COMP 396: Milestone 1

Data representation

Summary

1. Dataset description
2. Implementation and Data exploration
3. Baseline models (with hand engineered features)->https://github.com/bmbodj/COMP396 (or check colab notebook links below)
4. Next steps

The first objective of the research project is to understand the data that will be used, the methodology of the paper -> Neural Message Passing for Quantum Chemistry (https://arxiv.org/pdf/1704.01212.pdf) and attempt to reproduce the baseline models that were mentioned in the paper.

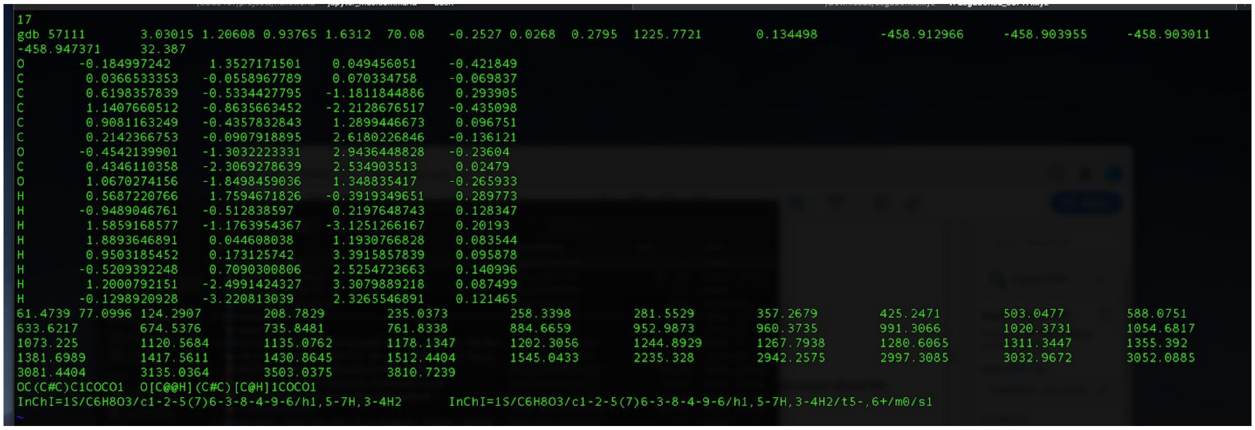
Dataset description

Data: QM9 (Quantum machine 9) is used as a benchmark for (Ramakrishnan et al., 2014) as well as for message passing neural networks. It is available on Figshare in an xyz folder format. The folder contains xyz files that have molecular model descriptions, including atom numbers, element symbols and X, Y, and Z coordinates

The complete description of the data is available at the following link - >Quantum chemistry structures and properties of 134 kilo molecules https://www.nature.com/articles/sdata201422.pdf

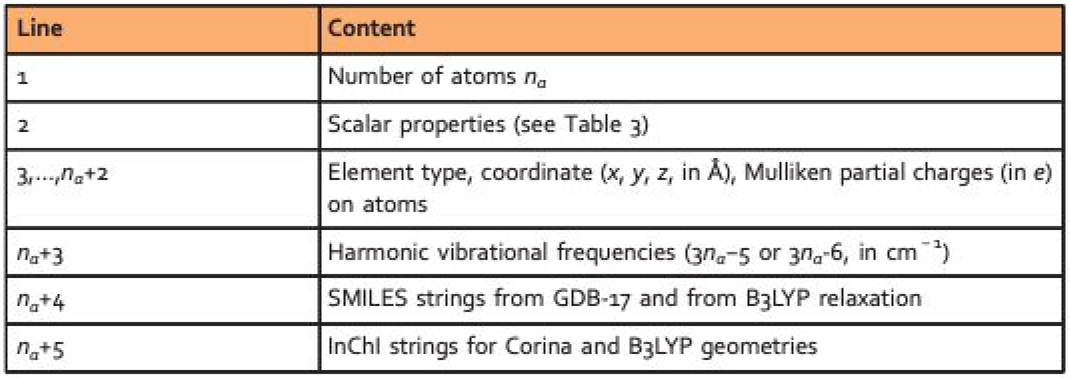
Direct download link : https://figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules/978904 :

The following is a screenshot of a single xyz file.



```
17
gdb 57111       3.03015 1.20608 0.93765 1.6312  70.08   -0.2527 0.0268  0.2795  1225.7721       0.134498        -458.912966     -458.903955     -458.903011
-458.947371     32.387
O       -0.184997242     1.3527171501    0.049456051    -0.421849
C        0.0366533353   -0.0558967789    0.070334758    -0.069837
C        0.6198357839   -0.5334427795   -1.1811844886    0.293905
C        1.1407660512   -0.8635663452   -2.2128676517   -0.435098
C        0.9081163249   -0.4357832843    1.2899446673    0.096751
C        0.2142366753   -0.0907918895    2.6180226846   -0.136121
O       -0.4542139901   -1.3032223331    2.9436448828   -0.23604
C        0.4346110358   -2.3069278639    2.534903513     0.02479
O        1.0670274156   -1.8498459036    1.348835417    -0.265933
H        0.5687220766    1.7594671826   -0.3919349651    0.289773
H       -0.9489046761   -0.512838597     0.2197648743    0.128347
H        1.5859168577   -1.1763954367   -3.1251266167    0.20193
H        1.8893646891    0.044608038     1.1930766828    0.083544
H        0.9503185452    0.173125742     3.3915857839    0.095878
H       -0.5209392248    0.7090300806    2.5254723663    0.140996
H        1.2000792151   -2.4991424327    3.3079889218    0.087499
H       -0.1298920928   -3.220813039     2.3265546891    0.121465
61.4739 77.0996 124.2907        208.7829        235.0373        258.3398        281.5529        357.2679        425.2471        503.0477        588.0751
633.6217        674.5376        735.8481        761.8338        884.6659        952.9873        960.3735        991.3066        1020.3731       1054.6817
1073.225        1120.5684       1135.0762       1178.1347       1202.3056       1244.8929       1267.7938       1280.6065       1311.3447       1355.392
1381.6989       1417.5611       1430.8645       1512.4404       1545.0433       2235.328        2942.2575       2997.3085       3032.9672       3052.0885
3081.4404       3135.0364       3503.0375       3810.7239
OC(C#C)C1COCO1  O[C@@H](C#C)[C@H]1COCO1
InChI=1S/C6H8O3/c1-2-5(7)6-3-8-4-9-6/h1,5-7H,3-4H2      InChI=1S/C6H8O3/c1-2-5(7)6-3-8-4-9-6/h1,5-7H,3-4H2/t5-,6+/m0/s1
~
```

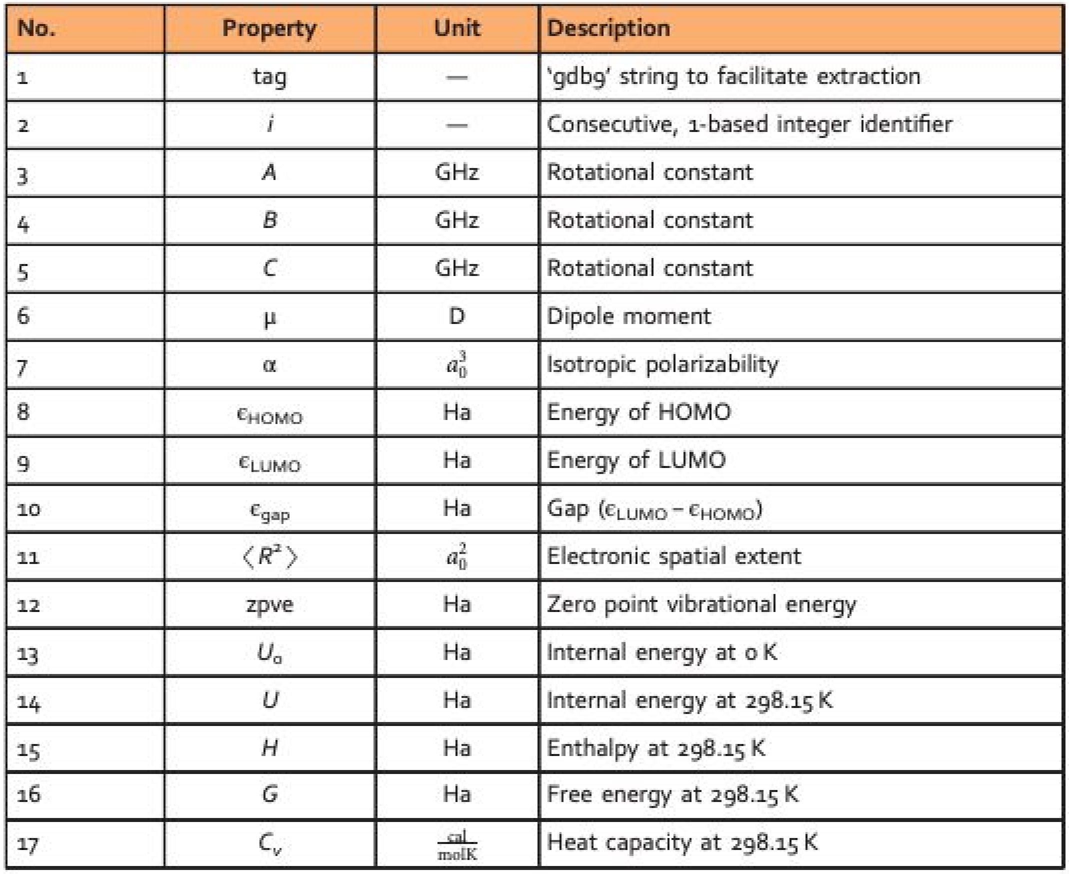
The tables below summarize the contents of the xyz files for each molecule.



Line	Content
1	Number of atoms n_a
2	Scalar properties (see Table 3)
3,…,n_a+2	Element type, coordinate (x, y, z, in Å), Mulliken partial charges (in e) on atoms
n_a+3	Harmonic vibrational frequencies (3n_a−5 or 3n_a-6, in cm^{-1})
n_a+4	SMILES strings from GDB-17 and from B3LYP relaxation
n_a+5	InChI strings for Corina and B3LYP geometries

Table 2. XYZ-like file format for molecular structure and properties. n_a = number of atoms.

The properties listed on table 3 represent the targets (1-13?) that are used for the neural message passing for quantum chemistry experiments.



No.	Property	Unit	Description
1	tag	—	'gdb9' string to facilitate extraction
2	i	—	Consecutive, 1-based integer identifier
3	A	GHz	Rotational constant
4	B	GHz	Rotational constant
5	C	GHz	Rotational constant
6	μ	D	Dipole moment
7	α	a_0^3	Isotropic polarizability
8	ϵ_{HOMO}	Ha	Energy of HOMO
9	ϵ_{LUMO}	Ha	Energy of LUMO
10	ϵ_{gap}	Ha	Gap ($\epsilon_{\mathrm{LUMO}} - \epsilon_{\mathrm{HOMO}}$)
11	$\langle R^2 \rangle$	a_0^2	Electronic spatial extent
12	zpve	Ha	Zero point vibrational energy
13	U_0	Ha	Internal energy at 0 K
14	U	Ha	Internal energy at 298.15 K
15	H	Ha	Enthalpy at 298.15 K
16	G	Ha	Free energy at 298.15 K
17	C_v	$\frac{\mathrm{cal}}{\mathrm{molK}}$	Heat capacity at 298.15 K

Table 3. Calculated properties. Properties are stored in the order given by the first column.

A more detailed description of the quantum properties copied from NPMC (https://arxiv.org/pdf/1704.01212.pdf).

NB: The properties used in the experiment are N0 6-17. The target Omega is also used, however, we couldn't find it 's corresponding property number in the table.

Atomization energies;

Atomization energy at 0K U0 (eV): This is the energy required to break up the molecule into all of its constituent atoms if the molecule is at absolute zero. This calculation assumes that the molecules are held at fixed volume.

- Atomization energy at room temperature U (eV): Like U0, this is the energy required to break up the molecule if it is at room temperature.

• Enthalpy of atomization at room temperature H (eV): The enthalpy of atomization is similar in spirit to the energy of atomization, U. However, unlike the energy this calculation assumes that the constituent molecules are held at fixed pressure

• Free energy of atomization G (eV): Once again this is similar to U and H, but assumes that the system is held at fixed temperature and pressure during the dissociation.

Fundamental vibration properties

• Highest fundamental vibrational frequency ω1 (cm−1): Every molecule has fundamental vibrational modes that it can naturally oscillate at. ω1 is the mode that requires the most energy. •

Zero Point Vibrational Energy (ZPVE) (eV): Even at zero temperature quantum mechanical uncertainty implies that atoms vibrate. This is known as the zero point vibrational energy and can be calculated once the allowed vibrational modes of a molecule are known.

States of electrons in the molecule

• Highest Occupied Molecular Orbital (HOMO) εHOMO (eV): Quantum mechanics dictates that the allowed states that electrons can occupy in a molecule are discrete. The Pauli exclusion principle states that no two electrons may occupy the same state. At zero temperature, therefore, electrons stack in states from lowest energy to highest energy. HOMO is the energy of the highest occupied electronic state.

• Lowest Unoccupied Molecular Orbital (LUMO) εLUMO (eV): Like HOMO, LUMO is the lowest energy electronic state that is unoccupied.

• Electron energy gap Δε (eV): This is the difference in energy between LUMO and HOMO. It is the lowest energy transition that can occur when an electron is excited from an occupied state to an unoccupied state. Δε also dictates the longest wavelength of light that the molecule can absorb.

Spatial distribution of electrons in the molecule:

• Electronic Spatial Extent hR2 i (Bohr2): The electronic spatial extent is the second moment of the charge distribution, ρ(r), or in other words hR2 i = R drr2ρ(r).

• Norm of the dipole moment µ (Debye): The dipole moment, p(r) = R dr0p(r 0)(r − r 0), approximates the electric field far from a molecule. The norm of the dipole moment is related to how anisotropically

Smile strings

Bonds are denoted as shown below:

Single bond	-
Double bond	=
Aromatic bond	#
Disconnected structures	.

Implementation and data exploration

Implementation

Main tools : In order to implement the baseline models with hand engineered molecular representations, we used RDKit which is an open source toolkit for cheminformatics, DeepChem, an integrated python library for chemistry and drug discovery as well as Pytorch and sci-kit learn.

Data exploration

We started by experimenting with the tools we mentioned above and explored the data in the xyz file.-> https://colab.research.google.com/drive/1N6wxzdwjb-35-gJniG0Tz2CC0LHsGmwy

Baseline models with hand engineered features

Coulomb matrix (CM)

Description: The coulomb matrix representation captures the geometry of a molecule by treating individual atoms as nodes and edge weights are computed from energetic interactions between pairs of node,

The coulomb matrix is calculated with the equation below:

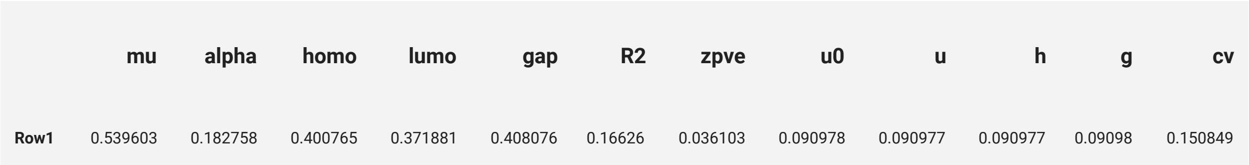
$$\boxed{M_{ij}^{\mathrm{Coulomb}} = \begin{cases} 0.5Z_i^{2.4} & \text{for } i = j \\ \dfrac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases}}$$

Results:

https://colab.research.google.com/drive/1q0zLgrTsSyzh4RBL69H5JrW7KlgakDAA

Dataset size : 10000 (80% training, 20% testing)

NB: This is the regular coulomb matrix



	mu	alpha	homo	lumo	gap	R2	zpve	u0	u	h	g	cv
Row1	0.539603	0.182758	0.400765	0.371881	0.408076	0.16626	0.036103	0.090978	0.090977	0.090977	0.09098	0.150849

-> experiment with eigenspectrum version
https://colab.research.google.com/drive/1U8VlPsfX2Uicj78JT9EnP7143XUqQeIn

Bag of Bonds (BoB)

https://pubs.acs.org/doi/pdf/10.1021/acs.jpclett.5b00831 (original paper)

Description: The bag of bonds is a descriptor that is similar in concept and inspired by the natural language processing bag of words. In order to get more accurate predictions throughout the chemical compound space, which is the space populated by all possible energetically stable molecules in composition, size and structure. The BoB is a vector composed of bags of particular bond types where each entry in every bag is computed in a similar fashion as in the coulomb matrix. All bag of bonds are concatenated in a specified order and empty bags are padded with zeros to give equal sizes across all molecules. This descriptor is invariant under molecular rotations and translations as well as row and column permutations.

Limitations: The main limitation of both the coulomb matrix (both regular and eigenspectrum type) and the bag of bonds model is its constant dimension as the length of the feature vectors depends on the number of atoms in the largest molecule of interest.

In order to reproduce this descriptor we used the implementation from mmltoolkit and added some modifications to correct the code.

Using sci-kit learn we trained and tested our data representation with kernel ridge regression and as in the original paper a laplacian kernel was defined. Laplacian kernels are often favored for their ability to optimally utilize information in non local chemical compound space.

Results:

https://colab.research.google.com/drive/1-alPAEcBrmECF_uF3Z9rCKzWcflwzyv4

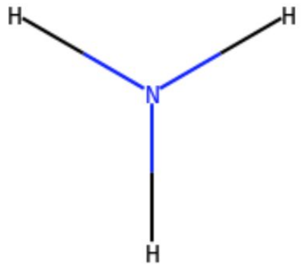
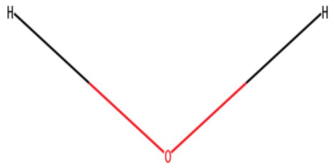
Dataset size : 10000 (80% training, 20% testing)

mu	alpha	homo	lumo	gap	R2	zpve	u0	u	h	g	cv
0.445946	0.097231	0.291975	0.23343	0.23343	0.098986	0.027917	0.020636	0.020636	0.020636	0.020637	0.092969

-> Experiment with 16000 molecules

https://colab.research.google.com/drive/1sCchqP_U7ejYqeiYEeWSDYX2Fvxfleca

The following is an example of how the bob model is implemented for NH3 and H20.

Molecule	**NH3**	**H20**
SMILES		
X,Y,Z Coordinates	[[-0.04042605 1.02410775 0.0625638] [0.01725746 0.01254521 -0.02737716] [0.91578937 1.35874519 -0.02875776] [-0.52027774 1.34353213 -0.77554261]]	[[-3.43604951e-02 9.77539571e-01 7.60159230e-03] [6.47664923e-02 2.05721989e-02 1.53463410e-03] [8.71790374e-01 1.30079240e+00 6.93133600e-04]]
Atom types	['N', 'H', 'H', 'H']	['O', 'H', 'H']

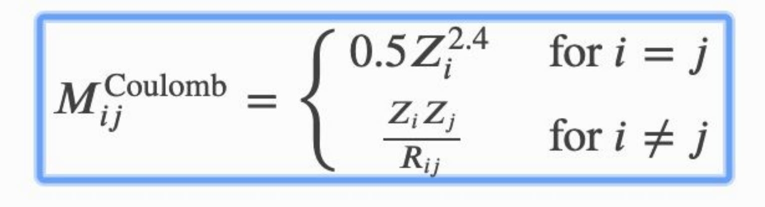
Charge array/ Atomic number	[7, 1, 1, 1]	[8, 1, 1]

1) **Initialize dictionary**

   {'C': [], 'N': [], 'O': [], 'F': [], 'H': [], 'CC': [], 'CN': [], 'CO': [], 'CF': [], 'CH': [], 'NN': [], 'NO': [], 'NF': [], 'NH': [], 'OO': [], 'OF': [], 'OH': [], 'FF': [], 'FH': [], 'HH': []}

2) **Retrieve coordinates atom types and charge array for both molecules (see table)**

3) **The energy for each bag is calculated with the same formula as in the coulomb matrix :**



$$M_{ij}^{\text{Coulomb}} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases}$$

   Z is the charge array

4) **At the end of this step we end up with the following bags;**

   1st bag... keys and list

   dict_keys(['C', 'N', 'O', 'F', 'H', 'CC', 'CN', 'CO', 'CF', 'CH', 'NN', 'NO', 'NF', 'NH', 'OO', 'OF', 'OH', 'FF', 'FH', 'HH'])

   {'C': [], 'N': [53.3587073998281], 'O': [], 'F': [], 'H': [0.5, 0.5, 0.5], 'CC': [], 'CN': [], 'CO': [], 'CF': [], 'CH': [], 'NN': [], 'NO': [], 'NF': [], 'NH': [6.881703335693628, 6.881722543904302, 6.881582376790935], 'OO': [], 'OF': [], 'OH': [], 'FF': [], 'FH': [], 'HH': [0.6178473546944685, 0.6177759125242487, 0.617777567529042]}

   2nd bag... keys and list

   dict_keys(['C', 'N', 'O', 'F', 'H', 'CC', 'CN', 'CO', 'CF', 'CH', 'NN', 'NO', 'NF', 'NH', 'OO', 'OF', 'OH', 'FF', 'FH', 'HH'])

   {'C': [], 'N': [], 'O': [73.51669471981023], 'F': [], 'H': [0.5, 0.5], 'CC': [], 'CN': [], 'CO': [], 'CF': [], 'CH': [], 'NN': [], 'NO': [], 'NF': [], 'NH': [], 'OO': [],

```
'OF': [], 'OH': [8.31508507867743, 8.31508507857124], 'FF': [], 'FH': [],
'HH': [0.6607822398398421]}
```

5) Keep track of the maximum length

6) Concatenate and sort the bags(to enforce permutational invariance)

The bonds used for both molecules are the following:

```
['N', 'O', 'H', 'H', 'H', 'NH', 'NH', 'NH', 'OH', 'OH', 'HH', 'HH', 'HH']
```

7) Final result : 2 feature vectors

```
  [[53.3587074 ,  0.        ,  0.5       ,  0.5       ,  0.5       ,
          6.88172254,  6.88170334,  6.88158238,  0.        ,  0.        ,
          0.61784735,  0.61777757,  0.61777591],
[ 0.        , 73.51669472,  0.5       ,  0.5       ,  0.        ,
          0.        ,  0.        ,  0.        ,  8.31508508,  8.31508508,
          0.66078224,  0.        ,  0.        ]])
```

Issues

We attempted to use the same metrics however we were not able to reproduce the same results. -

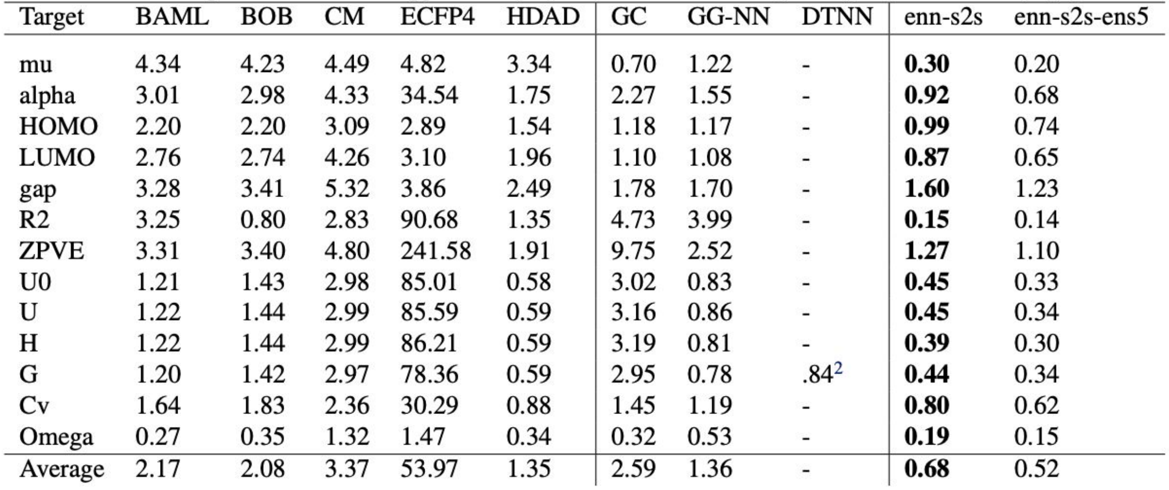


Table 2. Comparison of Previous Approaches (left) with MPNN baselines (middle) and our methods (right)

Target	BAML	BOB	CM	ECFP4	HDAD	GC	GG-NN	DTNN	enn-s2s	enn-s2s-ens5
mu	4.34	4.23	4.49	4.82	3.34	0.70	1.22	-	**0.30**	0.20
alpha	3.01	2.98	4.33	34.54	1.75	2.27	1.55	-	**0.92**	0.68
HOMO	2.20	2.20	3.09	2.89	1.54	1.18	1.17	-	**0.99**	0.74
LUMO	2.76	2.74	4.26	3.10	1.96	1.10	1.08	-	**0.87**	0.65
gap	3.28	3.41	5.32	3.86	2.49	1.78	1.70	-	**1.60**	1.23
R2	3.25	0.80	2.83	90.68	1.35	4.73	3.99	-	**0.15**	0.14
ZPVE	3.31	3.40	4.80	241.58	1.91	9.75	2.52	-	**1.27**	1.10
U0	1.21	1.43	2.98	85.01	0.58	3.02	0.83	-	**0.45**	0.33
U	1.22	1.44	2.99	85.59	0.59	3.16	0.86	-	**0.45**	0.34
H	1.22	1.44	2.99	86.21	0.59	3.19	0.81	-	**0.39**	0.30
G	1.20	1.42	2.97	78.36	0.59	2.95	0.78	.84[2]	**0.44**	0.34
Cv	1.64	1.83	2.36	30.29	0.88	1.45	1.19	-	**0.80**	0.62
Omega	0.27	0.35	1.32	1.47	0.34	0.32	0.53	-	**0.19**	0.15
Average	2.17	2.08	3.37	53.97	1.35	2.59	1.36	-	**0.68**	0.52

Next steps ?

-> reproduce the message passing neural networks baseline models and get a better understanding of the framework for graph feature learning and extraction.