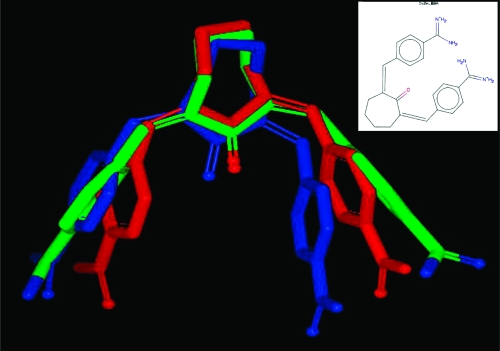
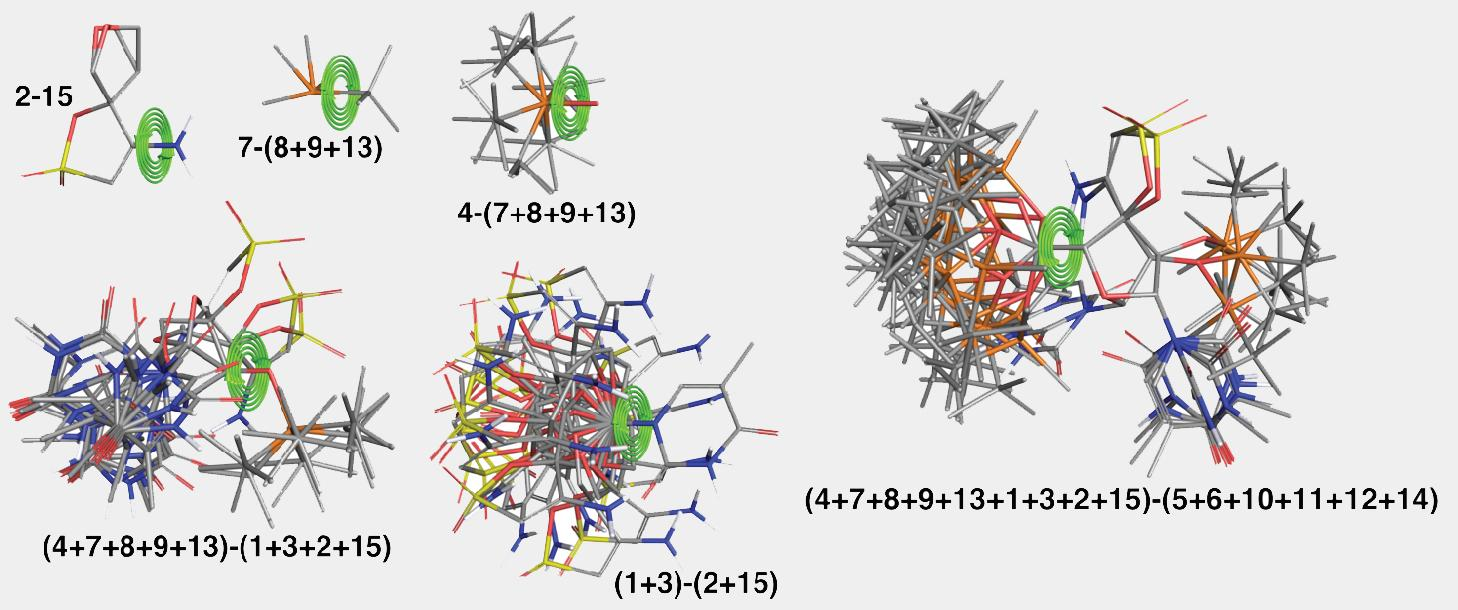
## **GeoMol Tests**

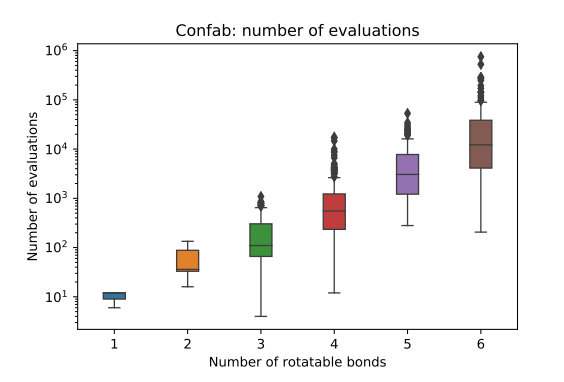


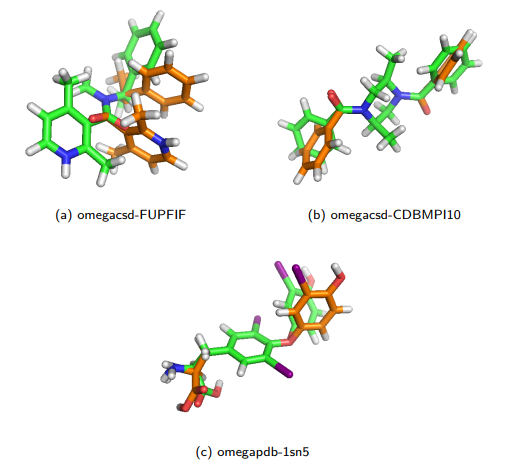


**Bayesian Optimization Results**

## 

## 





**Using the Platinum Diverse Dataset**

## **Generator Maximum ensemble size 50 Maximum ensemble size 250**

## **Mean Median Min. Max. Mean Median Min. Max.**

## **RMSD (Å)**

## **Balloon DG 0.967 0.788 0.027 4.564 0.806 0.613 0.025 3.996**

## **Balloon GA 0.989 0.858 0.027 4.606 0.833 0.711 0.027 4.336**

## **CONFORT**

## **Systematic Best**

## **0.669 0.486 0.030 3.919 0.559 0.416 0.030 3.675**

## **CONFORT**

## **Systematic**

## **Default**

## **0.683 0.548 0.036 3.140 0.616 0.524 0.036 2.793**

## **Conformator**

## **Best**

## **0.680 0.589 0.020 3.255 0.568 0.474 0.020 2.931**

## **Conformator**

## **Fast**

## **0.747 0.654 0.020 3.511 0.637 0.538 0.020 3.329**

## **iCon Best 0.701 0.588 0.030 3.662 0.639 0.569 0.030 3.662**

## **iCon Fast 0.719 0.535 0.030 3.921 0.598 0.474 0.030 3.662**

## **RDKit ETKDGv3 0.711 0.590 0.038 4.386 0.604 0.518 0.038 3.367**

## **RDKit KDG 0.738 0.596 0.038 4.156 0.621 0.515 0.038 3.638**

## 

## 

**Generator Total execution time (hh:mm:ss)**

**a Number of failed molecules**

**Balloon DG 14:31:35 9**

**Balloon GA 27:20:21 11**

**CONFORT Stochastic 56:30:54 0**

**Conformator Best 26:29:35 7**

**Conformator Fast 10:46:42 9**

**iCon Best 01:48:45 1**

**iCon Fast 01:40:27 1**

**RDKit ETKDGv3 39:09:47 0**

**RDKit KDG 18:03:17 0**

## **array([[ 6. , 2.33166293, -0.52962788, -0.48097309],**

## **[ 6. , 0.0948792 , 1.07597567, -1.33579553],**

## **[ 6. , -2.40436371, -0.29483572, -0.90388318],**

## **[ 1. , 2.18166462, -0.95639011, 1.569049 ],**

## **[ 1. , 4.1178375 , 0.51816193, -0.81949623],**

## **[ 1. , 2.39319787, -2.32844253, -1.56157176],**

## **[ 1. , 0.29919987, 1.51730566, -3.37889252],**

## **[ 1. , 0.08875543, 2.88229706, -0.26437996],**

## **[ 1. , -3.99100651, 0.92016315, -1.54358853],**

## **[ 1. , -2.66167993, -0.71627602, 1.136556 ],**

## **[ 1. , -2.45014726, -2.08833123, -1.99406318],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ],**

## **[ 0. , 0. , 0. , 0. ]])**

## **-------------------------------**

## **resnum Type TRUE/FALSE**

## **-------------------------------**

## **DEFAULT BB FALSE**

## **DEFAULT SC FALSE**

## **DEFAULT NU FALSE**

## **DEFAULT BRANCH FALSE**

## **050 BB TRUE**

## **051 BB TRUE**

## **-------------------------------**

## **jumpnum Type TRUE/FALSE**

## **-------------------------------**

## **DEFAULT JUMP FALSE**

## **-------------------------------**

## **resnum atomnum Type TRUE/FALSE**

## **-------------------------------**

## **DEFAULT PHI FALSE**

## **DEFAULT THETA FALSE**

## **DEFAULT D FALSE**

## **DEFAULT RB1 FALSE**

## **DEFAULT RB2 FALSE**

## **DEFAULT RB3 FALSE**

## **DEFAULT RB4 FALSE**

## **DEFAULT RB5 FALSE**

## **DEFAULT RB6 FALSE**

## **-------------------------------**

core.scoring.ScoreFunctionFactory: SCOREFUNCTION: ref2015

Mover name: MinMover, Mover type: MinMover, Mover current tag:NoTag

Minimization type: linmin

Scorefunction: ref2015

Score tolerance: 0.01

Nb list: True

Deriv check: False

Movemap:

-------------------------------

resnum Type TRUE/FALSE

-------------------------------

DEFAULT BB FALSE

DEFAULT SC FALSE

DEFAULT NU FALSE

DEFAULT BRANCH FALSE

040 BB TRUE

041 BB TRUE

042 BB TRUE

043 BB TRUE

044 BB TRUE

045 BB TRUE

046 BB TRUE

047 BB TRUE

048 BB TRUE

049 BB TRUE

050 BB TRUE

051 BB TRUE

052 BB TRUE

053 BB TRUE

054 BB TRUE

055 BB TRUE

056 BB TRUE

057 BB TRUE

058 BB TRUE

059 BB TRUE

060 BB TRUE

-------------------------------

jumpnum Type TRUE/FALSE

-------------------------------

DEFAULT JUMP FALSE

-------------------------------

resnum atomnum Type TRUE/FALSE

-------------------------------

DEFAULT PHI FALSE

DEFAULT THETA FALSE

DEFAULT D FALSE

DEFAULT RB1 FALSE

DEFAULT RB2 FALSE

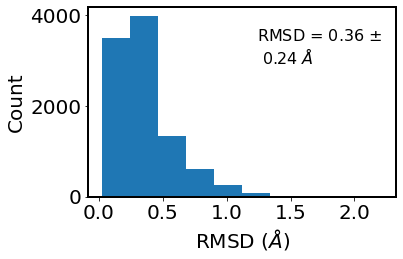
DEFAULT RB3 FALSE

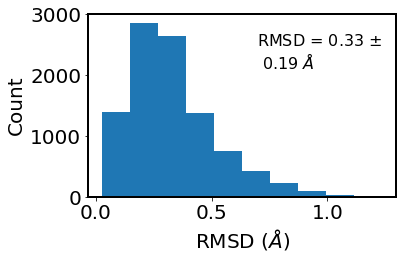
DEFAULT RB4 FALSE

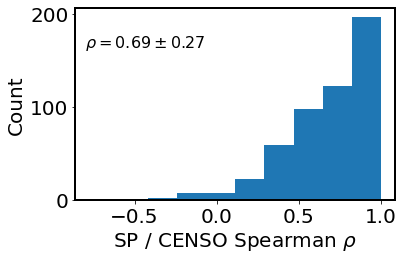
DEFAULT RB5 FALSE

DEFAULT RB6 FALSE

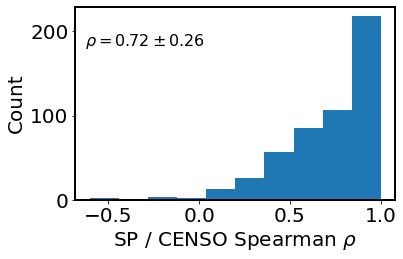
-------------------------------

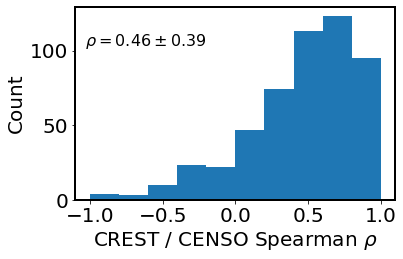
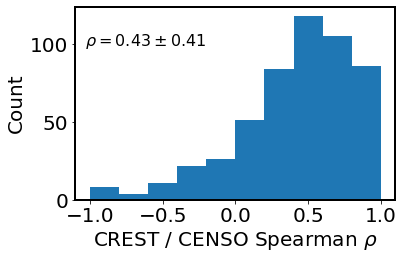






relative to closest CREST conf





## 

## **Descriptors**

### **Global descriptors**

The descriptor is a numerical representation (usually in a vector form) of a 3D molecular geometry used as input to the ML algorithm. One of the most obvious and simplest descriptors is the Cartesian (*XYZ*) coordinates traditionally used in computational chemistry. It can be classified as a global descriptor (GD) because it describes the entire molecule. *XYZ* coordinates were successfully applied for geometry optimizations accelerated by MLPs

However, such a descriptor does not guarantee by itself compliance with physical constraints, *e.g.*, that rotation and translation of an isolated molecule do not change the potential energy. Other global descriptors used in MLPs are internal coordinates, which are also traditional molecular representations in computational chemistry. They satisfy rotational and translational invariance, but their construction is not unique. This handicap motivated the development and use of specialized global descriptors for MLPs such as the Coulomb matrix (CM), inverse internuclear distances (IDs),their version normalized relative to equilibrium structure (RE descriptor), bag-of-bonds (BoB), and BAML (bonds, angles machine learning).

However, many of the above global descriptors (CM, ID, RE) are not invariant with respect to the permutation of chemically equivalent atoms. Addressing this problem is rather challenging. Some approaches (such as BoB) sort atoms making the representation permutationally invariant. Nevertheless, sorting atoms should be done with care. In many situations (*e.g.*, when using CM, ID, and RE), it may cause PES discontinuities and low accuracy because small geometry changes may lead to drastic changes in the descriptor. Another solution is summing up the terms arising from internuclear distances for each distinct atom pair as in encoded bonds (similarly to the approaches adopted in LDs). However, this approach leads to information loss as the structure cannot be uniquely reconstructed from such a descriptor. A more rigorous solution is to use the permutationally invariant polynomials or fundamental invariants as descriptors. Nevertheless, due to the quickly growing number of permutations with system size, they can only be applied to small molecules. Other solutions go beyond the descriptor itself. One can learn permutational invariance by expanding the training set with randomly sorted atoms or modifying the ML algorithm. This latter approach is adopted in permutationally invariant KREG (pKREG) and related approaches such as sGDML and RKHS + F (reproducing kernel Hilbert space using energies and forces), which use permutationally invariant kernel functions.

Finally, the global descriptors' dimensionality grows with the number of atoms. Although this is not an issue for the PES of small systems like those we discuss here, it may be a problem when dealing with chemically diverse datasets. Thus, special solutions with global representations have been suggested.

### **Local descriptors**

Alternatively to global descriptors, one can construct a suitable vector representation for MLPs based on an atom-centered approach. Such a molecular representation is designed to encode the local chemical environment around each atom through a set of basis functions that vanish outside of a cutoff sphere. These representations are called local descriptors (LDs).

Following the concept of many-body expansion, MLPs with local descriptors model the interatomic interactions by decomposing them into *n*-body terms, typically expressed as a combination of radial and angular distribution functions. The advantages of LDs are that they are permutationally invariant and size-extensive by construction and can achieve linear scaling. The list of developed LDs for molecules is extensive. It includes, among others, BP-ACSFs (Behler–Parrinello's atom-centered symmetry functions) and its ANI-AEV (atomic environment vectors) and wACSF (weighted ACSF) modifications, SOAP (smooth overlap of atomic positions), aSLATM (atomic spectrum of London and Axilrod–Teller–Muto), FCHL (Faber–Christensen–Huang–Lilienfeld), Gaussian moments, spherical Bessel functions, and descriptors used in DPMD (deep potential molecular dynamics) and DeepPot-SE (DPMD-smooth edition).

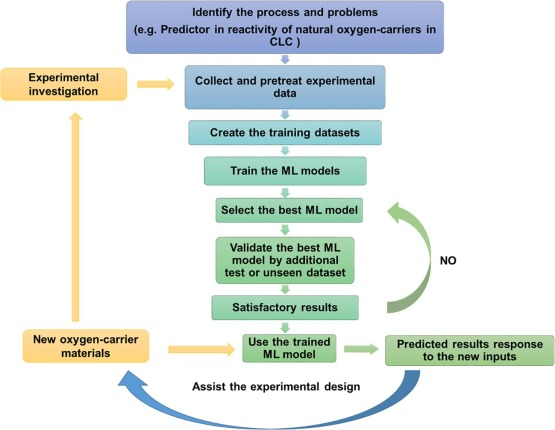
Local descriptors can be fixed before training an MLP. Alternatively, the ML algorithm can learn them as a part of the training process (discussed below in the context of NNs). To differentiate these two cases, we denote the first type as fixed LDs (fLDs) and the second type as learned LDs (lLDs).

### **Global *vs.* local descriptors**

Typically, an MLP using a local descriptor introduces an approximated total energy partitioning into contributions of each atom, whose environment is described only within a predefined cutoff radius. For a single molecule PES, this approximation may be considered conceptually inferior to direct learning of total energies with MLPs using global descriptors. Moreover, many global descriptors are complete descriptors from which the molecular geometry can be reconstructed unambiguously, while the reconstruction is a challenging issue for local descriptors, which can lead to accuracy loss. However, one should remember that the models based on local descriptors were developed with the motivation to make accurate predictions for larger systems even when trained only on smaller systems rather than solely apply such models to a single molecule PES. Thus, while the elements of global descriptors are often based only on the internuclear distances, local descriptors are usually augmented with additional structural parameters such as angles, which take into account many-body environmental factors to achieve the required accuracy. From a practical perspective, training global descriptors is also more straightforward as they typically do not have extra tunable parameters besides those already existing in the ML algorithms. Local descriptors have additional hyperparameters controlling the cutoff, shape, and size of the basis functions. They are usually system-dependent, and it is recommended to fine-tune them to improve the performance of the final MLP model.

Although global and local descriptors are conceptually different, a local descriptor effectively becomes a global one if no cutoff is used. Another way of constructing a global version of a local descriptor is simply by taking the average over all environments. In this sense, the distinction between local and global descriptors is fuzzy.

In principle, we can use the same descriptor with various ML algorithms. For example, CM was successfully used with both kernel methods and NNs. Thus, several specialized packages such as *DScribe* and *MolML* exist for generating structural descriptors from the molecular 3D geometry. However, many of the descriptors mentioned above have been either developed for or are most often used with a specific ML algorithm available in popular packages. For example, CM is typically used with KRR-based models, mainly because early benchmark studies showed that the combination of KRR-CM is superior to the combination of NN-CM.



Flexible molecules have many conformational minima which can be easily reached by torsional motions of the molecular framework in the potential energy surface (PES). For the last few years, there are methods, as the multi-structural harmonic-oscillator (MS-HO) approximation ([Zheng et al., 2011a](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B24)), which take into account the characteristics of all these equilibrium structures. Specifically, the MS-HO method incorporates the rotational and vibrational (rovibrational) partition function of each of the conformers within the rigid-rotor harmonic-oscillator approximation. This is a substantial improvement over the one-well harmonic oscillator (1W-HO) approximation in which the structure of the absolute minimum is the only one to be considered ([Ferro-Costas et al., 2018b](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B7)).

Locating all conformers is just the first step toward the evaluation of more accurate rovibrational partition functions. For instance, it has been shown that MS-HO partition functions improve over 1W-HO ones ([Ferro-Costas et al., 2018b](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B7)), additionally torsional anharmonicity should be also included ([Yu et al., 2011](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B19); [Zheng et al., 2011b](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B25); [Zheng and Truhlar, 2013](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B23)) to increase the accuracy of the results. The most reliable methods that incorporate torsional anharmonicity can only be applied to a reduced number of torsional degrees of freedom ([Fernández-Ramos, 2013](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B4)) and they require more information of the PES than just the minima. For instance, the extended two-dimensional torsional method (E2DT) ([Simón-Carballido et al., 2017](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B16)), implemented in the Q2DTor program ([Ferro-Costas et al., 2018a](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B6)), needs a fine grid of points for the construction of the torsional PES. The procedure also includes the location of all stationary points (i.e., minima, saddle points and maxima in the 2D-PES).

Therefore, the amount of information needed from the PES depends on the method, and it is crucial to devise algorithms that allow an efficient construction of such PES. For example, when the number of torsional degrees of freedom is only 2, so the E2DT method can be applied, geometry scans at a regular number of points along the PES can be carried out. These scans involve partial optimizations in which all degrees of freedom are optimized except the two torsional modes. When the torsional global PES is calculated by systematic mapping, if possible, it is essential to reduce the number of points to be calculated. This reduction depends on molecular geometry aspects as conformational enantiomerism, internal symmetry of the rotors and molecular symmetry. The rules to replicate points of a PES under some symmetry conditions are given in [Ferro-Costas et al. (2018a)](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B6). As the number of torsional degrees of freedom increases, the amount of information needed from the PES should be reduced in order to keep the problem tractable. For those cases, the multi-structural torsional method is a good choice ([Zheng et al., 2012](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B22), [2013](https://www.frontiersin.org/articles/10.3389/fchem.2020.00016/full#B21)), because the model is built assuming that the only information at hand is the set of conformational minima.

A molecular representation, also known as “descriptor” of feature vectors, encodes chemical identity of a molecular entity in terms of its chemical composition and atomic configuration. Only after the chemical identity is converted into a descriptor (an array of numbers), computer can efficiently process a large number of structures. Descriptor is a crucial ingredient for the development of predictive ML models over the input output mapping surface. There are three key invariances of physical systems those are preferably captured by a common molecular descriptor. Formally, these are as follows:

1. Rotational invariance: Representation must be invariant upon a rotation operation.
2. Translational invariance: Representation must be unchanged upon a translation in space.
3. Permutation invariance: Representation must be unaltered due to the change in the particular order of the atoms.

**Databases & Links Research**

The QM-based molecular modeling studies, where ML methods have been employed successfully are reported here, from recent contributions. Those can be considered as complementary, and give us an overall highlight of the applications. These include various QM observables: enthalpies of formation [[30](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0030), [72](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0072), [73](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0073)], atomization energies [[20](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0020)], reorganization energies [[74](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0074)], chemical reactivity [[75](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0075)], polymer properties [[76](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0076)], optimizing transition state surface which divides reactants from products [[77](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0077)], electronic ground state properties [[22](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0022), [31](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0031), [49](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0049)], electronic excitation energies [[78](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0078)], electron transmission coefficients [[79](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0079)], NMR nuclear shifts [[80](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0080)], frontier orbital eigenvalues [[81](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0081)], atomic charges, dipole and quadrupole moments [[82](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0082)], thermochemical properties (besides enthalpies, entropies and heat capacities) [[83](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0083)], infrared spectra [[84](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0084)], polarizabilities [[85](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0085), [86](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0086)], energies of methanol clusters [[87](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0087)], and transition metals [[88](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0088)]. ML has been playing potential roles in constructing approximate QM methods [[89](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0089)], as well as predicting electron densities from DFT [[61](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0061), [90](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0090)], improving the density functionals [[91](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0091)], basis set effects [[92](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0092)], finding approximate density functionals [[93](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0093), [94](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0094)], Møller–Plesset (MP) theory correction [[95](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0095)], prediction of coupled-cluster (CC) singles and doubles amplitudes from MP2-derived properties [[96](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0096)], also predicting MP2 and CC energies from Hartree–Fock orbitals [[97](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0097)-[99](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0099)].

The majority of DL algorithms currently relies on neural networks. DL methodologies have often been used physics-based calculations in order to determine properties of a given molecular system. This involves training the neural network to predict one of the key components/properties of the overall calculation. As a natural choice, neural networks were too used to predict DFT- and wavefunction-based energies [[95](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0095), [100](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0100)-[103](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0103)] like conventional ML models often did. One of the DL applications includes to predict PES. PES fitting caught attention in the early 1980s by the work done by Wagner, Schatz and Bowman. They demonstrated modern computing perspectives of, how to gain maximum information on the surface while using the smaller portion of the surface [[104](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0104), [105](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0105)]. Then, for almost two decades, neural network fits to PES pitched up with ample contribution to the field by various research groups [[26](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0026), [48](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0048), [63](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0063), [68](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0068), [106](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0106)-[114](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0114)].

While kernel-based ML models for fitting the potential were introduced in 1996 [[115](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0115)], over the following years, it started to surge in a variety of applications [[51](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0051), [56](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0056), [57](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0057), [116](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0116)-[119](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0119)]. Few of these approaches were used to construct molecular force fields for molecular dynamics simulations [[118](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0118), [120](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0120), [121](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0121)]. For MD simulations, one calculates energies and forces for a large number of atomic configurations in order to have adequate sampling of the phase space. That can be obtained by electronic structure calculations, often using the DFT or ab initio MD calculations (DFT-based), therefore restricted to a few 100 s of atoms with shorter simulation times, due to their time-consuming nature.

In case of conventional ML utilizing KRR, von Lilienfeld and co-workers used “Coulomb matrix” as descriptor in a so-called QML approach within the CCS, and got very impressive success in predicting quantum chemical properties without solving the Schrödinger equation. A randomized variant of the “Coulomb matrix” representation was later used to predict atomization energies [[122](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0122)], static polarizabilities, frontier orbital eigenvalues (HOMO, LUMO), ionization potential, and electron affinity [[22](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0022)]. The randomized variant was seen to improve the accuracy greatly. DL models have achieved quite satisfactory chemical accuracy in predicting PES using ACSFs and their modified versions. Neural network potential of protonated water clusters were recently studied by Behler and co-workers [[113](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0113)] with the CC accuracy, which is known as gold-standard wavefunction-based QM method. They have employed ACSFs as descriptors. However, SOAP-based ML model unifies predicting CC atomization energy of molecules, stability of molecular conformers, such as, conformers of glucose [[123](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0123)], receptor-ligand binding, ground state structures of silicon surfaces [[70](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0070)]. While calculating the excitation spectra, the CM representation was used with the conventional ML [[78](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0078)] as well as DL [[124](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0124)] models to predict the spectra of the same dataset. DL model gained similar accuracy to the ML approach, and Gghosh et al. stated therein DL will be able to work with the smaller data set also. In a gradient-domain machine learning (GDML) [[118](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0118)] approach to construct accurate molecular force fields using a restricted number of samples from ab initio molecular dynamics trajectories, Chmiela et al. used the reciprocal of the Euclidean distance of an atom-pair, inspired by the CM descriptor in order to define Cartesian geometries that are physically equivalent. Further, in a symmetrized GDML version, the same group used adjacency matrices (similar to CM) as representation to build molecular graphs [[125](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0125)] and constructed ML FF at CCSD(T) level of accuracy [[121](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0121)] for MD simulations. Being simple and efficient, the CM representation contributed in learning various components from QM calculations, also for constructing molecular force fields for MD simulations either using the conventional ML or DL methodologies.

| **References** | **Name** | **Links** |
| --- | --- | --- |
| *Packages for ML-based simulations of materials* | | |
| Khorshidi and Peterson [[166](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0166)] | AMP | <https://amp.readthedocs.io/en/latest> |
| <https://singroup.github.io/dscribe> |
| Bartók and Csányi [[167](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0167)] | GAP | <https://github.com/libAtoms/QUIP> |
| Thompson et al. [[57](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0057)] | SNAP | <https://lammps.sandia.gov/doc/pair_snap.html>, |
| <https://github.com/materialsvirtuallab/snap> |
| Artrith and Urban [[168](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0168)] | AENET | [http://ann.atomistic.net](http://ann.atomistic.net/) |
| Huan et al. [[169](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0169)] | AGNI | <https://lammps.sandia.gov/doc/pair_agni.html> |
| Kolb et al. [[170](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0170)] | PROPhet | <https://github.com/biklooost/PROPhet> |
| Yao et al. [[62](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0062)] | TensorMol | <https://github.com/jparkhill/TensorMol> |
| Smith et al. [[48](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0048)] | ANI | <https://github.com/isayev/ASE_ANI> |
| Smith et al. [[171](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0171)] | COMP6 | <https://github.com/isayev/COMP6> |
| Wang et al. [[172](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0172)] | DeePMD-kit95 | <https://github.com/deepmodeling/deepmd-kit> |
| Mardt et al. [[173](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0173)] | VAMPnet | <https://github.com/markovmodel/deeptime> |
| Xie and Grossman [[174](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0174)] | CGCNN | <https://github.com/txie-93/cgcnn> |
| Jha et al. [[175](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0175)] | ElemNet | <https://github.com/dipendra009/ElemNet> |
| Jha et al. [[176](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0176)] | OQMD-SC | <https://github.com/dipendra009/ElemNet> |
| *Data mining libraries of materials* | | |
| Curtarolo et al. [[177](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0177)] | AFLOWLIB | <http://aflow.org/aflow-ml> |
| Jain et al. [[178](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0178)] | Materials project | [http://materialsproject.org](http://materialsproject.org/) |
| Kirklin et al. [[179](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0179)] | OQMD | [http://oqmd.org](http://oqmd.org/) |
| Bartók et al. [[180](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0180)] | [libAtoms.org](http://libatoms.org/) | <http://www.libatoms.org/Home/DataRepository> |
| Ward et al. [[181](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0181)] | Matminer | <https://github.com/markovmodel/deeptime> |
| Gómez-Bombarelli et al. [[182](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0182)] | Chemical VAE | <https://github.com/aspuru-guzik-group/chemical_vae> |
| Choudhary et al. [[183](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0183)] | JARVIS-DFT | <https://github.com/usnistgov/jarvis> |
| Pun et al. [[112](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0112)] | DScribe, descriptors | <https://github.com/SINGROUP/dscribe>, |
| Draxl and Scheffler [[184](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0184)] | NOMAD | [https://analytics-toolkit.nomad-coe.eu](https://analytics-toolkit.nomad-coe.eu/) |
| Olsthoorn et al. [[185](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0185)] | OMDB | <https://omdb.mathub.io/dataset> |
| Chapman et al. [[186](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0186)] | Khazana | [https://khazana.gatech.edu](https://khazana.gatech.edu/) |
| Kayastha and Ramakrishnan [[187](https://onlinelibrary.wiley.com/doi/full/10.1002/qua.26870#qua26870-bib-0187)] | MolDis | <https://moldis.tifrh.res.in/index.html> |