SystemDescriptor file looks like this:

#F_data = short two water molecules.x

#F_data = short three water molecules.x

F_data = datafile1 from github gaussian process.x

#F_data = datafile2.x

#F_data = datafile3 2 water molecules.x

#F_data = datafile4 3 water molecules small.x

#F_data = datafile5 3 water molecules big.x

&FEATURES

&SingleDistances

SingleDistancesInclude = True

&SingleDistancesDescription

&endSingleDistancesDescription

&DefaultSingleDistances

SinglePowers: -1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-11,-12,-13

&endSingleDistances

&DoubleDistances

DoubleDistancesInclude=True

&DoubleDistancesDescription

O,O,intermolecular: -1,-2,-3,-4,-5,-6,-12

O,H,intermolecular: -1,-2,-3,-4,-5,-6,-12

H,H,intermolecular: -1,-2,-3,-4,-5,-6,-12

O,H,intramolecular:

H,H,intramolecular:

&endDoubleDistancesDescription

&DefaultDoubleDistances

DoublePowers: -1

IncludeSameType=True

IncludeAllExcept=True

ExcludeAllExcept=False

&IncludeExcludeList

&endIncludeExcludeList

&endDoubleDistances

&Harmonics

HarmonicsInclude=False

Order: 0

Degree: 0

symbol of atom to be a center of coordinate system to calculate harmonics

HarmonicCenter: O

HarmonicAtoms: O,H

&HarmonicDescription

O,O,intermolecular:

O,H,intermolecular: -1,-2,-3,-4,-5,-6

H,H,intermolecular:

O,H,intramolecular:

H,H,intramolecular:

&endHarmonicDescription

&DefaultHarmonics

HarmonicPowers: -1,-2

IncludeHarmonicSameType=True

IncludeHarmonicAllExcept=True

ExcludeHarmonicAllExcept=False

&IncludeExcludeHarmonicList

O,O,intermolecular; O,O,intermolecular

&endIncludeExcludeHarmonicList

&endHarmonics

&endFEATURES

&SYSTEM

Atom symbol (string), Molecule number (integer)

water 1

O,0

H,0

H,0

water 2

O,1

H,1

H,1

water 3

#O,2

#H,2

#H,2

&end SYSTEM

First of all, if the row starts with #, means that this is the remark.

At the beginning of the file there is a row that tells which database will be used. So, this record:

F_data = datafile1 from github gaussian process.x

means that file **"datafile1 from github gaussian process.x"** will be read and for making features. This file contains coordinates and looks like:

O: 13.2736963426 0.5424091274 10.271002223

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 12.6025205979 7.3576907186 13.3829059556

H: 12.4797559031 6.4098908778 13.4031217506

H: 13.5921441038 7.4665297563 13.2363446942

-0.0001261346

O: 13.2736963426 0.5424091274 10.271002223

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 11.6493522548 4.5897400107 12.7706071762

H: 10.949899632 4.5025919429 12.0549258793

H: 12.2626069322 3.8146469879 12.4250954796

-0.0013091863

O: 13.2736963426 0.5424091274 10.271002223

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 11.8264475793 7.5185476749 10.665474538

H: 11.1629751238 6.8277280792 10.653608656

H: 12.103205506 7.6004340976 11.6409427301

-0.0001968029

Which also means that file corresponds to two water molecules database. For three water molecules file looks like:

O: 13.2736963426 0.5424091274 10.2710022230

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 11.6493522548 4.5897400107 12.7706071762

H: 10.9498996320 4.5025919429 12.0549258793

H: 12.2626069322 3.8146469879 12.4250954796

O: 12.1003363063 1.6953807987 12.8096626538

H: 12.6322035368 1.2866316478 13.4766486029

H: 12.3701292777 1.2661586991 11.9896754592

-0.0004989883

O: 13.2736963426 0.5424091274 10.2710022230

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 11.7617267881 3.5066913264 15.3906910015

H: 10.9315757615 3.1344340959 15.6711825255

H: 11.7202649141 3.7765123915 14.3761205727

O: 12.1003363063 1.6953807987 12.8096626538

H: 12.6322035368 1.2866316478 13.4766486029

H: 12.3701292777 1.2661586991 11.9896754592

-0.0000267345

O: 13.2736963426 0.5424091274 10.2710022230

H: 14.2373366527 0.6810463935 10.3989958099

H: 13.1201498642 -0.0755229832 9.5440248299

O: 10.3961444299 0.9153703426 7.6052632004

H: 9.4345719775 0.7175939514 7.3745577692

H: 10.6327890923 0.0879677203 8.0324057509

O: 10.3659284593 2.4120291904 9.9458901332

H: 10.4484057756 1.8828847342 9.0791225677

H: 9.7991482226 1.8600709810 10.5202120134

0.0009437256

Than SystemDescriptor file contains two main sections: &FEATURES and &SYSTEM.

&SYSTEM part starts with row:

&SYSTEM

And ends with row

&endSYSTEM

It describes molecular system. For three water molecules system, it looks like:

Atom symbol (string), Molecule number (integer)

water 1

O,0

H,0

H,0

water 2

O,1

H,1

H,1

water 3

O,2

H,2

H,2

First symbol (or symbols) describes atom. In case of water only, all oxygens and hydrogens are equivalent. So, symbols are the same for all molecules. If we want to use CO_2 and H_2O in the same system for example, oxygen atoms are not the same anymore and in the description, they should look like O_Water and O_CO2 for example. The number followed by symbol is number of molecule for which this atom belongs. If we have 3 water molecules, then numbers are 0, 1 and 2. Each molecule must have unique number.

Section &FEATURES starts with key word &FEATURES and ends with &endFEATURES. It describes how features will be built. For now, it contains 3 subsystems: &SingleDistances, &DoubleDistances, &Harmonics. This means that each features can contain one distance raised to some power, product of two distances raised by some powers, product of two distances raised by some powers multiplied by product of corresponding spherical harmonics of come order and some degree (r1*r2*H1*H1).

&SingleDistances subsystem can look like this:

&SingleDistances

SingleDistancesInclude = True

&SingleDistancesDescription

O,O,intermolecular: -1,-2,-3,-4,-5,-6,-12

O,H,intermolecular: -1,-2,-3

H,H,intermolecular: -1,-2,-3,-4,-5

O,H,intramolecular:

H,H,intramolecular:

&endSingleDistancesDescription

&DefaultSingleDistances

SinglePowers: -1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-11,-12,-13

&endSingleDistances

It starts with row &SingleDistances and ends with &endSingleDistances.

Variable SingleDistancesInclude = True means that this type of features will be included in feature set. Sub-subsystem &SingleDistancesDescription starts with &SingleDistancesDescription and ends with &endSingleDistancesDescription. It assigns list of powers for each type of distance. Powers can be positive or

negative, integer of float (now in order to increase computation speed they are forced to be integer). System with only water molecules has 5 types of distances:

O,O,intermolecular

O,H,intermolecular

H,H,intermolecular

O,H,intramolecular

H,H,intramolecular

Where symbols correspond to atoms in the distance and intermolecular or intramolecular describes are they belong to the same molecule or not. For example, O,H,intermolecular means that this distance is between oxygen and hydrogen from different molecules. In this list

O,O,intermolecular: -1,-2,-3,-4,-5,-6,-12

O,H,intermolecular: -1,-2,-3

H,H,intermolecular: -1,-2,-3,-4,-5

O,H,intramolecular:

H,H,intramolecular:

For distances of type O,O,intermolecular, negative powers -1,-2,-3,-4,-5,-6,-12 will be assigned. For distances O,H,intramolecular and H,H,intramolecular powers will not be assigned at all. In other words, those types of distances will not be included in feature set. Next part of subsystem

&DefaultSingleDistances

SinglePowers: -1,-2,-3,-4,-5,-6,-7,-8,-9,-10,-11,-12,-13

describes default powers that will be assigned for each distance that does not have its own description in previous section. So, if subsystem looks like:

&SingleDistances

SingleDistancesInclude = True

&SingleDistancesDescription

O,O,intermolecular: -1,-2,-3,-4,-5,-6

&endSingleDistancesDescription

&DefaultSingleDistances

SinglePowers: -1,-2

&endSingleDistances

It means that distances of type O,O,intermolecular will have powers -1,-2,-3,-4,-5,-6. The rest (O,H,intermolecular; H,H,intermolecular; O,H,intramolecular; H,H,intramolecular) will have powers -1, -2.

In case:

&SingleDistances

SingleDistancesInclude = True

&SingleDistancesDescription

&endSingleDistancesDescription

&DefaultSingleDistances

SinglePowers: -1,-2

&endSingleDistances

All types of distances will have powers -1 and -2

&DoubleDistances part works in similar way but it also has additional information.

Variable IncludeSameType specifies whether same types of distances will be included in product. If IncludeSameType=True, then features like $(O\text{-}H)^m*(O\text{-}H)^n$ will be included in feature set. It works also for

$(H\text{-}H)^m*(H\text{-}H)^n$ and the rest. Also, there is a possibility to create include / exclude list of distance types in order to have more flexibility. It allows to not to include useless pairs in feature set. List starts with row &IncludeExcludeList, and ends with row &endIncludeExcludeList. Between, there can be a list of pares of distances like

O,O,intermolecular;O,H,intermolecular

O,H,intermolecular;H,H,intramolecular

If this list is not empty, there are two ways of using it. Variables IncludeAllExcept and ExcludeAllExcept specify how to interpret include / exclude list. If IncludeAllExcept=True and ExcludeAllExcept=False, then all possible pairs of distances will be included in feature set except the ones in the list. In this example, pairs O,O,intermolecular;O,H,intermolecular

O,H,intermolecular;H,H,intramolecular

Will be excluded. Variables IncludeAllExcept and ExcludeAllExcept cannot be both True or both False.

In the situation if IncludeAllExcept=False and ExcludeAllExcept=True, only pairs from the list will be included in the feature set. If this list is empty and IncludeAllExcept=True and ExcludeAllExcept=False, then all possible pairs will be included.

&Harmonics section describes spherical harmonic features. It is a product of two distances raised by some powers multiplied by product of corresponding spherical harmonics of come order and some degree (r1*r2*H1*H1). They look like $(O\text{-}O)^m*(O\text{-}H)^n*H1*H2$, where H1 and H2 are functions of 4 variables: order, degree, and angles theta and phi. Orders and degrees are specified in &Harmonics section:

Order: -3,-2,-1,0,1,2,3

Degree: 0,1,2,3

Where |Order| <= Degree

HarmonicCenter assigns center of coordinate system to calculate spherical harmonics

HarmonicCenter: O

Means that centers will be oxygen atoms.

HarmonicAtoms specifies atoms, for which spherical harmonics will be calculated

HarmonicAtoms: O,H

Means that harmonics will be calculated for oxygen and hydrogen atoms

Harmonics will be calculated only if center is the same atom for both harmonics.

Harmonics will not be calculated if harmonic1 atom and harmonic2 atom belong to the same molecule.

Harmonics will not be calculated for atoms within same molecule, so if center is oxygen of molecule 1, then there will be no harmonics for hydrogens which belong to molecule 1. The rest in this section is like section &DoubleDistances.