Data Science Group Meet Up

Min-Seok Kye

29th June, 2019

Agenda

1. Introduction of NMR

2. What is Scalar Coupling Constant

3. What else to know
    1. Dipole Moment
    2. Magnetic Shielding Tensor
    3. Mulliken Charge
    4. Potential Energy

4. Predicting Molecular Properties in Kaggle Competition

5. Additional Information for Competition

Introduction of NMR

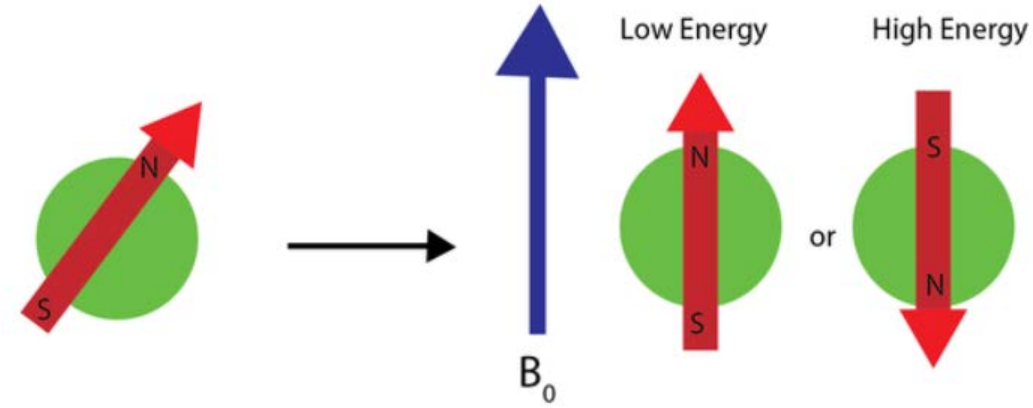
NMR Spectroscopy

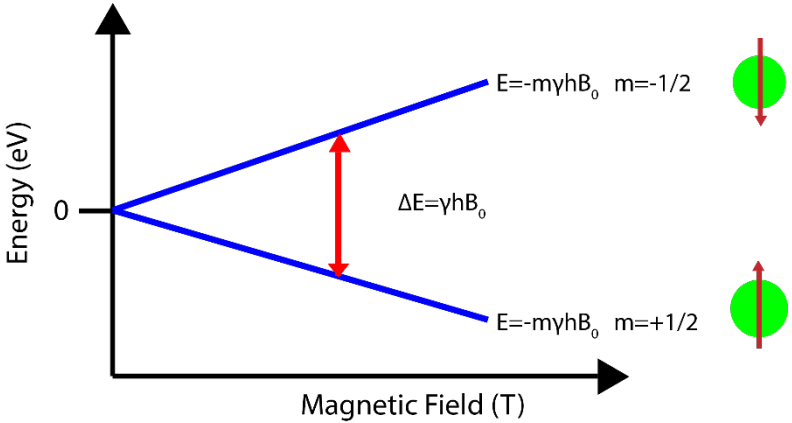
Nuclear magnetic resonance spectroscopy, most commonly known as **NMR spectroscopy** or **magnetic resonance spectroscopy** (**MRS**), is a spectroscopic technique to observe local magnetic fields around atomic nuclei. The sample is placed in a magnetic field and the NMR signal is produced by excitation of the nuclei sample with radio waves into nuclear magnetic resonance, which is detected with sensitive radio receivers. The intramolecular magnetic field around an atom in a molecule changes the resonance frequency, thus giving access to details of the electronic structure of a molecule and its individual functional groups. As the fields are unique or highly characteristic to individual compounds, in modern organic chemistry practice, NMR spectroscopy is the definitive method to identify monomolecular organic compounds. Similarly, biochemists use NMR to identify proteins and other complex molecules. Besides identification, NMR spectroscopy provides detailed information about the structure, dynamics, reaction state, and chemical environment of molecules. The most common types of NMR are proton and carbon-13 NMR spectroscopy, but it is applicable to any kind of sample that contains nuclei possessing spin.

A 900 MHz NMR instrument with a 21.1 T magnet at HWB-NMR, Birmingham, UK

Introduction of NMR

NMR Spectroscopy





^{1}H MRI of a human head showing the soft tissue such as the brain and sinuses. The MRI also clearly shows the spinal column and skull

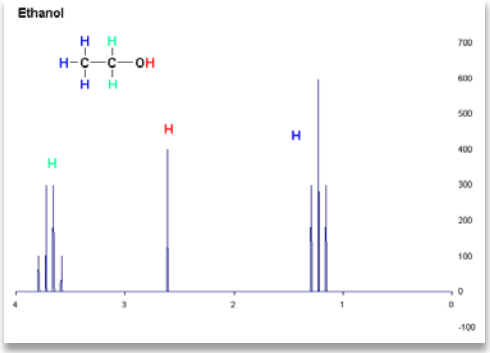
What is Scalar Coupling Constant

Some of the most useful information for structure determination in a one-dimensional NMR spectrum comes from **J-coupling** or **scalar coupling** (a special case of spin-spin coupling) between NMR active nuclei. This coupling arises from the interaction of different spin states through the chemical bonds of a molecule and results in the splitting of NMR signals. For a proton, the local magnetic field is slightly different depending on whether an adjacent nucleus points towards or against the spectrometer magnetic field, which gives rise to two signals per proton instead of one. These splitting patterns can be complex or simple and, likewise, can be straightforwardly interpretable or deceptive. This coupling provides detailed insight into the connectivity of atoms in a molecule.

Coupling to *n* equivalent (spin ½) nuclei splits the signal into a *n*+1 **multiplet** with intensity ratios following Pascal's triangle as described on the right. Coupling to additional spins will lead to further splittings of each component of the multiplet e.g. coupling to two different spin ½ nuclei with significantly different coupling constants will lead to a *doublet of doublets* (abbreviation: dd). Note that coupling between nuclei that are chemically equivalent (that is, have the same chemical shift) has no effect on the NMR spectra and couplings between nuclei that are distant (usually more than 3 bonds apart for protons in flexible molecules) are usually too small to cause observable splittings. *Long-range* couplings over more than three bonds can often be observed in cyclic and aromatic compounds, leading to more complex splitting patterns.

For example, in the proton spectrum for ethanol described above, the CH_3 group is split into a *triplet* with an intensity ratio of 1:2:1 by the two neighboring CH_2 protons. Similarly, the CH_2 is split into a *quartet* with an intensity ratio of 1:3:3:1 by the three neighboring CH_3 protons. In principle, the two CH_2 protons would also be split again into a *doublet* to form a *doublet of quartets* by the hydroxyl proton, but intermolecular exchange of the acidic hydroxyl proton often results in a loss of coupling information.

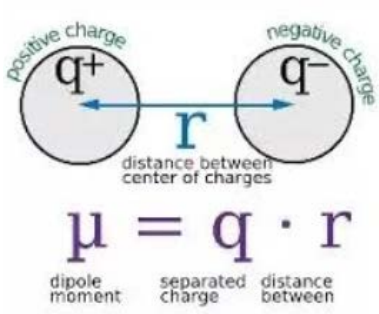
Multiplicity	Intensity Ratio
Singlet (s)	1
Doublet (d)	1:1
Triplet (t)	1:2:1
Quartet (q)	1:3:3:1
Quintet	1:4:6:4:1
Sextet	1:5:10:10:5:1
Septet	1:6:15:20:15:6:1

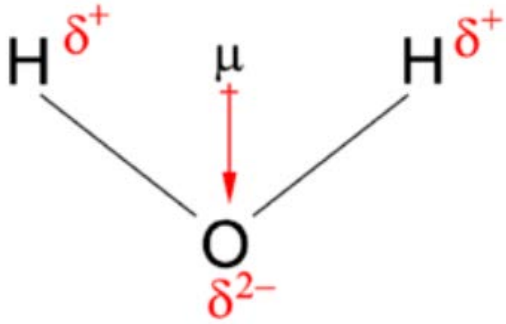


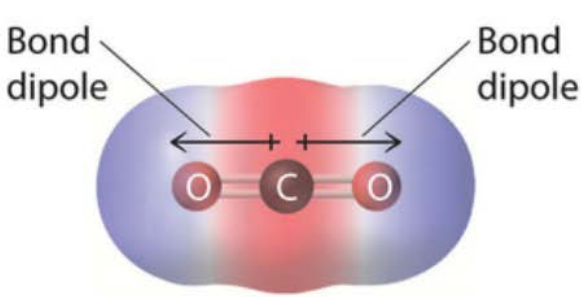
What else to know

Dipole Moment

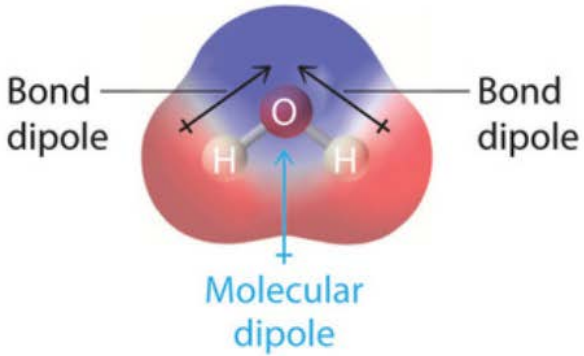
$$\vec{\mu} = \sum_i q_i \vec{r}_i$$







(a) No net dipole moment



(b) Net dipole moment

What else to know

Magnetic Shielding Tensor

The shielding tensor, σ is a second-order property depending on the external magnetic field, B and the nuclear magnetic spin momentum, m_k of nucleus k

$$\Delta E = -m_j(1-\sigma)B$$

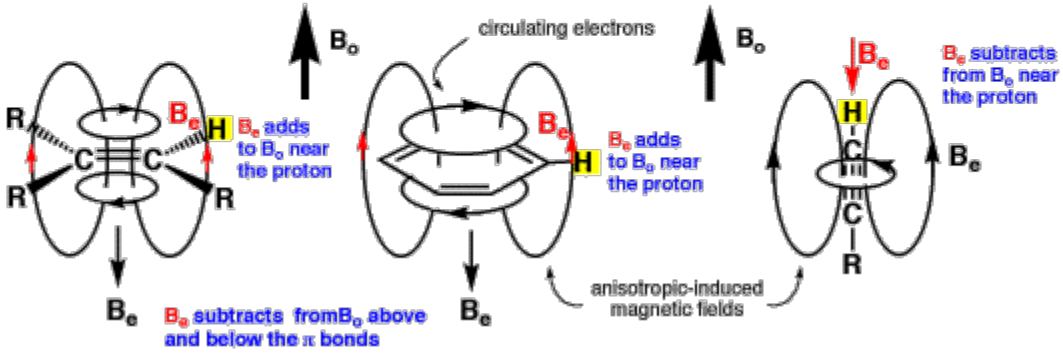
Using analytical derivative techniques to evaluate σ the components of this 3×3 tensor are computed as

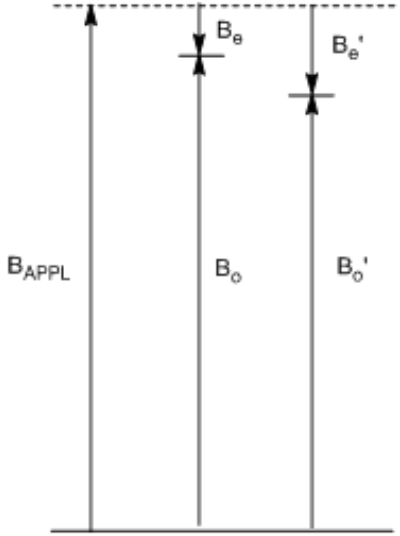
$$\sigma_{ij} = \sum_{\mu\nu} P_{\mu\nu} \frac{\partial^2 h_{\mu\nu}}{\partial B_i \partial m_{j,k}} + \sum_{\mu\nu} \frac{\partial P_{\mu\nu}}{\partial B_i} \frac{\partial h_{\mu\nu}}{\partial m_{j,k}}$$

What else to know

Magnetic Shielding by Electron

Figure 7. Components of the B_e field produced by the circulation of π-electrons in alkenes, aromatics, and alkynes as a function of orientation relative to the applied field B_o. The major anisotropic contribution to the total B_e field from circulation of π-electrons adds to (alkene or aromatic) or subtracts from B_o (alkyne).





What else to know

Mulliken Charge

Mulliken charges arise from the **Mulliken population analysis**[1][2] and provide a means of estimating partial atomic charges from calculations carried out by the methods of computational chemistry, particularly those based on the linear combination of atomic orbitals molecular orbital method, and are routinely used as variables in linear regression (QSAR[3]) procedures.[4] The method was developed by Robert S. Mulliken, after whom the method is named. If the coefficients of the basis functions in the molecular orbital are $\mathbf{C}_{\mu i}$ for the μ'th basis function in the i'th molecular orbital, the density matrix terms are:

$$\mathbf{D}_{\mu\nu} = 2\sum_{i} \mathbf{C}_{\mu i}\mathbf{C}^{*}_{\nu i}$$

for a closed shell system where each molecular orbital is doubly occupied. The population matrix $\mathbf{P}$ then has term

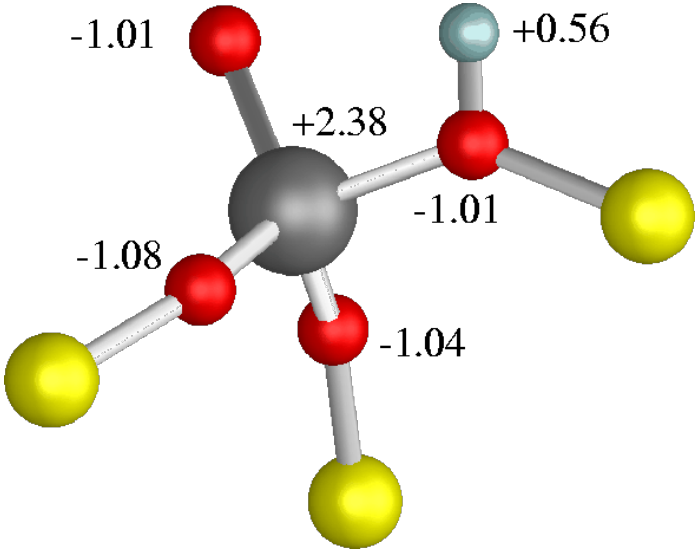
$$\mathbf{P}_{\mu\nu} = \mathbf{D}_{\mu\nu}\mathbf{S}_{\mu\nu}$$

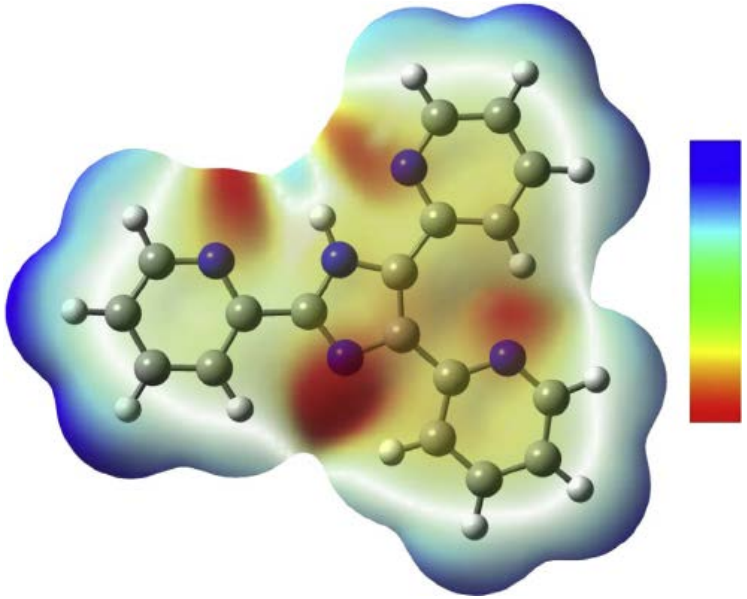
S is the overlap matrix of the basis functions. The sum of all terms of $\mathbf{P}_{\nu\mu}$ summed over u is the gross orbital product for orbital v - $\mathbf{GOP}_{\nu}$. The sum of the gross orbital products is **N** - the total number of electrons. The Mulliken population assigns an electronic charge to a given atom **A**, known as the gross atom population: $\mathbf{GAP}_{\mathbf{A}}$ as the sum of $\mathbf{GOP}_{\nu}$ over all orbital v belonging to atom A. The charge, $\mathbf{Q}_{\mathbf{A}}$ is then defined as the difference between the number of electrons on the isolated free atom, which is the atomic number $\mathbf{Z}_{\mathbf{A}}$, and the gross atom population:

$$\mathbf{Q}_{\mathbf{A}} = \mathbf{Z}_{\mathbf{A}} - \mathbf{GAP}_{\mathbf{A}}$$

What else to know

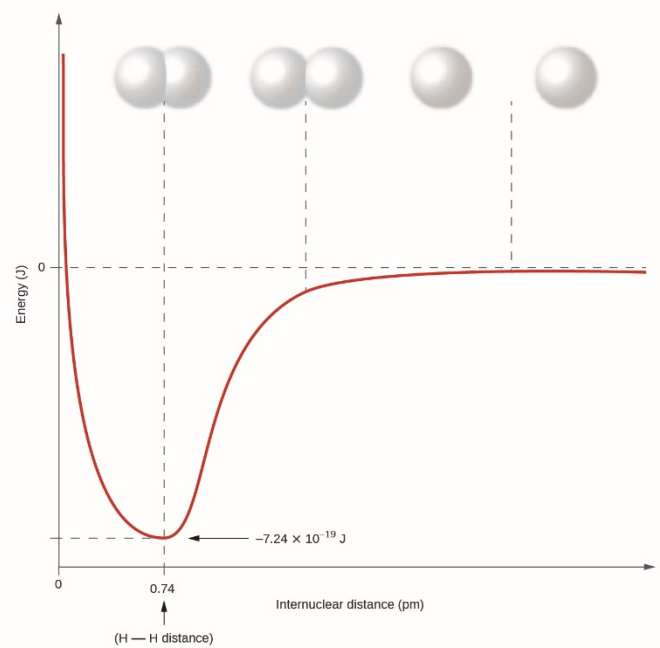
Mulliken Charge

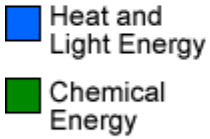


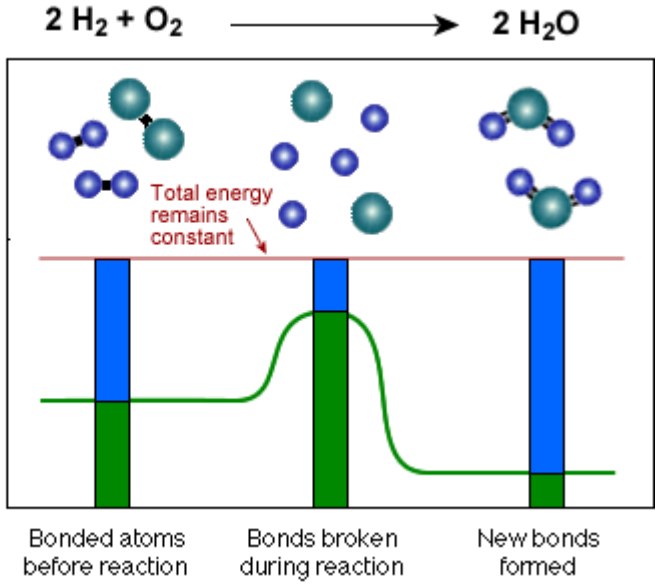


What else to know

Potential Energy







Kaggle Competition

Evaluation of Predicting Molecular Properties

Submissions are evaluated on the Log of the Mean Absolute Error, calculated for each scalar coupling type, and then averaged across types, so that a 1% decrease in MAE for one type provides the same improvement in score as a 1% decrease for another type.

$$score = \frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}|y_i - \hat{y}_i|\right)$$

Where:

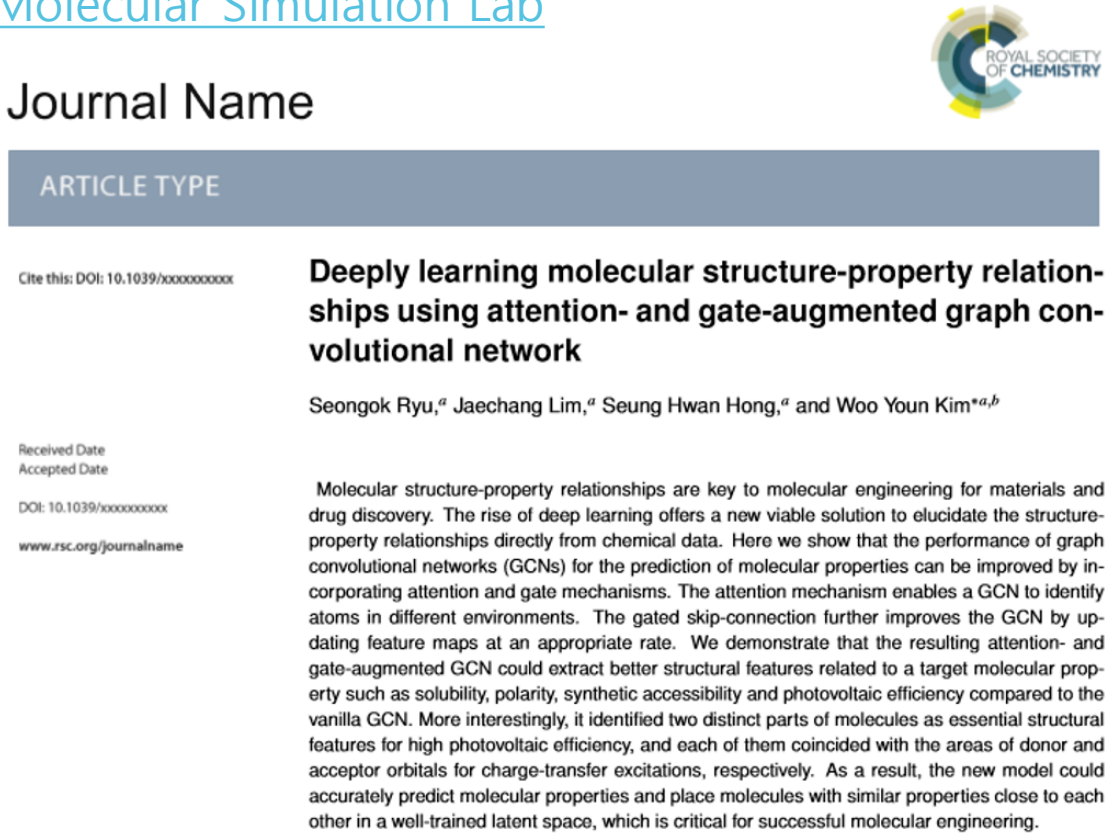
- T is the number of scalar coupling types
- n_t is the number of observations of type t
- y_i is the actual scalar coupling constant for the observation
- $\hat{y}_i$ is the predicted scalar coupling constant for the observation

For this metric, the MAE for any group has a floor of `1e-9`, so that the minimum (best) possible score for perfect predictions is approximately -20.7232.

Additional Information for Competition

Rank 2nd by Jaechang Lim in KAIST – Graduate people in Chemistry

ab initio Molecular Simulation Lab



Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network

Seongok Ryu,[a] Jaechang Lim,[a] Seung Hwan Hong,[a] and Woo Youn Kim*[a,b]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Molecular structure-property relationships are key to molecular engineering for materials and drug discovery. The rise of deep learning offers a new viable solution to elucidate the structure-property relationships directly from chemical data. Here we show that the performance of graph convolutional networks (GCNs) for the prediction of molecular properties can be improved by incorporating attention and gate mechanisms. The attention mechanism enables a GCN to identify atoms in different environments. The gated skip-connection further improves the GCN by updating feature maps at an appropriate rate. We demonstrate that the resulting attention- and gate-augmented GCN could extract better structural features related to a target molecular property such as solubility, polarity, synthetic accessibility and photovoltaic efficiency compared to the vanilla GCN. More interestingly, it identified two distinct parts of molecules as essential structural features for high photovoltaic efficiency, and each of them coincided with the areas of donor and acceptor orbitals for charge-transfer excitations, respectively. As a result, the new model could accurately predict molecular properties and place molecules with similar properties close to each other in a well-trained latent space, which is critical for successful molecular engineering.

Additional Information for Competition

One of Kaggle Discussion

For the last year, I have been focusing on bringing more scientific challenges to Kaggle. I personally believe it's a benefit to both our community as well as the scientific community to host these different types of problems. The obvious scientific benefit from this challenge is being able to predict molecular properties without doing the expensive quantum calculations. If you were doing this as a researcher, you wouldn't add in quantum computation properties as inputs to your model, as that defeats the purpose.

But what if a researcher wants to run a Kaggle competition for this problem? We know that some properties of the QM7 and QM9 molecules have been published, that they are expected to provide minimal-to-no benefit to the target for this competition, and using this data would be completely against the spirit of the scientific challenge. It would seem a prudent course of action to exclude this data from the competition.

That simple request was violated before we were even able to have a discussion about it on the forums. It is already extremely difficult to find practical scientific challenges that can't be abused via external data. And, unfortunately, there are some community members who feel entitled to abuse the rules "to make things fair".

The unfortunate reality of this situation is that we will have fewer opportunities to host scientific challenges on Kaggle. In addition, it means the scientific competitions we do launch are less likely to be Featured competitions, i.e., they will be launched as Research competitions without points or medals, and with smaller prize pools.

To address this current competition, *we are updating the rules to allow the use of QM7 and QM9 published data*, if participants wish to use it. Participants still may *not* use quantum calculations to obtain predictions.

With that said, I can't stress enough how wrong it is to break competition rules because you disagree with them or you want to force a particular course of action. Because this is such an important point, even though we are allowing the use of QM7 and QM9 going forward, any evidence that someone used the data to make a submission *before* this rule change will result in that individual's (or team's) removal from the competition once it has closed.