Universitat Autònoma de Barcelona

STUDY AND APPLICABILITY OF MACHINE LEARNING POTENTIALS

Sergi Ortiz Ropero

sergi.ortizr@autonoma.cat

Treball final del Grau en Física

Doble grau en Física i Química

March 10, 2025

Abstract

The aim of this work is to study Neural network potential (NNP) models, a novel approach to atomistic modeling in which, rather than performing time-consuming quantum mechanical calculations to obtain the energy and interactions of a given system, a Machine Learning (ML) model is trained on quantum mechanical results and used to predict these system properties solely from the atomic arrangement. In order to prove useful as physical models, these should capture the basic physics of any system, as well as the 'rules' of quantum mechanics to successfully describe atomic interactions. However, it is not clear how or up to what point the basic physical principles have been *learnt* or *understood* by the models. We begin by exploring and interpreting their basic physical knowledge, how it has been implemented and their limitations through a series of test cases. We then move on to showcase their applicability and usefulness in the context of chemistry, biophysics and materials science. posar el que s'ha obtingut com a resultats. ANI and MACE are physically adequate, while ORb is not, and has been found to have severe fundamental issues.

Introduction

Several fields of science such as novel material design, protein structure[1], drug discovery[2,3], and catalysis[REF NEEDED], have seen important development in the past few decades. To study these large atomic systems at a temperature and pressure of interest, an accurate description of atomic interactions is needed to correctly explore the phase space of the system, as in the context of statistical mechanics, through *in silico* simulations, to predict new materials, drugs and catalysts.

To model these interactions, one can take an *ab initio* (first principles) approach and solve the Schrödinger equation[4] of the system. This, however, is a complex task that can become prohibitively expensive with system size, which leads to necessary approximations of the description of the interactions and ultimately trading accuracy for computational efficiency[5].

ML has seen widespread adoption and increased applicability in many fields of science, including Physics[6,7], and is an active area of research to this day[8]. In this work, we will study ML interatomic potentials (MLIPs), models trained on quantum mechanical results to predict the atomic interactions solely based on the atomic representation, promising to bring near-QM accuracy at much lower computational cost[5] than *ab initio* approaches.

Using ML in physics differs from using ML in other fields in a very subtle way, it not only means training a model, but a *physically adequate* one: at its core, it must capture the underlying 'effective' physics (e.g. the 'rules' of quantum mechanics), but most importantly, conservation laws, symmetries and invariances of the system should be satisfied and exploited by the model.

We ... (explicar què faig). The aim of this project is two fold. First, to study three MLIP models to assess the physical suitability of these through a series of tests designed to understand what 'effective' physics each model has captured and how they have been implemented. Second, to showcase their practical potential and usefulness in an array of areas and motivate their adoption.

This project is organized as follows. Section 1 motivates the problem while providing a brief modeling and ML theoretical foundation. Section 2 introduces the models and associated descriptors. Section 3 designs the tests to prove the understanding of the models and Section 4 discusses its results. Finally, section 5 showcases their applicability through a set of examples in chemistry and materials science.

En la part del com, explicar approximacions estandard, volem susbtituir aquests calculs costosos amb un novel approach amb ML de forma més fàcil i simplificada.

1 Atomistic modeling and MLIPs

Atomistic modeling aims to describe the interactions within an atomic configuration. This is done by computing the potential energy of the system and from it, the forces. Fundamentally, as we are dealing with a set of atoms, but most importantly electrons, atomic interactions of a given system (forces and energies) and its properties can be derived directly from the system wavefunction $|\Psi\rangle$, obtainable from the time-dependent Schrödinger equation[4,9]

$$i\hbar\frac{\partial}{\partial t}|\Psi\rangle = \hat{H}|\Psi\rangle \ , \tag{1}$$

with $\hat{H}$ the system Hamiltonian. Note that in many cases, $\hat{H}$ is time independent and thus 'only' the time-independent Schrödinger equation (TISE), $\hat{H}|\Psi\rangle = E|\Psi\rangle$, needs to be solved. Numerical solution of the TISE is usually computationally unfeasible and not always applicable. Thus, alternative methods and approximations must be used to obtain the energies and system and electronic wavefunctions $|\Psi\rangle$ and $|\psi\rangle$, respectively, known in the context of Quantum Chemistry as *ab initio* Electronic Structure methods[10] (discussed in greater detail in Appendix A).

Most importantly, the computational scaling of these methods is $\mathcal{O}(n^4)$ with n the number of electrons (as seen in Fig. 1 for Hartree-Fock theory) or larger for high-level methods[11,12], making the evaluation of atomic interactions with ES methods unfeasible for large systems, such as proteins in solution or metallic surfaces with point defects. For these, compromised accuracy in exchange for better computational efficiency is needed, often requiring a $\mathcal{O}(N)$ scaling with the number of atoms, N. This alternative description of the interactions use *interatomic potentials* or *force fields* (known as molecular mechanics FFs), in which, given the N atomic coordinates, $\mathbf{r}^N$, the potential energy of the system is determined through a physics-inspired parametrized functional form[12,13], such as

$$\begin{aligned} U(\mathbf{r}^N) \equiv \overbrace{\sum_{b}^{\text{bonds}} k_b(r-r_b)^2 + \sum_{a}^{\text{angles}} k_a(\theta-\theta_a)^2 + \sum_{\phi}^{\substack{\text{proper}\\\text{dihedrals}}} k_\phi(1+\cos(n\phi-\gamma)) + \sum_{\Phi}^{\substack{\text{improper}\\\text{dihedrals}}} k_\Phi(r-r_\Phi)^2}^{\text{bonded interactions}} \\ \underbrace{+\sum_{i}^{N}\sum_{j>i}^{N}\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] + \sum_{i}^{N}\sum_{j>i}^{N}\frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}}_{\text{non-bonded interactions}} \ , \end{aligned} \tag{2}$$

from which the forces can be computed as the gradient of U. Note that FFs consist of two clearly distinguished contributions to the potential energy of the system.

Bonded interactions describe short-range interactions between bonded atoms. Bonds and angles are modeled with a harmonic potential, while proper (usual) 4-atom dihedrals are modeled as a Fourier series of cosines. Improper dihedrals represent 4-atom bonded contributions and are used to impose planarity between a central atom and its three bonded neighbors modeled with a harmonic potential of the solid angle.

Non-bonded terms consider pairwise combination of atoms (not directly linked) modeling electrostatics with Coulomb's law (where a point charge is assigned at each atom's position)

and dispersion with a Lennard-Jones[14] potential. Additional multipole and many-body expansion terms can be included to better model long-range and dispersion interactions, as in polarizable FFs[15,16].

In modeling, the description accuracy of U plays a crucial role in the quality of *in silico* simulations, and thus, of all the physical properties that can be extracted from them[17], *inter alia*, thermodynamic, structural and dynamic properties.

It is important to note that, while *ab initio* methods only require the geometry (and spin) of the system to compute the potential energy of the configuration, conventional MM-FFs require extensive system-dependent non-trivial parametrization (derived from QM reference calculations or empirically) of all bonded and non-bonded contributions of Eq. (2). Thus, these lack an 'out-of-the-box' applicability between systems (property known as 'transferability', where the same functional form is used for different systems) as a preconceived notion of the bonding patterns and correct parameters is needed, hindering its extension to complex ones.

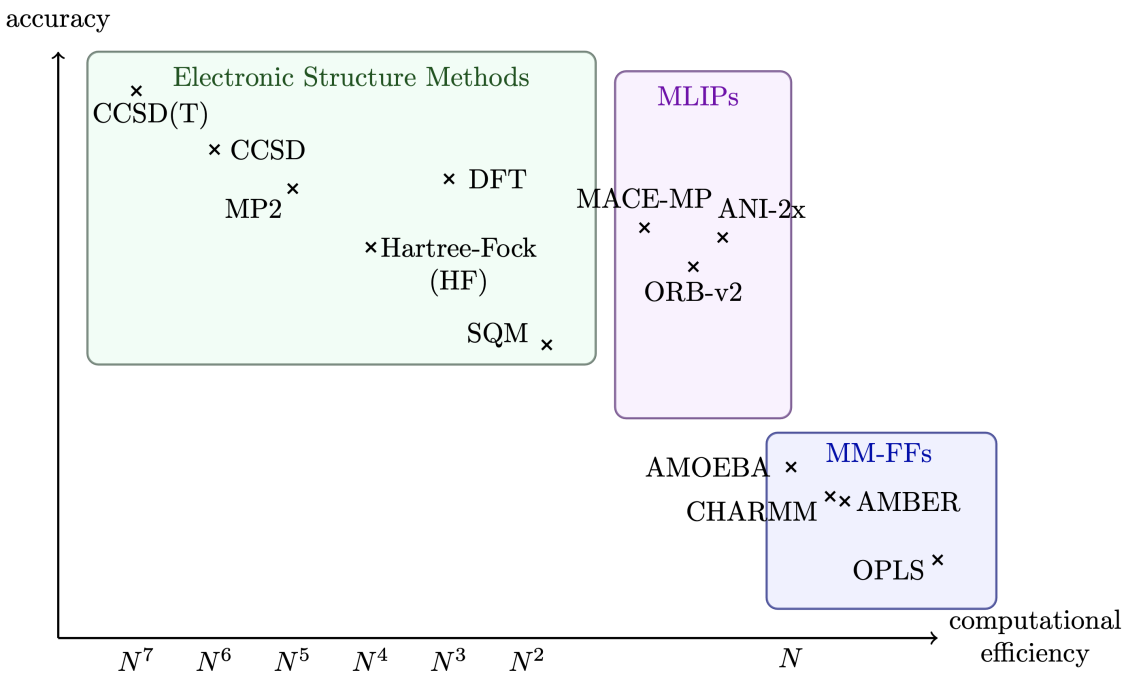


Figure 1: Accuracy of $U(\mathbf{r}^N)$ and approximate scalability of each method with respect to the number of atoms N [for ES methods, the scaling is formally with the number of basis functions...N provides a starting point]. Most common ES methods (AM1[18], MNDO[19], PM6[20], DBT-B[21]), FFs [REFS], and three MLIP models[22–24] are represented. Note that all the NNPs are substantially more accurate compared to MM-FFs while having similar computational cost. NNPs also surpass semiempirical QM (SQM) methods while at a fraction of the cost. PFFs such as AMOEBA[16] offer better long-range descriptions, but come at a higher cost.

Ideally, we would be interested in obtaining a transferable potential energy functional form able to describe the interactions at QM accuracy with FF computational cost. It is precisely in this task where ML can prove useful. MLIPs are data-driven ML transferable FFs trained on QM reference data that *learn* the mapping between a chemical representation (describing an atomic arrangement) and the potential energy without relying on a preconceived notion of fixed chemical bonds or knowledge about the relevant interactions, providing near-QM quality at near-FF cost.

Fig. 1 shows how transferable MLIPs bridge the gap between ES methods and conventional approximate MM-FF, with near linear scaling with system size, while being much more accurate than traditional FFs, due to them being trained on high-level QM data. These models show great applicability in theoretical studies involving modeling of large systems (where a hybrid ML/MM has been proposed[25–27] metals[22] and complex systems where parameters are unavailable or where a higher-level of accuracy is needed but performing ES calculations is prohibitively expensive.

Una de les limtiacions dels FFs: FF tenen fisica bàsica ja implementada (no crear ni trencar enllaços), fan servir coordenades internes, ja és rototranslacionalment invariant. En models de ML aquestes coses es poden perdre.... si fem servir NN potser.

2 Machine Learning and Deep Learning algorithms

A machine learning algorithm is one able to *learn* from data. The core concepts of a ML algorithm are the *task* to be performed, the *features* used and extracted from data, and the *model*, which is *trained* on the features to accomplish the task with a measured performance. In this sense, the mapping between input and output of the model is not hard-coded as in a traditional algorithm, but rather, it is *learnt from data*. The core concepts are defined as follows[28–30]

- **The task.** ML algorithms allow us to tackle tasks that are too difficult to solve with fixed hand-designed programs (e.g. image classification). The task is what the model is expected be able to perform after learning, rather than the learning itself. The models studied in this project will focus on regression.
- **The data.** To train a ML model, data is needed. The dataset is a collection of *examples* (or *data points*), consisting of a set of *features* that have been quantitatively measured from some object or event that we want the ML model to process.
- **The performance.** To assess the abilities of a ML algorithm, a task-specific quantitative measure of the performance must be designed, called a loss or objective function $\mathcal{L}$. For a classification task, that could be the model accuracy, or the Mean Squared Error (MSE) for regression.
- **The learning.** ML algorithms can be broadly categorized as *unsupervised* or *supervised* (although other categories exist). Supervised learning algorithms learn from example features from the dataset, but, additionally, each is also associated with a *label* (or target), provided by the 'teacher', which is to be correctly predicted by the model. In a regression task to predict molecular energies, an example consists of features (atomic positions, atomic number and other descriptors) and a label (the energy or forces).
- **The model**. Given the training data, a goal, and a training procedure; an architecture is chosen and trained to fulfill the task, producing a model. The choice of architecture is highly dependent on the *task*, and contains a series of trainable parameters, θ, often classified as weights w, and biases, b, which are optimized during training by maximizing the performance (minimizing the loss function).
- **Generalization**. The ultimate goal of the model is to correctly generalize to previously unseen data: if a model has been trained on predicting the energies for an atomic

arrangement, whether it also correctly predicts the energies for new (not in the training set) ones. For a regression task, the *generalization error* is defined to be the error between predicted and true value of previously unseen examples.

Deep Learning (DL) is a particular kind of machine learning in which the model learns a complex representation of the data based on a hierarchy of concept, each concept defined as a function of simpler concepts, which allows it to build an abstract representation computed in terms of less abstract ones[31].

2.1 Neural Networks and the Multilayer Perceptron

The quinetessential example of a ML / DL model is the artificial Neural Network (NN) and the multilayer perceptron (MLP) models[32–34], respectively. A NN is a non-linear model formed by a set of nodes (or neurons), each holding a value. The NN of Figure 2a consists of two visible input and output layers, with two hidden layers of $\{\ell_1, \ell_2\}$ nodes. The node values are computed in a feed-forward way from the previous layer values. For the first hidden layer,

$$h_i^{(1)} = \sigma\left(b_i^{(1)} + \sum_j^{\text{features}} w_{ij}^{(1)} x_j\right), \qquad \text{or equivalently,} \qquad \mathbf{h}^{(1)} = \sigma\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right), \tag{3}$$

as seen in Fig. 2b. The depth of the NN is the number of hidden layers (formally for $\ell \geq 3$, it is called a MLP) and deeper networks allow each layer to extract more abstract information out of the data from previous layers allowing it to learn highly complex feature spaces efficiently.

To assess the performance, a loss function is defined (e.g. MSE), and is minimized during training by optimizing the trainable parameters θ,

$$\mathcal{L}_{\text{MSE}}(\theta) = \min_\theta \sum_i^{\text{examples}} \left[\mathbf{y}_{\text{pred}}(\mathbf{x}_i, \theta) - \mathbf{y}_i\right]^2, \qquad \nabla_\theta \mathcal{L} \to 0\,. \tag{4}$$

Every model contains a set of **hyperparameters**, which are not defined in training, e.g. the number of nodes in a hidden layer or the number of hidden layers. These can affect the generalization error and must be optimized separate to training.

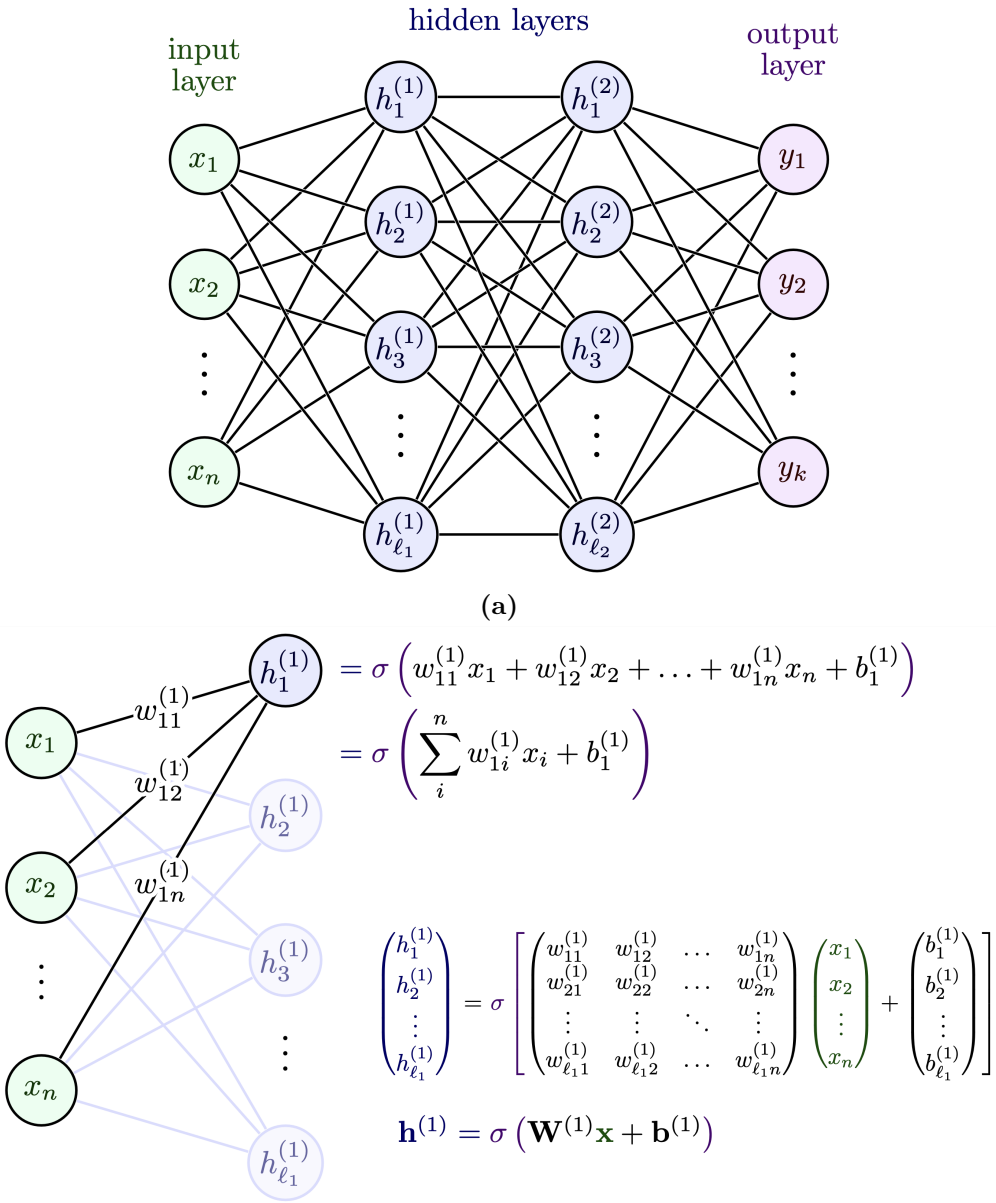


Figure 2: **(a)** Feed-forward NN with two hidden layers of $\{\ell_1, \ell_2\}$ nodes each. The input layer takes an example, consisting of n $\{x_i\}$ features. The result is passed to the output layer, consisting of k y_i labels. In a feed-forward way, the input is passed to the next layer until it reaches the output layer. The trainable parameters θ are trained until the loss function $\mathcal{L}$ optimizes the performance of the network. **(b)** Information passing mechanism in an artificial NN. Each node in the next layer takes the inputs of all the nodes of the previous layer and performs a non-linear transformation. In the image, $\mathbf{x}$ is the input feature vector, $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ the weight matrix and bias vector of the first hidden layer, respectively (trainable parameters of the model), and σ is a non-linear activation function, such as $\text{ReLU}(x)$.

2.2 Graph Neural Networks (GNNs)

Often, the architecture used in a ML model is linked to the data format or structure of a specific problem, with the objective of obtaining data-efficient models, ones that can take advantage and exploit the structure and symmetries in the data. GNNs take advantage of data that can be structured as a *graph*. A first intuitive example is molecules, which can be naturally modeled as graphs, where we can think of each atom as a node and each bond as an edge connecting the nodes, as seen in Fig. 3.

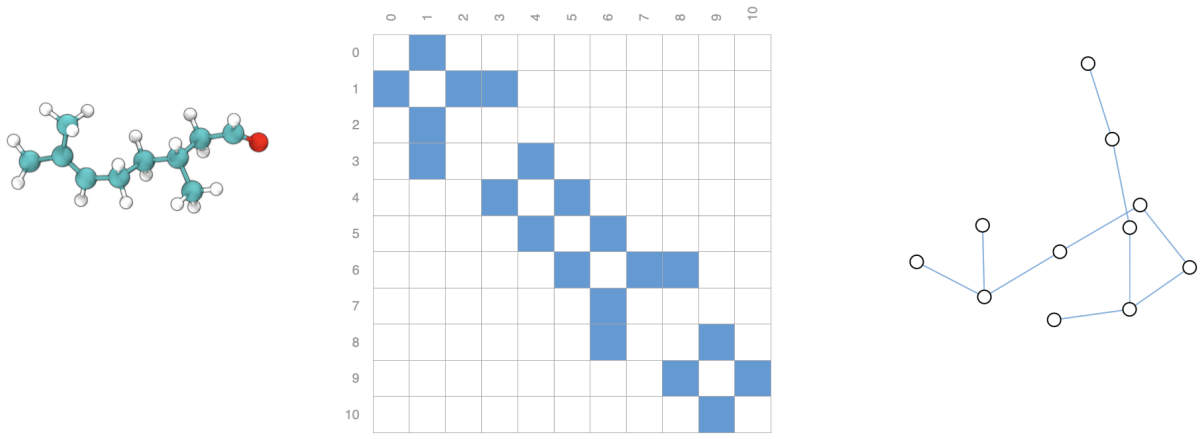


Figure 3: [foto provisional] An *a priori* perspective on how molecules can be naturally represented as graphs. Each node represents an atom and contains a node feature vector $\mathbf{x}_i$ with atomic attributes (e.g. atomic number, atomic mass, partial charge). Edges are identified with bonds, containing edge feature vectors $\mathbf{e}_{ij}$ with attributes about bond type (single, double, triple bond, etc.). The connectivity of the graph is captured from the atomic connectivity, and stored in the adjacency matrix $\mathbf{A}$.

We encourage the reader to refer to Appendix C to understand Graph Neural Networks (GNNs), and more specifically Message Passing Neural Networks, MPNNs, a state-of-the-art architecture of neural network, that is being implemented in a wide variety of fields in science with great success[REFS APPLICACIONS].

2.3 Typical ML process

The first step in any ML application, irrespective of the problem, task, dataset, architecture, and training procedure, is to define the problem and task. Rather than hard-coding the algorithm, the human intervention is in the *design of the model*. The nature of the problem will determine the data and features that can be extracted and ultimately influence the design process. In it, we must choose how the dataset is preprocessed, what data is fed to the model and how, the architecture, the loss function and devise the training procedure to ultimately minimize generalization error. Once the trainable parameters have been optimized, the model can be benchmarked on new unseen data to evaluate the final performance on the task.

2.4 Data preprocessing and choice of representation

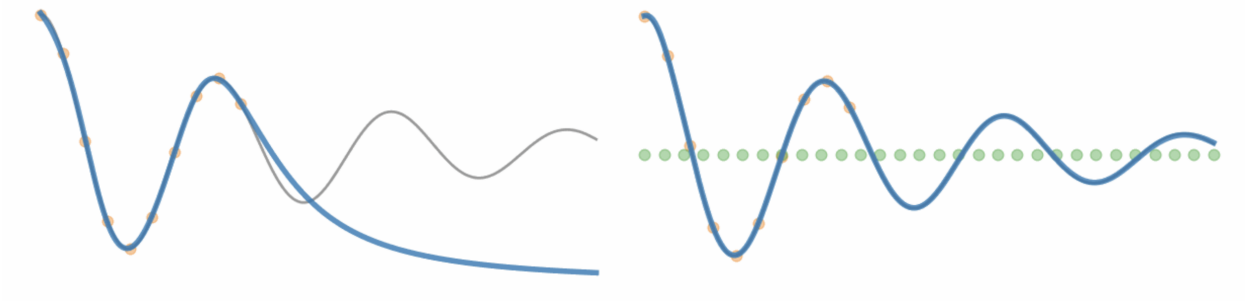
The collection and curation of the dataset the model will be fed is arguably as important as model design itself. It must contain both the right information and representation to facilitate the training and improve the generalization of the model.

In preprocessing, raw data is curated into a set of useful features (process known as 'feature engineering') and how these features are represented (i.e. the choice of representation) can have a large impact on training performance, effectively limiting what can be learnt from the data (e.g. performing an integral in a random set of coordinates vs. ones designed to make the problem easier). Ultimately, it is not only the collection of data that is important, but rather what features are selected and how they are represented such that the model trains as effectively as possible. Other practical considerations regarding model complexity can be found in Appendix B.

3 Physics Informed Neural Networks (PINNs)

això ho faria un apartat més dintre de ML, i igual amb els subapartats

NN models have been successfully applied to a variety of scientific applications. However, most of these approaches are 'data-driven', where a model is trained solely based on data.



(a) Data-driven naïve approach **(b)** Physics-informed approach

Figure 4: Despite having the same training set, a physics-informed model is able to generalize better to times outside training, as well as providing a more generalizable model than purely data-driven approaches. Note that, due to the lack of basic physics understanding, the data-driven approach violates energy conservation and presents an un-physical trajectory. Extracted from Ref. [ref].

Fig. 12c represents the trajectory prediction of an under-damped oscillator using the 'naïve' data-driven approach, in which, although the model correctly predicts the trajectory for times it was trained on, the generalization to other times is poor. This approach results on models with not only poor generalization, but also poor interpretability: if solely used as a black box, we cannot know (if the reference is unknown) if the model has understood the physics behind the data.

To solve this, PINNs[35–37] introduce some prior physics knowledge of the system (in this case, that the trajectory observed comes from an under-damped harmonic oscillator, and thus its corresponding PDE) as a form of training bias, altering the loss function to include both a 'data loss' (e.g. MSE) and 'physics loss' (deviation from the reference PDE) as seen in

$$\begin{aligned}\mathcal{L}(\theta) &= \mathcal{L}_{\text{data}}(\theta) + \lambda \mathcal{L}_{\text{physics}}(\theta) \\ &\equiv \underbrace{\frac{1}{N}\sum_{i}^{N}\left[y(x_i) - u_{\text{pred}}(x_i;\theta)\right]^2}_{\text{data loss, MSE}} + \lambda \underbrace{\frac{1}{M}\sum_{j}^{M}\left(\left[m\frac{d^2}{dx^2} + \mu\frac{d}{dx} + k\right]u_{\text{pred}}(x_j;\theta)\right)^2}_{\text{physics loss (prior physics knowledge needed)}}. \end{aligned} \tag{5}$$

with λ a regulation parameter [define N and M?]. This 'suggestion' on what the solution should physically resemble translates into more generalizable (see Fig. 12d) and interpretable models.

3.1 Transparency and interpretability of ML models

Even though the *design* basics of a ML model are clearly defined, a complex model can quickly become a dangerous 'black box', where the model designer looses track of 'what the model has learnt' about the underlying true representation of the data. In these cases, model

transparency and interpretability become of uttermost importance in determining what the model lacks and how it can be improved. As an example, a model that performs energy regression from an atomic arrangement, will output the energies, but it is important to understand 'what the model has understood' about the data to be able to interpret the energy output.

We, as ML practitioners but also physicists, are interested in models that can be applied to physical systems. As such, constrains derived directly from basic physics principles, such as symmetries and conservation laws must be imperative in order for the model to prove useful. For this, however, a first understanding of what physics has been captured is needed, and is why interpretable models are required.

4 Model architectures and representations

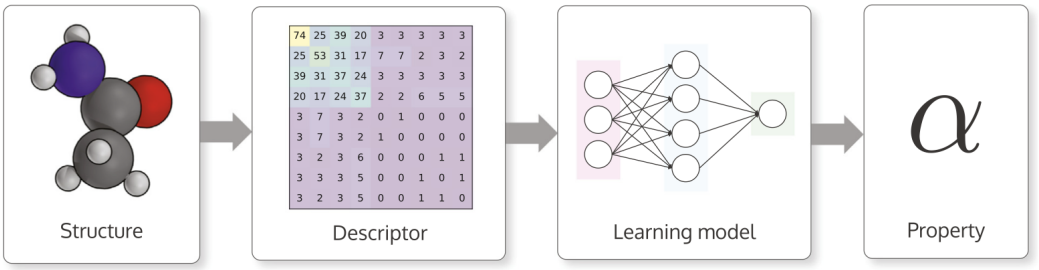


Fig. 1. Typical workflow for making machine learning based materials property predictions for atomistic structures. An atomic structure is transformed into a numerical representation called a descriptor. This descriptor is then used as an input for a machine learning model that is trained to output a property for the structure. There is also a possibility of combining the descriptor and learning model together into one inseparable step.

Figure 5

Aqui es tracta de definir el propi problema a tractar, el workflow de com passar de la representació a l'arquitectura a l'output. Introduir els models i l'arquitectura que es fa servir i plantejar la planificació del propi treball.

Fins aquí de moment no he introduit res de models de machine learning, només he explicat què són els MLIPs i per què models intereptables són interessants.

- Descriure el workflow general d'un MLIP
- Discutir la rellevància d'elaborar un model que invariant i que aprofiti les simetries.
- Discutir la localitat del sistema i com es pot descriure un model. Com es pot descriure l'atomic environment
- Quins models hi han (HDNNs i GNNs) (basics).
- Importance of body order descriptors in the description of chemical environements (papers CH4 i l'altre amb el gabor).

Aquí al text principal només es tracta d'introduir-ho de forma molt simplificada. Referenciar l'apèndix si es vol saber amb més detall.

MAX 3 PAGES. Fer una mica resum del que s'ha parlat a l'apèndix.

Three main aspects. In NNPs, we let the model learn the mapping, given a representation of the atomic positions, to the energy and forces of the whole molecule. It must be transferable...

Inductive biases (associated to the model itself, not the data nor the training) are meant to identify symmetries or regularities in the data and add modeling components that take advantage of these properties. In the case of MACE, we see that translations and rotations have the same energy due to a design choice (form of inductive bias).

Fer un extracte de l'apendix de descriptors.

fer molt clara la idea durant tot el text principal de per què és necessari que un model de ML sigui physically adequate. El seguent apartat de resultats té com a objectiu justificar i motivar les proves per intentar comprovar la adequacy dels models de ML.

4.1 Brief model introduction

Explicar quins models s'han fet servir

Despcriure les arquitectures que es poden fer servir, i els descriptors molt importants.

Si es vol, es pot explicar que hi ha més resultats amb més molècules i referenciar-ho a l'apèndix i o al github. "Altres molècules s'han testejat, obtenint resultats anàlegs. Aquests es poden trobar a l'Apèndix X i al github del projecte".

- ANI-family `ANI-1x`, `ANI-1ccx`, `ANI-1x`
- MACE-family `MACE-MP`, `MACE-OFF`
- Orbital-family `Orb-v2`, `Orb-D3-v2`

posar refs dels models. Necessito dir concretament quins elements suporten. Explicar cutoff distance and locality assumption.

NNPs not only must show a qualitatively correct description of atomic interactions, but must also be physically correct in order to be used as substitutes to proper physical models. Such physical adequacy checks are often neglected, which might lead to the usage of models that are not physically sound.

Symmetries of a system that, when applied, lead to essentially an equivalent one, must provide the same potential energy. This is the case for rotations and translations: performing a translation or rotation on any isolated system must leave the energy invariant.

Energy invariance can be achieved by constructing both invariant or equivariant models (explicar què és).

No se quan hauria de parlar de long-range interactions. Però hauria de. Amplicació de NNPs amb AMOEBA per tractar les missing long range interactions[38]

4.1.1 Effect of many-body correlations

A much more subtle consideration affects the local energy prediction due to how the chemical environment is *represented* (as discussed in Sec. 33d). Depending on what descriptors are used, two very distinct environments might be described as the same, severely impacting the energy prediction performance. Figure 26 shows how two sets of distinct chemical environments might be described as the same due to the body-order nature of the descriptors.

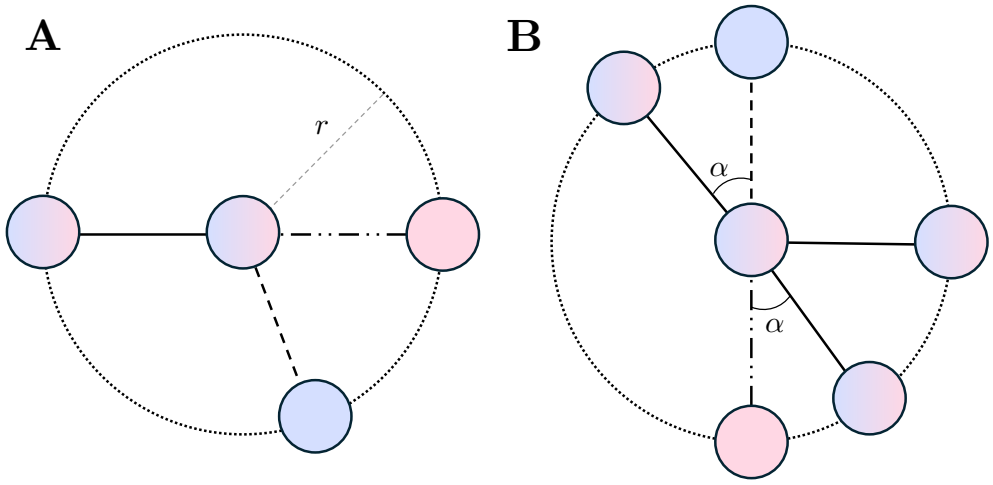


Figure 6: Different structures comprised of blue or red atoms. Atoms colored with a gradient are shared between both structures. (**A**) Both red and blue molecules are 3 atoms, the red one is linear while the blue one is angular-shaped. If only 2-body correlations are considered to represent the environment of the central atom, i.e. only distances, both structures would be identically described. Higher body-order correlations are needed to distinguish between them, such as angles (3-body order), which would show the linear vs. angular disparity. (**B**) Both red and blue structures are 5 atoms. From the central atom perspective, the environment is described by the same distances and angles (2 and 3-body order correlations). If distances and angles are used to describe the environment of the central atom, the structures are indistinguishable. These two examples **A** and **B** show the importance of the body-order of the interactions by a model, and its implications. Restricting the body-order limits the expessability of the model, ultimately hurting energy prediction. Example **B** is adapted from Refs. [39, 40]

Taula. Referenciar tot el que s'ha explicat a l'apendix i fer un breu resum. Fer la prova aquesta de H2O i CH4 en configuracions planes. ANI hauria de predir la mateixa energia (descriptors no suficients per distinguir, mentre que els altres dos haurien de poder).

No sé fins a quin punt seria rellevant fer aquestes proves, és més un test conceptual sobre models que no pas un resultat....

4.2 Applications

Citar moltes de les aplicabilitats

El que es pot fer, amb poc temps, és fer un scan amb el gaussian per comparar el temps de càlcul dft vs NNP, explicat abaix. Això em serviria per dir: ah! molt bé els NNPs són molt més rapids i ofereixen performance similar, que és la premisa que ha d'entendre el lector a banda de la necessitat d'un model physically adequate.

mesurar el temps que es triga a fer l'escan amb gaussian i amb els NNPs. A cada calcul gaussian et diu el CPU time. Python pots fer que et digui el temps, això serveix per dir: ab initio quality at half the cost.

4.3 Fundamental issues of NNPs

Explicar long range itneractions i el que deia l'article. Que tota la info de l'article estigui dins del treball.

5 Interpreting the models

The most important aspect of a ML model when applied to science is its physical adequacy and transferability, as ramarked previously, determined by how well it is able to 'capture' the fundamental constraints and symmetries of any given system. In this section, we aim to evaluate what 'basic physics' several models understand by performing simple yet relevant tests of increasing complexity.

The tests have been carried out using the `ASE`[41] python package, and the computational procedure and protocols followed throughout the tests can be found in the source code for the project, made available in the corresponding `GitHub` repository.

Important. Motivar cada exemple. Fonamentar molt bé cada càlcul, perquè l'he fet, què vull extreure i com el lligo amb tot. Si es tria un exemple, que no sigui random i que estigui fonamentat. Discutir molt bé els fonaments que hi han darrere.

5.1 Qualitative Analysis

A first important step to evaluate the adequacy and usability of NNPs as physical models is to qualitatively examine them to check for the presence of any fundamental flaws in the basic description of atomic interactions.

The following tests aim to find qualitative fundamental defects of interactions, in systems ranging from simple diatomic molecules to bond torsions of larger molecules, and to determine their range of applicability and where their description might fail and why.

5.1.1 2-body distortions. Distances

Two-body distortions can be described using its distance (internal coordinate), which only depends on two atoms. In Chemistry, bonds are formed due to a stabilization of the system. The general behavior for any type of bond is as follows. At $r \to \infty$ the stabilization is zero, as no formal interaction between the two fragments exists. As they come closer, stabilization increases due to delocalization of the electrons in a wider space, up to a minimum. Past the minimum, energy increases due to electron-electron and nuclear repulsion, increasing asimptotically as $r \to 0$. This behavior is qualitatively the same as a dispersion-based interaction described by a Lennard-Jones potential[14].

To assess the qualitative behavior of the NNPs on distances, a set of diatomic homonuclear molecules (X_2) have been deformed to obtain their qualitative potential shapes (Fig. 7), as well as other bond deformations of ethane and ethanol (Fig. 8).

Taking a glimpse at the qualitative potential shape of Fig. 8, all models correctly predict a minimum for ethane and ethanol, which is to be expected, as that is the region the models have been trained on. However, when deforming the bond, different behaviors appear. `MACE` and `Orbital` correctly predict the dissociation shape, which increases monotonously as the distance of the bond increases, providing a qualitatively correct dissociation limit. `ANI` models, however, either overestimate or underestimate the dissociation potential, missing the *essential* interactions. This qualitatively incorrect dissociation limit description is overly

apparent in homonuclears (Fig. 7). While `MACE` and `Orbital` capture the essential interaction, `ANI` provides different minima and completely misses the dissociation, including a large maximum between the correct minimum and the dissociated molecule.

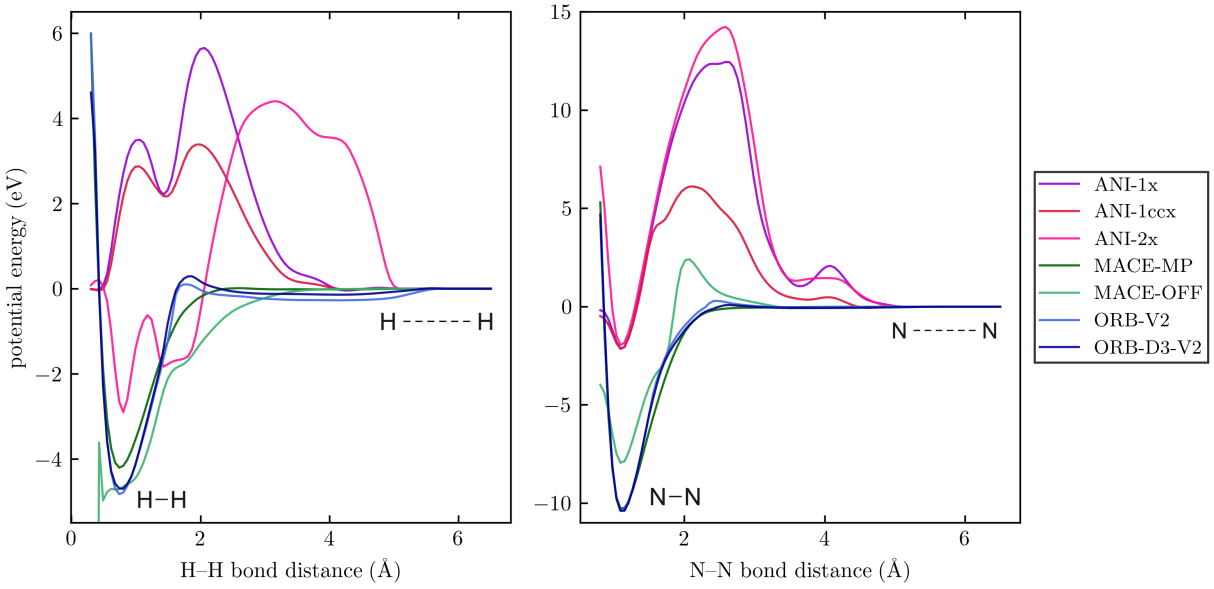
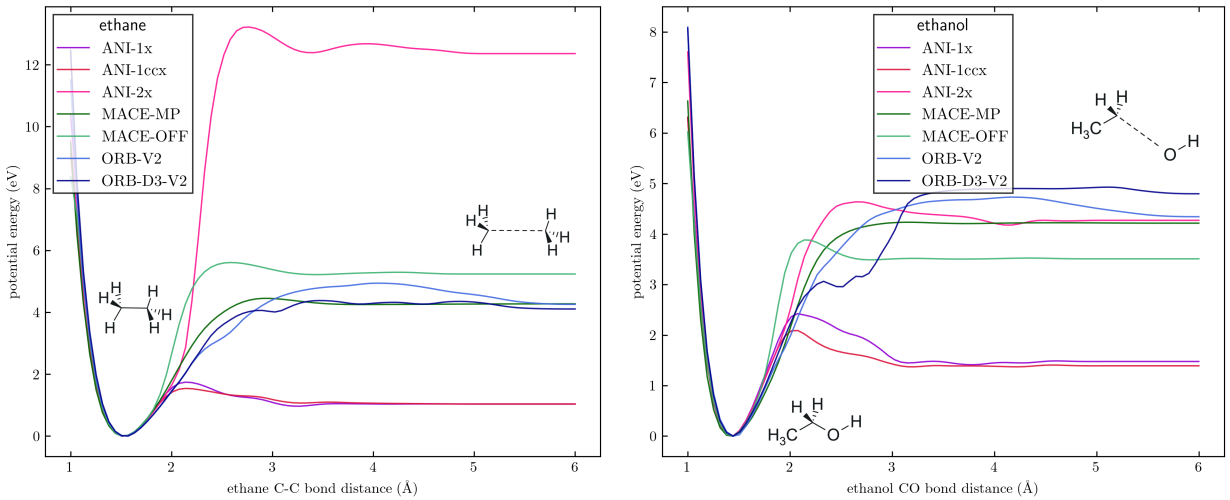


Figure 7: Dissociation descriptions of H_2 and N_2. Energy reference has been taken in the dissociation limit $r \to \infty$. `MACE` and `Orbital` models provide an overall qualitatively correct description of both the equilibrium distance of the molecule and its dissociation limit. `ANI` models, however, completely miss the dissociation of the molecule. Additional diatomic molecules, such as N_2, O_2, Si_2 and Cl_2 have been tested in Fig. 29 of Appendix G.



(a) ethane C–C bond deformation **(b)** ethanol C–O bond deformation

Figure 8: C–C and C–O bond deformation and dissociation of ethane and ethanol. Eenergies are referenced with respect to the energy minimum. The dissociation description is qualitatively correct, except for `ANI` potentials, where the dissociation energy is over- or underestimated.

The correct description of an equilibrium distance (deep minimum around it) and its dissociation is essential for applications such as modeling of chemical reactions, where bonds are formed and broken. In the case of bond formation, the great energy barrier `ANI` describes

in homonuclears would make it unusable for situations where non-equilibrium or deformed structures are relevant.

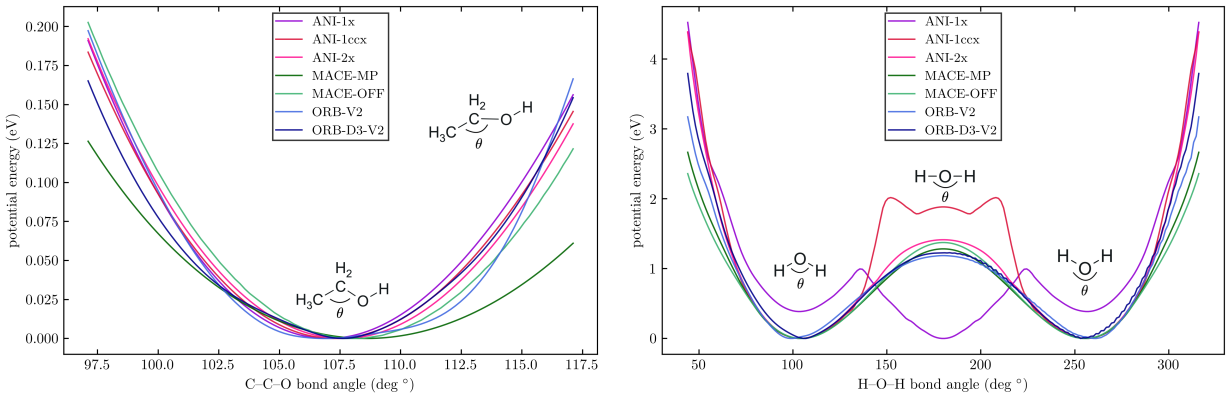
The qualitative misrepresentation of dissociations, however, is to be expected as none of the models have been trained on these high-energy deformed structures, only on the low-energy structures around the energy minimum (equilibrium distance).

Fig. 7 clearly shows the cutoff distance (design aspect of the models) to be around 5 Å. Past this point, according to the locality approximation, the two atoms 'do not see each other' and energy becomes analytically zero.

5.1.2 3-body distortions. Angles

A similar analysis can be carried out for three-body distortions, described in internal coordinates by an A-B-C angle θ. Small perturbations in internal coordinates with respect to the minimum energy structure are often described by a harmonic potential (up to second order Taylor expansion, ubiquitously used in physics to describe energy minima). Changes in angles can thus be modeled as $U \sim (\theta - \theta_{\text{ref}})^2$, with larger distortions breaking harmonicity.

Fig. 9 shows the qualitative behavior of the models when these angular distortions are applied to water and the CCO angle in ethanol. All models correctly describe small angular perturbations as a harmonic potential, as seen in Fig. 9a. When investigating the deformation of the HOH angle in water (Fig. 9b), two symmetric minimums appear, corresponding to the same, flipped, molecule, with small angular changes being similarly described as with ethanol. Larger angular deformations break harmonicity, as seen by the appearance of a quadratic potential. All models correctly describe two minima (angular waters) connected by an energy maximum (linear water), however, `ANI-1x` and `ANI-1ccx` completely misunderstand the linear HOH transition, predicting a linear water geometry, more stable than the angular one (`ANI-1x`), or exaggerating the maximum with small oscillations (`ANI-1ccx`). This failure is, once again, justified by the lack of training for large deformations, as linear water is far from equilibrium and not contained in the training set. It is interesting to note that `ANI-2x` correctly describes the linear transition, as this model improves torsional and angular deformations with respect to previous models.



(a) **(b)**

Figure 9: **(a)** Angular deformations for HOH angle in water and **(b)** CCO angle in ethanol.

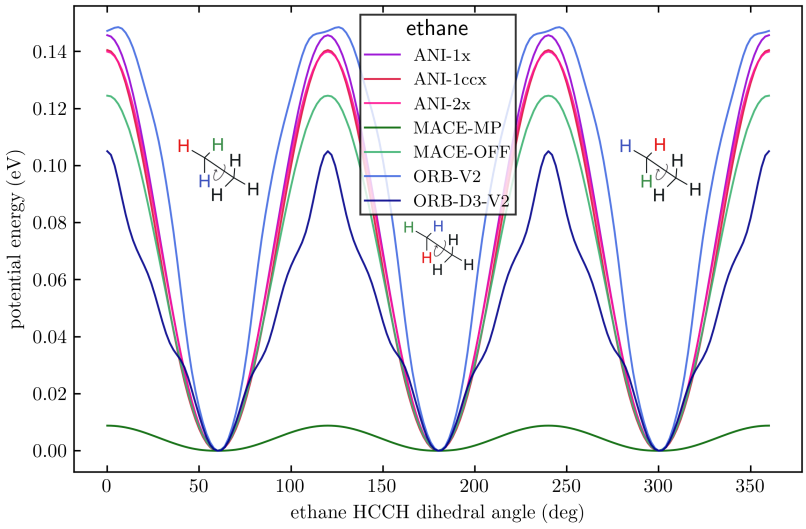
It is important to note that the potential energy shape must reflect the presence of certain

symmetries in the molecule, such as in water (Fig. 9b, $\alpha°$ and $(360 - \alpha)°$ are equivalent structures).

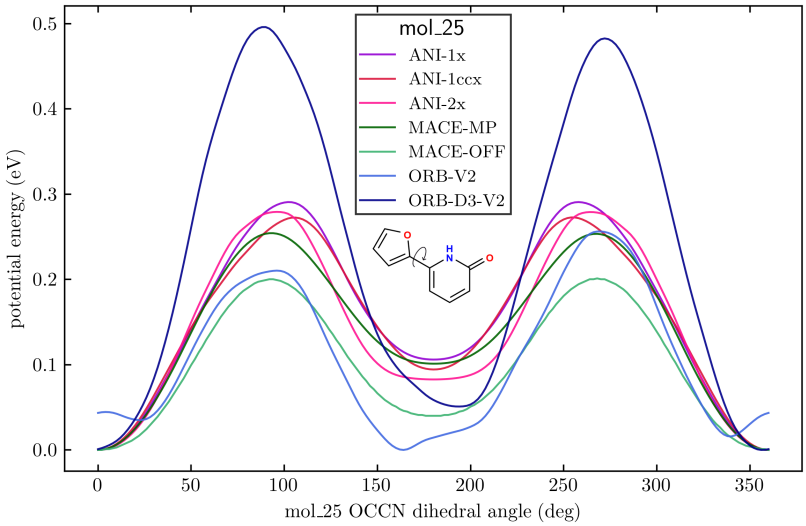
5.1.3 4-body distortions. Dihedrals

Lastly, a usual deformation of molecules are rotations along bond axes. These rotations are described using dihedral angles (four-body internal coordinate) formed by four atoms ABCD of the bond. The importance of correctly modeling these torsions stands from the fact that torsional minima are separated by small energy barriers, making molecules likely to 'jump' from one to another in simulations. Accurately describing these deformations is imperative, as they define the *flexibility* of a molecule, an especially important attribute for small organic molecules.

Fig. 10 showcases the torsional barrier descriptions of the HCCH dihedral (C–C bond) of ethane and CCCC dihedral of a byphenyl-like molecule, this later one inspired by Ref. [42]. As seen with angles, these potentials reflect the symmetries of the molecules, a correct description of which is extremely important for *in silico* simulations.



(a)



(b)

Figure 10: Dihedral angle potential energy scan of a dihedral angles of ethane and a biphenyl-like molecule. Additional dihedral scans for other biphenyl-like organic molecules can be found in Fig. 31 in Appendix G.

In the case of ethane (Fig. 10a), the model describe three equivalent minima, reflecting the molecule's symmetry, with comparable energy barriers. `MACE-MP`, however, neglects high-energy structures predicting free-rotation along the C–C bond. The dihedral in Fig. 31b is much more difficult, due to all the non-C atoms of the molecule. It is important to note that, essentially, the rotation clockwise or counterclockwise of the bond produces equivalent results, which translate to a symmetric torsional scan. While most models correctly identify this rotational symmetry, `Orbital` models do not, as reflected in the asymmetries of the potential curves, especially at $\sim 200°$. This incorrect symmetry description has meaningful consequences in simulations, favoring a rotation over another, potentially biasing the simulation.

Overall, structural deformations are mostly qualitatively well described by these NNPs, with the notable exception of dissociation limits in `ANI` models and some symmetries in `ORB` models, which will be explored in the following subsection.

5.2 Physical adequacy. Invariance and Equivariance

As previously stated in Sec. 4, NNPs not only must show a qualitatively correct description of atomic interactions, but must also be physically correct in order to be used as substitutes to proper physical models. The following tests aim to check that physical constraints and system symmetries are respected, ensuring that the models do not break basic physics.

5.2.1 Rotation and translation invariance

Rotations and translations of the system result in equivalent structures. It is important that energy is found to be invariant to these transformations, as otherwise, space isotropy will be broken resulting in a bias for certain positions or rotations during simulations, ultimately affecting all the information extracted from them in an unpredictable way.

Fig. 11 show the change in energy when different molecules (water and cyclobutadiene) are rotated and translated.

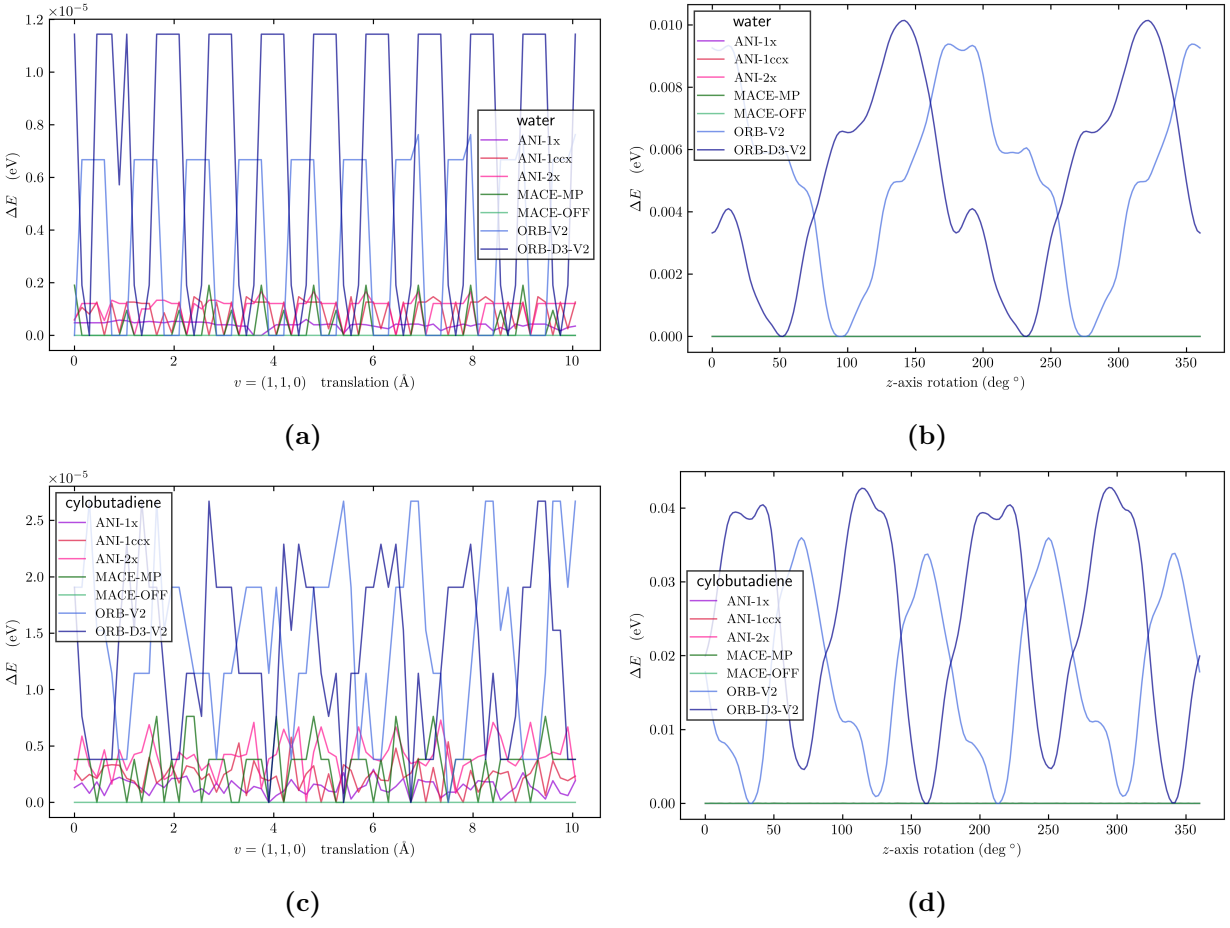


Figure 11: Energy differences when a water **(a)**–**(b)** and cyclobutadiene molecule **(c)**–**(d)** is rotated and translated. Note that for translations, the scale of the changes is augmented by a scale of 10^5. Variations in **(a)** and **(c)** correspond to numerical fluctuations, and invariance is seen as the fluctuations scale is negligible. For rotations, all except for `ORB` models are seen to be invariant. Other molecules can be found in Fig. 34.

We can see that `ANI` and `MACE` models 'understand' the symmetry of the system, which is to be expected as they are invariant and equivariant models, respectively. The case of `ORB` is very interesting: it is translationally invariant, as seen by the small fluctuations in Figs. 11a and **??**, but not rotationally invariant, as seen by the large periodic fluctuations of ~ 10

meV ($\sim$ 0.23 kcal mol^{-1}). `ORB` models, rather than having rotational invariance built into the architecture, opt to 'learn' rotational invariance through data augmentation, leading to an approximately invariant representation of the chemical environment. This is also the reason why angular and torsional deformations are not completely symmetric compared to other models, as seen in Figs. 9b, 10a and 11c.

While the data augmentation approach make the model faster when predicting energie, the lack of rotational invariance in `ORB` can have a biasing efect in simulations, favoring a specific set of atomic configurations in space. These effects should be studied futher to assess whether rotationally augmented models can be used over invariant or equivariant ones.

5.2.2 Conservation of linear and angular momentum

Besides energy invariance, NNPs must conserve total linear and angular momentum (see Appendix D), which translates to zero net force and torque. These constraints are particularly important in simulations where the dynamics of the system are of particular importance: if total force is different from zero, the system might start randomly accelerating to a particular direction. Similarly, if total torque is not zero, the molecule might start rotating, messing up the dynamics, and consequently, the information retrieved from them. These checks are, thus, of particular importance to ensure stable simulations.

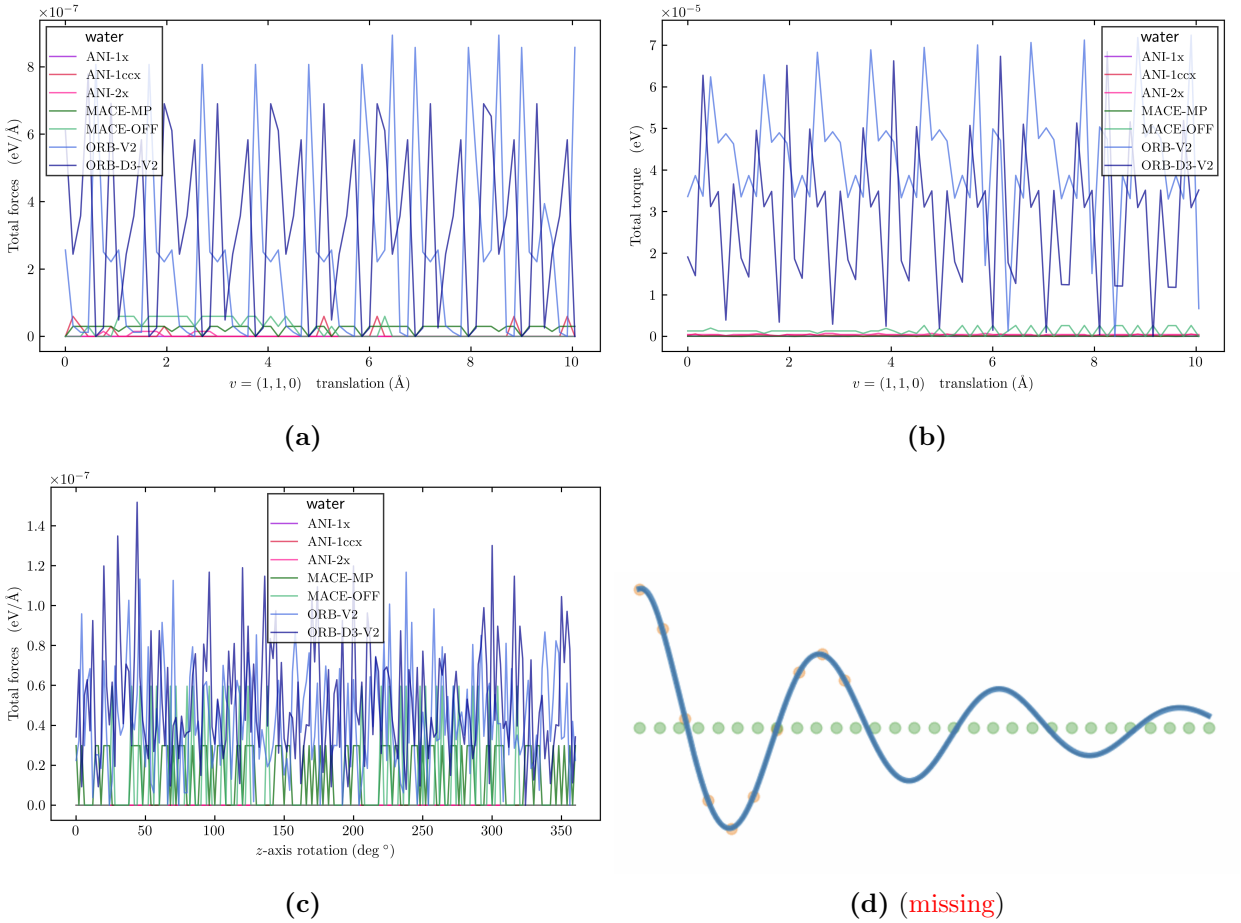


Figure 12: Total net force **(a)**-**(c)** and torque **(b)**-**(d)** for water along translations and rotations. All the models are shown to preserve linear and angular momentum, as shown the by scales of the fluctuations.

Fig. 12 shows the net force and torque when water is rotated or translated. It is interesting

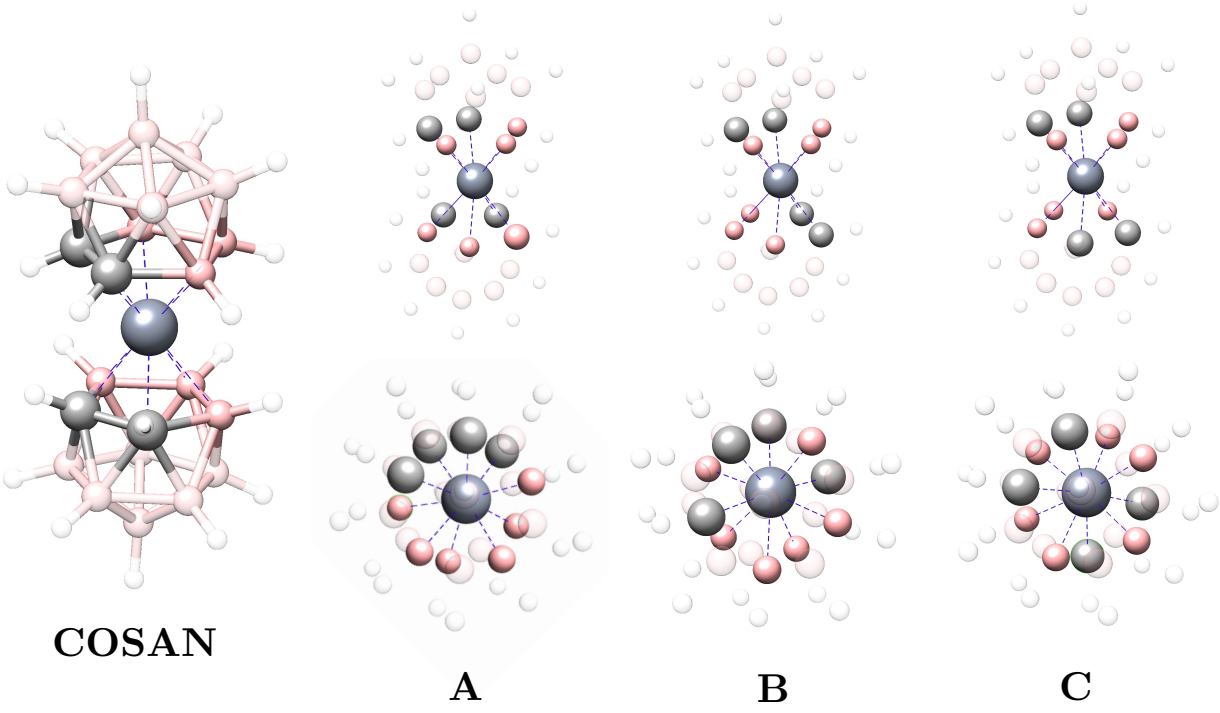
to see that all models preserve linear and angular momentum along these transformations, as shown by the scales of the numerical fluctuations, even though not all models compute forces in the same way. While `ANI` and `MACE` obtain forces as the analytical gradient of the potential energy, ensuring energy conservation, `ORB` models predict forces directly in a non-conservative way. As a result, this approach might contain a net force and torque. To satisfy physical contraints, the model solves corrects the final prediction by substracting both total force and torque.

5.3 Beyond symmetry: complex systems

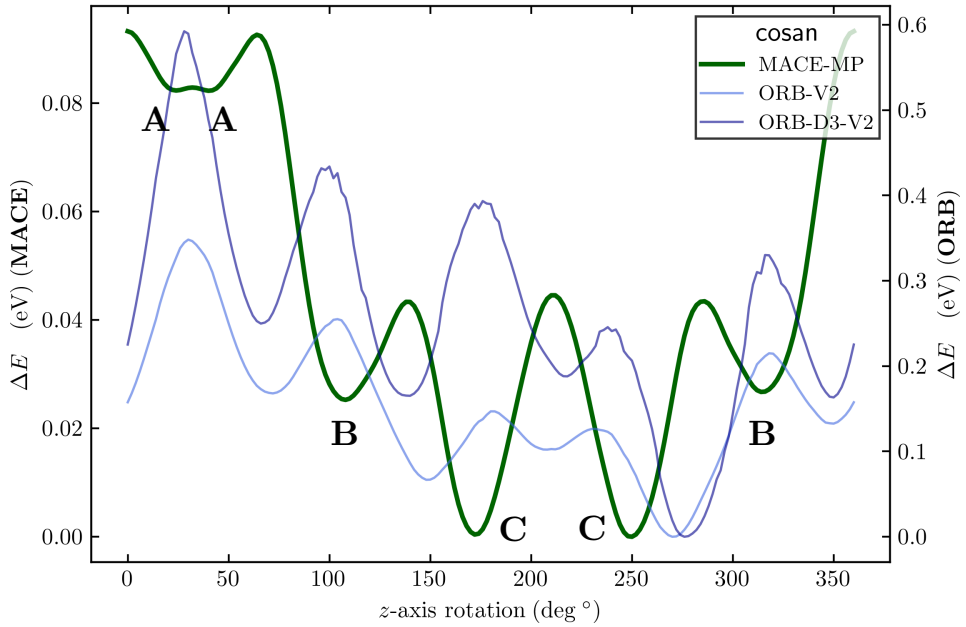
The final step after evaluating the quality of the atomic interactions and the physical adequacy of the models, is to see how these two aspects work together to predict the energy of complex systems. The following are two examples of systems with degenerate and non-degenerate symmetry states and aim to evaluate how the model ranks the stability and predicts energy barriers between them.

5.3.1 Non-degenerate symmetry system. COSAN.

COSAN (Fig. 13a) is a cobalt complex containing a total of 3 symmetrically different states (labeled **A**, **B** and **C** in Fig. 13a), obtained by rotating one of its carbon-boron ligands with respect to the other. A qualitatively correct model would predict these states, their symmetries, and smoothly connect them with high-energy 'transition states'. This is exactly what `MACE-MP` showcases in Fig. 13b, showing that conformation **C** is the most stable in vacuum, in agreement with Ref. [Jordi paper]. As more C atoms overlap (from a top view perspective), the energy of the minima increases, connected by high-energy transition states.



(a)



(b)

Figure 13: **(a)** COSAN, a cobalt-containing anionic complex with three distinct symmetry states, labeled **A**, **B** and **C**. The global structure, coordination sphere and top view are shown. The central atom corresponds to the Co center. Pink atoms are boron (B) and grey atoms carbon (C). **(b)** Potential energy along the rotation of one ligand with respect to the other. Note that all other models have not been used due to the Cobalt and Boron present in the structure. Two separate scales have been used for `MACE` and `ORB` models.

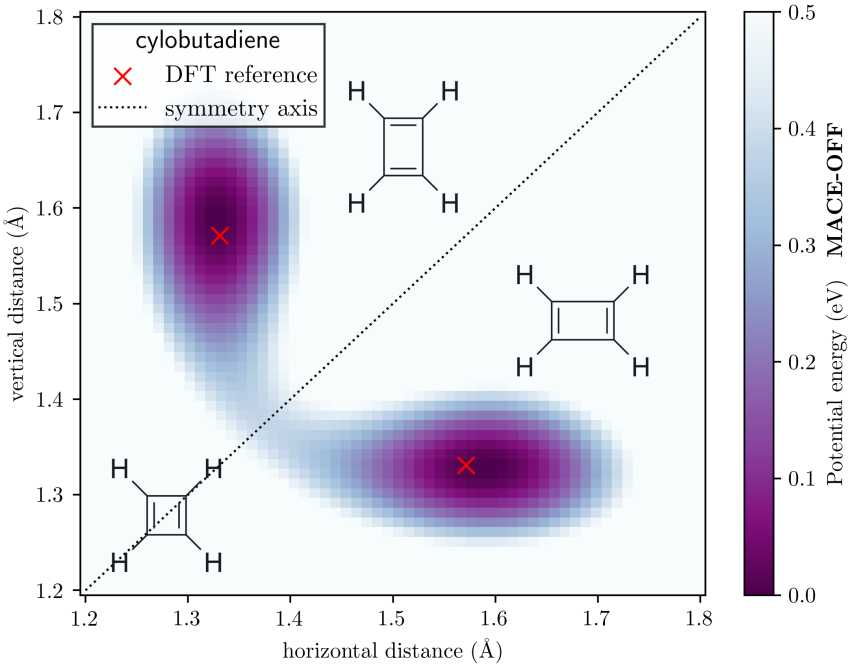
In contrast to `MACE-MP`, the description from `Orbital` models is qualitatively incorrect. First, the energy scale of these is vastly superior to MACE, overestimating the energy of transitions. Most importantly, however, the description does not capture the rotational symmetries of the system. This is a case where rotational invariance is essential to describe the system, and, as discussed in previous sections, `ORB` models are not. This system shows that *both*

a qualitatively correct description of atomic interactions and physical adequacy is needed to correctly describe the nuances of certain systems. While `Orbital` correctly captures interactions, its lack of symmetry understanding hiders its applicability substantially in real-world scenarios.

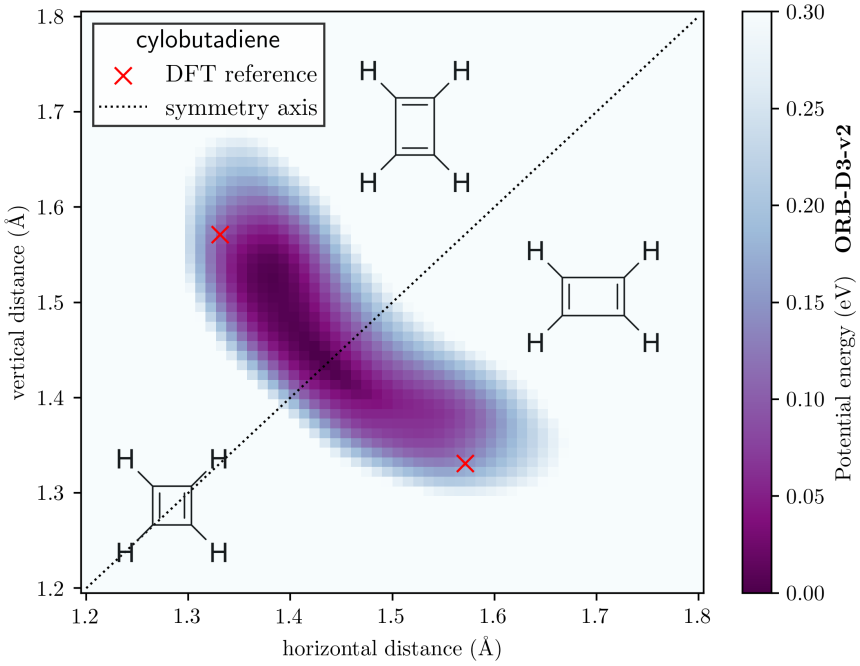
5.3.2 Molecular asymmetry. Cyclobutadiene.

Symmetries are the source of energy degeneracy. There are some cases where breaking this degeneracy can lead to a more stable system. Cyclobutadiene (Fig. 14) is an example where the most symmetric square geometry is naturally deformed to a less symmetric rectangular one, in order to break degeneracy and improve stability. The deformation can take place in both x and y directions, generating two rotationally-identical, deformed, molecules.

Thus, to correctly describe cyclobutadiene, the models must understand its asymmetry (deformation of the molecule), but also symmetry: x and y deformed molecules are identical. Fig. 14 contains the potential energy surface for each x and y distortion possible.



(a)



(b)

Figure 14: **(a)** `MACE-OFF` and **(b)** `ORB-D3-v2` energy surfaces for the distortion in the x- and y-axis of cyclobutadiene. The zero of energies corresponds to the minimum energy structure. The symmetry axis when the distortion in both dimensions is the same is represented with a black dotted line. Red crosses represent the geometry of a reference cyclobutadiene structure optimized with wB97XD/6-31G(d) DFT functional and basis set. The distance deformation scan in one direction can be found in Fig. 32. Fig. 33 contains other methods than the ones shown in this figure. The two molecules of the minima in **(b)** are related by a 90° rotation.

Most models, neglect the asymmetry of cyclobutadiene, as seen by the 1D deformation of Fig. 32 and the surfaces in Figs. 14b and 33 (centered around the symmetry line). The only model that describes the rectangular geometry is `MACE-OFF` (Fig. 14a), with two identical

minima situated exactly where the DFT reference is, with the square geometry being approximately 40 meV higher in energy.

On the other hand, `ORB-D3-v2` fails on both tasks. It predicts a symmetric square molecule, but most importantly, on of the deformations along the y axis is predicted to be more favorable, which breaks symmetry. This effectively means that rather than deforming equally on the x or y axis, the model introduces a bias for one of the two deformations, which can have great impact.

These two examples remark the fact that a qualitatively correct understanding of the atomic interactions is not sufficient, and a correct treatment of symmetries is necessary to achieve a good molecular description.

Referenciar que la diferència d'energia (approx 0.04 eV) entre les asimetries es correspon amb la diferència d'energia tal i com es mostra a la figura . 11d, efectivamnet 0.04 eV entre minim i màxim. Associem doncs aquesta diferència d'energia a la falta de rotational invariance.

6 Conclusions (i coses varies)

La conclusió hauria de fer referència als Nobel del 2024 i posar en context la importància del ML a la ciència en les pròximes dècades, dir que és un rapid moving field. Resumir el que s'ha vist als resultats.

A *Ab initio* electronic structure methods

A.1 Born-Oppenheimer approximation

One of the central aspects in theoretical and computational chemistry is the development of *ab initio* (first principles) quantum chemistry methods[10–13] to solve the time-independent schrödinger equation (TISE),

$$\hat{H}\,|\Psi\rangle \equiv \left[-\frac{\hbar}{2m}\hat{\nabla}^2 + \hat{V}\right]|\Psi\rangle = E\,|\Psi\rangle \ . \tag{6}$$

Let t be time, N and M the number of electrons and nuclei, respectively, and $\{\mathbf{R}\}$ and $\{\mathbf{r}\}$ the position of the M nuclei and N electrons, respectively. The system wavefunction (in coordinate representation) is

$$\Psi(t, \{\mathbf{r}\}, \{\mathbf{R}\}) \equiv \Psi(t, \mathbf{r}_1, \ldots, \mathbf{r}_N, \mathbf{R}_1, \ldots, \mathbf{R}_M) \ . \tag{7}$$

The TISE is only valid in case the potential is independent of time, $V(t, \{\mathbf{r}\}, \{\mathbf{R}\}) \to V(\{\mathbf{r}\}, \{\mathbf{R}\})$. In that case,

$$\Psi(t, \{\mathbf{r}\}, \{\mathbf{R}\}) = \phi(t)\psi(\{\mathbf{r}\}, \{\mathbf{R}\}) \,, \tag{8}$$

and the TISE simply becomes $\hat{H}\psi = E\psi$. However, solving it for a general system of arbitrary number of atoms and electrons is impossible, as nuclear and electronic degrees of freedom might be coupled.

An initial approximation based on the mass ratio of the proton and electron is called the **Born-Oppenheimer approximation**[43], in which the electronic and nuclear motion is decoupled based on the different time-scale of the 'motion' of both nuclei and electrons. This approximation implies that no mixing of different electronic stationary states happens due to the interaction between the nuclei (electrons do not undergo transitions between stationary states) and hence no heat exchange between nuclei and electrons is observed. The adiabatic approximation is only valid if the energy associated to nuclear vibrational degrees of freedom is not comparable to the electronic energy.

Under the BO approximation, the wavefunction becomes

$$\psi(\{\mathbf{r}\}, \{\mathbf{R}\}) \quad \longrightarrow \quad \psi(\{\mathbf{r}\}; \{\mathbf{R}\}) \,, \tag{9}$$

and only depends *parametrically* on the nuclear positions. The general Hamiltonian of the system is (in atomic units)

$$\begin{aligned} \hat{H} &= -\sum_{i}^{N} \hat{\nabla}_i^2(\mathbf{r}_i) - \sum_{\alpha}^{M} \frac{m_e}{m_\alpha}\hat{\nabla}_\alpha^2(\mathbf{R}_\alpha) - \sum_{i}^{N}\sum_{\alpha}^{M} \frac{Z_\alpha}{r_{i\alpha}} + \sum_{i}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} + \sum_{\alpha}^{M}\sum_{\beta>\alpha}^{M} \frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} \\ &\equiv \hat{T}_N(\{\mathbf{R}\}) + \hat{T}_e(\{\mathbf{r}\}) + \hat{V}_{Ne}(\{\mathbf{r}\}, \{\mathbf{R}\}) + \hat{V}_{ee}(\{\mathbf{r}\}) + \hat{V}_{NN}(\{\mathbf{R}\}) \,, \end{aligned} \tag{10}$$

with $\hat{T}_N$ and $\hat{T}_e$ the kinetic energy of the nuclei and the electrons, respectively, and $\hat{V}_{Ne}$, $\hat{V}_{ee}$, $\hat{V}_{NN}$ the potential energy due to nucleus-electron, electron-electron and nucleus-nucleus

interaction. The Hamiltonian can be separated into a nuclear and electronic contributions, separating the dependencies,

$$\begin{aligned}\hat{H} &= \left(\hat{T}_N(\{\mathbf{R}\}) + \hat{V}_{NN}(\{\mathbf{R}\})\right) + \left(\hat{T}_e(\{\mathbf{r}\}) + \hat{V}_{Ne}(\{\mathbf{r}\},\{\mathbf{R}\}) + \hat{V}_{ee}(\{\mathbf{r}\})\right) \\ &\equiv \hat{H}_{\text{nuclear}} + \hat{H}_{\text{electronic}} \,,\end{aligned} \tag{11}$$

where the nuclear Hamiltonian only depends on the coordinates of the nuclei. As ψ depends parametrically on $\{\mathbf{R}\}$, if nuclear coordinates are fixed, $\hat{H}_{\text{el}}$ will only depend on electronic coordinates. The potential energy of a given nuclear arrangement can be trivially determined from the nuclear repulsion term,

$$\hat{H}(\{\mathbf{R}\}) = \hat{H}_{\text{nuclear}} + \langle \hat{H}_{\text{el}} \rangle = \hat{T}_N + \left(\hat{V}_{NN} + \langle \hat{H}_{\text{el}} \rangle\right) = \hat{T}_N(\{\mathbf{R}\}) + U(\{\mathbf{R}\}) \,, \tag{12}$$

which is represents the movement of classical nuclei under the potential due to the average field of the electrons. This provides a way to obtain the PES for any given system: for a given atomic configuration (set of $\{\mathbf{R}\}$), the electronic TISE is solved and the eigenfunctions and eigenvalues are collected and the potential energy is obtained. Iteration over all nuclear arrangement provides the PES, which can be used to compute the nuclear wavefunction and eigenvalues corresponding to rotovibrational motion.

A.2 Hartree-Fock theory introduction

In order to solve the electronic TISE to obtain the wavefunction and energies, we must deal with $\hat{V}_{ee}(\{\mathbf{r}\})$, which introduces the so-called 'electron correlation', where the position of a single electron is determined by the positions of all other electrons, and these are determined likewise (many-body problem).

Hartree-Fock theory[10] tries to work around this issue to devise a method to obtain the electronic energies and wavefunctions. As a first approximation, the many-body problem is converted into a many-particle problem neglecting electron correlaction completely and introducing self-consistent field (SCF) theory (or mean-field theory), under which each electron is 'effectively' independent, only interacting with other electrons in an average way.

Note that, as electrons are fermions, the permutation of any pair of electrons must change the sign of the wavefunction. This, as a consequence, means that the probability density of finding two electrons of the same spin in the same space is zero (what is refered to as the Exclusion principle). Let Ψ be the N-electron wavefunction, $\Psi(\mathbf{r}_1,\ldots,\mathbf{r}_N)$, we can build it as an antisymmetrized product (Slater determinant) of a set of one-electron wavefunctions $\psi_j(\mathbf{r}_i)$,

$$\Psi(1,\ldots,N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(1) & \psi_2(1) & \ldots & \psi_N(1) \\ \psi_1(2) & \psi_2(2) & \ldots & \psi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(N) & \psi_2(N) & \ldots & \psi_N(N) \end{vmatrix} \tag{13}$$

with $i \equiv \mathbf{r}_i$ the coordinate of electron i. One last important approximation in the HF theory is the usage of basis functions φ_i such that each one-electron wavefunction is written as a linear combination of basis functions, such that

$$\psi_i = \sum_j^K c_{ij}\varphi_j \tag{14}$$

The basis set $\{\varphi_i\}$ contains K basis functions that provide a functional description of the one-electron wavefunctions. Note that each ψ_i might be delocalized over the whole molecule, as φ_j may be centered in any given atom of the molecule.

The final step to obtain the energy of the configuration (fixed $\{\mathbf{R}\}$) is to apply the variational theorem to optimize the set of coefficients $\{c_{ij}\}$ that minimize the electronic energy,

$$E_{\text{el}}^{*} \equiv \min_{\{c_{ij}\}\in\mathbb{R}} \frac{\langle \Psi(1,\ldots,N)|\, \hat{H}_{\text{el}}(1,\ldots,N)\, |\Psi(1,\ldots,N)\rangle}{\langle \Psi(1,\ldots,N)|\Psi(1,\ldots,N)\rangle} \geq E_{\text{el}}^{\text{real}} \tag{15}$$

which can be done solving the Roothan-Hall equations to provide the 'best' Slater Determinant for the molecule. Note that, to determine the coefficients of molecular orbital ψ_j, knowledge of all the other s is needed. This turns the SCF into an iterative procedure (Roothan-Hall equations are solved until convergence, or in other words, until the electronic density is self-consistent with the field it generates).

A.3 Electron correlation methods

In the limit of infinite basis set size (optimization over all functional space), the energy converges to the so-called Hartree-Fock limit. However, due to the lack of complete electron-electron correlation, the energy is larger than the 'exact' value due to the mean-field approximation: electrons avoid each other only on average, not as point particles. In relative terms, the correlation energy ranges from $0-2\%$, however, it is significant in absolute terms, making it a crucial energetic contribution for deriving molecular properties.

Two different approaches can be taken to introduce correlation[10] or 'recover' the missing correlation energy in the Hartree-Fock method. The computational cost of usual electronic structure methods can be found in Table 1.

- **Wavefunction methods.** Are based on expressing $\Psi(1,\ldots,N)$ as a linear combination of excited Slater Determinants, which allow for more flexibility when describing the real wavefunction and can be systematically improved including higher-order excitations. These methods include: Configuration Interaction methods (CI), Coupled Cluster methods (CC), Møller-Plesset perturbation theory methods (MP) and Multiconfiguratinal SCF methods (CASSCF, etc.)
- **Density functional methods.** Try to predict the energy of the configuration directly from the electron density using an exchange correlation functional within the Kohn-Sham formalism. Depending on what information of the density is used, several exhcange-correlation types arise: Local density approximation (LDA), Generalized gradient approximaiton (GGA), Hybrid functionals.

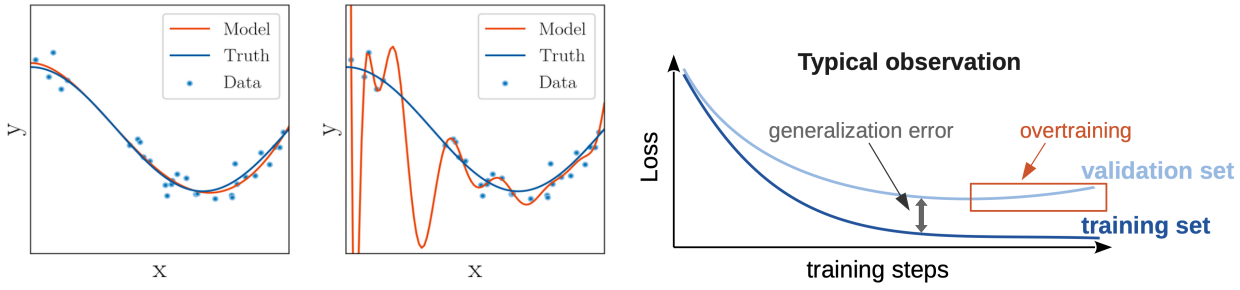
Table 1: Computational cost scaling in terms of the number of basis functions, K for common electronic structure methods[11]. For DFT, the exact scaling depends on the exchange-correlation functional used. Scaling shown is before code-implementation optimizations.

Electronic structure method	Computational cost scaling
Hartree-Fock	$\mathcal{O}(K^4)$
DFT	$\mathcal{O}(K^3)$
MP2	$\mathcal{O}(K^5)$
CCSD(T)	$\mathcal{O}(K^7)$

B Further ML considerations

B.1 Generalization error, underfitting and overfitting

The goal of the model is to try to learn the underlying representation of the input data (e.g. for a regression task, learn the underlying true function that generates the points, as seen in Fig. 15a). To track the generalization performance while and after training, subsets of the dataset are generated. For training, a *training* and *validation* set are used, the former to actually train the model and the later to assess its performance while training (but not to train it). The rest of the dataset is used to build a *test* set, used to assess the final performance after the model has been trained.



(a) (b)

Figure 15: **(a)** Overfitting [undefitting is missing] of a model. Underfitting is caused by the lack of parameter flexibility of the model, not being able to capture the true representation of the model. Overfitting, on the other hand, can be caused by excessive model complexity or overtraining. **(b)** Typical ML training observation. With training the model reaches a minimum generalization error, but this gap increases due to overtraining.

Model complexity plays a crucial role in determining the underlying representation: an overly simple model with very few trainable parameters will not be able to capture the underlying representation of the data due to the lack of 'flexibility' to represent the underlying function, even if the model is trained for many epochs[a], in which case the model is said to **underfit**. On the other hand, very complex models with a lot of trainable parameters are prone to create a complex function that minimizes training error but miss the underlying representation. In this case, the model 'memorizes' the training set and is unable to generalize to new data, which leads to **overfitting**.

Overfitting can be detected as an increase in generalization error (between the training and validation sets), where, the model starts to 'memorize' the training set due to overtraining. This can be mitigated by stopping the training where generalization error is minimum (as seen in Fig. 15b).

[a] An epoch of training ≡ training over all examples in the training set once.

C Graph Neural Networks (GNNs)

In ML, we often need to tailor the architecture to the data format of a specific task to obtain a data-efficient model, that is, one that can take advantage and exploit the structure and symmetries in the data. A simple case is image recognition: the MNIST dataset containing hand-written digits from 0 to 9 can be 'solved' with a MLP architecture, but is not data-efficient, as MLPs do not satisfy translational invariance of the features in images nor can be applied to images of variable dimensionality. Instead, Convolutional Neural Networks[REF NEEDED] (CNNs) are the go-to model for image recognition, as they are able to exploit the translational symmetry of the features in images and are extendible to large images, while being parameter-efficient.

The need for GNNs is apparent when we consider how many data structures can naturally be codified as graphs. A first intuitive example is molecules, where we can think of each atom as a node and each bond as an edge connecting the nodes, as seen in Fig. 3. Other type of data that can be modeled as a graph are images (each node is a pixel, each with a feature vector containing the corresponding RGB values), the citations in scientific papers, maps (where each intersection is a node, connected by roads), social interaction in social media, etc. Hence, it is interesting to develop a ML architecture that can efficiently exploit the underlying structure of the data.

This section aims to provide a solid, although not exhaustive, description of the GNN architecture based on Refs. [44–46]. The structure will be as follows. First, some math preliminaries about GNNs are given. Second, the barebones architecture is formulated to gain intuitive insight on it. Third, the message passing neural network (MPNN) and its variants are introduced. Finally, particular considerations when applying GNNs to computational chemistry are discussed.

C.1 Mathematical preliminaries

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is formed by nodes (or vertices) $\mathcal{V}$ which are connected through edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. These edges can be directed or undirected to specify the relationship between nodes. We will first start by considering graphs with no edges, and then move on to consider the connectivity of a graph.

Graphs in ML encode the features of a given example in a particular representation (embedding) in both nodes and edges. Let k be the number of node features, we define the node feature vector of node $i = 1, \dots, N$ as $\mathbf{x}_i \in \mathbb{R}^k$ and the corresponding feature matrix $\mathbf{X}$ as a stack of them

$$\mathbf{X} = \begin{bmatrix} — & \mathbf{x}_1 & — \\ — & \mathbf{x}_2 & — \\ — & \dots & — \\ — & \mathbf{x}_N & — \end{bmatrix}.$$

Note that, by stacking all feature vectors, we are assigning a label (a number) to each node, but these are arbitrary: all permutations of $\mathbf{X}$ are equally as valid and thus the model must return the same value for each permutation (see Fig. 16). If we define f as a function that operates on all the features of the graph, $\mathbf{X}$, we must impose f to be **permutation**

invariant such that $f(\mathbf{PX}) = f(\mathbf{X}) \quad \forall \mathbf{P} \in \Sigma$ with $\mathbf{P}$ the permutation matrix, $\mathbf{P} \in \mathbb{R}^{N\times N}$ (and $\Sigma \subset \mathbb{R}^{N\times N}$ the group that contains all $N!$ permutation matrices). This means that if we were to change the order in which we present the nodes to f, the result would be the same. A general permutation invariant function could be

$$f(\mathbf{X}) = \phi\left(\bigoplus_{i\in\mathcal{V}} \psi(\mathbf{x}_i)\right) , \tag{17}$$

where ψ operates on a single node, and all that information is aggregated in a permutation-invariant way with $\bigoplus$, which can be a sum, average or maximum operation. The resulting feature vector will be passed to ϕ to make the final prediction.

This, however, is only useful if we want to determine a property of the whole graph, i.e. a property given by the node features $\mathbf{X}$. However, it might be interesting to perform node-wise predictions, where feature vectors are updated to obtain a set of latent feature vectors $\mathbf{h}_i$ (in matrix form $\mathbf{H} = g(\mathbf{X})$, with $\mathbf{H}$ the latent feature matrix of the graph, defined similarly as to $\mathbf{X}$). If we let g be permutation invariant, the ordering of 'which part of the output belongs to which node' is lost. Thus, we must require g to be **permutation equivariant**, that is, the ordering of the output must be preserved independently of whether the permutation is performed before or after the operation, i.e. $g(\mathbf{PX}) = \mathbf{P}g(\mathbf{X})$.

We can combine both invariant and equivariant operations to provide a framework to obtain global features from per-node features. This can simply be done by defining a local, shared and equivariant operation ψ on each node, that updates each node's features $\mathbf{h}_i = \psi(\mathbf{x}_i)$ to form $\mathbf{H}$, from which we can obtain per-node properties. Then, we can combine all features with a permutation invariant operation, $\bigoplus$, to aggregate all latent features and apply ϕ over the final 'pooled' feature vector to obtain the final global property.

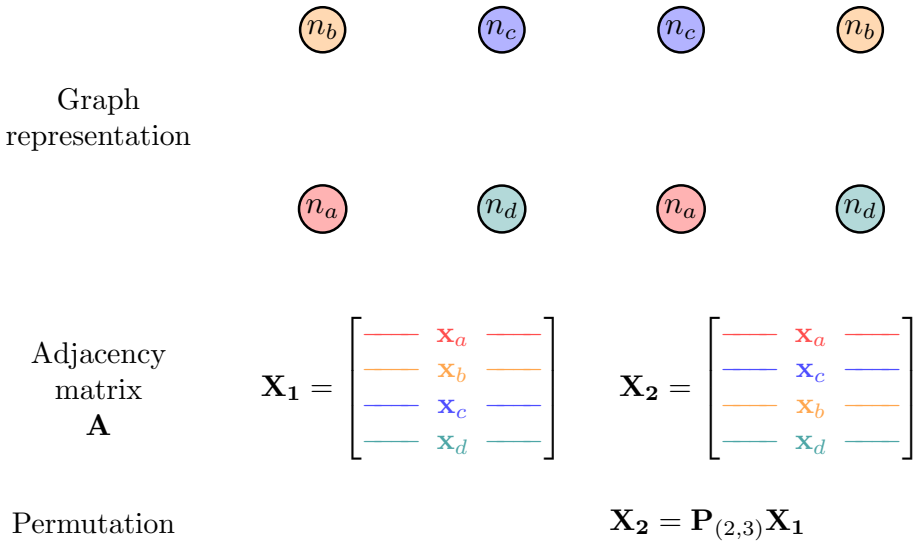


Figure 16: Two identical graphs, with only nodes $\mathcal{V}$ and no edges $\mathcal{E}$. The permutation matrix $\mathbf{X}$ is build by stacking the feature vectors $\mathbf{x}_i$ of each node, which necessarily imposes an arbitrary ordering of the nodes. Permutations (using a permutation matrix $\mathbf{P}$) of the rows of $\mathbf{X}$ contain the same information, and as such, a GNN model should return the same output. This property of graphs give rise to **permutation invariant** transformations on all the nodes, i.e. the condition $\mathbf{v} = f(\mathbf{X}) = f(\mathbf{PX})\ \forall \mathbf{P}$ where $\mathbf{v}$ is a global property of the graph obtained from $\mathbf{X}$. In the example, the labels of nodes n_b and n_c are swapped, as a result $\mathbf{X_2}$ is a permutation of the initial feature matrix $\mathbf{X_1}$.

Until now, we have not yet considered graph connecitivity. Connectivity is determined by edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, which can be represented using an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ such that

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if} \quad (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

which encodes which nodes are connected. We consider undirected edges, meaning that $\mathbf{A}$ is symmetric, and that each node is not connected to itself, i.e. $\mathbf{A}_{ii} = 0$. Similarly to nodes, each edge has an edge feature vector $\mathbf{e}_{ij}$, not necessarily the same dimensionality as k. If we consider a non $\mathbf{0}$ adjacency matrix (we have edges) the conditions for equivariance and invariance not only apply to nodes, but also to edges. This can be described in the following way,

$$\text{invariance} \qquad f(\mathbf{PX}, \mathbf{PAP}^\top) = f(\mathbf{X}, \mathbf{A}) \tag{19}$$

$$\text{equivariance} \qquad f(\mathbf{PX}, \mathbf{PAP}^\top) = \mathbf{P}f(\mathbf{X}, \mathbf{A}) \tag{20}$$

where $\mathbf{PAP}^\top$ means that the permutation is applied to both rows and columns of $\mathbf{A}$. Fig. 17 shows two isomorphic graphs, and although $\mathbf{A}$ is different, the connectivity remains the same and is why permutation invariance on nodes and edges is required for any graph $\mathcal{G}$.

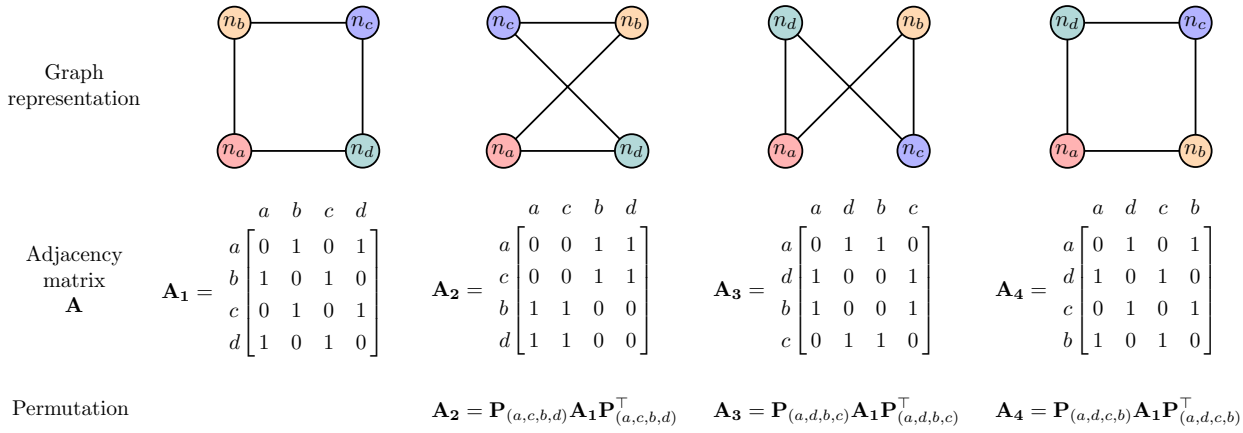


Figure 17: Set of four isomorphic graphs, all with the same connectivity but a different choice of labels to represent the adjacency matrix $\mathbf{A}$.

Another important definition is the 1-hop (immediate) neighborhood of node i

$$\mathcal{N}_i = \{u \mid \{u, v\} \in \mathcal{E}\} . \tag{21}$$

We can extract a multiset of features of the neighborhood

$$\mathbf{X}_{\mathcal{N}_i} = \{\mathbf{x}_j \mid j \in \mathcal{N}_i\} \tag{22}$$

which excludes the own node's $\mathbf{x}_i$. This allows us to define a local, shared, permutation invariant function h that operates on each node and its neigborhood, $h(\mathbf{x}_i, \mathbf{X}_{\mathcal{N}_i})$, to update each node's value as seen in Fig. 18.

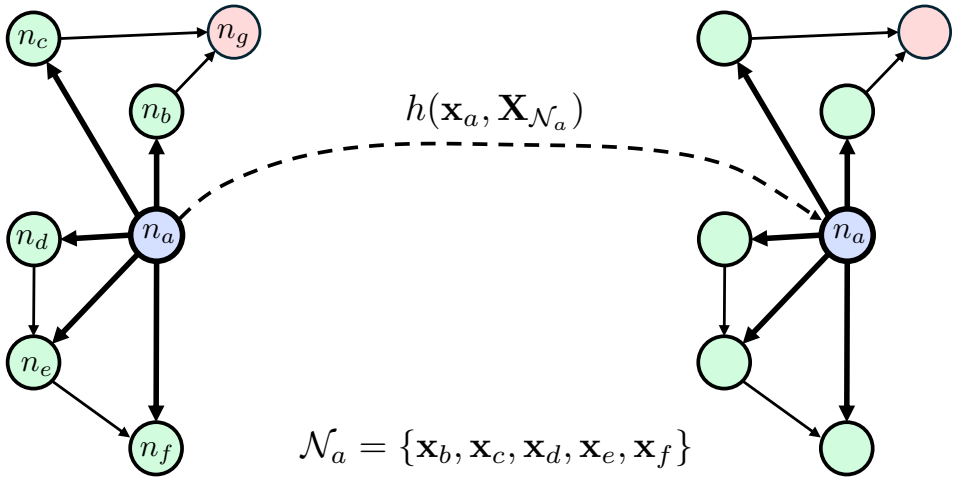


Figure 18: Visualization of the per-node update mechanism. The feature vector $\mathbf{x}_a$ of node n_a is updated using a shared local permutation invariant function h that operates on the own node (in blue) and its 1-hop neighborhood (in green). Nodes in red do not take part in the node update. This operation must be permutation invariant, as the labeling of the neighborhood is arbitrary.

We now return to search a blueprint for obtaining per-node and global properties, as done in the node-only case. To start, we need to find a function g that updates the features of all nodes, i.e. $g : (\mathbf{X}, \mathbf{A}) \to (\mathbf{H}, \mathbf{A})$. This transformation must be equivariant as discussed when we only had nodes. To do this, we can simply build $\mathbf{H}$ by stacking the latent feature vectors obtained via $\mathbf{h}_i = h(\mathbf{x}_i, \mathbf{X}_{\mathcal{N}_i})$, with h invariant and where the transformation includes its own node (as was for ψ in the only-nodes case), but also the 1-hop neighbors due to the connection through edges. This leads to

$$\mathbf{H} = g(\mathbf{X}, \mathbf{A}) = \begin{bmatrix} \text{—} & h(\mathbf{x}_1, \mathbf{X}_{\mathcal{N}_1}) & \text{—} \\ \text{—} & h(\mathbf{x}_2, \mathbf{X}_{\mathcal{N}_2}) & \text{—} \\ \text{—} & \cdots & \text{—} \\ \text{—} & h(\mathbf{x}_N, \mathbf{X}_{\mathcal{N}_N}) & \text{—} \end{bmatrix}, \tag{23}$$

with h the permutation invariant shared local function, and g the permutation equivariant 'update' function. Note that g returns a graph with the same connectivity as the input graph, $\mathbf{A}$.

C.2 Barebones GNN architecture

As seen in CNNs, ML models typically take rectangular or grid-like arrays as input[b]. For a graph $\mathcal{G}$, it is not obvious how we can process its data, as each node might be of different *degree* (contain a different number of nodes in its neighborhood). The GNN architectures we will introduce can take advantage of the graph structure and solve the different connectivites of each node.

The GNN architecture takes the graph connectivity ($\mathbf{A}$ or different representation), node and edge embeddings $\{\mathbf{x}_i\}$, $\{\mathbf{e}_{ij}\}$, respectively, and performs operations to update its features, from which per-node and global properties of the graph can be determined. This

[b]For non-numeric data, such as strings, a one-hot encoding is used (grid-like). If the dimensionality is large, the one-hot encoding becomes sparse and reduced representations are used (autoencoder architecture).

is a '*graph-in, graph-out*' approach that satisfies graph symmetries and exploits the graph representation, as seen in Fig. 19.

In a GNN architecture, the input graph is transformed by a sequence of GNN layers, producing an output graph with latent node and edge embeddings. To update the features, function $g : (\mathbf{X}, \mathbf{A}) \rightarrow (\mathbf{H}, \mathbf{A})$ must be equivariant. In the simplest form, each node and edge is transformed independently by an invariant function ψ only taking information of the own's node or edge, and not the neighborhood (which makes the invariant condition trivial). It is important to note that the *same* function ψ is applied for all nodes[c] (and or edges), decreasing the number of trainable parameters substantially.

Once the nodes have been updated, the latent feature matrix $\mathbf{H}$ can be built. As happens with deep architectures, these GNN layers can be stacked to update the features in each layer. It must be noted that ψ is often a MLP, taking the node or edge features and producing a latent feature output $\mathbf{h}_i \in \mathbb{R}^m$, with m the latent feature dimension, which can be different to k the initial input dimension. Each GNN layer containts a different ψ function.

Once the output graph is produced, per-node and global properties can be computed using permutation equivariant and invariant operations, respectively. For global properties, the blueprint of Eq. (17) is followed: first each node output is aggregated in a permutationally-invariant way and then ϕ (usually another MLP) is used for the final prediction.

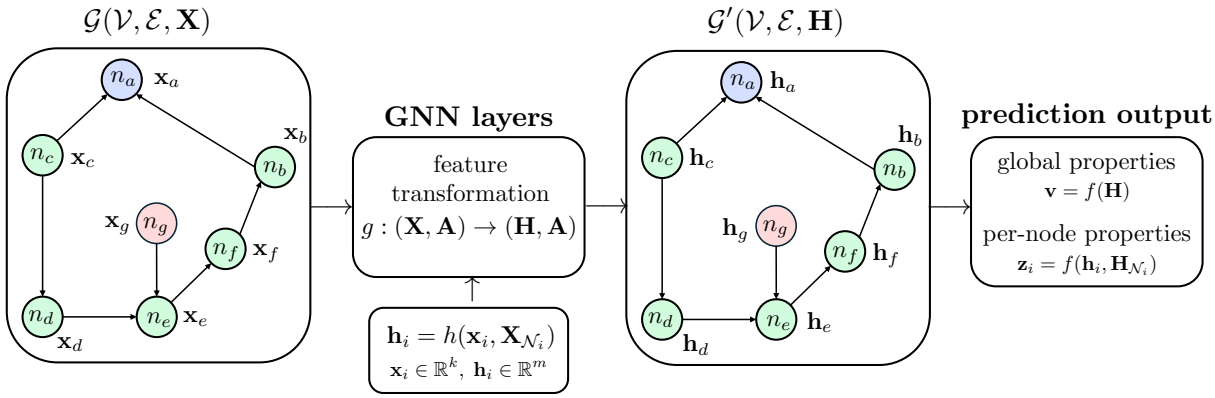


Figure 19: General workflow of a GNN architecture. First, an initial graph $\mathcal{G}$ is defined with connectivity $\mathbf{A}$ and node and edge embeddings $\mathbf{x}_i$ and $\mathbf{e}_{ij}$, respectively. This graph is transformed to a different graph $\mathcal{G}'$ by different GNN layers, which are fundamentally formed by a *permutation invariant* ' layer' update function $g : (\mathbf{X}, \mathbf{A}) \rightarrow (\mathbf{H}, \mathbf{A})$ formed of a per-node *permutation invariant* 'message passing' function $h(\mathbf{x}_i, \mathbf{X}_{\mathcal{N}_i})$. Depending on how h is defined, different types of GNNs arise. Once the output graph has been generated, global or per-node attributes can be computed, normally using an MLP for regression or classification. A more general pipeline overview can be found in Ref. [45].

The general blueprint for updating all the nodes of a graph is Eq. (23), with g often referred to as the 'propagation', 'update' or 'message passing' function. However, the most important step is to decide how each node is updated, i.e. define how $h(\mathbf{x}_i, \mathbf{X}_{\mathcal{N}_i})$ updates node i based on the features of the neighboring nodes. Regarding this function, there are three main 'flavors' of GNNs depending on how the neighboring information is used: GCNs[47], GATs[48]

[c] This function reuse is similar to *filters* in CNNs, which were reused for different images.

and MPNNs[49] (Fig. 20). The particular case where no neighbor information is used has been discussed in this barebones GNN architecture section.

C.3 Message-passing GNNs

The idea of message-passing GNNs is to define the function h of Eq. (23) to include the data (feature vectors) of neighboring nodes. This way, each node is made 'aware' of the data in its neighbors. Depending on how this mechanism is modeled we have convolutional, attentional and message-passing GNN architectures, as seen in Fig. 20.

The core mechanism for all architectures is the same and is depicted in Fig. 21. For each node, we prepare the features of neighboring nodes in some way ('prepare the messages') as well as the own node's features. All these features are aggregated by a permutation invariant operation and are mapped by ϕ to the updated feature vector, often by a MLP.

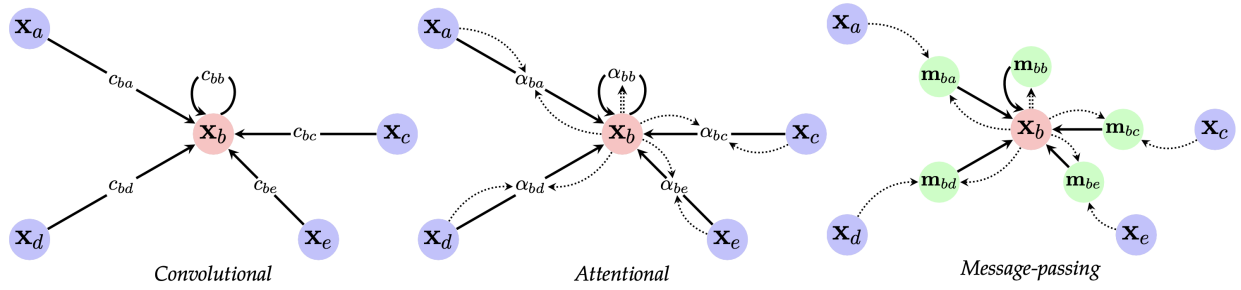


Figure 20: Figure extracted from Ref. [44]. Different ways of how the permutation invariant operation h can be defined. Depending on how the messages are produced, three main architecture types exist: convolutional (GCN), attentional (GAT), and message-passing (MPNN). α_{ij} refer to the normalized attention coefficients $a(\mathbf{x}_i, \mathbf{x}_j)$ over the whole neighborhood using a softmax function.

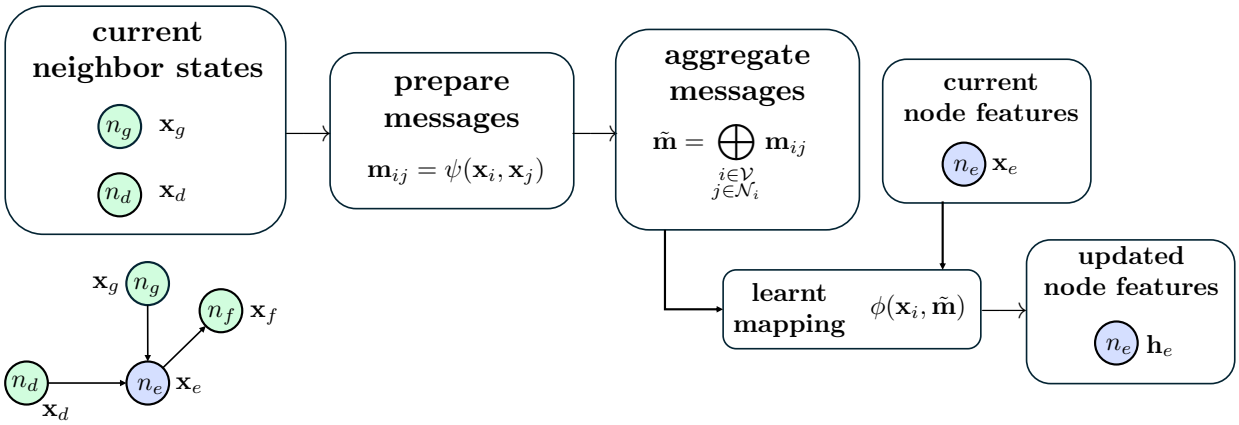


Figure 21: Core workflow of a general message-passing GNN. The general per-node propagation step can be seen in Fig. 18. Given a node n_e and its neighbors (n_g, n_d) with their respective features, the messages $\mathbf{m}_{ij}$ are prepared with a function ψ, taking information of the given node and its neighbors. The way messages are prepared are the main differentiating factor of different GNN architectures. The messages are then aggregated in a permutation-invariant way to ensure the whole nodeupdate process is permutation invariant.

C.3.1 Graph Convolutional Networks (GCNs)

For GCNs, h is defined as follows

$$\mathbf{h}_i = h(\mathbf{x}_i, \mathbf{X}_{\mathcal{N}_i}) = \phi\left(\mathbf{x}_i, \bigoplus_{j\in\mathcal{N}_i} c_{ij}\psi(\mathbf{x}_j)\right) \tag{24}$$

with ψ a shared local function that operates in isolation on each node of the neighborhood of node i, often a MLP. These 'messages' are then weighted by a fixed scalar c_{ij} and aggregated. Intuitively, the coefficients c_{ij} encode 'how much node i values the message from node j'. The node is updated using the node feature vector $\mathbf{x}_i$ and its local environment by ϕ, another MLP.

This procedure is reminiscent of a discrete convolution kernel in CNNs, where the weights of the kernel are fixed and discretized. As nodes graphs not necessarily have the same degree nor equidistant neighbors, we apply a continuous kernel to weigh the neighborhood nodes. This way, c_{ij} might be a function of node distance, $c_{ij} \sim ||\mathbf{r}_i - \mathbf{r}_j||$.

For further reading on this type of architecture, check Ref. [47] [classic GCN paper].

C.3.2 Graph Attention Networks (GATs)

For GATs, h is defined as follows

$$\mathbf{h}_i = \phi\left(\mathbf{x}_i, \bigoplus_{j\in\mathcal{N}_i} a(\mathbf{x}_i, \mathbf{x}_j)\psi(\mathbf{x}_j)\right) \tag{25}$$

where, compared to GCNs, the aggregation coefficient c_{ij} is replaced by a *learnable* scalar using an attention mechanism[50]. This attention function (often modeled as a MLP) is trained to take as inputs the two nodes for which the attention coefficient $a_{ij} \equiv a(\mathbf{x}_i, \mathbf{x}_j)$ is to be computed. a_{ij} are normalized among all of the neighborhood using a softmax function to qualitatively indicate how much attention node i should spend on the message from node j. Additionally, this mechanism can be extended to multiple attention heads, where different attention coefficients are computed using different attention functions.

Further information on GATs can be found in Ref. [48].

C.3.3 Message Passing Networks (MPNNs)

MPNNs are the most general type of GNNs that use the message passing mechanism, first applied to Quantum Chemistry [MPNN for QC[49]]. h is defined as follows

$$\mathbf{h}_i = \phi\left(\mathbf{x}_i, \bigoplus_{j\in\mathcal{N}_i} \psi(\mathbf{x}_i, \mathbf{x}_j)\right) \tag{26}$$

The idea is that, for each node i, messages of nodes $j \in \mathcal{N}_i$ are prepared through ψ, usually a MLP. This messages are transformations of the neighboring node's features taking into account also node i. The message preparation provides a general procedure for message-passing GNNs. If the messages are reduced to be a fixed scalar (computed scalar) times an

isolated transformation on neighboring nodes, we obtain GCNs (GATs).

As a last comment, note that MPNNs one message must be computed per edge, compared to the other two varieties where only transformations on nodes were considered and weighted by fixed or trainable scalars. This makes MPNNs the most expensive to compute, as there are many more edges than nodes, and only scale well for small graphs compared to the other two (cheaper) alternatives.

C.3.4 Equivariant MPNNs

In the context of physics and chemistry, MPNNs used to represent molecules for tasks such as molecular energy regression or per-node force predictions, not only must be permutation equivariant (due to the inherent graph representation), but must also satisfy the symmetries of Euclidean space, $E(3)$ (translation, rotation and reflection symmetries). This equivariance could, in principle, be learnt by either using data augmentation and letting the model learnt it through *data*, or using an $E(3)$-equivariant model architecture[51,52]. We will focus on the latter.

NNP models contain many information encoded in feature vectors, *inter alia* atom type, partial charge, velocities, and positions. It is desirable that, the latter are transformed in the same way as the molecule is transformed in space, i.e. are $E(3)$-equivariant[44,52–55]. To achieve this[44], positions $\boldsymbol{x}_i \in \mathbb{R}^3$ are treated independently to other scalar features $\mathbf{f}_i \in \mathbb{R}^{k-3}$. A GNN layer would transforms these feature vectors independently: it would be invariant with respect to $\mathbf{f}_i \to \mathbf{f}_i$, but equivariant with respect to positions $\boldsymbol{x}_i \to \mathcal{R}\boldsymbol{x}_i + \mathcal{T}$, with $\mathcal{R}$ an orthogonal matrix describing rotations and reflections, and $\mathcal{T}$ a vector describing translations[44].

Note that, due to invariance, the dependency of scalar features $\mathbf{f}_i$ with positions is only on the distance between nodes $||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$. However, there are important features which are not scalars (vector fields such as velocities or forces) which must be transformed under $E(3)$ symmetries[54,55]. The implementation of equivaraint models that can take advantage of higher-rank information are discussed in detail in Ref. [53], and a brief introduction is provided in Appendix E.

C.4 Prediction tasks for GNNs

Once the output graph has been obtained (resulting latent representation), the architecture must allow us to extract the information we need. There are three main types of predictions. **Node classification**, in which, we extract per-node information based on the resulting embedding of each node, such as the forces applied to each atom or the partial charges of each atom in a molecule. **Graph classification**, in which we extract per-graph (global) information, for example the total molecular energy or total charge. Finally, we can do **link prediction**, where the task is to predict the existance and properties of edges.

$$\text{Node classification:} \qquad \mathbf{z}_i = f(\mathbf{h}_i) \tag{27}$$

$$\text{Graph classification:} \qquad \mathbf{z} = f\left(\bigoplus_{i\in\mathcal{V}} \mathbf{h}_i\right) \tag{28}$$

$$\text{Link prediction:} \qquad \mathbf{z}_{ij} = f(\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}) \tag{29}$$

with $\mathbf{z}_i$ (or $\mathbf{z}_{ij}$ the outputs of the model (predictions) and $\mathbf{h}_i$ the output graph latent feature vectors.

C.5 Particular considerations for computational Chemistry

GNNs have great applicability in computational chemistry, as molecules can be naturally represented as graphs [A graph-convolutional neural network model for the prediction of chemical reactivity[56], MPNNs for QC[49]]. In the context of NNPs, we want models to be able to infer the molecular energy and forces without prior knowledge of the bonds, as is done in ES methods.

The most recent types of NNPs, such as MACE or Orbital models, are based on GNNs[d] and differ substantionally with respect to NN-based NNPs. All these models use a cutoff function $f_c(r)$ to limit the short-range interactions of the model (to exploit locality).

GNN NNPs define nodes as atoms, but edges are not identified with bonds (that would require prior knowledge of the type of bonds of the molecule). Instead, a node is connected to another atom if the distance is such that $||\mathbf{r}_i - \mathbf{r}_j|| \leq r_c$, with r_c the cutoff distance. The molecular energy is determined via the local contributions of each atom,

$$E_{\text{total}} = \sum_{i \in \mathcal{V}} E_i \ , \tag{30}$$

and as such, requires a per-node prediction: once the output node is obtained, energies are extracted from the latent features in each node. Forces follow the same per-node prediction procedure.

Graphs are particularly useful for computational chemistry as they are well suited to problems where the interaction between entities is important. Short range interaction between atoms can be modeled with high accuracy using the message-passing procedure, which provides a given atom information of its chemical environment and hence its interactions within the molecule.

C.5.1 Long-range interaction description

In principle, knowing that after ℓ GNN layers a given node obtains information about its ℓ-hop neighbors, we could naïvely stack enough layers to ensure every node receives information from every other node (with ℓ equal the graph *diameter*). This deep GNN architecture would, in principle, encode information from far away nodes, and thus describe long-range interactions. However, deep GNN architectures run into issues, namely **oversmoothing** and **oversquashing**[oversmoothing limiting factor of GNN depth[57], original paper[58]].

Most problems, such as molecular energy prediction, must to account for interaction between not directly connected nodes, which must be achieved (in the message-passing framework) by stacking multiple GNN layers. Different learning problems require different ranges of interaction between nodes in the graph to be described properly, which is known as the *problem radius* r. To account for this radius, the GNN architecture should contain $\ell \geq r$ layers, otherwise the model will suffer from **under-reaching**: distant nodes will not be aware of

[d] Another relevant model in molecular modeling based on GNNs is AlphaFold.

each other. However, as the number of layers increases, the number of nodes in each node's *receptive field* grows exponentially, $\mathcal{N}_i \sim \mathcal{O}(\exp \ell)$. This requires information from all these nodes to be compressed into a fixed size feature vector $\mathbf{h}_i$, which creates a bottleneck known as **over-squashing**. In this situation, the graph fails to effectively propagate messages from far away nodes, and nodes ultimately only learn short-range messages from the training data.

Deep GNN architectures also affect short-range problems. As the number of layers increases, node representations become indistinguishable, converging to a meaningless value. This is called **over-smoothing** and is mainly caused by using more layers than the problem required: if only short-range information is required, more layers generate inputs from far away nodes, irrelevant for the problem.

In the context of Quantum Chemistry, long-range interactions given by charge interactions are of uttermost importance for the description of chemical systems. Over-squashing is an inherent limitation of the GNN architecture, and should be mitigated to correctly encode long-range interactions.

D Atomic environment representation

Given the coordinates $\mathrm{r} = (\mathrm{r}_1, \ldots, \mathbf{r}_N)$ and identities $\mathbf{Z} = (Z_1, \ldots, Z_N)$ (usually the atomic number of each atom) of each component in an N-particle system, MLIPs aim to predict the energy of the configuration[5,59,60]. In this task, a ML model is trained on high-accuracy quantum mechanical results to predict energies of molecules it has not been trained on (regression task)[61]. The incentive of using an ML model is to 'skip' all the expensive quantum mechanical calculations of ES methods, while providing comparable accuracy.

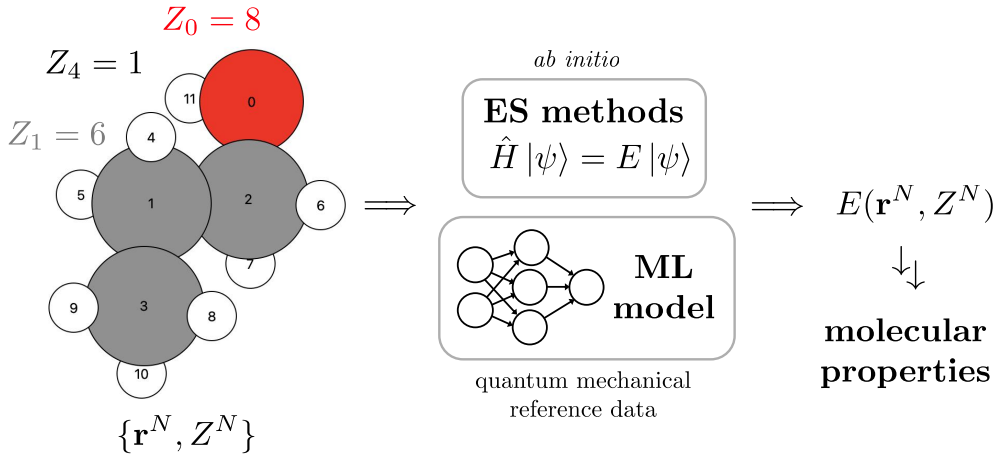


Figure 22: An atomic arrangement of N atoms can be uniquely determined by a set of atomic coordinates, $\mathbf{r}^N$, and their 'identity', given by the atomic number Z_i. This information is parsed to ES methods, which rely on standard well-known approximations to solve the TISE and ultimately obtain the energy of the configuration. On the other hand, MLIPs aim to use the same input together with a ML model trained on reference QM results to predict a molecular energy, from which predictions on molecular properties can be performed. The model effectively skips the QM calculations, 'learning the fundamental relations of , without learning the postulates'. This skip, however, is not completely physics-independent, the model still must be *physically adequate* and respect the symmetries of the system[5].

However, the task is deceptively simple compared to the design of a suitable model. This section aims to introduce the core aspects that must be considered regarding basic symmetry and physical constraints to produce a physically adequate model able generalize well on a wide variety of situations.

D.1 Physical constraints and symmetries

ML models differ substantially to traditional physics-based models in that the approach is purely data-driven, and no physical equations or priors are given to it (although one notable exception is PINNs[35,36,62]). As such, basic physics such as conservation laws or symmetries must be 'taught' to the model through data or through inductive biases in the model architecture. Conservation laws, typically derived from temporal invariance, provide strong constraints that can be used as guiding principles in search of physically adequate and plausible ML models[5]. Implications of conservation laws are as follows

- **Energy conservation** imposes restrictions on forces, namely, that forces must be the gradient of a potential. ML models can opt to predict forces according to the gradient of the potential energy prediction, or alternatively, predict forces directly. In the later case, however, as they are not computed from a potential, it is difficult to obtain a model with energy conservation.

– **Linear and angular momentum conservation** imposes a dependence of the potential energy on the *relative* positions of the atoms, thus translating into rotational and translational invariance of the potential energy. [rotational and translational equivariance of velocities and forces]

– Lastly, a subtile symmetry of the system is **permutation invariance**[5,61]: the potential energy must be symmetric to the exchange of two identical atoms (of the same element). Although trivial, it will have consequences when designing a model.

Failure to obey basic physics might result in models breaking symmetry, predicting a different energy for identical but rotated molecules, or breaking conservation laws, generating spurious forces under no external field[5]. In both of these cases, the consequences would prove catastrophic for *in silico* simulations and all the properties derived from them.

D.2 Special considerations of NNPs

The use of a ML model leads to practical considerations that define how NNPs are implemented. To introduce them, consider a model formed by a multi-layer perceptron (MLP), such as that of Fig. 2a. This model accepts a *fixed length*, *ordered*, input, the coordinates of the atoms $\mathbf{r}$ and outputs the potential energy prediction, $\hat{E}$. Additionally, we define the *representation* of the model as the description of the atomic arrangement that is parsed as input to the model[5,59,61,63]. In order for this model to be *physically adequate* (satisfy the basic physical constraints and system symmetries), the energy must be rotation-, translation- and permutation-invariant[59,63].

To ensure this, we must obtain a representation that is translation, rotation and permutation-invariant as well as *complete*[59] (must describe a molecule's conformation in a unique way). This would guarantee that, given any symmetry-modified coordinates, the same input is provided to the model and thus the same output (energy) is returned, satisfying invariance.

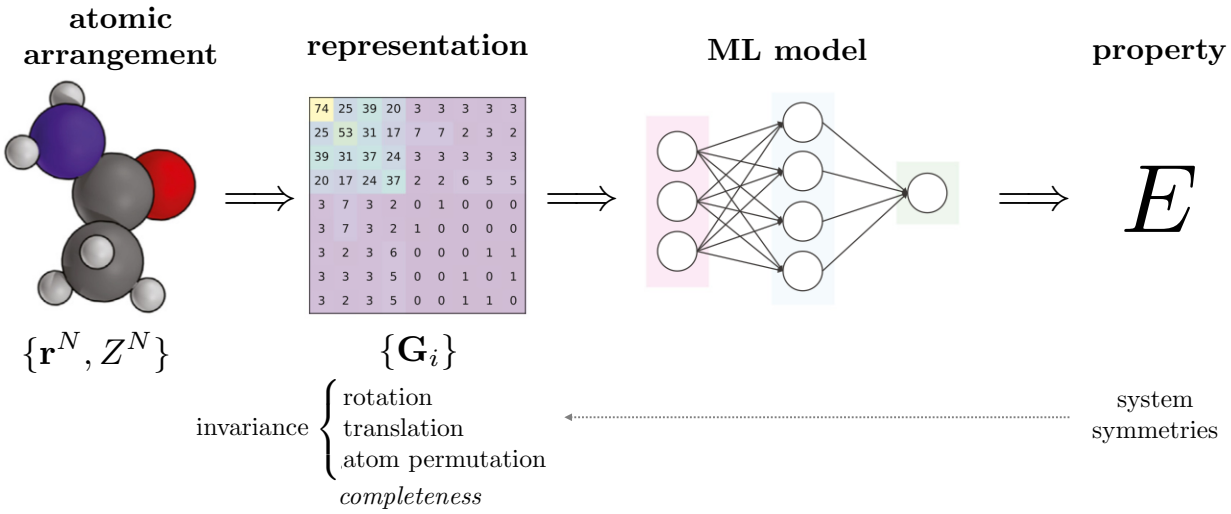


Figure 23: Overall NNP workflow. Given an atomic arrangement characterized by the atomic positions $\mathbf{r}^N = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ and identities $Z^N = (Z_1, \ldots, Z_N)$, a set of descriptors $\{\mathbf{G}_i\}$ for each atomic environment are computed and are used as input for the ML model, which predict the desired property, in the case of NNPs, the molecular energy. Physical constraints and system symmetries, however, impose a certain *structure* on the atomic representation, mainly invariances, as well as completeness: the representation must be univocal for each atomic arrangement, such that no two different structures have the same representation. Adapted from Ref. [64]

In order to be useful, the representation and model should be transferable[65]: independent to system size and applicable to very different systems. Note that a cartesian representation of the system, while simple and complete, is not suitable[61,63,66]: the list of coordinates may be arbitrarily ordered, and they are not rotation- and translation-invariant. Another option would be using internal coordinates (distances, angles, dihedrals, etc.), however, they may be arbitrarily ordered and the number of them increase drastically with the body-order, thus requiring a large and expensive model to train and evaluate[63]. [note that conventional FFs use a sum of internal coordinates up to 4-body dihedrals, making them rotation-, translation- and permutation-invariant and complete.]. In these two cases, the same but rotated molecule may be represented in a substantially different way, which would provide a different input to the model (representation not invariant), and consequently, the energy would not be invariant[5].

Most importantly, however, is the consequence of using a *fixed length* in the input of the model (Fig. 2a). As the input length is inherent to the model (cannot be changed), the applicability of the model is limited to a fixed number of atoms[5,59,60]. To make the model independent of system size, the system energy, E, is decomposed into local (site) contributions[59] as follows

$$E(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \sum_{i=1}^{N} E_i(\mathbf{r}_1, \ldots, \mathbf{r}_N)\,, \tag{31}$$

which additionally makes E permutation-invariant with respect to atomic permutations. The problem transforms into a different one: rather than predicting the energy of the whole molecule (considering all possible interactions), we predict the energy of an atom in the molecular environment. As such, closer atoms will have much more contribution to the local energy than atoms far away. This naturally introduces a *cutoff* on the distance, $f_c(r)$ to determine the locality approximation, i.e. 'how many interactions should be accounted for' at the expense of computational effort[5]. Interactions between atoms separated by more than the cutoff radius are truncated. Still, there might be interactions beyond this radius that should be included, like long-range electrostatic interactions[66], which will be discussed in a later sections.

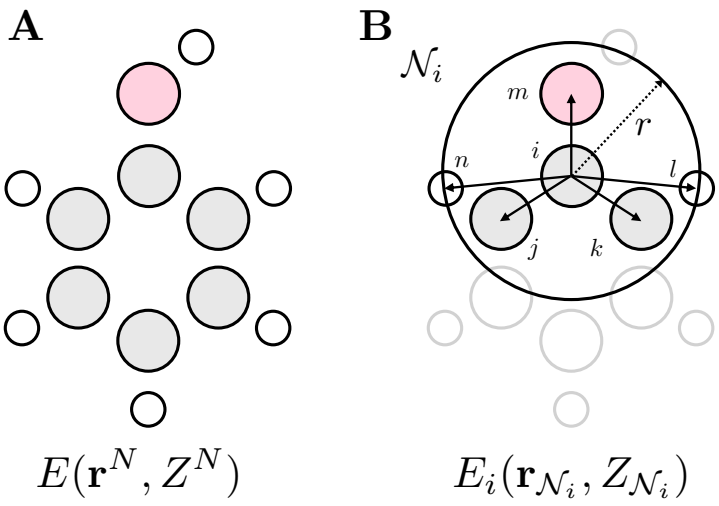


Figure 24: (**A**) atomic arrangement (positions and identites), of the phenol molecule. Its molecular energy is denoted by E. A severe limitation of using fixed length models, such as a MLPs, is that the number of atoms must remain fixed. To circumvent this issue, the energy is computed as site (local) contributions (see Eq. (31)). Each atom i is considered to only interact with its chemical neighborhood $\mathcal{N}_i$, defined by a certain cutoff radius r. The representation of the local environment is used to compute atomic energies E_i (**B**), which are then aggregated to obtain the total energy E. This locality approximation, however, might have great impact when long-range interactions are present, as they are truncated[66].

D.3 Description of the atomic environment

Once the permutation invariance and locality have been introduced, the next step is to design a suitable *representation* able to *describe* the atomic environment of each atom, and from it, derive the local atomic contributions E_i to the total energy.

The most important part of a NNP is the way the local atomic environment is represented (Fig. 24). This representation must be translation-, rotation- and permutation-invariant, *complete*, and independent of the number of atoms in the environment, due to fix length limitations. Depending on how this representations is performed, there two types of models[5,60,66], described in the following sections.

D.3.1 Descriptor-based NNs

In order to obtain the different site contributions, E_i, a first approach is to convert atomic coordinates $\mathbf{r}^N$ into an appropiate representation through the use of *descriptors*. These are many-body invariant symmetry functions that describe the chemical environment in a unique way[5,59,61,63,66], from which a model is able to predict the local energy contributions to the total energy. A general depiction of a descriptor-based model is shown in Fig. 25.

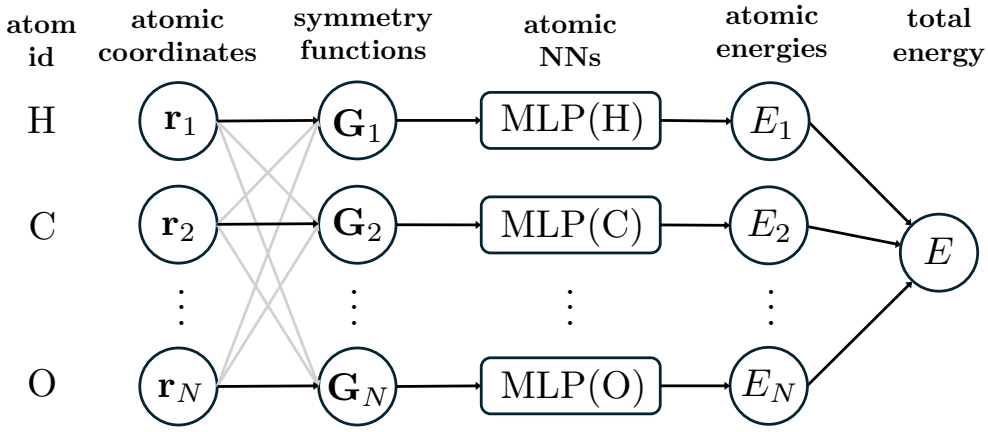


Figure 25: General view of the descriptor-based NNP workflow. The atomic environment (up to a certain cutoff) every atom is encoded into a fixed set of descriptors, $\mathbf{G}_i$ for atom i, which serve as input to a NN architecture to predict the local atomic energies E_i. The descriptors $\mathbf{G}_i$ are built using the cartesian coordinates $\mathbf{r}$ of the atoms inside the cutoff radius (see Fig. 24), as indicated by the grey arrows. To improve the model, we might choose to have different MLPs for each atomic element, i.e. one MLP for H, C, O, N, etc. The final molecular energy is determined from the local contributions. Adapted from Ref. [61].

In this sense, we can think of the representation of the atomic environment through a set of functions (descriptors) as a preprocessing step to obtain an invariant model, in which the predicted energy E is invariant to translations, rotations and permutations[63]. The next two sections provide an introduction to two sets of descriptors, atomic centered symmetry functions[59,61,66] (ACSFs), and the idea behind the atomic cluster expansion[67,68] (ACE).

D.3.2 Atomic centered symmetry functions (ACSFs)

The first use of chemical descriptors to obtain an invariant model used atomic centered symmetry functions[59] (ACSFs). In order to describe atomic environments, a fixed number M of symmetry functions is used (Fig. 25), which describe the 'fingerprint' of each atomic neighborhood i, $\mathbf{G}_i = (G_{i1}, \ldots, G_{iM})$, with G_{ij} the j-th symmetry function corresponding to the environment of atom i. Note that M is independent on the number of atoms inside the cutoff (neighborhood) to ensure that the input to the NN is fixed in length[66].

ACSFs (explained in detail in Refs. [59, 61, 63, 66] consist of two types of transferable functions to describe the environment: *radial* symmetry functions, which describe the radial distribution of neighbors up to the cutoff radius, and *angular* symmetry functions specifying their angular arrangement. Radial functions depend on the distances (2-body information) to all neighboring atoms, while angular functions depend on all possible angles (3-body information). The dependence on internal coordinates (distances and angles) guarantees translation-, and rotation-invariance, while the aggregation of all internal features guarantee permutation invariance.

In practice, a limited number of radial and angular functions are used to 'examine' the chemical environment, which imposes a limit on its resolution: more symmetry functions provide larger resolution to distinguish similar environments, but higher computational cost, as the NN model becomes larger[61].

ACSFs however, are just one type of descriptors used to represent atomic environments,

and other have been developed[63,67,69,70], which are summarized in Table 4 of Ref. [60]. A notable transferable NNP model that uses modified ACSF functions is the ANI family of models (ANI-1x[65], ANI-1ccx, ANI-2x[23]).

D.3.3 Atomic cluster expansion (ACE)

Recently, a new representation called atomic cluster expansion (ACE) has been introduced and proven to be a broad generalization of all before-mentioned sets of descriptors[67,68]. The idea behind ACE is to estimate the atomic energy E_i as many-body expansion

$$E_i = V^{(1)}(\mathbf{r}_i) + \frac{1}{2}\sum_j V^{(2)}(\mathbf{r}_i, \mathbf{r}_j) + \frac{1}{3!}\sum_{j,k} V^{(3)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \frac{1}{4!}\sum_{j,k,l} V^{(4)}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l) + \dots \quad (32)$$

with $V^{(K)}$ the contribution to the atomic energy E_i, which depends on $K-1$ neighbors. This is, at its core, what is used to express classical FFs in Eq. (2). Note that, as the body-order increases, the computational cost raises dramatically: there are many more angles (three-body) than distances (two-body), and in turn, many more dihedrals (four-body) than angles. Specific systems, such as metals, may require a many-body expansion of up to $K \sim 15$[67]. ACE is a generalization of ACSFs that generates a symmetry-aware ν-body expansion at low computational cost. In fact, when ACE is expanded up to three-body contributions, it can be shown that it is equivalent to ACSFs[68].

The importance of many-body descriptors is fundamental and has been proven a limiting factor in certain sets of descriptors, as there might exist different chemical environments which can only be distinguished by using higher-order descriptors[39,40], as depicted in Fig. 26.

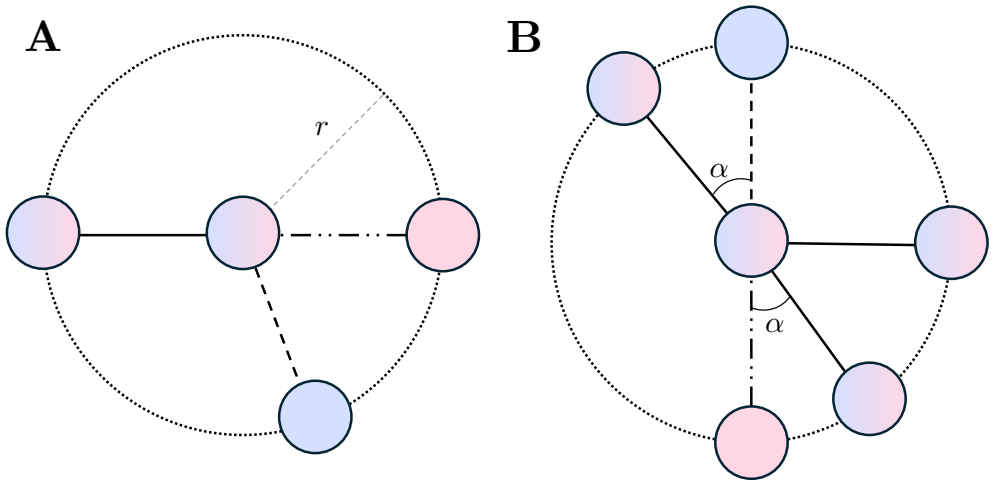


Figure 26: Different structures comprised of blue or red atoms. Atoms colored with a gradient are shared between both structures. (**A**) Both red and blue molecules are 3 atoms, the red one is linear while the blue one is angular-shaped. If only 2-body correlations are considered to represent the environment of the central atom, i.e. only distances, both structures would be identically described. Higher body-order correlations are needed to distinguish between them, such as angles (3-body order), which would show the linear vs. angular disparity. (**B**) Both red and blue structures are 5 atoms. From the central atom perspective, the environment is described by the same distances and angles (2 and 3-body order correlations). If distances and angles are used to describe the environment of the central atom, the structures are indistinguishable. These two examples **A** and **B** show the importance of the body-order of the interactions by a model, and its implications. Restricting the body-order limits the expessability of the model, ultimately hurting energy prediction. Example **B** is adapted from Refs. [39, 40]

D.3.4 End-to-end NNs

The need to use hand-picked symmetry functions in descriptor-based models to describe a given chemical environment has several downsides, as we (the designers) do not really know what functions will provide the best representation of the atomic environment: many functions can be used (with the invariance constraint), but not all can uniquely describe the environments[5,63].

To skip these decisions, a completely different end-to-end approach can be chosen. In it, rather than using hand-picked symmetry functions, we let the model *learn* a suitable invariant representation through data[5]. As such, the model figures out the best invariant representation of the chemical environment for the task, without the need for a hand-crafted functional forms of the descriptors[5,52]. This learnt representation is then used as input to predict site energies, E_i. Fig. 27 shows a comparison between to two architectures.

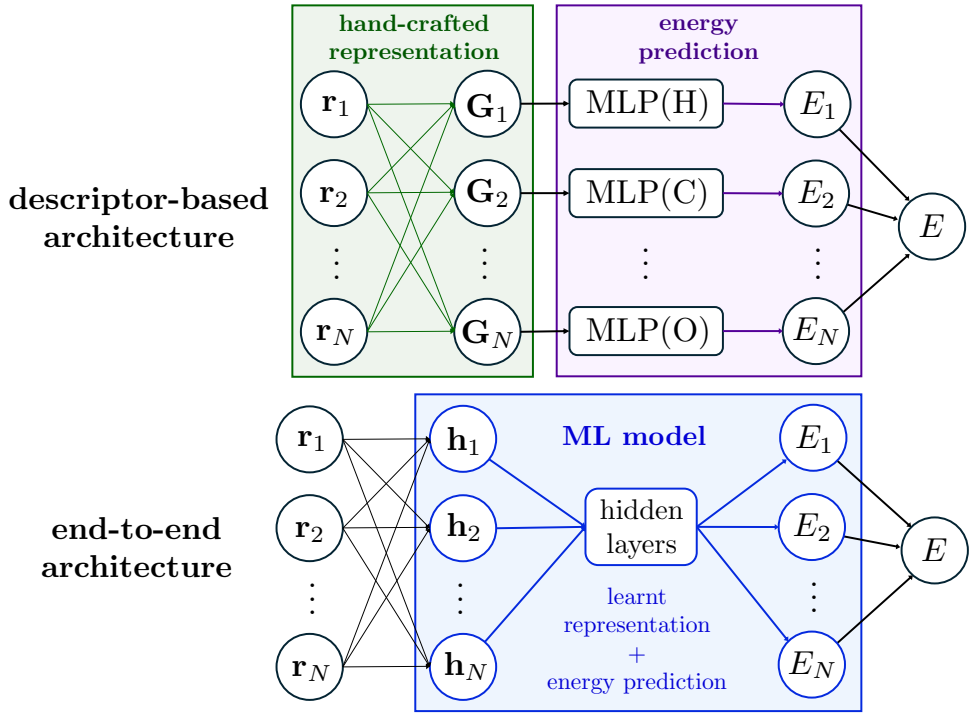


Figure 27: Visual comparison between descriptor-based and end-to-end models. (Top) Descriptor-based architectures rely on hand-crafted features to convert the initial atomic coordinates $\mathbf{r}^N = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ into descriptors $\{\mathbf{G}_i\}$, which can then be used as input to the NN-based model to obtain atomic energies. (Bottom) End-to-end architectures take a different approach and the representation is learnt by the model to produce internal representation feature vectors $\{\mathbf{h}_i\}$, which are then internally used by the model to predict the atomic energies, in an input-output closed fashion.

To obtain a learnt invariant representation, the model should ideally use only invariant internal features, such as distances or angles to ensure the invariance of the learnt model. Alternatively, we can lift this restriction and train the model through a data augmentation approach, where the training set is artificially enlarged by performing modifications such as rotations, in the hopes that the model learns a rotation-invariant representation. This last approach, however, besides being computationally costly (increasing training time substantially), it does not guarantee that the resulting representation is, in fact, rotation-invariant.

End-to-end models are usually MPNN[49], a type of GNN (explained in Appendix C) where a molecule is represented as a graph with atoms representing nodes. All atoms within the radial cutoff are connected through edges, representing interactions, and define the atomic environment of each node. In an end-to-end MPNN, each node stores the chemical information of the environment through the learnt representation, and is updated as in the message-passing formalism, using information of neighboring nodes to increase the receptive field to not just the local environment (1-hop neighbors), but beyond. A noteworthy invariant convolutional GNN model is SchNet[71].

D.3.5 Mixed approaches: MACE

A particularly interesting descriptor MPNN is the MACE (multi-ACE) architecture[22,51,72,73]. It structures the molecule as a GNN would, where nodes contain its atomic coordinates $\mathbf{r}_i$, fixed features $\boldsymbol{\theta}_i$ and the atomic environment description through generalized-ACE, $\mathbf{h}_i$. What truly makes it powerful is its ability to combine information of neighboring environments (within the message-passing formalism) to generate very-high order descriptions of the molecule, which provides an efficient way to encode the atomic interactions, and hence provide accurate energy predictions. MACE is, additionally, an *equivariant* model with respect to translations and rotations, which is a generalization of invariant models that allows it to use 3D structured data (vectors) for both training and prediction. For further details on equivariance and how it differs with respect to invariance, refer to Appendix E.

D.4 Going beyond: descriptor, symmetry and locality limitations [TODO]

Limitation of descriptor-based models, and in general, of NNPs, derive from the assumptions discussed throughout this section necessary to create a *physically adequate* model. These can be categorized into descriptor, symmetry and locality limitations.

- Completeness of the representation: two different structure must have a different representaiton.
- Invariant vs equivariant models, simply say that equivariant models are a generalization to invariant models that can use higher-rank data such as vectors and have deeper understanding on how different data such as vectors should change in rotations of the coordinate system.
- Discutir les limitacions de localitat: només contribucions shor-range... I si long range és necessari? REFS: [Scalable hybrid deep neural networks/polarizable potentials biomolecular simulations including long-range e ects, Machine Learning Interatomic Potentials and Long-Range Physics, ElectrostaticEmbeddingof MachineLearning Potentials]

E Equivariant Models. Spherical Tensor Embeddings

Neural network architectures that incorporate and process symmetry information have become recently very popular in the geometric deep learning space[51,52,54,55,74,75]. Correctly capturing the symmetry of a system is especially important when dealing with molecules (as is the case for NNPs), as all of their properties must transform in a predictable way under a transformation of the coordinate system through translations, rotations or reflections (Fig. 28). The inadequate description of the system symmetries can lead to severe consequences such as breaking fundamental physical constraints (e.g. conservation laws), thus resulting in fundamentally flawed models.

This section aims to introduce the importance of using rotation-equivariant models, able to identify features in any rotation, and discuss how equivariant models differ over invariant ones. Additionally, a brief overview on how equivariant models are implemented in the message-passing framework is provided.

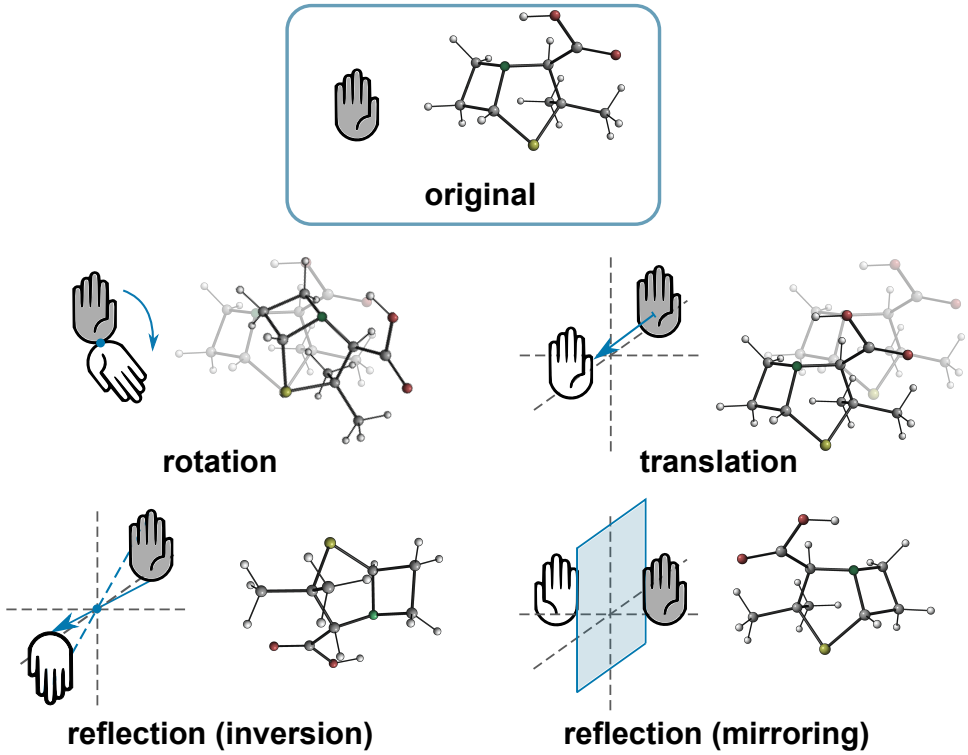


Figure 28: Symmetry elements of $E(3)$ (Euclidean symmetry group in 3D). Other relevant symmetry subgroups are $SE(3)$ (special Euclidean group in 3D), consisting of only translations and rotations, and $SO(3)$ (special Orthogonal group in 3D), consisting of only rotations. Extracted from Ref. [74]

E.1 The need for equivariant models

Consider an isolated molecule in the absence of any external field. Properties such as the potential energy of a molecular configuration, global charge, and other scalar attributes of the system should be *invariant* with respect to changes in relative orientation in space. Other order properties, e.g. vectorial properties such as dipole moment, forces and velocities of each atom, rather than remaining *invariant* with respect to, e.g. a rotation, should transform according to it: the prediction of the dipole moment vector should change if the atomic

coordinates are rotated, while its scalar value remains invariant with respect to the rotation.

Statements such as 'the potential energy of a molecule is invariant to rotations of the molecule' might appear trivial for physics-based models, where a system is modeled through a series of equations. For these models, the equations are such that basic physical constraints, such as conservation laws, and symmetries are always fundamentally respected. ML models, on the other hand, are not based on any physical priors, meaning that basic physical constraints (e.g. energy conservation) and how symmetry operations affect the system properties must be learnt through data. To circumnavigate this, a different approach has been to introduce these constraints as *inductive biases* in the architecture design[51,52,59,63,67,70,71,75].

The first models were designed such that the molecular energy prediction was *invariant* with respect to translation and rotation (thus a physically adequate model). Such models (descriptor-based[59,63,67,70] or GNN-based[71]) are made invariant through the use of scalar internal features (such as distances and angles), which do not change under rotations or translations and ensures the invariance of the output. However, limiting the internal representation with scalars means that they are unable to predict geometry-dependent objects, such as vectors, which change according to the rotation of the atomic coordinates.

Even though invariant models can only use scalar data for training, many molecular attributes are 3D-geometric data (vectors, matrices, etc.). This data, however, is sensitive to transformations of 3D space and as such, only models that 'understand' how they are transformed under rotations can use it[54]. The 'understanding' of symmetries comes from changing their internal representation according to rotations of the atomic coordinates. In this sense, equivariant models are more general than invariant ones, the later being a specific case of the former.

Thus, equivariant models lift the limitation of using scalar-only features by using internal features that transform under symmetry operations of the input, hence 'capturing'/'understanding' the rotation of the system[74]. The rotation-equivariant internal representation allows the model to use and predict higher-order data (e.g. vectors).

There are several benefits when it comes to using *equivariant* models rather than *invariant* ones[54,55,74]. First, as they are not limited to using scalar information, they are more information-rich and faithful in their predictions. Secondly, the use of scalars, vectors, and higher-rank geometric tensors in the internal representations, allow it to be more expressive in predictions and more efficient capturing features of the data. Additionally, it allows the model to be more data-efficient than invariant models[54], as the mapping to be learnt is restricted by rotation-equivariant functions.

E.2 Spherical tensors in equivariant models

The most important practical choice when implementing an equivariant neural network is to determine how the features can be embedded (expressed) and modified in an equivariant way under rotations. This section aims to provide an overview on the spherical tensor formalism and show how they can be used to create equivariant MPNNs[54,55,75].

Let f be a function such that $f : X \to Y$ with $\mathbf{r} \in X$ and G be a group with $g \in G$

the symmetry elements of that group. $\mathcal{D}(g)$ and $\mathcal{D}'(g)$ are the representations of g acting on X and Y, respectively. The equivariance condition[74] can be written as follows

$$f(\mathcal{D}(g)\mathbf{r}) = \mathcal{D}'(g)[f(\mathbf{r})] \,, \tag{33}$$

that is, predicting vector properties such as the dipole moment on the rotated set of atomic coordinates $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$, $\mathcal{D}(g)\mathbf{r}$, is the same as performing the prediction on the unrotated coordinates and then rotating the output of f. Note that the *representation* of the symmetry element can be very different. If g is a rotation, $\mathcal{D}(g)$ will be given by the usual rotation matrix in 3D, $\mathcal{R}$. However, if f returns the Hamiltonian matrix for $\mathbf{r}$, the representation of that exact rotation $\mathcal{D}'(g)$ will look very different and will be represented by a higher-rank geometric tensor. Note that the definition of invariance can be recovered from Eq. (33) by setting $\mathcal{D}'(g) \equiv \mathbb{I}$

Even though we are interested in E(3)-equivariance, we will only focus on SO(3) (3D rotations). Translation-equivariance can be trivially obtained using relative atomic displacement vectors[75], $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$.

The general idea to introduce equivariance into a model is to decompose the node feature vector into a set of rotationally equivariant geometric tensors called *spherical tensors*[54]. These are useful objects for dealing with rotation-equivariance due to their properties. In order to understand how higher dimensional spherical tensors transform under rotations, irreducible representations (irreps) of SO(3), a subset of rotation matrices that can be used to construct larger rotation matrices that operate on higher-dimensional tensors, are used. The irreps of SO(3) are the Wigner D matrices. Each type-ℓ Wigner D matrix, $\mathcal{D}^\ell$ with $\ell \geq 0$, defines how a type-ℓ spherical tensor rotates. These matrices provides a direct mapping between any rotation $\mathcal{R}$ of atomic coordinates, $\mathbf{r}$, in 3D space, $\mathcal{R}\mathbf{r}$, and a rotation in the embedded spherical tensor space via $\mathcal{D}^\ell_\mathcal{R} \in \mathbb{R}^{(2\ell+1)\times(2\ell+1)}$, the matrix corresponding to rotation $\mathcal{R}$.

Given a node feature vector, information might be stored as scalars, vectors or higher-rank cartesian tensors. These can be reexpressed spherical tensors, which can be directly rotated using $\mathcal{D}^\ell_\mathcal{R}$ matrices. The basis used for the projection are the (real) spherical harmonics, which represent a complete and orthonormal basis for rotations in SO(3).

Spherical harmonics $Y^m_\ell(\hat{\mathbf{r}})$, in this context, are functions defined on a sphere $Y^m_\ell : S^2 \to \mathbb{R}$ that project 3-dimensional vectors into spherical tensors that transform equivariantly and directly under Wigner-D matrices. Spherical tensors are classified acording to their degree ℓ (type-ℓ spherical tensors). Each type-ℓ spherical tensor transforms under the type-ℓ irreducible representation $\mathcal{D}^\ell$. Given a vector $\mathbf{r} \in \mathbb{R}^3$, we can project that vector into a type-ℓ spheric tensor $\mathbf{Y}^\ell$ via

$$\mathbf{Y}^\ell(\hat{\mathbf{r}}) = \begin{bmatrix} Y^{-\ell}_\ell(\hat{\mathbf{r}}) \\ \vdots \\ Y^0_\ell(\hat{\mathbf{r}}) \\ \vdots \\ Y^\ell_\ell(\hat{\mathbf{r}}) \end{bmatrix} \in \mathbb{R}^{2\ell+1} \,, \tag{34}$$

which transforms equivariantly via the corresponding Wigner D matrix

$$\mathbf{Y}^\ell(\mathcal{R}\hat{\mathbf{r}}) = \mathcal{D}^\ell_\mathcal{R}\mathbf{Y}^\ell(\hat{\mathbf{r}}) \,. \tag{35}$$

For $\ell = 0$, $\mathcal{D}^0_{\mathcal{R}}$ corresponds to 1, which describes the rotations of scalars (invariant under rotations). For $\ell = 1$, $\mathcal{D}^1_{\mathcal{R}}$ corresponds to the 3D rotation matrix R associated with the rotation $\mathcal{R}$ of vectors (equivariant under rotation). Higher degree Wigner D matrices describe the rotation of higher degree spherical tensors.

Interestingly, as spherical harmonics form a complete orthonormal basis, they are particularly useful to decompose any rotationally symmetric (SO(3)-equivariant) feature on the unit sphere, similarly to a Fourier decomposition of a periodic signal

$$f(\hat{\mathbf{r}}) \sim \sum_{\ell} \sum_{m=-\ell}^{\ell} \hat{f}^m_\ell Y^m_\ell(\hat{\mathbf{r}}) \; . \tag{36}$$

This function can be rotated identifying the rotation $\mathcal{R}$ on coordinates with its corresponding Wigner D matrix $\mathcal{D}$,

$$f(\mathcal{R}\hat{\mathbf{r}}) = \mathcal{D}_{\mathcal{R}}[f(\hat{\mathbf{r}})] \; , \tag{37}$$

which provides a blueprint to create rotationally equivariant functions. Overall, in order to produce an equivariant model, each operation should be equivariant, i.e.

$$\mathcal{F}(\mathbf{Y}^\ell(\mathcal{R}\hat{\mathbf{r}})) = \mathcal{D}^\ell_{\mathcal{R}}[\mathcal{F}(\mathbf{Y}^\ell(\hat{\mathbf{r}}))] \; , \qquad \forall \mathcal{R} \in SO(3) \; , \tag{38}$$

with any function $\mathcal{F}$ on spherical tensors of any degree.

E.3 Tensor product

Once we have a broad understanding of spherical tensors, we are interested in knowing how these tensors can interact among them and how they can be manipulated in an equivariant way. Spherical tensors 'interact' with each other through a tensor product[54] $\mathbf{s}^{\ell_1} \otimes \mathbf{t}^{\ell_2}$, with $\mathbf{s}^{\ell_1}$ and $\mathbf{t}^{\ell_2}$ spherical tensors of type-ℓ_1 and type-ℓ_2, respectively. This tensor product is bilinear and equivariant

$$\mathcal{T}^L_M \equiv \left(\mathbf{s}^{\ell_1} \otimes \mathbf{t}^{\ell_2}\right) \equiv \sum_{|m_1|\leq \ell_1} \sum_{|m_2|\leq \ell_2} \mathcal{C}^{(L,M)}_{(\ell_1,m_1),(\ell_2,m_2)} s^{\ell_1}_{m_1} t^{\ell_2}_{m_2} \; . \tag{39}$$

The resulting tensor, however, is not a spherical tensor but can be reduced (decomposed) into a set of spherical tensors of type ranging from $L = |\ell_1 - \ell_2|, \ldots, \ell_1 + \ell_2$, and $|M| \leq L$. This decomposition can be performed using Clebsch-Gordan coefficients, which appear naturally in the coupling of angular momentum in QM. This decomposition allows tensor transformations without breaking equivariance: if we have an L-equivariant function, it allows us to combine information of the two tensors, and retreive only the resulting L-degree tensor, which 'plays well' for the symmetry of the function, such that equivariance is mantained.

As an example[55], consider two spherical tensors of $\ell_1 = 1$ and $\ell_2 = 1$ (spherical harmonics projection of two 3D vectors). The resulting $L = 0$ and $L = 1$ spherical tensor comes from $1 \otimes 1 \to 0$ and $1 \otimes 1 \to 1$ which correspond to the scalar and cross product, respectively,

$$\mathcal{C}^{(L,M)}_{(\ell_1,m_1),(\ell_2,m_2)} \qquad \text{becomes} \qquad \mathcal{C}^{(0,0)}_{(1,i),(1,j)} \sim \delta_{ij} \; , \qquad \mathcal{C}^{(1,i)}_{(1,j),(1,k)} \sim \epsilon_{ijk} \; . \tag{40}$$

The tensor product is, thus, a way to multiply spherical tensors and obtain another set of tensors of specific symmetry, labeled by L. A more general expression of Eq. (39) exists for

products of k spherical harmonics that use generalized Clebsch-Gordan coefficients.

Finally, it is important to understand the *parametrization* of the tensor product. In equivariant models, features might be described by a set of tensors up to a certain degree $L_{\max}$, for example $\mathbf{s} = \{\mathbf{s}^0, \dots, \mathbf{s}^{L_{\max}}\}$ and $\mathbf{t} = \{\mathbf{t}^0, \dots, \mathbf{t}^{L_{\max}}\}$. The corresponding tensor product of these features, $\mathbf{s} \otimes \mathbf{t}$, must produce tensors of the same order to preserve equivariance i.e. up to degree $L_{\max}$. However, for a given degree of the resulting tensor ℓ, there might be different combinations of $\ell_1 = 0, \dots, L_{\max}$ and $\ell_2 = 0, \dots, L_{\max}$. Each of these *paths* are then parametrized by a given learnable weight, and tensors $L > L_{\max}$ are discarded.

E.4 Equivariant MPNNs

To discuss equivariant message-passing, we will use MACE[51] as architecture of reference and the ideas of Refs. [51, 52, 75]. Consider the usual MPNN (explained in Appendix C) where each node i at a layer t is denoted with a tuple $\sigma_i^{(t)} = (\mathbf{r}_i, \boldsymbol{\theta}_i, \mathbf{h}_i^{(t)})$, with $\mathbf{r}_i$ the Cartesian position vector of atom i, $\boldsymbol{\theta}_i$ a set of fixed attributes (such as the atomic number) and $\mathbf{h}_i^{(t)}$ its internal learnable features, updated at each layer t. We denote by $\mathbf{f}_i^{(t)}$ the features of node i (fixed, $\boldsymbol{\theta}_i$ and learnt, $\mathbf{h}_i^{(t)}$). Each message-passing layer consists of three steps.

First, in a message-passing phase, a learnable function $\mathcal{M}_t$ generates a message for each neighbor node $j \in \mathcal{N}_i$, $\mathbf{m}_{ij}^{(t)}$, which are aggregated in a permutationally invariant way

$$\mathbf{m}_i^{(t)} = \bigoplus_{j \in \mathcal{N}_i} \mathbf{m}_{ij}^{(t)} = \bigoplus_{j \in \mathcal{N}_i} \mathcal{M}_t(\sigma_i^{(t)}, \sigma_j^{(t)}) \, . \tag{41}$$

Note that these messages $\mathbf{m}_{ij}^{(t)}$ are two-body in nature (depend on two nodes). In a second phase, the aggregated message is used to update the features of the receiving node with a learnable update function $\mathcal{U}_t$

$$\sigma_i^{(t+1)} = (\mathbf{r}_i, \boldsymbol{\theta}_i, \mathbf{h}_i^{(t+1)}) = (\mathbf{r}_i, \boldsymbol{\theta}_i, \mathcal{U}_t(\sigma_i^{(t)}, \mathbf{m}_i^{(t)})) \, . \tag{42}$$

After all the message passing layers, the node embeddings at each iteration t are used to predict the local energy contribution of each node via a learnable readout function $\mathcal{R}_t$

$$E_i = \sum_t \mathcal{R}_t(\sigma_i^{(t)}) \, . \tag{43}$$

An equivariant model is one where its internal representation is able to transform in an equivariant way with rotations of the atomic coordinates, $\mathbf{r}$,

$$\mathbf{h}_i^{(t)}(\mathcal{R}\mathbf{r}) = \mathcal{D}(\mathcal{R})\mathbf{h}_i^{(t)}(\mathbf{r}) \, . \tag{44}$$

To ensure this, symmetry must be preserved throughout the MPNN procedure. The equivariant message-passing architecture is divided into three steps.

1. Node features $\mathbf{h}_i^{(t)}$ [in this case, the ACE input] are first expressed as spherical tensors. Tensors are segregated according to symmetry L, and same-symmetry tensors are stacked into different non-interacting channels k (the number of channels may depend on L). these features are obtained from a single ACE expansion on a single node, the benefit of MPNN is to be able to connect the ACE expansions, hence the Multi-ACE

2. For each neighboring node, a message of the same structure as $\mathbf{h}_i^{(t)}$ is created. This is done by combining the features of the two nodes using a parametrized tensor product.

3. Finally, to update the node, we allow same-symmetry channels k of the message to mix. The resulting tensors become the updated nodes features.

Internal node feature vectors, $\mathbf{h}_i^{(t)}$, can be embedded into different degree spherical tensors that transform in an equivariant through Wigner D matrices as seen in Eq. (35) (e.g. scalars and vectors transform under $L = 0$ and $L = 1$ Wigner D matrices, respectively). Thus, $\mathbf{h}_i^{(t)}$ can be decomposed into L-symmetry tensors, of which there may be k of them, stacked into k non-interacting (uncoupled) channels. The $2L + 1$ components of each spherical tensor indexed by the symmetry L and channel k, $\mathbf{h}_{i,kL}^{(t)} \in \mathbb{R}^{2L+1}$, are labeled with M, each transforming like a spherical harmonic of degree L and order M, Y_L^M.

To satisfy equivariance, each spherical tensor $\mathbf{h}_{i,kL}^{(t)}$ must transform according to their symmetry

$$\mathbf{h}_{i,kL}(\mathcal{R}\mathbf{r}) = \mathcal{D}^L(\mathcal{R})\mathbf{h}_{i,kL}(\mathbf{r}) \quad \text{or,} \quad h_{i,kLM}(\mathcal{R}\mathbf{r}) = \sum_{M'=-L}^{L} \mathcal{D}^L_{M'M}(\mathcal{R}) h_{i,kLM'}(\mathbf{r}) \; . \tag{45}$$

Messages are generated by combining features of two nodes through the use of a learnt parametrization of the tensor product. As we can combine different L symmetries of each feature in each of the nodes, there may be different combinations of the features that contain the same resulting symmetry. All these combinations are gathered and weighted to produce a final L-symmetry tensor, for each channel k. The details can be found in Refs. [51, 75].

Finally, to update the nodes features, all the k channels of messages of the same symmetry, i.e. $\mathbf{m}_{i,kL}^{(t)}$ (and their components $m_{i,kLM}^{(t)}$), are mixed. This is done with a learnable linear transformation $W_{Lk\tilde{k}}^{(t)}$, a square matrix for each symmetry L, each of dimension $k \times k$, with k the number of channels of that symmetry.

$$h_{i,kLM}^{(t+1)} = U_t(\sigma_i^{(t)}, \mathbf{m}_{i,L}^{(t)}) \equiv \sum_{\tilde{k}} W_{k\tilde{k}L}^{(t)} m_{i,\tilde{k}LM}^{(t)} \tag{46}$$

This can be reexpressed using a block diagonal matrix $W^{(t)}$, composed of $W_{Lk\tilde{k}}^{(t)}$ according to the symmetry and number of channels of the messages,

$$h_{i,kLM}^{(t+1)} = W^{(t)}\mathbf{m}_{i,kL}^{(t)} = \begin{bmatrix} W_0^{(t)} & & & 0 \\ & W_1^{(t)} & & \\ & & \ddots & \\ 0 & & & W_L^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{m}_{i,k0}^{(t)} \\ \mathbf{m}_{i,k1}^{(t)} \\ \vdots \\ \mathbf{m}_{i,kL}^{(t)} \end{bmatrix} \tag{47}$$

This framework ensures that, given a rotation of the atomic coordinates $\mathbf{r}$, thanks to the symmetry label segregation of the embedding, the model understands how internal features and messages should be modified, resulting in an equivariant message-passing model.

F Pretrained models

This section provides general information on the models used for testing in Sec. 5, such as the training dataset[e], chemical environment description and general applicability. References and GitHub pages and documentation (if available) are also provided. All models are easily integrated with `ASE` simulation environment[77]. Additional details on the installation for each model can be found in this project's `GitHub`.

It is important to note that in this project, only pretrained models have been used, as in order to generate a NNP, a lot of data for a wide variety of systems (if the model is to be transferable) is needed, and as such, computational time. Pretrained models are trained once on expensive and diverse datasets with the goal of capturing broad interaction patterns. One can use pretrained models directly to make predictions on similar systems the model was trained on, or instead, tweaking the model with system-specific additional data to increase its predictive performance on particular systems, process known as 'fine-tunning'. In this sense, pretrained models are also called 'foundational models', as they serve as a versatile foundation for fine-tunning.

F.1 Training datasets and applicability

`ANI` family of models consist of three models `ANI-1x`[65], `ANI-1ccx` and `ANI-2x`[23]. The pretrained models have been made available in Python by TorchANI[77] (`GitHub`, `Documentation`). These models are trained using `ANI-1` and `ANI-2` datasets, containing small organic molecules, more specifically:

`ANI-1x`. Trained on 5 million non-equilibrium small and midsized organic molecules, containing total energies, forces, dipoles and quadrupoles. Atomic coverage includes H, C, N, O.

`ANI-1ccx`. Fine-tunned model based on `ANI-1x`, with additional training data of the higher-level CCSD(T) ES method.

`ANI-2x`. Refined `ANI-1x` model with an expanded training set of 8.9 million non-equilibrium molecules, with special focus on torsions. Dataset properties are the same as in `ANI-1x`, with an expanded coverage to F, S and Cl.
These are invariant descriptor-based models (see Appendix D) using a modified version of ACSF of Ref. [59], which contain a two- and three-body order description of the chemical environment based on distances and angles. Forces are predicted in a conservative way as the analytical gradient of the potential energy with respect to nuclear displacements.

`MACE` family of models[22,51,72,73] consist of two sets of pretrained models, `MACE-MP` and `MACE-OFF`, each with several subvariants (`GitHub`, `Documentation`). More specifically:

- `MACE-MP`[22]. A foundational model focused on materials trained on the `MPtrj` dataset, with $\sim$ 145 thousand periodic DFT calculations including energies, forces and magnetic moments.
- `MACE-OFF`[72]. Focused on drug-like small organic molecules trained on the `SPICE` dataset, containing 1.1 million DFT calculations on small molecules interacting with proteins, with formation and total energies, forces and multipole expansion up to octupoles.

[e]Information on training datasets have are referenced from Ref. [76].

These are equivariant models (see Appendix E) based on an high-order generalized ACE[67] description of the chemical environment, with a MPNN architecture. Forces are predicted the as in the `ANI` family of models.

`ORB` (or simply `Orbital`) family of models[24] consist of two pretrained models, `ORB-v2` and `ORB-D3-v2` (`GitHub`).

- `ORB-v2`. A foundational model focused on materials trained on `MPtrj` and `Alexandia`, the later dataset consisting of 2.5 million DFT calculations of several materials.
- `ORB-D3-v2`. The same model as `ORB-v2` but trained with a dispersion-corrected D3 version of its datasets.

`ORB` models are not invariant/equivariant, instead they opt to 'teach' rotational invariance through augmenting data during training rather than strictly enforcing it as an inductive bias in the model. This increases prediction speed while using less memory, at the expense of physical adequacy. Additionally, it predicts forces in a non-conservative way: rather than using the gradient of the potential energy, forces are predicted by the model itself, resulting in a near but not strict conservation of energy, increasing speed at the expense of adequacy.

G Additional results

This section aims to list additional results following the same protocols and discussions as in Sec. 5. A more complete and organized set of results for this project can be found in the GitHub repository.

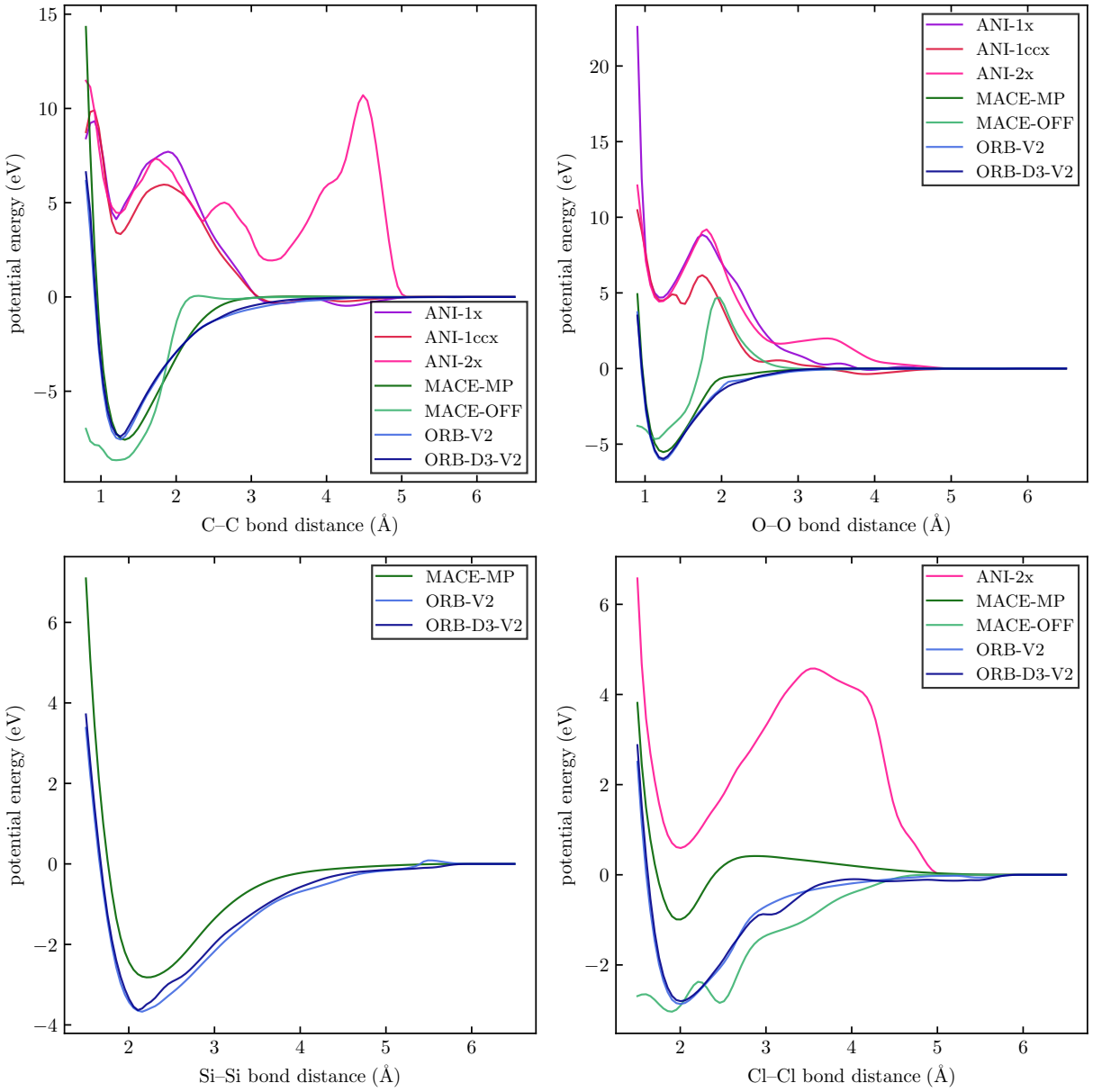


Figure 29

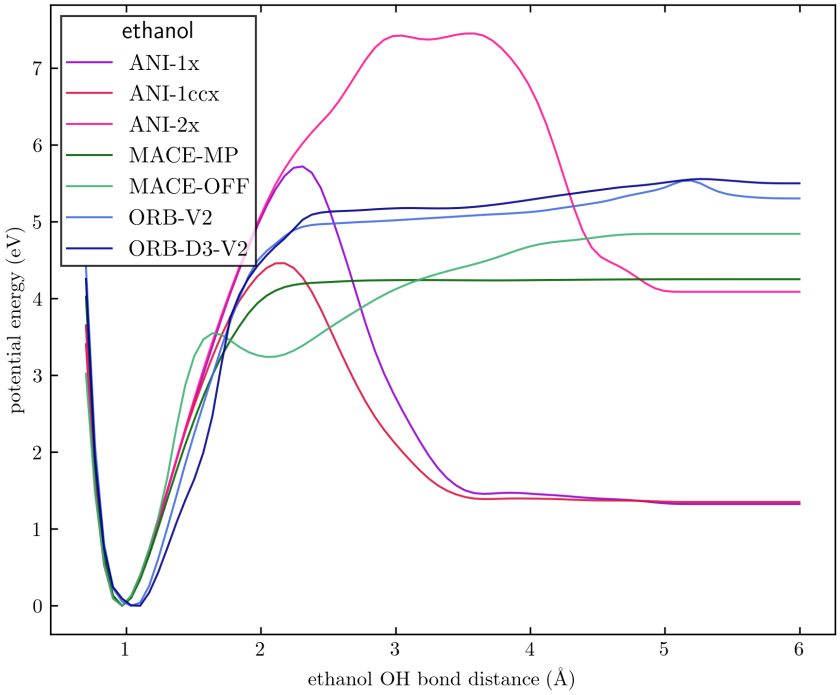
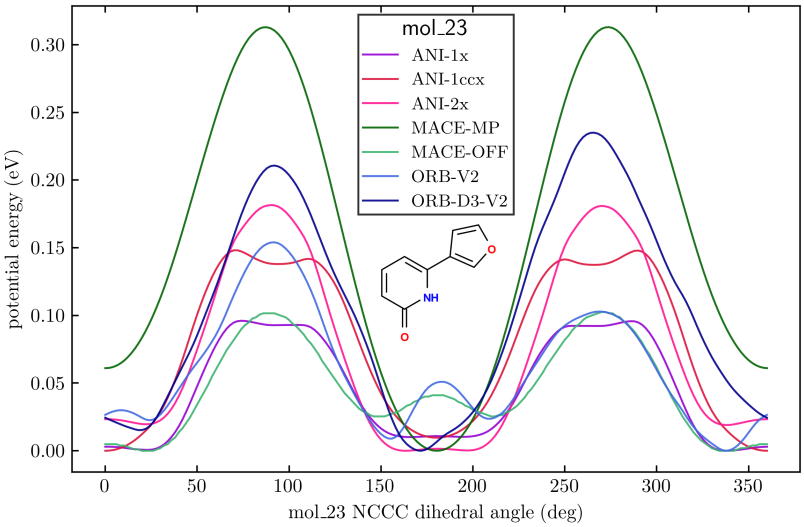
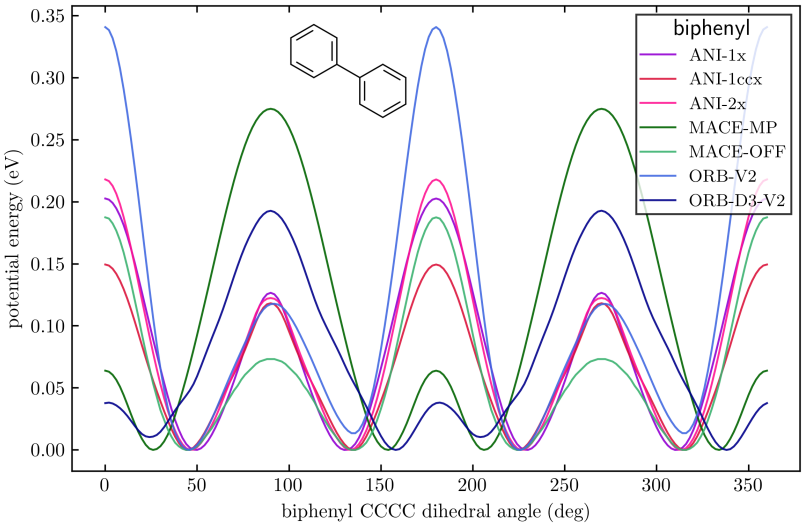

ethanol
ANI-1x
ANI-1ccx
ANI-2x
MACE-MP
MACE-OFF
ORB-V2
ORB-D3-V2
potential energy (eV)
ethanol OH bond distance (Å)
0
1
2
3
4
5
6
7
1
2
3
4
5
6

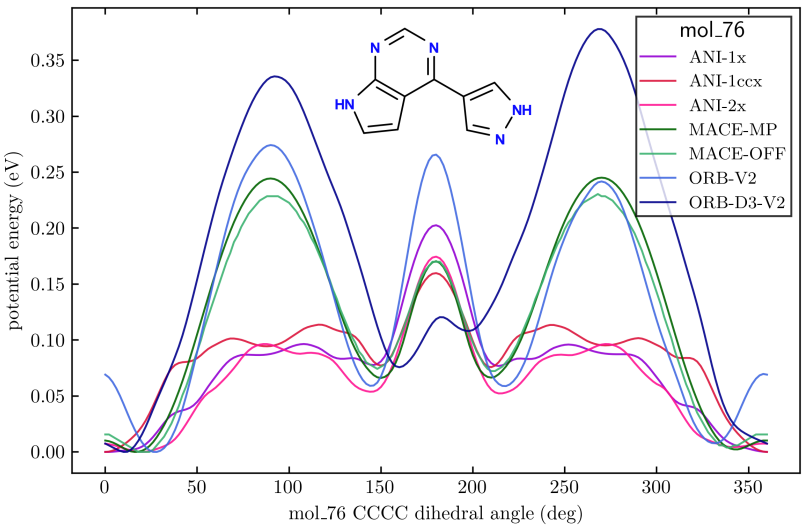
Figure 30



(a)

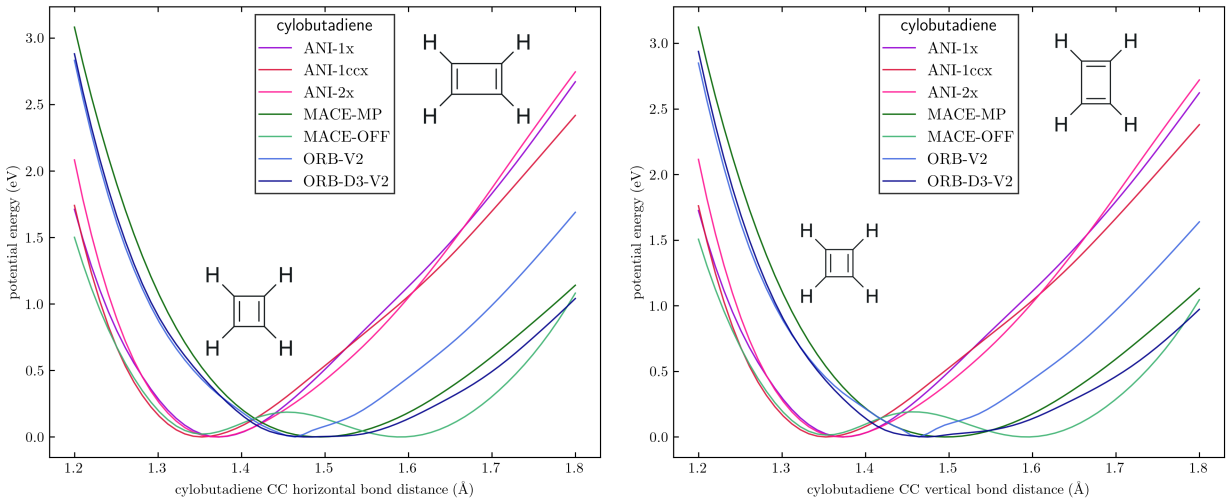


(b)



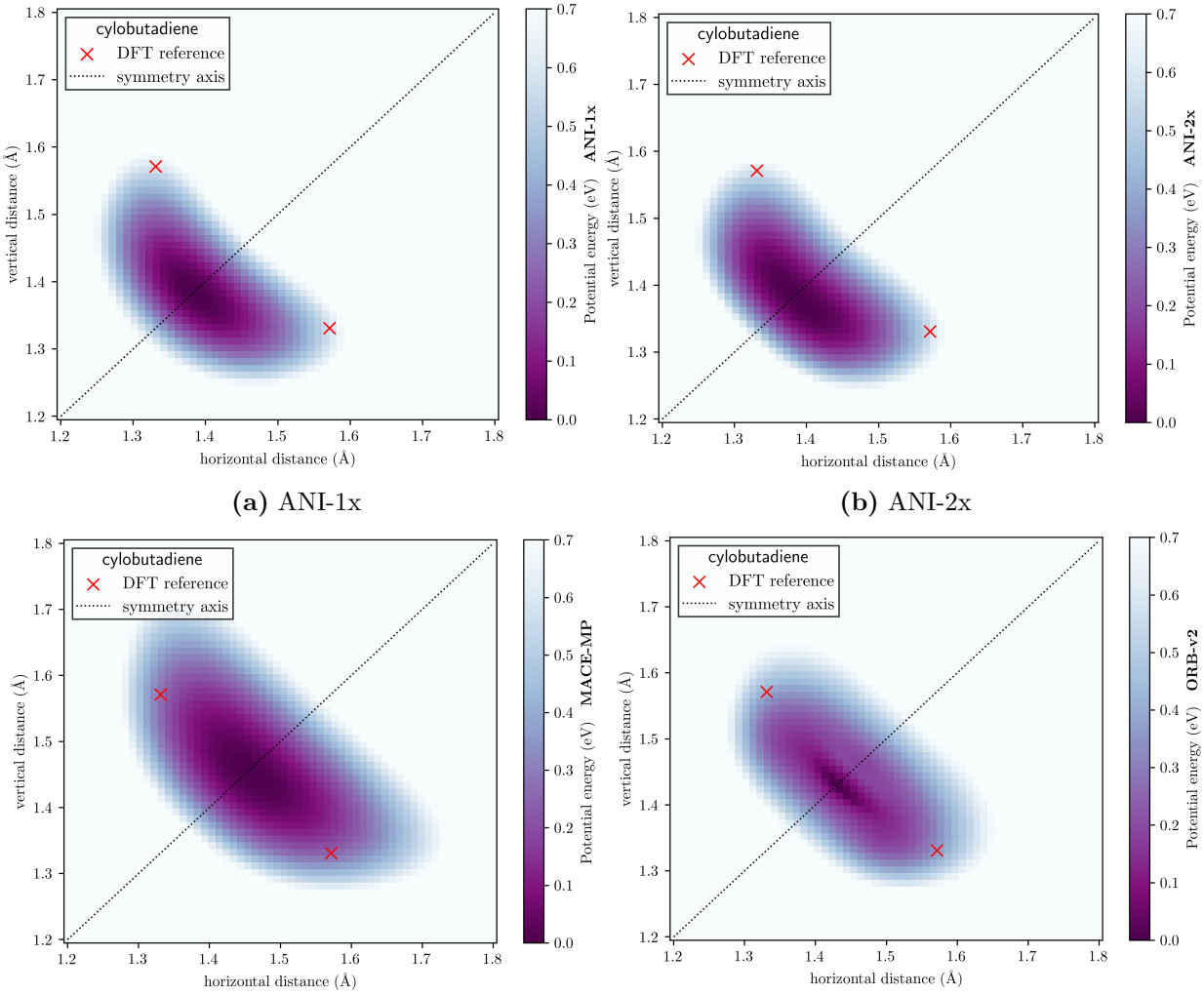
(c)

Figure 31



(a) horizontal **(b)** vertical

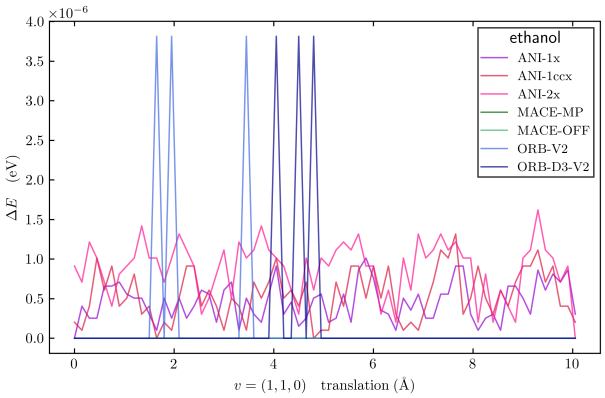
Figure 32



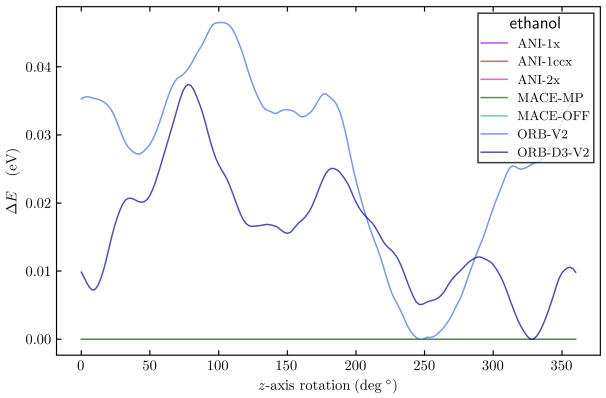
(a) ANI-1x **(b)** ANI-2x

(c) MACE-MP **(d)** ORB-v2

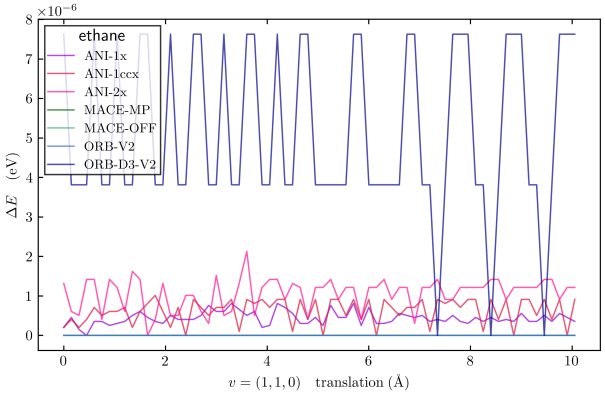
Figure 33



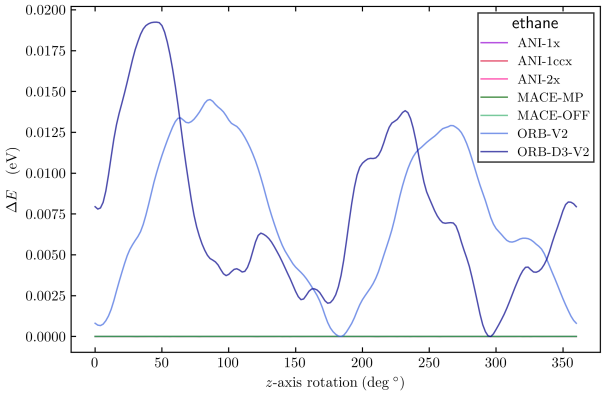
(a)



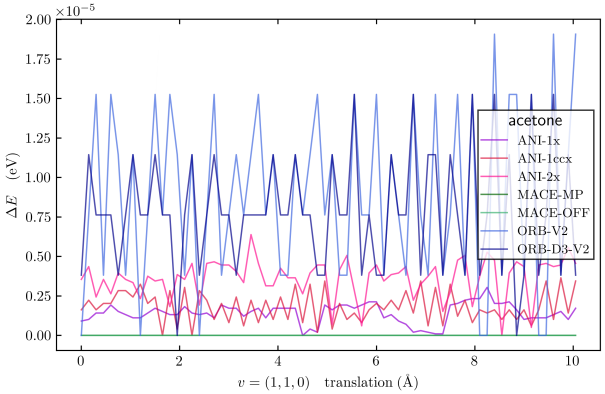
(b)



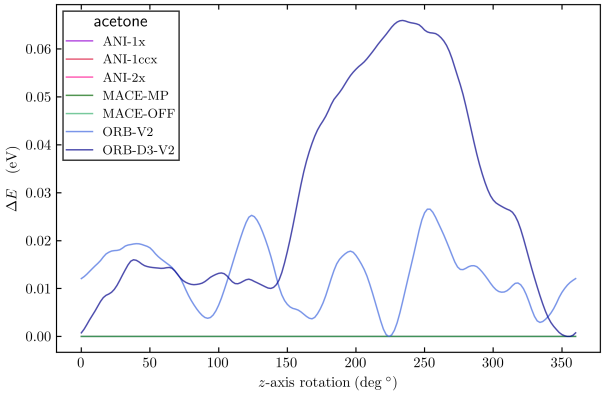
(c)



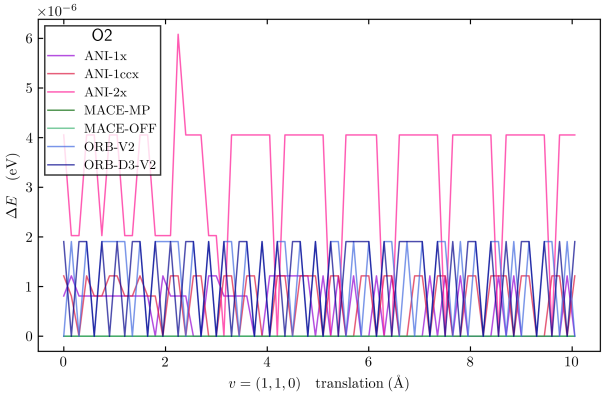
(d)



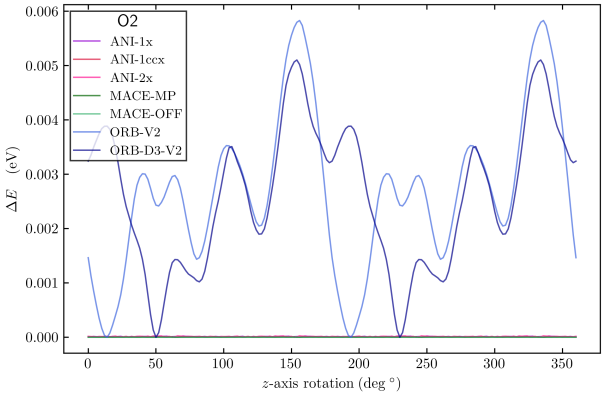
(e)



(f)



(g)



(h)

Figure 34

Acronyms

ACE Atomic Cluster Expansion 57

ACSF Atom Centered Symmetry Function 56

CASSCF Complete Active Space Self Consistent Field 30

CC Coupled Cluster 30

CCSD(T) Coupled Cluster with single, double and perturbational triples 30, 56

CI Configuration interaction 30

CNN Convolutional Neural Network 32, 35, 36

DFT Density Functional Theory 25, 26, 30, 56, 57

DL Deep Learning 6

ES Electronic Structure 3, 4, 5, 40, 42, 56

FF Forcefield 3, 4, 5, 47

GAT Graph Attention Network 36, 37, 38, 39

GCN Graph Convolution Network 36, 37, 38, 39

GGA Generalized Gradient Approximation 30

GNN Graph neural network 7, 8, 32, 33, 35, 36, 37, 38, 39, 40, 41, 49

HF Hartree-Fock 29

LDA Local Density Approximation 30

ML Machine Learning 1, 2, 4, 5, 6, 7, 8, 9, 10, 15, 31, 32, 42, 43

MLIP Machine Learning Interatomic Potential 2, 3, 4, 5, 42

MLP Machine learning potential / Multi-layer perceptron 6, 32, 43, 45, 46

MM Molecular Mechanics 4, 5

MP Møller-Plesset (perturbation theory method) 30

MP2 Møller-Plesset second order perturbation theory 30

MPNN Message Passing Neural Network 8, 32, 37, 38, 39, 49, 54, 57

MSE Mean Squared Error 5, 6, 9

NN Neural Network 6, 7, 9, 40, 45, 46, 48

NNP Neural network potential 1, 4, 11, 12, 15, 19, 20, 21, 39, 40, 43, 45, 46, 47, 49, 50, 56

PDE Partial Differential Equation 9

PES Potential Energy Surface 29

PFF Polarizable Force Field 4

PINN Physics Informed Neural Network 9, 42

QM Quantum Mechanics 2, 4, 5, 42, 53

SCF self-consistent field 29, 30

SQM Semiempirical Quantum Mechanics (in the context of ES methods) 4

TISE Time independent Schrödinger equation 3, 28, 29, 42

References

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with alphafold 3", Nature **630**, 493 (2024).

[2] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, "Protein–ligand scoring with convolutional neural networks", Journal of Chemical Information and Modeling **57**, PMID: 28368587, 942 (2017).

[3] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design", Science Advances **4**, eaap7885 (2018).

[4] E. Schrödinger, "An undulatory theory of the mechanics of atoms and molecules", Phys. Rev. **28**, 1049 (1926).

[5] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields", Chemical Reviews **121**, PMID: 33705118, 10142 (2021).

[6] J. Hermann, Z. Schätzle, and F. Noé, "Deep-neural-network solution of the electronic schrödinger equation", Nature Chemistry **12**, 891 (2020).

[7] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks", Science **355**, 602 (2017).

[8] F. Sabanés Zariquiey, R. Galvelis, E. Gallicchio, J. D. Chodera, T. E. Markland, and G. De Fabritiis, "Enhancing protein–ligand binding affinity predictions using neural network potentials", Journal of Chemical Information and Modeling **64**, 1481 (2024).

[9] P. A. M. Dirac and R. H. Fowler, "Quantum mechanics of many-electron systems", Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **123**, 714 (1929).

[10] A. Szabo and N. S. Ostlund, *Modern quantum chemistry*, Dover Books on Chemistry (Dover Publications, Mineola, NY, Jan. 1996).

[11] F. Jensen, *Introduction to computational chemistry*, 3rd ed. (John Wiley & Sons, Nashville, TN, Feb. 2017).

[12] C. J. Cramer, *Essentials of computational chemistry*, 2nd ed. (John Wiley & Sons, Chichester, England, Sept. 2004).

[13] A. R. Leach, *Molecular modelling*, 2nd ed. (Prentice-Hall, London, England, Jan. 2001).

[14] J. E. Jones and S. Chapman, "On the determination of molecular fields. —ii. from the equation of state of a gas", Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **106**, 463 (1924).

[15] T. A. Halgren and W. Damm, "Polarizable force fields", Current Opinion in Structural Biology **11**, 236 (2001).

[16] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, "Polarizable atomic multipole-based amoeba force field for proteins", Journal of Chemical Theory and Computation **9**, PMID: 24163642, 4046 (2013).

[17] F. Vitalini, A. S. J. S. Mey, F. Noé, and B. G. Keller, "Dynamic properties of force fields", The Journal of Chemical Physics **142**, 084101 (2015).

[18] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, "Development and use of quantum mechanical molecular models. 76. am1: a new general purpose quantum mechanical molecular model", Journal of the American Chemical Society **107**, 3902 (1985).

[19] M. J. S. Dewar and W. Thiel, "Ground states of molecules. 38. the mndo method. approximations and parameters", Journal of the American Chemical Society **99**, 4899 (1977).

[20] J. J. P. Stewart, "Application of the pm6 method to modeling proteins", Journal of Molecular Modeling **15**, 765 (2009).

[21] M. Elstner, "The scc-dftb method and its application to biological systems", Theoretical Chemistry Accounts **116**, 316 (2006).

[22] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, "A foundation model for atomistic materials chemistry", 2024, arXiv:2401.00096.

[23] 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 **16**, 4192 (2020).

[24] M. Neumann, J. Gin, B. Rhodes, S. Bennett, Z. Li, H. Choubisa, A. Hussey, and J. Godwin, "Orb: a fast, scalable neural network potential", 2024, arXiv:2410.22570.

[25] S.-L. J. Lahey and C. N. Rowley, "Simulating protein–ligand binding with neural network potentials", Chem. Sci. **11**, 2362 (2020).

[26] R. Galvelis, A. Varela-Rial, S. Doerr, R. Fino, P. Eastman, T. E. Markland, J. D. Chodera, and G. De Fabritiis, "Nnp/mm: accelerating molecular dynamics simulations with machine learning potentials and molecular mechanics", Journal of Chemical Information and Modeling **63**, PMID: 37694852, 5701 (2023).

[27] K. Zinovjev, L. Hedges, R. Montagud Andreu, C. Woods, I. Tuñón, and M. W. van der Kamp, "Emle-engine: a flexible electrostatic machine learning embedding package for multiscale molecular dynamics simulations", Journal of Chemical Theory and Computation **20**, PMID: 38804055, 4514 (2024).

[28] Y. Bengio, *Deep learning*, Adaptive Computation and Machine Learning series (MIT Press, London, England, Nov. 2016).

[29] P. Baldi, *Deep learning in science* (Cambridge University Press, Cambridge, England, July 2021).

[30] M. Erdmann, J. Glombitza, G. Kasieczka, and U. Klemradt, *Deep learning for physics research* (World Scientific Publishing, Singapore, Singapore, June 2021).

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature **521**, 436 (2015).

[32] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", Neural Networks **2**, 359 (1989).

[33] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop", in *Neural networks: tricks of the trade: second edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012), pp. 9–48.

[34] C. M. Bishop, *Neural networks for pattern recognition* (Oxford University Press, Nov. 1995).

[35] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning", Nature Reviews Physics **3**, 422 (2021).

[36] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", Journal of Computational Physics **378**, 686 (2019).

[37] G. P. P. Pun, R. Batra, R. Ramprasad, and Y. Mishin, "Physically informed artificial neural networks for atomistic modeling of materials", Nature Communications **10**, 2339 (2019).

[38] Y. Wang, T. J. Inizan, C. Liu, J.-P. Piquemal, and P. Ren, "Incorporating neural networks into the amoeba polarizable force field", The Journal of Physical Chemistry B **128**, PMID: 38445577, 2381 (2024).

[39] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations", Phys. Rev. Lett. **125**, 166001 (2020).

[40] Y. Zhang, J. Xia, and B. Jiang, "Physically motivated recursively embedded atom neural networks: incorporating local completeness and nonlocality", Phys. Rev. Lett. **127**, 156002 (2021).

[41] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms", Journal of Physics: Condensed Matter **29**, 273002 (2017).

[42] S.-L. J. Lahey, T. N. Thien Phuc, and C. N. Rowley, "Benchmarking force field and the ani neural network potentials for the torsional potential energy surface of biaryl drug fragments", Journal of Chemical Information and Modeling **60**, 6258 (2020).

[43] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln", Annalen der Physik **389**, 457 (1927).

[44] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: grids, groups, graphs, geodesics, and gauges", 2021, arXiv:2104.13478 [cs.LG].

[45] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: a review of methods and applications", AI Open **1**, 57 (2020).

[46] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model", IEEE Transactions on Neural Networks **20**, 61 (2009).

[47] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", 2017, arXiv:1609.02907 [cs.LG].

[48] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks", 2017, arXiv:1710.10903.

[49] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry", in Proceedings of the 34th international conference on machine learning - volume 70 (2017), pp. 1263–1272.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", 2023, arXiv:1706.03762 [cs.CL].

[51] I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, and G. Csányi, "Mace: higher order equivariant message passing neural networks for fast and accurate force fields", 2022, arXiv:2206.07697.

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

[53] V. G. Satorras, E. Hoogeboom, and M. Welling, "E(n) equivariant graph neural networks", 2022, arXiv:2102.09844 [cs.LG].

[54] M. Geiger and T. Smidt, "E3nn: euclidean neural networks", 2022, arXiv:2207.09453 [cs.LG].

[55] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: rotation- and translation-equivariant neural networks for 3d point clouds", 2018, arXiv:1802.08219 [cs.LG].

[56] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity", Chemical Science **10**, 370 (2019).

[57] T. K. Rusch, M. M. Bronstein, and S. Mishra, "A survey on oversmoothing in graph neural networks", 2023, arXiv:2303.10993 [cs.LG].

[58] U. Alon and E. Yahav, "On the bottleneck of graph neural networks and its practical implications", 2021, arXiv:2006.05205 [cs.LG].

[59] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces", Phys. Rev. Lett. **98**, 146401 (2007).

[60] 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 **121**, PMID: 34232033, 9816 (2021).

[61] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials", The Journal of Chemical Physics **134**, 074106 (2011).

[62] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials", Chemical Reviews **121**, PMID: 34310133, 9759 (2021).

[63] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments", Phys. Rev. B **87**, 184115 (2013).

[64] L. Himanen, M. O. 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 **247**, 106949 (2020).

[65] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost", Chem. Sci. **8**, 3192 (2017).

[66] J. Behler, "Constructing high-dimensional neural network potentials: a tutorial review", International Journal of Quantum Chemistry **115**, 1032 (2015).

[67] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials", Phys. Rev. B **99**, 014104 (2019).

[68] D. P. Kovács, C. v. d. Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner, and G. Csányi, "Linear atomic cluster expansion force fields for organic molecules: beyond rmse", Journal of Chemical Theory and Computation **17**, PMID: 34735161, 7696 (2021).

[69] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "Wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials", The Journal of Chemical Physics **148**, 241709 (2018).

[70] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons", Phys. Rev. Lett. **104**, 136403 (2010).

[71] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: a continuous-filter convolutional neural network for modeling quantum interactions", in Proceedings of the 31st international conference on neural information processing systems (2017), pp. 992–1002.

[72] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole, and G. Csányi, "Mace-off: transferable short range machine learning force fields for organic molecules", 2023, arXiv:2312.15211.

[73] D. P. Kovács, I. Batatia, E. S. Arany, and G. Csányi, "Evaluation of the mace force field architecture: from medicinal chemistry to materials science", The Journal of Chemical Physics **159**, 044118 (2023).

[74] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations", Nature Machine Intelligence **3**, 1023 (2021).

[75] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, "The design space of e(3)-equivariant atom-centered interatomic potentials", 2022, arXiv:2205.06643 [stat.ML].

[76] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, and S. Tretiak, "Data generation for machine learning interatomic potentials and beyond", Chemical Reviews **124**, PMID: 39572011, 13681 (2024), eprint: https://doi.org/10.1021/acs.chemrev.4c00572.

[77] 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 **60**, PMID: 32568524, 3408 (2020).