1 Molecular Dynamics Simulations

1.1 Introduction

Richard P. Feynman, in his pursuit of fundamental understanding in physics, famously stated: “If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms.” This foundational atomic perspective dictates that the function of any system, either chemical or biological, is encoded entirely within the motions and interactions of its constituent atoms.

This principle is the cornerstone of Molecular Dynamics (MD) simulations. MD is a powerful computational method for studying the evolution of atomic and molecular systems over time by numerically solving Newton’s equations of motion. By tracking the position and velocity of every atom, MD generates an atomistic "trajectory" that describes how a system evolves dynamically [1]. The forces governing these motions are derived from a force field [2], which translates atomic coordinates into the interatomic forces required for the time integration.

In a biological context, MD simulations bridge the gap between microscopic atomic interactions and macroscopic phenomena. The method provides important insights into complex processes such as protein folding, conformational flexibility, ligand binding, and membrane transport. Basically, MD offers a dynamic complement to static structural data obtained from experimental techniques like X-ray crystallography and cryo-electron microscopy. By pushing these static structures through MD simulations, researchers can observe how interatomic forces influence atoms, revealing the full dynamic and kinetic landscape of biological molecules [3, 4].

Because biomolecular systems operate across timescales ranging from femtoseconds for bond vibrations to milliseconds or longer for conformational rearrangements, MD enables a direct connection between fast quantum mechanical processes and slow biological behavior [5]. Its capacity to capture both structure and dynamics has established MD as a leading technique in modern structural biology, pharmacology, and materials biochemistry.

1.2 Historical Background

The origins of molecular simulations trace back to the rise of digital computing in the 20th century. Monte Carlo (MC) methods were introduced by Metropolis *et al.* (1953) [6] to estimate the thermodynamics of molecular configurations according to statistical mechanics ensembles. Around the same time, Alder and Wainwright (1957) [7] performed the first atomistic molecular dynamics simulations of hard-sphere fluids, demonstrating that macroscopic properties such as phase transitions could emerge directly from microscopic motion. Rahman (1964) [8] extended this approach to liquids, using the Lennard-Jones potential for liquid argon, while Verlet (1967) [9] introduced an efficient integration algorithm for Newton's equations that remains foundational today. Rahman and Stillinger (1971) [10] later achieved a landmark simulation of liquid water, revealing the dynamic hydrogen-bond network that characterizes its liquid state.

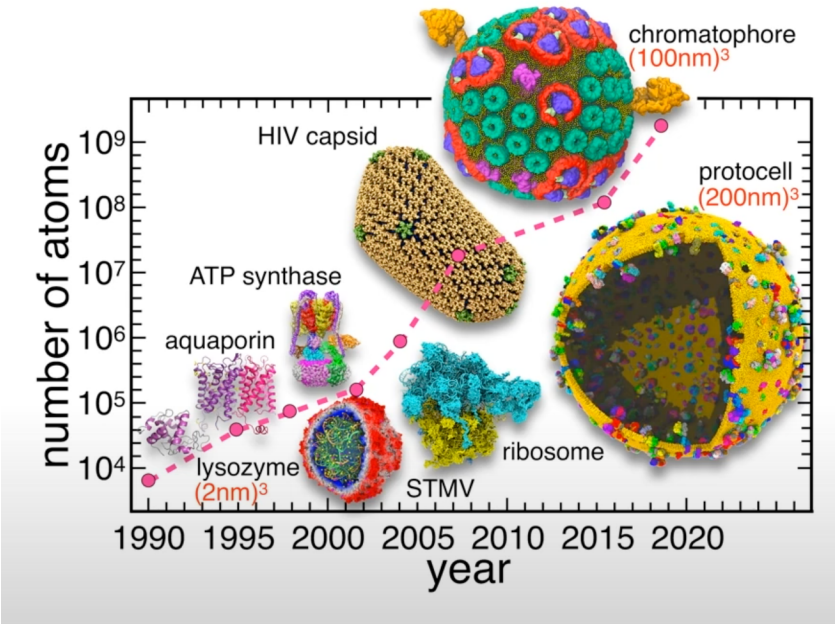

Figure 1.1: **MD evolution in time.** Figure derived from AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics

By the mid-1970s, MD enters the field of biomolecular science. McCammon, Gelin, and Karplus (1977) [11] reported the first simulation of a protein—bovine pancreatic trypsin inhibitor—showing that atomic fluctuations and conformational flexibility could be captured computationally. This breakthrough was soon supported by the development of empirical biomolecular force fields, including AMBER [12], CHARMM [13], and GROMOS [14], which enabled realistic simulations of peptides and nucleic acids. Continued advances in algorithms, hardware, and parallel computing extended timescales from picoseconds to microseconds and beyond, exemplified by the work of Shaw *et al.* (2010) [15], firmly establishing MD as an indispensable tool in computational chemistry and biophysics (Fig. 1.1).

1.3 Theoretical Foundations

1.3.1 Classical Mechanics

At its core, molecular dynamics (MD) is grounded in classical mechanics. Each atom i is treated as a point mass m_i subject to forces derived from a potential energy function U. The time evolution of the system follows Newton's equations of motion:

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i U(r_1, r_2, \ldots, r_N) \tag{1.1}$$

where r_i is the position vector of atom i. The potential energy U is typically an empirical function representing bonded and non-bonded interactions (see Section 1.4). Integrating these equations numerically over small time steps (~ 1-2 fs) yields a discrete trajectory $\{r_i(t), v_i(t)\}$ that approximates the continuous motion of atoms.

Classical mechanics assumes that atomic nuclei follow deterministic trajectories on a continuous potential energy surface (Fig. 1.2). Quantum effects such as zero-point energy, tunneling, and electronic excitations are generally neglected, which is reasonable for heavy atoms at physiological temperatures but may be significant for light atoms like hydrogen or at very low temperatures [3].

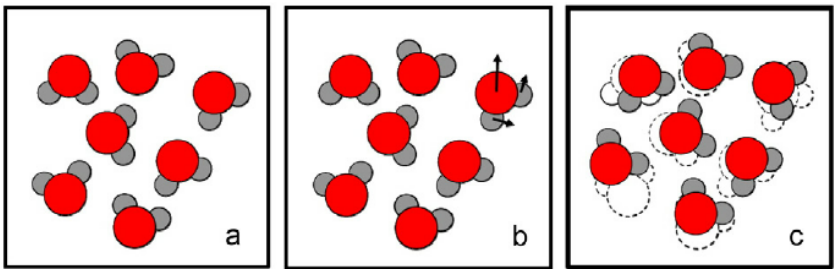

Figure 1.2: **Schematic representation of the molecular dynamics simulation process.** (**a**) Initial atomic positions and velocities are specified; (**b**) Forces acting on each atom are calculated; (**c**) Atomic positions are updated iteratively to generate the molecular trajectory [16].

Alternatively, MD can be formulated in the Hamiltonian framework:

$$H(p, r) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + U(r_1, r_2, \ldots, r_N) \tag{1.2}$$

where p_i is the momentum of atom i. Hamilton's equations provide an equivalent formulation of motion, useful for symplectic integration schemes that conserve energy over long trajectories [17].

The time evolution of the system is generated numerically by integrating Newton's equations of motion. Symplectic and time-reversible algorithms, such as the Verlet family of integrators (see Section 1.5), ensure stable propagation of atomic positions and velocities over long trajectories. A schematic representation of the velocity Verlet algorithm is shown in Figure 1.3.

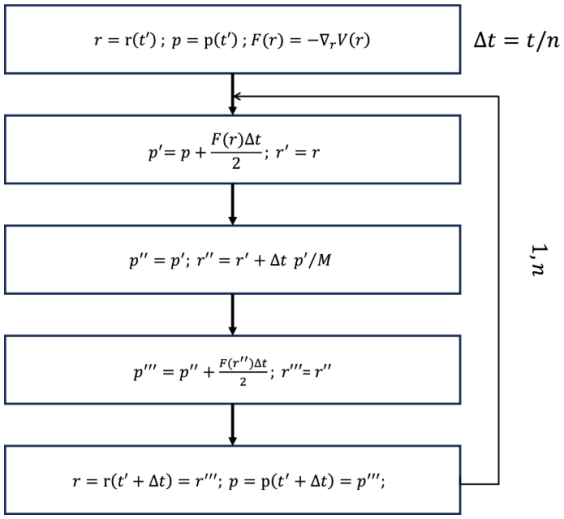

Figure 1.3: **Flowchart of the velocity Verlet algorithm for molecular dynamics simulations.** This scheme updates positions and velocities in a time-reversible and symplectic manner, ensuring stable and accurate integration over time [18].

1.3.2 Statistical Mechanics Connection

Although MD is deterministic at the microscopic level, its goal is to recover an ensemble of observables. According to statistical mechanics, the expectation value of any property A in an ensemble is given by:

$$\langle A \rangle = \int A(r^N, p^N)\rho(r^N, p^N)dr^N dp^N \tag{1.3}$$

where ρ is the phase-space probability distribution. In practice, MD replaces the ensemble average with a time average over a single trajectory, invoking the **ergodic hypothesis**—the assumption that over sufficient time, the system explores all accessible microstates consistent with the ensemble [17, 19]. This allows properties such as temperature, pressure, and radial distribution functions to be computed from a single long simulation.

Depending on the choice of integrator and thermostat/barostat algorithms, MD simulations can approximate different ensembles (NVE, NVT, NPT), connecting the microscopic trajectories to macroscopic thermodynamic conditions (see Section **??**). The trajectories also yield structural, energetic, and kinetic quantities such as radial distribution functions, RMSD,

hydrogen-bond lifetimes, diffusion coefficients, and free-energy landscapes. These observables provide molecular-level explanations for experimentally measured properties such as binding affinities, rate constants, and conformational equilibria [1].

1.4 Force Fields for Biomolecular Systems

In molecular dynamics simulations, the **force field** defines the potential energy function that governs interactions between atoms. Accurate force fields are essential for reproducing experimental structures, dynamics, and thermodynamic properties of proteins, nucleic acids, and biomolecular complexes [2, 20]. They provide the forces used to propagate the system in time and form the foundation for meaningful simulation results.

1.4.1 Components of Biomolecular Force Fields

The total potential energy of a biomolecular system is typically decomposed into **bonded** and **non-bonded** contributions:

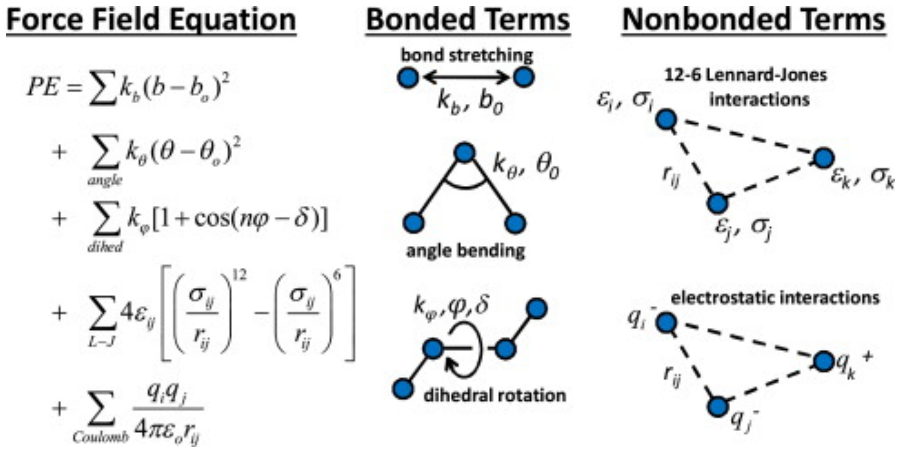

Figure 1.4: **Summary of bonded and non-bonded terms in biomolecular force fields.** Bonded terms include bond stretching, angle bending, and dihedral rotations. Non-bonded interactions include van der Waals and electrostatic interactions. This visual summarizes the forces that govern biomolecular dynamics [21].

Bonded Terms

Bonded interactions describe atoms connected by covalent bonds and maintain the molecular geometry. The most standard interactions described within the force fields parameters are:

- **Bond stretching** - atoms are held near equilibrium distances using harmonic potentials. Bond length equilibria are determined by the types of atoms involved in the bond and the bond type itself (single, double, triple). For example a single $C-H$ bond is always

around ~1.09Å. Since the bond resists streching or compressing, any deviation from standard values will increase the system's energy captured by the harmonic potential:

$$U_{bond} = \sum_{bonds} k_b(r - r_0)^2$$

where r_0 is the equilibrium bond length and k_b is the force constant.

- **Angle bending** - maintains bond angles. Three bonded atoms $i-j-k$ form a θ angle at the central atom i. Stereochemically, those angles tend to stay close to their equilibrium values. For example the $H-O-H$ bond angle in water is ~104.5°. Deviating from this angle will increase the energy of the system:

$$U_{angle} = \sum_{angles} k_\theta(\theta - \theta_0)^2$$

where θ_0 is the equilibrium bond angle and k_θ is the force constant.

- **Dihedral (torsional) rotations** - control rotation around bonds. A dihedral angle (or torsion angle) is defined by the angle between the planes formed by $i-j-k$ and $j-k-l$ atoms. This term is indeed crucial for backbone and side-chain conformations, since the rotations around a single bond affect the 3D shape of molecules. In proteins the ϕ and ψ dihedral angles define the local secondary structure. Certain angles favor α-helices and others β-sheets. Non-favorable rotations, are stretching the "spring" increasing the energy of the system:

$$U_{dihedral} = \sum_{dihedrals} \frac{1}{2} V_n[1 + \cos(n\phi - \delta)]$$

where ϕ the dihedral angle between the planes $i-j-k$ and $j-k-l$, captures rotations around the central bond $j-k$. V_n is the cost of deviation from preferred angles and n denotes the **periodicity** of the torsion (how many minima exist in a full rotation). Finally, δ shifts the phase of the harmonic to match the lowest-energy dihedral at the physically correct angle.

- **Improper torsions / out-of-plane terms** - maintain planarity in groups like aromatic rings, peptide bonds, or carbonyl groups, that should always lie in plane. This term adds a "spring" between the plane defined by three atoms and the fourth atom, which is tethered above or below that plane. Stretching or twisting that atom, hence the spring, increases the energy preventing it from leaving the plane and preserving correct stereochemistry. Similarly, at tetrahedral centers-like most carbons in amino acids-the 3D arrangement of substituents must remain in an R or S configuration, which can be disturbed by atom flipping.

However, there are more sophisticated fields that involve higher-order corrections and coupling terms to account for molecular flexibility and an-harmonicity. The **Bond-Bond and**

Bond-Angle Coupling is sensitive to the influences of angles and bonds to neighboring chemical bonds, while the **Urey-Bradley** term adds a spring at the base angle defined by 3 atoms, often referred to as 1,3 nonbonded interactions. The **CMAP (Correction Map)** term adds a smooth bias towards favorable combinations of $\phi - \psi$ angles, precomputed from QM calculations.

Non-Bonded Terms

Non-bonded interactions govern how atoms interact over distance:

- **Van der Waals interactions** - modeled with the Lennard-Jones potential. The repulsive term $(\frac{\sigma}{r})^{12}$ dominates at very short distances, due to overlapping electron clouds, following Pauli's exclusion principle. While the attraction term $-(\frac{\sigma}{r})^{6}$ dominates at longer distances, following the London's dispersion forces law of temporary dipoles created by the constant motion of electrons between atoms:

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

  Non-bonded interactions, especially van der Waals forces, decay with distance. To reduce computational cost, a **cutoff radius** r_c is applied, beyond which interactions are neglected:

  $$U_{vdW} = \sum_{i<j}^{r_{ij}<r_c} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{1.4}$$

- **Electrostatic interactions** — described by Coulomb's law, these forces govern charge–charge interactions crucial for hydrogen bonds, salt bridges, and solvation effects [20]:

  $$U_{elec} = \sum_{i<j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}$$

  Electrostatic interactions are inherently long-range and cannot be truncated without introducing artifacts. To efficiently compute them under periodic boundary conditions (see Section 1.5), modern force fields employ **Ewald summation** or its faster variant, the **Particle-Mesh Ewald (PME)** method [22]. PME separates electrostatic contributions into a short-range real-space term and a long-range reciprocal-space term, enabling accurate and scalable treatment of charged biomolecular systems.

Some older force fields (AMBER94) introduced explicit hydrogen bond potentials, while newer versions rely on electrostatics and Van der Waals to capture hydrogen bonds. More modern refined force fields also add dispersion corrections in the Lennard-Jones potential, since it is has been shown that it can underestimate attractive dispersion interactions, especially in

polarizable or large systems [23–25].

1.4.2 Solvent Models

Solvent plays a critical role in stabilizing protein structures, mediating hydrogen bonding, influencing electrostatics, and enabling ligand binding. Water and other solvents can be represented either explicitly or implicitly:

- **Explicit solvent** - atomistic water molecules. The solvent, typically water, is explicitly represented by its atoms and the interactions between solvent-system lies within the classical bonded/non-bonded terms of the force field. The landmark model for water is **TIP3P**. A three-site rigid water model widely used in AMBER [12] and CHARMM [13] force fields [26]. Similarly the **SPC/E** is an extenstion of the **TIP3P** model though efficient for large systems with improved dielectric properties [27]. The newest **TIP4P/ TIP4P-Ew** is a four-site model with improved water density and electrostatics reproduction. This model adds a "virtual site" localizing the negative charge to realize the dipole moment and hydrogen-bonding behavior. Its realistic solvation energetics encourages studies like protein unfolding/denaturation, especially in high-temperatures [28–30]. Explicit solvation ensures realistic structural dynamics but increases computational cost due to the large number of water molecules required to fully solvate the protein (typically 10-15 Å padding around the solute).
- **Implicit solvent** - approximates the solvent as a continuum. Unlike the explicit solvent, where every water molecule is represented atom by atom, the solvent's effects on the solute are averaged out as electrostatic and hydrophobic potentials. Electrostatic interactions are usually described by the **Generalized Born (GB) approximation** or the **Poisson-Boltzmann (PB)** equation and capture the interplay between solute's charges and solvent. The hydrophobic effects, explain the cost of energy need to create a hydrophobic cavity in presence of a solvent, are often approximated by the **solvent-accessible sufrace area (SA)**. By simplifying the solvation representation, these models significantly reduce computational costs and are useful for rapid conformational sampling of proteins or peptides. However, they lack explicit representation of solvent structure and dynamics, which can be critical for processes like water-meditated interactions and dynamic hydration effects related to folding and ligand binding [31].

1.4.3 Parameterization and Limitations

Force fields are empirical approximations calibrated against experimental or quantum mechanical data. Table 1.1 exhibits several widely used force fields for biomolecular systems, highlighting their target molecules, key features, and typical water models. Most classical force fields rely on fixed atomic charges and simplified functional forms for bonded and non-bonded interactions, which neglect electronic polarization and only approximate long-range

effects. As a result, transferability across different chemical environments can be limited, and certain interactions, such as hydrogen bonds and solvent-mediated effects, may not be fully captured. Polarizable force fields, such as the Drude model, explicitly account for the electronic response to the environment, partially mitigating these issues, but at a higher computational cost [32].

Table 1.1: Comparison of popular biomolecular force fields.

Force Field	Target Molecules	Key Features / Strengths	Limitations	Water Model
AMBER (ff14SB)[33]	Proteins, nucleic acids	Improved backbone dihedrals, well-validated for proteins	Fixed charges, limited transferability	TIP3P, SPC/E
CHARMM36[34]	Proteins, lipids, nucleic acids	Updated side-chain torsions, lipid support	Fixed charges, approximate long-range interactions	TIP3P, TIP4P
OPLS-AA[35]	Organic molecules, biomolecules	Optimized for proteins and small molecules	Fixed charges, limited transferability	TIP3P, SPC
Drude / Polarizable [32]	Proteins, nucleic acids	Electronic polarization, improved electrostatics	Computationally expensive	TIP3P, polarizable water models

Force fields form the backbone of biomolecular MD simulations. Choosing an appropriate force field and solvent model is critical for reproducing realistic dynamics, conformational equilibria, and interactions. Understanding the assumptions, parameterization, and limitations is essential for interpreting MD results in a biologically meaningful way.

1.5 Integration Algorithms and Simulation Setup

Molecular dynamics simulations of biomolecules require both numerical integration of the equations of motion and careful system preparation to ensure physically meaningful trajectories. This section describes the algorithms used to propagate atomic positions, the constraints and timestep considerations, and the practical steps for setting up protein or biomolecular simulations.

1.5.1 Integration Algorithms

MD relies on solving Newton's equations of motion for each atom (Eq. 1.1). Since analytical solutions are impossible for all but the simplest systems, finite-difference algorithms are used to integrate the equations numerically.

1. **Verlet Algorithm**
   The classical **Verlet algorithm** computes new positions using current and previous positions:

$$r_i(t+\Delta t) = 2r_i(t) - r_i(t-\Delta t) + a_i(t)\Delta t^2 \tag{1.5}$$

   where $a_i(t) = F_i(t)/m_i$ is the acceleration derived from the force F_i and δt the integration timestep. Thus, the new position is derived from the combination of current position and the momentum of the previous step, allowing for a smoother motion that

autocorrects small errors in velocity or acceleration. Verlet integration is time-reversible and conserves energy well over long trajectories with small time steps (~$1-2\,fs$), making it suitable for NVE ensemble simulations [9].

2. **Velocity-Verlet Algorithm**
   The **velocity-Verlet** algorithm extends the Verlet scheme by explicitly propagating velocities:

$$r_i(t+\Delta t) = r_i(t) + v_i(t)\Delta t + \frac{1}{2}a_i(t)\Delta t^2 \tag{1.6}$$

$$v_i(t+\Delta t) = v_i(t) + \frac{1}{2}[a_i(t) + a_i(t+\Delta t)]\Delta t \tag{1.7}$$

   In other words, it computes the new "predicted position" exactly like the Verlet algorithm. Then it updates using the velocity using the average of the current and new acceleration, which helps cancel numerical errors especially in small time steps. This allows easy computation of kinetic energy and temperature control, essential for canonical (NVT) and isothermal-isobaric (NPT) simulations [17].

3. **Leapfrog Algorithm**
   The **leapfrog algorithm** updates velocities at half-integer time steps, “leaping” over position updates.

$$v_i(t+\frac{\Delta t}{2}) = v_i(t-\frac{\Delta t}{2}) + a_i(t)\Delta t \tag{1.8}$$

$$r_i(t+\Delta t) = r_i(t) + v_i(t+\frac{\Delta t}{2})\Delta t \tag{1.9}$$

   This calculates the velocity "leapfrogged" forward to the next half-step, used to update the position at the next full integer step. This staggering improves numerical stability and keeps energy fluctuations small.

1.5.2 Constraints in Biomolecular Systems

The choice of timestep δt is constrained by the fastest vibrational motions in the system, typically bond stretching involving hydrogen atoms. For all-atom protein simulations, a timestep of 1-2 fs is standard. By applying bond constraints (e.g., SHAKE [36] or LINCS [37]), the highest-frequency bond vibrations are fixed, allowing slightly larger timesteps (2 fs) while maintaining numerical stability [36]. Many biomolecular simulations employ constraint algorithms to maintain rigid bond lengths or angles, reducing computational cost and improving stability:

- **SHAKE** - iteratively adjusts bond lengths to satisfy constraints at each time step. While a simple method and widely used, this iterative process of can be really slow for big systems, because adjusting one bond may affect others and SHAKE needs multiple rounds of corrections until all bonds are within tolerance [36].
- **LINCS** - a matrix-based constraint algorithm that unlike SHAKE, solves all constraints simultaneously. Although, this algorithm is more complex mathematically it can be faster and more stable for large systems [37].

Additionally, hydrogen mass repartitioning (HMR) redistributes mass from heavy atoms to bonded hydrogens, slowing down the fastest bond vibrations. This allows the use of larger timesteps (e.g., 4 fs) in combination with bond constraints like SHAKE or LINCS, improving computational efficiency without sacrificing stability [38].

Constraints are particularly important for X-H bonds because hydrogen vibrations occur on femtosecond timescales, which would otherwise require very small timesteps.

1.5.3 Simulation Setup Workflow

Proper setup of the biomolecular system is essential to obtain meaningful MD trajectories. The typical workflow includes:

1. **Preparation of Protein and Ligand Structures**

   (a) **Protein structures** are usually obtained from the Protein Data Bank (PDB) and processed to remove missing atoms, alternate conformations, and crystallographic water molecules (if not needed).

   (b) **Protonation states** are assigned based on pH using tools like rdkit/obabel [39, 40].

   (c) **Ligands or cofactors** are parameterized using compatible force field parameters.

2. **Solvation and Neutralization**
   Biomolecules are embedded in a solvent environment, typically explicit water models such as TIP3P, TIP4P, or SPC/E. A solvent box is generated with sufficient padding (usually ≥ 10Å) around the solute. Finite simulation boxes introduce artificial surfaces that can produce nonphysical behavior. **Periodic boundary conditions (PBCs)** mitigate these effects by surrounding the primary simulation box with infinite replicas of itself in all directions [17]. Under PBCs, an atom leaving one face of the box reenters from the opposite face, creating the illusion of an **infinite bulk system**. Distances between particles are computed using the **minimum image convention**, considering the nearest periodic image:

$$r_{ij} = \min(|r_i - r_j + nL|) \tag{1.10}$$

where L is the box length and n a vector of integers selecting the closest image. PBCs are essential for maintaining **bulk density** and **thermodynamic consistency** in biomolecular simulations. Ions (Na^+, Cl^-) are added to neutralize the system and mimic physiological ionic strength. Proper solvation and ion placement are crucial for stabilizing electrostatic interactions and reproducing experimental observables, such as protein folding equilibria and ligand binding affinities.

3. **Energy Minimization**
   Prior to dynamics, the system is energy-minimized to remove steric clashes or bad contacts. Minimization algorithms include steepest descent and conjugate gradient methods, which reduce potential energy to a local minimum without altering the overall structure [1, 19].

4. **Equilibration**
   After minimization, the system is gradually heated to the target temperature (e.g., 300 K). Usually the **canonical ensemble (NVT)** is employed which holds constant the number of particles N, volume V and temperature T. This thermostat (e.g., Langevin [41], Berendsen [27], Nose-Hoover [42, 43]) automatically reassigns atoms velocities at each time step so the instanteous temperature approaches the target value. The second phase of equilibration employs the **isothermal-isobaric ensemble (NPT)** (e.g., Berendsen [27], Parrinello-Rahman [44]) to rescale the box volume smoothly towards the target pressure. Equilibration ensures thermal stability and density adjustment before production runs. Restraints are often applied to heavy atoms initially and gradually removed.

   Statistical ensembles provide the theoretical framework linking atomic-level MD trajectories to macroscopic thermodynamic observables. Thermostats and barostats are essential for maintaining biologically relevant temperature and pressure conditions. The correct choice and implementation of these algorithms ensure that simulations of proteins and biomolecules are both physically realistic and statistically consistent with experimental systems.

5. **Production Simulation**
   Once equilibrated, unrestrained MD is performed for a sufficient duration to sample relevant protein motions, which may range from tens of nanoseconds to microseconds, depending on computational resources and the biological question. Trajectories are stored periodically for post-processing and analysis of structural, energetic, and dynamic properties.

1.6 Sampling, Equilibration, and Data Analysis

Molecular dynamics simulations produce detailed atomic trajectories that reflect the conformational dynamics of proteins and other biomolecules. However, extracting biologically meaningful information requires careful consideration of sampling adequacy, system equilibration, and trajectory analysis techniques [4, 5].

1.6.1 Confomrational Challenges

Proteins exhibit complex energy landscapes with multiple local minima separated by energy barriers. Conventional MD simulations may become trapped in a local minimum, preventing exploration of functionally relevant conformations. This issue is particularly acute for large proteins, slow domain motions, and rare events such as folding or ligand binding [3].

The total number of accessible microstates grows exponentially with system size, making ergodic sampling computationally challenging. Consequently, simulations must be long enough to adequately explore conformational space or employ enhanced sampling methods (see Section **??**).

1.6.2 Trajectory Analysis

1. **Structural Analysis**

   - **Root-Mean-Square Deviation (RMSD)** - measures the average deviation of atomic positions from a reference structure, indicating overall conformational changes:

   $$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i(t) - r_i^{ref})^2} \tag{1.11}$$

   In protein trajectories, usually we compare all the frames with the reference $t = 0$ conformation. Big changes in RMSD indicate either unfolding, high flexibility or exploration of alternative conformations. Typically, a simulation ends when RMSD plateaus, indicating the stability of the overall system.

   - **Root-Mean-Square Fluctuation (RMSF)** - quantifies per-residue flexibility by calculating the time-averaged deviation of each atom from its mean position:

   $$RMSF(i) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i(t) - \langle r_i \rangle)^2} \tag{1.12}$$

   Apart from the structuraly flexibility of regions like loops, termini or surface residues, or full mobile domains, RMSF could also highlight functional residues involved in binding, catalysis or conformational transitions. Comparing RMSF between different simulations (e.g. apo vs ligand-bound protein) helps accentuate stabilized or destabilized regions upon binding or releasing the ligand.

   - **Radius of Gyration (Rg)** - assesses the compactness of the protein structure over time:

   $$R_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_{cm})^2} \tag{1.13}$$

where r_{cm} is the center of mass of the protein. By measuring the average distance of atoms from the molecule's center of mass, it effectively summarizes the spread of the atoms within the structure. Monitoring the R_g over time might emphasize folding/unfolding events, domain movements, or similarly to RMSF, might reveal deviations between binding/unbinding states. A stable R_g is an indicator of a stable conformational change.

- **Secondary Structure Analysis** - tracks changes in secondary structure of a protein, marking folding/unfolding phenomena. Proteins rely on subtle rearrangements of domains for activity and secondary structure analysis might reveal stabilization or destabilization caused by mutations or binding incidents or even environemntal conditions. The most widely used method to analyse the secondary structure is the **Define Secondary Structure of Proteins (DSSP)** algorithm. It studies the hydrogen bonding patterns (energy criteria) and backbone geometry (phi and psi dihedrals) of each structure and outputs a per-residue assignment over the trajectory.

2. **Hydrogen Bond Analysis** Hydrogen bonds are critical for protein stability. Their number, occupancy, and lifetime can be computed using geometric criteria, typically a donor-acceptor distance < 3.5 Å and a donor-H-acceptor angle > 120°. These interactions are essential for maintaining protein secondary structures such as α-helices and β-sheets. The solvent-solute interactions, ligand binding events and mutational effects can be also analyzed by H-bond formation. However, the geometric criteria for defining a hydrogen bond are arbitrary, any deviation in angle or distance could shift the formation. Consequently, the H-bond lifetime of dynamic systems (e.g. water) is short (ps range) and their analysis is inherently statistical.

3. **Principal Component Analysis (PCA)** PCA analysis is a powerful tool in analyzing protein dynamics. Using the 3D atomic coordinates from an MD trajectory, a covariance matrix of atomic fluctuations is computed. By dentifying the **eigenvectors** and **eigenvectors** of this matrix we extract the **dominant collective modes**, that describe large-scale motions such as domain hinge movements. Clustering along the most significant $principalcomponents$ allows the identification of stable states (e.g. open-close protein domains) as well as flexible or random motions.

1.6.3 Enhanced Sampling Techniques

Conventional MD simulations can be limited by high-energy barriers and rare events, preventing full exploration of conformational space. Several enhanced sampling methods have been developed to overcome these limitations:

- **Replica Exchange Molecular Dynamics (REMD)** - Multiple replicas of the system are simulated simultaneously at different temperatures, allowing periodic exchange of configurations between replicas [45]. High-temperature replicas explore high-energy

regions more efficiently, while low-temperature replicas sample stable states accurately. This method improves sampling efficiency and helps overcome kinetic traps in protein folding, ligand binding, and other conformational transitions.

- **Metadynamics** - A history-dependent bias potential is added along selected collective variables (CVs) to discourage the system from revisiting already sampled states [46]. Over time, this fills the free energy wells and allows the system to escape local minima, enabling the exploration of rare events, conformational changes, and estimation of free energy surfaces.

- **Accelerated Molecular Dynamics (aMD)** - The potential energy surface is modified by adding a boost potential that reduces energy barriers above a chosen threshold [47]. This allows the system to cross barriers more easily and explore conformational space more rapidly while maintaining correct Boltzmann-weighted statistics after reweighting. aMD is particularly useful for studying slow processes such as protein folding, ligand unbinding, or large domain motions.

Enhanced sampling techniques are crucial for obtaining accurate thermodynamic and kinetic information in biomolecular systems, enabling the study of processes that are often inaccessible to conventional MD timescales.

1.7 Conclusion

Molecular dynamics (MD) simulations are a powerful computational tool for investigating the structural, dynamical, and thermodynamic properties of biomolecular systems. By numerically integrating Newton's equations of motion for atoms under the influence of empirically derived force fields, MD enables detailed exploration of protein conformational dynamics, solvent interactions, and ligand binding events over nanosecond to microsecond timescales.

This section has presented a comprehensive overview of MD, beginning with the theoretical foundations, including force fields and potential energy functions. The decomposition of interactions into bonded and non-bonded terms allows classical MD to approximate complex molecular forces with sufficient accuracy for most biochemical applications. Popular biomolecular force fields, such as AMBER and CHARMM, provide parameter sets tailored to proteins, nucleic acids, lipids, and small molecules, balancing accuracy with computational efficiency.

The chapter also highlighted the numerical integration schemes used to propagate atomic motions, including Verlet, velocity-Verlet, and leapfrog algorithms, as well as the use of constraints (SHAKE/LINCS) to stabilize fast vibrational modes. Proper system preparation, including protein and ligand modeling, solvation, ion placement, energy minimization, and equilibration, was emphasized as essential for producing reliable simulation trajectories.

Maintaining physically meaningful thermodynamic conditions requires careful selection of statistical ensembles and thermostats/barostats. NVE, NVT, and NPT ensembles provide frameworks for isolating energy, controlling temperature, and regulating pressure, respectively. Thermostats such as Nos*é* or Langevin and barostats like Parrinello-Rahman enable realistic modeling of biomolecular systems under experimental conditions.

Accurate representation of the solvent environment and boundary effects was discussed in Section 1.4. Explicit solvent models capture hydrogen bonding, solvation shells, and dynamic interactions with the protein, while implicit solvent models offer computational efficiency. Periodic boundary conditions (PBCs) and long-range electrostatics methods such as Particle-Mesh Ewald ensure that finite simulation boxes mimic bulk behavior, maintaining thermodynamic consistency.

Conformational sampling and data analysis are central to extracting biologically meaningful information from MD trajectories. Metrics such as RMSD, RMSF, hydrogen bond occupancy, secondary structure content, and principal component analysis provide insight into protein stability, flexibility, and functional motions. When conventional MD is insufficient, enhanced sampling techniques like replica exchange MD, metadynamics, and accelerated MD enable exploration of rare events and transitions between metastable states.

In summary, molecular dynamics simulations provide an indispensable framework for studying protein dynamics and biomolecular interactions with atomic-level detail. By integrating accurate force fields, robust integration algorithms, thermodynamic control, and solvent modeling, MD bridges the gap between structural biology and functional understanding, offering insights that are often inaccessible to experimental techniques alone. The combination of conventional MD with enhanced sampling and hybrid methods ensures that simulations remain a versatile and evolving tool in computational biochemistry.

References

[1] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[2] Alex D MacKerell Jr et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins". In: *The journal of physical chemistry B* 102.18 (1998), pp. 3586–3616.

[3] M. Karplus and J. A. McCammon. "Molecular dynamics simulations of biomolecules". In: *Nature Structural Biology* 9.9 (2002), pp. 646–652. DOI: 10.1038/nsb0902-646.

[4] S. A. Hollingsworth and R. O. Dror. "Molecular dynamics simulation for all". In: *Neuron* 99.6 (2018), pp. 1129–1143. DOI: 10.1016/j.neuron.2018.08.011.

[5] R. O. Dror et al. "Biomolecular simulation: A computational microscope for molecular biology". In: *Annual Review of Biophysics* 41 (2012), pp. 429–452. DOI: 10.1146/annurev-biophys-042910-155245.

[6] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[7] Berni J Alder and Thomas Everett Wainwright. "Studies in molecular dynamics. I. General method". In: *The Journal of Chemical Physics* 31.2 (1959), pp. 459–466.

[8] A. Rahman. "Correlations in the Motion of Atoms in Liquid Argon". In: *Phys. Rev.* 136 (2A 1964), A405–A411. DOI: 10.1103/PhysRev.136.A405. URL: https://link.aps.org/doi/10.1103/PhysRev.136.A405.

[9] L. Verlet. "Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules". In: *Physical Review* 159.1 (1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98.

[10] A. Rahman and F.H. Stillinger. "Molecular dynamics study of liquid water". In: *The Journal of Chemical Physics* 55.7 (1971), pp. 3336–3359.

[11] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. "Dynamics of folded proteins". In: *nature* 267.5612 (1977), pp. 585–590.

[12] S.J. Weiner et al. "New force field for molecular mechanical simulation of proteins and nucleic acids". In: *Journal of the American Chemical Society* 106.3 (1984), pp. 765–784.

[13] B.R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217.

[14] W.F. van Gunsteren and H.J.C. Berendsen. "GROMOS: A general molecular simulation program". In: *Biomolecular Simulation* 14 (1984), pp. 43–59.

[15] David E Shaw et al. "Atomic-level characterization of the structural dynamics of proteins". In: *Science* 330.6002 (2010), pp. 341–346.

[16] Rebecca Notman and Jamshed Anwar. "Breaching the skin barrier—Insights from molecular simulation of model membranes". In: *Advanced drug delivery reviews* 65.2 (2013), pp. 237–250.

[17] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. 2nd. Oxford University Press, 2017.

[18] Giovanni Ciccotti, Sergio Decherchi, and Simone Meloni. "Foundations of molecular dynamics simulations: how and what". In: *La Rivista del Nuovo Cimento* (2025), pp. 1–94.

[19] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. 2nd. Academic Press, 2002.

[20] David A. Case et al. "AMBER 2020". In: *University of California, San Francisco* (2020). URL: https://ambermd.org/Amber20.php.

[21] Robert A Latour. "Perspectives on the simulation of protein–surface interactions using empirical force field methods". In: *Colloids and Surfaces B: Biointerfaces* 124 (2014), pp. 25–37.

[22] T. Darden, D. York, and L. Pedersen. "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems". In: *The Journal of Chemical Physics* 98.12 (1993), pp. 10089–10092. DOI: 10.1063/1.464397.

[23] Pengfei Li, Lin Frank Song, and Kenneth M Merz Jr. "Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water". In: *The Journal of Physical Chemistry B* 119.3 (2015), pp. 883–895.

[24] Paolo Nicolini, Elvira Guardia, and Marco Masia. "Shortcomings of the standard Lennard-Jones dispersion term in water models, studied with force matching". In: *The Journal of Chemical Physics* 139.18 (2013), p. 184111. DOI: 10.1063/1.4829444.

[25] Chetan R Rupakheti, Alexander D MacKerell Jr, and Benoit Roux. "Global optimization of the Lennard-Jones parameters for the Drude polarizable force field". In: *Journal of chemical theory and computation* 17.11 (2021), pp. 7085–7095.

[26] W. L. Jorgensen, J. Chandrasekhar, and J. D. Madura. "Comparison of simple potential functions for simulating liquid water". In: *J. Chem. Phys.* 79 (1983), pp. 926–935.

[27] Herman JC Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[28] Dietmar Paschek, Ryan Day, and Angel E Garcia. "Influence of water–protein hydrogen bonding on the stability of Trp-cage miniprotein. A comparison between the TIP3P and TIP4P-Ew water models". In: *Physical Chemistry Chemical Physics* 13.44 (2011), pp. 19840–19847.

[29] Agusti Emperador, Ramon Crehuet, and Elvira Guardia. "Effect of the water model in simulations of protein–protein recognition and association". In: *Polymers* 13.2 (2021), p. 176.

[30] Changwon Yang, Soonmin Jang, and Youngshang Pak. "A fully atomistic computer simulation study of cold denaturation of a β-hairpin". In: *Nature communications* 5.1 (2014), p. 5773.

[31] Michael Feig and Charles L Brooks III. “Recent advances in the development and application of implicit solvent models in biomolecule simulations”. In: *Current opinion in structural biology* 14.2 (2004), pp. 217–224.

[32] Justin A. Lemkul, Benoit Roux, and Alexander D. MacKerell. “An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications”. In: *Chemical Reviews* 116.9 (2016), pp. 4983–5013.

[33] James A. Maier et al. “ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB”. In: *Journal of Chemical Theory and Computation* 11.8 (2015), pp. 3696–3713.

[34] Jing Huang and Alexander D. MacKerell. “CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data”. In: *Journal of Computational Chemistry* 34.25 (2013), pp. 2135–2145.

[35] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. “Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids”. In: *Journal of the american chemical society* 118.45 (1996), pp. 11225–11236.

[36] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. “Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes”. In: *Journal of Computational Physics* 23.3 (1977), pp. 327–341. DOI: 10.1016/0021-9991(77)90098-5.

[37] B. Hess et al. “LINCS: A linear constraint solver for molecular simulations”. In: *Journal of Computational Chemistry* 18.12 (1997), pp. 1463–1472. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[38] Chad W Hopkins et al. “Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning”. In: *Journal of Chemical Theory and Computation* 11.4 (2015), pp. 1864–1874. DOI: 10.1021/ct5010406. URL: https://doi.org/10.1021/ct5010406.

[39] *RDKit: Open-source cheminformatics*. https://www.rdkit.org. Accessed: 2025-10-15.

[40] Noel M O’Boyle et al. “Open Babel: An open chemical toolbox”. In: *Journal of cheminformatics* 3.1 (2011), p. 33.

[41] J. A. Izaguirre et al. “Langevin stabilization of molecular dynamics”. In: *The Journal of Chemical Physics* 114.5 (2001), pp. 2090–2098. DOI: 10.1063/1.1332996.

[42] S. Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519. DOI: 10.1063/1.447334.

[43] W. G. Hoover. “Canonical dynamics: Equilibrium phase-space distributions”. In: *Physical Review A* 31.3 (1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.

[44] M. Parrinello and A. Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190. DOI: 10.1063/1.328693.

[45] Y. Sugita and Y. Okamoto. “Replica-exchange molecular dynamics method for protein folding”. In: *Chemical Physics Letters* 314.1-2 (1999), pp. 141–151. DOI: 10.1016/S0009-2614(99)01123-9.

[46] A. Laio and M. Parrinello. “Escaping free-energy minima”. In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.

[47] Donald Hamelberg, John Mongan, and J Andrew McCammon. “Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules”. In: *The Journal of chemical physics* 120.24 (2004), pp. 11919–11929.