

# A basic electro-topological descriptor for the prediction of organic molecule geometries by simple machine learning

Carlos Manuel de Armas-Morejón,<sup>\*,†,‡,¶</sup> Ask Hjorth Larsen,<sup>\*,†</sup> Luis A. Montero-Cabrera,<sup>\*,¶</sup> Angel Rubio,<sup>\*,‡</sup> and Joaquim Jornet-Somoza<sup>\*,†,‡</sup>

<sup>†</sup>*Nano-Bio Spectroscopy Group and ETSF Scientific Development Centre, Department of Materials Physics, University of the Basque Country, CFM CSIC-UPV/EHU-MPC and DIPC, Tolosa Hiribidea 72, E-20018 Donostia-San Sebastián*

<sup>‡</sup>*Theory Department, Max Planck Institute for the Structure and Dynamics of Matter and Center for Free-Electron Laser Science, Luruper Chaussee 149, 22761 Hamburg, Germany*

<sup>¶</sup>*Laboratorio de Química Computacional y Teórica, Facultad de Química, Universidad de La Habana, 10400. La Habana, Cuba.*

E-mail: carlosdearmasm@gmail.com; asklarsen@gmail.com; lmc@fq.uh.cu;  
angel.rubio@mpsd.mpg.de; j.jornet.somoza@gmail.com

## Abstract

This paper proposes a machine learning (ML) method to predict stable molecular geometries from their chemical composition. The method is useful for generating molecular conformations which may serve as initial geometries for saving time during expensive structure optimizations by quantum mechanical calculations of large molecules. Conformations are found by predicting the local arrangement around each atom in the molecule after trained from a database of previously optimized small molecules. It works by dividing each molecule in the

database into minimal building blocks of different type. The algorithm is then trained to predict bond lengths and angles for each type of building block using an electro-topological fingerprint as descriptor. A conformation is then generated by joining the predicted blocks. Our model is able to give promising results for optimized molecular geometries from the basic knowledge of the chemical formula and connectivity. The method trends to reproduce interatomic distances within test blocks with RMSD under 0.05 Å.

## 1 Introduction

This work assesses the problem of generating reliable conformers of molecules from proposed chemical compositions. Realistic initial bond lengths and angles are essential for efficient geometry optimizations. They are normally the first step of the usual computational workflow of systematic variations of the atomic coordinates inside a molecule and the calculation of the potential energy and forces of the system in order to find a minimum value for the potential energy, which indicates a theoretically optimized conformation. Moreover, approximate and reliable molecular geometries serve for many modelling and process simulation purposes from docking to Molecular Dynamics.

The function combining all possible variations of the geometry and the potential energy forms a high-dimensional surface and it is the well known *potential energy surface* (**PES**). All possible conformers for a given compound are comprehended as minima in the appropriate PES. Two processes are required in order to obtain the model of the conformer with the lowest potential energy of a given molecule: (1) a procedure to attain its corresponding and plausible PES and (2) a method to navigate it to search for minima. Procedures to find PES's and then computing potential energies can vary in efficiency depending on several factors, and mostly the number of atoms of the compound. Options exist from the easily computed empirical force fields based on classical considerations of bodies in a molecule,<sup>1,2</sup> passing fast quantum mechanical semi-empirical calculations being parameterised for specific scenarios,<sup>3</sup> to more general and reliable but computationally expensive *ab-initio* and DFT calculations.<sup>4</sup> The method to navigate the PES can also be computationally expensive depending on how fast the global minimum can be found.<sup>5</sup> Obviously,

if the initial choice of a guess conformation is near to the final optimized structure, the corresponding PES’s minimum will be faster reached. This is crucial for accurate geometry optimizations of large molecules.

Machine learning (*ML*) has been recently used in a variety of topics inside the field of quantum mechanics.<sup>6–37</sup> For molecular geometry optimizations, progress has been made mostly by using neural networks to parameterise classical force fields.<sup>38,39</sup>

The proposal made in this work is to use ML, and specifically the Kernel Ridge Regression algorithm, for predicting molecular conformations by producing the local arrangement of each atom belonging to a molecule. It can be achieved by using a large and confident database of optimised small molecules<sup>40</sup> as a source for both the training and testing sets. For this purpose, certain molecular structural blocks are characterised and defined by an electro-topological descriptor<sup>18,41</sup> from the structures of the previously optimised molecules. Blocks are then reconstructed by applying ML tools. Using the ETKDG<sup>42–44</sup> (Experimental Torsion Knowledge Distance Geometry) present in RDKit<sup>45</sup> with a new *ab-initio* torsion angle database, we join all predicted blocks to produce the desired molecular structures with reasonable reliability. Results promise fast and confident predictions of molecular geometries and conformations from their formulas taken as structural graphs.

## 2 Learning data

The database used in this work is part of a larger one<sup>46,47</sup> of quantum PES’s minima geometries of small organic molecules containing up to 8 C, O, N and/or F atoms. The optimised geometries of this database are reported to be found using DFT/B3LYP<sup>48</sup> with the 6-31G(2df,p) basis set as a commonly accepted reliable PES. We will refer to this database as *8CONF*. The size of this resulting database subset is 21k molecules.<sup>40</sup>

To facilitate predictions based on this data, we seek a representation which minimizes the amount of redundant information in the learning set, and also could group together similar kinds of data.

First, each molecule is split into *blocks*. A block is the main building part of our model, and it is characterised by: (1) a central atom with more than one bond and (2) the first neighbors

of such central atom. Figure 1 shows a block decomposition for an example molecule from the 8CONF database. Note that each atom can normally be included in multiple blocks: the block centered around itself as well as each of the blocks surrounding its neighboring atoms. Atoms with only one neighbor are not considered to define a block. Blocks therefore have from two to four neighboring atoms in the selected molecular sets where all atoms belong to the first and second rows of the periodic table.

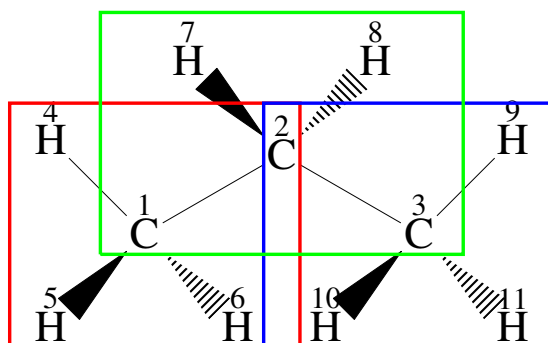


Figure 1: Block decomposition for the molecule  $C_3H_8$ . Each C atom is bonded to multiple atoms and hence defines a block. The molecule can therefore be divided in 3 blocks, two of which (for atoms 1 and 3) belong to the same *block-class*. In each block we have redundant atoms to indicate where blocks join together.

A unique Cartesian representation is not well suited for predictions because coordinate values depend on the chosen reference center. Instead, we represent local coordinates within a block by 1) subtracting the molecular Cartesian coordinates of the central atom position in that block to define it as the local coordinate origin, and 2) computing the matrix  $B$  of scalar products  $a_i \cdot a_j$  between each pair  $(i, j)$  of coordinates of local position vectors corresponding to the non-central atoms in the block. This matrix is the feature we use for training and predictions.

The matrix  $B$  of scalar products is symmetric and at most  $4 \times 4$  in size, and therefore has up to ten unique degrees of freedom.  $B$  contains enough information to rebuild the set of molecular Cartesian coordinates (see Appendix 8.1) for a block except for translations, rotations, and chirality, with which the matrix is invariant.

The non-central atoms in a block have no natural ordering. Hence a way must be chosen to assign an index  $i$  to each of them with a minimum of ambiguity for ML training and testing. To

this end we define an *equivalence relation* for the set of all blocks, i.e., each block belongs to a single, specific equivalence class or *block-class*. The ML algorithm is then independently trained for each block-class.

Two blocks belong to the same class if 1) the species of the central atom in each block is the same, 2) the species of each neighboring atom is the same (some ambiguity is solved by sorting by atomic numbers) and 3) the arrangement of the atoms in space is the same, i.e. either tetrahedral (*TH*), triangular (*TR*) or linear (*L*).

The definition of block-classes can be applied with different levels of restrictions. As a result, the number of different block-classes can vary, as blocks with the same atoms can appear in very different environments.

We denote a block-class by a series of chemical symbols followed by certain indicators of spatial configurations when necessary. The first symbol is that of the central atom, followed by the symbols of the neighboring atoms ordered with higher atomic numbers first. Figure 2 shows the distribution of blocks inside the database.

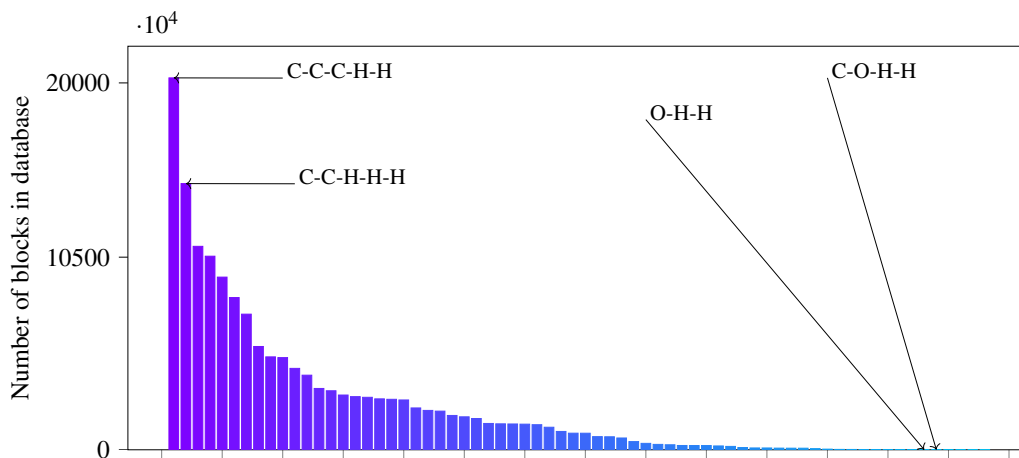


Figure 2: Distribution of block-classes within all molecules in the database. A few common and uncommon block-classes as O-H-H and C-O-H-H are referred.

When choosing the definition of block-classes there is a tradeoff: We can make predictions easier by maximizing the amount of chemical knowledge which defines a block. It can be achieved by dividing the blocks into a large number of classes each of which contains very similar blocks.

However doing so also decreases the size of the learning sets, resulting in block-classes with few or no members. In general block-classes must be defined in such a way to get neither too many nor too few.

### 3 Descriptors

To complete our learning data, a property is required that can correlate with the desired feature. This property is named as a *descriptor* or *fingerprint*. In this case, desired features are the scalar products of a block being directly related to the pursued molecular geometry predictions. *Descriptors* needs to fulfill a number of required characteristics such as: (1) easiness to establish or compute, (2) good representability and (3) low dimensionality.<sup>49</sup> We took the so-called electrotopological<sup>18,41</sup> state index (*e-state*), which is a combination of both electronic and topological characteristics of atoms in a molecule.

If  $Z^v$  is the number of valence electrons of a certain element and  $h$  the number of bonded hydrogen atoms according to the position in a molecule, then  $\delta^v$  could be defined as  $(Z^v - h)$ , the count of valence electrons of a certain atom left for being engaged in the skeleton of a molecule. Similarly,  $\delta$  could be defined as the count of engaged  $\sigma$  electrons ( $\sigma - h$ ). Then,  $I$  can be defined as the *intrinsic state value* of an atom in a molecule, given by  $I = \frac{\delta^v + 1}{\delta}$ . It gets related to the backbone valence of an atom other than Hydrogen.

Then, the *e-state* to relate bonded atoms with their positions in a molecule can be taken as  $S = I + \Delta I$ . It combines the previously defined intrinsic state value  $I$  and a certain  $\Delta I = \sum_{j=1}^N \frac{I_i + I_j}{r_{i,j}^2}$  where  $r_{i,j}$  is a rough expression of distances between atoms  $i$  and  $j$  given by the count of atoms in the shortest path between them, including themselves.  $\Delta I$  relates the intrinsic state value with the bonded environment. For each atom we use a vector of e-state composed by  $[I, \Delta I_0, \Delta I_n]$ , where  $\Delta I_0$  are the contributions of the first neighbours and  $\Delta I_n$  are the contributions of more distant neighbours. The choice of this descriptor fulfill to a fine degree all our desired characteristics. Figure 3 shows the distribution of e-state values of some elements in our database.

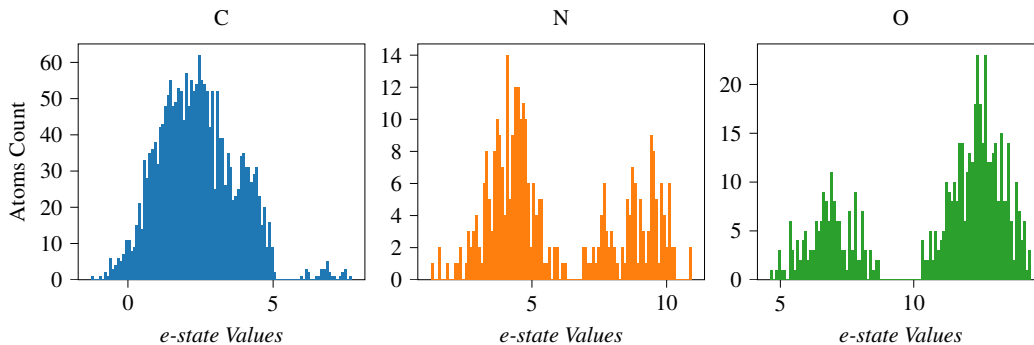


Figure 3: Distribution of e-state values for some atoms species present in our database.

## 4 Learning method

The kernel ridge regression (**KRR**) formulation proposed by *Ramakrishnan, R. et al.*<sup>50</sup> is our selected learning method. It has been tested with success in other *machine learning/Quantum Mechanical* applications.<sup>40,50</sup> It serves to provide a given property of any query molecule as a linear combination of similarity measures (as "distances") between the query's property and those of a finite set of training points. For the *tests* we randomly selected 450 molecules from 8CONF (C from now on). The rest of the molecules were chosen as the training set ( $\mathbb{T}$  from now on).

Our query's property is the scalar product matrix obtained as the result of applying the transformation  $B = XX^T$  to each block, where  $X$ 's are the nonzero position vectors resulting after a translation of the central atom of a block to the origin of coordinates. The KRR method needs training to grasp correlation between the descriptor and the property, and  $\mathbb{T}$  was used for that. Let  $R$  be the scalar product tensor of all blocks inside a block-class in the  $\mathbb{T}$ . In matrix notation the training process can be defined as:

$$R = (K + \lambda I)C \tag{1}$$

where  $C$  represents the coefficients to be computed,  $K_{ij} = K(|d_i - d_j|) = \exp(-\frac{|d_i - d_j|}{\alpha})$  are the kernel terms as calculated by using  $L_1$  norm distances  $|\cdot|$  between the descriptors  $d$  and  $\alpha$  is a regularization parameter to be commented below. From solving eq. 1, we obtain the coefficients  $C$ , used to get the prediction  $P$  by:

$$P = DC \quad (2)$$

where  $D$  is the matrix of distances between descriptors  $d$  of known  $j$  learning points and the new  $i$  points computed using  $K\left(-\frac{|d_{i \in \text{new}} - d_{j \in \text{known}}|}{\alpha}\right)$ .

The KRR method depends on two so-called hyper-parameters:  $\lambda$  to control the regularization factor of the kernel  $K$ , and  $\alpha$  which controls the radius of inclusion/similarity for the kernel function. Both have a tremendous impact on the process behaviour and the best combination for each application needs to be previously found. For this purpose we conduct a grid optimization for each block-class: using a linear distribution of 85 values of  $\alpha$  in the range  $[10^{-3}, 3000]$  and a linear distribution of 25 values of  $\lambda$  in the range  $[10^{-3}, 1]$ . Some blocks-class found their minimum at  $\alpha = 2500$ . This behaviour came from the fact that KRR gives the same weight to all blocks in a certain block-class as it needs all possible information to produce the best results. Most blocks require less data because they have a small deviation in their scalar product values. The values of  $\lambda$  behave as a compensation for how far from the optimum value the  $\alpha$  hyper-parameter is. But it can not correct by itself the deviation in the Root Mean Squared Distance (RMSD) produced by wrong  $\alpha$  values. As soon  $\alpha$  move closer to the optimum value,  $\lambda$  lose his influence as long as the values stays in the range  $(0, 1)$ . However, the values of  $\alpha = 10$  and  $\lambda = 0.9$  resulted from  $\mathbb{T}$  provided an appropriate trend of improvement of the RMSD for all block-classes. Figure 4 shows some results of this optimisation.

A new set of Cartesian coordinates can be predicted with our algorithm as outlined below. Scheme 1 shows the flow diagram. A SMILES<sup>51</sup> representation of the molecule is taken as input. It is then analyzed to extract connectivity, the blocks and their associated classes. The corresponding e-states are then computed. The learning data-set is built by using the block-classes to compute the scalar product and the e-states in order to train the KRR algorithm. The newly parameterized KRR is then used to predict each scalar product. Finally, the blocks are reconstructed again from



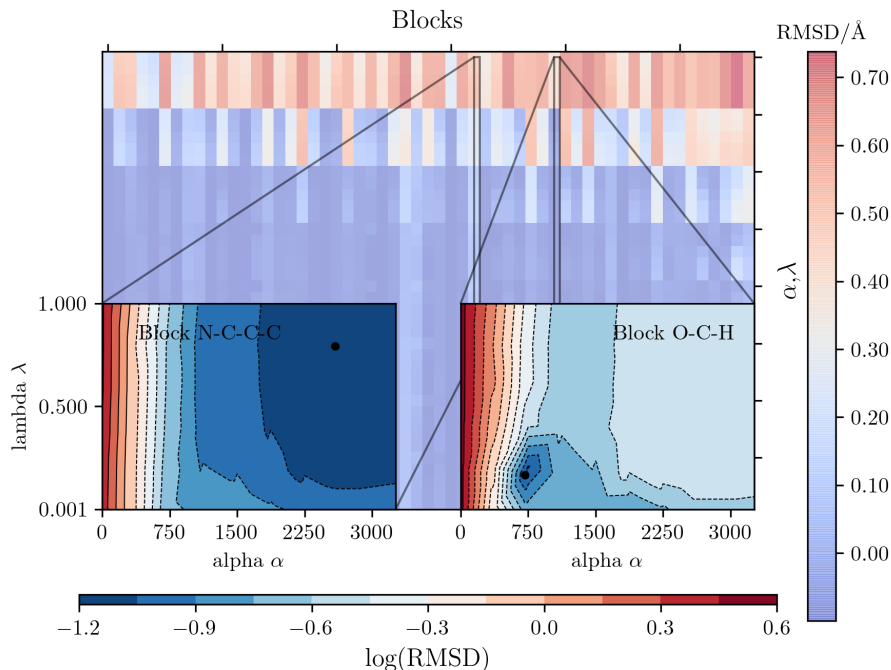
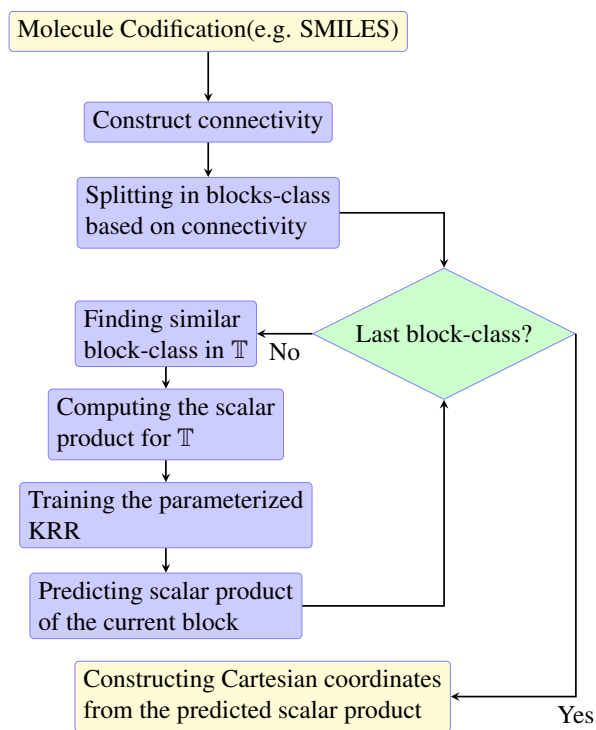


Figure 4: Summary of the KRR’s  $\alpha$  and  $\lambda$  hyper-parameter optimizations. The background image shows the RMSD value for all blocks-class by each combination of the hyper-parameter  $\alpha, \lambda$ . The color scale shows (on the right) in dark blue the found RMSD minimum. In the bottom two examples of hyper-surface formed with the hyper-parameters and  $\log(RMSD)$  for visibility purposes. A black dot (.) marks the minimum.

the scalar product predictions to a Cartesian coordinate output.<sup>52,53</sup> (The formulation used to reconstruct from scalar products to Cartesian coordinates is presented as an Appendix)

## 5 Joining blocks.

The RDKit<sup>45</sup> code was modified for the purpose of joining blocks. Geometry reconstruction by this kit implements the ETKDG<sup>42–44</sup> (Experimental Torsion Knowledge Distance Geometry) algorithm, which depends on a data base of torsion angles extracted from other sources, mostly experimental data. Using the same approach proposed by *Scharfer, C.; Schulz-Gasch, T., et al.*<sup>42</sup> we extracted a new torsion angle data base from the ground 8CONF. This newly *modified ETKDG* (Theoretical-Torsion Knowledge Distance Geometry and Machine Learning, TTKDG-ML from now on) was used to produce reasonable conformer geometries with the input of predicted ML blocks.



Scheme 1: Flow diagram for the prediction of a new coming molecule representation, from which the algorithm can extract connectivity (e.g. codified by SMILES), selects the necessary data from the training set  $\mathbb{T}$  to train the KRR method and produces blocks Cartesian coordinate.

## 6 Results and discussions

Two experiments were performed with the  $\mathbb{C}$  testing set. The first consists on block's prediction. Such control group was divided in block-classes to predict all scalar products followed by the reconstruction of their coordinates. Values of terms in the scalar product matrices  $B$  are obtained individually as each one gives a specific information. The diagonal describes bond lengths and off-diagonal terms provide information on angles. The first attempt to predict vectors, such as the upper triangular part of a matrix, resulted problematic, mostly because KRR does not particularize each unique component of the matrix and distribute the errors among all values.

Results of this computing experiment appear in Figure 5. RMSD histograms show how good the prediction/reconstruction of blocks really is, with most block-classes showing errors under 0.05 Å. However, methods must be also evaluated by their worst cases as some block-classes rose them to around 0.3 Å. Among the highest RMSD's values, two different kinds of problematic blocks are

found: those belonging to a group on which there are not enough data to properly train the KRR algorithm (e.g. *N-N-N-H* with only 7 occurrences) and those belonging to molecules where the blocks are located inside rings that tend to distort the bond lengths and angles (e.g. *C-C-C-C-C* which is a very abundant block-class with multiple combinations).

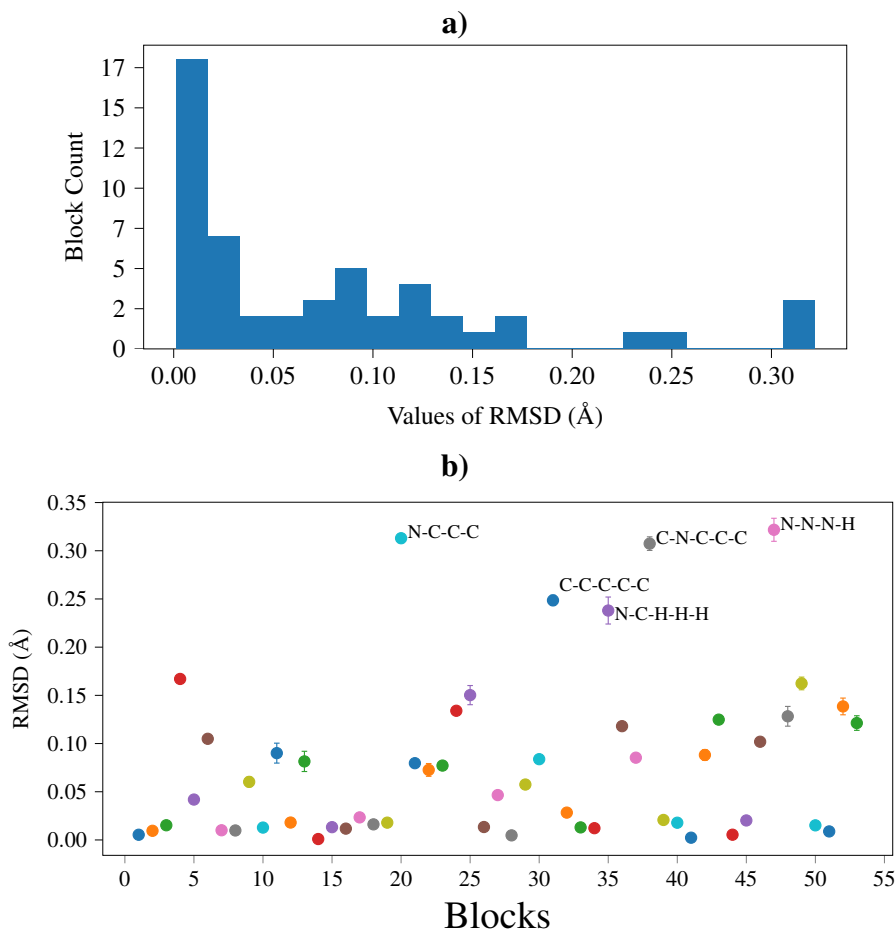


Figure 5: a) Histograms with the mean RMSD distribution for the block-classes present in the 450 molecules of  $\mathbb{C}$ . b) Details of block-classes where those with RMSD errors above 0.3 Å are explicit. Each color identify a block-class present in the test group  $\mathbb{C}$ .

The second experiment with the  $\mathbb{C}$  testing set involved the complete rebuilding of the molecules. After predicting the blocks and reconstructing the coordinates, TTKDG-ML procedure was fed with the distance matrices of each block. Some restrictions were put in place in the form of an arbitrary high number (the decimal value 10 was used) to be interpreted as a weight by the algorithm. So, the better conformers were left within those non modifying the predicted distances by

the KRR method. The best is selected by the minimum RMSD value.

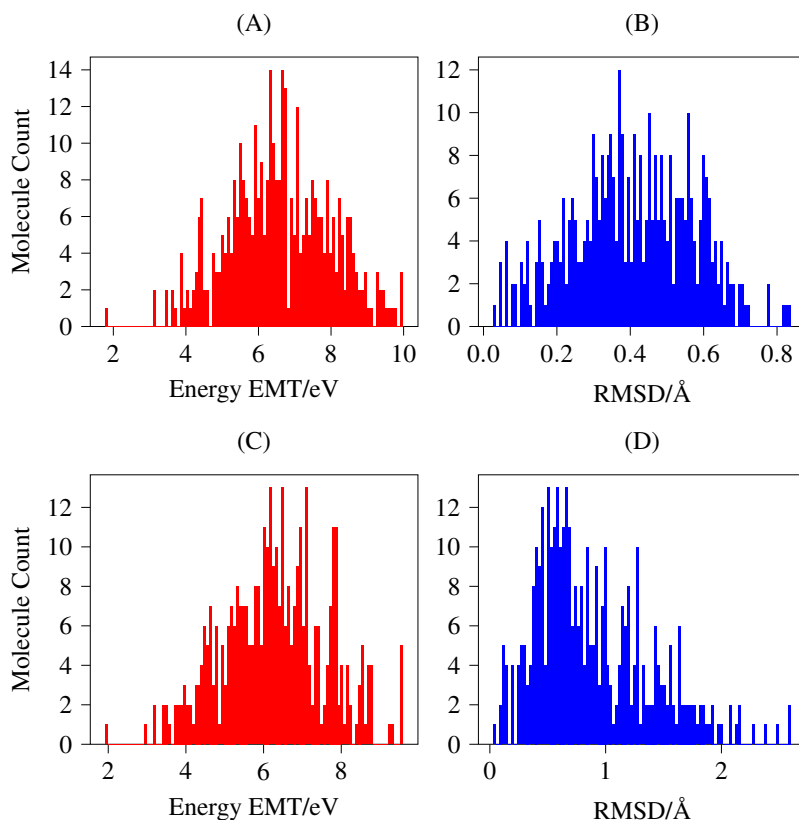


Figure 6: The first row (from the top) shows two histograms with the EMT (plot A) energies values and RMSD’s (plot B) obtained after the complete reconstruction of molecules contained in the test-set C. The similarities of the histogram shapes are noticeable as the EMT express bigger energy values when molecular assembling is more complex. As points of reference, the second row shows the EMT energy of the original molecules as present in the test-set C (plot C) to be compared with those in (plot A). The RMSD’s (plot D) between original molecules present in test-set C and their conformers obtained by the ETKDG in RDkit show very encouraging results.

Histograms comparing both RMSD results and potential energy calculations using a simple *Effective Medium Theory* (EMT)<sup>54,55</sup> for the 450 molecules of the test-set are shown in Figure 6. This method consists of an expression derived for obtaining a value related with the total energy of a system of interacting atoms. It is based on an ansatz for the total electron density of the system as a superposition of atom densities taken from calculations for the atoms embedded in a homogeneous electron gas. We use these energy values here to illustrate molecular complexity compared with total RMSD predictions. As EMT was developed only for solids and not for molecules, the resulting values of energy must be only taken as just a reference. It can be realized that there are

eight molecules from the test-set on which the TTKDG does not perform well. It must be originated in the fact that they contain one or more blocks with poor prediction results (See Figure 5) such as *N-C-C-C.Tr* and *C-C-C-C-C.Td*. The results show that the proposed TTKDG-ML performs better than the original ETKDG implemented in RDKit (See Figure 6 *B* and *D*).

## 7 Conclusions

To predict accurate geometries of molecules is a complex problem able to be multiple approached. A possible solution is described here by using a machine learning tool based on a Kernel Ridge Regression routine. The described procedure departing from bonded atom blocks results in a very flexible and easy-to-expand way to describe the canonical geometry of an atom in a molecule and its environment. The descriptor used showed a very good correlation with bond distances and angles when the coordinates were transformed into scalar products. RMSD values obtained from the experiments performed on blocks support our decision on this kind of descriptors and validate the block structure. To solve most of the *problematic block* predictions will be required to add more related training data because the involved atoms in such blocks are rare in common molecules and therefore scarce in the training set. For those atoms belonging to rings, a better data treatment could be a solution. But this needs to be exercised with caution, because too much pruning of the data base could lead to *over-fitting* the KRR method results.

For approaching the final purpose of getting a complete machine learning method, the selected descriptor was tested for torsion angles. It completely failed. The symmetries involved when joining two blocks were very difficult to grasp by the e-state with the only use of the KRR algorithm. Even the selection of the correct torsion angle for the learning process proved to be non-trivial. Several difficulties were faced in this area especially with linear carbon chains where the definition rules resulted to be ambiguous. Nevertheless, the TTKDG-ML from an *ab-initio* database proved to provide very consistent conformers, i.e. those related to extreme torsion angles. The used weights helped to maintain the obtained results by the KRR method.

All optimized geometries used can be downloaded from <http://www.quantum-machine.org/datasets/>,

## Acknowledgement

This work was supported by the European Research Council (ERC-2015-AdG694097), the Cluster of Excellence 'CUI: Advanced Imaging of Matter' of the Deutsche Forschungsgemeinschaft (DFG) - EXC 2056 - project ID 390715994, Grupos Consolidados (IT1249-19) and the SFB925 "Light induced dynamics and control of correlated quantum systems". We kindly recognize the partial support of the project ID PN223LH010-002 "Inteligencia Artificial Aplicada. Espectroscopía y Bioactividad" of the Cuban Ministry of Science, Technology and Environment as well as the overall support given to LAMC by the Universidad de La Habana and the Donostia International Physics Center.

## 8 Appendix

### 8.1 Scalar product formulations and reconstruction

For the sake of argument clarity, tetravalent blocks are taken here as an example for the scalar product formulation. Each block has five atoms, or a total of 15 coordinates. Rotations and translations account for six degrees of freedom (DOF's), leaving nine of them to consider.

Translations were automatically eliminated by measuring neighbouring atom coordinates as displacements from the central atom. The elimination of rotational dependencies were achieved by forming the matrix of all scalar products between the nonzero position vectors  $X_{ac}$ , where  $a = 1 \dots 4$  specifies direct neighbouring atoms and  $c$  is one of x, y, z:

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T = \begin{bmatrix} a^2 & a \cdot b & a \cdot c & a \cdot d \\ b \cdot a & b^2 & b \cdot c & b \cdot d \\ c \cdot a & c \cdot b & c^2 & c \cdot d \\ d \cdot a & d \cdot b & d \cdot c & d^2 \end{bmatrix} \quad (3)$$

The bond lengths can be obtained directly from the four diagonal elements. The information

about the angle is provided by the six distinct scalar products. The system is over determined by one DOF. However the matrix has only rank 3 because it was formed from a  $4 \times 3$  matrix. Therefore one eigenvalue is zero, leaving a three-dimensional eigenspace which spans the nine DOF's remaining as well determined.

The matrix can be rewritten by using its eigendecomposition as:

$$\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T = (\mathbf{Q}\mathbf{\Lambda}^{1/2})(\mathbf{\Lambda}^{1/2})^T\mathbf{Q}^T = (\mathbf{Q}\mathbf{\Lambda}^{1/2})(\mathbf{Q}\mathbf{\Lambda}^{1/2})^T = \bar{\mathbf{X}}'(\bar{\mathbf{X}}')^T \quad (4)$$

This defines a matrix  $\bar{\mathbf{X}}' = \mathbf{Q}\mathbf{\Lambda}^{1/2}$  which produces the same scalar products. The matrix  $\bar{\mathbf{X}}'$  will contain one row of zeros, corresponding to the eigenvalue 0.

This row/column is discarded, and the remaining matrix  $\mathbf{R}'$  will be an eligible set of reconstructed positions. In weeding out numerical garbage, the lowest eigenvalue is always discarded even if it is not exactly zero. The remaining eigenvalues must be positive since the scalar product is positive definite.

## References

- (1) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 490–519.
- (2) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. SOC* **1992**, *114*, 10024–10035.
- (3) Thiel, W. Semiempirical quantum–chemical methods. *WIREs Computational Molecular Science* **2014**, *4*, 145–157.
- (4) Vuckovic, S.; Burke, K. Quantifying and understanding errors in molecular geometries. *arXiv:2007.15076* **2020**,

- (5) Snyman, J. A.; Wilke, D. N. *Practical Mathematical Optimization Basic Optimization Theory and Gradient-Based Algorithms*, 2nd ed.; Springer, 2018.
- (6) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. SISSO: a compressed-sensing method for systematically identifying efficient physical models of materials properties. *Phys. Rev. Materials* **2018**, *2*, 083802.
- (7) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* **2016**, *18*, 13754–13769.
- (8) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.
- (9) Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *International Journal of Quantum Chemistry* **2015**, 1051–1057.
- (10) Marí, R. M.; Aguirre, N. F.; Daza, E. E. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. *Journal of Chemical Information and Modeling* **2008**, *48*, 109–118.
- (11) Guido, C. A.; Cortona, P.; Mennucci, B.; Adamo, C. On the Metric of Charge Transfer Molecular Excitations: A Simple Chemical Descriptor. *Journal of Chemical Theory and Computation* **2013**, *9*, 3118–3126, PMID: 26583991.
- (12) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *The Journal of Chemical Physics* **2013**, *139*, 184118.
- (13) Dong, J.; DS, C.; HY, M.; Liu, S.; BC, D.; YH, Y.; NN, W.; AP, L.; WB, Z.; AF, C. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J Cheminform* **2015**, *7*, 60.



- (14) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *The Journal of Chemical Physics* **2016**, *145*, 161102.
- (15) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *The Journal of Chemical Physics* **2018**, *148*, 241718.
- (16) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *The Journal of Physical Chemistry Letters* **2018**, *9*, 4495–4501.
- (17) Himanen, L.; Jäger, M. O. J.; Eiaki V. Morooka.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DDescribe: Library of Descriptors for Machine Learning in Materials Science. *Computer Physics Communications* **2020**, *247*, 106949.
- (18) Hall, L. H.; Mohny, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *Journal of Chemical Information and Computer Sciences* **1991**, *31*, 76–82.
- (19) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Physical Review Letters* **2018**, *120*.
- (20) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3309–3313.
- (21) Huang, B.; Symonds, N. O.; von Lilienfeld, O. A. The fundamentals of quantum machine learning. *arXiv:1807.04259* **2018**,
- (22) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Computational Chemistry* **2018**, *57*, 4164–4169.

- (23) Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multi-component systems: Applications to zinc oxide. *Physical Review B* **2011**, *83*, 153101.
- (24) Artrith, N.; Urban, A.; Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Physical Review B* **2017**, *96*, 014114.
- (25) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Scientific Advance* **2017**, *3*.
- (26) Behle, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *JOURNAL OF CHEMICAL PHYSICS* **2011**, *134*, 074106.
- (27) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *PHYSICAL REVIEW LETTERS* **2007**, *98*, 146401.
- (28) Bianucci, A. M.; Alessio Micheli, A. S.; Starita, A. Machine Learning Force Fields: Construction, Validation, and Outlook. *Applied Intelligence* **2000**, *12*, 117–146.
- (29) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *Computational Materials* **2019**, *22*.
- (30) Hughes, Z. E.; Thacker, J. C. R.; Wilson, A. L.; Popelier, P. L. A. Description of Potential Energy Surfaces of Molecules Using FFLUX Machine Learning Models. *Journal of Chemical Theory and Computation* **2019**, *15*, 116–126.
- (31) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *arXiv:1705.05907v1* **2017**,
- (32) Huang, B.; Symonds, N. O.; von Lilienfeld, O. A. *Quantum Machine Learning in Chemistry and Materials*; Springer International Publishing, 2018.

- (33) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. The Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *Journal of Physical Chemistry Letters* **2018**,
- (34) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.
- (35) Huo, H.; von Matthias Rupp, Unified Representation of Molecules and Crystals for Machine Learning. *arXiv:1704.06439v3* **2018**,
- (36) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *JOURNAL OF CHEMICAL PHYSICS* **2018**, *148*, 241730.
- (37) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model* **2018**, *58*, 27–35.
- (38) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* **2019**, *15*, 3678–3693.
- (39) Mansimov, E.; Kang, S.; Mahmood, O.; Cho, K. Molecular Geometry Prediction using a Deep Generative Graph Neural Network. *Sci Rep* **2019**, *9*, 20381.
- (40) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. *The Journal of Chemical Physics* **2015**, *143*, 084111.
- (41) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences* **1995**, *35*, 1039–1045.

- (42) Scharfer, C.; Schulz-Gasch, T. Torsion Angle Preferences in Druglike Chemical Space: A Comprehensive Guide. *Journal of Medicinal Chemistry* **2013**, *56*, 2016–2028.
- (43) Landrum, G. RDKit: Open-source cheminformatics. 2006; <http://www.rdkit.org>.
- (44) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Chem. Inf. Model.* **2015**, *55*, 2562–2574.
- (45) Landrum, G. RDKit: Open-Source Cheminformatics Software. **2016**,
- (46) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.
- (47) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.
- (48) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *ACS Publications* **1994**, *98*, 11623–11627.
- (49) Ghiringhelli, L.; Vybiral, J.; V. Levchenko, S.; Draxl, C.; Scheffler, M. Big Data of Materials Science - Critical Role of the Descriptor. *Physical Review Letters* **2015**, *114*, 105503.
- (50) Raghunathan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *CHIMIA International Journal for Chemistry* **2015**, *69*, 182–186.
- (51) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- (52) Crippen, G.; Havel, T. *Distance Geometry and Molecular Conformation*; 1988.
- (53) Blaney, J. M.; Dixon, J. S. *Distance Geometry in Molecular Modeling*; 1994.

- (54) Jacobsen, K.; Stoltze, P.; Norskov, J. A semi-empirical effective medium theory for metals and alloys. *ELSEVIER Surface Science* **1996**, 366, 394–402.
- (55) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, 29, 273002.