Invariance to atom permutations - Bag of Bonds Histogram

Viviana Petrescu

May 2015

1 Abstract

The high computational cost of quantum chemistry calculations have prompted the use of less expensive machine learning methods for predicting molecular properties in chemical compound space. Finding good feature representations for molecules is hard, in part because of the graph-like structure geometry of the molecules that need to be represented as high dimensional vectors. The desired properties of a descriptor are invariance to rotation and translation of the molecule and invariance with respect to the permutation of atom indexes. In this work we present a new histogram based descriptor which tries to overcome the above problems. While there are other descriptors which overcome the translation and rotation invariance using Coloumb repulsion information, the main property of the proposed descriptor is its invariance to atom indexing. Moreover, its dimensionality is independent of the size of the molecules and it varies with the number of *distinct* atom types in the dataset. We evaluate its predictive performance on two datasets (with approx 7k molecules each) which gives close to state of the art predictions of atomization energy using neural networks.

2 Introduction

The discovery of new molecular materials in chemistry has the potential of solving many of the problems we face today. Having a system which predicts both accurately and at a small computational cost the properties of new materials is highly desirable and has applications ranging from novel drugs discovery, water purification to efficient materials for high energy transmission and storage[1].

Any molecular property can be derived numerically by solving Schröedinger's equation, a setup which is computationally feasible only for small systems. Many approximation algorithms exists which have polynomial complexity in the number of atoms. However, they can also be prohibitive in practice, since predicting one property can take hours or even days for certain system sizes. If instead, the properties of molecules can be estimated using trained machine learning models, the prediction for a new property of a new unseen molecule can take a couple of milliseconds. To date, kernel ridge regression, Gaussian processes and neural networks have been successfully applied to predicting properties of molecules such as atomization energy, averaged molecular polarizability, HOMO and LUMO eigenvalues or ionization potentials [2].

Machine Learning models are as powerful as the feature descriptors used are. There is a long history of molecular descriptors [8] that aim at encoding the information in molecules in a discriminative manner. While some of them require extensive domain knowledge, recent approaches [6] use only the 3D position of atoms and their nuclear charges for describing a molecule.

In the following section we describe two popular descriptors based on Coloumb interactions and we introduce our novel descriptor. In section 3 we describe the datasets on which we evaluate the performance of our model and discuss our experimental results. We summarize the contributions of this work in the last section.

3 Molecular descriptors

To better describe the datasets and the properties of the molecular descriptors, we introduce the notations below:

- N maximum number of atoms in a molecule (present in our dataset)
- N_i maximum number of atoms of type i that constitute a molecule, where $i \in \{H, O, C, N, S\}$
- N_A number of different atom types in the dataset
- D_{max} maximum distance between two pairs of atoms in a molecule
- q quantization level, a number in $(0, 1]$ which defines the coarseness of the histogram of distances

The desired properties of a molecular descriptor are

- invariance to atom permutation
- invariance to translation and rotation

3.1 Coloumb Matrix

The Coloumb matrix[6] M^{NxN} is a descriptor based on Coloumb interaction terms between pairs of atoms. These are the terms that appear in equation.//TODO

$$M(i,j) = \begin{cases} 0.5 * Z_i^{2.4} & if i = j \\ Z_i * Z_j / |R_i - R_j| & otherwise \end{cases}$$

The size of the Coloumb descriptor is given by $\frac{N*(N-1)}{2}$ if we take into account the symmetry of the matrix.

While the Coloumb matrix solves the translation and rotation invariance, the permutation of atom indexes is solved using variants of the descriptor called Sorted Coloumb Matrix or Randomly Sorted Coloumb Matrices. Sorted Coloumb Matrix uses the ordering of the atoms obtained by sorting the rows according to their norm. Unlike the simple Coloumb matrix, any molecule has a unique Sorted Coloumb matrix representation. In practice, the norm of the rows of a matrix are very close to each other and therefore susceptible to small noise which can lead to a different ordering of the atoms for a molecule. Randomly Sorted Coloumb matrices have been introduced to cope with this issue by adding a small Gaussian noise to every row in the Coloumb Matrix and sort according to the new noisy norm value.

3.2 Bag Of Bonds

The descriptor which gives state of the art results in predicting atomization energy is called Bag of Bonds (BoB), whose name was inspired from natural language processing bag of words. In BoB, every bag contains all the pairwise interactions between two types of atoms (the types can be identical). Invariance to indexing of the atoms is obtained by sorting the values inside each bag according to their magnitude. One possible cause for the robustness of the descriptor is the fact that one bag is responsible for certain types of bonds only. Its dimensionality is given by $\sum_i \frac{N_i*(N_i-1)}{2} + \sum_{i,j,i!=j} N_i * N_j$. The first term counts the number of bonds between atoms of the same type and the second term counts the number of bonds between different atom types.

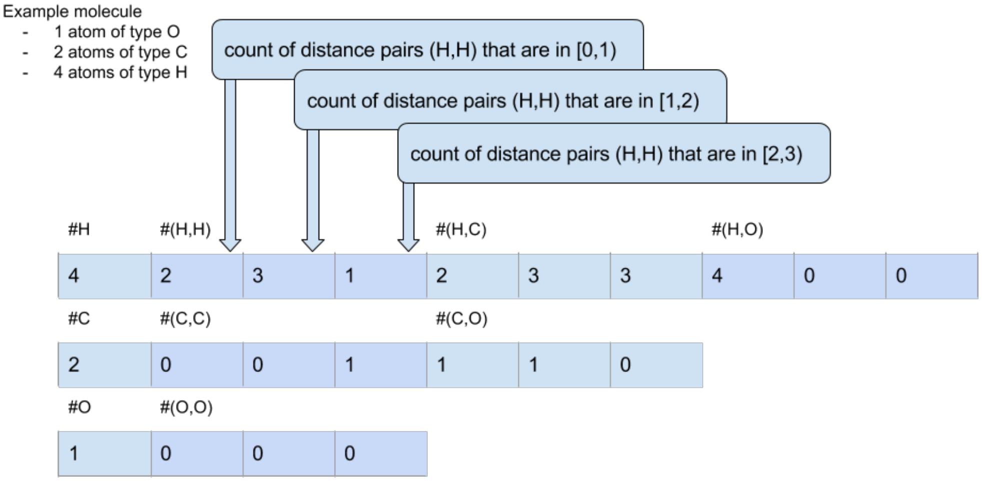
3.3 Bag of Bonds Histogram

We propose a new descriptor Bag of Bonds Histogram(BoBH), whose size is not dependent on the number of atoms in the molecules. Similarly with BoB, every bag encodes information about certain types of bonds. Unlike BoB, where the elements in every bag are sorted according to their magnitude, in BoBH case every bag is a histogram of quantized distances. The size of the bag histogram is given by the maximum distance between two pairs of atoms (pertaining to that bag type). The quantization level of the distances varies between bag types and we experimentally found that a quantization level of 0.2 or 0.25 perform well in practice, but this will vary for different datasets.

Thus, for a given molecule, the BoBH descriptor will encode

- for every different atom type present in the dataset, count how many times it appears in the molecule
- for every pair of atoms of certain type, count how many times the distance between them is in a given quantization interval

An example is given in Fig 1 for a molecule C_4H_2O with a quantization level of $q = 1$. In this example, first the bags containing H were concatenated, then the ones containing C and in the end the ones containing O, resulting in a descriptor of size 21. In general, the size of the descriptor is given by $NA + \sum_i \frac{N_A*(N_A+1)}{2} * \frac{D_{max}}{q}$



Example molecule
- 1 atom of type O
- 2 atoms of type C
- 4 atoms of type H
count of distance pairs (H,H) that are in [0,1)
count of distance pairs (H,H) that are in [1,2)
count of distance pairs (H,H) that are in [2,3)

#H	#(H,H)			#(H,C)			#(H,O)		
4	2	3	1	2	3	3	4	0	0

#C	#(C,C)			#(C,O)		
2	0	0	1	1	1	0

#O	#(O,O)		
1	0	0	0

Figure 1: Bag of Quantised Bonds. Sample computation of Bag of Bonds Histogram for C_2H_4O for a dataset which contains only atoms of type O, C and H. If the maximum distance between any two atoms in our dataset is 3, the size of the BoBH descriptor with quantization level 1 is 21, as shown above. The first entry encodes the number of H atoms, 4. The following 3 elements encode the number of (H,H) distances that fall in the interval [0,1), [1,2) and [2,3). //TODO -¿ add cool picture showin an actual histogram

To sum up, all descriptors achieve translation and rotation invariance through the use of distances between atoms instead of the actual 3D positions. Sorted Coloumb and Randomly Sorted Coloumb achieve invariance to index permutation by sorting the rows of the matrix according to their norm. Bag of Bonds sorts the bonds inside each bag and Bag of Bonds Histogram quantizes the distances into bins to bypass the sorting step present in other descriptors.

4 Experimental Results

4.1 Datasets description

We tested the performance of BoBH on two datasets for predicting atomization energy. The first curated dataset GDB-7 consists of 7165 molecules containing at most 23 atoms per molecules and at most 5 types of atoms (H,C,O,S,N). Five stratified folds for cross-validation were already given to us and are the same on which other people report their result. The other dataset consists of 7211 molecules containing at most 29 atoms per molecules and 6 types of atoms (H,C,O,S,N, Cl). Here we used 5000 training samples drawn randomly from the dataset and remaining 2211 for testing.

4.2 State of the art results

For the first dataset, [4] used a 2-layer neural networks with Random Coloumb matrices for predicting atomization energy with a MAE of 3.1 kcal/mol. It was noticed that both augumententing the dataset and binarizing the input was beneficial for the learning algorithm. Best performance of 1.5 kcal/mol was later obtained using BoB[3] features and Laplacian kernel regression.

Although not a purely machine learning approach, in [5] they combine chemical approximation methods with machine learning models to obtain results at the density functional theory level of accuracy.

For the second dataset,[4] trained a 3 layer NN with Random Coloumb matrices as input and successfully predicted simultaneously 14 properties of the molecules (in particular for atomization energy they obtained MAE of 3.68 kcal/mol).

4.3 BoBH Results

The data was normalised to have values between 0 and 1. Unlike standardization,we believe this helps more the network to learn, since it keeps the values in the same scale as the initialization of the weights and biases. We experimented with various quantization level, between [1.0.5,0.25,0.20, 0.125]. We noticed an improvement in performance with decreasing quantization level up to $q = 0.20$. We believe that further decreasing the quantization level did not improve the results because the size of the descriptor was quite large (almost 3000) and our dataset was relatively small.

Finding the best arhitecture for a dataset can be quite challenging since we need to choose the number of neurons for every hidden layer, the type of activation and the leanring rate. We used the Whetlab tool for finding best network setup TODO cite whetlab here. Unlike grid search, the tool uses gaussian processes to for finding optimal parameters of the network. The best setup was a fully connected NN, with 250 and 80 neurons for the first and second layers and with ReLU activation function. ReLU activation function was shown to work well on sparse descriptors seemed to perform better, possibly also due to its inherently nature of generating sparse representations and for the inherently sparse nature of our feature descriptor. TODO add citation to parse Dense Sparse ReLU They were trained with Whetlab [7]

This is the best neural network model and it comes at a fraction of cost in time, without even augumentening the dataset like in Random Sorted Coloumb with NN.

5 Future work

Learning invariant features, trying RBF networks or using ConvNet on 2D matrix. Compare BoBH with Laplacian kernels.

6 Conclusions

We proposed a new descriptor which solves the invariability to atom permutations problem. It gives best performance on both datasets using a neural network model and second after a BoB using Laplacian kernels. The downside of using kernel regression is the fact that it scales cubically with the number of samples in the training data. The comparison of its performance on the large dataset will be investigated.

We believe that besides vision and natural language processing tasks, chemoinformatics should start gaining more attention in the machine learning community, given the potentials benefits for our society.

References

[1] Johannes Hachmann, Roberto Olivares-Amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sánchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik. The harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, 2011.

[2] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.

[3] O.A. von Lilienfeld K.-R. Müller K. Hansen, F. Biegler and A. Tkatchenko. *Interaction Potentials in Molecules and Non-Local Information in Chemical Space.* Phys. Rev. Lett. (March 10, 2014)., 2014.

[4] Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V Lilienfeld, and

Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In *Advances in Neural Information Processing Systems*, pages 440–448, 2012.

[5] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The delta-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015.

[6] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[7] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

[8] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*, volume 11. Wiley-VCH, 2000.