Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

A machine learning approach to the simulation of bio-inorganic compounds

Author: Benjamin Stott, ID: 18336161
Supervisor: Alessandro Lunghi

SS Physics 2021/22 Trinity college Dublin

Contents

List of Figures

Acknowledgements

In this project I was sent sample Fitsnap script inputs by phd student Mr Valerio Briganti so that I could code up my own in a similar fashion. I was also aided by Ms Nguyen Vu Ha Anh, another phd student in Alessandro Lunghi's research group who showed me how to correctly compile fitsnap and include a necessary alteration in the lammps library files connecting to it. Finally, my supervisor aided me with some bug fixing, provided a training set for the Zinc complex to me because my device simply did not have the computing power to run the code in a good timeframe and read a draft of my final paper so that I could receive feedback.

Abstract

This study details various methods of simulating bio-inorganic compounds in particular those of use for pharmaceuticals like enzymes, namely the Human Carbonic Anhydrase II active site(30 atoms). The components of this enzyme(besides the zinc core), water(H_2O) and Imidazole($C_3H_4N_2$) were first tested using the GAFF force field parameters in molecular dynamics and compared to machine learning potentials(SNAP) using the fitsnap fortran code with DFT calculations as a quality test for each of these. This was then upscaled to the entire complex using the ZAFF force-field and SNAP machine learning. It was found that SNAP outperformed conventional molecular dynamics approaches by an order of magnitude. In the future, with proper care and rigor, combining machine learning and force-fields together may lead to an even more robust simulation platform.

note: references to the bibliography are denoted by superscript **|x|** where x corresponds to the index number of the bibliography entry.

:

I. Introduction

Bio-inorganic compounds are crucial to the function of living organisms. Their main role is as coordination complex centers(i.e Zinc, Iron,Copper etc) in catalytic enzymes. One that we shall focus upon in this study is the enzyme human carbonic anhydrase II[12]. Carbonic anhydrase is a well-studied, zinc-centred enzyme and its role is to catalyze the hydrolysis of carbon dioxide to bicarbonate ion. It is important to be able to make functioning computational models of such an enzyme's behaviour so that people can understand the effects of making pharmaceutical drugs upon this enzyme, to understand chemical reactions caused by them(to determine products of chemical reactions, reaction rates etc) and this at a low cost in a virtual environment. There are many approaches to this, some of which I shall introduce here to show the framework of this issue we tackle.

The first approach used was molecular dynamics. Detailed equations are in the next section but in brief it models atoms as particles which correspond to Newton's equations of motion. While this is representative of the overall movement of a system, it fails to capture quantum effects which are quite relevant when zooming into the level of an enzyme centre and its reactions. In order to provide Newton's equations with a much more mathematically detailed description of the potential, electronic structure and the atomic environment DFT(Density Functional Theory) is necessary. This is a theory which provides a quite complete explanation of electronic structure of an atom and the quantum effects upon it. It is the state of the art when it comes to accuracy in quantum simulations. Unfortunately, it is hamstringed by its high computational cost, scaling very poorly with the number of atoms per calculation. It can calculate numerous properties of interest to the material scientist such as the potential energy of a configuration which will be of most use to us. Clearly there must be a workaround to this.

Faced with this dilemna, we come to the crux of this paper.With the advent of machine learning, we're placed at a good time to solve this. Machine learning is a method by which certain properties of interest are stored as numerical values. This "training set" is used to teach an algorithm, provided a good enough sample size(with varying characteristics) to predict these properties exist. In light of our problem, this would mean training an algorithm with energies of molecular configurations predicted by DFT, and seeing how it performs. The benefit of such an approach is that it has a much lower computational cost

than DFT. This means that we can train a model, and once it is made, we can apply it to other situations on a grander scale with comparable accuracy.

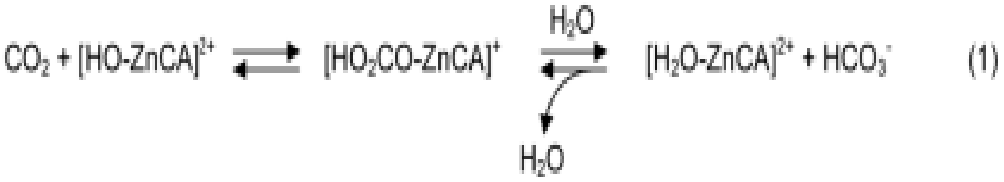
With this in mind, there is now a clear structure to this paper that will be ahered to. I will outline the current approaches in molecular dynamics, in particular, so-called "Force-fields" created by AMBER such as the GAFF(generalised Amber force field) and ZAFF(Zinc Amber force-field) later. The efficacy of molecular dynamics shall be tested(against DFT calculations) upon the inorganic components of the system(Zinc) and the organic components(water and imidazole). Data from DFT calculations will then be used as a training set for machine learning using the alogorithm(FITSNAP).

Finally, an attempt will be made to integrate force fields with machine learning to perhaps reach greater accuracy. This shall be elaborated upon in more detail in the following section.

II. Theory and Background

I. Carbonic anhydrase II

Human carbonic anhydrase II is an enzyme that catalyzes the hydrolysis of carbon dioxide to the bicarbonate ion according to the following reaction mechanism. This reaction is important in the respiratory system for the eventual exhalation of carbon dioxide that was intaken.



$$[H_2O - ZnCA]^{2+} \rightleftarrows [HO - ZnCA]^{+} + H^{+} \quad (2)$$

Figure 1: Carbonic anhydrase reaction mechanism[1]

The bound water is deprotonated rapidly, with the aid of a localized base, to regenerate the catalytic zinc-hydroxide species. We shall not delve too much into this as reaction dynamics are not covered.

The Zinc metallic enzyme centre takes the form seen in the following figure.

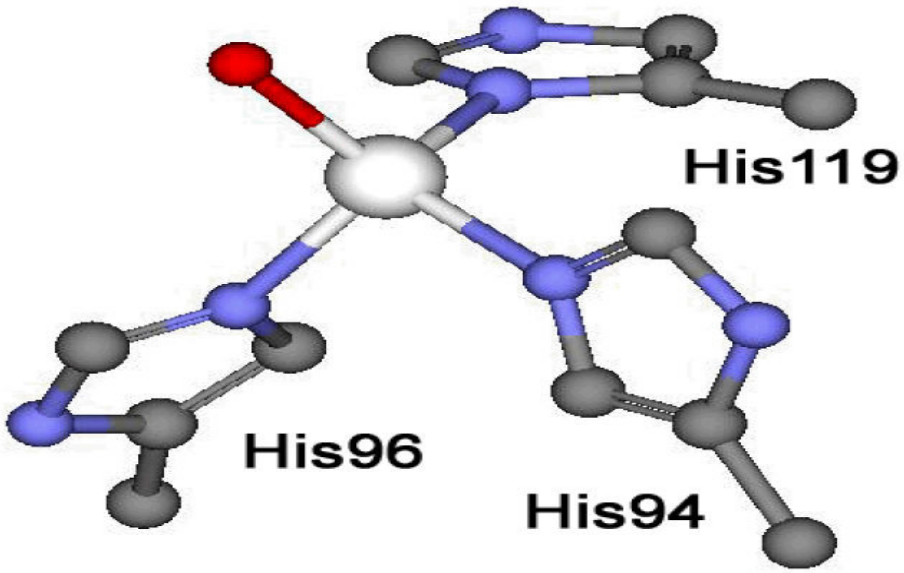


Figure 2: Enzyme metallic coordination centre(Zinc is white, Oxygen is red, carbon is grey and nitrogen is blue)[2]

We see a Zn^{+2} coordination compound coordinated to 3 histidine groups(with their residue IDs labelled) and an Oxygen to take the 4th free space. Hydrogens are not indicated here but they are indeed present, dangling on the carbons and the oxygen as usual.

For this study it is sufficient to examine components alone (water and imidazole) then work our way back to an analysis of the entire structure. Molecular dynamics is the first of such methods I shall speak of.

II. Molecular dynamics

II.1 Overview and historical background

Molecular dynamics is a method by which we can simulate the conformations of a molecule(different molecular configurations as it moves). It was developed in the aim of understanding ligand binding and such in order to create pharmaceutical drugs. X-ray crystallography can be performed to determine the atomic and molecular structure of a crystal. While this method is very concrete and practical, it is not pragmatic. It is laborious and costly, hence the need for simulation(although simulations will begin from crystallographic snapshots of structures and advance from there so it does have its uses in simulation if only as a preamble).

Molecular dynamics approximates atoms simply as particles, ignoring their electronic structure and treating them in a purely classical, Newtonian manner. The basic

algorithm is denoted below which I shall elaborate upon

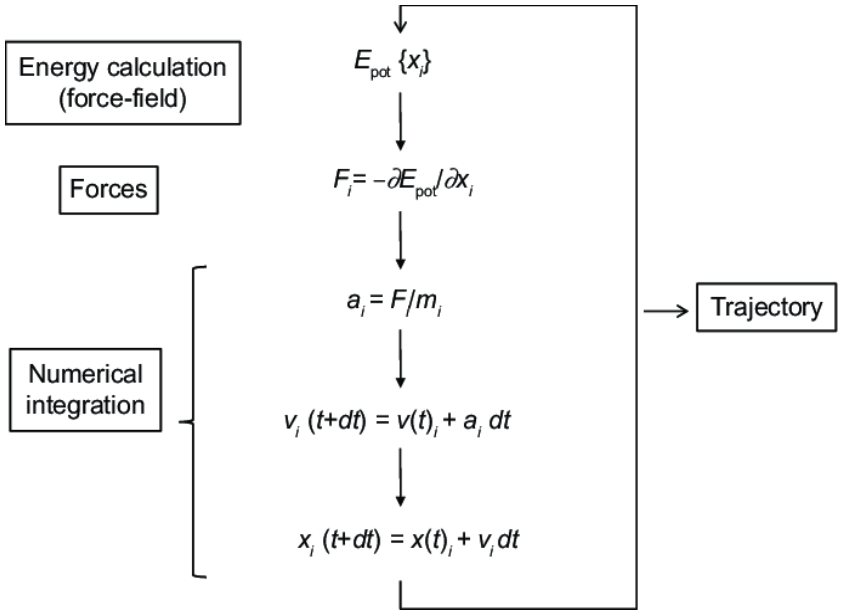


Figure 3: Molecular Dynamics Algorithm[3]

The algorithm is quite simple and indeed that is its strength since it means low computational cost. A potential energy value is calculated from a so called "force-field". A force field is a function that calculations potential energy as a function of atom positions in a certain conformation. It is normally down to the user to decide which force-field they would like to employ for a certain situation, many such as CHARRM, GAFF, etc existing for tailor made molecules at certain computational costs. The rest of the calculations are fairly self explanatory, the force is the negative gradient of energy, from force we gain acceleration, through which we can update our velocity and hence the position of each particle in the system. This is repeated iteratively for as many steps as the simulation is required.

Of course as with anything there are certain caveats to this simplicity. The first of which being that molecular dynamics does not account for quantum effects, making its description of complex reactions poor. The cost of molecular dynamics, while not as much as its competitors is still confined to small timescales for large systems. Fortunately this study deals with much smaller systems so this is not relevant here.

III. Force-fields

The definition of a force-field in a molecular dynamics context has been covered, I will explain two of the force fields that were utilised in this paper. Both of them belong to AMBER, a family of force fields used to study biomolecules

III.1 GAFF

GAFF, also known as the generalised amber force field, has parameters for almost all the organic molecules made of C, N, O, H, S, P, F, Cl, Br and I. It is very useful in pharmaceuticals for this reason. It is the force field that was used in this study to observe the water and Imidazole components of the Zinc complex. The functional form of GAFF is [4]

$$\mathrm{E}_{pair} = \sum_{Bonds} K_r(r - r_{eq})^2 + \sum_{Angles} K_\theta(\theta - \theta_{eq})^2 +$$

$$\sum_{Dihedrals} \frac{V_n}{2}[1 + cos(n\phi + \gamma)] + \sum_{i<j}[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}](1)$$

The energy is calculated by taking a simple harmonic energy for bond extension indicated by r and r_{eq} where the latter is the bond extension at equilibrium with a generic hookian "K" constant, angle change(θ and θ_{eq} used in a similar manner to the other *eqterm*) and dihedral angle(ϕ torsion angle). The last term is the sum of the Van der wals interactions between the atoms(these corresponding to the first two quantities in this sum; a R^{-12} and R^{-6} factor scaling) and a simple coulombic interaction(last quantity in the sum with a generic interaction between atom i and atom j at a set distance with added constant ϵ). Parameters corresponding to certain organic molecules are then collected and stored according to this model, either by experiment or calculation and GAFF had carved out quite a niche for itself.

III.2 ZAFF[5]

ZAFF, also known as the Zinc Amber force-field was tailor made specifically for the purpose of storing parameters for zinc compounds. This force field(with a mix of parameters such as bonds, angles etc calculated from ab initio and experiment) ZAFF uses the amber MCPB(metal center protein builder tool) to create its models. In brief, the metal center and its ligands are split from the PDB("protein databank file" which stores the crystallographically obtainted data of the coordinates, atoms and bonds involved in a protein but without hydrogens as these cannot be resolved well hence are added in later) into separate files. From these are made inputs for the software GAUSSIAN, a software that predicts spectroscopic properties, force-field parameters etc. This then performs quantum mechanical calculations for force constants and the like. The library of parameters was simply employed in this report.

IV. Density Functional Theory[6]

Density functional theory is a theory that describes the complete electronic structure of a molecule at a certain configuration, taking quantum effects into account. Its main tenets are described by the two Hohenberg-Kohn theorems which I shall elaborate upon.

IV.1 Theorem 1

The Hohenberg-Kohn theorems relate to any electronic system moving under the influence of an external potential $V_{ext}(r)$. The first theorem states that this external potential $V_{ext}(r)$ is a unique functional of the electron density p(r) i.e they have a direct numerical correlation which can be used in practice.

Consider the following scenario, we have two Hamiltonians with varying external potentials of the following forms

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext} \tag{2}$$

$$\hat{H}' = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext'} \tag{3}$$

These two different hamiltonians must also have two differing wavefunctions Ψ and Ψ'. The next step taken by Hohenberg and Kohn was to assume that both of these potentials, with varying wavefunctions and Hamiltonians, share the same electron density p(r).

It can then be shown by the variational principle that.

$$E < \langle \Psi' | \hat{H} | \Psi' \rangle = \langle \Psi' | \hat{H}' | \Psi' \rangle + \langle \Psi' | \hat{H} - \hat{H}' | \Psi' \rangle \tag{4}$$

$$E < E' + \langle \Psi' | \hat{T} + \hat{V}_{ee} + \hat{V}_{ext} - \hat{T} - \hat{V}_{ee} - \hat{V}'_{ext} \rangle \tag{5}$$

by shifting some components around,this can be changed to

$$E' < E - \int p(r)[V_{ext} - V'_{ext}]$$

(6)

integral of the density over all space associated with the potentials is just the energy, we obtain

$$E + E' < E' + E \tag{7}$$

This is obviously contradictory and means that two different densities cannot have the same external potential. This means we can focus on electron density. We then have a form composed of terms related to the configuration of the system and terms that would be true regardless

$$E(p) = \int p(r)V dr + T(p) + E_{ee}(p) \tag{8}$$

These last two terms can be taken to be a functional, i.e

$$E(p) = \int p(r)V_{ext} dr + F(p) \tag{9}$$

Since the first terms are dependent on the system, this means if we know the functional and its form we can calculate exact properties of the system. Unfortunately we do not know its form and there are workaround to this.

IV.2 Theorem 2

The groundstate energy can be obtained variationally and any physical quantity must be a property of density. We have already seen above how we would integrate a functional into an equation for the energy. Remember

$$F(p) = \langle \Psi | \hat{F} | \Psi \rangle \tag{10}$$

$$E(p(r)) = \int pV_{ext}(r) dr + F(p(r)) \tag{11}$$

Using the variational principle we have the following, where the prime notation $'$ denotes another density that is not the ground state version

$$\langle \Psi' | \hat{F} | \Psi' \rangle + \langle \Psi' | \hat{V}_{ext} | \Psi' \rangle > \langle \Psi | \hat{F} | \Psi \rangle + \langle \Psi | V_{ext} | \Psi \rangle \tag{12}$$

substituting these terms for what we have already determined we obtain

$$\int p'(r)V_{ext}(r)dr + F[p'(r)] >$$

$$\int p(r)V_{ext}(r)dr + F[p(r)] \tag{13}$$

Meaning

$$E[p'(r)] > E[p(r)] \tag{14}$$

i.e any trial density attempted will be greater than the true one, so the true functional exists and can be obtained by the functional that minimises the electron density. This is a step in the right direction but it left people with questions on how to use this theory in practice since the functional is by nature unknown.

IV.3 Kohn-Sham DFT[7]

Kohn and Sham proposed a more easily solvable problem. They proposed to work with a system of non-interacting electrons in such a manner that their density is the same as the interacting ones. This results in a cleaner and more manageable schrödinger equation and Hamiltonian

$$H_{eff} = \sum_{i=1}^{N_{electrons}} \frac{-1}{2}\nabla_i^2 + \sum_{i=1}^{N_{electrons}} V_{ext}(r_i) + \sum_{i=1}^{N_{electrons}} V_{av}(r_i) \tag{15}$$

This V_{av} is the "average" potential which replaces the direct interaction by a one-electron operator. The Hamiltonian can then be written as a sum of one electron interactions.

$$\hat{H} = \sum_{i=1}^{N_{electrons}} \hat{h}(r_i) \tag{16}$$

therefore,

$$\hat{h}(r_i)\Psi(r_i) = \epsilon\Psi(r_i) \tag{17}$$

More detailed mathematics may be found in the appendix as this is adjacent to the point. The following flowchart gives us a way to apply the Kohn-sham method through iteratively replacing trial densities until a self consistent energy minimisation is obtained.

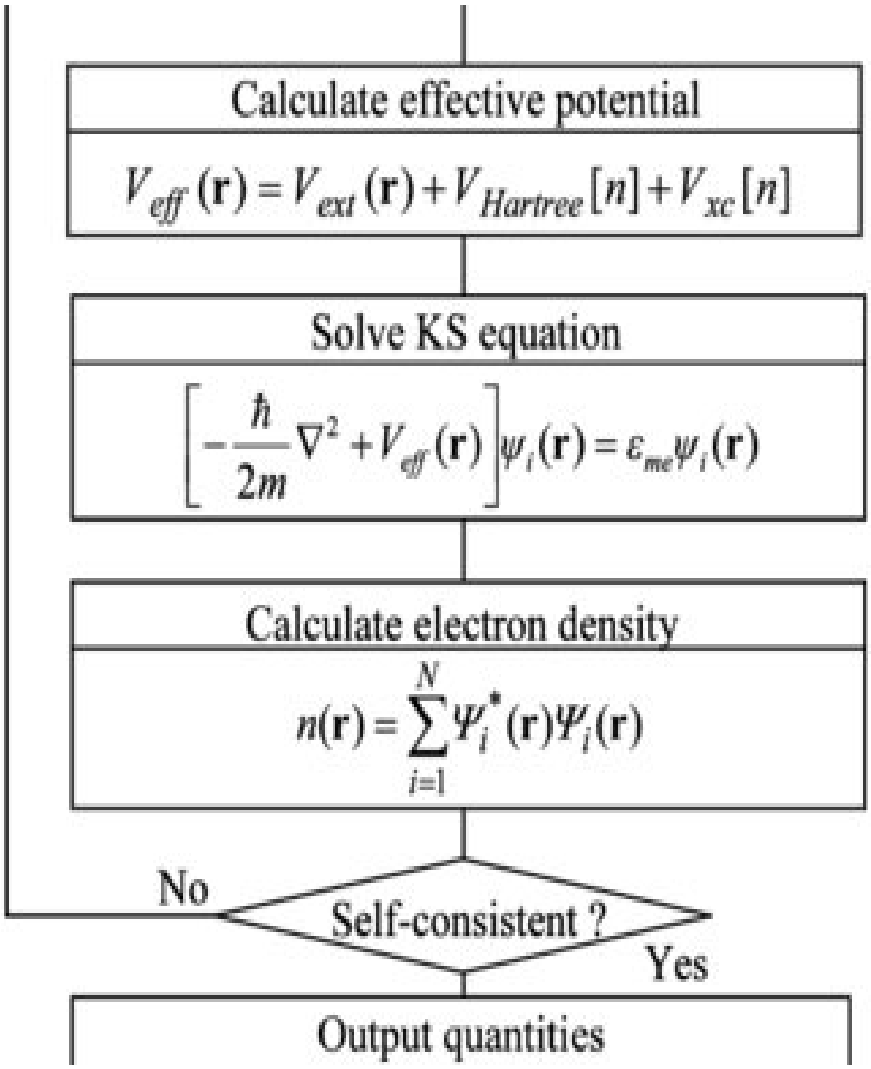


Figure 4: Kohn-Sham method flowchart[8]

The iteration of trial densities through this sequence of steps is repeated until the equations are self consistent, at which point we can now calculate energies, forces and stresses. This is a very useful and accurate method. While it is not an exact functional, it is very close to one and so suits most purposes we would have.

V. Machine learning

V.1 Machine learning logic

Machine learning is revolutionising many areas of science as we know them and the field of molecular dynamics is no exception to this. The approach allows one to use existing data, whether from experiment or a previous simulation to create a model for the behaviour of a system of interest. The size of the training set and the relevance of it(i.e if it is varied

enough to be useful) is particularly important. A training set of data values is compiled from existing data, followed by a test set(verifying the new model produces results similar to the data fed into it), and a validation set(fine tuning parameters). The workflow for such a process is like such.

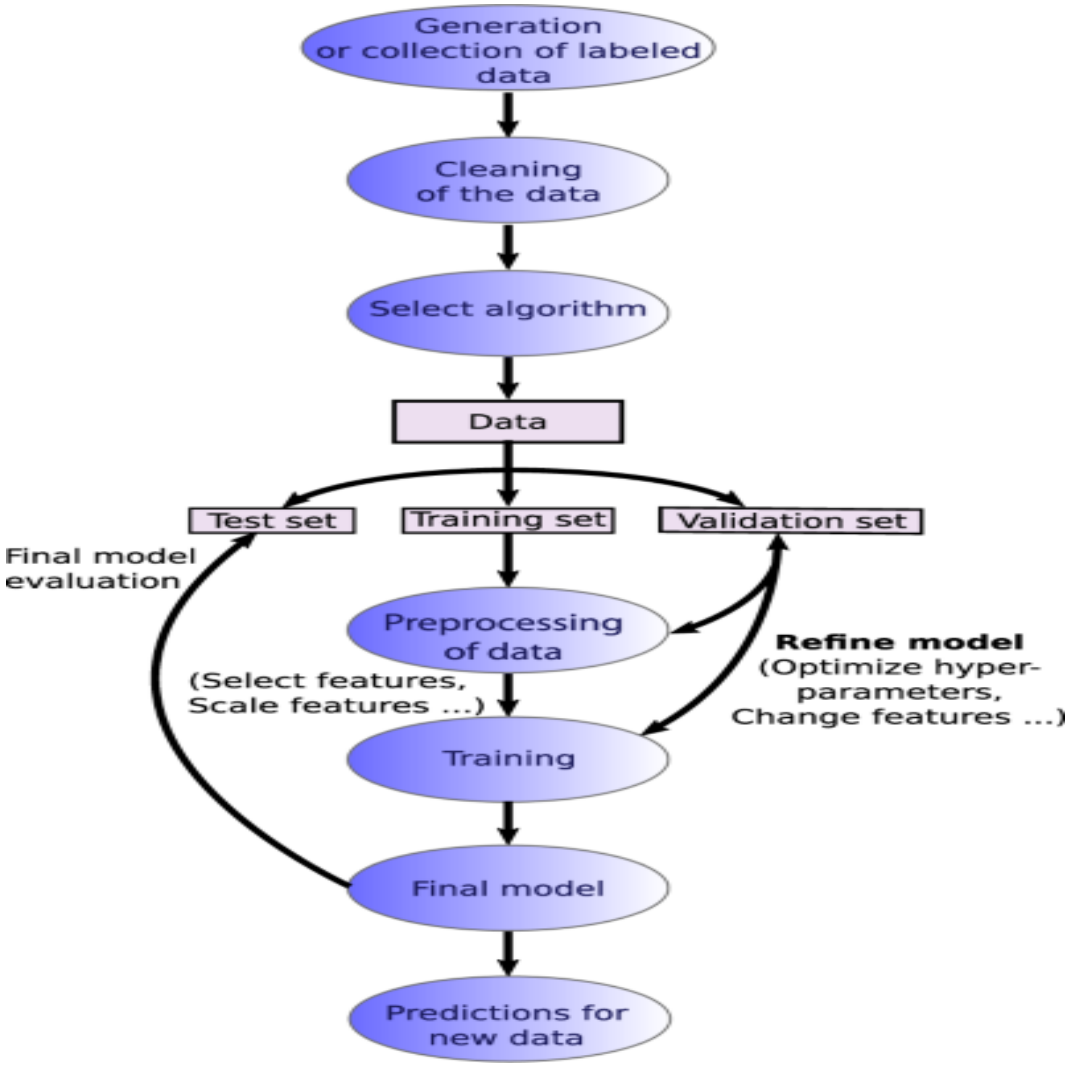


Figure 5: Machine learning basic workflow[9]

The operation of this will become clearer once I explain the mathematics of the SNAP Spectral Neighbor Analysis Potential.

VI. SNAP

SNAP or spectral neighbor analysis potential is a machine learning model used for interatomic potentials. The local environment of each atom is characterized by a set of bispectrum

components of the local neighbor density projected on to a basis of hyperspherical harmonics in four dimensions. These bispectrum components or coefficients represent the behaviour of the system being studied but invariant to rotation and translation.

The reason for this is that one could have a group of atoms translated 2 units to the left but that system would fundamentally still be the same corresponding to its neighbors and would therefore be a redundant configuration. Or, say we rotate a molecule by 30°. If everything in the system is rotated, the energy will remain the same.

VI.1 The bispectrum[10]

We will begin by describing an equation that calculates the density of atoms around a focal one at location r. These discrete points can be thought of as delta functions

$$p_i(r) = \delta(r) + \sum_{rii'<R_{cut}} f_c(r_{ii'}) w_{i'} \delta(r - r_{ii'}) \tag{18}$$

The value $r_{ii'}$ joins the neighbor atom i' to central atom i, the w value is a weight, and the sum is until a cutoff radius R_{cut}.It is then possible to take the angular component of density and create a spherical harmonic equation. We may use the projection of a 3d sphere onto 4d spherical harmonics like such

$$p(r) = \sum_{j=0,1/2}^{\infty} \sum_{m=-j}^{j} \sum_{m'=-j}^{j} u^j_{m,m'} U^j_{m,m'}(\theta_0, \theta, \phi) \tag{19}$$

summing over -j to j can be thought of as rotating around the sphere. We have the three angular terms denoted by a 4d hypersphere while u and U are functions like such

$$u^j_{m,m'} = U^j_{m,m'}(0,0,0) + \sum_{r_{ii'}} f_c(r_{ii'}) w_{i'} U^j_{m,m'}(\theta_0, \theta, \phi) \tag{20}$$

more detailed mathematics is included in the appendix(see description of spherical harmonics and clebsch gordan coefficients), assuming we can add clebsch-gordan coefficients for rotations of a 2d sphere in a triple scalar product we obtain the following

$$B_{j_1,j_2,j} = \sum_{m_1,m_1'=-j_1}^{j_1} \sum_{m_2,m_2'=-j_2}^{j_2} \sum_{m,m'=-j}^{j} (u^j_{m,m'}) H^{jmm'}_{\substack{j_1 m_1 m_1' \\ j_2 m_2 m_2'}} u^{j_1}_{m_1,m_1'} u^{j_2}_{m_2,m_2'} \tag{21}$$

The energy can then be taken as a function of these coefficients in a linear form like such

$$E_{SNAP}(r^N) = N\beta_0 + \beta \cdot \sum_{i=1}^{N} B^i \tag{22}$$

the k vector of snap coefficients is β with β_0 being a constant energy contribution given by n atoms. β_i is then the K vector of coefficients for a given atom "i". This equation is rather simple and easily interpreted by a computer. From this we can gain many properties of the system such as force which is just the negative gradient of energy, which(as seen in figure 3) leads to a host of other properties by association.

VII. Software packages used

VII.1 LAMMPS

LAMMPS is an open source code used to simulate classical molecular dynamics, the theory of which we have covered. It is the main package which will be used in this study for simulation.

VII.2 ORCA

ORCA is a quantum chemistry program used for ab initio calculations such as DFT. It was developed in the research group of Frank Neese and will be the quality test of this report. Any data obtained from simulation software that finds a workaround to DFT will have an error against DFT as the benchmark.

VII.3 FITSNAP

FITSNAP is a fortran code developed by researcher Alessandro Lunghi(the supervisor of this project). Its purpose is to find a set of SNAP bispectrum coefficients describining an input set of energies and configurations fed to it. The number of snap coefficients can be varied but the number used in this report is 56, this is a reasonable number of coefficients to describe a system.

III. Computational method

I. compilation

All programs used in this study(such as ORCA, LAMMPS and FITSNAP were compiled beforehand and detailed documentation can be found on their respective parent websites).

II. Studying the efficacy of AMBER upon the simulation of Water and Imidazole

II.1 Cutting the necessary structures

A pdb file of the 1CA2(Human carbonic anhydrase II protein) was downloaded. This file was subsequently cleaned, and from it were extracted the organic components of the zinc active-site, i.e water and imidazole. These were extracted as simple xyz files retaining the coordinates within the original. A visualisation of the two resulting structures can be observed below.

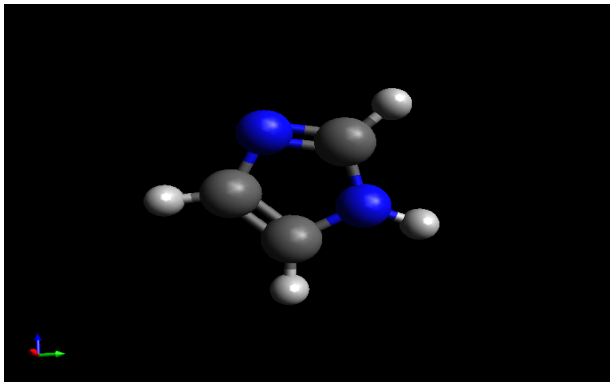


Figure 6: Imidazole xyz visualised structure in Avogadro(blue atoms are nitrogens, grey are carbons and white are hydrogens)

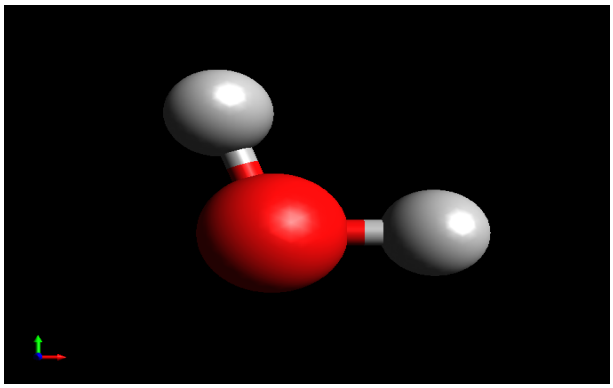


Figure 7: water xyz visualised structure in Avogadro(red atoms are Oxygen and white are hydrogens)

II.2 Creating a force field with AMBER

The next step was to superimpose the GAFF force-field onto these molecules using AMBER. Firstly, the files elaborated upon were put into antechamber, the AMBER file conversion program, and converted to mol2 files with charge calculated in the bcc format. Following this, an .frcmod file of the molecule was made. This file details whether any bond, angle or dihedral parameters are missing and tries to fill in the blanks.

Retaining this file for use, the AMBER LEAP console was opened. This is an interactive environment for setting up force fields and simulations in AMBER beyond simple file conversion. The GAFF force field source was provided to the console, the "check" command was then used to verify this functioned. With this done, the water or imidazole molecule could be converted into Topology(.top) and coordinate files(.crd). The energy landscape and the coordinates of

atoms within the system.

III. Amber2lammps tool

To run detailed molecular dynamics simulations on these force fields that were created in LAMMPS, it was necessary to use the dated but still reliable tool "Amber2lammps.py"[11] tool. This converted the two files generated(per molecule) in the previous step into a lammps datafile which can be read with ease. As some final house cleaning, the box size of the lammps datafile had to be increased to 100x100x100 to make the box boundless, this removes periodicity effects.

III.1 Energy minimisation and the NVT ensemble

Before running any worthwhile simulations upon these structures, it was necessary to minimise their energy i.e get the system to a configuration with the lowest and most stable energy possible so that we can observe real behaviour and not a chaotic cauldron of atoms.

There are multiple parameters to input to a lammps script before running a simulation which I will specify here and for future reference. The units "real" were specified for the simulation dump values i.e distance in angstroms, energy in kcal/mol(see documentation). The "full" atom style was used to include bonds, angles, dihedrals and impropers into our force field. These styles were given a harmonic dependence as in the GAFF equation. The pair style sets the formula for pairwise interactions in a molecule. The pairstyle utilised was lj/cut/coul/cut with a cutoff radius of 10.0 angstroms(a lennard jones potential, with a coulombic term at a certain cutoff distance). A cg "conjugate gradient" style of minimisation was employed.

The results of minimisation were saved to a separate datafile which was run into an NVT ensemble. In the NVT ensemble, a thermostat is set up, and the energy fluctuations of a system are monitored with the number of atoms, volume and temperature of the system kept constant. In reality there will be some slight fluctuation of the temperature also, depending on the number of atoms $\propto 1/N$. The temperature of the simulation was set to be 298.15K with 3 thermostats, a timestep of 0.5fs per simulation step and a Tdamp value of 100fs(the time which it takes to relax the temperature).

The simulation was run for 500ps, dumping xyz coordinates and potential energies to a lammps dumpfile every 0.5ps for further analysis of 1000 discrete datasets. Some run of the mill file parsing and python graphing was done to graph energy(kcal/mol) vs time(ps) which could be used in comparison with DFT.

III.2 Running DFT calculations

With the xyz snapshots from molecular dynamics in mind, I could perform DFT calculations.

Using the dft software ORCA, I calculated the single point energy of each configuration(corresponding to each timestep) in a batch job. It should be noted that for all DFT calculations done in this work, the def2-TZVPP basis set was used with a PBE functional. This involved doing a single point calculation and looping over it 1000 times, storing the results in different files in one folder, and using "grep" to extract the final single point energy from this collection of files. The energy values were then stored in a text file and converted to kcal/mol.

From this point it was a simple matter of graphing the energies of dft in kcal/mol vs the energies obtained using the GAFF force field, using the DFT values as the state of the art for accuracy. This process was repeated for both molecules. The aim here is to test if AMBER is sufficient for organic molecules.

III.3 Training a model using Fitsnap

The next aim was to compare the efficacy of machine learning to the results previously obtained. To do so, the DFT energies and their corresponding molecular energies were used as a training set.

Alessandro Lunghi's fortran fitsnap code was compiled on my device and I subsequently inputted the dft energies and configurations into fitsnap, with a simple lammps datafile specifying the SNAP pair style as well as the atoms to which I was fitting a potential. This was done for 1000 configurations(1000 for water and 1000 for imidazole, though the accuracy peaked at about 800 configurations which lines up with his previous works).

Fitsnap has an inbuilt function allowing one to calculate the accuracy of the model against the training set. Two columns of data are returned to the user, one being the training set input energies for a certain configuration and the other column being fitsnap's predictions according to its bispectrum components. These could be and indeed were plotted against each other for the two respective models to obtain the RMSE in energies and the correlation.

IV. Creating a test set

While this is a fair test of accuracy, it is more rigorous to create a "test set" to make sure no overfitting has occurred. Overfitting is when a machine learning model is a good fit for the training data it was fed but unrealistically so, so it doesn't mimic reality. An example of overfitting would be fitting a many termed polynomial to a straight line behaviour data. It fits the data but is a poor representation.

A test set was made by taking the minimised structure of either imidazole or water, and running a simulation in the NVT ensemble like before but changing the pair style to a SNAP potential, and the atom style to atomic(this removes bonds, angles and dihedrals terms since the bispectrum coefficients are now representative of the system). The data for this simulation

was extracted with a similar timescale to the training set, and as before, DFT calculations were run upon the configurations and compared to the potential energies calculated by SNAP. The rest came down to simple graphing of these values and calculation of the RMSE energy to determine if the test set was accurate.

V. Studying the entire zinc complex

Now that comparisons have been made between SNAP, DFT and molecular dynamics on the components of the zinc coordination complex, I was able to move onto simulating the entire complex. An image of the 1ca2 active site was supplied earlier in this paper, and that is what we shall be working with. There were 30 atoms in total, being a mixture of Nitrogen, Carbon, Hydrogen and Oxygen.

V.1 Using random displacements on the complex and calculating energies using DFT

With the initial pdb file of the zinc active site created, a sample of 200 configurations was made for testing using the random displacements method. This is a very simple method that involves shifting each xyz coordinate in the active site xyz file by a random number between -0.1 and 0.1 angstroms. This allows good enough deformation that we can analyse different energy configurations but not so much so that the system no longer makes sense.

The energies of these configurations were then calculated using a batch script as before in ORCA. We now have a training set with quantum levels of accuracy for fitsnap. Once again, the configurations and energies were run through fitsnap and the RMSE of energies calculated with the model was measured against the actual values. However, since DFT scales poorly with the number of atoms, it was infeasible to gain more DFT data upon the hardware provided. Due to this quandary the test set is merely a set of 50 random configurations from the 200 total.

V.2 Setting up the ZAFF force field upon the entire complex

The next step was to setup the ZAFF force field. This is a very similar process to that of setting up GAFF bar sourcing some ZAFF parameter files and specifying ligand connections to the central zinc atom and using some values particular to carbonic anhydrase in the zaff library so there is not too much to go into.

The topology and coordinate files were generated and passed through Amber2lammps.py as before to obtain a lammps datafile. The aim here is to determine how current approaches to analysing the zinc complex compare to the state of the art(but demanding) DFT.

To do this, a lammps script with harmonic dependencies and the lj/cut/coul/cut pair style was set up, but ran for 0 steps(effectively just calculating the

energy of whichever configuration is fed into it and dumping that value). This script with the implemented force field was run upon all configurations extracted from random displacements and the potential energy(according to ZAFF) was calculated. This was then graphed against DFT , extracting the RMSE in energies between the two methods and the correlation.

V.3 Training the coordination complex model and attempting to combine the Amber force fields with machine learning

After the comparisons between multiple approaches to simulating bio-inorganic compounds were made, an attempt was made at a novel approach; combining the AMBER force field ZAFF with SNAP.

In theory this would mean that SNAP does not have to calculate coefficients to represent the entire behaviour of the system since the force-field will aid in this. It should be a way of error correcting SNAP to be even more accurate. This was done by inputting the force-field lammps datafile into Fitsnap, as opposed to the generic "atomic" style one normally used. It was then necessary to specify a hybrid snap lj/cut/coul/cut potential to be fit. The AMBER special bonds function was also turned off to stop this from interfering. Once these steps were carried out, the 200 zinc complex configurations and their energies were used as a training set with this. RMSE and correlation were extracted.

IV. Results and Discussion

I. AMBER GAFF vs DFT

I.1 Water

The first test involved seeing if GAFF(the organic parameters force-field library) was sufficient for the organic components of the system individually. We shall examine the case of water first.

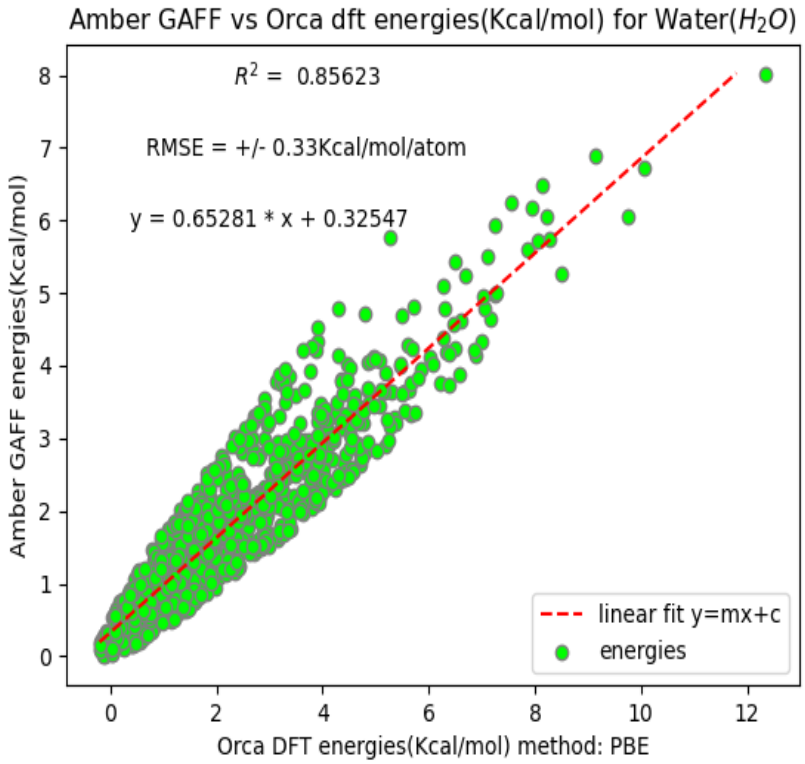


Figure 8: AMBER GAFF energies(kcal/mol) calculated from a set of 1000 md configurations at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for H_2O

As can be seen that GAFF calculated potential energies have a deviation of $\sim$ 1 kcal/mol, a linefit with a slope of 0.65281 with a scaling correction between the two energies of 0.19490. The R^2 correlation value of $\sim$ 0.86 lets us put reasonable faith in this fit. This is a decent correlation with DFT and mimics most of the energy behaviour observed(though it can be seen to fall off in that regard as higher energies are achieved). The RMSE value of 0.33 kcal/mol/atom is in good agreement with the average energy deviation of GAFF from

reality which is 0.69 kcal/mol averaged over 55 compounds[13]. Since that is a large sample collection and I dealt with a very simple molecule, my resulting error was lower.

I.2 Imidazole

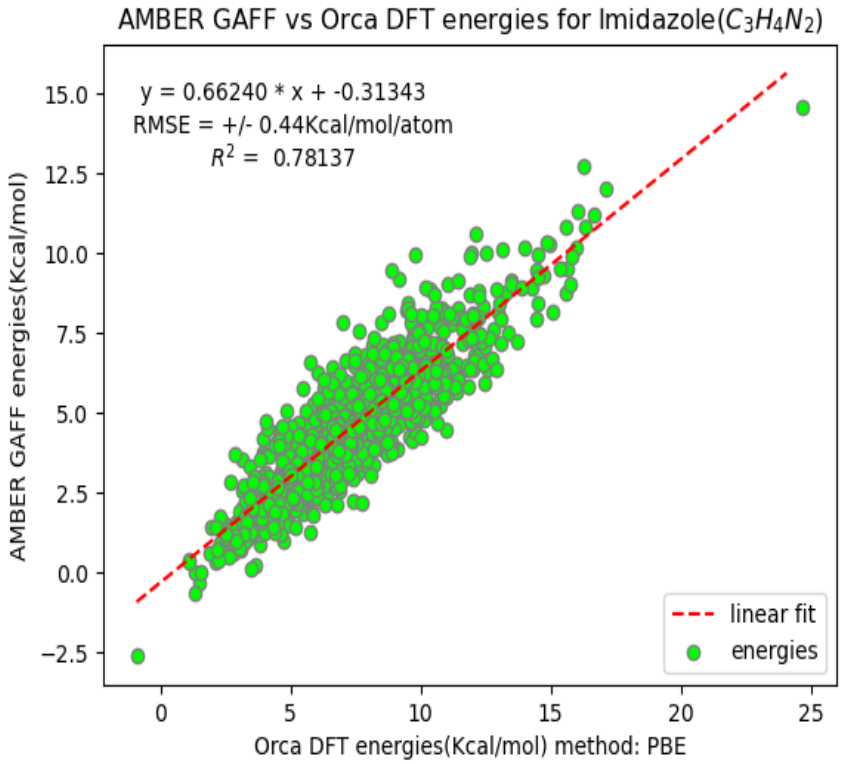


Figure 9: AMBER GAFF energies(kcal/mol) calculated from a set of 1000 md configurations at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for $C_3H_4N_2$

A slightly higher error and slightly lower R^2 value than the case of water is obtained here but that is likely due to the system being more complicated, mimicing the behaviour of 9 atoms as opposed to a mere 3 atoms.

It is clear that the generalised amber force-field is sufficient for the purpose of modelling water and imidazole within a reasonable degree of accuracy which lines up with literature, however, there are more accurate approaches(nearing DFT levels of accuracy) inspired by machine learning.

II. SNAP VS DFT

II.1 Training set and test sets

II.2 Water

Using Lunghi's FITSNAP code as discussed in the computational method section, we obtain a graph of the training input data vs the energies predicted by the model we have created. This is a crude initial test but allows the researcher to make further steps confidently.

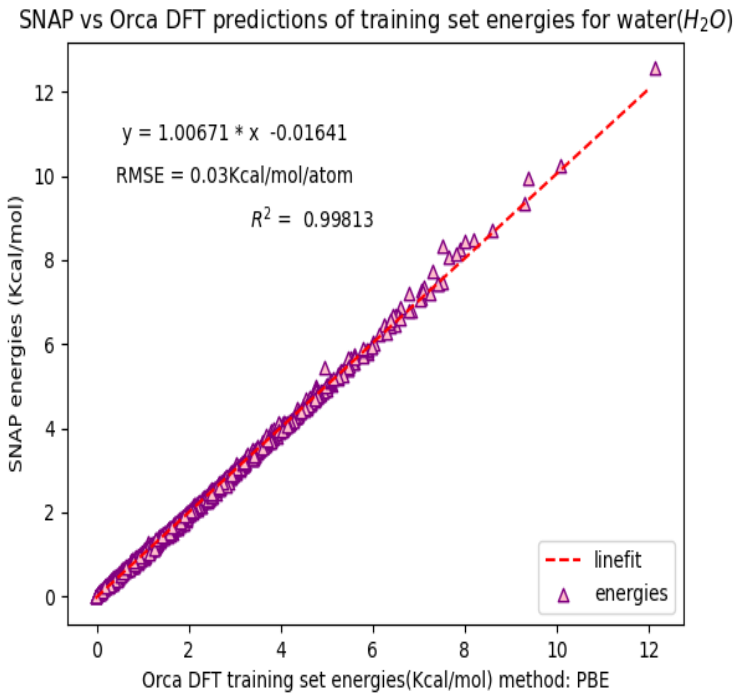


Figure 10: SNAP predicted energies(kcal/mol) calculated for a training set of 1000 md configurations at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for H_2O

As can be seen from the figure above, the fitsnap algorithm predicted data for the input training set configurations is nearly a perfect match with an RMSE of 0.03Kcal/mol/atom(far less than that exhibited by GAFF) and a near perfect slope and R^2 value with respective values of 1.00671 and 0.99. The model performs well against the training data but a test set with a separate collection of data was also made to make sure this model performs well in all situations.

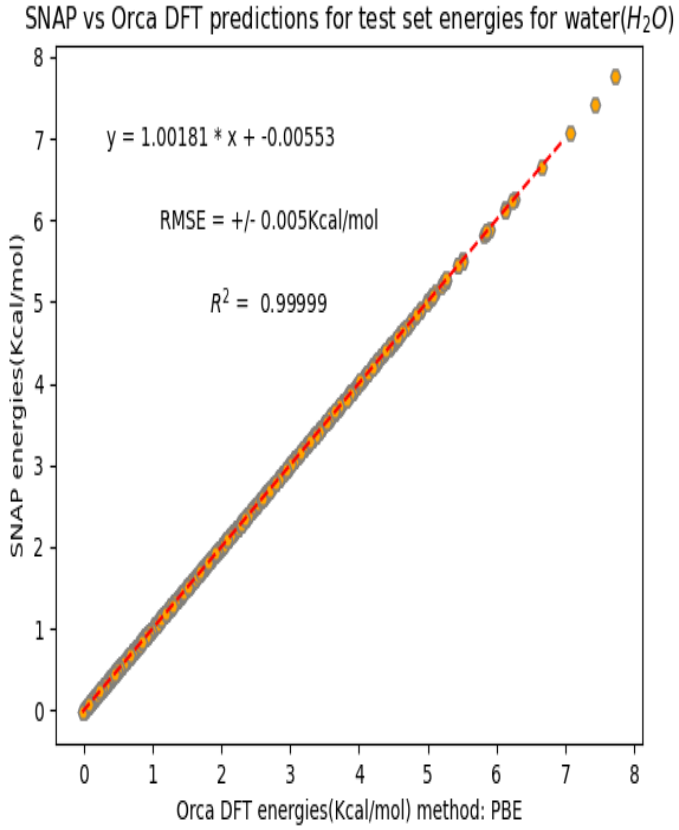


Figure 11: SNAP predicted energies(kcal/mol) calculated for a test set of 1000 md configurations(a separate batch from those used in training) at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for H_2O

The test set performed quite

well, having near perfect values in all categories as seen above. This shows the effectiveness of SNAP as a model(at least for the case of the simple 3 membered water molecule).

II.3 Imidazole

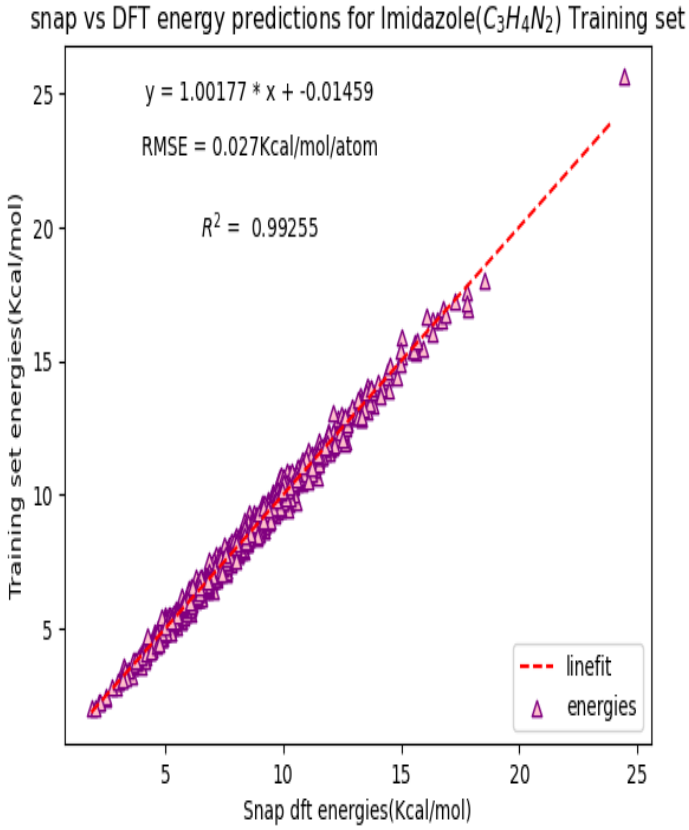


Figure 12: SNAP predicted energies(kcal/mol) calculated for a training set of 1000 md configurations at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for $C_3H_4N_2$

The training set for the imidazole molecule is modelled quite well by SNAP coeffiients. If we then turn to a test set of this in the NVT ensemble of 400 more configurations.

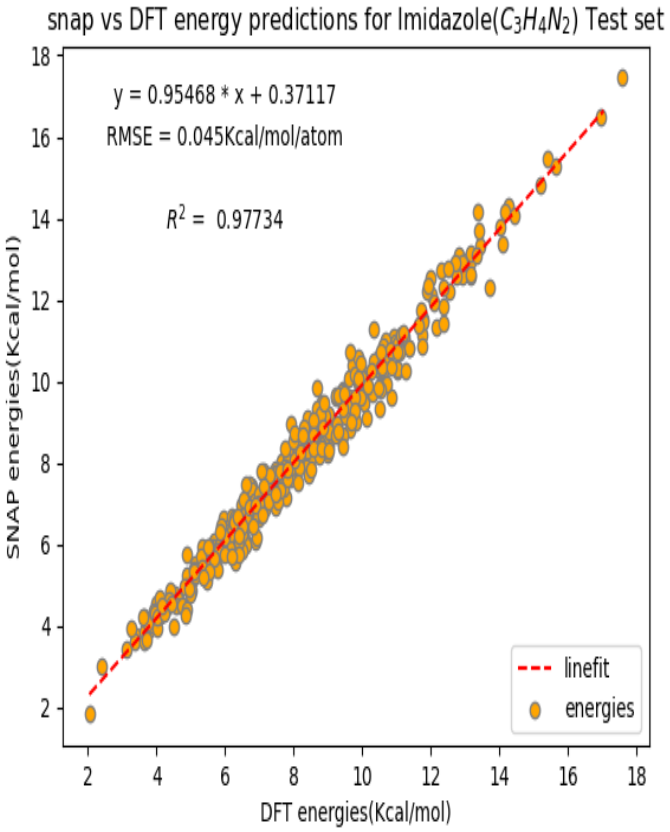


Figure 13: SNAP predicted energies(kcal/mol) calculated for a test set of 400 md configurations(a separate batch from those used in training) at intervals of 0.5ps in the NVT ensemble vs the DFT energy calculations of those same snapshots for $C_3H_4N_2$

The test set exhibited here also performs well with a high r-value, good linefit correlation and a low rmse though somewhat higher than that of the training set. This is likely due to minor changes in the system once a new simulation is begun, so the coefficients are ever so slightly off.

III. Comparison of the two modelling methods

IV. SNAP performance vs DFT Zinc complex

IV.1 Training set Zinc complex

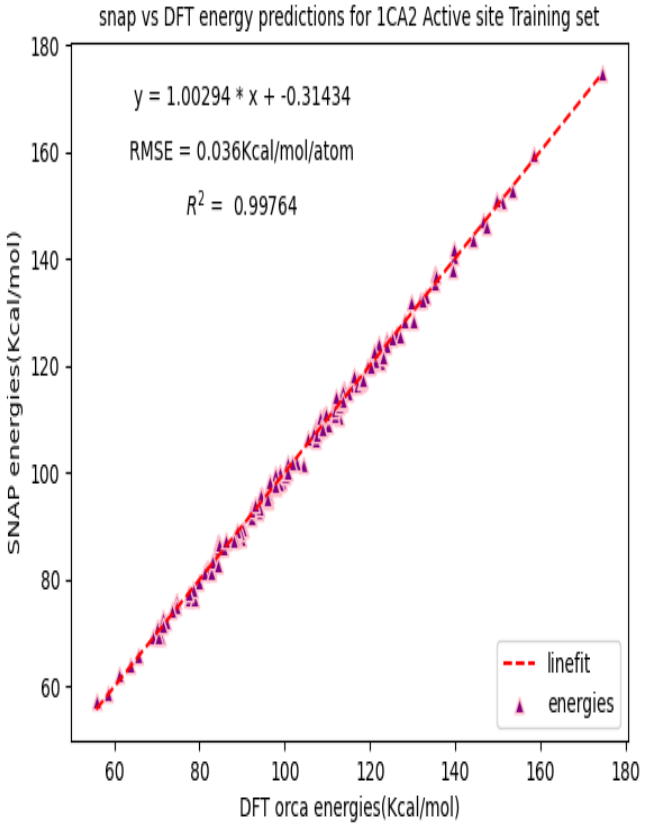


Figure 14: SNAP predicted energies(kcal/mol) calculated for a test set of 150 random displacement configurations vs the DFT energy calculations of those same snapshots for the 1CA2 active site

The training set for the Zinc complex performed quite well with a slope of nearly 1, strong R^2 value and an RMSE of 0.036Kcal/mol/atom which is an order of magnitude less than that of DFT for the other two samples though molecular dynamics tests were run on the complex too which shall be discussed.

IV.2 Test set Zinc complex

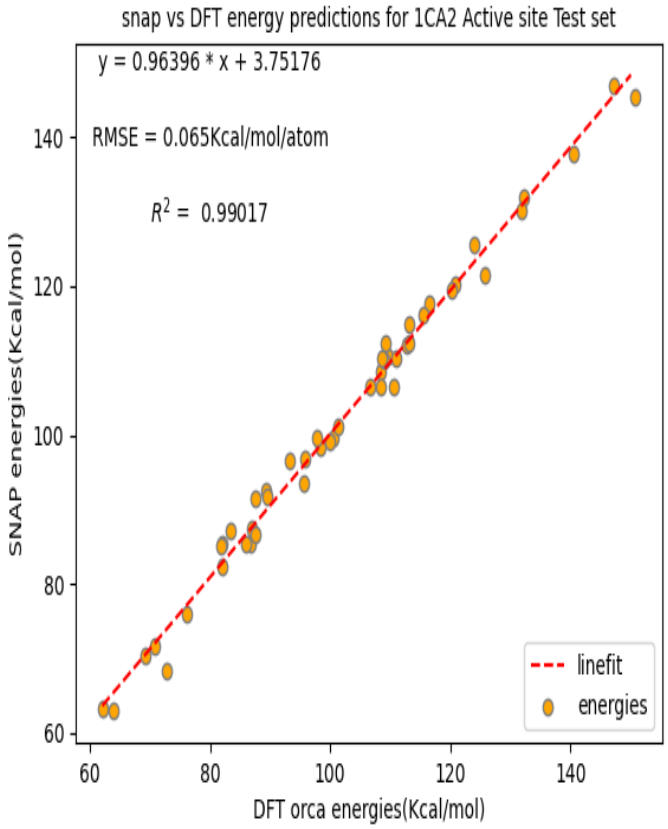


Figure 15: SNAP predicted energies(kcal/mol) calculated for a test set of 50 random displacement configurations vs the DFT energy calculations of those same snapshots for the 1CA2 active site

The test set of 50 samples from the 200 random configurations generated also produced good results that allow us to have a reasonable level of certitude that the results ring true. Slightly higher RMSE on this

sample since ridge-regression(slightly) in an effort to mitigate overfitting a small set.

V. ZAFF performance vs DFT

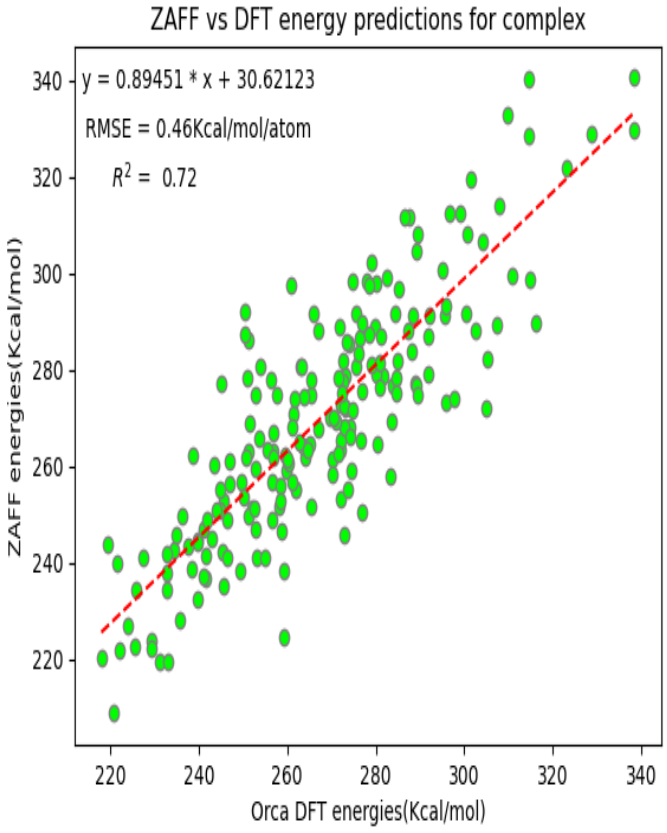


Figure 16: SNAP predicted energies(kcal/mol) calculated for a training set of 150 random displacement configurations vs the DFT energy calculations of those same snapshots for the 1CA2 active site

While the ZAFF force-field calculates the energy to a reasonable level, it has a linefit correlation of around 0.89 with an RMSE of 0.46kcal/mol/atom. This is an order of magnitude higher than that exhibited by SNAP. Granted the test set wasn't ideal for the 1CA2 active site in that it was simply sliced from the random configuration energies and not produced from molecular dynamics simulations as would be done properly due to time constraints.

The fitsnap code outperforms its competition on a protein active site simulation which can be of interest to the future student/researcher looking to upscale these ideas to larger simulations with more accurate and also low computational cost results. This is a proof of concept that machine learning can be used to model inorganic metallic complexes in proteins for better efficiency than current systems.

VI. integrated ZAFF-SNAP force-field attempt

An attempt was made to combine existing force-field library ZAFF with the SNAP coefficients to correct SNAP so that it can operate at even greater efficiency. However, the results were unsatisfactory, with fitsnap producing random numbers in a trial run.

A graph is not included as it would be nonsensical but energy values exploding are a red flag. There are some a few ways in which this could have happened. Firstly, there is the obvious suspect of human error in assigning charges, angles, bonds and dihedrals in AMBER and converting to LAMMPS. With the previous graph detailing ZAFF's performance, this is unlikely. It could also be a technical error which I overlooked when applying the hybrid pair

style to the lj/cut/coul/cut and SNAP styles. Due to time constraints in the project runtime, further code debugging was left to future work. Some suggestions for further work in this avenue would be to first of all use a larger training set for the zinc complex(150 configurations were used in the experiment), to make sure that LAMMPS is indeed properly attempting to read SNAP coefficients and ZAFF combination.

V. Conclusion

In conclusion,it was found that pre-existing molecular dynamics approaches to simulating small organic molecules do function but have a fundamental discrepancy with ab initio calculations which limits their accuracy. This is not the case for SNAP which mimics them well in training and test sets.

Upon upscaling this to the Zinc active site of Human Carbonic anhydrase II it was found that the machine learning approach to training data was had an error an order of magnitude lower than that of molecular dynamics which agrees quite well with tests upon the smaller water and imidazole molecules. Finally, a novel test was performed upon the zinc complex in an attempt to combine force-fields and machine learning. While this did not function correctly, it is clear there are many fruits to be wrought from using machine learning in molecular dynamics simulations which beat out their competitors and warrant further study.

References

[1] He Song†, David L. Wilson†, Erik R. Farquhar‡, Edwin A. Lewis†, and Joseph P. Emerson†,* †Department of Chemistry, Mississippi State University, Mississippi State, MS 39762, USA [*"Revisiting Zinc Coordination in Human Carbonic Anhydrase II"*] page 2

[2] He Song†, David L. Wilson†, Erik R. Farquhar‡, Edwin A. Lewis†, and Joseph P. Emerson†,* †Department of Chemistry, Mississippi State University, Mississippi State, MS 39762, USA [*"Revisiting Zinc Coordination in Human Carbonic Anhydrase II"*] Figure 1

[3] Adam Hospital1 Josep Ramon Goñi2,3 Modesto Orozco1–4 Josep L Gelpí2–4 1 Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology [*"Molecular dynamics simulations: advances and applications"* Figure 3 page 39 Dove Press Journal 19 November 2015]

[4] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. 25, 1157 (2004). [*"Development and testing of a generalised AMBER force-field"*.

[5] Martin B. Peters, Yue Yang, Bing Wang, La´szlo´ Fu¨sti-Molna´r, Michael N. Weaver, and Kenneth M. Merz, Jr.* J. Chem. Theory Comput. 2010, 6, 2935–2947 [*"Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF)"*]

[6] P. HOHENBERGt Ecole Xornzale Superzeure, I'aris, France AND W. KonNt Ecole Xonnale Superieure, I'aris, Prance and I'aculte des Sciences, Orsay, France and University of Calzfonia at San Diego, La Jolla, Calzfornia (Received 18 June 1964)[*Inhomogenieous electron gas*] Physical review

[7] Kohn, Walter; Sham, Lu Jeu (1965). "Self-Consistent Equations Including Exchange and Correlation Effects". Physical Review. 140

[8] Eiji Nakamachi · Yasutomo Uetsuji · Hiroyuki Kuramae · Kazuyoshi Tsuchiya · Hwisim Hwang [*"Process Crystallographic Simulation for Biocompatible Piezoelectric Material Design and Generation"*Arch Comput Methods Eng (2013]

[9] Jonathan Schmidt 1 , Mário R. G. Marques 1 , Silvana Botti2 and Miguel A. L. Marques1 ["*Recent advances and applications of machine learning in solidstate materials science*"] Computational materials Figure 1

[10] A.P. Thompsona, , L.P. Swilerb , C.R. Trottc , S.M. Foilesd , G.J. Tuckerd ["*A Spectral Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials*"] Condensed matter and materials science 12 Sep 2014

[11] Keir E. Novik amber2lammps.py https://github.com/CFDEMproject/LAMMPS/blob/master/tools/amber2lmp/amber2lammps.py

[12] A E Eriksson 1, T A Jones, A Liljas ["*Refined structure of human carbonic anhydrase II at 2.0 A resolution*" 1988;4(4):274-82. doi: 10.1002/prot.340040406.]

[13] http://ambermd.org/antechamber/gaff.html Table VI

[14] http://resonanceswavesandfields.blogspot.com/2016/05/spherical-harmonics.html Table 2a Spherical Harmonics

[15] David J.Griffiths ["*Introduction to quantum mechanics 3rd edition*" chapter 4.1 spherical coordinates pages 132-136]

A. Appendix

I. Extra Theory

I.1 Spherical Harmonics

Spherical harmonics describe the angular momentum of a particle, and they require a slight background for this paper as Clebsch Gordan coefficients are involved in the bispectrum of SNAP coefficients. Laplace's equation for angular momentum is given as

$$\nabla^2 = \frac{1}{r^2}\frac{\delta}{\delta r}(r^2\frac{\delta}{\delta r}) + \frac{1}{r^2 sin^2(\theta)}\frac{\delta}{\delta\theta}(sin(\theta)\frac{\delta}{\delta\theta}) + \frac{1}{r^2 sin^2(\theta)}\frac{\delta^2}{\delta\phi^2}$$

The full derivation may be found elsewhere(see reference **[15]**) but for our purposes it is enough to know that it can be solved in a useful fashion dependent on the magnetic quantum number(m) and the angular momentum quantum number(l).

$$\gamma_l^m(\theta,\phi) = (-1)^m\sqrt{\frac{(2l+1)(l-m)!}{(4\pi)(l+m)!}}P_l^m(cos(\theta))e^{im\phi}$$

for l=0,1,2,3... and for m=-l,-l+1,...,l-1,l

Figure 17 on the next page details values for the spherical harmonics(depending on l and m) for spherical, cartesian and rotating forms for the first few terms. It is of interest to take the product of multiple spherical harmonics to determine properties of many different atoms. Spherical harmonics form an orthonormal basis set and can be written as

$$\gamma_1^0(\theta,\phi)\cdot\gamma_1^0(\theta,\phi) = c_0^0\gamma_0^0(\theta,\phi) + C_2^0\gamma_2^0(\theta,\phi)$$

i.e the product depends on the initial form of the function with two distinct terms multiplied each by a separate "c" value. This means that products of spherical harmonics have their own coefficients called clebsch-gordan coefficients which are related to figure 17 but distinct due to their matrix nature. This is written in the following form with distinct values of l and m forming matrix components

$$\gamma_{l1}^{m1}(\theta,\phi)\cdot\gamma_{l2}^{m2}(\theta,\phi) = \sum_{l,m}\sqrt{\frac{(2l_1+1)(2l_2+1)(2l+1)}{4\pi}}\times\begin{pmatrix} l1 & l2 & l \\ m_1 & m2 & m\end{pmatrix}\gamma_l^m(\theta,\phi)\begin{pmatrix} l1 & l2 & l \\ 0 & 0 & 0\end{pmatrix}$$

This logic is useful in creating bispectrum components which describe a molecule's behaviour based on properties such as its spherical harmonics.

	Spherical	Cartesian	Rotating
$\ell=0$	$Y_{\ell=0}^{m=0}(\theta,\phi)=\sqrt{\frac{1}{4\pi}}$	$\sqrt{\frac{1}{4\pi}}$	$\sqrt{\frac{1}{4\pi}}$
$\ell=1$	$Y_{\ell=1}^{m=-1}(\theta,\phi)=\sqrt{\frac{3}{8\pi}}\sin\theta e^{-i\phi}$	$\sqrt{\frac{3}{8\pi}}\frac{x-iy}{r}$	$\sqrt{\frac{3}{8\pi}}\sin\theta e^{i(-\omega t-\phi)}$
	$Y_1^0(\theta,\phi)=\sqrt{\frac{3}{4\pi}}\cos\theta$	$\sqrt{\frac{3}{4\pi}}\frac{z}{r}$	$\sqrt{\frac{3}{4\pi}}\cos\theta$
	$Y_1^{+1}(\theta,\phi)=-\sqrt{\frac{3}{8\pi}}\sin\theta e^{i\phi}$	$-\sqrt{\frac{3}{8\pi}}\frac{x+iy}{r}$	$-\sqrt{\frac{3}{8\pi}}\sin\theta e^{i(-\omega t+\phi)}$
$\ell=2$	$Y_2^{-2}(\theta,\phi)=\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{-2i\phi}$	$\sqrt{\frac{15}{32\pi}}\frac{(x-iy)^2}{r^2}$	$\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{i(-\omega t-2\phi)}$
	$Y_2^{-1}(\theta,\phi)=\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{-i\phi}$	$\sqrt{\frac{15}{8\pi}}\frac{(x-iy)z}{r^2}$	$\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{i(-\omega t-\phi)}$
	$Y_2^0(\theta,\phi)=\sqrt{\frac{5}{16\pi}}(3\cos^2\theta-1)$	$\sqrt{\frac{5}{16\pi}}\left(\frac{3z^2}{r^2}-1\right)$	$\sqrt{\frac{5}{16\pi}}(3\cos^2\theta-1)$
	$Y_2^{+1}(\theta,\phi)=-\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{i\phi}$	$-\sqrt{\frac{15}{8\pi}}\frac{(x+iy)z}{r^2}$	$-\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta e^{i(-\omega t+\phi)}$
	$Y_2^{+2}(\theta,\phi)=\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{2i\phi}$	$\sqrt{\frac{15}{32\pi}}\frac{(x+iy)^2}{r^2}$	$\sqrt{\frac{15}{32\pi}}\sin^2\theta e^{i(-\omega t+2\phi)}$

Figure 17: Table of Spherical Harmonics[14]