Chapter 2

Methodology

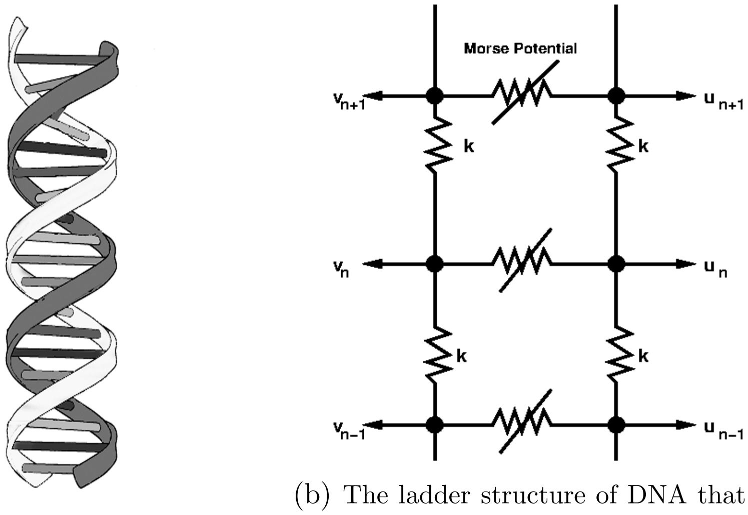
2.1 Statistical Model

Michel Peyrard and Alan Bishop proposed a statistical model of DNA in 1989 [65]. The works by E. Prohofsky and his team [58, 129, 130] provided the basic principle of making this new model. The main objective of the PB (Peyrard-Bishop) model is to explain the DNA denaturation. At first, it was limited to the thermal ensemble, later on, it went to one step further, and DNA denaturation in force ensemble was included. The model can successfully explain DNA unzipping for long as well as short DNA chains. It treated the interaction of the nearest neighbor coupling as a harmonic spring and later the anharmonicity part also was added by T. Dauxois and since it is known as PBD model (Peyrard-Bishop-Dauxois) [64]. Several groups have been using PBD model to study the DNA unzipping process at different framework [18, 38, 131–134].

Basic assumptions in the PBD-model :

- The model works in the center of the mass frame, and it is a quasi-one-dimensional model.
- The model considers equal reduced mass for every type of base pair. However, the heterogeneity effect is assimilated in the basic model later on.
- Tha basic PBD model lies in a plane, and it describes the ladder model of DNA. Later helicoidal geometry is added also [14, 135, 136].
- As the PBD model is interested mainly in DNA unzipping, so the longitudinal movements of bases are not so much significant and their amplitude of

vibrations are neglected. Hence, the transverse stretching of the hydrogen bonds between complementary bases is considered exclusively.



(a) DNA as double helix

(b) The ladder structure of DNA that the PBD model take cares.

Figure 2.1: Graphical presentation of PBD model. DNA double helix is simplified as a ladder. The filled circle represents the nucleotides, and the arrow shows that transverse motions are only considered.

2.1.1 Model Hamiltonian

The diagrammatic representation of the PBD model is shown in Figure 2.1(b). Every base pair holds two chains, each base pair keeps two degrees of freedom. In the diagram, u_n represents the displacement of the n^{th} base pair in one chain and v_n represents the displacement of corresponding n^{th} base pair of the second chain. The displacement of bases is taken from their mean position. The model represents the hydrogen bonding between the bases in a pair through the Morse potential. Morse potential was first used to represent the Hydrogen bond in 1985 by E. Prohofsky *et al.* [129]. The stacking interaction between adjacent base pairs is represented by a harmonic potential first and later on it is modified as anharmonic potential. The Hamiltonian of the system is [65, 137, 138],

$$H = \sum_n \left[\frac{1}{2} m \{ \dot{u}_n^2 + \dot{v}_n^2 \} + \frac{1}{2} k [(u_n - u_{n+1})^2 + (v_n - v_{n+1})^2] + D (\exp[-a(u_n - v_n)] - 1)^2 \right] \tag{2.1}$$

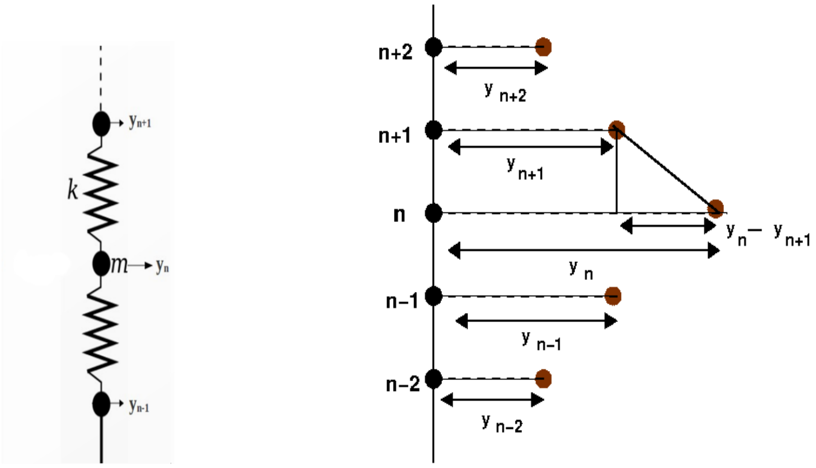
In the Hamiltonian equation, m is the effective nucleotide mass of the base pair and k represents the elasticity of the DNA strand. The parameter D represents the depth of the potential well, and a represents the inverse of the width of the potential well. There are two motions of base pairs, one is *in-phase*, and the other one is *out-of-phase*. They are:

$$x_n = \frac{u_n + v_n}{\sqrt{2}} \quad ; \quad y_n = \frac{u_n - v_n}{\sqrt{2}} \tag{2.2}$$

The Hamiltonian can be re-written as,

$$\begin{aligned} H = \sum_n \Big[\{\frac{1}{2} m \dot{x}_n^2 + \frac{k}{2}\left(x_n - x_{n+1}\right)^2\} + \{\frac{1}{2} m \dot{y}_n^2 + \frac{k}{2}\left(y_n - y_{n+1}\right)^2\} \\ + D(\exp[-a y_n \sqrt{2}] - 1)^2 \Big] \end{aligned} \tag{2.3}$$

As it is said in the “Basic assumptions in the PBD-model”, the model is interested only in the separation of the chains not in the movements of DNA chain as a whole. Since *out-of-phase* motion is solely responsible for hydrogen bond stretching; hence, the equation of Hamiltonian is described by the scalar variable y_n exclusively. Figure 2.2 shows the transverse motion of the hydrogen bonds. So the functional



(a)

(b)

Figure 2.2: (a) PBD model is viewed as a one-dimensional monotonic lattice model. (b) displacements of each nucleotides from their equilibrium positions.

Hamiltonian is written as,

$$H = \sum_{n=1}^{N} \left[\frac{p_n^2}{2m} + V(y_n) \right] + \sum_{n=1}^{N-1} [W(y_n, y_{n+1})] \tag{2.4}$$

where $V(y_n) = D(e^{-a y_n} - 1)^2$ *and* $W(y_n, y_{n+1}) = \frac{k}{2}(y_n - y_{n+1})^2$
The onsite potential V is function of y_n and the stacking potential W is function

of variable y_n and y_{n+1}. The stacking potential is a crucial factor in the Hamiltonian. The PB model with this potential harmonic stacking provided a good value of melting temperature (T_m), but the values were around 150 K more than actual experimental results [139]. Dauxois *et al.* [64, 139] modified this harmonic potential, and a nonlinear term was introduced to describe stacking interactions more appropriately. The new stacking potential is:

$$W(y_n, y_{n+1}) = \frac{k}{2}(y_n - y_{n+1})^2[1 + \rho e^{-b(y_n+y_{n+1})}]. \tag{2.5}$$

The parameter b is the decay constant for stacking interaction, and ρ is a dimensionless parameter. Both these parameters account the range of anharmonicity. When the DNA is in zipped state then $y_n = 0$, and the force constant becomes $k(1 + \rho)$. The force constant starts to decrease from $k(1 + \rho)$ to k according to the two interacting base pair's stretching. This decrease of force constant delivers large entropy and that helps the unzipping process of DNA. The outcome of this PBD model showed good agreement with experimental results, and the transition *per se* was preferably sharp. In his paper, T. Dauxois [64] has shown that how anharmonicity of stacking helps to get a sharp transition in the PBD model.

2.1.2 Partition Function

Using the model Hamiltonian the canonical partition function is calculated, and through this partition function (Z_c) the essential thermodynamics properties can be studied. The partition function of the system :

$$Z = \frac{1}{h^N} \int \prod_{n=1}^{N} \{dy_n dp_n \exp\left(-\beta H\right)\} = Z_p Z_c, \tag{2.6}$$

The partition function holds momentum part (Z_p) and configurational part(Z_c). The number N says about the number of base pairs in the DNA chain. The parameter $\beta = \frac{1}{k_B T}$. The momentum part can be integrated through Gaussian integral and for the N number of base pairs it is :

$$Z_p = \int_{-\infty}^{\infty} \left[\prod_{n=1}^{N} dp_n \exp\{-\beta \left[\frac{p^2}{2m}\right]\} \right]$$

$$Z_p = (2\pi m k_B T)^{N/2}, \tag{2.7}$$

The configurational part of the partition function is :

$$Z_c = \int_{-\infty}^{\infty} \left[\prod_{n=1}^{N-1} dy_n \exp\{-\beta \left[W(y_n, y_{n+1}) + V(y_n)\right]\} \right] dy_N V(y_N), \tag{2.8}$$

The Z_c part of the partition function holds the coupled terms so the solution of this part is not straight forward. If we want to follow the standard TI(Transfer Integral) method then the solution would follow these steps:
A kernel $K(y_n, y_{n+1})$ is defined to solve it [140].

$$K(y_n, y_{n+1}) = \exp\left[-\beta H(y_n, y_{n+1})\right] \tag{2.9}$$

It is evident that for the homogeneous chain, $K(x, y) = K(y, x)$. Z_c can then be written as :

$$Z_c = \int_{-\infty}^{\infty} \prod_{n=1}^{N} dy_n K(y_n, y_{n+1}) \tag{2.10}$$

Following the periodic boundary condition,we can write :

$$Z_c = \int_{-\infty}^{\infty} \prod_{n=1}^{N} dy_1 dy_2 ....dy_N K(y_1, y_2) K(y_2, y_3).....K(y_N, y_1) \tag{2.11}$$

Integral equation is introdued now and we are able to write,

$$\int dy K(x, y)\phi(y) = \lambda\phi(x) \tag{2.12}$$

There is a assumption that should be taken now on the basis of the two facts $K(x, y) > 0$ and the symmetry of the kernel.

$$\|K(x, y)\| = \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{K(x, y)\}^2 \right]^{1/2} < \infty \tag{2.13}$$

the integral equation holds positive Hilber Schmidt type kernel [141] so positive eigenvalues and orthonormal eigenfunctions are supposed to be the two aspect of it. Let us consider the eigenvalues are λ_i and the corresponding eigenvectors are ϕ_i then

$$\int dx \phi_n(x)\phi_m(x) = \delta_{nm} \qquad \text{and} \qquad \sum_{n=1}^{\infty} \phi_n(x)\phi_n(y) = \delta(x - y). \tag{2.14}$$

The kernel $K(x, y)$ can be written [141] ,

$$K(x, y) = \sum_n \lambda_n \phi_n(x)\phi_n(y) \tag{2.15}$$

Now we put the $K(x, y)$ into Eq.2.11 and implement the orthonormality conditions (Eq.2.14). Then the configurational partition becomes,

$$Z_c = \sum_{n=1}^{\infty} \lambda_n^N \tag{2.16}$$

For the open boundary condition the Z_c is :

$$Z_c = \sum_n \left(\int dy \phi_n(y) \exp\left\{ -\beta \frac{V(y)}{2} \right\} \right)^2 \lambda_n^{N-1} \tag{2.17}$$

Through diagonalizing the output matrices in Eq.2.9, the eigenvalues and corresponding eigenstates can be determined.
There are some technical issues that we have to address in the standard TI(Transfer Integral) method. The TI method asks for the unbound property of all the on-site potentials through the existence of Eq.2.13. On the contrary, the Morse potential in the Hamiltonian is bound [142] therefore, it violates the condition of Eq.2.13. So the kernel that we define in Eq.2.9 does not fall under the category of a Hilbert-Schmidt type kernel and it becomes a singular kernel for which integration limit would be $[-\infty, +\infty; -\infty, +\infty]$. The properties of this kernel lead to the divergence of the partition function. if we have to use the TI method, an upper limit of y_n has to be set up. This upper bound will limit the kernel on a finite space [a, b ; a, b], so that its norm exists.
If the heterogeneity is introduced in the sequence, then the calculation of partition function is not so straight forward like homogeneous sequence. For the heterogeneous sequence the n^{th} site may not be the same nature as its $n-1$ and $n+1$ sites. To overcome this issue there are different approaches that have been proposed like Cule [143] suggested to treat the heterogeneity as quench disorder. Y Zhang *et al.* [140] suggested extended transfer matrix approach (ETMA). With open boundary condition the matrix multiplication method can address the calculation of partition function [38, 132, 144]. The crucial task of choosing proper cut-offs for the integration has been introduced by T S Van Erp *et al.* [145]. A large cut-off can increase the probability of the complete denaturation of a finite chain. This probability comes to be one in the limit of infinite cut-off. This signifies the

same divergent issue of Eq.2.8. A double-stranded ensemble was introduced in the numerical calculations of partition function by T S Van Erp *et al.* [145] to avoid this divergent issue. Their group-work showed an upper cut-off of ≈ 144 Å and a lower cut-off of -0.4 Å for a set of model parameters at $T = 300$ K. To calculate partition function, we generate matrices using Eq.2.10 and multiply the obtained matrices one by one. The integration in the Eq.2.8 can be written as:

$$Z_c = \int_{-\infty}^{\infty} \exp\left(-\beta\frac{V(y_1)}{2}\right) dy_1 \prod_{n=1}^{N-1} dy_n \exp\left[-\frac{\beta}{2}\{V(y_n)+V(y_{n+1}) + 2W(y_n,y_{n+1})\}\right] \exp\left(-\beta\frac{V(y_N)}{2}\right) dy_N \tag{2.18}$$

After the proper cut-offs, the next task is to discretise the integral. In order to get a precise value of melting temperature (T_m) we have observed that Gaussian quadrature is the most effective quadrature. We have found that discretization of the space with 900 points(this is the dimension of the matrix 900×900 also) is sufficient to get an accurate value of T_m [146].

Once we calculate the partition function, then we can calculate the other thermodynamic properties through the partition function. For finding out the melting temperature T_m, we have to calculate specific heat because the peak of the specific head is the melting temperature point of the DNA chain. So first we calculate free energy (Helmholtz free energy of the system) per base pair then it will proceed toward entropy S and finally specific heat C_v using the following relations,

$$f(T) = -\frac{1}{2}k_BT\ln(2\pi mk_BT) - \frac{k_BT}{N}\ln Z_c. \tag{2.19}$$

$$S(T) = -\frac{\partial f}{\partial T} \tag{2.20}$$

$$C_v(T) = -T\frac{\partial^2 f}{\partial T^2}. \tag{2.21}$$

2.1.3 Limitations of the PBD model

Although the PBD model has the ability to illustrate the denaturation process of DNA chain delightfully still, the model can not be said as a complete model. It allows to study both homogeneous and heterogeneous sequence of long as well as short DNA chain, notwithstanding it ignores many structural and dynamical features of DNA chain. The model is a quasi-one-dimensional model, so it ignores the real three-dimensional structure of the molecule. The water and the solution effects are considered as the background effect [20]. The model under-estimate the entropy of the dsDNA since it ignores the entropy which is due to the motion of the dsDNA strand. The effect of hydration of open base pairs is also an issue in the model. Researchers have been trying to overcome these issues of the PBD model [38, 133, 147, 148].

2.2 Molecular Dynamics

Molecular dynamics (MD) is an algorithm to study the dynamics of atoms and molecules for a fixed time period. The trajectories of these atoms and molecules in phase space as a function of time which belong to the same ensemble, are calculated through the MD simulation. The forces that act on the atoms or molecules are either taken from the classical potential frame or from the quantum potential frame, and these two approaches separate classical and quantum molecular dynamics. Statistical mechanics plays an important role in MD calculation since it deals with the microscopic parameters, and through these parameters, it gives the observable macroscopic properties. In the late of 1950s, it was proposed within the field of theoretical physics, and later it has been started to apply in all the branches of general science research mostly in chemical physics, materials science, and biological sciences.

2.2.1 Classical Molecular Dynamics

We are interested mainly in the classical molecular dynamics, and in classical MD, the Newton equation of motion play the role to calculate the dynamics of the system. To know the position of each atom with respect to time we have to solve the newton's equation of motion, and that is,

$$F_i = m_i a_i = m_i \frac{d^2 r_i}{dt^2} \tag{2.22}$$

where m_i is the mass of the i^{th} atom and a_i is the acceleration so it can be written to $\frac{d^2 r_i}{dt^2}$. The force on each atom can be written in terms of potential energy,

$$F_i = -\frac{\partial V}{\partial r_i}. \tag{2.23}$$

The accuracy of the MD simulation mainly depends on how accurately we consider every potential energy that acts on the system. The potential energy and initial coordinates of the system generate a new set of coordinates and a force that is acting on the new set of coordinates. Repetition of the procedure makes the trajectory depending on how the system will evaluate with time. There is another mathematical expression required now, which instructs how the particles are supposed to interact. In the MD simulation, this mathematical functional form is called the Force field. The force field dwells on the interatomic potentials and the set of parameters that belong to these potentials. The parameters have been defined in the force field by mechanical calculations or through fitting the experimental data. There are many force fields depending on the details of the simulation system. If we want to write down a typical expression of a force field, then it holds two types of interaction energy, and they are E_{bonded} and $E_{nonbonded}$.

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}$$

$$E_{nonbonded} = E_{eletrostatic} + E_{vanderWalls}$$

So a typical force field looks like this,

$$V = \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} [1 + cos(n\phi - \delta)] + \sum_{improper} V_{imp} + \sum_{LJ} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}}, \tag{2.24}$$

where the first four potentials are respectively bond stretching, angle bending, and dihedral and improper torsions so they are basically represent bonded interactions and the last two potentials show repulsive (Coulombic interactions) and Van der Waals interactions (Lennard-Jones potential) therefore they represent the nonbonded interactions(see fig.2.3). Finite-size can cause problems with boundary effects in MD simulation also. Periodic boundary conditions (PBC) are used to solve this boundary effect problem. PBC can make the system an infinite one though the periodicity effects. If any corner atom which is not fit in the box and leaves the simulation box by right-hand face, then according to PBC, that atom is supposed to appear in the next simulation box by the left-hand face.

The next part is the algorithms. We have already said that the equation of motion can not be solved without numerical technique. So discretization of the trajectory

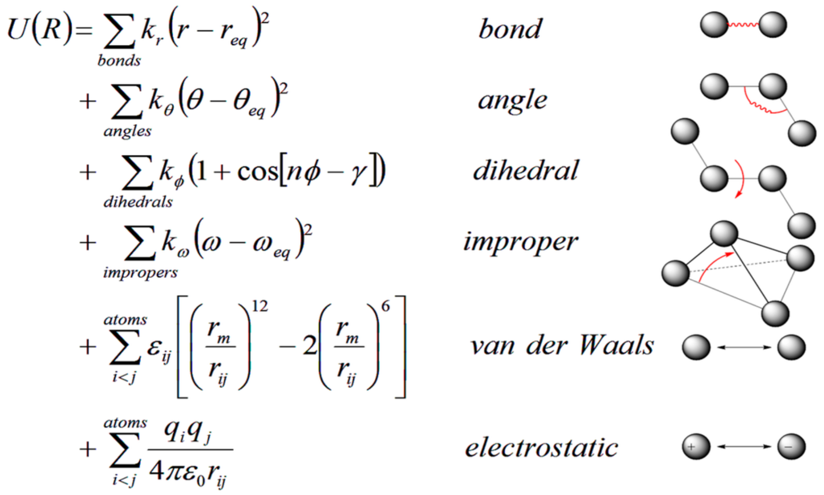

$U(R) = \sum_{bonds} k_r (r - r_{eq})^2$
bond
$+ \sum_{angles} k_\theta (\theta - \theta_{eq})^2$
angle
$+ \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma])$
dihedral
$+ \sum_{impropers} k_\omega (\omega - \omega_{eq})^2$
improper
$+ \sum_{i<j}^{atoms} \varepsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right]$
van der Waals
$+ \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$
electrostatic
+
−

Figure 2.3: Typical potential terms in a force field. Figures are taken from [1]

is the first step then use an integrator to proceed over small time steps like,

$$\mathbf{r}_i(t_0) \rightarrow \mathbf{r}_i(t_0 + \Delta t) \rightarrow \mathbf{r}_i(t_0 + 2\Delta t) \rightarrow ....\mathbf{r}_i(t_0 + n\Delta t)$$

If we follow the Taylor expansion with a starting time t_0 and a known initial positions, positions and forces then we can proceed to time $t_0 + \Delta t$ as we have discussed above. The Taylor expansion as follows:

$$\mathbf{r}_i(t_0 + \Delta t) = \mathbf{r}_i(t_0) + \frac{d\mathbf{r}_i(t_0)}{dt}\Delta t + \frac{1}{2}\frac{d^2\mathbf{r}_i(t_0)}{dt^2}\Delta t^2 + O(\Delta t^3) \tag{2.25}$$

Now for better numerical precision, Verlet proposed that let add the Taylor expansion for $\mathbf{r}_i(t_0 + \Delta t)$ and $\mathbf{r}_i(t_0 - \Delta t)$ then odd powers will be cancelled and the output is,

$$\mathbf{r}_i(t_0 + \Delta t) + \mathbf{r}_i(t_0 - \Delta t) = 2\mathbf{r}_i(t_0) + \mathbf{a}_i(t_0)\Delta t^2 + O(\Delta t^4) \tag{2.26}$$

Therefore the velocities can be written as,

$$\mathbf{v}_i(t_0) = \frac{1}{2\Delta t}\left[\mathbf{r}_i(t_0 + \Delta t) - \mathbf{r}_i(t_0 - \Delta t)\right] \tag{2.27}$$

And the velocity-Verlet algorithm says,

$$\mathbf{r}_i(t_0 + \Delta t) = \mathbf{r}_i(t_0) + \mathbf{v}_i(t_0)\Delta t + \frac{1}{2}\mathbf{a}_i(t_0)\Delta t^2 \tag{2.28}$$

$$\mathbf{v}_i(t_0 + \Delta t) = \mathbf{v}_i(t_0) + \frac{1}{2}\left[\mathbf{a}_i(t_0) + \mathbf{a}_i(t_0 + \Delta t)\right] \Delta t \qquad (2.29)$$

Those algorithms are reasonably accurate and therefore they are a good choice as integrators for an MD simulation.

Since the MD simulation research has been moving from simple to more and more complicated systems, hence the force field also has been developing accordingly. There are some popular force fields: CHARMM [149],AMBER [150],GROMOS [151], OPLS [152],and COMPASS [153]. These force fields are continuously developing, so there are many versions of each force field that are available for us.

In addition to those general force fields, there are also some specific interaction potentials that also play a role in MD simulation. The water model is a great example of this because of its importance [154]. There are many models that have been developed, and they have their own strengths and weaknesses [155]. Vega *et al.* has shown a comparison results among five popular water models(TIP3P, TIP4P, TIP5P, SPC, SPC/E) in their research work [156].

2.2.2 Limitations of the MD simulations

Like every method, MD also has some limitations. When we want to use a technique, then it is good to be aware of the limitations of this technique. Despite all the shortcomings, classical MD can be considered mature [157]. Microsecond simulation is done still simulation time-scales is an issue, and developing computer power is the pathway to it. Active research is going on to get longer-timescale simulation results, and it will be possible through algorithmic improvements, parallel computing, and developing specialized hardware. The next limitation that we face is force field accuracy. Force fields are written by considering many approximations, but researchers have been trying to improve over the last decade, but still, there are many limitations.