Predicting Scalar Couplings among Molecules with NMR using Machine Learning Algorithms

John Q. Li
York University
Toronto, Canada
johnwill4g@gmail.com

ABSTRACT

Using Nuclear Magnetic Resonance (NMR) to gain insight into a molecule's structure and dynamics depends on the ability to accurately predict so-called "scalar couplings". These are effectively the magnetic interactions between a pair of atoms. The strength of this magnetic interaction depends on intervening electrons and chemical bonds that make up a molecule's three-dimensional structure. Using state-of-the-art methods from quantum mechanics, it is possible to accurately calculate scalar coupling constants with a 3D molecular structure as input. However, these quantum mechanics calculations are extremely expensive. Sometimes these calculations take days or weeks to complete and therefore have limited applicability in day-to-day workflows. A fast and reliable method to predict these interactions will allow medicinal chemists to gain structural insights in a faster and cheaper way, enabling scientists to understand how the 3D chemical structure of a molecule affects its properties and behavior. Ultimately, such tools will enable researchers to make progress in a range of important problems, like designing molecules to carry out specific cellular tasks or designing better drug molecules to fight disease[6].

ACM Reference Format:
Muhammad Kamran and John Q. Li. 2018. Predicting Scalar Couplings among Molecules with NMR using Machine Learning Algorithms. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

ChEMBL is an open source database which comprises of binding and functional information of a huge number of drug-like bio active compounds .[5] The dataset includes train.csv, test.csv, structure.csv and some other data files, the difference between train and test file is the scalar coupling constant is given as the supervised value in train.csv file. Supervised Learning (Discrete Variable Prediction) would be selected for our later predictive analysis. In 1 we can see the five first rows of structures.csv file.The first column (molecule_name) is the name of the molecule, the second column (atom_index) is the index of the atom, the third column contains the atomic element (H for hydrogen, C for carbon etc.) and the remaining columns contain the X, Y and Z Cartesian coordinates (a standard format for chemists and molecular visualization programs).

	molecule_name	atom_index	atom	x	y	z
0	dsgdb9nsd_000001	0	C	-0.0126981359	1.0858041580	0.0080009958
1	dsgdb9nsd_000001	1	H	0.0021504160	-0.0060313176	0.0019761204
2	dsgdb9nsd_000001	2	H	1.0117308430	1.4637511620	0.0002765748
3	dsgdb9nsd_000001	3	H	-0.5408150690	1.4475266140	-0.8766437152
4	dsgdb9nsd_000001	4	H	-0.5238136345	1.4379326440	0.9063972942

Figure 1: structures

In 2 we can see the five first rows of train.csv file. This is the the training set, where the first column (molecule_name) is the name of the molecule where the coupling constant originates (the corresponding XYZ file is located at ./structures/.xyz), the second (atom_index_0) and third column (atom_index_1) is the atom indices of the atom-pair creating the coupling and the fourth column (scalar_coupling_constant) is the scalar coupling constant that we want to be able to predict.

Since our purpose is to identify the coupling type by predicting coupling constant and the scalar-coupling-constant as a target value has been given There are 4659076 rows and 85012 distinct molecules in train data. There are 2505190 rows and 45777 distinct molecules in test data. Test set is 2 times smaller than train set. We have 5 unique atoms and 8 coupling types. After stating all this, we are likely to assume the train.csv has sufficient data record for us to apply supervised ML techniques for us to predict the bond type in test data. The dataset we obtained is well organized, which saves our effort to deal with missing data and encoding data, but the atom's Cartesian coordinates information are predictor features which are stored separately with the train data and test data. Combining the data and identifying the predictor and target features is required. Beyond the data combination, we could also create new features by using data normalization and standardization. The coordinates of the molecules can be

	id	molecule_name	atom_index_0	atom_index_1	type	scalar_coupling_constant
0	0	dsgdb9nsd_000001	1	0	1JHC	84.807599999999994
1	1	dsgdb9nsd_000001	1	2	2JHH	-11.257000000000000
2	2	dsgdb9nsd_000001	1	3	2JHH	-11.254799999999999
3	3	dsgdb9nsd_000001	1	4	2JHH	-11.254300000000001
4	4	dsgdb9nsd_000001	2	0	1JHC	84.807400000000001

Figure 2: train data

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

re engineered into new distance features such as (coordinates of $\text{atom}_1 - coordinatesofatom_0$), $(atomx_1 - atomx_0) * *2$

for the later predictive analysis processing. From the dataset, each atom pair scalar couplings structure is uniquely identified. The molecule's couple structures are given in 3 Dimensional axis as x,y, and z. We will use the supervised data with Scalar_coupling_constant to differentiate each structure and will train the Machine learning models for the bond constant based on the couple atoms. So we can later predict the bond based on the constant value. Through this process, the chemical structure can be detected and understood.

With the given coupling constant and structures we will come up with a network graph to describe the relationship among the listed atoms.

We will run machine learning models with the training data and predict the scalar_coupling_constant value for the test result and calculate the Mean Absolute Error value from prediction models to measure solution accuracy.

2 RELATED WORK

Drug exploratory analysis demands to integrate the perception of pharmacology network while keeping in mind the difficulties of drug-target interactions.[3] DTIs and PPIs can be modeled as a complete network if we consider nodes to be consisting of compounds and targets while edges will be connected as relationships. So an edge connecting a target with a compound can be a candidate to show as experimentally calculated bioactivity value while the edge connecting two targets can be thought to be as a regulatory link between the targets. Also, prediction of link can be used for solving problems in the homogeneous or heterogeneous multiplex network. [7] Various measures of drug-to-drug-like similarity and multiple aspects, such as the emergence of collections and chemical structures can be defined. Different similarity calculations may give additional information for medications. If different measures of similarities are brought together, the effect can be magnified precisely. The approach for predicting DTI as measures of similarity will directly support various measures of similarity as inputs, as well as information on global structures. A more important manual that requires information on the pair of other drugs, at the edge of a dataset that will be used for drilling on the model, linear relationships between resources, which result in a lower inferior performance. Or, Multi-modal Deep Auto-Encoder (MDA) can combine multiple semantics and dynamically learn high level features. The low-dimensional resources are extracted from the Deep Neural Network (DNN) to predict DTI.[4]

Several researchers have used chemogenomics in combination with modern deep learning algorithms. Basic architecture consists of different blocks including abstract learner for molecules in relation with their structure, the abstract amino acids sequence learner, a neural network and a multiplex perceptrons layers to predict the molecules pairs interaction. [2] BindingDB is a database for calculated affinity assessments. for bindings of minuscule molecules to proteins. In the past few years, researchers have been focused on phenotype screening during drug discovery project. [1]

3 METHODOLOGY

Due to the target value is a discrete variable, we are going to run the data with the following ML models and select the best method depends on the performance. 1) Logistic Regression Classifier 2) Support Vector Machine Classifier 3) K-nearest neighbor (KNN) Classifier 4) Decision Tree Classifier 5) Random Forest Classifier

Sci-kit Learn package would be suitable for our predictive analysis. And the ML simulation result accuracy and mse will be used for model validation and model verification. From other aspects, we have molecules, atom pairs, so this means data, which is interconnected. Network graphs should be useful to visualize such data. The molecular atoms and bonds could be represented as the graph network. According to our project's unique purpose, although the complex graph network algorithms such as shortest path, sub communities are not part of our study, to demonstrate the unique bonds with the atoms by type would be meaningful for further understanding the molecules. We Plot the network graph by type but the graph We can see that atom connections have different shapes for different types.

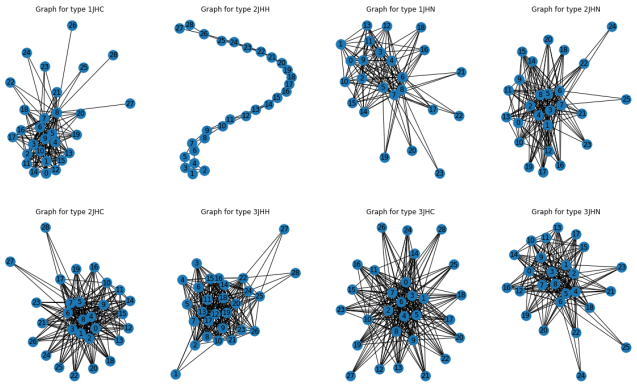

Figure 3: Graph Representation

As a result graphs will be skewed. Which seriously affects our conception regards to the atom's connection. One of the ways to improve the visualization is to get rid of the least relevant information. In order to optimize the graph data visualization, we need to perform data normalization and drop the rare frequent atom nodes from the atom connection graph.

4 CONCLUSION

Living things are made up of atoms, but in most cases, those atoms are not just floating around individually. Instead, they're usually interacting with other atoms (or groups of atoms). For instance, atoms might be connected by strong bonds and organized into molecules or crystals. Or they might form temporary, weak bonds with other atoms that they bump into or brush up against. Both the strong bonds that hold molecules together and the weaker bonds that create temporary connections are essential to the chemistry of our bodies, and to the existence of life itself. Why form chemical bonds? The basic answer is that atoms are trying to reach the most stable (lowest-energy) state that they can. Many atoms become stable when their valence shell is filled with electrons or when they

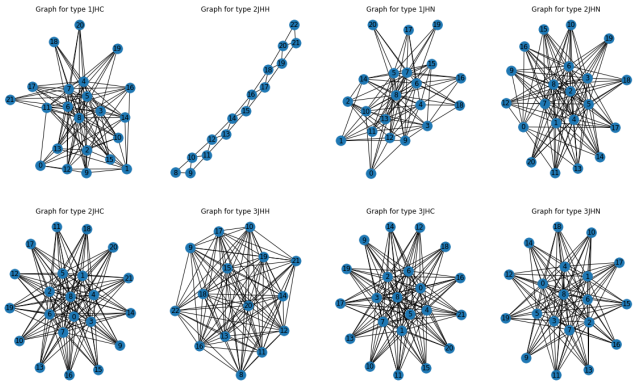

Figure 4: Graph Representation after Data Normalization

satisfy the octet rule (by having eight valence electrons). If atoms don't have this arrangement, they'll "want" to reach it by gaining, losing, or sharing electrons via bonds. So we tried to know the connection and arrangement of molecules with NMR and by using machine learning which will be useful for future of the mankind.

REFERENCES

[1] A. Gaulton, A. Hersey, A. P. Nowotka, M.and Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R Leach. 2017. The ChEMBL database in 2017. *Nucleic acids research* (2017), 945–954.

[2] S. Lee, B.and Zhang, A. Poleksic, and L Xie. [n.d.]. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysisr. *Frontiers in genetics* 10 ([n. d.]).

[3] Benoit Playe and Veronique Stoven. [n.d.]. *Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity.*

[4] H. Wang, J. Wang, C. Dong, Y. Lian, D. Liu, and Z. Yan. [n.d.]. A Novel Approach for Drug-Target Interactions Prediction Based on Multimodal Deep Autoencoder. *Frontiers in pharmacology* 10 ([n. d.]).

[5] E.L Willighagen. [n.d.]. .

[6] A yanenko. 2020. *Molecular Properties EDA and models.* Retrieved Jan 30, 2020 from https://www.kaggle.com/artgor/molecular-properties-eda-and-models/notebook

[7] Zahoránszky-Kőhalmi, T. G., Sheils, and T. I Oprea. 2020. SmartGraph: a network pharmacology investigation platform. *Journal of Cheminformatic* 12, 5 (2020).