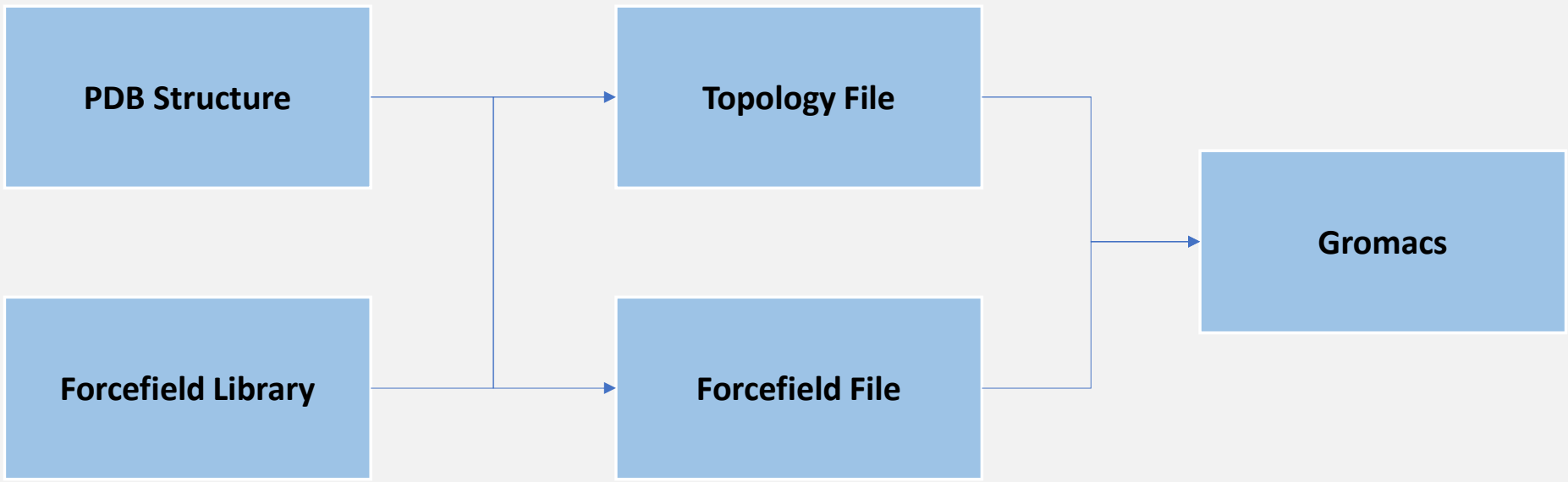
PDB & Forcefield

Naf Guo
2025

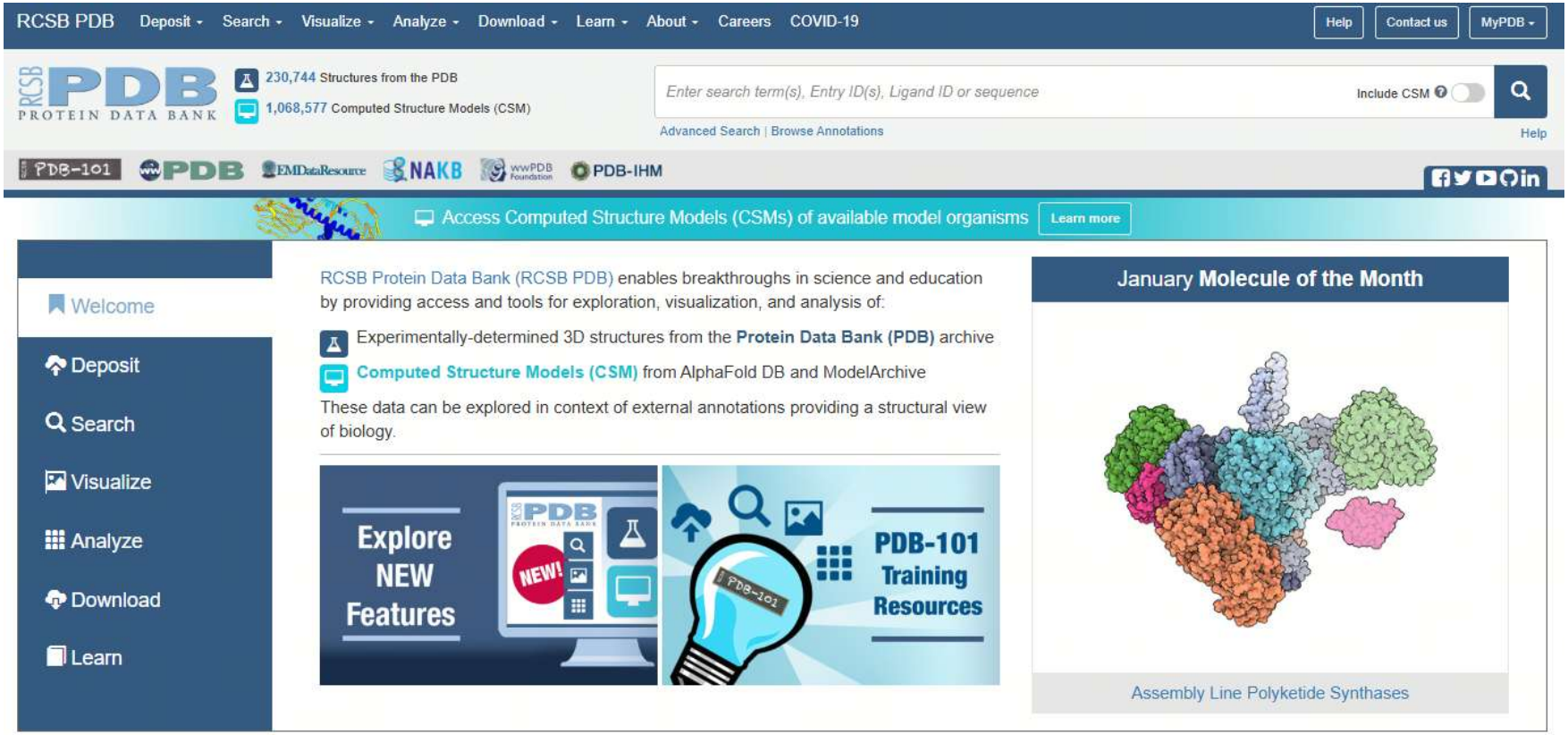
What we need to know about our input



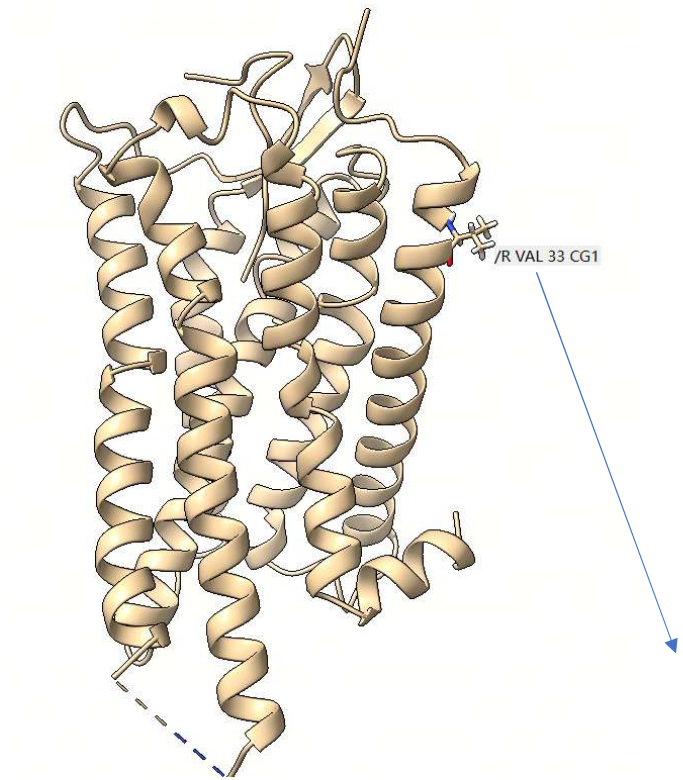
PDB: Protein Data Bank

https://www.rcsb.org/

RCSB PDB 数据库存储了人类几十年解析的数十万条结构数据。而结构是动力学模拟的输入。



PDB File Format



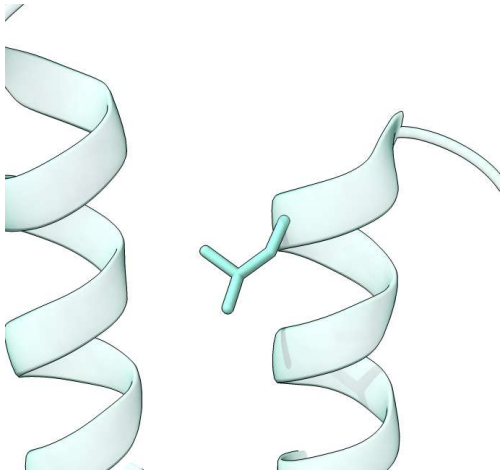
Atom index | Atom name | Residue name | Chain name | Residue index | x | y | z

ATOM	148	N	GLN	R	31	117.965	133.304	182.451	1.00	90.97	N
ATOM	149	CA	GLN	R	31	119.316	133.051	181.934	1.00	90.97	C
ATOM	150	C	GLN	R	31	119.295	132.547	180.486	1.00	90.97	C
ATOM	151	O	GLN	R	31	120.101	133.020	179.686	1.00	90.97	O
ATOM	152	CB	GLN	R	31	120.100	132.057	182.815	1.00	90.97	C
ATOM	153	CG	GLN	R	31	120.446	132.590	184.215	1.00	90.97	C
ATOM	154	CD	GLN	R	31	121.427	131.697	184.984	1.00	90.97	C
ATOM	155	OE1	GLN	R	31	122.029	130.775	184.435	1.00	90.97	O
ATOM	156	NE2	GLN	R	31	121.599	131.979	186.276	1.00	90.97	N
ATOM	184	N	VAL	R	33	116.920	133.176	178.236	1.00	84.77	N
ATOM	185	CA	VAL	R	33	116.531	134.353	177.454	1.00	84.77	C
ATOM	186	C	VAL	R	33	117.703	135.313	177.152	1.00	84.77	C
ATOM	187	O	VAL	R	33	117.780	135.805	176.027	1.00	84.77	O
ATOM	188	CB	VAL	R	33	115.355	135.114	178.142	1.00	84.77	C
ATOM	189	CG1	VAL	R	33	115.158	136.594	177.744	1.00	84.77	C
ATOM	190	CG2	VAL	R	33	114.033	134.357	177.918	1.00	84.77	C

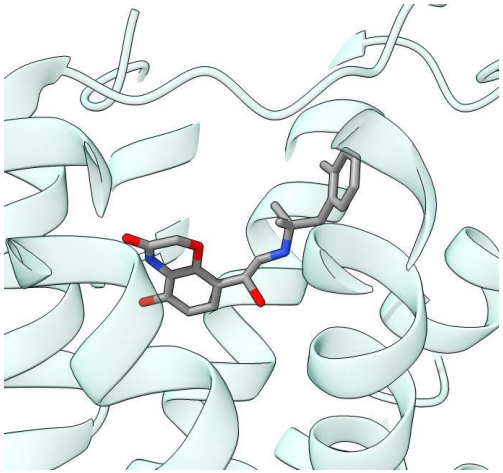
PDB文件通过记录蛋白质每一个原子的原子名，所属的氨基酸残基名，序号，空间坐标（x，y，z）来存储其结构信息，而专门的软件能读取PDB格式的文件来展示蛋白质的3D结构。

Preprocess of PDB before MD

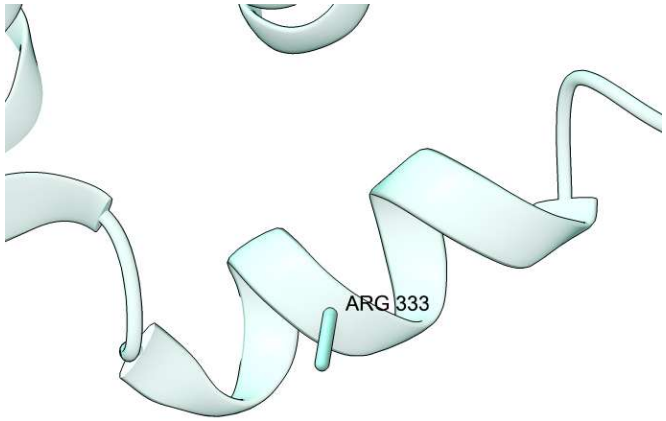
PDB文件一般无法直接用于动力学模拟，因为它可能会缺失一些必要的细节



PDB文件一般没有氢的位置信息，因为在大多数情况下，实验手段看不清氢原子的位置



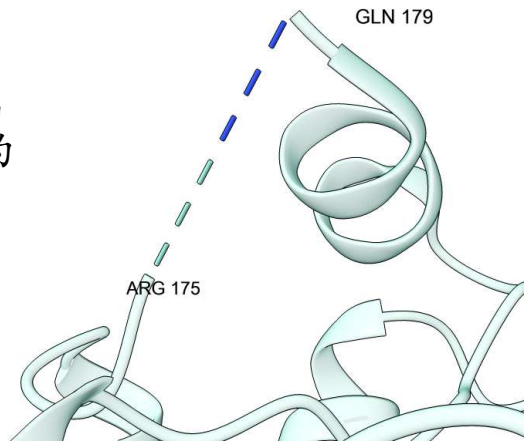
PDB文件中非氨基酸的部分键级信息，质子化状态，氢原子常常是缺失的



PDB文件有的氨基酸残基可能有一部分原子的位置未能被解析出来

Missing sidechain

Missing backbone



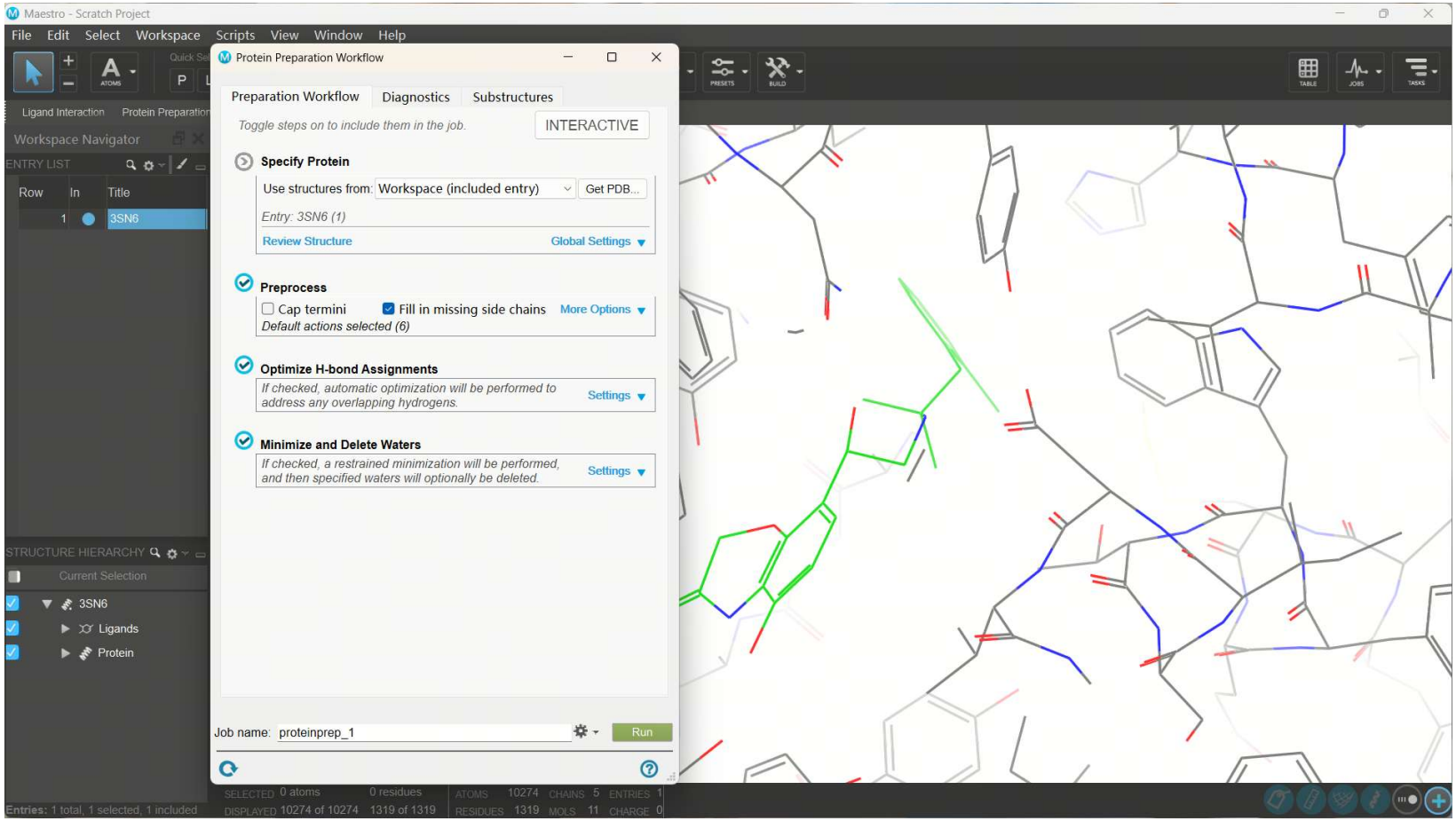
PDB文件中可能有单个或多个氨基酸残基整个未能被解析出来

Missing Residues

Missing Loop

Preprocess of PDB before MD

一般可以用maestro中的protein preparation进行蛋白质准备修复所有问题



Preprocess of PDB before MD

也可以用swiss-model，用结构对该蛋白质本身进行建模，以此修复缺失的残基，残基侧链等

https://swissmodel.expasy.org/interactive#structure

Preprocess of PDB before MD

- 使用python包PDBFixer修复蛋白质

https://github.com/openmm/pdbfixer

PDBFixer

1. Introduction

Protein Data Bank (PDB or PDBx/mmCIF) files often have a number of problems that must be fixed before they can be used in a molecular dynamics simulation. The details vary depending on how the file was generated. Here are some of the most common ones:

1. If the structure was generated by X-ray crystallography, most or all of the hydrogen atoms will usually be missing.
2. There may also be missing heavy atoms in flexible regions that could not be clearly resolved from the electron density. This may include anything from a few atoms at the end of a sidechain to entire loops.
3. Many PDB files are also missing terminal atoms that should be present at the ends of chains.
4. The file may include nonstandard residues that were added for crystallography purposes, but are not present in the naturally occurring molecule you want to simulate.
5. The file may include more than what you want to simulate. For example, there may be salts, ligands, or other molecules that were added for experimental purposes. Or the crystallographic unit cell may contain multiple copies of a protein, but you only want to simulate a single copy.
6. There may be multiple locations listed for some atoms.
7. If you want to simulate the structure in explicit solvent, you will need to add a water box surrounding it.
8. For membrane proteins, you may also need to add a lipid membrane.

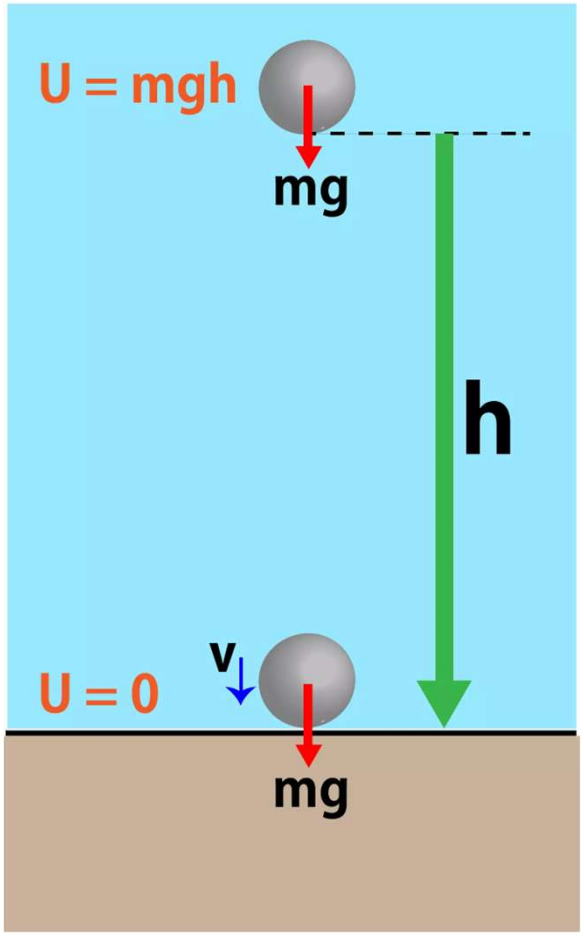
PDBFixer can fix all of these problems for you in a fully automated way. You simply select a file, tell it which problems to fix, and it does everything else.

PDBFixer can be used in three different ways: as a desktop application with a graphical user interface; as a command line application; or as a Python API. This allows you to use it in whatever way best matches your own needs for flexibility, ease of use, and scriptability. The following sections describe how to use it in each of these ways.

- 使用Rosetta Common修复蛋白质

https://rosettacommons.org/

Forcefield



- 假设有一个质量为$\boldsymbol{m}$的球位于高度$\boldsymbol{h}$处
- 其向下的重力为$\boldsymbol{mg}$
- 其重力势能为$\boldsymbol{U = mgh}$
- 注意到$\boldsymbol{mg} = \frac{\boldsymbol{mgh}}{\boldsymbol{h}} = \frac{\boldsymbol{d}\mathrm{U}}{\boldsymbol{d}\mathrm{h}}$
- 在这里我们不加证明的给出，势能在对某各方向坐标的偏导数，其绝对值等于该方向上物体受的力，方向相反。
- $F_x = \frac{dU}{dx}$ $\quad F_y = \frac{dU}{dy}$ $\quad F_z = \frac{dU}{dz}$
- 有了受力以后，可以计算加速度，设初速度为0或者随机一个初速度，就可以计算物体在时间$\boldsymbol{t}$后的位置

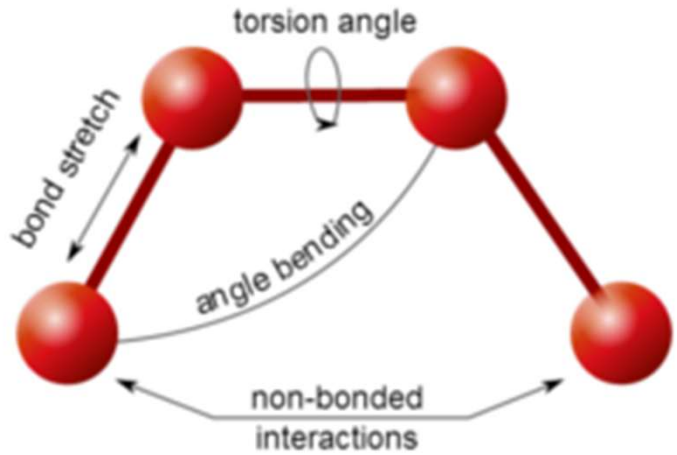
$$F = ma \qquad v_t = at \qquad S_t = v_0 t + \frac{1}{2}at^2$$

Forcefield Parameters

在动力学模拟中，我们要做的就是:

1. 用一套力场参数计算某一套位置坐标下蛋白质体系的势能
2. 基于势能可以得到受力
3. 基于受力得到加速度以及一个很短的Δt以后的每一个原子的位移
4. 基于位移更新体系所有原子的坐标，再计算新的势能以及受力……依此循环

$$
\begin{aligned}
V(r) = & \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \\
& + \sum_{\text{dihedrals}} k_\phi \big(1 + \cos(n\phi - \phi_0)\big) + \sum_{\text{impropers}} k_\psi (\psi - \psi_0)^2 \\
& + \sum_{\substack{\text{non-bonded} \\ \text{pairs}(i,j)}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{\substack{\text{non-bonded} \\ \text{pairs}(i,j)}} \frac{q_i q_j}{\varepsilon_D r_{ij}}.
\end{aligned}
$$



Mapping between PDB & forcefield

根据PDB文件里每一个原子的名字，动力学模拟软件会在力场参数文件中查表找到这个原子名(PDB atom name)对应的原子类型(atom type)，每一个原子类型都具有一系列参数，如带电量，质量等。

```
!entry.ALA.unit.atoms table  str name
 "N" "N" 0 1 131072 1 7 -0.415700
 "H" "H" 0 1 131072 2 1 0.271900
 "CA" "CX" 0 1 131072 3 6 0.033700
 "HA" "H1" 0 1 131072 4 1 0.082300
 "CB" "CT" 0 1 131072 5 6 -0.182500
 "HB1" "HC" 0 1 131072 6 1 0.060300
 "HB2" "HC" 0 1 131072 7 1 0.060300
 "HB3" "HC" 0 1 131072 8 1 0.060300
 "C" "C" 0 1 131072 9 6 0.597300
 "O" "O" 0 1 131072 10 8 -0.567900
```

```
!entry.ALA.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 3 4 1
 3 5 1
 3 9 1
 5 6 1
 5 7 1
 5 8 1
 9 10 1
```

PDB atom name

Amber atom type

charge

How atoms are connected, for e.g, first line is 1 2 1, means atom1 is connect with atom2. Forget the third “1”. 1 2 is enough.

For example, atom1 which is N, is connect with atom2, which is H

可以注意到，在上面的amber力场里，虽然同为碳原子，PDB里的丙氨酸的CA被分配的atom type是CX，PDB的CB被分配了CT，还能注意到CX的电荷是0.0337，CT的电荷是-0.1825

Mapping between PDB & forcefield

根据atom type，软件进一步可以去查表找到其它的力场参数，用于计算整个体系的势能

```
MASS
CO 12.01          0.616   !          sp2 C carboxylate group
2C 12.01          0.878              sp3 aliphatic C with two (duo) heavy atoms
3C 12.01          0.878              sp3 aliphatic C with three (tres) heavy atoms
C8 12.01          0.878              sp3 aliphatic C basic AA side chain
```

Amber Atom types

```
BOND
C -2C   317.0     1.5220
C*-2C   317.0     1.4950
C8-C8   310.0     1.5260
C8-CX   310.0     1.5260
```

```
ANGL
N -C -2C     70.0      116.60
O -C -2C     80.0      120.40
OH-C -2C     80.0      110.00
CB-C*-2C     70.0      128.60
```

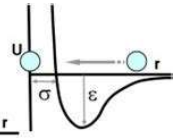
```
DIHE
C8-CX-N -C    1    0.000        0.0          -4.    phi'
C8-CX-N -C    1    0.800        0.0          -3.
C8-CX-N -C    1    1.800        0.0          -2.
C8-CX-N -C    1    2.000        0.0           1.
CT-CX-N -C    1    0.000        0.0          -4.
CT-CX-N -C    1    0.800        0.0          -3.
```

```
NONB
  2C          1.9080   0.1094          Spellmeyer
  3C          1.9080   0.1094          Spellmeyer
  C8          1.9080   0.1094          Spellmeyer
  CO          1.9080   0.0860          OPLS
```

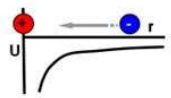
Force constant k

equilibrium bond length r_0

$$U = \sum_{i<j}\sum 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$



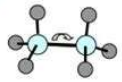
$$+ \sum_{i<j}\sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$



$$+ \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2$$

$$+ \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$

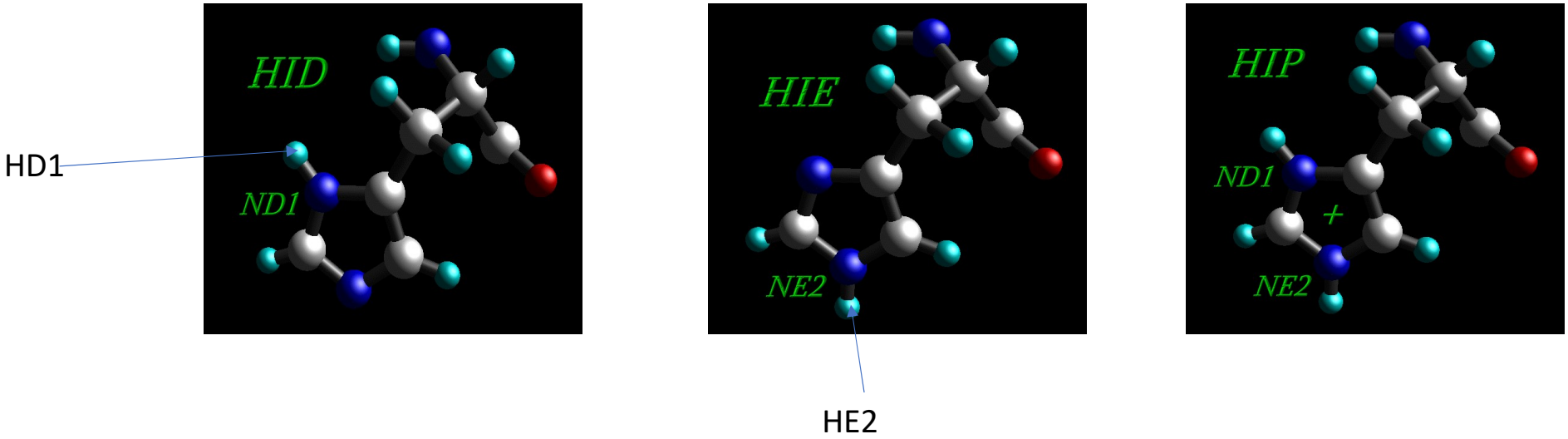
$$+ \sum_{torsions} k_\phi [1 + \cos(n\phi - \delta)]$$



所以，动力学模拟软件是通过PDB里原子的名字来识别并分配力场参数的，那么，如果出现了软件不认识的原子名，就会报错

Amino Acid Residue Name

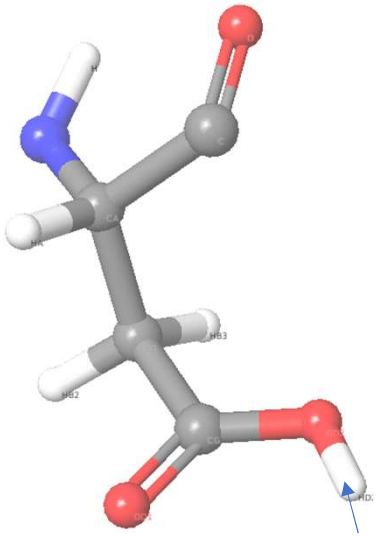
- Histidine (HIS in normal pdb files) is really one of three possible residues:
  HID: Histidine with hydrogen on the delta nitrogen
  HIE: Histidine with hydrogen on the epsilon nitrogen
  HIP: Histidine with hydrogens on both nitrogens; this is positively charged.



Amino Acid Residue Name

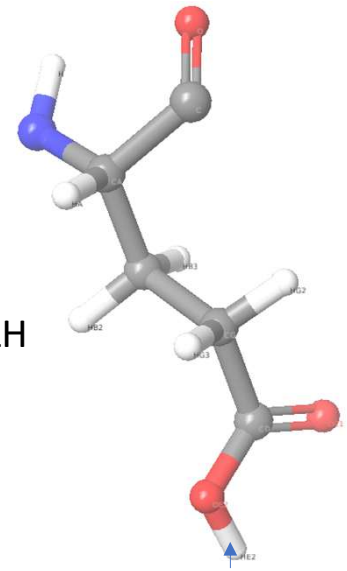
- Cys: PDB residues named “CYS” are automatically converted into a free cysteine with an SH side chain end. If the cysteine is known to be in a S-S bridge, the residue name must be “CYX”. That SH would be called “HG” in PDB.
- Asp, Glu, Lys: charged form would be ASP, GLU, LYS. Uncharged form must be “ASH”, “GLH”, “LYN”.

ASH



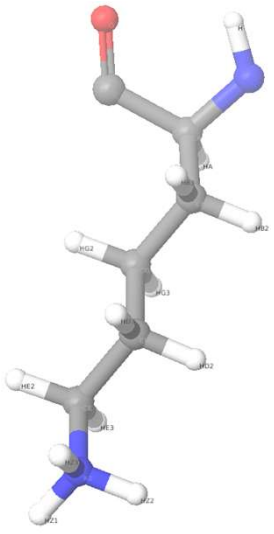
HD2

GLH



HE2

LYS



HZ1 would lost in LYN

Mapping between PDB & forcefield