the PWPB95-D4 level, and the number of such cases in NNP is 529. However, in the semiempirical methods, the MAE is larger than 2kcal/mol, and the number of incorrect low energy conformations is more than 700, among which PM7 yields the largest MAE of 2.41 kcal/mol, while GFN2-xtb yields the largest number of incorrect low energy conformations (970 out of 3838 conformations). Interestingly, the GFNn-xtb method has a bias to lower the relative energies, as shown in **Figure 16**, the y-axis range in GFNn-xtb is around 10 kcal/mol, while the reference relative energy range, *i.e.,* the x-axis range is around 25 kcal/mol, but other tested methods generally match this range well. Overall, the results indicate that our ML method outperforms these semiempirical methods in both qualitative and quantitative evaluations.

Finally, the computational efficiency of each method is generally evaluated. our proposed ML method can greatly reduce the computation cost. It only took 10 seconds to predict these 3999 conformations on a single GPU, while the average time of r^{2}SCAN-3c in our test set was 4.46 minutes running on 4 CPU processors for each geometry. For the PM6-D3H4X method, the average time is 0.97s, while it took 0.94s for PM7 method. And the GFN1-xtb method took 6.65s on average, while GFN2-xtb took 10.82s for each geometry.

2.6 Conclusions

In the present study, a high-quality data set, Zinc_60 which includes both configurationally and conformationally diverse zinc structures was curated. The metadynamics was used to generate the conformations, aiming at covering a wide potential energy space. This data set was then used to train a ML model to model the PES of zinc complexes. To accurately model long-range interactions, EwaldMP_Q was proposed to introduce partial charges to a neural network based on Ewald summation. The results indicate the usefulness of this method in both types of models which shows the generality of this method and the possibility to apply this method to other baseline models in the future. Moreover, this NNP outperforms some widely applied semiempirical methods but at much less cost in both qualitative and quantitative evaluations.

CHAPTER 3 MODELING Fe (II) COMPLEXES USING NEURAL NETWORKS

3.1 Fe (II) Complexes

TMCs with d^4 to d^7 configurations have multiple electron configurations. Depending on the ligand types, coordination number, and the transition metal itself, TMCs can exist in either low-spin (LS) or high-spin (HS) state. And sometimes transition metals with d^5 or d^6 electron configuration also have an intermediate-spin state. The spin state indicates the quantity of unpaired electrons the central metal has and the energetically favorable electron configuration. An important contribution to the energy difference between different spin states is vibrational effects. The stretching vibrations between metal and ligands usually have lower frequencies in HS state than in LS state. Compared to LS state, the HS state also has higher entropy but lower zero-point energy. Both energy terms also give rise to the free energy difference.[171] But since the energy difference is minimal, both HS state and LS state can interconvert flexibly, which is called thermal spin crossover (SCO).

Many SCO complexes with Fe (III),[172] Mn (III),[173] Co (II),[174] Fe (II)[175] as the metal center have been synthesized. Among these reported complexes, Fe (II) compounds are prevalent because the variations of their physicochemical property along with spin state change is more prominent than other TMCs. For example, during the spin state exchange, the Fe-N distance varies a lot (Δd_{Fe-N}~0.2Å), resulting in a dramatic change in the unit-cell volume. In addition, the color of materials may abruptly change from light color, e.g. yellow, green to dark color, such as blue, red, or pink. And the magnetic property changes between diamagnetic ($t_{2g}^{6}e_{g}^{0}$, S = 0) and paramagnetic ($t_{2g}^{4}e_{g}^{2}$, S = 2).[176] The SCO-active Fe (II) complexes usually possess a N_6 coordination sphere in octahedral shape.[177] Some frequently observed ligands, such as 1*H*-1,2,4-triazole[178], 2,6-bis(pyrazol-1-yl) pyridine,[179] tris (2- pyridylmethyl) amine,[180] are widely used to synthesize the Fe (II) SCO complexes. Such N-coordinating ligands are reasonable choices because they are abundant and easy to synthesize and functionalize.[181] The SCO process appears in a variety of chemistry areas.[182,183] For example, it occurs in metal-ligand bond dissociations, e.g. O_2 binding to hemoglobin.[184] And small ligands, such as H_2O, NO, CO only binds to heme group under specific spin state.[185] Meanwhile, it can also induce emergent superconductivity in iron-based honeycomb lattic.[186] In addition, the optical behavior is also observed.[187,188] And the SCO process can be activated under various external stimulus, such as pressure perturbation, temperature variations, light radiation.[189,190]

Precisely calculating the spin splitting energy, i.e., the energy gap between both spin states is crucial. First, the spin-state energetics determines the ground spin state, magnetic properties, and the possibility of SCO. More importantly, different spin states result in different ligand-activation propensities[191] and chemical reactivities.[192] However, the spin-state energetics heavily depends on the electron correlation effects. For instance, the Hartree-Fock (HF) theory, where only includes exchange without any other correlation effects, strongly biases HS state. And the electron correlation included DFT, or wave function theory (WFT) methods can effectively mitigate this bias but at a much more expensive computational cost. Another underlying issue is the effect of conformations on the ground spin state. Usually, researchers randomly select one geometry for each spin state, and calculate the relative energy to determine whether the investigated complex has SCO property. But such simple calculation is not enough, or it can only decide that the investigated conformations of the given complex have SCO property or not but it could not conclude that the complex in the given configuration has SCO property or not since the relative energy of different conformations in this given configuration may vary a lot. For example, in the present work, we used CREST to generate 78 conformations for a Fe (II) complex (CSD code: WIWBEK). And among them, 44 conformations are in HS state, while the remaining are in LS state. For more details, please see section 3.2. **Figure 18** shows the relative energy of these 1496 pairs (44×34). As we can see, the relative energy fluctuates from -10.21 kcal/mol to 54.92 kcal/mol. Therefore, the ground spin state is heavily influenced by the conformation. To expedite the high-throughput screening of Fe (II) SCO complexes, this present work designed a neural network model to efficiently and accurately locate the ground spin state.

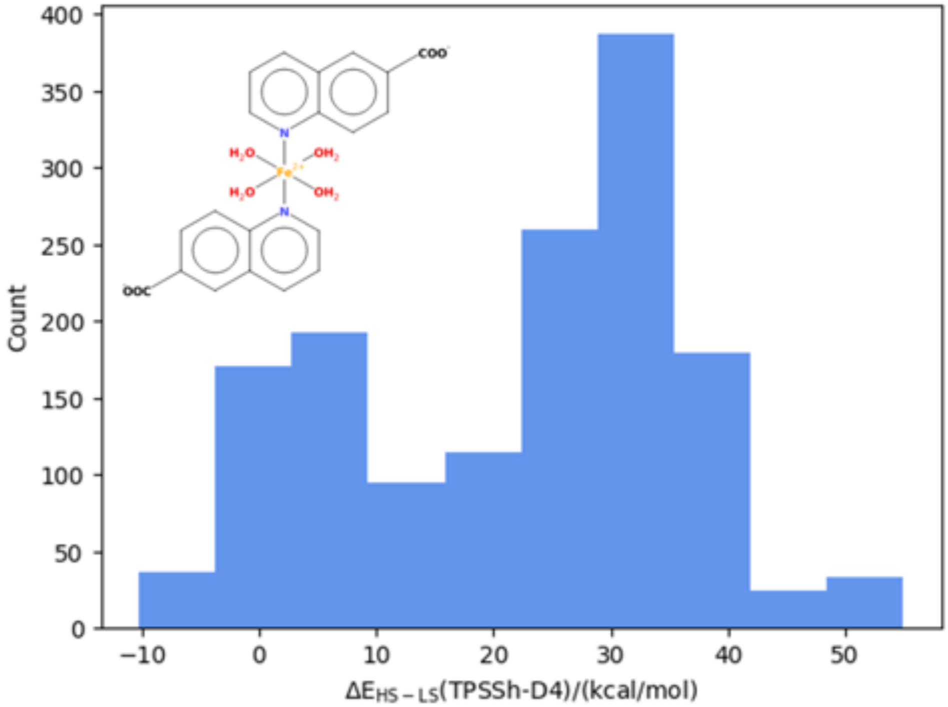


Figure 18. The relative energy of WIBEK complex.

3.2 Fe (II) Data Set Curation

A data set of over 240,000 crystallized mononuclear TMCs extracted from The Cambridge Structural Database (CSD)[147] was reported by Aditya and coworkers.[193] Well-defined Fe (II) complexes were collected from this data set by following the procedures below: (i) both oxidation states and charges were predetermined by the structure uploader; (ii) no hydrogen atoms were lost in the structure. Eventually, a subset of 383 unique Fe (II) complexes with 80 atoms or less was curated in this present work. Some representative structures are given in **Figure 19**. The size distribution of these 383 complexes is given in **Figure 20**. As shown in **Figure 21**, the element types in this subset include H, C, N, O, S, Cl, P, Fe. Most complexes have +2 charge (**Figure 22**). These complexes also cover a variety of coordination patterns (**Table 4**).

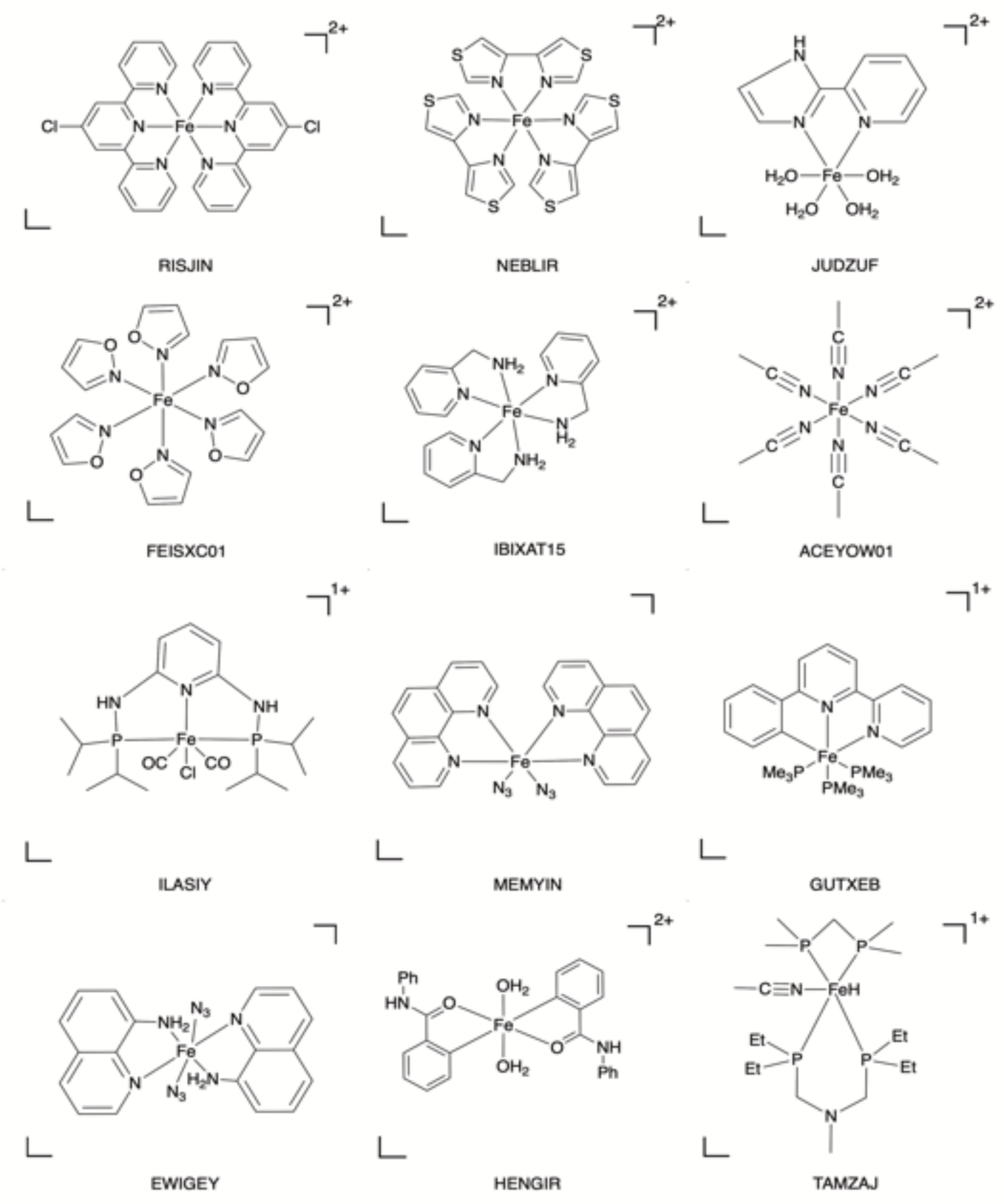


Figure 19. Fe (II) complex examples in Fe (II)_80 dataset.

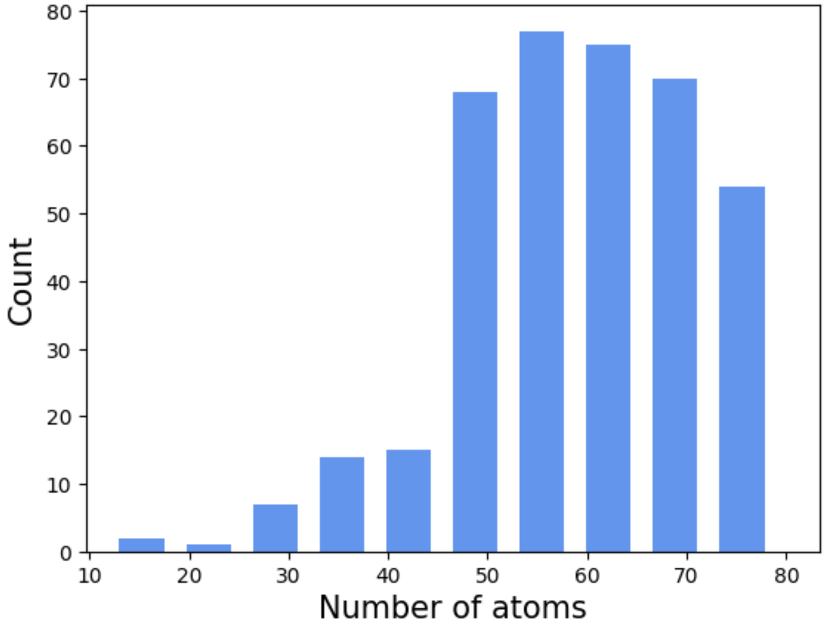


Figure 20. The molecular size of the 383 complexes.

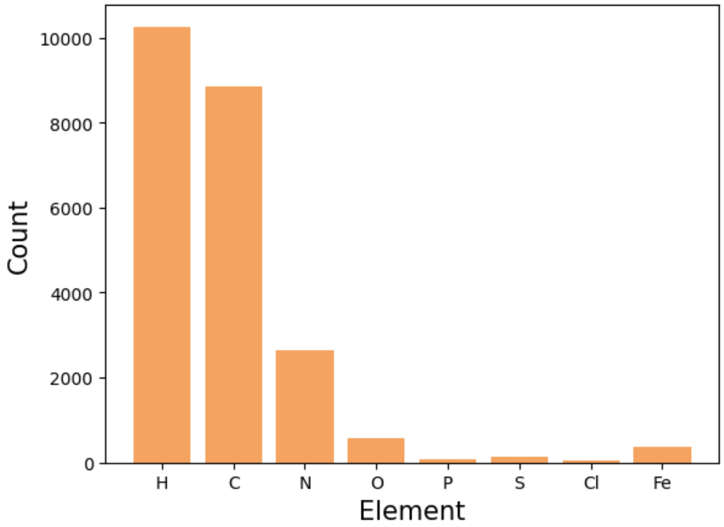


Figure 21. The element distribution in the 383 complexes.

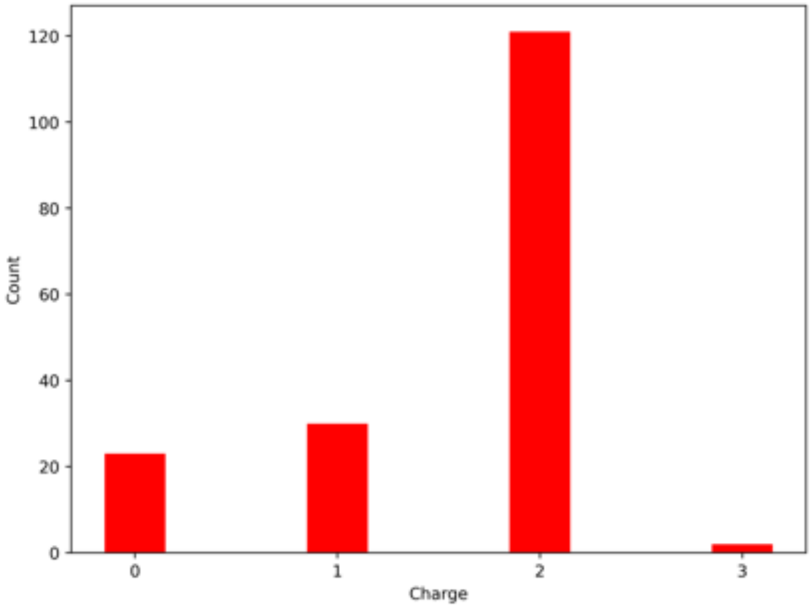


Figure 22. The charge distribution of the 383 complexes.

Table 4. The denticity types of the 383 complexes.

Denticity type	Counts
6	51
5,1	53
4,2	15
4,1,1	27
3,3	115
3,2,1	1
3,1,1,1	15
2,2,2	27
2,2,1,1	35
2,1,1,1,1	4
1,1,1,1,1,1	40

We then followed the strategy we developed in Chapter 2 to generate conformers for each Fe (II) complex. Specifically, we designated a HS state and a LS state for every configuration, then employed CREST to obtain conformers specific to each spin state. Conformations with a RMSD of 0.1 Å or less were excluded. And each pair was recentered and then rotated unto each other to get the real structural differences. The B97-3c method[194] was utilized to optimize the geometries. All optimizations were with the *TightSCF, DEFGRID3*, *SOSCF*, and *SlowConv* settings using ORCA 5.0.4. Geometries were removed if they met any of the following criteria: (1) convergence was not achieved during optimization, (2) the presence of an imaginary frequency was detected after optimization, (3) the discrepancy between the anticipated $\langle \hat{S}^2 \rangle$ and the actual value exceeded $1\mu_B$. Finally, the curated dataset for Fe (II) complexes, designated as Fe (II)_80, comprised a total of 15568 geometries in HS state and 13266 geometries in LS state. The size distribution and the element distribution in each spin state are given in **Figure 23** and **Figure 24**, respectively. A representative conformation ensemble is given in **Figure 25**. Certain DFT methods might demonstrate a bias towards HS or LS states, arising from the specific formulation of each functional. In this present work, TPSSh functional[195] was utilized as the reference method due to the exceptional cabilities it demonstrated across various evaluations.[196-199] The ultimate energy assessments for the structures were executed with the TPSSh-D4 functional with the def2-TZVP[200] basis set via ORCA 5.0.4 with the *TightSCF* setting. To expedite the computational process, the RI-J approximation[201] was employed in conjunction with the def2/J[202] auxiliary basis set. The totally 28,834 geometries included in the Fe (II)_80 data set was divided at random into three distinct sets: a training set comprising 23,834 samples, a validation set with 2,500 samples, and a test set also containing 2,500 samples.

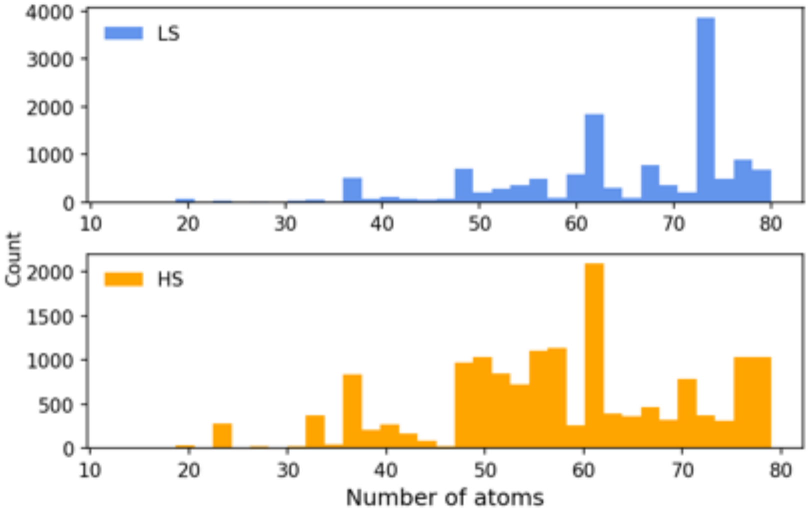


Figure 23. The molecular size distribution of 28834 geometries.

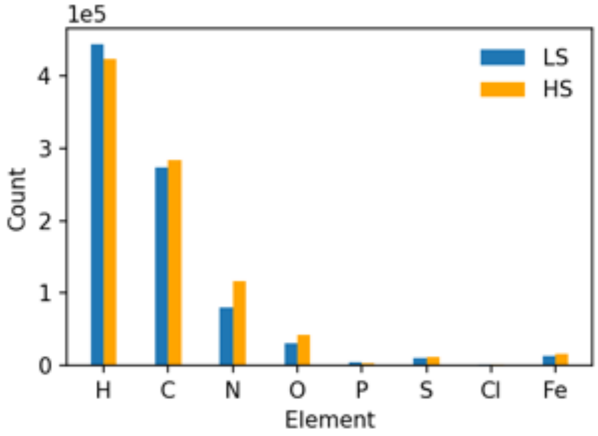


Figure 24. The element distribution of 28834 geometries.

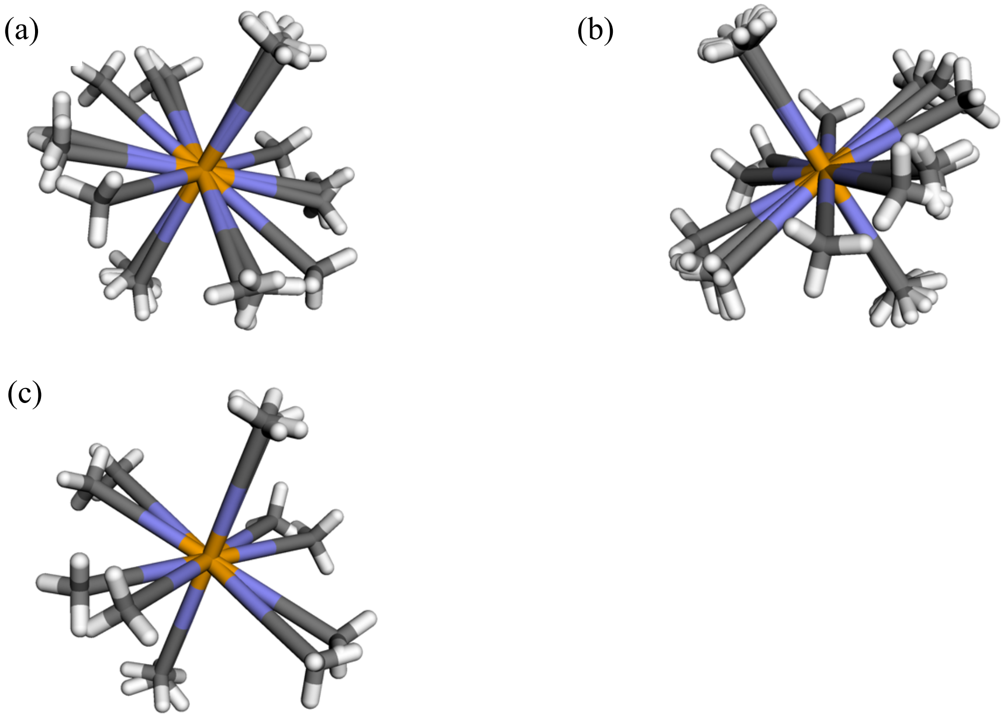


Figure 25. (a) 3 HS state conformers (refcode: ACEYOW01) (b) 4 LS state conformers (refcode: ACEYOW01). (c) The minimal energy conformation for both high-spin and low-spin state. ΔE_{HS-LS} = 12.45kcal/mol. (refcode: ACEYOW01).

The unique CSD code of 383 complexes and the number of conformations in Fe (II)_80 data set is given in **Table 13**(See APPENDIX A: TABLES).

3.3 Fe_NNPs

In the present work, both charge and spin state were incorporated into our proposed model to better predict the nonlocal interactions. The inputs contain the nuclear charge $Z_i \in N$, the atomic coordinates $r_i \in R^3$, the total charge $Q \in Z$ and the spin state $S \in Z$. Electronic properties Q and S as well as nuclear charge were subsequently converted into high-dimensional features. The atomic representation $x^0 \in R,$ has two components: (1) the nuclear embedding $x_N^0 = x_z^0 + x_{e_z}^0$, where the atom-type embedding x_z^0 as well as the atomic electron-configuration embedding $x_{e_z}^0$ depend on the atom types; (2) the electronic embedding $x_E^0 = x_Q^0 + x_S^0,$ where x_Q^0 is the charge embedding and x_S^0 is the spin state embedding. The complete atomic embedding is

$$x^0 = x_z^0 + x_{e_z}^0 + x_Q^0 + x_S^0 \tag{19}$$

where x_z^0 and $x_{e_z}^0$ are embedded via a look-up table based on the atom types. For x_Q^0 and x_S^0, SpookyNet[203] uses the attention mechanism.[204] In this work, we simplified the mappings through only scaling the spin state embeddings and charge embeddings,

$$q = MLP(x_z^0 + x_{e_z}^0) \tag{20}$$

$$e = Softplus(q * MLP(s)) \tag{21}$$

$$\tilde{e} = e + (s - \sum_i^N e_i)/N \tag{22}$$

$$x_s^0 = residual(\tilde{e}) \tag{23}$$

where s = Q is charge embeddings, and s = S is spin state embeddings. A detailed workflow is shown in **Figure 26**. The total charge is initially distributed equally among all atoms as the partial charges, which are subsequently multiplied by x_N^0 to differentiate the significance of partial charge. At the final step, the partial charges are scaled to ensure their summation matches Q. The spin state S is processed similarly to derive the spin state embeddings.

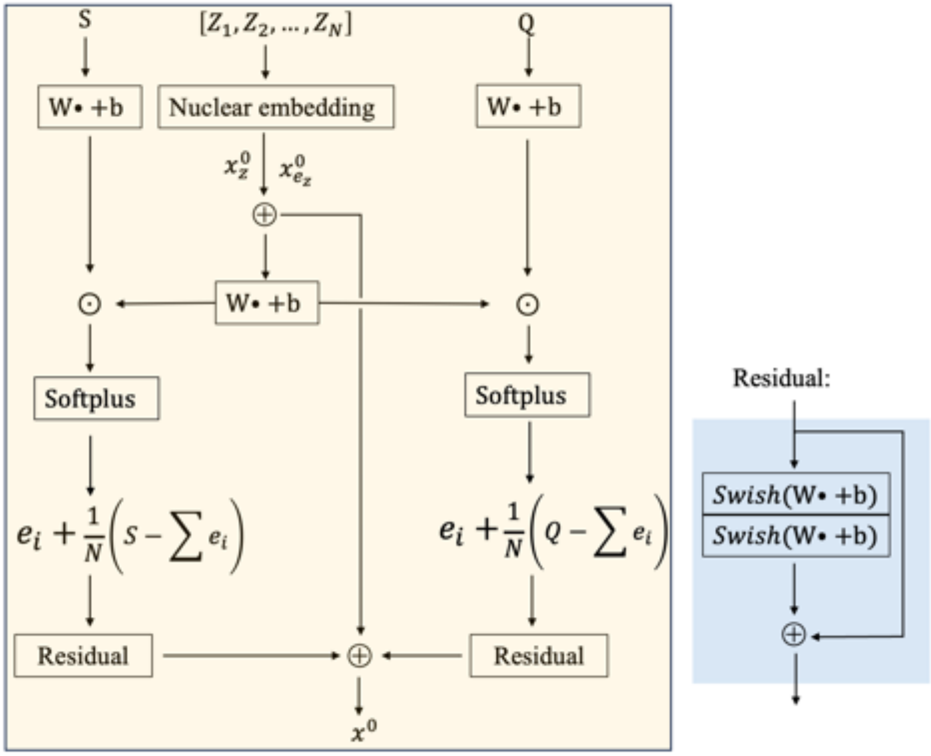


Figure 26. The workflow of the initial embeddings x^0.

Several types of models were run in the present work. First, the electronic embeddings x_E^0 were tested. Three types of atomic embeddings including the SpookyNet electron embeddings, the scaled embeddings as well as the sole nuclear embedding were compared. For the baseline model, which only covers short-range interactions, we tested SchNet and PAINN. Compared to SchNet, PAINN uses extra vector representations to model the PES. Third, as Zinc_NNPs, EwaldMP was compared with the baseline models to test whether it can model the nonlocal interactions in Fe (II) complexes. We evaluated it in the following ways: (1) the entire embedding x^0 was fed into both models. EwaldMP is a standalone module to combine with any short-range model, resulting in a connected model that encompasses both short-range interactions and long-range interactions; (2) the nuclear embeddings x_N^0 were input to base model, either SchNet or PAINN, while the electron embeddings x_E^0 were passed into the EwaldMP. In this scenario, both models underwent independent updates at each iteration.

The batch size of 16 was used in all models with learning rate, lr = 5×10^{-4}. The warmup technique was used with a warmup factor of 0.1 up to the first 50000 steps and decreased at steps of 150000, 250000, 350000 with a decay factor of 0.1. For the PAINN-related models, the Adam optimizer was used with the plateau scheduler.

3.4 Results and Discussion

Each model's capability of estimating the total energy and the splitting energy was tested. Among 2500 Fe (II) conformations in the test set, 121 configurations have both HS state and LS state conformations. Specifically, totally 1075 conformations are in HS state and the number of conformations in LS state is 654. To compare the splitting energy (SE), totally 23446 pairs from the test set, i.e., every pair contains a HS spin state conformation and a LS spin state conformation, and both have the identical structural configuration. The mean absolute error (MAE) of each model is listed in **Table 5**.

Table 5. MAE for the total energy and the splitting energy predictions in eV, respectively. Best result in bold.

Model[a]	With x_E^0				Without x_E^0	
	SpookyNet		Scaled		Only x_z^0	
	energy	ΔE_{HS-LS}	energy	ΔE_{HS-LS}	energy	ΔE_{HS-LS}
SchNet	0.045	0.036	**0.037**	**0.030**	0.140	0.118
SchNet+EwaldMP	0.083	0.068	0.083	0.070	0.128	0.099
SchNet, EwaldMP	0.048	0.038	0.050	0.039	–	
PAINN	0.189	0.108	0.173	0.127	0.128	0.120
PAINN+EwaldMP	0.192	0.127	0.176	0.113	0.119	0.097
PAINN, EwaldMP	0.149	0.125	0.106	0.094	–	

[a] '+' indicates base model and EwaldMP have the identical embedding and ',' indicates the nuclear embeddings are input to base model and the electronic embedding are fed to EwaldMP.

The results show that the introduction of electronic embedding x_E^0 significantly enhances the capabilities of these models. Moreover, the scaled embedding outperforms the SpookyNet embedding. In terms of SchNet-related models, the scaled embedding results in the lowest MAE of 0.037 eV for the total energy and 0.030 eV for SE. The attention-focuesed electronic embedding results in a marginally higher MAE of 0.045 eV and 0.036 eV, respectively. But both types of embeddings enhance the accuracy of modelling Fe (II) complexes, as evidenced by the fact that removing them results in significantly larger MAEs of 0.140 eV and 0.118 eV, respectively. When the electronic embedding x_E^0 is not utilized, the baseline PAINN model slightly outperforms SchNet in predicting total energy, reducing the MAE from 0.140 eV to 0.128 eV. But with regard to SE, both baseline models achieve almost the same MAE of 0.120 eV. Finally, if only x_Z^0 is considered, the combined models, i.e., the baseline+EwaldMP decrease the MAE by around 0.01 eV and 0.02eV in terms of SchNet and PAINN, respectively. For instance, for the total energy predictions, the SchNet+EwaldMP combined model yields a MAE of 0.128 eV while pure SchNet yields a MAE of 0.140 eV. With the electronic embeddings x_E^0, simply add the Ewald message passing to the baseline model as another contribution, is not the best choice for modelling the PES of Fe (II) complexes. Since the electronic embeddings x_E^0 already cover the long-range interactions, simply concatenate two models together and sharing the same complete embedding can cause the interactions to overlap. To overcome this issue, these electronic embeddings x_E^0 should be passed into EwaldMP separately. As a result, the nuclear embeddings x_N^0 models the short-range interactions, while the electronic embeddings x_E^0 reproduce the nonlocal interactions. For instance, with the scaled embeddings, the MAE value of SE decreases from 0.070 eV (SchNet+EwaldMP) to 0.039 eV (SchNet, EwaldMP), along with the total energy error from 0.083 eV (SchNet+EwaldMP) to 0.050 eV (SchNet, EwaldMP). These comparisons indicate that EwaldMP can model the long-range interactions well. But the most efficient way is to just pass the complete embedding x^0 into SchNet. With the scaled embeddings, this model can cover the long-range interactions even better than EwaldMP but at much less cost, give the lowest MAE of 0.037 eV for the total energy predictions, along with 0.030 eV for SE error. The computational time of each type of model is given in **Table 6**.

Table 6. The training time of each model.

Model	Time/h		
	With x_E^0		Without x_z^0
	SpookyNet	Scaled	Only x_z^0
SchNet	5.97	5.88	5.11
SchNet+EwaldMP	11.48	11.43	10.81
SchNet, EwaldMP	11.69	11.60	–
PAINN	8.60	8.51	7.54
PAINN+EwaldMP	13.05	12.97	12.43
PAINN, EwaldMP	13.00	13.05	–

As shown in **Table 6**, although the additional scaled embeddings are incorporated into SchNet, since the embeddings are quite simple to compute and do not need many extra parameters, our adapted method (5.88hr) does not take much longer time than the original SchNet (5.11hr). Hence, our method is very efficient.

Next, we compared the ML method with a couple of widely used semiempirical methods since the computational cost of all these methods is roughly at the same level. Recently, Hagen and co-workers designed the newly spin-polarized (sp)GFNn-xTB(n=1,2)[205] as an extension of the GFNn-xTB (n=1,2) methods to differentiate the spin states of TMCs. We also tested PM6-D3H4 as well as the PM7 method. (sp)GFNn-xTB(n=1,2) calculations were conducted using *xtb* version 6.6.1. The PM6-D3H4 and PM7 calculations were performed using MOPAC, version 22.0.6. All results are given in **Table 7**. We report the MAE of the splitting energy in eV as well as the number of correct spin states predicted as a qualitative analysis. In these semiempirical methods, some geometries were excluded due to job failures. In this extensive test, we found that the semiempirical methods did not predict the splitting energy nor the correct spin state very well. The splitting energy errors are consistent with the results tested on the TM90S benchmark set.[205] In contrast, the SchNet model with the scaled embeddings only predicted 8 incorrect ground spin states with a MAE of 0.030 eV. The splitting energy predictions of each method is given in **Figure 27**. An overall trend is given in **Figure 39** (see APPENDIX B: FIGURE). A detailed example is given in **Figure 28**. This complex (CSD code: CODQAO) has 6 conformations in HS state, while 3 conformations in LS state. For these totally 18 pairs, the ML method make good predictions for the splitting energy which match the reference method well. However, these semiempirical methods increase the errors by orders of magnitude.

Table 7. Performance of the ML model and all tested semiempirical methods on the spin state splitting.

	Count[b]	ΔE_{HS-LS}[c]
SchNet[a]	23438/23446	0.030
PM6	6724/23307	2.8904
PM7	9757/23428	2.1062
spGFN1	5539/23428	3.5372
spGFN2	4407/23446	3.7195

[a]The SchNet baseline model with the scaled electronic embeddings is used as a comparison with these semiempirical methods. [b]The number of correct spin states predicted. Since some systems could not run successfully in these semiempirical methods, the total numbers differ. [c]The MAE value is given in eV.

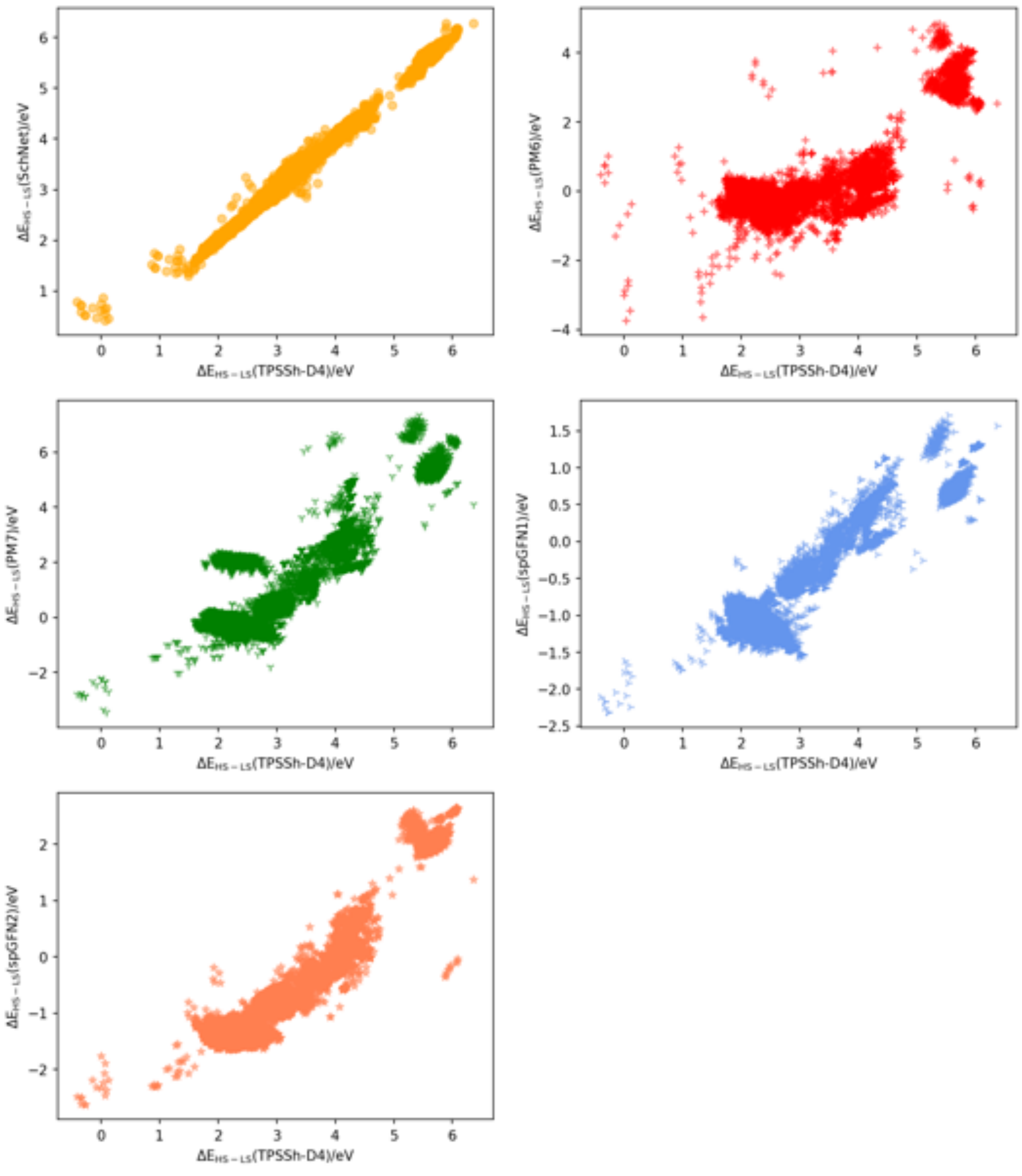


Figure 27. The splitting energy predicted by each tested method with regard to TPSSh-D4.

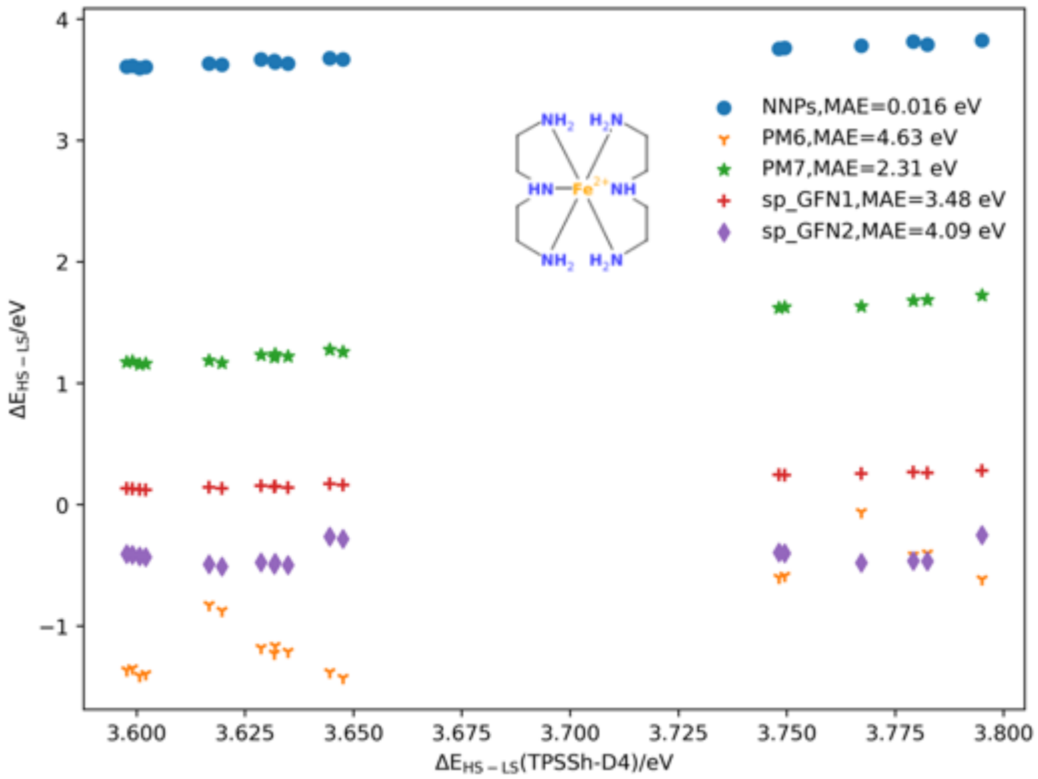


Figure 28. The splitting energy of Fe (II) complex (CSD code: CODQAO).

3.5 Conclusions

In the present study, a high-quality data set, Fe (II)_80 which includes both configurationally and conformationally diverse iron structures was curated. More importantly, the spin-state-specific conformations were generated by using metadynamics. This data set was then used to train a ML model to model the PES of Fe (II) complexes. To accurately model long-range interactions, the electronic properties which include the total charge, and the spin state were introduced to the baseline model. The results indicate the usefulness of this method in both types of models which shows the generality of this method and the possibility to apply this method to other baseline models in the future. Moreover, this NNP outperforms some widely applied semiempirical methods but at much less cost in both qualitative and quantitative evaluations.

CHAPTER 4 3D TRANSITION METAL COMPLEXES GENERATION

4.1 Generative Models

Generative models, a type of unsupervised ML model, aim at producing unseen, new representation of a targeted variable based on probability distribution. The goal of generative modeling is to learn the joint probability distribution of a data set. For example, in terms of language generation, a generative model learns from the joint distribution to determine the likelihood of the occurrence of a particular set of words and phrases in specific context. From the probability distribution, generative models learn the patterns and structures in a given dataset and synthesize new content.

Generative models have a wide range of applications, such as image synthesis, language translation, text-image translation, inpainting, *etc*. For instance, one latest class of generation AI for context generation is large language models (LLMs). The LLMs can create artificial text context and large chunks of sentences. But these generated sentences are not incoherent rambling or irrelevant words strung together. Instead, they can understand intricate concepts about this physical world and conduct some challenging tasks, like summarizing a book, writing great essays, solving complicated math problems, and so on. A primary example of LLMs is ChatGPT. It is trained based on ~13T tokens compiled from

Wikipedia, books, academic publications, historic documents, webpages, etc. With such extensive context sources, ChatGPT can quickly give comprehensive and insightful responses to a wide array of prompts, encompassing topics from the annals of history to the complexities of philosophy, from the creative nuances of the arts to the empirical rigor of science. Another primary application of generative models is image creation. For instance, DALL·E is an advanced AI model that can synthesize intricate, detailed, and realistic images from textual descriptions. It enables users to transform their textual ideas into vivid visual representations. By bridging the gap between words and images, DALL·E can accelerate the imaginative exploration and artistic creation.

Diffusion models represent a novel category of generative models. They have ended the longstanding supremacy of generative adversarial networks (GANs) in a variety of fields, such as image synthesis,[206-208] natural language processing,[209, 210] computer vision,[211, 212] temporal data modeling,[213, 214] multi-modal modeling.[215, 216] More importantly, diffusion models also show remarkable performance in life science from small organic molecule generation[217, 218] to medical image reconstruction.[219, 220] Denoising Diffusion Probabilistic Models (DDPMs)[208] , inspired by non-equilibrium thermodynamics, are becoming a predominant class of diffusion models. The key to DDPMs is to progressively destroy data by adding noise randomly and subsequently new samples could be generated by successively removing noise. A DDPM works in a dual-phase mechanism with two Markov chains. In the first stage, random noise is sampled and then added to the clean input data point x_0 with predefined steps *T*. This process aims at transforming any complex data distribution of the data set into a simple predefined distribution, e.g., Gaussian. The technique behind this idea is that noise is sampled from a Gaussian distribution and is progressively added to the input data. With enough steps, the input data is covered by all sampled noise and lose the structural patterns. This first Markov chain is called the forward diffusion in DDPMs. And it generates an array of random variables $x_1, x_2, \cdots, x_{T.}$ At a specific step, denoted as $t = 0, \ldots, T,$ the intermediate noised data state x_t given the previous state is derived by the multivariate normal distribution,

$$q(x_t|x_{t-1}) = N(x_t|(\bar{\alpha}_t x_{t-1}, \bar{\sigma}_t^2 I) \tag{24}$$

where $\bar{\alpha}_t \in \mathbb{R}^+$ signifies the retained information, and $\bar{\sigma}_t \in \mathbb{R}^+$ controls the added noise. Furthermore, Sohl-Dickstein *et al.*[221]determined $\bar{\sigma}_t^2 = 1 - \bar{\alpha}_t$. With the chain rule of probability and the Markov property, the joint distribution of $x_1, x_2, \cdots, x_T$ given on x_0, denoted as $q(x_1, x_2, \cdots, x_T \; |x_0)$, is derived as

$$q(x_1, x_2, \cdots, x_T \; |x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{25}$$

Finally, a simple and unified formula for the intermediate state x_t given x_0 is derived as

$$q(x_t|x_0) = \mathcal{N}(x_t|\alpha_t x_0, \sigma_t^2 I) \tag{26}$$

where $\alpha_t = \prod_{i=1}^{t} \bar{\alpha}_i$. This closed-form formula indicates that any intermediate state can be obtained directly from x_0 instead of iteratively adding from x_0, which greatly simplifies the forward diffusion. Usually, the noise schedule α_t is set in advance, and it gradually shifts from $\alpha_0 \approx 1$ to $\alpha_T \approx 0$. And with sampling a Gaussian distribution $\epsilon \in \mathcal{N}(0, I)$, the transition state x_t is derived as

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon \tag{27}$$

Intuitively, x_T is sole noise, devoid of any embedded structural details. The second Markovian chain reverses the diffusion process by removing the added noise at each step. This reverse chain involves recognizing the specific noise patterns introduced at each step and denoising the data accordingly. The denoising process is also a closed-form formula, defined as

$$q(x_{t-1}|x_0, x_t) = \mathcal{N}(x_{t-1}|\mu_t(x_0, x_t), \sigma_{t\to t-1}^2 I) \tag{28}$$

along with the mean and variance are as follows

$$\mu_t(x_0, x_t) = \frac{\alpha_{t|t-1}\sigma_{t-1}^2}{\sigma_t^2} x_t + \frac{\alpha_{t-1}\sigma_{t|t-1}^2}{\sigma_t^2} x_0 \tag{29}$$

$$\sigma_{t\to t-1} = \frac{\sigma_{t|t-1}\sigma_{t-1}}{\sigma_t} \tag{30}$$

$$\alpha_{t|t-1} = \frac{\alpha_t}{\alpha_{t-1}} \tag{31}$$

$$\sigma_{t|t-1}^2 = \sigma_t^2 - \alpha_{t|t-1}^2 \sigma_{t-1}^2 \tag{32}$$

Eq. 28 indicates that any intermediate state x_t is an interpolation between x_0 and x_T in a diffusion trajectory. Then the true denoising/generative process starts with a predefined initial distribution $p(x_T)$,

$$p(x_T) = \mathcal{N}(x_T; 0, I) \tag{33}$$

and the goal of this process is to invert the diffusion trajectory, however x_0 in eq. 28 is not known. A neural network ϕ is then utilized. Following eqs.28 and 33 mentioned above, the denoising transition state can be derived as

$$p(x_{t-1}|x_t) = q(x_{t-1}|\hat{x}, x_t) \tag{34}$$

where $\hat{x} = \phi(x_t, t)$, an approximation of x_0. For better performance, ϕ is modified to estimate the noise, $\hat{\epsilon}_t = \phi(x_t, t)$.[208] The approximated $\hat{x}$ is obtained as

$$\hat{x} = (1/\alpha_t)x_t - (\sigma_t/\alpha_t)\hat{\epsilon}_t \tag{35}$$

The aim of this model is to reduce $\mathcal{L}(t) = \|\epsilon - \hat{\epsilon}_t\|^2$ by applying gradient descent techniques. After proper training of the model, new data points can be synthesized. By starting with any noise vector x_T drawn from a normal distribution $\mathcal{N}(0, \mathrm{I})$, the denoising process described by eq 34 is performed iteratively for steps *t* = *T*, . . ., *0.*

4.2 LigandDiff

Molecular generation is crucial for drug design and new materials discovery. Developing *de novo* drugs is an essential research field in chemistry, biology and life science. However, traditional theoretical methods are costly and time-consuming. For instance, the approximate cost of developing a new drug varies from \$314 million to \$ 2.8 billion and continues growing.[222] And the drug development process typically spans almost 12 years.[223] Estimates suggest that the synthesizable chemical space contains nearly 10^{23}~10^{60} potential drug-like molecules,[224] but so far, only about 10^{8} to 10^{10} of these molecules have been synthesized.[225] Identifying new drugs through exhaustive high-throughput screening is an incredibly time-consuming and laborious process. Additionally, relying solely on human intuition for searching small molecules can introduce biases, potentially overlooking novel compounds with optimal properties.

In recent years, generative models have proposed novel avenues for molecule design. A variety of models have been developed, including SMILES strings oriented generative models, such as the Sequence Autoencoder (seq2seq AE)-based[226] and the variational autoencoder (VAE)-based[227]; 3D full-molecule generation models, such as GraphRNN,[228] molGAN[229] and scaffold-based generative models, such as EMPIRE[230] and DeepScaffold.[231] All previous work achieved remarkable performance on generating small organic molecules. In this present study, a generative model, called LigandDiff was developed to generate novel ligands for TMCs. The design of novel TMCs is vital to a variety of applications. In medicine, TMCs are important metallodrugs. The first metallodrug, cisplatin, and its derivatives have been widely used for various disease treatments.[35] In material science, TMCs are good catalysts for organic synthesis and electrochemical reactions.[46, 47]

The importance of organometallic complexes motivates researchers to explore the chemical space of TMCs. However, the currently available methods simply extract ligands from existing TMCs in CSD and attach the random ligands to transition metals under different combinations to design new TMCs.[232, 233] Such methods are very limited to explore the diversity of novel ligands, thereby restricting the discovery of novel TMCs. As the ensemble of available ligands in CSD is already fixed, the diversity of generated TMCs is limited by the combinations of these ligands. Moreover, in this workflow, plenty of work, such as the selecting and assembling ligands, still requires much manual input and discernment, which hampers the pace of the overall procedure.

LigandDiff, proposed in this work, is a generative model which can automatically design numerous unique and novel ligands from scratch. LigandDiff is scaffold-oriented since exclusively one ligand is diffused or generated while both the central metal and other ligands remain static in the whole process. Such 'scaffold modeling' has been extensively applied in organic synthesis and drug discovery. The idea is to keep the majority of the molecule unchanged while selectively altering certain segments.[234] Maintaining the main structure usually enables to keep the core properties of the molecule, meanwhile adjusting minor functional groups leads to the enhancement of overall properties.[235] Generative models have potential to accelerate the targeted exploration with its powerful flexibility which do not require any human intervention. In addition, LigandDiff can be applied to investigate the ligand substitution reactions for new material discovery.[236, 237] Overall, LigandDiff offers a tool to explore the structure-activity relationship (SAR) in the context of metal-ligand interactions. Ferrocene is an instance of an organometallic 'sandwich' compound, characterized by two stable cyclopentadienyl rings. Either functional group can be easily modified by inserting new segments to the existing functional groups or supplanting the entire group with other new organic segments, giving rise to a bunch of diverse derivatives. Indeed, Fc analogues are promising drug candidates for the treatment of cancer and malaria. Each redesigned cyclopentadienyl moiety has shown unique mechanism to interact with biomolecules, leading to an overall improvement of the therapeutic efficacy.[238] And we believe LigandDiff has great potential in exploring TMCs derivatives with promising properties.

In LigandDiff all TMCs are considered as three-dimensional point clouds within a given space. And any point cloud x can be represented as $x = [r, h, h_L]$, where r is the Cartesian coordinates $r = (r_1, \dots, r_N) \in \mathbb{R}^{N\times 3}$, the variable h denotes the one-hot encoded representations that identify type of each atom $h = (h_1, \dots, h_N) \in$

$\mathbb{R}^{N\times m}$, where *N* is the size of a given complex, *m* is the count of the corresponding atom types and h_L serves as a one-hot vector that decodes ligand group information and it indicates the specific ligand group to which a given atom is associated, h_L= $\left(h_{L_1}, \ldots, h_{L_N}\right) \in \mathbb{R}^{N\times l}$, where *l* is the quantity of ligands. Prior to the input being fed into ϕ, random noise is applied exclusively to the ligand undergoing diffusion for both Cartesian coordinates and atom types, keeping the ligand-group information intact. In addition, even though the neural network updates the entire atomic embedding $x = [r, h, h_L]$, we only focus on the estimated Cartesian coordinates as well as the discrete atom type attributes.

Every designated ligand x^L undergoes a process of diffusion or noise reduction under a static context *u, i.e.*, the metal center and other ligands. *u* has the same component as *x*. On condition of this fixed context, both sampling steps eqs. 34 and 35 are modified accordingly as

$$p(x_{t-1}^L | x_t^L, u) = q(x_{t-1}^L | \hat{x}^L, x_t^L) \tag{36}$$

$$\hat{x}^L = (1/\alpha_t) x_t^L - (\sigma_t/\alpha_t)\, \phi(x_t^L, u, t) \tag{37}$$

The entire process is illustrated as **Figure 29**.

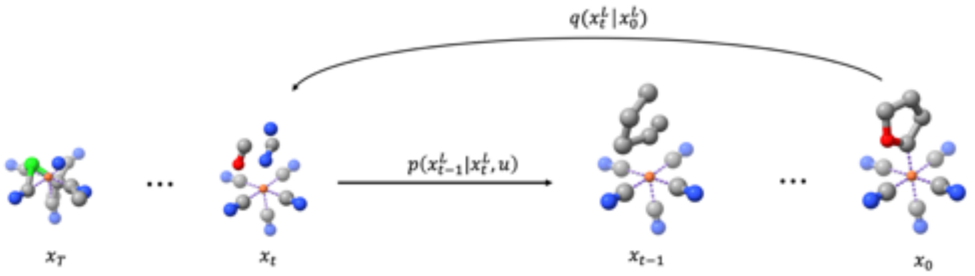


Figure 29. LigandDiff's overall scheme. The entire procedure begins from the forward diffusion process q, transitioning from the original ligand state x_0^L to the noised state x_t^L, in order to obtain the noised datapoint of a specific ligand x^L. After it has been trained, any new ligands are synthesized starting from a noised state x_T^L, which is sample from $\mathcal{N}(0,I)$, and progressively refined by denoising x_t^L through the conditional distributions *p.*

The dynamics of the diffusion model, specifically the adapted function ϕ are captured using Geometric Vector Perceptrons (GVPs).[239] GVPs are based on GNNs and define scalar and vector embedding as nodes and edges. The edges $e = (s, V)$ include the relative position between two nodes $s \in \mathbb{R}^{N\times 1}$ and a normalized direction vector $V \in \mathbb{R}^{N\times 1\times 3}$, where *N* is the number of edges. And they are further transformed as $e' = (s', V')$, where $s' \in \mathbb{R}^{N\times F}$ and $V' \in \mathbb{R}^{N\times 1\times 3}$, F is the number of hidden features. The update mechanism for the nodes initiates with scalar attributes and follows a similar procedure, $h = (s)$, where $s \in \mathbb{R}^{N\times m}$ and m is the number of features. Every node is updated as $h' = (s', V')$, where $s' \in \mathbb{R}^{N\times F}$ and $V' \in \mathbb{R}^{N\times (F/2)\times 3}$. In the subsequent part, except where specifically stated, *e* will represent the updated edges, and *h* will represent the updated nodes for ease of understanding. Within LigandDiff, every graph is completely interconnected, with all atomic interactions being considered during the message passing process, characterized as

$$m_{ij} = \phi_e\left(h_i, h_j, e_{ij}\right) \tag{36}$$

$$\tilde{e}_{ij} = \phi_{att} m_{ij} \tag{37}$$

$$m_i = \sum_j^{N-1} \tilde{e}_{ij} m_{ij} \tag{38}$$

$$h_i = \phi_h(h_i, m_i) \tag{39}$$

where h_i represents the central node's embedding, h_j corresponds to the adjacent node's embedding, e_{ij} signifies the attributes of the edge between them, and ϕ_e employs three GVPs to integrate messages from neighboring nodes. The ϕ_{att} is an attention-based neural network crafted with a single GVP for edges, while ϕ_h uses multiple GVPs to refine each central node. To thoroughly capture geometric details from the molecular structure, this message-passing procedure is iteratively applied. In the last step, an additional GVP reconverts scalar and vector features into a three-dimensional data point in the format of x_0, from which the estimated noise $\hat{\epsilon} = [\hat{\epsilon}^r, \hat{\epsilon}^h]$ is deducted, where $\hat{\epsilon}^r$ represents the positional noise and $\hat{\epsilon}^h$ denotes the atomic type noise.

4.3 Data Set Curation and Evaluation Metrics

Recently, Naveen et al. compiled a data set including nearly 86k mononuclear octahedral TMCs.[240] Organometallic complex containing M metal, M = Cr, Mn, Fe, Co, Ni, Cu, Zn with 100 atoms or fewer were extracted from this data set. Furthermore, nonmetal elements were constrained to {H, C, N, O, F, P, S, Cl, Br}. TMCs that lacked hydrogen atoms or exhibited disorder were further removed, and finally 23308 TMCs were collected and each of them has at least two ligands. **Figure 30** shows the distribution of metals. **Figure 31** shows the size distribution. **Figure 32** gives the denticity type of each ligand in these complexes. **Table 8** lists the denticity type of each complex. **Table 9** lists the distribution of heavy atoms in the masked ligands.

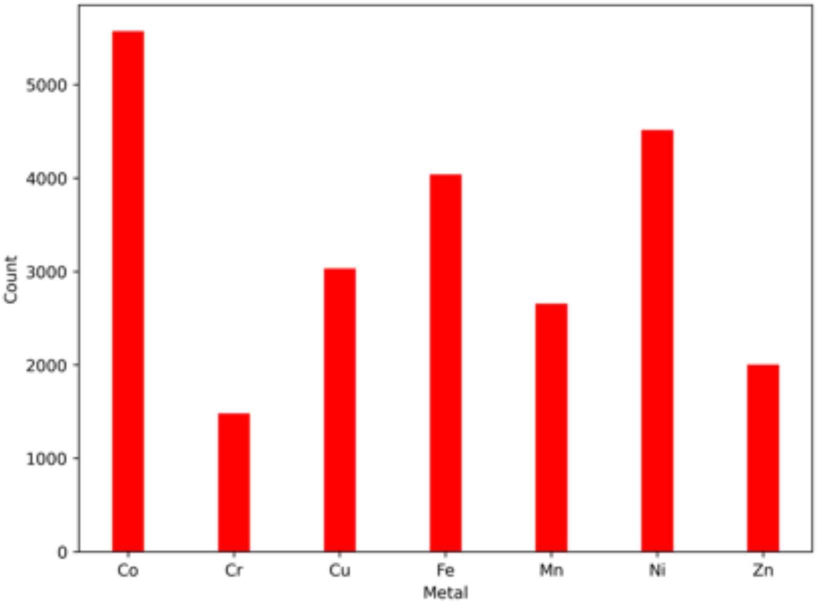


Figure 30. The distribution of metals in 23308 complexes.

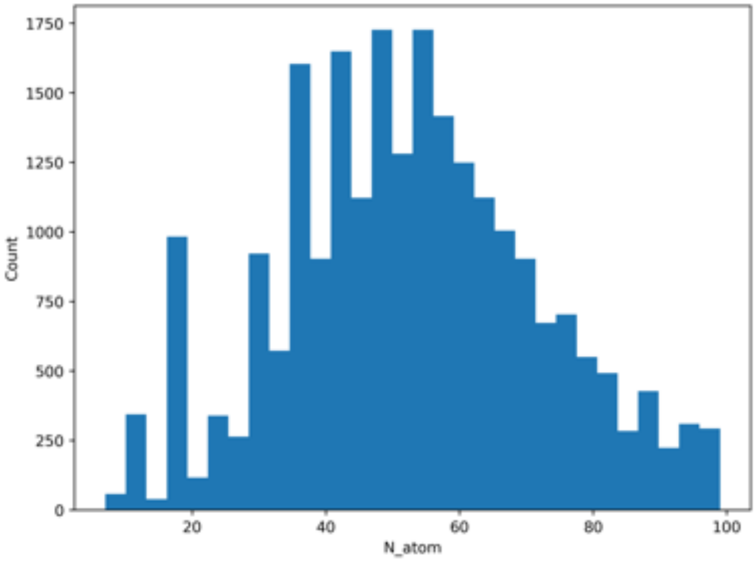


Figure 31. The size distribution of 23308 TMCs.

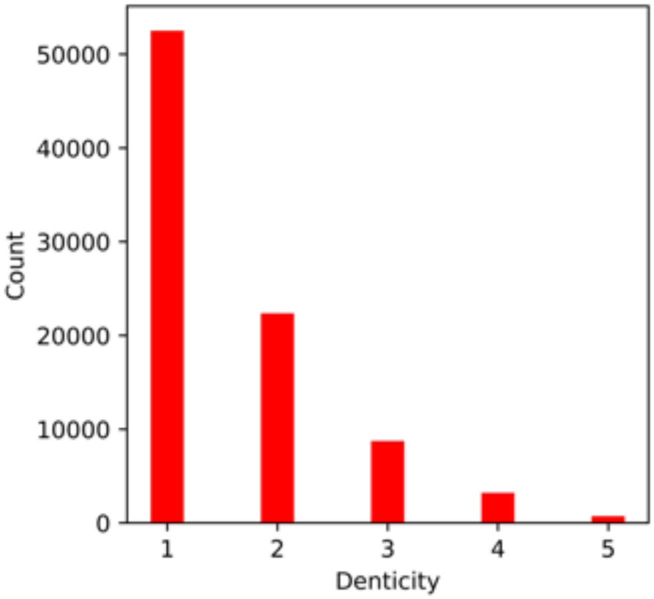


Figure 32. The denticity type of a single ligand in 23308 TMCs.

Table 8. The denticity types of the 23308 complexes.

Denticity type	Counts
5,1	719
4,2	702
4,1,1	2500
3,3	3539
3,2,1	512
3,1,1,1	1146
2,2,2	3318
2,2,1,1	5022
2,1,1,1,1	1151
1,1,1,1,1,1	4699

By using molSimplify,[241,242] each complex was deconstructed to extract the information about its ligands. And every ligand within each complex was selected for the diffusion or generation process. For instance, from a complex with six ligands, six variations can be derived, each with a different ligand targeted for modification. This process resulted in a total of 87,531 samples. To streamline the computational effort, all hydrogen atoms were excluded from the data. For the purpose of validation and testing, two groups of 400 samples were set aside, with the rest employed for the training set.

Table 9. The number of heavy atoms in the masked ligands.

N_{heavy_atom}	Count	N_{heavy_atom}	Count
1	22871	33	51
2	8089	34	119
3	5467	35	31
4	5017	36	77
5	3639	37	19
6	3343	38	42
7	3072	39	20
8	2878	40	107
9	2466	41	17
10	2538	42	28
11	1832	43	2
12	4586	44	20
13	2094	45	10
14	4218	46	10
15	1549	47	5
16	1852	48	52
17	1400	49	3
18	1974	50	4
19	1128	51	2
20	1270	52	14
21	734	53	1
22	1101	54	4
23	626	55	2
24	931	56	8
25	513	58	1
26	489	59	1
27	211	60	2
28	345	64	2
29	144	65	1
30	249	72	1
31	73	76	1
32	174	80	1

To comprehensively evaluate the performance of LigandDiff, a range of metrics were used. First, OpenBabel[243] was utilized to introduce chemical bonds into the denoised data points x^L. We used RDKit[244] to first examine the validity of the generated ligands,

$$p_l^{val} = \frac{N_l^{valid}}{N_{total}} \tag{40}$$

where N_l^{valid} is the number of valid ligands, N_{total} is the number of total generated ligands. The second metrics is connectivity, i.e., to check whether all atoms in the valid ligands are fully connected, calculated as

$$p_l^{con} = \frac{N_l^{valid\&connected}}{N_l^{valid}} \tag{41}$$

where $N_l^{valid\&connected}$ is the number of valid and connected ligands. The uniqueness and novelty are also evaluated as

$$p_l^{uniq} = \frac{N_l^{unique}}{N_l^{valid\&connected}} \tag{42}$$

$$p_l^{nov} = \frac{N_l^{nov}}{N_l^{valid\&connected}} \tag{43}$$

where N_l^{unique} is the number of unique ligands among outputs and N_l^{nov} is the number of the ligands outside the training dataset. Finally, the validity of the whole complex is check by molSimplify, calculated as

$$p_c^{val} = \frac{N_c^{valid}}{N_{total}} \tag{44}$$

4.4 Results and Discussion

In the evaluated set, 25% of the samples resulted in generated ligands comprising solely 1 heavy atom. Furthermore, 50% of the complexes contain fewer than five heavy atoms. Such implementation facilitates the generation of valid ligands by LigandDiff due to their simplicity. To strictly assess LigandDiff, we first did some random sampling, i.e., the size of generated ligands was randomly assigned. We set the sampling range from 6 to 20 because it only covers around 42% of the ligand size found within our training data set. Notably, half of the ligands diffused during training are characterized by having no more than five heavy atoms. Therefore, the sampling range we chose is challenging for LigandDiff. The results are given in **Table 10**. Even though with this tricky sampling, LigandDiff demonstrates strong capabilities. It inherently learns valency rules that other models[245, 246] require to be predefined, allowing for the automatic generation of valid ligands. This leads to a high rate of valid and interconnected complexes, achieving a 90% validity rate.

Table 10. Performance of LigandDiff.

	N_{atom}	p_l^{val}	p_l^{con}	p_l^{uniq}	p_l^{nov}	p_c^{val}
Random sample	6 ~ 20	0.94	0.96	0.97	0.96	0.90
Fix the ligand size	6	0.97	0.94	0.56	0.81	0.91
	7	0.97	0.95	0.70	0.83	0.92
	8	0.97	0.95	0.89	0.94	0.91
	9	0.96	0.95	0.90	0.98	0.90
	10	0.96	0.95	0.92	0.98	0.91
	11	0.96	0.96	0.95	0.99	0.91
PPR_100	11 ~ 40	0.94	0.94	0.92	1.0	0.87

To further evaluate LigandDiff's ability to understand chemical principles rather than simply memorizes the training set ligands, the size of the generated ligands was fixed, i.e., the generated ligand of each TMC in the test set had the same assigned size. The testing began with ligands composed of 6 atoms and incrementally expanded to 11. As shown in **Table 10**, consistently high validity rates were maintained for both ligands and complexes throughout the evaluation, with ligands showing 0.96 validity and complexes 0.91. In addition to the high validity, the connectivity metric is also impressive, with over 94% of the valid ligands being fully connected. This significant surge in uniqueness and novelty strongly suggests that our model can 'learn chemistry' from the given data. For the smaller size n = 6, uniqueness is relatively low at 0.56, indicating a high repetition rate in the ligands generated. However, as the size increases to n = 11, uniqueness soars to 0.95, underscoring a broadened structure variety. This trend is in agreement with chemical principles where larger systems inherently possess greater structural diversity. LigandDiff effectively utilizes its learning capabilities to create a diverse array of ligands. This applies to novelty as well. LigandDiff displays a tendency to innovate by creating new ligands rather than merely replicating those present in the training set, indicating its capability to venture into unexplored regions of the chemical space and contribute novel structures for potential applications. Even with a smaller size of n = 6, LigandDiff successfully generates 81% novel ligands, illustrating its capacity to innovate beyond the structures contained in the training set. As the ligand size increases to n = 11, the model almost exclusively crafts ligands that are novel, highlighting its robustness in designing unique TMCs. All the results mentioned above show that in this extreme situation where all the generated ligands must have the same size, LigandDiff still can generate different ligands under the given context. Some examples are given in **Figure 33**.

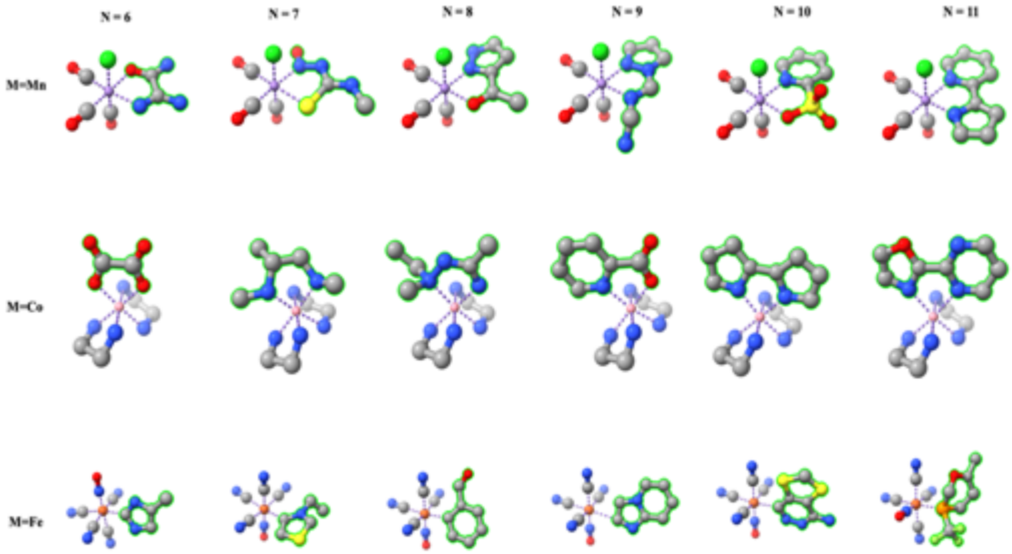
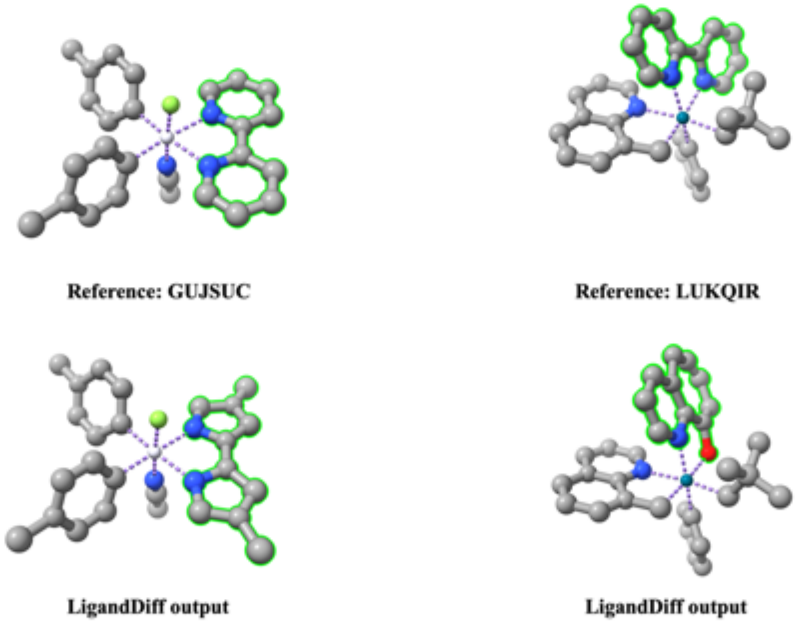


Figure 33. Various complexes created by LigandDiff with fixed ligand sizes. Each column represents complexes generated from ligands of the same size but varied contexts, while each row represents an increase in ligand size within the same context. The generated ligands are outlined in green for emphasis. Atoms include C: gray, N: blue, O: red, F: greenyellow, P: orange, S: yellow and Cl: green.

Finally, to test the transferability of LigandDiff, a trickly data set named PPR_100 was created from a larger original data set.[240] This subset includes 100 TMCs that specifically contain Pt, Pd and Ru, chosen due to their prevalence in the database after excluding the TMCs already present in the training dataset. Every complex in the PPR_100 set has more than 50 atoms, which presents a challenge due to their size and complexity. Ligands comprising more than ten heavy atoms were selected for masking, resulting in 148 samples for the generative task. This selective masking was because a subset of the complexes contained multiple ligands that met the criteria. However, since about 68% of the ligands in the training set contained ten or fewer heavy atoms, it is tricky for LigandDiff to generate valid complexes for PPR_100 data set. The results are given in **Table 10**. Even in the absence of Pt, Pd and Ru transition metals in the training set, LigandDiff demonstrates an impressive capability to generate new and structurally diverse ligands. it showcases a 94% success rate in creating valid and connected ligands that are entirely novel, with a very low repetition rate of only 8%. Furthermore, 87% of the generated complexes meet the validity criteria set by molSimplify. These achievements highlight LigandDiff's effectiveness in exploring the chemical space of ligands for various transition metals, despite the complexity and novelty of the task. Some successful examples of the newly generated TMCs and the reference TMCs are given in **Figure 34**. The LigandDiff framework utilizes the metal component merely as a static reference point to facilitate the prediction of variations in the surrounding ligand. This strategic limitation ensures that the metal does not participate in the diffusion dynamics. With this flexible framework, LigandDiff is capable of crafting novel ligands tailored to various transition metals without modifying the metal core.



Reference: GUJSUC

Reference: LUKQIR

LigandDiff output

LigandDiff output

Figure 34. Newly generated complexes (bottom) and the corresponding reference complexes (up) in the PPR_100 set. The CSD code is given.

In evaluating the ease of synthesis for the ligands produced, the average synthetic accessibility (SA) score[247] was employed as a measure. **Table 11** reveals that the ligands generated by LigandDiff not only bear high SA scores, denoting their realistic synthetic potential, but also maintain these high scores consistently, even as the size of the ligand increases.

Table 11. The SA scores of LigandDiff.

	N_{atom}	SA[a]
Random sample	6 ~ 20	0.69 ± 0.008
Fix the ligand size	6	0.80 ± 0.005
	7	0.77 ± 0.020
	8	0.74 ± 0.005
	9	0.74 ± 0.003
	10	0.73 ± 0.008
	11	0.72 ± 0.008
PPR_100	11 ~ 40	0.68 ± 0.008

4.5 Conclusions

In this present study, we proposed a generative model, LigandDiff, to design novel and unique ligands for TMCs from scratch. We first curated tens of thousands of TMCs from available database. To enrich the diversity of ligand samples, we masked each ligand in each complex as a unique sample. And inspired by diffusion models, we designed our scaffold-oriented generative model. Our results shows that LigandDiff is able to generate unlimited, diverse and easily synthesizable ligands on the condition of given context.

BIBLIOGRAPHY

1. Heiserman, D. L. Exploring Chemical Elements and Their Compounds.; Blue Ridge Summit: PA, 1940.

2. Popov, I. A.; Jian, T.; Lopez, G. V.; Boldyrev, A. I.; Wang, L.-S. Cobalt-Centred Boron Molecular Drums with the Highest Coordination Number in the CoB16− Cluster. *Nat Commun* **2015**, 6, 8654.

3. Soni, P. L.; Soni, V. The Chemistry of Coordination Complexes and Transition Metals; Taylor & Francis Group: Milton, 2021.

4. Malinowski, J.; Zych, D.; Jacewicz, D.; Gawdzik, B.; Drzeżdżon, J. Application of Coordination Compounds with Transition Metal Ions in the Chemical Industry—a Review. *Int. J. Mol. Sci.* **2020**, 21, 5443.

5. Haas, K. L.; Franz, K. J. Application of Metal Coordination Chemistry to Explore and Manipulate Cell Biology. *Chem. Rev.* **2009**, 109, 4921–4960.

6. Renfrew, A. K. Transition Metal Complexes with Bioactive Ligands: Mechanisms for Selective Ligand Release and Applications for Drug Delivery. Metallomics **2014**, *6*, 1324–1335.

7. Lee, L. C.-C; Lo, K. K.-W. Luminescent and Photofunctional Transition Metal Complexes: From Molecular Design to Diagnostic and Therapeutic Applications. *J. Am. Chem. Soc.* **2022**, *144*, 14420–14440.

8. Xiao, M.; Wu, Q.; Li, L.; Mu, S.; Sørensen, M. N.; Wang, W.; Cui, C. Regenerable Catalyst for Highly Alkaline Water Oxidation. *ACS Energy Lett.* **2021**, *6*, 1677–1683.

9. Liu, Y.; Bai, Y. Design and Engineering of Metal Catalysts for Bio-Orthogonal Catalysis in Living Systems. *ACS Applied Bio Materials* **2020**, *3*, 4717–4746.

10. Monro, S.; Colón, K. L.; Yin, H.; Roque, J.; Konda, P.; Gujar, S.; Thummel, R. P.; Lilge, L.; Cameron, C. G.; McFarland, S. A. Transition Metal Complexes and Photodynamic Therapy from a Tumor-Centered Approach: Challenges, Opportunities, and Highlights from the Development of TLD1433. *Chem. Rev.* **2018**, *119*, 797–828.

11. Kenny, R. G.; Marmion, C. J. Toward Multi-Targeted Platinum and Ruthenium Drugs—a New Paradigm in Cancer Drug Treatment Regimens? *Chem. Rev.* **2019**, *119*, 1058–1137.

12. Raj, P.; Singh, A.; Singh, A.; Singh, N. Syntheses and Photophysical Properties of Schiff Base Ni (II) Complexes: Application for Sustainable Antibacterial Activity and Cytotoxicity. *ACS Sustainable Chem. Eng.* **2017**, *5*, 6070–6080.

13. Adhikari, S.; Nath, P.; Das, A.; Datta, A.; Baildya, N.; Duttaroy, A. K.; Pathak, S. A Review on Metal Complexes and Its Anti-Cancer Activities: Recent Updates from in Vivo Studies. *Biomedicine & Pharmacotherapy* **2024**, *171*, 116211.

14. Ugwu, D. I.; Conradie, J. Anticancer Properties of Complexes Derived from Bidentate Ligands. *Journal of Inorganic Biochemistry* **2023**, *246*, 112268.

15. Fernández-Moreira, V.; Thorp-Greenwood, F. L.; Coogan, M. P. Application of d^6 Transition Metal Complexes in Fluorescence Cell Imaging. *Chem. Commun.* **2010**, *46*, 186–202.

16. Malviya, R.; Singh, A. K.; Yadav, D. Advances in Metallodrug-driven Combination Therapy for Treatment of Cancer. *Multi-Drug Resistance in Cancer* **2023**,155–170.

17. Banerjee, S.; Banerjee, S. Metal-Based Complexes as Potential Anti-Cancer Agents. *Anti-Cancer Agents in Medicinal Chemistry* **2022**, *22*, 2684–2707.

18. Kar, K.; Ghosh, D.; Kabi, B.; Chandra, A. A Concise Review on Cobalt Schiff Base Complexes as Anticancer Agents. *Polyhedron* **2022**, *222*, 115890.

19. Tisovský, P.; Donovalová, J.; Kožíšek, J.; Horváth, M.; Gáplovský, A. Reversible on/off and off/on, Light-Stimulated Binding, or Release Processes of Metal Cations from Isatin Diarylhydrazone Complexes in Solution. *Journal of Photochemistry & Photobiology, A: Chemistry* **2022**, *427*, 113827.

20. Anwar, M. U.; Al-Harrasi, A.; Rawson, J. M. Structures, Properties and Applications of Cu (II) Complexes with Tridentate Donor Ligands. *Dalton Trans.* **2021**, *50*, 5099–5108.

21. Lo, K. K.-W.; Choi, A. W.-T.; Law, W. H.-T. Applications of Luminescent Inorganic and Organometallic Transition Metal Complexes as Biomolecular and Cellular Probes. *Dalton Trans* **2012**, *41*, 6021.

22. Baggaley, E.; Weinstein, J. A.; Williams, J. A. G. Lighting the Way to See inside the Live Cell with Luminescent Transition Metal Complexes. *Coord. Chem. Rev*. **2012**, *256*, 1762–1785.

23. Mimura, I.; Nangaku, M. The suffocating kidney: tubulointerstitial hypoxia in end-stage renal disease. *Nat. Rev. Nephrol*. **2010**, *6*, 667–678.

24. Arden, G. B. & Sivaprasad, S. Hypoxia and oxidative stress in the causation of diabetic retinopathy. *Curr. Diabetes Rev*. **2011**, *7*, 291–304.

25. Vaupel, P.; Mayer, A. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev.* **2007**, *26*, 225–239.

26. Hara, D.; Umehara, Y.; Son, A.; Asahi, W.; Misu, S.; Kurihara, R.; Kondo, T.; Tanabe, K. Tracking the Oxygen Status in the Cell Nucleus with a Hoechst-tagged Phosphorescent Ruthenium Complex. *ChemBioChem* **2018**, *19,* 956–962.

27. Dmitriev, R. I.; Ropiak, H. M.; Ponomarev, G. V.; Yashunsky, D. V.; Papkovsky, D. B. Cell-Penetrating Conjugates of Coproporphyrins with Oligoarginine Peptides: Rational Design and Application for Sensing Intracellular O2. *Bioconjugate Chem.* **2011**, *22*, 2507–2518.

28. Spencer, J. A.; Ferraro, F.; Roussakis, E.; Klein, A.; Wu, J.; Runnels, J. M.; Zaher, W.; Mortensen, L. J.; Alt, C.; Turcotte, R.; Yusuf, R.; Côté, D.; Vinogradov, S. A.; Scadden, D. T.; Lin, C. P. Direct Measurement of Local Oxygen Concentration in the Bone Marrow of Live Animals. *Nature* **2014**, *508*, 269–273.

29. Marengo-Rowe, A. J. Structure-Function Relations of Human Hemoglobins. *Proc (Bayl Univ Med Cent)*.**2006**, *19*, 239–245.

30. Blumenthal, I. Carbon Monoxide Poisoning. *J R Soc Med*. **2001**, *94*, 270–272.

31. Outten, F. W.; Theil, E. C. Iron-Based Redox Switches in Biology. *Antioxid Redox Signal.***2009**, *11*, 1029–1046.

32. González, A.; Sevilla, E.; Bes, M. T.; Peleato, M. L.; Fillat, M. F. Pivotal Role of Iron in the Regulation of Cyanobacterial Electron Transport. *Adv Microb Physiol*. **2016**, 169–217.

33. Day, D. A.; Smith, P. M. Iron Transport across Symbiotic Membranes of Nitrogen-Fixing Legumes. *Int J Mol Sci*. **2021**, *22*, 432.

34. Rosenberg, B.; Van Camp, L.; Krigas, T. Inhibition of Cell Division in Escherichia Coli by Electrolysis Products from a Platinum Electrode. *Nature* **1965**, *205*, 698–699.

35. Shah N, Dizon DS. New-generation platinum agents for solid tumors. *Future Oncol*. **2009**, *5*, 33–42.

36. Gill, M. R.; Thomas, J. A. Ruthenium (II) Polypyridyl Complexes and DNA—from Structural Probes to Cellular Imaging and Therapeutics. *Chem. Soc. Rev.* **2012**, *41*, 3179.

37. Lee, S. Y.; Kim, C. Y.; Nam, T.-G. Ruthenium Complexes as Anticancer Agents: A Brief History and Perspectives. *Drug Des Devel Ther*. **2020**, *14*, 5375–5392.

38. Thota, S.; Rodrigues, D. A.; Crans, D. C.; Barreiro, E. J. Ru(II) Compounds: = Next-Generation Anticancer Metallotherapeutics? *J Med Chem* **2018**, *61*, 5805–5821.

39. Cervantes-Cervantes, M. P.; Calderón-Salinas, J. V.; Albores, A.; Muñoz-Sánchez, J. L. Copper Increases the Damage to DNA and Proteins Caused by Reactive Oxygen Species. *Biol Trace Elem Res* **2005**, *103*, 229–248.

40. Trejo-Solís, C.; Palencia, G.; Zuñiga, S.; Rodríguez-Ropon, A.; Osorio-Rico, L.; Torres Luvia, S.; Gracia-Mora, I.; Marquez-Rosado, L.; Sánchez, A.; Moreno-García, M. E.; Cruz, A.; Bravo-Gómez, M. E.; Ruiz-Ramírez, L.; Rodríquez-Enriquez, S.; Sotelo, J. Cas Ilgly Induces Apoptosis in Glioma C6 Cells in Vitro and in Vivo through Caspase Dependent and Caspase-Independent Mechanisms. *Neoplasia* **2005**, *7*, 563–574.

41. Abosede, O. O.; Vyas, N. A.; Singh, S. B.; Kumbhar, A. S.; Kate, A.; Kumbhar, A. A.; Khan, A.; Erxleben, A.; Smith, P.; de Kock, C.; Hoffmann, F.; Obaleye, J. A. Copper (II) Mixed-Ligand Polypyridyl Complexes with Doxycycline – Structures and Biological Evaluation. *Dalton Trans* **2016**, *45*, 3003–3012.

42. Shi, X.; Chen, Z.; Wang, Y.; Guo, Z.; Wang, X. Hypotoxic Copper Complexes with Potent Anti-Metastatic and Anti-Angiogenic Activities against Cancer Cells. *Dalton Trans* **2018**, *47*, 5049–5054.

43. Spoerlein, C.; Mahal, K.; Schmidt, H.; Schobert, R. Effects of Chrysin, Apigenin, Genistein and Their Homoleptic Copper (II) Complexes on the Growth and MetastaticPotential of Cancer Cells. *J Inorganic Biochem* **2013**, *127*, 107–115.

44. González-Ballesteros, M. M.; Mejía, C.; Ruiz-Azuara, L. Metallodrugs: An Approach against Invasion and Metastasis in Cancer Treatment. *FEBS Open Bio* **2022**, *12*, 880–899.

45. Ndamse, C. C.; Masamba, P.; Kappo, A. P. Bioorganometallic Compounds as Novel Drug Targets against Schistosomiasis in Sub-Saharan Africa: An Alternative to Praziquantel? *Adv Pharm Bull.* **2021**, *12*, 283-297.

46. Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis.*Chem.Rev.***1977**,*77*, 313−348.

47. Van Leeuwen, P. W. N. M.; Kamer, P. C. J.; Reek, J. N. H.; Dierkes, P. Ligand Bite Angle Effects in Metal-Catalyzed C-C Bond Formation. *Chem. Rev.* **2000**, 100, 2741−2770.

48. Zimmerman, J. B.; Anastas, P. T.; Erythropel, H. C.; Leitner, W. Designing for a Green Chemistry Future. *Science* **2020**, *367*, 397–400.

49. Francke, R.; Schille, B.; Roemelt, M. Homogeneously Catalyzed Electroreduction of Carbon Dioxide—Methods, Mechanisms, and Catalysts. *Chem. Rev.* **2018**, *118*, 4631-4701.

50. Kinzel, N. W.; Werlé, C.; Leitner, W. Transition Metal Complexes as Catalysts for the Electroconversion of CO_2: An Organometallic Perspective. *Angew. Chem. Int. Ed.* **2021**, *60*, 11628–11686.

51. Michelin, C.; Hoffmann, N. Photosensitization and Photocatalysis—Perspectives in Organic Synthesis. *ACS Catal.* **2018**, *8*, 12046–12055.

52. Manav, N.; Kesavan, P. E.; Ishida, M.; Mori, S.; Yasutake, Y.; Fukatsu, S.; Furuta, H.; Gupta, I. Phosphorescent Rhenium-Dipyrrinates: Efficient Photosensitizers for Singlet Oxygen Generation. *Dalton Trans* **2019**, *48*, 2467–2478.

53. Sun, J.; Zhao, J.; Guo, H.; Wu, W. Visible-Light Harvesting Iridium Complexes as Singlet Oxygen Sensitizers for Photooxidation of 1,5-Dihydroxynaphthalene. *Chem. Commun.* **2012**, *48*, 4169-4171.

54. Nazeeruddin, M. K.; Kay, A.; Rodicio, I.; Humphry-Baker, R.; Mueller, E.; Liska, P.; Vlachopoulos, N.; Graetzel, M. Conversion of Light to Electricity by Cis-X2bis(2,2'-Bipyridyl-4,4'-Dicarboxylate) Ruthenium (II) Charge-Transfer Sensitizers (X = Cl-, Br-, I-, CN-, and Scn-) on Nanocrystalline Titanium Dioxide Electrodes. *J. Am. Chem. Soc.* **1993**, *115*, 6382–6390.

55. Behm, K.; McIntosh, R. D. Application of Discrete First-row Transition-metal Complexes as Photosensitisers. *ChemPlusChem* **2020**, *85*, 2611–2618.

56. Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machines Applied to Chemical Problems. Molecular Formula Determination from Low Resolution Mass Spectrometry. *Anal. Chem.* **1969**, *41*, 21–27.

57. Tsou, L. K.; Yeh, S.-H.; Ueng, S.-H.; Chang, C.-P.; Song, J.-S.; Wu, M.-H.; Chang, H.-F.; Chen, S.-R.; Shih, C.; Chen, C.-T.; Ke, Y.-Y. Comparative Study between Deep Learning and QSAR Classifications for TNBC Inhibitors and Novel GPCR Agonist Discovery. *Sci Rep* **2020***, 10*, 16771.

58. Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, H.-N.; Kim, M. S.; No, K. T.; Wang, G. Comprehensive Strategies of Machine-Learning Based Quantitative Structure-Activity Relationship Models. *iScience* **2021**, *24*, 103052.

59. Bursch, M.; Mewes, J.; Hansen, A.; Grimme, S. Best-practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem. Int. Ed*. **2022**, *134*, e202205.

60. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**,98, 146401.

61. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**,104, 136403.

62. Thompson, A., Swiler, L., Trott, C., Foiles, S. & Tucker, G. Snap: automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2014**,285, 316–330.

63. Zong, H., Pilania, G., Ding, X., Ackland, G. J. & Lookman, T. Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *npj Comput. Mater.* **2018**,4, 48.

64. Morawietz, T., Singraber, A., Dellago, C. & Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl Acad. Sci. USA* **2016**,113, 8368–8373.

65. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* 2017,8, 3192–3203.

66. Eckhoff, M. & Behler, J. From molecular fragments to the bulk: development of a neural network potential for MOF-5. *J. Chem. Theory Comput.* **2019**,15, 3793–3809.

67. Elias, J. S. et al. Elucidating the nature of the active phase in copper/ceria catalysts for CO oxidation. *ACS Catalysis* **2016**,6, 1675–1679.

68. Artrith, N. & Behler, J. High-dimensional neural network potentials for metal surfaces: a prototype study for copper. *Phys. Rev. B* **2012**,85, 045439.

69. Sun, G. & Sautet, P. Metastable structures in cluster catalysis from first-principles: structural ensemble in reaction conditions and metastability triggered reactivity. *J. Am. Chem. Soc.* **2018**,140, 2812–2820.

70. Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. Big–Deep–Smart Data in Imaging for Guiding Materials Design. *Nature Mater* **2015**, *14*, 973–980.

71. Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; Wei, L. Deep Generative Models in *De Novo* Drug Molecule Generation. *J. Chem. Inf. Model.* **2023**.

72. Vallee, B. L.; Galdes, A. The Metallobiochemistry of Zinc Enzymes. *Adv Enzymol Relat Areas Mol Biol* **1984**, *56,* 283–430.

73. Vallee, B. L.; Auld, D. S. Active-Site Zinc Ligands and Activated H_2O of Zinc Enzymes. *Proc. Natl. Acad. Sci USA* **1990**, *87*, 220–224.

74. Turner, A. J. Exploring the Structure and Function of Zinc Metallopeptidases: Old Enzymes and New Discoveries. *Biochem Soc Trans* **2003**, *31*, 723–727.

75. Coleman, J. E. Zinc Enzymes. *Curr. Opin. Chem. Biol.* **1998**, *2*. 222–234.

76. Roohani, N, Hurrell, R, Kelishadi, R, Schulin, R. Zinc and its importance for human health: An integrative review. J Res Med Sci 2013, 18, 144-157.

77. Adhikari, S.; Bhattacharjee, T.; Butcher, R.J.; Porchia, M.; De Franco, M.; Marzano, C.; Gandin, V.; Tisato, F. Synthesis and characterization of mixed-ligand Zn (II) and Cu (II) complexes including polyamines and dicyano-dithiolate(2-): In vitro cytotoxic activity of Cu(II) compounds. *Inorg. Chim. Acta* **2019**, *498*, 119098.

78. Auld, D. S. The ins and outs of biological zinc sites. *BioMetals* **2009**, *22*, 141−148.

79. Parkin, G. The bioinorganic chemistry of zinc: synthetic analogues of zinc enzymes that feature tripodal ligands. *Chem. Commun*. **2000**, 1971−1985.

80. Vallee, B. L.; Auld, D. S. Zinc: biological functions and coordination motifs. *Acc. Chem. Res.***1993**, *26*, 543−551.

81. Maret, W. Zinc and Sulfur: A Critical Biological Partnership. *Biochemistry* **2004**, *43*, 3301−3309.

82. Padjasek, M.; Kocyła, A.; Kluska, K.; Kerber, O.; Tran, J. B.; Krężel, A. Structural zinc binding sites shaped for greater works:Structure-function relations in classical zinc finger, hook and claspdomains. *J. Inorg. Biochem.* **2020**, *204*, 110955.

83. Cousins, R. J. A Role of Zinc in the Regulation of Gene Expression. *Proc Nutr Soc.* **1998**, *57*, 307–311.

84. Falchuk, K. H. The molecular basis for the role of zinc in developmental biology. *Mol Cell Biochem*. **1998**, *188*,41-48.

85. Korkmaz-Icöz, S.; Atmanli, A.; Radovits, T.; Li, S.; Hegedüs, P.; Ruppert, M.; Brlecic, P.; Yoshikawa, Y.; Yasui, H.; Karck, M.; Szabó, G. Administration of Zinc Complex of Acetylsalicylic Acid after the Onset of Myocardial Injury Protects the Heart by Upregulation of Antioxidant Enzymes. *J Physio Sci* **2015**, *66*, 113–125.

86. Li, H.-T.; Jiao, M.; Chen, J.; Liang, Y. Roles of Zinc and Copper in Modulating the Oxidative Refolding of Bovine Copper, Zinc Superoxide Dismutase. *Acta Biochim Biophys Sin* **2010**, *42*, 183–194.

87. Skalny, A. A.; Tinkov, A. A.; Medvedeva, Y. S.; Alchinova, I. B.; Karganov, M. Y.; Skalny, A. V.; Nikonorov, A. A. Effect of Short-Term Zinc Supplementation on Zinc and Selenium Tissue Distribution and Serum Antioxidant Enzymes. *Acta Sci Pol Technol Aliment.* **2015**, *14*, 269–276.

88. Pellei, M.; Del Bello, F.; Porchia, M.; Santini, C. Zinc Coordination Complexes as Anticancer Agents. *Coord. Chem. Rev.* **2021**, *445*, 214088.

89. Di Vaira, M.; Bazzicalupi, C.; Orioli, P.; Messori, L.; Bruni, B.; Zatta, P. Clioquinol, a drug for Alzheimer's disease specifically interfering with brain metal metabolism: Structural characterization of its zinc (II) and copper (II) complexes. *Inorg. Chem.* **2004**, *43*, 3795–3797.

90. D'Angelo, J.; Morgant, G.; Ghermani, N.E.; Desmaële, D.; Fraisse, B.; Bonhomme, F.; Dichi, E.; Sghaier, M.; Li, Y.; Journaux, Y.; et al. Crystal structures and physico-chemical properties of Zn (II) and Co (II) tetraaqua(3-nitro-4-hydroxybenzoato) complexes: Their anticonvulsant activities as well as related (5-nitrosalicylato)-metal complexes. *Polyhedron* **2008**, *27*, 537–546.

91. Nakayama, A.; Hiromura, M.; Adachi, Y.; Sakurai, H. Molecular mechanism of antidiabetic zinc-allixin complexes: Regulations of glucose utilization and lipid metabolism. *J. Biol. Inorg. Chem*. **2008**, *13*, 675–684.

92. Sakurai, H.; Yoshikawa, Y.; Yasui, H. Current state for the development of metallopharmaceutics and anti-diabetic metal complexes. *Chem. Soc. Rev.* **2008**, *37*, 2383–2392.

93. Sakurai, H.; Kojima, Y.; Yoshikawa, Y.; Kawabe, K.; Yasui, H. Antidiabetic vanadium (IV) and zinc (II) complexes. *Coord. Chem. Rev*. **2002**, *226*, 187–198.

94. Perdew, J. P. & Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc*. **2001**, *577*, 1–20.

95. Nagai, R., Akashi, R. & Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Comput Mater* **2020**, *6*, 43.

96. González, M. Force Fields and Molecular Dynamics Simulations. *École thématique de la Société Française de la Neutronique* **2011**,*12*, 169−200.

97. Lennard-Jones, J. E. On The Determination of Molecular Fields. − II. From The Equation of State of A Gas. *Proc. R. Soc.London A* **1924**,*106*, 463−477.

98. Vitalini, F.; Mey, A. S.; Noé, F.; Keller, B. G. Dynamic Properties of Force Fields. *J. Chem. Phys.* **2015**, *142*, 084101.

99. Schütt, K. T.; Chmiela, S.; von Lilienfeld, O. A.; Tkatchenko, A.; Tsuda, K.; Müller, K.-R.*Machine Learning Meets Quantum Physics*; Springer, 2020.

100.Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies withMachine Learning. *Phys. Rev. Lett.* **2012**,*108*, 058301.

101.Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.;Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**,*15*, 095003.

102.Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.***2017**, *8*, 3192−3203.

103.Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics.*Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**,*1*, 826−843.

104.Truhlar, D. G. *Potential Energy Surfaces and Dynamics Calculations*, 1st ed.; Springer: Boston, MA, 1981.

105.Bushnell, E. A. C.; Huang, W.; Gauld, J. W. Applications of potential energy surfaces in the study of enzymatic reactions. *Adv.Phys. Chem.* **2012**, *2012*,1.

106.Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural Network Models of Potential Energy Surfaces. *J. Chem. Phys.* **1995**, *103*, 4129–4137.

107.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.

108.Drautz, R. Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials. *Phys. Rev. B* **2019**, *99,* 014104.

109.Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials. *Journal of Computational Physics* **2015**, *285*, 316–330.

110.Wood, M. A.; Thompson, A. P. Extending the Accuracy of the Snap Interatomic Potential Form. *J. Chem. Phys.* **2018**, *148,* 241721.

111.Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.***2010**, *104*, 136403.

112.Bartók, A. P.; Csányi, G. Erratum. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quant. Chem.* **2016**, *116*, 1049–1049.

113.Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153–1173.

114.Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K.-R.; Tkatchenko, A. sGDML: Constructing Accurate and Data Efficient Molecular Force Fields Using Machine Learning. *Comput. Phys. Commun.* **2019**, *240*, 38–45.

115.Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037–10072.

116.Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural Network Models of Potential Energy Surfaces. *J. Chem. Phys.* **1995**, *103*, 4129–4137.

117.Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

118.Behler, J. Atom-Centered Symmetry Functions for Constructing High Dimensional Neural Network Potentials. *J. Chem. Phys*. **2011**,*134*, 074106.

119.Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. **2017.** Neural message passing for quantum chemistry. *In Proceedings of the 34th International Conference on Machine Learning*, pp. 1263–72. Sydney, Aust.: JMLR.

120. Scarselli, F.; Gori, M.; Ah Chung Tsoi; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* **2009**, *20*, 61–80.

121. Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – a Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

122. Ramakrishnan, R., Dral, P., Rupp, M. et al. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **2014**, *1*, 140022.

123. Schütt, K. T.; Unke, O. T.; Gastegger, M.Equivariant message passing for the prediction of tensorial properties and molecular spectra. 2021. arXiv preprint arXiv: 2102.03150. https://arxiv.org/abs/2102.03150 (accessed October 26, 2023).

124. Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., & Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun*. **2022**, *13*, 2453.

125. Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds. 2018. arXiv preprint arXiv:1802.08219. https://arxiv.org/abs/1802.08219 (accessed October 26, 2023).

126. Fuchs, F. B.; Worrall, D.E.; Fischer, V.; Welling, M. SE (3)-transformers: 3d Roto-translation equivariant attention networks. 2020. arXiv preprint arXiv:2006.10503. https://arxiv.org/abs/2006.10503 (accessed October 26, 2023).

127. Kronik, L.; Tkatchenko, A. Understanding Molecular Crystals with Dispersion-Inclusive Density Functional Theory: Pairwise Corrections and Beyond. *Acc. Chem. Res.* **2014**, *47*, 3208–3216.

128. Artrith, N.; Morawietz, T.; Behler, J. High-Dimensional Neural-Network Potentials for Multicomponent Systems: Applications to Zinc Oxide. *Phys. Rev. B* **2011**, *83*, 153101.

129. Morawietz, T.; Sharma, V.; Behler, J. A Neural Network Potential-Energy Surface for the Water Dimer Based on Environment-Dependent Atomic Energies and Charges. *J. Chem. Phys.* **2012**, *136*, 064103.

130.Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

131.Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic Potentials for Ionic Systems with Density Functional Accuracy Based on Charge Densities Obtained by a Neural Network. *Phys. Rev. B.* **2015**, *92*, 045131.

132.Rappe, A. K.; Goddard, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358–3363.

133.Faraji, S.; Ghasemi, S. A.; Rostami, S.; Rasoulkhani, R.;Schaefer, B.; Goedecker, S.; Amsler, M. High Accuracy andTransferability of a Neural Network Potential Through ChargeEquilibration for Calcium Fluoride. *Phys. Rev. B: Condens. MatterMater. Phys.* **2017**, *95*, 104105.

134.Faraji, S.; Ghasemi, S. A.; Parsaeifard, B.; Goedecker, S.Surface Reconstructions and Premelting of the (100)CaF_2 Surface. *Phys. Chem. Chem. Phys*. **2019**, *21*, 16270−16281.

135.Eivari, H. A.; Ghasemi, S. A.; Tahmasbi, H.; Rostami, S.; Faraji, S.; Rasoulkhani, R.; Goedecker, S.; Amsler, M. Two-Dimensional Hexagonal Sheet of TiO_2. *Chem. Mater*. **2017**, *29*, 8594−8603.

136.Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *J. Chem. Theory Comput.* **2020**, *16*, 4256−4270.

137.Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **2021**, *54*, 808–817.

138.Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun*. **2021**, *12*, 398.

139.Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-conserving Molecular Force Fields. *Sci. Adv*. **2017**, *3*, No. e1603015.

140. Smith, J., Isayev, O. & Roitberg, A. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci Data* **2017**, *4*, 170193.

141. Balcells, D.; Skjelstad, B. B. TMQM Dataset—Quantum Geometries and Properties of 86K Transition Metal Complexes. *J. Chem. Inf. Model.* **2020***, 60*, 6135–6146.

142. Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513-524.

143. 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.

144. Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

145. He, X.; Ni, D.; Zhang, H.; Li, X.; Zhang, J.; Fu, Q.; Liu, Y.; Lu, S. Zinc-Mediated Conformational Preselection Mechanism in the Allosteric Control of DNA Binding to the Zinc Transcriptional Regulator (ZitR). *Sci Rep* **2020**, *10*, 13276.

146. 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.

147. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta. Crystallogr. B. Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171. The Cambridge Structural Database.

148. 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.

149. Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. R^{2}Scan-3c: A “Swiss Army Knife” Composite Electronic-Structure Method. *J. Chem. Phys.* **2021**, *154*, 064103.

150. Neese, F. Software Update: The Orca Program System—Version 5.0. *WIREs Comput Mol Sci.* **2022**, 12, No. e1606.

151. 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*, 154104.

152. 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*, 154122.

153. Kosmala, A.; Gasteiger, J.; Gao, N.; Günnemann, S. Ewald-based Long-Range Message Passing for Molecular Graphs. 2023. arXiv preprint arXiv: 2303.04791. https://arxiv.org/abs/2303.04791 (accessed October 26, 2023).

154. Wells, B. A.; Chaffee, A. L. Ewald Summation for Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 3684–3695.

155. Ewald, P. P. Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. *Annalen der Physik* **1921**, *369*, 253–287.

156. Toukmaji, A. Y.; Board, J. A. Ewald Summation Techniques in Perspective: A Survey. *Computer Physics Communications* **1996**, *95*, 73–92.

157. Rackers, J. A.; Liu, C.; Ren, P.; Ponder, J. W. A Physically Grounded Damped Dispersion Model with Particle Mesh Ewald Summation. *J. Chem. Phys.* **2018**, *149*, 084115.

158. Aguado, A.; Madden, P. A. Ewald Summation of Electrostatic Multipole Interactions up to the Quadrupolar Level. *J. Chem. Phys.* **2003**, *119*, 7471–7483.

159. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; pp 770−778.

160. Manz, T. A.; Sholl, D. S. Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.* **2010**, *6*, 2455–2468.

161. 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.

162.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.

163.Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model* **2007**, *13*, 1173–1213.

164.Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2011**, *8*, 141–151.

165.Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model* **2013**, *19*, 1–32.

166.Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended Tight-binding Quantum Chemistry Methods. *WIREs Comput Mol Sci.* **2021**, *11*, No: e1493.

167.Stewart, J. J. P. MOPAC2016, Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net.

168.Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals—Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

169.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.

170.Weigend, F.; Kattannek, M.; Ahlrichs, R. Approximated Electron Repulsion Integrals: Cholesky Decomposition versus Resolution of the Identity Methods. *J. Chem. Phys.* **2009**, *130,* 164106.

171.Kepp, K. P. Consistent Descriptions of Metal–Ligand Bonds and Spin-Crossoverin Inorganic Chemistry. *Coord. Chem. Rev.* **2013**, *257*, 196–209.

172.Nihei, M.; Shiga, T.; Maeda, Y.; Oshio, H. Spin Crossover Iron (III) Complexes. *Coord. Chem. Rev*. **2007**, *251*, 2606–2621.

173.Martinho, P. N.; Gildea, B.; Harris, M. M.; Lemma, T.; Naik, A. D.; Müller-Bunz, H.; Keyes, T. E.; Garcia, Y.; Morgan, G. G. Cooperative Spin Transition in a Mononuclear Manganese (III) Complex. *Angewandte Chemie International Edition* **2012**, *51* (50), 12597–12601. DOI:10.1002/anie.201205573.

174.Guo, Y.; Yang, X.-L.; Wei, R.-J.; Zheng, L.-S.; Tao, J. Spin Transition and Structural Transformation in a Mononuclear Cobalt (II) Complex. *Inorganic Chemistry* **2015**, *54*, 7670–7672.

175.Yao, Z.-S.; Tang, Z.; Tao, J. Bistable molecular materials with dynamic structures. *Chem. Commun.* **2020**, *56*, 2071−2086.

176.Scott, H. S.; Staniland, R. W.; Kruger, P. E. Spin cross over in homoleptic Fe (II) imidazolylimine complexes. *Coord. Chem. Rev*. **2018**, *362*, 24−43.

177.Bao, X.; Shepherd, H. J.; Salmon, L.; Molnár, G.; Tong, M.-L.; Bousseksou, A. The Effect of an Active Guest on the Spin Crossover Phenomenon. *Angew. Chem., Int. Ed*. **2013**, *52*, 1198−1202.

178.Nguyen, T.-A.D.; Veauthier, J. M.; Angles-Tamayo, G. F.;Chavez, D. E.; Lapsheva, E.; Myers, T. W.;Nelson, T. R.; Schelter, E. J. Correlating Mechanical Sensitivity with Spin Transition in the Explosive Spin Crossover Complex$[Fe(Htrz)_3]_n[ClO_4]_{2n}$. *J. Am. Chem. Soc.* **2020**,*142*, 4842−4851.

179.Kershaw Cook, L. J.; Kulmaczewski, R.; Mohammed, R.; Dudley, S.; Barrett, S. A.; Little, M. A.; Deeth, R. J.; Halcrow, M. A. A Unified Treatment of the Relationship Between Ligand Substituents and Spin State in a Family of Iron (II) Complexes. *Angew.Chem., Int. Ed.* **2016**, *55*, 4327−4331.

180.Li, B.; Wei, R.-J.; Tao, J.; Huang, R.-B.; Zheng, L.-S.; Zheng, Z. Solvent-induced transformation of single crystals of a spin-crossover (SCO) compound to single crystals with two distinct sco centers. *J. Am. Chem.Soc*. **2010**, *132*, 1558−1566.

181.Atmani, C.; El Hajj, F.; Benmansour, S.; Marchivie, M.; Triki, S.; Conan, F.; Patinec, V.; Handel, H.; Dupouy, G.; Gómez-García, C. J. Guidelines to design new spin crossover materials. *Coord. Chem. Rev*. **2010**, *254*, 1559−1569.

182.Real, J. A.; Gaspar, A. B.; Muñoz, M. C. Thermal, Pressure and Light Switchable Spin-Crossover Materials. *Dalton Trans.* **2005**, *12*, 2062–2079.

183.Gütlich, P.; Goodwin, H. A. *Spin Crossover in Transition Metal Compounds II*; Springer Berlin Heidelberg, 2004.

184.Jensen, K. P.; Ryde, U. How O_2 Binds to Heme. *J. Biol. Chem.* **2004**, *279*, 14561 14569.

185.Strickland, N.; Harvey, J. N. Spin-Forbidden Ligand Binding to the Ferrous−heme Group: Ab Initio and DFT Studies. *J. Phys. Chem. B* **2007**, *111*, 841–852.

186.Wang, Y.; Ying, J.; Zhou, Z.; Sun, J.; Wen, T.; Zhou, Y.; Li, N.; Zhang, Q.; Han, F.; Xiao, Y.; Chow, P.; Yang, W.; Struzhkin, V. V.; Zhao, Y.; Mao, H. Emergent Superconductivity in an Iron-Based Honeycomb Lattice Initiated by Pressure- Driven Spin-Crossover. *Nat. Commun*. **2018**, *9*, 1914.

187.Delgado, T.; Tissot, A.; Guénée, L.; Hauser, A.; Valverde-Muñoz, F. J.; Seredyuk, M.; Real, J. A.; Pillet, S.; Bendeif, E.-E.; Besnard, C. Very Long-Lived Photogenerated High-Spin Phase of a Multistable Spin-Crossover Molecular Material. *J. Am. Chem. Soc.* **2018**, *140*, 12870–12876.

188.Lochenie, C.; Schötz, K.; Panzer, F.; Kurz, H.; Maier, B.; Puchtler, F.; Agarwal, S.; Köhler, A.; Weber, B. Spin-Crossover Iron (II) Coordination Polymer with Fluorescent Properties: Correlation between Emission Properties and Spin State. *J. Am. Chem. Soc.* **2018**, *140*, 700–709.

189.Hoshino, N.; Iijima, F.; Newton, G. N.; Yoshida, N.; Shiga, T.; Nojiri, H.; Nakao, A.; Kumai, R.; Murakami, Y.; Oshio, H. Three-Way Switching in a Cyanide-Bridged [CoFe] Chain. *Nature Chem* **2012**, *4*, 921–926.

190.Zheng, H.; Meng, Y.; Zhou, G.; Duan, C.; Sato, O.; Hayami, S.; Luo, Y.; Liu, T. Simultaneous Modulation of Magnetic and Dielectric Transition via Spin-crossover-tuned Spin Arrangement and Charge Distribution. *Angew. Chem. Int. Ed.* **2018**, *57*, 8468–8472.

191.Góra-Marek, K.; Stępniewski, A.; Radoń, M.; Broclawik, E. Ammonia-Modified Co (II) Sites in Zeolites: IR Spectroscopy and Spin-Resolved Charge Transfer Analysis of No Adsorption Complexes. *Phys. Chem. Chem. Phys.* **2014**, *16*, 24089–24098.

192. Shaik, S.; Chen, H.; Janardanan, D. Exchange-enhanced reactivity in bond activation by metal–oxo enzymes and synthetic reagents. *Nature Chem* **2011**, *3*, 19–27.

193. 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.

194. Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3C: A Revised Low-Cost Variant of the B97-D Density Functional Method. *J. Chem. Phys*. **2018**, *148*, 064104.

195. Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative Assessment of a New Nonempirical Density Functional: Molecules and Hydrogen-Bonded Complexes. *J. Chem. Phys*. **2003**, *119*, 12129.

196. Cirera, J.; Via-Nadal, M.; Ruiz, E. Benchmarking Density Functional Methods for Calculation of State Energies of First Row Spin-Crossover Molecules. *Inorg. Chem.* **2018**, *57*, 14097–14105.

197. Cirera, J.; Ruiz, E. Computational Modeling of Transition Temperatures in Spin Crossover Systems. *Comments Inorg. Chem*. **2019**, *39*, 216-241.

198. Jensen, K. P.; Cirera, J. Accurate Computed Enthalpies of Spin Crossover in Iron and Cobalt Complexes. *J. Phys. Chem. A*. **2009**, *113*, 10033–10039.

199. Radoń, M. Benchmarking Quantum Chemistry Methods for Spin-State Energetics of Iron Complexes against Quantitative Experimental Data. *Phys. Chem. Chem. Phys*. **2019**, *21*, 4854–4870.

200. Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to RN: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys*. **2005**, *7*, 3297-3305.

201. Neese, F. An Improvement of the Resolution of the Identity Approximation for the Formation of the Coulomb Matrix. *J. Comput. Chem.* **2003**, *24*, 1740–1747.

202. Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057-1065.

203. Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, K.-R. SpookyNet: Learning Force Fields with Electronic Degrees of Freedom and Nonlocal Effects. *Nat. Commun.* **2021**, *12*, 7273.

204. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. 2017. arXiv preprint arXiv: 1706.03762, https://arxiv.org/abs/1706.03762 (accessed May 6, 2023).

205. Neugebauer, H.; Bädorf, B.; Ehlert, S.; Hansen, A.; Grimme, S. High-throughput Screening of Spin States for Transition Metal Complexes with Spin-polarized Extended Tight-binding Methods. *J. Comput. Chem.* **2023**, *44*, 2120–2129.

206. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. *Adv Neural Inf Process Syst*. **2021**, *34*, 8780–8794.

207. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Adv Neural Inf Process Syst*. **2019**, *32*.

208. Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst*. **2020**, *33*, 6840–6851.

209. Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; van den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Adv Neural Inf Process Syst*. **2021**.

210. Baranchuk, D.; Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A. Label-Efficient Semantic Segmentation with Diffusion Models. 2021.arXiv preprint arXiv: 2112.03216, https://arxiv.org/abs/2112.03126 (accessed May 6, 2023).

211. Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; Babenko, A. Label-Efficient Semantic Segmentation with Diffusion Models. 2021. arXiv preprint arXiv: 2112.03126, https://arxiv.org/abs/2112.03126 (accessed May 6, 2023).

212. Amit, T.; Nachmani, E.; Shaharbany, T.; Wolf, L. Segdiff: Image segmentation with diffusion probabilistic models. 2021. arXiv preprint arXiv:2112.00390. https://arxiv.org/abs/2112.00390 (accessed May 6, 2023).

213. Lopez Alcaraz, J. M.; Strodthoff, N. Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models. 2022. arXiv preprint arXiv:2208.09399 https://arxiv.org/abs/2208.09399 (accessed May 6, 2023).

214.Chen, X.; Zhang, X.; Zen, H.; Weiss, R. J.; Norouzi, M.; Chan, W. WaveGrad: Estimating gradients for waveform generation. 2020. arXiv preprint arXiv:2009.00713 https://arxiv.org/abs/2009.00713 (accessed May 6, 2023).

215.Avrahami, O.; Lischinski, D.; Fried, O. Blended diffusion for text-driven editing of natural images. 2022. arXiv preprint arXiv: 2111.1481 https://arxiv.org/abs/2111.14818 (accessed May 6, 2023).

216.Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. 2022. arXiv preprint arXiv:2204.06125 https://arxiv.org/abs/2204.06125 (accessed May 6, 2023).

217.Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. 2022. arXiv preprint arXiv: 2203.17003 https://arxiv.org/abs/2203.17003 (accessed May 6, 2023).

218.Igashov, I., Stark, H., Vignac, C., Satorras, V. G., Frossard, P., Welling, M., Bronstein, M., and Correia, B. Equivariant 3d-conditional diffusion models for molecular linker design. 2022. arXiv preprint arXiv: 2210.05274 https://doi.org/10.48550/arXiv.2210.05274 (accessed May 6, 2023).

219.Cao, C.; Cui, Z-, X.; Liu, S; Liang, D.; Zhu, Y. High-Frequency Space Diffusion Models for Accelerated MRI. 2022. arXiv preprint arXiv: 2208.05481 https://doi.org/10.48550/arXiv.2208.05481 (accessed May 6, 2023).

220.Xie, Y.; Li, Q. Measurement-conditioned Denoising Diffusion Probabilistic Model for Under-sampled Medical Image Reconstruction. 2022. arXiv preprint arXiv:2203.03623 https://arxiv.org/abs/2203.03623 (accessed May 6, 2023).

221.Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. arXiv preprint arXiv: 1503.03585 https://arxiv.org/abs/1503.03585 (accessed May 6, 2023).

222.Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323*, 844-853.

223.Mohs, R. C.; Greig, N. H. Drug Discovery and Development: Role of Basic Biological Research. *Alzheimers Dement (N Y)* **2017**, *3*, 651–657.

224. Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided. Mol. Des.* **2013**, *27*, 675–679.

225. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44,* D1202–D1213.

226. Gao, K.; Duc Duy Nguyen; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682–5698.

227. Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci*. **2018**, *4*, 268–276.

228. You, J.; Ying, R.; Ren, X.; Hamilton, W. L.; Leskovec, J. GraphRNN: Generating Realistic Graphs with Deep Auto-regressive Models. 2018. arXiv preprint arXiv: 1802.08773 https://doi.org/10.48550/arXiv.1802.08773 (accessed May 6, 2023).

229. Cao, N. D.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. 2018. arXiv preprint arXiv: 1805.11973 https://doi.org/10.48550/arXiv.1805.11973 (accessed May 6, 2023).

230. Kaitoh, K.; Yamanishi, Y. Scaffold-Retained Structure Generator to Exhaustively Create Molecules in an Arbitrary Chemical Space. *J. Chem. Inf. Model.* **2022**, *62*, 2212-2225.

231. Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based de Novo Drug Discovery Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *60*, 77–91.

232. Duan, C.; Nandy, A.; Terrones, G.; Kastner, D. W.; Kulik, H. J. Active Learning Exploration of Transition-Metal Complexes to Discover Method-Insensitive and Synthetically Accessible Chromophores. *JACS Au* **2022**, *3*, 391–401.

233.Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. New Strategies for Direct Methane-To-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts. *JACS Au* **2022**, *2*, 1200–1213.

234.Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *J. Med. Chem.* **2016**, *5*
235.*9*, 4062–4076.

236.Hu, Y.; Stumpfe, D.; Bajorath, J. Recent Advances in Scaffold Hopping. *J. Med. Chem.* **2016**, *60*, 1238–1246.

237.Cornia, A.; Fabretti, A. C.; Garrisi, P.; Mortalò, C.; Bonacchi, D.; Gatteschi, D.; Sessoli, R.; Sorace, L.; Wernsdorfer, W.; Barra, A.-L. Energy-Barrier Enhancement by Ligand Substitution in Tetrairon(III) Single-Molecule Magnets. Angewandte Chemie **2004**, *116*, 1156–1159.

238.Langley, S. K.; Chilton, N. F.; Moubaraki, B.; Murray, K. S. Single-Molecule Magnetism in { $\mathrm{Co_2^{III}Dy_2^{III}}$ }-Amine-Polyalcohol-Acetylacetonate Complexes: Effects of Ligand Replacement at the Dy (III) Sites on the Dynamics of Magnetic Relaxation. *Inorg. Chem. Front.* **2015**, *2*,867–875.

239.Patra, M.; Gasser, G. The Medicinal Chemistry of Ferrocene and Its Derivatives. *Nat Rev Chem* **2017**, *1*, 1–12.

240.Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. 2020*,* arXiv preprint arXiv: 2009.0141 https://doi.org/10.48550/arXiv.2009.01411 (accessed May 6, 2023).

241.Arunachalam, N.; Gugler, S.; Taylor, M.; Duan, C.; Nandy, A.; Jon Paul Janet; Meyer, R.; Jonas Albrecht Oldenstaedt; Daniel; Kulik, H. J. Ligand Additivity Relationships Enable Efficient Exploration of Transition Metal Chemical Space. *J. Chem. Phys.* **2022**, *157*, 184112-184127.

242.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.

243.Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973–13986.

244. Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org/ (accessed May 6, 2023).

245. Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. J. Chem. Inf. Model. **2020**, *60*, 1983–1995.

246. Huang, Y.; Peng, X.; Ma, J.; Zhang, M. 3DLinker: An E (3) Equivariant Variational Autoencoder for Molecular Linker Design. 2022. arXiv preprint arXiv: 2205.07309 https://doi.org/10.48550/arXiv.2205.07309 (accessed May 6, 2023).

247. Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J Cheminform* **2009**, *1*, 8.

APPENDIX A: TABLES

Table 12. The 771 complexes in Zinc_60 data set.

CSD Code	Formula	N_con
SESWET	C16H14O8Zn	39
VIGFIA	C14H30N4O2Zn	51
KEYVES	C27H22N2O7Zn	59
BEZDER	C14H12N2O6Zn	35
BUDBEI	C6H8N6O6Zn	27
GUKZOD	C19H13N7O6Zn	46
ATAFEG	C20H18N12Zn	51
WEGHOH	C12H10N2O6Zn	31
HATCEN	C21H26N2O5Zn	55
LUGTAJ	C13H18N4O3Zn	39
MPEZNC	C16H12N4Zn	33
HAVBUC	C26H20N6O2Zn	55
QITRUH	C13H13N5O6Zn	38
VUYBEX	C16H12N4O5Zn	38
QISMUD	C18H22N4O6Zn	51
TOXMIE	C13H14N4O8Zn	40
GEPDAH	C26H18N4O2Zn	51
GARQUM	C22H20N10O2Zn	55
KERLOJ	C10H10N4O10Zn	35
QEBFOV	C26H24N2O7Zn	60
SANVOT	C24H18N6OZn	50
UGIBAP	C8H10N4O8Zn	31
BONRIF	C20H22N4Zn	47
AMOXZN	C4H8N2O8Zn	23
NOBHIX	C24H16N6O13Zn	60
BOMQEZ	C22H24N6O4Zn	57
ISUYAW	C30H20N2O4Zn	57
CEXPOM	C22H18O8Zn	49
EBAXUD	C17H14N4O6Zn	42
AHITUI	C16H22N4O4Zn	47
AHUTED	C17H19N3O4Zn	44
QOBTAE	C11H22N2OZn	37
QOBSOR	C12H24N2OZn	40
XUNDAM	C26H20N4O4Zn	55

Table 12. (cont’d)

VILBOK	C14H12N2O6Zn	35
YELXAR	C26H22N8O2Zn	59
VEVVID	C16H21N3O6Zn	47
SEKXOY	C20H26N8O4Zn	59
RIKXUH	C18H24N2Zn	45
LOSLUB	C26H20N6O4Zn	57
WENQEO	C12H12N2O6Zn	33
LOPYOE	C15H25N3O4Zn	48
IRATOL	C18H24N2O4Zn	49
IDANOS	C16H16N10O10Zn	53
MALFIO	C12H14O8Zn	35
FAYMIC	C17H13N3O4Zn	38
BUXZUQ	C7H11NO8Zn	28
SOTCAG	C28H24N2O2Zn	57
GUWGIP	C8H14N8O4Zn	35
SIHFEX	C18H18N10Zn	47
UCIFOC	C28H20N6Zn	55
GANZIG	C22H18N12O2Zn	55
RUDWIX	C17H17N3Zn	38
GILWOP	C24H17N5O8Zn	55
MUDMAA	C18H14N4Zn	37
LAJMUE	C16H38N2Zn	57
HAYNOM	C14H14N12O4Zn	45
BUYBAZ	C13H13N3O5Zn	35
ZERHEL	C18H18N12O8Zn	57
CORCAP	C18H16N2O10Zn	47
XIYREC	C10H22N2O4Zn	39
SOSHIS	C15H22N4O4Zn	46
SUQCIR	C6H12N6O10Zn	35
TEWNEO	C10H10N4O6Zn	31
VIPJOU	C24H25N5O4Zn	59
IMIMEY	C14H18O10Zn	43
FORKED	C15H22N4O4Zn	46
CIXZIV	C26H23N3O4Zn	57
BIYJAU	C14H22N4O4Zn	45
ERIKIA	C14H16N10O2Zn	43
MOLTEM	C20H18N6O9Zn	54

Table 12. (cont’d)

BIGXOH	C12H20N2O2Zn	37
CAPWAU	C25H26N2O5Zn	59
KUJRAI	C16H20O2Zn	39
HAGBIA	C14H6N2O8Zn	31
YENSEQ	C22H12N12Zn	47
FOKXIO	C18H23N5O6Zn	53
GOXFII	C10H14O10Zn	35
BUZQUK	C12H12N2O8Zn	35
KIMWUA	C12H19N5O5Zn	42
UGINEE	C33H21N5Zn	60
RILBIA	C26H26N4Zn	57
QUYDIZ	C24H16N6O12Zn	59
EHEWIY	C11H9N5O9Zn	35
TIHMUV	C18H26N2Zn	47
QAHTOK	C27H18N4O4Zn	54
PESHEB	C18H14N4O2Zn	39
QEXQAO	C24H20N6O4Zn	55
WUZNIP	C16H18O10Zn	45
INOCIY	C10H14N4O8Zn	37
OXUVAF	C16H18O10Zn	45
QAYGOM	C14H13N3O7Zn	38
PIKFOF	C20H22N8O4Zn	55
KESTAF	C18H14N4O8Zn	45
ZACZOS	C12H12N10O4Zn	39
UQAFOI	C18H18N4O6Zn	47
RUQDAK	C18H18N2O4Zn	43
MIDKUF	C26H22N4O2Zn	55
LATKIB	C21H32N2O4Zn	60
HERGIU	C16H28N6O8Zn	59
WEGBEP	C12H24N6O12Zn	55
HOTFAZ	C10H10N4O10Zn	35
DEZGOH	C21H35N3Zn	60
DEXSUX	C24H30N4OZn	60
KEGJUC	C14H18N4O4Zn	41
BUYBED	C13H12N2O6Zn	34
XUGJAL	C20H18N4O6Zn	49
HOPTIS	C6H13N17O2Zn	39

Table 12. (cont’d)

CEWBUC	C26H18N4O9Zn	58
XIWTUU	C28H20N6Zn	55
QIMRIQ	C12H12N2O6Zn	33
DENSIB	C24H16N4O6Zn	51
ODIGEO	C20H18N4O4Zn	47
GIYLIM	C16H18N2Zn	37
CIBLIK	C14H18N8O6Zn	47
KOKVIQ	C12H20N8O6Zn	47
AFEJAX	C22H20N2O3Zn	48
OBOJUM	C12H14N4O7Zn	38
HUVTAT	C21H17N7O6Zn	52
AYIFUL	C16H14N4O4Zn	39
YABNIA	C18H20N6O6Zn	51
IDATOZ	C14H10N4O6Zn	35
ONAFIT	C21H22N2O2Zn	48
XACMIY	C20H24N4O2Zn	51
GIWLOO	C18H17N5O4Zn	45
ECOZAA	C20H22N4O6Zn	53
WIKZOI	C14H22N12O8Zn	57
KUYGIV	C8H12N8O7Zn	36
EBAVOU	C22H24N2Zn	49
VOGFIG	C12H26N2Zn	41
YAMXAO	C17H15N5O6Zn	44
LELFOZ	C17H18N2Zn	38
SAPNOO	C20H16N2O3Zn	42
VIQFAE	C24H14N6O12Zn	57
QERHEB	C22H16N4O4Zn	47
IZAVEK	C26H18N4O9Zn	58
KUBVOU	C16H22N2O5Zn	46
URONUL	C16H28N2O2Zn	49
QEZLUD	C8H18Zn	27
YEBYAI	C16H18N4O6Zn	45
OLOZEV	C14H26N4Zn	45
FUHYOX	C8H14O12Zn	35
EJEBIF	C12H14N4O10Zn	41
LALTID	C25H21N3O4Zn	54
GIRMEA	C16H30N6O4Zn	57
JUXLUL	C12H14N4O8Zn	39

Table 12. (cont'd)

PUVWEK	C12H12N4O10Zn	39
WABSAV	C12H16N10O4Zn	43
TIHDAR	C18H14N4O8Zn	45
POCVAG	C20H14N10Zn	45
YOXFEX	C25H32N2Zn	60
KUJREM	C18H24O3Zn	46
VOXNED	C26H22N4O2Zn	55
WEZGEO	C18H12N4O5Zn	40
FAHCUO	C16H38N2Zn	57
EHOYIK	C18H26N4O4Zn	53
DURFUS	C10H22N2O4Zn	39
NICAZN	C12H16N2O8Zn	39
NEXKUX	C15H19NO5Zn	41
OZIMOB	C17H13N11OZn	43
YUVPIQ	C25H26N4O4Zn	60
UGOKAC	C25H16N4O3Zn	49
UCEDEL	C18H18N4O4Zn	45
IJOFIY	C20H21N7OZn	50
BENLAJ	C20H20N6O8Zn	55
SATKOO	C14H25NO5Zn	46
BONXAE	C21H17N3O4Zn	46
GIJMAO	C24H19N3O4Zn	51
GIRMEB	C6H14N12O4Zn	37
GOSXET	C18H31N3Zn	53
ICIZUQ	C18H22N2Zn	43
QOLDEB	C16H15N3O5Zn	40
PEHDAK	C21H29N5OZn	57
ACOCIF	C15H20N2O4Zn	42
YEGQAC	C20H22N6O6Zn	55
ECUQAV	C7H9N3O4Zn	24
VUQCEQ	C20H18N2O6Zn	47
GIFDEH	C2H10N12O8Zn	33
ETUBAX	C8H10N4O6Zn	29
PUVPAY	C26H22N4O6Zn	59
RIKYIW	C24H22N4Zn	51
JIFKAN	C14H18N2O6Zn	41
YOWCAQ	C11H15NO8Zn	36
XOKGIQ	C16H16N2O10Zn	45

Table 12. (cont'd)

XOTVOS	C18H24N2O12Zn	57
ICEXUK	C25H29N3O2Zn	60
FORKAZ	C14H22N4O2Zn	43
RAHPIC	C20H14O8Zn	43
QERMIK	C5H10N2O8Zn	26
XICBUI	C24H18N2O11Zn	56
WOFNUC	C20H36N2OZn	60
BOFWOK	C16H18N4O6Zn	45
RUXGOG	C8H12N6O6Zn	33
NASFIW	C6H14O8Zn	29
PAWGOK	C16H26N4Zn	47
MATLEB	C11H10N4O6Zn	32
MAZTEO	C20H24N2Zn	47
POWDOV	C15H15NO5Zn	37
JEKMUH	C16H22N4Zn	43
EJIPAP	C10H15N9Zn	35
PUVDUF	C10H16N2O8Zn	37
GENWAY	C12H16N10O6Zn	45
ZOMWIJ	C16H18N8O6Zn	49
ABASAA	C13H19N9OZn	43
NUGRAK	C16H22N4O6Zn	49
JAMPIZ	C18H22N4O6Zn	51
LUFHOL	C6H12N10O8Zn	37
XEJZUJ	C19H19N3O5Zn	47
SOSHAK	C18H24N4O4Zn	51
CUDXIJ	C16H28N2Zn	47
AZOLAD	C24H20N4O4Zn	53
OXOXEG	C22H26N2O8Zn	59
EBAYEO	C21H20N2O4Zn	48
FOWHEF	C24H18N4O10Zn	57
POKPUD	C10H20N4O5Zn	40
FUGLOJ	C26H22N2O4Zn	55
TEJHUM	C16H20N2O8Zn	47
GIMWAC	C12H28N2O4Zn	47
TICXEK	C30H18N2O6Zn	57
GEVXEK	C28H22N2O2Zn	55
XETHEK	C10H10N12O2Zn	35
ATULUY	C18H22N4O12Zn	57

Table 12. (cont’d)

YEFYUF	C20H14N4O5Zn	44
YIXSAB	C16H18N4O6Zn	45
GOMMUR	C20H16N8O4Zn	49
MAHZUR	C25H20N2O2Zn	50
YOXDUL	C19H24N2Zn	46
DEVTEE	C12H16N10O4Zn	43
NUZXOX	C16H16N2O16Zn	51
RURFOB	C22H32N4Zn	59
EQEFOW	C17H15N5O6Zn	44
ATUMUZ	C14H18N4O4Zn	41
ECAQIL	C26H20N2O10Zn	59
EFOXAB	C22H14N6O4Zn	47
ZEXHOZ	C6H14N4O6Zn	31
LIVYOE	C20H30N2Zn	53
BOCRUG	C12H6N2O6Zn	27
QELPAB	C28H22N2O4Zn	57
LODZOV	C22H18N12O2Zn	55
CADHIY	C18H26N4Zn	49
IROLAC	C22H20N2O6Zn	51
OYAKAB	C26H16N6O4Zn	53
XUCGEI	C21H20N6O6Zn	54
WELSOW	C18H32N2O4Zn	57
VUWJIH	C13H11N5O4Zn	34
FESBOV	C17H26N4Zn	48
UMOVIC	C18H16N4O5Zn	44
ROQQIZ	C22H34N2Zn	59
XIYNOJ	C26H22N2O4Zn	55
RUVDET	C23H21N5O8Zn	58
GILSIF	C12H12N6O6Zn	37
KEKHEP	C16H18N4O9Zn	48
NAQWEK	C32H18N2O5Zn	58
OBIHEN	C20H18N4O6Zn	49
INICZN	C12H16N2O8Zn	39
IGOPAW	C9H12N4O6Zn	32
UWIQEX	C13H16N6O5Zn	41
MIJVAF	C24H22N4O4Zn	55
OZOQIF	C16H28N2O2Zn	49
TAPVOY	C20H26N2O4Zn	53

Table 12. (cont'd)

LELFAL	C19H20O3Zn	43
LUGREK	C18H23N3O2Zn	47
HAVBOW	C26H18N4O2Zn	51
HECMEI	C14H10N2O9Zn	36
RUSDIS	C23H28N4Zn	56
OJUJAE	C12H16N4O5Zn	38
INOCEU	C10H14N4O8Zn	37
FOJVIL	C20H18N4O8Zn	51
DIYYAL	C6H12N6O10Zn	35
OCOZAJ	C14H24N4Zn	43
ILONIG	C14H18N4O4Zn	41
ROPZOL	C20H30N2Zn	53
KOVQOD	C18H14N4O6Zn	43
XACMOE	C22H28N4O4Zn	59
TUMQOI	C24H16N10Zn	51
LICMEQ	C25H19N3O2Zn	50
LEWFAV	C20H18N2O3Zn	44
PYACZN	C14H16N2O6Zn	39
LEPVUA	C18H26O8Zn	53
FIDWUL	C6H12N2O8Zn	29
SOSHOY	C15H24N4O4Zn	48
XALWUF	C10H14N2O8Zn	35
KAGLOU	C16H18N4O4Zn	43
QIPXUI	C16H20N6O2Zn	45
RIRVIY	C6H8N6O6Zn	27
GEKVUQ	C24H20N6O2Zn	53
YIDSEK	C14H12N2O10Zn	39
IWOPUF	C16H19N5O4Zn	45
LUFKOM	C10H18N4O8Zn	41
PEGQEX	C18H22O13Zn	54
VOJMEP	C8H17NO6Zn	33
EQECOT	C23H21N5O6Zn	56
LELJAO	C20H30N2Zn	53
NIWNOX	C26H20N2O7Zn	56
DORGUM	C14H16N2O9Zn	42
MIQVAL	C8H15N5O4Zn	33
MAHRAP	C10H10N4O6Zn	31
YUSZIX	C20H21N3O11Zn	56

Table 12. (cont'd)

QERPAE	C22H24N2O4Zn	53
QIQLAD	C14H16N6O8Zn	45
POMMOV	C4H10N10O8Zn	33
TISBUS	C16H14N2O6Zn	39
PIZLAN	C12H16N10Zn	39
NAFGOT	C34H18N2O4Zn	59
YADNUO	C30H20N2O7Zn	60
CEGFIE	C25H25N3O2Zn	56
DAFLIG	C28H20N10Zn	59
DEXYEL	C24H18N12Zn	55
PUTPUP	C18H36N2Zn	57
JUVCEK	C18H18O16Zn	53
ROGXAM	C14H22O12Zn	49
CEHCAT	C14H12N2O10Zn	39
OZOQOL	C18H32N2O4Zn	57
RUVDIX	C20H21N5O6Zn	53
KIWTOA	C20H22N2O4Zn	49
MAVLEB	C22H20N2O6Zn	51
XOXYOZ	C18H30N2OZn	52
SEZVID	C12H14N4O6Zn	37
ATUPEM	C14H12N6O4Zn	37
DAPQAN	C18H18N8O4Zn	49
HODTEZ	C23H19N7O3Zn	53
SESPIQ	C28H20N2O7Zn	58
DUHJUM	C19H13N3O2Zn	38
YUMWOT	C12H16N6O6Zn	41
CAPXEY	C21H16N4O8Zn	50
GOSMOS	C10H8N10O4Zn	33
BIPFAI	C20H16N10Zn	47
YAZXAZ	C26H20N6O4Zn	57
QURTED	C18H22N4O4Zn	49
VOXNIH	C26H22N4O2Zn	55
SIZPOJ	C15H16N4O4Zn	40
DUVSIY	C25H22N4Zn	52
YEXJOB	C32H18N8OZn	60
BPPZNH	C16H8N10O12Zn	47
ERURAL	C20H32N2Zn	55
NOWHEO	C26H23N3O6Zn	59

Table 12. (cont’d)

DEHKIM	C11H13N5O6Zn	36
GERTEE	C22H14N4O4Zn	45
XIBYOX	C20H14N6O5Zn	46
KAGMAH	C17H14N4O6Zn	42
POWVON	C24H20N2O7Zn	54
IDAVER	C24H16N6O4Zn	51
CUMBES	C16H15N3O5Zn	40
HOXYUQ	C20H15N3O5Zn	44
TAHYOS	C12H26O6Zn	45
VIFTAI	C24H22O6Zn	53
XECNUQ	C18H19N3O4Zn	45
PIKGAS	C22H24N6O6Zn	59
KAZFIC	C20H14O8Zn	43
IGATUF	C12H28N2O6Zn	49
EXUYOL	C10H9N9Zn	29
WAPMUW	C28H24N2O2Zn	57
TIBPEB	C22H16N2O10Zn	51
MASRED	C24H22N4O4Zn	55
ZOPQUS	C8H18N2O4Zn	33
KIYSEQ	C18H18O6Zn	43
WISNET	C14H16N4O9Zn	44
UNOPUJ	C26H24N2O6Zn	59
SAJDAK	C4H10Zn	15
SOGLOS	C26H19N3O8Zn	57
EFOKIW	C22H24N6O4Zn	57
UGINOO	C29H19N5Zn	54
OHORIM	C12H12N10O2Zn	37
KEFNUF	C25H28N4OZn	59
IKOYUE	C26H26N4O3Zn	60
KIKNAT	C16H26N4O8Zn	55
ULICOH	C23H19N7O2Zn	52
SIRMOX	C22H22N2O8Zn	55
FASBEG	C5H10N2O8Zn	26
SAGYEH	C18H18N2O4Zn	43
DENSAT	C19H20N4O7Zn	51
PATSOT	C28H20N6O2Zn	57
ZETDOT	C26H20N2O2Zn	51
MIRSEM	C16H18O6Zn	41

Table 12. (cont'd)

MIQVEP	C8H15N5O4Zn	33
TEFDOY	C18H18N2O6Zn	45
SUNNAR	C27H17N7O2Zn	54
TARWIT	C24H20N4O4Zn	53
MESZAM	C22H18N4O6Zn	51
FOMKOH	C22H28N2O2Zn	55
UZIFAL	C24H21N5O4Zn	55
FOJTUV	C20H18N6O6Zn	51
IXOCIH	C10H14N8O6Zn	39
MIRXUI	C24H20N10O2Zn	57
PAGFUA	C24H26O2Zn	53
QAFHAH	C12H12N2O8Zn	35
UYAQAO	C18H14N10Zn	43
QEJLEX	C10H14N4O4Zn	33
HUCHOD	C12H12N4O5Zn	34
ISAVAZ	C28H22N2O4Zn	57
XAHGIY	C12H8N10O2Zn	33
NEYTES	C17H15NO6Zn	40
BAXROH	C10H20N2O10Zn	43
SUXREJ	C14H18N4O8Zn	45
BIYZUF	C18H22N8O8Zn	57
NALKUI	C12H24N2O8Zn	47
WECKUN	C26H20N4O8Zn	59
TEKJEA	C18H34N4O2Zn	59
WIWCUB	C18H14N4Zn	37
MEXFOM	C14H20N10Zn	45
REWFAB	C20H16N2O7Zn	46
MOVBUV	C24H21N5O4Zn	55
IKADIJ	C26H20N4O4Zn	55
MALFOU	C14H18N2O5Zn	40
IKEYOM	C10H18O6Zn	35
RUQLIA	C25H19N5O6Zn	56
ONATED	C10H22O6Zn	39
GEPLOD	C10H13NO7Zn	32
PAXZOG	C6H16O2Zn	25
JIZCAW	C14H36N6Zn	57
NUFJAB	C16H16N10O14Zn	57
ERUQUE	C18H28N2Zn	49

Table 12. (cont'd)

QANSAB	C12H11N5O7Zn	36
REWHIN	C12H14N8O2Zn	37
AYISEH	C20H24N2O5Zn	52
BEKPUC	C8H24N6O4Zn	43
XISBOR	C12H12N2O8Zn	35
QABYAV	C12H16N6Zn	35
QURSOM	C18H22N4O4Zn	49
FEHBAY	C22H28N2O2Zn	55
RATXER	C14H20N2O3Zn	40
GEGQUE	C12H9N5Zn	27
KAPDIQ	C20H14N8O12Zn	55
QIRFUV	C12H14N4O10Zn	41
LUSVEA	C19H30N4O2Zn	56
EKIGEM	C12H18N8O6Zn	45
UNELEF	C10H16N2Zn	29
GILSOL	C12H12N6O6Zn	37
QEMDOB	C11H11NO6Zn	30
EXOFAZ	C21H24N2O7Zn	55
SAQHUP	C12H17NO8Zn	39
HUXVAX	C19H30N4O2Zn	56
WUXCUO	C20H16N4O4Zn	45
MUSPUL	C13H18N4O4Zn	40
RIKZIX	C28H26N4Zn	59
JAMNIX	C10H22N4O6Zn	43
ICOBEI	C14H30N2Zn	47
HIQBIS	C28H24N2O4Zn	59
ZIQBIM	C16H16N2O10Zn	45
ONISUA	C15H19N7O6Zn	48
SIDNAW	C20H22N4O4Zn	51
PAKXEG	C17H21N7O6Zn	52
MIWSOC	C24H22N4O6Zn	57
COLTUT	C18H17N3O9Zn	48
DENRUM	C14H16N2O8Zn	41
LAJRIZ	C22H16N4O10Zn	53
FOMLAU	C22H28N2O4Zn	57
KAZFAU	C20H14O8Zn	43
WUDMEN	C18H20N2O4Zn	45
ABOWOF	C16H20N2O5Zn	44

Table 12. (cont’d)

POMMEL	C4H8N8O8Zn	29
GOSMIM	C12H8N10O2Zn	33
CIBTIS	C25H17N5Zn	48
WUDMAJ	C11H15NO5Zn	33
UQEBEZ	C18H18N2O6Zn	45
YABNAS	C16H16N6O6Zn	45
UNOKIS	C18H20O11Zn	50
FORJUS	C15H26N4Zn	46
QAHZOQ	C24H19N7O7Zn	58
SAJDEO	C2H6Zn	9
ICIMAJ	C26H20N4O4Zn	55
DOSQUA	C7H11NO9Zn	29
QOLMEK	C19H23NO5Zn	49
PAXPOU	C12H30N2Zn	45
NAPMUM	C16H36N2Zn	55
ERIMIC	C15H19NO4Zn	40
RIKGEA	C16H16N2O2Zn	37
MASYAH	C25H23N5O3Zn	57
CAPDIG	C12H14N4O4Zn	35
XAXBIJ	C10H21NO7Zn	40
AGEJOM	C20H21N3OZn	46
QOLDIF	C18H16N4O5Zn	44
PESKEE	C24H18N2O4Zn	49
FULSOV	C10H12N6O16Zn	45
GOQQAG	C16H16N10O6Zn	49
WILHOR	C14H16N10O6Zn	47
DUVKOU	C8H12N10O6Zn	37
AXUQER	C12H20N10Zn	43
UZICIQ	C16H20N6O6Zn	49
XEZTII	C4H14N8O8Zn	35
HIQBEO	C26H18N4O6Zn	55
XIZVAG	C16H22N4O4Zn	47
YOWMAY	C26H22N4O6Zn	59
QAHVAY	C20H14O10Zn	45
HOSXUJ	C8H10N12O2Zn	33
AXILUO	C12H16N10O6Zn	45
PIKVAJ	C17H19N5O5Zn	47
ZUTNUZ	C22H24N2O5Zn	54

Table 12. (cont'd)

TIGJEZ	C6H14N2O4Zn	27
YAFZEL	C16H14N2O6Zn	39
VORNOF	C12H18N8O8Zn	47
QUQVAB	C12H16N6O2Zn	37
GEZLAA	C26H24N4O2Zn	57
ICOBAE	C16H18N2Zn	37
MUZNIG	C24H16N4O8Zn	53
GISQAA	C14H32N10Zn	57
XILGAB	C6H14N12O2Zn	35
REZDEI	C10H12N6O6Zn	35
NENCOZ	C14H15NO6Zn	37
VOLWAV	C16H16N4O4Zn	41
ULASUW	C21H29N3Zn	54
BIGXIB	C15H25N3OZn	45
CUDXOP	C13H22N2Zn	38
SOJXUK	C6H13NO9Zn	30
CELKAF	C14H16N2O10Zn	43
OFUFEA	C15H14N4O6Zn	40
USIVEZ	C20H12N4Zn	37
EHOYAC	C18H26N4O4Zn	53
MIDKOZ	C24H18N4O2Zn	49
DICWAP	C20H17N5O5Zn	48
DUMLIH	C20H22N8O4Zn	55
JIHROK	C16H20N2O12Zn	51
WELSIQ	C18H32N2O2Zn	55
YEVNOD	C24H24N2O9Zn	60
OLOWIV	C9H14N6O6Zn	36
DISWUA	C24H22N4O8Zn	59
TERQEM	C14H14N2O5Zn	36
FIDWOF	C4H8N2O8Zn	23
AZILOK	C22H28N2O2Zn	55
ICEWOD	C24H32N2OZn	60
XECRAZ	C10H10N4O6Zn	31
CODRIX	C14H18N2Zn	35
QOLWUJ	C17H24N2O4Zn	48
VIKMEK	C19H12N6O6Zn	44
PUKLUE	C26H22N4O4Zn	57
ASOCOC	C10H14O4Zn	29

Table 12. (cont'd)

OGANAM	C24H16N4O8Zn	53
YUZROC	C16H17N7O8Zn	49
ILAZIG	C28H22N2O4Zn	57
YAYYED	C20H18N4O6Zn	49
ATUMAF	C18H22N4O12Zn	57
KIWTIU	C22H22N2O4Zn	51
LIYXIA	C28H16N10Zn	55
EXIFUN	C21H24N4O6Zn	56
COLCUD	C28H24N2O5Zn	60
TENGEB	C22H28N2OZn	54
HINLIC	C11H17N9Zn	38
EYIXEQ	C21H19N5O8Zn	54
UQEBID	C28H22N2O4Zn	57
VUJTEA	C26H16N6O4Zn	53
UNEQUB	C12H16N2O10Zn	41
ASOZEP	C28H24N2O4Zn	59
XILHAB	C18H22N2O6Zn	49
GICFUW	C18H20N6O8Zn	53
LOFYIP	C26H20N6O2Zn	55
WEWVIG	C16H18N8O8Zn	51
KOVRAQ	C24H28N2O4Zn	59
SIGLUP	C17H24N6O2Zn	50
LUYBAI	C26H22N8O2Zn	59
DAMLAG	C20H26N4O5Zn	56
PAXPIO	C14H34N2Zn	51
DUCPAT	C14H26O6Zn	47
ECULAQ	C19H13N3O5Zn	41
BEZLID	C15H19N3O4Zn	42
MEHHIS	C9H10N6O9Zn	35
MAHQAO	C12H22O6Zn	41
FENVEA	C26H18N2O8Zn	55
YORRUU	C14H18N4O10Zn	47
PAXPUA	C19H36N2Zn	58
XACGUD	C14H16N6O6Zn	43
VIQMEO	C20H16N2O5Zn	44
VIRTIC	C14H26N2OZn	44
HILMOG	C18H20N6O8Zn	53
MATLIF	C12H12N4O6Zn	35

Table 12. (cont'd)

RACYAW	C20H14N2O5Zn	42
YAFZIP	C25H25N3O5Zn	59
METVIR	C12H30N2Zn	45
IWOLEN	C8H20N10O8Zn	47
WUXCOI	C22H16N4O4Zn	47
ROKDIF	C18H28N4Zn	51
EQIMEY	C10H14N2O8Zn	35
LIKXAF	C11H18Zn	30
VAQZUK	C17H19N5O4Zn	46
EZICAS	C24H16N4O6Zn	51
FIPRUR	C5H11N5O4Zn	26
DAXGUF	C19H16N2O5Zn	43
UGINII	C26H18N4Zn	49
SORBOQ	C18H22N8O11Zn	60
SIGXOX	C18H23N3OZn	46
MELYAE	C20H18N2O6Zn	47
IDIFIN	C22H22N4O8Zn	57
LEHCUY	C20H30N4O4Zn	59
DAMJUY	C19H24N4O4Zn	52
IJEDOR	C19H14N2O8Zn	44
BENDAB	C16H18N2O4Zn	41
PIFFEQ	C22H20N2O5Zn	50
XACNAR	C10H16N2O9Zn	38
SOBRAD	C14H12N6O6Zn	39
IDAVAN	C19H12N6O4Zn	42
CEHNAH	C15H18N10OZn	45
TAZYUP	C14H22N2Zn	39
YIGGAZ	C10H19N3O7Zn	40
DIYDUK	C18H20O12Zn	51
POFCUJ	C20H16N4O5Zn	46
JEVMIH	C18H18N4O2Zn	43
BEPJOY	C25H22N2O9Zn	59
DEMJIR	C18H18N10O3Zn	50
YAMXOA	C16H18O8Zn	43
SUPDIQ	C16H32N2Zn	51
NISVAO	C16H29NZn	47
COLWIK	C29H16N4O5Zn	55
TEDGUF	C20H22N2O4Zn	49

Table 12. (cont’d)

FODMEQ	C10H24N2Zn	37
LOBDIS	C28H22N4O2Zn	57
TOMTAR	C12H14N2O5Zn	34
RIJHAV	C15H27NZn	44
CIWDOC	C16H14O10Zn	41
YIHMAD	C24H18N4O2Zn	49
ISEXOU	C16H28N2O2Zn	49
ALOPAS	C14H14N2O5Zn	36
RIKGIE	C28H24N2O2Zn	57
ZUGVEC	C12H17NO6Zn	37
KUFBAP	C28H18N12Zn	59
GOGPAU	C17H26N4Zn	48
DUVMUD	C29H21N5O3Zn	59
VUGBOP	C24H26N4O4Zn	59
UYAJIO	C13H17N3O6Zn	40
QOBTOS	C18H32N4OZn	56
SELXEM	C22H30N2OZn	56
QITROB	C12H11N5O6Zn	35
TEFDIS	C18H18N2O6Zn	45
UNODOR	C18H20N2O12Zn	53
HOLTUX	C10H14O12Zn	37
FOSJIJ	C18H28N2O2Zn	51
OMUVUO	C22H18N6O4Zn	51
HIQSIK	C22H18N4O6Zn	51
NUNYAX	C22H24N2O4Zn	53
BOFROF	C16H20N6O6Zn	49
DISWOU	C24H24N4O7Zn	60
COBKOW	C18H10N8O4Zn	41
CAKGIH	C26H26N4Zn	57
WOBGAV	C16H20N6O4Zn	47
AKAPIN	C23H31N3Zn	58
PAMBEM	C20H22N6O8Zn	57
KAYPUW	C16H16N2O5Zn	40
MAZTAK	C16H16N2Zn	35
PEHDEO	C22H31N5OZn	60
JILGUJ	C24H18N4O2Zn	49
WIBHEW	C13H14N4O5Zn	37
LAYLAB	C22H18N2O7Zn	50

Table 12. (cont'd)

NABVIY	C14H20N2Zn	37
QANTOP	C24H16N10Zn	51
KAWSAD	C12H20N2Zn	35
RUSDAK	C22H26N4Zn	53
PEGQIB	C18H22O13Zn	54
LIVRUD	C25H22N2O7Zn	57
TISBOM	C14H14N2O6Zn	37
BIBBEU	C14H22N4O6Zn	47
HEFJEJ	C14H26N2OZn	44
FABZUF	C21H33N3O2Zn	60
QUHNIS	C12H14N4O10Zn	41
GIHSIA	C18H22N4Zn	45
TIPXIZ	C14H16N2O14Zn	47
NIQXER	C22H20N6O4Zn	53
FEQTUS	C5H10N2O8Zn	26
IFICEG	C20H16N4O5Zn	46
COVTEN	C22H30N4O2Zn	59
CAPDEC	C10H10N4O4Zn	29
GODFEK	C18H18N2O4Zn	43
CEFZUL	C14H16N2O5Zn	38
FEKGAE	C24H20N4O3Zn	52
REHDIS	C26H18N2O4Zn	51
ICOWED	C14H17N5O5Zn	42
BENLAH	C6H16N4O4Zn	31
FEVFAQ	C24H20N6O2Zn	53
CIBLIL	C30H20N6O2Zn	59
GOWQOY	C14H15N3O4Zn	37
HIVHIG	C14H22N4O4Zn	45
RACYEA	C20H16N2O6Zn	45
USIVAU	C14H17N3O7Zn	42
GEZLEE	C24H20N4O2Zn	51
DUCPUN	C24H22N4O2Zn	53
PIZTOK	C22H14N4O4Zn	45
BAQROA	C22H22N4Zn	49
WOGFEE	C22H26N2Zn	51
OMEWOU	C8H19N3O4Zn	35
MPLZNC	C16H12N4Zn	33
RUVDAP	C15H17N5O8Zn	46

Table 12. (cont’d)

BOVNOQ	C16H16N10O6Zn	49
XEBVOQ	C14H24N4O4Zn	47
ERIRAZ	C4H12N4O7Zn	28
LOFCEP	C20H22N2O4Zn	49
EMUKEC	C30H22N4Zn	57
WUKNAR	C14H12N2O10Zn	39
TEYDUZ	C28H24N2O4Zn	59
FEXWEM	C23H19N5O4Zn	52
IXOFUY	C28H22N2O6Zn	59
NIDLEU	C22H24N6O6Zn	59
EFOXEF	C28H18N6O4Zn	57
HUBMUM	C12H16N6O6Zn	41
KUMTUI	C24H32N2Zn	59
JIPBEP	C12H27N3O4Zn	47
NOJPUA	C20H22N4O2Zn	49
PIHKUP	C20H26N2O4Zn	53
CORLOM	C21H17N3O4Zn	46
IDATUF	C24H18N8O4Zn	55
LENFOB	C22H30N2OZn	56
VOBXER	C8H8N6O8Zn	31
TOGSUF	C22H18N4O8Zn	53
XILGUU	C26H20N4O4Zn	55
NENMAU	C16H14O4Zn	35
ABEGOG	C18H16N4O5Zn	44
ULAVUY	C19H26N2O4Zn	52
VOPVAY	C16H34N4Zn	55
OCOYUC	C14H29N3Zn	47
WESRAN	C10H10N4O10Zn	35
XUCGAE	C10H16N6O6Zn	39
CIBJUU	C26H24N2O6Zn	59
KIYRAN	C20H22N4O10Zn	57
PAHSOH	C24H18N2O6Zn	51
DITWEL	C25H17N3O5Zn	51
ZUDLUF	C10H16N6O6Zn	39
SOGLUY	C24H19N3O8Zn	55
ACUDUA	C20H20N12Zn	53
FOZGEH	C15H24N10O2Zn	52
YOTFOD	C16H16N18Zn	51

Table 12. (cont’d)

EYOTER	C6H12N2O8Zn	29
YUJMAT	C14H24N2O4Zn	45
WEDLUM	C17H32N2Zn	52
MIPYAO	C19H22N2O6Zn	50
NUGFEA	C14H16N6O4Zn	41
ERUREP	C20H32N2Zn	55
SOSHEO	C18H22N4O4Zn	49
ICEXIY	C26H28N2O2Zn	59
GAQYOM	C18H20N6O4Zn	49
AREWOJ	C10H10N4O6Zn	31
TASZAP	C22H19N5O4Zn	51
TECTUR	C18H18O16Zn	53
APURZN	C8H8N6O11Zn	34
NIDFEM	C22H18N4Zn	45
TOXDIV	C19H14N6O5Zn	45
LACGAA	C10H16N8O3Zn	38
KABWUH	C14H12N2O12Zn	41
VIJTUE	C8H12N2O10Zn	33
QITKEL	C18H14N12Zn	45
LUTPOG	C16H18N8O10Zn	53
VENGUR	C20H26N4O6Zn	57
MOHZAL	C18H18N4O4Zn	45
DEZNOM	C8H10N4O6Zn	29
BEZKOH	C8H12N6O4Zn	31
OPOLUC	C6H8N8O5Zn	28
IDAWAN	C16H28N2O4Zn	51
XUBPAM	C14H14N4O10Zn	43
TOPYEE	C26H22N4O2Zn	55
LARYOS	C21H26N4O5Zn	57
AYIYOX	C16H18O10Zn	45
WAQSIR	C21H21N5O8Zn	56
CERPAS	C17H28N4OZn	51
PIMTAI	C8H10N4O6Zn	29
INILOH	C20H32N2Zn	55
XISTAX	C10H18O6Zn	35
APUDIA	C24H20N6Zn	51
RUSDEO	C21H24N4Zn	50
ONUNER	C14H34N6Zn	55

Table 13. The 383 Fe (II) complexes in Fe (II)_80 data set.

refcode	index	N_atoms	N_con
GORMOS	5	59	3
LOWXON	8	37	56
NOZZOU	9	37	7
YOYVEQ	12	53	569
AMAVOB	14	19	99
APAFEH02	2	49	691
AXAKIT	20	55	306
AZOFOL	23	57	7
AZUHUY	24	55	4
BANSEQ	3	53	4
BEPZIF	28	53	2
BINZUW	30	58	78
BOMFER	32	31	16
BOVMEF	33	53	14
BUKRAB	34	53	18
CAQFEH	37	55	4
CARQUH	38	39	64
CETFAI	41	35	29
CEVTUT	42	55	29
CEYRAA	43	54	32
CIRTII	45	56	20
CODQAO	4	41	75
COMKUL	48	47	3
CUCVIH	51	57	9
CUDVUT01	53	57	186
DAJRAH	54	53	10
DAQHIO	56	57	3
DAQHOU	57	48	2
DETTOL	60	49	29
DIBJED	63	43	76
DIDXEV	65	49	15
DIHCUV	66	57	101
DOQRAC	69	52	22
DOYKIN	70	56	51
ECAJOH	72	38	42
ECOHOT	73	59	9
EKAKUY	74	51	1

Table 13. (cont'd)

EMIPIZ01	75	49	4
ESOSOW05	76	59	11
ESUQOY	77	53	3
ETOGEA	79	60	16
ETOHOL	80	55	71
EWIGEY	81	45	66
EXOTAN	84	58	25
FEGKEK	86	57	2
FEISXC01	90	49	323
FEJBOM	91	53	14
FEJJAG	93	54	10
FEWREG	95	57	9
FOGFIT	97	51	4
FOGFOZ	98	57	3
GAVDEN	103	39	96
GEDJUX	105	41	67
GEHBEB	108	56	12
GEJZAX	109	57	22
GLYCFE01	6	33	350
GUWTEX	114	56	10
HAKDOO	117	47	101
HENGIR	121	57	30
HEYMIH	122	43	26
HEYMON	123	49	86
HIKPEZ	124	59	7
HIVQEJ	126	59	4
HOLMOK	39	13	3
HOMXIT02	111	51	5
HOPJAY	127	57	17
HUQFAA	129	46	5
HUYDUB	130	47	9
IFIPAP	132	47	315
IGIHEL	133	59	184
IGURAF	134	59	10
IPIWAF	135	57	20
ISULAJ	137	53	22
IXOZOK	138	46	12
IXUZOQ01	139	59	8

Table 13. (cont'd)

JAQQIB	142	52	18
JITWUF	144	59	85
JOWGEH	145	43	5
JUDZUF	146	31	8
JURBEF	148	55	34
JUXPUN	149	49	23
KETMED	150	58	49
KETMON	151	55	25
KINQUT	155	55	5
KISRUA	156	40	55
KIWWUL	157	51	8
KIWXAS	158	48	25
KIXVUJ	159	46	13
LAXNON	162	49	291
LAXNUT	163	49	37
LONMIN	166	52	2
LUGWEQ	167	57	52
MEMYIN	172	51	23
MENXAF	173	51	153
MEQVEM01	175	59	5
MIKJAS	180	37	78
MUCREJ	184	59	5
NAJKIS	187	52	52
NANMIZ	188	58	46
NAXSOV	189	57	24
NEBHUX	190	49	90
NEBLIR	7	43	2
NEJBOT	191	35	15
NELGUG	192	53	16
NENTAB	193	55	3
NEVROW	194	51	65
NIPNUW	10	53	7
NOBYOU	195	37	423
NUDJED	196	51	109
NULGIL	197	54	2
OBUYIT	198	54	3
ODAJUB	199	50	18
OJIZUC	201	43	132

Table 13. (cont'd)

OWIHIM	207	48	3
PASFAR01	209	57	2
PAZXAP	211	49	4
PEGZIM	212	23	307
PEWHAC	214	55	50
PEWHEG	215	55	22
PURYIK	219	55	8
QERDIB	223	59	17
QETDOI	224	47	578
QIFBIR	226	58	7
QOQHEK	230	57	10
QOQHIO	231	60	31
QUWGUM	235	51	159
RAXWUK	237	52	17
RIPZAS01	239	59	11
RIYTIC	240	54	25
RIYTOI	241	54	5
RIZSOI	243	55	2
SAJHIX	245	47	4
SAPYIU	246	60	4
SAVQIR	247	49	18
SAVYUM	248	55	5
SEDJIU	249	51	28
SEGXUZ	250	49	9
SEKMEA	251	49	2
SIDMEA	256	55	725
SIXMES	257	51	217
SOFNEG	258	57	4
SOYVEK	262	55	5
SUJVAW	263	45	32
TAGLAO10	265	52	9
TAPUFE	266	33	2
TAWFIH01	267	55	10
TAYBUT	268	53	78
TEKJIE	269	57	883
TEVWUN	273	37	153
TIHTAF	275	53	38
TITRAS	277	59	381

Table 13. (cont’d)

ULEHUO	284	59	3
UMAXIQ	285	54	22
URABIZ	286	51	75
USIMOA	287	37	621
VIHCOF	294	59	32
VUGQAR	296	33	71
WAHPEC	298	37	5
WEWQOH	299	31	18
WEYWEC	301	53	7
WIGPOR	302	53	13
WIWBEK	306	51	78
WIYGOA	307	41	138
WULSEB	309	49	20
WUSBOD	310	54	20
XABSOJ	311	53	89
XAXNIW	313	51	4
XENBEX03	314	51	5
XEVKIU	315	52	14
XILLEK	317	27	29
XOPSEB	319	39	71
XOVQIK	320	39	11
YAQVIY11	322	50	9
YAZJEP	323	55	3
YILFEH	11	57	8
YOQKOG	327	51	2
ZEBSAA	330	55	26
ZEKGAZ	332	49	27
ZIMLUC	334	45	42
MUHMOU	335	58	9
DULDAS	338	55	6
DUBSUR	401	73	26
QOSZOQ	404	73	49
ACAHUJ02	405	79	476
ADEQIL	406	61	11
AFANEB	407	63	6
ALILEN	409	77	1019
AVINUO	411	62	4
AWUKEK	412	79	11

Table 13. (cont'd)

BAKGUR	416	73	2842
BEHCEY	419	61	38
BEQKOY	420	73	6
BOJLUI	422	71	84
BUDKUG	425	61	48
BUKDUH	426	75	2
BUNSIN	427	63	35
BUSVAO	428	67	24
CAYQOI	429	73	8
CIDDUR	432	71	24
CIDXAR	433	73	1
COTZUH	434	62	86
COWFAW	435	67	25
CUCPUM	402	67	12
DENYAZ	437	64	21
DEYNAX	438	62	10
DEYNIF	439	74	162
DEZMIF	440	73	458
DOBRIY	442	65	19
DOMQON	444	77	1
DOZROA	445	73	102
DUCFOW	446	61	12
DUDDUD	447	63	4
DUDHUH	448	75	4
DUVQAN	450	63	3
DUXGUA	451	63	82
EBORIW	452	62	10
ECAKUP	453	79	2
EHABAQ	454	65	136
EJONEW	455	63	8
EXOMOU	457	61	6
EXOVUI	458	69	2
EXOWIX	459	73	8
FEDSAL	460	75	30
FEHPYO	462	73	193
FEXWUC	464	71	295
FILCAF	465	67	25
FOGFIS	466	63	113

Table 13. (cont'd)

FOZKAH	469	75	26
FUFMAW	470	77	3
GEDFIE	473	61	19
GIMDEL	476	68	55
GIZZUL	478	63	4
GOLVEK	480	75	3
GUJJUQ	481	62	5
HATGIU	482	64	11
HIKPID02	486	65	10
HIPXOW	487	61	253
HIQFOE	488	79	19
HOGFER	489	80	78
IBUVUW	492	67	27
ICAHUP	493	79	113
ILASIY	495	61	294
IQIJAT	497	67	153
IQIWOW	498	79	4
ITOKEI	499	75	54
IWOTEV	500	71	13
JAFBOI03	501	61	17
JAMYOO	502	78	57
JAMZIJ	503	71	103
JAVVAG	506	67	3
JEFBOO	509	79	7
JOHCOZ	510	61	2618
JUVHIR02	513	61	72
KABWAO	514	79	20
KAFWAQ	515	67	3
KAHNAK	516	63	47
KALWUR	517	63	3
KAQYUZ	518	63	2
KATZUB	519	67	84
KEPDIT	520	69	4
KETLUS	521	75	27
KEXCAU	522	75	18
KITFAV	523	77	1
KOFMUP	525	75	3
KOKKUR02	526	61	7

Table 13. (cont'd)

LEXLIK	529	70	88
LINQUV	530	68	29
LINRAC	531	65	47
LINREG	532	65	47
LINROQ	533	64	40
LIQCEW	534	79	9
LOCTEF	535	66	5
MAGCUV	536	69	10
MAKJUD	537	78	670
MALGIQ	538	67	173
MELLOF	539	63	41
MENTEF	541	73	368
MUTMIZ	543	64	33
MUTXEG	544	62	9
NAVYEP01	546	62	50
NEBJAF	547	63	79
NERNAZ	550	63	8
NESVUD01	551	63	2
NEXLOS	552	77	4
NEXMAF	553	71	13
NIQFUP01	554	68	41
NISBIA	555	68	186
NOKFID	557	79	37
NOYJEQ	558	67	8
NUVJOE	559	68	87
OBATIW	560	73	81
OGELAQ	564	63	4
OWIGUX	569	63	5
PASDOD01	570	62	14
PEPWIS	572	61	180
PEXMIO	573	62	21
POBCEQ	577	67	70
POKCUO	578	61	28
POMYIC	579	75	195
PYAMFE	581	63	2
QAHGIQ	582	77	7
QAWMAE	583	63	10
QAXQAJ	584	71	9

Table 13. (cont’d)

QOSNIW	590	79	462
RANDAN	592	64	150
REVCIG	594	79	34
REXROE	595	67	199
RIRHAB	596	68	10
RISSUI	597	76	190
ROCDUK	598	76	44
RONPIT08	599	75	78
SANRIL	601	75	9
SAVZEX	603	63	4
SAVZIB	604	63	4
SAXFIH	605	75	18
SUFLOU	606	64	26
TADQUN	607	71	194
TAMZAJ	608	70	229
TILREL	612	70	25
TODSAJ	613	76	47
TUDVUK01	615	63	6
UHEFOC	619	72	53
UHEGOD	621	67	90
VAHGES	622	69	30
VASZIY	623	64	12
VESNEO03	624	71	18
VIHCUL	626	76	91
VONYEF01	627	67	13
VURXUB01	629	67	73
WAWGIK	630	69	25
WEPDED	631	76	17
WEZVIH	632	67	27
WEZVON	633	68	13
WIZNEB	634	61	119
WOGLUA	635	77	59
WOGMEL	636	77	35
WOQXEF	638	77	27
WUJCUA	639	77	2
WUPKEX	640	63	16
WUPKOH	641	65	15
WUPKUN	642	73	9

Table 13. (cont’d)

XETFUA	645	73	2
XIFKOP	647	73	2
XIFVAM	648	77	60
XIGVUG	649	71	9
XIGWAN	650	70	9
XIQFEJ	651	78	4
XISXIJ	652	62	2
XITDIO	653	70	6
XITLAQ	654	71	1
XUKDOY	656	65	7
XURWOX	657	69	59
YACNEW	659	79	60
YAZJOZ	662	75	2
YIKNUC01	666	73	2
YIMYEZ	667	71	10
YIMYUP	668	77	2
YUVBOH	670	65	3
ZASMAK	672	79	11
ZAVRUM	673	63	2
ZEKDOK01	674	71	51
ZERFEK	675	70	64
ZUYQOB02	677	71	42
BOZQUF	678	79	23
MUHMUA	679	61	1
DULCOF	680	79	7
GUTXEB	681	69	5
OSABOB	203	50	14
OWIHEI	206	55	3
IYUFOX	140	59	12
IZAFOG	141	55	11
TIYPIC	278	49	28
UDULOU	281	57	66
AYOFEZ	414	65	9
BACVOR	415	67	2
FOGLEV	467	63	36
FOYYAU02	468	61	3
KUWGUE	527	64	22
LABMUY	528	77	149

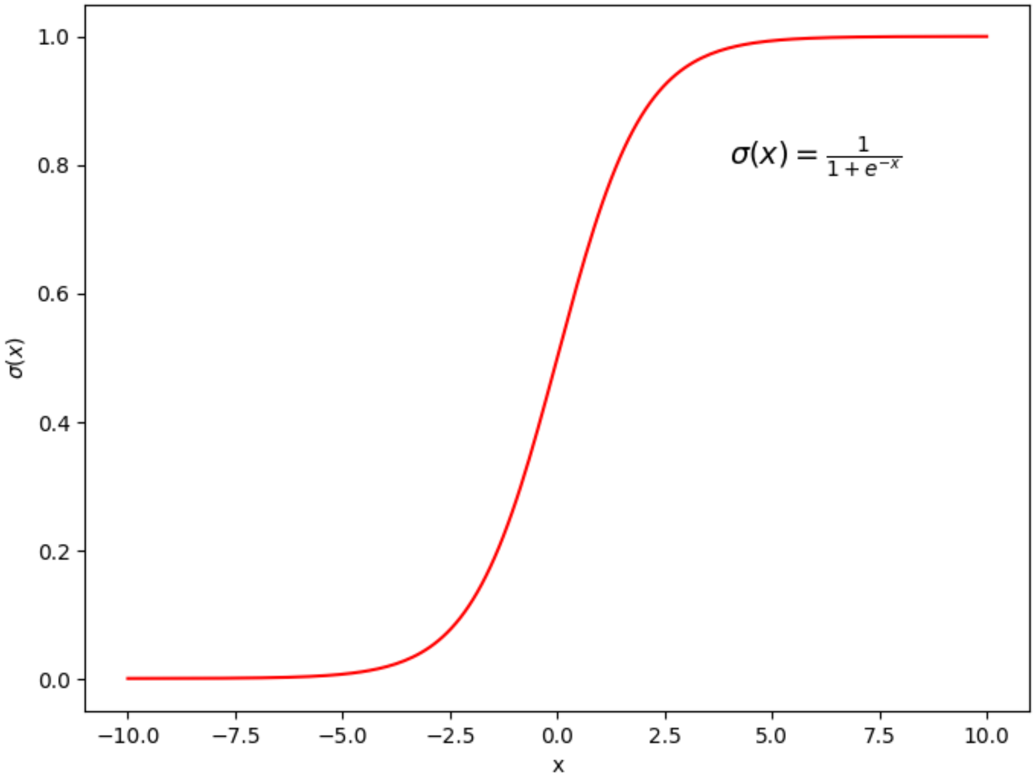


Figure 35. The Sigmoid function.

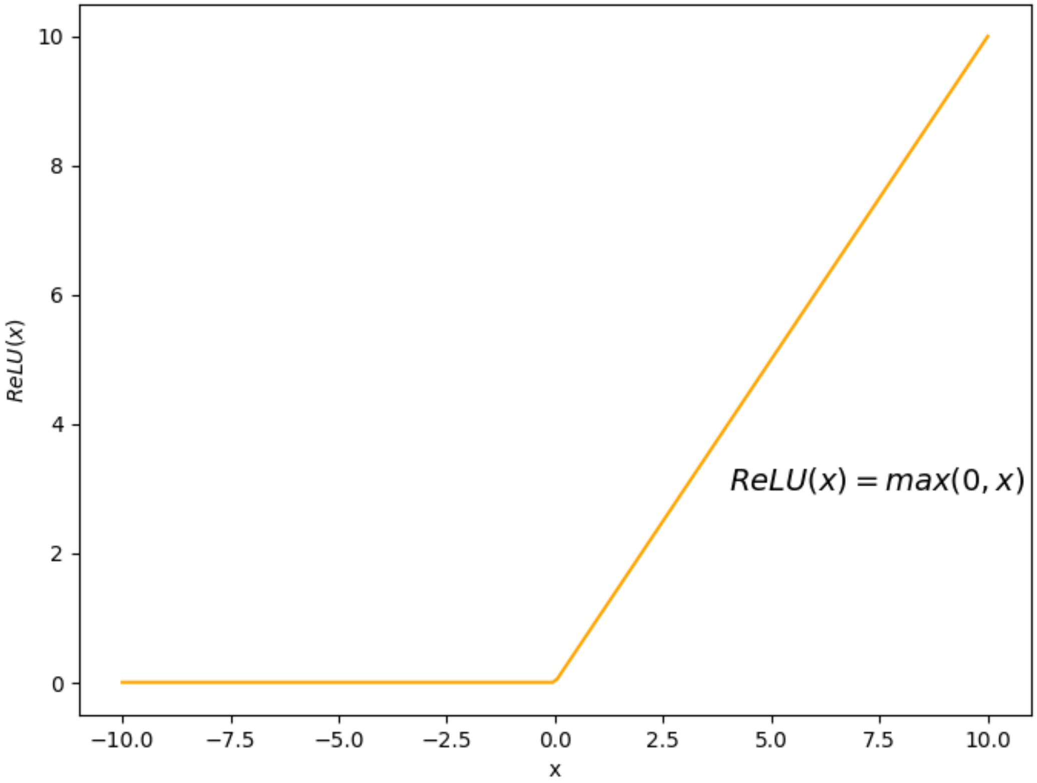


Figure 36. The ReLU function.

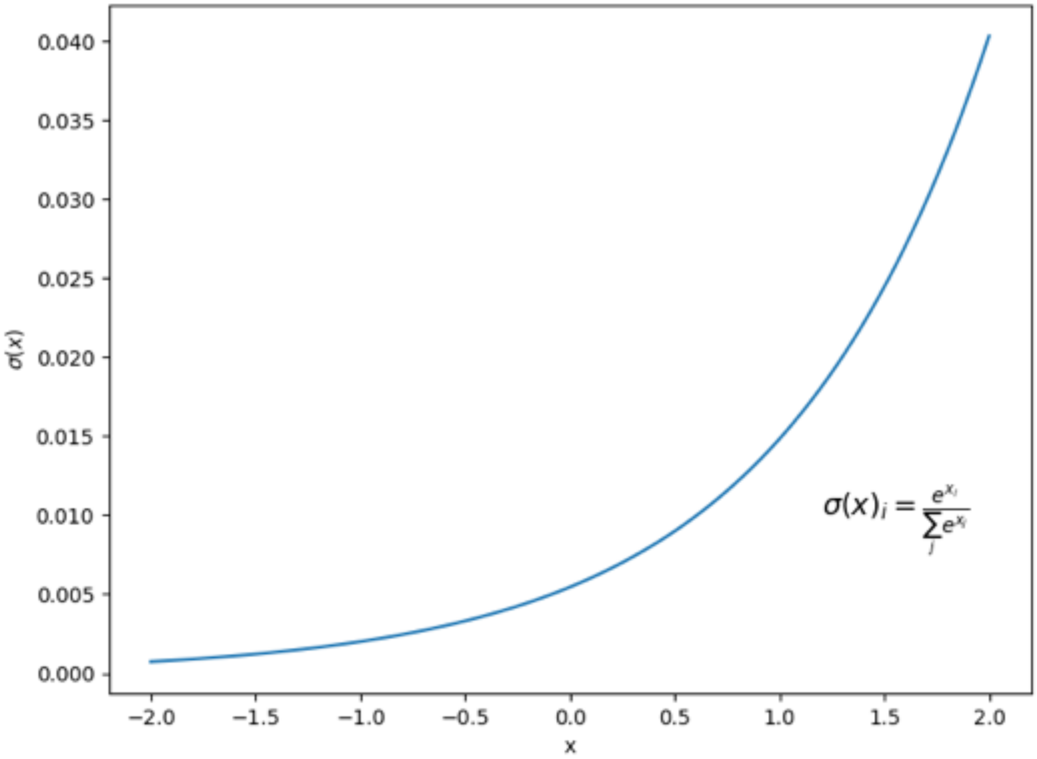


Figure 37. The softmax function.

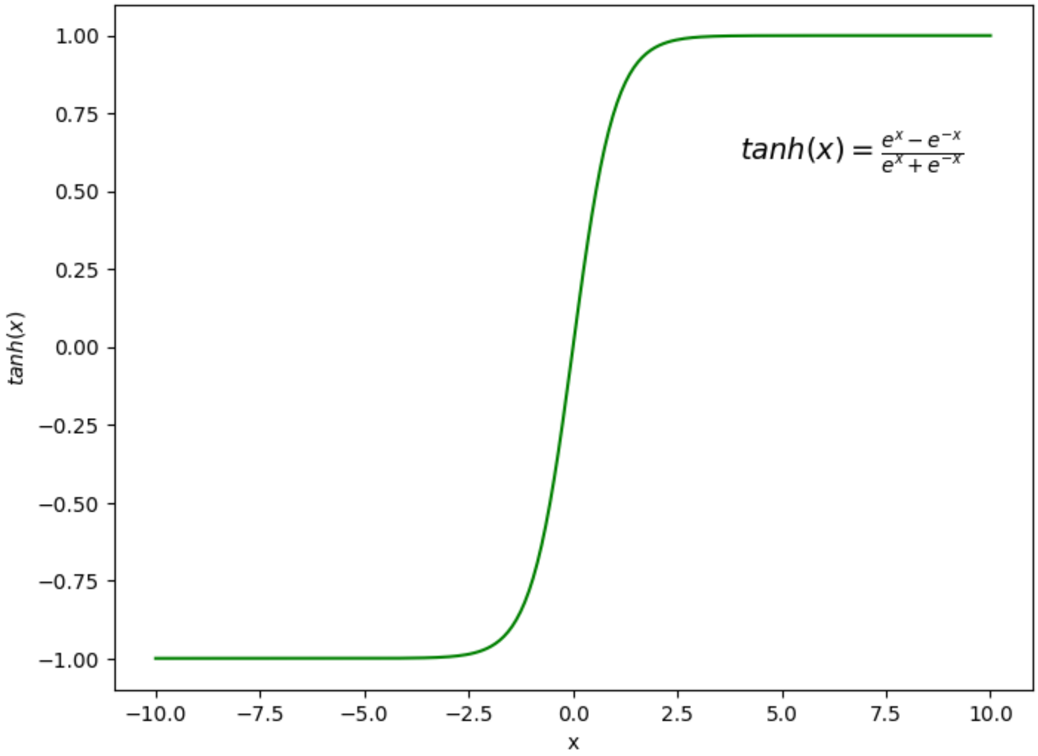


Figure 38. The Tanh function.

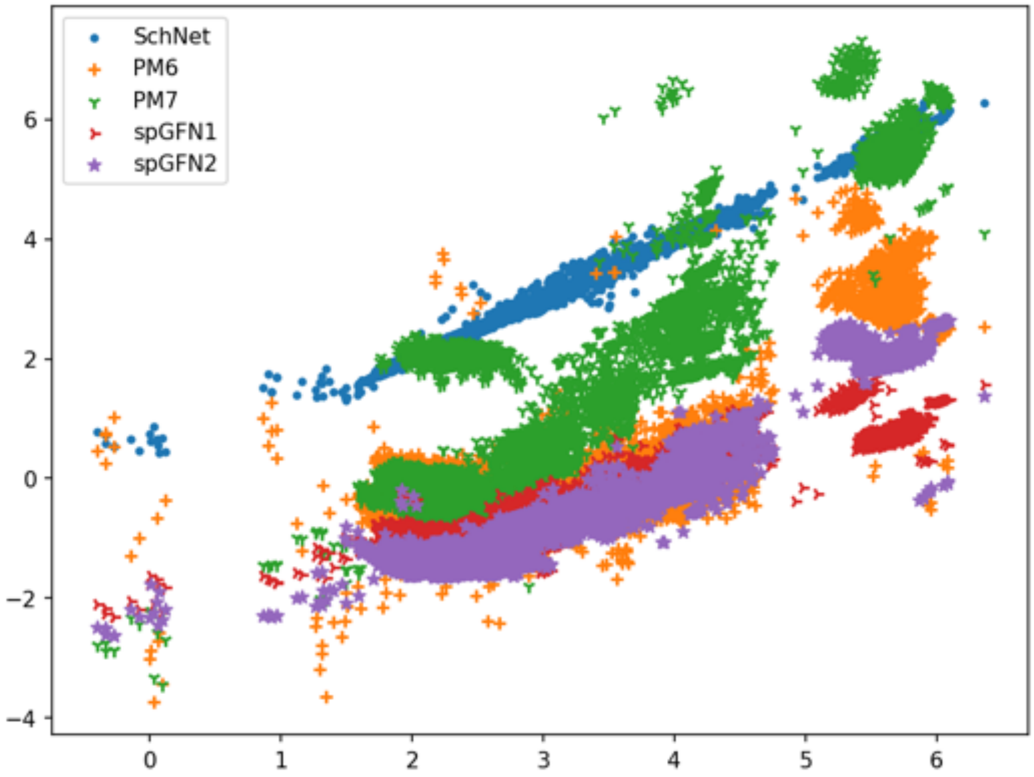


Figure 39. The overall trend of splitting energy predictions.