Physics Reports 934 (2021) 1–71

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Machine learning accelerates quantum mechanics predictions of molecular crystals

Yanqiang Han, Imran Ali, Zhilong Wang, Junfei Cai, Sicheng Wu, Jiequn Tang, Lin Zhang, Jiahao Ren, Rui Xiao, Qianqian Lu, Lei Hang, Hongyuan Luo, Jinjin Li *

National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao Tong University, Shanghai, 200240, China
Key Laboratory for Thin Film and Microfabrication of Ministry of Education, Department of Micro/Nano-electronics, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Article history:
Received 26 July 2021
Accepted 3 August 2021
Available online 25 August 2021
Editor: D.K. Campbell

Keywords:
Machine learning
Quantum mechanics prediction
Molecular crystal

ABSTRACT

Quantum mechanics (QM) approaches (DFT, MP2, CCSD(T), etc.) play an important role in calculating molecules and crystals with a high accuracy and acceptable efficiency. In recent years, with the development of artificial intelligence technology, machine learning (ML) has played an increasingly essential role in accelerating the QM calculations and predictions of molecular crystals, as well as the discovery of novel materials. This review provides state-of-the-art information and prospects for QM theories, fragment-based methods and ML methods, as well as their up-to-date applications in predicting small inorganic molecules, large drug molecules and relevant molecular crystals. The discussed applications include ML potential energy surface (PES) construction, crystal structure prediction (CSP), chemical reaction prediction and predictions of a series of properties, such as structure, energy, atomic force, bond length, chemical shift, superconductivity, super-hardness, vibrational spectra, phase transition and diagram. This work also reviews software and packages built recently based on ML methods for property predictions and PES constructions in the field of physics and chemistry. For the three discussed methods, the most time-consuming one is the high-level all-atom QM method, which is capable of describing electronic structures with high accuracy and thus predicts properties that are consistent with the experimental results. The second one, fragment-based QM method, requires less computational time than all-atom QM, which can accelerate all-atom QM calculations for large systems by dividing the entire system into subsystems, presenting a considerable efficiency increase. The computational complexities for fragment-based QM and all-atom QM are N - N^2 and N^5-N^7 (N is the size of the system), respectively. A well-trained ML model can make the above predictions within seconds while ensuring a high prediction accuracy, where its prediction cost and accuracy are determined by the training data and the training process. Therefore, it is challenging for ML applications in physics and chemistry to generate highly accurate and powerful ML models while ensuring sufficient datasets. This work not only provides an overview of the recent progress in QM theories, fragment-based methods, ML methods and several ML-based software programs and applications on small inorganic molecules, large drug molecules and relevant crystals, but also shed light on ML methods in accelerating QM prediction, optimization and novel crystal material design.

* Corresponding author at: Key Laboratory for Thin Film and Microfabrication of Ministry of Education, Department of Micro/Nano-electronics, Shanghai Jiao Tong University, Shanghai, 200240, China.
E-mail address: lijinjin@sjtu.edu.cn (J. Li).

https://doi.org/10.1016/j.physrep.2021.08.002
0370-1573/

Contents

1. Introduction

For decades, numerous new materials and novel phenomena have been predicted and determined under extreme pressure and temperature conditions. These harsh experimental conditions necessitate high-quality theoretical investigations, which could provide a principal explanation for physical and chemical phenomena. Predictions of molecules and molecular crystals under extreme conditions have been a significant area in both physics and chemistry. The large number of possible molecules and materials and the numerous ways for physical and chemical transformations make high-quality approaches required for understanding the fundamentals of physics and chemistry.

For decades, substantial effort has been made to develop theoretical methods that advance quantum mechanics (QM) methods, such as Hartree–Fock (HF), density functional theory (DFT), second-order Møller–Plesset perturbation (MP2) theory and Coupled Cluster Single-Double and perturbative Triple (CCSD(T)) theory. Despite the enormous computational costs, these all-atom QM methods are considered the most reliable and precise methods for predicting the crystal structures and properties under extreme conditions, such as energies, atomic forces, equations of state, vibrational spectra, chemical shifts and phase transitions. However, the most popular electronic structure method, DFT, suffers the disability of accurately describing long-range electronic correlation effects and London dispersion corrections. In recent years, this topic has received increasing attention, and tremendous efforts have been made to develop dispersion corrections for DFT methods. In addition, several post-Hartree–Fock (post-HF) *ab initio* QM methods (such as MP2 and CCSD(T)) have been proposed and used in computational chemistry and physics for molecules of different sizes, and they are capable of describing dispersion effects. Progress has provided the basis for wide applications of QM methods in both physics and chemistry.

However, as the accuracy improved, the computational cost scaling of these high-level QM methods and the relevant dispersion corrections have become much steeper compared with the underlying DFT models, which significantly limits their efficient applications, especially for large molecular systems. In addition, high-level QM methods are usually performed with large basis sets, which also dramatically increases the computational cost. In that case, to solve the low efficiency issues of all-atom QM methods, a series of fragment-based QM schemes has been proposed to accurately predict the structures, energies, and phase transitions of molecular crystals (such as carbon dioxide, ice, carbon monoxide, and solid ammonia) at extreme pressures. The fragment-based QM method has a similar accuracy but a much higher efficiency than the all-atom QM calculation. Fragment-based QM methods have been widely used to calculate energy [1], force and Hessians, normal mode phonon analysis [2,3], NMR chemical shift calculations [4–6], and symmetry prediction of space groups [2] for a series of molecular and crystalline systems, including carbon monoxide, carbon dioxide [7–15], solid ice [16–21], ammonia [22], and formic acid [1,23].

Although the computational complexity of the fragment-based QM methods is much lower than that of all-atom QM calculations, it increases rapidly with the size of the system, which makes these fragment-base methods less efficient and less applicable for large molecules and large crystal systems. In that case, developing fast, accurate and less expensive approach has become a lasting global challenge. In recent years, the rise of artificial intelligence (AI) algorithms has attracted significant attention [24–26] and has been widely applied in the fields of physics, chemistry, and material science, including translation, web search, medical diagnostics [27–30], brain–computer interfaces [31], social media analysis [32], board games [33–36], robotics [37,38], nano sciences [39], bioinformatics [40–44] and particle physics [45]. Particularly in chemistry and physics, machine learning (ML) methods improve the calculation speed by hundreds of thousands of times while maintaining the prediction accuracy of traditional *ab initio* calculations, which greatly shortens the distance between human cognition and nature. In general, ML methods use pattern recognition algorithms to train mathematical relationships between structures/coordinates and observed physical/chemical quantities and further extend these models to predict the physical, chemical and biological properties for novel compounds, which is distinctly different from physical models that rely on explicit physical equations, such as molecular dynamics and QM methods. In addition, compared with physical models, ML methods are more efficient and can be applied to large datasets without much of an extra increase in computational cost.

Fig. 1.1 shows an illustration of the QM, fragmental, and ML methods and the corresponding applications. The comparison of the computational costs and efficiencies of all-atom QM, fragment-based QM and ML-driven QM methods is shown in Fig. 1.1(a). All-atom QM methods correspond to traditional *ab initio* calculations, such as DFT, MP2, and CCSD(T). Fragment-based QM methods are capable of accelerating the *ab initio* calculations on molecular crystal systems with a high accuracy and relatively high efficiency. The last ML-driven methods can speed up *ab initio* calculations to the greatest extent with a high accuracy and high efficiency. For crystal systems with a large number of molecules, the ML-driven QM method is capable of calculating the energies, atomic forces and force constants within ten seconds, while it takes more than 10^7 seconds for all-atom QM calculations. As shown in Fig. 1.1(b), the all-atom QM, fragment-based QM and ML-driven QM methods all have the potential for PES construction applications in molecules and crystals [46–56] and the predictions of crystal structure [57–60], physical and chemical properties [61–78] (such as energy, atomic force, chemical shift, and phase transition), and chemical reactions [79–81]. The high-speed development of ML methods in physics and chemistry has made it possible to not only predict structures and properties but also to design compounds or materials that have particular properties for a given purpose.

Motivated by the progress in QM methods, fragment-based QM approaches, and ML-driven QM methods, and especially the substantial advance in ML approaches and the wide range of applications, we will review the recent progress of these widely used methods and their relevant applications in physics and chemistry. In Section 2, we will review the three types of computational methods and the relevant progress, including the all-atom QM methods (DFT, MP2, and CCSD(T)), fragment-based QM approaches (many-body expansion scheme and inclusion–exclusion scheme) and ML-driven QM methods, including ML algorithms, descriptors and model optimizations. In Section 3, we will discuss the recently reported applications of these QM and ML methods on small molecules and relevant crystals to predict structures and a series of properties (including energy, atomic forces, EOS, vibrational spectra, superconducting, super-hard, chemical shift, and phase transition) and the ML-based PESs with the QM level of accuracy. In Section 4, we will review the QM calculations and ML-driven methods that are suitably developed for large drug molecules and crystals in property prediction, crystal structure prediction (CSP) and PES/force field construction. In Section 5, a series of developed software and packages based on ML methods for physical and chemical applications will be reviewed.

2. Quantum mechanical (QM) methods

The considerably small energy difference between different phases and polymorphs of molecular crystals makes accurately predicting the structure, spectra and phase transition challenging. Motivated by these challenges, for decades tremendous efforts have been made to develop various computational methods for the investigation of molecular crystals, including molecular mechanics methods (Monte Carlo and molecular dynamics), wavefunction-based QM methods (HF and semi-empirical), electron density-based QM methods (DFT, MP2 and CCSD), ML methods, such as neural networks (NN), graph neural networks (GNN) and decision tree (DT), and modern ML-driven QM approaches/software, such as PES-Learn, NN-FQM, UADDCR, and ChemML, as shown in Fig. 2.1.

Since the 1980s, QM methods have made significant improvements in terms of accuracy, including DFT, MP2 and CCSD(T). These electronic structure methods have been widely used in computational chemistry, physics, and materials. In recent years, van der Waals dispersion has been introduced into DFT, and periodic conditions have been introduced into DFT and MP2, which make these *ab initio* calculation theories more accurate and applicable. However, for large systems these methods are computationally expensive and suffer significant efficiency restrictions. In this context, some fragment-based QM methods [4,82–86], such as binary interaction models, embedded methods, incremental methods and hybrid many-body methods, have been proposed to more accurately predict the structures and energies of large molecular systems. Fragment-based QM approaches have also been widely and successfully applied to calculating the energy, structure, phase transition and spectra of large molecular crystals with remarkable precision compared to all-atom QM calculations, but still have a restriction of low computational efficiency. In recent years, the crossing and integration of computational science, chemistry, physics, and material science has promoted the development and improvement

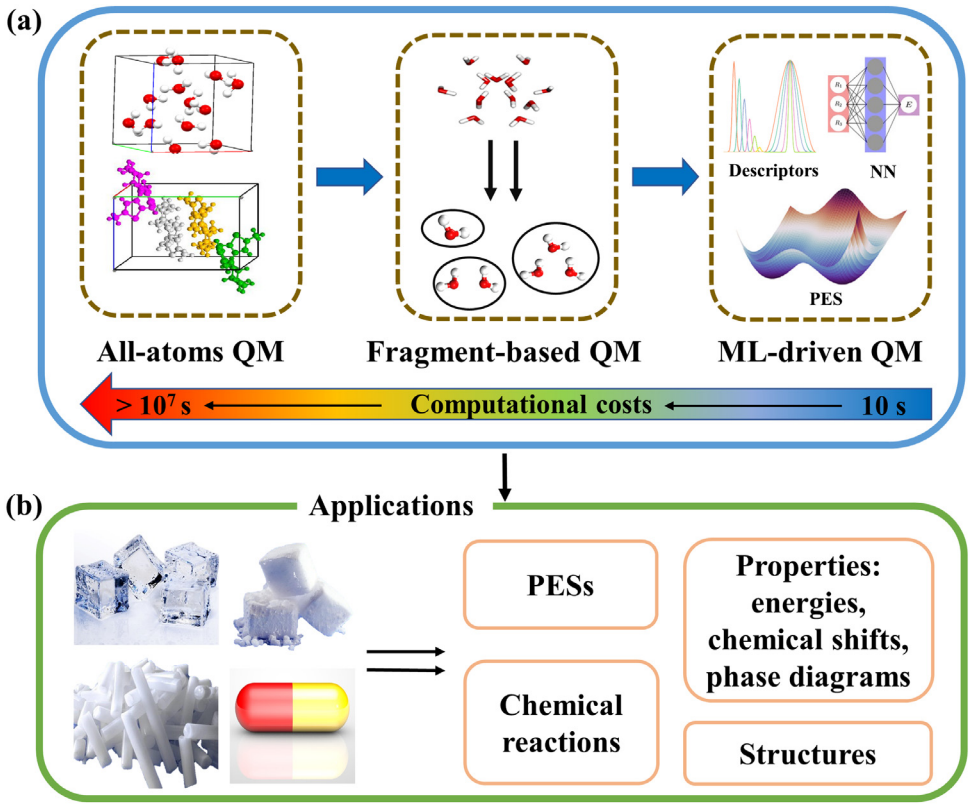

Fig. 1.1. Illustration of all-atom QM, fragment-based QM, and ML-driven QM methods and the corresponding applications. (a) Comparison of computational costs and efficiencies of all-atom QM, fragment-based QM and ML-driven QM methods. (b) Corresponding applications of the methods shown in (a), including PESs construction, properties prediction, chemical reaction prediction and crystal structure prediction.

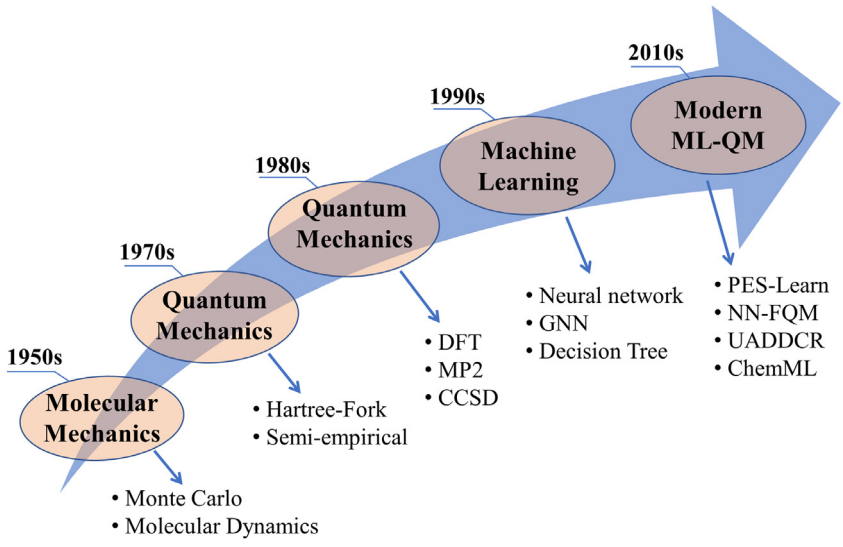

Fig. 2.1. Development of computation methods in molecular crystal, including molecular mechanics (MM), wavefunction-based quantum mechanics (QM), electron density-based QM, machine learning (ML) methods and modern ML-driven QM approaches.

of a series of computational approaches for accelerating QM calculations on molecular crystals. In particular, ML has made significant progress and been widely used in the field of physics, chemistry and materials, which has led to the development of ML-driven QM methods with high accuracy and efficiency.

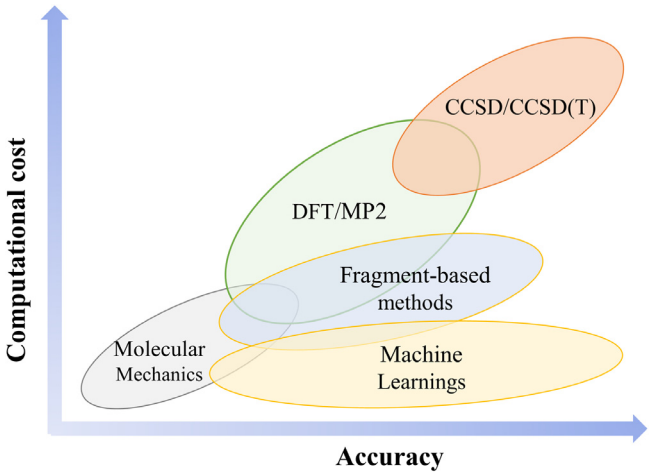

Fig. 2.2. Comparison of the accuracy and computational cost for molecular mechanics, DFT, MP2, CCSD(T) and ML methods.

The accuracy and computational cost of molecular mechanics, DFT, MP2, CCSD(T), fragment-based methods and ML methods are roughly shown in Fig. 2.2. Compared with the all-atom QM method, the acceleration effect of ML-driven QM methods is up to hundreds of thousands of times, which shows its potentially important application prospects. The ML-driven QM methods have been widely used in the field of computational physics and chemistry, including CSP, chemical reaction prediction, PES construction, and properties prediction, such as energy, atomic force, chemical shift and thermal coefficient.

Faster and more accurate calculations are the goals of theoretical and computational physics and chemistry. Solving this contradiction will allow structural simulations and designs to achieve leap-forward progress. Addressing this problem, this section will review the three representative QM methods, including the traditional *ab initio* theories (such as DFT, MP2 and CCSD(T)), the advanced fragment-based QM methods, the NN-fragment method, and several ML-driven QM methods proposed by different research teams. All-atom QM is the earliest *ab initio* calculation method with the highest accuracy. The fragment-based QM method was proposed to solve the low calculation speed of QM. While retaining the calculation precision of all-atom QM, the fragment-based QM method greatly shortens the calculation time via fragment partitioning. In addition, the recently developed ML-driven QM methods can greatly shorten the calculation time by hundreds of thousands of times while retaining the accuracy of traditional calculations.

2.1. All-atom QM theories (DFT, MP2, CCSD(T))

All-atom QM theories, such as HF, Kohn–Sham DFT, MP2, and CCSD(T), are widely used in chemistry, physics and material science, especially crystal systems. For decades, *ab initio* theories have progressed considerably in dispersion correlation, periodic conditions, basis sets, complete basis set (CBS) extrapolation and counterpoise correction for predicting molecular crystal energies and the optimization of crystal structures.

DFT

Electronic structure methods, such as HF and DFT, are the most popular methods in computational physics and chemistry. However, these methods with semi-local density functional approximations cannot deal with electronic correlation interactions and thus fail to consider London dispersion corrections [87,88], which are also known as van der Waals (vdW) dispersion corrections. The vdW dispersion interactions arise from correlated fluctuations between regions with different electron densities and contribute significantly to the packing of molecular crystals. In recent years, this topic has received increasing attention, and tremendous efforts have been made to develop dispersion corrections for the DFT methods. These improved dispersion corrections are shown in Table 2.1, which fall into three major categories: semi-classical treatments, nonlocal (NL) density-based schemes and effective one-electron-based dispersion corrections.

The semi-classical treatment performs the corrections by adding the dispersion energy between pairs of atoms to the electronic energy of the DFT methods. Although dispersion interactions between atoms or molecules are intrinsically quantum mechanical interactions, they are treated as classical interactions, which are known as vdW or London interactions. In 2004, Grimmer et al. [89] proposed the DFT-D1 method for accurately describing vdW complexes, including empirical corrections. Subsequently, Grimmer et al. [90] (2006) introduced the DFT-D2 method as an update of the DFT-D1 method, which has become an understandable starting point for developing other dispersion corrections. In 2010, Grimmer et al. [91–93] presented the DFT-D3 approach, in which the dispersion coefficients are explicitly evaluated by the chemical environment with the empirical concept of fractional coordination numbers (CNs). As an update of DFT-D1 and DFT-D2, the DFT-D3 model is superior and has presented an excellent performance in the prediction of

Table 2.1
List of some notable dispersion correction models, including correction categories, computational complexity compared to underlying DFT models, higher many-body interactions, authors and first reported years.

Models	Categories	Computational complexity	Many-body	Author
vdW-DF	Nonlocal density-based	High	Yes	Dion et al. [107–111] (2004)
Minnesota	Effective one-electron based	Medium	–	Zhao et al. [112–114] (2005)
DFT-D2	Semi-classical	Low	No	Grimmer et al. [90] (2006)
XDM	Semi-classical	Medium	Yes	Becke et al. [102,103] (2006)
TS	Semi-classical	Low	No	Tkatchenko and Scheffler [93] (2009)
VV10	Nonlocal density-based	Medium	Yes	Vydrov and Van Voorhis [115–118] (2009)
DFT-D3	Semi-classical	Low	Yes	Grimmer et al. [91,92] (2010)
MBD	Semi-classical	Medium	No	DiStasio et al. [106] (2012)
DCP	Effective one-electron based	Medium	–	van Santen et al. [119] (2015)

small molecules [94–97], large molecular benchmarks [98,99] and molecular crystals [100,101]. In 2009, Tkatchenko and Scheffler [93] proposed a density-dependent scheme of dispersion correction (TS) between atoms and molecules, in which the dispersion energy is restricted to the two-body interaction and the chemical environment is evaluated by the electron density. Other state-of-the-art dispersion corrections with semi-classical treatment include the exchange-dipole moment model (XDM) [102,103], the local-response dispersion model (LRD) [104,105], and the many-body dispersion model (MBD) [106], all of which have demonstrated a good performance in various benchmarks.

The nonlocal density-based dispersion corrections obtain the dispersion energy and the corresponding integral kernel with only the electron density. In 2004, Dion et al. developed the first nonlocal density-based dispersion correction model, vdW density functionals (vdW-DF1) [107–109], in which the correlation energy is obtained by summing the local part evaluated with the LDA correlation and the nonlocal part. Despite the various successful applications, vdW-DF1 is demonstrated to underestimate hydrogen bond interactions and to overestimate separations between molecules. Substantially, several modern dispersion corrections have been introduced, including vdW-DF2 by Lee et al. [111] and vdW-DF-09 [110], VV9 [116–118], and VV10 by Vydrov and Van Voorhis (VV) [115]. These nonlocal functionals have presented a good performance in a variety of benchmarks, including ices, benzene, aspirin, and hexamine. Compared with semi-classical functionals, nonlocal density-based dispersion corrections include self-consistent vdW interactions in electron density but are typically more demanding in terms of computational cost.

The London dispersion interaction is a two-particle interaction arising from correlated electron movements. In effective one-electron dispersion corrections, the dispersion interactions are empirically described, while all nonlocal density information and dynamical properties are ignored. In recent years, a series of one-electron potentials have been introduced, including dispersion-corrected atom-centered potentials (DCACPs) [120], DCACPs for a particular DFA or basis set combination (DCPs) [119], Minnesota functionals [112–114] (M05, M06, MN12, etc.) and the related variants. These functionals have presented an excellent performance for kinetics and thermochemistry on large datasets, but with much higher computational cost. As seen above, these dispersion corrections of DFT functionals are continuously developed and improved, and there is no one model that is significantly better than the others.

MP2

Decades ago, MP2 was developed for the calculation of periodic systems, such as polymers and other one-dimensional systems. In recent years, periodic MP2 in three dimensions has made great progress and has become an increasingly essential approach for predicting molecular crystals. By adopting Laplace transform techniques, Scuseria's group [121–123] (2001–2010) introduced the atomic orbital-based algorithm of periodic MP2, which has been applied to polymers. In 2007, Maschio et al. [124,125] presented a density-fitted local MP2 algorithm using a direct space method and the Pulay/Werner-type local correction implemented by projected atomic orbital domains, which is the first periodic MP2 algorithm applied to molecular crystals. Subsequently, many other periodic MP2 algorithms have been proposed, including the plane wave periodic MP2 algorithm with the projector-augmented wave approach [126,127], the localized resolution of identity approach for periodic MP2 [128], the orbital specific virtual (OSV)-based MP2 algorithm [129] (which eliminates the need for careful domain selection and discontinuities in the potential energy surface and reduces the computational costs), and the extremely parallel Gaussian and pale wave MP2 (GPW-MP2) [130,131] approach, with an 80% parallel efficiency. Notably, these periodic MP2 approaches can also be implemented in double-hybrid density functionals and in methods with random phase approximation (RPA).

CCSD(T)

As the demand for computational accuracy increases, much effort has been made to develop more efficient QM methods compared with DFT. Several post-HF *ab initio* QM methods have been proposed and used in computational chemistry and physics for molecules of different sizes. Among these methods, CCSD(T) is the most important, in which the single and double contributions are fully treated, but the connected triple contribution is treated noniteratively using many-body perturbation theory. CCSD(T) is an upgraded version of CCSD that includes only single and double excitations. CCSD(T) has shown excellent performance for a wide range of applications and is known as the gold standard in quantum chemistry [132]. However, the computational cost scaling of CCSD(T) is extremely steep at $O(N^7)$, where N denotes the molecular size, which makes it a challenge for CCSD(T) to be efficiently applied to large molecules.

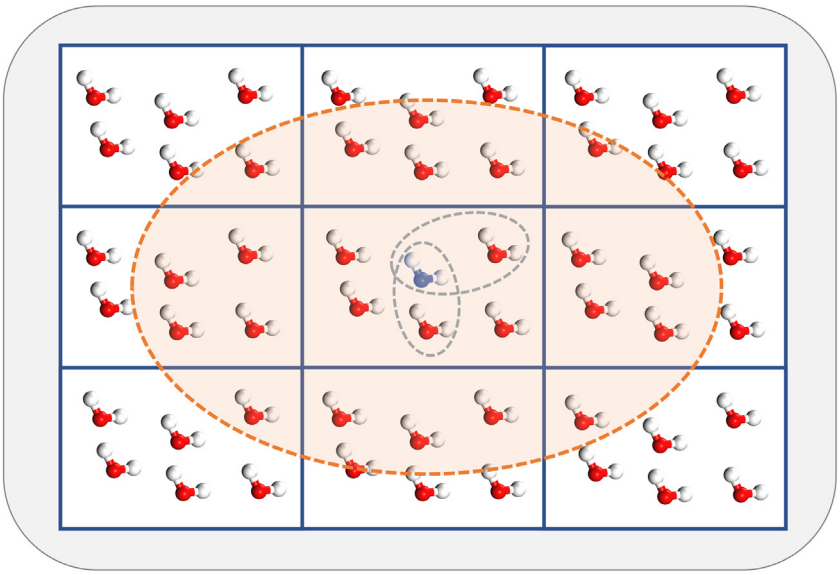

Fig. 2.3. Illustration of fragment-based QM methods for molecular crystals, with the approximation that the chemical environment of a local region in a large system is mainly influenced by the nearby regions (orange shade), where the energies are calculated by QM methods.

The improvements in modern computer hardware and the parallel program have made periodic MP2/CCSD(T) approaches much more applicable for crystal systems containing tens of atoms. However, despite the improvements in terms of accuracy, the periodic MP2 and CCSD(T) methods increase the computational effort faster as the system grows compared with DFT models. The computational cost scaling of MP2, CCSD, and CCSD(T) is $O(N^5)$, $O(N^6)$ and $O(N^7)$, respectively, while it is $O(N^3)$ for the underlying DFT models. Moreover, to meet the accuracy requirements, these QM models are usually performed with large basis sets, which substantially increases the computational cost and makes them more challenging for applications in large molecular crystal systems. In practice, well-designed algorithms and careful numerical treatment can be helpful to address this challenge. The introduction of fragment-based methods and deep neural networks (DNNs) also makes these high-level QM methods suitable for large molecular crystals, which will be discussed in Sections 2.2 and 2.3.

2.2. Fragment-based QM methods

Ab initio theories (DFT, MP2, CCSD(T)) can accurately describe the electronic structure of molecules and are thus considerably precise for molecular crystal calculations and understanding the fundamentals at the atomic level. However, the extreme computational cost limits their applications for large-scale crystal systems. In this case, some efforts have been made to develop linear-scaling methods to address the computational limitations. Among these methods, fragment-based QM methods have been popular for decades as a powerful tool for QM calculations of large molecular systems due to their ability to effectively reduce the computational cost scale. In the fragment scheme, a large system is divided into a series of smaller subsystems for QM calculations, and the total energy and properties of the entire system can be calculated by the summation of these small subsystems. Many fragmental schemes have been proposed for molecular crystals with different formations of fragments and interactions, including many-body expansion (MBE) schemes [82,83], inclusion–exclusion principle (IEP) methods [133–136] and hybrid many-body interaction (HMBI) methods [84].

Fragment-based QM methods assume that the chemical environment of a local region in a large system is mainly influenced by nearby regions and weakly influenced by faraway atoms, as shown in Fig. 2.3. Thus, a macromolecule or a large molecular system is divided into a series of subsystems, and the properties of the entire system are evaluated by combining the corresponding properties of the small individual fragments. For molecular crystal systems, each individual molecule is treated as a unique, non-overlapping fragment in most typical fragment-based methods. Therefore, a MBE scheme can be used to handle the interactions among the fragments, where the total energy of a molecular crystal system is expressed in terms of one-body, two-body, three-body, and higher-order terms. For periodic systems, the contributions of molecules in periodic cells interacting with molecules in the central cell are involved in two-body and higher terms. The pairwise interactions usually contribute most of the lattice energy [137] (80%–90%). However, in practice, MBE schemes cannot always present rapid convergence, especially for crystal systems with strong hydrogen bonding interactions. The high-order terms in the many-body expansion will lead to some challenges. For example, the fragments are usually calculated with MP2 and even CCSD(T), which have very steep computational cost scaling. Therefore, the computational cost for fragments of the higher-order terms will increase rapidly. In addition, the number of fragments for higher-order

terms will also increase rapidly with the number of orders (n). Corresponding to the rapid growth of the computational cost, the contribution of high-order terms will rapidly decrease as the number of orders increases. Moreover, in MBE schemes, basis set superposition errors (BSSEs) have become a great challenge for higher-order terms [138,139].

To make MBE schemes more practical, various fragment-based methods have been proposed based on two main types of options. Based on the IEP scheme, the first treatment is to approximately evaluate the contribution of high-order terms with lower-level theories, such as HF. Generally, the contribution of high-order terms is implicitly evaluated by performing the calculation for the full system and subtracting out the contribution of low-order terms (e.g., one-body and two-body), which will be calculated using high-level methods (e.g., MP2, CCSD(T)). Another treatment (MBE) is to truncate the many-body expansion (e.g., two-body), in which the high-order terms are neglected. In this treatment, the contribution of these neglected high-order terms can be evaluated by embedding the low-order terms in the electronic field, which is generated using efficient methods.

MBE scheme

In 2005, Hirata et al. [82] introduced the binary interaction method based on MBE scheme, which is the first fragment-based QM method for molecular systems. Subsequently, the binary interaction method was improved and applied to periodic crystalline systems [1,83]. In the binary interaction method, the one-body and two-body terms in the many-body expansion are calculated with high-level methods, while all of the higher-order terms are neglected. Then, these one-body and two-body terms are embedded in the electrostatic field of the remaining systems represented by atomic charges computed by low-level methods, which approximately evaluate the contribution of the many-body polarization and long-range pairwise polarization and electrostatics. The binary interaction method has been extended to a series of methods suitable for calculating the energy, force and Hessians [1], normal mode phonon analysis [2], NMR chemical shift calculations [4–6], and symmetry prediction of space groups [2]. The variants of binary interaction methods have been widely used to calculate the energy, vibrational spectra, equation of state (EOS) and phase transition for a series of crystalline systems, including carbon monoxide, carbon dioxide [7,9,10,12,13,15], solid ice [18,19,140], ammonia [22], and formic acid [1]. Soon after, the binary interaction model was also extended to the ternary interaction variant, which explicitly includes three-body terms [83]. However, the ternary interaction variant suffers substantial increases in computational cost, which limits its application to a wide range of crystalline systems.

IEP scheme

Different from the binary interaction methods, the IEP scheme includes all many-body terms by approximately evaluating the contributions of high-order terms with a low-level QM approach [133–135]. These contributions of high-order terms are implicitly summed by calculating the energy of the entire system and then using the energy of the high-level theoretical calculation of the low-order term in the expansion to replace the value of the low-level calculation. The IEP scheme has been widely applied to calculate the lattice energy of many crystals. In addition, this scheme can be easily proposed for other calculations, such as chemical shifts, phonons and crystal structures, with minimal modifications of the existing code. The choice of electronic structure theories for low-level calculations and high-level calculations will significantly influence the computational accuracy and cost. In general, the HF is used for the low-level method calculations. In recent years, some DFT functionals have been used as a replacement for HF to improve the performance of IEP methods, showing a good performance in calculating the lattice energy for benzene [141], urea [136], and *para*-diiodobenzene [142]. As discussed in Section 2.1, the underlying DFT cannot evaluate the many-body dispersion correctly, which can be addressed using dispersion corrections, such as the D3 correction. Therefore, DFT functionals with dispersion corrections can be a better choice to improve the accuracy. With progress in periodic and parallel algorithms, MP2 can also be used as a low-level calculation, in which the many-body polarization and exchange effects can be reasonably incorporated.

HMBI model

Apart from electronic structure methods (HF, DFT and MP2), the force field can also be a choice to calculate the contribution of the many-body terms [137]. In 2009, Beran et al. [84] proposed the HMBI model, which treats the one-body and short-range two-body terms with QM calculations, but evaluates the many-body and long-range two-body terms with a polarizable force field instead of a low-level electronic structure method. The inclusion of polarizability, exchange and many-body dispersion terms in the force fields will significantly increase the accuracy of this model. The HMBI method has been widely used for calculating the lattice energy, Hessians, structure optimizations and property predictions in solid ice [16], carbon dioxide [11], aspirin and other molecular crystals.

Using an embedded approach, many other related fragment-based QM methods have been proposed in recent years. In 2012, Bygrave et al. [85] introduced an embedded MBE scheme in which exchange–repulsion contributions are included in the embedding potential. The embedded method was successfully used for predicting the energy and structure on several crystalline systems, including solid ice, carbon dioxide, hydrogen fluoride and clathrate hydrates, with the results consistent with experiments but at a low computational cost. In 2014, Hertman et al. [4] also implemented an embedded model based on HMBI to calculate NMR chemical shifts. With electrostatic embedding, Fang et al. [86] (2015) proposed a generalized energy-based fragmentation (GEBF) scheme for molecular crystals.

Table 2.2 shows several notable fragment-based QM methods based on MBE, IEP and HMBI schemes, most of which incorporate an electrostatic embedded field. These embedded schemes have demonstrated successful applications on molecular crystals in their own field, including structure optimization and prediction, calculation of energy, EOS, vibrational spectroscopy, and phase transition. These embedded fragment methods are being continuously developed and improved for molecular crystals, and there is no one fragment approach that is significantly superior to the others.

Table 2.2
Notable fragment-based QM methods including the many-body truncation, category, electrostatic embedding and year of first report.

Method	Truncation	Category	Embedding	Year
Hirata et al. [1,82]	Two-body	MBE	No/yes	2005
Dahlke and Truhlar [143]	Three-body	IEP	Yes	2007
Bludsky et al. [141]	Two-body	IEP	No	2008
Kamiya et al. [83]	Three-body	MBE	Yes	2008
Bygrave et al. [85]	Two-body	MBE	Yes	2012
Beran et al. [84]	Two-body	HMBI	Yes	2009
Hertman et al. [4]	Two-body	HMBI	Yes	2014
Fang et al. [86]	–	GEBF	Yes	2014

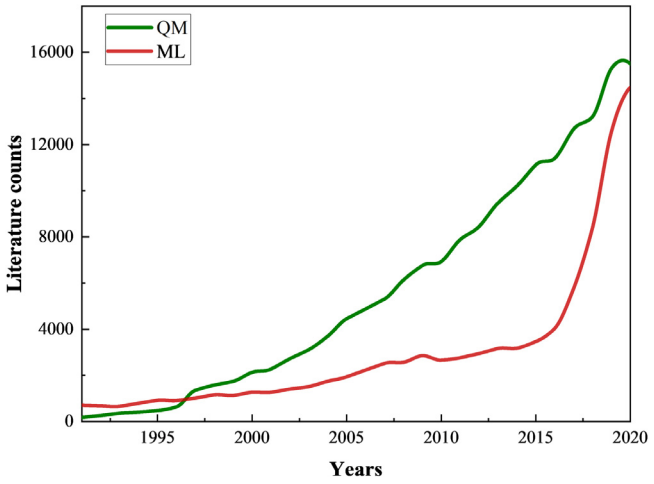

Fig. 2.4. Literature counts from 1991 to 2020 in chemistry and physics. The information is extracted from the Web of Science in April, 2021, using "DFT, MP2, CCSD" and "machine learning, neural networks, artificial intelligence" as the key words, respectively.

2.3. Machine learning (ML)-based QM methods

Although the improvements in modern supercomputers have made electronic structure methods much less expensive in terms of computational cost and applicable in many fields, these QM methods still face a challenge since the system size and complexity have also increased rapidly. To catch up with the growing demands, abandoning ideal model systems is significantly required to find a larger series of structures, compositions and materials, to track the evolution of the system with a longer time scale and to resolve even subtle details of atomic interactions with a high degree of accuracy, such as reaction barriers.

In recent years, ML approaches have made substantial progress and have been widely applied to physics, chemistry and biology with considerable precision and efficiency, including potential energy surface calculations, force field construction and rapid predictions of NMR chemical shifts, mass spectra and three-dimensional (3D) structures and pathways of chemical reactivity and catalysis. As shown in Fig. 2.4, since 2016 the number of academic articles related to ML with the keywords "ML, neural network (NN), and artificial intelligence (AI)" in the field of chemistry and physics has grown exponentially. Compared with QM, the keywords "DFT, MP2, CCSD" has linearly increased. The popularity of ML has demonstrated that ML is being used as a powerful tool in chemistry, physics, and material science for predicting the energy, structure and property. Particularly, based on QM calculations the neural network-based potential energy surface has a great potential in predicting molecular crystals and substituting electronic structures.

ML is omnipresent in everyday life and has a long and storied history that originated from the exploration of artificial intelligence [61]. Since the 1950s, acquiring knowledge by machine has drawn much attention, and a series of symbolic methods have been proposed for this purpose [144]. Subsequently, connection principles, such as NN and perceptron, were widely used to propose new ML methods for information storage and organization in the brain [145]. Based on statistical learning theory (SLT), various ML methods have been proposed, such as DTs and support vector machines (SVMs). In recent years, some modern ML methods have received much attention in academia and industry, such as GNNs, Gaussian process regression (GPR), convolutional neural networks (CNNs) and deep learning (DL) for big data investigations [26]. Generally, ML focuses on researching approaches to make computers learn automatically for knowledge acquisition and to improve the performance continuously without an explicit program design. In 1980, as the first ML seminar hosted at Carnegie Mellon University of the United States, ML became a discipline in its own right and began to take shape

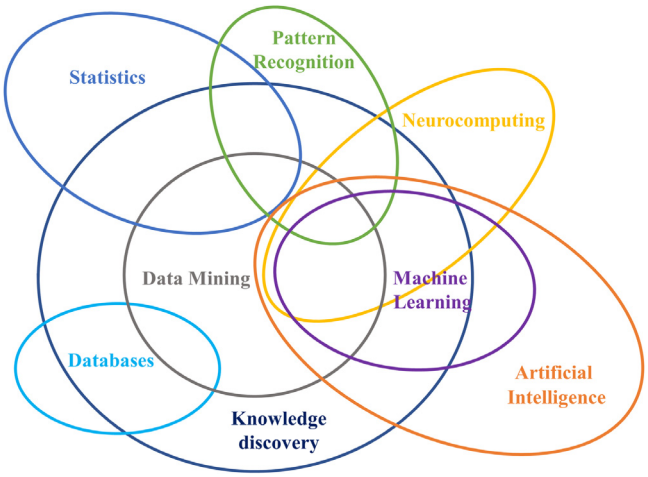

Fig. 2.5. Overall relationships of machine learning (ML) and current hot topics in computer science, including databases, knowledge discovery, data mining, artificial intelligence, statistics, pattern recognition, and neurocomputing.

rapidly. Fig. 2.5 shows the relationship of ML and current prominent research areas in computer science. As a branch of AI, ML has become an essential approach for machine knowledge acquisition of machines. ML is also an interdisciplinary discipline with close ties to current popular methods in computer science. In addition, ML can also be applied to pattern recognition and data mining. ML learns knowledge, finds insights from data and outputs reliable and repeatable results by learning from previous consistent data. ML has presented an excellent performance in terms of regression, classification and other topics, and has been demonstrated to be an essential tool in various fields, including computational physics, chemistry, bioinformatics, material science, speech recognition, image recognition [67], natural language processing (NLP) and information security.

ML workflow

Typically, ML is defined as $\langle P, T, E\rangle$, where P denotes performance, T is task and E is experience. The definition is that for a program, if its performance presented by P on task T improves with experience E, then the program is considered capable of learning for task T [146] from experience E. In computational physics and chemistry, task T mainly refers to the prediction of energy, related derivatives, property and structure. For computational physics and chemistry, the widely used ML approaches have evolved from the previous algorithms to now include nonparametric statistical learning, such as DNN, CNN, GNN GPR, DL and kernel ridge regression (KRR). Fig. 2.6 shows the workflow of the application of ML approaches to chemical research, which presents the three elements of a successful process: size and quality of the dataset, feature representations, and ML methods. First, for a model training process, a large dataset with an appropriate form is extremely important for the ML method to be effective. These datasets can be obtained from experimental results, calculations and some existing databases. In practice, abundant training datasets or small datasets with well-formulated representations [63] with ML models can lead to faithful results in chemical predictions, such as molecular representations [65], enzyme classification [66] and predictions of electronic structure correlation energies [63]. In contrast, a training process with a dataset from inexpensive calculations or experimental results that are error prone and noisy will strongly depend on the choice of ML models to achieve a convincing performance. Second, data cleaning and feature engineering (including feature extraction and representation) are proposed to convert the original data to representations or descriptors suitable for feeding to ML models. Then, an appropriate ML model will be chosen to train the relationship between the conditional attributes and the decision attribute by tuning the optimal hyperparameter. Finally, the proposed relationship model can be applied to predict the properties of interest.

Datasets

Since the training and testing sets are usually generated from same dataset, some of the issues existing in the dataset cannot be noticed when selecting the ML models and performing training process. Despite small errors during the cross-validation process, the ML model may still perform poorly in real chemical/physical problem. Besides, the case where the training and testing sets are from different datasets is called covariate shift, which can adopt new approaches for validation [147] such as importance weighted cross validation (IWCV) [148].

A robust model can usually be constructed from comprehensive datasets, appropriate feature representation and data-efficient ML methods. Therefore, a careful data collection process is essential for ML model construction, which may incorporate with an initio preprocessing for identifying and deal with the missing or spurious elements [149]. The Inorganic Crystal Database (ISCD) contains more than 190000 corrected data, which still contains measurement and human errors. To build robust ML model and avoid being misled, identifying and handling these errors are of great

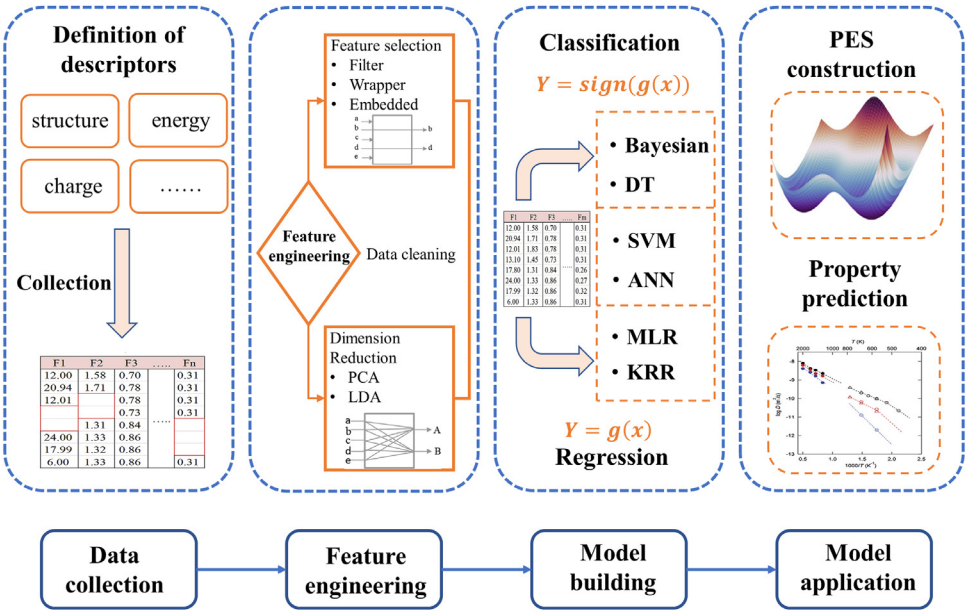

Fig. 2.6. Workflow of the application of ML-based QM method in computational chemistry and physics, including data collection, feature engineering, model building and model application. The model can be constructed by various ML methods, such as Bayesian theory, decision tree (DT), supporting vector machines (SVM), artificial neural network (ANN), multiple linear regression (MLR) and kernel ridge regression (KRR).

important [150]. The error propagation and low reproducibility of experimental datasets from peer-reviewed scientific researchers have been a public concern in ML model construction.

The datasets are of great importance for ML applications. Typically, the form of the database determines the choice of machine learning model. Supervised learning requires a large amount of training data containing inputs and corresponding outputs to build a function that can predict the output with a given input. The dataset consists of only inputs can be applied to unsupervised learning for pattern identifying and clustering. Besides, semi-supervised learning is suitable for the dataset that containing a large amount of input but only a small amount of output values. Among the three types of ML models, supervised learning is most powerful and is commonly used in physical and chemical sciences. On the other hand, the unsupervised learning is usually applied to large datasets for classification and analysis of data [151].

Feature representations

In practice, the ML model requires the dataset in a particular numerical representation that allow the algorithm to extract meaningful information from these data [152–156]. The process of encoding original data into the numerical format that suitable for a ML model is known as feature engineering. Generally, a well-defined representation of training data can improve the performance of ML models. However, fundamental understanding of the focused physical/chemical problem and the ML algorithms are strongly required to constructing best representation of data, which makes feature representation remaining a challenge and hot topic for ML models in chemical and physical problems.

Currently, a series of representation methods has been proposed. For example, on-hot encoding is the most naïve approach that treats molecules as distinct categorical variables and represents such variables with Boolean vectors. ML models with on-hot encoding are only capable for predicting of structures that have been encoded. Another example is the Coulomb matrix [156], which presents the information in atomic nuclear repulsion and potential energy and is invariant to molecular translations and rotations. Other representation approaches include SMILES, attributed graph, voxel, fingerprint, expert descriptors, image, and spatial coordinates [157]. For molecular systems, the major challenge is to reconcile all invariances into a descriptor without sacrificing its uniqueness and computability. For example, different geometries may be converted to same representation in some representation approaches. In that case, much effort has been made to address this challenge, such as fingerprints approaches [158–162], density representations [163], parameter sharing [164–166] and invariant integration [167]. Recent years, some ML methods have been capable to operate directly on 2D molecular graphs and even 3D molecular conformers, such as message-passing NNs [68] and atom-centered convolutional networks [168]. Fig. 2.7(a) shows several approaches for representing molecular structures, including one-hot, fingerprint, Coulomb matrix, expert descriptor, attributed graph, image and voxel, which can be fed into different ML methods. Noteworthy, since the current representation approaches are with compromise, careful selecting of descriptors is essential for every ML model.

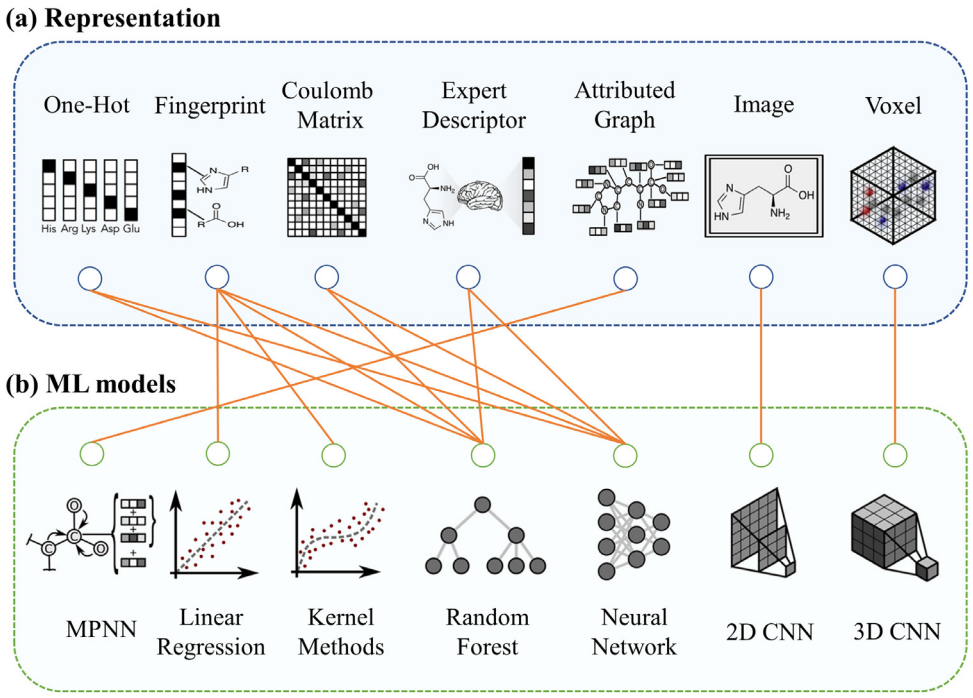

Fig. 2.7. Illustration of feature representations and ML methods. (a) Approaches for representing molecular structures, including one-hot, fingerprint, Coulomb matrix, expert descriptor, attributed graph, image and voxel. (b) A series of ML methods that can be represented in different ways, including linear regression, kernel methods, random forest, neural network, message passing NN (MPNN), 2D convolutional neural network (CNN) and 3D CNN.

ML methods

After collection and representation, the data are fed into ML models for classification or regression tasks. According to the task and the type of data, a series of ML algorithms can be applied to model training. For example, based on Bayes' theory, Naïve Bayes classifiers [169] are a series of classification algorithms for identifying the most probable hypothesis. K-nearest-neighbor (KNN) methods [170] can be applied to both classification and regression tasks, which relies on the k nearest neighbors in the data. Based on the operation of the brain, NNs and relevant variants are the most common ML methods in physical/chemical problems, including artificial NNs (ANNs) [171], DNNs [25], message passing NNs (MPNNs) [172], CNN [173], and 3D CNN [174]. In addition, kernel methods are the best-known ML methods, including SVM and KRR [175], which use a kernel function to transform input data into a higher-dimensional representation that makes the model solving easier. Besides, other commonly used ML methods in physics and chemistry include DT, random forest (RF) [176], and atomistic network [177]. Fig. 2.7(b) presents seven ML models (linear regression, kernel methods, random forest (RF), NNs, MPNN, 2D CNN and k3D CNN) that are capable with different representations.

Generally, using a range of different algorithms or similar algorithms with different parameters can be helpful to construct a robust ML model. The hyperparameters require to be estimated beforehand, since even small changes in hyperparameters can significantly influence the performance of ML model. For a ML training process, the model errors mainly remain in model bias (incorrect assumptions in the ML method), model variance and irreducible errors. High bias usually results from the inflexible model or insufficient data that cannot adequately describe the relationship between inputs and outputs. On the other hand, a large number of parameters or a complex model may result in high variance. To evaluate the accuracy, the ML models usually require to be applied to unseen data. The cross-validation is a commonly used procedure for accuracy testing, which reliable when the training and validation sets are representative for the whole population. Carefully selecting methods to evaluate the applicability and transferability of ML models is strongly required in most tasks [149].

Categories of ML applications

Although ML methods were proposed primarily for tasks in computer science, such as image recognition, a series of novel feature representations and descriptors has been developed for applications in chemical science. There are three main categories for ML being applied to physics and chemistry: forward models that predict chemical properties directly; potential energy surface (PES) models; and hybrid models that construct exchange and correlation functionals of electron density or wavefunctions. Fig. 2.8 shows the schematic diagram of the forward model, PES model, and hybrid model. In the forward model (Fig. 2.8(a)), the inputs and outputs are the representation of the property of interest, making this

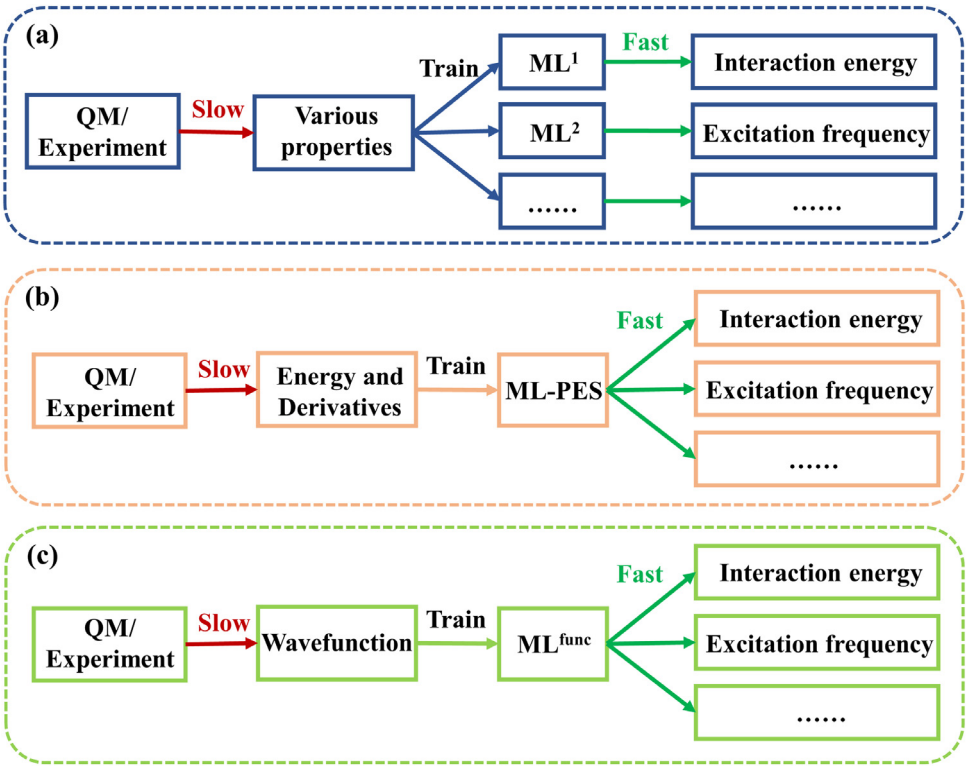

Fig. 2.8. Three categories of ML models applicable to chemistry and physics. (a) forward model that directly predicts the chemical property, in which different ML models (ML^1, ML^2, etc.) are constructed for different predictions; (b) potential energy surfaces (PESs) model that predicts properties through PESs; and (c) hybrid model that constructing ML-based exchange and correlation functionals (ML^{func}) of electron density or wavefunction for predicting ground-state properties.

process simple and easiest for understanding and applications. For decades, under the forward model process ML methods have been successfully applied to a wide range of predictions. For example, very recently Liu et al. (2019) proposed a deep learning algorithm to predict the chemical shift using an atom-centered Gaussian density model as the data representation, which presents good consistency in the chemical shifts of ^{13}C, ^{15}N, and ^{17}O compared with the QM methods [69]. In the same year, Kelley et al. reported a graph convolutional model with fixed molecular descriptors and previous graph neural architectures that outputs reliable results at the level of experimental reproducibility [68]. Subsequently, Morita et al. (2020) introduced a supervised ML model to predict the optical dielectric constants of crystals by combining the SVM, DNN and DFT methods. In addition, with an appropriate process design, ML methods have been successfully used to predict 3D structures, spectroscopy, chemical reactivity and molecular properties [68–72].

Based on QM or experimental results, ML advances computational physics and chemistry by quickly predicting properties. However, as shown in Fig. 2.8(a), the transferability leads to great limitations on the efficiency of the forward model strategy. Different property predictions require particular ML models (ML^1, ML^2, etc.) and datasets, resulting in enormous computational or experimental costs. On the other hand, these direct property predictions do not explicitly capture fundamental chemical concepts, which makes it challenging to explain chemical and physical discoveries. In this case, PES models (Fig. 2.8(b)) and hybrid models (Fig. 2.8(c)) have drawn substantial attention. All chemical and physical properties can be obtained by the electronic Schrödinger equation and derived from the grand-state wavefunction. Therefore, an electronic structure or wavefunction ML model is very practical for predicting ground-state properties (Fig. 2.8(c)). In 2017, Hegde and Bowen introduced a transferable ML model using representations of atomic neighborhoods and KRR to predict DFT Hamiltonians, which were used to calculate the band structure and ballistic transmission with a considerable accuracy [178]. Two years later, Townsend and Vogiatzis [179] built a model to predict the converged coupled-cluster singles and doubles (CCSD) amplitudes from the electronic properties inherent to MP2. More recently, other models based on supervised ML or DP have been proposed for predicting QM wavefunctions as well as exchange and correlation functionals.

The buildings and applications of hybrid models require extensive knowledge of quantum mechanics, which limits the wide-range applications, not to mention the model's ultrahigh requirements for accuracy. In that case, PESs are introduced for predicting the energy and properties without electronic structure calculations (Fig. 2.8(b)). Among the various ML methods, NNs, which are biology-inspired functions, are the most widely used methods for constructing PESs.

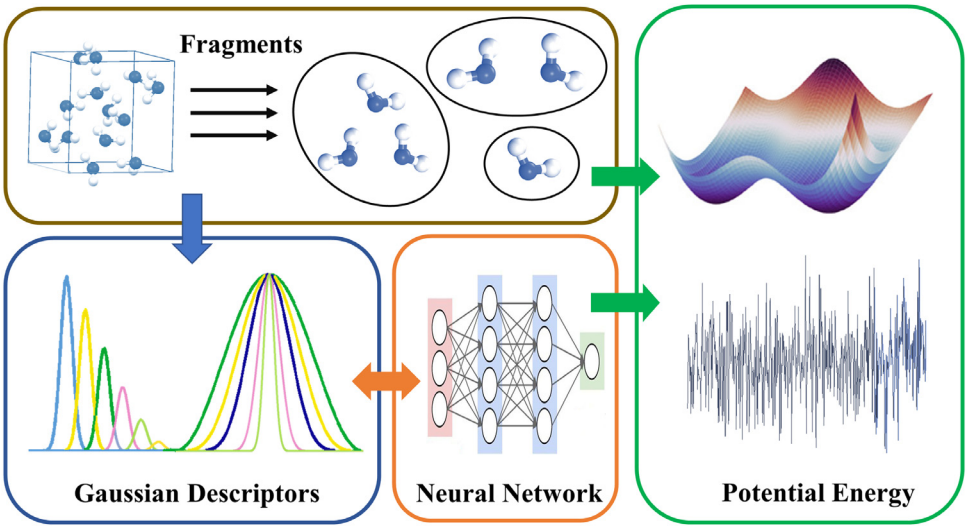

Fig. 2.9. Process of constructing PESs for one-, two-, and three-body terms with NN from molecular crystal structures. *Source:* Reproduced from Han et al.'s work [183].

In 1995, Blank et al. introduced feed-forward NN (FFNN) to represent the DFT PESs for molecule-surface scattering [180], which is known as the beginning of modern ML potential (MLPs). As shown in Fig. 2.9, the construction of MLPs is very modular, containing only two domain components: the structural representation or descriptor and the ML method. They are proposed using a consistent dataset of DFT energies and forces obtained from atomic structures. Then, these data are converted to ML representations, such as Gaussian descriptors via feature engineering. With these representations, the MLP can be trained using an appropriate ML method. The structure-energy relationship in MLPs is presented by an ML method that does not bring any approximations or assumptions, apart from the DFT methods used for the generation of a consistent dataset. Since then, all NN PESs (NNPs) have been based on feed-forward NNs (FFNNs). However, these early NNs are more suitable for low-dimensional systems and suffer great restrictions when applied to high-dimensional systems. For example, the number of input nodes for an FFNN is fixed, which corresponds to system freedom, allowing only one FFNN PES to be applied to one particular system. In addition, the rotational and translational invariance of energy and permutation symmetry [181,182] make the development of appropriate inputs a major challenge.

In 2007, Behler and Parrinello [184] proposed a novel NNP method suitable for high-dimensional systems with thousands of atoms, in which a separate FFNN is used for each atom in the system, and the total energy is calculated by combining all the atomic contributions. Fig. 2.10 shows an illustration of a high-dimensional NNP (HDNNP) for the three-component system, which contains three elements (A, B and C). Using a separate FFNN for each atom, this HDNNP overcomes the restriction of input nodes in early NNPs. For each element, the atomic NNs use a similar design in setting the number of hidden layers and neurons and the same weight parameters. In HDNNP, many-body atom-centered symmetry functions are used to describe the atomic environments, and a series of atomic NNs is used to determine the relationship between the descriptors and energies. Smith et al. [185] proposed an extensible neural network potential (ANI) for molecular energy in 2017, followed by a general-purpose neural network potential based on transfer learning. Subsequently, Yao et al. [186] introduced TensorMol-0.1: a neural network augmented with long-range physics. Soon after, Wang et al. [187] (2018) presented a deep learning package, DeePMD-kit, for many-body PES representation and quantum molecular dynamics for molecules, extended systems, metallic systems and bonded systems. Using transfer learning (TL), Smith et al. [77] (2019) introduced a general-purpose PES approaching coupled cluster (CC) accuracy. Very recently, Nandi et al. [188] (2021) proposed the permutationally invariant polynomial (PIP) method to train high-dimensional PESs that reach the accuracy of the CCSD(T) level. The PIP method is a "Δ-machine learning" approach that can bring a property (such as PES) based on DFT energies and gradients to the accuracy of the CC level. Recently, other MLP models have also been proposed with considerable accuracy, such as TorchANI [189] and AP-Net [190]. A list of constructions of MLPs is shown in Table 2.3, including the model names, the testing property and error, and the ML method used for potential training.

For years, numerous descriptors, which are one of the main components for constructing MLPs, have been proposed for HDNNPs to describe atomic and molecular environments, such as Coulomb matrices, molecular local frames, NNs for electrostatic multiples, the combination of Gaussian processes and four-dimensional spherical harmonics, the smooth overlap of atomic positions (SOAP), Fourier series of atomic radial distribution functions, and the bag of bonds approach. The improvements in the ML methods and descriptors have significantly promoted the improvements of MLPs. With the substantial progress of these descriptors and methods, a wide range of successful HDNNPs have been reported for different

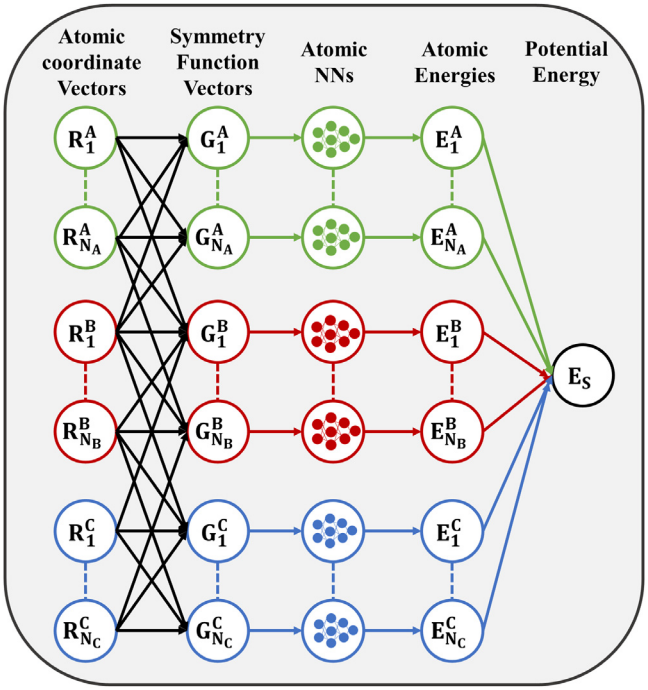

Fig. 2.10. Illustration of the high-dimensional NN (HDNN) suitable for the three-component system, which contains elements A, B and C.

Table 2.3
List of applications of ML methods for constructing PESs.

Year	Models	Testing property	Testing error	ML method
2007 [184]	HDNNP	Energy and force	5–6 meV and 0.2 eV/Å	NN
2017 [185]	ANI-1	Energy	0.4 kcal mol^{-1}	DNN
2017 [186]	TensorMol	Energy and force	0.054 kcal mol^{-1} $atom^{-1}$ and 0.49 kcal mol^{-1} $Å^{-1}$	NN
2018 [187]	DeepMD-kit	Energy and force	4.3% and 2.9%	NN
2019 [77]	ANI-1ccx	Energy	0.23 kcal mol^{-1}	TL
2020 [189]	TorchANI	coefficient of determinations	0.96 and 0.99	DL
2020 [190]	AP-Net	Interaction energy	0.37 kcal mol^{-1}	NN
2021 [188]	PIP	Harmonic frequencies	31 cm^{-1}	Δ-ML

systems, including small molecules (TiO_2, ZnO, and water), molecular clusters, bulk materials, metal cluster surfaces and aqueous electrolyte solutions.

As discussed above, direct property prediction is the most efficient and practical way for ML methods to be applied to chemical science. Since the performance of ML training strongly depends on the quality of datasets, experimental results or expensive QM calculations are usually essential for faithful predictions. However, each predicted property requires re-collecting the dataset and training the ML model, leading to a great challenge for ML methods to be transferable and efficient. In contrast, ML methods can also be used for constructing exchange and correlation functionals of the electron density or wavefunction, which are capable of calculating a wide range of properties. Despite the broad applicability and transferability, ML constructed functionals still suffer limitations in terms of accuracy and system size. Nevertheless, ML-constructed exchange and correlation functionals of electron density or wavefunctions are worthy of further development and are the most promising solution for widespread applications in QM physics and chemistry. Finally, with the capability of predicting various properties that relate to energy, ML-constructed PESs are also a reliable choice for molecular crystalline systems. More importantly, PESs can be combined with fragment-based schemes, greatly expanding their applicability for a wide range of systems with different compositions and sizes.

2.4. Discussion: improvements and outlook

In recent decades, since the first development and application, dispersion corrections for the vdW dispersion energy in standard electronic structure methods (e.g., HF or the underlying DFT) have reached maturity and are being continuously applied in computational physics and chemistry. In line with theoretical development and improvement, London

dispersion interactions have been gradually accepted as an essential concept in computational physics and chemistry. The dispersion interactions influence the chemical properties only for large systems with more than 30 atoms, which results in a long time for establishing widely used DFT-D methods. For a long time, large molecules could not be treated with QM in computational physics and chemistry, whereas DFT calculations can currently be performed on molecules with 100–200 atoms with periodic boundary conditions. Moreover, these post-HF methods (e.g., MP2-related methods and CCSD(T)) have been greatly improved to describe electronic structures well, and thus achieve a higher accuracy. Improvements in modern computer hardware and parallel versions of programs have made periodic MP2/CCSD(T) approaches more applicable for large crystal systems containing tens of atoms. However, despite the improvements in accuracy, the computational scaling of periodic MP2 and CCSD(T) methods are much steeper compared with underlying DFT models, which limits these electronic structure-based methods from being widely applied to large systems.

On the one hand, the previously introduced fragmental approaches have presented a good performance in computational chemistry and physics, especially for molecular crystals. These fragmental methods all have advantages, and there is no clear evidence that any certain fragmental method is better than the others. For example, many body schemes with an embedded charge can be simply achieved and performed in energy calculations with considerable performance, in which the polarization effects are evaluated by the embedded charges. On the other hand, the incorporation of embedded charges makes each energy gradient element in the Hessian matrix influenced by the contribution of all of the remaining atoms, and thus much attention is required for dealing with the energy gradient or the Hessian matrix. In addition, the dispersion interaction and many-body exchange are essential effects in many-body embedded schemes, which must be considered in further research.

Since the report of the first HDNN is suitable for high-dimensional systems, many improvements have been made in the construction of ML PESs, especially NN ESPs. Constantly constructed MLPs have reached a considerable accuracy for computational physics and chemistry. A main advantage of ML approaches is the capability to be introduced to a series of systems in the same way, and thus extensive experience is not required for new model construction. A list of MLPs that have received wide attention is shown in Table 2.3, in which the MLPs are suitable for a wide range of systems and are easily applied for various properties. The computational cost of ML predictions is determined by the acquisition of training data and the training process, which also significantly influence the prediction accuracy. Therefore, large and high-quality datasets, new powerful ML methods that are capable of making predictions with a small amount of data and careful validation are strongly required and have become challenges for ML applications in the field of physics and chemistry. Other challenges remain in a substantial effort for ML model development, and the selection bias encoded in many training datasets is another serious problem lurking behind rigorous and robust statistical learning curves. In addition, several challenges remain in constructing transferable and accurate ML-based QM models of electron densities, determining the irreducible set of variables and assuring constant prediction errors. Currently, apart from improving the ML methods and descriptors, integrating ML-QM models across different levels of theories is expected to be a promising research direction. Currently, the development of ML-driven QM models has become a popular research field, leading to various reports for new models. Predictably, the rapid development and improvement in ML methods and descriptors can extend the range of applications of ML-driven QM models in the near future.

3. Predictions of inorganic small molecules and crystals

Predictions of molecular crystals under high pressures have been a significant topic for numerous new materials and novel phenomena. For years, substantial efforts have been made to develop efficient methods with accurate DFT, MP2 or even CCSD(T) levels, which advances the electronic structure methods with dispersion corrections, post-HF models, fragment-based QM methods and ML-based QM methods. With a high precision and efficiency, fragment-based QM methods have been widely applied in the predictions and explanations of inorganic small molecule crystals at a high pressure. For example, fragment-based QM DFT/MP2 methods were successfully applied in the prediction of crystal structures, EOS, vibrational spectra and phase transitions for small molecules, such as carbon dioxide (CO_2) [13–15], ice (H_2O) [17–21], carbon monoxide (CO) and solid ammonia (NH_3) [22]. High-pressure conditions also have the chance to make small molecules superconducting and super-hard. Apart from high pressure conditions, new phases, structures, and properties under negative pressures are also interesting. For example, in recent years several new phases of ice under a negative pressure (virtual ices) have been reported and widely investigated using QM methods. As discussed in Section 2, the periodic DFT/MP2 methods require a significant computational cost with an extremely steep complexity scaling. Fragment-based QM methods significantly reduce the computational cost while maintaining a high level of accuracy. However, the requirement of high-level electronic structure methods for maintaining a sufficient accuracy and rapid growth of the number of fragments result in a great increase in the computational cost for large crystal systems. In that case, with considerable accuracy and extremely low computational cost, ML methods, especially NNs, have been widely applied to computational physics and chemistry to accelerate the prediction of small molecular properties, PESs, and phase transitions.

3.1. Small molecules at high pressures

For decades, numerous new phases and novel phenomena of crystals have been predicted and determined under extremely high pressures. Small molecules and corresponding crystals have attracted considerable attention for their physics and abundant polymorphs at high pressures. For years, all-atom QM methods, fragment-based QM methods, and ML-driven methods have been widely used for the predictions of crystal structures, structural stability, EOS, vibrational spectra, and phase transitions. These methods can be applied to a wide range of small molecules and corresponding crystals at various conditions, such as two-atoms (H_2, N_2, CO, etc.), three-atoms (CO_2, H_2O, H_2S, etc.), four-atoms (NH_3, HCHO, etc.) and multi-atoms (CH_4, etc.) compounds. For example, these QM methods have present considerable success when being applied to solid CO_2, H_2O, CO and NH_3 for structures and properties predictions.

CO_2 is the major ingredient of the atmospheres of terrestrial planets (such as Mars and Venus) and is usually found in crystal form in asteroids and planets [191–193]. Despite decades of extensive experiments and theoretical simulations, the structure, properties, and phase diagram of carbon dioxide under extreme pressures and temperatures are yet to be properly understood. In addition, H_2O is of great importance to the astrophysics and geophysics of planets and satellites, many of which apparently exist in satellites of giant planets, such as Ganymede and Callisto, as well as the nuclei of comets [194,195]. The structures and properties of H_2O under extreme conditions have exploded in popularity for decades due to its richness of structures and complex phase diagrams [18,195]. In addition, the physics fundamentals of CO crystals and structural differences are of great fundamental and practical significance, not only because of their high toxicity but also because of their therapeutic potential and important signaling capabilities in physiological and pathophysiological situations. Despite decades of theoretical research, the crystal structures of solid carbon monoxide polymorphs and the transitions between polymorphs at the atomic level are still far from being well understood. In addition, NH_3 is one of the most basic components on the planet, and its high-pressure characteristics play an important role in planetary science [196,197]. Solid ammonia crystals frequently adopt multiple distinct polymorphs exhibiting different properties. Predicting the crystal structure of these polymorphs and under what thermodynamic conditions these polymorphs are stable would be of great value to the environmental industry and other fields. Decades ago, lattice dynamics, infrared spectroscopy and Raman spectra were used to determine the crystal structures and phase transitions of small molecules under high-pressure conditions [198–201]. However, these methods suffer a challenge in providing accurate and detailed descriptions and explanations of crystal properties and phase transition conditions. In recent years, periodic QM methods and fragment-based QM methods have been widely used for predicting small molecular crystals under high-pressure conditions, including EOS, vibrational spectra, crystal structure and phase transition.

Equation of state (EOS)

Small molecules, such as CO_2, H_2O, CO, NH_3, CO and N_2, usually possess very rich phase diagrams, containing a great number of phases under a high pressure. For a long time, however, the knowledge of these phase diagrams was limited, with crystal structures and phase transition boundaries far from being clearly determined. With the development of fragment-based QM methods, much effort has been made toward predicting and determining the crystal structures, structural parameters, properties and phase transitions of small molecular crystals under extreme conditions. In 2019, Li et al. [7,15] introduced an *ab initio* approach that combines an improved fragment-based method and electronic structural method (MP2) for the prediction of solid CO_2 at high pressure. After full structural optimization, this approach successfully confirmed the crystal structure of solid CO_2 phases I, II and III and predicted their phase transitions, which matched the experimental results very well. Fragment-based QM methods have also been applied to other small molecules, such as solid NH_3 and ice, for the prediction of lattice parameters and EOSs. In 2019, Huang et al. [14] performed a theoretical investigation on solid CO_2 phase VII and solid NH_3 phases I and IV [22] for the prediction of EOSs using an embedded generalized molecular fractionation (EE-GMF)-based QM method with an accuracy of the MP2 level. Very recently, Xu et al. [17] predicted the EOSs of ice phases IX and XIII in the pressure range of 0–0.5 GPa using the MP2-based fragment method. In the same year, Xiao et al. [19] presented the lattice parameters and EOSs of ice phases XV, XIV, and VIII. In addition, Lu et al. [18] predicted the lattice parameters and EOSs for a series of ice crystals at high pressure, including phases II, VI, VII, VIII, IX and XV. Han et al. [183] predicted the EOSs of ice phases IX and XV using an NN-based fragmental method with the accuracy of the MP2 level. The crystal structures and predicted EOSs of solid CO_2 (phases I, II, III, and VII) and ice (phases IX, XV, VIII and XIV) are shown in Fig. 3.1.

Phase transition

Apart from EOSs, the prediction of phase transition boundaries is also a main application of fragment-based QM methods. Fig. 3.2 shows the predicted and observed phase diagrams of solid CO_2 and H_2O, which are taken from the work by Han et al. [15] and Lu et al. [18]. In 2019, Han et al. successfully determined the crystal structure of the solid CO_2 phase II and predicted the phase boundaries of solid CO_2 phases I–III and II–III at a pressure range of 10–20 GPa. As early as 2017, solid CO_2, phases III and VII, were suggested by Sontising et al. [202] to be identical with a comparison of theoretical predictions and previous experimental results. In 2019, however, Huang et al. [14] predicted the phase transition between solid CO_2 phases I and VII with the existing crystal structures. Currently, the crystal structure of CO_2 phase VII has not been determined, and further investigation is required. In 2020, much effort was made to predict phase transitions between solid ices under high pressure. Xiao et al. [19] investigated ice phases XV, XIV and VIII and predicted that a triple point of phase transition occurs at approximately 1.258 GPa and 112 K. Soon after, Lu et al. [20] predicted

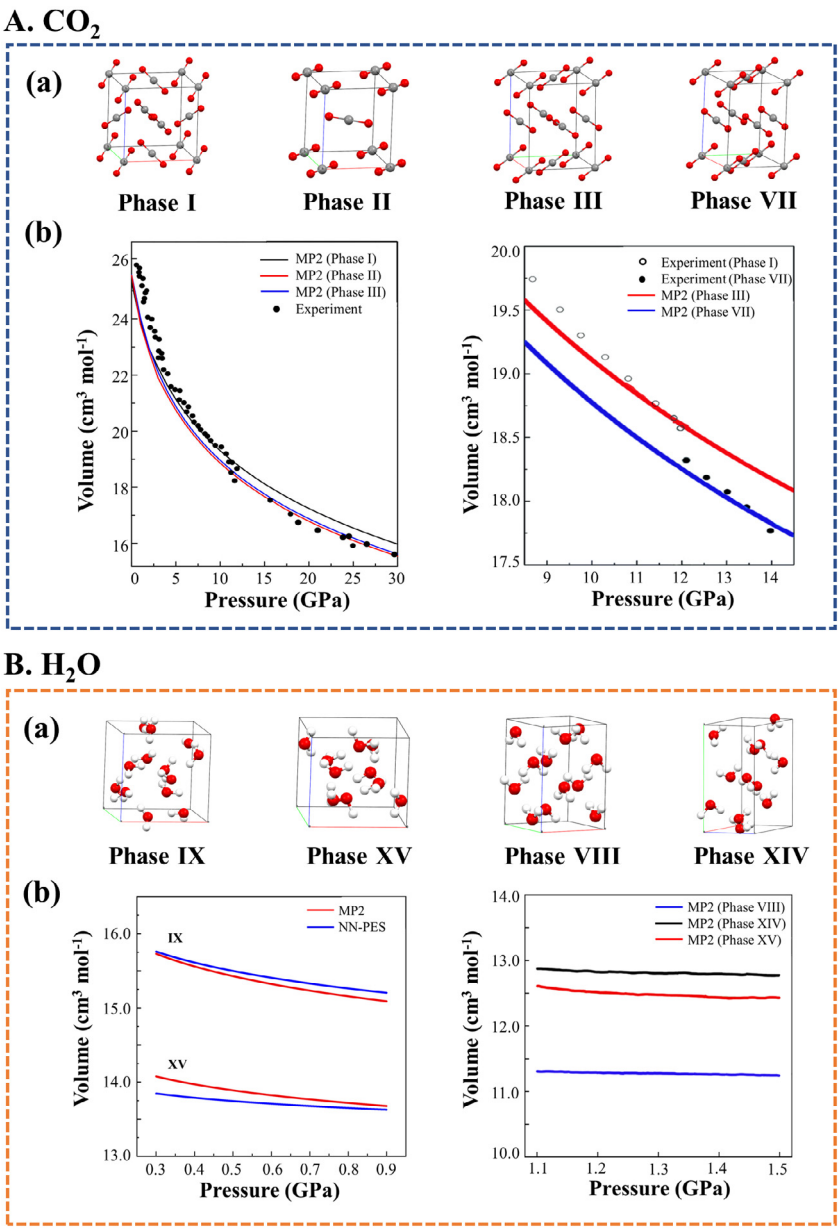

Fig. 3.1. Crystal structures and volume–pressure relationships of solid CO_2 and ice phases. A. (a) Crystal structures and (b) volume–pressure relationships of solid CO_2 phases I, II, III and VII. B. (a) Crystal structures and (b) volume–pressure relationships of ice phases IX, XV, VIII and XIV, reproduced from the works by Han et al. [183] (© 2021 American Physical Society), and Xiao et al. [19] (© 2020 Elsevier), respectively.
Source: A. Reproduced from the works by Han et al. [15] and Huang et al. [14], respectively.

the phase transition of a series of ices, including ordered phases II, IX, VIII and XV and hydrogen-disordered phases VI and VII. In addition, Huang et al. [22] also presented the phase transition of solid NH_3 between phases I and IV using a fragment-based QM method along with a complete basis set correction (CBS) at the CCSD(T) level. In these theoretical works, the phase transition between two crystal structures is determined using a Gibbs free energy comparison, which was calculated by the fragment-based QM method. Fig. 3.3 shows the Gibbs free energy surfaces of solid CO_2 phases I–III (Fig. 3.3(a)), II–III (Fig. 3.3(b)), and I–IV(Fig. 3.3(c)) and ice phases VIII, XIV and XV (Fig. 3.3(d)), which were reported by Han et al. [15], Huang et al. [14], and Xiao et al. [19], respectively. With these energy surfaces, the phase transition boundary can be obtained from the intersection of the two surfaces.

Vibrational spectra

As a distinct chemical fingerprint from a particular crystal or molecule, Raman vibrational spectra can be used to rapidly identify the crystal and distinguish it from others. Several theoretical works for predicting phase transitions also

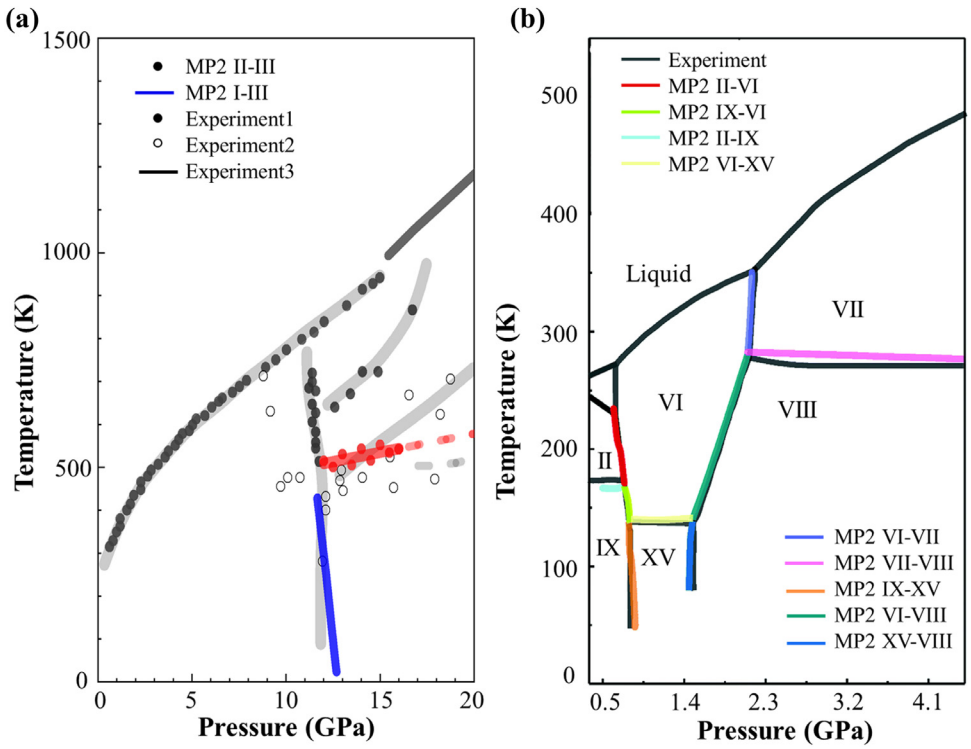

Fig. 3.2. Calculated phase transition boundaries and measured phase diagram of (a) solid CO_2 (phases I, II, and III) and (b) ice (phases II, VI, VII, VIII, IX, and XV).
Source: Reproduced from the works of Han et al. [15] and Lu et al. [18], respectively.

reproduced the Raman spectra. Along with phase transition prediction, Han et al. [15] reproduced the Raman spectra in the librational regions of solid CO_2 phase I at 11.7 GPa, phase II at 26 GPa, and phase III at 26 GPa, which were then compared with the experimentally observed spectra and presented good consistency with these experimental results in terms of the number and positions of peaks. For CO_2 phase VII, Huang et al. [14] reproduced the Raman spectra at 12.6 GPa with the optimized crystal structure as evidence for the correctness of the existing structure. In ice prediction works, Raman spectra were also used to confirm the optimized crystal structure. For example, Xiao et al. [19] compared the calculated Raman spectra of ice phase XV (0.9 GPa), XIV (1.2 GPa), and VIII (2.8 GPa) with the related observed results, which shows a good agreement between them. Similarly, Lu et al. [18] calculated the Raman spectra of ice II at 0.27 GPa and 0.28 GPa and compared them with experiments in the librational region and stretching region. The calculated and observed Raman spectra of solid CO_2 and ice are shown in Fig. 3.4, including CO_2 phases I, II, and III (Han et al. [15]), CO_2 phases I and VII (Huang et al. [14]), H_2O phase I (Lu et al. [18]), and H_2O phases XV and VIII (Xiao et al. [19]). In addition, several works also provided the calculated and observed frequencies of Raman bonds depending on the pressure. For example, Huang et al. [14] demonstrated maximum errors of 5.9% for CO_2 phase I and 6.5% for phase VI between their calculated frequency and experimental results. The predicted frequencies of Raman bands depending on the pressure for solid CO_2 are shown in Fig. 3.5, which are taken from the works of Han et al. [15] and Huang et al. [14].

As discussed above, fragment-based QM methods have been widely applied in predicting small molecular crystals (such as CO_2, H_2O and NH_3) at a high pressure for the significant decrease in the computational cost as a powerful research tool that complements the experiment. Notably, the mentioned works were performed using the fragment-based method and electronic structure method (MP2). Huang et al. [22] also performed a CBS correction at the CCSD(T) level to improve the quality of the calculations. Within the framework of these fragment-based QM methods, any electronic structure method (such as DFT, MP2, CCSD(T)) can be used as the QM method for the calculation of different orders of many bodies. Therefore, the accuracy of these fragment-based QM methods can be increased by using high-level electronic structure methods, which will result in a significant increase in the computational cost. Attention strongly needs to be paid to the balance between accuracy and efficiency.

3.2. Small molecules at negative pressures

Ice and clathrate ice are omnipresent in the solar system, including Earth, comets, icy moons of the giant planets, and asteroids. Depending on the pressure and temperature conditions, ice presents an extremely rich and complicated

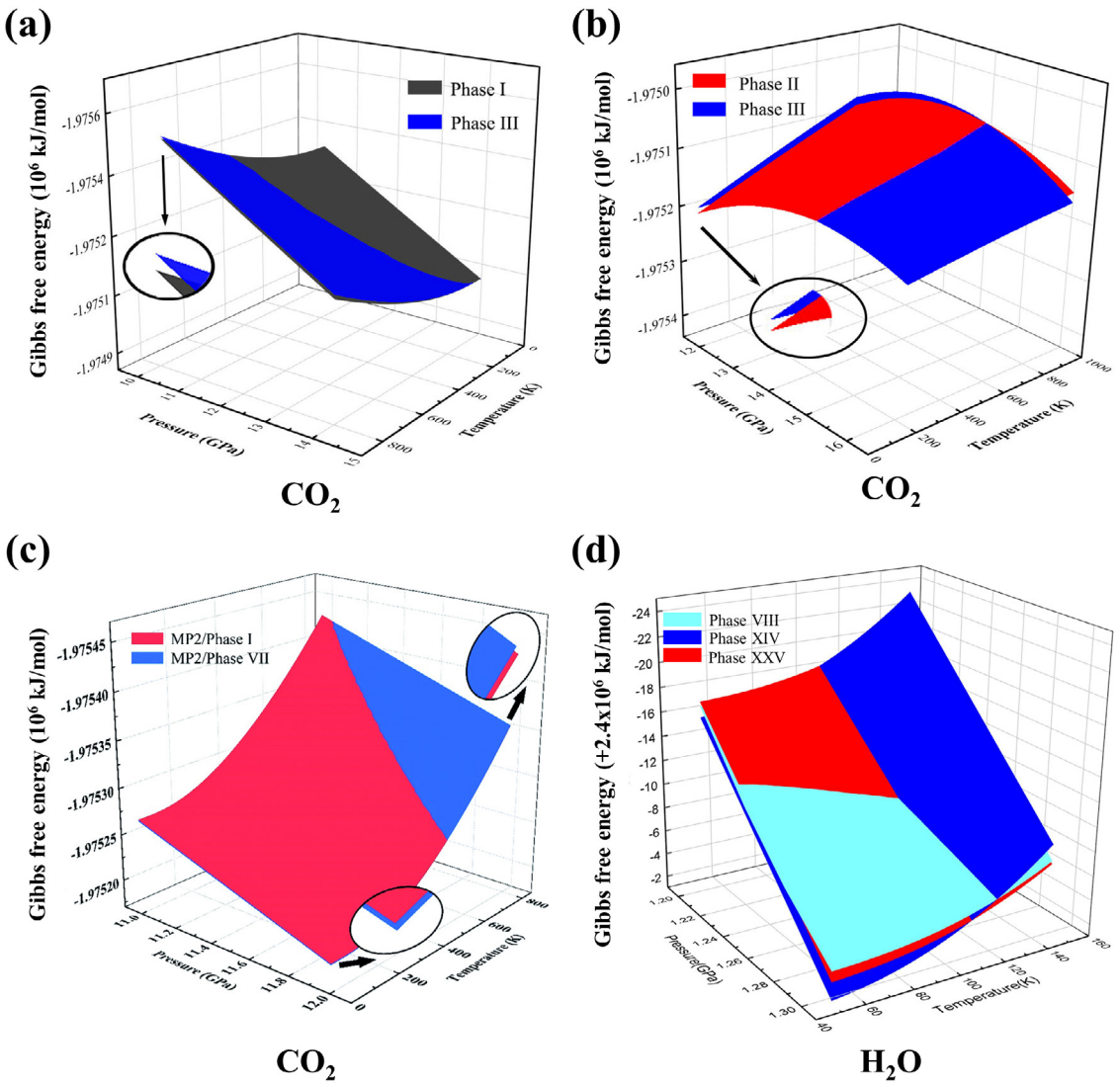

Fig. 3.3. Calculated Gibbs free energy surfaces of solid CO_2 (a) phase I (black surface) and III (blue surface), (b) phase II (red surface) and III (blue surface), (c) phase I (red surface) and VII (blue surface) and (d) ice phase XV (cyan surface), XIV (blue surface) and VIII (red surface).
Source: (a) (b) are reproduced from Han et al.'s work [15], (c) is reproduced from Huang et al.'s work [14], and (d) is reproduced from Xiao et al.'s work [19].

phase diagram with seventeen crystalline polymorphs. Among these ices, phases XI and Ih [203] are located at a low pressure and slightly negative pressure in the phase diagram [204] with a low mass density (0.94 g/cm^3). Clathrate ice, a clathrate compound with the inclusion of guest molecules, also exhibits a series of different polymorphs that are stable under the Earth's oceans and abundant on comets and asteroids of the solar system [205]. As co-crystals, the crystallization of clathrate hydrates requires guest atoms or molecules. The unique inner cavities and low mass density mean the clathrate hydrates are expected to remain stable under a negative pressure. Early clathrate hydrates usually require guest molecules to keep the structure stable via vdW interactions with host water molecules, such as tetragonal structure III [206], cubic hydrate structures sI [206,207] and sII [208], hexagonal hydrate structure sH [209], and tetragonal hydrate structure sT [210] at a high pressure. In recent years, guest-free clathrate hydrates have drawn significant attention since Falenty et al. [205] (2014) introduced ice XVI, a guest-free clathrate obtained by excluding the Ne atoms from a type sII clathrate hydrate. Subsequently, several guest-free clathrate hydrates were reported to complete the phase diagram of ice at a negative pressure, including phases sIII [211], sIV [212], sL [213], ITT [214], dtc [215] and EMT [204]. As research progresses, a series of new structures of clathrate ices have been determined [216], but their locations on the phase diagram are still unclear. Fig. 3.6 shows the eight structures of clathrate ices at a negative pressure and one ice crystal XI. Recently, QM methods and MM simulations have been used to investigate the crystal structure

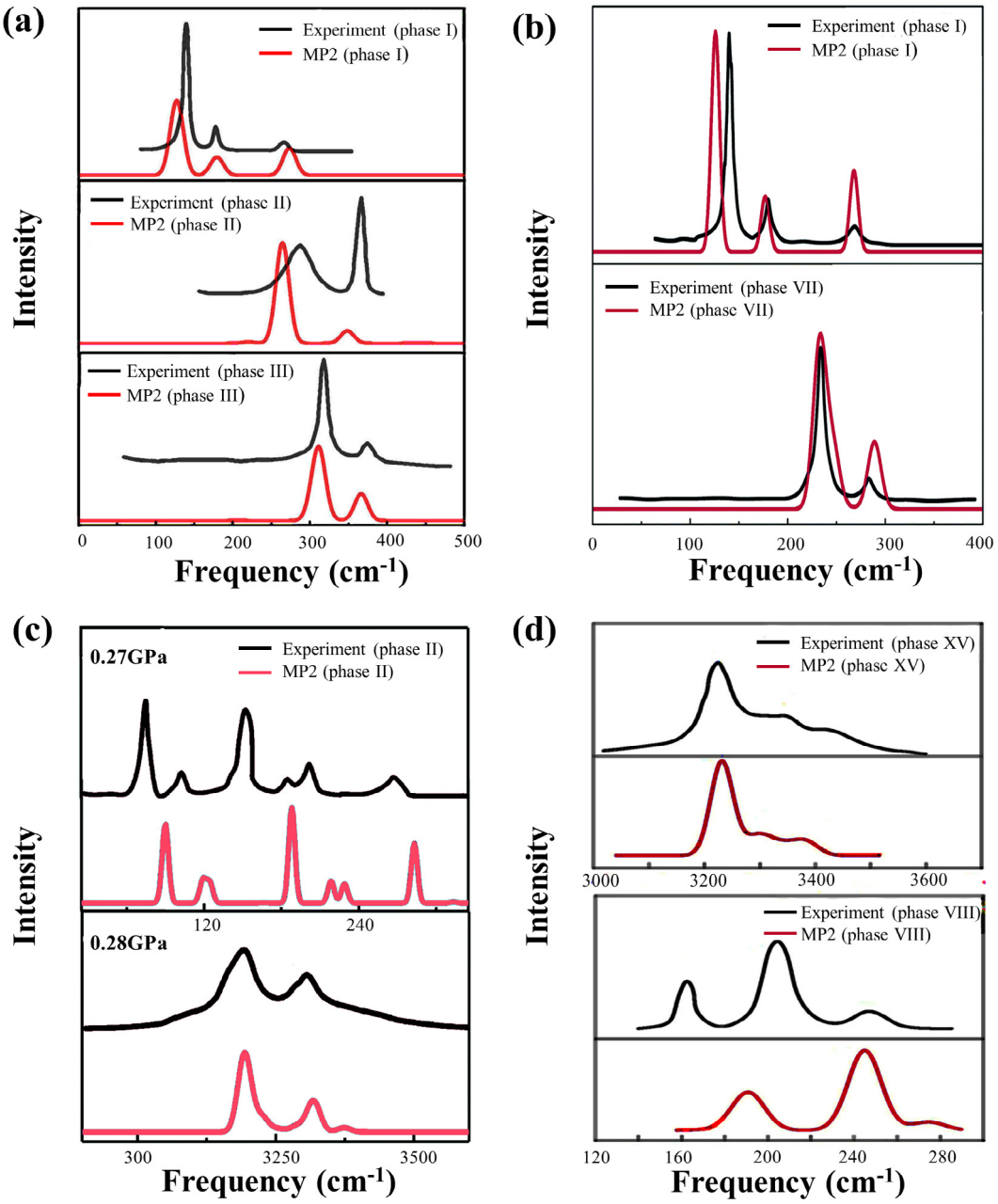

Fig. 3.4. Calculated and observed Raman spectra of (a) solid CO_2 phase I–II–III, (b) solid CO_2 phase I–VII, (c) ice phase I, and (d) ice phases XV, VIII. *Source:* Reproduced from the works of Han et al. [15], Huang et al. [14], Lu et al. [18], and Xiao et al. [19], respectively.

and phase diagram of clathrate ice under negative pressures. In 2016 [211] and 2017 [212], Huang et al. reported two new clathrates (sIII and sIV) and predicted phase diagrams using DFT calculations and Monte Carlo simulation with the TYP4P model. In 2017, Matsui et al. [214] examined three types of low-density ices (zeolitic ices, space fullerene ices, and aeroices) and demonstrated that the ITT and sII hydrates are the most stable phases for zeolitic ices and space fullerene ice, respectively, and aeroices are the most stable solid phases of water near the absolute zero temperature under a negative pressure. Soon later, with first-principal calculations Liu et al. [213] (2018) reported a new phase of ice (sL) with an ultralow density (0.6 g/cm^3), which is predicted to be stable under a low negative pressure. Next year, Liu et al. [204] determined ultralow-density porous ice (EMT, ∼60% of the mass density of ice XVI) with the largest internal cavity using both DFT calculations and Monte Carlo simulations. In 2019, Matsui et al. [215] demonstrated that the limit of mechanical stability for the crystalline phases will terminate the region of the most thermodynamically stable phase in the phase

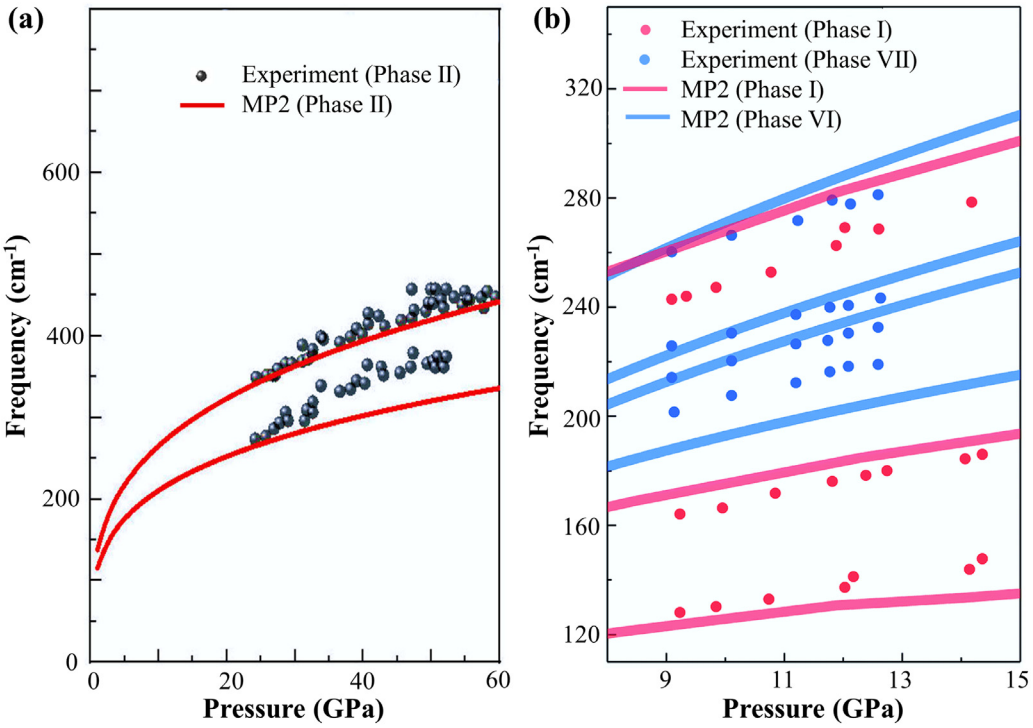

Fig. 3.5. Predicted frequencies of Raman bands of solid CO_2 (a) phase II, and (b) phase I and VII. *Source:* Reproduced from the works of Han et al. [15], and Huang et al. [14], respectively.

diagram and presented a very complicated diagram of low-density ice phases at negative pressure. In the same year, Yagasaki et al. [217] presented a novel ice crystalline structure with hypothetical dtc zeolite topology using an all-atom MD simulation. Very recently, high-level QM methods have been used for clathrate hydrates. Using the MP2 and CCSD(T) methods combined with fragment-based methods, Lu et al. [20] (2021) investigated nine empty clathrate hydrates of ice (sII, sIII, sIV, sH, sL, dtc, EMT, ITT and ice XI) at negative pressures. Based on structure optimization and Gibbs free energy calculations, they presented a renewed phase diagram of ice at negative pressures, as shown in Fig. 3.7.

As discussed above, the accurate crystal structure and complete phase diagram at a negative pressure are still not fully determined, and further in-depth research is also required. As fragment-based methods progress, high-level QM methods combined with fragment-based methods will substantially benefit from investigations.

3.3. Superconducting and super-hard molecules at high pressures

Superconducting and super-hard materials are essential for a myriad of scientific, biomedical, and industrial applications. The contradiction between covalent bonds in super-hard materials and metallic bonds in superconductors makes superconductivity and super-hardness in the same material a very interesting and precious effect. Their zero-resistance and anti-pressure abilities stem from the relationship between the crystal structure, chemical composition, and microstructure. The complexity of this interdependence limits researchers to conduct comprehensive experimental investigations but can be supported by theoretical calculations. Generally, low-temperature superconductivity [218] is more practical than high-temperature superconductivity. However, these low-temperature superconductors require liquid helium temperature conditions, which significantly limits their applications. Years ago, with the establishment of the electron–phonon mechanism of superconductivity based on the Bardeen–Cooper–Schrieffer (BCS) [219] theory, the multiband effect [220,221] based on electronic structure, and the McMillan equation, the prediction of superconducting has become practical. In particular, ultrahigh pressure is considered to be physical doping by influencing the electronic structure of covalent molecules. The breaking of bipolarons in the flat/sleep band model is a good description of superconductors induced by ultrahigh pressure [222,223].

In 2017, based on the CSP method and QM calculations, Liu et al. [224] demonstrated that the stable hydrogen-rich phases exhibited unusually high superconducting temperatures at a high pressure, such as La–H and Y–H systems. For example, LaH10 is predicted to be stable above 200 GPa with a sodalite-like face-centered cubic (fcc) structure, with a very high superconducting transition temperature of 274–286 K at 210 GPa. This study suggests that density hydrides may represent a new class of potential very high-temperature superconductors. In the same year, Peng et al. [225] predicted

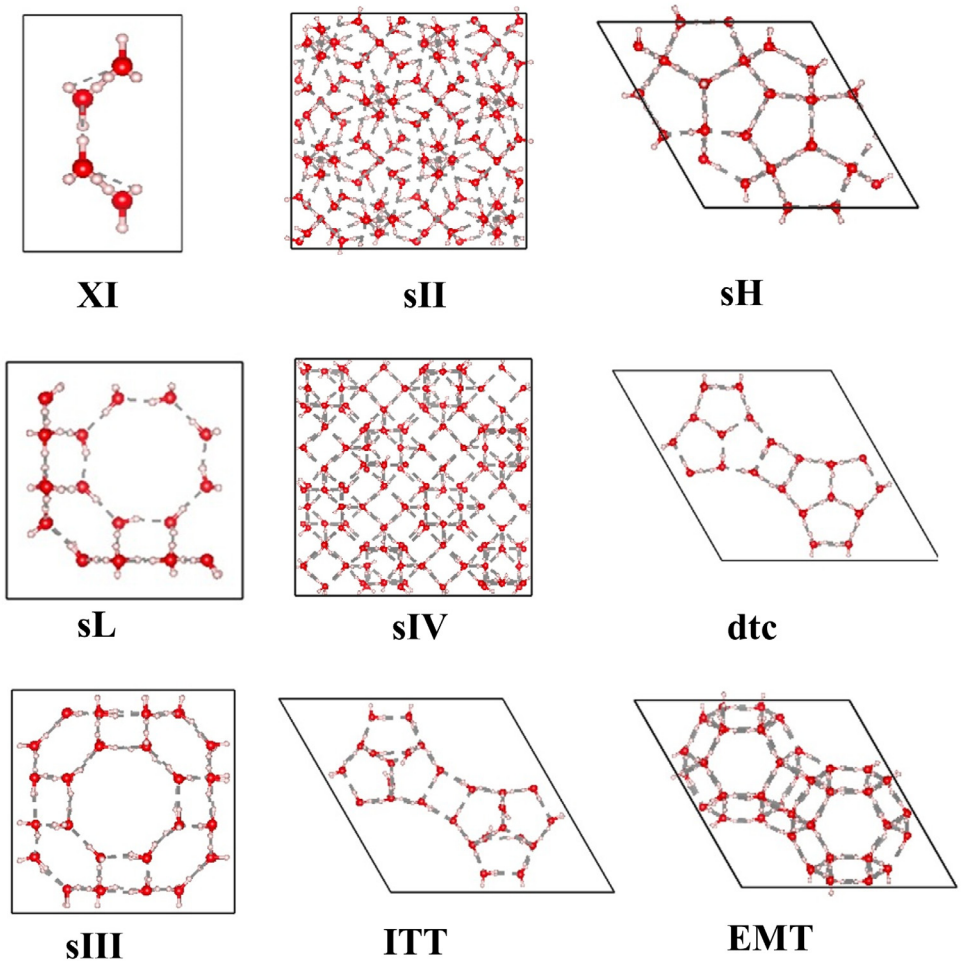

Fig. 3.6. Structures of eight clathrate ices (in plane view) at negative pressures, including phases sII, sIII, sIV, sH, sL, ITT, dtc and EMT, compared with one high-pressure ice phase XI.
Source: Reproduced from the work by Lu et al. [20].

that a YH32 clathrate structure of stoichiometry YH10 is a potential room-temperature superconductor with a predicted Tc of up to 303 K at 400 GPa using first principles. In 2019, Zhao et al. [226] studied Li-rich phosphides using first-principles swarm structure calculations and predicted that a pressure-induced stable Li6P electride was a superconductor with a superconducting transition temperature Tc of 39.3 K.

Recently, *ab initio* methods have been applied to the investigation of superconducting and super-hard materials, especially the lightest oxygen group hydride H_2O. With a general *ab initio* method, Lu et al. [21] (2020) investigated the superconductivity of the structures [227] of ice with space groups $Pmc2_1$, $P2_1$ and $C2/m$ at terapascal pressures and demonstrated that the structure of $Pmc2_1$ is energetically stable above 930 GPa and will transform to a $P2_1$ structure above 1.3 TPa. As the pressure increases, the $P2_1$ structure will remain stable until 4.28 TPa and turn into a metallic structure $C2/m$ above 4.28 TPa. They also calculated the electronic band structures and density of states (DOSs) at the Fermi level and the electron localization functions (ELFs) of the three predicted structures at the terapascal pressure, as shown in Fig. 3.8, which were then used for the analysis of superconductivity and super-hardness. They predicted that the ice was super-hard with space group $P2_1$or $C2/m$ above 1.3 TPa and turned into a superconductor (Tc: 1.782 K) with space group $C2/m$ as the pressure increased to 5.0 TPa. Generally, superconductors are metallically bonded, but super-hard materials are covalently bonded with strong orientations and large bond energies. -Superconductivity and super-hardness existing in the same structure ($C2/m$) of ice is a very interesting phenomenon that is worth further in-depth studies with the progress of QM methods.

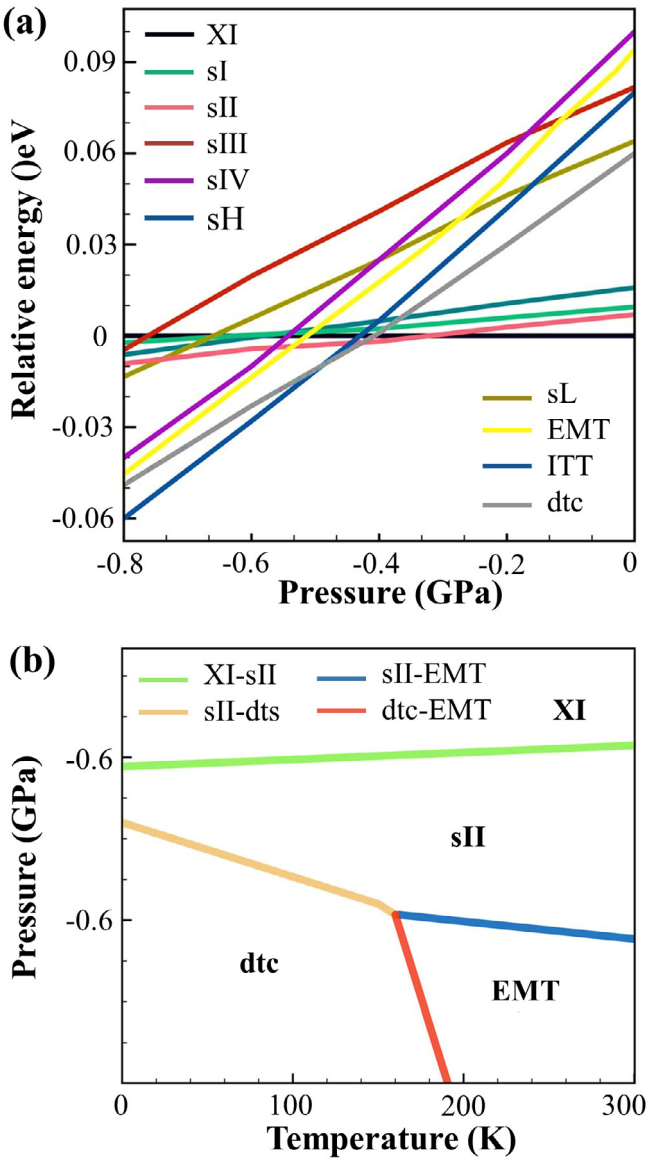

Fig. 3.7. Predicted phase diagram of clathrate ices at negative pressures in the range of −0.8 GPa to −0.1 GPa. (a) Calculated relative enthalpies (referenced to the enthalpy of ice XI) of clathrate ices sH, sL, sII, sIII, sIV, EMT, ITT and dtc depending on pressure. (b) Calculated phase diagram of ice XI and clathrate ices sII, dtc and EMT.
Source: Reproduced from the work by Lu et al. [20].

3.4. ML potential energy surfaces for small molecules

As discussed in Section 2.3, a series of ML methods for constructing PESs have been proposed, such as HDNNP [184], ANI [185], TensorMol [186], DeepMD [187], TorchANI [189], AP-Net [190] and PIP [188]. ML PESs have become a significantly important tool for accelerating QM investigations in computational physics and chemistry, especially for molecular crystals. In recent years, with the great abundance of data and the continuous development of ML methods and chemical descriptors, ML PESs and force fields have been widely used for predicting energy-related properties or reactions for various inorganic small molecules, such as vibrational energies, H-bonds, spectra, transmission coefficients, rate coefficients of reactions, and excitation frequencies.

In 2006, Manzhos et al. [229] successfully fitted NN PESs for several three- and four-atom molecules, including H_2O, HOOH and H_2CO, with root-mean-square errors (RMSEs) less than 2 cm^{-1}. The low-lying levels have 1 cm^{-1} of the exact results for all molecules, which is achieved with multiple NNs by first fitting a rough NN PES and then fitting an NN to

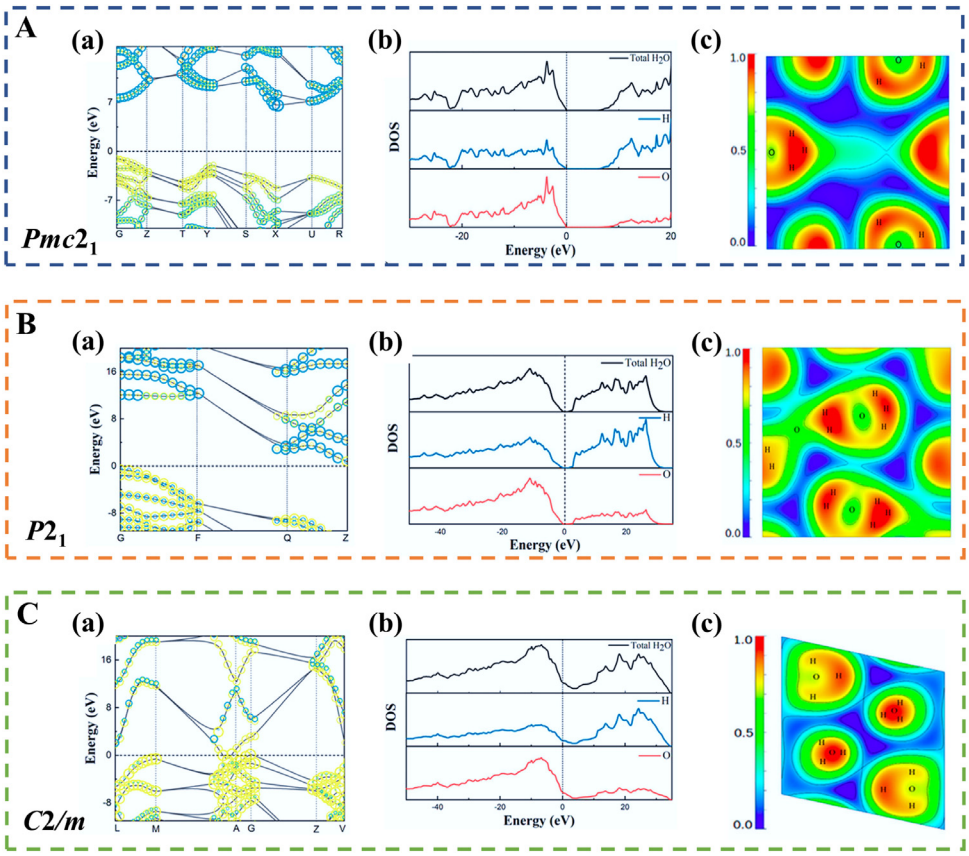

Fig. 3.8. Calculated electronic band structures and density of states (DOSs) at the Fermi level, and electron localization functions (ELFs) of ice at terapascal pressures. A. shows the (a) electronic band structures, (b) DOSs and (c) the ELFs with (010) plane of the $Pmc2_1$ structure (at 1 TPa). B. shows the (a) electronic band structures, (b) DOSs and (c) the ELFs with (001) plane of the $P2_1$ structure (at 4 TPa). C shows the (a) electronic band structures, (b) DOSs and (c) the ELFs with (001) plane of the *C*2/*m* structure (at 5 TPa).
Source: Reproduced from the work of Lu et al. [21].

the energetic difference at fitting points and values on the rough PES. This small PES is considered to be the beginning for multiple NNs being applied to PES constructions. Although the accuracy of the potential surface fitting can be improved by increasing the number of neurons, this may lead to overfitting. In a multidimensional space, this potential energy surface accuracy is usually not representative of the overall accuracy. As shown in Manzhos's work, the adoption of multiple NNs is an effective method for addressing this problem.

In 2017, using 638 data points calculated with CCSD(T)/cc-pVQZ(-g), Pradhan and Brown [230] fitted a six-dimensional (6D) NN PES of HONO in a sum-of-products form, which contains the cis- and trans-isomers and the related transition states. They also proposed another PES based on QM data calculated using the CCSD(T) approach with the complete basis set (CBS) correction. For this model, the vibrational energies were used for model validation, which were then calculated by the block improved relaxation combined with the multiconfiguration time dependent Hartree (MCTDH) method, which were compared with the experimental results, and previous calculations with the MP2/aug-cc-pVTZ and CCSD(T)/aug-cc-pVTZ levels of accuracy. In the same year, Guan et al. [231] reported an NN model for solving the mixing angle for the diabatization of the lowest two electronic states of LiFH based on *ab initio* adiabatic energies and derivative couplings, which is only suitable for two electronic states without crossing. Two years later, Guan et al. [232] introduced a modified NN model for predicting adiabatic energies, energy gradients and derivative couplings accurately based on the Zhu–Yarkony diabatization strategy, which is potentially applicable for multiple-state diabatization with avoided crossings. In addition, in 2018 Yuan et al. [233] also proposed accurate derivative-based diabatic NN PESs for the H_3 system by solving the three-dimensional Poisson equation. In 2019, Yin et al. [228] developed a procedure for fitting coupled two-state diabatic PESs with conical intersections based on nonadiabatic couplings using a general NN. The outputs of NN PESs and the diabatic potential energy matrix (DPEM) are calculated by first solving the adiabatic-to-diabatic transformation (ADT) angle for each geometry and then fitting individual NNs for the three terms of the DPEM. Then, they used this procedure to construct the two-state diabatic PES of ClH2, which presents a high fitting accuracy and a reliable representation of

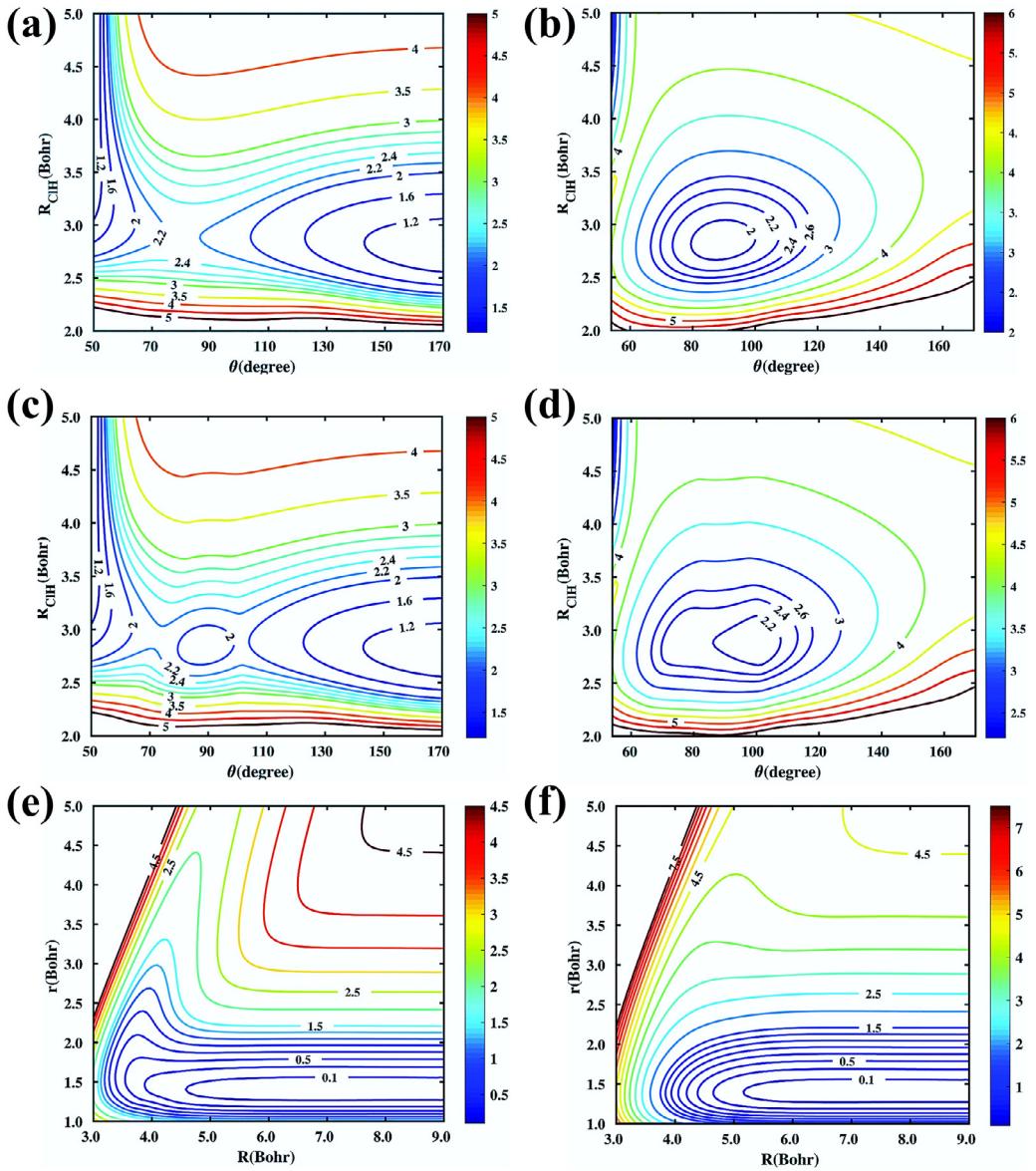

Fig. 3.9. Contour plots of the fitted diabatic PESs and adiabatic PESs. (a) (b) The diabatic PESs (W_{11}, and W_{11}) and (c) (d) the adiabatic PESs as functions of R_{ClH} and the enclosed angle of the two equal R_{ClH} in C_{2v} geometries. (e) (f) The diabatic PESs as functions of Jacobi R and r in the collinear geometries.
Source: Reproduced from Yin et al.'s work [228].

the vicinity of conical intersections. Fig. 3.9 shows the fitted diabatic PESs and adiabatic PESs (W_{11} and W_{11}) as functions of R_{ClH} and the enclosed angle of the two equal R_{ClH} values in C_{2v} geometries as well as the diabatic PESs as functions of Jacobi R and r in collinear geometries, which is reproduced from Yin's work.

In 2019, Sauceda et al. [234] introduced the construction of molecular force fields for small molecules containing no more than 25 atoms using the symmetrized gradient domain machine learning (sGDML) [167,235] model, which is a fully data-driven universal approximator for describing any kind of quantum interaction. High-dimensional manifolds in the training dataset are obtained from a few samples, which allows for calculating these data with high-level QM methods such as CCSD(T). The accuracy of these force fields was evaluated by ubiquitous and challenging features of general interest in chemical physics, including intramolecular hydrogen bonds, proton tunneling effects, electrostatic interactions, electron lone pairs, and other electronic effects. An overview of different types of molecules and the corresponding PESs along

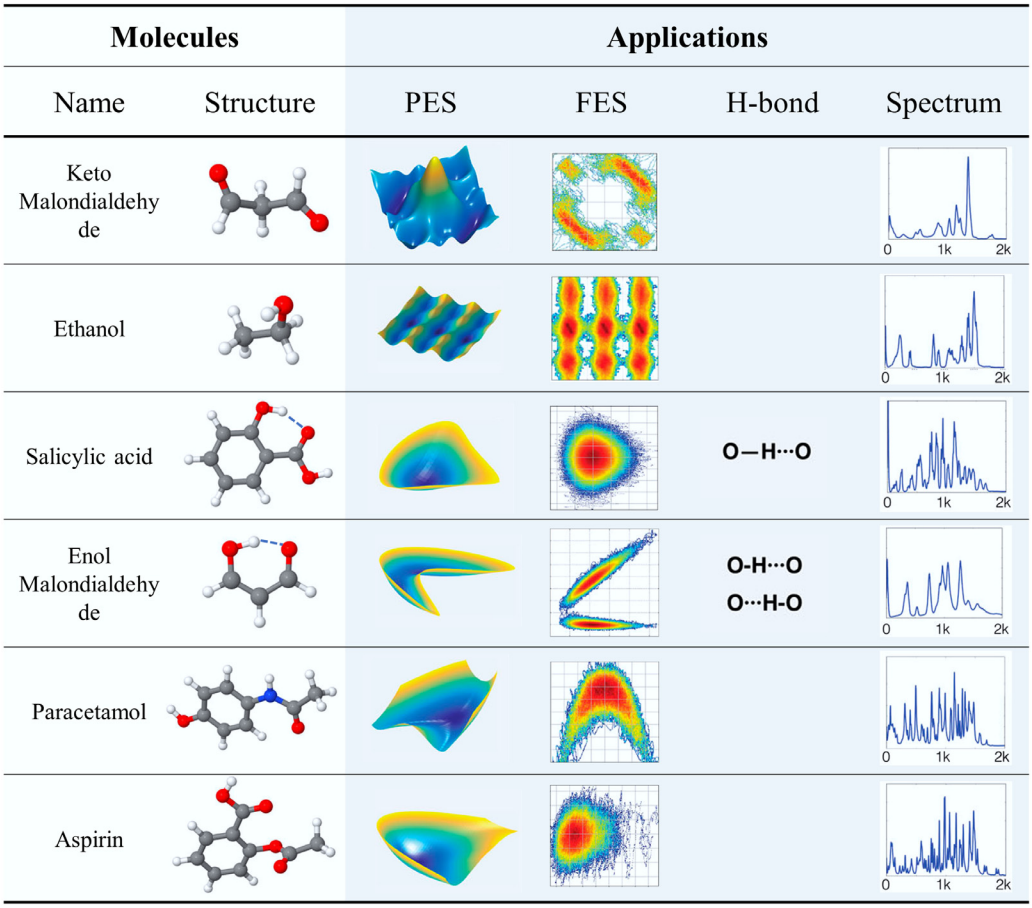

Fig. 3.10. An overview of different types of molecules and corresponding PESs along two relevant torsional degrees of freedom, free energy surface (FES) at 300 K, type of intramolecular hydrogen bonds, vibrational spectrum at 300 K, and type of electronic effect.
Source: Reproduced from Sauceda et al.'s work [234].

two relevant torsional degrees of freedom, free energy surface (FES) at 300 K, type of intramolecular hydrogen bonds, vibrational spectrum at 300 K, and the type of electronic effect are shown in Fig. 3.10.

In addition, NN PESs can also be applied to the MD study of chemical reactions. For example, using the permutationally invariant polynomial NN (PIP-NN) method, Liu and Li [236] (2019) proposed a globally accurate full-dimensional PES for the reaction $Cl + CH_4 \rightarrow HCl + CH_3$ based on 74000 QM data with a CCSD(T) level of accuracy. To evaluate the performance of this PES model, kinetic isotope effects (KIEs) and thermal rate coefficients were calculated for the $Cl + CH_4 \rightarrow HCl + CH_3$ and $Cl + CD_4 \rightarrow DCl + CD_3$ reactions using the ring polymer MD (RPMD) approach with the proposed PES, which is consistent with the experimental results. Fig. 3.11 shows the transmission coefficients (Fig. 3.11(a) (c)) and the rate coefficients (Fig. 3.11(b) (d))of reactions $Cl + CH_4 \rightarrow HCl + CH_3$ and $Cl + CD_4 \rightarrow DCl + CD_3$.

Apart from the widely used NN approach, other ML methods have also been applied for constructing PESs. For example, the GPR algorithm combined with the adaptive density guided approach (ADGA) has been used for PES construction based on physical information and statistical analysis of the data at the MP2/CCSD(T) level. Schmitz et al. [237] evaluated the performance of this GPR-ADGA model on nine three- and four-atomic molecules, such as H_2O, H_2CO, and H_2S, and one five-atomic molecule (formic acid). The fundamental excitation frequency calculations on conventional and GPR-ADGA PES show good agreement with the RMSD below 2 cm^{-1}. In 2018, Hu et al. [238] constructed an on-the-fly PES using KRR, and Dral et al. [239] also used KRR for excited state dynamics involving surface hopping.

As we discussed above, ML methods have been a powerful tool for constructing PESs and further chemical studies of small molecules [240]. We cannot list all applications of various ML methods in PES construction in this review, but only present some representative applications in recent years, as shown in Table 3.1. Other successful applications include PIP-GPR constructing high-dimensional PESs [241], Behler–Parrinello NNs (BP-NNs) for predicting interaction energy with symmetry adapted perturbation theory (SAPT) [242], GPR reconstructing free energy surface (FES) [243], NNs fitting PESs for studying Diels–Alder reaction [244], fundamental invariants (FI) NNs constructing PES for the H2 +

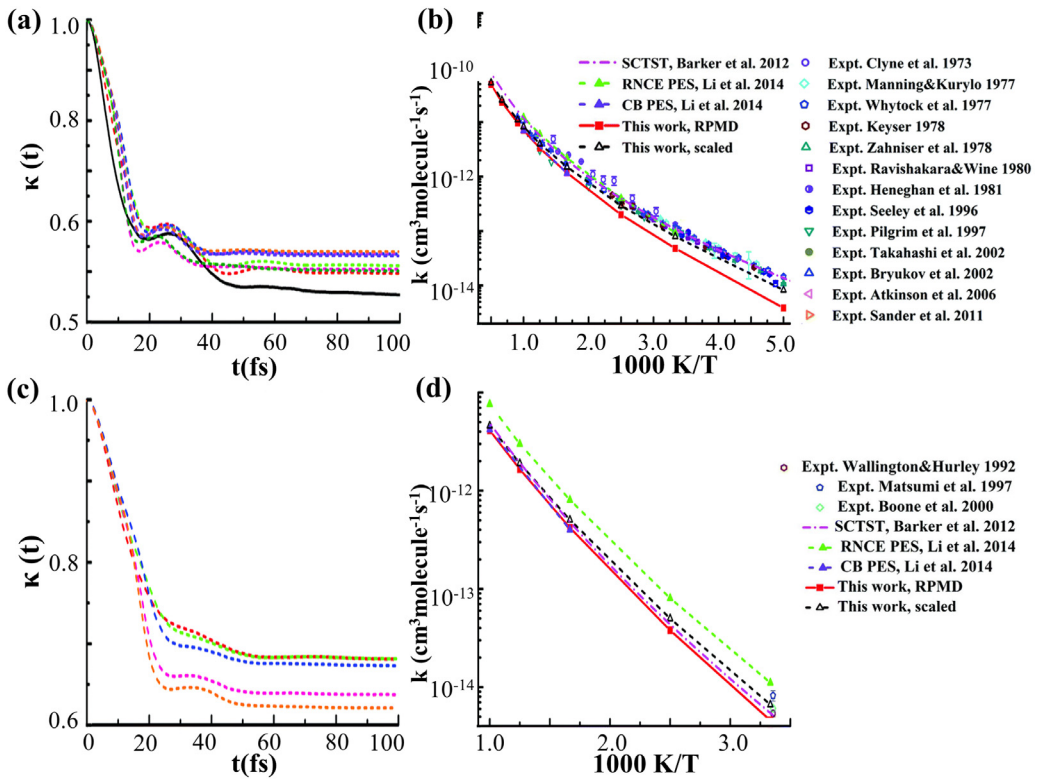

Fig. 3.11. Transmission coefficients at different temperatures and rate coefficients for different reactions. (a) The transmission coefficients (at the temperature conditions of 200, 300, 400, 600, 800, 1000, 1100, 1500, 2000 K, respectively), and (b) the rate coefficients for the reaction of Cl + CH_4 → HCl + CH_3. (c) The transmission coefficients (at the temperature conditions of 200, 300, 400, 600, 800, 1000, 1100, 1500, 2000 K, respectively), and (d) the rate coefficients for the reaction of Cl + CD_4 → DCl + CD_3.
Source: Reproduced from the work by Liu and Li [236].

HS → H2S + H reaction [245], high-dimensional NN PESs for CO_2–Pt(111) interaction [246], full-dimensional NN PESs for the rate coefficients of reactions [247], and GPR constructed PESs for free energy landscape and reconstructive phase transition [248]. With the rapid development of ML methods, PES construction via ML has become a popular research area promoting chemical research, with potential applications in predicting the potential energy, free energy, vibrational energy, diabatic potential energy, thermal rate coefficients and KIEs in chemical reactions and intermolecular interaction energy.

On the other hand, using an ML model to train a PES of a system usually requires numerous and high-quality chemical data either calculated by high-level QM methods or obtained by experiments. The dataset should cover the relevant pressure and temperature conditions, compositions, polymorphs, and reaction pathways for full PES construction [46,49, 249–251]. In addition, much attention needs to be paid to avoiding numerical noise, which will influence the smoothness of PES and result in discontinuities [252–254]. Most importantly, ML PESs usually suffer restrictions on the time scale and system size, which significantly limits their application to large systems. Nevertheless, the continuous development of ML methods and descriptors will advance the applications of PESs in computational physics and chemistry.

3.5. ML accelerates the QM predictions of properties and phase transitions

In Section 3.4, we discussed the PESs constructed using various ML methods, such as GPR, sGDML NN and related variants, for different usages. ML PESs are usually designed for small molecules and clusters. In recent years, apart from PESs, some ML applications include the prediction of properties, such as energies and structures, with previous experimental or computational results feeding in the models.

In 2019, Tawfik et al. [255] introduced complementary ML methods combined with property-labeled material fragment descriptors to rapidly and reliably predict the thermal properties of crystals, such as entropy, specific heat, effective polycrystalline dielectric function and nonvibrational properties. In practice, the RF, SVM [256], the relevance vector machine (RVM) [257], the Huber regression algorithm [258], the XGBOOST algorithm [259], and the feed-forward back-propagation NN [260] are used for the prediction. As Table 3.2 shows, the ML methods exhibit good performance for

Table 3.1
List of notable PESs constructed with ML methods, including the author and year of first report, research system, type of methods, and applications of the PESs.

Year	Systems	Method	Applications
Manzhos et al. [229] (2006)	H_2O, HOOH, and H_2CO	6-D NNs	Potential energy and vibrational spectra
Pradhan and Brown [230] (2017)	HONO	6-D NNs	Vibrational energies and frequencies
Guan et al. [231] (2017)	LiFH	NNs	Diabatic PESs and mixing angles
Yuan et al. [233] (2018)	Reaction of $H+HD \rightarrow H_2+D$	NNs	Total energy and geometric phase (GP)
Guan et al. [232] (2019)	LiFH	NNs	Adiabatic energies, energy gradients and interstate couplings
Liu and Li [236] (2019)	Reactions of $Cl+CH_4 \rightarrow HCl+CH_3$ and $Cl+CD_4 \rightarrow DCl+CD_3$	PIP-NN	Thermal rate coefficients and kinetic isotope effects
Yin et al. [228] (2019)	ClH_2	NNs	Two-state diabatic PESs, adiabatic energies and derivative couplings
Sauceda et al. [234] (2019)	small molecules containing less than 25 atoms	sGDML	PES, FES, H-bond, spectrum and electronic effects
Schmitz et al. [237] (2019)	Nine three- and four-atomic molecules	GPR-ADGA	Excitation frequencies
Metcalf et al. [242] (2020)	Hydrogen-bonded dimers	BP-NNs	Interaction energies
del Cueto et al. [246] (2020)	CO_2–Pt (111) interaction	BP-NNs	Geometry and energy
Zuo et al. [247] (2020)	Reactions of H+O3 and HO2+O	PIP-NN	Thermal rate coefficients
Tong et al. [248] (2021)	GaN	GPR	Free energy landscape and reconstructive phase transition

Table 3.2
The mean absolute relative errors (MAREs), RMSEs and the coefficient of determination (R^2) for the specific heat (C_V) in units of meV/K/atom, the entropy (S) in units of meV/K/atom, and the trace of the effective polycrystalline dielectric function (ε_{eff}) in different ML models.
Source: Reproduced from Ref. [255].

ML model	Dataset	C_V			S			ε_{eff}		
		MARE (%)	RMSE	R^2	MARE (%)	RMSE	R^2	MARE (%)	RMSE	R^2
RF	Test	10.9	0.19	0.88	18.2	0.39	0.85	17.1	4.85	0.66
	Train	4.6	0.09	0.97	6.5	0.16	0.97	7.1	2.46	0.93
SVM	Test	14.4	0.24	0.81	15.7	0.36	0.87	18.4	4.70	0.68
	Train	8.3	0.11	0.95	9.1	0.17	0.97	9.1	3.40	0.86
RVM	Test	12.4	0.22	0.84	13.8	0.32	0.90	17.6	4.57	0.70
	Train	2.9	0.04	0.99	3.4	0.08	0.99	7.6	1.58	0.97
Huber	Test	13.1	0.23	0.83	13.2	0.34	0.89	17.9	5.35	0.58
	Train	8.1	0.16	0.90	7.5	0.23	0.94	11.6	4.33	0.77
XGBOOST	Test	11.1	0.18	0.89	14.9	0.33	0.89	16.8	4.87	0.65
	Train	0.0	0.00	1.00	0.0	0.00	1.00	0.0	0.01	1.00
NN	Test	12.4	0.20	0.88	14.5	0.31	0.90	17.9	4.41	0.72
	Train	0.8	0.06	0.99	1.1	0.09	0.99	1.2	0.89	0.99

predicting the QM properties with RMSEs of 0.31 meV/K/atom for entropy, 0.18 meV/K/atom for specific heat, 0.5 eV for the band gap and 4.41 for the trace of the dielectric tensor, which shows potential applications for screening a large number of crystals for properties of interest. Very recently, Loftis et al. [261] (2021) proposed a genetic programming-based symbolic regression (SR) model to predict the lattice thermal conductivity (κ_L), which shows a higher accuracy than the traditional Slack formula [262]. Compared with PESs, these direct predictive ML models are more straightforward and easier to understand, and even have a better performance for a particular property but are usually less transferable.

On the other hand, as discussed in Section 3.1 the crystal structures, properties, and phase diagram of solid crystals under extreme conditions have been a topic of extensive study for decades. Substantial progress has been made on the predictions of solid–solid phase transitions with QM methods and fragment-based QM methods. However, the extreme computational cost greatly restricted the fast in-depth determination of the crystal phase diagram. In a general fragmental scheme, for accuracy consideration the low-order (such as one- and two-order) terms need to be calculated using high-level electronic methods, such as MP2 and CCSD(T), which is the main reason for the extreme computational cost. Motivated by this problem, NN PESs with an accuracy at the level of MP2 or CCSD(T) are introduced to the rapid calculations of low-order terms in the fragment-based method for predicting large crystalline systems.

In 2020, Han et al. [183] introduced a newly efficient approach based on NN PESs and an improved fragmental method to predict the Gibbs free energy, structural parameters, EOS, and thus the phase transition of molecular crystals. Fig. 3.12

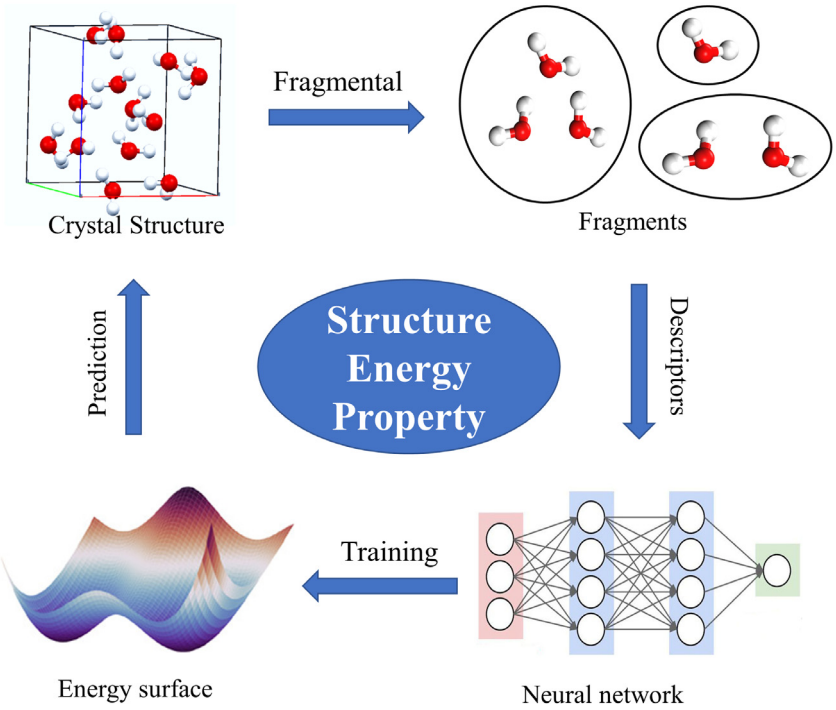

Fig. 3.12. Flow diagram of the approach combining NN PES and fragment-based method for the prediction of energy and various properties of molecular crystals.
Source: Reproduced from Han et al.'s work [183].

shows the flow diagram of the proposed approach. The PESs are trained using DeepMD based on data of one- and two-body energy and forces calculated at the MP2 or CCSD(T) level. Then, these PESs are integrated into an improved fragment-based model that can be used to calculate the energy, force, and other properties of a molecular crystal.

Then, this approach is evaluated by performing the predictions and structure optimizations of ice at a high pressure. The PESs of one- and two-body H_2O were constructed based on the data calculated at the MP2 level of accuracy. These NN PESs exhibited considerably low RMSEs (0.0029 kcal/mol for one-body energy, 0.0353 kcal/mol/Å for one-body force, 0.0429 for two-body energy and 0.9782 for two-body force), which provides assurance of the accuracy of subsequent predictions by the fragment-based method. The one- and two-body PESs were then integrated into the fragmental method, in which three-body terms were obtained using low-level QM models. The structural optimization and prediction of the Gibbs free energy, lattice parameters, EOS, covalent bonds, and phase transition of solid ice IX and XV present a good agreement with previous experimental results, with RMSEs of 0.55 and 0.32 kcal/mol for the two ice crystal systems. More importantly, the greatest advantages of this approach are the computational efficiency, which is shown in Table 3.3 As the table shows, the prediction of the phase transition for ice crystals can reach the MP2 level of accuracy but is tens of times faster than fragment-based QM methods. Notably, the restriction of computational efficiency mainly comes from the low-level QM calculations of the three-body terms, providing ideas for further improvements by integrating three-body PES. In addition, a newly designed fragment-based method suitable for PESs will be helpful for increasing the accuracy and efficiency. In short, a new way of predicting the phase transition and properties of large molecular crystals using a combination of NN PESs and fragment-based methods is opened.

4. Predictions of drug molecules and crystals

Energy and structure determination of polymorphs of pharmaceutical molecules are considerable significant in the chemical and pharmaceutical industries. However, because drug molecules are usually larger than small inorganic molecules, it is difficult to use all-atomic QM methods. Fragment-based QM methods can divide large crystal systems into a series of small fragments and can be easily applied to large systems, such as drug crystals with a high precision and efficiency, including olanzapine, sulfathiazole, di-p-tolyl disulfide, and β-lactam antibiotics. Fragment-based QM methods have been successfully applied to predicting crystal structures, EOSs, vibrational spectra and phase transitions of drug molecules and crystals. In recent years, ML methods have also been widely applied to the prediction of drug molecules and crystals, such as force field construction, chemical shift prediction and CSP. In this section we will introduce the structure,

Table 3.3
Computational costs (the average of three calculations) for monomers, dimers, enthalpy (energy, force) and free energy (energy, force and force constant) of ices IX and XV.
Source: Reproduced from Han et al.'s work [183].

Calculations	MP2 (s)	NN-PES (s)	Times
100 monomers[a]	228	0.8	285
1000 monomers[a]	1978	2.2	899
100 monomers[b]	3305	27.6	120
100 dimers[a]	503	1.7	296
1000 dimers[a]	5073	4.5	1127
100 dimers[b]	11971	193	62.0
Enthalpy[a] (IX)	3066	439	7.0
Enthalpy[a] (XV)	2075	324	6.4
Gibbs free energy[b] (IX)	36500	2241	16.3
Gibbs Free energy[b] (XV)	22828	1621	14.1

[a]Only calculate energy and force.
[b]Calculate energy, force and force constant.

Table 4.1
Crystal structural parameters of olanzapine form I, II, III and IV from experiments and calculations. All computations were performed on DFT level of ωB97XD/6-31G*.

Form	Method	Temp. (K)	a (Å)	b (Å)	c (Å)	Beta (deg)
Form I	Experiment (Reutzel–Edens et al. 2003) [274]	293	10.383	14.826	10.560	100.616
Form I	DFT (Luo et al. 2019) [276]	298	10.056	14.702	10.426	
Form I	DFT (Tang et al. 2021) [277]	298	10.008	14.402	10.420	
Form II	Experiment (Thakuria et al. 2011) [278]	298	9.913	16.533	9.999	98.023
Form II	DFT (Luo et al. 2019) [276]	298	9.737	16.399	9.861	
Form II	DFT (Tang et al. 2021) [277]	298	9.685	16.314	9.810	
Form III	Modeled (Bhardwaj et al. 2013) [279]		10.3454	19.5267	16.5280	90
Form III	DFT (Tang et al. 2021) [277]	298	9.97508	19.0714	16.1121	
Form IV	Experiment (Askin et al. 2019) [280]		8.6555	15.4441	2.5558	95.284
Form IV	DFT (Tang et al. 2021) [277]	298	8.37334	15.0032	11.9860	

stability, vibrational spectra and transformation predictions of several drug molecules using fragment-based QM and the recent applications of ML methods in large molecules and relevant crystals, including chemical shift predictions, force field constructions, and CSP.

4.1. Structure, stability, and spectra predictions

Under different physical and chemical conditions, a crystal with a certain chemical composition will crystallize to various forms, which is known as polymorphism. The polymorphism of drug molecules usually results in extremely different chemical and physical properties, including stability, solubility, compressibility and flexibility, which can influence the clinical efficacy [263–266] and safety of the related drugs. During drug design and production, investigating the polymorphism and determining the structural stability of different polymorphs of drugs were considered to be of great importance. Substantial progress has been made in CSP, crystal growth [267–270] and property prediction of polymorphisms [271–273].

Olanzapine

For years, substantial attention has been given to polymorphs of olanzapine, a widely used drug molecule with great efficacy in the treatment of schizophrenia, which is a serious mental illness that afflicts patients for life. Approximately two decades prior, the crystal structural parameters of olanzapine form-I were introduced [274]. Based on QM methods and fragment-based methods, considerable efforts have been made for crystal structure prediction, polymorph transformation, crystal growth and property prediction [268,275–277].

Approximately two decades prior, the crystal structural parameters of olanzapine form I were introduced [274]. The crystal structural parameters of form II and form III were reported [278] and modeled [279], respectively, early in the preceding decade [279], two years before Askin et al. reported crystal structural parameters of olanzapine form IV [280]. The observed crystal structural parameters of these four forms are given in Table 4.1. The calculated crystal structure parameters given in Table 4.1 were optimized by the DFT/ωB97XD/6-31G* level. The experimentally observed crystal structure parameters of olanzapine are well matched with those calculated by the DFT/ωB97XD/6-31G* level given in Table 4.1.

The vibrational spectra are considered a fingerprint to recognize the crystal structure of molecules. In polymorphic molecules, different polymorphic forms have different Raman spectra; hence, each form can be identified by their characteristic peaks [276,277]. Fig. 4.1A shows the calculated and experimentally observed Raman spectra of olanzapine

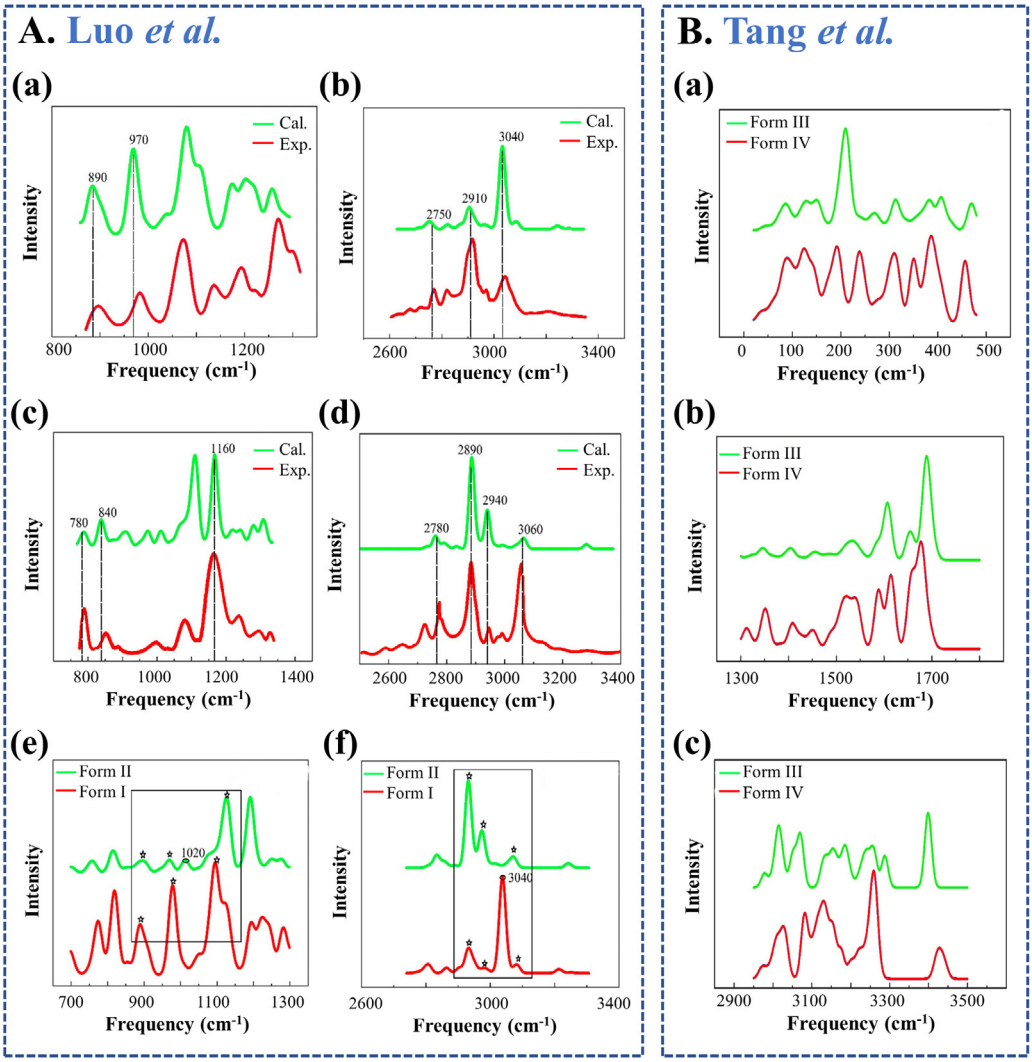

Fig. 4.1. Calculated Raman spectra of olanzapine forms at atmospheric pressure condition. A. (a–f) shows the calculated and observed Raman spectra of (a) (b) form I and (c) (d) form II, and (e) (f) the comparison of calculated Raman spectra of olanzapine forms I and form II (© 2019 American Physical Society). B. shows the calculated Raman spectra of olanzapine polymorphs III and IV at different regions of frequency, which were reproduced from the paper of Tang et al. [277] (© 2021 Elsevier).
Source: A. Reproduced from the research papers of Luo et al. [276].

forms I and II at atmospheric pressure conditions from Luo et al.'s work [276]. The Raman spectra were calculated via the DFT/ωB97XD/6-31G* level, and the experimental results were taken from Bhardwaj et al.'s work [279]. In Fig. 4.1A, the curves of the calculated Raman spectra of olanzapine forms I and II are colored red and green, respectively, and the curves of Raman spectra from experimental observation of olanzapine forms I and II are colored blue and yellow, respectively. As shown in Fig. 4.1A, the calculated Raman spectra match well with the experimentally determined Raman spectra for both forms I and II. Fig. 4.1A(a) presents the Raman spectra of olanzapine form I in the region of medium frequency. We can observe that characteristic peaks from the experimental results are accurately predicted at 890 cm^{-1} and 970 cm^{-1}, as shown in Fig. 4.1A(a). Fig. 4.1A(b) presents the Raman spectra of olanzapine form I at high frequencies. The characteristic peaks from the experimental results are accurately predicted at 2750 cm^{-1}, 2910 cm^{-1}, and 3040 cm^{-1}, as shown in Fig. 4.1A(b).

Next, we will study the Raman spectra of olanzapine polymorphs III and IV. To date, the Raman spectra of olanzapine polymorphs III and IV have not been experimentally observed. A similar method (Luo's paper [276]) was used to calculate the Raman spectra of olanzapine polymorphs III and IV. The calculated Raman spectra at different frequency regions can be used as a further reference, which is shown in Fig. 4.1B. As shown in Fig. 4.1B(a), the Raman spectra of polymorphs III and IV have eight and nine peaks, respectively, at low frequencies. In Fig. 4.1B(b), polymorphs III and IV have eight and seven peaks, respectively, at medium frequencies. Notably, the Raman spectra of polymorphs III and IV present obvious

Table 4.2
Calculated and observed crystal structural parameters of sulfathiazole forms I, II, III, IV and V. All calculations were performed on DFT level of ωB97XD/6-31G*.

Form	Method	Temp. (K)	a (Å)	b (Å)	c (Å)	CCDC Ref code
Form I	Experiment		10.554	13.220	17.050	SUTHAZ01
Form I	DFT /B97XD/6-31G* [281]	300	10.355	13.185	16.842	
Form II	Experiment		8.235	8.550	15.558	SUTHAZ
Form II	DFT /B97XD/6-31G* [281]	300	8.119	8.419	15.472	
Form III	Experiment		17.570	8.574	15.583	SUTHAZ02
Form III	DFT /B97XD/6-31G* [281]	300	17.333	8.306	15.550	
Form IV	Experiment		10.867	8.543	11.456	SUTHAZ04
Form IV	DFT /B97XD/6-31G* [281]	300	10.643	8.251	11.398	
Form V	Experimental		14.330	15.273	10.443	SUTHAZ06
Form V	DFT /B97XD/6-31G* [281]	300	14.096	14.854	10.319	

differences in the region of high frequencies shown in Fig. 4.1B, in which polymorph III presents five peaks and polymorph IV has seven peaks. In Raman spectroscopy, differences in the number and position of peaks will be fundamental concepts for identifying a particular form.

Sulfathiazole

In 1939, the organosulfur compound sulfathiazole (also known as N'-2-thiazolylsulphanilamide) was determined and has become a typical representative bacteriostatic drug and one of the most effective sulfonamides. Sulfathiazole has shown polymorphism with five polymorphs. The crystal structure and stability of these polymorphs have been researched for quite some time. Recently, fragment-based QM methods have been applied to the prediction of sulfathiazole polymorphs [281], such as stability determination and property prediction. Sulfathiazole, a compound that forms five known crystal structures, has been examined to understand its different polymorphs and their stabilities. The main differences between the polymorphs of sulfathiazole lie in the bonding of hydrogen and its effects on the arrangement of the molecules in the crystal structure [282]. A comparison of the observed and calculated crystal structure parameters of five different polymorphs of sulfathiazole is presented in Table 4.2, which shows a good consistency between the optimized crystal structure parameters and the experimentally observed results. The deviation in the average crystal structure parameters between the calculated and observed values is approximately 0.1 to 0.3 Å.

Hao et al. computed the Raman spectra of five forms of sulfathiazole by using the DFT level of ωB97XD/6-31G* [281]. Additionally, they compared their computational Raman spectra with the experimentally observed results performed by Munroe et al. [283]. Hao et al. noted the computationally obtained Raman spectra and compared them with the experimental results given in Fig. 4.2. From 1100 cm^{-1} to 1500 cm^{-1} in Fig. 4.2, form I and form II both have six distinct Raman peaks, while form III, form IV and form V present five different Raman peaks. Fig. 4.2 presents the positions of Raman peaks for the calculation, which indicate a minor difference in the position when compared with the observed position. The calculated Raman frequencies are slightly larger than the observed spectra, which mainly results from the harmonic approximation in the implemented computational approach that ignores the nonharmonic effect. Generally, the Raman spectra from theoretical calculations match well with the experimentally observed results of Munroe et al. [283], which establishes the correctness of their calculation. Moreover, the computational Raman spectra given in Fig. 4.2 for form I and form III are also comparable with the experimental work done by Hu et al. [284].

To determine the stability of different polymorphs, Hao et al. calculated the Gibbs free energy and compared their differences for the five sulfathiazole forms using the embedded fragmental approach at the DFT/B97XD/6-31G* level at a standard atmospheric pressure and 300 K [281]. They found that at 300 K, the calculated order of stability based on the Gibbs free energy is form I < form V < form IV < form II < form III [281]. Generally, a polymorph with the lowest Gibbs free energy is considered the most stable form, indicating a clear correlation between the structural stability and Gibbs free energy. In that case, form III is the most stable crystal of the sulfathiazole polymorphs at 300 K. Munroe et al. studied the relative structural stabilities of the five sulfathiazole polymorphs using solid-state and solution-based methods [283]. Based on a combination of solubility measurements and isothermal suspension equilibration experiments, they identified that in the temperature range of 10–50 °C form I and form V are less stable than form II, form III and form IV, with only small differences in stability among the latter three forms [283]. Based on the various experimental results, they proposed that in the lower temperature range of 10–50 ° C the stability order of sulfathiazole is form I < form V < form IV < form II < form III [283], where this experimentally observed stability order of sulfathiazole matches well with the computational results given above proposed by Hao et al. Moreover, their stability computational results also match well with the experimental results [282,283,285–287].

Di-p-tolyl disulfide

With the rapid growth of energy demand and the depletion of existing energy resources, new materials with superior performances, low costs and environmental friendliness for energy production and storage are being explored. Di-p-tolyl disulfide is a typical lubricating material that has been used in the field of energy storage. The crystal structures, conformational properties and phase transformations of di-p-tolyl disulfide have been investigated for years with fragment-based DFT and MP2 methods. In this section we will review the QM studies on di-p-tolyl disulfide for crystal

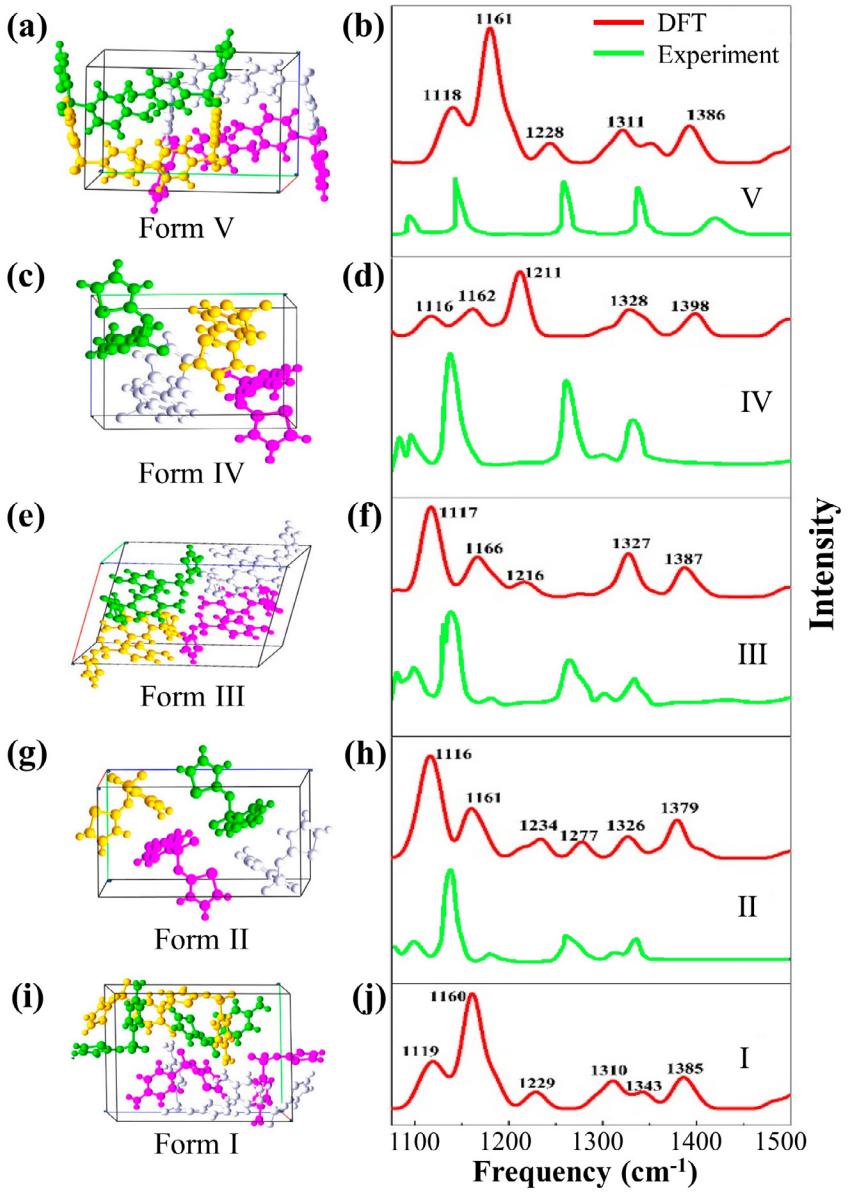

Fig. 4.2. Crystal structures and calculated and observed Raman spectra of sulfathiazole (a) (b) form V, (c) (d) form IV, (e) (f) form III, (g) (h) form II and (i) (j) form I, where red curves represent calculated Raman spectra and black color represent observed Raman spectra.
Source: These figures were taken from Hao's research paper [281]. The observed Raman spectra were from Munroe et al.'s work [283].

structure screening, prediction and optimization, stable structure determination and property (such as Raman and IR spectra) predictions [288].

Hao et al. [288] used the different crystal structure prediction tools of MOLPAK [289,290] (MOLecular PAcKing) and USPEX [291–293] to investigate crystal structure screening, prediction and optimization of di-p-tolyl disulfide. In total, 9000+ possible hypothetical crystal structures [288] were produced via USPEX and MOLPAK. They selected 38 structures (18 from USPEX prediction and 20 from MOLPAK prediction) with the lowest lattice energies for further optimization, where crystal structure optimization was performed with the DFT/ωB97XD/6-31G* theory, and the single-point energy was obtained from MP2 calculations based on the stable crystal structures after optimization. In their study, MP2 theory was adopted to perform the calculations [288] more accurately. They then employed Gibbs free energy, which is more accurate than traditional lattice energy calculations [288]. Gibbs free energy includes the contribution of entropy and temperature in the energy calculations. By comparing the Gibbs free energies between the predicted and experimental structures, they found that phase α is the most stable structure for di-p-tolyl disulfide crystals at an ambient temperature

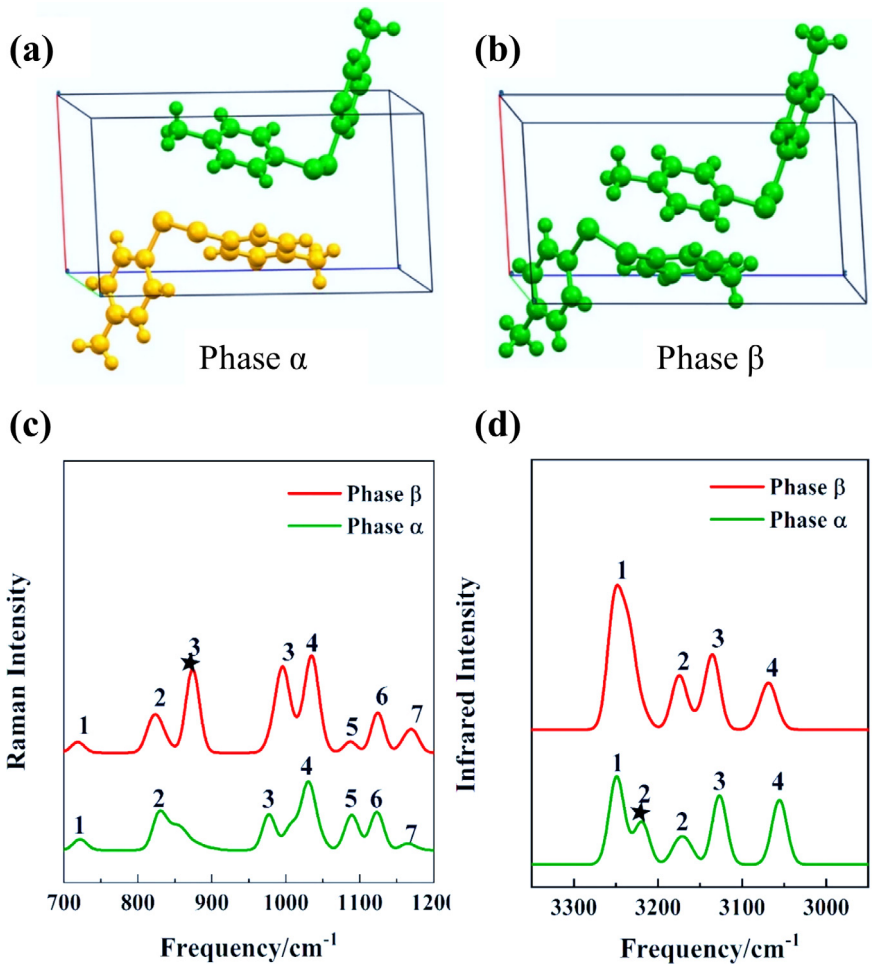

Fig. 4.3. Crystal structures and vibrational spectroscopy of di-p-tolyl disulfide. (a) and (b) represent the crystal structure of phase α and β, (c) and (d) represent the Raman and IR spectra, respectively. The curve with green and red color represents phase α and phase β, respectively.
Source: Reproduced from Hao et al.' work. [288].

and standard atmospheric pressure. The crystal structures of di-p-tolyl disulfide phases α and β are shown in Fig. 4.3(a) (b), respectively.

Generally, vibrational spectroscopy [294,295] can be used to identify the crystal structure of a certain molecule. To determine the differences between di-p-tolyl disulfide phases α and β, Hao et al. [288] presented the IR and Raman spectra from the theoretical calculations of di-p-tolyl disulfide crystals. The IR and Raman spectra are presented in Fig. 4.3(c) and (d), respectively. In Fig. 4.3(c), the Raman spectra are presented in the frequency region of 700 to 1200 cm^{-1}, which shows that seven and eight remarkable Raman peaks can be identified from the spectra of phase α (colored in green) and phase β (colored in red), respectively. For di-p-tolyl disulfide phase β (colored in red), the Raman peak 3* ($\sim$874 cm^{-1}) is the domain characteristic peak. In addition, in Fig. 4.3(d), five and four IR peaks for phase α (colored in green) and phase β (colored in red) can be identified, respectively. IR peak 2*, with a frequency of $\sim$3220 cm^{-1}, is a domain characteristic peak for phase α (colored in green). In summary, the di-p-tolyl disulfide phases α and β can be distinguished by Raman spectra with peak 3* and IR spectra with peak 2*.

β-lactams

As a class of antibiotics, β-lactams contain all antibiotic agents with β-lactam rings in their molecular structures. They have the advantages of a strong bactericidal activity, a good clinical efficacy, a low toxicity, and wide indications. Isostructural β-lactams (trans-13-azabicyclo[10.2.0]tetradecan-14-one) were demonstrated to have three polymorphs that have been widely used as components of common antibiotics, such as cephalosporin and penicillin. Luo et al. [296] calculated the crystal structure parameters and predicted the Raman spectra of β-lactam forms I and II using QM theories (DFT and MP2) combined with the proposed embedded fragmental QM method [296]. They investigated the crystal structures and calculated the Gibbs free energies of two β-lactam polymorphs. The two polymorphs of β-lactam were

Table 4.3
Calculated and experimentally observed crystal structure parameters (Å) of polymorphs I and II at atmospheric pressure. The calculation is performed based on the crystal structure optimization [296] with DFT/B97XD/6-31G* theory.

Parameters	Expt. Form I [298]	DFT Form I	Expt. Form II [298]	DFT Form II
a/Å	5.858	5.645	5.962	5.781
b/Å	7.629	7.451	7.267	7.327
c/Å	28.237	28.519	28.689	28.402

identified with the Raman spectra and a comparison of Gibbs free energy, which further shows a temperature-induced phase transition at approximately 308 K.

Sultan et al. investigated two derivatives of β-lactam antibiotics (i.e., amoxicillin and ampicillin), which exhibit a similar antibacterial spectrum [297]. They used quantum mechanical methods for this study. These compounds were confirmed by an XRD analysis, and optimized bond parameters were calculated using DFT at the B3LYP/6-31G(d) level. The optimized geometrical parameters were in good agreement with the crystal data [297]. Luo et al. [296] also used quantum mechanical methods to optimize the crystal structure of β-lactam. Table 4.3 shows a comparison of the crystal structure parameters of polymorphs I and II from theoretical calculations and experimental observations. The optimized crystal structure parameters are consistent with the experimental observations. Specifically, for both β-lactam polymorphs I and II, the deviations between lattice constants **a** and **b** from the calculations and experiments are 0.2 Å, which are 0.282 Å and 0.287 Å = for polymorphs I and II, respectively.

Sultan et al. [297] optimized the molecular geometries of β-lactam in the gas phase using the QM method at the B3LYP/6-31G (d) level. The IR spectrum was obtained from optimized structures, and the frequencies obtained as harmonics were converted to anharmonic frequencies with a scale factor of 0.9600. The vibrational spectra obtained from the optimized structures calculated at the B3LYP/6-31G(d) level in the gas phase of amoxicillin and ampicillin were compared [297]. Fig. 4.4 shows the calculated IR (Fig. 4.4(a)) and Raman spectra (Fig. 4.4(b)–(d)) of β-lactam polymorphs I and II in the low frequency (b) and high frequency (d) regions, respectively, which were done by Luo et al. [296]. In Fig. 4.4, the spectra of polymorphs I and II are colored red and green, respectively. The QM method (DFT/B97XD/6-31G*) was used for structure optimization and energy calculation [296]. In Fig. 4.4(d), the Raman spectra of the two polymorphs present the same number of spectral peaks in the region of high frequency. Meanwhile, the first Raman peak of both polymorphs is located at 3090 cm^{-1}, and the second Raman peak of form II shows a slight shift compared to that of form I. For the region of low frequency, however, the number of peaks is very different for the two polymorphs (Fig. 4.4(b)), in which polymorph I shows five peaks while polymorph II shows seven peaks. In Fig. 4.4(b), the two extra peaks are labeled with black star signs and are capable of distinguishing the crystal structure of β-lactam polymorph II. In conclusion, the significant discrepancy in the Raman spectra, including the number and frequency of the peaks, between the two structures is capable of indicating the existence of two different β-lactam polymorphs. The Raman and IR spectra can present the features of crystal structures very well [296].

4.2. Low temperature phase transition prediction

Olanzapine

Polymorphic molecules may transfer from the parent phase to the secondary phase and vice versa by changing their temperature. Different polymorphic forms of the same compound may possess different chemical and physical properties, including stability, flexibility and compressibility, thus affecting the safety and overall performance. Therefore, it is very important for chemists, physicists, and scientists to know the effect of temperature on polymorphic molecules. Here we study the phase transition of olanzapine polymorphs induced by temperature [276,277,296]. Fig. 4.5(a)–(c) shows the Gibbs free energy difference per unit cell between different olanzapine polymorphs. The Gibbs free energies in Fig. 4.5(b) and (c) were calculated by computing the single point energy of enthalpy. Then, MP2 theory was performed with the 6-31G* basis set based on the DFT ωB97XD/6-31G* of optimized crystal structures, and the zero points of energies and entropies were obtained using ωB97XD/6-31G*. Fig. 4.5(c) shows that the Gibbs free energy of polymorph I is lower than that of polymorph II over the entire temperature range; hence, the crystal structure of olanzapine polymorph I is more stable than that of polymorph II [276]. As shown in Fig. 4.5(c), the difference between the Gibbs free energy of the two polymorphs increases with an increasing temperature, demonstrating that polymorph II becomes less stable as the temperature [276] increases. In Fig. 4.5, the difference between the Gibbs free energies of polymorphs I and II increases from 0.68 kcal/mol to 1.14 kcal/mol by varying the temperature from 5 K to 350 K, respectively [276]. As Fig. 4.5(b) and (c) use the MP2/6-31G* level for Gibbs free energy computation, they have minor differences in the Gibbs free energies, but the variation in the Gibbs free energy difference is similar. Fig. 4.5(a) was computed by only using density functional theory (DFT). DFT provides a reasonable accuracy with acceptable computational costs, whereas MP2 provides a higher accuracy with high computational costs. Therefore, the Gibbs free energy difference computation in Fig. 4.5(b) is more accurate than in (a). Furthermore, it also confirms that olanzapine form I is the most stable form, as believed. More interestingly, Fig. 4.5(b) and (c) demonstrate the phase transition in forms III and IV. Fig. 4.5(b) and (c) show the difference in the Gibbs

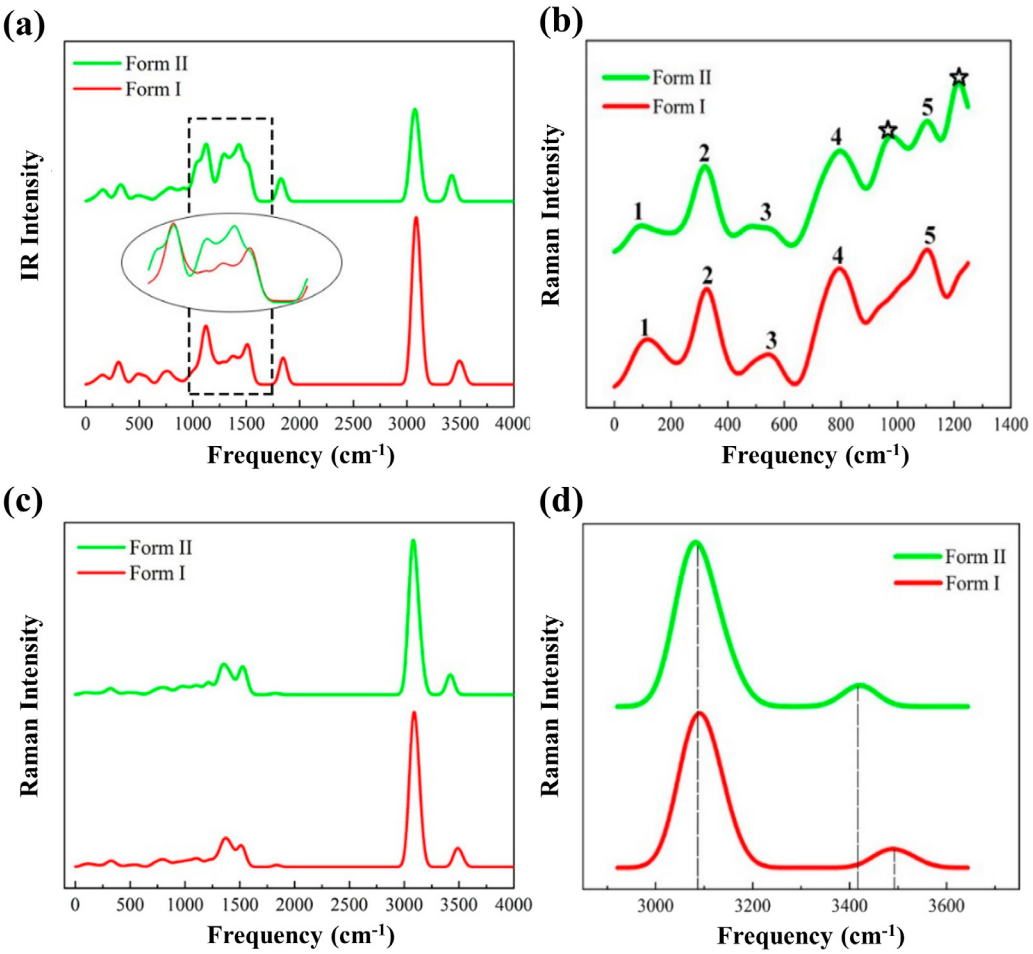

Fig. 4.4. The comparison of (a) IR and (b)–(d) Raman spectra of β-lactam polymorphs I (colored in red) and II (colored in green) at atmospheric pressure in the region of (b) low frequency and (d) high-frequency, respectively. The characteristic Raman peaks of polymorph II are labeled as black stars.
Source: Reproduced by the work of Luo et al. [296].

free energy between the four olanzapine forms in the temperature range of 0 to 450 K using ωB97XD/6-31G* (b) and MP2/6-31G* (c), respectively. The Gibbs free energy curves of olanzapine polymorphs I, II, III, and IV are colored black, red, blue, and green, respectively. In the temperature range of 0 to 450 K, polymorph I remains the most stable structure, while polymorph II exhibits the most unstable structure. Notably, the energy lines of olanzapine polymorphs III and IV show an intersection at 200 K. More specifically, the intersection temperature is 195 K from the DFT calculations and 189 K from the MP2 calculations, as shown in Fig. 4.5(a) and (b) [277], respectively. Therefore, based on the intersection of the Gibbs free energy between polymorphs II and IV, we can demonstrate that a temperature-induced polymorphic transformation occurs at 200 K, which means that polymorph IV remains stable at low temperature conditions (less than 200 K) and can probably transform to polymorph III as the temperature increases above 200 K.

β-lactam

Fig. 4.5(d) presents the difference in the DFT-calculated Gibbs free energy between form I and form II [296]. The enthalpy was calculated using the MP2/6-31G* theory based on the crystal structure optimized by B97XD/6-31G*, while the zero-point energy (ZPE) and entropy contribution were calculated using the B97XD/6-31G* theory [296]. Fig. 4.5(d) shows that the Gibbs free energy of polymorph I is obviously larger than that of polymorph II at low-temperature conditions (less than 308 K) and decreases as the temperature increases above 308 K. Therefore, the phase transition temperature of β-lactam polymorphs I and II is 308 K, above which polymorph II can transform into polymorph I [296]. Previously, neither experimental nor theoretical work presented this polymorphic transformation of β-lactam, and thus a transformation prediction is expected to be instructive for future experimental investigations [296].

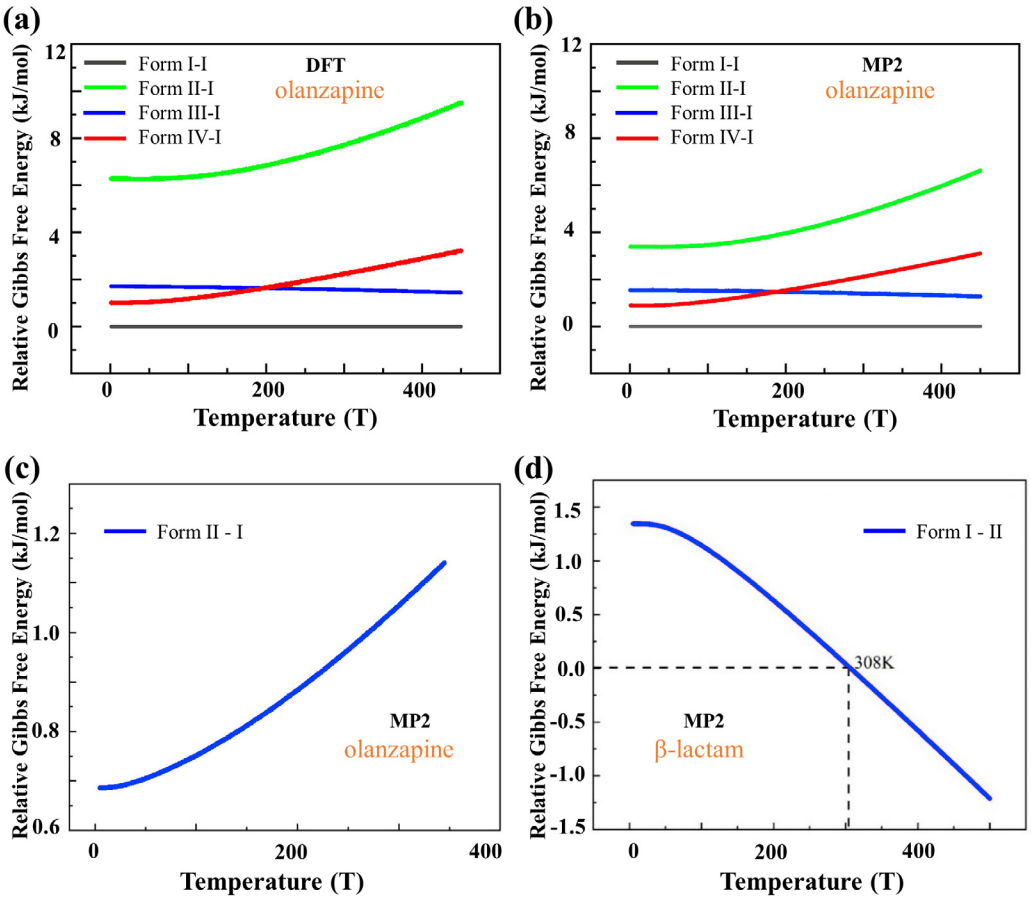

Fig. 4.5. Differences of calculated Gibbs free energies among polymorphs for olanzapine and β-lactam at standard atmospheric pressure, presenting the phase transformation between polymorphs. (a) and (b) show differences of Gibbs free energy between four forms of olanzapine calculated by DFT and MP2, respectively [277] (© 2021 Elsevier). (c) shows Gibbs free energy difference between forms II and I of olanzapine calculated by MP2 [276] (© 2019 American Physical Society). (d) shows differenced of Gibbs free energy between form II and I for β-lactam calculated by MP2, where the phase transformation occurs at about 308 K [296].

4.3. ML predictions for chemical shift and force fields

In computational physics and chemistry, the large number of possible molecules and materials and the numerous methods for chemical transformations make QM approaches required for fundamental understanding of physics and chemistry. For a fixed molecule or molecular geometry, QM methods are capable of accurately calculating microscopic properties, such as energy, atomic forces, polarizability, and electrostatic multiples. As we discussed in Section 2, many ML models have been proposed for the property calculation of small molecules with a better efficiency. For organic molecules, especially drug molecules and relevant crystals, various ML methods have been used for predicting properties, such as NMR chemical shifts [299,300], refractive indices [301], and other molecular properties [302].

For example, in 2020 Scalia et al. [302] quantitatively evaluated the performance of three state-of-the-art methods for the uncertainty estimation of graph convolutional neural networks (GCNNs) [303] for predicting molecular properties, including deep ensembles [304], bootstrapping and Monte Carlo dropout (MC-dropout) [305] with a concrete dropout [306]. Fig. 4.6 presents the illustration for predicting molecular properties with a GCNN.

Chemical shifts

Among these properties, NMR chemical shifts, strongly dependent on local atomic environments, have become one of the most powerful tools for structure elucidation of powdered solids or amorphous materials. Despite the great accuracy of chemical shift calculations, QM methods encounter great difficulties in widespread applications due to the extremely large computational cost. In that case, ML method has become a choice for predicting the chemical shift of solids at the QM level [307]. For a long time, SPARTA+ [308], proposed by the single-layer feed-forward network, has been one of the most popular methods for chemical shift prediction. Based on sequence homology, SHIFTX2 [309] has become a more powerful tool with a better predictive value. In particular, the gauge-including projector-augmented waves (GIPAW) [310] method

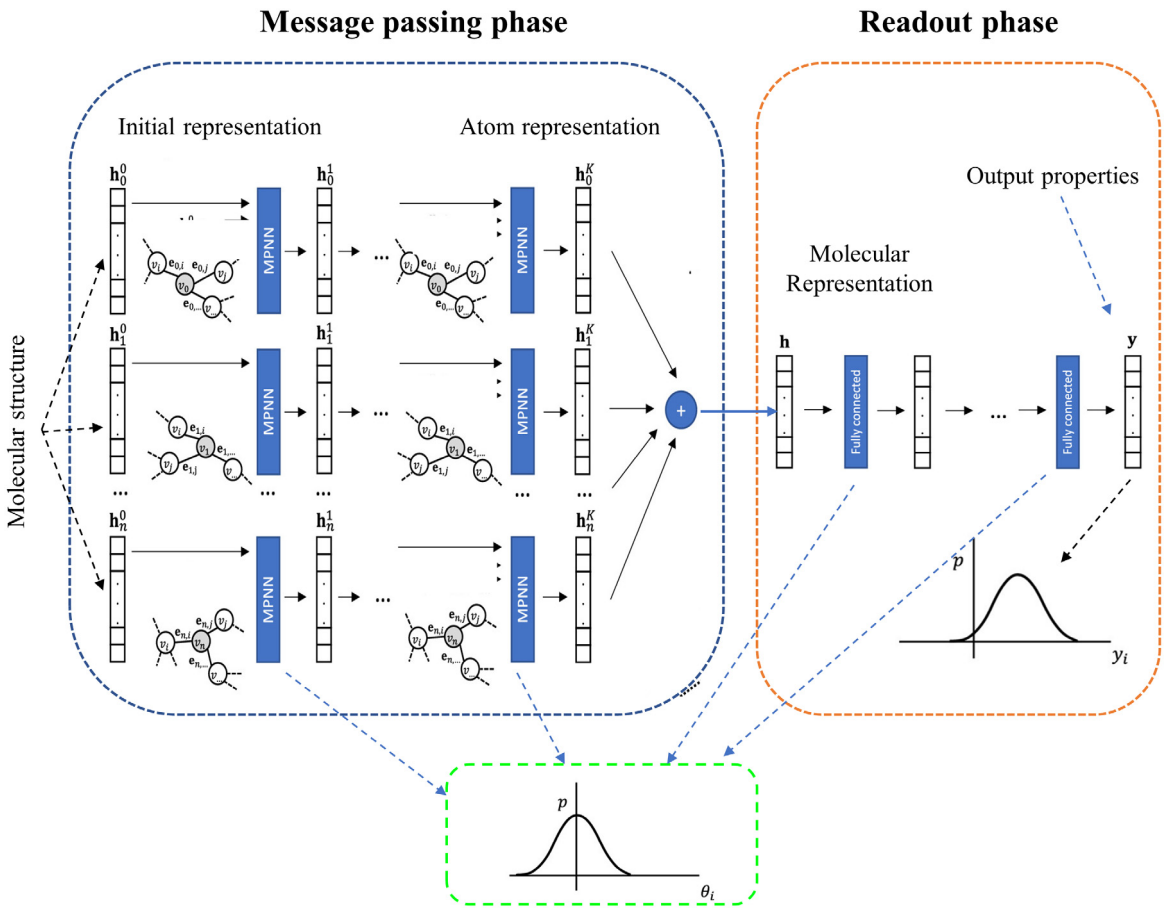

Fig. 4.6. Illustration of molecular property prediction with a GCNN method.
Source: Reproduced with permission from the work of Scalia et al. [302].
© 2020 American Physical Society.

is widely used to identify NMR chemical shifts for solid-state crystals with DFT calculations. To avoid computationally intensive DFT calculations, Paluzzo et al. [311] (2018) proposed an ML model (shiftML) for a chemical shift prediction based on 3D structures and rotational symmetry using KRR and the SOAP kernel [312], which exhibits a good performance for molecular crystal systems. Fig. 4.7 shows the chemical shift calculation times (DFT and shiftML) and NMR chemical shift predictions (shiftML) of six large molecular crystals (CSD reference codes: CAJVUH, RUKTOI, EMEMUE, GOKXOV, HEJBUW and RAYFEF) with more than 700 atoms per unit cell. As shown in Fig. 4.7(a), the CPU times of shiftML predictions (turquoise) are dramatically decreased than DFT calculations (blue and orange).

Despite the considerable acceleration effect of shiftML for chemical shift prediction, training data generation using DFT is still the bottleneck, which makes shallow ANN essential, and the complexity of computing and inverting the kernel matrix makes it impractical for KRR to handle large datasets. In 2019, Liu et al. [300,307] developed a deep learning model for chemical shift prediction for molecular crystals without DFT calculations. They proposed the multiresolution 3D-DensseNet architecture (MR-3D-DenseNet) by using multiple channels to describe different spatial resolutions for each atom type with cropping, pooling, and concatenation. This model presents good results for ^{13}C, ^{15}N, and ^{17}O chemical shifts compared with the QM methods. Fig. 4.8 shows the overview of the MR-3D-DenseNet architecture and the testing RMSEs for ^{1}H, ^{13}C, ^{15}N, and ^{17}O from KRR, 3D-DenseNet without data augmentation and 3D-DenseNet with data augmentation.

In addition, another DFT-based ML model was proposed by Gao et al. [319] (2020), which presented a significant accuracy increase for ^{13}C and ^{1}H NMR chemical shift prediction for a variety of organic molecules. In the same year, Gerrard et al. [320] proposed an intelligent machine prediction of shift and scalar information of a nuclei (IMPRESSION) system for efficiently predicting the NMR parameters of 3D molecular structures at the QM level of accuracy.

Force fields

Classical force fields and molecular dynamic simulations constitute the cornerstone of contemporary atomistic modeling in physics, materials, chemistry and biology. However, based on the interatomic potential, classical force fields have

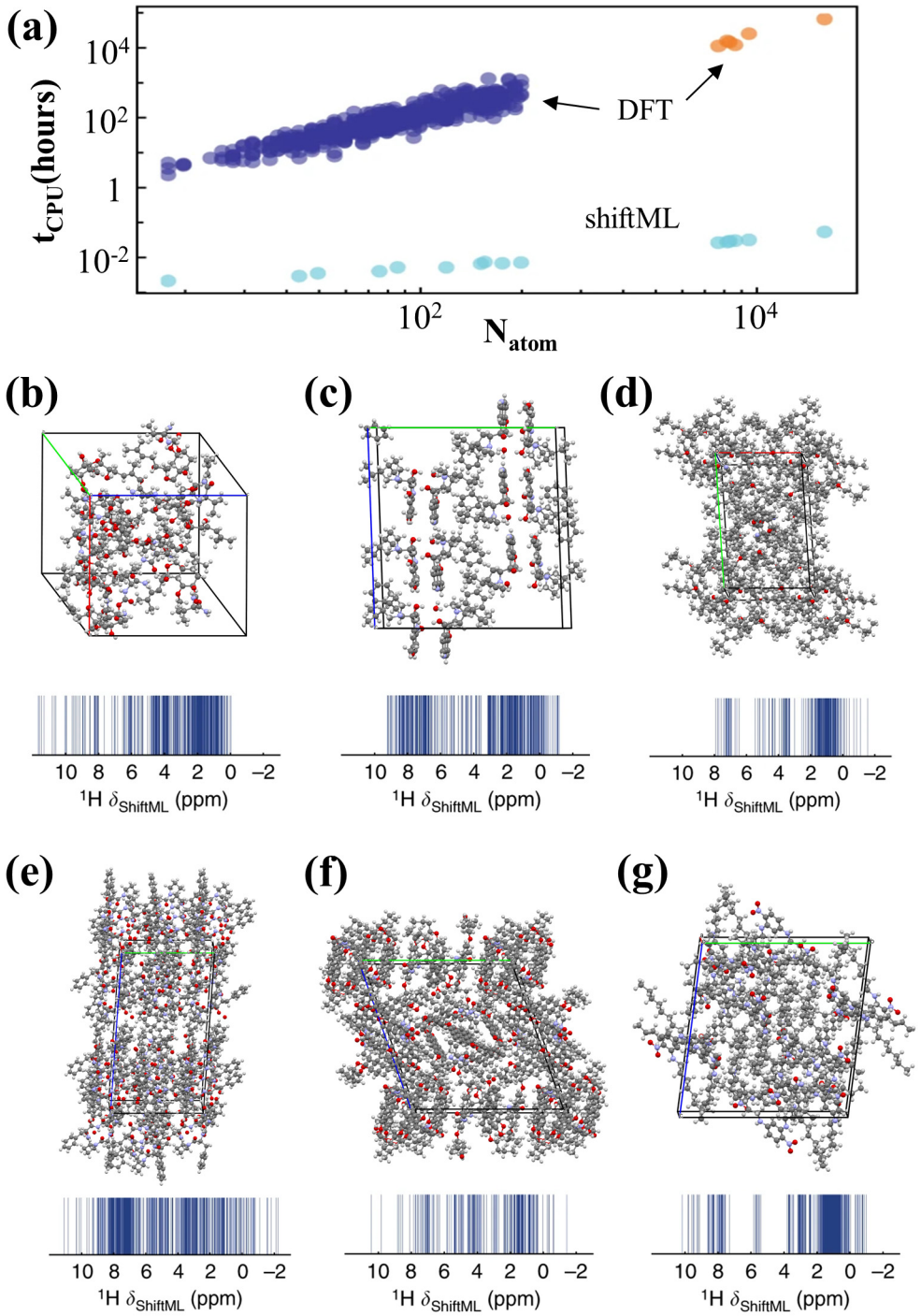

Fig. 4.7. Chemical shift calculation times and large structures. (a) DFT (GIPAW) calculation time (blue and orange) and ShiftML prediction time (turquoise) for different system sizes. (b)–(g) 3D-schemes and ^{1}H NMR spectra predicted with ShiftML, of the six large molecular crystals with CSD refcodes: (b) CAJVUH [313], $N_{atoms} = 828$, (c) RUKTOI [314], $N_{atoms} = 768$, (d) EMEMUE [315], $N_{atoms} = 860$, (e) GOKXOV [316], $N_{atoms} = 945$, (f) HEJBUW [317], $N_{atoms} = 816$, and (g) RAYFEF [318], $N_{atoms} = 1584$.
Source: Reproduced from the work of Paluzzo et al. [311].

a series of limitations and cannot present key quantum effects in molecules. In recent years, ML methods have been used to construct molecular force fields based on high-level QM calculations for a series of small molecules, as discussed in Section 3.4. In addition, for several drug molecules the global ML-based force field constructed by these ML models has shown a great potential for applications in molecular dynamic simulations with fully quantized electrons and nuclei.

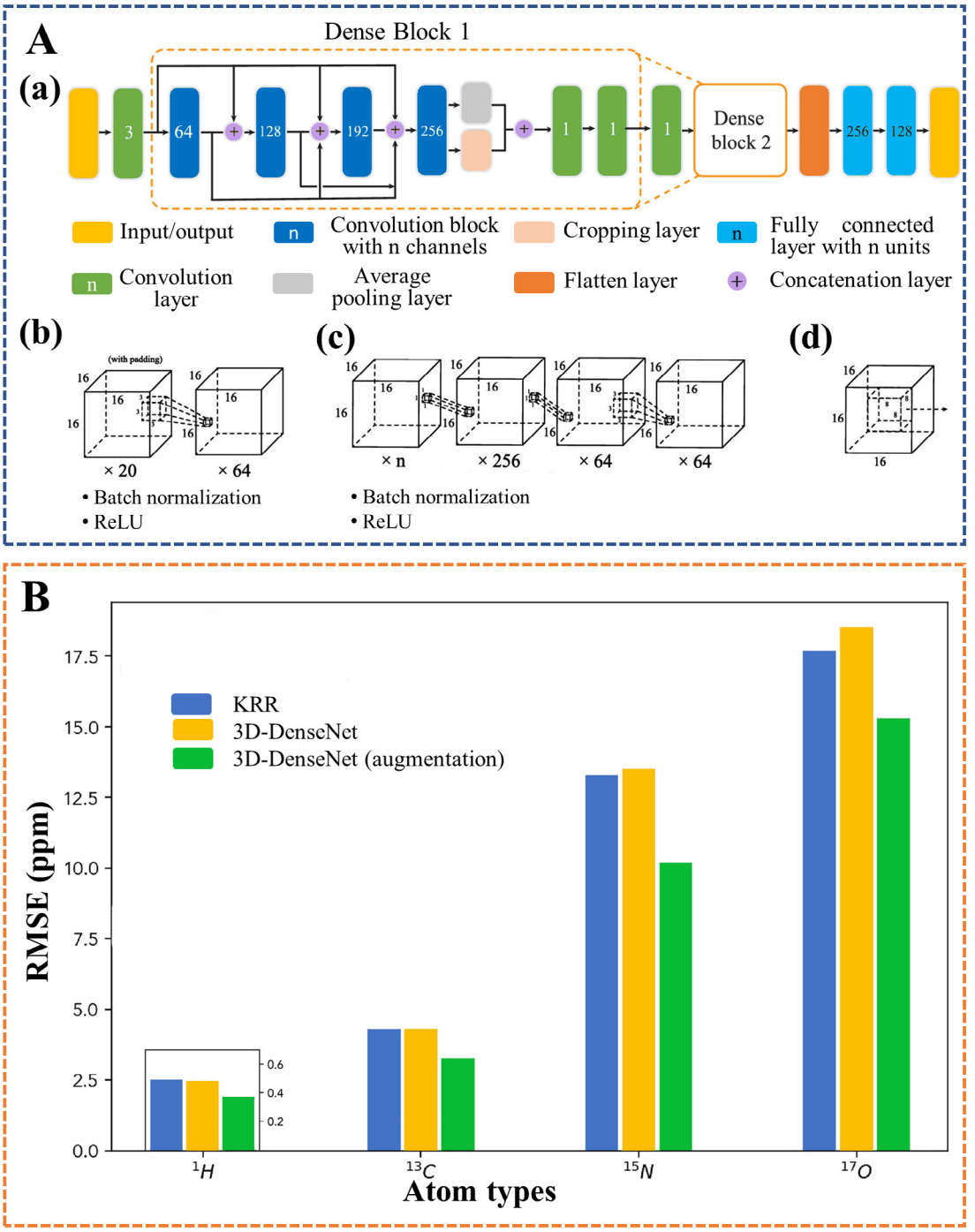

Fig. 4.8. Illustration and the performance of MR-3D-DenseNet architecture. (A) presents the (a) overall workflow of the network, (b) illustration of 3 × 3 × 3 convolution layer prior to the first dense block, (c) illustration of the repeating unit in DenseNet block containing two 1 × 1 × 1 convolution layers followed by a 3 × 3 × 3 convolution layer, and (d) illustration of the cropping layer from the center of the feature map. (B) presents the testing RMSEs for 1H, 13C, 15N, and 17O from KRR, 3D-DenseNet without data augmentation and 3D-DenseNet with data augmentation. (© 2019 American Physical Society) and (B) is taken from Haghighatlari et al. [307] (© 2020 Elsevier).
Source: (A) is taken from the papers of Liu et al. [300].

In 2018, Chmiela et al. [167] proposed an sGDML [235] model for constructing flexible and global molecular force fields based on spatial and temporal physical symmetries. Since the data for model training were obtained from high-level *ab initio* calculations, this ML model is capable of constructing force fields with a CCSD(T) level of accuracy. During model

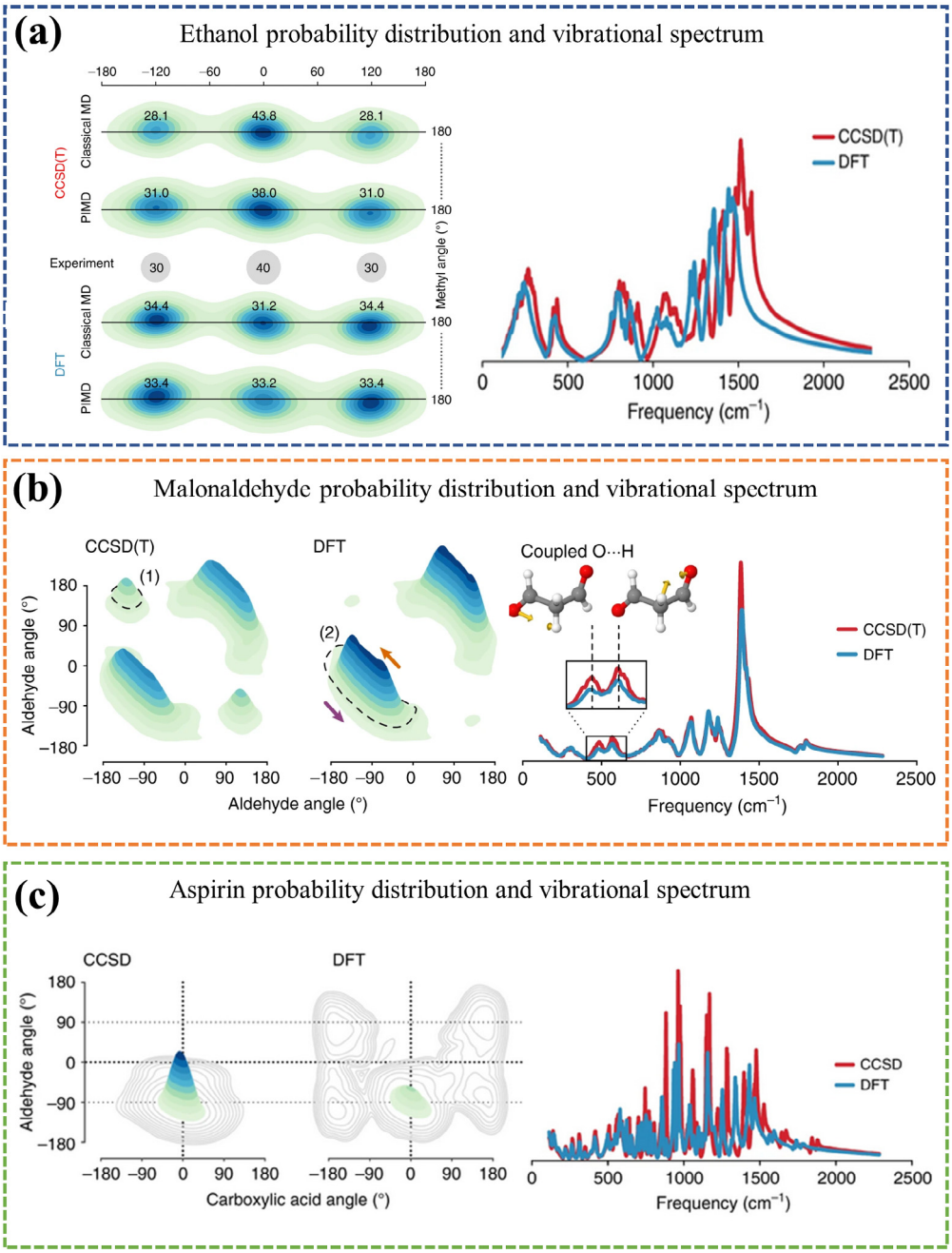

Fig. 4.9. MD simulations using sGDML model for (a) ethanol, (b) malonaldehyde and (c) aspirin. (a) The joint probability distribution function for the two dihedral angles, and the vibrational spectra (velocity–velocity autocorrelation function) using PIMD simulations with sGDML-CCSD(T) and sGDML-DFT at 300 K. (b) The joint probability distributions of the dihedral angles in malonaldehyde of both aldehyde groups using classical MD simulations with sGDML-CCSD(T) and sGDML-DFT. (c) The joint probability distributions for the dihedral angles in aspirin of the ester and carboxylic acid groups using PIMD simulations with sGDML-CCSD and sGDML-DFT at 300 K.
Source: Reproduced from the work of Chmiela et al. [167].

construction, the computational complexity is significantly reduced by a data-based investigation of related physical symmetries in space and time and increasing the information content of data entries by implementing the indicated static and dynamic symmetries, and thus implicitly increasing the data volume. The constructed force fields can be applied to molecular dynamics simulations at the CCSD(T) level with fully quantized molecular electrons and nuclei. Fig. 4.9 shows the performance of these force fields in molecular dynamics simulations for ethanol (Fig. 4.9(a)), malonaldehyde

(Fig. 4.9(b)) and aspirin (Fig. 4.9(c)). The ML simulations for ethanol, malonaldehyde and aspirin demonstrate the necessity of adopting a force field with high accuracy to obtaining quantitative and reliable understanding of molecular systems. The accuracy of the proposed forced fields satisfies the stringent demands of spectroscopic accuracy for molecular simulations in the range of wavenumbers (∼0.03 kcal/mol), compared with 0.1–0.2 kcal/mol of the energy difference between molecular conformers. The ML force fields have been demonstrated to be powerful for predictions with a high accuracy for a series of atomistic and molecular systems [321–323], which usually have common features and relatively flat free energy surfaces.

The accuracy of such a force field strongly depends on the quality and characteristics of the data, such as atomic configurations and relevant energies. Another example is that Plazinski et al. [324] (2020) proposed a new approach for constructing ML force fields based on MM/MD simulations and QM calculations for selected structures. The biased subsampling of the configurations, which can increase the population of hardly accessible states, is used to address the challenge of a decreasing accuracy that results from the Boltzmann distribution and the absence of high-energy states. The applications of the proposed force fields on two flexible, heterocyclic molecules (2-fluorotetrahydrofuran (THP-F) and 2-fluorotetrahydropyran (THF-F)) exhibit a great performance. These ML force fields exhibit a high potential for applications in MD simulations at the QM level of accuracy as ML model progress.

As discussed above, significant progress has been made in constructing force fields and predicting chemical shifts and other properties of organic molecules and molecular crystals using various ML methods. However, there are many challenges for property predictions using ML methods with the accuracy of high-level QM methods. For example, much more large and high-quality datasets are required for chemical space exploration and property prediction, such as the Materials Project [325], MD17 [235], QM7 [156,326], QM9 [327] and OQMD [328]. On the other hand, new powerful ML methods rely on only a small amount of data and are essential for computational savings and wide applications. Proposing reliable models based on small datasets is challenging, which makes it essential for achieving data efficiency with a considerable accuracy and robustness for ML methods by including prior physical knowledge and invariance information. Another issue is the selection bias existing in a series of training datasets in this field, which can influence the robustness and rigorousness of the statistical learning process, such as k-fold cross-validation and convergent learning curves. Stability or attribute distributions are often unknown, thus hindering a rigorous assessment of the extent to which a particular dataset is fully representative of the broad chemical space. Predictably, progress in ML methods, high-quality data acquisition and data representation will advance the increase in the accuracy and efficiency for chemical shifts and other property predictions.

4.4. ML accelerates crystal structure prediction (CSP)

In the previous sections, we introduced the applications of ML methods for predicting the properties of molecules and crystals. For molecular crystals, the properties depend on the molecular constituents and the relative arrangement of molecules in the crystal unit cell. Weak intermolecular interactions [329,330] usually result in polymorphisms in molecular crystals, presenting multiple forms [331,332] at different pressure and temperature conditions. However, direct crystal structure determination via power X-ray diffraction patterns is resource and time consuming and is not always possible. In addition, the predicted structures can provide useful information for determining the structure from incomplete experimental data, such as electron diffraction [333], solid-state NMR [334] and power X-ray diffraction [335]. In that case, CSP has been of great importance and draws much attention in pharmaceutical research and the design of high-performance organic electronics [336–338]. The CSP only relies on basic connectivity information of a molecule (chemical diagram), and the different polymorphs often differ by only a few kJ/mol, leading to a significant challenge for computational physics and chemistry [339,340]. As the starting point and critical component of CSP, random crystal structure generation requires efficient methods to generate new structures for sampling the high-dimensional configuration space related to molecular crystals, which has become part of the Cambridge Crystallographic Data Centre (CCDC) CSP blind test [341–346]. Much effort has been made to propose such an algorithm to randomly generate crystal structures. For example, several methods have been proposed, such as relaxing a few handmade structures [347], using the morphology of the molecule [348] and adding up atomic volumes [349]. A random structure generator for molecular crystals (Genarris 2.0) was developed by Tom et al. [350] in 2020, which is a new version of Genarris [351] with many improvements. Apart from random structure generation, algorithms for structure optimization have also become essential components of CSP blind tests [341], such as evolutionary algorithms [352,353], random search [349,354], quasi-random search (Sobol sequence) [355,356], MC parallel tempering [357] and simulated annealing [358].

In recent years, progress in ML methods has significantly promoted the improvement in CSP for overcoming the limitations in accuracy for force fields and in extreme computational cost for full QM methods. In 2018, Yamashita et al. [57] developed a selection-type CSP model based on a random search and Bayesian optimization [359–361], which is distinguished from evolutionary algorithms [292,362,363] and particle swarm optimization [364,365]. With the ML algorithm, the proposed approach demonstrated a high efficiency for selecting the most stable structure from a large number of random structures when applied to known systems, such as Y_2Co_{17} and NaCl. The number of searching trials for finding the global minimum structure was reduced by 30%–40%. Based on the evolutionary algorithm (USPES) [292,362,363], Podryabinkin et al. [58] (2019) introduced an approach for automated construction of interatomic interaction models from scratch using ML interatomic potentials and the active learning on-the-fly algorithm [366].

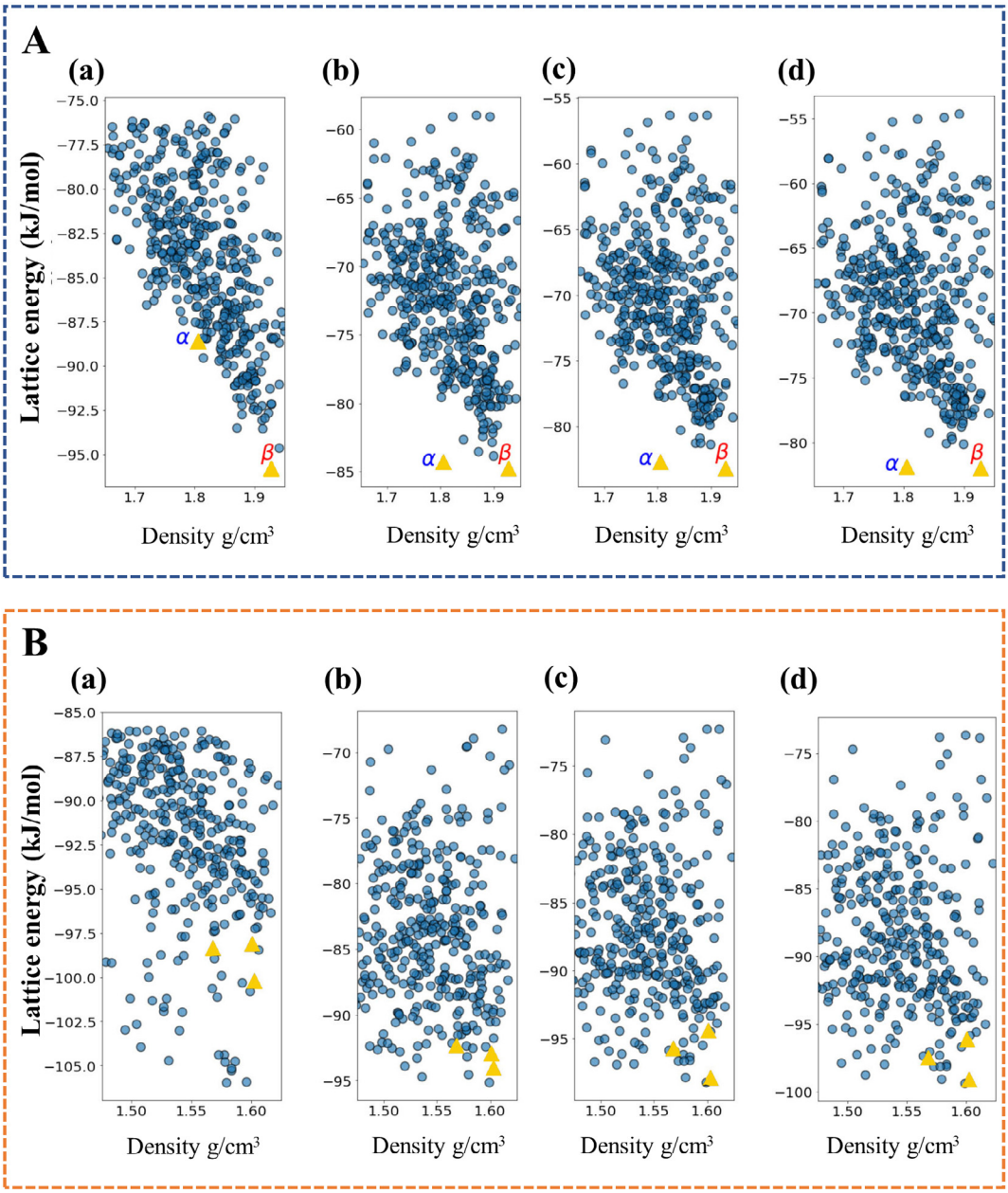

Fig. 4.10. Lattice energy depending on density for predicted crystal structures of A. oxalic acid and B. maleic hydrazide, using (a) the FIT+DMA force field, with (b) the GP ML model using 10% training data, (c) GP ML model using 20% training data, and (d) MP2 (target) landscape.
Source: The pictures were taken from the paper of McDonagh et al. [59].

The proposed model then successfully reproduced all the main allotropes for the testing system, such as carbon, the high-pressure phase of sodium, and boron allotropes. In the same year, McDonagh et al. [59] proposed an approach for improving the force field lattice energy calculations during CSP via two-body corrections with a high-level DFT/MP2 and Gaussian process (GP) ML model. For accuracy evaluation, Fig. 4.10 shows the lattice energy depending on the density for predicted crystal structures of oxalic acid (Fig. 4.10A) and maleic hydrazide (Fig. 4.10B) using the FIT+DMA force field with different corrections, namely, GP ML with 10% data, GP ML with 20% data and MP2, indicating that the GP ML model with only a 20% training data can produce a faithful reproduction of the MP2 correction.

Recently, Egorova et al. [60] (2020) introduced a multifidelity statistical ML (GP model) to predict expensive hybrid DFT functional (PBE0) calculations in CSP for molecular crystals, which presented a good performance for reproducing the crystal structure landscapes for oxalic acid, urazole, and maleic hydrazide. In 2021, based on data mining and ML,

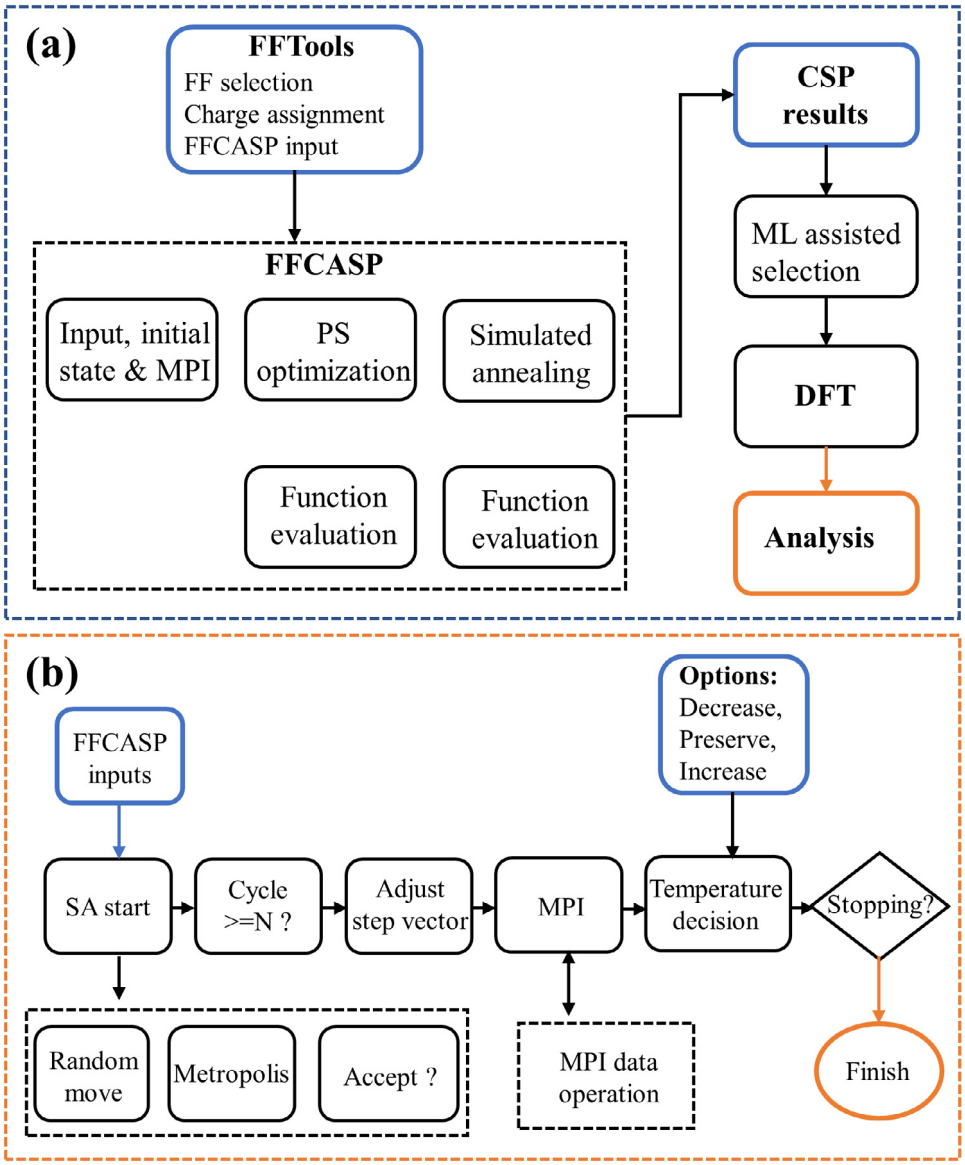

Fig. 4.11. The (a) workflow of FFCASP and (b) the improved SA algorithm equipped in FFCASP.
Source: The pictures were taken from the paper of Demir and Tekin [367].

Demir and Tekin [367] developed the fast and flexible crystal structure predictor (FFCASP) algorithm to predict the structure of molecular crystals, which is an enhanced version of the DFT-based algorithm (CASPESA) [368]. The parallel acceleration provides the ability for global optimization with 1000 separate parameters and the prediction of structures with more than 200 atoms in a unit cell. Fig. 4.11(a) presents the workflow of the FFCASP algorithm, in which global optimization is performed using an improved simulated annealing (SA) [369] algorithm with several new features, such as periodic parameters, task forming parallelism and a new temperature reduction method, as shown in Fig. 4.11(b). This algorithm has been applied to the CSP of pyrazinamide, cytosine, and coumarin by generating more than 20000 structures and selecting unique structures for further optimization, which successfully reproduced the reported crystal structures, indicating a further possibility to investigate the polymorphic nature for molecular crystals of interest and important, such as various drugs. Fig. 4.12 shows the results of predicting the cytosine structure with FFCASP, including the comparison of structures from FFCASP and DFT optimizations, the matminer [370] calculated dendrogram diagram based on the structure distance matrix, and the phonon band structures of the predicted crystal structures.

As we discussed above, many ML methods along with various CSP algorithms, QM methods, and fragment-based methods have been widely used for developing new approaches for efficiently and accurately predicting crystal structures.

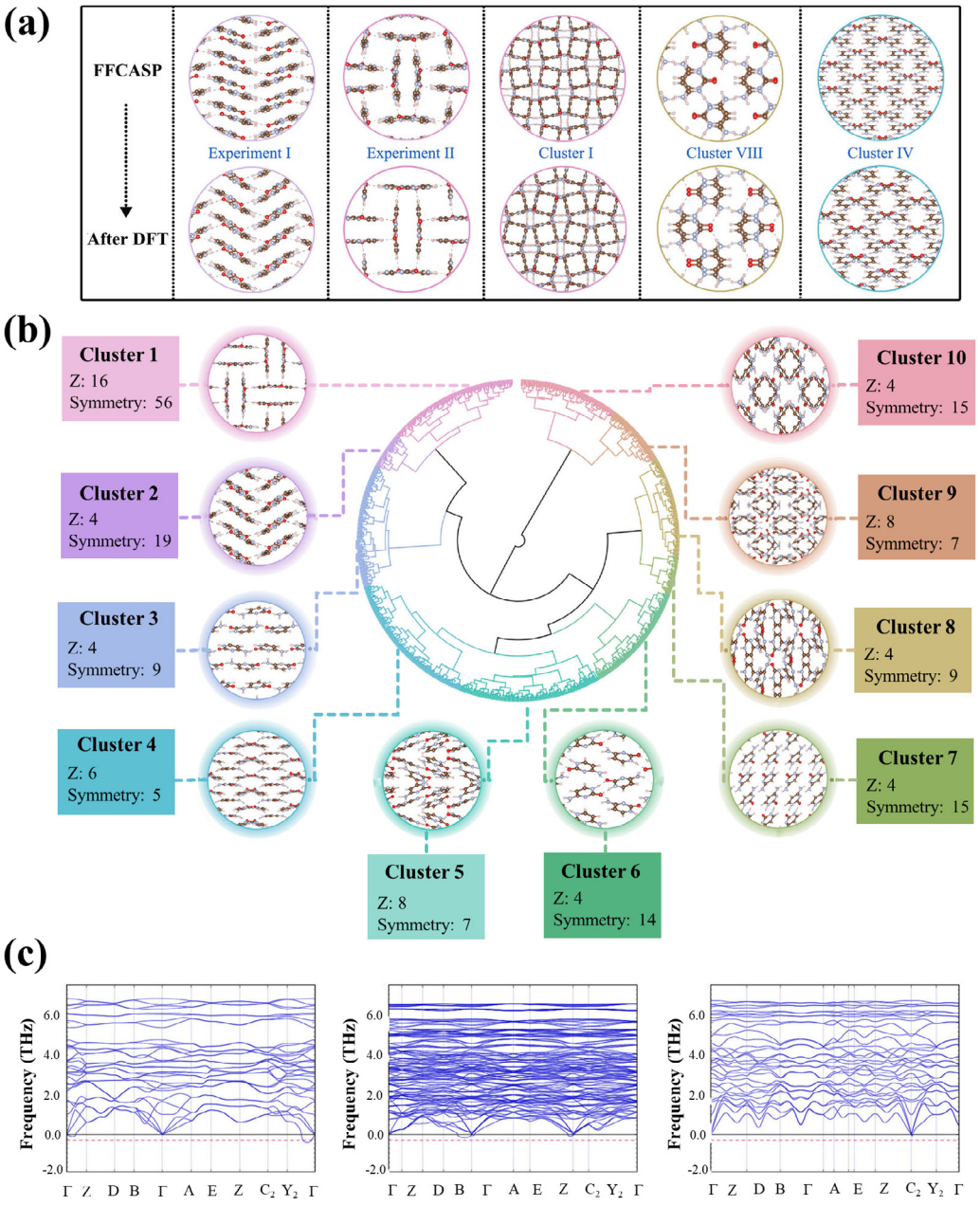

Fig. 4.12. The results of predicting cytosine structure with FFCASP, including (a) the comparison of structures from FFCASP and DFT optimizations, (b) the matminer calculated dendrogram diagram based on the structure distance matrix, and (c) the phonon band structures of three predicted crystal structures, including monoclinic structure ($P2_1/c$, Z = 4), tetragonal structure ($I4d$, Z = 16), and orthorhombic structure ($Fdd2$, Z = 16). *Source:* The pictures were taken from the paper of Demir and Tekin [367].

It is foreseeable that with the rapid development of ML methods, structural representation methods, stochastic algorithms and global optimization algorithms, the accuracy, reliability and efficiency of CSP will be greatly improved.

5. ML-driven software for accelerating the QM calculation

In the above sections, we introduced several QM methods, fragment-based QM methods and ML models for accelerating QM calculations in computational QM calculations, especially for molecules and molecular crystals. ML methods have been widely used for PES constructions of molecules and molecular crystals, resulting in a series of algorithms, models and packages, such as Amp [371], sGDML [167,235], AP-Net [190], DeePMD [187], and PES-Learn [372]· as discussed in

Table 5.1
Software and packages for computational chemistry/physics with ML methods, including the program name, author and year of first report, program type, methods used in the program, and the main function.

Name	Author	Type	Method	Function
Ochem_predic_nn	Coley et al. [79] (2017)	Python program	ML	Predicting organic reaction outcomes
Mol2vec	Jaeger et al. [373] (2018)	Python package	Unsupervised ML	Learning vector representations of molecular substructures
XtalOpt(r12)	Avery et al. [374,375] (2018)	Software	ML	Predicting crystal structure, and hardness
PES-Learn	Abbott et al. [372] (2019)	Python package	GP and FFNN	Constructing PESs
MLatom	Dral [376] (2019)	Program	KRR	Atomistic simulations
ChemML	Haghighatlari et al. [377] (2020)	Program and platform	AL, TL, auto ML	Analysis, mining, and modeling of chemical and materials data
UADDCR	Zhang et al. [378] (2020)	Software	One-hot, NN and K-means	Predicting chemical reaction
QSAR-Co-X	Halder and Cordeiro [379] (2021)	Toolkit	kNN, NB, SVC, RF, GB, MLPNN	modeling mt-QSAR
NN-FQM	Han et al. [183] (2021)	Software	NN	Optimizing structure, and predicting properties and phase transition

Sections 2.3, 3.4 and 4.3. For wide applications, several ML models or algorithms have been compiled into visualization software or packages to provide quick and easy access for researchers, such as Mol2vec, MLatom, UADDCR, QSAR-Co-X and NNFQM. Table 5.1 presents a list of notable programs developed using ML methods for molecules and crystals. Based on the different ML models and data representations, these programs were designed for a series of applications, including molecule representation, PES construction, QSAR modeling, atomic simulation, and the prediction of structure, property, chemical reaction, and phase transition. Generally, the use of these visualization programs does not require extensive experience with ML, programming, or scripting, which benefits researchers focusing their attention on chemical studies with ML. In addition, open-source python packages are usually more flexible and extensible for a wide range of applications, allowing users to design the programs on their own investigations, and even modify the packages using different ML models.

Molecule representation

The appropriate descriptions of molecules are of great importance for ML methods being applied to computational physics and chemistry. In 2018, Jaeger et al. [373] introduced an unsupervised ML model (Mol2vec) for learning vector representations of molecular substructures. Mol2vec can encode compounds as information-rich vectors by summing the vectors of the individual substructures, which are obtained from the unsupervised ML model for a corpus of compounds containing all of the available chemical matter. The encoded vector representations can then be fed into the ML models for further ML training and property prediction. The evaluation of Mol2vec on common substructures and amino acids demonstrates that derived substructure vectors of chemically related substructures and compounds occupy similar vector space, and thus a new compound can be described by summing these substructure vectors obtained from a previous Mol2vec model. An illustration of Mol2vec is shown in Fig. 5.1, in which a Mol2vec model is generated by embedding-unsupervised pretraining, and the generated vectors are fed into supervised ML for applications. ML models for chemical representation, such as Mol2vec, provide a basis for a wide range of ML applications for predicting chemical properties.

Property prediction

With molecular representation, ML methods can be easily used to predict a wide range of properties, such as the energy, atomic force, chemical shift, hardness, and phase transition of crystals, which have been introduced in the previous sections. In addition to the models for a particular system, several software programs have been proposed for a wide range of systems and properties. In 2018, Avery et al. introduced Xtalopt [374] version r12, an open-source evolutionary software for crystal structure prediction, in which an automatic flow for materials discovery-ML (AFLOW-ML) [375] model is incorporated for hardness calculations and the prediction of hard structures. In 2020, Han et al. [183] proposed a newly efficient approach for the predictions of Gibbs free energy, structural characteristics and thus phase transition of solid crystal structures. Based on a combination of QM calculations, fragment-based methods, and NNs, this approach is capable of accelerating high-level *ab initio* calculations with the MP2 level of accuracy and presents a good performance in the evaluation of ice crystals. Based on this approach, they also proposed NN-based fragmental QM (NN-FQM) software for molecular crystal calculations with a GUI interface for usability. In the NN-FQM procedure, the PES models for low-order many-body terms can be obtained via model training in this software or by incorporating external PESs. With these PESs, the NN-FQM software is capable of predicting the Gibbs free energy and atomic forces, optimizing the structure of a molecular crystal under a series of pressure–temperature conditions and predicting the phase transition between two crystal phases. Quantitative structure–activity relationship (QSAR) construction has become one of the most essential and effective tools for a wide range of applications, such as chemical data mining and analysis, predicting properties and investigating potential information via virtual screening from libraries [380,381]. Previously, conventional QSAR models were constructed using a small number of experimentally observed or theoretically calculated data points, which significantly limited the model accuracy. In recent years, a series of ML models was proposed for QSAR modeling [382,383].

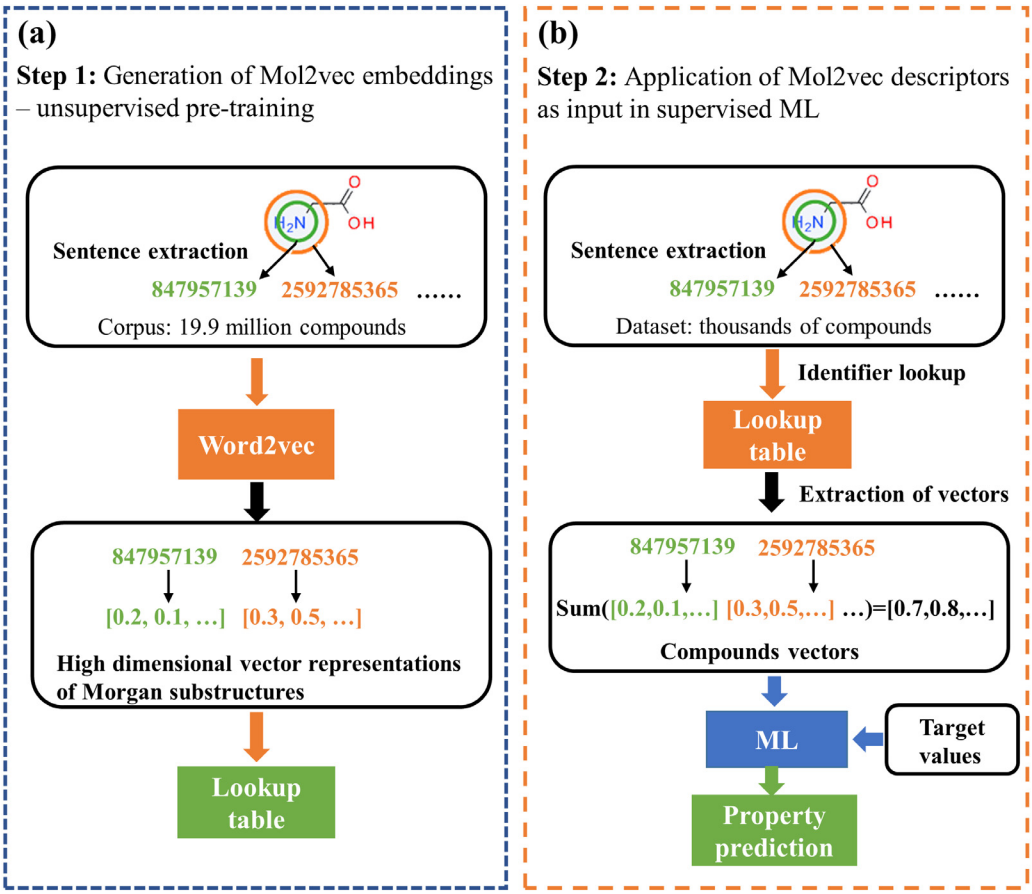

Fig. 5.1. Illustration of (a) generation and (b) application of Mol2vec.
Source: Reproduced from Jaeger et al.'s work [373].

In 2019, Ambure et al. [384] introduced open-source visualization software (QSAR-Co) for constructing classification-based QSAR models and allowed mining of the response data from multiple conditions by adopting a genetic algorithm [385,386] (GA) based on a Linear Discriminant Analysis (GA-LDA) [387] or the RF [388] classifier. Fig. 5.2(a) (b) shows the workflow of the QSAR-Co software, which contains two modules: the model development module (module 1) and the screen/predict module (module 2). The QSAR-Co software has been demonstrated to be extremely user-friendly and efficient, which only requires setting up parameters and techniques and simply clicking a button.

Very recently, Halder and Cordeiro [379] (2021) moved a step forward from QSAR-Co software and developed an open-source toolkit (QSAR-Co-X) for multitarget QSAR (mt-QSAR) modeling based on the Box–Jenkins moving average approach. By integrating diverse chemical and biological data from various conditions into a certain model and simultaneously predicting the targeted response variables [389–391], mt-QSAR modeling is capable of extending and improving the reliability of QSAR modeling. The QSAR-Co-X software contains four modules, including the linear modeling (LM) module, the non-linear modeling with grid (NLG) search module, the non-linear modeling with user specific parameters (NLU) module, and the condition-wise prediction (CWP) module, in which various descriptors, feature selection algorithms, ML methods, validation strategies and analysis techniques are incorporated. Fig. 5.2(c) presents an illustration of the workflow of the QSAR-Co-X toolkit, in which different modules are highlighted with different colors (blue, orange, black, and green for module 1, module 2, module 3, and module 4, respectively, and red for modules 1–3). This toolkit has presented a high efficiency for handling large datasets from a series of experimental and theoretical conditions [392–398], presenting its potential to become a widely used toolkit for easily constructing mt-QSAR models.

Automatic model construction

As we discussed in the previous sections (Sections 2.3, 3.4 and 4.3), PES construction is an essential application for ML methods in computational physics and chemistry, and several ML models for constructing PESs have been introduced. However, constructing PESs using complex ML methods could become a challenge and a problem since extensive programming experience and ML knowledge are usually required for model implementation. In addition, most PESs are

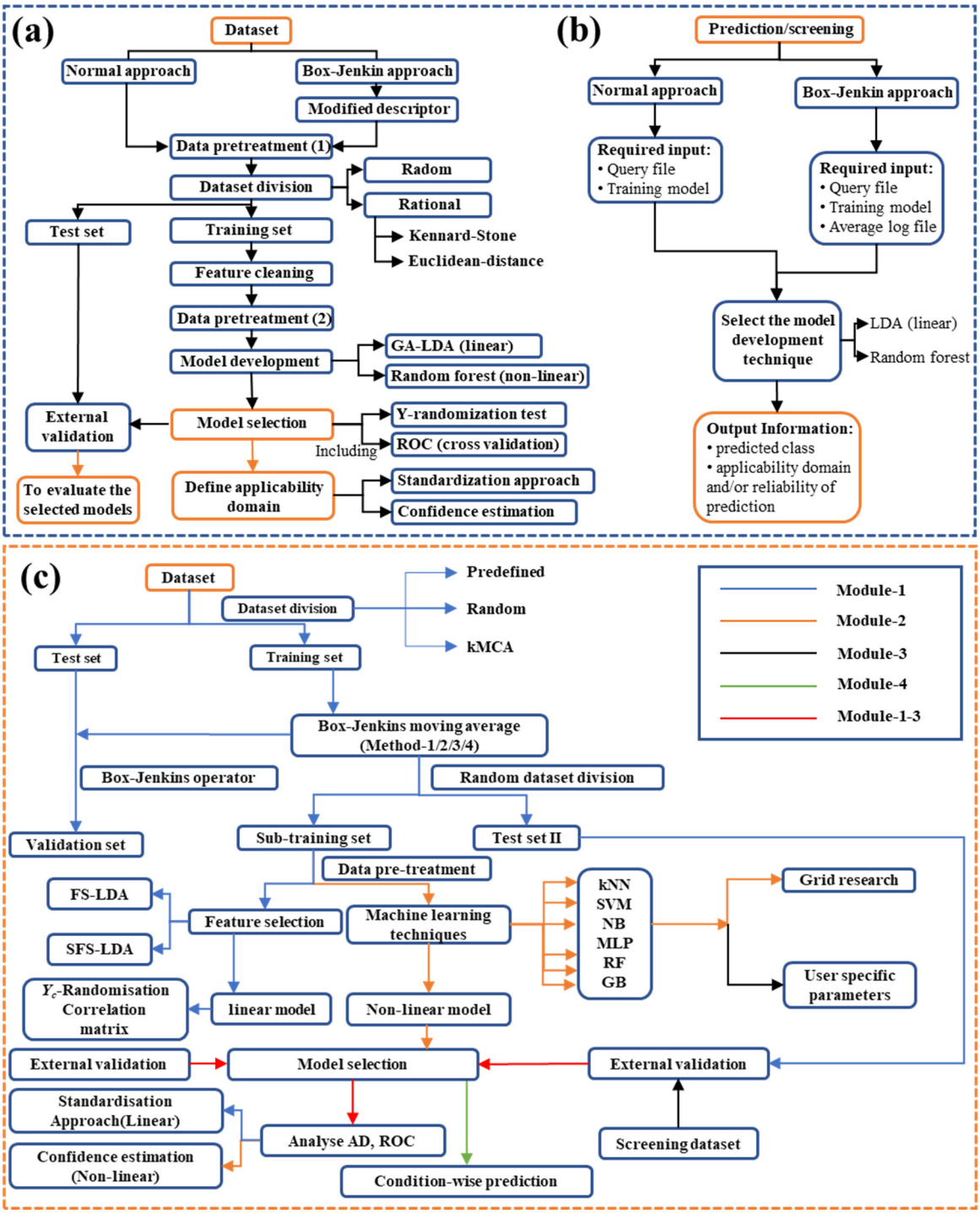

Fig. 5.2. Illustration of overall functionalities of QSAR-Co and software QSAR-Co-X toolkit. The QSAR-Co contains two modules, including (a) model development (Module 1) and (b) prediction/screening of query chemicals/database (Module 2). The QSAR-Co-X toolkit (c) contains four modules, in which different modules are highlighted with different colors (blue for module 1, orange for module 2, black for module 3, green for module 4, and orange for modules (1–3).
Source: These figures were reproduced from the reports of Ambure et al. [384] and Halder et al. [379].

constructed for different cases, which makes them less transferable and verifiable. In that case, automatically constructing ML models and PESs is of great importance, and several software and platforms have been developed. Fig. 5.3 shows the

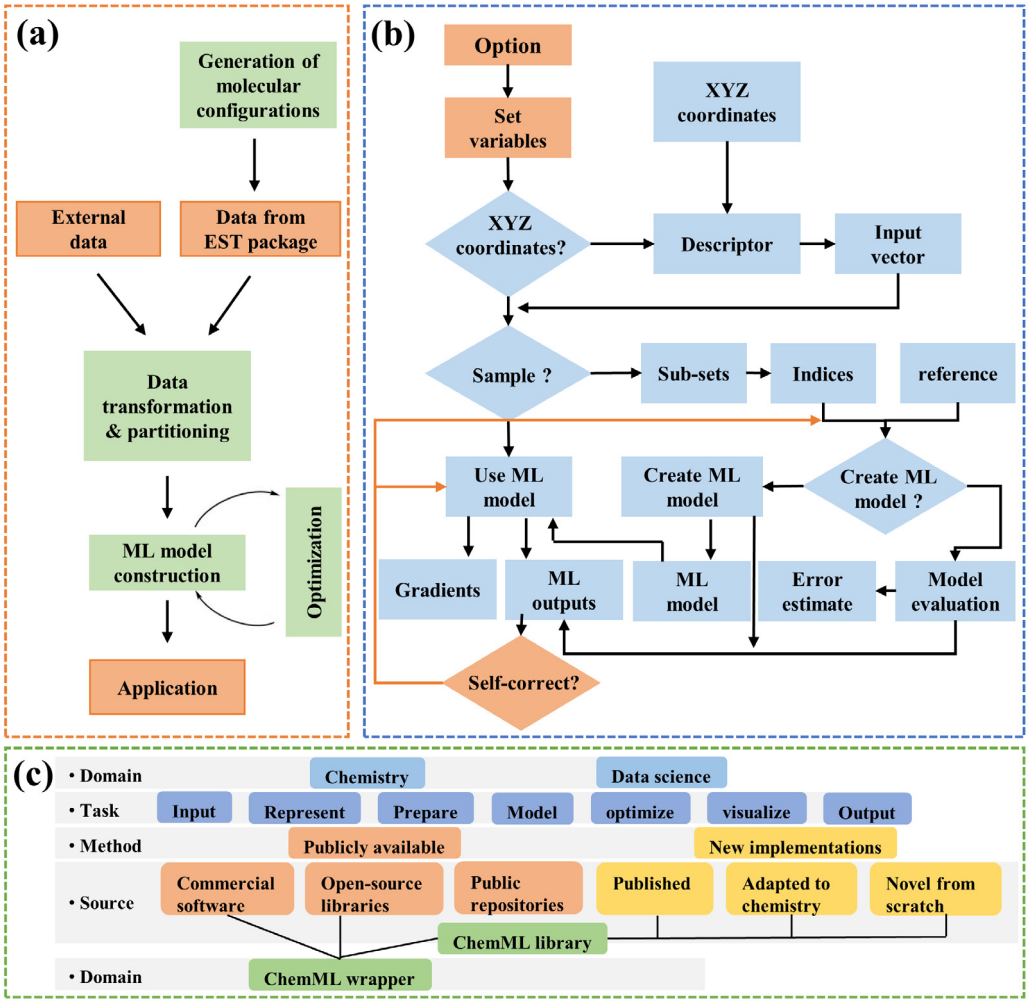

Fig. 5.3. Workflow of three ML software packages. (a) PES-Learn including three automatic steps: generation of molecular configurations, data preprocessing, transforming and partitioning, and ML model construction, (b) MLatom, in which tasks in blue are run by MLatomF command and tasks in orange are run by MLatom.py script, and (c) ChemML software package, containing ChemML library and ChemML Wrapper. (© 2019 American Physical Society), Dral [376] (© 2019 John Wiley and Sons), and Haghighatlari et al. [377], respectively.
Source: The figures were reproduced from the reports of Abbott et al. [372].

workflow schematic for three software packages (PES-Learn, MLatom, and ChemML). In 2019, Abbott et al. [372] developed an open-source and free software package (PES-Learn) for the automatic development of high-quality ML models of system-specific Born–Oppenheimer molecular PESs. As shown in Fig. 5.3(a), the PES-Learn package automates many steps that are required for ML PES construction, including the generation of molecular configurations and corresponding electronic energies, data transformations and partitioning, as well as ML model training.

In addition, PES-Learn provides two ML algorithms for PES training: GP [241,399] and FFNN [155,400], which can also use externally supplied PES data. In practice, PES-Learn presents a good performance for fitting several semi-global PESs and conducting high-level vibrational configuration interaction computations for H_2O and H_2CO, with the CCSD(T) level of accuracy. In the same year, Dral [376] proposed a ML-based program package (MLatom) for computationally efficient simulations of atomistic systems, which can be used out-of-the-box as a standalone program without extensive experience of ML, and programming. The workflow of MLatom is shown in Fig. 5.3(b), in which three types of tasks are implemented, including converting molecular coordinates into descriptors, sampling points from the dataset with built-in sampling procedure, and perform the ML operations. The MLatom is proposed with various physics/chemistry-specific features, including converting molecular composition and geometry to ML input vectors, enforcing invariance to atom permutations, self-correction, and efficient parallel implementation of KRR [401,402], farthest-point and structure-based sampling, model selection and evaluation. The MLatom also provides a series of molecular descriptors and various kernel functions of KRR, such as Gaussian, Laplacian and Matérn [51,239,249,403,404]. Currently, the MLatom has been used for

PESs construction, energy gradients calculation, nonadiabatic excited-state dynamics running, quantum chemical methods designing, and chemical space exploration.

Subsequently, Haghighatlari et al. [377] (2020) introduced an open-source ML and informatics software package (ChemML) for performing various data-driven researches in the chemical and materials domain, including the analysis, mining, validation and modeling. Fig. 5.3(c) presents the workflow of ChemML, in which a host of methods are incorporated to support the core tasks. The ChemML is implemented with three ML methods, including active learning (AL) [405] for minimizing the dataset size, TL [406] for generating high-quality data derived prediction models with a combination of a large set lower-quality data and a small set of high-quality data, and auto ML for automatically selecting the hyperparameters. These ML methods promise the key features, such as automation, general-purpose utility, versatility, and user-friendliness, which ChemML a viable and widely accessible tool in computational physics and chemistry. Apart from various applications in computational physics and chemistry, these software packages (PES-Learn, MLatom, and ChemML) are also designed to facilitate methodological innovation, and have become the cornerstones of the software ecosystem for data-driven in silico research.

Chemical reaction prediction

The process of identifying a suitable reaction pathway which transforms a series of available reactants into a target compound is usually achieved by expert chemists with years or decades of experience. For decades, much effort has been made for computer-aided synthesis design. Since Corey and Wipke [407,408] introduced the Logic and Heuristics Applied to Synthetic Analysis (LHASA) [409], a computer-assisted approach for codifying retrosynthesis involved the explicit identification of molecular structures which lend themselves to disconnection or can be produced by known reactions in the forward direction, a series of retrosynthetic planning approaches has been proposed, such as the Computer-Assisted Mechanistic Evaluation of Organic Reactions (CAMEO) [410], SOPHIA [411], and Eros [412]. However, these retrosynthetic planning approaches require reaction templates, which are submolecular patterns (pattern-matching rules) that encode changes in atom connectivity and are recursively applied to a target molecule to produce a candidate synthesis tree. The application of retrosynthetic templates does not always lead to successful forward synthesis. Manual encoding of these templates strongly relies on the intuition and experience of a small number of chemists and is not scalable. Because of this, modern descriptors and ML methods have been used to design syntheses, such as graph-based representations used by Kayala et al. [80,413], NNs used by Wei et al. for predicting reaction outcomes [414], and the knowledge-graph approach used by Segler and Waller for generating possible products [415]. In 2017, Coley et al. [79] proposed a model framework (“Ochem_predic_nn” model) for predicting reaction outcomes, which incorporates traditional reaction templates and NNs for flexibility pattern recognition. Fig. 5.4A shows the model framework, which contains two main steps: the forward enumeration step for applying overgeneralized forward reaction templates to a pool of reactants to generate a series of chemically plausible products, and the candidate ranking step for estimating which candidate product is the major product as a multiway classification problem using ML methods. Based on 15000 experimental reaction records from the USPTO database, this model is evaluated by training a model and predicting the major product, which presents the major product rank 1 in 71.8% of cases. ML plays an essential role in computer-aided synthesis design, not only as a key component of automated inverse synthesis programs but also as a stand-alone tool for chemists to assess the feasibility of reactions. In 2020, based on unsupervised assisted NNs, Zhang et al. [378] introduced an unsupervised assisted directional design of chemical reactions (UADDCR) software for determining the probability of a chemical reaction to be conducted smoothly at certain conditions, including the chemical equations, facets and surface compositions. Fig. 5.4B presents the workflow of UADDCR software, which contains four steps: the inputs step (step 1), containing the reactant, product, facet, surface composition, and reaction energy for optional, one-hot encoding step (step 2) for transforming the inputs into encoding information, K-means clustering and automatically classifying or the SVM classification step (step 3), and the prediction step (step 4 for) of activation energies.

As we discussed above, ML methods have been widely used for PES construction, property prediction, automatic ML model proposal, and reaction prediction and design, leading to a series of relevant software and packages. In the following sections, three representative software programs (NN-FQM, UADDCR, and QSAR-Co-X) with the GUI interface are introduced in detail.

5.1. NN-FQM: Neural network (NN)-based fragmental QM software for molecular crystal prediction

As we discussed in Sections 3.1 and 4.1, high-level QM methods and fragment-based methods have been widely used in molecular crystal systems for structure optimization, predictions of Gibbs energy, atomic forces, bond lengths, vibrational spectra and the phase transition under extreme pressure conditions. In recent years, ML approaches have made rapid and substantial progress in the applications of property prediction, PES construction, and force field construction. Based on this, Han et al. [183] (2020) proposed a new NN-based fragmental QM (NN-FQM) approach to efficiently and accurately construct the PESs of low-order many-body terms and the energy derivative and to further optimize the crystal structures and predict the properties at high pressure, thus predicting the phase transition between two crystal phases. Fig. 5.5 shows an illustration of the NN-FQM workflow, which contains the PES construction module (Fig. 5.5(a)) based on a combination of fragmental methods and NNs, and the application module (Fig. 5.5(a)) for optimizing the crystal structure, predicting the energy, atomic forces, bond length, and phase transition. In the PES construction module, a particularly designed fragmental method is used to decompose the crystal structures into a series of one- and two-body fragments, whose

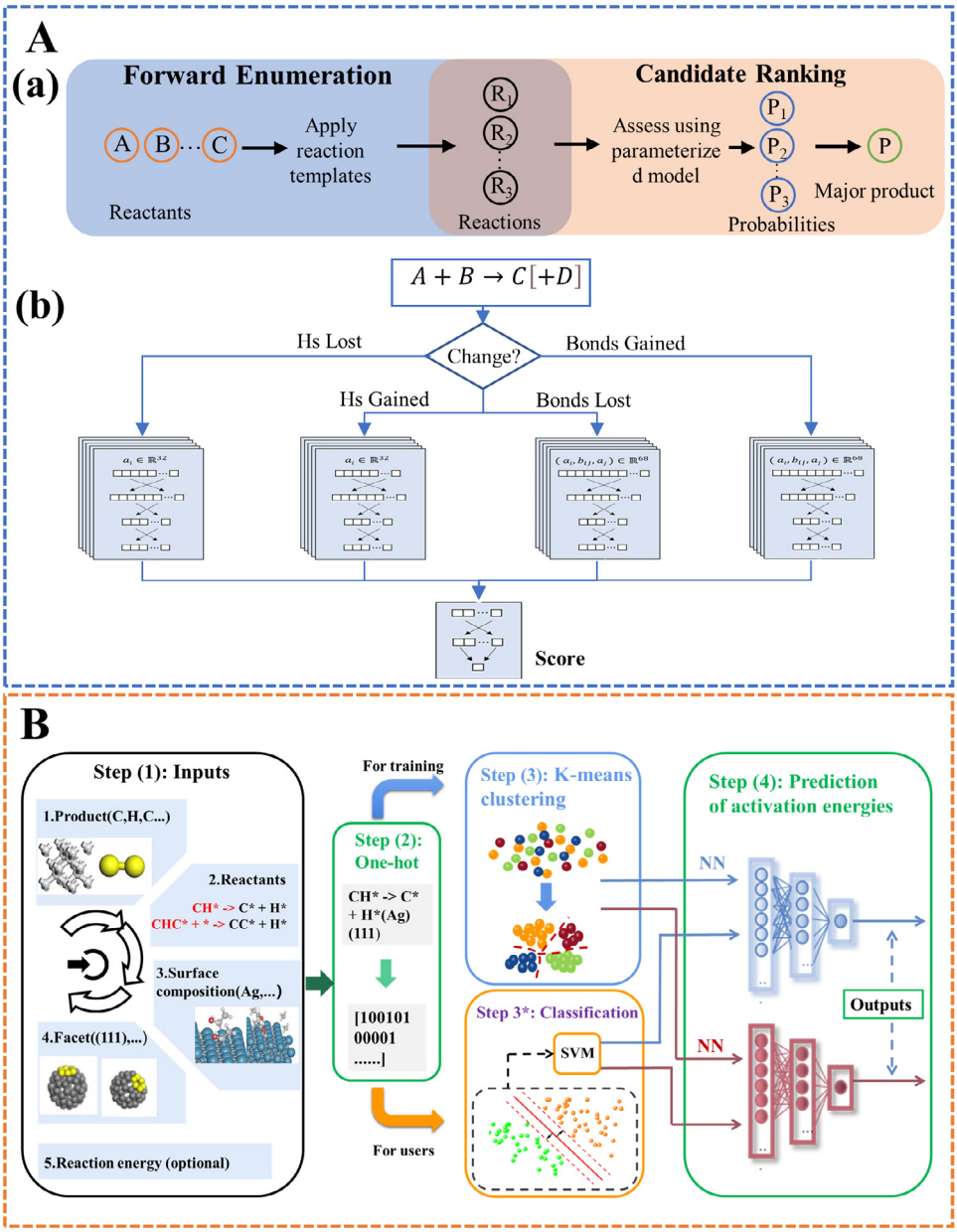

Fig. 5.4. Workflow of (A) "Ochem_predic_nn" model framework, and (B) UADDCR software. The candidate ranking in "Ochem_predic_nn" model framework A(a) is achieved by the edit-based model architecture (b).
Source: These figures were reproduced from these reports of Coley et al. [79], and Zhang et al. [378].

energy and derivation are calculated by high-level QM methods such as MP2 and CCSD(T). MD simulations are used for structure sampling to generate a large number of structures, while Gaussian descriptors are used for the transformation of the atomic positions that typically produces feature vectors (fingerprints) to describe the atomic system, which are suitable to be fed into various ML models to obtain a specified output [416,417]. The descriptors are then fed into an NN model for PES construction, which is implemented by the Atomistic Machine-learning Package (Amp) [371]. During model training, the two hidden layers, each with 15 nodes, are fully connected, which is determined by the NN to predict the potential energies via molecular descriptors. The activation function of the hyperbolic tangent function is used in each node, with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) [418] optimizer to improve the training speed while strengthening the optimization. Notably, for crystals composed of the same molecule, the PES model only needs to be trained once, or replaced by external PESs. In the application module, a crystal structure is divided into a series of fragments that are

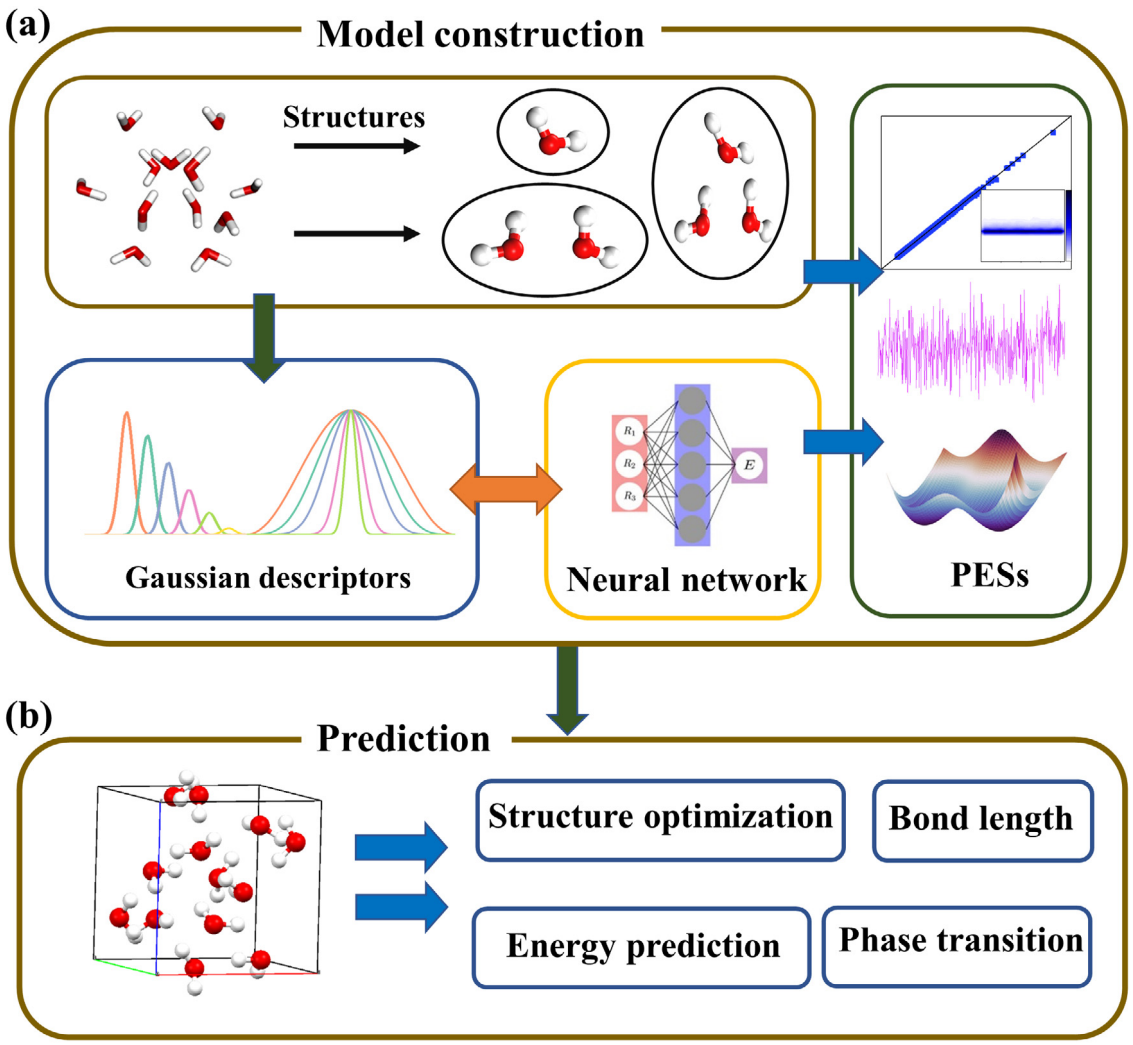

Fig. 5.5. Illustration of NN-FQM workflow. (a) the PESs construction module based on NNs, and (b) the property prediction module for optimizing crystal structure, predicting energy, atomic forces, bond length, and phase transition.
Source: Reproduced from Han et al.'s work [183].

treated with the PESs, and thus the energies, atomic forces, and properties can be obtained by summarizing the results of these fragments.

The performance of NN-FQM is evaluated by applying it to ice crystals (IX and XV) under high-pressure conditions. In total, 105000 one-body and 109025 two-body data of ice molecules generated at a pressure range of 0.1 GPa to 2.0 GPa are used to construct the PESs, which present considerably small RMSEs (0.0029 kcal/mol and 0.0429 kcal/mol for one-body and two-body, respectively). In further crystal predictions, NN-FQM successfully reproduced the crystal structure parameters, Gibbs free energies, and covalent bond lengths and predicted the boundary of the phase transition between ice IX and XV at a high pressure, which agrees well with the experimental results. NN-FQM is designed as an easy-to-use open-source visualization software, the interface of which is shown in Fig. 5.6.

In particular, the PESs of ice molecules are embedded in NN-FQM software. For an investigation of ice crystals, users are required to load the crystal structures, which can be viewed in the structure window. After the parameter setting, several properties and phase transitions can be obtained by clicking on the button. Notably, to improve the accuracy of prediction, users are also allowed to load external PESs from other ML models. Apart from ice, the NN-FQM software is capable for the structure optimization and property predictions of a wide range of molecules and molecular crystals (such as CO_2, NH_3, N_2,) by loading corresponding PESs without any change in the software. Furthermore, the NN-FQM software is designed based on an MBE fragmental scheme, and can be easily modified with other fragmental schemes for wider application.

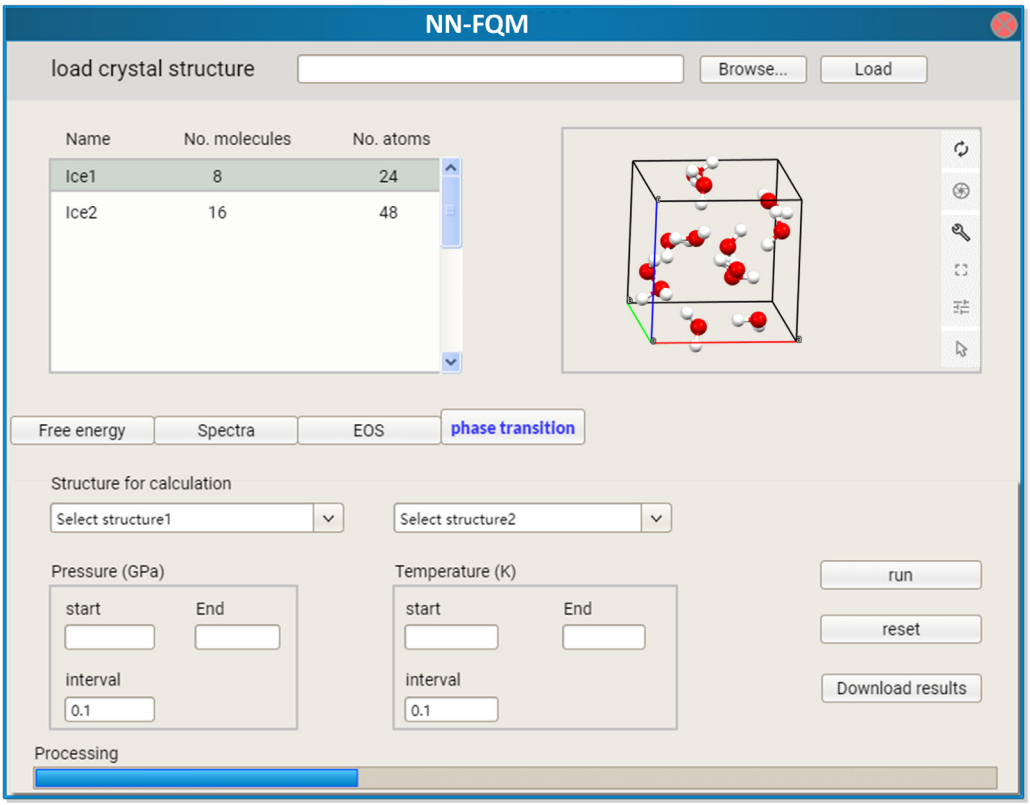

Fig. 5.6. Interface of NN-FQM software. Computer software copyright registration certificate: No. 08593048, China, by Jinjin Li et al.

5.2. UADDCR: unsupervised assisted directional design of chemical reactions

For decades, controlling and designing chemical reactions has remained a great challenge due to the high trial and error cost in experiments and the extreme computational cost of QM calculations. The process of identifying a suitable reaction pathway for transforming a series of available reactants into a target compound is usually implemented by expert researchers with extensive relevant experience. For years, electronic structure methods such as DFT have been commonly used to obtain the activation energies of chemical reactions. [419] However, high-level QM methods suffer large computational complexity scaling, which limits the applicability for large datasets with a large number of reactions. In recent years, ML methods have been successfully applied to accelerate QM calculations, such as the prediction of energy, atomic forces, chemical shifts and the construction of PESs. Efforts have also been made to predict the activation energy of reactions based on ML methods. Several works have reported using NN methods for activation energy prediction with the input features calculated using DFT [81,420]. To avoid DFT calculations for reaction features, a series of descriptors are proposed to express the features of chemical reactions [421–424], including molecular fingerprints, such as the widely used Morgan fingerprint, a Coulomb matrix and extended circular fingerprint, as well as physical representations, such as partial charge, the number of rings and molecular weight.

In 2020, Zhang et al. [378] proposed UADDCR visualization software to determine the possibility of chemical reactions with inputs of reactants, reaction equations, products, facets and surface compositions. UADDCR is designed to automatically predict the possibility of chemical reactions by activation energy prediction based on the ML method and statistical analysis, including NNs, one-hot [425] encoding, K-means clustering [426,427], and SVM [428]. Fig. 5.4B presents the overall workflow of the UADDCR software, which contains four steps: inputs, one-hot encoding, clustering or classification, and the prediction of activation energies. In step 1, the reaction equation, reactant, product, facet, and surface composition are the standard input parameters for the rapid prediction process, while the input of the reaction energy is an optional parameter for improving the prediction performance. The reaction energy can then be obtained from experiments or *ab initio* calculations. In step 2, the input parameters, such as the reactant, reaction equation, product, and surface composition, are transformed by one-hot encoding from letter strings to digital form for further model training. In step 3, based on the input features, the K-means approach is adopted to train the model and automatically cluster data into different categories. In step 4, based on the classifications NNs are used for predicting the activation energies and thus the possibility of the reactions.

To evaluate the performance of this UADDCR approach, an activation energy prediction model was proposed based on 886 reaction data points. Fig. 5.7A shows the analysis of these reaction data. The activation energy frequency distribution (Fig. 5.7A(a)) shows that 95.59% of the data are located in the activation energy range from −1 eV to 4 eV, with an average value of 1.24 eV, a minimum value of −2.47 eV, and a maximum value of 7.00 eV. The Brønsted–Evans–Polanyi (BEP) relationship (Fig. 5.7A(b)) shows that a high activation energy is usually accompanied by a high reaction energy and that surface compositions and facets also significantly influence the activation energy and reaction energy. Based on these analyses, the reactant, product, surface compositions and facets are selected as the model inputs. In addition, to avoid model instability resulting from excessive features, K-means clustering and SVM classification are performed before NN model training and activation energy prediction. Fig. 5.7B presents the analysis of the clustering results, including the two-dimensional t-distributed stochastic neighbor embedding (t-SNE) [429,430] visualization of groups A and B (Fig. 5.7B(a)), the Gaussian distribution of activation energies in different groups (Fig. 5.7B(b)), the cumulative possibility of group A and group B depending on the surface composition (Fig. 5.7B(c)), and the cumulative possibility of group A and group B depending on the facet (Fig. 5.7B(d)). In practice, UADDCR presents a good performance in predicting the activation energy. For example, the activation energy of the reaction $CH_4 \rightarrow CH_3 + H$ predicted by UADDCR software is 0.69 eV, which is much lower than that predicted by DFT calculations [431] (0.76 eV), but close to the experimental results [432] (0.54 eV).

The UADDCR software is developed as an easy-to-use visualization software, which contains four standard inputs, including reactant, product, facet and surface composition, and an advanced input (reaction energy) for optional. The interface of UADDCR is shown in Fig. 5.8, in which the predictions of activation energy and chemical reaction possibility are shown on the right side of the interface and the reaction with activation energy larger than 1.2 eV is considered as low probability. With the reliability and usability, this software is expected to be a power tool for reaction prediction and design. Furthermore, benefit from the one-hot encoding, the model in UADDCR can be simply improved by incorporating additional data of reactions, such as reaction temperature, environmental acidity. The authors also state that the UADDCR model can be further improved by expanding the dataset, considering the relationship between two chemical reactions, and incorporating more features. In addition, the UADDCR software was originally designed to predict catalytic reactions, and can be extended to other chemical reactions and even biological reactions, such as enzymatic reactions.

5.3. QSAR-Co-X: an open-source toolkit for multitarget QSAR (mt-QSAR) modeling

Quantitative structure–activity relationship (QSAR) modeling is a well-known and essential computational technique that has been proven to be extremely powerful with a wide range of applications in research fields, including physics, chemistry and materials science [433,434]. QSAR can be used for screening desirable lead chemicals and providing hints to improve the physical and chemical properties of interest. For years, ML methods [383] (such as RF, NN, and DL) and the Monte Carlo method [435–437] have been successfully applied to perform QSAR modeling. Despite the significant progress, QSAR modeling still suffers some limitations in practice [438]. In general, QSAR models can be divided into two types: classification-based models for establishing a relationship between the descriptors and the categorical values of the response variables and regression-based models for finding the relationship between the descriptors and the quantitative values of the response variables [439]. In 2019, Ambure et al. [384] developed open-source standalone software (QSAR-Co) for constructing classification-based QSAR models, which allows mining the response data coming from multiple conditions. The overall workflow of QSAR-Co is shown in Fig. 5.2A, which contains two main modules, including module 1 (Fig. 5.2A(a)) for model development and module 2 (Fig. 5.2A(b)) for screening and prediction. During the model development module (module 1), the users are provided a total of eight steps for developing a classification-based QSAR model, including step 1 for the selection between the Normal approach [440,441] and the Box–Jenkins approach [391,442–446] for model building, an optional step 2 for the data retreatment to remove noninformative descriptors that may not have a significant contribution for model building, step 3 for the dataset division based on random or rational approaches (such as Kennard-Stone's algorithm [447] and Euclidean distance-based algorithm [448]), step 4 for removal of less-discriminating descriptors identified by the molecular spectrum analysis approach [448], step 5 for variable selection of the genetic algorithm (GA) [385,449–451], step 6 for the selection of model development approach between the two-class linear discriminant analysis (LDA) [387] technique and RF [388] method, step 6 for model selection and model validation, and step 8 for defining the applicability domain (AD) based on the standardization technique [452] or confidence estimation approach [448,453]. In the screening and prediction module (module 2), users can perform three steps for screening the query chemicals, including step 1 for providing the required input, step 2 for selecting the appropriate model development approaches, and step 3 for screening that provides the predicted class for query compounds and the applicability domain status for every query compound. QSAR-Co software has been demonstrated to be significantly predictive for previously reported datasets [391,444,454].

In general, mt-QSAR modeling is strongly dependent on the strategies (such as methods and descriptors) used for model construction, which mainly results from the fact that the number of input descriptors grows with the experimental and theoretical conditions. However, employing numerous strategies for mt-QSAR modeling will significantly improve the usefulness and scope. In that case, Halder and Cordeiro [379] (2021) developed open-source visualization software (QSAR-Co-X) for modeling mt-QSAR based on the Box–Jenkins approach [390,391,454]. The QSAR-Co-X toolkit, an improved version of QSAR-Co, provides several functionalities for dataset selection, curation plus computation of descriptors, linear

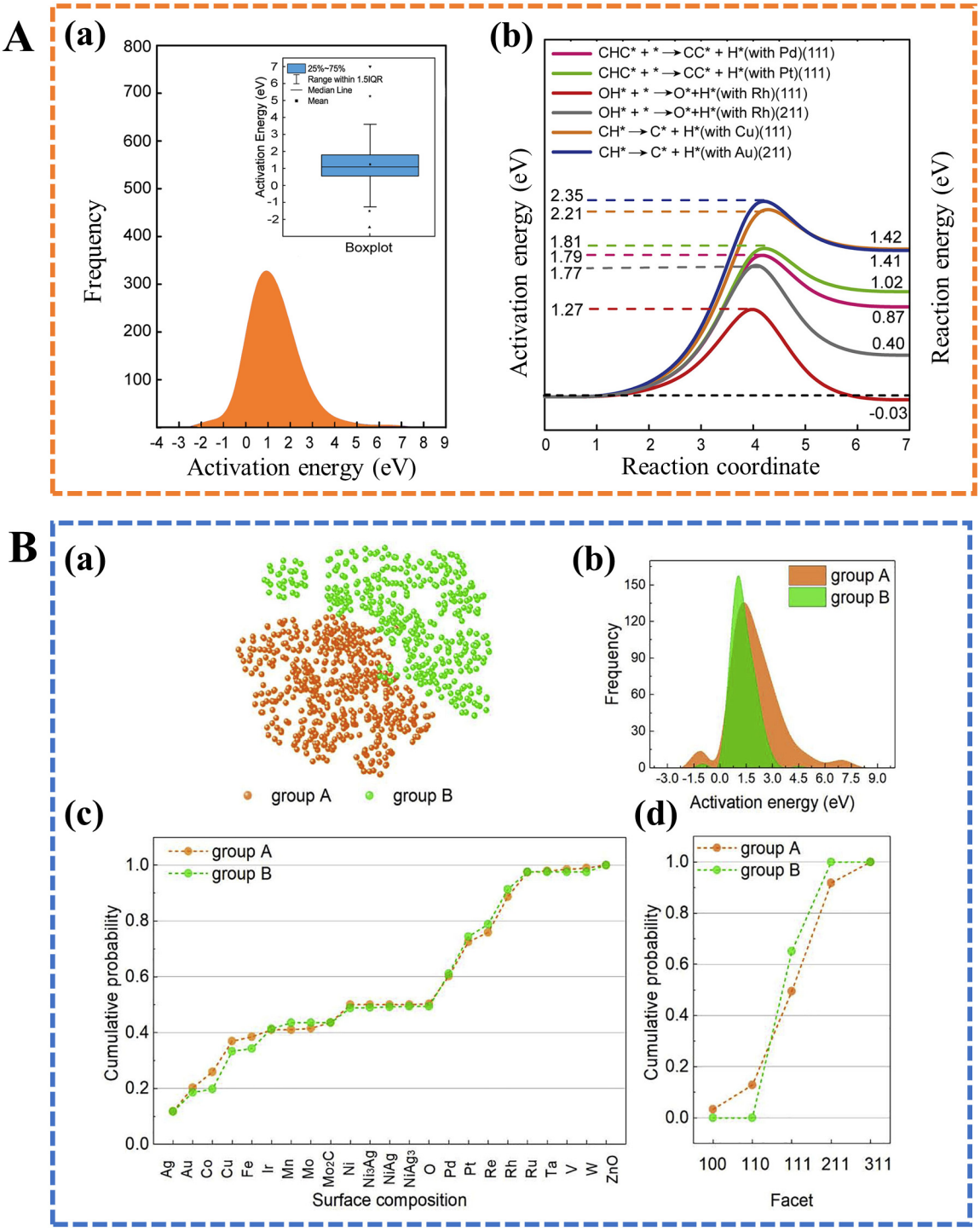

Fig. 5.7. Data analysis of training dataset and clustering result in UADDCR software. (A) Analysis of reaction data, including (a) the activation energy frequency distribution, and (b) the Brønsted–Evans–Polanyi (BEP) relationship for six different reactions. (B) Analysis of clustering results, including (a) the two-dimensional t-distributed stochastic neighbor embedding (t-SNE) visualization of groups A and group B, (b) the activation energies distribution for different groups, (c) the cumulative possibility of the two groups (A and B) depending on the surface composition, and (d) the cumulative possibility of the two groups (A and B) depending on the facet.
Source: These figures were reproduced from Zhang et al.'s work [378].

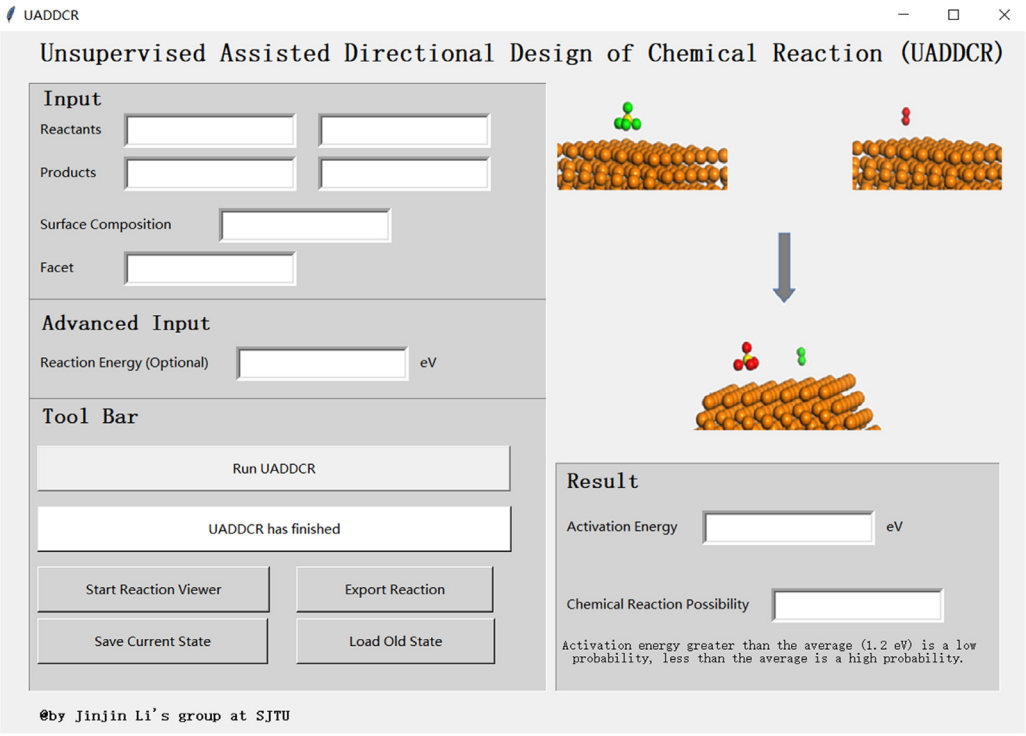

Fig. 5.8. Visualization interface for UADDCR software, which contains four standard inputs, including reactant, product, facet and surface composition, and an advanced input (reaction energy) for optional. The predictions of activation energy and chemical reaction possibility are shown on the right side of the interface. The reaction with activation energy larger than 1.2 eV is considered as low probability. Computer software copyright registration certificate: No. 6048003, China, by Jinjin Li et al.

or non-linear model construction, and a comprehensive results analysis. Fig. 5.9(a) shows the function of this QSAR-Co-X software, which contains four separating modules, including module 1 for linear modeling (LM), module 2 for non-linear modeling with a grid (NLG) search, module 3 for non-linear modeling with user-specific parameters (NLU), and module 4 for condition-wise prediction (CWP). In module 1, six steps are adopted to construct the linear models, including dataset segmentation, calculation of the input deviation descriptors, data pretreatment, LDA model training with the sequential stepwise (SFS) and fast stepwise (FS) feature selection algorithms [455], model validation by goodness-of-fit [456] and by internal and external validation criteria [452], and performance evaluation by Y-randomization with conditions scheme. Module 2 is designed to construct nonlinear models using the grid search with the hyperparameter optimization approach, which provides six ML methods for the implementation, including kNN [457], Bernoulli naïve Bayes (NB) classifier [458], support vector classifier (SVC) [459], RF [176], gradient boosting (GB) [460], and multilayer perceptron (MLP) NNs [461]. The NLU module (module 3) allows users to construct nonlinear models (NLMs) with specific parameter settings that can be used for fast generation of NLMs. Module 4 is designed as an automatic and simple tool for checking the mt-QSAR obtained results. An illustration of the workflow of QSAR-Co-X software is shown in Fig. 5.2(b).

As a visualization toolkit, QSAR-Co-X also provides a GUI interface for easy use, as shown in Fig. 5.9(b). The QSAR-Co-X toolkit provides various features, including dataset division options, Box–Jenkins moving average operators, feature representation approaches, ML models, hyperparameter tuning for ML algorithms, *Yc*-randomization, correlation matrix analyses, and condition-wise-prediction. To evaluate the functionalities, the QSAR-Co-X toolkit is applied to mt-QSAR modeling on four previously reported datasets, which presents a considerably good performance. The QSAR-Co-X toolkit implements a series of additional functions and provides a well-designed and useful platform for mt-QSAR modeling and is expected to make significant contributions to mt-QSAR modeling with a wide range of applications.

6. Conclusion

As discussed above, all-atom QM methods and fragment-based QM methods have undergone tremendous progress in the field of computational physics and chemistry. Thanks to the significant development of ML methods and chemical descriptors, in recent years numerous ML-driven models, algorithms, software and platforms have been developed for various applications in physics and chemistry, including PES construction, CSP, chemical reaction prediction, and property prediction, such as chemical shift, interaction energy, and thermal properties. In this review we introduced the recent

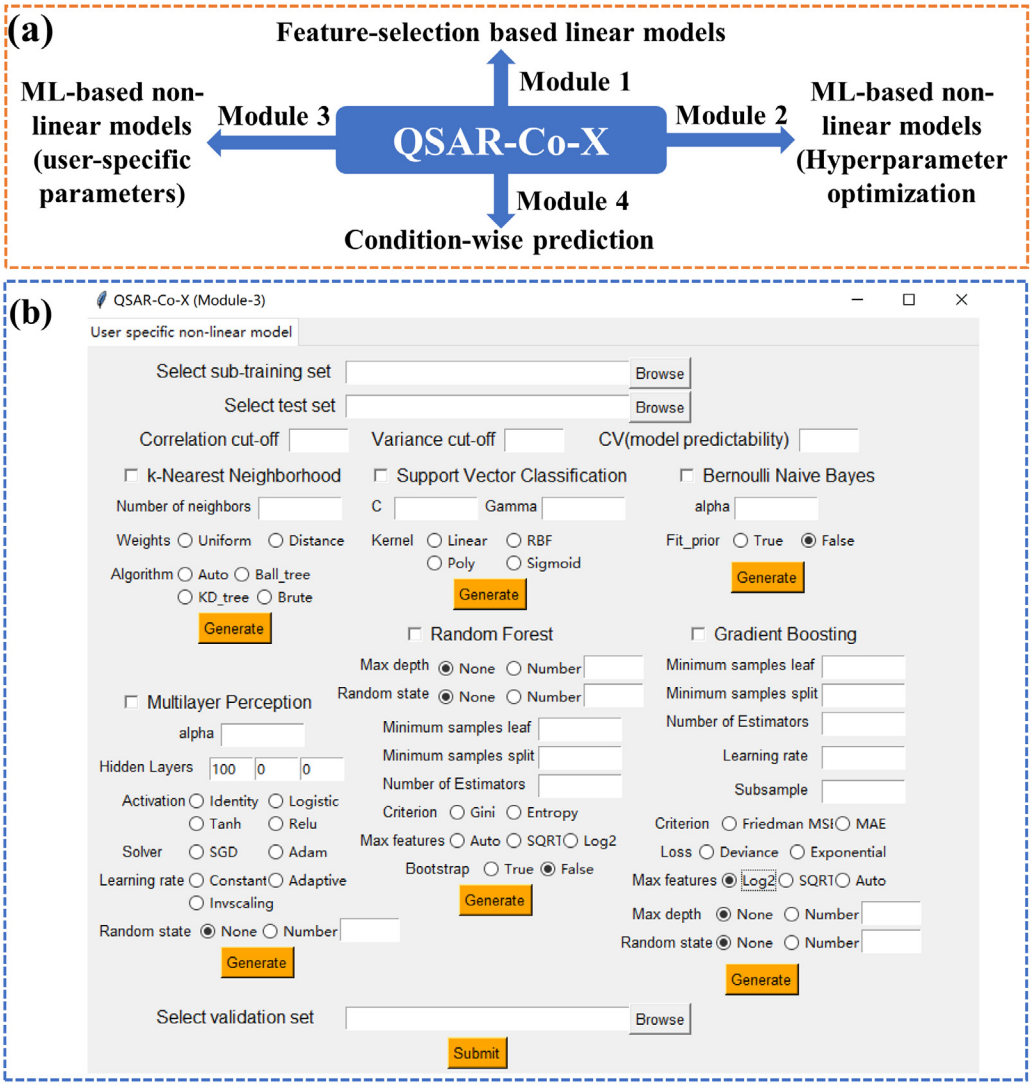

Fig. 5.9. Illustration and graphic user interface of QSAR-Co-X toolkit. (a) The functions of four modules and (b) the graphic user interface of module 3 (NLU).
Source: Figure (b) was a screenshot of software from Halder et al.'s work [379].

progress in computational methods for physical and chemical calculations, including all-atom QM methods (DFT, MP2 and CCSD(T)), fragment-based methods, and ML-based QM methods. We also reviewed the wide range of applications of these methods, including predictions of small inorganic molecules and crystals (such as property predictions of small molecules and crystals at a high pressure and negative pressure based on QM methods, ML-driven PESs, and property prediction based on ML), predictions of drug molecules (such as property and phase transformation predictions based on QM methods, property prediction based on ML methods, ML PES construction and ML-driven CSP), and ML-driven software and packages for implementing ML in physical and chemical applications. ML algorithms have become an essential and powerful tool for accelerating QM calculations in computational physics and chemistry.

However, there are many challenges for property predictions using ML methods with the accuracy of high-level QM methods. For example, several more large and high-quality datasets are required for PES construction and property prediction. On the other hand, new powerful ML methods only rely on a small amount of data and are essential for computational savings and wide applications. Proposing reliable models based on small datasets is challenging, which makes it crucial to achieve data efficiency with a considerable robustness and accuracy for ML methods by including prior physical knowledge and invariance information. In addition, the selection bias encoded in many of the training sets used in the field is another serious problem lurking behind rigorous and robust statistical learning procedures, such as k-fold cross-validation and convergent learning curves. Stability or attribute distributions are often unknown, thus hindering a rigorous assessment of the extent to which any given dataset is truly representative of the broader chemical space.

Processing big data at high volume, velocity and veracity with great versatility is also a challenge for ML algorithms. Predictably, progress in ML methods, high-quality data acquisition and data representation will advance the accuracy and efficiency for chemical shifts and other property predictions, as well as PES constructions. Furthermore, a series of ML-driven software and packages will promote the development of ML models and applications in a wide range of fields.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work is supported by the National Key R&D Program of China (No. 2021YFC2100100), the National Natural Science Foundation of China (No. 21901157), and the SJTU Global Strategic Partnership Fund (2020 SJTU-HUJI).

References

[1] S. Hirata, Fast electron-correlation methods for molecular crystals: An application to the α, $\beta 1$, and $\beta 2$ modifications of solid formic acid, J. Chem. Phys. 129 (2008) 204104, http://dx.doi.org/10.1063/1.3021077.
[2] O. Sode, S. Hirata, Second-order many-body perturbation study of solid hydrogen fluoride under pressure, Phys. Chem. Chem. Phys. 14 (2012) 7765–7779, http://dx.doi.org/10.1039/C2CP40236J.
[3] O. Sode, S. Hirata, Second-order many-body perturbation study of solid hydrogen fluoride, J. Phys. Chem. A 114 (2010) 8873–8877, http://dx.doi.org/10.1021/jp102721j.
[4] J.D. Hartman, G.J.O. Beran, Fragment-based electronic structure approach for computing nuclear magnetic resonance chemical shifts in molecular crystals, J. Chem. Theory Comput. 10 (2014) 4862–4872, http://dx.doi.org/10.1021/ct500749h.
[5] J.D. Hartman, S. Monaco, B. Schatschneider, G.J.O. Beran, Fragment-based 13C nuclear magnetic resonance chemical shift predictions in molecular crystals: An alternative to planewave methods, J. Chem. Phys. 143 (2015) 102809, http://dx.doi.org/10.1063/1.4922649.
[6] J.D. Hartman, R.A. Kudla, G.M. Day, L.J. Mueller, G.J.O. Beran, Benchmark fragment-based ^{1}H, ^{13}C, ^{15}N and ^{17}O chemical shift predictions in molecular crystals, Phys. Chem. Chem. Phys. 18 (2016) 21686–21709, http://dx.doi.org/10.1039/C6CP01831A.
[7] J. Li, O. Sode, G.A. Voth, S. Hirata, A solid–solid phase transition in carbon dioxide at high pressures and intermediate temperatures, Nature Commun. 4 (2013) 2647, http://dx.doi.org/10.1038/ncomms3647.
[8] O. Sode, M. Keçeli, K. Yagi, S. Hirata, Fermi resonance in solid CO_2 under pressure, J. Chem. Phys. 138 (2013) 074501, http://dx.doi.org/10.1063/1.4790537.
[9] G. Cardini, V. Schettino, Comment on "Fermi resonance in solid CO_2 under pressure" [J. Chem. Phys. 138, 074501 (2013)], J. Chem. Phys. 140 (2014) 177101, http://dx.doi.org/10.1063/1.4873690.
[10] J. Li, O. Sode, S. Hirata, Second-order many-body perturbation study on thermal expansion of solid carbon dioxide, J. Chem. Theory Comput. 11 (2015) 224–229, http://dx.doi.org/10.1021/ct500983k.
[11] Y.N. Heit, K.D. Nanda, G.J.O. Beran, Predicting finite-temperature properties of crystalline carbon dioxide from first principles with quantitative accuracy, Chem. Sci. 7 (2016) 246–255, http://dx.doi.org/10.1039/C5SC03014E.
[12] S. Hirata, O. Sode, M. Keçeli, K. Yagi, J. Li, Response to comment on 'Fermi resonance in solid CO_2 under pressure' [J. Chem. Phys. 140, 177101 (2014)], J. Chem. Phys. 140 (2014) 177102, http://dx.doi.org/10.1063/1.4873692.
[13] Y. Han, J. Liu, J. Li, Molecular structure determination of solid carbon dioxide phase IV at high pressures and temperatures based on Møller-Plesset perturbation theory, Int. J. Quantum Chem. 120 (2020) e26397, http://dx.doi.org/10.1002/qua.26397.
[14] L. Huang, Y. Han, X. He, J. Li, Ab initio-enabled phase transition prediction of solid carbon dioxide at ultra-high temperatures, RSC Adv. 10 (2020) 236–243, http://dx.doi.org/10.1039/C9RA06478H.
[15] Y. Han, J. Liu, L. Huang, X. He, J. Li, Predicting the phase diagram of solid carbon dioxide at high pressure from first principles, NPJ Quantum Mater. 4 (2019) 1–7, http://dx.doi.org/10.1038/s41535-019-0149-0.
[16] K.D. Nanda, G.J.O. Beran, What governs the proton ordering in ice XV? J. Phys. Chem. Lett. 4 (2013) 3165–3169, http://dx.doi.org/10.1021/jz401625w.
[17] J. Xu, J. Liu, J. Liu, W. Hu, X. He, J. Li, Phase transition of ice at high pressures and low temperatures, Molecules 25 (2020) 486, http://dx.doi.org/10.3390/molecules25030486.
[18] Q. Lu, I. Ali, J. Li, Prediction of properties from first principles with quantitative accuracy: Six representative ice phases, New J. Chem. 44 (2020) 21012–21020, http://dx.doi.org/10.1039/D0NJ04687F.
[19] R. Xiao, L. Huang, Y. Han, J. Liu, J. Li, Ab initio phase transition prediction for ices XV/XIV/VIII at high pressures and low temperatures, Chem. Phys. Lett. 760 (2020) 138015, http://dx.doi.org/10.1016/j.cplett.2020.138015.
[20] Q. Lu, J. Ren, J. Li, Structures, stabilities and phase diagram assessments of clathrate ices at negative pressures, Phys. Lett. A 401 (2021) 127330, http://dx.doi.org/10.1016/j.physleta.2021.127330.
[21] Q. Lu, J. Li, Superconducting and superhard ice, ChemPhysChem 21 (2020) 2012–2018, http://dx.doi.org/10.1002/cphc.202000582.
[22] L. Huang, Y. Han, J. Liu, X. He, J. Li, Ab initio prediction of the phase transition for solid ammonia at high pressures, Sci. Rep. 10 (2020) 7546, http://dx.doi.org/10.1038/s41598-020-64030-3.
[23] T. Yagasaki, S. Saito, I. Ohmine, A theoretical study on decomposition of formic acid in sub- and supercritical water, J. Chem. Phys. 117 (2002) 7631–7639, http://dx.doi.org/10.1063/1.1509057.
[24] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[25] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117, http://dx.doi.org/10.1016/j.neunet.2014.09.003.
[26] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, MIT press Cambridge, 2016.
[27] D. Wong, S. Yip, Machine learning classifies cancer, Nature 555 (2018) 446–447, http://dx.doi.org/10.1038/d41586-018-02881-7.
[28] F. Klauschen, K.-R. Müller, A. Binder, M. Bockmayr, M. Hägele, P. Seegerer, S. Wienert, G. Pruneri, S. de Maria, S. Badve, S. Michiels, T.O. Nielsen, S. Adams, P. Savas, F. Symmans, S. Willis, T. Gruosso, M. Park, B. Haibe-Kains, B. Gallas, A.M. Thompson, I. Cree, C. Sotiriou, C. Solinas, M. Preusser, S.M. Hewitt, D. Rimm, G. Viale, S. Loi, S. Loibl, R. Salgado, C. Denkert, Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning, Sem. Cancer Biol. 52 (2018) 151–157, http://dx.doi.org/10.1016/j.semcancer.2018.07.001.

[29] P. Jurmeister, M. Bockmayr, P. Seegerer, T. Bockmayr, D. Treue, G. Montavon, C. Vollbrecht, A. Arnold, D. Teichmann, K. Bressem, U. Schüller, M. von Laffert, K.-R. Müller, D. Capper, F. Klauschen, Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases, Sci. Transl. Med. 11 (2019) http://dx.doi.org/10.1126/scitranslmed.aaw8513.
[30] D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nature Med. 25 (2019) 954–961, http://dx.doi.org/10.1038/s41591-019-0447-x.
[31] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, K. Muller, Optimizing spatial filters for robust EEG single-trial analysis, IEEE Signal Process. Mag. 25 (2008) 41–56, http://dx.doi.org/10.1109/MSP.2008.4408441.
[32] B. Perozzi, R. Al-Rfou, S. Skiena, DeepWalk: Online learning of social representations, in: Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Association for Computing Machinery, New York, NY, USA, 2014, pp. 701–710, http://dx.doi.org/10.1145/2623330.2623732.
[33] M. Lewis, Moneyball: The Art of Winning an Unfair Game, WM Norton & Company, Inc, N. Y, 2003.
[34] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489, http://dx.doi.org/10.1038/nature16961.
[35] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, D. Hassabis, Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359, http://dx.doi.org/10.1038/nature24270.
[36] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, E.T. Mueller, Watson: Beyond jeopardy! Artificial Intelligence 199–200 (2013) 93–105, http://dx.doi.org/10.1016/j.artint.2012.06.009.
[37] W. Burgard, D. Fox, S. Thrun, Probabilistic Robotics, MIT Press, 2005.
[38] D.-O. Won, K.-R. Müller, S.-W. Lee, An adaptive deep reinforcement learning framework enables curling robots with human-like performance in real-world conditions, Sci. Robot. 5 (2020) http://dx.doi.org/10.1126/scirobotics.abb9764.
[39] P. Leinen, M. Esders, K.T. Schütt, C. Wagner, K.-R. Müller, F.S. Tautz, Autonomous robotic nanofabrication with reinforcement learning, Sci. Adv. 6 (2020) eabb6987, http://dx.doi.org/10.1126/sciadv.abb6987.
[40] T. Lengauer, O. Sander, S. Sierra, A. Thielen, R. Kaiser, Bioinformatics prediction of HIV coreceptor usage, Nature Biotechnol. 25 (2007) 1407–1410, http://dx.doi.org/10.1038/nbt1371.
[41] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710, http://dx.doi.org/10.1038/s41586-019-1923-7.
[42] K.A. Shastry, H.A. Sanjay, Machine learning for bioinformatics, in: K.G. Srinivasa, G.M. Siddesh, S.R. Manisekhar (Eds.), Stat. Model. Mach. Learn. Princ. Bioinforma. Tech. Tools Appl., Springer, Singapore, 2020, pp. 25–39, http://dx.doi.org/10.1007/978-981-15-2445-5_3.
[43] A. Serra, P. Galdi, R. Tagliaferri, Machine learning for bioinformatics and neuroimaging, WIREs Data Min. Knowl. Discov. 8 (2018) e1248, http://dx.doi.org/10.1002/widm.1248.
[44] J. Gauthier, A.T. Vincent, S.J. Charette, N. Derome, A brief history of bioinformatics, Brief. Bioinform. 20 (2019) 1981–1996, http://dx.doi.org/10.1093/bib/bby063.
[45] P. Baldi, P. Sadowski, D. Whiteson, Searching for exotic particles in high-energy physics with deep learning, Nature Commun. 5 (2014) 4308, http://dx.doi.org/10.1038/ncomms5308.
[46] B. Jiang, J. Li, H. Guo, Potential energy surfaces from high fidelity fitting of ab initio points: The permutation invariant polynomial - neural network approach, Int. Rev. Phys. Chem. 35 (2016) 479–506, http://dx.doi.org/10.1080/0144235X.2016.1200347.
[47] K. Shao, J. Chen, Z. Zhao, D.H. Zhang, Communication: Fitting potential energy surfaces with fundamental invariant neural network, J. Chem. Phys. 145 (2016) 071101, http://dx.doi.org/10.1063/1.4961454.
[48] A. Kamath, R.A. Vargas-Hernández, R.V. Krems, T. Carrington, S. Manzhos, Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy, J. Chem. Phys. 148 (2018) 241702, http://dx.doi.org/10.1063/1.5003074.
[49] J. Li, K. Song, J. Behler, A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry, Phys. Chem. Chem. Phys. 21 (2019) 9672–9682, http://dx.doi.org/10.1039/C8CP06919K.
[50] H.M. Le, S. Huynh, L.M. Raff, Molecular dissociation of hydrogen peroxide (HOOH) on a neural network ab initio potential surface with a new configuration sampling method involving gradient fitting, J. Chem. Phys. 131 (2009) 014107, http://dx.doi.org/10.1063/1.3159748.
[51] A.P. Bartók, G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, Int. J. Quantum Chem. 115 (2015) 1051–1057, http://dx.doi.org/10.1002/qua.24927.
[52] A.E. Wiens, A.V. Copan, H.F. Schaefer, Multi-fidelity Gaussian process modeling for chemical energy surfaces, Chem. Phys. Lett. X 3 (2019) 100022, http://dx.doi.org/10.1016/j.cpletx.2019.100022.
[53] A. Christianen, T. Karman, R.A. Vargas-Hernández, G.C. Groenenboom, R.V. Krems, Six-dimensional potential energy surface for NaK–NaK collisions: Gaussian process representation with correct asymptotic form, J. Chem. Phys. 150 (2019) 064106, http://dx.doi.org/10.1063/1.5082740.
[54] Z.E. Hughes, J.C.R. Thacker, A.L. Wilson, P.L.A. Popelier, Description of potential energy surfaces of molecules using FFLUX machine learning models, J. Chem. Theory Comput. 15 (2019) 116–126, http://dx.doi.org/10.1021/acs.jctc.8b00806.
[55] C.M. Handley, G.I. Hawe, D.B. Kell, P.L.A. Popelier, Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning, Phys. Chem. Chem. Phys. 11 (2009) 6365–6376, http://dx.doi.org/10.1039/B905748J.
[56] B. Jiang, H. Guo, Permutation invariant polynomial neural network approach to fitting potential energy surfaces, J. Chem. Phys. 139 (2013) 054112, http://dx.doi.org/10.1063/1.4817187.
[57] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, T. Oguchi, Crystal structure prediction accelerated by Bayesian optimization, Phys. Rev. Mater. 2 (2018) 013803, http://dx.doi.org/10.1103/PhysRevMaterials.2.013803.
[58] E.V. Podryabinkin, E.V. Tikhonov, A.V. Shapeev, A.R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, Phys. Rev. B 99 (2019) 064114, http://dx.doi.org/10.1103/PhysRevB.99.064114.
[59] D. McDonagh, C.-K. Skylaris, G.M. Day, Machine-learned fragment-based energies for crystal structure prediction, J. Chem. Theory Comput. 15 (2019) 2743–2758, http://dx.doi.org/10.1021/acs.jctc.9b00038.
[60] O. Egorova, R. Hafizi, D.C. Woods, G.M. Day, Multifidelity statistical machine learning for molecular crystal structure prediction, J. Phys. Chem. A 124 (2020) 8065–8078, http://dx.doi.org/10.1021/acs.jpca.0c05006.
[61] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2002.
[62] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408, http://dx.doi.org/10.1037/h0042519.
[63] M. Welborn, L. Cheng, T.F. Miller, Transferability in machine learning for electronic structure via the molecular orbital basis, J. Chem. Theory Comput. 14 (2018) 4772–4779, http://dx.doi.org/10.1021/acs.jctc.8b00636.

[64] W. Torng, R.B. Altman, 3D deep convolutional neural networks for amino acid environment similarity analysis, BMC Bioinformatics 18 (2017) 302, http://dx.doi.org/10.1186/s12859-017-1702-0.
[65] D. Kuzminykh, D. Polykovskiy, A. Kadurin, A. Zhebrak, I. Baskov, S. Nikolenko, R. Shayakhmetov, A. Zhavoronkov, 3D molecular representations based on the wave transform for convolutional neural networks, Mol. Pharmacol. 15 (2018) 4378–4385, http://dx.doi.org/10.1021/acs.molpharmaceut.7b01134.
[66] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, E.I. Zacharaki, EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation, PeerJ 6 (2018) e4750, http://dx.doi.org/10.7717/peerj.4750.
[67] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (1989) 541–551, http://dx.doi.org/10.1162/neco.1989.1.4.541.
[68] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (2019) 3370–3388, http://dx.doi.org/10.1021/acs.jcim.9b00237.
[69] S. Liu, J. Li, K.C. Bennett, B. Ganoe, T. Stauch, M. Head-Gordon, A. Hexemer, D. Ushizima, T. Head-Gordon, Multiresolution 3D-densenet for chemical shift prediction in NMR crystallography, J. Phys. Chem. Lett. 10 (2019) 4558–4565, http://dx.doi.org/10.1021/acs.jpclett.9b01570.
[70] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C.A. Hunter, C. Bekas, A.A. Lee, Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction, ACS Cent. Sci. 5 (2019) 1572–1583, http://dx.doi.org/10.1021/acscentsci.9b00576.
[71] K. Shakouri, J. Behler, J. Meyer, G.-J. Kroes, Accurate neural network description of surface phonons in reactive gas–surface dynamics: N2 + Ru(0001), J. Phys. Chem. Lett. 8 (2017) 2131–2136, http://dx.doi.org/10.1021/acs.jpclett.7b00784.
[72] S. Brickel, A.K. Das, O.T. Unke, H.T. Turan, M. Meuwly, Reactive molecular dynamics for the [Cl–CH_3–Br] - reaction in the gas phase and in solution: A comparative study using empirical and neural network force fields, Electron. Struct. 1 (2019) 024002, http://dx.doi.org/10.1088/2516-1075/ab1edb.
[73] M. AlQuraishi, End-to-end differentiable learning of protein structure, Cell Syst. 8 (2019) 292–301.e3, http://dx.doi.org/10.1016/j.cels.2019.03.006.
[74] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (2018) 360–365, http://dx.doi.org/10.1126/science.aat2663.
[75] Y. Wang, J.M. Lamim Ribeiro, P. Tiwary, Machine learning approaches for analyzing and enhancing molecular dynamics simulations, Curr. Opin. Struct. Biol. 61 (2020) 139–145, http://dx.doi.org/10.1016/j.sbi.2019.12.016.
[76] S. Amabilino, L.A. Bratholm, S.J. Bennie, A.C. Vaucher, M. Reiher, D.R. Glowacki, Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality, J. Phys. Chem. A 123 (2019) 4486–4499, http://dx.doi.org/10.1021/acs.jpca.9b01006.
[77] J.S. Smith, B.T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A.E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, Nature Commun. 10 (2019) 2903, http://dx.doi.org/10.1038/s41467-019-10827-4.
[78] S. Chmiela, H.E. Sauceda, I. Poltavsky, K.-R. Müller, A. Tkatchenko, sGDML: Constructing accurate and data efficient molecular force fields using machine learning, Comput. Phys. Comm. 240 (2019) 38–45, http://dx.doi.org/10.1016/j.cpc.2019.02.007.
[79] C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, K.F. Jensen, Prediction of organic reaction outcomes using machine learning, ACS Cent. Sci. 3 (2017) 434–443, http://dx.doi.org/10.1021/acscentsci.7b00064.
[80] M.A. Kayala, P. Baldi, ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning, J. Chem. Inf. Model. 52 (2012) 2526–2540, http://dx.doi.org/10.1021/ci3003039.
[81] A.R. Singh, B.A. Rohr, J.A. Gauthier, J.K. Nørskov, Predicting chemical reaction barriers with a machine learning model, Catal. Lett. 149 (2019) 2347–2354, http://dx.doi.org/10.1007/s10562-019-02705-x.
[82] S. Hirata, M. Valiev, M. Dupuis, S.S. Xantheas, S. Sugiki, H. Sekino, Fast electron correlation methods for molecular clusters in the ground and excited states, Mol. Phys. 103 (2005) 2255–2265, http://dx.doi.org/10.1080/00268970500083788.
[83] M. Kamiya, S. Hirata, M. Valiev, Fast electron correlation methods for molecular clusters without basis set superposition errors, J. Chem. Phys. 128 (2008) 074103, http://dx.doi.org/10.1063/1.2828517.
[84] G.J.O. Beran, Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields, J. Chem. Phys. 130 (2009) 164115, http://dx.doi.org/10.1063/1.3121323.
[85] P.J. Bygrave, N.L. Allan, F.R. Manby, The embedded many-body expansion for energetics of molecular crystals, J. Chem. Phys. 137 (2012) 164102, http://dx.doi.org/10.1063/1.4759079.
[86] T. Fang, W. Li, F. Gu, S. Li, Accurate prediction of lattice energies and structures of molecular crystals with molecular quantum chemistry methods, J. Chem. Theory Comput. 11 (2015) 91–98, http://dx.doi.org/10.1021/ct500833k.
[87] F. London, Zur theorie und systematik der molekularkräfte, Z. Phys. 63 (1930) 245–279.
[88] F. London, The general theory of molecular forces, Trans. Faraday Soc. 33 (1937) 8b–26.
[89] S. Grimme, Semiempirical GGA-type density functional constructed with a long-range dispersion correction, J. Comput. Chem. 27 (2006) 1787–1799, http://dx.doi.org/10.1002/jcc.20495.
[90] S. Grimme, Accurate description of van der Waals complexes by density functional theory including empirical corrections, J. Comput. Chem. 25 (2004) 1463–1473, http://dx.doi.org/10.1002/jcc.20078.
[91] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132 (2010) 154104, http://dx.doi.org/10.1063/1.3382344.
[92] S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory, J. Comput. Chem. 32 (2011) 1456–1465, http://dx.doi.org/10.1002/jcc.21759.
[93] A. Tkatchenko, M. Scheffler, Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data, Phys. Rev. Lett. 102 (2009) 073005, http://dx.doi.org/10.1103/PhysRevLett.102.073005.
[94] W. Hujo, S. Grimme, Performance of the van der Waals density functional VV10 and (hybrid) GGA variants for thermochemistry and noncovalent interactions, J. Chem. Theory Comput. 7 (2011) 3866–3871, http://dx.doi.org/10.1021/ct200644w.
[95] K.S. Thanthiriwatte, E.G. Hohenstein, L.A. Burns, C.D. Sherrill, Assessment of the performance of DFT and DFT-D methods for describing distance dependence of hydrogen-bonded interactions, J. Chem. Theory Comput. 7 (2011) 88–96, http://dx.doi.org/10.1021/ct100469b.
[96] L. Goerigk, S. Grimme, A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions, Phys. Chem. Chem. Phys. 13 (2011) 6670–6688, http://dx.doi.org/10.1039/C0CP02984J.
[97] L. Goerigk, H. Kruse, S. Grimme, Benchmarking density functional methods against the S66 and S66x8 datasets for non-covalent interactions, ChemPhysChem 12 (2011) 3421–3433, http://dx.doi.org/10.1002/cphc.201100826.
[98] T. Risthaus, S. Grimme, Benchmarking of London dispersion-accounting density functional theory methods on very large molecular complexes, J. Chem. Theory Comput. 9 (2013) 1580–1591, http://dx.doi.org/10.1021/ct301081n.
[99] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, P. Hobza, Accuracy of quantum chemical methods for large noncovalent complexes, J. Chem. Theory Comput. 9 (2013) 3364–3374, http://dx.doi.org/10.1021/ct400036b.

[100] S. Gohr, S. Grimme, T. Söhnel, B. Paulus, P. Schwerdtfeger, Pressure dependent stability and structure of carbon dioxide—A density functional study including long-range corrections, J. Chem. Phys. 139 (2013) 174501, http://dx.doi.org/10.1063/1.4826929.
[101] J. Moellmann, S. Grimme, DFT-D3 study of some molecular crystals, J. Phys. Chem. C 118 (2014) 7615–7621, http://dx.doi.org/10.1021/jp501237c.
[102] A.D. Becke, E.R. Johnson, A density-functional model of the dispersion interaction, J. Chem. Phys. 123 (2005) 154101, http://dx.doi.org/10.1063/1.2065267.
[103] E.R. Johnson, A.D. Becke, A post-Hartree–Fock model of intermolecular interactions, J. Chem. Phys. 123 (2005) 024101, http://dx.doi.org/10.1063/1.1949201.
[104] T. Sato, H. Nakai, Density functional method including weak interactions: Dispersion coefficients based on the local response approximation, J. Chem. Phys. 131 (2009) 224104, http://dx.doi.org/10.1063/1.3269802.
[105] T. Sato, H. Nakai, Local response dispersion method. II. Generalized multicenter interactions, J. Chem. Phys. 133 (2010) 194101, http://dx.doi.org/10.1063/1.3503040.
[106] A. Tkatchenko, R.A. DiStasio, R. Car, M. Scheffler, Accurate and efficient method for many-body van der Waals interactions, Phys. Rev. Lett. 108 (2012) 236402, http://dx.doi.org/10.1103/PhysRevLett.108.236402.
[107] M. Dion, H. Rydberg, E. Schröder, D.C. Langreth, B.I. Lundqvist, Van der Waals density functional for general geometries, Phys. Rev. Lett. 92 (2004) 246401, http://dx.doi.org/10.1103/PhysRevLett.92.246401.
[108] M. Dion, H. Rydberg, E. Schröder, D.C. Langreth, B.I. Lundqvist, Erratum: Van der Waals density functional for general geometries [Phys. Rev. Lett. 92, 246401 (2004)], Phys. Rev. Lett. 95 (2005) 109902, http://dx.doi.org/10.1103/PhysRevLett.95.109902.
[109] T. Thonhauser, V.R. Cooper, S. Li, A. Puzder, P. Hyldgaard, D.C. Langreth, Van der Waals density functional: Self-consistent potential and the nature of the van der Waals bond, Phys. Rev. B 76 (2007) 125112, http://dx.doi.org/10.1103/PhysRevB.76.125112.
[110] O.A. Vydrov, T. Van Voorhis, Improving the accuracy of the nonlocal van der Waals density functional with minimal empiricism, J. Chem. Phys. 130 (2009) 104105, http://dx.doi.org/10.1063/1.3079684.
[111] K. Lee, É.D. Murray, L. Kong, B.I. Lundqvist, D.C. Langreth, Higher-accuracy van der Waals density functional, Phys. Rev. B 82 (2010) 081101, http://dx.doi.org/10.1103/PhysRevB.82.081101.
[112] Y. Zhao, N.E. Schultz, D.G. Truhlar, Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions, J. Chem. Theory Comput. 2 (2006) 364–382, http://dx.doi.org/10.1021/ct0502763.
[113] Y. Zhao, N.E. Schultz, D.G. Truhlar, Exchange–correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions, J. Chem. Phys. 123 (2005) 161103, http://dx.doi.org/10.1063/1.2126975.
[114] Y. Zhao, D.G. Truhlar, Comparative DFT study of van der Waals complexes: Rare-gas dimers, alkaline-earth dimers, zinc dimer, and zinc-rare-gas dimers, J. Phys. Chem. A 110 (2006) 5121–5129, http://dx.doi.org/10.1021/jp060231d.
[115] O.A. Vydrov, T. Van Voorhis, Nonlocal van der Waals density functional: The simpler the better, J. Chem. Phys. 133 (2010) 244103, http://dx.doi.org/10.1063/1.3521275.
[116] O.A. Vydrov, T. Van Voorhis, Nonlocal van der Waals density functional made simple, Phys. Rev. Lett. 103 (2009) 063004, http://dx.doi.org/10.1103/PhysRevLett.103.063004.
[117] O.A. Vydrov, T. Van Voorhis, Vydrov and Van Voorhis reply, Phys. Rev. Lett. 104 (2010) 099304, http://dx.doi.org/10.1103/PhysRevLett.104.099304.
[118] D.C. Langreth, B.I. Lundqvist, Comment on nonlocal van der Waals density functional made simple, Phys. Rev. Lett. 104 (2010) 099303, http://dx.doi.org/10.1103/PhysRevLett.104.099303.
[119] J.A. van Santen, G.A. DiLabio, Dispersion corrections improve the accuracy of both noncovalent and covalent interactions energies predicted by a density-functional theory approximation, J. Phys. Chem. A 119 (2015) 6703–6713, http://dx.doi.org/10.1021/acs.jpca.5b02809.
[120] R.P. Feynman, Forces in molecules, Phys. Rev. 56 (1939) 340–343, http://dx.doi.org/10.1103/PhysRev.56.340.
[121] F. Manby, Accurate Condensed-Phase Quantum Chemistry, CRC press, 2010.
[122] A.F. Izmaylov, G.E. Scuseria, Resolution of the identity atomic orbital Laplace transformed second order Møller–Plesset theory for nonconducting periodic systems, Phys. Chem. Chem. Phys. 10 (2008) 3421–3429, http://dx.doi.org/10.1039/B803274M.
[123] P.Y. Ayala, K.N. Kudin, G.E. Scuseria, Atomic orbital Laplace-transformed second-order Møller–Plesset theory for periodic systems, J. Chem. Phys. 115 (2001) 9698–9707, http://dx.doi.org/10.1063/1.1414369.
[124] L. Maschio, D. Usvyat, F.R. Manby, S. Casassa, C. Pisani, M. Schütz, Fast local-MP2 method with density-fitting for crystals. I. Theory and algorithms, Phys. Rev. B 76 (2007) 075101, http://dx.doi.org/10.1103/PhysRevB.76.075101.
[125] D. Usvyat, L. Maschio, F.R. Manby, S. Casassa, M. Schütz, C. Pisani, Fast local-MP2 method with density-fitting for crystals. II. Test calculations and application to the carbon dioxide crystal, Phys. Rev. B 76 (2007) 075102, http://dx.doi.org/10.1103/PhysRevB.76.075102.
[126] M. Marsman, A. Grüneis, J. Paier, G. Kresse, Second-order Møller–Plesset perturbation theory applied to extended systems. I. Within the projector-augmented-wave formalism using a plane wave basis set, J. Chem. Phys. 130 (2009) 184103, http://dx.doi.org/10.1063/1.3126249.
[127] A. Grüneis, M. Marsman, G. Kresse, Second-order Møller–Plesset perturbation theory applied to extended systems. II. Structural and energetic properties, J. Chem. Phys. 133 (2010) 074107, http://dx.doi.org/10.1063/1.3466765.
[128] A.C. Ihrig, J. Wieferink, I.Y. Zhang, M. Ropo, X. Ren, P. Rinke, M. Scheffler, V. Blum, Accurate localized resolution of identity approach for linear-scaling hybrid density functionals and for many-body perturbation theory, New J. Phys. 17 (2015) 093020, http://dx.doi.org/10.1088/1367-2630/17/9/093020.
[129] D. Usvyat, L. Maschio, M. Schütz, Periodic local MP2 method employing orbital specific virtuals, J. Chem. Phys. 143 (2015) 102805, http://dx.doi.org/10.1063/1.4921301.
[130] M. Del Ben, J. Hutter, J. VandeVondele, Electron correlation in the condensed phase from a resolution of identity approach based on the Gaussian and plane waves scheme, J. Chem. Theory Comput. 9 (2013) 2654–2671, http://dx.doi.org/10.1021/ct4002202.
[131] M. Del Ben, J. Hutter, J. VandeVondele, Second-order Møller–Plesset perturbation theory in the condensed phase: An efficient and massively parallel Gaussian and plane waves approach, J. Chem. Theory Comput. 8 (2012) 4177–4188, http://dx.doi.org/10.1021/ct300531w.
[132] R.O. Ramabhadran, K. Raghavachari, Extrapolation to the gold-standard in quantum chemistry: Computationally efficient and accurate CCSD(T) energies for large molecules using an automated thermochemical hierarchy, J. Chem. Theory Comput. 9 (2013) 3986–3994, http://dx.doi.org/10.1021/ct400465q.
[133] H. Stoll, On the correlation energy of graphite, J. Chem. Phys. 97 (1992) 8449–8454, http://dx.doi.org/10.1063/1.463415.
[134] H. Stoll, The correlation energy of crystalline silicon, Chem. Phys. Lett. 191 (1992) 548–552, http://dx.doi.org/10.1016/0009-2614(92)85587-Z.
[135] B. Paulus, The method of increments—a wavefunction-based ab initio correlation method for solids, Phys. Rep. 428 (2006) 1–52, http://dx.doi.org/10.1016/j.physrep.2006.01.003.
[136] S. Tsuzuki, H. Orita, K. Honda, M. Mikami, First-principles lattice energy calculation of urea and hexamine crystals by a combination of periodic DFT and MP2 two-body interaction energy calculations, J. Phys. Chem. B 114 (2010) 6799–6805, http://dx.doi.org/10.1021/jp912028q.
[137] S. Wen, G.J.O. Beran, Accurate molecular crystal lattice energies from a fragment QM/MM approach with on-the-fly ab initio force field parametrization, J. Chem. Theory Comput. 7 (2011) 3733–3742, http://dx.doi.org/10.1021/ct200541h.

[138] J.F. Ouyang, M.W. Cvitkovic, R.P.A. Bettens, Trouble with the many-body expansion, J. Chem. Theory Comput. 10 (2014) 3699–3707, http://dx.doi.org/10.1021/ct500396b.
[139] R.M. Richard, K.U. Lao, J.M. Herbert, Achieving the CCSD(T) basis-set limit in sizable molecular clusters: Counterpoise corrections for the many-body expansion, J. Phys. Chem. Lett. 4 (2013) 2674–2680, http://dx.doi.org/10.1021/jz401368u.
[140] K. Gilliard, O. Sode, S. Hirata, Second-order many-body perturbation and coupled-cluster singles and doubles study of ice VIII, J. Chem. Phys. 140 (2014) 174507, http://dx.doi.org/10.1063/1.4873919.
[141] O. Bludský, M. Rubeš, P. Soldán, Ab initio investigation of intermolecular interactions in solid benzene, Phys. Rev. B 77 (2008) 092103, http://dx.doi.org/10.1103/PhysRevB.77.092103.
[142] C.R. Taylor, P.J. Bygrave, J.N. Hart, N.L. Allan, F.R. Manby, Improving density functional theory for crystal polymorph energetics, Phys. Chem. Chem. Phys. 14 (2012) 7739–7743, http://dx.doi.org/10.1039/C2CP24090D.
[143] E.E. Dahlke, D.G. Truhlar, Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller-Plesset perturbation theory energies for large water clusters, J. Chem. Theory Comput. 3 (2007) 1342–1348, http://dx.doi.org/10.1021/ct700057x.
[144] G. Briscoe, T. Caelli, A Compendium of Machine Learning, Intellect Books, 1996.
[145] F. Rosenblat, The perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408.
[146] T.M. Mitchell, Machine learning and data mining, Commun. ACM. 42 (1999) 30–36.
[147] M. Sugiyama, M. Kawanabe, Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT press, 2012.
[148] M. Sugiyama, M. Krauledat, K.-R. Müller, Covariate shift adaptation by importance weighted cross validation., J. Mach. Learn. Res. 8 (2007).
[149] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547–555, http://dx.doi.org/10.1038/s41586-018-0337-2.
[150] D. Fourches, E. Muratov, A. Tropsha, Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research, J. Chem. Inf. Model. 50 (2010) 1189–1204, http://dx.doi.org/10.1021/ci100176x.
[151] N. Kireeva, I.I. Baskin, H.A. Gaspar, D. Horvath, G. Marcou, A. Varnek, Generative topographic mapping (GTM): Universal tool for data visualization, structure-activity modeling and dataset comparison, Mol. Inform. 31 (2012) 301–312, http://dx.doi.org/10.1002/minf.201100163.
[152] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B 99 (2019) 014104, http://dx.doi.org/10.1103/PhysRevB.99.014104.
[153] A.S. Christensen, L.A. Bratholm, F.A. Faber, O. Anatole von Lilienfeld, FCHL revisited: Faster and more accurate quantum machine learning, J. Chem. Phys. 152 (2020) 044107, http://dx.doi.org/10.1063/1.5126701.
[154] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (2015) 2326–2331, http://dx.doi.org/10.1021/acs.jpclett.5b00831.
[155] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (2011) 074106, http://dx.doi.org/10.1063/1.3553717.
[156] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (2012) 058301, http://dx.doi.org/10.1103/PhysRevLett.108.058301.
[157] L. Pattanaik, C.W. Coley, Molecular representation: Going long on fingerprints, Chem 6 (2020) 1204–1207, http://dx.doi.org/10.1016/j.chempr.2020.05.002.
[158] Z. Li, S. Wang, W. Shan Chin, L.E. Achenie, H. Xin, High-throughput screening of bimetallic catalysts enabled by machine learning, J. Mater. Chem. A 5 (2017) 24131–24138, http://dx.doi.org/10.1039/C7TA01812F.
[159] E.S.R. Ehmki, R. Schmidt, F. Ohm, M. Rarey, Comparing molecular patterns using the example of SMARTS: Applications and filter collection analysis, J. Chem. Inf. Model. 59 (2019) 2572–2586, http://dx.doi.org/10.1021/acs.jcim.9b00249.
[160] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (2016) 170901, http://dx.doi.org/10.1063/1.4966192.
[161] J. Lim, S. Ryu, J.W. Kim, W.Y. Kim, Molecular generative model based on conditional variational autoencoder for de novo molecular design, J. Cheminformatics 10 (2018) 31, http://dx.doi.org/10.1186/s13321-018-0286-7.
[162] R. Schmidt, E.S.R. Ehmki, F. Ohm, H.-C. Ehrlich, A. Mashychev, M. Rarey, Comparing molecular patterns using the example of SMARTS: Theory and algorithms, J. Chem. Inf. Model. 59 (2019) 2560–2571, http://dx.doi.org/10.1021/acs.jcim.9b00250.
[163] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. 104 (2010) 136403, http://dx.doi.org/10.1103/PhysRevLett.104.136403.
[164] K.T. Schütt, P.-J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, in: Proc. 31st Int. Conf. Neural Inf. Process. Syst., Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 992–1002.
[165] K.T. Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nature Commun. 8 (2017) 13890, http://dx.doi.org/10.1038/ncomms13890.
[166] E. Kocer, J.K. Mason, H. Erturk, A novel approach to describe chemical environments in high-dimensional neural network potentials, J. Chem. Phys. 150 (2019) 154102, http://dx.doi.org/10.1063/1.5086167.
[167] S. Chmiela, H.E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, Nature Commun. 9 (2018) 3887, http://dx.doi.org/10.1038/s41467-018-06169-2.
[168] K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – A deep learning architecture for molecules and materials, J. Chem. Phys. 148 (2018) 241722, http://dx.doi.org/10.1063/1.5019779.
[169] D.J. Hand, K. Yu, Idiot's Bayes—Not so stupid after all? Int. Stat. Rev. 69 (2001) 385–398, http://dx.doi.org/10.1111/j.1751-5823.2001.tb00465.x.
[170] G. Shakhnarovich, T. Darrell, P. Indyk, Nearest-neighbor methods in learning and vision, in: Neural Inf. Process., 2005.
[171] B. Yegnanarayana, Artificial Neural Networks, PHI Learning Pvt. Ltd., 2009.
[172] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: Int. Conf. Mach. Learn., PMLR, 2017, pp. 1263–1272, http://proceedings.mlr.press/v70/gilmer17a.html. (Accessed 19 May 2021).
[173] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 Int. Conf. Eng. Technol., ICET, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICEngTechnol.2017.8308186.
[174] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 221–231, http://dx.doi.org/10.1109/TPAMI.2012.59.
[175] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.
[176] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[177] K.T. Schütt, P. Kessel, M. Gastegger, K.A. Nicoli, A. Tkatchenko, K.-R. Müller, SchNetPack: A deep learning toolbox for atomistic systems, J. Chem. Theory Comput. 15 (2019) 448–455, http://dx.doi.org/10.1021/acs.jctc.8b00908.
[178] G. Hegde, R.C. Bowen, Machine-learned approximations to density functional theory Hamiltonians, Sci. Rep. 7 (2017) 42669, http://dx.doi.org/10.1038/srep42669.

[179] J. Townsend, K.D. Vogiatzis, Data-driven acceleration of the coupled-cluster singles and doubles iterative solver, J. Phys. Chem. Lett. 10 (2019) 4129–4135, http://dx.doi.org/10.1021/acs.jpclett.9b01442.
[180] T.B. Blank, S.D. Brown, A.W. Calhoun, D.J. Doren, Neural network models of potential energy surfaces, J. Chem. Phys. 103 (1995) 4129–4137, http://dx.doi.org/10.1063/1.469597.
[181] H. Gassner, M. Probst, A. Lauenstein, K. Hermansson, Representation of intermolecular potential functions by neural networks, J. Phys. Chem. A 102 (1998) 4596–4605, http://dx.doi.org/10.1021/jp972209d.
[182] S. Lorenz, A. Groz, M. Scheffler, Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks, Chem. Phys. Lett. 395 (2004) 210–215, http://dx.doi.org/10.1016/j.cplett.2004.07.076.
[183] Y. Han, Z. Wang, J. Li, Neural networks accelerate the ab initio prediction of solid–solid phase transitions at high pressures, J. Phys. Chem. Lett. 12 (2021) 132–137, http://dx.doi.org/10.1021/acs.jpclett.0c03101.
[184] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (2007) 146401, http://dx.doi.org/10.1103/PhysRevLett.98.146401.
[185] J.S. Smith, O. Isayev, A.E. Roitberg, ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost, Chem. Sci. 8 (2017) 3192–3203, http://dx.doi.org/10.1039/C6SC05720A.
[186] K. Yao, J.E. Herr, D.W. Toth, R. Mckintyre, J. Parkhill, The tensorMol-0.1 model chemistry: A neural network augmented with long-range physics, Chem. Sci. 9 (2018) 2261–2269, http://dx.doi.org/10.1039/C7SC04934J.
[187] H. Wang, L. Zhang, J. Han, W. E., DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Comput. Phys. Comm. 228 (2018) 178–184, http://dx.doi.org/10.1016/j.cpc.2018.03.016.
[188] A. Nandi, C. Qu, P.L. Houston, R. Conte, J.M. Bowman, δ-Machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory, J. Chem. Phys. 154 (2021) 051102, http://dx.doi.org/10.1063/5.0038301.
[189] X. Gao, F. Ramezanghorbani, O. Isayev, J.S. Smith, A.E. Roitberg, TorchAni: A free and open source PyTorch-based deep learning implementation of the ANI neural network potentials, J. Chem. Inf. Model. 60 (2020) 3408–3415, http://dx.doi.org/10.1021/acs.jcim.0c00451.
[190] Z.L. Glick, D.P. Metcalf, A. Koutsoukas, S.A. Spronk, D.L. Cheney, C.D. Sherrill, AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials, J. Chem. Phys. 153 (2020) 044112, http://dx.doi.org/10.1063/5.0011521.
[191] K.D. Litasov, A.F. Goncharov, R.J. Hemley, Crossover from melting to dissociation of CO_2 under pressure: Implications for the lower mantle, Earth Planet. Sci. Lett. 309 (2011) 318–323, http://dx.doi.org/10.1016/j.epsl.2011.07.006.
[192] B. Boates, A.M. Teweldeberhan, S.A. Bonev, Stability of dense liquid carbon dioxide, Proc. Natl. Acad. Sci. 109 (2012) 14808–14812, http://dx.doi.org/10.1073/pnas.1120243109.
[193] A.R. Oganov, R.J. Hemley, R.M. Hazen, A.P. Jones, Structure, bonding, and mineralogy of carbon at extreme conditions, Rev. Mineral. Geochem. 75 (2013) 47–77, http://dx.doi.org/10.2138/rmg.2013.75.3.
[194] T. Jahnke, H. Sann, T. Havermeier, K. Kreidi, C. Stuck, M. Meckel, M. Schöffler, N. Neumann, R. Wallauer, S. Voss, A. Czasch, O. Jagutzki, A. Malakzadeh, F. Afaneh, T. Weber, H. Schmidt-Böcking, R. Dörner, Ultrafast energy transfer between water molecules, Nat. Phys. 6 (2010) 139–142, http://dx.doi.org/10.1038/nphys1498.
[195] A. Kilaj, H. Gao, D. Rösch, U. Rivero, J. Küpper, S. Willitsch, Observation of different reactivities of para and ortho- water towards trapped diazenylium ions, Nature Commun. 9 (2018) 2096, http://dx.doi.org/10.1038/s41467-018-04483-3.
[196] F. Schüth, R. Palkovits, R. Schlögl, D.S. Su, Ammonia as a possible element in an energy infrastructure: Catalysts for ammonia decomposition, Energy Environ. Sci. 5 (2012) 6278–6289, http://dx.doi.org/10.1039/C2EE02865D.
[197] G. Marnellos, M. Stoukides, Ammonia synthesis at atmospheric pressure, Science 282 (1998) 98–100, http://dx.doi.org/10.1126/science.282.5386.98.
[198] M. Lipp, W.J. Evans, V. Garcia-Baonza, H.E. Lorenzana, Carbon monoxide: Spectroscopic characterization of the high–pressure polymerized phase, J. Low Temp. Phys. 111 (1998) 247–256, http://dx.doi.org/10.1023/A:1022267115640.
[199] D.C.B. Whittet, A.J. Adamson, W.W. Duley, T.R. Geballe, A.D. McFadzean, Infrared spectroscopy of dust in the Taurus dark clouds: Solid carbon monoxide, Mon. Not. R. Astron. Soc. 241 (1989) 707–720, http://dx.doi.org/10.1093/mnras/241.4.707.
[200] P.F. Fracassi, R. Righini, R.G. Della Valle, M.L. Klein, Lattice dynamics of solid α-carbon monoxide, Chem. Phys. 96 (1985) 361–369, http://dx.doi.org/10.1016/0301-0104(85)85099-0.
[201] S.M. El-Sheikh, K. Barakat, N.M. Salem, Phase transitions of methane using molecular dynamics simulations, J. Chem. Phys. 124 (2006) 124517, http://dx.doi.org/10.1063/1.2179422.
[202] W. Sontising, Y.N. Heit, J.L. McKinley, G.J.O. Beran, Theoretical predictions suggest carbon dioxide phases III and VII are identical, Chem. Sci. 8 (2017) 7374–7382, http://dx.doi.org/10.1039/C7SC03267F.
[203] T. Bartels-Rausch, V. Bergeron, J.H.E. Cartwright, R. Escribano, J.L. Finney, H. Grothe, P.J. Gutiérrez, J. Haapala, W.F. Kuhs, J.B.C. Pettersson, S.D. Price, C.I. Sainz-Díaz, D.J. Stokes, G. Strazzulla, E.S. Thomson, H. Trinks, N. Uras-Aytemiz, Ice structures, patterns, and processes: A view across the icefields, Rev. Modern Phys. 84 (2012) 885–944, http://dx.doi.org/10.1103/RevModPhys.84.885.
[204] Y. Liu, Y. Huang, C. Zhu, H. Li, J. Zhao, L. Wang, L. Ojamäe, J.S. Francisco, X.C. Zeng, An ultralow-density porous ice with the largest internal cavity identified in the water phase diagram, Proc. Natl. Acad. Sci. 116 (2019) 12684–12691, http://dx.doi.org/10.1073/pnas.1900739116.
[205] A. Falenty, T.C. Hansen, W.F. Kuhs, Formation and properties of ice XVI obtained by emptying a type sii clathrate hydrate, Nature 516 (2014) 231–233, http://dx.doi.org/10.1038/nature14014.
[206] K.W. Allen, G.A. Jeffrey, On the structure of bromine hydrate, J. Chem. Phys. 38 (1963) 2304–2305.
[207] R.K. McMullan, G.A. Jeffrey, Polyhedral clathrate hydrates. IX. Structure of ethylene oxide hydrate, J. Chem. Phys. 42 (1965) 2725–2732, http://dx.doi.org/10.1063/1.1703228.
[208] T.C.W. Mak, R.K. McMullan, Polyhedral clathrate hydrates. X. Structure of the double hydrate of tetrahydrofuran and hydrogen sulfide, J. Chem. Phys. 42 (1965) 2732–2737, http://dx.doi.org/10.1063/1.1703229.
[209] J.A. Ripmeester, J.S. Tse, C.I. Ratcliffe, B.M. Powell, A new clathrate hydrate structure, Nature 325 (1987) 135–136, http://dx.doi.org/10.1038/325135a0.
[210] A.V. Kurnosov, A.Y. Manakov, V.Y. Komarov, V.I. Voronin, A.E. Teplykh, Y.A. Dyadin, A new gas hydrate structure, in: Dokl. Phys. Chem., Kluwer Academic Publishers-Plenum Publishers, 2001, pp. 303–305.
[211] Y. Huang, C. Zhu, L. Wang, X. Cao, Y. Su, X. Jiang, S. Meng, J. Zhao, X.C. Zeng, A new phase diagram of water under negative pressure: The rise of the lowest-density clathrate s-III, Sci. Adv. 2 (2016) e1501010, http://dx.doi.org/10.1126/sciadv.1501010.
[212] Y. Huang, C. Zhu, L. Wang, J. Zhao, X.C. Zeng, Prediction of a new ice clathrate with record low density: A potential candidate as ice XIX in guest-free form, Chem. Phys. Lett. 671 (2017) 186–191, http://dx.doi.org/10.1016/j.cplett.2017.01.035.
[213] Y. Liu, L. Ojamäe, Clathrate ice sL: A new crystalline phase of ice with ultralow density predicted by first-principles phase diagram computations, Phys. Chem. Chem. Phys. 20 (2018) 8333–8340, http://dx.doi.org/10.1039/C8CP00699G.
[214] T. Matsui, M. Hirata, T. Yagasaki, M. Matsumoto, H. Tanaka, Communication: Hypothetical ultralow-density ice polymorphs, J. Chem. Phys. 147 (2017) 091101, http://dx.doi.org/10.1063/1.4994757.
[215] T. Matsui, T. Yagasaki, M. Matsumoto, H. Tanaka, Phase diagram of ice polymorphs under negative pressure considering the limits of mechanical stability, J. Chem. Phys. 150 (2019) 041102, http://dx.doi.org/10.1063/1.5083021.

[216] M.M. Conde, C. Vega, G.A. Tribello, B. Slater, The phase diagram of water at negative pressures: Virtual ices, J. Chem. Phys. 131 (2009) 034510, http://dx.doi.org/10.1063/1.3182727.
[217] T. Yagasaki, M. Yamasaki, M. Matsumoto, H. Tanaka, Formation of hot ice caused by carbon nanobrushes, J. Chem. Phys. 151 (2019) 064702, http://dx.doi.org/10.1063/1.5111843.
[218] W. Si, S.J. Han, X. Shi, S.N. Ehrlich, J. Jaroszynski, A. Goyal, Q. Li, High current superconductivity in $FeSe_{0.5}Te_{0.5}$ -coated conductors at 30 tesla, Nature Commun. 4 (2013) 1347, http://dx.doi.org/10.1038/ncomms2337.
[219] V.L. Ginzburg, Once again about high-temperature superconductivity, Contemp. Phys. 33 (1992) 15–23.
[220] H. Suhl, B.T. Matthias, L.R. Walker, Bardeen-Cooper-Schrieffer theory of superconductivity in the case of overlapping bands, Phys. Rev. Lett. 3 (1959) 552–554, http://dx.doi.org/10.1103/PhysRevLett.3.552.
[221] M.-H. Whangbo, S. Deng, J. Köhler, A. Simon, Interband electron pairing for superconductivity from the breakdown of the Born–Oppenheimer approximation, ChemPhysChem 19 (2018) 3191–3195, http://dx.doi.org/10.1002/cphc.201800738.
[222] S. Deng, A. Simon, J. Köhler, The flat/steep band condition created in Te-II, Physica C 460–462 (2007) 1020–1021, http://dx.doi.org/10.1016/j.physc.2007.03.204.
[223] S. Deng, A. Simon, Lone pairs, bipolarons and superconductivity in tellurium, High Tc Supercond. Relat. Transit. Met. Oxides (2007) 201, Spec. Contrib. Honor K Alex Müller Occas. His 80th Birthd.
[224] H. Liu, I.I. Naumov, R. Hoffmann, N.W. Ashcroft, R.J. Hemley, Potential high-Tc superconducting lanthanum and yttrium hydrides at high pressure, Proc. Natl. Acad. Sci. 114 (2017) 6990–6995, http://dx.doi.org/10.1073/pnas.1704505114.
[225] F. Peng, Y. Sun, C.J. Pickard, R.J. Needs, Q. Wu, Y. Ma, Hydrogen clathrate structures in rare earth hydrides at high pressures: Possible route to room-temperature superconductivity, Phys. Rev. Lett. 119 (2017) 107001, http://dx.doi.org/10.1103/PhysRevLett.119.107001.
[226] Z. Zhao, S. Zhang, T. Yu, H. Xu, A. Bergara, G. Yang, Predicted pressure-induced superconducting transition in electride Li_6P, Phys. Rev. Lett. 122 (2019) 097002, http://dx.doi.org/10.1103/PhysRevLett.122.097002.
[227] A. Hermann, N.W. Ashcroft, R. Hoffmann, High pressure ices, Proc. Natl. Acad. Sci. 109 (2012) 745–750, http://dx.doi.org/10.1073/pnas.1118694109.
[228] Z. Yin, Y. Guan, B. Fu, D.H. Zhang, Two-state diabatic potential energy surfaces of ClH_2 based on nonadiabatic couplings with neural networks, Phys. Chem. Chem. Phys. 21 (2019) 20372–20383, http://dx.doi.org/10.1039/C9CP03592C.
[229] S. Manzhos, X. Wang, R. Dawes, T. Carrington, A nested molecule-independent neural network approach for high-quality potential fits, J. Phys. Chem. A 110 (2006) 5295–5304, http://dx.doi.org/10.1021/jp055253z.
[230] E. Pradhan, A. Brown, A ground state potential energy surface for hono based on a neural network with exponential fitting functions, Phys. Chem. Chem. Phys. 19 (2017) 22272–22281, http://dx.doi.org/10.1039/C7CP04010E.
[231] Y. Guan, B. Fu, D.H. Zhang, Construction of diabatic energy surfaces for LiFH with artificial neural networks, J. Chem. Phys. 147 (2017) 224307, http://dx.doi.org/10.1063/1.5007031.
[232] Y. Guan, D.H. Zhang, H. Guo, D.R. Yarkony, Representation of coupled adiabatic potential energy surfaces using neural network based quasi-diabatic Hamiltonians: 1, 2 $^2A'$ states of LiFH, Phys. Chem. Chem. Phys. 21 (2019) 14205–14213, http://dx.doi.org/10.1039/C8CP06598E.
[233] D. Yuan, Y. Guan, W. Chen, H. Zhao, S. Yu, C. Luo, Y. Tan, T. Xie, X. Wang, Z. Sun, D.H. Zhang, X. Yang, Observation of the geometric phase effect in the $H + HD \rightarrow H_2 + D$ reaction, Science 362 (2018) 1289–1293, http://dx.doi.org/10.1126/science.aav1356.
[234] H.E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, A. Tkatchenko, Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces, J. Chem. Phys. 150 (2019) 114102, http://dx.doi.org/10.1063/1.5078687.
[235] S. Chmiela, A. Tkatchenko, H.E. Sauceda, I. Poltavsky, K.T. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, Sci. Adv. 3 (2017) e1603015, http://dx.doi.org/10.1126/sciadv.1603015.
[236] Y. Liu, J. Li, An accurate potential energy surface and ring polymer molecular dynamics study of the $Cl + CH_4 \rightarrow HCl + CH_3$ reaction, Phys. Chem. Chem. Phys. 22 (2020) 344–353, http://dx.doi.org/10.1039/C9CP05693A.
[237] G. Schmitz, D.G. Artiukhin, O. Christiansen, Approximate high mode coupling potentials using Gaussian process regression and adaptive density guided sampling, J. Chem. Phys. 150 (2019) 131102, http://dx.doi.org/10.1063/1.5092228.
[238] D. Hu, Y. Xie, X. Li, L. Li, Z. Lan, Inclusion of machine learning kernel ridge regression potential energy surfaces in on-the-fly nonadiabatic molecular dynamics simulation, J. Phys. Chem. Lett. 9 (2018) 2725–2732, http://dx.doi.org/10.1021/acs.jpclett.8b00684.
[239] P.O. Dral, M. Barbatti, W. Thiel, Nonadiabatic excited-state dynamics with machine learning, J. Phys. Chem. Lett. 9 (2018) 5660–5663, http://dx.doi.org/10.1021/acs.jpclett.8b02469.
[240] S. Manzhos, T. Carrington, Neural network potential energy surfaces for small molecules and reactions, Chem. Rev. (2020) http://dx.doi.org/10.1021/acs.chemrev.0c00665.
[241] C. Qu, Q. Yu, B.L. Van Hoozen, J.M. Bowman, R.A. Vargas-Hernández, Assessing Gaussian process regression and permutationally invariant polynomial approaches to represent high-dimensional potential energy surfaces, J. Chem. Theory Comput. 14 (2018) 3381–3396, http://dx.doi.org/10.1021/acs.jctc.8b00298.
[242] D.P. Metcalf, A. Koutsoukas, S.A. Spronk, B.L. Claus, D.A. Loughney, S.R. Johnson, D.L. Cheney, C.D. Sherrill, Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory, J. Chem. Phys. 152 (2020) 074103, http://dx.doi.org/10.1063/1.5142636.
[243] L. Mones, N. Bernstein, G. Csányi, Exploration, sampling, and reconstruction of free energy surfaces with Gaussian process regression, J. Chem. Theory Comput. 12 (2016) 5100–5110, http://dx.doi.org/10.1021/acs.jctc.6b00553.
[244] U. Rivero, O.T. Unke, M. Meuwly, S. Willitsch, Reactive atomistic simulations of Diels–Alder reactions: The importance of molecular rotations, J. Chem. Phys. 151 (2019) 104301, http://dx.doi.org/10.1063/1.5114981.
[245] X. Xu, J. Chen, S. Liu, D.H. Zhang, An ab initio-based global potential energy surface for the SH3 system and full-dimensional state-to-state quantum dynamics study for the $H_2 + HS \rightarrow H_2S + H$ reaction, J. Comput. Chem. 40 (2019) 1151–1160, http://dx.doi.org/10.1002/jcc.25746.
[246] M. del Cueto, X. Zhou, L. Zhou, Y. Zhang, B. Jiang, H. Guo, New perspectives on CO_2–Pt(111) interaction with a high-dimensional neural network potential energy surface, J. Phys. Chem. C 124 (2020) 5174–5181, http://dx.doi.org/10.1021/acs.jpcc.9b10883.
[247] J. Zuo, Q. Chen, X. Hu, H. Guo, D. Xie, Theoretical investigations of rate coefficients for $H + O_3$ and $HO_2 + O$ reactions on a full-dimensional potential energy surface, J. Phys. Chem. A 124 (2020) 6427–6437, http://dx.doi.org/10.1021/acs.jpca.0c04321.
[248] Q. Tong, X. Luo, A.A. Adeleke, P. Gao, Y. Xie, H. Liu, Q. Li, Y. Wang, J. Lv, Y. Yao, Y. Ma, Machine learning metadynamics simulation of reconstructive phase transition, Phys. Rev. B 103 (2021) 054107, http://dx.doi.org/10.1103/PhysRevB.103.054107.
[249] P.O. Dral, A. Owens, S.N. Yurchenko, W. Thiel, Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels, J. Chem. Phys. 146 (2017) 244108, http://dx.doi.org/10.1063/1.4989536.
[250] A.J. Ballard, R. Das, S. Martiniani, D. Mehta, L. Sagun, J.D. Stevenson, D.J. Wales, Energy landscapes for machine learning, Phys. Chem. Chem. Phys. 19 (2017) 12585–12603, http://dx.doi.org/10.1039/C7CP01108C.
[251] B. Fu, D.H. Zhang, Ab initio potential energy surfaces and quantum dynamics for polyatomic bimolecular reactions, J. Chem. Theory Comput. 14 (2018) 2289–2303, http://dx.doi.org/10.1021/acs.jctc.8b00006.

[252] J. Chen, X. Xu, X. Xu, D.H. Zhang, A global potential energy surface for the $H_2 + OH \leftrightarrow H_2O + H$ reaction using neural networks, J. Chem. Phys. 138 (2013) 154301, http://dx.doi.org/10.1063/1.4801658.
[253] X. Xu, J. Chen, D.H. Zhang, Global potential energy surface for the $H+CH_4 \leftrightarrow H_2+CH_3$ reaction using neural networks, Chin. J. Chem. Phys. 27 (2014) 373–379, http://dx.doi.org/10.1063/1674-0068/27/04/373-379.
[254] D.F.R. Brown, M.N. Gibbs, D.C. Clary, Combining ab initio computations, neural networks, and diffusion Monte Carlo: An efficient method to treat weakly bound molecules, J. Chem. Phys. 105 (1996) 7597–7604, http://dx.doi.org/10.1063/1.472596.
[255] S.A. Tawfik, O. Isayev, M.J.S. Spencer, D.A. Winkler, Predicting thermal properties of crystals using machine learning, Adv. Theory Simul. 3 (2020) 1900208, http://dx.doi.org/10.1002/adts.201900208.
[256] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.
[257] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.
[258] J. Fox, S. Weisberg, Robust Regression, R -Plus Companion Appl. Regres. 91, 2002.
[259] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.
[260] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[261] C. Loftis, K. Yuan, Y. Zhao, M. Hu, J. Hu, Lattice thermal conductivity prediction using symbolic regression and machine learning, J. Phys. Chem. A 125 (2021) 435–450, http://dx.doi.org/10.1021/acs.jpca.0c08103.
[262] D.T. Morelli, G.A. Slack, High lattice thermal conductivity solids, in: S.L. Shindé, J.S. Goela (Eds.), High Therm. Conduct. Mater., Springer, New York, NY, 2006, pp. 37–68, http://dx.doi.org/10.1007/0-387-25100-6_2.
[263] D. Singhal, W. Curatolo, Drug polymorphism and dosage form design: A practical perspective, Adv. Drug Deliv. Rev. 56 (2004) 335–347, http://dx.doi.org/10.1016/j.addr.2003.10.008.
[264] S.L. Morissette, Ö. Almarsson, M.L. Peterson, J.F. Remenar, M.J. Read, A.V. Lemmo, S. Ellis, M.J. Cima, C.R. Gardner, High-throughput crystallization: Polymorphs, salts, co-crystals and solvates of pharmaceutical solids, Adv. Drug Deliv. Rev. 56 (2004) 275–300, http://dx.doi.org/10.1016/j.addr.2003.10.020.
[265] S. Datta, D.J.W. Grant, Crystal structures of drugs: Advances in determination, prediction and engineering, Nat. Rev. Drug Discov. 3 (2004) 42–57, http://dx.doi.org/10.1038/nrd1280.
[266] M. Saifee, N. Inamda, D. Dhamecha, A. Rathi, Drug polymorphism: A review, Int. J. Health Res. 2 (2009) http://dx.doi.org/10.4314/ijhr.v2i4.55423.
[267] J. Li, C.J. Tilbury, S.H. Kim, M.F. Doherty, A design aid for crystal growth engineering, Prog. Mater. Sci. 82 (2016) 1–38, http://dx.doi.org/10.1016/j.pmatsci.2016.03.003.
[268] Y. Sun, C.J. Tilbury, S.M. Reutzel-Edens, R.M. Bhardwaj, J. Li, M.F. Doherty, Modeling olanzapine solution growth morphologies, Cryst. Growth Des. 18 (2018) 905–911, http://dx.doi.org/10.1021/acs.cgd.7b01389.
[269] A.G. Shtukenberg, M.D. Ward, B. Kahr, Crystal growth with macromolecular additives, Chem. Rev. 117 (2017) 14042–14090, http://dx.doi.org/10.1021/acs.chemrev.7b00285.
[270] L. Lee, J. Baek, K.S. Park, Y.-E. Lee, N.K. Shrestha, M.M. Sung, Wafer-scale single-crystal perovskite patterned thin films based on geometrically-confined lateral crystal growth, Nature Commun. 8 (2017) 15882, http://dx.doi.org/10.1038/ncomms15882.
[271] L. Yu, S.M. Reutzel, G.A. Stephenson, Physical characterization of polymorphic drugs: An integrated characterization strategy, Pharm. Sci. Techn. Today 1 (1998) 118–127, http://dx.doi.org/10.1016/S1461-5347(98)00031-5.
[272] M. Ingelman-Sundberg, Human drug metabolising cytochrome P450 enzymes: Properties and polymorphisms, Naunyn. Schmiedebergs Arch. Pharmacol. 369 (2004) 89–104, http://dx.doi.org/10.1007/s00210-003-0819-z.
[273] Y.A. Abramov, K. Pencheva, Thermodynamics and relative solubility prediction of polymorphic systems, in: Chem. Eng. Pharm. Ind., John Wiley & Sons, Ltd, 2019, pp. 505–518, http://dx.doi.org/10.1002/9781119600800.ch22.
[274] S.M. Reutzel-Edens, J.K. Bush, P.A. Magee, G.A. Stephenson, S.R. Byrn, Anhydrates and hydrates of olanzapine: crystallization, solid-state characterization, and structural relationships, Cryst. Growth Des. 3 (2003) 897–907, http://dx.doi.org/10.1021/cg034055z.
[275] Q. Lu, I. Ali, Z. Wei, J. Li, Crystal morphology prediction of olanzapine forms III and IV, Cryst. Res. Technol. 2000215. http://dx.doi.org/10.1002/crat.202000215.
[276] H. Luo, X. Hao, Y. Gong, J. Zhou, X. He, J. Li, Rational crystal polymorph design of olanzapine, Cryst. Growth Des. 19 (2019) 2388–2395, http://dx.doi.org/10.1021/acs.cgd.9b00068.
[277] J. Tang, Y. Han, I. Ali, H. Luo, A. Nowak, J. Li, Stability and phase transition investigation of olanzapine polymorphs, Chem. Phys. Lett. 767 (2021) 138384, http://dx.doi.org/10.1016/j.cplett.2021.138384.
[278] R. Thakuria, A. Nangia, Polymorphic form IV of olanzapine, Acta Crystallogr. C 67 (2011) o461–o463, http://dx.doi.org/10.1107/S0108270111043952.
[279] R.M. Bhardwaj, L.S. Price, S.L. Price, S.M. Reutzel-Edens, G.J. Miller, I.D.H. Oswald, B.F. Johnston, A.J. Florence, Exploring the experimental and computed crystal energy landscape of olanzapine, Cryst. Growth Des. 13 (2013) 1602–1617, http://dx.doi.org/10.1021/cg301826s.
[280] S. Askin, J.K. Cockcroft, L.S. Price, A.D. Gonçalves, M. Zhao, D.A. Tocher, G.R. Williams, S. Gaisford, D.Q.M. Craig, Olanzapine form IV: Discovery of a new polymorphic form enabled by computed crystal energy landscapes, Cryst. Growth Des. 19 (2019) 2751–2757, http://dx.doi.org/10.1021/acs.cgd.8b01881.
[281] X. Hao, J. Liu, H. Luo, Y. Han, W. Hu, J. Liu, J. Li, X. He, Crystal structure optimization and Gibbs free energy comparison of five sulfathiazole polymorphs by the embedded fragment QM method at the DFT level, Crystals 9 (2019) 256, http://dx.doi.org/10.3390/cryst9050256.
[282] N. Blagden, R.J. Davey, H.F. Lieberman, L. Williams, R. Payne, R. Roberts, R. Rowe, R. Docherty, Crystal chemistry and solvent effects in polymorphic systems sulfathiazole, J. Chem. Soc. Faraday Trans. 94 (1998) 1035–1044, http://dx.doi.org/10.1039/A706669D.
[283] Á. Munroe, Å.C. Rasmuson, B.K. Hodnett, D.M. Croker, Relative stabilities of the five polymorphs of sulfathiazole, Cryst. Growth Des. 12 (2012) 2825–2835, http://dx.doi.org/10.1021/cg201641g.
[284] Y. Hu, A. Erxleben, A.G. Ryder, P. McArdle, Quantitative analysis of sulfathiazole polymorphs in ternary mixtures by attenuated total reflectance infrared, near-infrared and Raman spectroscopy, J. Pharm. Biomed. Anal. 53 (2010) 412–420, http://dx.doi.org/10.1016/j.jpba.2010.05.002.
[285] F.C. Chan, J. Anwar, R. Cernik, P. Barnes, R.M. Wilson, Ab initio structure determination of sulfathiazole polymorph V from synchrotron X-ray powder diffraction data, J. Appl. Crystallogr. 32 (1999) 436–441.
[286] G.J.O. Beran, Modeling polymorphic molecular crystals with electronic structure theory, Chem. Rev. 116 (2016) 5567–5613, http://dx.doi.org/10.1021/acs.chemrev.5b00648.
[287] J.-D. Chai, M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections, Phys. Chem. Chem. Phys. 10 (2008) 6615–6620, http://dx.doi.org/10.1039/B810189B.
[288] X. Hao, J. Liu, I. Ali, H. Luo, Y. Han, W. Hu, J. Liu, X. He, J. Li, Ab initio determination of crystal stability of di-p-tolyl disulfide, Sci. Rep. 11 (2021) 7076, http://dx.doi.org/10.1038/s41598-021-86519-1.
[289] A.T. Anghel, G.M. Day, S.L. Price, A study of the known and hypothetical crystal structures of pyridine: Why are there four molecules in the asymmetric unit cell? CrystEngComm 4 (2002) 348–355, http://dx.doi.org/10.1039/B202084J.
[290] J.R. Holden, Z. Du, H.L. Ammon, Prediction of possible crystal structures for C-, H-, N-, O-, and F-containing organic compounds, J. Comput. Chem. 14 (1993) 422–437, http://dx.doi.org/10.1002/jcc.540140406.

[291] C.W. Glass, A.R. Oganov, N. Hansen, USPEX—Evolutionary crystal structure prediction, Comput. Phys. Comm. 175 (2006) 713–720, http://dx.doi.org/10.1016/j.cpc.2006.07.020.
[292] A.O. Lyakhov, A.R. Oganov, H.T. Stokes, Q. Zhu, New developments in evolutionary structure prediction algorithm USPEX, Comput. Phys. Comm. 184 (2013) 1172–1182, http://dx.doi.org/10.1016/j.cpc.2012.12.009.
[293] A.R. Oganov, C.W. Glass, Evolutionary crystal structure prediction as a tool in materials design, J. Phys.: Condens. Matter. 20 (2008) 064210, http://dx.doi.org/10.1088/0953-8984/20/6/064210.
[294] M. Khalil, N. Demirdöven, A. Tokmakoff, Obtaining absorptive line shapes in two-dimensional infrared vibrational correlation spectra, Phys. Rev. Lett. 90 (2003) 047401, http://dx.doi.org/10.1103/PhysRevLett.90.047401.
[295] S. Hirata, K. Gilliard, X. He, J. Li, O. Sode, Ab initio molecular crystal structures, spectra, and phase diagrams, Acc. Chem. Res. 47 (2014) 2721–2730, http://dx.doi.org/10.1021/ar500041m.
[296] H. Luo, J. Liu, X. He, J. Li, Low-temperature polymorphic transformation of β-lactam antibiotics, Crystals 9 (2019) 460, http://dx.doi.org/10.3390/cryst9090460.
[297] S.E. Kariper, Spectroscopic and quantum chemical studies on some β-lactam inhibitors, Turk. Comput. Theor. Chem. 1 (2017) 13–26.
[298] L. Fábián, A. Kálmán, G. Argay, G. Bernáth, Z. Cs. Gyarmati, Two polymorphs of a β-lactam (trans-13-azabicyclo[10.2.0]tetradecan-14-one). Concomitant crystal polymorphism and isostructurality, Chem. Commun. (2004) 2114–2115, http://dx.doi.org/10.1039/B408505A.
[299] J. Li, K.C. Bennett, Y. Liu, M.V. Martin, T. Head-Gordon, Accurate prediction of chemical shifts for aqueous protein structure on real world data, Chem. Sci. 11 (2020) 3180–3191, http://dx.doi.org/10.1039/C9SC06561J.
[300] S. Liu, J. Li, K.C. Bennett, B. Ganoe, T. Stauch, M. Head-Gordon, A. Hexemer, D. Ushizima, T. Head-Gordon, Multiresolution 3D-densenet for chemical shift prediction in NMR crystallography, J. Phys. Chem. Lett. 10 (2019) 4558–4565, http://dx.doi.org/10.1021/acs.jpclett.9b01570.
[301] M. Haghighatlari, G. Vishwakarma, M.A.F. Afzal, J. Hachmann, A physics-infused deep learning model for the prediction of refractive indices and its use for the large-scale screening of organic compound space, ChemRxiv (2019).
[302] G. Scalia, C.A. Grambow, B. Pernici, Y.-P. Li, W.H. Green, Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction, J. Chem. Inf. Model. 60 (2020) 2697–2717, http://dx.doi.org/10.1021/acs.jcim.9b00975.
[303] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, in: Proc. 28th Int. Conf. Neural Inf. Process. Syst.- Vol. 2, MIT Press, Cambridge, MA, USA, 2015, pp. 2224–2232.
[304] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Proc. 31st Int. Conf. Neural Inf. Process. Syst., Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6405–6416.
[305] J. Mukhoti, Y. Gal, Evaluating Bayesian deep learning methods for semantic segmentation, 2019, ArXiv181112709Cs. (Accessed 3 May 2021).
[306] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[307] M. Haghighatlari, J. Li, F. Heidar-Zadeh, Y. Liu, X. Guan, T. Head-Gordon, Learning to make chemical predictions: The interplay of feature representation, data, and machine learning methods, Chem 6 (2020) 1527–1542, http://dx.doi.org/10.1016/j.chempr.2020.05.014.
[308] Y. Shen, A. Bax, SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network, J. Biomol. NMR 48 (2010) 13–22.
[309] B. Han, Y. Liu, S.W. Ginzinger, D.S. Wishart, SHIFTX2: Significantly improved protein chemical shift prediction, J. Biomol. NMR 50 (2011) 43.
[310] C.J. Pickard, F. Mauri, All-electron magnetic response with pseudopotentials: NMR chemical shifts, Phys. Rev. B 63 (2001) 245101, http://dx.doi.org/10.1103/PhysRevB.63.245101.
[311] F.M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, L. Emsley, Chemical shifts in molecular solids by machine learning, Nature Commun. 9 (2018) 4501, http://dx.doi.org/10.1038/s41467-018-06972-x.
[312] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (2013) 184115, http://dx.doi.org/10.1103/PhysRevB.87.184115.
[313] C.C. Arico-Muendel, H. Blanchette, D.R. Benjamin, T.M. Caiazzo, P.A. Centrella, J. DeLorey, E.G. Doyle, S.R. Johnson, M.T. Labenski, B.A. Morgan, G. O'Donovan, A.A. Sarjeant, S. Skinner, C.D. Thompson, S.T. Griffin, W. Westlin, K.F. White, Orally active fumagillin analogues: Transformations of a reactive warhead in the gastric environment, ACS Med. Chem. Lett. 4 (2013) 381–386, http://dx.doi.org/10.1021/ml3003633.
[314] H.T. Dao, C. Li, Q. Michaudel, B.D. Maxwell, P.S. Baran, Hydromethylation of unactivated olefins, J. Am. Chem. Soc. 137 (2015) 8046–8049, http://dx.doi.org/10.1021/jacs.5b05144.
[315] D. Garozzo, G. Gattuso, F.H. Kohnke, A. Notti, S. Pappalardo, M.F. Parisi, I. Pisagatti, A.J.P. White, D.J. Williams, Inclusion networks of a Calix[5]arene-based exoditopic receptor and long-chain alkyldiammonium ions, Org. Lett. 5 (2003) 4025–4028, http://dx.doi.org/10.1021/ol035310b.
[316] J.W. Bats, CSD commun, 2010, http://dx.doi.org/10.5517/ccvxvk4.
[317] G.-B. Huang, W.-E. Liu, A. Valkonen, H. Yao, K. Rissanen, W. Jiang, Selective recognition of aromatic hydrocarbons by endo-functionalized molecular tubes via C/N-H$\cdots\pi$ interactions, Chin. Chem. Lett. 29 (2018) 91–94, http://dx.doi.org/10.1016/j.cclet.2017.07.005.
[318] M.J. Plater, W.T.A. Harrison, L.M.M. de los Toyos, L. Hendry, The consistent hexameric paddle-wheel crystallisation motif of a family of 2, 4-bis(n-alkylamino)nitrobenzenes: Alkyl = pentyl, hexyl, heptyl and octyl, J. Chem. Res. 41 (2017) 235–238, http://dx.doi.org/10.3184/174751917X14902201357356.
[319] P. Gao, J. Zhang, Q. Peng, J. Zhang, V.-A. Glezakou, General protocol for the accurate prediction of molecular 13C/1H NMR chemical shifts via machine learning augmented DFT, J. Chem. Inf. Model. 60 (2020) 3746–3754, http://dx.doi.org/10.1021/acs.jcim.0c00388.
[320] W. Gerrard, L.A. Bratholm, M.J. Packer, A.J. Mulholland, D.R. Glowacki, C.P. Butts, IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy, Chem. Sci. 11 (2020) 508–515, http://dx.doi.org/10.1039/C9SC03854J.
[321] M.R.G. Marques, J. Wolff, C. Steigemann, M.A.L. Marques, Neural network force fields for simple metals and semiconductors: Construction and application to the calculation of phonons and melting temperatures, Phys. Chem. Chem. Phys. 21 (2019) 6506–6516, http://dx.doi.org/10.1039/C8CP05771K.
[322] R. Jinnouchi, F. Karsai, G. Kresse, On-the-fly machine learning force field generation: Application to melting points, Phys. Rev. B 100 (2019) 014105, http://dx.doi.org/10.1103/PhysRevB.100.014105.
[323] C. Zeni, K. Rossi, A. Glielmo, F. Baletto, On machine learning force fields for metallic nanoparticles, Adv. Phys. X 4 (2019) 1654919, http://dx.doi.org/10.1080/23746149.2019.1654919.
[324] W. Plazinski, A. Plazinska, A. Brzyska, Efficient sampling of high-energy states by machine learning force fields, Phys. Chem. Chem. Phys. 22 (2020) 14364–14374, http://dx.doi.org/10.1039/D0CP01399D.
[325] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 011002, http://dx.doi.org/10.1063/1.4812323.
[326] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, New J. Phys. 15 (2013) 095003, http://dx.doi.org/10.1088/1367-2630/15/9/095003.

[327] R. Ramakrishnan, P.O. Dral, M. Rupp, O.A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data 1 (2014) 140022, http://dx.doi.org/10.1038/sdata.2014.22.
[328] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), JOM 65 (2013) 1501–1509, http://dx.doi.org/10.1007/s11837-013-0755-4.
[329] A.M. Reilly, A. Tkatchenko, Van der Waals dispersion interactions in molecular materials: Beyond pairwise additivity, Chem. Sci. 6 (2015) 3289–3301, http://dx.doi.org/10.1039/C5SC00410A.
[330] J. Hermann, R.A. DiStasio, A. Tkatchenko, First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications, Chem. Rev. 117 (2017) 4714–4758, http://dx.doi.org/10.1021/acs.chemrev.6b00446.
[331] G.P. Stahly, Diversity in single- and multiple-component crystals. The search for and prevalence of polymorphs and cocrystals, Cryst. Growth Des. 7 (2007) 1007–1026, http://dx.doi.org/10.1021/cg060838j.
[332] A.Y. Lee, D. Erdemir, A.S. Myerson, Crystal polymorphism in chemical process development, Annu. Rev. Chem. Biomol. Eng. 2 (2011) 259–280, http://dx.doi.org/10.1146/annurev-chembioeng-061010-114224.
[333] M.D. Eddleston, K.E. Hejczyk, E.G. Bithell, G.M. Day, W. Jones, Determination of the crystal structure of a new polymorph of theophylline, Chem. Eur. J. 19 (2013) 7883–7888, http://dx.doi.org/10.1002/chem.201204369.
[334] M. Baias, J.-N. Dumez, P.H. Svensson, S. Schantz, G.M. Day, L. Emsley, De novo determination of the crystal structure of a large drug molecule by crystal structure prediction-based powder NMR crystallography, J. Am. Chem. Soc. 135 (2013) 17501–17507, http://dx.doi.org/10.1021/ja4088874.
[335] M.-A. Perrin, M.A. Neumann, H. Elmaleh, L. Zaske, Crystal structure determination of the elusive paracetamol form III, Chem. Commun. (2009) 3181–3183, http://dx.doi.org/10.1039/B822882E.
[336] O.D. Jurchescu, D.A. Mourey, S. Subramanian, S.R. Parkin, B.M. Vogel, J.E. Anthony, T.N. Jackson, D.J. Gundlach, Effects of polymorphism on charge transport in organic semiconductors, Phys. Rev. B 80 (2009) 085201, http://dx.doi.org/10.1103/PhysRevB.80.085201.
[337] A.N. Sokolov, S. Atahan-Evrenk, R. Mondal, H.B. Akkerman, R.S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S.C.B. Mannsfeld, A.P. Zoombelt, Z. Bao, A. Aspuru-Guzik, From computational discovery to experimental characterization of a high hole mobility organic crystal, Nature Commun. 2 (2011) 437, http://dx.doi.org/10.1038/ncomms1451.
[338] H. Chung, Y. Diao, Polymorphism as an emerging design strategy for high performance organic electronics, J. Mater. Chem. C 4 (2016) 3915–3933, http://dx.doi.org/10.1039/C5TC04390E.
[339] S. (Sally) L. Price, Computed crystal energy landscapes for understanding and predicting organic crystal structures and polymorphism, Acc. Chem. Res. 42 (2009) 117–126, http://dx.doi.org/10.1021/ar800147t.
[340] G.M. Day, Current approaches to predicting molecular organic crystal structures, Crystallogr. Rev. 17 (2011) 3–52, http://dx.doi.org/10.1080/0889311X.2010.517526.
[341] A.M. Reilly, R.I. Cooper, C.S. Adjiman, S. Bhattacharya, A.D. Boese, J.G. Brandenburg, P.J. Bygrave, R. Bylsma, J.E. Campbell, R. Car, D.H. Case, R. Chadha, J.C. Cole, K. Cosburn, H.M. Cuppen, F. Curtis, G.M. Day, R.A. DiStasio Jr, A. Dzyabchenko, B.P. van Eijck, D.M. Elking, J.A. van den Ende, J.C. Facelli, M.B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T.S. Gee, R. de Gelder, L.M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D.W.M. Hofmann, J. Hoja, R.K. Hylton, L. Iuzzolino, W. Jankiewicz, D.T. de Jong, J. Kendrick, N.J.J. de Klerk, H.-Y. Ko, L.N. Kuleshova, X. Li, S. Lohani, F.J.J. Leusen, A.M. Lund, J. Lv, Y. Ma, N. Marom, A.E. Masunov, P. McCabe, D.P. McMahon, H. Meekes, M.P. Metz, A.J. Misquitta, S. Mohamed, B. Monserrat, R.J. Needs, M.A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A.R. Oganov, A.M. Orendt, G.I. Pagola, C.C. Pantelides, C.J. Pickard, R. Podeszwa, L.S. Price, S.L. Price, A. Pulido, M.G. Read, K. Reuter, E. Schneider, C. Schober, G.P. Shields, P. Singh, I.J. Sugden, K. Szalewicz, C.R. Taylor, A. Tkatchenko, M.E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R.E. Watson, G.A. de Wijs, J. Yang, Q. Zhu, C.R. Groom, Report on the sixth blind test of organic crystal structure prediction methods, Acta Crystallogr. B 72 (2016) 439–459, http://dx.doi.org/10.1107/S2052520616007447.
[342] D.A. Bardwell, C.S. Adjiman, Y.A. Arnautova, E. Bartashevich, S.X.M. Boerrigter, D.E. Braun, A.J. Cruz-Cabeza, G.M. Day, R.G. Della Valle, G.R. Desiraju, B.P. van Eijck, J.C. Facelli, M.B. Ferraro, D. Grillo, M. Habgood, D.W.M. Hofmann, F. Hofmann, K.V.J. Jose, P.G. Karamertzanis, A.V. Kazantsev, J. Kendrick, L.N. Kuleshova, F.J.J. Leusen, A.V. Maleev, A.J. Misquitta, S. Mohamed, R.J. Needs, M.A. Neumann, D. Nikylov, A.M. Orendt, R. Pal, C.C. Pantelides, C.J. Pickard, L.S. Price, S.L. Price, H.A. Scheraga, J. van de Streek, T.S. Thakur, S. Tiwari, E. Venuti, I.K. Zhitkov, Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test, Acta Crystallogr. B 67 (2011) 535–551, http://dx.doi.org/10.1107/S0108768111042868.
[343] G.M. Day, T.G. Cooper, A.J. Cruz-Cabeza, K.E. Hejczyk, H.L. Ammon, S.X.M. Boerrigter, J.S. Tan, R.G. Della Valle, E. Venuti, J. Jose, S.R. Gadre, G.R. Desiraju, T.S. Thakur, B.P. van Eijck, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, M.A. Neumann, F.J.J. Leusen, J. Kendrick, S.L. Price, A.J. Misquitta, P.G. Karamertzanis, G.W.A. Welch, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, J. van de Streek, A.K. Wolf, B. Schweizer, Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test, Acta Crystallogr. B 65 (2009) 107–125, http://dx.doi.org/10.1107/S0108768109004066.
[344] W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D.W.M. Hofmann, F.J.J. Leusen, J.P.M. Lommerse, W.T.M. Mooij, S.L. Price, H. Scheraga, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer, D.E. Williams, Crystal structure prediction of small organic molecules: A second blind test, Acta Crystallogr. B 58 (2002) 647–661, http://dx.doi.org/10.1107/S0108768102005669.
[345] G.M. Day, W.D.S. Motherwell, H.L. Ammon, S.X.M. Boerrigter, R.G. Della Valle, E. Venuti, A. Dzyabchenko, J.D. Dunitz, B. Schweizer, B.P. van Eijck, P. Erk, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, F.J.J. Leusen, C. Liang, C.C. Pantelides, P.G. Karamertzanis, S.L. Price, T.C. Lewis, H. Nowell, A. Torrisi, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, P. Verwer, A third blind test of crystal structure prediction, Acta Crystallogr. B 61 (2005) 511–527, http://dx.doi.org/10.1107/S0108768105016563.
[346] J.P.M. Lommerse, W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Gavezzotti, D.W.M. Hofmann, F.J.J. Leusen, W.T.M. Mooij, S.L. Price, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer, D.E. Williams, A test of crystal structure prediction of small organic molecules, Acta Crystallogr. B 56 (2000) 697–714, http://dx.doi.org/10.1107/S0108768100004584.
[347] P.G. Karamertzanis, C.C. Pantelides, Ab initio crystal structure prediction—I. Rigid molecules, J. Comput. Chem. 26 (2005) 304–324, http://dx.doi.org/10.1002/jcc.20165.
[348] D.H. Case, J.E. Campbell, P.J. Bygrave, G.M. Day, Convergence properties of crystal structure prediction by quasi-random sampling, J. Chem. Theory Comput. 12 (2016) 910–924, http://dx.doi.org/10.1021/acs.jctc.5b01112.
[349] C.J. Pickard, R.J. Needs, Ab initiorandom structure searching, J. Phys.: Condens. Matter. 23 (2011) 053201, http://dx.doi.org/10.1088/0953-8984/23/5/053201.
[350] R. Tom, T. Rose, I. Bier, H. O'Brien, Á. Vázquez-Mayagoitia, N. Marom, Genarris 2.0: A random structure generator for molecular crystals, Comput. Phys. Comm. 250 (2020) 107170, http://dx.doi.org/10.1016/j.cpc.2020.107170.
[351] X. Li, F.S. Curtis, T. Rose, C. Schober, A. Vazquez-Mayagoitia, K. Reuter, H. Oberhofer, N. Marom, Genarris: Random generation of molecular crystal structures and fast screening with a Harris approximation, J. Chem. Phys. 148 (2018) 241701, http://dx.doi.org/10.1063/1.5014038.
[352] F. Curtis, X. Li, T. Rose, Á. Vázquez-Mayagoitia, S. Bhattacharya, L.M. Ghiringhelli, N. Marom, GAtor: A first-principles genetic algorithm for molecular crystal structure prediction, J. Chem. Theory Comput. 14 (2018) 2246–2264, http://dx.doi.org/10.1021/acs.jctc.7b01152.
[353] S. Kim, A.M. Orendt, M.B. Ferraro, J.C. Facelli, Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field, J. Comput. Chem. 30 (2009) 1973–1985, http://dx.doi.org/10.1002/jcc.21189.

[354] R.J. Needs, C.J. Pickard, Perspective: Role of structure prediction in materials discovery and design, APL Mater. 4 (2016) 053210, http://dx.doi.org/10.1063/1.4949361.
[355] I. Sugden, C.S. Adjiman, C.C. Pantelides, Accurate and efficient representation of intramolecular energy in ab initio generation of crystal structures. I. Adaptive local approximate models, Acta Crystallogr. B 72 (2016) 864–874, http://dx.doi.org/10.1107/S2052520616015122.
[356] M. Habgood, I.J. Sugden, A.V. Kazantsev, C.S. Adjiman, C.C. Pantelides, Efficient handling of molecular flexibility in ab initio generation of crystal structures, J. Chem. Theory Comput. 11 (2015) 1957–1969, http://dx.doi.org/10.1021/ct500621v.
[357] D.J. Earl, M.W. Deem, Parallel tempering: Theory, applications, and new perspectives, Phys. Chem. Chem. Phys. 7 (2005) 3910–3916, http://dx.doi.org/10.1039/B509983H.
[358] E. Paquet, H.L. Viktor, Molecular dynamics, Monte Carlo simulations, and langevin dynamics: A computational review, BioMed Res. Int. 2015 (2015) e183918, http://dx.doi.org/10.1155/2015/183918.
[359] O. Chapelle, L. Li, An empirical evaluation of thompson sampling, Adv. Neural Inf. Process. Syst. 24 (2011) 2249–2257.
[360] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. 13 (1998) 455–492, http://dx.doi.org/10.1023/A:1008306431147.
[361] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, 2012, ArXiv12062944CsStat. (Accessed 3 May 2021).
[362] A.R. Oganov, A.O. Lyakhov, M. Valle, How evolutionary crystal structure prediction works—and why, Acc. Chem. Res. 44 (2011) 227–237, http://dx.doi.org/10.1021/ar1001318.
[363] A.R. Oganov, C.W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, J. Chem. Phys. 124 (2006) 244704, http://dx.doi.org/10.1063/1.2210932.
[364] Y. Wang, J. Lv, L. Zhu, Y. Ma, Crystal structure prediction via particle-swarm optimization, Phys. Rev. B 82 (2010) 094116, http://dx.doi.org/10.1103/PhysRevB.82.094116.
[365] Y. Zhang, H. Wang, Y. Wang, L. Zhang, Y. Ma, Computer-assisted inverse design of inorganic electrides, Phys. Rev. X 7 (2017) 011017, http://dx.doi.org/10.1103/PhysRevX.7.011017.
[366] E.V. Podryabinkin, A.V. Shapeev, Active learning of linearly parametrized interatomic potentials, Comput. Mater. Sci. 140 (2017) 171–180, http://dx.doi.org/10.1016/j.commatsci.2017.08.031.
[367] S. Demir, A. Tekin, FFCASP: A massively parallel crystal structure prediction algorithm, J. Chem. Theory Comput. 17 (2021) 2586–2598, http://dx.doi.org/10.1021/acs.jctc.0c01197.
[368] A. Emdadi, S. Demir, Y. Kışlak, A. Tekin, Computational screening of dual-cation metal ammine borohydrides by density functional theory, J. Phys. Chem. C 120 (2016) 13340–13350, http://dx.doi.org/10.1021/acs.jpcc.6b01833.
[369] A. Corana, M. Marchesi, C. Martini, S. Ridella, Minimizing multimodal functions of continuous variables with the simulated annealing algorithm-corrigenda for this article is available here, ACM Trans. Math. Software 13 (1987) 262–280, http://dx.doi.org/10.1145/29380.29864.
[370] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I. Foster, A. Jain, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69, http://dx.doi.org/10.1016/j.commatsci.2018.05.018.
[371] A. Khorshidi, A.A. Peterson, Amp: A modular approach to machine learning in atomistic simulations, Comput. Phys. Comm. 207 (2016) 310–324, http://dx.doi.org/10.1016/j.cpc.2016.05.010.
[372] A.S. Abbott, J.M. Turney, B. Zhang, D.G.A. Smith, D. Altarawy, H.F. Schaefer, PES-learn: An open-source software package for the automated generation of machine learning models of molecular potential energy surfaces, J. Chem. Theory Comput. 15 (2019) 4386–4398, http://dx.doi.org/10.1021/acs.jctc.9b00312.
[373] S. Jaeger, S. Fulle, S. Turk, Mol2vec: Unsupervised machine learning approach with chemical intuition, J. Chem. Inf. Model. 58 (2018) 27–35, http://dx.doi.org/10.1021/acs.jcim.7b00616.
[374] P. Avery, C. Toher, S. Curtarolo, E. Zurek, XtalOpt Version r12: An open-source evolutionary algorithm for crystal structure prediction, Comput. Phys. Comm. 237 (2019) 274–275, http://dx.doi.org/10.1016/j.cpc.2018.11.016.
[375] E. Gossett, C. Toher, C. Oses, O. Isayev, F. Legrain, F. Rose, E. Zurek, J. Carrete, N. Mingo, A. Tropsha, S. Curtarolo, AFLOW-ML: A RESTful API for machine-learning predictions of materials properties, Comput. Mater. Sci. 152 (2018) 134–145, http://dx.doi.org/10.1016/j.commatsci.2018.03.075.
[376] P.O. Dral, MLatom: A program package for quantum chemical research assisted by machine learning, J. Comput. Chem. 40 (2019) 2339–2347, http://dx.doi.org/10.1002/jcc.26004.
[377] M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B.U. Kota, A. Sonpal, S. Setlur, J. Hachmann, ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data, WIREs Comput. Mol. Sci. 10 (2020) e1458, http://dx.doi.org/10.1002/wcms.1458.
[378] L. Zhang, Z. Wang, Z. Wei, J. Li, Unsupervised assisted directional design of chemical reactions, Cell Rep. Phys. Sci. 1 (2020) 100269, http://dx.doi.org/10.1016/j.xcrp.2020.100269.
[379] A.K. Halder, M.N. Dias Soeiro Cordeiro, QSAR-Co-X: An open source toolkit for multitarget QSAR modelling, J. Cheminformatics 13 (2021) 29, http://dx.doi.org/10.1186/s13321-021-00508-0.
[380] E.N. Muratov, J. Bajorath, R.P. Sheridan, I.V. Tetko, D. Filimonov, V. Poroikov, T.I. Oprea, I.I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D.A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, QSAR without borders, Chem. Soc. Rev. 49 (2020) 3525–3564, http://dx.doi.org/10.1039/D0CS00098A.
[381] R.A. Lewis, D. Wood, Modern 2D QSAR for drug discovery, WIREs Comput. Mol. Sci. 4 (2014) 505–522, http://dx.doi.org/10.1002/wcms.1187.
[382] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure–activity relationships, J. Chem. Inf. Model. 55 (2015) 263–274, http://dx.doi.org/10.1021/ci500747n.
[383] Y.-C. Lo, S.E. Rensi, W. Torng, R.B. Altman, Machine learning in chemoinformatics and drug discovery, Drug Discov. Today 23 (2018) 1538–1546, http://dx.doi.org/10.1016/j.drudis.2018.05.010.
[384] P. Ambure, A.K. Halder, H. González Díaz, M.N.D.S. Cordeiro, QSAR-Co: An open source software for developing robust multitasking or multitarget classification-based QSAR models, J. Chem. Inf. Model. 59 (2019) 2538–2544, http://dx.doi.org/10.1021/acs.jcim.9b00295.
[385] V. Venkatasubramanian, A. Sundaram, Genetic algorithms: Introduction and applications, in: Encycl. Comput. Chem., American Cancer Society, 2002, http://dx.doi.org/10.1002/0470845015.cga003.
[386] M. Ferrand, B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J.M. Trommenschlager, M. Brochard, Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression, Chemom. Intell. Lab. Syst. 106 (2011) 183–189, http://dx.doi.org/10.1016/j.chemolab.2010.05.004.
[387] L.-F. Chiu, P.-Y. Huang, W.-F. Chiang, T.-Y. Wong, S.-H. Lin, Y.-C. Lee, D.-B. Shieh, Oral cancer diagnostics based on infrared spectral markers and wax physisorption kinetics, Anal. Bioanal. Chem. 405 (2013) 1995–2007, http://dx.doi.org/10.1007/s00216-012-6625-z.
[388] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, http://dx.doi.org/10.1007/BF00058655.
[389] A. Speck-Planche, Recent advances in fragment-based computational drug design: Tackling simultaneous targets/biological effects, Future Med. Chem. 10 (2018) 2021–2024, http://dx.doi.org/10.4155/fmc-2018-0213.

[390] S.-P. Alejandro, N.D.S.C. Maria, Advanced in silico approaches for drug discovery: Mining information from multiple biological and chemical data through mtk-QSBER and pt-QSPR strategies, Curr. Med. Chem. 24 (2017) 1687–1704.
[391] V.V. Kleandrova, J.M. Ruso, A. Speck-Planche, M.N. Dias Soeiro Cordeiro, Enabling the discovery and virtual screening of potent and safe antimicrobial peptides. Simultaneous prediction of antibacterial activity and cytotoxicity, ACS Comb. Sci. 18 (2016) 490–498, http://dx.doi.org/10.1021/acscombsci.6b00063.
[392] A.K. Halder, M.N.D.S. Cordeiro, Probing the environmental toxicity of deep eutectic solvents and their components: An in silico modeling approach, ACS Sustain. Chem. Eng. 7 (2019) 10649–10660, http://dx.doi.org/10.1021/acssuschemeng.9b01306.
[393] A.K. Halder, M.N.D.S. Cordeiro, Development of multi-target chemometric models for the inhibition of class I PI3K enzyme isoforms: A case study using QSAR-Co tool, Int. J. Mol. Sci. 20 (2019) 4191, http://dx.doi.org/10.3390/ijms20174191.
[394] A. Speck-Planche, Multicellular target QSAR model for simultaneous prediction and design of anti-pancreatic cancer agents, ACS Omega 4 (2019) 3122–3132, http://dx.doi.org/10.1021/acsomega.8b03693.
[395] A. Speck-Planche, M.T. Scotti, BET bromodomain inhibitors: Fragment-based in silico design using multi-target QSAR models, Mol. Divers. 23 (2019) 555–572, http://dx.doi.org/10.1007/s11030-018-9890-8.
[396] V.V. Kleandrova, M.T. Scotti, L. Scotti, A. Nayarisseri, A. Speck-Planche, Cell-based multi-target QSAR model for design of virtual versatile inhibitors of liver cancer cell lines, SAR QSAR Environ. Res. 31 (2020) 815–836, http://dx.doi.org/10.1080/1062936X.2020.1818617.
[397] A.K. Halder, A.K. Giri, M.N.D.S. Cordeiro, Multi-target chemometric modelling, fragment analysis and virtual screening with ERK inhibitors as potential anticancer agents, Molecules 24 (2019) 3909, http://dx.doi.org/10.3390/molecules24213909.
[398] A.K. Halder, A. Melo, M.N.D.S. Cordeiro, A unified in silico model based on perturbation theory for assessing the genotoxicity of metal oxide nanoparticles, Chemosphere 244 (2020) 125489, http://dx.doi.org/10.1016/j.chemosphere.2019.125489.
[399] C.E. Rasmussen, C. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.
[400] R. Reed, R.J. MarksII, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, Mit Press, 1999.
[401] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009.
[402] M. Rupp, Machine learning for quantum mechanics in a nutshell, Int. J. Quantum Chem. 115 (2015) 1058–1073, http://dx.doi.org/10.1002/qua.24954.
[403] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, Assessment and validation of machine learning methods for predicting molecular atomization energies, J. Chem. Theory Comput. 9 (2013) 3404–3419, http://dx.doi.org/10.1021/ct400195d.
[404] P.O. Dral, O.A. von Lilienfeld, W. Thiel, Machine learning of parameters for accurate semiempirical quantum chemical calculations, J. Chem. Theory Comput. 11 (2015) 2120–2125, http://dx.doi.org/10.1021/acs.jctc.5b00141.
[405] B. Settles, Active learning, Synth. Lect. Artif. Intell. Mach. Learn. 6 (2012) 1–114, http://dx.doi.org/10.2200/S00429ED1V01Y201207AIM018.
[406] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.
[407] E.J. Corey, General methods for the construction of complex molecules, Pure Appl. Chem. 14 (1967) 19–38.
[408] E.J. Corey, W.T. Wipke, Computer-assisted design of complex organic syntheses, Science 166 (1969) 178–192.
[409] E.J. Corey, A.K. Long, Computer-assisted synthetic analysis. Performance of long-range strategies for stereoselective olefin synthesis, J. Organic Chem. 43 (1978) 2208–2216, http://dx.doi.org/10.1021/jo00405a027.
[410] T.D. Salatin, W.L. Jorgensen, Computer-assisted mechanistic evaluation of organic reactions. 1. Overview, J. Organic Chem. 45 (1980) 2043–2051.
[411] H. Satoh, K. Funatsu, SOPHIA, a knowledge base-guided reaction prediction system-utilization of a knowledge base derived from a reaction database, J. Chem. Inf. Comput. Sci. 35 (1995) 34–44.
[412] P. Röse, J. Gasteiger, Automated derivation of reaction rules for the EROS 6.0 system for reaction prediction, Anal. Chim. Acta 235 (1990) 163–168, http://dx.doi.org/10.1016/S0003-2670(00)82071-1.
[413] M.A. Kayala, C.-A. Azencott, J.H. Chen, P. Baldi, Learning to predict chemical reactions, J. Chem. Inf. Model. 51 (2011) 2209–2222, http://dx.doi.org/10.1021/ci200207y.
[414] J.N. Wei, D. Duvenaud, A. Aspuru-Guzik, Neural networks for the prediction of organic chemistry reactions, ACS Cent. Sci. 2 (2016) 725–732, http://dx.doi.org/10.1021/acscentsci.6b00219.
[415] M.H.S. Segler, M.P. Waller, Modelling chemical reasoning to predict and invent reactions, Chem. Eur. J. 23 (2017) 6118–6128, http://dx.doi.org/10.1002/chem.201604556.
[416] V.L. Deringer, M.A. Caro, G. Csányi, Machine learning interatomic potentials as emerging tools for materials science, Adv. Mater. 31 (2019) 1902765, http://dx.doi.org/10.1002/adma.201902765.
[417] D. Yoo, K. Lee, W. Jeong, D. Lee, S. Watanabe, S. Han, Atomic energy mapping of neural network potential, Phys. Rev. Mater. 3 (2019) 093802, http://dx.doi.org/10.1103/PhysRevMaterials.3.093802.
[418] J.D. Head, M.C. Zerner, A Broyden—Fletcher—Goldfarb—Shanno optimization procedure for molecular geometries, Chem. Phys. Lett. 122 (1985) 264–270, http://dx.doi.org/10.1016/0009-2614(85)80574-1.
[419] B. Hammer, J.K. Nørskov, Theoretical surface science and catalysis—calculations and concepts, Adv. Catal. 45 (2000) 71–129, http://dx.doi.org/10.1016/S0360-0564(02)45013-4.
[420] X. Ma, Z. Li, L.E.K. Achenie, H. Xin, Machine-learning-augmented chemisorption model for CO_2 electroreduction catalyst screening, J. Phys. Chem. Lett. 6 (2015) 3528–3533, http://dx.doi.org/10.1021/acs.jpclett.5b01660.
[421] Z.W. Ulissi, M.T. Tang, J. Xiao, X. Liu, D.A. Torelli, M. Karamad, K. Cummins, C. Hahn, N.S. Lewis, T.F. Jaramillo, K. Chan, J.K. Nørskov, Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction, ACS Catal. 7 (2017) 6600–6608, http://dx.doi.org/10.1021/acscatal.7b01648.
[422] W. Beker, E.P. Gajewska, T. Badowski, B.A. Grzybowski, Prediction of major regio-, site-, and diastereoisomers in Diels–Alder reactions by using machine-learning: The importance of physically meaningful descriptors, Angew. Chem. Int. Ed. 58 (2019) 4515–4519, http://dx.doi.org/10.1002/anie.201806920.
[423] C.W. Coley, W. Jin, L. Rogers, T.F. Jamison, T.S. Jaakkola, W.H. Green, R. Barzilay, K.F. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity, Chem. Sci. 10 (2019) 370–377, http://dx.doi.org/10.1039/C8SC04228D.
[424] Z. Wang, Y. Han, J. Li, X. He, Combining the fragmentation approach and neural network potential energy surfaces of fragments for accurate calculation of protein energy, J. Phys. Chem. B 124 (2020) 3027–3035, http://dx.doi.org/10.1021/acs.jpcb.0c01370.
[425] A.V. Uriarte-Arcia, I. López-Yáñez, C. Yáñez Márquez, One-hot vector hybrid associative classifier for medical data classification, PLoS One 9 (2014) e95715, http://dx.doi.org/10.1371/journal.pone.0095715.
[426] J.A. Hartigan, M.A. Wong, A K-means clustering algorithm, J. R. Stat. Soc. Ser. C. Appl. Stat. 28 (1979) 100–108, http://dx.doi.org/10.2307/2346830.
[427] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666, http://dx.doi.org/10.1016/j.patrec.2009.09.011.
[428] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, http://dx.doi.org/10.1145/1961189.1961199.
[429] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008).

[430] L. Van Der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.
[431] M. Sacchi, D.J. Wales, S.J. Jenkins, Mode-specificity and transition state-specific energy redistribution in the chemisorption of CH_4 on Ni{100}, Phys. Chem. Chem. Phys. 14 (2012) 15879–15887, http://dx.doi.org/10.1039/C2CP42345F.
[432] I. Chorkendorff, I. Alstrup, S. Ullmann, Xps study of chemisorption of CH_4 on Ni(100), Surf. Sci. 227 (1990) 291–296, http://dx.doi.org/10.1016/S0039-6028(05)80017-2.
[433] T. Puzyn, A. Gajewicz, D. Leszczynska, J. Leszczynski, Nanomaterials–the next great challenge for QSAR modelers, in: Recent Adv. QSAR Stud., Springer, 2010, pp. 383–409.
[434] K. Roy, Advances in QSAR modeling, Appl. Pharm. Chem. Food Agric. Environ. Sci. Springer Cham Switz. 555 (2017) 39.
[435] A. Kumar, S. Chauhan, QSAR differential model for prediction of SIRT1 modulation using Monte Carlo method, Drug Res. 67 (2017) 156–162.
[436] E. Amata, A. Marrazzo, M. Dichiara, M.N. Modica, L. Salerno, O. Prezzavento, G. Nastasi, A. Rescifina, G. Romeo, V. Pittalà, Comprehensive data on a 2D-QSAR model for Heme Oxygenase isoform 1 inhibitors, Data Brief. 15 (2017) 281–299, http://dx.doi.org/10.1016/j.dib.2017.09.036.
[437] J.F. Aranda, D.E. Bacelo, M.S.L. Aparicio, M.A. Ocsachoque, E.A. Castro, P.R. Duchowicz, Predicting the bioconcentration factor through a conformation-independent QSPR study, SAR QSAR Environ. Res. 28 (2017) 749–763, http://dx.doi.org/10.1080/1062936X.2017.1377765.
[438] S. Tazuke, S. Kurihara, H. Yamaguchi, T. Ikeda, Photochemically triggered physical amplification of photoresponsiveness, J. Phys. Chem. 91 (1987) 249–251.
[439] K. Roy, S. Kar, R.N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts, Springer, 2015.
[440] Eriksson Lennart, Jaworska Joanna, P.Worth Andrew, T.D.Cronin Mark, M.McDowell Robert, Gramatica Paola, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375, http://dx.doi.org/10.1289/ehp.5758.
[441] T. Kalliokoski, C. Kramer, A. Vulpetti, P. Gedeck, Comparability of mixed IC50 data – A statistical analysis, PLoS One 8 (2013) e61007, http://dx.doi.org/10.1371/journal.pone.0061007.
[442] F.J. Romero-Durán, N. Alonso, M. Yañez, O. Caamaño, X. García-Mera, H. González-Díaz, Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives, Neuropharmacology 103 (2016) 270–278, http://dx.doi.org/10.1016/j.neuropharm.2015.12.019.
[443] A. Speck-Planche, V.V. Kleandrova, J.M. Ruso, M.N.D.S. Cordeiro, First multitarget chemo-bioinformatic model to enable the discovery of antibacterial peptides against multiple gram-positive pathogens, J. Chem. Inf. Model. 56 (2016) 588–598, http://dx.doi.org/10.1021/acs.jcim.5b00630.
[444] A. Speck-Planche, M.N.D.S. Cordeiro, Fragment-based in silico modeling of multi-target inhibitors against breast cancer-related proteins, Mol. Divers. 21 (2017) 511–523, http://dx.doi.org/10.1007/s11030-017-9731-1.
[445] G.M. Casañola Martin, H. Le-Thi-Thu, F. Pérez-Giménez, Y. Marrero-Ponce, M. Merino-Sanjuán, C. Abad, H. González-Díaz, Multi-output model with Box–Jenkins operators of linear indices to predict multi-target inhibitors of ubiquitin–proteasome pathway, Mol. Divers. 19 (2015) 347–356, http://dx.doi.org/10.1007/s11030-015-9571-9.
[446] T. Hill, P. Lewicki, P. Lewicki, Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining, StatSoft, Inc., 2006.
[447] P. Ambure, R.B. Aher, A. Gajewicz, T. Puzyn, K. Roy, NanoBRIDGES software: Open access tools to perform QSAR and nano-QSAR modeling, Chemom. Intell. Lab. Syst. 147 (2015) 1–13, http://dx.doi.org/10.1016/j.chemolab.2015.07.007.
[448] P. Ambure, J. Bhat, T. Puzyn, K. Roy, Identifying natural compounds as multi-target-directed ligands against Alzheimer's disease: An in silico approach, J. Biomol. Struct. Dyn. 37 (2019) 1282–1306, http://dx.doi.org/10.1080/07391102.2018.1456975.
[449] E. Besalú, X. Gironés, L. Amat, R. Carbó-Dorca, Molecular quantum similarity and the fundamentals of QSAR, Acc. Chem. Res. 35 (2002) 289–295, http://dx.doi.org/10.1021/ar010048x.
[450] J.J. Sutherland, L.A. O'Brien, D.F. Weaver, Spline-fitting with a genetic algorithm: A method for developing classification structure-activity relationships, J. Chem. Inf. Comput. Sci. 43 (2003) 1906–1915, http://dx.doi.org/10.1021/ci034143r.
[451] P. Ambure, K. Roy, Understanding the structural requirements of cyclic sulfone hydroxyethylamines as hBACE1 inhibitors against Aβ plaques in Alzheimer's disease: A predictive QSAR approach, RSC Adv. 6 (2016) 28171–28186, http://dx.doi.org/10.1039/C6RA04104C.
[452] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemom. Intell. Lab. Syst. 145 (2015) 22–29, http://dx.doi.org/10.1016/j.chemolab.2015.04.013.
[453] M. Mathea, W. Klingspohn, K. Baumann, Chemoinformatic classification methods and their applicability domain, Mol. Inform. 35 (2016) 160–180, http://dx.doi.org/10.1002/minf.201501019.
[454] A. Speck-Planche, M.N.D.S. Cordeiro, De novo computational design of compounds virtually displaying potent antibacterial activity and desirable in vitro ADMET profiles, Med. Chem. Res. 26 (2017) 2345–2356, http://dx.doi.org/10.1007/s00044-017-1936-4.
[455] T. Menzies, E. Kocaguneli, B. Turhan, L. Minku, F. Peters, Sharing Data and Models in Software Engineering, Morgan Kaufmann, 2014.
[456] S.S. Wilks, Certain generalizations in the analysis of variance, Biometrika (1932) 471–494.
[457] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory. 13 (1967) 21–27, http://dx.doi.org/10.1109/TIT.1967.1053964.
[458] A. McCallum, K. Nigam, A comparison of event models for Naive Bayes text classification, in: AAAI-98 Workshop Learn. Text Categ., Citeseer, 1998, pp. 41–48.
[459] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proc. Fifth Annu. Workshop Comput. Learn. Theory, Association for Computing Machinery, New York, NY, USA, 1992, pp. 144–152, http://dx.doi.org/10.1145/130385.130401.
[460] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[461] G.-B. Huang, H.A. Babri, Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions, IEEE Trans. Neural Netw. 9 (1998) 224–229, http://dx.doi.org/10.1109/72.655045.