Data-driven approach of discovering organic photocatalysts and developing molecular force field by machine learning

Dissertation submitted to

University of Liverpool

in partial fulfilment of the requirement

for the degree of

Doctor of Philosophy

in

Chemistry

by

Yu Che

Primary Supervisor:	Professor Andrew I. Cooper
Secondary Supervisor:	Professor Edward O. Pyzer-Knapp
Secondary supervisor:	Doctor Linjiang Chen

May, 2023

PGR Policy on Plagiarism and Dishonest Use of Data
PGR CoP Appendix 4 Annexe 1

PGR DECLARATION OF ACADEMIC HONESTY

NAME (Print)	**Yu Che**
STUDENT NUMBER	**201158882**
SCHOOL/INSTITUTE	**Department of Chemistry**
TITLE OF WORK	**Data-driven approach of discovering organic photocatalysts and developing molecular force field by machine learning**

This form should be completed by the student and appended to any piece of work that is submitted for examination. Submission by the student of the form by electronic means constitutes their confirmation of the terms of the declaration.

Students should familiarise themselves with Appendix 4 of the PGR Code of Practice: PGR Policy on Plagiarism and Dishonest Use of Data, which provides the definitions of academic malpractice and the policies and procedures that apply to the investigation of alleged incidents.

Students found to have committed academic malpractice will receive penalties in accordance with the Policy, which in the most severe cases might include termination of studies.

<u>STUDENT DECLARATION</u>
I confirm that:

- I have read and understood the University’s PGR Policy on Plagiarism and Dishonest Use of Data.

- I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.

- I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached material.

- I have not copied material from another source, nor colluded with any other student in the preparation and production of this material.

- If an allegation of suspected academic malpractice is made, I give permission to the University to use source-matching software to ensure that the submitted material is all my own work.

SIGNATURE………………………………………...……………

DATE…………………………...

ABSTRACT

Machine learning techniques are becoming more prevalent in chemistry research as they offer an effective approach for handling large, complex chemical datasets generated from high-throughput experiments and molecular simulations. To gain a comprehensive understanding of datasets, it is crucial to employ efficient methods for data representation and analysis. This PhD project utilized classical machine learning algorithms to effectively visualize high-dimensional chemical data, ascertain connections between chemical structure and properties, facilitate the discovery of novel organic catalysts, and developing a machine learning potential to describe intermolecular interactions.

The primary application of machine learning involves examining a large, intricate database of multidimensional organic molecular crystal structures generated through previous research conducted by our group. By applying unsupervised learning techniques to calculated pore descriptors, a set of 'landmark' structures was identified systematically from the dataset. The two-dimensional embedding of these structures became human interpretable after employing dimensionality reduction algorithms. Additionally, an interactive web explorer was developed to facilitate data sharing and to showcase the findings of the study. To speed up the discovery of organic molecular crystal structure calculations, a machine learning force field based on TorchANI was developed to describe the intermolecular interactions.

Then, a combination of machine learning and high throughput experimentation of organic photocatalysts was used to visualize, interpret, and reveal the feature-activity correlations. This study visually mapping the relationship between structure and hydrogen evolution activity correlations by using unsupervised learning techniques. A virtual experiment was conducted on the aforementioned measured dataset, demonstrating that the use of algorithms as an experimental advisor can significantly reduce the cost of conducting experiments. In order to leverage the benefits of algorithmic assistance in catalyst discovery, a closed-loop discovery strategy was applied by integrating Bayesian optimization to identify promising organic photoredox catalysts (OPCs) from a set of 560 designed candidate molecules and to determine their optimal reaction conditions. The identified OPC formulation was found to be competitive with iridium catalysts at high nickel concentrations and outperformed iridium catalysts at lower nickel concentrations.

Through the completion of the above research projects, I developed an interactive web explorer that utilizes machine learning techniques to view large organic molecular datasets and extended the ANI machine learning potential to describe the organic intermolecular interactions. I also explored the structure-properties relationships of organic photocatalysts and employed Bayesian optimization in routine laboratory procedures to guide the discovery of catalysts. Overall, these studies make a step forward the machine learning assisted discovery of organic materials allowing chemists to systematically consider a much broader chemical space than than previously possible.

Keywords: machine learning; machine learning potential; data visualization; Bayesian optimization; Gaussian processes; photocatalyst

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESUME

Publication list and their links to chapters

1. C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day and A. I. Cooper, Digital navigation of energy-structure-function maps for hydrogen-bonded porous molecular crystals, Nature Communications, 2021, 12, 817
2. X. Wu, Y. Che, L. Chen, E. J. Amigues, R. Wang, J. He, H. Dong and L. Ding, Mapping the Porous and Chemical Structure-Function Relationships of Trace CH3I Capture by Metal-Organic Frameworks using Machine Learning, ACS Applied Materials & Interfaces, 2022, 14, 47209-47221
3. X. Li, P. M. Maffettone, Y. Che, T. Liu, L. Chen and A. I. Cooper, Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules, Chemical Science, 2021, 12, 10742-10754
4. X, Li, Y. Che, L. Chen, T. Liu, E. O. Pyzer-Knapp, and A. I. Cooper, Sequential Bayesian Optimization for the Discovery of Metallophotocatalysis Formulations (in preparation)

- Chapter 2 of this work includes the development of an interactive web explorer and the identification of ‘landmark’ structures, which was published in paper 1. The web application was subsequently expanded and employed for visualization purposes in papers 2 and 3. Chapter 3 focuses on the machine learning analysis of electronic photocatalyst features, which was published in paper 3. Chapter 4 outlines the closed-loop photocatalyst discovery project, currently undergoing preparations for publication as paper 4.
- All experiments were conducted by collaborator Xiaobo Li, while Tao Liu made significant contributions to the quantum calculations, particularly in the calculation of photocatalytic properties and molecular structure optimization.

CHAPTER 1 INTRODUCTION

1.1 Machine learning sparks in chemistry

Looking for a pattern of mapping chemical structures to desired properties is one of the major expectation of computational chemists. Based upon derivations of classic chemical theory, the background knowledge from experiments, and chemical intuition, traditional chemical discovery relied increasingly on the rapid evolution of experimental data and the use of computers.[1-3] As large dataset generated by experiments or computational, an experienced chemist may have the insight into the complex chemical results, but it needs decades training and research whereas machine learning (ML) model as a statistic tool has the ability to accelerate this discovery and determine psychical laws without human intuition.[4-10] The classic material research based on the knowing chemical knowledge can be considered as the 'First generation' of exploration approach.[11] Future development of equipments drives the high throughput experiment and computation to boost the discovery process as the 'Second generation', whereas the emerging ML workflow can corporate prior knowledge and the advantage of statistic leads the 'Third generation' procedure,[2,11] as illustrated below:

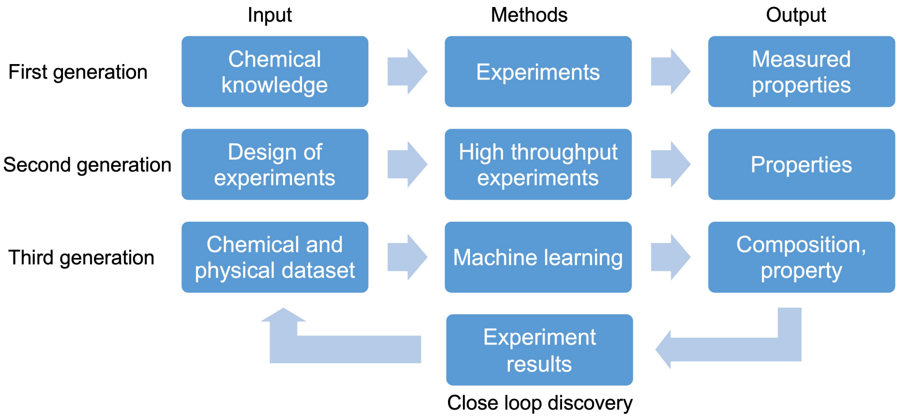


Figure 1.1 Evolution of the research workflow in material discovery

With increased availability of datasets generated and collected in high-throughput simulations and experiments, machine learning techniques have emerged as promising avenues of probing and understanding the rich information[4,12-13] and structure-property relationships,[14] as well as assisting experiments,[15-19] by discovering hidden patterns in the vast amount of data. A typical workflow of data-driven material discovery, comprising data acquisition (either generated on the fly or retrieved from existing databases), fertilisation, machine learning and visualization. Such exercises can be conveniently (semi-)automated by scripting and using open-source software such as matminer.[20] These data-driven approaches hold the promising towards speeding up both the experimental chemical discovery, and theoretical modelling, so that chemists can quickly gain insights into the designed or screened chemical spaces,[21] discover structure-properties relationships,[22] and be guided by closed-loop autonomous evolution beyond their general chemical intuition.[23-27]

1.1.1 Machine learning predictions based on data from quantum calculations

There always exist two contradictory requirements in computational chemistry: accuracy simulation of electronic features and short timescales. The high level electronic structure calculation requests heavy computation resources and fast modelling techniques such as force filed only have rough estimation of the molecular potential energy surface.[28] Such gap can be fill by developing ML techniques that combine electronic structure calculations and statistical analysis tools. For example, von Lilienfeld and co-workers[29-31] have demonstrated that machined learned models, using large datasets, can recover what is absent between two different levels of theory, delivering a computational framework that achieves quantum chemical accuracy at a fraction of its normal computational cost. One major development is artificial neural network potential (ANN),[32-34] which increased the accuracy of the force filed by transfer learning approaching CCSD(T) level.[35] Additionally, various machine learning (ML) algorithms have been employed in the development of machine learning force fields, such as 'Gaussian-Approximated Potential' (GAP) method,[36-37] based on Gaussian process regression (GPR). Another approach of simulating interatomic potential is achieved by physically informed neural network that increase the transferability of ML potentials to unknown structures.[38] Those algorithms require advanced representation of chemical structure to capture the intricate interatomic interactions.[39-40]

In addition to the development of an accuracy machine learning force field, another area where these methodologies find application is in the field of in-silico crystal structure prediction (CSP).[41-42] This application aims to support process risk management and facilitate the rapid screening of potential materials, particularly during the initial stages of crystallization process development in the pharmaceutical industry. Machine learning techniques have the potential to effectively address the challenge posed by the vast search space resulting from the increasing number of configurations.[43] In general, an ML model can predict the crystal lattice energies from crystal structures using 1/5[44] to 1/10000[45] time compared with expensive density functional theory (DFT) calculations based on the symmetry functions and ANN model.[32,34,46] Recently, Wang, Wang, Zhao, Du, Xu, Gu, and Duan developed a molecular dynamics graph neural network that incorporates high-order terms of interatomic distances. This network successfully reproduces the force fields of molecules and crystals derived from both classical and first-principles molecular dynamics simulations.

Another application including the prediction of functional materials properties using machine learning can be driven back to 1990s, where new magnetic and electro-optical materials with specific crystal structures were predicted.[48] However, there has been a substantial increase in the number of studies in this field since 2010.[49-50] Graph based neural networks show great promise for property predictions of crystalline materials.[51-52] The prediction of lattice parameters,[53] formation energies,[54] structure-energy-property relationships of molecular crystals,[55] and infrared spectra[56] can be achieved via *in-silico*. For example, Xie and Grossman used crystal graph convolutional neural networks to predict eight different properties of crystals including formation energy, absolute energy, Fermi energy, band gap etc. Similarity, various properties of inorganic crystals were predicted using a gradient boosting decision tree.[57] These properties include metal/insulator classification, bandgap energy, bulk/shear moduli, Debye temperature, and heat capacities. The developed model achieved a high prediction accuracy ranging from 88% to 95%.

The majority of machine learning studies on crystal solids have primarily focused on a specific type of crystal structure. This is primarily due to the challenge of representing crystalline solids in a format that can be readily utilized by statistical learning algorithms. By concentrating on a single structure type, the representation is inherently integrated into the model. However, the development of flexible and transferable representations is an important research area within machine learning for crystalline solids.

1.1.2 Machine learning for high throughput materials discovery

Except *in-silico* design of material discovery, ML methods have been increasingly applied to complement experiments when studying complex chemical systems, as exemplified by accelerated discoveries of drugs,[58] catalysts,[14,59-64] adsorbent materials,[65] and batteries.[66] Both generating predictive models and gaining physical insights are essential tasks: the former allows for fast, in silico pre-evaluation of potential candidates through quantitative prediction, qualitative ranking or coarse-grained filtering; the latter is important for probing and understanding the relationship between a material's structure or physicochemical characteristics and its functional properties or activities. Classic ML models help to get reliable relationships by extracting the features importance and make algorithms interpretable. Since AlphaGo, a computer program that plays the board game Go, defeated the word champion Lee Sedol in March 2016,[67] the explosive development and deployment of deep learning, also in the form of deep neural networks, has catalysed technological revolutions in many research areas. Computer assisted synthesis planning with neural networks demonstrates the success of artificial intelligence (AI) in guiding chemists to find better synthetic routes.[5] DL models are trained to predict chemical reactions from reactants to products,[68] following the principles of language translation, using sequence-to-sequence neural networks.

The timescale of developing new materials from chemistry laboratory to a practical application can easily be decades.[7,69] A typical, current, integrated approach to functional material discovery starts with researchers coming up with a new material concept and predicting the performance metric, then proceeds to laboratory synthesis and characterization, and finally feeds information back to the researchers to think about possible improvement or re-design of the material. Each step of this cycle can take such long time that researchers are developing computational methods to boost the material design cycle, reducing the time cost for each step. Efforts to accelerate this process are focused on methodologies of high throughput virtual screening (HTVS),[70] in which computations are performed on large numbers of possible candidates to identify the best-performing ones, often leveraging the large throughputs using robotic platforms. For example, computational calculation assisted HTVS contributes to the discovery of water-splitting photocatalysts.[71] Bai *et al.* screened 6354 candidates co-polymers computationally and identified 99 of them, assisted by machine learning, for the latter experimental, robotic make-and-measure workflow.

Beyond these high-throughput virtual screening, active learning has become a very powerful avenue of research to help this acceleration.[72-75] A recent study demonstrated a machine learning-assisted workflow for optimizing catalyst designs, resulting in the discovery of new and optimal catalysts at a faster rate.[14] Bayesian optimization was used to drive this discovery of new catalysts in a designed chemical space consisting of more than 800 catalysts. Another systematic study of Bayesian reaction optimization proved the algorithm outperforms human decision in efficiency and consistency.[76] This close loop discovery also links other techniques, such as inverse design,[69,77] to accelerate and enable the end-to-end material discovery. The inverse design directly generates chemical structures with desired properties by searching the functional space constructed by machined learned existing data.[78-79] The main difference between HTVS and active learning is that ML models observe new chemical knowledge of both structure and property beyond the given data whereas HTVS cannot. Inverse design has been demonstrated by, for example, using variational autoencoders,[80-82] and generative models.[77,83]

An additional crucial step in experimental material discovery involves determining the appropriate reaction conditions, including the crystallization conditions for porous materials and the synthesis conditions for organic compounds. Considering the vast amount of experimental procedures available for synthesizing porous materials, numerous efforts have been made to mine or extract this collective knowledge to training ML models for prediction. The prediction of zeolite synthesis has been achieved using decision trees that rely on parameters describing the synthetic conditions.[84] Likewise, Jensen *et al.* developed a workflow to extract information on conditions from a dataset of 70,000 published papers. This dataset was then utilized to build a model capable of predicting the density of materials. However, extracting reaction conditions directly from published papers may introduce bias since many unsuccessful or failed synthesis routes are often omitted, and only successful ones are typically reported.[15] One example to prof this is Moosavi *et al.* shown how to learn from the failed and partially successful experiments to synthesis metal-organic frameworks(MOFs). The challenges associated with ML-driven reaction optimization are substantial, and the key requirement is the need for publicly accessible sharing of all synthesis attempts.

1.2 Encoding chemical data

1.2.1 Topological methods

In the field of chemical and molecular science, an early 'data-driven' philosophy emerged in response to the vast combinatorial space of possible molecules[87] and the relatively straightforward synthetic strategies available for exploring this space. This philosophy gave rise to the development of Quantitative Structure-Property Relationship (QSPR) techniques,[58,88-89] which attempt to map descriptors of molecular structures to the behaviours of a chosen compound. These descriptors can be derived from various approaches, including extended connectivity fingerprints,[90] chemical-intuition driven descriptors,[91-92] molecular graphs,[93-94] or indicators obtained from quantum chemical calculations.[95] The focus is typically on predicting properties of direct applicative interest,[92] such as solubility, toxicity,[96] or reaction activity.[14]

Before being fed into machine learning algorithms, chemical information related to the structure of a molecule or crystal structure needs to undergo pre-processing to make it machine-readable. While the ultimate goal of obtaining absolute accuracy in describing chemical properties lies in solving the Schrödinger equation to determine electronic properties and geometry, this approach is computationally challenging and resource-intensive. In most cases, 3D Cartesian coordinators are not an efficient way, or a correct one, of representing chemical structures when coupled with ML algorithms, because one such combination does not contain any chemical or physical theory (Schrödinger equation). A good representation needs to capture as much chemical information as possible, in addition to satisfying other principles such as being invariant to rigid motions. Therefore, scientists have made significant efforts in developing both engineered descriptors and numerical representations, most of which fall into three categories depending on the data formation:[69]

- discrete (string),
- continuous (tensors),
- weighted graphs

These engineered cheminformatics descriptors are constructed in an *ad hoc* manner, combining descriptors related to molecular structure, composition, and easily estimable molecular properties. They typically require a significant amount of *prior* knowledge, are often specific to a particular system or problem, and are designed to label a compound rather than a specific configuration of its atoms. This is because QSPR aims to provide an end-to-end description of a thermodynamic property, which is not an attribute of a single configuration but of a thermodynamic state of matter.[39]

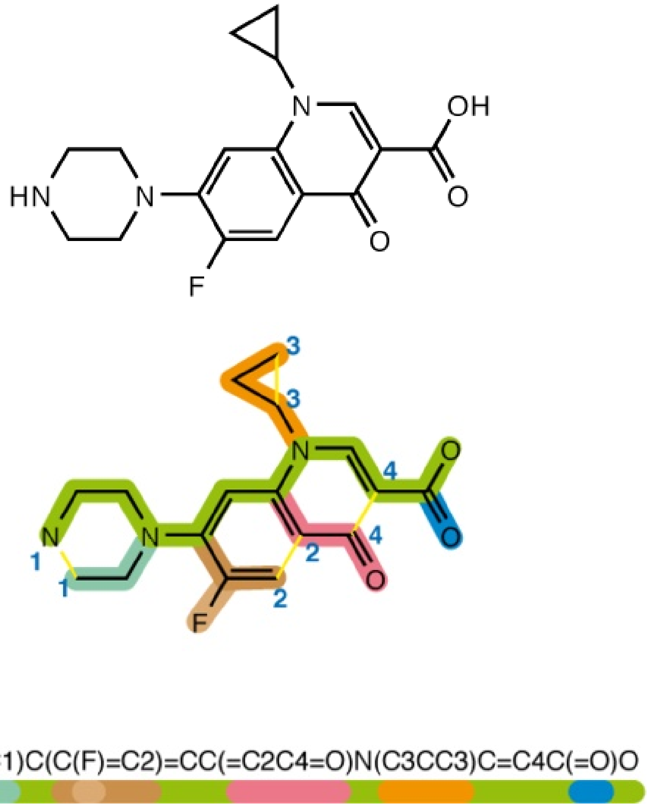


Figure 1.2 A ciprofloxacin molecule represented by SMILES and coloured by branches. Original by Fdardel, slight edit by DMacks (https://commons.wikimedia.org/wiki/File:SMILES.png). Copyright CC BY-SA 3.0

The most straightforward method to encode cheminformatics is using their 2D structures to evaluate the similarity between the topological structure. The simplified molecular-input line-entry system (SMILES),[97-99] invented by David Weiniger, is a string representation for molecular structure, which encodes chemical species into a sequence notation using a depth-first graph traversal. It is a simple, yet powerful, tool to enumerate molecular structures from a library of building blocks (Fig 1.2). For example, a molecule comprising multiple fragments connected via single covalent bonds can be expressed in SMILES as a simple concatenation of the fragment SMILES in the same ordering. Some basic rules are:

- Atoms are written by standard abbreviation of their element names.
- Bonds are represented using -, =, #, and $ for single, double, triple and quadruple bonds, respectively.
- A backbone is arbitrary chosen as the main written string in a molecule (coloured by green in Fig 1.2).
- Branches are described in paired bracket and right behind the connection atom in the backbone (multiple coloured in Fig 1.2).
- Rings are broken at an arbitrary point as two branches and adding numerical ring closure labels to mark the connectivity between two breakpoints (numbered as 1, 2, 3, and 4 in Fig 1.2).

Except these basic rules, aromatic atoms can be written in lower-case forms and the aromatic bond notation is commonly omitted as well as the single bond and hydrogen atom notation. Chemists can use SMILES format as a starting point to build a chemical database because of simplicity and compatibility. Therefore, the character string can be converted to 2D chemical structure and 3D Cartesian coordinate by applying standard force field optimization in many packages such as RDkit.[100]

Besides, SMILES arbitrary target specification (SMARTS), related to the SMILES line notation, is a more flexible chemical language to describe molecules. The chemical reaction string utility (e.g., in RDKit) written in SMARTS is a particularly powerful tool to design a group of chemicals by input different reactants. However, the limitation of SMILES is many fractions of strings do not correspond to a valid molecule which could block the application on generative models to design novel functional materials. Krenn *et al.* developed a robust method, Self-referencing embedded strings (SELFIES), to extend the application of such string-based representation.[101] Those string format can be converted to a one-hot-vector as the input features of many linear deep neural networks.[70]

The design of SMILES shares similarities with certain cheminformatics methods based on graph theory. SMILES can be viewed as a graph-based representation where each character in the SMILES string corresponds to a node, and the bonds between atoms are represented as edges. To convert the graph representation into numerical data, the chemical structure graph is fragmented, and pairwise similarity is measured by counting the number of common substructures or pieces shared by two chemicals.[94,102] The original design of these algorithms, such as Morgan fingerprints,[103] is used for substructure searching, but later for analysing molecular similarity. By dividing the chemical structure into fragments and comparing the presence of substructures, these algorithms enable the assessment of molecular similarity based on shared features.

Extended-connectivity fingerprints (ECFPs)[90] is a modified algorithm from Morgan fingerprints for molecular characterization. Morgan algorithm is an iterative process assigning numerical numbers as the atomic environment identifier. The initial process encodes the invariant atomic information into an atomic identifier, and followed by using the generated identifiers from previous iteration. The iteration process continues until every atomic identifier is unique. The final result constitute the invariant fingerprints of the molecule, and the intermediate results are discarded during the iteration.

Because the absolute disambiguation strategy in Morgan algorithm, each substructure in every iteration is carefully assigned to avoid mathematical 'collision', that is two different environments are given the same identifier. Therefore, this process has the side effect that can assign different identifiers to two identical atom environment. ECFPs abandons this expensive encoding process by a fast hash scheme resulting a saving of computational effort during the fingerprint generation compared with Morgan algorithm.[90] Also, ECFPs terminate at a user determined number of iterations instead of achieving maximum disambiguation of identifiers. The predefined iteration number is the maximum radius of counted substructure, and named as ECFP4 or ECFP6 for 2 and 3 iteration, receptively. The initial assignment process in ECFPs contains seven atomic information (labelled as $a_1, a_2, \ldots, a_7$ in Algorithm 1.1) and hash this numbers as an integer identifier for each atom. The hash function is used to make sure as much as unique the integer assigned for each atom. The initial atom identifiers are collected as the fingerprint array. Then, an iterative update stage is performed to capture substructures in molecule and including bond types. The bond to hydrogen and hydrogen atoms are ignored in each stage. Finally, duplicate identifiers are removed after the final iteration, and the remaining identifiers within the initial fingerprint array define the ECFP fingerprint (Algorithm 1.2). Because each identifier represent a unique substructure, the similarity calculation comes as counting how many identifiers are same and different in two fingerprint sets.

Algorithm 1.1 Initial assignment of atomic identifier[90]

```
Input: An atom from a molecule
Output: $\eta$, hashed identifier
begin
    $a_1 \leftarrow$ The number of heavy atom neighbours;
    $a_2 \leftarrow$ The valence minus the number of hydrogen;
    $a_3 \leftarrow$ The atomic number;
    $a_4 \leftarrow$ The atomic mass;
    $a_5 \leftarrow$ The atomic charge;
    $a_6 \leftarrow$ The number of attached hydrogen;
    $a_7 \leftarrow$ Weather the atom is in a ring;
    $\eta \leftarrow HASH([a_1, a_2, \ldots, a_7])$;
end
```

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcap B|}{|A| + |B| - |A \bigcap B|} \tag{1.1}$$

Algorithm 1.2 ECFP generation process[90]

```
Input: A molecule structure
Data: $R$, The radius of the largest feature
Output: F
F ← an empty array // The list of fingerprints;
forall atoms $\alpha$ of molecule do
    $\eta$ ← using Algorithm 1.1 ($\alpha$);
    if $\eta$ not in F then
        $\eta$ append to F;
    end
end
for $r = 1$ to $R$ do
    foreach $\alpha$ in $F$ do
        $\mathbf{\Omega} \leftarrow (r, \alpha)$;
        $B$ ← All attached atoms of structure $\alpha$, $B \subset F$;
        for $\beta$ in $B$ do
            $\sigma$ ← The bond order for the bond between atom $\beta$ and centre $\alpha$;
            // 1, 2, 3, and 4 for single, double, triple, and
            aromatic bonds, respectively;
            $(\sigma, \beta)$ append to $\mathbf{\Omega}$;
        end
        $\eta \leftarrow HASH(\mathbf{\Omega})$;
    end
    if $\eta$ not in F then
        $\eta$ append to F;
    end
end
```

In most instances, the similarity for fingerprints is defined by Jaccard (or Tanimoto) similarly (Equation (1.1)), mathematically. After definition of pairwise similarity, the initial applications of ECFPs are the area of high-throughput virtual screening by combining with Bayesian optimization(BO) and Gaussian processes(GPs). Except this, fingerprints can be used in kernel methods for structure-activity relationship analysis and manifold learning for data visualization.

1.2.2 Simulated and measured chemical features

A straightforward approach in the design of molecular descriptors is to utilize calculated or measured chemical properties, such as activities or electronic features, obtained from quantum mechanical (QM) calculations[95,104] using methods such as *ab-initio* techniques, density functional theory (DFT), or empirical simulations. It is worth noting that there is typically a trade-off between computational efficiency and accuracy. Highly accurate quantum methods tend to be computationally expensive compared to empirical potential-based simulation methods. The choice of method depends on the available computational resources and the size of the chemical system under investigation. The utilization of quantum calculations involves the initial determination of the relevant geometry of the system, as well as its total ground state energy. From this step, various physical properties of interest can be obtained using quantum mechanics and statistical mechanics principles. These calculations provide valuable insights into molecular systems and enable the determination of important electronic properties. Some examples of calculated electronic properties include atomic charges, molecular orbital energies, dipole moments, polarity indices, electronic charge distributions, and energy band gaps between exciton and ground states. In the case of organic photocatalysts, several factors contribute to their catalytic activity, including the efficiency of light absorption, thermodynamically driven force, and the ability of electronic charge transfer. These factors collectively influence the overall performance of the photocatalyst. Apart from electronic properties, the physical characteristics of porous materials play a crucial role in determining their adsorption properties. Features such as surface area, pore size distribution, and pore volume have a significant impact on the adsorption capacity and efficiency of porous materials. These physical characteristics directly affect the available surface area for adsorption, the accessibility of pores to target molecules, and the overall adsorption performance of the material.

Descriptors derived from electronic and geometric properties have the potential to offer a comprehensive characterization of materials, making them well-suited for machine learning models due to their continuity and the ability to incorporate prior chemical knowledge obtained through simulation approaches.[105] However, there are potential risks associated with the use of computational methods to design and apply descriptors pertaining to photocatalysts. An instance of such risk is encountered when attempting to predict the electronic and spectroscopic properties of dyes, which play a crucial role in highly active photocatalytic reactions like water splitting and photoredox synthesis.[106-107] For the most efficient metal-based dyes, the DFT and time-dependent density functional theory (TD-DFT) provide accurate results. These computational methods successfully reproduce the optical properties of diverse Ru(II) complexes,[108-109] as well as accurately predict their ground and excited-state oxidation potentials. This capability enables the predictions and screen novel synthetic approaches.[110] For organic molecules, the reliable calculation of excitation energies still represents an open issue, as a definite and effective computational approach has not yet been defined.[111] In the case of molecules with Long-range charge-transfer and spatially extended π systems, it is often observed that many TD-DFT calculations significantly underestimate the excitation states.[112-113] A benchmark study was conducted on five different dye-sensitized solar cells, evaluating the performance of various wave function methods, including B3LYP,[114] Coulomb-attenuating B3LYP(CAM-B3LYP),[115] PBE0,[116] MPW1K,[117] and MP2.[118] The results of this benchmark indicate that the CAM-B3LYP approach is the most accurate TD-DFT method among the ones tested in this study. When examining the photocatalytic properties of covalent organic framework (COF) photocatalysts, the conclusions regarding the accuracy of computational methods can differ. A study comparing the performance of the benchmark B3LYP method against two range-separated functionals (CAM-B3LYP and wB97X) in predicting the standard reduction potentials of half-reactions for free electrons/holes and excitons using molecular fragments representing certain COF photocatalysts revealed that all three functionals produced nearly identical results.[119] Similar agreement was also observed for conjugated polymers, where range-separated functionals such as CAM-B3LYP only provided marginal improvements over B3LYP in predicting the optical gap.[105] Gómez-Bombarelli *et al.* performed extensive benchmarking of their chosen TD-B3LYP level of theory against experiments, and against range-separated functionals and other hybrid functionals, concluding that B3LYP offers a cost-effective accuracy while performing on-par or better than the other functionals tested. Therefore, the choice of DFT calculations still relies on the specific research system, and there is no universal rule or standard approach.

Additionally, since simulated descriptors are derived from molecular structures, they can exhibit correlations with each other, leading to redundancy. This redundancy may impact the performance of machine learning models. Besides, chemical reactions involve multiple molecular species, multiple reaction pathways, and numerous intermediate steps, which is challenging to represent in a number of calculated descriptors. Nonetheless, further research is needed to enhance the speed and efficiency of feature generation methods. This includes exploring new approaches that can capture more intricate molecular properties while addressing the challenges of redundancy and representation of chemical reactions.

1.2.3 Atomic centred descriptors and machine learning force field

An alternative method of constructing QSPR involves the use of atomic constituents to simulate and predict atomic-scale properties, which is known as the ‘bottom-up’ approach.[39] This approach starts with the chemical structure to predict the energy, forces, or a specific molecular configuration, and then uses these predictions to search for chemical behaviours by building an accurate molecular dynamics model.[120] The atomic 3D Cartesian coordinates is a simple and unequivocal representation of chemicals in classic physical and quantum chemical calculations. However, it is not suitable for machine learning applications because the comparison between coordinates does not reflect the difference of two chemicals. For example, two different coordinates can represent same molecule because the list of coordinates has an arbitrarily assigned order, and two structures can be mapped by space symmetry operations such as rotation, reflection, or translation. A good atomic centred descriptor need to be invariant to represent the symmetric operation.[39] This mapping associate chemical structure with a point in designed feature space, which then used to represent a QSPR by machine learning models. Therefore, several commonly used descriptors were developed to transform the coordinates in a way that fulfils physically informed requirements: smoothness and symmetry with respect to isometrics, such as Coulomb matrices,[121] atom-centred symmetry functions (ACSFs),[34,122] and the smooth overlap of atomic positions (SOAP).[123] Most of these descriptors are used for modelling atomization energies and forces, as they are directly related to molecular geometries.

Since the introduction of symmetric functions, there have been numerous studies conducted on neural network potentials. One notable example is the Accurate Neural Network Engine for Molecular Energies (ANI),[124] which modifies the original symmetry function to create an atomic environment vector (AEV) and trains NNPs. Despite these, other atomic centred descriptors are developed by exploiting many mathematical tricks.

One popular example is the Smooth Overlap of Atomic Positions (SOAP) descriptor,[123] which encodes local regions of atomic geometries by combining spherical harmonics and radial Gaussian functions to describe the angular and radial environment around the central atom. SOAP descriptors have been shown to achieve high accuracy in predicting both atomic energies and advanced chemical electronic properties.[125] One major difference between SOAP and ACSFs is in the way the descriptors are constructed. SOAP descriptors are based on radial and angular distribution functions that describe the electron density of atoms in a material, while symmetry function descriptors are based on a set of pre-defined mathematical functions that quantify the local symmetry and coordination of atoms. SOAP method has been shown to be effective in computational chemistry, including fitting the Gaussian Approximation Potentials (GAPs) model,[120,123] and comparing the similarity of molecular structures. On the other hand, symmetry function descriptors are often more interpretable and intuitive, as they are based on explicit geometric features of the atomic environment.

1.3 Fundamental of machine learning

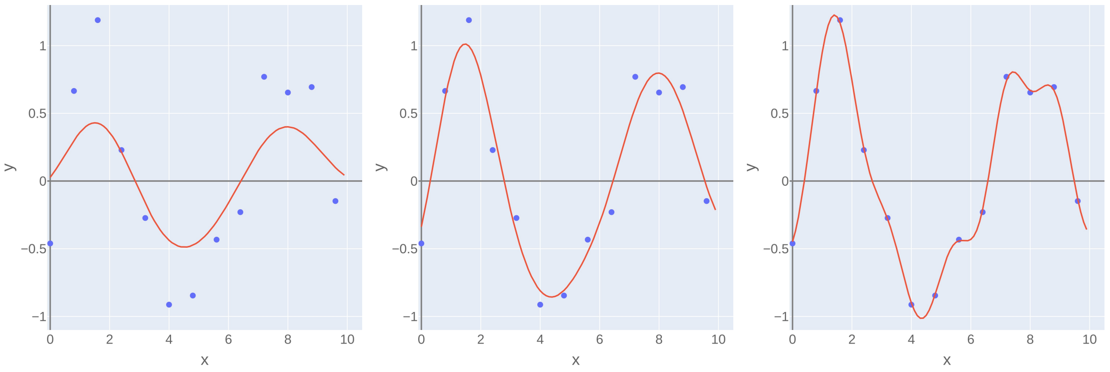
1.3.1 Model selection and evaluation

Machine learning is a series of modelling methods and statistic models by computing the cost function when given enough amount data with a property algorithm. A general process of building an ML model is training the model with known dataset and test it in a new dataset. Models are always classified into several branches depending on their approach: supervised learning, unsupervised learning, reinforcement learning and so on.[126] Supervised learning algorithms aim at mapping the input data to the given output data whereas unsupervised learning algorithms take the dataset only containing inputs without any targets. Furthermore, machine learning techniques can be categorized based on their specific purposes. Regression models are employed to predict a continuous-valued attribute associated with the input data. On the other hand, classification models are designed to identify the category to which an object belongs. They assign input data to specific predefined classes or categories. Additionally, clustering models, which fall under the category of unsupervised classification models, are utilized to automatically group similar input data into distinct categories or clusters. Unlike classification models, clustering models do not have predefined categories and instead seek to discover patterns or similarities in the data without explicit labels.

During the training process of machine learning models, numerous parameters need to be optimized. To accomplish this, a cost function, also known as a loss function or error function, is designed to evaluate the mapping between the input data and the output predictions.[126] Simultaneously, an optimization algorithm is introduced to minimize this cost function. In regression models, the cost function typically computes the mean squared error, while in classification models, cross-entropy is commonly used as the cost function during parameter estimation.[126]

In addition to the optimized parameters, machine learning models also have hyperparameters that control the learning process. To evaluate the performance of a trained model and optimize hyperparameters, the dataset is divided into three sets: the training set, the validation set, and the test set.[126] The training set is used to train the algorithm whereas the test set is used to evaluate the model performance on never-before-seen data. The necessity of the validation set is that many complex models need to search a good configuration in hyperparameter space, for example, the number of layers in deep learning and the regularization parameter in regression model. To avoid 'cheating', the best hyperparameter space is chosen by using as feedback signal the performance on the validation set, then the test set is implemented to evaluate machine learning model.[126]

Another technique known as cross-validation can be used to achieve the same objective without requiring a separate validation set. In cross-validation, the training set is randomly partitioned into several subsets, and an iterative process selects one of these subsets as the validation set while training the model using the remaining data. The performance evaluation is then based on the average values computed across the iterations. While cross-validation can be time-consuming, it has the advantage of utilizing the available data more efficiently, which is particularly beneficial when dealing with small datasets.



(a) Under fitting (b) Fitting (c) Over fitting

Figure 1.3 Under fitting, fitting and overfitting function(red line) training on the same noised dataset(blue points)

Other common issues in machine learning are over-fitting and under-fitting (Figure 1.3) due to variance reasons. For example, the undue complexity (Figure 1.3(c)) causes the learning model fits the training set perfectly but fits the test set poorly. Information leaks lead over-fitting quickly when the validation set is involved in the hyperparameter configuration step whereas inadequate training set and simple model (Figure 1.3(a)) cause under-fitting.

1.3.2 Linear regression and classification

The fundamental method of a machine learning, named as simple linear regression, is mapping the linear relationship between two one dimensional data. The hypotheses of this algorithm is that there exist a linear function to describe the given data by calculating the dataset $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ and yield property $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$. The simplest mathematical function of a one dimensional dataset is given by

$$f(x) = \mathbf{w}x + \epsilon \tag{1.2}$$

where parameters **w** and need to be optimized to give the best predication of the property $\mathcal{Y}$. The most commonly used function to evaluate the model called cost function is the sum of squared error function as below:

$$J(\mathbf{w}) = \sum_{i=1}^{n}(f(x_i) - \hat{y}_i)^2 \tag{1.3}$$

The gradient descent technique is utilized to optimize the model and minimize the cost function. As the dimensionality of the features, denoted as x_i, increases, the complexity of the parameters **w** also increases in order to effectively fit the given dataset. However, this can lead to overfitting issues. To address this concern, an additional hyperparameter, λ, known as the regularization parameter, is introduced.[126] This parameter acts as a penalty term to shrink the parameters and mitigate the problem of overfitting. By adjusting the value of λ, one can control the balance between the model's complexity and its ability to generalize to unseen data. The cost function is given by

$$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{n}\left\{(f(x_i) - \hat{y}_i)^2\right\} + \lambda \left|\left|\mathbf{w}\right|\right|^2 \tag{1.4}$$

where the penality term can be either $||\mathbf{w}||^2$ or $||\mathbf{w}||$. The former is called ridge regression or l_2 norm and the latter is called lasso regression or l_1 norm.[126] The minimum of the cost function of ridge regression also has a mathematical solution called the normal equation[126] which is given by

$$\mathbf{w} = (X^T X + \lambda I)^{-1} X^T \mathbf{y} \tag{1.5}$$

where X and $\mathbf{y}$ correspond to the matrix form of data set $\mathcal{X}$ and $\mathcal{Y}$. The parameter λ works as same as in the cost function of ridge regression, which controls the amount of shrinkage: the larger value of λ, the smaller parameter **w** and thus the complexity of the model is reduced to prevent the overfitting problem.

In addition to linear regression, another commonly used application in machine learning is classification. In classification tasks, the property value (represented as $\mathcal{Y}$) is discrete or categorical, typically in the form of integers, rather than continuous numbers as in regression problems. The logistic function is often employed in classification tasks to calculate the binary target variable y_i, which can take values of either 0 or 1. This represents the simplest form of classification problems.

$$f(x) = \frac{1}{1 + \exp(-\omega x)} \tag{1.6}$$

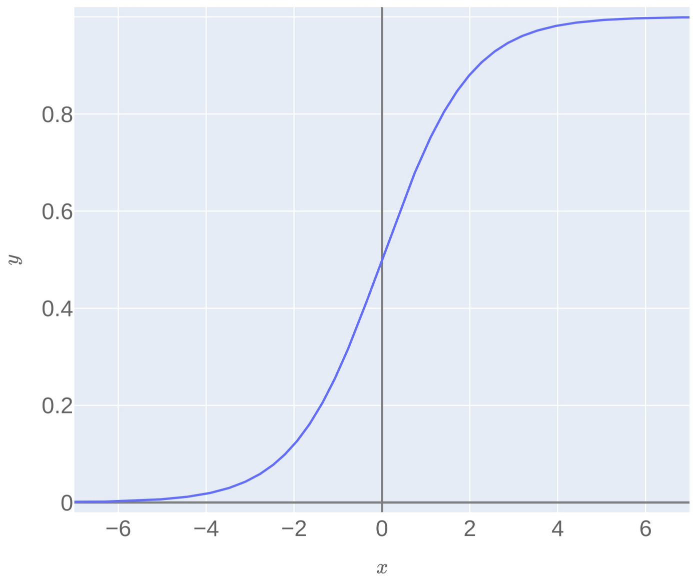
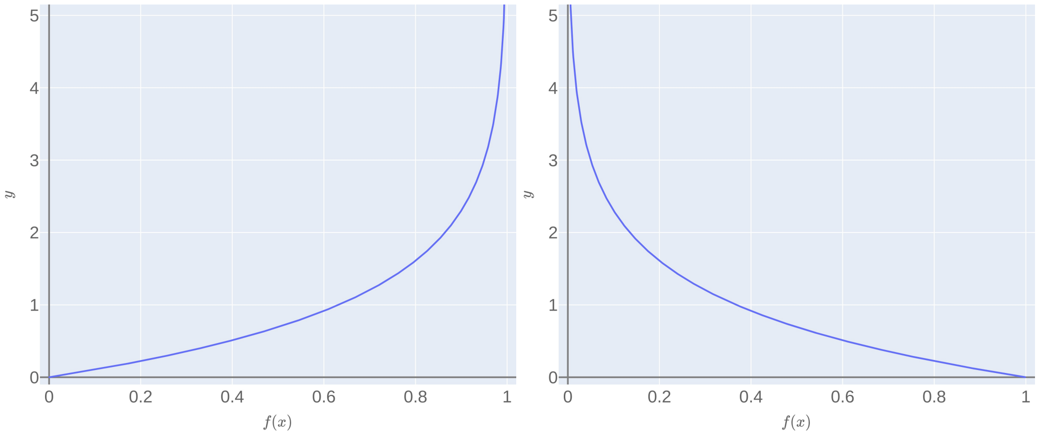


Figure 1.4 The standard logistic function where $\omega = 1$

that the predication of this function is the probability of the positive class $P(y_i = 1|x_i)$. For a binary classification, it is easy to conclude that the output follow the requirement $P(y_i = 1|x_i)+P(y_i = 0|x_i) = 1$ and if $f(x) > 0.5$ then the predication label is 1. The cost function of logistic function need to capture the error of the probability, which defined as the negative log-likelihood of logistic model:[126]

$$Cost_{log}(f(x), y) = -(y \log f(x) + (1 - y) \log(1 - f(x))) \tag{1.7}$$

where y is the true label in the binary classification so that it is always be 0 or 1. Figure 1.5 present the two section of the loos, and it is commonly called cross entropy in statistic.



(a) The log loss with true label equal to 0 (b) The log loss with true label equal to 1

Figure 1.5 Example graph of the two part of cross entropy loss function

Both squared error and cross entropy loos are continues so that they are able to apply the gradient descent as the optimization method to find the best model during training.

1.3.3 Kernel methods

The linear regression and logistic regression models are mainly defined by itself parameters. With increasing the complexity of data, more parameters need to be introduced to explain the model reasonably. Such methods that use the fixed number of parameters are called parametric methods whereas another technique can overtake this disadvantage without fixed parameters are called non-parametric methods. Another type of object such as binary set of bit fingerprint also does not have a fixed size vectors to best represent them. One of the approach to such problems is to assume a method to calculate the similarity between objects that does not require the representation as the feature vector format. The most commonly used technique here is kernel function which measures the similarity between any given training points.

Kernel methods refer a class of machine learning algorithms by utilizing the advantage of 'kernel trick' to deal with high dimensional dataset.[126] This function operates the data with a cheap evaluation to calculate the similarity in original space yield an $n \times n$ kernel matrix. For all x and x' in given space $\mathcal{D}$, the kernel function can be expressed as an inner product in another space $\mathcal{V}$, and it also can be written in the form of mapping $\varphi : \mathcal{D} \rightarrow \mathcal{V}$ as follows

$$\kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \tag{1.8}$$

It turns out that many algorithms can be kernelized in this way, if the inner products of the form $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is replaced by a kernel function. Applying the kernel trick to the fundamental ridge regression results the kernel ridge regression(KRR).[126]

$$\mathbf{w} = (K + \lambda I)^{-1} X^T \mathbf{y} = X^T (X X^T + \lambda I)^{-1} \mathbf{y} \tag{1.9}$$

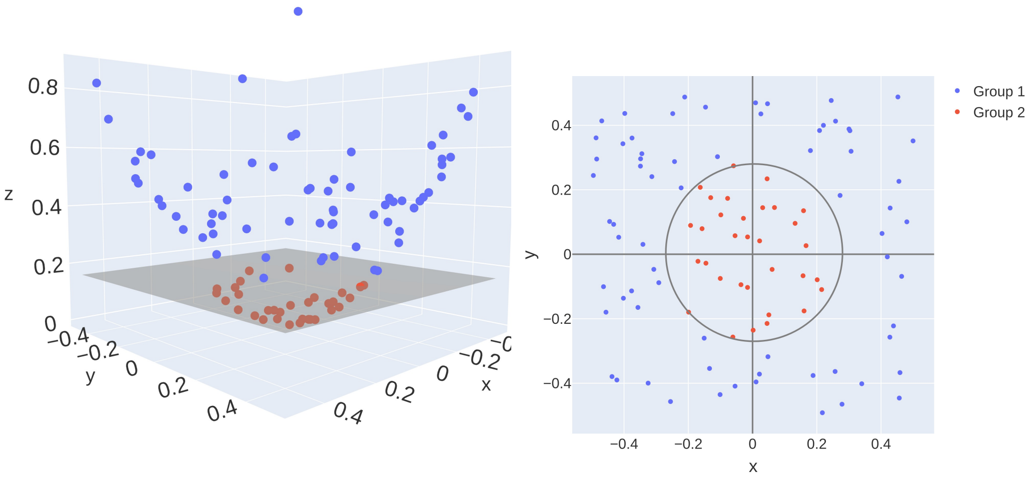
The solution of ridge regression is given by Equation 1.5.[126] The XX^T term can be replaced by matrix K which also called positive definite kernel after the matrix inversion(Equation 1.9).[126] Following dual variables are defined as follows

$$\alpha = (K + \lambda I)^{-1} \mathbf{y} \tag{1.10}$$

Then the solution of Equation 1.5 and the prediction of new points $\mathbf{x}$ with N training vectors can be rewritten as follows

$$\begin{aligned} \mathbf{w} &= X^T \alpha = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \\ \hat{f}(\mathbf{x}) &= \mathbf{w}^T \mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x} = \sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}, \mathbf{x}_i) \end{aligned} \tag{1.11}$$

The hypotheses of the 'kernel trick' is that there exist a higher denominational space (latent space) where a linear model can fit the given data point when project them from the original space to the latent space.[126] A visualization understand of this can be expressed using support vector machine(SVM) classification in a 2D space. After project the training points to the 3D latent space(Figure 1.6(a)), there is a linear surface to separate the group 1 and group 2. This boundary can be mapped back to the 2D feature space (Figure 1.6(b)) as a non-linear circle. Therefore, by applying this transformation, some linear model can fit the complex training dataset such as principal component analysis(PCA),



(a) The training points mapped to a 3D space (b) The points in the original 2D space

Figure 1.6 Example of SVM classification with kernel given by $\kappa(x, y) = xy + x^2y^2$. The linear boundary function in 3D space(the grey plane) is presented in (a), and the projected function in 2D space (the grey circle) is presented in (b).

GPs and KRR, SVM mentioned before.

There are many kernel functions used in machine learning. Deriving a kernel directly from the feature vector can yield the linear kernel($\phi(\mathbf{x}) = \mathbf{x}$),[126] defined by

$$\kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}') = \mathbf{x}^T \mathbf{x}' \tag{1.12}$$

The squared exponential kernel or Gaussian kernel[126] is defined by

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp - \frac{\left\|\mathbf{x} - \mathbf{x}'\right\|^2}{2\sigma^2} \tag{1.13}$$

It is also called a radial basis function (RBF) kernel since it is only a function of radial $\left\|\mathbf{x} - \mathbf{x}'\right\|$.[126] The parameter σ is knowns as the bandwidth and the term $-\frac{1}{2\sigma^2}$ often written as length scale l to simplify the format of Equation 1.13.[126] The measurement radial can be recognized as the Euclidean distance between two points in the original space. Another different distance measurement can also be implemented here with different type of features such as Jaccard distance for binary data.

1.3.4 Multilayer perception

Deep learning, as a subfield of machine learning, also process the input to targeted output mapping calculations, but the data are transformed via a deep sequence layers, referred to the artificial neural network (ANN). The deep in deep learning does not refer to any deeper understanding, rather, it implies the complexity of learning model with many successive layers and nodes. The simplest neural network is constituted by one hidden layer with one node, one dimension of input, and an output layer, which is similar to the one dimension linear regression (Equation 1.2). The computation for output contains can be decomposed into two steps

$$z = wx + b \tag{1.14}$$

$$a = \delta(z) \tag{1.15}$$

where δ is the activation function and a equal to the output $\hat{y}$. When this network is extended to multiple hidden layers and nodes with high dimension input features, all parameters are stacked up horizontally into a matrix or vector as follows

$$Z^{[l]} = \begin{bmatrix} W_{11}^{[l]} & W_{12}^{[l]} & \cdots & W_{1n}^{[l]} \\ W_{21}^{[l]} & W_{22}^{[l]} & \cdots & W_{2n}^{[l]} \\ \vdots & \vdots & \ddots & \vdots \\ W_{n_x1}^{[l]} & W_{n_x2}^{[l]} & \cdots & W_{n_xn}^{[l]} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_x} \end{bmatrix} + \begin{bmatrix} b_1^{[l]} \\ b_2^{[l]} \\ \vdots \\ b_n^{[l]} \end{bmatrix} \tag{1.16}$$

$$Z^{[l]} = W^{[l]^T} X + B^{[l]} \tag{1.17}$$

$$A^{[l]} = \delta(Z^{[l]}) \tag{1.18}$$

where $W_{n_xn}^{[l]}$ represent the distribution parameter w which multiplied by the n_xth dimension of input x in lth hidden layer and nth neuron. The last term $B^{[l]}$ in Equation 1.17

refers to the bias matrix in Equation 1.16. The second step of ANN Equation 1.18 is same as Equation 1.15 using an activation function δ such as rectified linear unit (ReLU), Sigmoid function and Tanh function. The output matrix of hidden layer l ($A^{[l]}$) is propagated forward to next layer as their input matrix X that organized layers stacked on top of each other. Besides, these forward propagation steps can also be visualized as a graph network

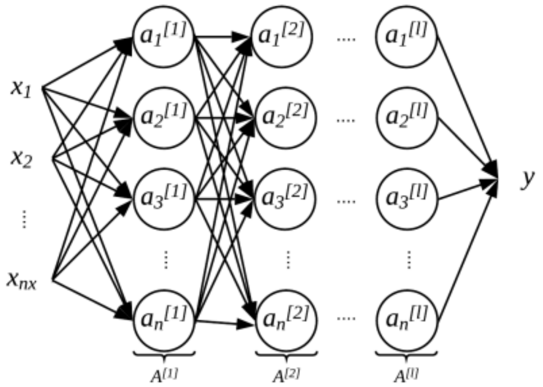


Figure 1.7 The structure of a multilayer perceptron

where arrows depicting the dependences between nodes and circular node represents the two mentioned computation steps. The ANN is also called multilayer perception (MLP) when the nodes in each layer are fully connected by the propagation function (Figure 1.7).

The training process for deep learning involves optimizing the weight parameters of the network so that it can accurately predict the output data. This optimization process is performed by minimizing a loss function, which measures the difference between the predicted output and the true output. To find the minimum position for the loss function, gradient descent is the basic method for the optimization. However, in neural networks, there are so many weight parameters that calculating the derivatives is a time-consuming step in the training loop. The algorithm called backpropagation[127] distributes the loss term from the final value and backward through layers by applying the chain rule.

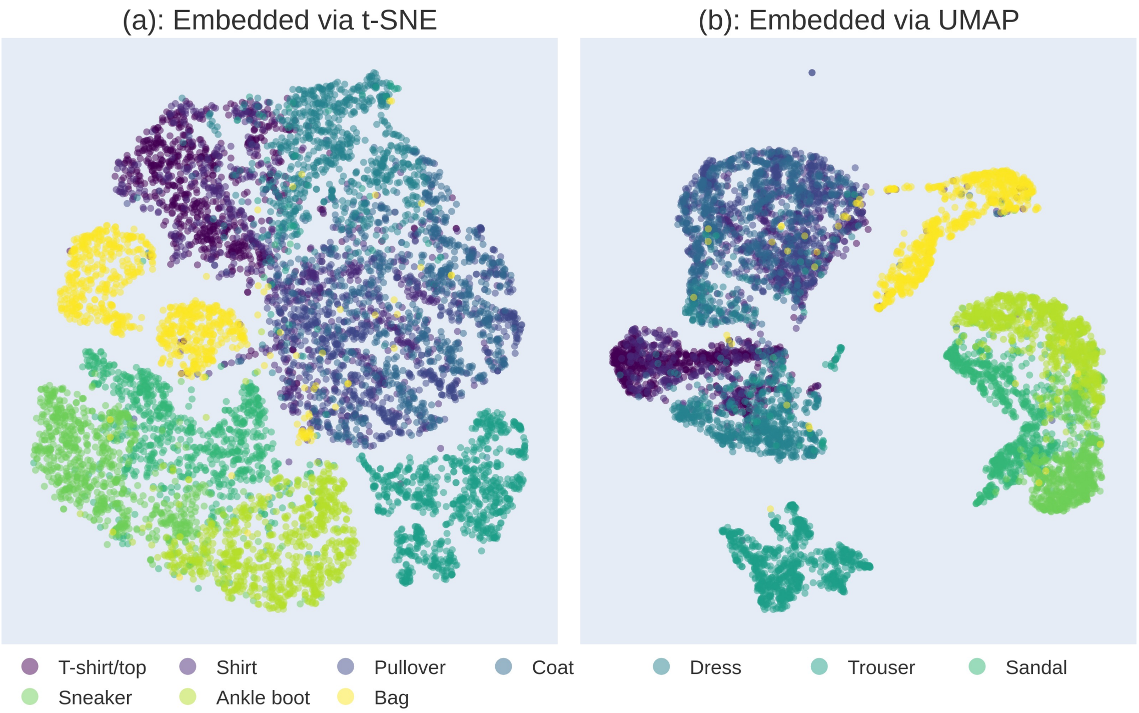
The application of deep learning algorithms often requires a large amount of data to be used as the training set. In many cases, this data is obtained by searching through published scientific papers, dataset, and conducting quantum calculations. The combination of different DL models with chemical targets has the potential to enable a wide range of applications in materials science. For example, generative model, e.g. the variational auto encoder is applied to inverse design;[69] sequence deep neural network is used to predicting the organic synthesis route by introducing string representation as ‘chemical language’ and using published chemical reaction datasets;[68] graph neural network is also a reasonable for investigating chemical structure;[94] deep neural network potential trained with millions calculated organic molecular structure to simulate the interatomic interaction.[124]

1.4 Visualization of high dimensional data

The field of chemical sciences is facing a significant challenge as it generates a large volume of complex data that includes chemical structures and their related properties. For instance, the ChEMBL[128] database contains more than 1 million bioactive molecules from literature. Other chemical database, GDB serials,[87,129] generated from theoretical calculation also contain up to 166 billion organic molecules. Visualization of high dimensional data can be a major requirement in data science and high throughput chemical discovery.[130] It allows chemists to gauge the distribution and relative relations of compounds in chemical space that can help with manual inspection of structure activity or property relationships.

Thus, the dimensionality reduction becomes a popular application of machine learning algorithm, which project data points from the original high dimensional space into a human-readable two or three-dimensional space and also keep the similarity information as much as possible.[131] To reach human intuition, data set are normally displayed in a 2D or 3D scatter plot, but it is difficult to reserve the original features exactly after the representation. Various techniques involved for this problem proposed many algorithms depending on the preserved features. Classic dimensionality reduction method such as PCA was invented nearly a century ago as a linear technique to decompose the original features into a new set of orthogonal components which keep the maximum variance in each dimension.[132] Another linear transform algorithm, multidimensional scaling (MDS),[133] was proposed that focus on keeping the dissimilar data points apart in the low dimensional representations.[131] However, it is usually more useful to keep points close together if they are very similar in the original space, which linear mapping algorithms are not possible to achieve.[134]

Many non-linear dimensionality reduction methods have been proposed to preserve the local structure of the original space based on the hypothesis that the dimensionality of the dataset is only artificially high. T-distributed stochastic neighbour embedding (t-SNE) probability is the most popular manifold learning technique in the past decades since it was proposed by Maaten and Hinton in 2008.[134] It is a stochastic neighbour embedding and very sensitive to local structure to extract local clustered groups that is beneficial to visually distinguish a dataset, but t-SNE also have some disadvantages such as computationally expensive and absence of global structure. Therefore, uniform manifold approximation and projection (UMAP) algorithm was proposed by McInnes *et al.* recently as a new technique to overcome these defects.[135] The ideas of UMAP came from the t-SNE process and topological data analysis to construct a high dimensional graph representation so that the global structure is preserved before the optimization of low dimensional embedding. Here is a comparison of these two algorithms on the Fashion MNIST[136] which is a dataset of clothing images consisting of 70k examples in 784 dimensional space with label from 10 categories (Fig 1.8). Both algorithms successfully exhibit local clustering features and group similar items (gathered different footwear) together on the 2D figure. UMAP preserves more global structure such as clearly separates footwear (light blue and yellow) away from coat (cyan) and shirt (dark purple) whereas t-SNE can only distinguish different cluster without the distance information between each group. It is also worth


(a): Embedded via t-SNE

(b): Embedded via UMAP

Figure 1.8 Dimensionality reduction applied to the Fashion MNIST dataset. Coloured by the category of clothing items.

noting that UMAP cost 21 seconds in comparison to 200 seconds with t-SNE. Both of these techniques are stochastic that multiple restarts can yield different embeddings. Future details and applications of these dimensionality reduction techniques are discussed in chapter 2. However, these techniques still has time-consuming limitation with large datasets (10^7), so that other implementation TMAP[137] was proposed to visualization very large chemical database. The TMAP algorithm can visualize the relation between clusters because the tree format includes edges and branches. However, the explanation of the tree graph is still challenge for large chemical datasets.

1.5 Active learning and Bayes optimization

1.5.1 Bayesian optimization

In real word optimization problem, such as finding the best hyperparameters in training of ML models and searching the reaction condition in chemistry experiment, the fundamental challenge is the expensive measuring step and large exploration of reaction and parameters space for many high throughput techniques.[138] In a typical laboratory, chemists can only estimate a small subset of experiments results during the optimization process due to the time and materials limitation. With the help of modern high throughput experimentation,[139] the capability of experiments design is extended to thousands dataset under limited configurations. The design of these experiments is mainly relying on scouring chemical literatures, the success of previous experimental experience, chemical intuition, and mechanical understanding of data. Future development of systematic approach design of experiments (DOE) identify the importance of parameters by sampling experimental conditions systematically.[140] DOE enable to explore the prior information gained from previous sampling and guide the evolution of next selection of experimental design. However, the sampling experiments increased exponentially with the number of searching dimensions makes DOE ineffective.[76] Such challenge is also existed in machine learning during the optimization of hyperparameters, which drives the empirical optimization to algorithm guided approach.

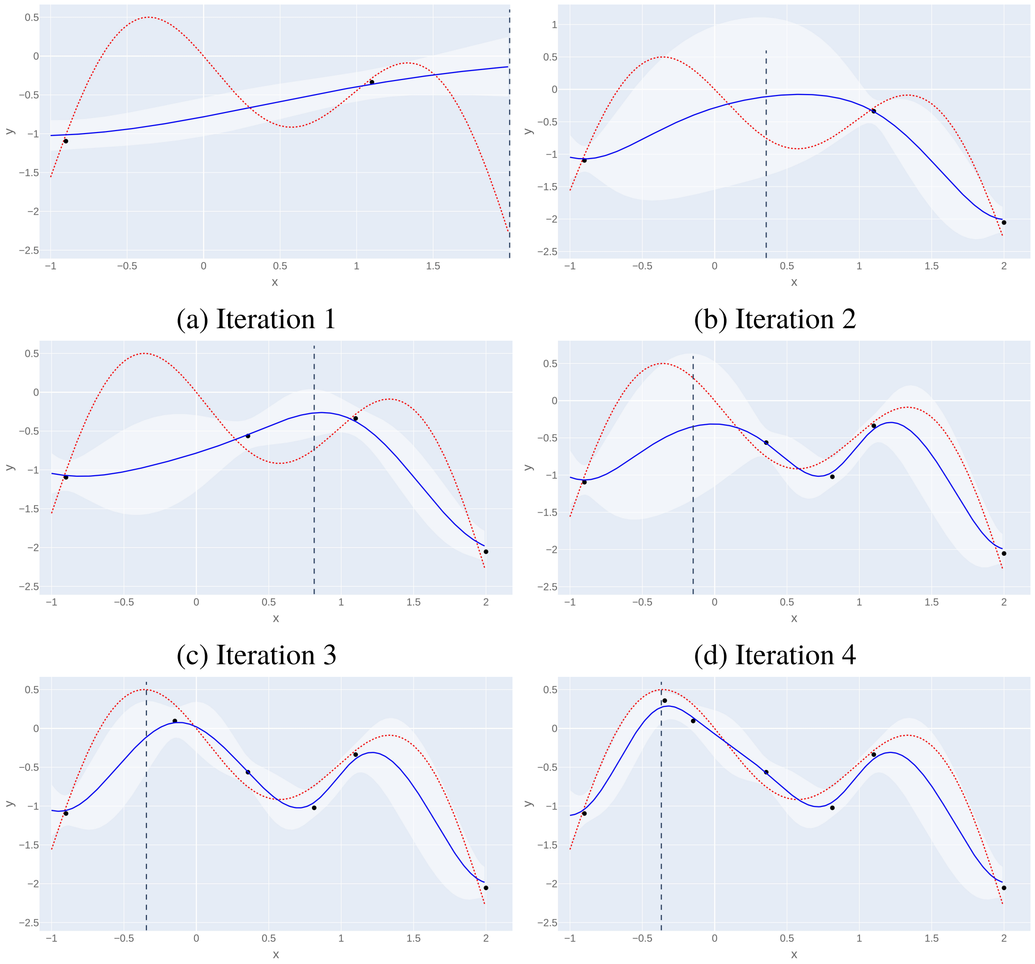
So-called active learning or close loop optimization attempt to find the global optimum in designed space in a minimum number of steps. The advantage of this strategy is that the algorithm can incorporate both prior and posterior knowledge during the experiment process to propose next sampling. Bayesian optimization (BO),[138] an uncertainty guided strategy, has shown excellent performance in the optimization of expensive objective functions. It is a sequence design for global optimization of black box functions that incorporates the prior function with samples draw from the fitted model to get a posterior that better than the prior approximation. The process uses a surrogate model to simulate the objective function and propose next sampling by acquisition function.

The most commonly used surrogate model is GPs because it is flexible and relative cheap to estimate the objective function.[141] The surrogate model gives both the mean and variance so that the acquisition function can balance the exploitation and exploration to propose the next sampling. Both drive to high acquisition function values, corresponding to high potential of high value of the objective function, and the sampling position can be determined by maximizing the acquisition function. There are many designed functions to achieve this purpose. Fore example, probability improvement (PI), expected improvement (EI), upper confidence boundary (UCB) and lower confidence boundary (LCB). Some of them are purely exploitation (PI), some introduces extra parameter to control the balance of mean $\mu(x)$ and variance $\delta(x)$ such as UCB and EI:[141]

$$\begin{aligned} f_{UCB}(x) &= \mu(x) + \beta\delta(x) \\ f_{EI}(x) &= (\mu(x) - f(x^{+}) - \xi)\Phi(Z) + \delta(x)\varphi(Z), if \sigma > 0 \end{aligned} \tag{1.19}$$

where $Z = (\mu(x) - f(x^{+}) - \xi)/\delta(x)$, Φ and φ are the cumulative density function and probability density function of the standard normal distribution. The β and ξ are the parameters to control the amount of exploration. The first term of both function are the exploitation term and the second term is the exploration term, receptively. After defined both surrogate model and acquisition function, the proposed sampling position is calculated by maximizing the acquisition function using gradient descent.

Figure 1.9 is a simple example to show how the algorithm works on a one dimensional objective function with two prior sampled points (Iteration 1). In the context of Bayesian optimization, the algorithm's behaviour can be observed in terms of its proposed positions or solutions over iterations. Typically, during the early steps (Iteration 1-4 in Figure 1.9), the algorithm tends to prioritize exploration by proposing positions with high variance. This means it explores different regions of the search space to gain a better understanding of the landscape and potential global maximum. As the optimization progresses to later iterations (Iteration 5-6), the algorithm shifts its focus towards exploitation. It starts to prioritize positions with high mean values, indicating a preference for regions that have shown promising results so far. By concentrating on these high mean positions, the algo-



(a) Iteration 1

(b) Iteration 2

(c) Iteration 3

(d) Iteration 4

(e) Iteration 5

(f) Iteration 6

Figure 1.9 Bayesian optimization for 6 iterations in graphs including the noise free objective function(in red), the tested point(black dots), the surrogate function(in blue), the 95% confidence interval of the mean(the white shaded region), and the proposed sampling location(the dash line). The GPs with squared exponential kernel is used here as the surrogate model. The EI with $\xi = 0.01$ is the acquisition function.

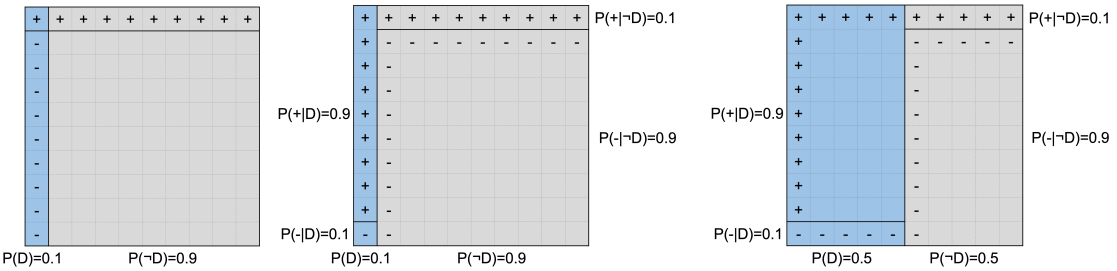
rithm aims to refine its search and converge towards the global maximum, leveraging the knowledge gained during the exploration phase. This combination of exploration and exploitation is an effective strategy in escaping local maxima and ensuring that the algorithm discovers the global maximum.

1.5.2 Bayes' theorem and Gaussian processes

Behind the Bayesian optimization is the art of design of GPs supported by Bayes' rule. The standard Bayes' theorem proposed the estimation of posterior by prior and likelihood.[141]

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}; p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{1.20}$$

The simplest understanding of Bayes' theorem is predicting the probability of an event after observation of related information, such as estimating the probability of a patient cached virus (event D) after receiving a positive test result (event $+$). The first hypothesis is both event D and event $+$ are independent(no true in real word), and the following two examples is more realistic that the test accuracy for patient $p(+|D)$ and the accuracy for health people $p(-|\neg D)$ are both equal to 90%. The graph explanation is shown below:



(a) Two event D and $+$ are independents (b) the prior $p(D)$ of example (b) is 0.1 (c) the prior $p(D)$ of example (b) is 0.5

Figure 1.10 Three geometric explanations of the Bayes' theorem. Events including positive test result($+$); negative test result($-$); and a patient cached virus(D).$p(+|D)$ is the probability of a patient given a positive test result as the likelihood. $p(D)$ is the probability of a patient in the population as the prior. Each cell of the table represent 1% of population and 100% in total, where the blue box represented the patient and the grey box represent the health people. The $+$ denote the positive test result and $-$ denote negative test result. $p(+|\neg D)$ represents the probability of a positive result given that a person is not diseased, which is commonly referred to as the false positive rate.

The posterior $p(D|+)$ is easy to calculated as 0.1 for example (a)(Figure 1.10(a)), because the likelihood is equal to the marginal likelihood when $p(D)$ and $p(+)$ are two independents event. However, the test method should identify patients and health people so that the $p(D)$ and $p(+)$ are related. By apply Bayesian inference, the marginal likelihood is transformed to a different equation as the real probability of the positive result in population ($p(+)$) is difficult to measure, but the likelihood $p(+|D)$ and the accuracy for health people $p(-|\neg D)$ are easy to estimate since the test method is solid. Therefore, the probability of a people catch virus when he got a positive result $p(D|+)$ is mainly determined by the prior and the Bayes' theorem equation is written as follows:

$$p(D|+) = \frac{p(+|D)p(D)}{p(+)} = \frac{p(+|D)p(D)}{p(D)p(+|D) + p(\neg D)p(+|\neg D)} \tag{1.21}$$

As shown in Figure 1.10(b) and Figure 1.10(c) with different prior but same test accuracy, the posterior can be calculated using the Equation 1.21. The probability of people caught the virus when received a positive test $p(D|+)$ is 83% in example c, but it reduced to only 50% in example b if the prior $p(D)$ changed from 0.5 to 0.1. The influence of the prior is remarkable for example b and c even using the same test method with same accuracy.

The Gaussian processes (GPs) is a random process where the distribution of fitted regression models is the posterior function and prior function is the joint distribution of observed data (training data).[141] The Bayes' theorem described the probability of a posterior event $p(\omega|\boldsymbol{y}, X)$ based on the prior probability $p(\omega)$ and the likelihood $p(\boldsymbol{y}|X, \omega)$.[141]

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood}; p(\omega|\boldsymbol{y}, X) = \frac{p(\boldsymbol{y}|X, \omega)p(\omega)}{p(\boldsymbol{y}|X)} \tag{1.22}$$

where the ω is parameters or weight of the GPs model. The dataset $\mathcal{D}$ has n observations, $\mathcal{D} = \{(x_i, y_i)|i = 1, 2, \ldots, n\}$, where x donates an input vector of dimension D and y is the measured target value. X is the aggregated vector of the input dataset in the $D \times n$ matrix and $\boldsymbol{y}$ donates the collection of targets. A GPs assume that the prior is jointly Gaussian, with mean and variance function.[141] The prior joint distribution is written as follows:

$$p(f|X) = \mathcal{N}(f|\mu, K) \tag{1.23}$$

where the probability p is $p(f(x_1), f(x_2), \dots f(x_n))$ with given points and $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. μ is the mean function, and it is common to use $\mu = 0$ when there is no prior information. The function κ is a positive definite kernel function or covariance function in GPs as in section 1.3.3 discussed before. Therefore, a GPs is a distribution of functions whose shape is defined by matrix K_{ij}. If object x_i and x_j are considered to be similar by the kernel function, $f(x_i)$ and $f(x_j)$ are expected to be similar as well.

Consider a training dataset with noisy function $y = f(\mathbf{x}) + \epsilon$, where the noise $\varepsilon \sim \mathcal{N}(0, \sigma_y^2)$ is independent with each observation.[141] The model comes close the observed data because of noise. Suppose there is a training set X of size $N \times D$ and the test set X_* of size $N_* \times D$, where the observation is $\mathbf{f}$ and predication outputs is $\mathbf{f}_*$. By definition of the GP, the join distribution of $\mathbf{f}$ and $\mathbf{f}_*$ is again a Gaussian with the following form

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right) \tag{1.24}$$

where $\mathbf{K}_* = \kappa(\mathbf{x}, \mathbf{x}_*)$ and $\mathbf{K}_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$.[141] The posterior of Gaussian with noise is given by

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_*|\mu_*, \Sigma_*) \\ \mu_* &= \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y} \\ \Sigma_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_* \end{aligned} \tag{1.25}$$

where $\mathbf{K}_y = \mathbf{K} + \sigma_y^2 I$.[141] The μ_* and Σ_* in equation 1.25 is the main function to calculate the mean prediction and variance for each predicted points.[141] The parameter, σ, controls the noise level including the range of variance, and fitting between over fitting and under fitting.

Since kernel parameters are introduced in Equation 1.25 and 1.13, the estimation of these values can be quite slow if using an exhaustive search such as grid search. (This is commonly used in kernel SVM and KRR.) The advantage of GPs approach is utilizing the Bayes rule that allow continuous optimization methods to search parameters in a faster way. In this case, the optimal parameter values can be estimated by maximizing the log marginal likelihood.[141]

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^T\mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_y| - \frac{N}{2}\log(2\pi) \tag{1.26}$$

The first term is a data fit term, the second term is the complexity term which is similar with the regularization term in linear regression, and the third term is a constant. This function represents the probability of generating observation from a prior so that the optimization of parameters of kernel in GPs is also the maximizing the probability of observations.[141] Finally, the kernel parameters can be estimated by this equation and its derivative using a standard gradient-based optimizer. To avoid local minimal problem, multiple starting points are randomly selected to do the optimization.

Algorithm 1.3 Gaussian process regression[141]

```
Input: $\mathbf{X}$, training set
       $\mathbf{y}$, targets
       $\kappa$, kernel function
       $\mathbf{x}_*$, test-set
Data: $n$, the number of restarts optimization
      $\theta$, the kernel parameters
      $\mathbb{R}$, the space of each parameter
      $\sigma_y$, the noise level
Output: $\mu_*$, predicted targets
        $\Sigma_*$, variance
for $i = 0$ to $n$ do
    $\hat{\theta} \in \mathbb{R}$ // Random select the start point of parameters;
    $\mathbf{K} = \kappa(\mathbf{x}, \mathbf{x})$;
    $\mathbf{K}_y = \mathbf{K} + \sigma_y^2 I$;
    $\log p(\mathbf{y}|\mathbf{X}) \leftarrow$using Equation 1.26;
    $\theta_i \leftarrow \min(-\log p(\mathbf{y}|\mathbf{X}), \hat{\theta})$ // Gradient-based optimization;
end
$\theta_* \leftarrow \min(-\log p(\mathbf{y}|\mathbf{X}), \theta_i)$;
$\mathbf{K}_* = \kappa(\mathbf{x}, \mathbf{x}_*)$;
$\mathbf{K}_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$;
$\mu_*$ and $\Sigma_* \leftarrow$using Equation 1.25
```

To apply GPs algorithm in data analysis, various kernel function are available including previous mentioned RBF kernel (Equation 1.13), linear kernel (Equation 1.12), Matérn kernel, and polynomial kernel.[141] Future more, the distance measurement also has several forms of estimation that allow the implementation of cheminformatics features into numerical equation. By integrating the advantage of kernel function, Gaussian processes, and Bayesian optimization, the active learning process is attainable to searching both exploitation and exploration in designed chemical space.[24]

1.6 The object of thesis

The subsequent sections of this PhD project delve into various aspects of chemical research, including chemical data visualization, data-driven methodologies for catalyst discovery, and the utilization of machine learning techniques. The subsequent sections of this PhD project delve into various aspects of chemical research, including chemical data visualization, data-driven methodologies for catalyst discovery, and the utilization of machine learning force fields. By harnessing the power of data-driven methodologies, advanced computational techniques, and interdisciplinary collaboration, researchers can unlock new opportunities for innovation and accelerate the development of novel chemicals and materials.

CHAPTER 2 DIGITAL NAVIGATION OF HIGH-DIMENSIONAL CHEMICAL DATASETS

2.1 Introduction

Thanks to the rapid development of algorithms and computing hardware, many cheminformatics methods and machine learning algorithms for data visualization are well established in both chemical and computing science community. Using modern algorithms and simulation methods, large chemical dataset, containing tens of thousands of molecules, are routinely calculated at quantum mechanical levels on supercomputers. Computational methodologies and toolkit are of paramount importance in facilitating the design and discovery of novel materials with unparalleled performance, guiding and providing inspirations for laboratory efforts.[20,142] The high throughput measure equipments also increased the availability of experimental data. This leads the formation of large, diverse machine learning community in chemistry which are developing and utilizing the dataset (by experiments or calculation) to explore chemical knowledge. Here, cheminformatics methods, dimensionality reduction techniques, and visualization toolkit are presented to show how a typical high throughput virtual screening(HTVS) and data visualization workflow is performed and designed for chemists.

2.2 Dimensionality reduction algorithms

A general request for data driven research topic is how to make a human interpretable format of a complex, high dimensional dataset.[143-144] A major question concerning the application of HTVS is how to find an efficient way to visualize the generated datasets. In this case, dimensionality reduction is an important section in data science, being a fundamental technique in both visualization and pre-processing for machine learning. Most dimensionality reduction algorithms fall into two categories: those that favour the preservation of the distance in original space (principal component analysis and multidimensional scaling) and those that prefer to preserve local over global topology of the data; for example, t-distributed stochastic neighbour embedding (t-SNE).[134] The t-SNE algorithm has been a popular choice for visualization of high-dimensional data. However, it is difficult for t-SNE to preserve the global topology of the data as it uses Gaussian joint probabilities to represent the affinities in the original space and Student's t-distributions in the embedded space. The t-distribution, like the Gaussian distribution, is bell-shaped and symmetric, but it has heavier tails, meaning that it tends to produce values that fall far from its mean. Therefore, to minimize the cross entropy between the two distance matrices in their spaces, the t-SNE algorithm 'shrinks' and 'rescales sequence' of the Gaussian distribution matrix in the original space to fit the t-distribution matrix. The algorithm also needs to arbitrarily add the missing information about long distances to the embedded space, due to the Gaussian distributions not preserving such information from the original space. Recently, a novel manifold learning algorithm, Uniform Manifold Approximation and Projection (UMAP),[135] was devised and has been shown to often outperform t-SNE in both preserving global information and computational costs of calculations. UMAP is the same as t-SNE, in terms of manifold learning, but uses topology theories to construct the cluster group instead of Gaussian functions. Then, UMAP minimizes the cross-entropy between the topological representation and the layout of representation in the low dimensional space to improve the retention of distance between clusters.

The t-SNE algorithm[134] is modified from stochastic neighbour embedding (SNE) which starts by encoding the distance of neighbours in the original high dimensional space into a conditional probability representing similarities and converts it to a Student t-distribution in the output low dimensional space. The similarity of data point x_j to x_i is the conditional probability $p_{j|i}$ if point x_j is considered as a neighbour of x_j. Mathematically, the conditional probability $p_{j|i}$ is given by

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)} \tag{2.1}$$

where σ_i is the variance of the Gaussian centred on point x_i. The property of Gaussian distribution gives high $p_{j|i}$ if the distance of neighbour $\left\|x_i - x_j\right\|^2$ is small, whereas $p_{j|i}$ will be almost infinitesimal for widely separated data points. The parameters σ_i in Equation 2.1[134] determine shape of the Gaussian distribution as a key control of the SNE techniques. Thus, hyperparameter called perplexity is introduced to evaluate the value of σ_i for different point x_i and control how many and smooth of neighbours is counted in SNE. For the low dimensional mapped points y_j and y_i from points x_j and x_i in the high dimensional space, the similarity of y_j and y_i is also computed by

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right)} \tag{2.2}$$

as the conditional probability $q_{j|i}$. If this mapping is precisely described the similarity between the points in high dimensional space and low dimensional space, the conditional probability $p_{j|i}$ and $q_{j|i}$ will be equal. The aim of the dimensionality reduction algorithm SNE is finding a representation to minimize the difference between $p_{j|i}$ and $q_{j|i}$.[134] This difference can be calculated by Kullback-Leibler (KL) diverge as the cost function which describes the statistic similarity between two probability distributions.

There is a limitation using KL diverge on the gradient descent optimization because of the complexity and non-symmetry of the conditional probability $p_{j|i}$ and $q_{j|i}$.[134] Therefore, a symmetric joint probability $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$ is used here to reduce the complexity of the gradient calculation of the KL diverge. The pairwise similarities of joint probability p_{ji} and q_{ji} have similar mathematical form with conditional probability (Equation 2.1 and

2.2)[134] so that the KL diverge $KL(P||Q)$ is

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.3}$$

and the gradient of symmetric SNE is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \tag{2.4}$$

The so-called 'crowding problem' is also introduced by using the distance to measure the similarity in the high dimensional space since the neighbour region around point i with moderate distance is much larger in high dimensional space than represented in low dimensional space. Here, a Student t-distribution with single degree of freedom (Equation 2.5)[134] is introduced as the joint probability in low dimensional space q_{ij} to solve this problem.

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}} \tag{2.5}$$

Overall, the t-SNE algorithm is proposed below:

Algorithm 2.1 t-Distributed Stochastic Neighbour Embedding[134]

```
**Input:** dataset $\chi = \{x_1, x_2, \dots, x_n\}$
cost function parameter: $Perp$ perplexity
optimization parameters: number of iterations $T$, learning rate $\eta$, momentum $\alpha(t)$
**Output:** $\mathcal{Y}^T = \{y_1, y_2, \dots, y_n\}$
**begin**
    calculate $p_{j|i} \leftarrow$ with $Prep$ using Equation (2.1);
    $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$;
    sample initial $\mathcal{Y}^0 = \{y_1, y_2, \dots, y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$;
    **for** $t \leftarrow 1$ **to** $T$ **do**
        calculate $q_{ij}$ using Equation (2.5);
        calculate the gradient $\frac{\delta C}{\delta \mathcal{Y}}$; Update $\mathcal{Y}^t$ by counting the gradient with a momentum
          term $\mathcal{Y}^t = \mathcal{Y}^{t-1} + \eta\frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)(\mathcal{Y}^{t-1} - \mathcal{Y}^{t-2})$;
    **end**
**end**
```

As mentioned in Chapter 1, the UMAP[135] algorithm provides more advantages compared with t-SNE. The mathematical process of UMAP[135] is similar with t-SNE, but topological analysis technique is introduced to evaluate the similarity of neighbours in high dimensional space. The first step of UMAP can be considered as constructing a K-neighbour graph using the nearest neighbour algorithm.[135] A hyperparameter k is introduced here to set the k nearest neighbours as a group for each point x_i.

$$\rho_i = \min\left\{ d(x_i, x_{i_j}) | 1 \le j \le k, d(x_i, x_{i_j}) > 0 \right\} \tag{2.6}$$

where the metric d is the distance measure for given dataset $X = \{x_1, x_2, \ldots, x_N\}$. The selection of ρ_i ensures that every x_i connects at least one neighbour with an edge. The weighted graph $G = (X, E, \omega)$ can be defined with edges $E = \{(x_i, x_{i_j}) | 1 \le j \le k, 1 \le i \le N\}$ and weight function ω by Gaussian function.[135]

$$\omega((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i} \right) \tag{2.7}$$

where the σ_i corresponds to a smoothed factor. The weight of edge can be considered as the probability of given edge exists. A similar approach is given in the low dimensional space to construct the representation by measuring cross-entropy. With introducing the graph theory to describe the similarity, the UMAP is capable to separate clusters as tightly packed and try to reach maximal separation compared with t-SNE.[135]

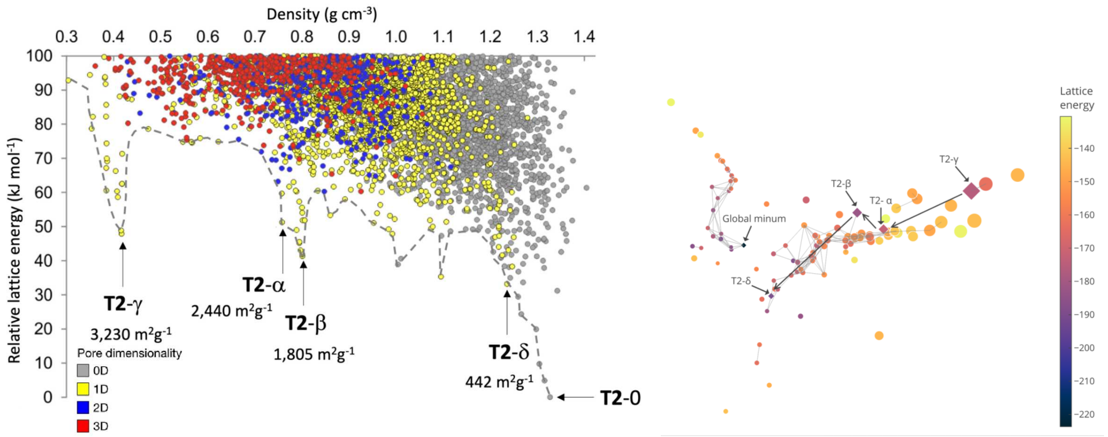
2.3 Plotting chemical structure activities relationships

2.3.1 Visualization of crystal structure dataset

Large-scale computational screening has become an indispensable tool for functional material discovery. It, however, remains a challenge to adequately interrogate the large amount of data generated by a screening study. For example, Pulido *et al.* have demonstrated a step-change strategy for how new, functional molecular materials can be discovered: by carrying out a priori prediction of both the crystal structure and its functional properties: Energy-structure-function (ESF) maps are created to aid researchers, without computational expertise, in realizing several remarkable porous materials promising for different possible applications. The *in-silico* crystal structure prediction (CSP)[41-42] is a class of methods to determine the stable crystalline arrangements that are available to a molecule. Here, the process of generating a stability-ranked list of crystal structures involves several steps. Initially, the geometries of all molecules are optimized using the B3LYP/6-311G(d,p) level of theory. These optimized molecular geometries are kept fixed during both crystal structure generation and lattice energy minimization. A quasi-random sampling procedure is employed to define the chemical system, which includes specifying information about the unit cell, molecular positions, orientations, and lattice parameters within each space group. The application of space-group symmetry allows for the generation of trial crystal structures, and a geometric test is performed to check for any overlap between molecules. Subsequently, lattice energy calculations are conducted using an anisotropic atom-atom potential. To ensure the uniqueness of the generated structures, structural relaxation techniques are applied to remove duplicate structures that may have been generated during the process. This helps to refine the list of distinct crystal structures. CSP can generate a list of predicted crystal structures ranked by stability of a given material without perform the time/labour expensive experiments.

In the published dataset,[145] 5679 crystal structures generated by CSP and are presented on a 2D diagram(Figure 2.1(a)) using density against relative lattice energy (build block T2, Figure 2.2(d)), with each point corresponding to a computed crystal structure. Thus, chemists need efficient design to overview the big data (thousands of crystal structures in ESF maps), narrow down the possibilities, and identify promising candidates for further investigation. Except this, it is also a challenge of interpretation and visualization with large-scale of generated CSP dataset.

While projecting an ESF map onto individual dimensions is a useful way of exploring data, it can be laborious when many structural and functional properties are associated with 1000s to 10,000s of structures typically on a single ESF map, even with the help of the interactive ESF Explorer(Chapter 2.3). It is therefore desirable to devise a simple and general approach to represent the high-dimensional data of ESF maps, allowing researchers to systematically inspect 'landmark' structures on the map, be they either energetically favourable or functionally interesting structures. With the implementation of the published CSP data (Figure 2.1(a)), machine learning algorithms can help to identify local representing structures and reducing the screening time. The screening requirement of such structure space is reduced from viewing 5679 structures to less than 100 structures. Such 2D diagrams, with additional information easily conveyed by colour coding and symbol size, are a powerful tool to allow for direct visualization of high-dimensional data and to help with deciphering the multivariate structure-property relationships.

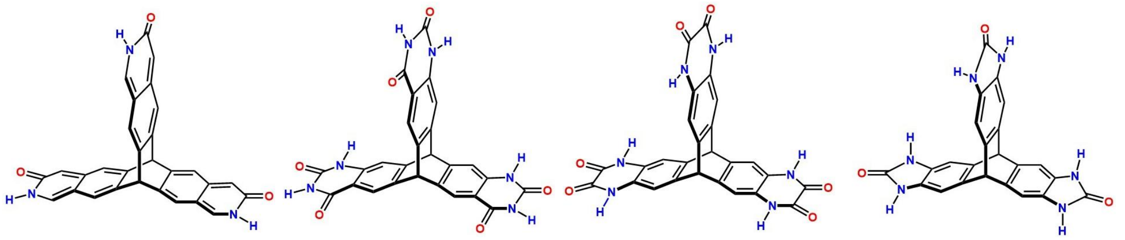


(a) Crystal structure prediction (CSP) energy-density plots for T2

(b) 2D embeddings of the porosity space of T2

Figure 2.1 Transforming the CSP landscape from energy-density space to algorithm embedded space. The symbols in (a) are colour coded by the dimensionality of the pore channels, assessed using a CH4 probe radius, 1.7 Å. The arrow in (b) represents the observed polymorph transformations from T2-γ to T2-δ.

After investigating the published T2 CSP dataset, a series of awkwardly shaped molecules with different hydrogen-bonding functionalities (Figure 2.2) were selected to analysis their CSP landscape, with the input of collaborators.[22] To influence crystal packing, the molecular cores were functionalized by different hydrogen-bonding moieties.



(a) TH2 (b) TH4 (c) TH5 (d) T2

Figure 2.2 The four candidate building blocks for porous solids and invested by CSP

To identify 'landmark' structures of these CSP data, each of the crystal structures was encoded by a number of pore descriptors as applied on T2 data. In the end, the UMAP algorithm was implemented to project the 'landmark' structures on a 2D representation and coloured by the number of pore dimensionality (Figure 2.3). In addition to the existing colour axis, it is possible to change it to other properties, such as relative lattice energies. This modification would enable researchers to identify the position of the global minimum structure on the 2D map. Implementing such features in the subsequent online interactive application would allow for a more comprehensive analysis and interpretation of the data.

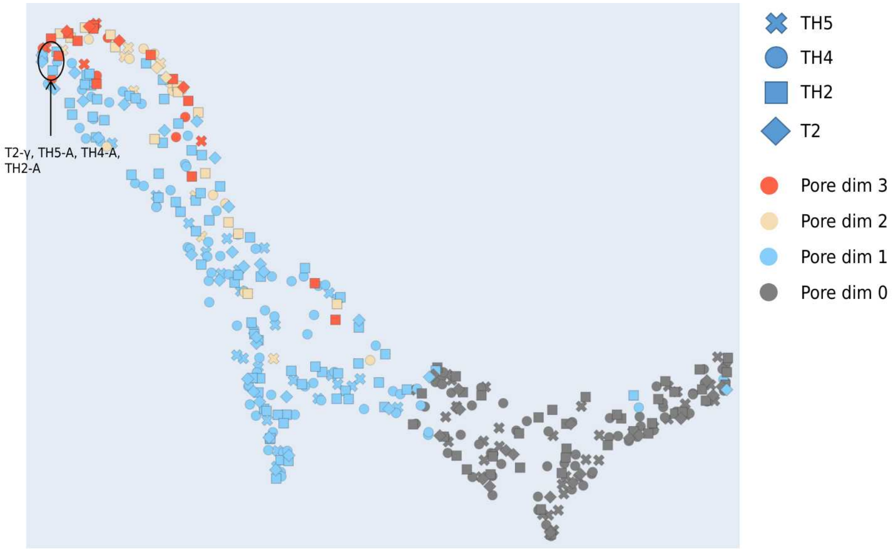


Figure 2.3 2D UMAP embeddings of the porosity spaces of TH2(diamond), TH4(circle), TH5(cross) and T2(diamond), colour coded by the dimensionality of the pore channels, assessed using a CH4 probe radius, 1.7 Å.

Since the selected porosity descriptors were agnostic to the molecular structure, 'landmark' structures can be compared across the different molecules in a single projection. The structures that have isostructural pore channels—for example, TH2-A, TH4-A, TH5-A and T2-γ all have hexagonal pore channels—are located in proximity on the 2D UMAP representation (left top corner on Figure 2.3). The 2D embedding approach shown here makes ESF maps machine-readable. To give one use case: it is often desirable to make comparisons between ESF maps for different molecules to assess whether two molecules will be functionally similar or not. This unified embedding process will be useful for comparing multiple CSP datasets and identifying functionally similar structures using the encoding representation. This might be used, for example, to select the most synthetically accessible molecule in a set of candidates that is likely to express the property of interest, such as a specific pore size. This approach automatically and systematically identifies a small set of 'landmark' structures (typically, 10s to 100s) from the whole CSP landscape (typically, 1000s to 10,000s structures), which allows chemists to focus more expensive calculations on a smaller set of structures.

2.3.2 Methods of clustering and identifying 'landmark' structures

The simulated porous features including pore diameters, surface areas and some variants of these in order to capture, to some extent, the heterogeneity of pore/channel sizes within a given map. To enhance the efficiency of visualization and screening, a machine learning guided workflow is built for working with the CSP dataset. This method involves the following steps: (I) featuring the CSP dataset; in this case engineered descriptors characterizing porous structures are used; (II) calculating the similarity matrix of the whole dataset, using pairwise Euclidean distances between the structures in the high-dimensional space (III) using clustering algorithms (here being affinity propagation[146]) to group similar structures, and select the lowest-energy structure in each cluster to represent the cluster; (IV) Applying MDS techniques to transform the high-dimensional feature space into a 2D diagram (Figure 2.1(b)), or using the UMAP algorithm to project the ‘landmark’ structures, in which the distances between data point respect their counterparts in the original, high-dimensional space.

2.3.3 Web interactive application

In order to improve the interoperability and accessibility of the ESF map embedding results and other cheminformatics figures, a web based shareable interactive analytic application was developed for direct visualization of most information obtained from the dataset. A general interactive application includes data operation as the back-end and HTML components as the front-end. Here, Dash is chosen as the low-code front-end framework to rapid develop apps in Python and deployed on Heroku server. This visualization tool allows users to scan any selected columns up to five dimensions simultaneously. For instance, three descriptors can be selected as the axes for visualization in the 3-dimensional (3D) diagram against other two descriptors as colour coding and symbol size. Furthermore, additional textual information is labelled for each data point, which is activated by hovering the cursor over the plot. Within some python frameworks, such as Pandas,[147] Plotly[148] and Dash,[149] such visualization apps are customise-able and can be rendered on the web by a few hundred lines of scripting.

By integrating the 2D UMAP embedding map and traditional 2D ESF maps, the web application allows researchers to interactive inspect identify ‘landmark’ structures on different structure property relationship figures by either changing the chart type or colour type. Figure 2.4 is a snapshot of one such web-based application (https://www.interactive-esf -maps.app/), showing an interactive 2D scatter plot with three dimensions of displayable information chosen by the user. The visualization is not only interactive but also highly user-friendly; for example, each plotting dimension is defined via a drop-down menu giving the user full control at ease, while slider bars allow the user to choose the value range

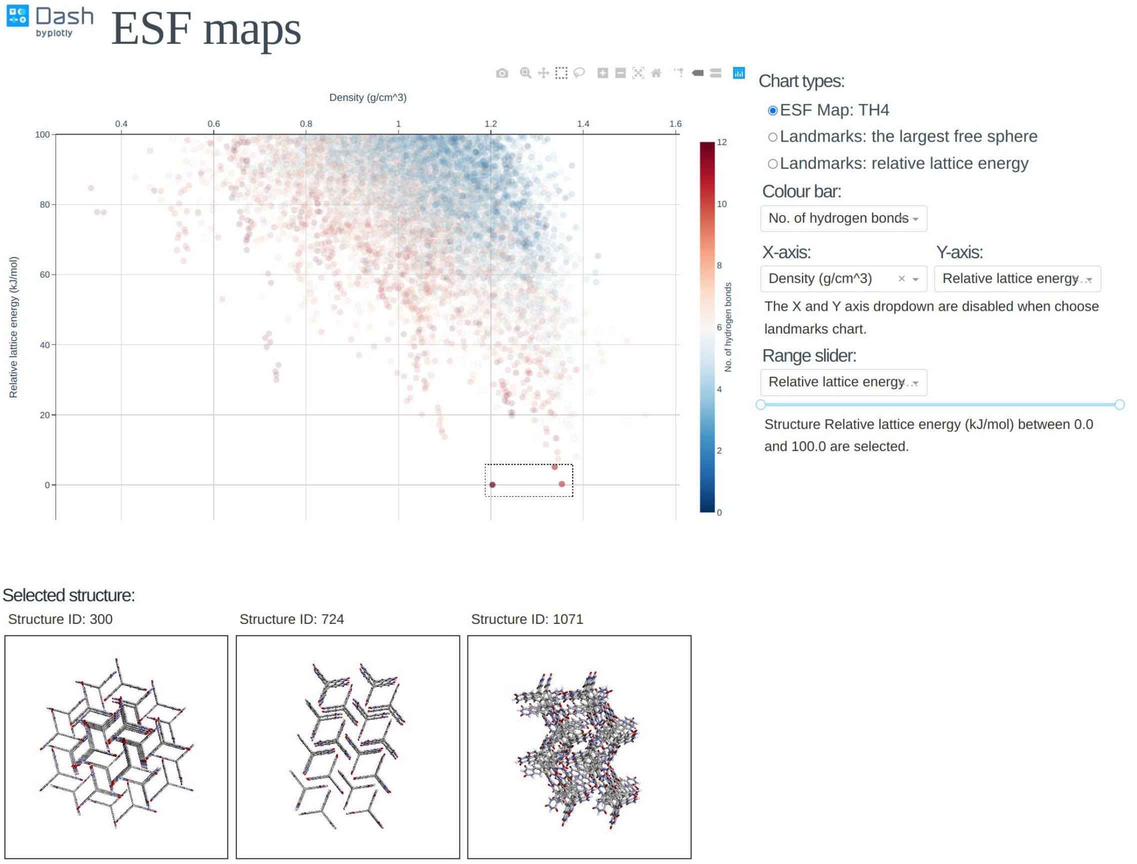


Chart types:

● ESF Map: TH4

○ Landmarks: the largest free sphere

○ Landmarks: relative lattice energy

Colour bar: No. of hydrogen bonds

X-axis: Density (g/cm^3)

Y-axis: Relative lattice energy

The X and Y axis dropdown are disabled when choose landmarks chart.

Range slider: Relative lattice energy

Structure Relative lattice energy (kJ/mol) between 0.0 and 100.0 are selected.

Selected structure:

Structure ID: 300

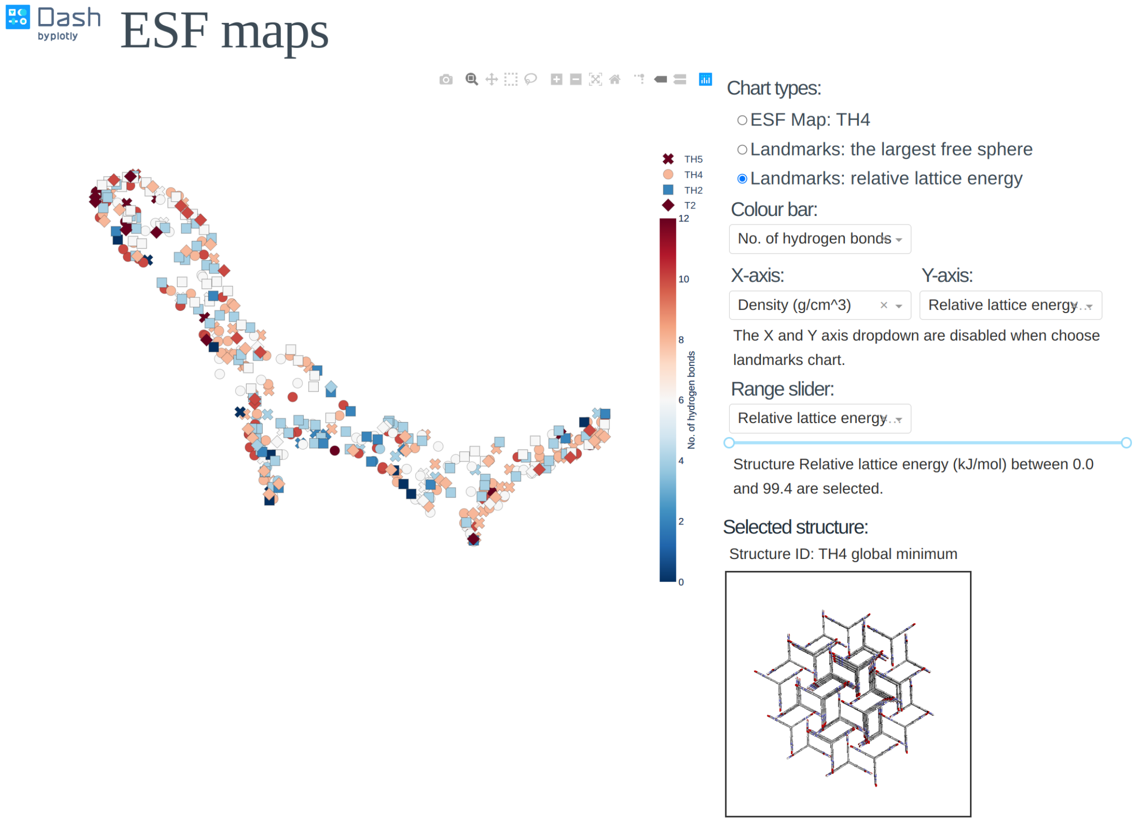
Structure ID: 724

Structure ID: 1071

Figure 2.4 The energy-structure-function (ESF) maps of TH4 molecules on web application for plotting/display. Chemical structures can be interactively displayed upon selection of data points on the plot to allow comparison of different crystals and molecules.

More generally, the visualization interface is highly customizable and extendable. For instance, the 2D embeddings of the porosity spaces of TH2, TH4, TH5 and T2 landmarks figure can also be represented online with flexible of colour bar and 3D structure visualization. Users can switch the chart type between ESF map and 2D UMAP embeddings figure to inspect crystal structures of each sample. Furthermore, pre-built components, such as HTML images and JavaScript plugin models for 3D chemical structure visualization(Jsmol),[150] can be easily linked up with the interface. The Dash package provides many flexible web components to interactive the plots such as live updating new data, cross filter data between multiple figures, and searching/filtrating samples by input text.

The above approach is interchangeable and can be readily adapted to other classes of



ESF maps

Chart types:

○ ESF Map: TH4

○ Landmarks: the largest free sphere

● Landmarks: relative lattice energy

Colour bar: No. of hydrogen bonds

X-axis: Density (g/cm^3)

Y-axis: Relative lattice energy

The X and Y axis dropdown are disabled when choose landmarks chart.

Range slider: Relative lattice energy

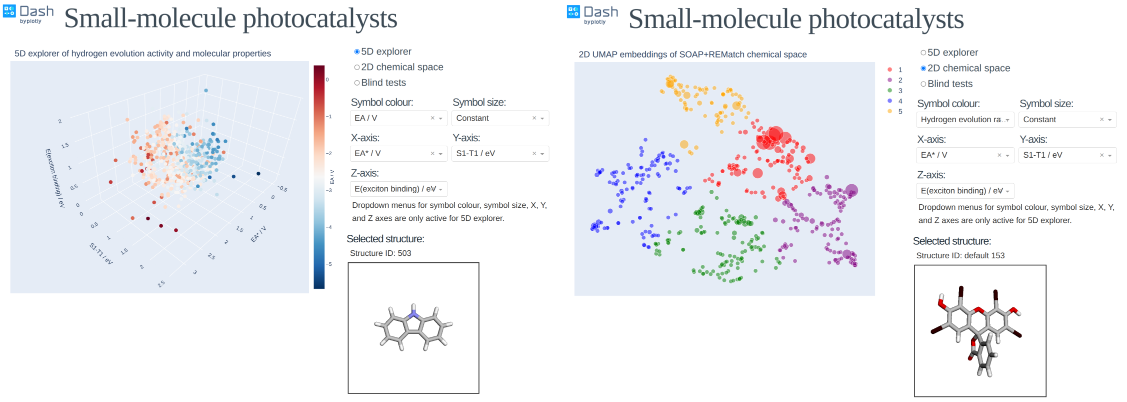
Structure Relative lattice energy (kJ/mol) between 0.0 and 99.4 are selected.

Selected structure:

Structure ID: TH4 global minimum

Figure 2.5 The web application of 2D UMAP embeddings of the porosity spaces of TH2, TH4, TH5 and T2

materials; for example, Figure 2.6 is a screenshot of an interactive, web-based application implemented a library of candidates organic photocatalysts (https://www.molecular-photocatalysts-library.app/).[151] Drawing on the laboratory's existing chemical stocks, my collaborators identified 572 aromatic molecules and investigated their performance for photocatalytic hydrogen evolution rate (HER) activity using high-throughput property measurements. A range of key optoelectronic, excited-state properties, and energy levels are determined computationally using density functional theory (DFT). Such features presented on a 5D interactive application (Figure 2.6(a)) allows for new physical insights and/or new design principles to be developed by other catalysis researchers. Similarly, all the molecules were first encoded, here using the SOAP descriptor,[123] with their pairwise distances determined by the REMatch kernel[152] implemented by DScribe package.[153] Then, they were spatially arranged onto a 2D diagram (Figure 2.6(b)) respecting their similarities, using the UMAP dimensionality reduction method, and colour-coded by



Symbol colour: EA / V
Symbol size: Constant
X-axis: EA* / V
Y-axis: S1-T1 / eV
Z-axis: E(exciton binding) / eV
Dropdown menus for symbol colour, symbol size, X, Y, and Z axes are only active for 5D explorer.
Selected structure:
Structure ID: 503

Symbol colour: Hydrogen evolution ra...
Symbol size: Constant
X-axis: EA* / V
Y-axis: S1-T1 / eV
Z-axis: E(exciton binding) / eV
Dropdown menus for symbol colour, symbol size, X, Y, and Z axes are only active for 5D explorer.
Selected structure:
Structure ID: default 153

(a) 5D explorer of HER and molecular properties

(b) 2D UMAP embeddings of SOAP and REmatch chemical space

Figure 2.6 Structure-activity relationship map of the molecular photocatalyst library

their photo-catalytic activities. This interactive explorer was future migrated to visualize another MOFs dataset (https://www.ch3i-capture-by-mofs.app).[154]

Another future challenge for HTVS is the infrastructure for data storage and retrieval that is safe, flexible and efficient. One of the most common approaches is the use of Structured Query Language (SQL)[155] in relational dataset management system; in other words, the structured orthogonal data always have the same dimensional features. Other related alternatives include NoSQL (originally referring to "non-SQL" but often to "not-only-SQL" as well) databases that are often found to be better suited to chemical data, because of their flexibility and capability of handling rapidly changing data types. In this project, MongoDB,[156] a cross-platform document-oriented database programme, is chosen to manage the storage and retrieval of the HTVS data. Briefly, each molecule or crystal structure is stored as an entry in a JavaScript Object Notation (JSON) format document—either directly or, for crystal structures, converted by the Pymatgen toolkit[157]—and stored in a MongoDB database, allowing the user to retrieve data via its python API and to interface it with other python functions. The standardized data storage format allows collecting chemical information across different research groups and will be beneficial to build larger chemical dataset for the research community.

2.4 Summary

Overall, the demonstrated process provides a simple and general framework for representing the high-dimensional data of ESF maps and for systematically identifying 'landmark' structures on the map. By applying machine learning algorithms to pore features, calculated electronic features, as well as SOAP representations, 2D embedded chemical feature space could be learned, which are human interpretable. This approach of encoding, learning, and representing high throughput chemical dataset enables an efficient navigation of the complex ESF space within a unified framework, allowing researchers to identify energetically favourable or functionally interesting structures across different systems, as well as revealing complex structure-function correlations that are hidden when inspecting individual structural features. This makes a step forward an automated analysis of HTVS, which will be beneficial in facilitating autonomous searches for chemical materials. Besides, the online interactive explorer that were developed here (https://www.interactive-esf-maps.app/, https://www.molecular-photocatalysts-library.app/, and https://www.ch3i-capture-by-mofs.app) might allow for new physical insights and/or new design principles to be developed by other researchers.

CHAPTER 3 FUNCTIONAL PHOTOCATALYTIC MOLECULES SCREENING BY MACHINE LEARNING AND EXPERIMENTS

3.1 Introduction

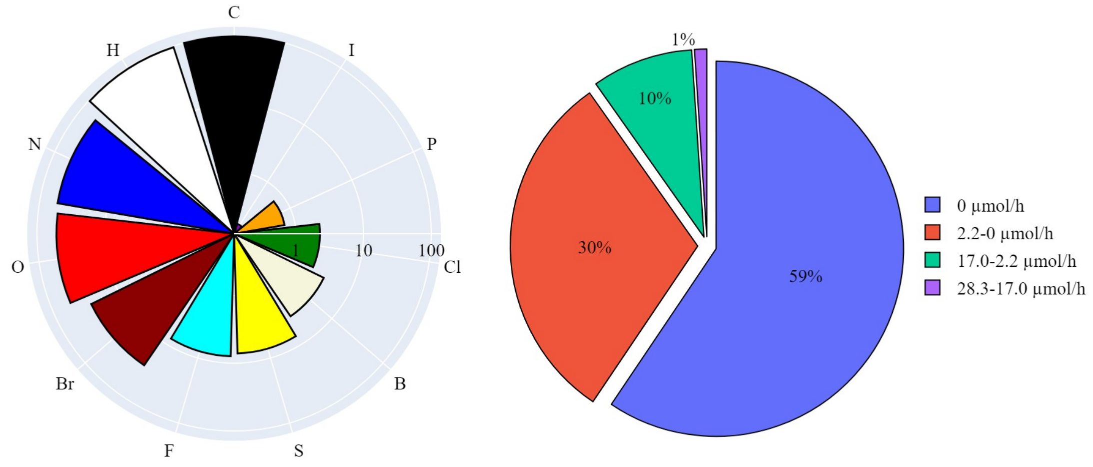
There is a continuous interest in organic materials, including conjugated polymers, covalent organic frameworks, and π-conjugated molecules, as alternatives to inorganic materials for the photocatalytic generation of solar fuels.[158-161] However, it remains a challenge to predict the activity of an organic photocatalyst, based on intuitions or guided by computations. The performance is influenced by a host of factors spanning multiple length scales, such as light absorption, thermodynamic driving force, exciton binding energy, charge carrier mobility, and so on.[162] To de-convolute such complex, multivariate relationships, sufficient data are required, which should ideally be collected consistently to diminish uncontrollable sources of errors, for example, from different labs. However, most studies in the literature are individually focused on only a handful of catalysts, hence making it significantly hard, if not impossible, to establish a predictive model considering more than a couple of factors.

ML-based approaches are often referred to as data-driven approaches, implying that they require sizeable datasets to train on. This could make them unattainable when the acquisition of experimental data is time-consuming and expensive. For photocatalysts, this challenge is further compounded by the fact that mining the literature is rarely an option due to data inconsistencies originating from the lack of a field-wide standard for data collection and reporting. Currently, the largest library of organic photocatalysts which was measured under identical conditions by high throughput equipment contained 175 conjugated linear polymers.[64] To extend the application of ML techniques on photocatalyst, a more large and diverse organic molecule library was created using a high-throughput, automated method by experimental collaborators. This library of candidate organic molecules is diverse, as it includes many molecules that were originally obtained or synthesized for other purposes, such as the production of porous organic cages, conjugated microporous polymers, and covalent organic frameworks. Additionally, it contains some molecules that had been previously investigated for photocatalytic related issues. The aim is to establish en extensive and diverse library of potential organic compounds that could be promptly obtained and evaluated as a training set for sacrificial photocatalytic hydrogen evolution.

With creating the largest dataset of hydrogen evolution by organic photocatalyst, the density functional theory calculations and machine learning was applied to analyse and predict the activity of the molecules. Through unsupervised learning, the identified correlations between molecular structure and photocatalytic activity, which were human interpretable to some extent. These correlations were also found to be machine learnable and were used to predict the photocatalytic activity through supervised classification algorithms and calculated molecular descriptors. Finally, the potential of machine learning to assist chemists in discovering new photocatalysts was proved through in silico virtual experiments and experimental blind tests.

3.2 Building a library of candidate organic photocatalysts

Drawing on the laboratory's existing chemical stocks, experimental collaborators identified 572 aromatic molecules and investigated their performance for photocatalytic hydrogen evolution activity to build the large dataset of organic photocatalyst activities in hand. This library presents a great opportunity for applying ML methods to interpret the data and to establish possible structure-activity relationships for organic photocatalysts. To ensure that no prior human knowledge of the molecular structure was utilized in assessing photocatalytic performance, every organic molecule available in stock was tested, to minimize ‘intuitions’ skewing the true structure-activity correlation. A total of 11 elements occurred in this library of molecules; the frequencies of their occurrence is shown in Figure 3.1(a). Figure 3.1(b) shows a statistical summary of the photocatalytic hydrogen evolution rates (HERs) of the dataset. In comparison with two benchmark conjugated polymers PCN29 and CTF-1.



(a) Polar bar chart

(b) Statistical summary of the HER

Figure 3.1 A library of 572 aromatic candidate organic photocatalyst molecules. (a) Polar bar chart showing the percentage of molecules in the library containing the 11 different chemical elements that occur: 100% contain C, 96% contain H, and so on. The radial coordinates are on a logarithmic scale. (b) Statistical summary of the photocatalytic hydrogen evolution performance of the candidate molecular catalysts in the library. The hydrogen evolution rate (HER) was classified against two conjugated polymers as a benchmark: carbon nitride PCN29 (2.2 µmol/h)[163] and a covalent triazine framework CTF-1 (ref. 30) (17.0 µmol/h).[164]

To assess the chemical diversity of molecules in terms of the chemical space coverage, it was compared with the linear polymer photocatalysts reported by Bai *et al.*, which is the largest library of organic photocatalysts in a single study to date. The Smooth Overlap of Atomic Positions (SOAP) descriptor[123] was used to encode atomic neighbour environments for the H, C, N and O elements in both libraries. For visual comparison, the Uniform Manifold Approximation and Projection (UMAP)[135] technique was applied to learn a mapping from the high-dimensional SOAP vectors to a two-dimensional (2D) representation (Figure 3.2). This showed that the library of 572 molecules covers a significantly

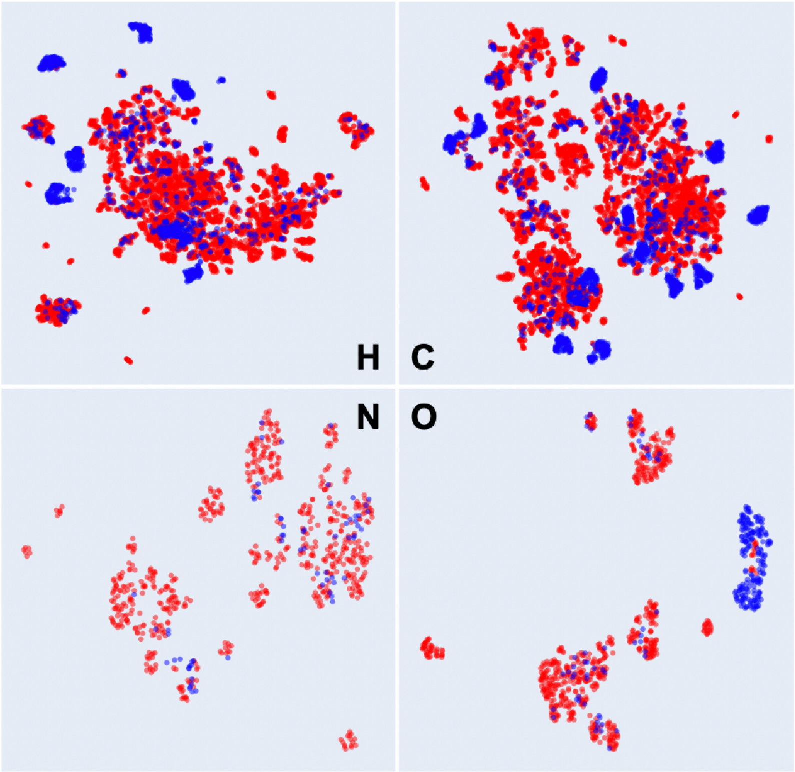


Figure 3.2 Comparison of the diversity of atomic neighbour environments for H, C, N and O elements found in this molecular library (red points) and in the library of 99 conjugated polymers(blue points)[64]

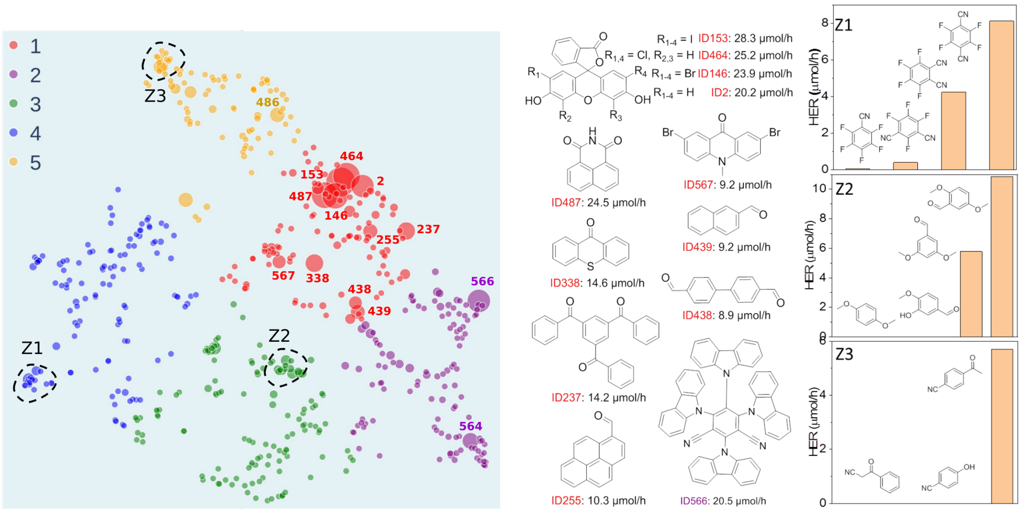
larger chemical space than the polymer library,[64] as expressed by the four key elemental types (Figure 3.2). The higher chemical diversity of the molecular library stems from the larger number of occurring elements (11 vs. 8), as well as this library containing a significantly larger number of different molecules than the total number of unique monomers in the polymer library (572 vs. 99). Also, that polymer library was constructed by using chemical knowledge: specifically, it was biased to include comonomers such as dibenzosulfone that were already known to promote photocatalytic activity.[165]

The photocatalytic hydrogen evolution performance for the small molecule library was investigated using a high-throughput parallel photocatalysts screening platform that utilizes a solar simulator, as described previously.[64] Figure 3.1(b) shows a statistical summary of the photocatalytic hydrogen evolution rates (HERs) of the dataset. The generalized catalytic mechanism proposed here is that the photo-generated excitons on the molecule can either undergo a single-electron reduction or oxidation, mediated by the sacrificial electron donor (TEA) and the proton reduction catalyst (Pt), respectively. In comparison with two benchmark conjugated polymers PCN29[163] and CTF-1,[164] synthesized in-house and measured under exactly the same conditions, 63 molecules showed HERs higher than for PCN (2.2 µmol/h), and 6 molecules surpassed the HER for CTF-1 (17.0 µmol/h). The highest HER among these molecular photocatalysts (ID153; see Figure 3.3 for structure) was 28.3 µmol/h (5660 µmol/g/h), which is comparable to the highest HER (around 6000 µmol/g/h) measured for the 175 conjugated polymers using the same experimental setup using a more design-led approach.[64]

3.3 Mapping structure-activity relationships of experimental data

3.3.1 Mapping structure-activity correlations

To investigate possible structure-activity correlations in this library, the SOAP descriptor is used to encode the molecules and, together with a regularized entropy match (REMatch) kernel,[152] to quantify the similarity between all pairs of molecules. The resulting similarity matrix was then projected onto a 2D space by a UMAP embedding, as shown in Figure 3.3(a), where each point represents a molecule. The size of each point relates to the photocatalytic activity of the molecule (the HER). The points are arranged spatially such that the closer the two points are on the plot, the more similar the two molecules are, as described by SOAP. Then, the k-means algorithm is used to identify clusters on the 2D UMAP space, showing that the 572 molecules can be broadly clustered into five groups (colour coded in Figure 3.3(a)), based on their chemical and structural similarity. This figure shows that there are correlations between molecular structure and hydrogen evolution activity in the dataset. For example, the molecules in the library with high HERs are mostly located in group 1 (red points on plot). Within each of the five larger sub-groupings, molecules with relatively high catalytic performance tend to form smaller,



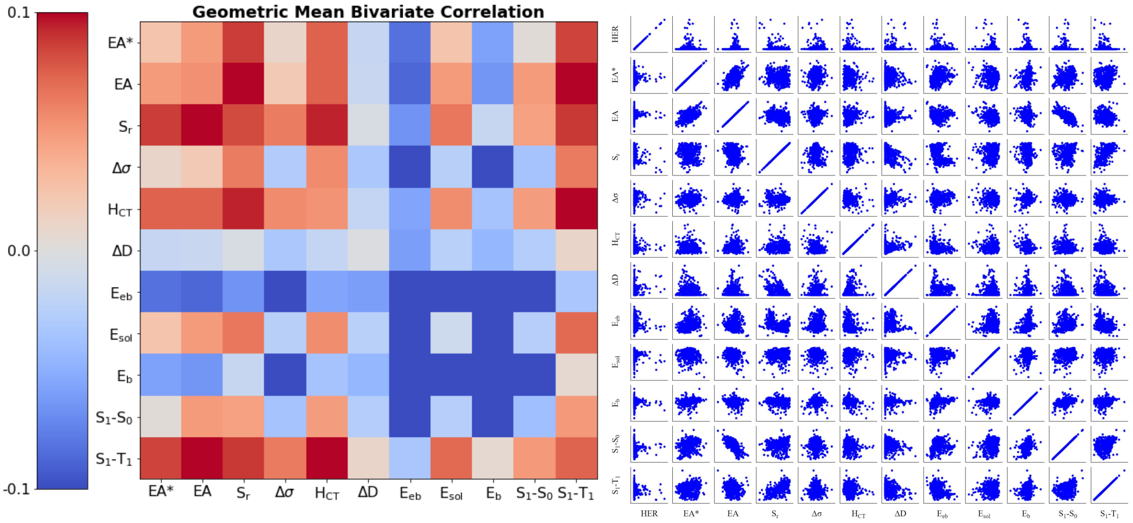
(a) (b)

Figure 3.3 Structure-activity map of the molecular photocatalyst library. (a)2D UMAP embedding of the chemical space of the photocatalyst library, colour-coded by k-means clusters identified using the 2D UMAP coordinates; symbol size denotes the experimentally measured hydrogen evolution rate (HER). (b)Plots showing relative HERs for groups of structurally similar molecules; the locations of these three groups (Z1-Z3) are circled on the UMAP plot in (a).

local clusters. Structural analysis of the molecules with the highest activities (>9 µmol/h) revealed that all but one examples (ID566) shared the common structural feature of having at least one aryl carbonyl moiety. However, it is worth noting that molecules with similar structures can show large differences in hydrogen evolution activities; for example, the structural isomers shown in sub-cluster Z1-Z3(Figure 3.3(b)). In group Z1, the molecules feature fluorine and cyanide substituent on the benzene ring, while group Z2 consists of molecules sharing a methoxy group. Similarly, group Z3 comprises molecules that share a cyanide group. Despite their structural similarities, these molecules exhibit distinct hydrogen evolution activities.

3.3.2 Machine learning the hydrogen evolution activity

For an organic molecule to act as an efficient hydrogen evolution photocatalyst, it must absorb light efficiently and drive thermodynamically the reduction of protons and the oxidation of water or, in this study, a sacrificial agent (TEA). To achieve this, the density functional theory (DFT) calculation was performed by collaborators using Gaussian 16 software[166] to represent the catalytic activity by 11 molecular electronic features, which includes the electron affinity (EA), the exciton electron affinity ($EA*$), the exciton binding energy (E_{eb}), the solvation energy of the molecule in water (E_{sol}), the self-binding (in a dimer) energy (E_b), and a range of key optoelectronic and excited-state properties: (I) light absorption (optical gap, $\Delta E_{S_1 \to S_0}$), (II) change in dipole moment between S_1 and S_0 (ΔD), (III) degree of spatial extension of hole and electron distribution in the charge-transfer direction (H_{CT}), (IV) the difference in the extent of spatial distribution between electron and hole ($\Delta\sigma$), (V) electron-hole overlap (S_r), (VI) the energy gap between the first singlet (S_1) state and the first triplet (T_1) state ($\Delta E_{S_1 \to T_1}$). To gain insight into the dependence of HER on these various calculated descriptors, the Pearson's correlation coefficients for individual features and their binary combinations were explored firstly (Figure 3.4(a)).



(a) (b)

Figure 3.4 Visualization of the feature correlation. (a)Bivariate Pearson's correlation between the HER and all pairs of the calculated molecular descriptors, where the scale runs between -0.1 and +0.1; the diagonal running from the top-left corner to the bottom-right corner shows the correlation between the HER and individual descriptors. (b)One-to-one correlation between all pairs of the calculated molecular descriptors and the measured HER.

The cells on the diagonal (top-left to bottom-right) of Figure 3.4(a) shows the extent of linear correlation of the HER with individual variables, while the off-diagonal cells contain the geometric mean of the correlation of HER with each of the two descriptors. The absolute value of the Pearson correlation coefficient is less than 0.1 for all variables and variable pairs, indicating a weak linear correlation, if any, between the HER and single descriptors or binary combinations of them. This shallow statistical analysis will not capture any complicated or non-linear behaviours dependent on multiple features, but it confirms that any possible structure-property-activity relationship in our dataset is of a non-linear, multivariate nature. Figure 3.4(b) shows that the HERs are not linearly dependent on any individual descriptors, nor is any pair of the calculated molecular descriptors correlated in a simple way.

Next, a number of machine learning (ML) models were evaluated for their suitability to construct predictive models together with the computed molecular descriptors. This included k-nearest neighbours (KNN), random forests (RF), support vector machines (SVM), Gaussian processes (GP), gradient boosted decision trees (GB-DT), and multi-layer perceptron (MLP), all of which have been used in various areas of chemistry and materials science.[11,167-168] The models are trained for tiered classification tasks based on optimized HER thresholds. By transforming a regression problem into a lower-resolution classification problem, the models act as a filtration step for flagging potentially photoactive candidate molecules. For binary classification, this resulted in one class being assigned to HER values smaller than 1.07 µmol/h, with the other class assigned to values larger than 1.07 µmol/h. The class thresholds (in µmol/h) for ternary classification were 1.07 and 12.5; that is, low: $\mathrm{HER} \leqslant 1.07$, medium: $1.07 < \mathrm{HER} \leqslant 12.5$, high: $\mathrm{HER} > 12.5$. The quaternary classification was also attempted, as well as regression tasks, but no satisfactorily predictive models could be achieved(Figure 3.5).

Leave-one-out results showed that the calculated molecular descriptors were successful at producing binary and ternary classifications with greater than 87% accuracy, independent

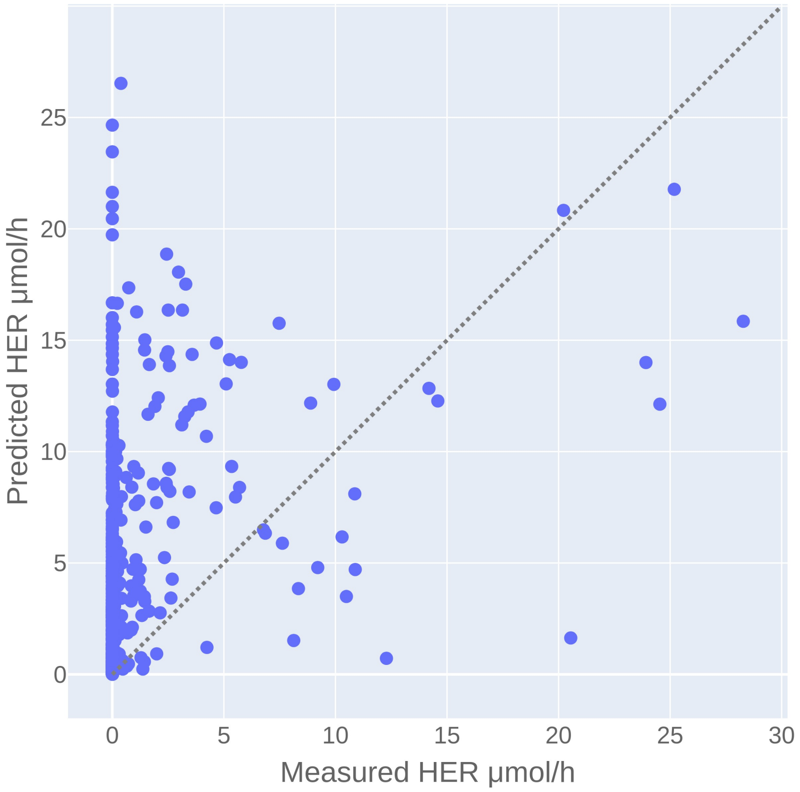


Figure 3.5 The 5-fold cross validation result of KNN regression. Number of neighbours=5; using SOAP REMatch kernel as the precomputed metrics.

of the model type (Tables 3.1). The use of 10-fold cross-validation affords computational efficiency but fails to produce high F1-scores. This was a result of class imbalance: there are far more data points in the ‘low’ performance class than in the ‘high’ performance class (492 vs. 80), and 59% of molecules in the library produced no hydrogen at all, thus exposing the classifier to more information related to the low-performance case. This occurs when the dataset is sampled uniformly for each fold of cross validation; this issue remains when using biased sampling to force each fold to have a constant amount of each class. Class imbalance is a core challenge in applying machine learning to a wide range of research problems in the physical sciences, such as diversity-oriented screening for new photocatalysts, where there are often far more zeros in a dataset than non-zero values. Overall, our results show that the use of molecular descriptors that quantify a range of photochemical and electronic features of the molecule, in conjunction with ML models, can predictively assign HER performance levels (low, medium or high) to candidate pho-

Table 3.1 Binary and ternary classification metrics across all models, obtained by 10-fold and leave-one-out (LOO) cross-validation procedures

Model	Binary			Ternary		
	10-fold		LOO	10-fold		LOO
	Accuracy[a]	F1-score[b]	Accuracy	Accuracy	F1-score	Accuracy
KNN	0.89	0.69	0.89	0.89	0.61	0.89
GP	0.87	0.57	0.87	0.87	0.42	0.87
RF	0.89	0.69	0.88	0.88	0.57	0.89
GB-DT	0.89	0.69	0.88	0.88	0.57	0.89
SVM	0.87	0.68	0.87	0.88	0.58	0.87
MLP	0.89	0.71	0.89	0.89	0.56	0.88

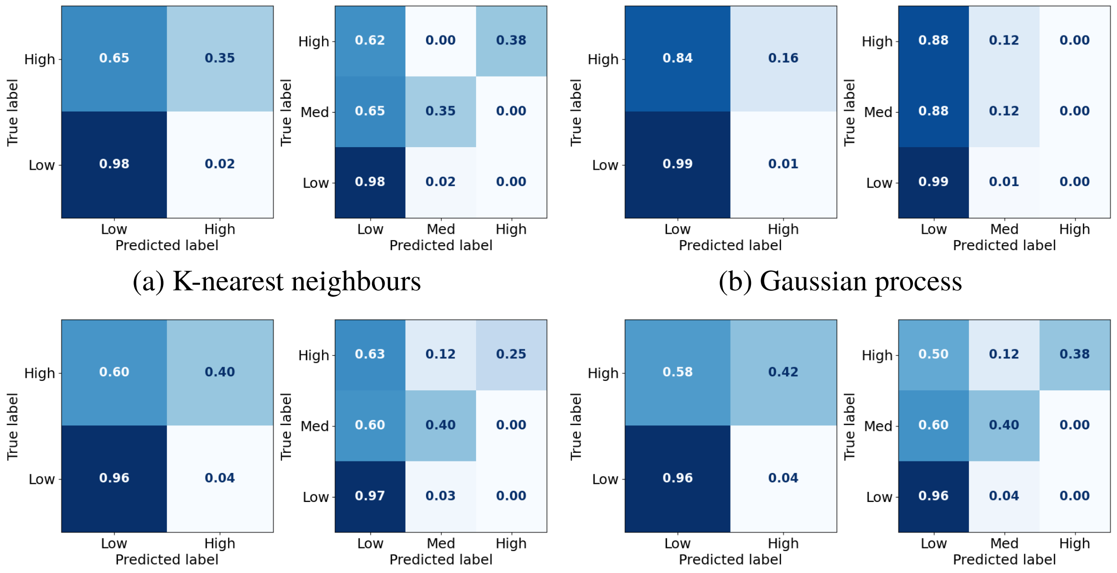
[a] The sum of the number of true positives (TP) and true negatives (TN) divided by the sum of the number of true positives, true negatives, false positives (FP), and false negatives (FN).

[b] Weighted harmonic mean of precision and recall, where precision is the number of true positives divided by the sum of the number of true positives and false positives; recall is the number of true positives divided by the number of true positives and false negatives. For ternary classification, metrics are computed independently for each class and then averaged (macro average).

tocatalysts, albeit with limitations.

Comparing KNN with the other ML models (Tables 3.1), it is more important than the ML model itself that the relative similarity of the models and the largely interpolative behaviour of them, with characterization. This is unsurprising because of the class imbalance challenge described above as well as missing any mesoscale experimental factors, which are not captured by the current set of molecular descriptors, as discussed further in Chapter 3.5.

To assess the practical utility of these models, how the models failed was explored. From the confusion matrices shown in Figure 3.6, it is clear that the experimentally high-performing catalysts are more often mislabelled as being ‘low’ performers than the op-



(a) K-nearest neighbours

(b) Gaussian process

(c) Gradient boosted decision trees

(d) Multilayer perceptrons

Figure 3.6 Confusion matrix for binary and ternary classifiers based on different models

posite case. This is a result of the significant class imbalance, discussed above. The current models are robust against producing false positives (more than 95% ‘low’ performers are correctly labelled by all the models), and hence useful to screen out candidates that would show zero or low hydrogen evolution activities. Some ‘high’ performing molecules will also be eliminated because they are mislabelled as ‘low’ performing, but this behaviour could be acceptable when the cost of experiment is high and evaluating an excess of candidates becomes expensive—for example, to guide investigations that cannot access high-throughput screening facilities, as used here. To minimize the model’s proneness to false negatives—that is, mislabelling ‘high’ performers as ‘low’—more data points in the ‘high’ HER class would be required to improve the model’s performance. By examining these confusion matrices and the performance metrics in Table 3.1, the MLP models was identified as the strongest binary and ternary classifiers.

Binary and ternary classification tasks were also performed for the 572 molecules, using only the molecular structure as input representation. To encode the molecules for machine learning, both Morgan fingerprints and SOAP descriptors were tested, together with using the Tanimoto index or the REMatch kernel as the similarity measure (further details are given in Table 3.1). KNN and SVM models were evaluated for both structural representations, using their respective, precomputed distance metrics. All the KNN models outperformed their SVM counterparts, in both binary and ternary classifications, for both structural representations (Table 3.2). The SOAP-based KNN model was identified as the strongest binary classifier, while the Morgan fingerprints-based KNN model was identified as the strongest ternary classifier; both of these models performed well in both binary and ternary classification tasks. These results show that the structure-activity correlations that are only somewhat human-interpretable in Figure 3.3 are machine-learnable, achieving equivalent levels of predictive ability to the strongest ML models using engineered descriptors (Table 3.1). It would be particularly advantageous to use structure-based ML models to guide large-scale experimental screening of molecular photocatalysts, as expensive descriptor calculations would otherwise become the bottleneck to increasing the throughput. Naturally, such models do not intuitively highlight physical features of high-performance photocatalysts, which could then be used to guide the design of better cata-

Table 3.2 Binary and ternary classification metrics across models based molecular fingerprints or SOAP descriptors, obtained by 10-fold cross-validation procedures

Representation	Model	Binary[a]		Ternary	
		Accuracy	F1-score	Accuracy	F1-score
Fingerprints[b]	KNN[d]	0.88	0.68	0.88	0.63
Fingerprints	SVM[e]	0.77	0.48	0.77	0.34
SOAP[c]	KNN	0.87	0.73	0.88	0.60
SOAP	SVM	0.75	0.47	0.84	0.32

[a] The class thresholds were the same as those in Table 3.1.

[b] Morgan fingerprints with a radius=2, generated by RDKit; similarity measure: Tanimoto index.

[c] SOAP descriptors with r=6.0, n=8, l=6, generated by DScribe; similarity measure: regularized entropy match (REMatch) kernel. The similarity matrix for the 572 molecules used here is the same as the one used for Figure 3.2.

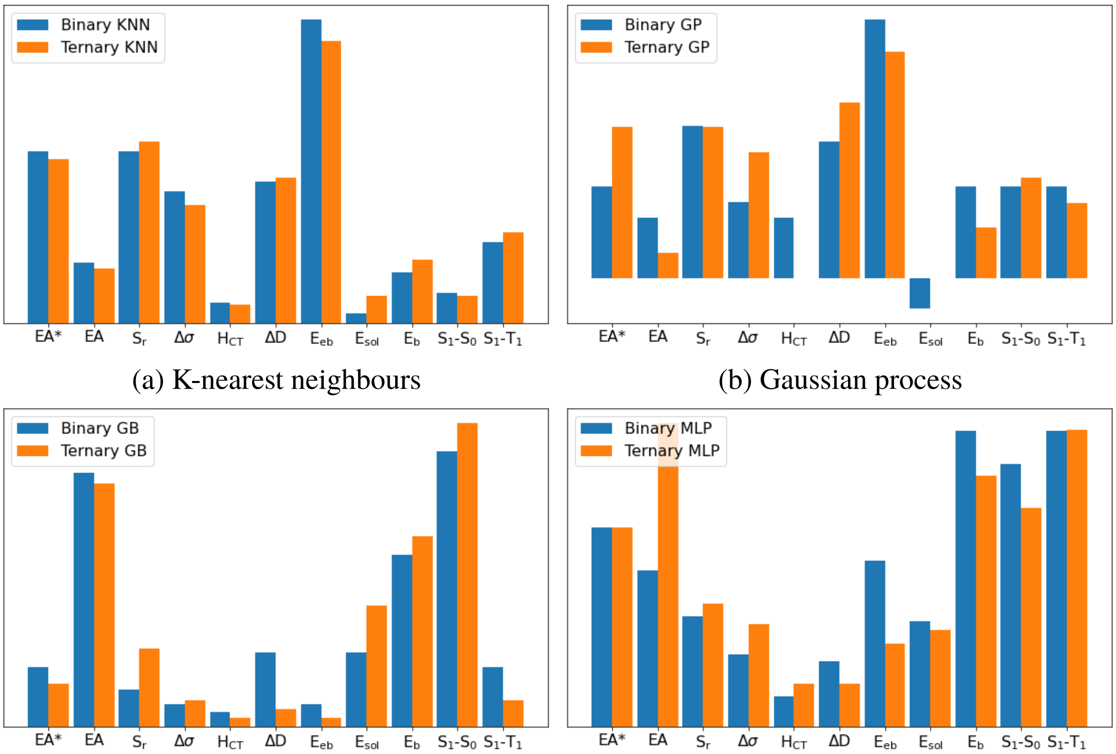
[d] number of neighbours=5; metric=precomputed.

[e] C=15.6; metric=precomputed.

lysts, nor do they reveal directly the structural features that may have correlated with the photocatalytic activity.

3.3.3 Understanding the important molecular features for photocatalytic activity

Beyond their predictive ability, interpretability is a key goal for ML models to understand the importance of each descriptor and to obtain physical insights into structure-property-activity relationships. Permutation importance was calculated for four of the models presented in Figure 3.7. It works by randomly permuting the values of a particular feature and then evaluating the decrease in a model score of these changes on feature. Permutation importance can vary across different machine learning models due to differences in model architecture, training algorithms, data characteristics, feature interactions, and randomness. The emphasis on features can vary depending on the model structure. For instance, the training techniques are also different for these ML models. GP using the log marginal likelihood as the cost function to assign the hyperparameters, whereas MLP using the softmax function for the classification task. Except this, in a tree-based model, the importance of features is often determined based on their frequency of use for splitting. On the other hand, a linear model assigns importance to features based on the magnitude of their coefficients. The MLP models, the strongest binary and ternary classifiers, assign high relative importance to exciton electron affinity (EA^*), electron affinity (EA), exciton binding energy (E_{eb}), optical gap ($\Delta E_{S_1 \to S_0}$), and singlet-triplet energy gap ($\Delta E_{S_1 \to T_1}$) for both binary and ternary classification tasks. EA^* estimates the thermodynamic driving force for the molecular photocatalyst to oxidize the sacrificial agent, TEA. EA estimates the thermodynamic driving force for proton reduction. $\Delta E_{S_1 \to T_1}$ estimates the optical gap of the molecular photocatalyst. EA^*, EA and $\Delta E_{S_1 \to S_0}$ are intuitively important, because they are essential optoelectronic requirements for a molecule to act as a photocatalyst: that



(a) K-nearest neighbours

(b) Gaussian process

(c) Gradient boosted decision trees

(d) Multilayer perceptrons

Figure 3.7 Extracted permutation feature importance based on different models in binary and ternary classification tasks

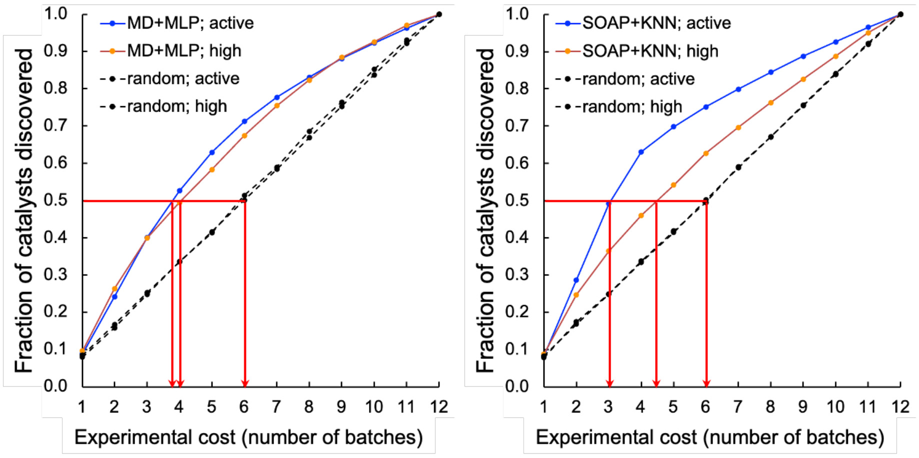
is, the molecule must absorb light efficiently over a broad range in the visible spectrum as well as having enough thermodynamic driving force to oxidize TEA (EA^*) or to reduce protons (EA). Importantly, the MLP models identified two additional molecular properties, E_{eb} and $\Delta E_{S_1 \to T_1}$, that correlate strongly with a high photocatalytic activity.

Molecules ID146, ID153, ID255, ID338 and ID487 are among the most active photocatalysts in this study and share the common structural feature of having at least one aryl carbonyl moiety (Figure 3.3). Their high HERs might be attributed, at least in part, to their ability to generate triplet excitons.[123,169] However, an attempt to correlate HERs for some molecules with their reported triplet-state yields in isolation failed to produce any correlation. This is perhaps unsurprising since the hydrogen evolution activity of a photocatalyst is rarely governed by a single physicochemical or optoelectronic property but rather by a host of molecular and mesoscale factors. As such, more sophisticated approaches—such as the structure-based or the descriptor-based machine learning demonstrated here—hold the promise for using data-driven strategies to probe the complex structure-property-activity relationship for molecular photocatalysts. In addition to the intrinsic challenge of classifying reactions that are dictated by a complex, interrelated set of factors, there are other experimental factors that may contribute to the difficulty of this classification task. While all reactions were conducted under the same experimental conditions, the generalized mechanism of hydrogen evolution makes various assumptions: the hydrogen produced is generated from the water, rather than from the organic molecule itself. Normally, this can be conformed for each reaction via isotopic labelling experiments, but this is more challenging for such a large library of reactions. Another consideration is solubility: while the molecules selected have, on the whole, low aqueous solubility, some molecules in the library might have finite solubility in the water/TEA/MEOH mixture, and this feature was not accounted in the descriptors. Also, the interaction between the organic molecules and the Pt co-catalyst is important—which could be influenced by particle size, surface properties, or the Pt loading method—but these factors were not captured explicitly by any of the descriptors used. That said, the objective of this work was to build a useful classifier with affordable experimental cost, and in this respect, a balance must be struck between exactness and complexity of the experiments.

3.4 Virtual experiments and blind tests

To assess the potential for using an ML ‘advisor’ to discover molecular photocatalysts, the in silico experiments was designed on the 572 molecules using their measured HERs as the ground truth to evaluate the search performance. Figure 3.8(a) shows that it took, on average, about 3.8 and 4.0 batches to discover 50% of the active and high-activity catalysts, respectively, using the ML advisor. Using the random selection approach, it took 6 batches to discover the same proportion of active and highly active photocatalysts. Similarly, using binary and ternary classifiers both built on SOAP-based KNN models

(Figure 3.8(b)), the adaptive approach was able to discover 50% of the active and high-activity catalysts within, on average, about 3.0 and 4.5 batches, respectively. The use of this adaptive ML advisor to assist the chemist could therefore significantly reduce the experimental cost for finding promising photocatalysts, thus providing a predictive method to explore large molecular search spaces.

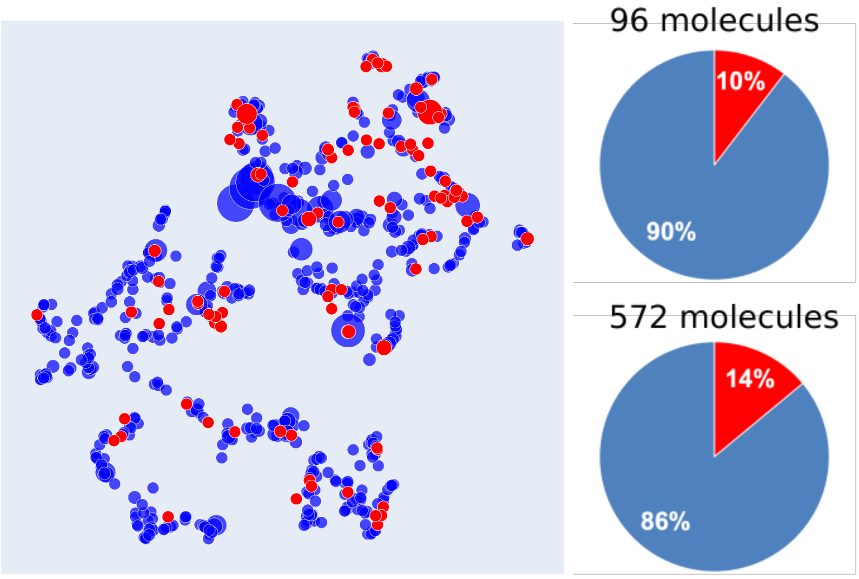


(a) Molecules encoded by the molecular descriptors

(b) Molecules encoded by the SOAP descriptors

Figure 3.8 Virtual experiments comparing an adaptive machine learning approach with random sampling: the 572 molecules were encoded by the molecular descriptors(MD) and trained with MLP models (a) or encoded by the SOAP descriptors and trained with KNN models (b). Active samples were defined as having HERs > 1.07 µmol/h and high-activity samples as having HERs > 12.5 µmol/h. The average number of batches taken to find 50% of the active and highly active catalysts is marked by the red arrows. A total of 200 in silico experiments was carried out for both the ML approach and for the random sampling method, each with a different random starting point, to obtain these average results.

To better assess its potential in real-world applications, 96 extra molecules that were not included in the initial 572-molecule photocatalyst library were tested on the ML advisor. The 96 molecules, referred to as the blind-test set, were selected considering only their aromaticity and (again) availability in the lab, as for the first 572 molecules. They were measured in two batches using our high-throughput parallel photo-catalysis screening platform by collaborators. The blind-test set falls within the chemical space of the 572-molecule library (Figure 3.9(a)) and has a similar percentage (10%) of active samples to that of the 572-molecule library (14% in Figure 3.9(b))-— 10 out of the 96 molecules

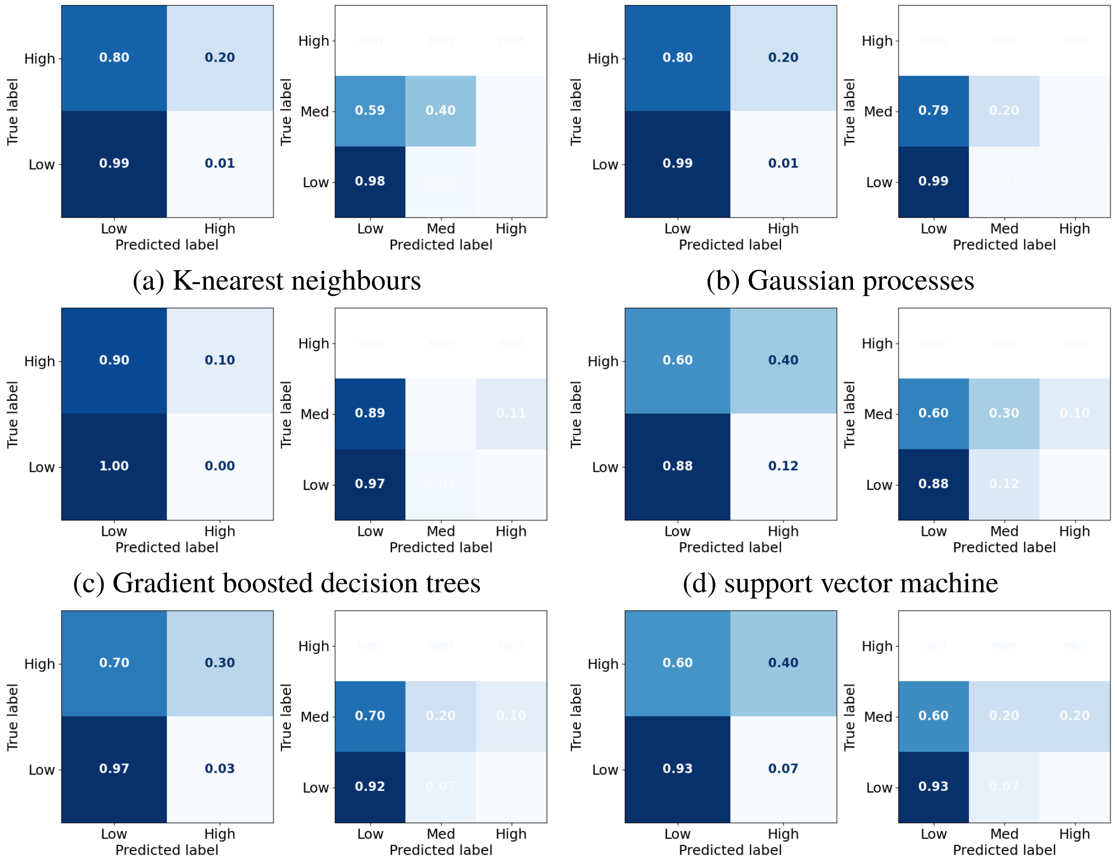


(a) (b)

Figure 3.9 Comparison of the 572-molecule library and the blind-test set. (a) 2D UMAP embedding of the chemical space (encoded by SOAP) of the 572-molecule library (in blue) and the blind-test set (in red); the symbol size is scaled by the experimentally measured HER. (b) Percentages (in red) of the active samples (HERs > 1.07 µmol/h) in the 572-molecule library and the blind-test set.

had HERs larger than 1.07 µmol/h, none of which was greater than 12.5 µmol/h.

In predicting for the blind-test set, the MLP model was again identified as the strongest binary classifier, when combined with the calculated molecular descriptors (Figure 3.10(e)); the MLP model was ranked second for ternary classification, with a slightly inferior performance to the KNN model (Figure 3.10(a)). Binary classification for the blind-test samples directly from their molecular structures (Figure 3.10(f)), encoded by SOAP descriptors, using KNN yielded an equivalent level of predictive accuracy to that achieved by the strongest binary classifier using molecular descriptors. For ternary classification, the descriptor-based KNN (Figure 3.10(a)) markedly outperformed the structure-based KNN (Figure 3.10(f)). These blind-test results confirmed the potential of using an ML advisor to assist the chemists in the discovery of new molecular photocatalysts, as well as highlighting the promise for structure-based ML models to facilitate large-scale high-throughput screening.



(a) K-nearest neighbours

(b) Gaussian processes

(c) Gradient boosted decision trees

(d) support vector machine

(e) Multilayer perceptrons

(f) K-nearest neighbours of SOAP features

Figure 3.10 Confusion matrices for the predictions of the blind-test set by models based on electronic features and structure features

Looking forward, the predictive ability of machine learning for molecular photo-catalysis might be improved by capturing additional information for the higher-activity molecules. For example, efficient charge transfer between the molecular photocatalyst and the sacrificial agent or the co-catalyst is key to catalytic performance, but such intermolecular effects are not considered explicitly in this study. Future work in engineering descriptors might focus on better capturing the charge-transfer characteristics of the system, as well as the exciton lifetime and transport properties. Second, populating the dataset in the high-activity region is essential for training robust, predictive machine-learning models. Besides, model assembling might increase the robustness of the ML models discovery. Ensembles should include models trained against a variety of descriptors, for example, including those derived from the molecular structure and those abstracted from graph neural networks.[102] Transfer learning may be a particularly promising strategy because the acquisition of large experimental datasets can be time-consuming and expensive; here, models are pre-trained on large datasets with relevant or surrogate properties, followed by task-specific fine-tuning for predictive modelling. The experimental study presented here was a single batch process; that is, all the experiments were done prior to model building, because it was tractable to attempt measuring HERs for all 572 molecules in the available library using the high-throughput automated methods in the lab. For much larger libraries, or where such automation is not available, a more efficient approach would be to build the model 'on the fly', as in the virtual experiments above, and to recommend the next batch of molecules as the model evolves. This could also tackle the class imbalance problem that is discussed above. In this respect, a closed-loop autonomous search would be particularly attractive.[75,170-172]

3.5 Methods of machine learning and virtual experiments

For data visualization, the Uniform Manifold Approximation and Projection (UMAP) technique was used for dimensionality reduction for mapping high-dimensional data to 2D representations, while preserving both global and local topological structures of the data in the high-dimensional space as much as possible. For Figure 3.2, all atoms of one of the four elemental types were used to learn the 2D UMAP embedding of their atomic neighbour environments, in which atoms of any elemental types may be present. In the resulting UMAP-learned 2D representation, points will overlay in the 2D space if they are at the same position in the original high-dimensional space. All machine-learning models were implemented using the scikit-learn package[173] except for the MLPs, which were implemented in PyTorch.[174] Hyperparameters were optimized using a discrete Bayesian optimization[170] and the scikit-optimise package.[175] During model training and optimization, the dataset was split between 80% training and 20% test across 10 different folds. The target metric—accuracy and F1-score for classification—of the resulting 10 models is averaged across all folds during hyperparameter optimization.

To do the virtual experiments, an adaptive ML advisor was used to compare with random sampling. In these in silico virtual experiments, 48 samples were ‘measured’ in each batch, which matches the batch-size of the real high-throughput experiments. In the ML advisor approach, an MLP binary classifier and an MLP ternary classifier were trained on all known data after each batch, and then used to predict a class for each of the remaining untested molecules. The next batch was then chosen from the untested molecules until the 48 slots were filled by selecting, in the following priority order: (i) molecules predicted by the ternary classifier to have high-activity (HER > 12.5 µmol/h); (ii) molecules predicted by both the ternary classifier and the binary classifier to be active (HER > 1.07 µmol/h); and (iii) molecules predicted by the binary classifier to be active. When necessary, the batch of 48 molecules was completed with molecules selected randomly from the non-active class. The classification models were then rebuilt after each batch. For the random sampling approach, each batch of 48 molecules was simply chosen randomly from the untested molecule pool.

3.6 Conclusions

Here the largest library of organic photocatalysts was assembled and tested experimentally to date and tested all 572 molecules under identical experimental conditions using a high-throughput testing methodology. Further tested 96 molecules as a blind-test set for evaluating the trained ML models were added, bring the combined total of experimentally measured molecules to 668. This is a tiny fraction of the total available chemical space, but large enough to construct useful ML structure-property-activity models. Unsupervised learning and supervised classification were used to reveal the structural features and optoelectronic properties that positively impact the activity of these molecular photocatalysts for sacrificial hydrogen production. This suggests further exploration of molecules known for inter-system crossing. Despite being sourced simply on the basis of availability in the laboratory, rather than any more sophisticated rationale, 1% of the molecules in the library (5 in total) performed comparably (4040-5660 µmol/g/h) to the highest HER (around 6000 µmol/g/h) measured for the 175 conjugated polymers using the same experimental setup from a more design-led but much more synthetically expensive approach.

Virtual experiments show that an adaptive ML-assisted selection approach outperforms random sampling (Figure 3.8), significantly reducing the experimental cost of identifying the active photocatalysts in the library. A further evaluation of the trained ML advisor on a blind test set of 96 molecules confirmed its potential in assisting the discovery of new molecular photocatalysts. While some active catalysts discovered could have been prioritized based on existing literature reports (e.g., ID67, ID146, ID153, and ID566),[176-178] others were unknown and non-intuitive, such as ID183 and ID237. As such, these fast screening methods can create new inspiration for future research directions. The ML-assisted rapid screening method could be particularly helpful for problems where there is little or no prior literature to draw upon—for example, in the search for photocatalysts that illicit new, unknown reactivity in organic transformations, where the initial hit rate will be low by definition. In summary, this is one of a relatively small number of studies where machine learning methods have been integrated with high-throughput property measurements across a sizeable and diverse set of materials (668 organic molecules in total). This makes an important step towards for the acceleration of the discovery of molecular photocatalysts by considering a much broader chemical space than previously explored.

CHAPTER 4 CLOSE LOOP DISCOVERY OF PHOTOCATALYSTS

4.1 Introduction

As the studied photocatalysts hydrogen evolution in previous chapter, it is challenging to predict the catalytic activities from first principles, either by expert knowledge or by using a *priori* calculations.[179] This is because the collected molecular library covered variety of organic compounds, and the hydrogen evolution activity depends on a complex range of interrelated properties,[64,160] which are difficult captured by several simulated descriptors. However, the virtual experiment of the photocatalysts hydrogen evolution proved that the close loop discovery is feasible of screening catalysts from a designed chemical space. Here, with the input of experimental collaborators, a two-step data-driven approach to the targeted synthesis of organic photoredox catalysts (OPCs) and the subsequent reaction optimization for metallophotocatalysis, as demonstrated for decarboxylative sp^3-sp^2 cross-coupling of amino acids with aryl halides.[180]

The activation of organic substrates via single-electron transfer using photoredox catalysts is a powerful tool in organic synthesis.[162,181-184] Metallophotocatalysis merges photoredox catalysis with transition-metal catalysis to allow organic reactions that are challenging with a single catalyst.[180,185-188] The photoredox catalyst (PC) must exhibit suitable redox potentials in both the excited and ground states to allow for electron transfer to the substrates/transition-metal catalysts. So far, PCs have mostly been discovered through a mix of design, trial and error, and serendipity.[189] In some cases, high throughput synthesis and testing have been used, particularly when the PCs can be generated *in situ* and do not require an elaborate purification procedure, as demonstrated for the discovery of transition metal complexes as PCs.[190] However, photoredox catalysis is by nature a multivariate problem, involving the intersection of many molecular and mesoscale factors. Besides selecting the best photoredox catalysts, the optimization of reaction conditions—that is, the pairing of photoredox catalysts and transition-metal catalysts, reaction concentrations, and so on—can yield significant improvements in metallophotocatalysis activity. Both of the optimization process can be achieved by the close loop discovery in the defined high dimensional search space.

This data-driven approach comprises two sequential closed-loop optimization workflows, both integrating predictive machine learning with experiments under algorithmic control. The algorithm uses Bayesian optimization (BO) to explore the search space and to inform subsequent experiments.[76,170] First, to identifying promising OPCs, a virtual library of 560 novel yet potentially synthesisable organic molecules was designed using a common molecular scaffold with different functional groups. A batched BO was used to build a model that could be updated and queried to guide the experimental search for the most valuable catalysts. This strategy explored led the experimental collaborator to synthesize 55 molecules out of the total library of 560 candidates, achieving reaction yields of 67% for the target reaction. In the second step, 18 of these synthesized molecules were selected for optimization of the reaction conditions, along with varied concentration of the nickel catalyst and its coordinating ligands. After screening a small fraction of the available catalyst formulation space (107 / 4500 possible sets of reaction conditions), the target reaction yield reached 88%, which surpasses the well-studied organic photocatalyst, 4CZiPN, and is comparable to or better than the performance of iridium photocatalysts, depending on the nickel cocatalyst concentration. This research shows that Bayesian, data-driven experimental design is a promising approach for the discovery of new metallophotocatalysis formulations, and by extension for other research challenges where there is a large search space and limited prior knowledge.

4.2 In silico design of the candidate organic photoredox catalysts

The Hantzsch pyridine synthesis is a multi-component organic reaction that produces pyridine compounds from an aldehyde, two equivalents of a β-ketoester and a nitrogen donor, such as ammonia (Figure 4.1).[191] This reaction is metal-free and features high atom efficiency, using common reactants and facile reaction conditions, which is used here to synthesis the designed OPCs. A virtual library of 560 molecules that all share a common cyanopyridine (CNP) core, functionalized by different chemical moieties is constructed by combining 20 benzoylacetonitrile derivatives and 28 aromatic aldehydes—these functional groups are denoted hereafter as Ra and Rb, respectively. All of these 560 CNP molecules are in principle synthesisable (Figure 4.1). These chemically and structurally

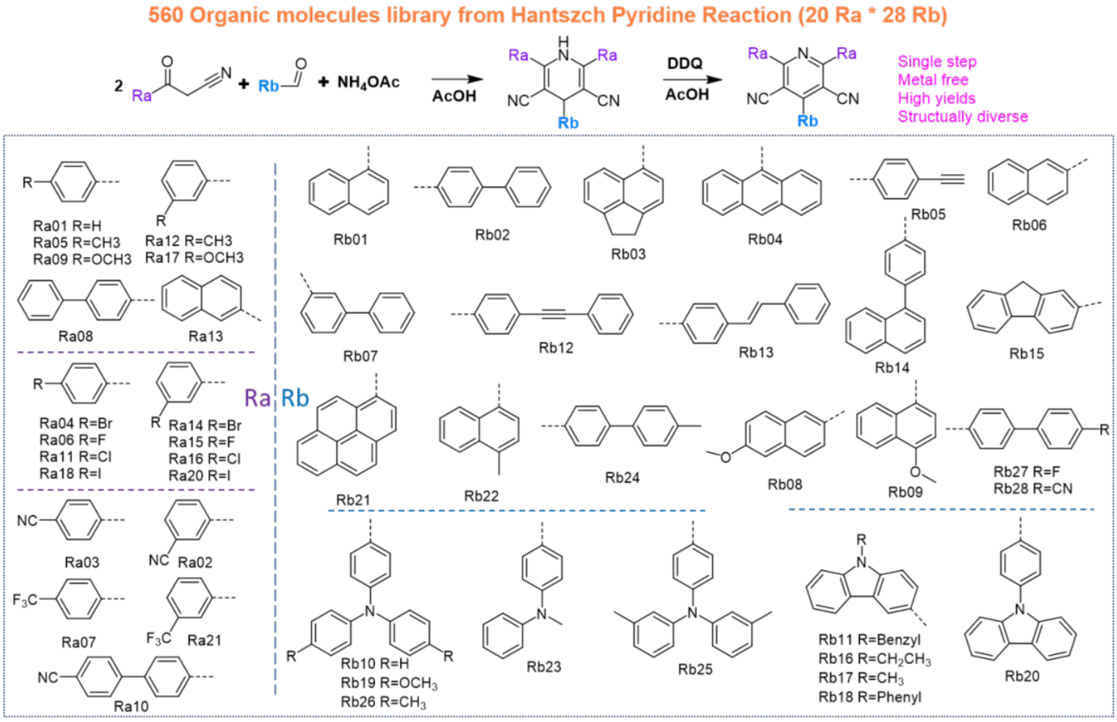


Figure 4.1 A virtual library of 560 candidate CNPs as potential photoredox catalysts. The reaction scheme for the Hantzsch pyridine synthesis is shown along with the various chemical moieties that may be attached to the cyanopyridine (CNP) core at the Ra or Rb positions. The different combinations of Ra and Rb moieties leads to 560 potential CNP molecules in this chemical space. Dashed purple and blue lines indicate sub-groupings of the Ra and Rb moieties based on their structural and chemical similarities.

diverse Ra and Rb moieties generate 560 different binary Ra/Rb combinations when combined into CNP molecules, offering the potential to tune the optoelectronic properties and the redox potentials of the CNPs over a broad range. The cyanopyridine core of the CNPs is analogous to cyanoarenes, many of which are known to be active photocatalysts.[162,192] This diverse library of CNP molecules might contain promising OPCs for both reductive and oxidative photoredox reactions, as photoredox/nickel dual catalysed cross-coupling reactions.[193]

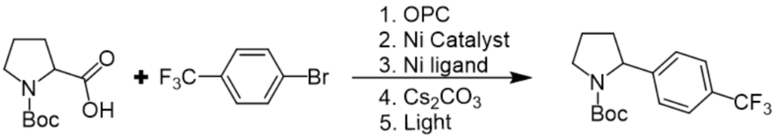


Figure 4.2 The target reaction: a photoredox/Ni dual catalytic sp^3-sp^2 cross-coupling reaction

However, initially there were no clear physical principles to follow when selecting molecules from this library for the target reaction depicted shown in Figure 4.2. Synthesizing and testing all 560 molecules was unrealistic. Therefore, an active learning approach was designed for the selection of CNPs for experiment, which made the use of a closed-loop BO workflow with real-time feedback between experiment and prediction.

4.3 Close loop discovery of organic photoredox catalysts

4.3.1 Encoding the chemical space of CNP photocatalysts

The decarboxylative sp^3-sp^2 cross-coupling reaction that was considered here (Figure 4.2) involves two interwoven catalytic cycles: one is photoredox catalysis and the other is nickel catalysis. This dual catalysis mechanism was discussed in detail previously.[180] Briefly, the photo-excited CNP, CNP*, oxidizes the α-amino acid substrate via a single electron transfer (SET) event, generating an α-amino radical and the corresponding CNP^{-}. Concurrently, the nickel catalytic cycle involves oxidative addition of the Ni(0) species into the aryl halide substrate, producing a Ni(II)-aryl intermediate, which captures the -

amino radical and produces an alk-N(III)-aryl adduct. The desired C(sp^3)-C(sp^2) bond is subsequently forged via reductive elimination. A second SET event between the CNP$^-$ species and the Ni(I) species expelled after the C(sp^3)-C(sp^2) bond formation completes both the photoredox cycle and the nickel catalytic cycle simultaneously, regenerating the CNP photocatalyst and the Ni(0) species.

Thermodynamically, the excited-state photocatalyst, CNP*, must be a strong oxidant for the α-amino acid substrate, Boc-Pro-OH, which has a reduction potential, E^{red}, of 1.19 V versus standard hydrogen electrode (vs SHE). Using density functional theory (DFT) and time-dependent (TD) DFT calculations, the reduction potential for the CNP*→CNP$^-$ half-reaction, $E_{1/2}^{red}$[CNP*/CNP$^-$], was determined for all the 560 CNP molecules. All but four of these CNPs had a calculated value of $E_{1/2}^{red}$[CNP*/CNP$^-$] that exceeded 1.19 V vs SHE, meaning that the oxidation of Boc-Pro-OH by CNP* should be thermodynamically favourable for most of the candidate CNPs. The resulting CNP$^-$, from the oxidation process, must then act as a strong reductant to regenerate the Ni(0) species, for which $E_{1/2}^{red}$[Ni^I/Ni^0] = -1.17 V vs SHE. All but one of the 560 CNPs were calculated to have an $E_{1/2}^{red}$[CNP/CNP$^-$] < -1.17V vs SHE, suggesting that they should be able to drive the Ni reduction process. These computational results suggest that many of the candidate CNP molecules could potentially serve as photocatalysts for the target reaction. However, these findings also demonstrate that the use of redox potentials as selection criteria is not sufficient, since the catalytic activity of these species is not solely determined by their thermodynamic redox potential.

Since there is no straightforward predictive guideline for selecting CNPs for experiment, the selection process was approached as an exploration of the chemical space available to maximize an objective metric, which was defined as the reaction yield for the cross-coupling reaction. Similar to previous features design of photocatalysts, the 560 CNPs were encoded by 16 molecular descriptors that captured a range of thermodynamic, optoelectronic, and excited-state properties with input of quantum chemistry collaborators. To be thermodynamically viable in the reaction, the electron affinity (EA, equivalent to $E_{1/2}^{red}$[CNP/CNP$^-$]) and the exciton electron affinity (EA*, equivalent to $E_{1/2}^{red}$[CNP*/CNP$^-$]) of the CNP molecule must straddle the Ni(I) reduction and the Boc-Pro-OH oxidation potentials. Also, the optoelectronic and excited-state properties of the molecule may strongly influence its photocatalytic activity, which were also used to encode the chemical space. The molecular features include: (I) light absorption (first singlet excited state, $\Delta E_{S_1 \rightarrow S_0}$, together with the oscillator strength, f, of this transition); (II) excited-state charge distribution (change in dipole moment between S_1 and S_0, ΔD; degree of spatial extension of hole and electron distributions in the charge-transfer direction, H_{CT}, H index, t index); (III) excited-state charge separation (difference in the extent of spatial distribution between electron and hole, $\Delta\sigma$; electron-hole overlap, S_r; distance between the centres of the electron and hole, D index; Coulomb attraction between the electron and hole, E_C); (IV) the energy gap between the first singlet state and the first triplet state ($\Delta E_{S_1 \rightarrow T_1}$); and (V) the internal reorganization energy of CNP acting as a reductant for the Ni(I) compound.

4.3.2 Selecting and measuring *priori* samples

To implement Bayesian optimization and Gaussian processes, the *priori* knowledge need to be sampled systematically, because the *priori* in GPs is set to zero if there is no *priori* knowledge. Starting from the complete, unexplored chemical space, a small subset of six CNPs (molecular structures in Appendix A1) was selected across the feature space using the Kennard-Stone (KS) algorithm (Algorithm 4.1).[194] This sampling technique is a sequence method to collect N number of representation subset uniformly over the entire dataset based on the pairwise distance in the feature space. The first step started by choose two objects which have the longest distance between them. The following subsequent sample is added by computing the distance of a candidate object from the selected subset and requiring this distance to be the longest. The KS technique, which uniformly samples points in the search space, provides a general overview of the 'full picture'.

Algorithm 4.1 Kennard-Stone (KS) algorithm

```
**Input:** $k$, the number of samples to select
**Data:** $\kappa(\mathbf{x})$, distance matrix
**Output:** $A$, the selected subset
$D(x_i, x_j) = \max_{x_i, x_j \in \mathbf{x}} \kappa(\mathbf{x})$ // The init-selection;
$A = \{x_i, x_j\}$;
for $n = 2$ to $k$ do
    $\max\left(\min_{x_n \in \mathbf{x} \setminus A} \kappa(x_n, A)\right)$;
    $A+ = x_n$;
end
```

Data sampling by the KS algorithm is unique given a predefined number of samples to pick, yielding a deterministic starting set of points in the chemical space. These six CNPs were then synthesized and tested for the target cross-coupling reaction, forming step 0 in the optimization process. All CNPs were tested under identical reaction conditions: 4 mol% CNP photocatalyst, 10 mol% $NiCl_2$ glyme (glycol ether), 15 mol% dtbbpy (4,4-di-tert-butyl-2,2-bipyridine), 1.5 equivalents Cs_2CO_3 base, DMF, and blue LED irradiation source. All catalysis measurements were repeated three times; the resulting average reaction yield is reported here. The highest reaction yield achieved in step 0 was 39% for CNP-129, which combines Ra05 and Rb18. The yields achieved in step 0 gave confidence that some but not all the CNPs in the virtual library had the potential to facilitate a synergistic combination of photoredox catalysis and nickel catalysis for the target C(sp^2)-C(sp^3) cross-coupling reaction.

4.3.3 Identifying synthetic targets using Bayesian optimization

For target selection from the virtual library of 560 CNPs (Figure 4.1), a batched, constrained, discrete Bayesian optimization was used to explore the encoded chemical space of the CNPs, driving forward sequential experiments to improve the reaction yield. The BO scheme comprised two main steps: first, a surrogate model based on Gaussian processes was trained on all available observations; that is, measured reaction yields for all the synthesized CNPs at that point; second, a new set of CNPs was proposed for subsequent experiments, based on predictions by the surrogate model. This equates to the BO predicting the performance of candidate CNPs using available data and requesting new CNPs to be synthesized to verify its predictions. The parallel sampling strategy is an intuitive and inexpensive approach to proposing multiple points (forming a batch) in the search space, using a portfolio of acquisition functions favouring exploitation or exploration of the search space. This BO implementation follows closely that used previously in a robotic workflow used to find improved photocatalysts for hydrogen production.

The Bayesian optimization started by building a Gaussian-processes-based surrogate model using the six data points in step 0. The pre-calculated 16 electronic molecular properties are the feature space of the designed 560 CNPs. The Matérn kernel was used to combine the GPs algorithm with multiple length scales respective feature dimension and the smooth parameter $\nu = 2.5$. Since there were only 6 samples after step 0, the PCA was applied to reduce the dimension of features from 16 to 5 allowing GP model fitted properly at the beginning of the close loop optimization. The sum of the percentage of variance of the 5 dimensions is 0.9897 shows the reduced features still maintains the major variance of the original feature space.

After fitting the GPs model, based on the current, predicted mean and uncertainty, an acquisition function is required to hypothesize the most promising setting for the next experiment. To take the advantage of BO, the ideal condition is adding one sample sequentially by the acquisition function. Subsequent sampling of 12 points per optimization step was done using sets of 12 upper confidence bound (UCB) functions (Equation 1.19), a weighted sum of the posterior mean $\mu(x)$, and uncertainty $\delta(x)$, controlled by a hyper-

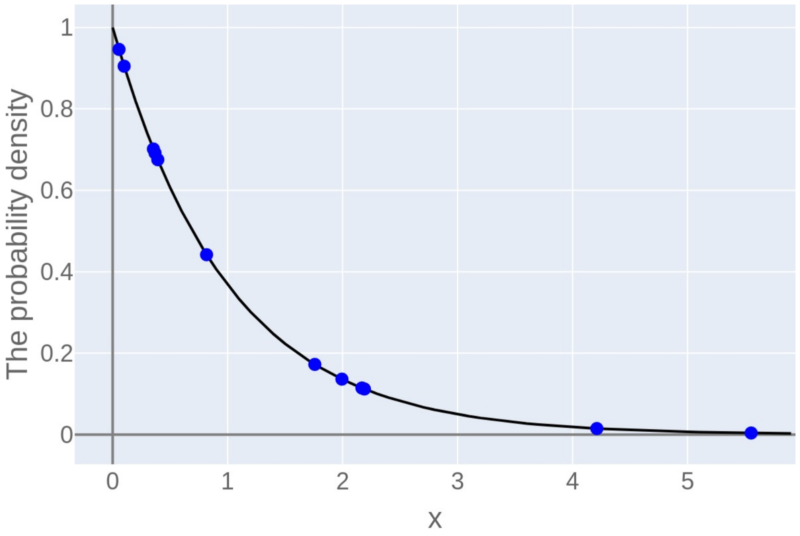
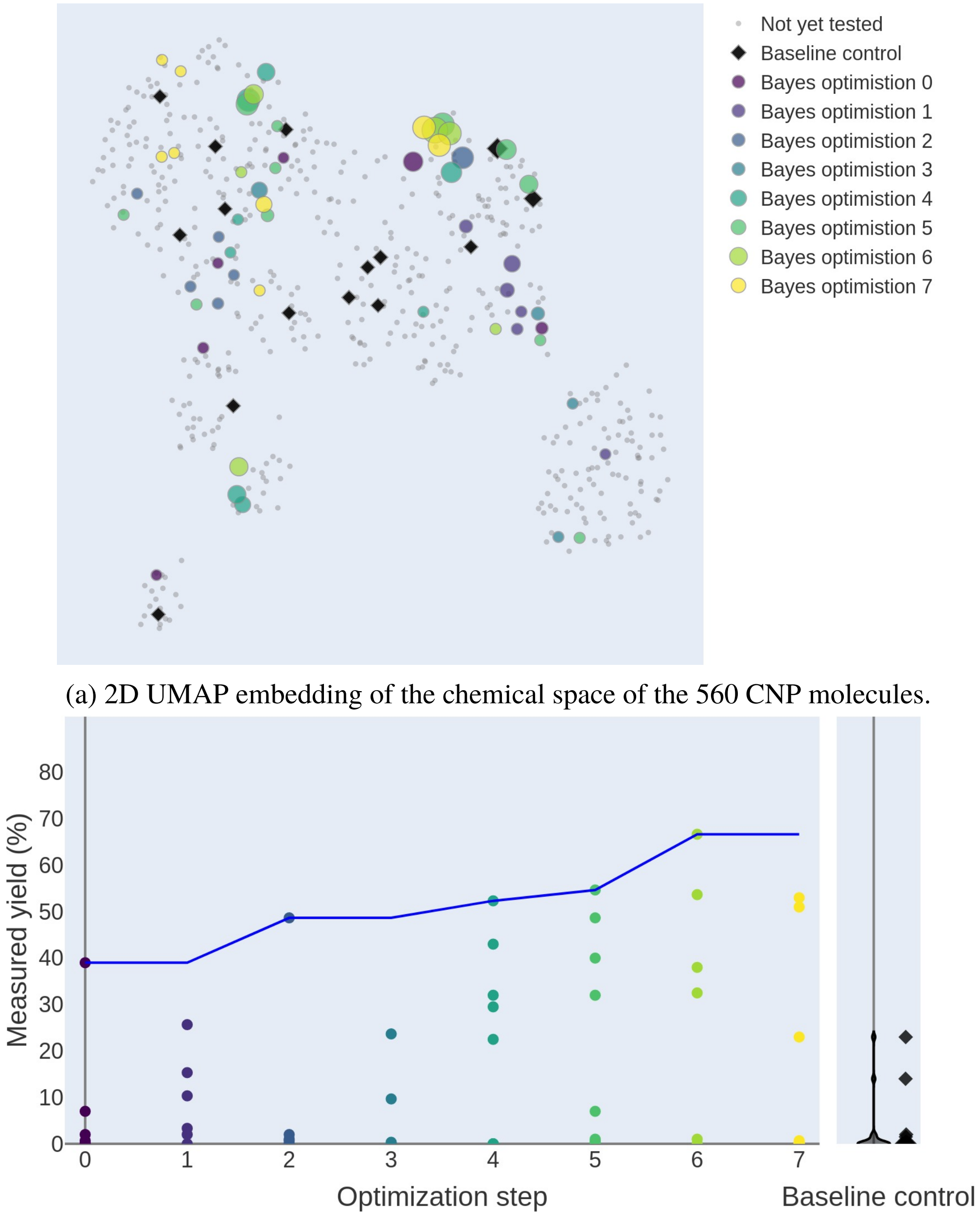


Figure 4.3 The value and density of 12 samples(blue points) on an exponential distribution(the black curve).

parameter β. For each step, the set of 12 β values was generated on a random exponential distribution (Figure 4.3), with small β values favouring predicted high performance, $\mu(x)$, or exploitation, and large β values favouring high uncertainties, $\delta(x)$, or exploration. From these 12 BO-proposed CNPs, a subset of around 6-8 CNPs per suggested batch were selected to do experiments, ensuring that the selected CNPs exploited a trade-off between exploitation and exploration, ensued by their different β values. The experimental collaborator can choose 2-3 samples from either the high β suggestion or low β suggestion, which allows for intuitive biasing toward exploitation or exploration of the search space by assigning different β values to the acquisition functions of different BO instances at each step. This protocol of joint decision-making for candidate CNPs selection combines both BO and insight from the chemist.

By integrating the UMAP method and the encoded electronic feature space, 560 CNPs molecules are projected to a 2D space. The synthesized CNP molecules are colour-coded by experimental batches, using the same colour scheme as in Figure 4.4(b), the molecules that were not synthesized are coloured in grey Figure 4.4(a). Symbol size denotes the experimentally measured reaction yield for the target reaction. The black points refer to a baseline control experiment conducted for a set of 15 molecules chosen in a way that



(a) 2D UMAP embedding of the chemical space of the 560 CNP molecules.

(b) Measured yield for the target cross-coupling reaction plotted against the experiment batch

Figure 4.4 Targeted synthesis of CNPs for organic photoredox catalyst discovery. The CNP molecules synthesized in this work are colour-coded by experimental batches in (a), using the same colour scheme as in (b); the molecules that were not synthesized are coloured in grey. Symbol size in (a)denotes the experimentally measured reaction yield for the target reaction shown in (Figure 4.2). The blue line in (b) represents the yield of the highest selected sample at different steps. The highest yield attained after 8 batches (Optimization steps 0-7) was 67%. The black points refer to a baseline control experiment conducted for a set of 15 molecules chosen in a way that maximized the structural diversity of the set.

maximized the structural diversity of the set. Seven such batches (Figure 4.4) resulted in a total of 49 additional CNPs that were synthesized and tested. The number CNPs tested by experiment was 6, 6, 4, 8, 11, 6, and 8 in steps 1 to 7, respectively (molecular structures in Appendix A1). The highest reaction yield attained increased from 39% at step 0 to 67% by step 7 (Figure 4.4(b)), which was achieved in step 6 using CNP-127 (Figure 4.5). The iteration was stopped after these seven optimization steps because a yield of 67% was considered acceptable in the absence of reaction condition optimization.

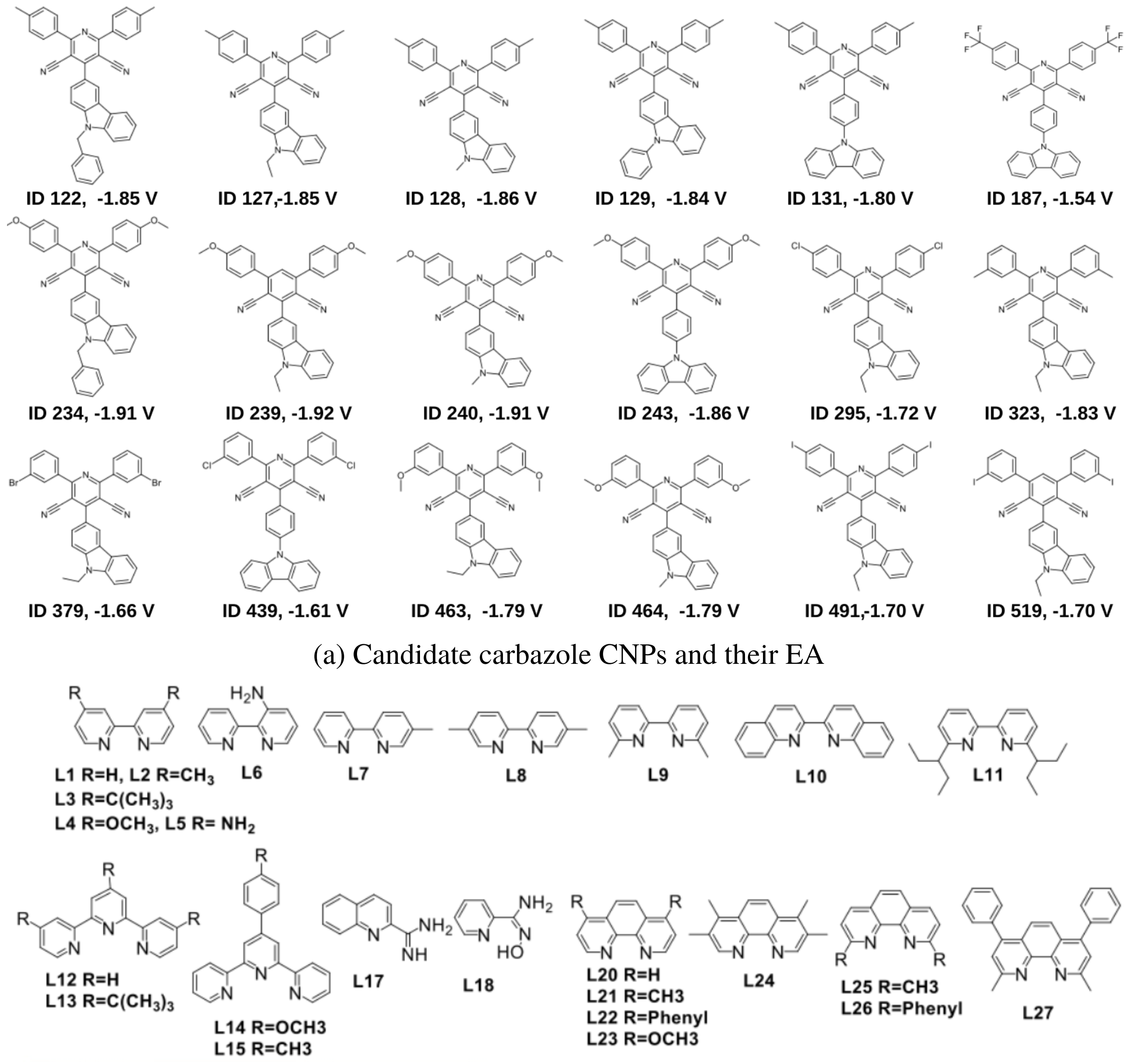
One might ask whether the 'sweet spots' discovered by the BO search within this virtual library of 560 CNPs constitutes a global optimum, or at least close to one. It could not be guaranteed without synthesizing the entire library, which was impractical. However, to probe this further, a diverse set of 20 CNPs was picked from the 2D structural space of the 560 CNPs encoded by Morgan fingerprints, using the KS algorithm, as was used to pick the CNPs for step 0. Of the selected 20 CNP samples, three were un-synthesizable within the time available and two (CNP-459 and CNP-244) were already picked up by the BO algorithm; as such, 15 additional CNPs (molecular structures in Appendix A1) were synthesized and tested for the photoredox reaction (black points in Figure 4.4(a)). The results obtained were in line with the structure-activity relationship summarized for the CNP samples arising from the BO search. The CNPs with Rb moieties, Rb04, Rb06, Rb08, Rb13 and Rb15, which were not explored by BO search, showed no or little activity (Yield < 3%). The highest yield attained by the molecules in this structurally selected baseline control set was 32%.

4.4 Close loop discovery of the reaction conditions

4.4.1 Encoding reaction conditions

After identified high performance CNPs from the first BO search, a similar Bayesian strategy was set out to optimize the reaction overall conditions. The target decarboxylative C(sp^3)-C(sp^2) cross-coupling reaction requires a photoredox catalyst (in this case CNPs) and an organometallic nickel catalyst. As discussed above, these two catalysts must work synergistically in completing two interwoven catalytic cycles. The diffusional electron transfer between CNP$^-$ and CNP^1 not only depends on the thermodynamic driving force but also on the concentration of Ni catalysts and Ni ligands. The maximum observed yield of 67% was achieved with fixed Ni loading and Ni ligands. Varying the concentrations of the Ni catalyst (10 mol% or 1 mol%) affected reaction yields for the target cross-coupling reaction markedly.

To further optimize the reaction yield, three key variables were studied: (I) the choice of CNP photocatalyst (Figure 4.5(a)); (II) the concentration of the Ni catalyst, and (Figure 4.5(b)); (III) the choice of the Ni-coordinating pyridyl ligands. Here, 18 carbazole-containing CNPs that were selected in the first BO (Figure 4.4), were chosen to exhibit widely varying catalytic performance. The reason of choose this range of carbazole catalysts, rather than simply the best material from the first BO selection (CNP-127), since it was initially unclear that this catalyst would also be optimal at all Ni concentrations and with all Ni ligands. A range of 25 pyridyl compounds was selected with different coordination environments and different molecular shape, size, and degree of bulkiness. Because these three dimensions comprise both discrete and continuous features, the conventional gradient optimization of the acquisition function is incompatible with these designed key variables when applying the BO optimization. Thus, regarding the continues variables, the concatenation of the NI catalyst was studied between 1 mol% and 10 mol%, with 1



(a) Candidate carbazole CNPs and their EA

(b) Candidate pyridyl ligands

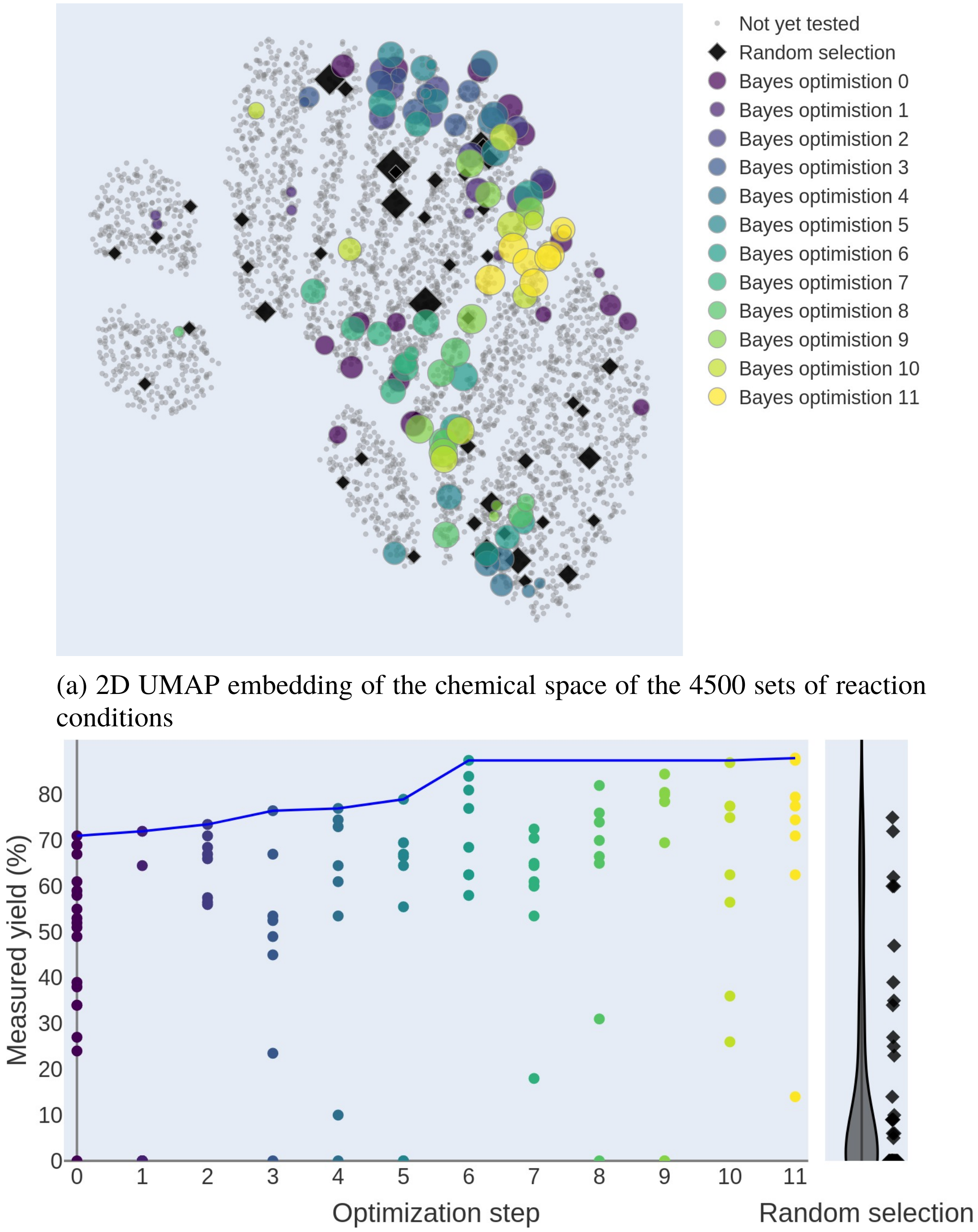
Figure 4.5 A total of 18 candidate carbazole CNPs and 25 candidate pyridyl ligands were considered in these experiments

mol% intervals, resulting in 10 distinct Ni concatenation values. These three variables gave rise to a total of 4500 (18 × 25 × 10) unique potential experiments. After this transformation, the gradient optimization is no longer required, and the acquisition function's value is evaluated at each sampling point to determine the optimal solution.

Then, the 4500 sets of reaction conditions were encoded into a chemical space as follows. First, each combination of a CNP, a pyridyl ligand, and a Ni concentration was encoded by the concatenation of (I) the experimentally measured reduction potential $E_{1/2}^{red}$[CNP/CNP$^-$] (labelling in Figure 4.5(a) under each CNPs), a measure of the reducing ability of the CNP$^-$ species to regenerate Ni(0); (II) the Morgan fingerprint of the CNPs; (III) the Morgan fingerprint of the pyridyl ligand, and; (IV) the concentration of the Ni(II) source. Second, the distance between two sets of reaction conditions was given by a combined distance from these four encoding elements, which is the summation of (I) the scalar difference between the reduction potentials; (II) the Tanimoto distance between the CNPs' fingerprints (radius = 2; 2048 bits); (III) the Tanimoto distance between the pyridyl ligand fingerprints, and; (IV) the scalar distance between the Ni concentrations. All four component distances were normalized before being added together to give the combined distance. Last, the chemical space encoding the 4500 sets of reaction conditions took the form of a 4500 × 4500 distance matrix, containing pairwise distances for all sets of reaction conditions. Figure 4.6(a) shows the resulting chemical space as a 2D UMAP embedding of the distance matrix.

4.4.2 Identifying the best reaction condition by Bayesian optimization

Similar with the first close loop optimization in chapter 4.3.2, the initial 19 sets of reaction conditions (step 0, Figure 4.6(b)) were selected by the KS algorithm as the *priori* knowledge of this optimization, and then they were used to training Gaussian process-based surrogate models to suggest the first batch of experiments (step 1) in the optimization



(a) 2D UMAP embedding of the chemical space of the 4500 sets of reaction conditions

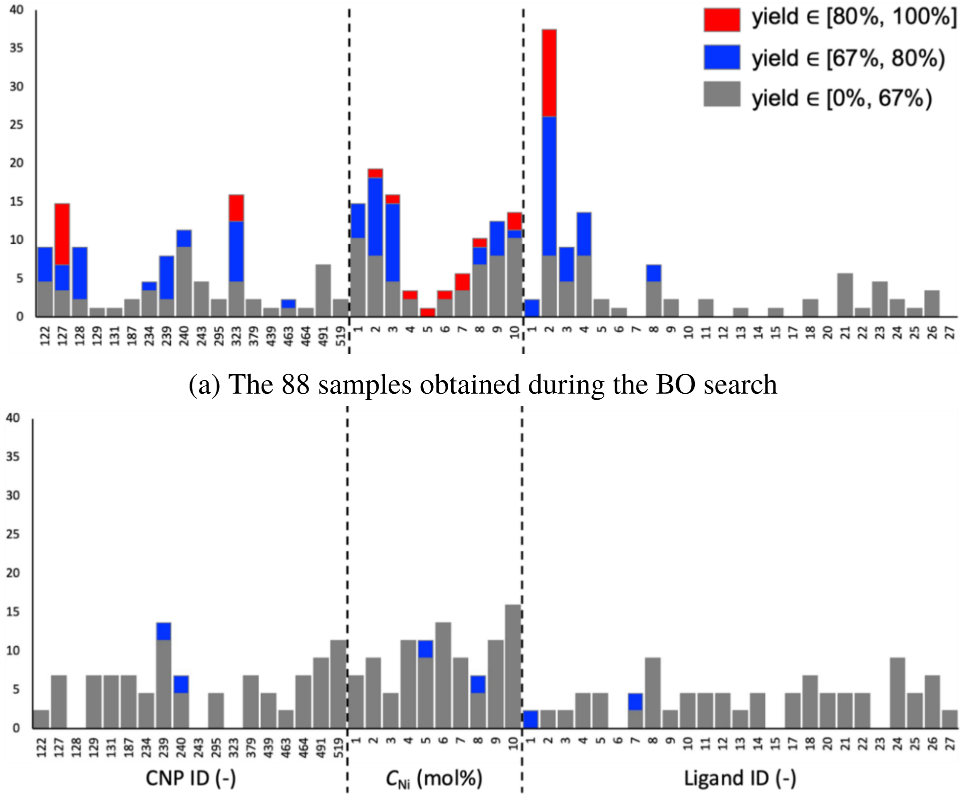
(b) Measured yield (average of 3 repeats) for the target cross-coupling reaction plotted against the experiment batch, optimization steps or baseline

Figure 4.6 Reaction condition screening of CNPs, pyridyl ligands, and the amount of Ni. The tested set of conditions are colour-coded by experimental batches in (a), adopting the same colouring scheme as in (b), with all the untested conditions coloured in grey. Symbol size denotes the experimentally measured reaction yield. Number of samples: 19 samples at step 0; 8 samples at steps 1-11 each; 44 samples were included in a random selection as a baseline (black points). The blue line in (b) represents the yield of the highest selected sample at different steps.

workflow. The same BO parallel sampling approach was used as for the synthetic candidate selection workflow (Figure 4.4). Eight samples were acquired at each BO step, covering a portfolio of upper confidence bound functions with varying degrees of balance between exploitation and exploration. From step 0 to step 6, the maximum reaction yield achieved at each step continuously increased from 71% to 88%. No further improvement in the maximum yield was attained in the subsequent five steps (40 reactions); the optimization was therefore terminated at step 11 having evaluated 88 sets of reaction conditions. The highest yield achieved during the 11 BO steps was 88%; this occurred when CNP-127 (Figure 4.5(a)) was used at 2, 4, or 5 mol% Ni concentration, in all cases with the ligand (L2 in Figure 4.5(b)). In step 0, CNP-239 (Figure 4.5(a)) was the highest-performing photocatalyst, reaching a yield of 71%, together with the ligand L3 in Figure 4.5(b) and a 1 mol% Ni concentration. As such, CNP-127 is ‘rediscovered’ in this second BO search, but the reaction conditions are re-optimized to improve the photocatalytic yield significantly (from 71% to 88%).

4.4.3 Comparison between random and algorithm searched conditions

For baseline comparison, 44 sets of conditions for catalysis measurements were randomly selected (Figure 5d); only two sets attained a yield above 67% (the blue bars in Figure 4.7(b)): these reactions gave yields of 75% (CNP-240, L1, and 5 mol% Ni) and 72% (CNP-239, L7, and 8 mol% Ni), respectively. By comparison with the BO-acquired samples (Figure 4.7), a more uniform sample distribution was generated by random sampling for the candidate CNPs, Ni concentrations, and pyridyl ligands. The BO search markedly outperformed the random sampling, attaining a higher maximum reaction yield (88% versus 75%). Also, the BO method gave a much larger proportion of high-activity samples; for example, 39/88 reaction conditions (44%) gave yields of more than 67% for the BO



(a) The 88 samples obtained during the BO search

(b) The 44 samples obtained by random selection

Figure 4.7 Histograms of measured reaction conditions over three different variable ranges: candidate CNPs(CNP ID), Ni concentrations (C_{Ni}), and pyridyl ligands(Ligand ID). These histograms were calculated separately for each value range in either the BO set or the random set. The bars are coloured red, blue, or grey to indicate a yield greater than 80%, 67-80%, or lower than 67%, respectively.

search whereas just 2/44 (4.5%) of conditions gave a comparable yield in the random selection. This shows that BO explores the high-performing areas of the chemical space much more effectively than random sampling.

Overall, the 88 sets of conditions requested by the BO algorithm covered all the 18 CNPs (that is, each was selected at least once), all the 10 Ni concentrations, and 18 out of the 25 available pyridyl ligands (Figure 4.7(a)). Only 10 out of the 88 sets of conditions gave a reaction yield greater than or equal to 80% (the red bars in Figure 4.7(a)), all of which involved CNP-127 or CNP-323 as the photocatalyst and L2 as the Ni-coordinating ligand. It should be noted that CNP-323 is a structural isomer of CNP-127 (Figure 4.5(a)).

4.5 Analysing structure-properties relationships

To gain insight into the relationship between calculated descriptors and measured catalytic activities, multiple machine learning algorithms was evaluated (Table 4.1). These models were trained for regression task based on the average yield of each catalyst reaction in first BO search. In the training and validation process, 70 objects (55 samples from steps 0-7 and 15 samples from the baseline control) were utilized, with 16 electronic features that had been scaled being used as the training dataset. The kernel function was the same as the one used during the BO search for kernel methods. However, due to the small size of the training set, the application of an MLP model was still limited.

Table 4.1 Regression metrics across different models of catalysts, obtained by 5-fold and leave-one-out(LOO) cross-validation procedures

	LOO		5-fold	
Model	R2	MAE	R2	MAE
GP	0.906	0.046	0.863	0.051
KRR	0.912	0.041	0.844	0.053
GBRT	0.820	0.059	0.824	0.059
RF	0.818	0.065	0.767	0.072
SVR	0.801	0.070	0.633	0.095

The 5-fold cross validation results showed the GP model has the best performance and second strongest in leave-one-out (LOO) validation with a slightly inferior performance to the KRR model. These results show that the descriptors can predictively assign the cross coupling reaction activity under a homogenous reaction condition. It is not surprised that GP outperformed other machine learning tasks since the training set was suggested by BO optimization with GP as the surrogate model. This could introduce some systemic bias to reduce the error of GP model.

Furthermore, similar tests were also performed for the reaction condition task with the pre-defined RBF kernel function (Equation 4.2) as the input and measured reaction yield as the target during the reaction condition optimization. Due to the limitation of encoded descriptors (fingerprints for molecular and ligands), only kernel methods were tested here in Table 4.2. The parity plots

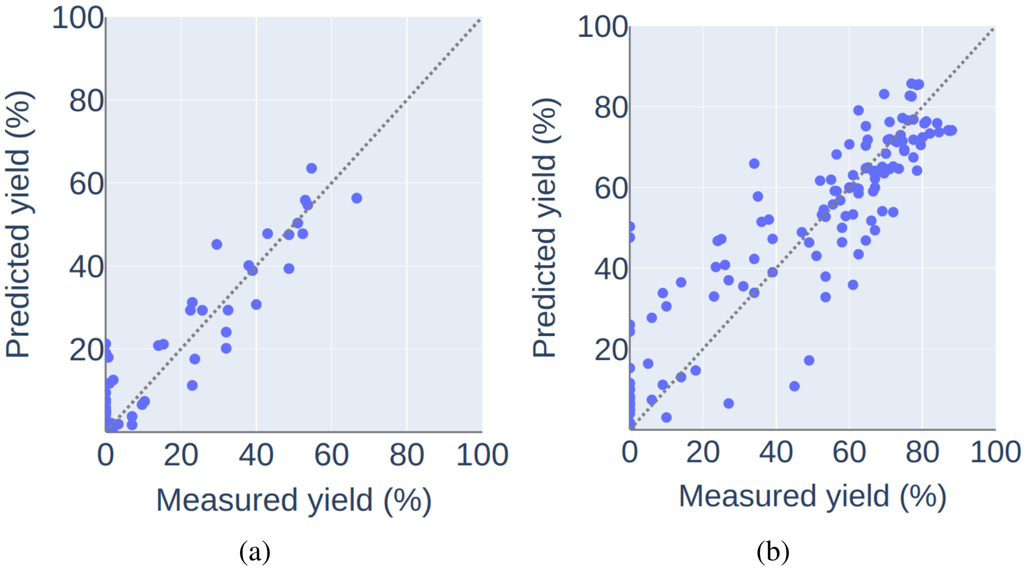
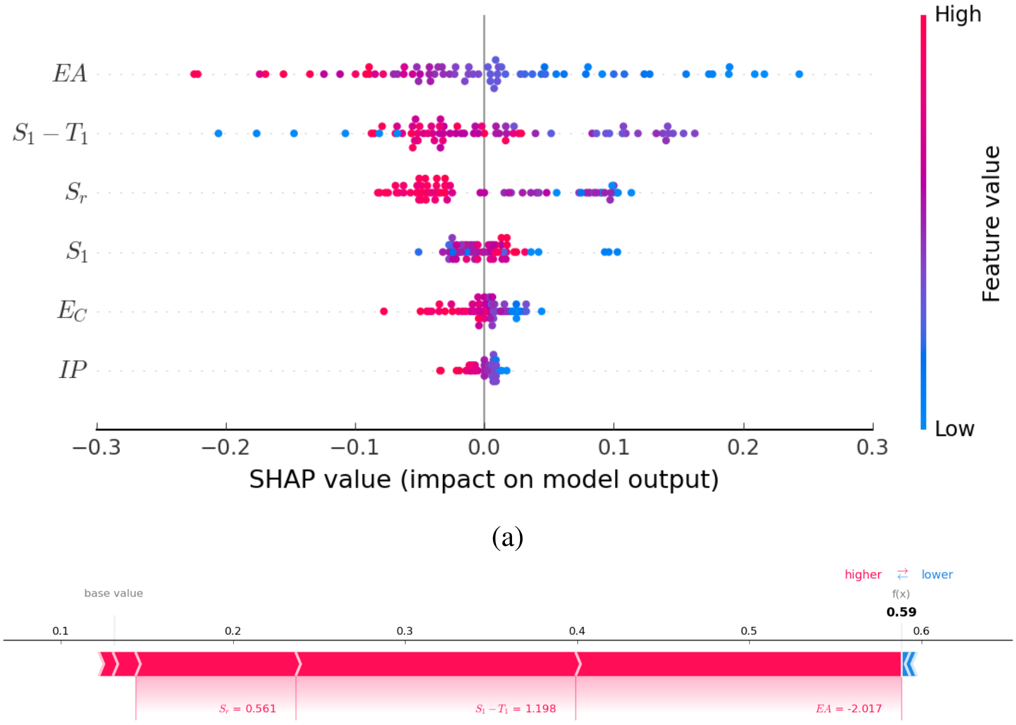


Figure 4.8 Leave-one-out(LOO) validation of regression task on different experimental dataset. (a)Evaluating on the optimization of organic photocatalysts data by GP. (b)Evaluating on the optimization of reaction condition data by KRR.

Table 4.2 Regression metrics across different models of conditions, obtained by 5-fold and leave-one-out(LOO) cross-validation procedures

	LOO		5-fold	
Model	R2	MAE	R2	MAE
GP	0.773	0.114	0.673	0.140
KRR	0.836	0.091	0.715	0.123
SVR	0.765	0.114	0.677	0.134

The aforementioned result instils confidence in further exploring the relationships between features and properties in greater detail. To make the model explainable, a machine learning interpretability technique, SHapley Additive exPlanations(SHAP),[195] was introduced and assigned each feature a Shapley value. The Shapley values are computed by comparing the model's predictions with and without the feature present, over all possible feature combinations. The resulting Shapley values are then used to produce a feature importance ranking, showing the relative importance of each feature in the model's predictions.

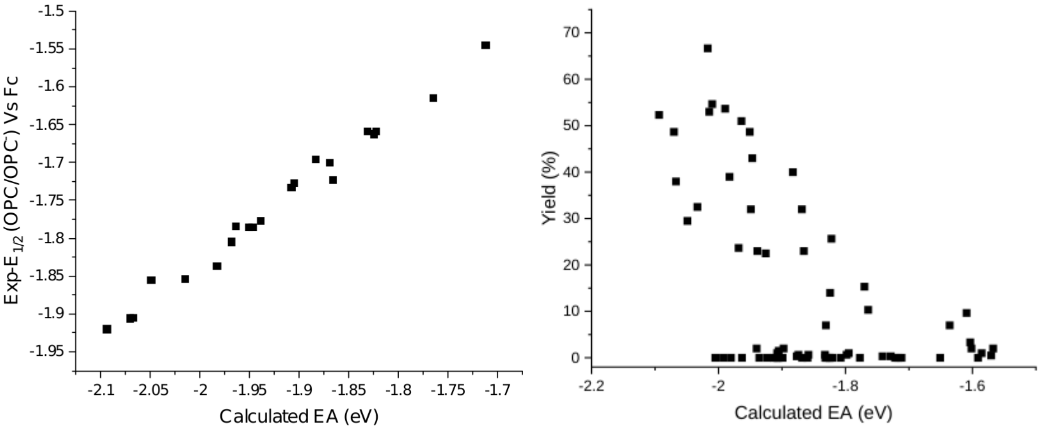


(a)

(b)

Figure 4.9 The SHAP explanation of feature importance. (a)A beeswarm plot of top 6 important features, showing the distribution of SHAP values for each input feature across all instances in the dataset. (b)A force plot explaining the ML model's prediction for the best-forming catalyst, CNP-127, showing how each input feature contributes to the prediction. Each feature's contribution is represented by an arrow, with the length of the arrow proportional to the magnitude of the SHAP value. Red arrows pointing to the right indicate positive contributions, while blue arrows pointing to the left indicate negative contributions.

By fitting the molecular features with GPs model, the SHAP values were evaluated for all training samples in Figure 4.9(a) to get an overview of the features' importance. Features were sorted by the sum of SHAP value magnitudes over all samples, and colour coded by the feature value in Figure 4.9(a). The feature contribution of the best CNP catalyst (CNP-127) in Figure 4.9(b) shows both EA, $\Delta E_{S_1 \to T_1}$ and S_r contribute to push the prediction ($f(x)$) higher. The two most important features (EA, and $\Delta E_{S_1 \to T_1}$) proved again the discovery of feature importance of photocatalyst in chapter 3.3.3. Besides, The EA has negative relationship with the SHAP value in Figure 4.9(a), which indicates the potential correlation with the cross-coupling reaction activity. In general, donating groups (on Ra) raise the reduction potential to more negative energy level, increasing the driving force of the electron transfer from the reduced CNP$^-$ to Ni catalysts, a key step that couples the PC and nickel catalyst catalytic cycles.



(a) Linear correlation between calculated EA and experimentally measured $E_{1/2}^{red}$[CNP/CNP$^-$]

(b) Reaction yield versus calculated reduction potential for the experimentally measured CNPs

Figure 4.10 The correlation between reduction potentials and reaction yield

The weak negative correlation between EA of CNPs and the SHAP values suggests that it may be possible to search for CNPs with more negative calculated reduction potentials in the virtual library of 560 CNPs. CNP-239 and CNP-234, were both explored experimentally by BO but did not show the best performance under the explored conditions, despite having the strongest reduction potentials in the library (-1.92V and -1.91V in Figure 4.5(a), respectively). Following this, a computationally-led search of 100 additional CNP molecules (in Appendix A2) comprising donating Ra groups and carbazole Rb groups was performed. It was found that CNP-624 has a calculated reduction potential of -2.17 eV vs SHE. However, no activity was observed for the cross-coupling reaction using CNP-624. In addition, it is worth to emphasize here that the correlation of reduction potential with yield observed fails to apply to the whole CNP molecules (Figure 4.10(b)). Again, these observations indicate that the performance of CNPs is determined by a range of factors, rather than a single photophysical feature, rationalizing the use of a BO-led search strategy rather than more classical computational design.

4.6 Methods of Bayesian optimization and machine learning

Bayesian optimization (BO) is a sequential hypothesis testing approach to global optimization of 'black-box' functions, i.e., functions that do not have a closed-form representation and does not provide function derivatives, thus only allowing for point-wise evaluation. Here, it equates to finding the highest reaction yield in the chemical space of CNPs or reaction conditions.

Gaussian processes were used as the surrogate model, together with the Matérn similarity kernel. A Gaussian process maintains a belief over the design space, by simultaneously predicting the mean, and the uncertainty, at any point in the input space, given existing observations. To hypothesize the most promising setting for the next experiment, based on the current, predicted mean and uncertainty, an acquisition function is required; here, the upper confidence bound (UCB) function was used, which is given by

$$f_{UCB}(x) = \mu(x) + \beta\delta(x) \tag{4.1}$$

where $\mu(x)$ is the posterior mean, $\delta(x)$ is the uncertainty, and β is a hyperparameter. For each optimization step, the highest value of the acquisition function (Equation 4.1) was used as the next experimental suggestion.

The syntheses and photocatalytic measurements of CNPs were time-consuming, but were amenable to parallelization; that is, they could be made and tested in batches. To facilitate an efficient parallel search, a batched, discrete BO approach was adopted; that is, multiple BO instances were run in parallel, all using the same existing observations and contributing to the subsequent steps. Here, a set of 12 BO instances were run at each optimization step. This parallel sampling strategy allowed for intuitive biasing toward exploitation or exploration of the search space by assigning different β values to the acquisition functions of different BO instances at each step. Small values of β prioritized areas where the mean was expected to be largest (i.e., exploitation), while large values prioritized areas where the model was most uncertain (i.e., exploration); a random exponential distribution function was used to generate β values within a batch.

For optimization of reaction conditions, a customized RBF kernel was defined as follows:

$$\kappa(P_i, P_j) = \alpha \exp -(\theta_1 d_{EA}(M_i, M_j)^2 + \theta_2 d_{fps}(M_i, M_j)^2 + \theta_3 d_{fps}(L_i, L_j)^2 + \theta_4 d_{Ni}(C_i, C_j)^2) \tag{4.2}$$

where P_i and P_j are two sets of reaction conditions, each involving a CNP molecule (M_i or M_j), a Ni-coordinating ligand (L_i or L_j), and a Ni concentration (C_i or C_j). $d_{EA}(M_i, M_j)$ refers to the Euclidean distance between the values of electron affinity (EA) of the two CNPs (M_i and M_j), $d_{fps}(M_i, M_j)$ refers to the fingerprints distance of CNPs, $d_{fps}(L_i, L_j)$ represents the fingerprints distance of pyridyl ligands, and $d_{Ni}(C_i, C_j)$ is the Euclidean distance between the values of Ni concentration of the two reaction condition sets. Four scaling hyperparameter θ_1, θ_2, θ_3, and θ_4 regulated the relative weighting of the four distances and were tuned during the training of Gaussian processes. The RBF kernel is a robust kernel to fit the measured experimental yield without introducing too much noise section during the fitting of GP model. All four component distances were normalized before being added together to give the combined RBF kernel. Last, the chemical space encoding the 4500 sets of reaction conditions took the form of a 4500 dimensions distance matrix, containing pairwise distances for all sets of reaction conditions.

The SHapley Additive exPlanations (SHAP) algorithm to provide explanations for the machine learning model predictions. We used the python implementation of SHAP, version 0.35.0, available via the conda-forge channel (https://anaconda.org/conda-forge/shap).

SHAP combines game theory with local explanation, enabling accurate interpretations on how the model predicted a particular value for a given sample. The explanations are called local explanations and reveal subtle changes and interrelations that are otherwise missed when these differences are averaged out. Local explanations allow the inspection of samples that have extreme phenotypes values (e.g., a high or low photocatalytic reactivity).

4.7 Conclusion

A Bayesian optimization strategy was used to identify promising OPCs from a virtual library of 560 candidate molecules while exploring a small fraction of the available chemical space (55 / 560 organic photoredox catalysts in the first BO search; 107 / 4500 reaction conditions in the second reaction condition optimization). This identified OPCs with reaction yields for a cross-coupling reaction of up to 88% that could match iridium catalysts at high nickel concentrations and outperform iridium catalysts at lower nickel concentrations. BO is a promising approach for the discovery of metallophotocatalyst formulations, and by extension for other research challenges where there is a large search space and limited prior knowledge. The Bayesian optimization approach also has some limitations; for example, the Hanztsch synthesis is broadly generalizable, but it is not ubiquitous for all combinations of Ra and Rb functionalities. This was addressed by fusing BO algorithms with human decisions as to which molecules to pursue in each batch. It would be desirable to fully automate such workflows, but this would require some technical developments; for example, to carry out trial syntheses for candidate OPCs and to make autonomous decisions about which OPCs to carry forward into catalysis testing.

CHAPTER 5 EXTENDING THE APPLICABILITY OF THE ANI POTENTIAL TO INTERMOLECULAR INTERACTIONS

5.1 Introduction

The application of machine learning can also be success in the area of quantum property predication to achieve higher accuracy predication of such interatomic information and maintain a computational cost comparable to classical force field.[8,28] Such ML based force field techniques can calculate molecular atomization energies, forces, potential energy surfaces(PES), and even including atomic partial charges and dipoles.[196] The speed, accuracy of these model are beneficial by the rapid development of modern machine learning algorithms and the computing hardware such as deep learning and accelerated calculation by GPU.

One of the successful application is the ANAKIN-ME (ANI) method[124] for building transferable machine learning potential using symmetry function as the atomic descriptor (Behler and Parrinello-type descriptors[34]). The original ANI-1[124] model was developed by training 4 multilayer neural networks on a dataset of 22 million randomly selected conformers of small HCNO-only organic molecules. The reference force and energy information were calculated using the wB97X/6-31G* DFT level. Tests of molecular dihedral rotation and bond stretch demonstrated that ANI-1 accuracy is superior to that of two popular semi-empirical methods (DFTB and PM6) and comparable to the reference data level. Subsequently, several future studies extended the applicability of ANI by increasing its accuracy to a higher level (ANI-1ccx)[35] through transfer learning and by adding new molecules to the training dataset to support additional elements, such as sulfur and halogens (ANI-2x).[197] The development of a fast and transferable machine learning potential model holds great potential for the prediction of crystal structures in molecules, particularly in the context of crystal structure prediction (CSP).[145] This is because the current methodology employed in our research group for predicting molecular crystal structures is limited to rigid organic molecules. Additionally, the simulated CSP datasets often consist of thousands of polymorphs, making it computationally challenging to perform high-accuracy simulations for each individual structure. However, calculating the lattice energy of the predicted CSP dataset using ANI-2x and ANI-1 model failed due to the lack of intermolecular interaction features in these models.

Thus, this study aims to use similar methodologies to extend the applicability of ANI to describe intermolecular interactions, such as hydrogen bonding. The TorchANI software,[198] which was recently released, presents a standardized approach for creating and training new ANI models with PyTorch, which is a valuable tool for researchers. To create additional training data, the S66a8 dimer dataset[199-200] was sampled to generate new geometries, then these structures were calculated under DFT level with dispersion correction. Expect this, modifications were introduced to TorchANI's workflow to facilitate this process. By augmenting the symmetry function's hyperparameters and incorporating the new dimer dataset to account for intermolecular interactions, the newly trained ANI model can capture the desired information, although improvements are still required to enhance the accuracy and expand the range of applications of this model.

5.2 The structure of ANI model

The atomic centred symmetry functions, developed by Behler and Parrinello in 2007,[34] was used in ANI as the atomic environmental descriptor. The original design of ACSFs introduced a given centre atom with a cutoff radius R_C to describe the atomic neighbour environment, so that it can avoid the permutation and symmetry problems of 3D coordinates. The restriction of using a cutoff reduces the computational requirement for large chemical systems. Besides, the atomic environment represented by pairwise distances in the cutoff range is invariant when doing symmetry operation. A frequently used cutoff form in symmetry function is the decaying of cosine function

$$f_c(R_{ij}) = \begin{cases} 0.5 \times \cos(\frac{\pi R_i j}{R_C}) + 1 & R_{ij} < R_C \\ 0 & R_{ij} > R_C \end{cases} \tag{5.1}$$

where R_{ij} is the distance between centre atoms i and neighbour j, whereas R_C is the cutoff radius boundary. If R_{ij} is larger than R_C the cutoff function and its derivative become zero. The cutoff function must decay to zero at cutoff radius and be differentiable in all region since the atomic force is calculated by derivative of energy.[201] Despite the cosine function, other cutoff function can be used here if they match the above requirement, such as hyperbolic tangent function (Tanh), exponential function, and polynomial function.

To simulate atomic radius interaction behaviour, the radial function G_m^R(Equation 5.2) is designed to slope at the cutoff radius by multiplying one or more cutoff functions $f_c(R_{ij})$(Equation 5.1). Therefore, atoms beyond the cutoff distance are not counted in the atomic energy calculation. Function G_m^R below is a sum of Gaussian multiplied by cutoff functions for all neighbouring atoms

$$G_m^R = \sum_{j \neq i}^{all\ atoms} \exp^{-\eta(R_{ij}-R_S)^2} f_c(R_{ij}) \tag{5.2}$$

where parameter η control the width of the Gaussian decaying and extending to the cutoff radius. The centre of the Gaussian is shifted to a certain distance from the central atom by parameter R_S, which forms a diffusion sphere around the central atom to describe neighbour atoms only located at the predefined distance from the central atom (Figure 5.1). The summation over the Gaussian functions provides a single value to describe the environment of the centred atom i, so that independents with the number of neighbouring atoms within the cutoff distance R_C. This feature fixes the vector size of ACSFs satisfying one

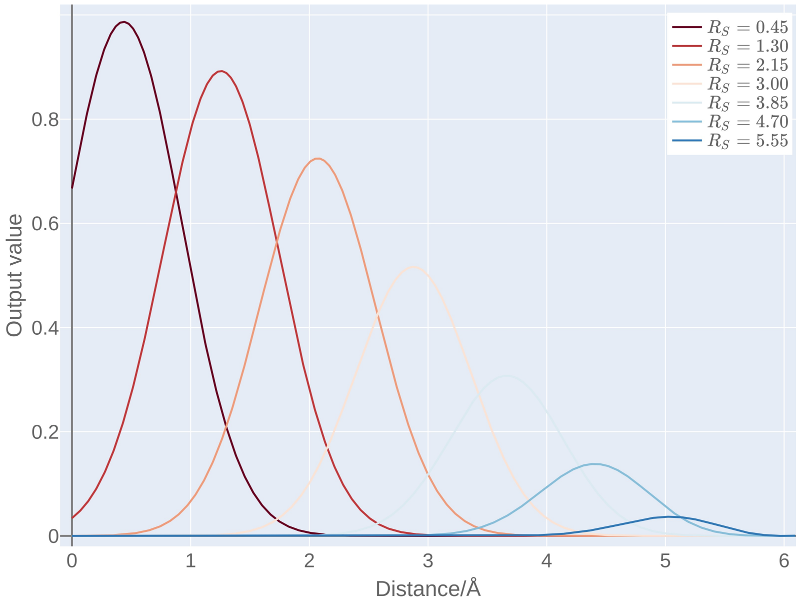


Figure 5.1 Examples of the radial function (Equation 5.2) with different parameter of R_S. $R_C = 6$, $\eta = 2$

of important requirements in machine learning algorithms. In a real application, multiple parameter R_S are defined to ensure a reasonable resolution around the radial distance. The size of a radial section of a ACSFs is $N(species) \times N(R_S) \times N(\eta)$ so that single η is used to reduce the vector to a reasonable size.

The radial functions are constructed only for two-atom environments whereas multi-atom environments, such as the difference between tetrahedral and square planar position, are not distinguished if the neighbour has same distance. Therefore, the angular environment also need to be captured by ACSFs as neighbour atoms are spread in 3D space. Following function G_m^A is defined as the angular function, a sum over all cosines multiplied by Gaussian of the interatomic distances in the triplet atoms and the respective cut off functions,[124]

$$G_m^A = 2^{1-\zeta} \sum_{j,k\neq i}^{all\ atoms} (1 + \lambda \cos(\theta_{ijk}))^{\zeta} \exp[-\eta(R_{ij} + R_{ik} + R_{jk})^2] f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{5.3}$$

where given atoms ijk selected to measure their angular θ_{ijk} centred on atom i and paired distance R_{ij}, R_{ik},and R_{jk}. The Gaussian parameter η control the width of the Gaussian peaks in the angular part. For complex chemical structures, different values of η is used to form several shells around the central atom as same as in radial function (5.2). The exponent parameter ζ defines the angular resolution in 3D space, which normally needs multiple values to represent environment. The angular degree parameter λ, is used to centre the pick of cosine terms at either 0 or 360 degree by +1 and -1.

Equation 5.2 and 5.3 are determined by several parameters that dictate the spatial shape of the ACSFs, making it essential to carefully select the appropriate values for these parameters. To systematically generate the necessary parameters and select suitable functions for a given dataset, optimization approaches can be employed. Alternatively, one may opt to use default ACSFs definitions based on previous experience. While the former method can provide an accurate description of the dataset, the latter method aims to create an unbiased ACSFs function.

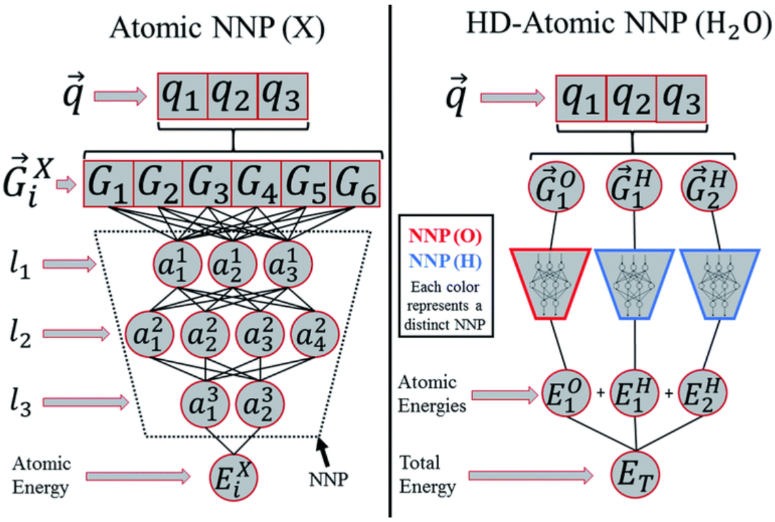


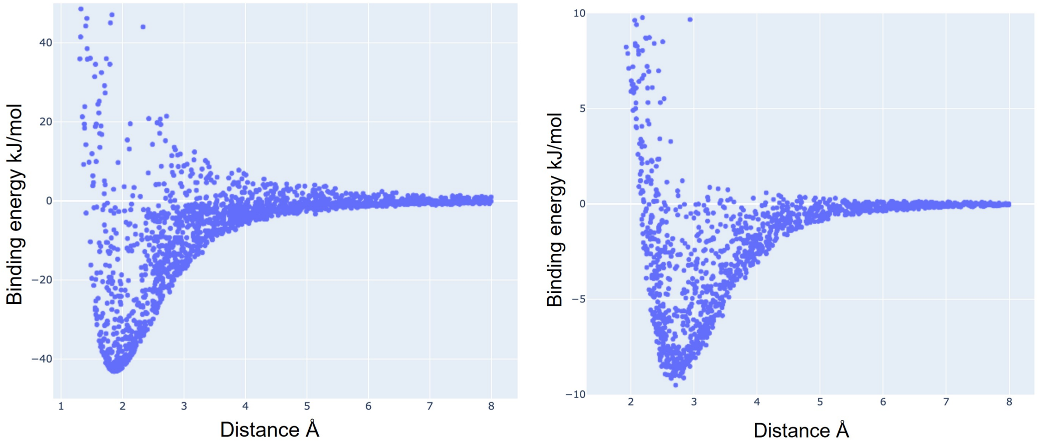
Figure 5.2 Examples of the ANI neural network structure from the ANI paper.[124] Copyright 2017 Royal Society of Chemistry. The left one represents the algorithmic structure of an atomic number specific neural network potential, while the right-hand side illustrates the high-dimensional atomic neural network potential (HD-NNP) model for a water molecule. The input atomic coordinates $\vec{q}$ are used to calculate their ACSFs vector $\vec{G}_i^X$ for atom i with element X.

Once the atomic environmental descriptors were created, a fully connected multilayer neural network was employed to obtain atomic energies. The total energy of a molecule was then calculated as the sum of the individual atomic energies. In ANI, the ACSFs were processed by distinct neural network leads for each of the elements H, C, N, and O(Figure 5.2). Consequently, a total of four neural networks were employed, one for each of the HCNO elements.

5.3 Generating new dimer reference data

The database of dimer molecular systems used to build the new ANI model training dataset is composed of small organic dimer molecules from the widely used SS66 benchmark.[200] The dataset is named after the 66 molecular complexes it contains, which span a diverse range of chemical species and interaction types, including electrostatic dominated (hydrogen bounding), dispersion dominated (- stacking and van der Waals interaction), as well as mixed interactions. In addition to equilibrium geometries, extra extension of this dataset, S66x8,[200] was also made by same research group to describe the non equilibrium position, 8 geometries were provided along their dissociation curve of each dimer. Since the dimer binding geometries in S66 were accuracy calculated at their equilibrium position, it is straightforward to samples one of the molecule to different position to representing the intermolecular interactions.

To systematically generate training datasets, a Python program was developed to identify monomers and provide an interface to allow steric variation including transformation and rotation. Additionally, the program was designed to calculate the shortest distance between two monomers within a dimer system, which serves as a measure of their relative position. Since the molecular relative position in crystal structure is not always along the path of the dissociation of binding dimer, the training set need include wide range of position away form the path in the 3D space. Here, more than 1500 geometries were obtained through random sampling from various directions, with the condition that the shortest distance between two monomers falls within the range of 0.6 to 3 times their distance at equilibrium position. Then, structures where the distance between two monomers was longer than 8 Angstroms were discarded since ACSFs are designed to capture the neighbouring environment up to a certain distance, and larger distances may not be adequately represented by these descriptors. In addition to using the S66 dataset, the sampling process also included the S66a8 dataset. The S66a8 dataset was utilized due to its provision of more rotated geometries, which include rotations in both directions (±) within the molecular plane, and rotations perpendicular to it by 30°. Finally, the total energy and atomic force of all geometries were calculated at the density functional theory level, using the ωB97XD/6-31G*[202] level of theory to cover dispersion interactions, as implemented in Gaussian 16.[166] Two examples of the dimer dataset are presented in Figure 5.3



(a) Between water and acetamide

(b) Between ethene and ethane

Figure 5.3 The binding energy of two examples of sampled dimer structures

where the binding energy of each geometry was used to against their shortest dimer distance. The curve described the Lennard-Jones potential properly since many of geometries located in the range of intermolecular interactions. The size of this dimer dataset increased to approximate 1M, and was merged with the 5.5M size ANI-1x[203] dataset to prepare training new ANI model.

5.4 Modifying TorchANI API and training the model

The recent availability of the TorchANI program has significantly expedited the development of new ANI models. This is partly due to the use of PyTorch and ASE Python implementation, which have made TorchANI a lightweight, user-friendly, cross-platform, and easily modifiable software tool. The initial creation of ANI-1, ANI-1x, and ANI-2x involved the use of the NeuroChem package. Consequently, there is no direct Application Programming Interface (API) within TorchANI for reading PyTorch neural network files and using them as a force field calculator for the ASE. To overcome this limitation, several modifications were made to the TorchANI program in order to enable the development of the training, implementation, and benchmark processes. The ACSFs calculator, data loader, and the ASE interface were kept the same as in the original TorchANI program.

Modern deep learning frameworks, such as PyTorch,[174] incorporate an automatic differentiation engine. This feature is particularly advantageous when performing certain physical properties calculations, such as force calculations. Therefore, incorporating both total energy and atomic force as the cost function in the ANI model is feasible without incurring significant computational costs. Furthermore, this approach has the potential to improve the stability of the ANI model in molecular dynamics since the geometry and energy of molecules are determined by force. Integrating force training into PyTorch is a straightforward process, requiring only a few lines of code to be added to the energy trainer.

To effectively capture intermolecular interactions in the ANI model, it is necessary to increase the cutoff distance (R_C in Equation 5.1) hyperparameter of the ACSFs. Accomplishing this requires the addition of more shift parameters (R_S), which in turn enhances the resolution in the atomic neighbourhood environment. A serial of R_S was generated using an arithmetic progression generator. Other parameters such as η, ζ were also included in the optimization of ACSFs hyperparameters. The training of the model was conducted using the high-performance computing system (HPC) available at the University of Liverpool. In order to accelerate the calculations, an Nvidia V100 graphics processing unit (GPU) was employed.

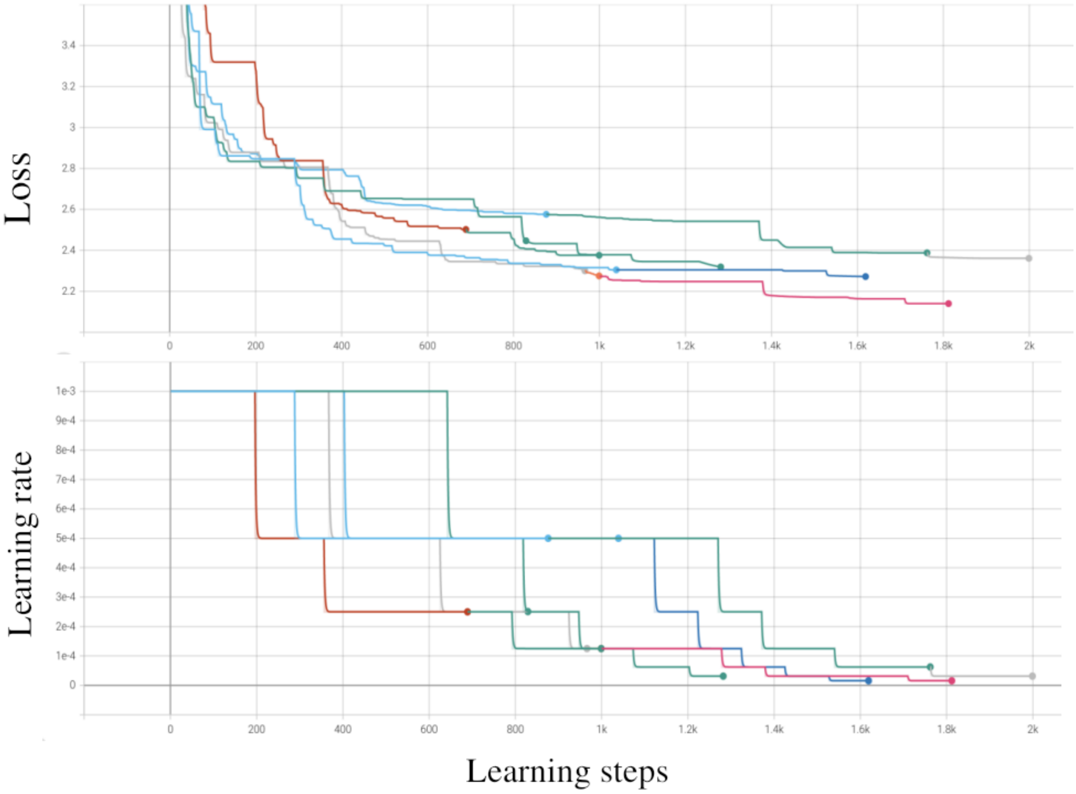


Figure 5.4 The loss and learning rate curves of ANI during training process. Different colour represents different running jobs on HPC for searching hyperparameters of ACSFs.

Figure 5.4 displays several training curves, with each colour representing a distinct training process. Some model training processes had to be split into multiple runs and considered as continuous training due to the running time limitation on Barkla. The training curves depict how the performance of the model changes as training progresses over time. The x-axis represents the number of training epochs, and the y-axis shows the loss (top) and learning rate (bottom). A learning rate scheduler was utilized to facilitate learning rate decay and regulate the learning rate.

5.5 Benchmark of the new trained ANI model

To illustrate the performance improvements of the above changes on the training process, several case studies were conducted including testing dimer Lennard-Jones potential, trimer molecules binding energy, and the lattice energy of published CSP dataset. All the selected compounds only contains HCNO elements. The dimer benchmark structures were sourced from both the S66x8 and S12L[204] datasets, which encompass small organic molecules and large supramolecular complexes, respectively. To represent a relative many-body non-covalent interaction, the trimer structures were selected from the 3B-69[205] dataset. Additionally, the simulated CSP datasets were generated in accordance with a previous publication from our research group.

5.5.1 Testing the Lennard-Jones potential on dimer complexes

In order to assess the effectiveness of the newly trained ANI model, the binding energies of four distinct systems sourced from the S66x8 dataset were computed. These systems include water-water and methanol-nethylamine, which serve to represent electrostatically dominated interactions, as well as benzene-benzene (face to perpendicular) and benzene-cyclopentane (face to face), which are representative of dispersion dominated interactions. The binding energy is determined by calculating the difference between the energy of the dimer complex (E_{ab}) and the combined energies of the individual molecules (E_a and E_b) at corresponding geometries (Equation 5.4).

$$E = \frac{1}{2}(E_{ab} - (E_a + E_b)) \tag{5.4}$$

The Figure 5.5 illustrates the assessment of binding energy as a function of distance along the dissociation curve for each complex. The displaced complexes in S66x8 are generated by adjusting the intermolecular distance in the optimized structure. Here, more 75 points for each complex are created to ensure a precise reconstruction of the dissociation curve

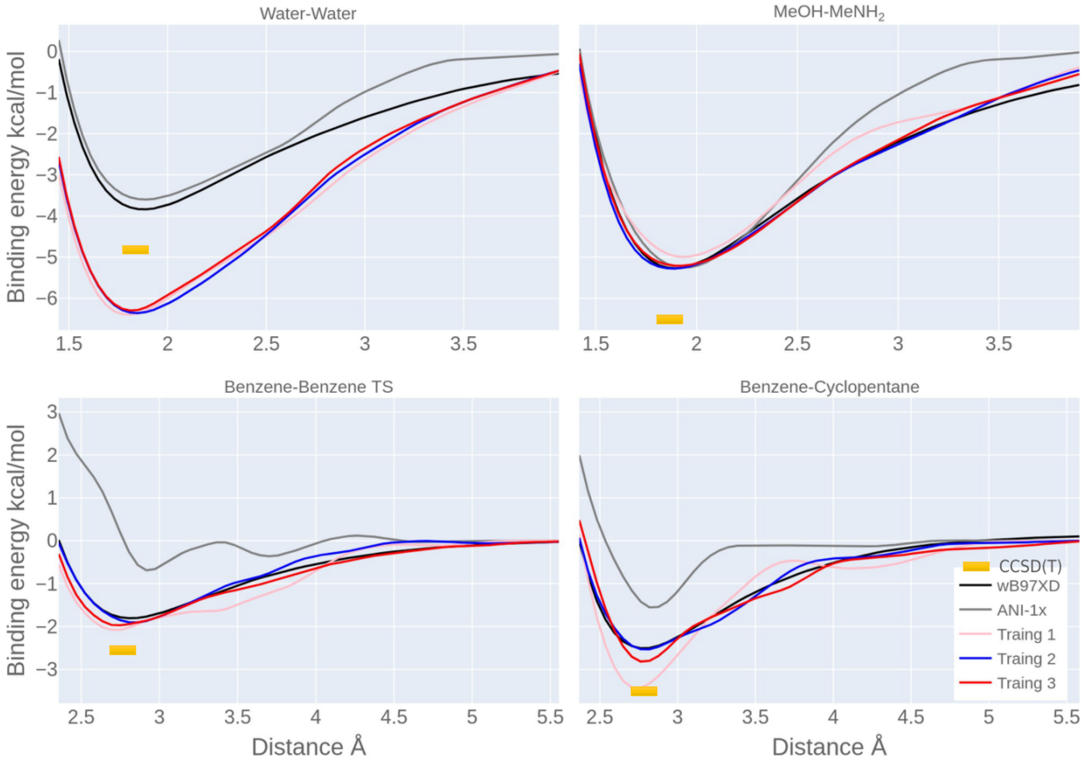


Figure 5.5 The binding energy test for four different systems from S66. New trained ANI model are labeled as training 1, 2, and 3 using different hyperparameters. CCSD(T) data came from S66.[200]

through interpolation. The newly trained ANI models 1, 2, and 3, as depicted in Figure 5.5 exhibit an enhancement in binding energy as compared to ANI-1x. It is important to note that the reference data used in this study (ωB97XD/6-31G* level) is not the most accurate hybrid functionals for describing hydrogen bonds, as indicated by a benchmark study.[206] According to this benchmark, the error for a smaller set of 16 hydrogen-bonded complexes using this level of theory is reported to be 0.7 kcal/mol when compared to the highly accurate CCSD(T) method, which is often considered the gold standard in computational chemistry for single-reference systems.[206] When comparing the results with the S66 data, which also provides CCSD(T) level binding energy, it is observed that the reference functional underestimates the binding energy of hydrogen bonds between water molecules. On the other hand, the three new ANI models trained in this study tend to overestimate the binding energy. This discrepancy could potentially be attributed to the increased cut-off distance of ACSFs utilized during the training of these ANI models.

Table 5.1 shows the MAE of the binding energies calculated with different ANI model comparing with DFT level. The training 2 model has the best error compared with other ANI model, although both new trained model have larger MAE compared with ANI-1x in the water dimer test.

Table 5.1 The MAE between ANI-1x, three new trained ANI model against ωB97XD/6-31G* on the binding energy test

Dimer name	ANI-1x	Training 1	Training 2	Training 3
Water-Water	0.49	1.74	1.59	1.56
MeOH-MeNH$_2$	0.71	0.89	0.34	0.49
Benzene-Benzene TS	2.79	0.53	0.51	0.69
Benzene-Cyclopentane	1.70	0.98	0.32	0.56

[a] All in kcal/mol.

In addition to the aforementioned dimer small molecule test, larger molecular complexes were also included in the benchmark. Two supramolecular complexes from S12L and three trimer examples from 3B-69 were selected to assess their binding energies against DFT calculations (Figure 5.6). As expected, the ANI model demonstrated satisfactory performance on the trimer uracil and benzene tests, as these monomers were part of the training dataset. However, it still tended to overestimate the binding energy of the trimer water system. Regarding large molecular complexes, the ANI model exhibited a reasonable degree of transferability with limited improvements compared to ANI1-x, although the values remained comparable. Among these ANI models, the training 2 model exhibits the best performance, and it is selected for the subsequent molecular crystal structure tests consequently.

5.5.2 Testing the lattice energy landscape on CSP dataset

To further stress the application of ANI machine learning potential on organic crystal, simulated CSP datasets were chosen to show the performance of the new ANI model

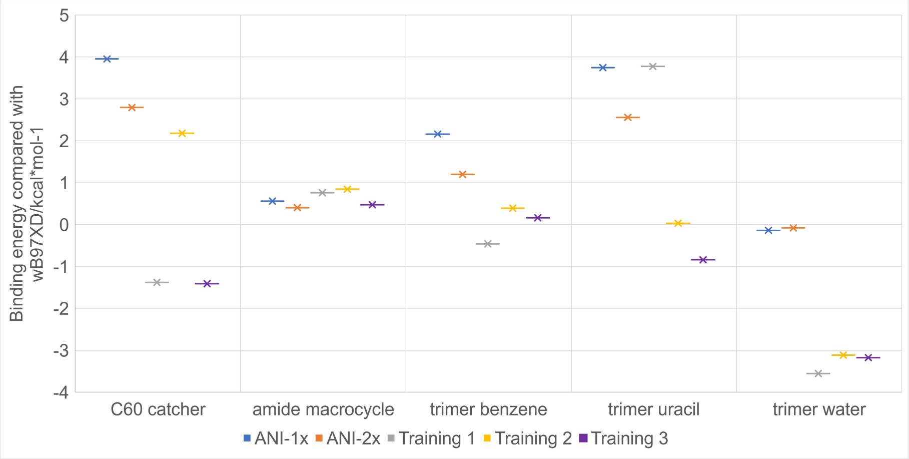
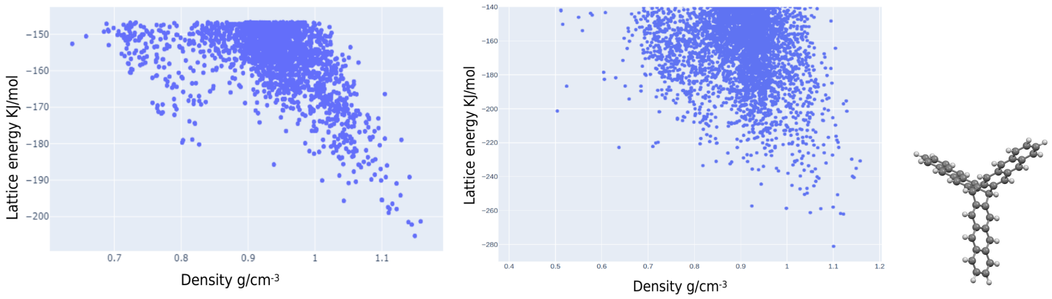


Figure 5.6 The binding energy test for large molecular complexes and trimer system against ωB97XD/6-31G* level

in these complex systems. To visually represent the landscape of crystal structures, a commonly used approach is to plot the lattice energy or relative lattice energy against the density of each structure.

As a preliminary step, a dispersion dominated system (Figure 5.7(c)) was tested, as depicted in Figure 5.7 provided below. The corresponding structure was obtained through CSP, and subsequently, ANI was utilized to calculate the lattice energy without conducting any structure optimization. The training 2 model (in Figure 5.7(b)) successfully captured the observed negative relationship between lattice energy and density, when compared to the CSP landscape. However, accurately evaluating the performance is challenging due to the variation in the ranking of minimal structures. This discrepancy arises from the limitations of the employed CSP technique, which solely incorporates full structure optimization of individual molecules while keeping them rigid during crystal structure generation and lattice energy minimization to reduce the cost.

Further investigations were undertaken using our research group's recently published dataset of organic cage crystals (cage-3-NH_2 in Figure 5.8(e)).[207] In this study, the authors not only performed CSP calculations but also conducted high-level structure optimization, enabling a comparison with structures of relatively high accuracy. From the



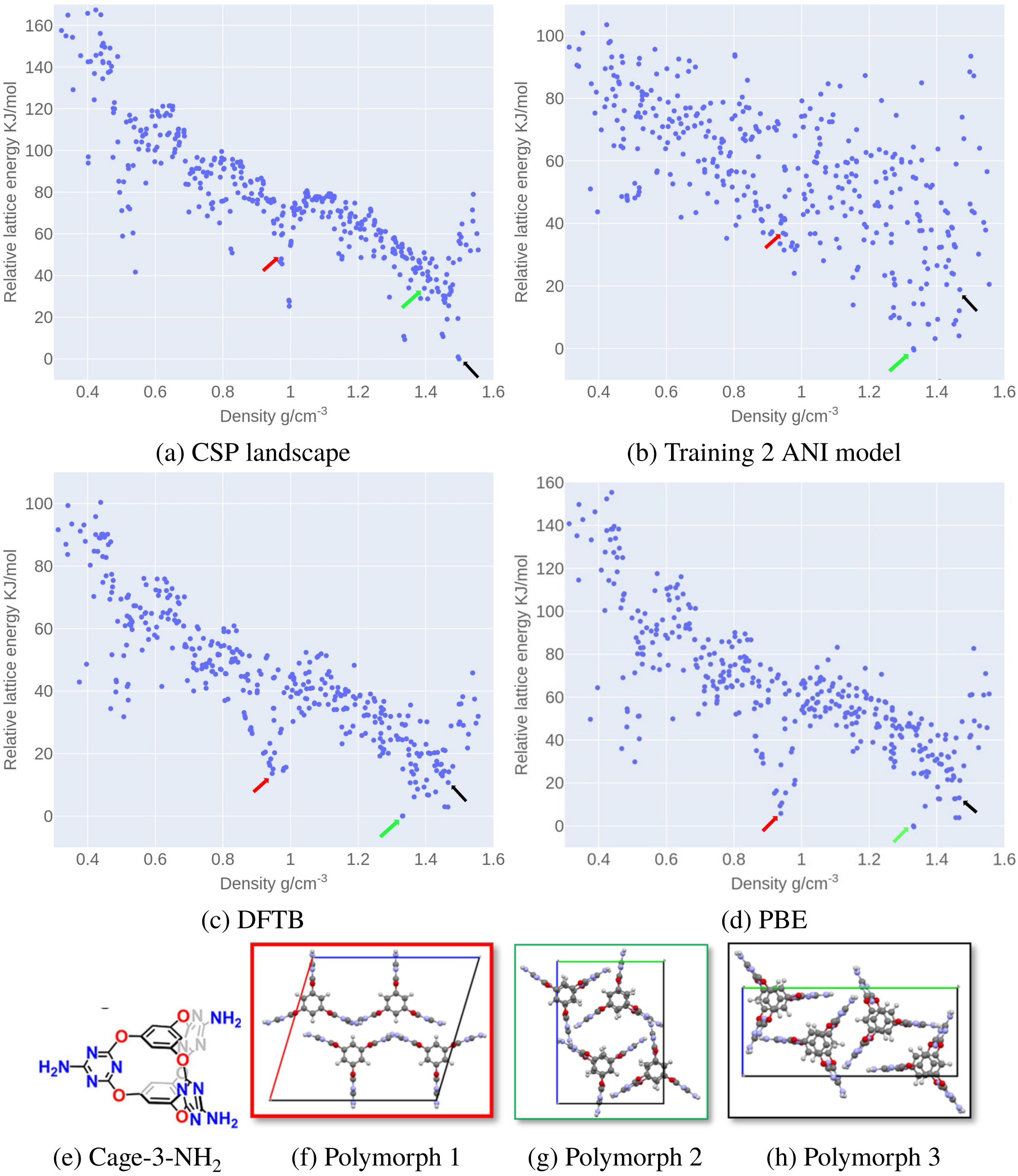
(a) CSP landscape (b) Training 2 ANI model landscape (c) Monomer

Figure 5.7 The lattice energy test of a no-hydrogen bound CSP dataset

energy landscape of the CSP, structures located on the ‘leading edge’ (Figure 5.8(a)) were selected and re-optimized using density functional theory-based tight binding (DFTB)[208] (Figure 5.8(c)). Furthermore, the relative lattice energy of these crystal structures was calculated using both the training 2 ANI model (Figure 5.8(b)) and the PBE method[209] (Figure 5.8(d)), allowing for a comprehensive comparison.

In Figure 5.8, three highlighted polymorphs were marked as red (Figure 5.8(f)), green (Figure 5.8(g)) and black (Figure 5.8(h)), receptively. The polymorph 3 represents the minimum lattice energy structure identified within the CSP landscape, while the polymorph 2 represents the minimum lattice energy structure identified by the DFTB, PBE, or ANI calculations. Additionally, the polymorph 1 represents the minimum lattice energy structure of a distinct feature referred to as a ‘spike’. This valuable configuration exhibits the potential to be synthesized as a highly porous material. The DFTB, PBE, and ANI calculations were carried out utilizing the crystal structure optimized through DFTB. This optimized structure, along with its corresponding density, provides a higher level of accuracy when compared to the CSP method.

While the ANI model successfully identifies the minimum energy structure, it tends to underestimate the energy of several other structures, including the highly porous polymorph (depicted in red). One possible hypothesis is that the ACSFs descriptor may not adequately capture longer-range interactions beyond the cut-off sphere (typical 8 Å), whereas electronic-dominated interactions continue to influence and contribute to the lattice en-



(a) CSP landscape

(b) Training 2 ANI model

(c) DFTB

(d) PBE

(e) Cage-3-NH$_2$

(f) Polymorph 1

(g) Polymorph 2

(h) Polymorph 3

Figure 5.8 Energy-density distributions of the leading edge structures on the CSP landscape using different calculation methods

ergy in crystals within this range. The lattice energy of many high-porosity, low lattice energy polymorphs in the CSP dataset is primarily influenced by intermolecular hydrogen bonding. This limitation is inherent in the design of the ANI model and other symmetry function-based machine learning force fields, as they were primarily developed to simulate interatomic interactions rather than intermolecular interactions.

5.6 Conclusion and future works

The rapid advancement and refinement of machine learning force fields present a promising approach to molecular simulation. In this study, a new training model for the ANI has been introduced, focusing on enhancing the accuracy of intermolecular interactions in molecules containing HCNO elements. The improvement was achieved by incorporating a new dimer training dataset, calculating energy using a dispersion-corrected method (ωB97XD/6-31G*), and modifying the hyperparameters of ACSFs.

The modified ANI model was tested on both small dimer organic molecules, which are considered as classic benchmarks for intermolecular interactions, and two larger supramolecular complexes. In comparison to the ANI1x model, the modified ANI model exhibited improved accuracy, reaching a level comparable to density functional theory (DFT) calculations, for small dimer organic molecules. However, for the larger molecular complexes, the performance was not as satisfactory as initially anticipated. Except this, this fast and flexible force field offers a potential method for optimizing organic molecular crystal structures generated through CSP techniques. While the ANI model can identify structures with minimum lattice energy from the extensive CSP dataset, it still falls short in capturing the full spectrum of features present in the CSP landscape due to the lack of long-range information in its current description.

Moving forward, it is apparent that the inclusion of long-range interactions through the addition of a dimer dataset and modifications to the ACSFs alone are insufficient to fully address the missing long-range interaction issue. Future studies should therefore concentrate on implementing advanced sampling methods to generate reference data and explore the utilization of additional deep learning networks specifically designed to capture interactions beyond the cutoff sphere. Notably, Behler has proposed the concept of 'third generation' high-dimensional neural network potentials,[33] which could serve as a promising direction for these investigations. These advancements hold the potential to significantly enhance the accuracy and capability of machine learning force fields in capturing long-range interactions and improving their overall predictive power.

CHAPTER 6 SUMMARY AND OUTLOOK

6.1 Data driven material discovery by *in-silico* design

While high throughput virtual screening (HTVS) has been topical in materials discovery and applied to a wide range of materials classes and functional properties, automated (big-)data analytics are still rarely used to assist the visualization and interpretation of HTVS results that are often high-dimensional and convoluted. With the modern dimensionality reduction algorithms and online visualization tools, high dimensional chemical datasets, calculated by organic crystal structure prediction, are capable for screening ‘landmark’ structures and visualizing their structures on the interactive data explorers on the fly with dynamical data input. The visualization and ML tools developed herein were then used to help accelerate the discovery of molecular photocatalysts by concerting efforts from experiment and computation.

To expedite the energy calculation process in crystal structure prediction, machine learning potential offers a promising approach for evaluating and optimizing the structures of organic molecules. An extension and refinement of Torch-ANI and ANI-1x models were conducted to effectively capture long-range intermolecular interactions, which dominated the lattice energy surface of organic crystal structures. By adding new dimer dataset to the ANI-1x training set and optimizing the ACSFs hyperparameters, the new trained ANI model is capable to describe the interactions in dimer complexes. However, the testing of the newly trained ANI model on larger and more complex systems, such as the CSP dataset, did not yield satisfactory results due to the limitations of the cut-off sphere. Despite the substantial growth in the size of simulated chemical data, the neural network potential model still requires further enhancement to effectively capture intermolecular interactions. Moreover, its transferability might be constrained when applied to large-sized molecules during testing. Future development could involve the introduction of additional neural networks and descriptors.

It is worth noting that those porous materials has the potential application of adsorption which related to the host-guest interactions.[210] Although this study did not investigate the application of porous organic molecular crystals, the accurate prediction of crystal structures is crucial because it allows researchers to understand the arrangement of atoms within the material. By accurately predicting the structure of porous organic molecular crystals, researchers can gain insights into the arrangement of pores and channels within the material. This information is vital for determining the suitability of the material for specific adsorption applications, such as gas storage, separation processes, or catalysis.[211-212] One limitation of this estimate could be the difficulty in accurately calculating the electron density distribution for organic molecular crystals.[213]

6.2 Screening organic photocatalysts by machine learning

Organic molecules present a promising avenue for the photocatalytic production of hydrogen from water. By applying both unsupervised learning and supervised regression/classification to a large library of organic photocatalysts tested experimentally, structural and electronic features that positively impact the catalytic performance were identified. For example, the formation of triplet excitons has been suggested to be a key, beneficial effect on HERs. The chosen of DFT methods for calculating optoelectronic and physicochemical properties is indeed crucial and should be tailored to the specific type of molecules under investigation. While the regression task of targeting these properties to hydrogen evolution reaction (HER) activity have been unsuccessful, the application of binary and ternary classification still allows for the examination of important features related to HER. For instance, one important feature that may emerge as influential for HER is the energy difference ($\Delta E_{S_1 \to T_1}$) between the first singlet excited state(S_1) and the first triplet excited state (T_1). Furthermore, conducting virtual experiments with an adaptive ML-assisted selection approach can significantly reduce experimental costs. By leveraging machine learning techniques, the selection of experiments can be optimized, leading to a more efficient use of resources and reducing the overall experimental burden. Additionally, in an effort to make this library accessible to all, an online web application has been developed, enabling users to derive independent insights from the data. The machine learning models trained through this process can be deployed in future studies and will be continuously refined with the inclusion of new experimental data as they become available.

Based on the high throughput screening of organic photocatalysts of hydrogen evolution, a Bayesian optimization strategy was used to identify promising OPCs and their ideal reaction conditions from a virtual library of 560 candidate molecules while exploring a small fraction of the available chemical space. The results revealed OPC formulations with target reaction yields of up to 88%, which were found to be comparable to iridium catalysts at high nickel concentrations and superior to iridium catalysts at lower nickel concentrations. It is shown that Bayesian, data-driven experimental design is a promising approach for the discovery of new photocatalysts formulations, and by extension for other research challenges where there is a large search space and limited prior knowledge. Except the discovery of highly active catalysts, the valuable photoredox results obtained can be utilized in a binary classification task by SVM, GBDT etc. to predict the potential activity of newly designed organic CNPs. This approach is similar to the discovery made by Raccuglia *et al.*. The ‘failed’ attempts were also incorporated during the training of GPs in the Bayesian optimization loop. This means that unsuccessful reactions were considered as part of the training data for the GPs, providing valuable information on the factors that contribute to lower activity. The chemical intuition regarding the substituent Rb was initially identified through the first close loop optimization. The CNPs consisting of carbazole Rb moieties paired with donating Ra moieties provide optimal yields, but the computationally-led search of 100 additional CNP molecules did not discover new high activity catalysts, highlights the complexity of the photoredox reaction.

In addition to applying calculated optoelectronic properties to small organic molecules for photocatalysis, extending such methodologies to other porous materials like COFs (covalent organic frameworks) and MOFs (metal-organic frameworks) can be valuable for screening potential photoactive materials. This is because the photocatalytic mechanisms in these materials are still closely related to processes such as light absorption and electron transfer between donors and acceptors.[160,212] In the case of COFs and MOFs, computational calculations can be simplified by representing the material using a single representative piece of the framework, thereby reducing the computational time required. Additionally, the substructure of the building blocks within these porous materials has been found to influence the photocatalytic activity.[119] Therefore, combining substructure fingerprints as representatives and graph representations of the frameworks may aid in the design of efficient porous photocatalysts.

As large-scale computational screening studies become routinely carried out by chemists,[11] new opportunities have arisen for accelerating materials discovery by taking advantage of the availability of big data, and spanning from computer-generated data to recorded laboratory results. However, it remains a challenge to properly understand and use the vast amount of data generated by simulations or experiments.[140] Obtaining a big, homogeneous, experimental dataset and their reproducibility is challenging due to the various physical factors that can influence experimental outcomes even with high throughput testing methodology.[8]

Another important step towards making experimental data accessible, interoperable, and reusable is to collect and publish the data using a standardized ecosystem. In addition to the challenge of collecting chemical data, chemists are often constrained by limitations in the featurization of materials for properties that require numerical representation. For instance, substructure characterization for organic molecules, electronic behaviours for photocatalysts, and geometric measurements for forcefield development all require appropriate numerical representation. Such insights are particularly important to enable materials discovery and design paradigms to go beyond serendipitous discoveries and even the currently most successful predict-make-measure approach, which is nevertheless sometimes undesirably linear.

ACKNOWLEDGEMENTS

After years of training and research in the fields of machine learning and its applications in chemistry, this doctoral thesis has finally been completed and presented here. Through participation and leadership in various research projects, I have experienced the allure of scientific research, accompanied by numerous challenging journeys.

First, I would like to express my gratitude to my supervisor team from the University of Liverpool and IBM for providing a world-class research environment that allowed me to explore topics of my interest. Prof. Andy Cooper and Ed Pyzer-Knapp guided the entire doctoral project, offering constructive insights and keeping me updated on cutting-edge scientific research.

I would like to give special thanks to Dr. Linjiang Chen. Prior to the commencement of my doctoral project, I had no knowledge in coding and machine learning. It was his guidance and encouragement that steered me from traditional experimental chemistry research towards data-driven computational research. He assisted me in various aspects, from macroscopic views to in-depth discussions on research directions.

Furthermore, my research during the doctoral period required the assistance and participation of numerous individuals. I am grateful to my experimental collaborators Dr. Xiaobo Li for providing valuable data and research insights. My computational chemistry colleagues Dr. Tao Liu helped me establish models and assisted in numerous quantum chemical calculations. Thank you, Xenofon, for reviewing this thesis and providing me with valuable advice. I appreciate the other members of the research group for creating a warm and vibrant research atmosphere, such as Chengxi, Zac, Phil, Xue, Zhongfu. Additionally, I would like to acknowledge external collaborator Prof. Olexandr Isayev from Carnegie Mellon University, for their help in resolving various issues related to machine learning force fields.

REFERENCES

[1] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha and T. Buonassisi, Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing, *Joule*, 2018, **2**, 1410–1420, DOI: 10.1016/j.joule.2018.05.009.

[2] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, Best practices in machine learning for chemistry, *Nature Chemistry*, 2021, **13**, 505–508, DOI: 10.1038/s41557-021-00716-z.

[3] A. Tkatchenko, Machine learning for chemical discovery, *Nature Communications*, 2020, **11**, 4125, DOI: 10.1038/s41467-020-17844-8.

[4] M. H. S. Segler, M. Preuss and M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature*, 2018, **555**, 604–610, DOI: 10.1038/nature25978.

[5] M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks, *ACS Central Science*, 2018, **4**, 120–131, DOI: 10.1021/acscentsci.7b00512.

[6] L. Zhang, J. Han, H. Wang, R. Car and W. E, Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics, *Physical Review Letters*, 2018, **120**, 143001, DOI: 10.1103/PhysRevLett.120.143001.

[7] P. Z. Moghadam, S. M. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. V. Speybroeck and D. Fairen-Jimenez, Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning, *Matter*, 2019, **1**, 219–234, DOI: 10.1016/j.matt.2019.03.002.

[8] S. M. Moosavi, K. M. Jablonka and B. Smit, The Role of Machine Learning in the Understanding and Design of Materials, *Journal of the American Chemical Society*, 2020, **142**, 20273–20287, DOI: 10.1021/jacs.0c09105.

[9] K. Champion, B. Lusch, J. N. Kutz and S. L. Brunton, Data-driven discovery of coordinates and governing equations, *Proceedings of the National Academy of Sciences*, 2019, **116**, 22445–22451, DOI: 10.1073/pnas.1906995116.

[10] L. Lu, M. Dao, P. Kumar, U. Ramamurty, G. E. Karniadakis and S. Suresh, Extraction of mechanical properties of materials through deep learning from instrumented indentation, *Proceedings of the National Academy of Sciences*, 2020, **117**, 7052–7062, DOI: 10.1073/pnas.1922210117.

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**, 547–555, DOI: 10.1038/s41586-018-0337-2.

[12] D. McDonagh, C.-K. Skylaris and G. M. Day, Machine-Learned Fragment-Based Energies for Crystal Structure Prediction, *Journal of Chemical Theory and Computation*, 2019, **15**, 2743–2758, DOI: 10.1021/acs.jctc.9b00038.

[13] C. Chen, Y. Zuo, W. Ye, X. Li and S. P. Ong, Learning properties of ordered and disordered materials from multi-fidelity data, *Nature Computational Science*, 2021, **1**, 46–53, DOI: 10.1038/s43588-020-00002-x.

[14] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning, *Science*, 2019, **363**, eaau5631, DOI: 10.1126/science.aau5631.

[15] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, Machine-learning-assisted materials discovery using failed experiments, *Nature*, 2016, **533**, 73–76, DOI: 10.1038/nature17439.

[16] A. E. Gongora, B. Xu, W. Perry, C. Okoye, P. Riley, K. G. Reyes, E. F. Morgan and K. A. Brown, A Bayesian experimental autonomous researcher for mechanical design, *Science Advances*, 2020, **6**, eaaz1708, DOI: 10.1126/sciadv.aaz1708.

[17] S. Masubuchi, M. Morimoto, S. Morikawa, M. Onodera, Y. Asakawa, K. Watanabe, T. Taniguchi and T. Machida, Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices, *Nature Communications*, 2018, **9**, 1413, DOI: 10.1038/s41467-018-03723-w.

[18] R. E. Brandt, R. C. Kurchin, V. Steinmann, D. Kitchaev, C. Roat, S. Levcenco, G. Ceder, T. Unold and T. Buonassisi, Rapid Photovoltaic Device Characterization through Bayesian Parameter Estimation, *Joule*, 2017, **1**, 843–856, DOI: 10.1016/j.joule.2017.10.001.

[19] Z. Allahyari and A. R. Oganov, Coevolutionary search for optimal materials in the space of all possible compounds, *npj Computational Materials*, 2020, **6**, 55, DOI: 10.1038/s41524-020-0322-9.

[20] L. Ward, A. Dunn, A. Faghaninia, N. E. 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 and A. Jain, Matminer: An open source toolkit for materials data mining, *Computational Materials Science*, 2018, **152**, 60–69, DOI: 10.1016/j.commatsci.2018.05.018.

[21] P. Z. Moghadam, T. Islamoglu, S. Goswami, J. Exley, M. Fantham, C. F. Kaminski, R. Q. Snurr, O. K. Farha and D. Fairen-Jimenez, Computer-aided discovery of a metal-organic framework with superior oxygen uptake, *Nature Communications*, 2018, **9**, 1378, DOI: 10.1038/s41467-018-03892-8.

[22] C. Zhao, L. Chen, Y. Che, Z. Pang, X. Wu, Y. Lu, H. Liu, G. M. Day and A. I. Cooper, Digital navigation of energy-structure-function maps for hydrogen-bonded porous molecular crystals, *Nature Communications*, 2021, **12**, 817, DOI: 10.1038/s41467-021-21091-w.

[23] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson and A. Aspuru-Guzik, Accelerating the discovery of materials for clean energy in the era of smart automation, *Nature Reviews Materials*, 2018, **3**, 5–20, DOI: 10.1038/s41578-018-0005-z.

[24] J. Chang, P. Nikolaev, J. Carpena-Núñez, R. Rao, K. Decker, A. E. Islam, J. Kim, M. A. Pitt, J. I. Myung and B. Maruyama, Efficient Closed-loop Maximization of Carbon Nanotube Growth Rate using Bayesian Optimization, *Scientific Reports*, 2020, **10**, 9040, DOI: 10.1038/s41598-020-64397-3.

[25] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto and B. Maruyama, Autonomy in materials research: a case study in carbon nanotube growth, *npj Computational Materials*, 2016, **2**, 16031, DOI: 10.1038/npjcompumats.2016.31.

[26] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments, *Science Advances*, 2018, **4**, eaaq1566, DOI: 10.1126/sciadv.aaq1566.

[27] K. Hippalgaonkar, Q. Li, X. Wang, J. W. Fisher, J. Kirkpatrick and T. Buonassisi, Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics, *Nature Reviews Materials*, 2023, 1–20, DOI: 10.1038/s41578-022-00513-1.

[28] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems, *Chemical Reviews*, 2021, **121**, 9816–9872, DOI: 10.1021/acs.chemrev.1c00107.

[29] O. A. von Lilienfeld, Quantum Machine Learning in Chemical Compound Space, *Angewandte Chemie International Edition*, 2018, **57**, 4164–4169, DOI: 10.1002/anie.201709686.

[30] R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach, *Journal of Chemical Theory and Computation*, 2015, **11**, 2087–2096, DOI: 10.1021/acs.jctc.5b00099.

[31] O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, *Nature Reviews Chemistry*, 2020, **4**, 347–358, DOI: 10.1038/s41570-020-0189-9.

[32] P. Friederich, F. Häse, J. Proppe and A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, *Nature Materials*, 2021, **20**, 750–761, DOI: 10.1038/s41563-020-0777-6.

[33] J. Behler, Four Generations of High-Dimensional Neural Network Potentials, *Chemical Reviews*, 2021, **121**, 10037–10072, DOI: 10.1021/acs.chemrev.0c00868.

[34] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, *Physical Review Letters*, 2007, **98**, 146401, DOI: 10.1103/PhysRevLett.98.146401.

[35] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning, *Nature Communications*, 2019, **10**, 2903, DOI: 10.1038/s41467-019-10827-4.

[36] A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, *Physical Review Letters*, 2010, **104**, 136403, DOI: 10.1103/PhysRevLett.104.136403.

[37] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, Machine learning unifies the modeling of materials and molecules, *Science Advances*, 2017, **3**, e1701816, DOI: 10.1126/sciadv.1701816.

[38] G. P. P. Pun, R. Batra, R. Ramprasad and Y. Mishin, Physically informed artificial neural networks for atomistic modeling of materials, *Nature Communications*, 2019, **10**, 2339, DOI: 10.1038/s41467-019-10343-5.

[39] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, Physics-Inspired Structural Representations for Molecules and Materials, *Chemical Reviews*, 2021, **121**, 9759–9815, DOI: 10.1021/acs.chemrev.1c00021.

[40] T. Zubatiuk and O. Isayev, Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence, *Accounts of Chemical Research*, 2021, **54**, 1575–1585, DOI: 10.1021/acs.accounts.0c00868.

[41] S. L. Price, Predicting crystal structures of organic compounds, *Chemical Society Reviews*, 2014, **43**, 2098–2111, DOI: 10.1039/C3CS60279F.

[42] S. M. Woodley and R. Catlow, Crystal structure prediction from first principles, *Nature materials*, 2008, **7**, 937–946, DOI: 10.1038/nmat2321.

[43] M. K. Dudek and K. Drużbicki, Along the road to crystal structure prediction (CSP) of pharmaceutical-like molecules, *CrystEngComm*, 2022, **24**, 1665–1678, DOI: 10.1039/D1CE01564H.

[44] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti and M. A. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, *Chemistry of Materials*, 2017, **29**, 5090–5103, DOI: 10.1021/acs.chemmater.7b00156.

[45] A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, Structure prediction drives materials discovery, *Nature Reviews Materials*, 2019, **4**, 331–348, DOI: 10.1038/s41578-019-0101-8.

[46] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre and J. Parkhill, The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics, *Chemical science*, 2018, **9**, 2261–2269, DOI: 10.1039/C7SC04934J.

[47] Z. Wang, C. Wang, S. Zhao, S. Du, Y. Xu, B.-L. Gu and W. Duan, Symmetry-adapted graph neural networks for constructing molecular dynamics force fields, *Science China Physics, Mechanics & Astronomy*, 2021, **64**, 117211, DOI: 10.1007/s11433-021-1739-4.

[48] N. Kiselyova, V. Gladun and N. Vashchenko, Computational materials design using artificial intelligence methods, *Journal of Alloys and Compounds*, 1998, **279**, 8–13, DOI: 10.1016/S0925-83889800606-9.

[49] G. Hautier, C. C. Fischer, A. Jain, T. Mueller and G. Ceder, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, *Chemistry of Materials*, 2010, **22**, 3762–3767, DOI: 10.1021/cm100795d.

[50] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, Accelerating materials property predictions using machine learning, *Scientific reports*, 2013, **3**, 1–6, DOI: 10.1038/srep02810.

[51] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Physical review letters*, 2018, **120**, 145301, DOI: 10.1103/PhysRevLett.120.145301.

[52] C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chemistry of Materials*, 2019, **31**, 3564–3572, DOI: 10.1021/acs.chemmater.9b01294.

[53] X. Zheng, P. Zheng and R.-Z. Zhang, Machine learning material properties from the periodic table using convolutional neural networks, *Chemical science*, 2018, **9**, 8426–8432, DOI: 10.1039/C8SC02648C.

[54] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld and R. Armiento, Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals, *Physical review letters*, 2016, **117**, 135502, DOI: 10.1103/PhysRevLett.117.135502.

[55] F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day and M. Ceriotti, Machine learning for the structure–energy–property landscapes of molecular crystals, *Chemical science*, 2018, **9**, 1289–1300, DOI: 10.1039/C7SC04665K.

[56] M. Gastegger, J. Behler and P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra, *Chemical science*, 2017, **8**, 6924–6935, DOI: 10.1039/C7SC02267K.

[57] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, *Nature communications*, 2017, **8**, 15679, DOI: 10.1038/ncomms15679.

[58] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard and A. Tropsha, QSAR Modeling: Where Have You Been? Where Are You Going To?, *Journal of Medicinal Chemistry*, 2014, **57**, 4977–5010, DOI: 10.1021/jm4004285.

[59] A. K. Singh, J. H. Montoya, J. M. Gregoire and K. A. Persson, Robust and synthesizable photocatalysts for CO_2 reduction: a data-driven materials discovery, *Nature communications*, 2019, **10**, 443, DOI: 10.1038/s41467-019-08356-1.

[60] H. Masood, C. Y. Toe, W. Y. Teoh, V. Sethu and R. Amal, Machine Learning for Accelerated Discovery of Solar Photocatalysts, *ACS Catalysis*, 2019, **9**, 11774–11787, DOI: 10.1021/acscatal.9b02531.

[61] Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov and T. F. Jaramillo, Combining theory and experiment in electrocatalysis: Insights into materials design, *Science*, 2017, **355**, eaad4998, DOI: 10.1126/science.aad4998.

[62] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, Predicting reaction performance in C–N cross-coupling using machine learning, *Science*, 2018, **360**, 186–190, DOI: 10.1126/science.aar5169.

[63] D. Padula and A. Troisi, Concurrent Optimization of Organic Donor-Acceptor Pairs through Machine Learning, *Advanced Energy Materials*, 2019, **9**, 1902463, DOI: 10.1002/aenm.201902463.

[64] Y. Bai, L. Wilbraham, B. J. Slater, M. A. Zwijnenburg, R. S. Sprick and A. I. Cooper, Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory, *Journal of the American Chemical Society*, 2019, **141**, 9063–9071, DOI: 10.1021/jacs.9b03591.

[65] P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou and B. Smit, Data-driven design of metal–organic frameworks for wet flue gas CO2 capture, *Nature*, 2019, **576**, 253–256, DOI: 10.1038/s41586-019-1798-7.

[66] M.-F. Ng, J. Zhao, Q. Yan, G. J. Conduit and Z. W. Seh, Predicting the state of charge and health of batteries using data-driven machine learning, *Nature Machine Intelligence*, 2020, **2**, 161–170, DOI: 10.1038/s42256-020-0156-7.

[67] 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 and D. Hassabis, Mastering the game of Go without human knowledge, *Nature*, 2017, **550**, 354–359, DOI: 10.1038/nature24270.

[68] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models, *Chemical Science*, 2018, **9**, 6091–6098, DOI: 10.1039/C8SC02339E.

[69] B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, *Science*, 2018, **361**, 360–365, DOI: 10.1126/science.aat2663.

[70] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach, *Nature Materials*, 2016, **15**, 1120–1127, DOI: 10.1038/nmat4717.

[71] Y. Wu, P. Lazic, G. Hautier, K. Persson and G. Ceder, First principles high throughput screening of oxynitrides for water-splitting photocatalysts, *Energy & environmental science*, 2013, **6**, 157–168, DOI: 10.1039/C2EE23482C.

[72] B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram and J. M. Gregoire, Benchmarking the acceleration of materials discovery by sequential learning, *Chemical Science*, 2020, **11**, 2696–2706, DOI: 10.1039/C9SC05999G.

[73] K. Korovina, S. Xu, K. Kandasamy, W. Neiswanger, B. Poczos, J. Schneider and E. P. Xing, ChemBO: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations, *arXiv e-prints*, 2019, DOI: 10.48550/arXiv.1908.01425.

[74] E. O. Pyzer-Knapp, G. N. Simm and A. A. Guzik, A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials, *Materials Horizons*, 2016, **3**, 226–233, DOI: 10.1039/C5MH00282F.

[75] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon and W. C. Chueh, Closed-loop optimization of fast-charging protocols for batteries with machine learning, *Nature*, 2020, **578**, 397–402, DOI: 10.1038/s41586-020-1994-5.

[76] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, Bayesian reaction optimization as a tool for chemical synthesis, *Nature*, 2021, **590**, 89–96, DOI: 10.1038/s41586-021-03213-y.

[77] Z. Ren, S. I. P. Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun, X. Wang, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung and T. Buonassisi, An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties, *Matter*, 2022, **5**, 314–335, DOI: 10.1016/j.matt.2021.11.032.

[78] D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, Deep learning for molecular design–a review of the state of the art, *arXiv e-prints*, 2019, DOI: 10.1039/C9ME00039A.

[79] A. Nouira, N. Sokolovska and J.-C. Crivello, CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks, *arXiv e-prints*, 2018, DOI: 10.48550/arxiv.1810.11203.

[80] D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, *arXiv e-prints*, 2013, DOI: 10.48550/arXiv.1312.6114.

[81] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Central Science*, 2018, **4**, 268–276, DOI: 10.1021/acscentsci.7b00572.

[82] V. Kondratyev, M. Dryzhakov, T. Gimadiev and D. Slutskiy, Generative model based on junction tree variational autoencoder for HOMO value prediction and molecular optimization, *Journal of Cheminformatics*, 2023, **15**, 11, DOI: 10.1186/s13321-023-00681-4.

[83] C. Rupakheti, A. Virshup, W. Yang and D. N. Beratan, Strategy to discover diverse optimal molecules in the small molecule universe, *Journal of chemical information and modeling*, 2015, **55**, 529–537, DOI: 10.1021/ci500749q.

[84] K. Muraoka, Y. Sada, D. Miyazaki, W. Chaikittisilp and T. Okubo, Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials, *Nature communications*, 2019, **10**, 4459, DOI: 10.1038/s41467-019-12394-0.

[85] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma and E. Olivetti, A machine learning approach to zeolite synthesis enabled by automatic literature data extraction, *ACS central science*, 2019, **5**, 892–899, DOI: 10.1021/acscentsci.9b00193.

[86] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, Capturing chemical intuition in synthesis of metal-organic frameworks, *Nature communications*, 2019, **10**, 539, DOI: 10.1038/s41467-019-08483-9.

[87] L. C. Blum and J.-L. Reymond, 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, *Journal of the American Chemical Society*, 2009, **131**, 8732–8733, DOI: 10.1021/ja902302h.

[88] T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, Quantitative structure–property relationship modeling of diverse materials properties, *Chemical reviews*, 2012, **112**, 2889–2919, DOI: 10.1021/cr200066h.

[89] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl and V. Svetnik, Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships, *Journal of Chemical Information and Modeling*, 2015, **55**, 263–274, DOI: 10.1021/ci500747n.

[90] D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, 2010, **50**, 742–754, DOI: 10.1021/ci100050t.

[91] 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 and R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction, *Journal of Chemical Information and Modeling*, 2019, **59**, 3370–3388, DOI: 10.1021/acs.jcim.9b00237.

[92] T. J. Wills, D. A. Polshakov, M. C. Robinson and A. A. Lee, Impact of Chemist-In-The-Loop Molecular Representations on Machine Learning Outcomes, *Journal of Chemical Information and Modeling*, 2020, **60**, 4449–4456, DOI: 10.1021/acs.jcim.0c00193.

[93] G. Schneider and U. Fechner, Computer-based de novo design of drug-like molecules, *Nature Reviews Drug Discovery*, 2005, **4**, 649–663, DOI: 10.1038/nrd1799.

[94] M. Aldeghi and C. W. Coley, A graph representation of molecular ensembles for polymer property prediction, *Chemical Science*, 2022, **13**, 10486–10498, DOI: 10.1039/D2SC02839E.

[95] M. Karelson, V. S. Lobanov and A. R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chemical Reviews*, 1996, **96**, 1027–1044, DOI: 10.1021/cr950202r.

[96] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, MoleculeNet: a benchmark for molecular machine learning, *Chemical Science*, 2018, **9**, 513–530, DOI: 10.1039/C7SC02664A.

[97] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Modeling*, 1988, **28**, 31–36, DOI: 10.1021/ci00057a005.

[98] D. Weininger, A. Weininger and J. L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, *Journal of Chemical Information and Computer Sciences*, 1989, **29**, 97–101, DOI: 10.1021/ci00062a008.

[99] D. Weininger, SMILES. 3. DEPICT. Graphical depiction of chemical structures, *Journal of Chemical Information and Modeling*, 1990, **30**, 237–243, DOI: 10.1021/ci00067a005.

[100] G. Landrum *et al.*, *RDKit: Open-source cheminformatics*, https://www.rdkit.org, 2015, Accessed: March 12, 2022.

[101] M. Krenn, F. Häse, A. K. Nigam, P. Friederich and A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation, *Machine Learning: Science and Technology*, 2020, **1**, 045024, DOI: 10.1088/2632-2153/ABA947.

[102] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik and R. P. Adams, Advances in Neural Information Processing Systems, 2015.

[103] H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service, *Journal of Chemical Documentation*, 1965, **5**, 107–113, DOI: 10.1021/c160017a018.

[104] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nature Communications*, 2017, **8**, 13890, DOI: 10.1038/ncomms13890.

[105] L. Wilbraham, E. Berardo, L. Turcani, K. E. Jelfs and M. A. Zwijnenburg, High-throughput screening approach for the optoelectronic properties of conjugated polymers, *Journal of chemical information and modeling*, 2018, **58**, 2450–2459, DOI: 10.1021/acs.jcim.8b00256.

[106] W. J. Youngblood, S.-H. A. Lee, Y. Kobayashi, E. A. Hernandez-Pagan, P. G. Hoertz, T. A. Moore, A. L. Moore, D. Gust and T. E. Mallouk, Photoassisted overall water splitting in a visible light-absorbing dye-sensitized photoelectrochemical cell, *Journal of the American Chemical Society*, 2009, **131**, 926–927, DOI: 10.1021/ja809108y.

[107] D. A. Nicewicz and T. M. Nguyen, *Recent applications of organic dyes as photoredox catalysts in organic synthesis*, 2014, https://pubs.acs.org/doi/full/10.1021/cs400956a, DOI: 10.1021/cs400956a.

[108] S. Fantacci, F. De Angelis, A. Sgamellotti, A. Marrone and N. Re, Photophysical properties of $[Ru(phen)_2(dppz)]^{2+}$ Intercalated into DNA: an integrated Car-Parrinello and TDDFT study, *Journal of the American Chemical Society*, 2005, **127**, 14144–14145, DOI: 10.1021/ja054368d.

[109] A. Vlček Jr and S. Záliš, Modeling of charge-transfer transitions and excited states in d6 transition metal complexes by DFT techniques, *Coordination Chemistry Reviews*, 2007, **251**, 258–287, DOI: 10.1016/j.ccr.2006.05.021.

[110] F. De Angelis, S. Fantacci and A. Selloni, Alignment of the dye's molecular levels with the TiO2 band edges in dye-sensitized solar cells: a DFT–TDDFT study, *Nanotechnology*, 2008, **19**, 424002, DOI: 10.1088/0957-4484/19/42/424002.

[111] M. Grätzel, Recent advances in sensitized mesoscopic solar cells, *Accounts of chemical research*, 2009, **42**, 1788–1798, DOI: 10.1021/ar900141y.

[112] D. Jacquemin, E. A. Perpete, G. E. Scuseria, I. Ciofini and C. Adamo, TD-DFT performance for the visible absorption spectra of organic dyes: conventional versus long-range hybrids, *Journal of chemical theory and computation*, 2008, **4**, 123–135, DOI: 10.1021/ct700187z.

[113] A. Dreuw, J. L. Weisman and M. Head-Gordon, Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange, *The Journal of chemical physics*, 2003, **119**, 2943–2946.

[114] A. D. Becke, A new mixing of Hartree–Fock and local density-functional theories, *The Journal of chemical physics*, 1993, **98**, 1372–1377.

[115] T. Yanai, D. P. Tew and N. C. Handy, A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP), *Chemical physics letters*, 2004, **393**, 51–57, DOI: 10.1016/j.cplett.2004.06.011.

[116] C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *The Journal of chemical physics*, 1999, **110**, 6158–6170.

[117] B. J. Lynch, P. L. Fast, M. Harris and D. G. Truhlar, Adiabatic connection for kinetics, *The Journal of Physical Chemistry A*, 2000, **104**, 4811–4815, DOI: 10.1021/jp000497z.

[118] C. Møller and M. S. Plesset, Note on an approximation treatment for many-electron systems, *Physical review*, 1934, **46**, 618, DOI: 10.1103/PhysRev.46.618.

[119] X. Wang, L. Chen, S. Y. Chong, M. A. Little, Y. Wu, W.-H. Zhu, R. Clowes, Y. Yan, M. A. Zwijnenburg, R. S. Sprick *et al.*, Sulfone-containing covalent organic frameworks for photocatalytic hydrogen evolution from water, *Nature Chemistry*, 2018, **10**, 1180–1189, DOI: 10.1038/s41557-018-0141-5.

[120] V. L. Deringer, M. A. Caro, R. Jana, A. Aarva, S. R. Elliott, T. Laurila, G. Csányi and L. Pastewka, Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory, *Chemistry of Materials*, 2018, **30**, 7438–7445, DOI: 10.1021/acs.chemmater.8b02410.

[121] M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, *Physical review letters*, 2012, **108**, 058301, DOI: 10.1103/PhysRevLett.108.058301.

[122] J. Behler, Perspective: Machine learning potentials for atomistic simulations, *The Journal of Chemical Physics*, 2016, **145**, 170901, DOI: 10.1063/1.4966192.

[123] A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Physical Review B*, 2013, **87**, 184115, DOI: 10.1103/PhysRevB.87.184115.

[124] J. S. Smith, O. Isayev and A. E. Roitberg, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, *Chemical Science*, 2017, **8**, 3192–3203, DOI: 10.1039/C6SC05720A.

[125] B. Huang and O. A. von Lilienfeld, Ab Initio Machine Learning in Chemical Compound Space, *Chemical Reviews*, 2021, **121**, 10001–10036, DOI: 10.1021/acs.chemrev.0c01303.

[126] K. P. Murphy, *Machine learning: a probabilistic perspective*, the MIT press, 2012.

[127] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors, *Nature*, 1986, **323**, 533–536, DOI: 10.1038/323533a0.

[128] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Research*, 2011, **40**, D1100–D1107, DOI: 10.1093/nar/gkr777.

[129] L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *Journal of chemical information and modeling*, 2012, **52**, 2864–2875, DOI: 10.1021/ci300415d.

[130] T. Sander, J. Freyss, M. von Korff and C. Rufener, DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis, *Journal of Chemical Information and Modeling*, 2015, **55**, 460–473, DOI: 10.1021/ci500588j.

[131] J. B. Tenenbaum, V. d. Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, 2000, **290**, 2319–2323, DOI: 10.1126/science.290.5500.2319.

[132] N. Halko, P.-G. Martinsson and J. A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM review*, 2011, **53**, 217–288, DOI: 10.1137/090771806.

[133] J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, 1964, **29**, 1–27, DOI: 10.1007/BF02289565.

[134] L. van der Maaten and G. Hinton, Visualizing Data using t-SNE, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

[135] L. McInnes, J. Healy and J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv e-prints*, 2018, DOI: 10.48550/arxiv.1802.03426.

[136] H. Xiao, K. Rasul and R. Vollgraf, Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, *arXiv e-prints*, 2017, DOI: 10.48550/ARXIV.1708.07747.

[137] D. Probst and J.-L. Reymond, Visualization of very large high-dimensional data sets as minimum spanning trees, *Journal of Cheminformatics*, 2020, **12**, 12, DOI: 10.1186/s13321-020-0416-x.

[138] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, Taking the Human Out of the Loop: A Review of Bayesian Optimization, *Proceedings of the IEEE*, 2016, **104**, 148–175, DOI: 10.1109/JPROC.2015.2494218.

[139] A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet *et al.*, Nanomole-scale high-throughput chemistry for the synthesis of complex molecules, *Science*, 2015, **347**, 49–53, DOI: 10.1126/science.1259203.

[140] Z. Yao, Y. Lum, A. Johnston, L. M. Mejia-Mendoza, X. Zhou, Y. Wen, A. Aspuru-Guzik, E. H. Sargent and Z. W. Seh, Machine learning for a sustainable energy future, *Nature Reviews Materials*, 2023, **8**, 202–215, DOI: 10.1038/s41578-022-00490-5.

[141] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006.

[142] C. S. Smith, K. Jouravleva, M. Huisman, S. M. Jolly, P. D. Zamore and D. Grunwald, An automated Bayesian pipeline for rapid analysis of single-molecule binding data, *Nature Communications*, 2019, **10**, 272, DOI: 10.1038/s41467-018-08045-5.

[143] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath and A. Varnek, Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge, *Journal of Chemical Information and Modeling*, 2015, **55**, 84–94, DOI: 10.1021/ci500575y.

[144] G. Ivosev, L. Burton and R. Bonner, Dimensionality Reduction and Visualization in Principal Component Analysis, *Analytical Chemistry*, 2008, **80**, 4933–4944, DOI: 10.1021/ac800110w.

[145] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, Functional materials discovery using energy-structure-function maps, *Nature*, 2017, **543**, 657–664, DOI: 10.1038/nature21419.

[146] B. J. Frey and D. Dueck, Clustering by Passing Messages Between Data Points, *Science*, 2007, **315**, 972–976, DOI: 10.1126/science.1136800.

[147] T. pandas development team, *Pandas: powerful Python data analysis toolkit*, 2020, DOI: 10.5281/zenodo.3509134.

[148] P. T. Inc., *Plotly*, https://plotly.com/, 2012, Accessed: March 12, 2022.

[149] P. T. Inc., *Dash*, https://dash.plotly.com/, 2017, Accessed: March 12, 2022.

[150] *Jmol: an open-source Java viewer for chemical structures in 3D*, https://jmol.sourceforge.net/, 2021, Accessed: March 12, 2022.

[151] X. Li, P. M. Maffettone, Y. Che, T. Liu, L. Chen and A. I. Cooper, Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules, *Chemical Science*, 2021, **12**, 10742–10754, DOI: 10.1039/D1SC02150H.

[152] S. De, A. P. Bartók, G. Csányi and M. Ceriotti, Comparing molecules and solids across structural and alchemical space, *Physical Chemistry Chemical Physics*, 2016, **18**, 13754–13769, DOI: 10.1039/C6CP00415F.

[153] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, DScribe: Library of descriptors for machine learning in materials science, *Computer Physics Communications*, 2020, **247**, 106949, DOI: 10.1016/j.cpc.2019.106949.

[154] X. Wu, Y. Che, L. Chen, E. J. Amigues, R. Wang, J. He, H. Dong and L. Ding, Mapping the Porous and Chemical Structure-Function Relationships of Trace CH_3I Capture by Metal-Organic Frameworks using Machine Learning, *ACS Applied Materials & Interfaces*, 2022, **14**, 47209–47221, DOI: 10.1021/acsami.2c10861.

[155] ISO/IEC, *SQL: Structured Query Language*, https://www.iso.org/standard/63555.html, 2016, Accessed: March 12, 2022.

[156] M. Inc., *MongoDB*, https://www.mongodb.com/, 2007–2021, Accessed: March 12, 2022.

[157] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Computational Materials Science*, 2013, **68**, 314–319, DOI: 10.1016/j.commatsci.2012.10.028.

[158] X. Wang, K. Maeda, A. Thomas, K. Takanabe, G. Xin, J. M. Carlsson, K. Domen and M. Antonietti, A metal-free polymeric photocatalyst for hydrogen production from water under visible light, *Nature Materials*, 2009, **8**, 76–80, DOI: 10.1038/nmat2317.

[159] M. Z. Rahman, M. G. Kibria and C. B. Mullins, Metal-free photocatalysts for hydrogen evolution, *Chemical Society Reviews*, 2020, **49**, 1887–1931, DOI: 10.1039/C9CS00313D.

[160] Y. Wang, A. Vogel, M. Sachs, R. S. Sprick, L. Wilbraham, S. J. A. Moniz, R. Godin, M. A. Zwijnenburg, J. R. Durrant, A. I. Cooper and J. Tang, Current understanding and challenges of solar-driven hydrogen generation using polymeric photocatalysts, *Nature Energy*, 2019, **4**, 746–760, DOI: 10.1038/s41560-019-0456-5.

[161] Y. Wang, X. Wang and M. Antonietti, Polymeric Graphitic Carbon Nitride as a Heterogeneous Organocatalyst: From Photochemistry to Multipurpose Catalysis to Sustainable Chemistry, *Angewandte Chemie International Edition*, 2012, **51**, 68–89, DOI: 10.1002/anie.201101182.

[162] N. A. Romero and D. A. Nicewicz, Organic Photoredox Catalysis, *Chemical Reviews*, 2016, **116**, 10075–10166, DOI: 10.1021/acs.chemrev.6b00057.

[163] X. Li, S. T. Melissen, T. Le Bahers, P. Sautet, A. F. Masters, S. N. Steinmann and T. Maschmeyer, Shining light on carbon nitrides: leveraging temperature to understand optical gap variations, *Chemistry of Materials*, 2018, **30**, 4253–4262.

[164] S. Ren, M. J. Bojdys, R. Dawson, A. Laybourn, Y. Z. Khimyak, D. J. Adams and A. I. Cooper, Porous, Fluorescent, Covalent Triazine-Based Frameworks Via Room-Temperature and Microwave-Assisted Synthesis, *Advanced Materials*, 2012, **24**, 2357–2361, DOI: https://doi.org/10.1002/adma.201200751.

[165] R. S. Sprick, Z. Chen, A. J. Cowan, Y. Bai, C. M. Aitchison, Y. Fang, M. A. Zwijnenburg, A. I. Cooper and X. Wang, Water Oxidation with Cobalt-Loaded Linear Conjugated Polymer Photocatalysts, *Angewandte Chemie International Edition*, 2020, **59**, 18695–18700, DOI: 10.1002/anie.202008000.

[166] M. J. Frisch *et al.*, *Gaussian 16, Revision A.03*, 2016.

[167] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K.-i. Shimizu, Machine learning for catalysis informatics: recent applications and prospects, *ACS Catalysis*, 2019, **10**, 2260–2297, DOI: 10.1021/acscatal.9b04186.

[168] K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, Big-data science in porous materials: materials genomics and machine learning, *Chemical reviews*, 2020, **120**, 8066–8129, DOI: 10.1021/acs.chemrev.0c00004.

[169] T. Lazarides, T. McCormick, P. Du, G. Luo, B. Lindley and R. Eisenberg, Making hydrogen from water using a homogeneous system without noble metals, *Journal of the American Chemical Society*, 2009, **131**, 9192–9194, DOI: 10.1021/ja903044n.

[170] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, A mobile robotic chemist, *Nature*, 2020, **583**, 237–241, DOI: 10.1038/s41586-020-2442-2.

[171] C. W. Coley, N. S. Eyke and K. F. Jensen, Autonomous discovery in the chemical sciences part I: Progress, *Angewandte Chemie International Edition*, 2020, **59**, 22858–22893, DOI: 10.1002/anie.201909987.

[172] C. W. Coley, N. S. Eyke and K. F. Jensen, Autonomous discovery in the chemical sciences part II: outlook, *Angewandte Chemie International Edition*, 2020, **59**, 23414–23436, DOI: 10.1002/anie.201909989.

[173] F. Pedregosa, G. Varoquaux, A. Gramfort, B. M. V., Thirion, O. Grisel, M. Blondel, R. P. P., Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

[174] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035.

[175] T. Head, M. Kumar, H. Nahrstaedt, G. Louppe and I. Shcherbatyi, *scikit-optimize/scikit-optimize*, 2021, https://doi.org/10.5281/zenodo.5565057, DOI: 10.5281/zenodo.5565057.

[176] C. K. Graetzel and M. Graetzel, Hydrogen evolution from the photolysis of alcoholic benzophenone solutions via redox catalysis, *Journal of the American Chemical Society*, 1979, **101**, 7741–7743, DOI: 10.1021/ja00520a032.

[177] A.-H. Mau, O. Johansen and W. Sasse, Xanthene dyes as sensitizers for the photoreduction of water, *Photochemistry and Photobiology*, 1985, **41**, 503–509, DOI: 10.1111/j.1751-1097.1985.tb03519.x.

[178] Z.-J. Yu, W.-Y. Lou, H. Junge, A. Päpcke, H. Chen, L.-M. Xia, B. Xu, M.-M. Wang, X.-J. Wang, Q.-A. Wu *et al.*, Thermally activated delayed fluorescence (TADF) dyes as efficient organic photosensitizers for photocatalytic water reduction, *Catalysis Communications*, 2019, **119**, 11–15, DOI: 10.1016/j.catcom.2018.09.018.

[179] H. Mai, T. C. Le, D. Chen, D. A. Winkler and R. A. Caruso, Machine learning for electrocatalyst and photocatalyst design and discovery, *Chemical Reviews*, 2022, **122**, 13478–13515, DOI: 10.1021/acs.chemrev.2c00061.

[180] Z. Zuo, D. T. Ahneman, L. Chu, J. A. Terrett, A. G. Doyle and D. W. C. MacMillan, Merging photoredox with nickel catalysis: Coupling of α-carboxyl sp^3-carbons with aryl halides, *Science*, 2014, **345**, 437–440, DOI: 10.1126/science.1255525.

[181] M. H. Shaw, J. Twilton and D. W. MacMillan, Photoredox catalysis in organic chemistry, *The Journal of organic chemistry*, 2016, **81**, 6898–6926, DOI: 10.1021/acs.joc.6b01449.

[182] L. Marzo, S. K. Pagire, O. Reiser and B. König, Visible-light photocatalysis: does it make a difference in organic synthesis?, *Angewandte Chemie International Edition*, 2018, **57**, 10034–10072, DOI: 10.1002/anie.201709766.

[183] D. A. Nicewicz and D. W. MacMillan, Merging photoredox catalysis with organocatalysis: the direct asymmetric alkylation of aldehydes, *Science*, 2008, **322**, 77–80, DOI: 10.1126/science.1161976.

[184] A. Studer and D. P. Curran, The electron is a catalyst, *Nature Chemistry*, 2014, **6**, 765–773, DOI: 10.1038/nchem.2031.

[185] J. C. Tellis, D. N. Primer and G. A. Molander, Single-electron transmetalation in organoboron cross-coupling by photoredox/nickel dual catalysis, *Science*, 2014, **345**, 433–436, DOI: 10.1126/science.1253647.

[186] R. W. Pipal, K. T. Stout, P. Z. Musacchio, S. Ren, T. J. Graham, S. Verhoog, L. Gantert, T. G. Lohith, A. Schmitz, H. S. Lee *et al.*, Metallaphotoredox aryl and alkyl radiomethylation for PET ligand discovery, *Nature*, 2021, **589**, 542–547, DOI: 10.1038/s41586-020-3015-0.

[187] X. Jiang, W. Xiong, S. Deng, F.-D. Lu, Y. Jia, Q. Yang, L.-Y. Xue, X. Qi, J. A. Tunge, L.-Q. Lu *et al.*, Construction of axial chirality via asymmetric radical trapping by cobalt under visible light, *Nature Catalysis*, 2022, **5**, 788–797, DOI: 10.1038/s41929-022-00831-1.

[188] Z. Dong and D. W. MacMillan, Metallaphotoredox-enabled deoxygenative arylation of alcohols, *Nature*, 2021, **598**, 451–456, DOI: 10.1038/s41586-021-03920-6.

[189] A. McNally, C. K. Prier and D. W. MacMillan, Discovery of an α-amino C–H arylation reaction using the strategy of accelerated serendipity, *Science*, 2011, **334**, 1114–1117, DOI: 10.1126/science.1213920.

[190] V. Mdluli, S. Diluzio, J. Lewis, J. F. Kowalewski, T. U. Connell, D. Yaron, T. Kowalewski and S. Bernhard, High-throughput synthesis and screening of iridium (III) photocatalysts for the fast and chemoselective dehalogenation of aryl bromides, *ACS Catalysis*, 2020, **10**, 6977–6987, DOI: 10.1021/acscatal.0c02247.

[191] A. Hantzsch, Condensationsprodukte aus Aldehydammoniak und ketonartigen Verbindungen, *Berichte der deutschen chemischen Gesellschaft*, 1881, **14**, 1637–1638.

[192] A. Tlili and S. Lakhdar, Acridinium salts and cyanoarenes as powerful photocatalysts: opportunities in organic synthesis, *Angewandte Chemie*, 2021, **133**, 19678–19701, DOI: 10.1002/ange.202102262.

[193] J. Luo and J. Zhang, Donor–acceptor fluorophores for visible-light-promoted organic synthesis: Photoredox/Ni dual catalytic C(sp^3)– C(sp^2) cross-coupling, *ACS Catalysis*, 2016, **6**, 873–877, DOI: 10.1021/acscatal.5b02204.

[194] R. W. Kennard and L. A. Stone, Computer aided design of experiments, *Technometrics*, 1969, **11**, 137–148, DOI: 10.1080/00401706.1969.10490666.

[195] S. M. Lundberg and S.-I. Lee, Advances in Neural Information Processing Systems, 2017.

[196] J. Westermayr and P. Marquetand, Machine learning for electronically excited states of molecules, *Chemical Reviews*, 2021, **121**, 9873–9926, DOI: 10.1021/acs.chemrev.0c00749.

[197] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens, *Journal of Chemical Theory and Computation*, 2020, **16**, 4192–4202, DOI: 10.1021/ct200523a.

[198] X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials, *Journal of chemical information and modeling*, 2020, **60**, 3408–3415, DOI: 10.1021/acs.jcim.0c00451.

[199] J. Řezáč, K. E. Riley and P. Hobza, Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries, *Journal of Chemical Theory and Computation*, 2011, **7**, 3466–3470, DOI: 10.1021/ct200523a.

[200] J. Řezáč, K. E. Riley and P. Hobza, S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures, *Journal of chemical theory and computation*, 2011, **7**, 2427–2438, DOI: 10.1021/ct2002946.

[201] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, *The Journal of Chemical Physics*, 2011, **134**, 074106, DOI: 10.1063/1.3553717.

[202] J.-D. Chai and M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections, *Physical Chemistry Chemical Physics*, 2008, **10**, 6615–6620, DOI: 10.1039/B810189B.

[203] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, Less is more: Sampling chemical space with active learning, *The Journal of chemical physics*, 2018, **148**, 241733, DOI: doi.org/10.1063/1.5023802.

[204] T. Risthaus and S. Grimme, Benchmarking of London dispersion-accounting density functional theory methods on very large molecular complexes, *Journal of chemical theory and computation*, 2013, **9**, 1580–1591, DOI: 10.1021/ct301081n.

[205] J. Řezáč, Y. Huang, P. Hobza and G. J. Beran, Benchmark calculations of three-body intermolecular interactions and the performance of low-cost electronic structure methods, *Journal of chemical theory and computation*, 2015, **11**, 3065–3079, DOI: 10.1021/acs.jctc.5b00281.

[206] A. D. Boese, Density functional theory and hydrogen bonds: are we there yet?, *ChemPhysChem*, 2015, **16**, 978–985, DOI: 10.1002/cphc.201402786.

[207] Q. Zhu, J. Johal, D. E. Widdowson, Z. Pang, B. Li, C. M. Kane, V. Kurlin, G. M. Day, M. A. Little and A. I. Cooper, Analogy powered by prediction and structural invariants: computationally led discovery of a mesoporous hydrogen-bonded organic cage crystal, *Journal of the American Chemical Society*, 2022, **144**, 9893–9901, DOI: 10.1021/jacs.2c02653.

[208] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties, *Physical Review B*, 1998, **58**, 7260, DOI: 10.1103/PhysRevB.58.7260.

[209] J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Physical review letters*, 1996, **77**, 3865, DOI: 10.1103/PhysRevLett.77.3865.

[210] T. D. Bennett, F.-X. Coudert, S. L. James and A. I. Cooper, The changing state of porous materials, *Nature Materials*, 2021, **20**, 1179–1187, DOI: 10.1038/s41563-021-00957-w.

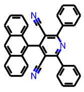
[211] J. D. Dunitz and A. Gavezzotti, Molecular recognition in organic crystals: directed intermolecular bonds or nonlocalized bonding?, *Angewandte Chemie International Edition*, 2005, **44**, 1766–1787, DOI: 10.1002/anie.200460157.

[212] L. Liu and A. Corma, Confining isolated atoms and clusters in crystalline porous materials for catalysis, *Nature Reviews Materials*, 2021, **6**, 244–263, DOI: 10.1038/s41563-021-00957-w.

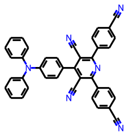
[213] D. H. Bowskill, I. J. Sugden, S. Konstantinopoulos, C. S. Adjiman and C. C. Pantelides, Crystal structure prediction methods for organic molecules: State of the art, *Annual Review of Chemical and Biomolecular Engineering*, 2021, **12**, 593–623, DOI: 10.1146/annurev-chembioeng-060718-030256.

APPENDIX A

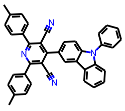
A.1 Molecular structures of the experimentally tested CNPs



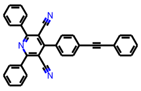
ID: 3 Step: 0



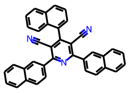
ID: 65 Step: 0



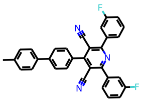
ID: 129 Step: 0



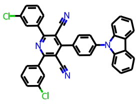
ID: 11 Step: 0



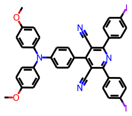
ID: 336 Step: 0



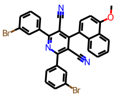
ID: 415 Step: 0



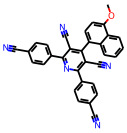
ID: 439 Step: 1



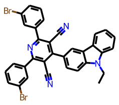
ID: 494 Step: 1



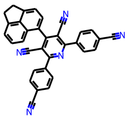
ID: 372 Step: 1



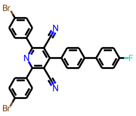
ID: 64 Step: 1



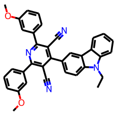
ID: 379 Step: 1



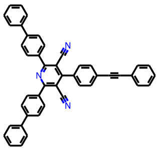
ID: 58 Step: 1



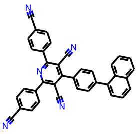
ID: 110 Step: 2



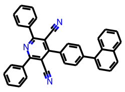
ID: 463 Step: 2



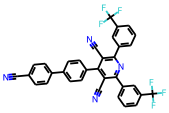
ID: 207 Step: 2



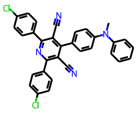
ID: 69 Step: 2



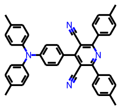
ID: 13 Step: 2



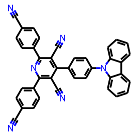
ID: 559 Step: 2



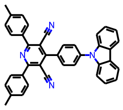
ID: 302 Step: 3



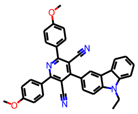
ID: 137 Step: 3



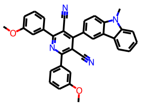
ID: 75 Step: 3



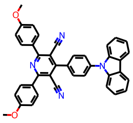
ID: 131 Step: 3



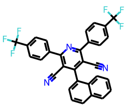
ID: 239 Step: 4



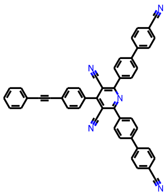
ID: 464 Step: 4



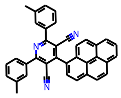
ID: 243 Step: 4



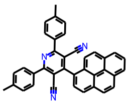
ID: 168 Step: 4



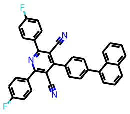
ID: 263 Step: 4



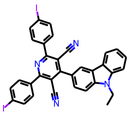
ID: 328 Step: 4



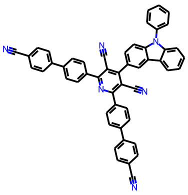
ID: 132 Step: 4



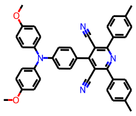
ID: 153 Step: 4



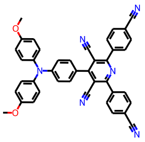
ID: 491 Step: 5



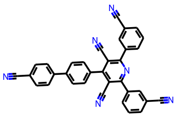
ID: 269 Step: 5



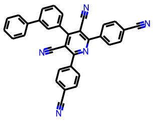
ID: 130 Step: 5



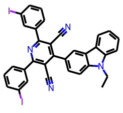
ID: 74 Step: 5



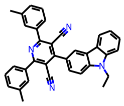
ID: 55 Step: 5



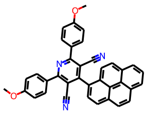
ID: 62 Step: 5



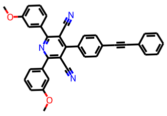
ID: 127 Step: 6



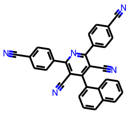
ID: 323 Step: 6



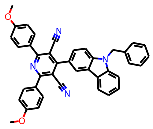
ID: 244 Step: 6



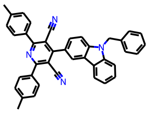
ID: 459 Step: 6



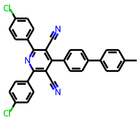
ID: 56 Step: 6



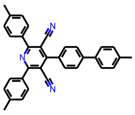
ID: 234 Step: 6



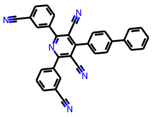
ID: 122 Step: 7



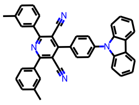
ID: 303 Step: 7



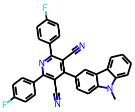
ID: 135 Step: 7



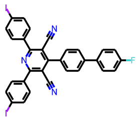
ID: 29 Step: 7



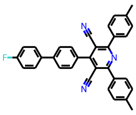
ID: 327 Step: 7



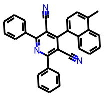
ID: 156 Step: 7



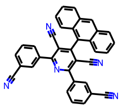
ID: 502 Step: 7



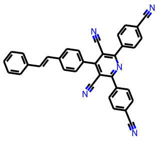
ID: 138 Step: 7



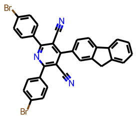
ID: 21 Step: Structure selection



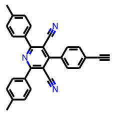
ID: 31 Step: Structure selection



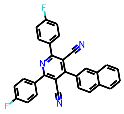
ID: 68 Step: Structure selection



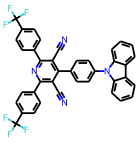
ID: 98 Step: Structure selection



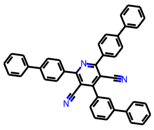
ID: 116 Step: Structure selection



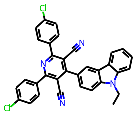
ID: 145 Step: Structure selection



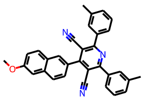
ID: 187 Step: Structure selection



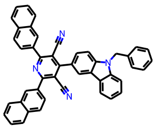
ID: 202 Step: Structure selection



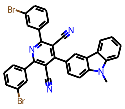
ID: 295 Step: Structure selection



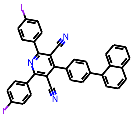
ID: 315 Step: Structure selection



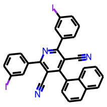
ID: 346 Step: Structure selection



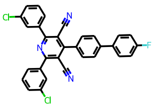
ID: 380 Step: Structure selection



ID: 489 Step: Structure selection

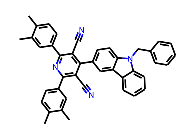


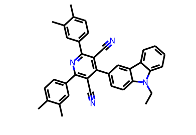
ID: 504 Step: Structure selection

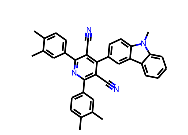


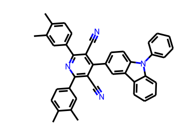
ID: 446 Step: Structure selection

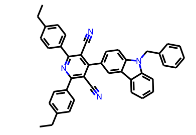
A.2 Molecular structures of 100 CNPs designed for the search of selected CNP with stronger negative reduction potentials

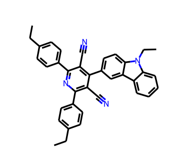

ID: 560

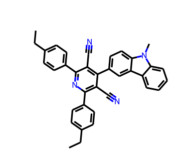

ID: 561

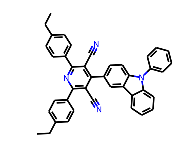

ID: 562

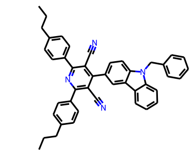

ID: 563

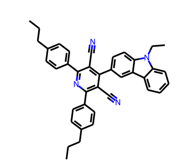

ID: 564

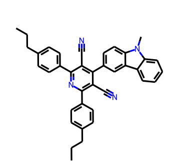

ID: 565

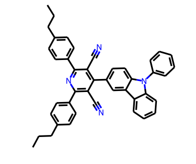

ID: 566

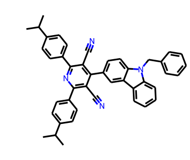

ID: 567

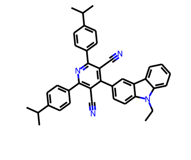

ID: 568

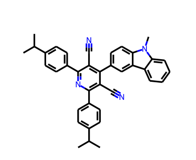

ID: 569

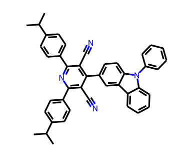

ID: 570

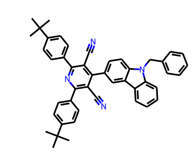

ID: 571

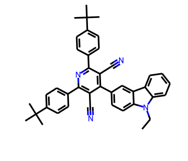

ID: 572

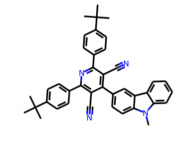

ID: 573

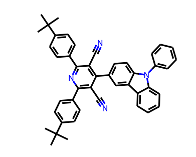

ID: 574

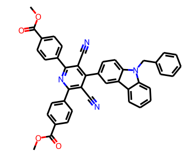

ID: 575

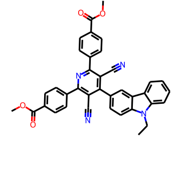

ID: 576

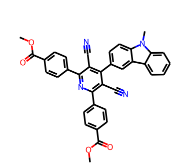

ID: 577

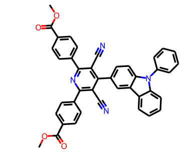

ID: 578

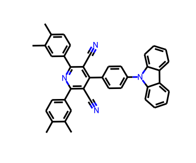

ID: 579

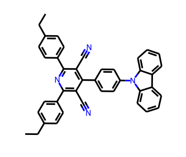

ID: 580

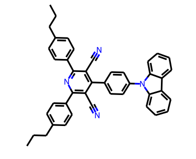

ID: 581

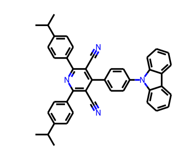

ID: 582

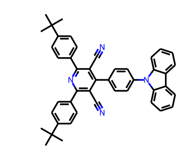

ID: 583

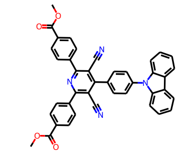

ID: 590

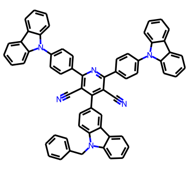

ID: 591

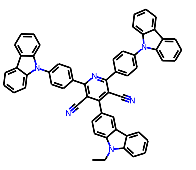

ID: 592

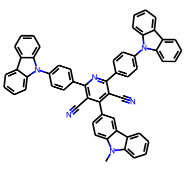

ID: 593

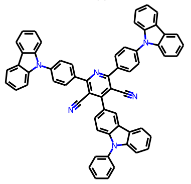

ID: 594

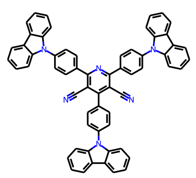

ID: 595

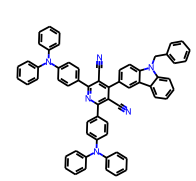

ID: 596

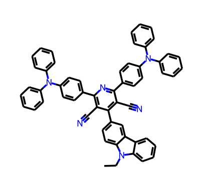

ID: 597

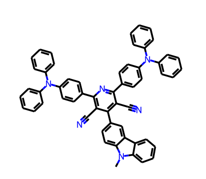

ID: 598

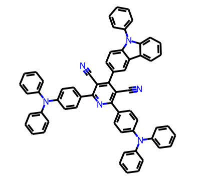

ID: 599

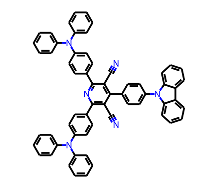

ID: 600

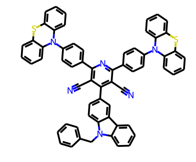

ID: 601

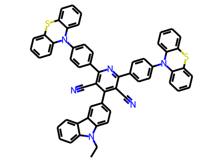

ID: 602

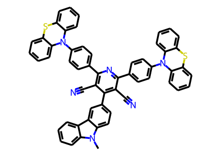

ID: 603

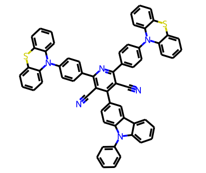

ID: 604

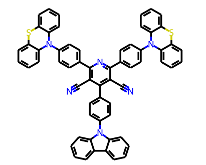

ID: 605

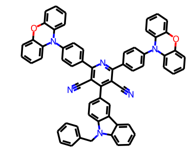

ID: 606

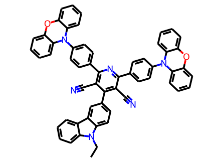

ID: 607

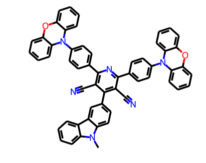

ID: 608

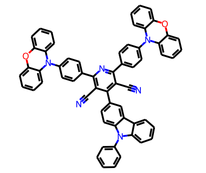

ID: 609

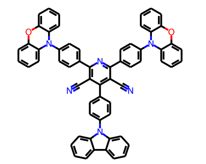

ID: 610

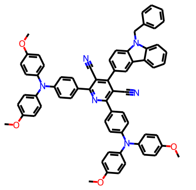

ID: 611

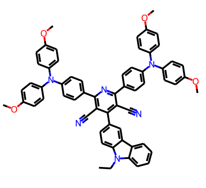

ID: 612

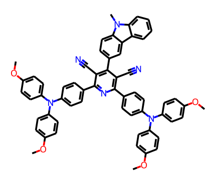

ID: 613

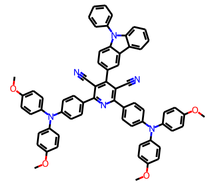

ID: 614

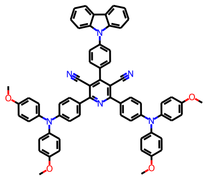

ID: 615

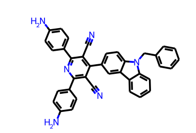

ID: 616

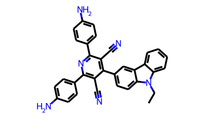

ID: 617

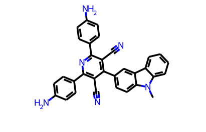

ID: 618

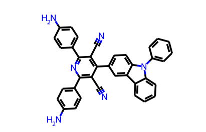

ID: 619

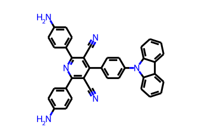

ID: 620

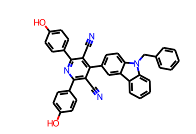

ID: 621

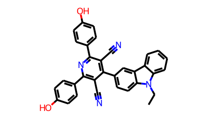

ID: 622

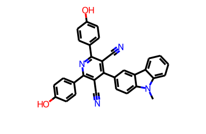

ID: 623

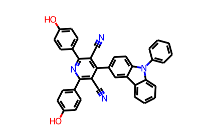

ID: 624

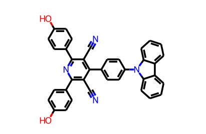

ID: 625

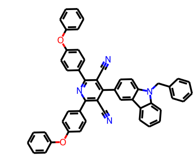

ID: 626

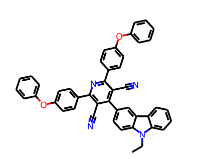

ID: 627

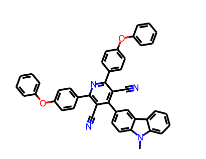

ID: 628

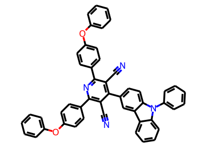

ID: 629

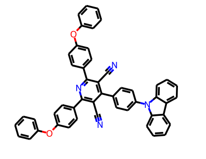

ID: 630

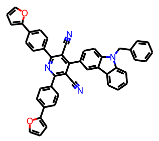

ID: 631


ID: 632

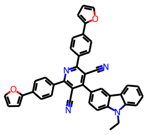

ID: 633


ID: 634

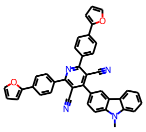

ID: 635



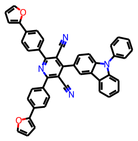
ID: 636



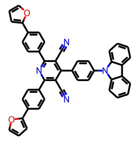
ID: 637



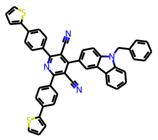
ID: 638



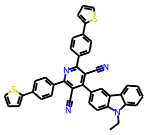
ID: 639



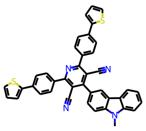
ID: 640



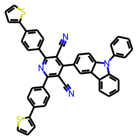
ID: 641



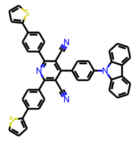
ID: 642



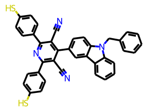
ID: 643



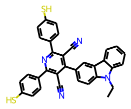
ID: 644



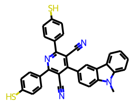
ID: 645



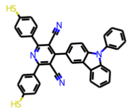
ID: 646



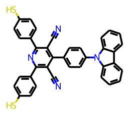
ID: 647



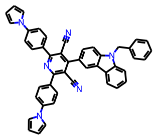
ID: 648



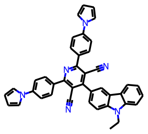
ID: 649



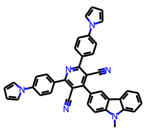
ID: 650



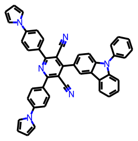
ID: 651



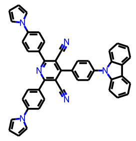
ID: 652



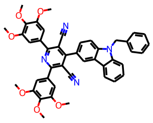
ID: 653



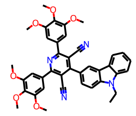
ID: 654



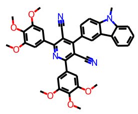
ID: 655



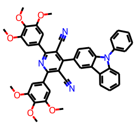
ID: 656



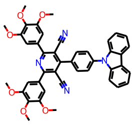
ID: 657



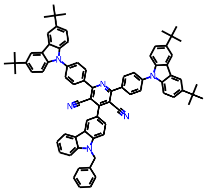
ID: 658



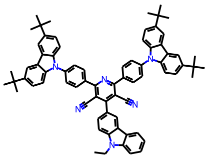
ID: 659



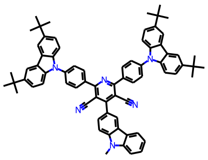
ID: 660



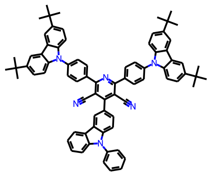
ID: 661



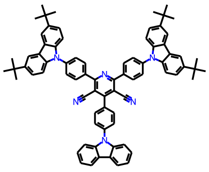
ID: 662



ID: 663



ID: 664



ID: 665