University of Warsaw
Faculty of Physics

Jan Malinowski
Record book number: 372202

Comparison of conformational variability within protein kinase catalytic subunits based on molecular dynamics simulations

Master's thesis
in the field of Applications of Physics in Biology and Medicine

The thesis written under the supervision of
Dr Piotr Setny
Centre of New Technologies
University of Warsaw

Warsaw, December 2024

Summary

Kinases constitute family of enzymes responsible for regulation of vast amount of processes inside the cell. Their uniqueness comes from the ability to operate as molecular switches due to reversibility of their activation state. Despite the diversity of their sequences that appeared during the course of evolution, architecture of their active states remained fairly conserved. In contrast, much lower conservation is observed for inactive structures. This observation raises a question, whether similar relations hold with respect to the dynamics exhibited by kinase catalytic subunits. In this work differences between conformational ensembles of these two states were examined based on the example of two tyrozine and two serine/threonine kinase structures. Global characteristics of the dynamics were analyzed by means of principal component analysis alongside with RMSD comparison. To capture differences on local level, RMSF and Jensen-Shannon divergence measures were used. Through their analysis, it was shown that differences between family members are proportional to their evolutionary distance. Furthermore, active states' dynamics occured to be more conserved likely due to the functional role they perform. Finally it was concluded that large scale dynamics is encoded by tertiary protein structure, whereas local dynamic characteristic is dependent on the primary structure and remains unique for each kinase irrespective of its activation state.

Key words

protein kinases, kinase activation, serine-threonine kinases, tyrozine kinases, conformation ensembles, molecular dynamics, PCA, Jensen-Shannon divergence

The title of the thesis in Polish language

Porównanie zmienności konformacyjnej podjednostek katalitycznych różnych kinaz białkowych na podstawie symulacji dynamiki molekularnej

Contents

1. Introduction 4
1.1. Proteins 4
1.2. Protein kinases 5
1.3. The work purpose 7

2. Methodology and system 9
2.1. Molecular dynamics 9
2.2. Descriptors 13
2.2.1. Root-Mean-Square Deviation and Fluctuations (RMSD, RMSF) 13
2.2.2. Principal Components Analysis (PCA) 13
2.2.3. Jensen-Shannon Divergence 14
2.2.4. Statistical hypotheses testing 15
2.3. Systems under study 17
2.3.1. Structures 17
2.3.2. Simulations protocols 18
2.3.3. Analysis protocols 19
2.4. Software and additional informations 22

3. Results and analysis 23
3.1. Stability of simulations 24
3.2. Global descriptors results 26
3.2.1. RMSD analysis 26
3.2.2. Subspace overlap analysis 26
3.2.3. Overall space overlap analysis 28
3.3. Local descriptors results 28
3.3.1. RMSF analysis 28
3.3.2. Jensen-Shannon Divergence analysis 31
3.4. Statistical testing 34

4. Conclusions 35

Chapter 1

Introduction

1.1. Proteins

Proteins are one of the few fundamental types of bricks that constitue life on earth. Inside every cell of each organism there are billions of them. They roles vary from performing DNA multiplication and transporting ions to catalyzing reactions and transferring chemical information. Specific tasks require them to obtain different structures, which is done in two-stage process. Firstly the DNA, consisting of pairs of nucleotides, is transcribed to RNA, which is followed by translation the latter into amino acid chains. In this process different orders of subsequent nucleotide triplets give rise to various sequences of the aforementioned biomolecules. Finally polypeptide chain is bended until final functional structure is achieved. [1]

Protein functions are mainly realised by tertiary structure (or quaternary in case of multiple chain proteins), which describes relative positions of successive secondary structures like α-helices and β-sheets and other important areas in space. Aforementioned fact shows how important is to know the protein structure in order to infer about its function. There are few experimental methods that allow discovering protein structures e.g. NMR, crystallography or cryo-EM, but although by their means structural models with reliable dose of accuracy could be obtained, they can't make detailed insights about dynamic properties of proteins [2–4]. To find out how chains move in dynamic processes like ligand binding or signal transduction, one has to use computational methods one of which is molecular dynamics simulation [5]. This approach allows to extend the research area and in consequence to test new hypothesies. One can for example consider what is the connection between the sequence, structure and dynamic of specific protein. More precisely, does resemblance of structure imply dynamics similarity [6]?

In the course of evolution some ares of the protein chains proved to be more or less prone to changes. It turned out that proteins sharing the same function but different sequences have some conserved areas that were not changed through time. That phenomenon became the basis of natural division of proteins into groups. One example of such group are protein kinases, which are interesting to examine for few reasons. Firstly, it consists of large amount of proteins - in human kinome there are 538 kinases, sharing similar structure with different sequences, what makes the sampling population big enough for different testing purposes. Secondly they possess two modes of action, active and inactive one with differences in some secondary structures locations. Lastly because of their function they are target for many designing drugs, thus enabling possible novel therapies [7–9].

1.2. Protein kinases

Kinases are class of enzymes responsible for catalyzing phosphorylation reaction, that results in transferring the phosphoryl group from ATP into one of the residues of the protein acceptor. This process is important for variety of events such as activation/deactivation, joining/splitting of other proteins or controlling their transportation inside the cell. The importance of kinases comes from the ability to act as the molecular switches, that allows more precise regulation of vast amount of signalling and metabolic pathways. It is worth noting that they cover nearly 2% of human genome, indicating their huge role for all the cell's regulation processes [10].

Catalytic domains of protein kinases consists of two lobes possessing distinct characteristics and named from their corresponding locations in protein chain. Smaller one starts at the N-terminal end and is built mainly from β-sheets, responsible for the ATP bonding and a helical fragment named C-helix. Proper positioning of the latter structure is central to activity regulation. The larger lobe is placed on the C-terminal end, consisting mostly of α-helices and one β-sheet. Helical subdomain here forms the rigid core of the kinase and functions as a tethering area for protein and peptide substrates. Furthermore an important, centrally located F-helix anchors all fragments crucial for kinase's functioning [11]. Although many human kinases evolved independently, their multiple sequence alignment revealed some commonly shared features. These fragments appearing in all kinases occured to code a number of conserved structural motifs. Of particular note are DFG and HRD motifs that are responsible for precise positioning of ATP substrate and transfer of its phosphoryl group to protein acceptor. Important structural parts of kinases were marked on the example of PKA structure model shown in Fig. 1.1.

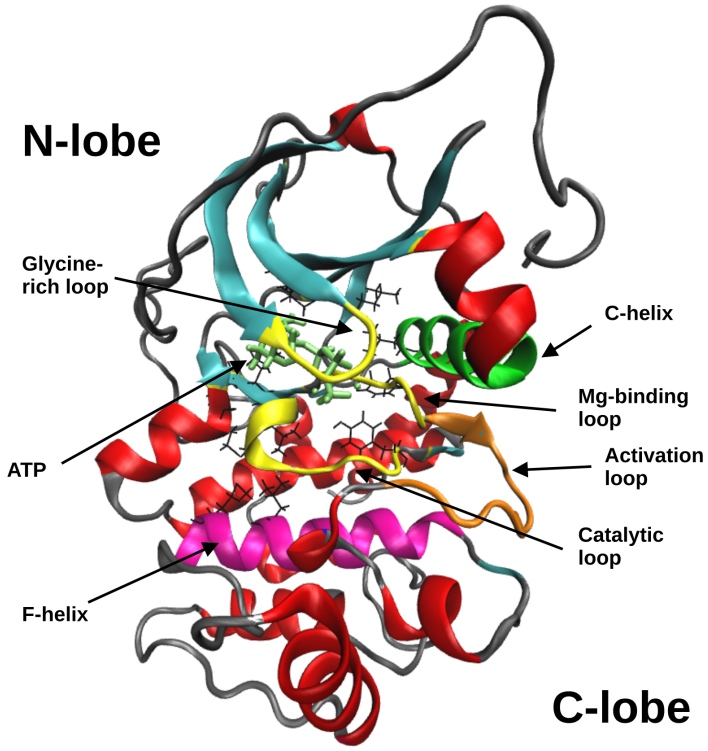


Figure 1.1: Structure of PKA kinase in active state, modelled based on PDB record 3FJQ with important features marked and pointed to. Sidechains of residues forming C- and R-spines were presented with black lines.

Kinases are unique due to their ability to be reversibly activated. This process starts with phosphorylation of activation loop, that structures its previously disorganized form allowing it to anchor to the rest of the protein through attraction with its highly conserved arginine. In consequence DFG motif can be positioned properly, which allows the nucleation of both C- and R-spines (catalytic and regulatory). These two are the groups of certain spatially conserved residues forming hydrophobic spines in the kinases' active states, named from their respective functions [12]. Aforementioned event leads in turn to the closing of the N-lobe alongside with the turning of C-helix. Such steps establish hydrophobic core in the center of the structure that alllows to evaporate water molecules from there and create pocket for ATP binding. It is worth to note that only the complete formation of R-spine is sufficient indicator of active state [11]. Structural differences between both states can be inspected through the Fig. 1.2. It is clearly visible there that active states maintain higher degree of architecture conservation.

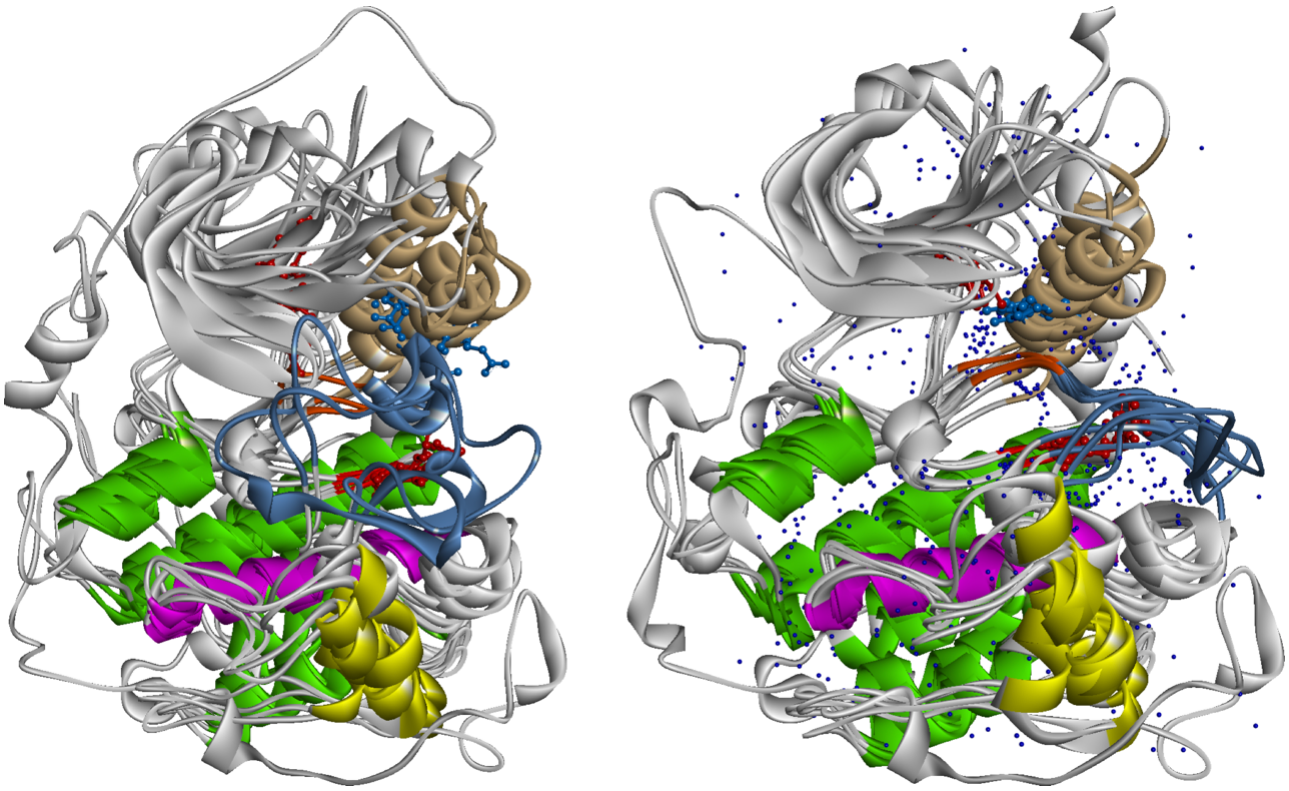


Figure 1.2: Multiple kinase structures (PKA, CDK2, EGFR, IRK, SRC) sumperimposed together in theirs respective inactive states on the left and active ones on the right figure [15].

Complete set of genes encoding over 500 protein kinases forms the kinome. There are a few major groups within it that can be distinguished based on kinases sequence distance including: tyrosine kinases (TK), tyrosine kinase-like (TKL), AGC (named from member families PKA, PKG and PKC), CAMK (Calcium and Calmodulin-regulated kinases), Casein kinase 1 (CK1), CMGC (named after the initials of some of its members including: the MAPK growth- and stress-response kinases, the cell cycle cyclin dependent kinases and kinases involved in splicing and metabolic control), STE (activators of MAPK family) and RGC (receptor guanylate cyclases) that contain eukaryotic protein kinase domains that all appear to be inactive, all differing in broad substrate site specificity [7]. There are also some kinases not classified into any major group or named atypical, lacking sequence similarities shared within the whole

kinome. Protein kinases can be also divided into families by both sequence similarity and broad biological functions and further grouped into subfamilies based on finer functional and structural similarities. Evolutionary relations between these groups are shown in the Fig. 1.3.

Despite demonstrating relatively low structural similarity in inactive states, active states reveal higher degree of structure conservation. It seems that significance of the function that they maintain lead to preservation of spatial organisation of key amino acids [12, 14]. Two groups that would be focused in this work are serine-threonine and tyrozine kinases named from the amino acid type receiving phosphate group from ATP. Two representative kinases for each of them were chosen for further analysis: PKA and CDK serine-threonine kinases, IRK and EGFR from tyrozine group [13].

1.3. The work purpose

Protein dynamics is directly driven by its structure rather than sequence, therefore knowing that kinase catalytic subunits share rather conserved tertiary structures, despite varying evolutionary distances, it is interesting to examine if they maintain similarity in dynamic behavior in spite of sometimes considerably different sequences. There are few interesting hypotheses that will be tested in this work regarding that topic:

1. The differences in dynamics between family members are proportional to their evolutionary distance.
2. The existence of two distinct activity states, means that kinases posses likewise two different dynamics states. The active state being important for catalytic function should behave more conservely due to its important role, hence checking if the activation process leads to unification of dynamic properties would be the second goal of this work,
3. Similarites between different kinases can manifest in various ways. It is interesting to ask if the global resemblance of the architecture translate into similarity of global dynamics. Analogous question can be asked with regard to the local (sequence-resolved) level differences by examining the relations between them and local dynamics differences.

Acknowledging these problems would grant a unique insight into differences between protein kinases dynamic behavior in their distinct activation states. This recognition about the details of protein structure dynamics, that accounts for both global conformational movements and variability at the atomic level exceeds the capabilities of the experimental methods but it's possible to be gained by the means of molecular dynamics simulations.

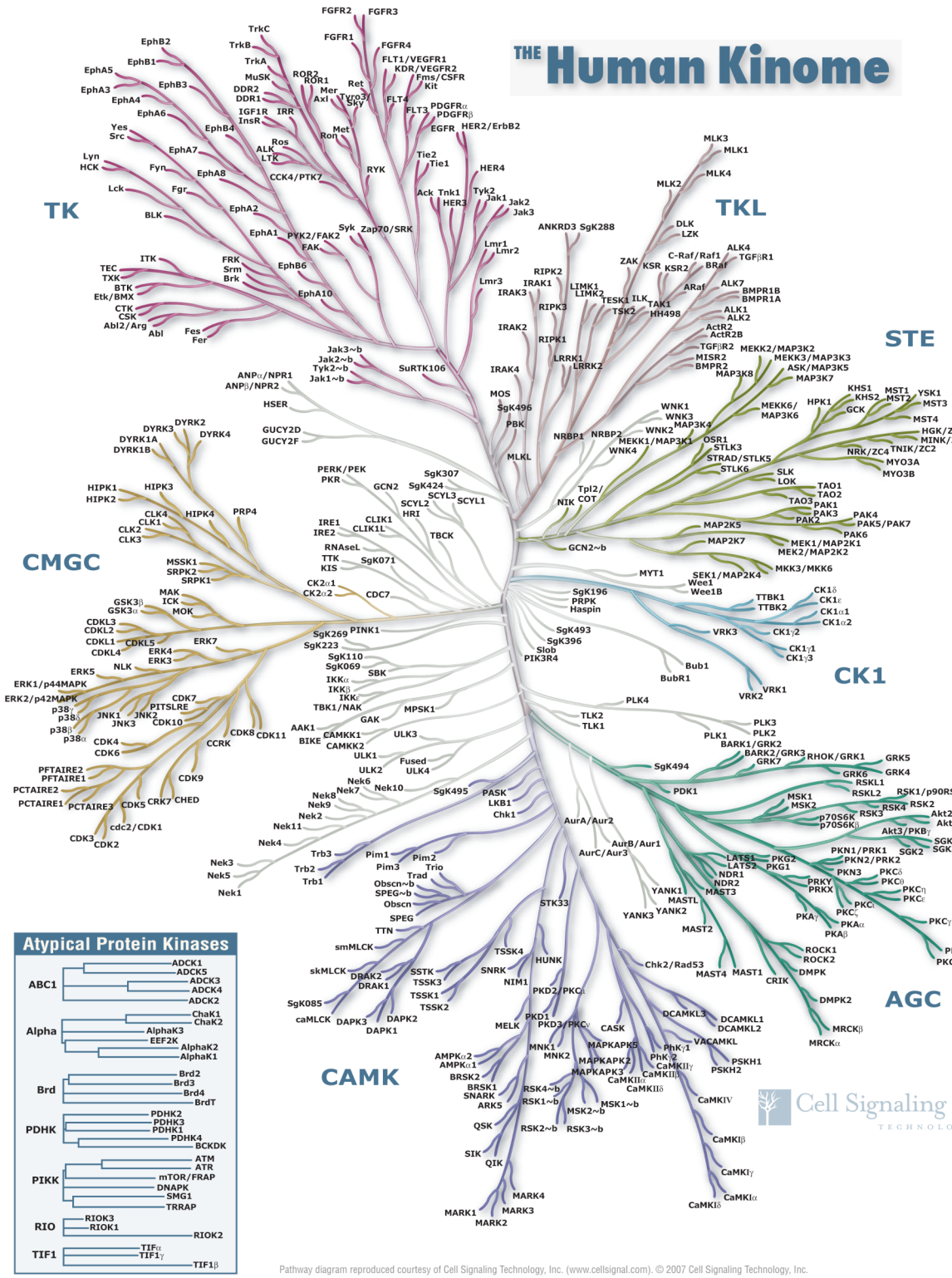


Figure 1.3: Dendrogram of human kinome based on sequence similarity between protein kinases derived from Manning et. al [7]. Kinome illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).

Chapter 2

Methodology and system

2.1. Molecular dynamics

There are many computational methods that allow examination of physical quantities of interest of biomolecular system. They can be ranked according to their complexicity as a consequence of different underlying assumptions, cost demand and scope of phenomena they can be applied to. The most accurate one is **explicit molecular dynamics** simulation that attempts to recreate particles behavior in real time addressing all major interactions between them and their surroundings. This can be carried on twofold: 1) applying quantum principles explicitly including all electron dependant potentials, 2) approximating electrons behavior with classical parameters. Although the first approach undoubtedly provides the best precision, it comes with a price of high computational cost, therefore it is commonly applied to small molecules. On the other hand, the latter can handle systems consisting of hundreds residues in practically feasible, creating sufficient enough results.

Classical molecular dynamics utilizes simple Newtonian's equation for calculating atoms' accelerations as in the following formula

$$\vec{\mathbf{F}}_i = m_i \vec{\mathbf{a}}_i, \tag{2.1}$$

where $\vec{\mathbf{F}}_i$ is a cumulative force acting on atom i, m_i and $\vec{\mathbf{x}}_i$ are its mass and acceleration, respectively. Solving this equation allows finding the accelerations, velocities and positions of all elements in the system in the subsequent moment.

Forces acting on each can be derived through potential energy gradient $\vec{\mathbf{F}} = -\nabla U$, thus assuming that all non negligble interactions between atoms origin from conservative forces. The exact formula of this potential can be determined by means of quantum mechanics simulation, but it is too computationally demanding for the protein-sized systems. Owing to the Born-Oppenheimer approximation it is possible to assume that the electron density depends on the momentary positions of nuclei by integration of the electronic degrees of freedom in calculations, and thus focusing on the nuclei translations [16]. Moreover the usage of Hellmann-Feynmann theorem allows the calculation of forces acting on the atomic nuclei, using classical Coulomb's electrostatic law, from electron density and the rest atomic nuclei [17–20]. Subsequent assumption that molecules remain close to their energy minima and that the chemical bonds are stable and thus can not be broken during the simulation, forms the basis for the introduction of simple potential energy models as a function of nuclei positions. In this manner atoms are approximated as charged points posessing certain masses. Overall potential energy of the system can be then decomposed into several contributions representing different types

of attraction as presented below

$$U_{\text{total}}(\vec{\mathbf{r}}) = U_{\text{bond}}(\vec{\mathbf{r}}) + U_{\text{angle}}(\vec{\mathbf{r}}) + U_{\text{dihedral}}(\vec{\mathbf{r}}) + U_{\text{elec}}(\vec{\mathbf{r}}) + U_{\text{vdW}}(\vec{\mathbf{r}}). \tag{2.2}$$

Here, subsequent potential contributions correspond to U_{bond} - chemical bond stretching, U_{angle} and U_{dihedral} - deformations of the planar angles between atoms and dihedral angles respectively, U_{elec} - electrostatic energy, U_{vdW} - attraction and dispersive interaction of atom nuclei [21]. Electrostatic interaction between atoms i and j is accounted through U_{elec} potential, calculated using the classical Coulomb's formula

$$U_{\text{elec}} = \sum_{i \neq j} -k \frac{q_i q_j}{r_{ij}}, \tag{2.3}$$

,where q_i and q_j are partial charges of respective atoms and k is Coulomb's constant.
Short range repulsive and attractive dispersion interactions are typically modelled with Lennard-Jones potential

$$U_{\text{LJ}} = \sum_{i \neq j} 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \tag{2.4}$$

where ε is the depth of the potential well and σ represents the distance at which potential has a zero place [22]. Other potentials are responsible for approximation of covalence bonding effects. First two terms utilize harmonic potential, whereas the third one is defined by simple periodic function. Examples of these potentials are listed below

$$U_{\text{bond}} = \sum_{i} \frac{1}{2} k_b (r_i - r_0)^2, \tag{2.5}$$

$$U_{\text{angle}} = \sum_{i} \frac{1}{2} k_a (\theta_i - \theta_0)^2, \tag{2.6}$$

$$U_{\text{dihedral}} = \sum_{i} \frac{1}{2} \sum_{j=1}^{n} V_j [1 + \cos(j\omega - \gamma_j)], \tag{2.7}$$

where k_b, k_a are the force constants, $r_i - r_0$ and $\theta_i - \theta_0$ are corresponding deviations from the potential minima, j depicts one of the n minima with the V_j rotational barrier, ω the torsional angle, γ the phase factor and i denotes the index of respected covalence term. Exact mathemathical formulas describing them vary in different constants values. These parameters along with effective atoms' charges are determined experimentally or by means of quantum mechanics *ab inito* calculations. The final description of all modelled interactions with corresponding constants is called force field. There are a few force fields that differ in manner of parametrization of aforementioned formulas e.g. AMBER, CHARMM and GROMOS [23–25].

Following the choice of that desirable parametrization, atoms' accelerations can be obtained from equation 2.1. Their movement is then found by means of numerical integration using one of several algorithms. A popular one is the Verlet algorithm, that approximates acceleration as constant over the duration of a time step as in formula below

$$\vec{\mathbf{x}}_i(t + \Delta t) = \vec{\mathbf{x}}_i(t) + \vec{\mathbf{v}}_i(t)\Delta t + \frac{1}{2}\vec{\mathbf{a}}_i(t)\Delta t^2, \tag{2.8}$$

$$\vec{\mathbf{v}}_i(t + \Delta t) = \vec{\mathbf{v}}_i(t) + \frac{\vec{\mathbf{a}}_i(t) + \vec{\mathbf{a}}_i(t + \Delta t)}{2}\Delta t, \tag{2.9}$$

where $\vec{\mathbf{x}}_i(t)$, $\vec{\mathbf{v}}_i(t)$, $\vec{\mathbf{a}}_i(t)$ denote subsequently position, velocity and acceleration of the particle i at time t and Δt depicts the length of simulation's timestep [26].

This assumption is valid for short enough timesteps that could be set manually according to specific needs (comonly used one is 2fs). There are few conditions to be met for simulation to reflect physical conditions, one of them being proper system definition. Firstly borders around structures have to be drawn, thus creating fixed cell that would be subject to calculations. There are several shapes that can be chosen to tackle specific problems. Secondly biomolecules' surrondings should imitate chemical environment, which is solved by placing solvent, typically in form of water particles and ions to even the system charges. With such definition, problem emerges at the edges of the simulation box that create non-physical artifacts due to arbitrary limited space e.g. crossing the box' boundaries. It is solved by surrounding the cell with many copied and translated versions of itself, exampled in the Fig 2.1.

Finally particles are allowed to cross boundaries resulting in moving to neighboring cell image.

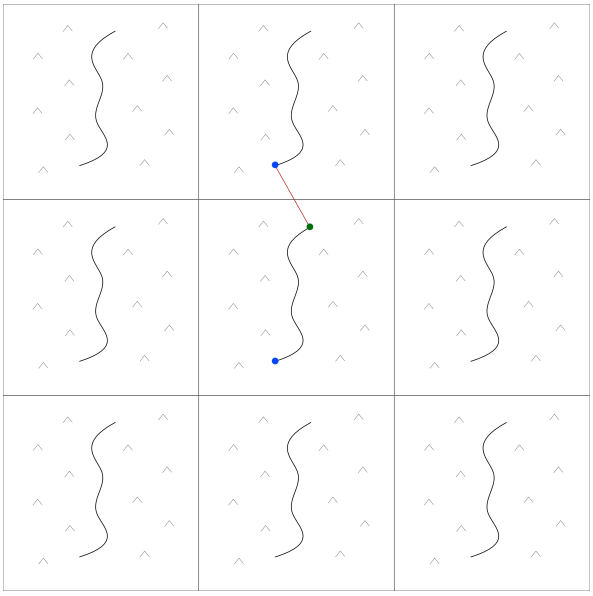


Figure 2.1: A 2-dimensional depiction of a simulated cell in the center surrounded by its 8 copies. Bronze line represents interaction between green atom and closest copy of blue atom. Rest of the blue atoms is excluded from calculation of such interaction.

Following the minimum image convention, only the closest pairs of atoms are considered for short-range non-bonded interaction terms, unlike the long-range electrostatic interactions that are typically calculated with one of the lattice sum methods e.g. Ewald Sum or Particle Mesh Ewald [27–29].

Moreover to comply with thermodynamics' laws some properties need to be kept at fixed values. Sets of such parameters determine statistical ensembles. Three of these typically considered ensebles for MD simulations are:

- microcanonical (NVE) with fixed number of particles N, volume V and energy E of the system,
- canonical (NVT) with fixed temperature T, volume and number of particles,
- isothermal-isobaric (NpT) with fixed themperature, pressure p and number of particles.

Without any further interference simulated system would tend to represent microcanonical ensemble, but generally imitation of laboratory conditions in molecular dynamics requires use of canonical or isothermal-isobaric ensembles. This is accomplished by using additional algorithms representing specific thermostats and barostats applied to the system. One of the historical approaches for keeping constant temperature was proposed by Berendsen [30]. This

procedure imitates heat bath with given temperature T_0 applied to the system. Effective adjustement to cell's temperature T can be described by formula below

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau}, \tag{2.10}$$

where τ is time constant set manually. Adjusting the value of τ leads to change in coupling strength - wider timescales result in weaker corrections. Modification of this parameter by including degrees of freedom and heat capacity of the system leads to derivation τ_T, that is followingly utilized to obtain time-dependent factor λ

$$\lambda = \left[1 + \frac{n_{TC}\Delta t}{\tau_T}\left\{\frac{T_0}{T(t - \frac{1}{2}\Delta t)} - 1\right\}\right]^{1/2}, \tag{2.11}$$

where n_{TC} is number of steps after which scaling occurs. Finally it's value is applied for velocity adjustment such that to modify kinetic energy E_k as following

$$\Delta E_k = (\lambda^2 - 1)E_k. \tag{2.12}$$

Nowadays velocity rescaling algorithm is more recommended to use, although it differs from Berendsen's merely by addition of stochastic term for energy adjustment, to ensure production of correct canonical ensemble [31].

In a similar way, pressure can be kept at constant level using one of the few possible barostats. Although Berendsen proposed solution to this problem as well, it couldn't replicate correctly NpT ensemble. Instead one of the better alternatives, example being Parinello-Rahman barostat, can be utilized. This approach rescales simulation box' vectors described by the matrix $\mathbf{b}$ using parameter matrix $\mathbf{W}$ according to formula

$$(\mathbf{W}^{-}1)_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_p^2 L}. \tag{2.13}$$

Here inversion of the $\mathbf{W}$ matrix determines the strength of the coupling and defines box deformation through pressure time constant τ_p and isothermal compressibilities $\boldsymbol{\beta}$ parameters with largest box matrix element L.
Applying this matrix to the equation of motion allows pressure coupling description by equation

$$\frac{\mathrm{d}\mathbf{b}^2}{\mathrm{d}t^2} = V\mathbf{W}^{-1}\mathbf{b'}^{-1}(\mathbf{P} - \mathbf{P}_{ref}), \tag{2.14}$$

where V denotes volume of the box, $(\mathbf{b}^T)^{-1}$ transposed box vectors' matrix, $\mathbf{P}$ and $\mathbf{P}_{ref}$ are current and reference pressure tensors respectively.
Procedure requires that also equations of motions have to be changed through conservation of modified Hamiltonian $\hat{H}$

$$\hat{H} = E_{pot} + E_{kin} + \sum_i P_{ii}V + \sum_{i,j}\frac{1}{2}W_{ij}\left(\frac{\mathrm{d}b_{ij}}{\mathrm{d}t}\right)^2, \tag{2.15}$$

where i, j point to specific elements of aforementioned matrices. This change allows determination of atoms' movements from the equation

$$\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}^2t} = \frac{\mathbf{F}}{m} - \mathbf{M}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}, \tag{2.16}$$

$$\mathbf{M} = \mathbf{b}^{-1}[\mathbf{b}\frac{\mathrm{d}\mathbf{b}'}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{b}}{\mathrm{d}t}\mathbf{b}']\mathbf{b}'^{-1}, \tag{2.17}$$

where $\mathbf{M}$ represents friction looking term that is rather artifically imposed due to equations of motions being defined relatively to the box vectors. [32, 33].

Lastly simulation should be consistent with statistical physics laws. It is achieved through the ergodic theorem, that states that mean values of the properties of the stationary process for the sufficiently long sampling converge to their respective statistical ensemble averages and thus respect them via well defined thermodynamic properties [34]. This finalizes validation of whole molecular simulation based approach.

2.2. Descriptors

2.2.1. Root-Mean-Square Deviation and Fluctuations (RMSD, RMSF)

RMSD is the simplest way to measure similarity between different models of one structure. It is calculated followingly, using two sets of corresponding atomic positions

$$\mathrm{RMSD}(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big[(a_{i_x}-b_{i_x})^2+(a_{i_y}-b_{i_y})^2+(a_{i_z}-b_{i_z})^2\Big]}. \tag{2.18}$$

Here a and b denotes two different models of one structure having n atoms, $a_{i_{(x,y,z)}}$ and $b_{i_{(x,y,z)}}$ coordinates of atom i originating from them.
Calculating RMSD of product of two sets' of frames allows to generate **RMSD matrix**, which elements are defined as following

$$\mathbf{RMSD}(\mathbf{A},\mathbf{B})_{i,j} = \mathrm{RMSD}(A_i, B_j), \tag{2.19}$$

where A_i and B_j represent i-th and j-th frames from trajectories $\mathbf{A}$ and $\mathbf{B}$ respectively.
Finally, two different trajectories can be then globally compared by means of averaging all elements from such matrix to measure distance d between them

$$d(\mathbf{A},\mathbf{B}) = \overline{\mathbf{RMSD}(\mathbf{A},\mathbf{B})_{i,j}}. \tag{2.20}$$

RMSD concept can be easily extended to the scope of whole simulation leading to construction of RMSF, that measures the fluctuation of individual atoms throughout its duration

$$\mathrm{RMSF}(\mathbf{x},i) = \sqrt{\big\langle \mathbf{x}_i - \langle \mathbf{x}_i\rangle\big\rangle^2}. \tag{2.21}$$

Here $\mathbf{x}_i$ represents vector containing positions of atom i through whole simulation and $\langle \mathbf{x}_i\rangle$ its ensemble average respectively. Whereas first measure quantifies distance from chosen reference structure and can be helpful in identification of closeness, the second one illustrates mobility of different parts of the structure, which could point out important areas of the protein.

2.2.2. Principal Components Analysis (PCA)

Machine learning provides several data exploration methods, one of them being PCA that is also commonly referred as Covariance Analysis. It shows main directions of data's variability by decomposing covariance matrix of certain quantities into vectors with specific weights. In biomolecular systems the main point of interest is covariance matrix of atomic coordinates x_i from certain time trajectory. Additionaly diagonal matrix $\mathbf{M}$, containing masses of the atoms can be included in it as presented below

$$A_{ij} = \big\langle M_{ii}^{1/2}(x_i - \langle x_i\rangle)M_{jj}^{1/2}(x_j - \langle x_j\rangle)\big\rangle. \tag{2.22}$$

This makes $\mathbf{A}$ a symmetric $3N \times 3N$ matrix that is subsequently diagonalized by orthonormal transformation matrix $\mathbf{R}$ as followingly

$$\mathbf{R}^T \mathbf{A} \mathbf{R} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_{3N}). \tag{2.23}$$

where $\lambda_1 > \lambda_2 > ... > \lambda_{3N}$ are eigenvalues associated with successive eigenvectors $\mathbf{R}_i$.
The modes obtained this way represent directions of atomic movements decreasing in importance - from the level of correlated movements of large parts of the structure to less local correlations.

Overall space overlap In order to compare principal modes of two trajectories, mathematically correct distance measure needs to be constructed. To begin with, simple difference d between covariance matrices $\mathbf{A}$ and $\mathbf{B}$ can be defined using their traces, that could be reformulated utilizing 2.23.

$$d(\mathbf{A}, \mathbf{B}) = \left(\sum_i^n (\lambda_i^A + \lambda_i^B) - 2 \sum_i^n \sum_j^n \sqrt{\lambda_i^A \lambda_j^B} \left(\mathbf{R}_i^A \cdot \mathbf{R}_j^B \right)^2 \right)^{1/2}. \tag{2.24}$$

Such expression lacks interpretability, but can be furtherly modified to define **overall space overlap** distance s as followingly

$$s(\mathbf{A}, \mathbf{B}) = 1 - \frac{d(\mathbf{A}, \mathbf{B})}{\sqrt{\mathrm{tr}\mathbf{A} + \mathrm{tr}\mathbf{B}}}, \tag{2.25}$$

where tr$\mathbf{A}$ and tr$\mathbf{B}$ denote traces of indicated matrices [35].
Its value ranges continuously from 0 to 1, reaching maximum for identical modes and minimum for completely orthogonal ones, which makes it valid measure for comparison.

Subspace overlap Due to global nature of protein functional motions, most relevant information is typically stored in first few principal modes. For such cases overlap between these components would be sufficient to observe desired phenomenon. Truncated measure for such two sets of vectors $\{\mathbf{v}_i\}$ and $\{\mathbf{w}_j\}$ of equal size n is commonly defined as

$$\mathrm{overlap}(\mathbf{v}, \mathbf{w}) = \frac{1}{n} \sum_i^n \sum_j^n (\mathbf{v}_i \cdot \mathbf{w}_j)^2, \tag{2.26}$$

Both for increasing number of principial modes and increasing amount of variance explained by first components, subspace overlap converges to overall space one.

2.2.3. Jensen-Shannon Divergence

Previous descriptor allowed examination of differences on global - trajectory scale, whereas current one gives insight into individual residues behavior. Comparison of protein dynamics can be carried on also through conformational ensembles. In this approach distribution of some parameters representing trajectory have to be selected. A possible candidate for this is **torsion angle space** due to two reasons: 1) torsion angles indicate biologically relevant functional motions and 2) they lack frame-fitting issues existsing in Cartesian analysis. Their distributions can be compared utilizing Kullback-Leibler Divergence [36, 37], also called relative entropy, that is fundamental quantity in information theory and could be formulated as presented below

$$\mathrm{KL}(x_1, ..., x_m || x_1^*, ..., x_m^*) = \int J(x_1, ..., x_m) \rho(x_1, ..., x_m) \ln \frac{\rho(x_1, ..., x_m)}{\rho^*(x_1^*, ..., x_m^*)} \mathrm{d}x_1, ..., \mathrm{d}x_m, \tag{2.27}$$

where ρ, $\rho*$ are probability density functions (p.d.f.) of perturbed and reference ensembles respectively, $\mathbf{x} \in \mathbb{R}^m$ variables described by these distributions and J Jacobian determinant.

Applying this form to distribution $\rho(\phi, \psi, \chi)$, which is function of two backbone and the first side chain's torsion angles of aminoacids in a protein, and utilizing Generalized Kirkwood Superposition Approximation [38,39] allows to obtain Kullback-Leibler Divergence expansion over m degrees of freedom in few steps:

$$\mathrm{KL} = \sum_{n=1}^{m} \sum_{i}^{n\mathrm{bins}} p_i \ln \frac{p_i}{p_i*} - \sum_{n=1}^{m} \sum_{n' \neq n} \sum_{ij}^{n\mathrm{bins}^2} p_{ij} \ln \frac{p_i p_j p_{ij}*}{p_i * p_j * p_{ij}} + ... \tag{2.28}$$

Here p_i and p_j represent bin probabilities of specific degrees of freedom, p_{ij} their joint histograms for marginal p.d.f's ivolving both of them and $pi*, p_j*, p_{ij}*$ their reference counts. This expansion allows for well-defined truncation at desired complexity. Specifically for a single protein residue, the first-order term approximation can be used to obtain divergences for the ϕ, ψ and χ torsion angles as following

$$\mathrm{KL}_{\mathrm{res}_n} = \sum_{\phi,\psi,\chi} \sum_{i}^{n\mathrm{bins}} p_i \ln \frac{p_i}{p_i*}. \tag{2.29}$$

The term obtained this way can be finally used to compare distributions by means of **Jensen-Shannon Divergence** that is defined below

$$\mathrm{JS}(x_1, ..., x_m || y_1, ..., y_m) = \frac{1}{2}\mathrm{KL}(x_1, ..., x_m || (x_1, ..., x_m) + (y_1, ..., y_m)) \tag{2.30}$$

$$+ \frac{1}{2}\mathrm{KL}(y_1, ..., y_m || (x_1, ..., x_m) + (y_1, ..., y_m)). \tag{2.31}$$

The Jensen-Shannon Divergence is slightly modified verion of Kullback-Leibler Divergence that is symmetrical in ensembles treatment. Due to Kullback-Leibler Divergence expansion being general for any "reference" distribution, statistical test for Jesnsen-Shannon Divergences could be created with the null hypothesis as the average of the null hypothesis distributions of the Jensen-Shannon Divergences within the separate "reference" and "target" ensembles [40]. In this work values of JS_{tot} were analyzed, which are the total residue-residue Jensen-Shannon divergences

$$\mathrm{JS}_{i_{tot}} = \sum_{\phi,\psi} \mathrm{JS}_i, \tag{2.32}$$

where i is the residue number.

2.2.4. Statistical hypotheses testing

The difference between two sets of observations can be examined for the statistical significance with either parametrical or non-parametrical test depending if their distributions have a defined analytical form. Moreover all the samples should be drawn independently meaning that they don't correlate with each other. Problem appears when observations are transformed in a manner that generate correlated terms. Solution for fixing this issue is suggested below for the sake of comparison of molecular dynamics trajectories.

Let's assume that there are two sets of frames from different MD simulations and a related to them descriptor f (e.g. RMSD) operating on their subsets. Let each set consist of n blocks of equal size, from which corresponding descriptor values could be calculated and then put into $n \times n$ matrices. For pair of simulations' frames sets three products between n blocks, one per

each simulation could be created, consisting of descriptor values of: 1) two products between n blocks of the same trajectory, 2) product between sets of blocks from different trajectories. Let denote $\boldsymbol{A}$, $\boldsymbol{B}$, self-product matrices from first and second simulation subsequently and $\boldsymbol{C}$ descriptor matrix of product between two different simulations. In that notation $a_{i,j}$ denotes the descriptor calculated from i and j block of first and second set used to build $\boldsymbol{A}$ respectively. Let also $\{a_i\}$ represent set of frames from block i of simulation no. 1.

Obvious approach to this problem would be gathering all descriptor values and examininig statistical differences of means of matrices. Because $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices are products of the same sets of blocks in general each of them consists of $n(n-1)/2+n$ different values - in other words they are symmetrical, whereas $\boldsymbol{C}$ have n^2 unique values, that makes up a total of $(3n^2+n)/2$ sized sample. Lower triagular left part of $\boldsymbol{A}$ could be then joined with upper triangular right one of $\boldsymbol{B}$ to create matrix $\boldsymbol{D}$ possessing unique values. Its diagonal has to be rescaled by the factor of $\frac{1}{2}$ to ensure physicality of the results. This process is illustrated in the Fig 2.2. Two trajectories can be then examined for differences using statistical tests between

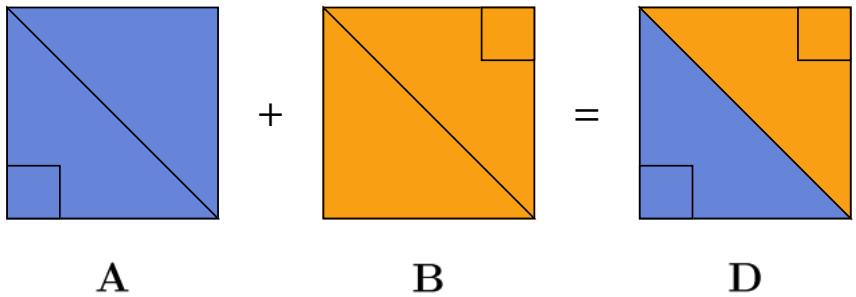


Figure 2.2: Depiction of process of creation of conjoined matrix $\boldsymbol{D}$ from two self-product matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

the conjoined matrix $\boldsymbol{D}$ and $\boldsymbol{C}$ using all their values. This approach would be justified only if the results sampled from simulations were independent.

Recalling the procedure of generating $\boldsymbol{A}$ matrix, one can find its element from following formula

$$a_{i,j} = f(\{a_i\}, \{a_j\}), \tag{2.33}$$

which shows that the result is obtained by some transformation of sets of frames. By expanding this observation for other elements it's easy to notice that $a_{i,j}$ is correlated to some degree to element $a_{i,k}$ for every $j \neq k$, because $a_{i,k}$ is also transformation of set $\{a_i\}$ and some another set. In another words following relation is valid

$$(\forall j \in \{1, ..., n\})(r_{\{a_{i,j}\},\{b_{i,j}\}} \not\approx 0), \tag{2.34}$$

where $r_{\{a_{i,j}\},\{b_{i,j}\}}$ is Pearson's correlation coefficient of sets of elements $a_{i,j}$ and $b_{i,j}$ with fixed index i. Although sets of frames from different simulations can be treated as independent, the same doesn't apply to pairwise descriptor values that utilize them. The reason behind this is that every block of simulation is used for calculation of n results, which results in non-zero correlation between them, thus some of matrix' elements should not be treated as independantly drew samples.

According to the statistical inference some of the elements should thus be treated as pairs of correlated results instead in order to fulfil the condition of independance for different pairs. Recalling the equation 2.34 one finds that $a_{i,j}$ is correlated to all elements which were calculated using block i - in our case $\forall k\{a_{i,k}, b_{i,k}\}$, because $b_{i,k}$ were also obtained by some

transformation of $\{a_i\}$. This reasoning leads to the problem of construction of independant random variables. Let's consider sample of $2n$ results X_i generated as in the following formula

$$\mathrm{X}_i = \left(\sum_{m=1}^{n} a_{i,m}, \sum_{m=1}^{n} c_{i,m} \right) \text{ for } i <= n \tag{2.35}$$

$$\mathrm{X}_i = \left(\sum_{m=1}^{n} b_{m,i-n}, \sum_{m=1}^{n} c_{m,i-n} \right) \text{ for } i > n, \tag{2.36}$$

where all elements of each X_i are transformation of one block i (for first n results comes from frames $\{a_i\}$, for last from $\{b_i\}$) and $\{c_i\}$ are the values of product between these two sets. In such manner both elements of pair X_i are correlated, but results X_i and X_j are independant with the agreement to at most single components inside sums which repeats in both of them (e.g. for $n = 8$, $a_{1,2}$ will be a part of X_1 and X_2, because $\boldsymbol{A}$ is symmetrical and for $b_{1,2}$ in X_1 and X_{10}), thus the correlation between the different X_i's should be minimized. Furthermore this formula ensures that all the values from aforementioned matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ (recalling that it's conjoined matrix from self-products of $\boldsymbol{A}$ and $\boldsymbol{B}$) are utilized the least amount times, what is also the reason for reversed indexation in the equation 2.36. Finally it should be possible to estimate it by calculating proper coefficient between the sets of values, which were used to X_i construction similar to the formula below

$$\begin{aligned} r(\mathrm{X}_i, \mathrm{X}_j) = r(&\{a_{i,m}\, b_{i,m} : m \in \{1, ..., n\}\}, \\ &\{a_{j,m}\, b_{j,m} : m \in \{1, ..., n\}\}) \end{aligned} \tag{2.37}$$

In this example $i, j <= n$.
The Student's t-test for paired samples can be then performed on dataset

$$\mathbf{X} = \{\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, ..., \mathrm{X}_{2n}\},$$

assuming that random variable X_i is distributed normally. Otherwise Wilcoxon's non-paramteric signed-rank test [41] can be carried out. Latter one examines if the distribution of observations are symmetrical around 0. It explains that $\boldsymbol{C}$ elements can be used for generation of samples X_i, because examining differences between simulations assumes that under the null hypothesis distributions of $\boldsymbol{A}$ and $\boldsymbol{B}$ are identical thus their elements can be treated in the same manner.

2.3. Systems under study

2.3.1. Structures

Two kinase groups were chosen for experiment - serine-threonin and tyrosine ones. From them both, two different proteins were selected, each in inactive and active form. Structures were gathered from PDB and their record ids are listed in Table 2.1.

Table 2.1. Kinase types, families and PDB ids

type	**substrate**	**inactive model**	**active model**
PKA	ser-thr	4DFY	3FJQ
CDK2	ser-thr	6YL6	1JST
IRK	tyr	1IRK	4XLV
EGFR	tyr	3POZ	3VJN

Models for PKA and CDK2 kinases were prepared during previous work [15]. Due to missing parts in original PDB 3POZ record, its structure was generated through automated homology-modelling SWISS-MODEL server [42]. It was then joined with crystallic water and minimized in vacuum with atoms from PDB record restrained using GROMACS. After relaxation, model was protonated using PROPKA 3 package, that predicts pK_a values for titratable residues heuristically [43]. Finally obtained structure was analyzed for any abnormalities with PDB2PQR software suite [44]. Similar procedure was applied to 3VJN and 4XLV modelling due to the same reasons. All active state kinase systems' included ATP particle with two magnesium ions positioned at structure's binding site. Models of active and inactive structures superimposed pair-wise of PKA, CDK2, IRK and EGFR kinases are shown in Fig. 2.3.

2.3.2. Simulations protocols

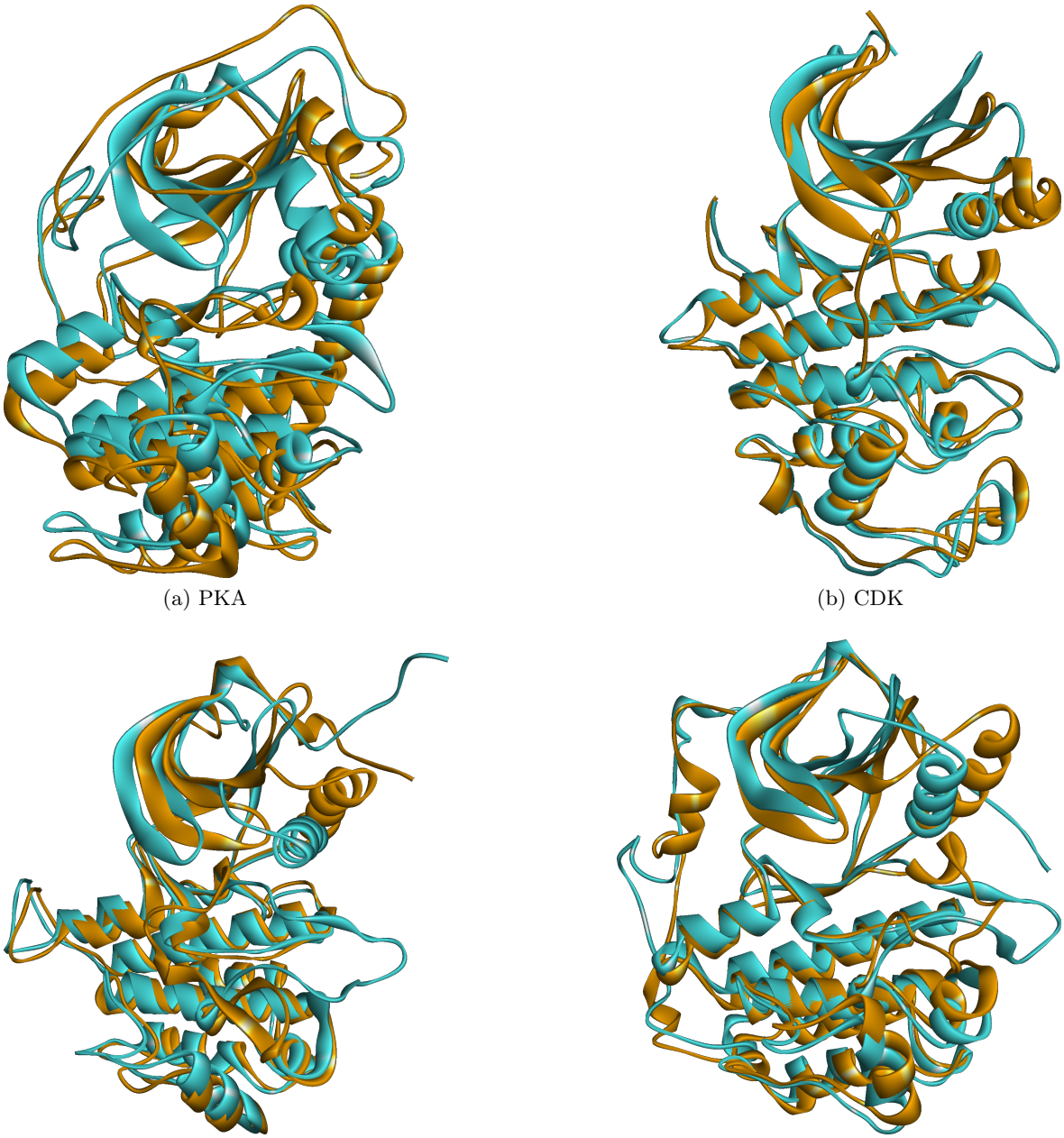
Dodecahedron shape of simulation box was chosen with distance from its borders to any of the protein's atom to be at least 13Å. Aqueous solvent represented by rigid TIP3P water model [45, 46] was added to the simulation box using the gmx solvate tool. Sodium and chlorine ions were inserted to neutralise the system and reach concentration of $0.15\frac{\text{mol}}{\text{l}}$ using the gmx genion program. The cutoff for van der Waals interactions was set to 10 Å and electrostatic potentials were calculated explicitly up to 10 Å. Beyond this distance the Particle Mesh Ewald method was used to determine the long-range electrostatic interactions [47, 48]. All simulations were carried out using the GROMACS package [49–56] with AMBER99SB-ILDN force field [57] and additional ATP parametrization [58], periodic boundary conditions and 2 fs timestep. Constraints on atom's bonds between heavy atoms and hydrogens were placed using LINCS algorithm [59].

For each structure two independent simulations were conducted according to the following protocol. Energy of the systems were minimised using the steepest descent algorithm for 1000 steps. Initial velocities were randomly assigned from Maxwell-Boltzmann distribution at $T_0 =$ 200K. During the first dynamics stage velocity rescaling thermostat with temperature $T_{ref} =$ 310K [31] was applied to recreate NVT ensemble. Second stage was run with additional use of Parinello-Rahman barostat [32,33] to ensure fixed pressure at $p_{ref} = 1$bar. Both equilibration and production stages were performed with Nose-Hoover [60, 61] instead of velocity rescaling thermostat. Durations of particular stages along with atom groups restrained throughout them were listed in Table 2.2.

Table 2.2. Simulation stages with their restraints and durations

stage	ensemble	restraints	duration
1	NVT	non-**H**	0.1 ns
2	NpT	non-**H**	1 ns
3	NpT	$\mathbf{C}_\alpha$	5 ns
4	NpT	$\mathbf{C}_\alpha$	5 ns
5	NpT	-	2500 ns

Homologically-modelled structures followed slightly different protocol to ensure stability of simulations and physicality of results. They were firstly minimized with fixed crystallic atoms positions (including water) and then their equilibration stage was split into two parts: 1) lasting 50 ns with restraints on all heavy atoms and 2) lasting 50 ns with restraints only on $\mathbf{C}_\alpha$s.



(a) PKA (b) CDK

(c) IRK (d) EGFR

Figure 2.3: Active and inactive state structures of considered kinases superimposed together in cyan and orange, respectively.

2.3.3. Analysis protocols

Simulations were visually inspected for undesirable events e.g. ATP dissociating from active site and RMSDs anomalies. For each kinase, structurally equivalent 'core' residues (defined as a common crucial aminoacids involved in catalytic activity shared within kinases family) were identified based on functional multiple alignment and then extracted [10]. These residues are listed below in Table 2.3.

Table 2.3. Aligned sequences of core residues in different kinases

№	PKA	CDK2	IRK	EGFR	№	PKA	CDK2	IRK	EGFR
1	PHE43	PHE4	ILE996	PHE712	53	TYR108	ASP68	GLY1064	GLY779
2	ASP44	GLN5	THR997	LYS713	54	SER109	VAL69	VAL1065	ILE780
3	ARG45	LYS6	LEU998	LYS714	55	PHE110	ILE70	VAL1066	CYS781
4	ILE46	VAL7	LEU999	ILE715	56	LYS111	HIS71	SER1067	LEU782
5	LYS47	GLU8	ARG1000	LYS716	57	ASP112	THR72	LYS1068	THR783
6	THR48	LYS9	GLU1001	VAL717	58	ASN113	GLU73	GLY1069	SER784
7	LEU49	ILE10	LEU1002	LEU718	59	ASN115	LYS75	PRO1071	THR785
8	GLY50	GLY11	GLY1003	GLY719	60	LEU116	LEU76	THR1072	VAL786
9	THR51	GLU12	GLN1004	SER720	61	TYR117	TYR77	LEU1073	GLN787
10	GLY52	GLY13	GLY1005	GLY721	62	MET118	LEU78	VAL1074	LEU788
11	SER53	THR14	SER1006	ALA722	63	VAL119	VAL79	VAL1075	ILE789
12	PHE54	TYR15	PHE1007	PHE723	64	MET120	PHE80	MET1076	THR790
13	GLY55	GLY16	GLY1008	GLY724	65	GLU121	GLU81	GLU1077	GLN791
14	ARG56	VAL17	MET1009	THR725	66	TYR122	PHE82	LEU1078	LEU792
15	VAL57	VAL18	VAL1010	VAL726	67	VAL123	LEU83	MET1079	MET793
16	MET58	TYR19	TYR1011	TYR727	68	GLY126	GLN85	GLY1082	GLY796
17	LEU59	LYS20	GLU1012	LYS728	69	GLU127	ASP86	ASP1083	CYS797
18	VAL60	ALA21	GLY1013	GLY729	70	MET128	LEU87	LEU1084	LEU798
19	LYS61	ARG22	ASN1014	LEU730	71	PHE129	LYS88	LYS1085	LEU799
20	HIS62	ASN23	ALA1015	TRP731	72	SER130	LYS89	SER1086	ASP800
21	GLY66	GLY27	GLU1024	LYS739	73	HIS131	PHE90	TYR1087	TYR801
22	ASN67	GLU28	THR1025	ILE740	74	LEU132	MET91	LEU1088	VAL802
23	HIS68	VAL29	ARG1026	PRO741	75	ARG137	GLY98	PRO1103	ASN808
24	TYR69	VAL30	VAL1027	VAL742	76	PHE138	ILE99	PRO1104	ILE809
25	ALA70	ALA31	ALA1028	ALA743	77	SER139	PRO100	THR1105	GLY810
26	MET71	LEU32	VAL1029	ILE744	78	GLU140	LEU101	LEU1106	SER811
27	LYS72	LYS33	LYS1030	LYS745	79	PRO141	PRO102	GLN1107	GLN812
28	ILE73	LYS34	THR1031	GLU746	80	HIS142	LEU103	GLU1108	TYR813
29	LEU74	ILE35	VAL1032	LEU747	81	ALA143	ILE104	MET1109	LEU814
30	ASP75	ARG36	ASN1033	ARG748	82	ARG144	LYS105	ILE1110	LEU815
31	ILE85	PRO45	ARG1041	ASN756	83	PHE145	SER106	GLN1111	ASN816
32	HIS87	THR47	GLU1043	GLU758	84	TYR146	TYR107	MET1112	TRP817
33	THR88	ALA48	PHE1044	ILE759	85	ALA147	LEU108	ALA1113	CYS818
34	LEU89	ILE49	LEU1045	LEU760	86	ALA148	PHE109	ALA1114	VAL819
35	ASN90	ARG50	ASN1046	ASP761	87	GLN149	GLN110	GLU1115	GLN820
36	GLU91	GLU51	GLU1047	GLU762	88	ILE150	LEU111	ILE1116	ILE821
37	LYS92	ILE52	ALA1048	ALA763	89	VAL151	LEU112	ALA1117	ALA822
38	ARG93	SER53	SER1049	TYR764	90	LEU152	GLN113	ASP1118	LYS823
39	ILE94	LEU54	VAL1050	VAL765	91	THR153	GLY114	GLY1119	GLY824
40	LEU95	LEU55	MET1051	MET766	92	PHE154	LEU115	MET1120	MET825
41	GLN96	LYS56	LYS1052	ALA767	93	GLU155	ALA116	ALA1121	ASN826
42	ALA97	GLU57	GLY1053	SER768	94	TYR156	PHE117	TYR1122	TYR827
43	VAL98	LEU58	PHE1054	VAL769	95	LEU157	CYS118	LEU1123	LEU828
44	ASN99	ASN59	THR1055	ASP770	96	HIS158	HIS119	ASN1124	GLU829
45	PHE100	HIS60	CYS1056	ASN771	97	SER159	SER120	ALA1125	ASP830
46	PRO101	PRO61	HIS1057	PRO772	98	LEU160	HIS121	LYS1126	ARG831
47	PHE102	ASN62	HIS1058	HIS773	99	ASP161	ARG122	LYS1127	ARG832
48	LEU103	ILE63	VAL1059	VAL774	100	LEU162	VAL123	PHE1128	LEU833
49	VAL104	VAL64	VAL1060	CYS775	101	ILE163	LEU124	VAL1129	VAL834
50	LYS105	LYS65	ARG1061	ARG776	102	TYR164	HIS125	HIS1130	HIS835
51	LEU106	LEU66	LEU1062	LEU777	103	ARG165	ARG126	ARG1131	ARG836
52	GLU107	LEU67	LEU1063	LEU778					

№	PKA	CDK2	IRK	EGFR	№	PKA	CDK2	IRK	EGFR
104	ASP166	ASP127	ASP1132	ASP837	156	TRP222	TRP187	TRP1193	TRP898
105	LEU167	LEU128	LEU1133	LEU838	157	ALA223	SER188	SER1194	SER899
106	LYS168	LYS129	ALA1134	ALA839	158	LEU224	LEU189	PHE1195	TYR900
107	PRO169	PRO130	ALA1135	ALA840	159	GLY225	GLY190	GLY1196	GLY901
108	GLU170	GLN131	ARG1136	ARG841	160	VAL226	CYS191	VAL1197	VAL902
109	ASN171	ASN132	ASN1137	ASN842	161	LEU227	ILE192	VAL1198	THR903
110	LEU172	LEU133	CYS1138	VAL843	162	ILE228	PHE193	LEU1199	VAL904
111	LEU173	LEU134	MET1139	LEU844	163	TYR229	ALA194	TRP1200	TRP905
112	ILE174	ILE135	VAL1140	VAL845	164	GLU230	GLU195	GLU1201	GLU906
113	ASP175	ASN136	ALA1141	LYS846	165	MET231	MET196	ILE1202	LEU907
114	GLN176	THR137	HIS1142	THR847	166	ALA232	VAL197	THR1203	MET908
115	GLN177	GLU138	ASP1143	PRO848	167	ALA233	THR198	SER1204	THR909
116	GLY178	GLY139	PHE1144	GLN849	168	GLY234	ARG199	LEU1205	PHE910
117	TYR179	ALA140	THR1145	HIS850	169	TYR235	ARG200	GLU1207	SER912
118	ILE180	ILE141	VAL1146	VAL851	170	PRO236	ALA201	GLN1208	LYS913
119	GLN181	LYS142	LYS1147	LYS852	171	PRO237	LEU202	PRO1209	PRO914
120	VAL182	LEU143	ILE1148	ILE853	172	PHE238	PHE203	TYR1210	TYR915
121	THR183	ALA144	GLY1149	THR854	173	PHE239	PRO204	GLN1211	ASP916
122	ASP184	ASP145	ASP1150	ASP855	174	ALA240	GLY205	GLY1212	GLY917
123	PHE185	PHE146	PHE1151	PHE856	175	ASP241	ASP206	LEU1213	ILE918
124	GLY186	GLY147	GLY1152	GLY857	176	ILE246	GLN211	VAL1218	ILE923
125	PHE187	LEU148	MET1153	LEU858	177	TYR247	LEU212	LEU1219	SER924
126	ALA188	ALA149	THR1154	ALA859	178	GLU248	PHE213	LYS1220	SER925
127	LYS189	ARG150	ARG1155	LYS860	179	LYS249	ARG214	PHE1221	ILE926
128	ARG190	ALA151	ASP1156	LEU861	180	ILE250	ILE215	VAL1222	LEU927
129	VAL191	PHE152	ILE1157	LEU862	181	VAL251	PHE216	MET1223	GLU928
130	LYS192	GLY153	XXX1158	GLY863	182	SER252	ARG217	ASP1224	LYS929
131	THR195	ARG157	XXX1162	LYS867	183	LYS254	LEU219	GLY1226	GLU931
132	LEU198	GLU162	GLY1169	GLY874	184	VAL255	GLN246	LEU1228	LEU933
133	CYS199	VAL163	LEU1170	LYS875	185	ARG256	ASP247	ASP1229	PRO934
134	GLY200	VAL164	LEU1171	VAL876	186	PHE257	PHE248	GLN1230	GLN935
135	THR201	THR165	PRO1172	PRO877	187	PRO258	VAL251	PRO1231	PRO936
136	PRO202	LEU166	VAL1173	ILE878	188	HIS260	PRO254	ASN1233	ILE938
137	GLU203	TRP167	ARG1174	LYS879	189	PHE261	LEU255	CYS1234	CYS939
138	TYR204	TYR168	TRP1175	TRP880	190	SER262	ASP256	PRO1235	THR940
139	LEU205	ARG169	MET1176	MET881	191	SER263	GLU257	GLU1236	ILE941
140	ALA206	ALA170	ALA1177	ALA882	192	ASP264	ASP258	ARG1237	ASP942
141	PRO207	PRO171	PRO1178	LEU883	193	LEU265	GLY259	VAL1238	VAL943
142	GLU208	GLU172	GLU1179	GLU884	194	LYS266	ARG260	THR1239	TYR944
143	ILE209	ILE173	SER1180	SER885	195	ASP267	SER261	ASP1240	MET945
144	ILE210	LEU174	LEU1181	ILE886	196	LEU268	LEU262	LEU1241	ILE946
145	LEU211	LEU175	LYS1182	LEU887	197	LEU269	LEU263	MET1242	MET947
146	SER212	GLY176	ASP1183	HIS888	198	ARG270	SER264	ARG1243	VAL948
147	LYS213	LYS178	GLY1184	ARG889	199	ASN271	GLN265	MET1244	LYS949
148	GLY214	TYR179	VAL1185	ILE890	200	LEU272	MET266	CYS1245	CYS950
149	TYR215	TYR180	PHE1186	TYR891	201	LEU273	LEU267	TRP1246	TRP951
150	ASN216	SER181	THR1187	THR892	202	GLN274	HIS268	GLN1247	MET952
151	LYS217	THR182	THR1188	HIS893	203	VAL275	TYR269	PHE1248	ILE953
152	ALA218	ALA183	SER1189	GLN894	204	ASP276	ASP270	ASN1249	ASP954
153	VAL219	VAL184	SER1190	SER895	205	THR278	ASN272	LYS1251	ASP956
154	ASP220	ASP185	ASP1191	ASP896	206	LYS279	LYS273	MET1252	SER957
155	TRP221	ILE186	MET1192	VAL897	207	ARG280	ARG274	ARG1253	ARG958

208	PHE281	ILE275	PRO1254	PRO959	215	ASN293	ALA282	ASN1261	ILE966
209	GLY282	SER276	THR1255	LYS960	216	HIS294	HIS283	LEU1262	GLU967
210	VAL288	ALA277	PHE1256	PHE961	217	TRP296	PHE285	LEU1263	PHE968
211	ASN289	LYS278	LEU1257	ARG962	218	PHE297	PHE286	LYS1264	SER969
212	ASP290	ALA279	GLU1258	GLU963	219	ALA298	GLN287	ASP1265	LYS970
213	ILE291	ALA280	ILE1259	LEU964	220	THR299	ASP288	ASP1266	MET971
214	LYS292	LEU281	VAL1260	ILE965					

Subsequent atoms from these residues were used to calculate: 1) RMSD matrices and principal modes were based on C_α atoms, 2) RMSF values were obtained with all theirs heavy atoms. RMSDs used for the examination of simulations' stability was obtained using all kinases' heavy atoms instead of just the ones from the core. Trajectory only from last 2000ns of each simulation's production stage, were considered for subsequent calculations of RMSD, RMSF, PCA and Jensen-Shannon Divergence. For each simulation, an average core structure was produced using gmx covar tool and subsequently used to perform PCA procedure. Both RMSD values between each pair of frames and RMSF for all residues were obtained using Python3 MDAnalysis library [62, 63]. Subspace overlap between first 10 principal modes and overall space overlap of all eigenvectors were implemented manually according to equations 2.18 and 2.21 utilizing NumPy library [64, 65]. Jensen-Shannon divergence was calculated between simulation-based distributions of phi and psi angles of amino acids that are found at equivalent locations within the conserved core structure, common to all considered kinase [40] using Python3 BioPython toolkit [66].

For statistical testing each simulation was split into 8 blocks, resulting in 16 independant results in significance examination of putative differences. Due to low number of samples non-parametric test was chosen for better relevancy. Tested matrices were prepared accordingly to procedure proposed in subsection 2.2.4. Overall 84 tests were carried on in this manner (independant matrices count multiplyed by the number of descriptors).

2.4. Software and additional informations

Structures models were visualized using VMD software [67] and BIOVIA Discovery Studio Visualizer [68]. Plots were created using Python3 Matplotlib [69] and seaborn [70] libraries and Mathematica software [71]. Fig. 2.1 was made in GIMP program [72]. Post-processing was done using Python3 NumPy and pandas [73, 74] libraries. Statistical testing was performed with Python3 SciPy libraries [75].

Chapter 3

Results and analysis

Analysis presentation was split into two parts to provide more logical order: firstly global characteristics of simulation trajectories were analyzed, secondly local dynamics at residue level was discussed. Lastly values of all descriptors were tested for significant differences using statistical tests.

3.1. Stability of simulations

Trajectories obtained from each kinase's simulations were used to calculate RMSDs of heavy atoms, against the first frames of their respected production stages, which are plotted in Fig. 3.1 and Fig. 3.2. Only last 2000ns of each production stage, visible on aforementioned plots, were considered for further analysis.

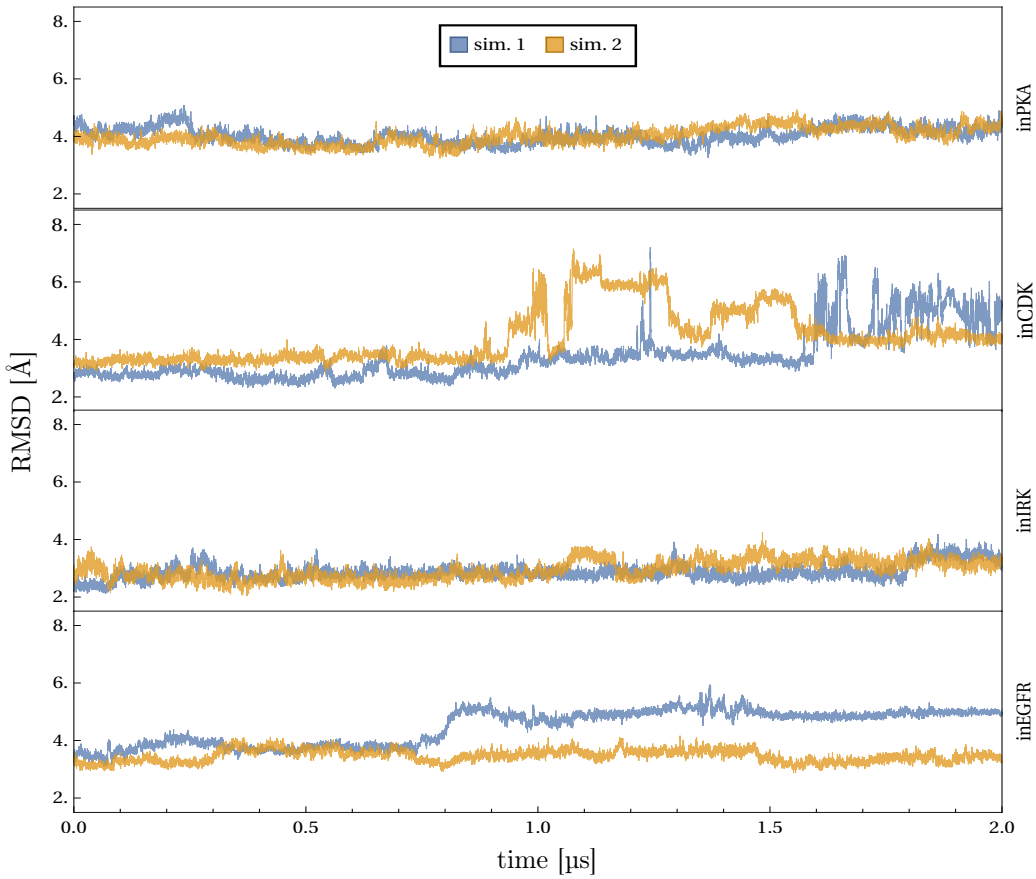


Figure 3.1: RMSDs of inactive kinases heavy atoms against first frames of their respective production stages during two production stage's fragments considered for further analysis ('in' and 'a' prefixes denote inactive and active forms respectively).

Here inactive versions of CDK and EGFR probe sometimes distinguished conformational basin, whereas PKA and IRK are mainly restricted to one macrostate, which is also noticeable in the degree of deviation from mean values (>2 Å for the former vs <1 Å for the latter). On the contrary all active versions of kinases seem to remain in one conformational basin with smaller degree of RMSD variations than inactive versions. These observations are in agreement with the hypothesis that active structures should have more restricted conformational space. RMSDs of all kinases during all simulations remained sufficently stable to validate the use of the obtained trajectories in further analysis.

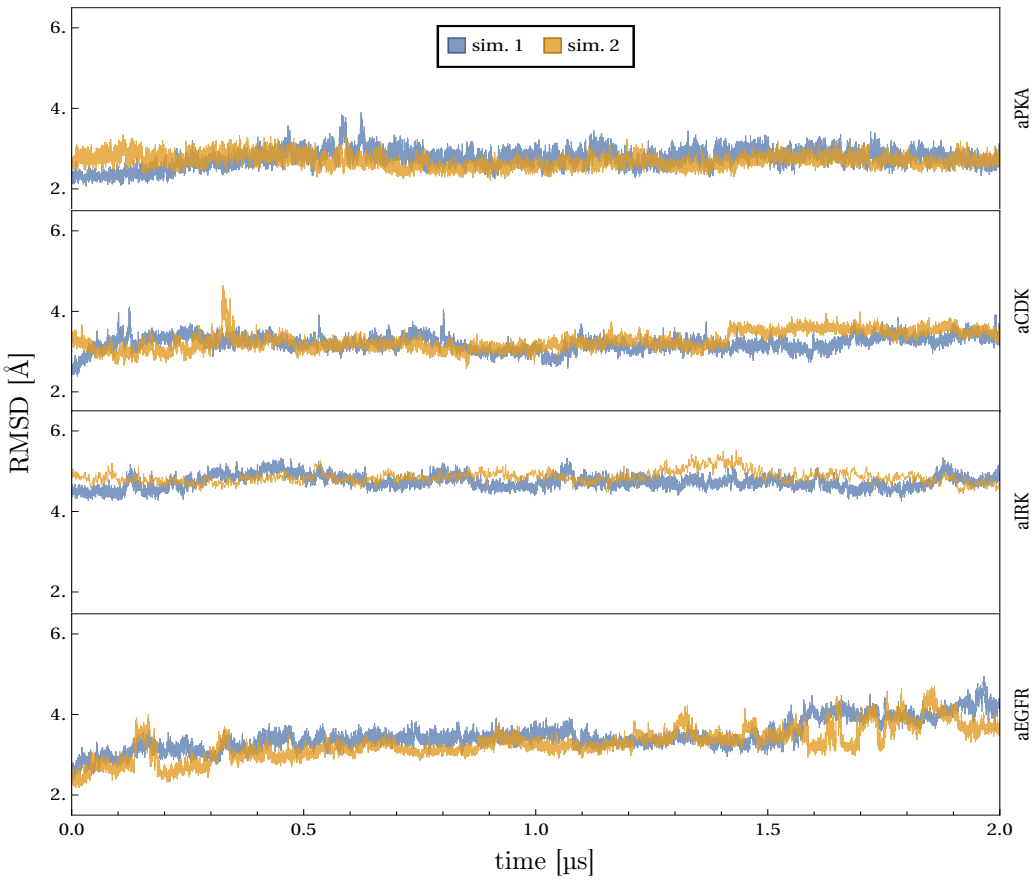


Figure 3.2: RMSDs of active kinases heavy atoms against first frames of their respective production stages during two production stage's fragments considered for further analysis.

3.2. Global descriptors results

3.2.1. RMSD analysis

For each pair of simulations RMSD matrices were generated as explained in Subsection 2.2.1. Their values were then averaged and presented in Fig 3.3.

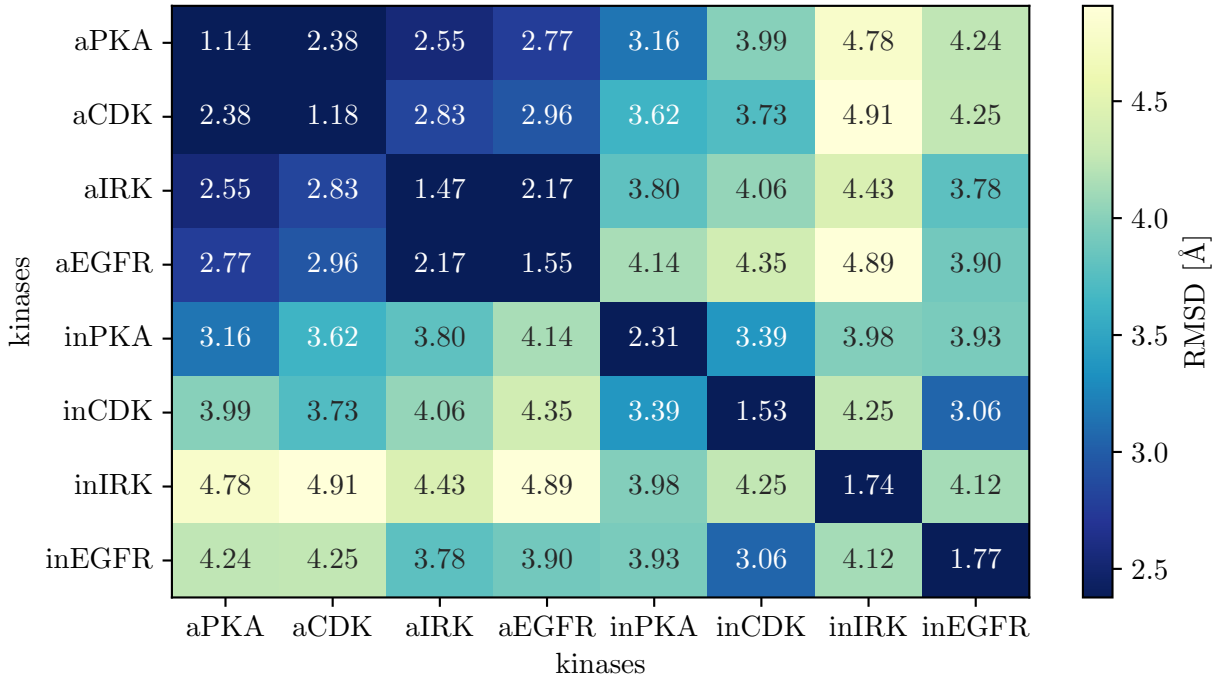


Figure 3.3: Average RMSD between kinase structures.

The first thing to notice is the structure in the upper-left corner of the matrix where the RMSDs of active kinases with respect to themselves are gathered. Differences in dynamics of activated intragroup kinases are less than intergroup ones, which means that the trajectories of closely related kinases preserves their structures. Though the above relation doesn't hold for inactive structures. Mutual RMSD values in active kinases group are lower than in correspoding inactive lower-right corner group, meaning that active kinases adopt more specific and stable conformations. Moreover RMSD of active kinases against inactive ones of the same type are of similar magnitude to those between active against inactive of different models, which indicates that conformationsl change upon activation of individual kinases is considerable, being greater than differences between distinct active structures. Distances between inactive models exhibit similarity to those between actives-inactives and those between different active kinases are smaller than those between different inactive ones and smaller than these between active and inactive states of the same kinase. This allows conclusion that large conformational change happens upon activation, that brings diverse geometries to a common structure.

3.2.2. Subspace overlap analysis

Similar analysis was performed using subspace overlap of 10 first principal modes. Matrix containing values of this descriptor between each pair of simulations is shown in Fig. 3.4.

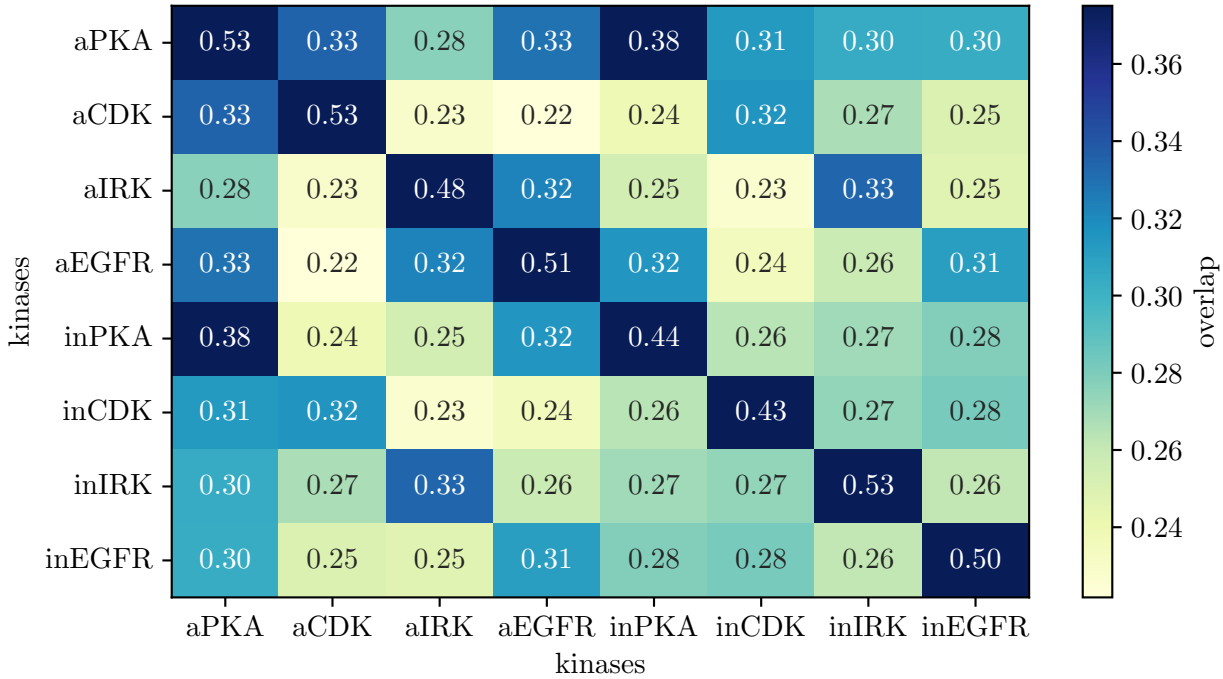


Figure 3.4: Subspace overlap values between each simulation pairs.

Here it's easy noticeable that in the case of active structures there's higher overlap within groups than between groups, thus indicating some degree of evolutionally conserved similarities in global dynamic modes. On the other hand intergroup active states overlap is generally lower than respective one between active and inactive forms, which suggests that there exist some motions determined by sequence irrespective of large conformational change occuring during the activation process. Furthermore there are no visible patterns for intragroup or intergroup differences between inactive structures, hence leading to reasoning that specific movement pattern arises only upon activation. Finally, active forms exhibit slightly higher self-overlap than inactive ones meaning that kinases dynamic behaviors in functional forms are more restricted.

3.2.3. Overall space overlap analysis

Analogous matrix was created for overall space overlap that is presented below in Fig 3.5.

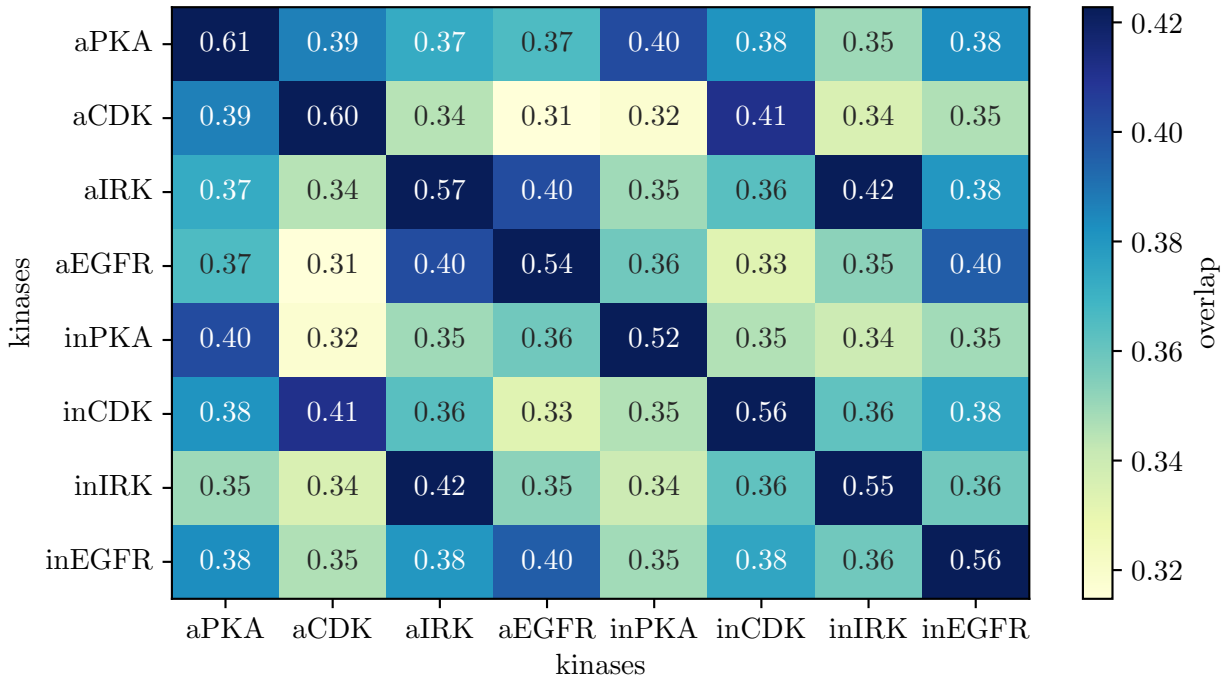


Figure 3.5: Overall space overlap values between each simulation pairs.

As a more general version of the previous descriptor, overall space overlap exhibits some similar characteristics:

- there is a higher degree of overlap within active kinases intragroup than intergroup and activated kinases trajectories highly self overlap, indicating relatively narrow dynamic space upon activation,
- inactive kinases don't demonstrate this relation,
- active and inactive versions within one kinase type have similar degree of overlap as respective intragroup overlap between active states.

Despite these similarities, intergroup overlap for activated structures is similar to the intergroup one for inactive ones. Other conclusions are identical to the ones already stated in Subsection 3.1.2. Results here are akin to these from subspace overlap analysis, which means that they are dominated by protein's global movements.

3.3. Local descriptors results

3.3.1. RMSF analysis

In order to inspect the fluctuations of residues along protein chain, plot was drawn in Fig 3.6. Two bars were added to it presenting different secondary structures coloristically and indicating areas of important core motifs.

For whole structure there are generally lower fluctuations observed in activated state. Biggest differences occur in 2nd-5th and 7th motif regions (C-helix, hinge, HRD, DFG, APE) and

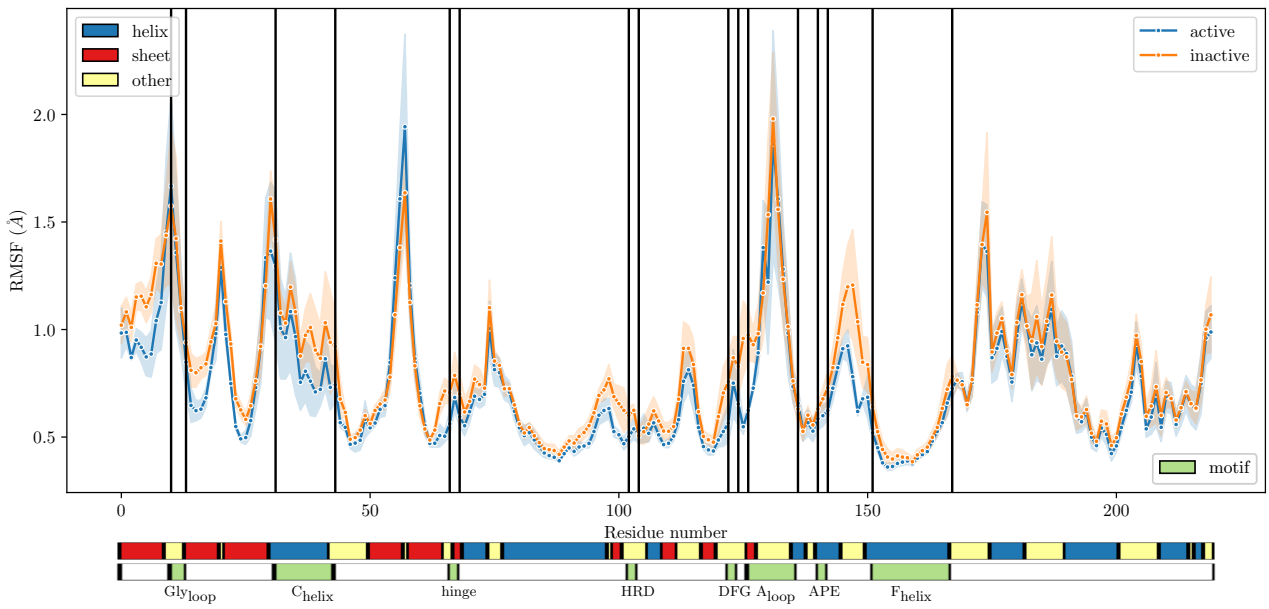


Figure 3.6: Mean RMSF values within active and inactive kinases plotted against bars depicting secondary structures and important core motifs. Shaded area corresponds to interval within one standard deviation range from mean.

similarly in 1st and 2nd sheet and 1st helix. It indicates that these fragments became stabilized upon activation within whole kinase family. Notably aspartic acid residues in both HRD and DFG motifs are directly involved in catalytic reaction, thus their fixation upon activation is in agreement with functional role.

To inspect such effects more precisely, ratio between RMSFs of active and inactive kinases were plotted in Fig 3.7.

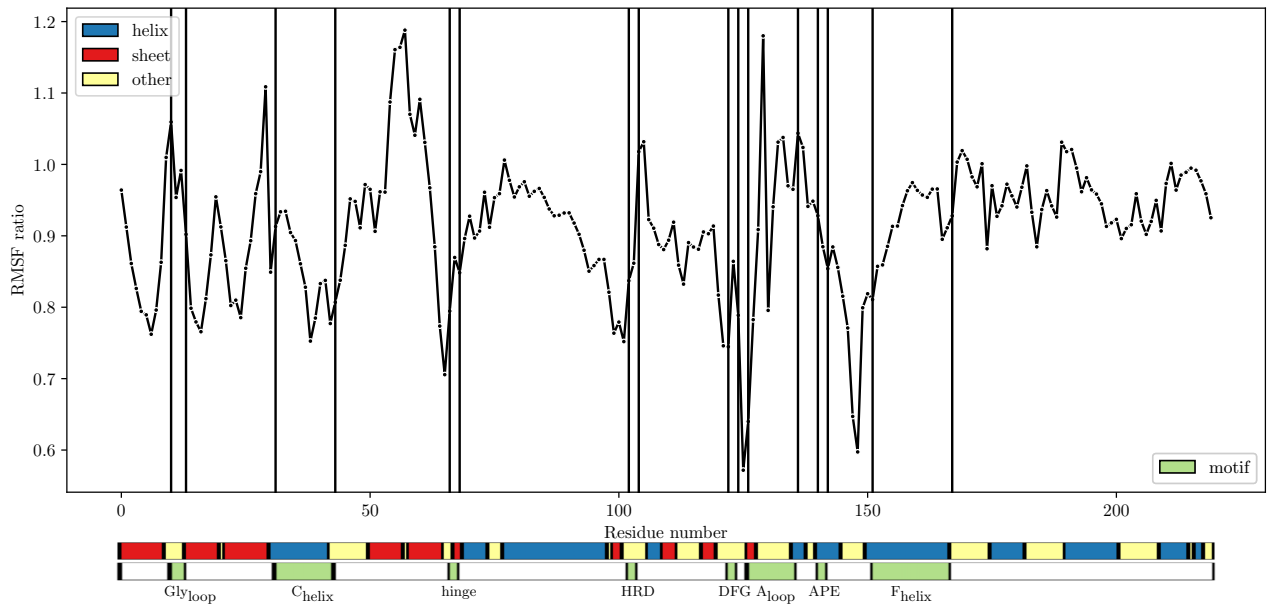


Figure 3.7: Ratio between RMSF means of active and inactive kinases plotted against bars depicting secondary structures and important core motifs.

Alike previous Figure, there are areas characterized by more significant changes, which confirm aforementioned inferences.

To inspect more bulk properties of RMSF, values for each model were averaged and presented in the form of barplot in Fig 3.8.

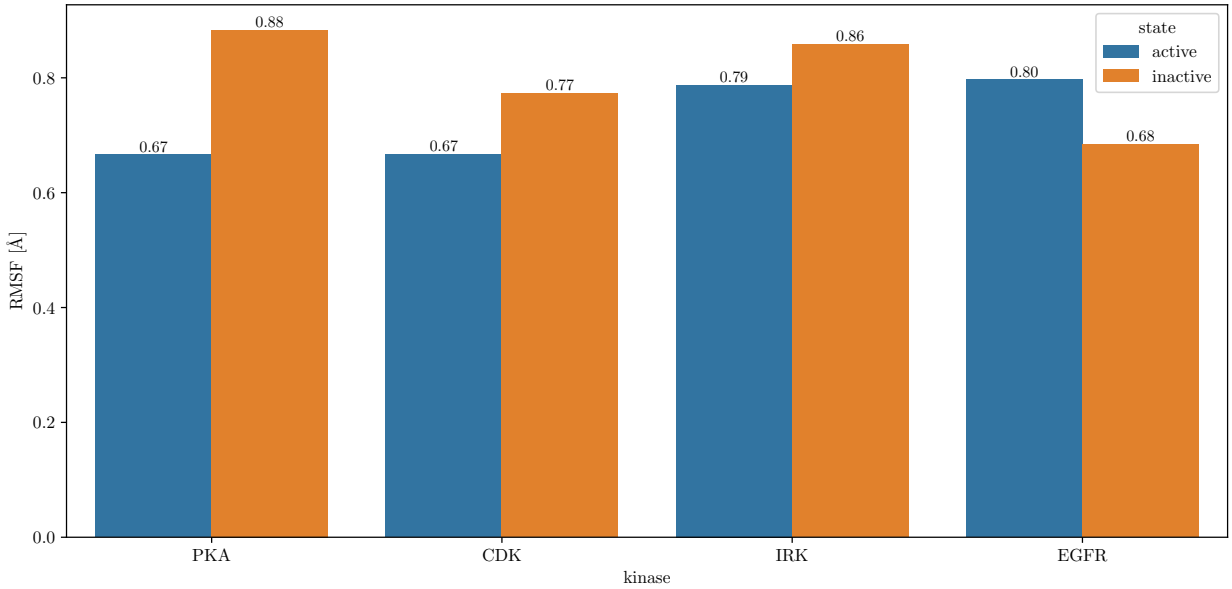


Figure 3.8: Average RMSF for all kinase structures.

Here active kinases from the same family exhibit similar level of fluctuations, which indicates that local motions could be conserved within them. In addition, residues from active serine/threonine kinases probe smaller space than their inactive structures, showing possible stabilizating effect of activation. This relation doesn't equally hold for tyrosine kinases probably due to issues with active EGFR model - due to the novel interaction with modelled loop, ATP kept diffusing out of the binding site, leading to several simulation's restarts.

Lastly, values from active and inactive structures were plotted in form of histogram, with smoothing continuous visualization filter applied, in Fig. 3.9.

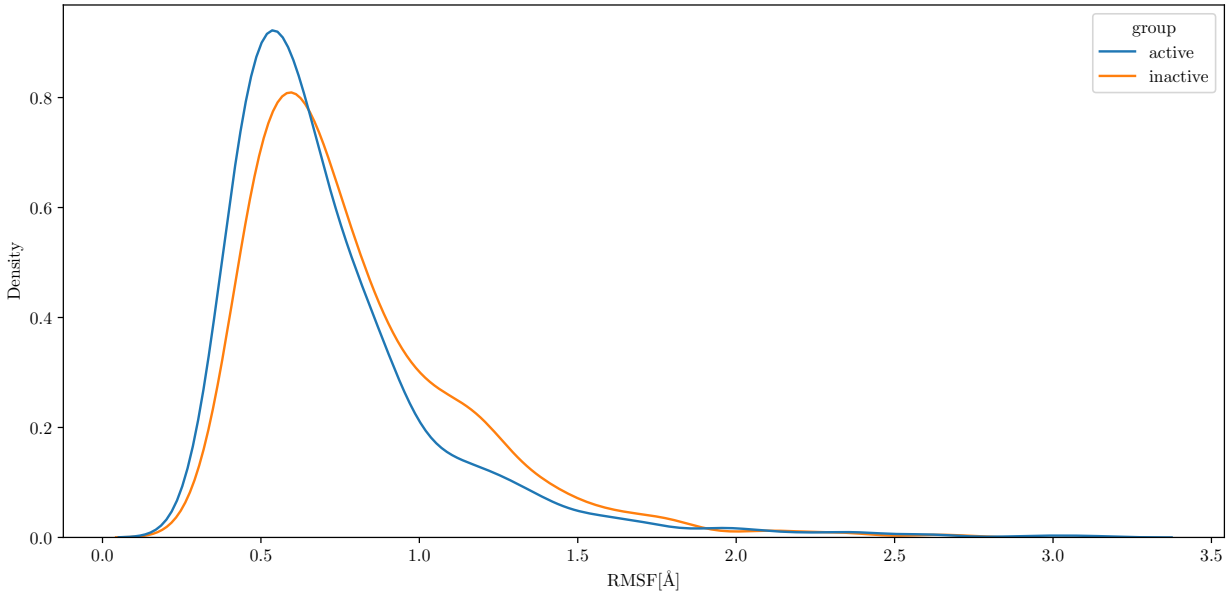


Figure 3.9: RMSF histograms for active and inactive kinase forms.

Shapes of both distributions are similar, which could indicate that the nature of local movements is conserved regardless of activation state. However active kinases tend to have less fluctuations comparing to inactive kinases, which means that there is less space for probing, thus activation stabilizes their movements.

3.3.2. Jensen-Shannon Divergence analysis

To examine local movement patterns in torsion angle space Jensen-Shannon Divergences between each pair of active kinase structures and followingly for pairs of inactive models were calculated for each residue. Means and standard deviations within these groups were then plotted along secondary structures and important core areas in Fig. 3.10. For closer inspection, differences between JS-div values of active and inactive kinases were plotted in Fig. 3.11.

Firstly, it's worth to note that around functionally significant fragments of HRD and DFG motifs, Jensen-Shannon divergence is generally lower between active states than between inactive ones, which means that activation results in narrowing down the probing spaces of these areas to similar local conformations, in conclusion functionally important. Inverse effect can be observed for areas of Glycine rich-loop, C-helix and A-loop. JS-divs of these fragments increase after the activations, meaning that these structures adopt unique configurations in active state, possibly linked to the specific regulatory mechanisms.

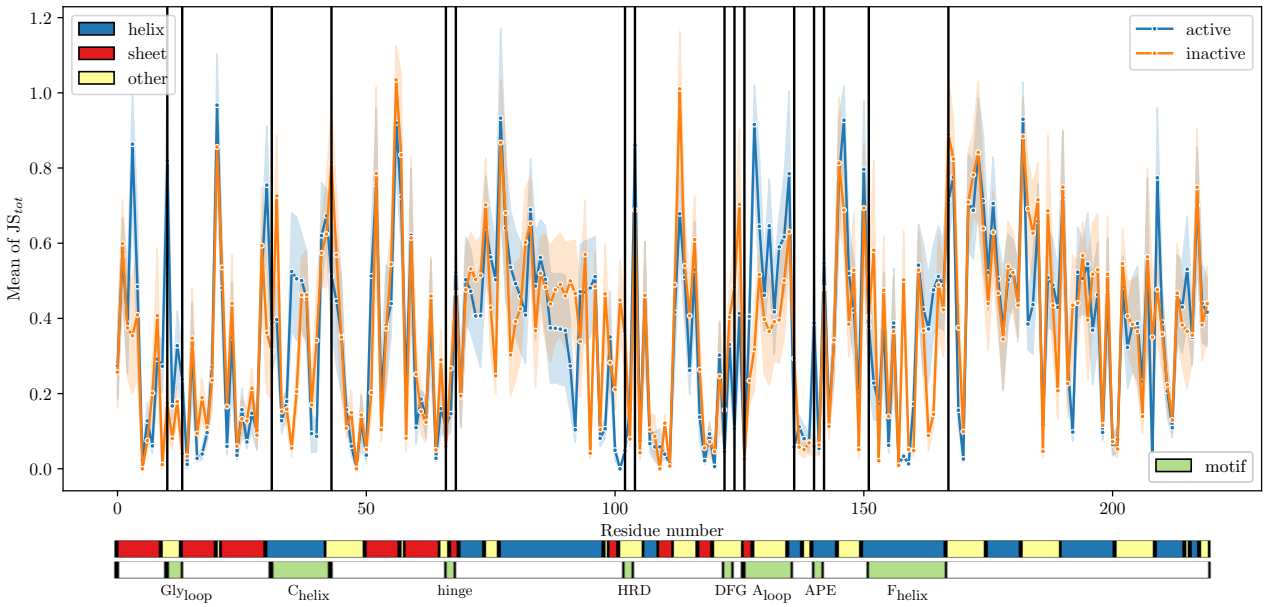


Figure 3.10: JS-div means within active and inactive kinases plotted against bars depicting secondary structures and important core motifs.

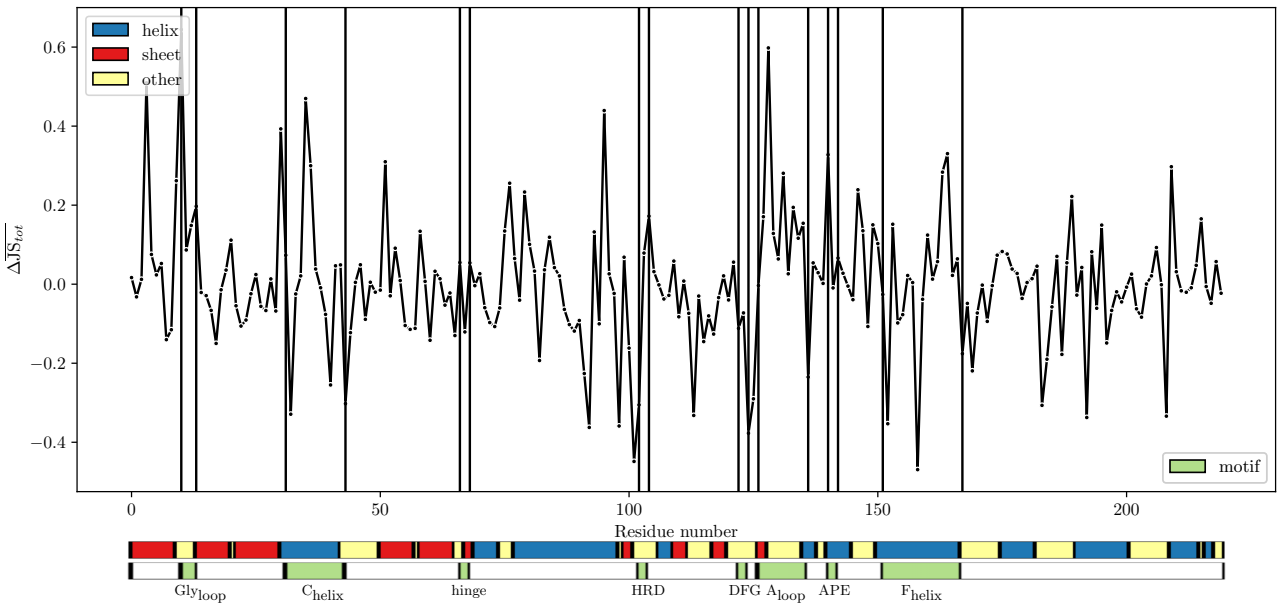


Figure 3.11: Difference between active and inactive structures JS-div means plotted against bars depicting secondary structures and important core motifs.

In order to gain more detailed insight into local characteristics, plot of mean Jensen-Shannon divergences between pairs of active and inactive states of the same type kinase was drawn in Fig. 3.12.

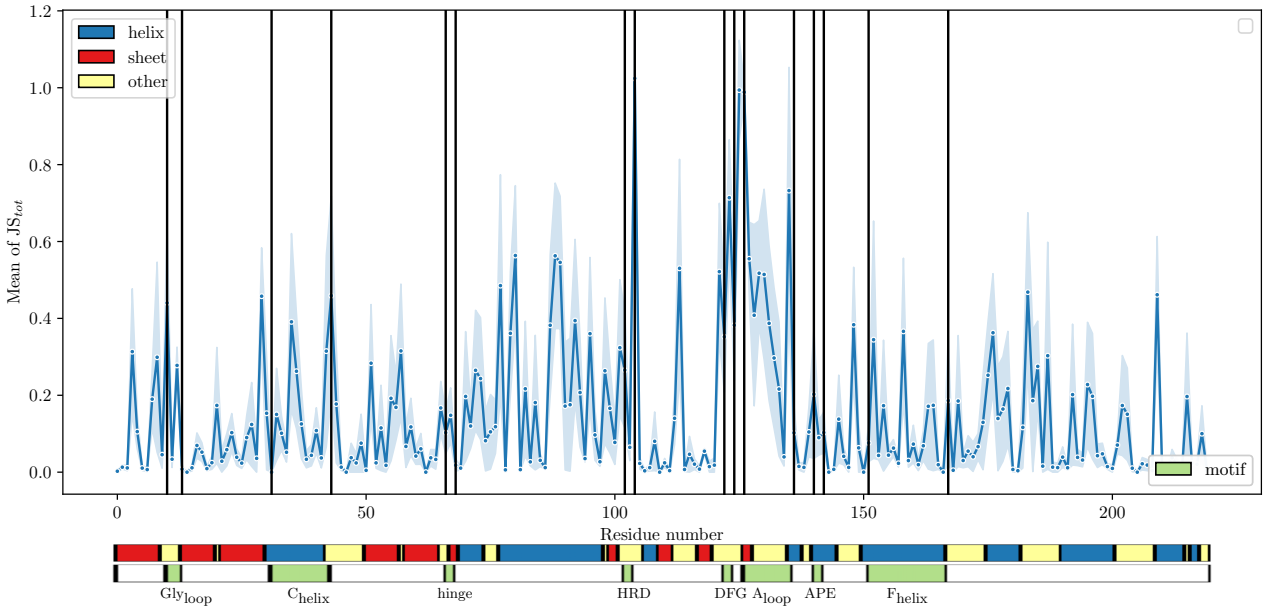


Figure 3.12: Jensen-Shannon Divergence between pairs of active and inactive states averaged over all 4 considered kinases, plotted against bars depicting secondary structures and important core motifs.

Highest observable differences occur in the vicinity of A-loop and DFG-motif confirming large conformational changes of these fragments during transition between active and inactive state. It is in line with the details of the activation process, which include phosphorylation-induced structuring of the activation loop and repositioning of the DFG motif leading to formation of the R-spine. Of note is also large difference in case of Asp residue from the HRD motif which is likely related to its adopting specific structure in the active state and more relaxed in the inactive state.

Finally Jensen-Shannon divergence between pairs of simulations values were averaged and gathered in the matrix as presented in Fig. 3.13.

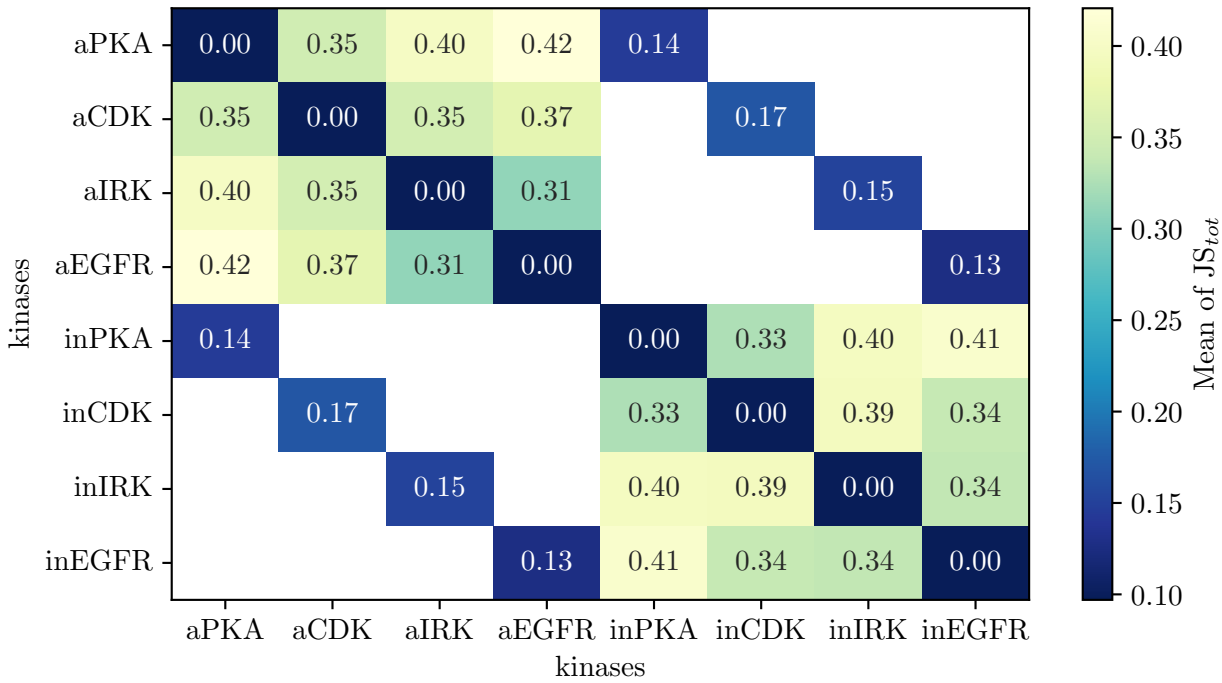


Figure 3.13: Mean of JS$_{tot}$ for each kinase.

The main fact to notice here is that active and inactive kinases of the same type have much lower JS-div than any any other pair of different models, revealing existence of specific, individual protein movement patterns. It indicates that local dynamics in torsion space at residue level is much more dependent on protein sequence than global motions.

3.4. Statistical testing

Simulation, for each considered system, was split into 8 blocks of 500ns that were used to generate descriptor matrices according to procedure presented in Subsection 2.2.4. For each pair of compared kinases set of 16 samples of random variables defined in Eqs 2.35 and 2.36 was created. Significance of differences between structures were examined using Wilcoxon's rank test for paired results due to lack of underlying analytical distribution. Overally 84 tests were performed examining: 1) different states of the same kinases, 2) differences between structures within active and inactive groups respectively. All tests indicated **significance** with $p \ll 0.01$ (all p-values were equal due to the fact that sufficient level of difference was reached for such non-parametric character of the test). There was no need to test differences between blocks of the same trajectory as that would turned out negatively due to the chosen testing method. No multiple-hypotheses testing correction was applied because of lack of differentiation between any results and thus to avoid any arbitrality in decision which ones should be recognized as non-significant.

Chapter 4

Conclusions

In this work we examined differences in dynamic properties of two representants from tyrosine and serine/threonine kinase groups. Two independent molecular dynamics simulations were conducted for all considered structures. Multiple descriptors were chosen to gain insight of both local and global characteristics of their motions.

Analysis of the global differences shown that active kinase catalytic subunits share highly conserved geometry, with level of similarity correlated to evolutionary distance space. On the other hand distances between inactive models exhibit similarity to those between different states within one kinase type. Moreover distances between different active kinases are smaller than those between different inactive ones and smaller than these between active and inactive states of the same kinase, leading to conclusion that large conformational change happens upon activation, bringing diverse geometries to a common structure. These conclusions are in agreement with hypotheses stated at the beginning of this work that:

1. there are evolutionary preserved structure and global dynamics similarities within kinase families and
2. despite the considerably large differences in structure and dynamics of inactive forms, activation leads to the adoption of alike geometries possessing similar global movements, which is the evidence of the functional specialization of active state.

Analysis of local changes occuring during activation has further specified this view indicating ordering into common geometry of fragments that are known to be involved in catalytic reaction like HRD, DFG motifs. On the contrary, regulatory regions (C-helix, Glycine rich-loop and activation loop) adopt distinct configurations upon activation, likely indicating their importance for specific kinase characteristic. In contrast to global motions, dynamics at residue level show relatively little degree of similarity between different kinases, but are highly similar between active and inactive structures of the same protein, indicating large sequence-dependence. This, with the previous notices, confirms the view that large scale dynamics is encoded by tertiary protein structure, whereas local dynamic characteristic is dependent on the primary structure and remains unique for each kinase irrespective of its activation state.

Bibliography

[1] Lubert Stryer, Jeremy M. Berg, Gregory J. Gatto, Justin Hines, and John L. Tymoczko. *Biochemistry*. Palgrave Macmillan, Austin Boston New York Plymouth, 10 edition, May 2023.

[2] Quincy Teng. *Structural Biology: Practical NMR Applications*. Springer Nature, January 2013.

[3] Bernhard Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, New York, October 2009.

[4] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, November 2020.

[5] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, December 2007. Publisher: Nature Publishing Group.

[6] Scott A. Hollingsworth and Ron O. Dror. Molecular Dynamics Simulation for All. *Neuron*, 99(6):1129–1143, September 2018. Publisher: Elsevier.

[7] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The Protein Kinase Complement of the Human Genome. *Science*, 298(5600):1912–1934, December 2002. Publisher: American Association for the Advancement of Science.

[8] Gerard Manning, Gregory D. Plowman, Tony Hunter, and Sucha Sudarsanam. Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, 27(10):514–520, October 2002. Publisher: Elsevier.

[9] S. K. Hanks and T. Hunter. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 9(8):576–596, May 1995.

[10] Eric D Scheeff and Philip E Bourne. Structural Evolution of the Protein Kinase–Like Superfamily. *PLoS Computational Biology*, 1(5):e49, October 2005.

[11] Susan S. Taylor and Alexandr P. Kornev. Protein kinases: evolution of dynamic regulatory proteins. *Trends in Biochemical Sciences*, 36(2):65–77, February 2011.

[12] Alexandr P. Kornev, Nina M. Haste, Susan S. Taylor, and Lynn F. Ten Eyck. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17783–17788, November 2006.

[13] Juliette Martin, Krishanpal Anamika, and Narayanaswamy Srinivasan. Classification of Protein Kinases on the Basis of Both Kinase and Non-Kinase Regions. *PLoS ONE*, 5(9):e12460, September 2010.

[14] Alexandr P. Kornev and Susan S. Taylor. Dynamics-Driven Allostery in Protein Kinases. *Trends in Biochemical Sciences*, 40(11):628–647, November 2015.

[15] Piotr Setny. Conserved internal hydration motifs in protein kinases. *Proteins*, 88(12):1578–1591, December 2020.

[16] M. Born and J. R. Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys. (Leipzig)*, 389(20):457–484, 1927.

[17] P. Güttinger. Das verhalten von atomen im magnetischen drehfeld. *Z. Physik*, 73(3):169–184, 1932.

[18] W. Pauli. Principles of wave mechanics. In *Handbuch der Physik*, volume 24, page 162. Springer, Berlin, 1933.

[19] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56(4):340–343, 1939.

[20] H. Hellmann. *Einführung in die Quantenchemie*. Franz Deuticke, Leipzig, 1937.

[21] A. R. Leach. *Molecular modelling: principles and applications*. Prentice Hall, 2001. OCLC: 45008511.

[22] J. E. Jones. On the determination of molecular fields. —II. from the equation of state of a gas. *Proc. Royal Soc. Lond.*, 106(738):463–477, 1924.

[23] David A. Pearlman, David A. Case, James W. Caldwell, Wilson S. Ross, Thomas E. Cheatham, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1):1–41, September 1995.

[24] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540040211.

[25] Walter R. P. Scott, Philippe H. Hünenberger, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, and Wilfred F. van Gunsteren. The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A*, 103(19):3596–3607, May 1999. Publisher: American Chemical Society.

[26] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98–103, 1967.

[27] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19213690304.

[28] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N.\log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, June 1993.

[29] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, November 1995.

[30] H. J. C. Berendsen et al. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[31] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[32] M. Parrinello and A. Rahman. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.*, 45(14):1196–1199, 1980.

[33] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[34] Karl E. Petersen. *Ergodic Theory.* Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1983.

[35] Berk Hess. Convergence of sampling in protein simulations. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 65(3 Pt 1):031910, March 2002.

[36] Hong Qian. Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Physical Review E*, 63(4):042103, March 2001. Publisher: American Physical Society.

[37] Michael E. Wall. Ligand Binding, Protein Fluctuations, And Allosteric Free Energy. *AIP Conference Proceedings*, 851(1):16–33, August 2006.

[38] H. Matsuda. Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 62(3 Pt A):3096–3102, September 2000.

[39] Benjamin J. Killian, Joslyn Yundenfreund Kravitz, and Michael K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *The Journal of Chemical Physics*, 127(2):024107, July 2007.

[40] Christopher L. McClendon, Lan Hua, abriela Barreiro, and Matthew P. Jacobson. Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *Journal of chemical theory and computation*, 8(6):2115–2126, July 2012.

[41] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945. Publisher: [International Biometric Society, Wiley].

[42] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(Web Server issue):W296–W303, July 2018.

[43] Mats H. M. Olsson, Chresten R. Søndergaard, Michal Rostkowski, and Jan H. Jensen. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, February 2011. Publisher: American Chemical Society.

[44] Chresten R. Søndergaard, Mats H. M. Olsson, Michał Rostkowski, and Jan H. Jensen. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, July 2011. Publisher: American Chemical Society.

[45] D. van der Spoel, P. J. van Maaren, and H. Berendsen. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *J. Chem. Phys.*, 108:10220–10230, 1998.

[46] A. Wallqvist and R. D. Mountain. Molecular Models of Water. I. Derivation and Description. *Rev. Comp. Ch.*, 13(Chap. 4), 1999.

[47] C. L. Wennberg et al. Lennard-jones lattice summation in bilayer simulations has critical effects on surface tension and lipid properties. *J. Chem. Theory Comput.*, 9(8):3527–3537, 2013.

[48] C. L. Wennberg et al. Direct-space corrections enable fast and accurate lorentz–berthelot combination rule lennard-jones lattice summation. *J. Chem. Theory Comput.*, 11(12):5737–5746, 2015.

[49] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.*, 91(1):43–56, 1995.

[50] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7(8):306–317, 2001.

[51] D. van der Spoel et al. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[52] B. Hess et al. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[53] S. Pronk et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[54] S. Páll et al. Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In Stefano Markidis and Erwin Laure, editors, *Solving Software Challenges for Exascale*, Lecture Notes in Computer Science, pages 3–27. Springer International Publishing, 2015.

[55] M. J. Abraham et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[56] S. Páll and B. Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.*, 184(12):2641–2650, 2013.

[57] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, June 2010.

[58] Kristin L. Meagher, Luke T. Redman, and Heather A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *Journal of Computational Chemistry*, 24(9):1016–1025, July 2003.

[59] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%28199709%2918%3A12%3C1463%3A%3AAID-JCC4%3E3.0.CO%3B2-H.

[60] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81:511–519, 1984.

[61] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[62] N. Michaud-Agrawal et al. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32(10):2319–2327, 2011.

[63] R. J. Gowers et al. MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*, pages 98–105, 2016.

[64] T. E. Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[65] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Comput Sci Eng*, 13(2):22–30, 2011.

[66] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009.

[67] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 27–28, February 1996.

[68] BIOVIA, Dassault Systèmes, Discovery Studio Visualizer, v24.1.0.23298, San Diego: Dassault Systèmes, 2024.

[69] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[70] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, April 2021.

[71] Wolfram Research, Inc. Mathematica, Version 11.3. Champaign, IL, 2018.

[72] The GIMP Development Team. Gimp, 2019.

[73] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[74] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[75] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.