DISCOVERY OF NEW POTENTIAL PYRIDO[2,3-D]PYRIMIDIN-7-ONE BASED CDK4 INHIBITORS USING MOLECULAR MODELLING APPROACHES

OMAR HUSHAM AHMED AL ATTRAQCHI

UNIVERSITI SAINS MALAYSIA

2023

DISCOVERY OF NEW POTENTIAL PYRIDO[2,3-D]PYRIMIDIN-7-ONE BASED CDK4 INHIBITORS USING MOLECULAR MODELLING APPROACHES

by

OMAR HUSHAM AHMED AL ATTRAQCHI

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

February 2023

ACKNOWLEDGEMENT

Alhamdulillah, all praise to Allah (SWT) who has blessed me and granted me the ability to complete my Master project, I am most grateful to him. I would like to thank everyone who has supported me in my study, as I would not have been able to complete my project without their help and contribution.

I would like to also express my gratitude to my parents and my sisters for supporting me throughout my study, I owe them deeply for their continuous love and support.

I am also grateful to the previous director of the Centre for Drug Research, Prof. Dr. Vicknasingam Balasingam who showed great care and assistance for me. Finally, I would like to thank my supervisor Prof. Dr. Mohd Nizam Mordi for his given assistance and corrections to the project.

Omar Husham Ahmed Al Attraqchi

August 2022

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

2D	2-Dimentional
3D	3-Dimentional
4D	4-Dimentional
5D	5-Dimentional
Å	Angstrom
ADME	Absorption, distribution, metabolism and excretion
ALA	Alanine
ANN	artificial neural networks
ASP	Aspartate
BBB	Blood brain barrier
CADD	Computer-aided drug design
CAK	Cdk-activating kinase
CDK	Cyclin-dependent kinase
CDK4	Cyclin-dependent kinase 4
CNS	Central nervous system
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity index analysis
CYP	Cytochrome P450
DUD-E	Directory of Useful Decoys-Enhanced
ETKDG	Torsional Knowledge-Distance Geometry
fs	Femtosecond
GA	Genetic algorithm
GAFF	general amber forcefield
GA-MLR	Genetic algorithm-multiple linear regression
GI	Gastrointestinal

GLY	Glycine
GPU	Graphics processing unit
HIS	Histidine
HR+	Hormone receptor +
IC50	Inhibitory concentration
ILE	Isoleucine
k	Kelvin
kDa	Kilo Dalton
LEU	Leucine
LogP	Lipophilicity hydrophilicity coefficient
LYS	Lysine
MAE	Mean absolute error
MD	Molecular dynamics
MIF	molecular interaction field
MLR	Multiple-linear regression
MMFF94	Merck molecular forcefield 94
MOE	Molecular Operating Environment
NPT	Constant number of structures, constant pressure, and constant temperature
Ns	Nanoseconds
PDB	Protein Databank
PLS	Partial least squares
ps	Picosecond
Q^2	Correlation coefficient of cross validation
QSAR	Quantitative structure-activity relationship
R^2	Coefficient of determination
RMSD	Root mean squared deviation

RMSE	Root mean squared error
RMSF	root mean squared fluctuations
SA	Synthetic accessibility
VAL	Valine
VMD	Visual molecular dynamics
Y_{obs}	Observed activity
Y_{pred}	Predicted activity

PENEMUAN PERENCAT CDK4 BAHARU BERPOTENSI BERASASKAN PIRIDO[2,3-D]PIRIMIDINA-7-ON MENGGUNAKAN PENDEKATAN PEMODELAN MOLEKUL

ABSTRAK

Kinase yang bergantung kepada siklin (CDK) adalah kelas enzim pengawalaturan yang memodulasi pelbagai sifat biokimia sel, antaranya pembahagian sel. Enzim CDK yang berbeza telah menjadi tapak sasaran biologi yang berguna untuk memerangi pelbagai jenis kanser. Enzim CDK4 telah meningkat ekspresinya dalam beberapa jenis kanser, antaranya kanser payudara. Penggunaan molekul drug kecil bagi menghalang aktiviti enzim CDK4 telah dibuktikan sebagai pendekatan yang sah secara klinikal. Walau bagaimanapun, pada masa ini pendekatan ini mengalami pelbagai kekurangan antaranya aktiviti yang tidak mencukupi. Antara perencat penting bagi enzim CDK4 adalah sebatian yang berasaskan kerangka pirido[2,3-D]pirimidina-7-on. Dalam kajian ini, kaedah reka bentuk drug berbantuan komputer (CADD) termasuklah kaedah yang berdasarkan ligan dan yang berdasarkan struktur, telah diaplikasikan kepada sebatian yang diterbitkan daripada kerangka ini. Kaedah CADD yang digunakan menunjukkan korelasi antara sifat fizikokimia dan aktiviti, juga memberi pemahaman mengenai proses pengikatannya. Kaedah CADD berasaskan ligan yang digunakan dalam kajian ini termasuk hubungan struktur-aktiviti kuantitatif (QSAR) dan pemodelan farmakofor. Kedua-dua model 2D dan 3D-QSAR telah dibangunkan dan disahkan dengan ketat, menghasilkan model yang boleh digunakan untuk meramalkan aktiviti sebatian yang belum disintesis atau bagi mengutamakan sebatian calon bagi penilaian biologi. Model ini juga boleh memberikan kefahaman berkenaan keperluan struktur bagi menghasilkan aktiviti yang tinggi. Model yang terbaik menghasilkan

parameter validasi berikut: pekali penentuan (R^2=0.913), pekali korelasi pengesahan silang (Q^2=0.745) dan pekali korelasi aktiviti ramalan ($R^2_{(pred)}$=0.764). Model farmakofor untuk sebatian aktif menunjukkan keperluan ciri farmakofor yang terdiri daripada satu penerima ikatan hidrogen, satu hidrofobik dan dua gelang aromatik. Kaedah berasaskan struktur antaranya pengedokan molekul juga telah dijalankan. Prosedur pengedokan telah disahkan melalui eksperimen pengedokan semula iaitu berdasarkan pengukuran sisihan punca min kuasa dua (RMSD) bagi membuktikan bahawa hasil dok adalah boleh dipercayai. Pengedokan juga menunjukkan interaksi pengikatan penting yang berlaku dalam tapak sasaran, antaranya ikatan hidrogen dengan residu HIS95 dan VAL96. Akhirnya, penyaringan maya pangkalan data CHEMBL yang mengandungi 2,086,898 sebatian telah dijalankan dengan menggunakan model QSAR yang telah dibangun dan disahkan, dan prosedur pengedokan molekul untuk mengenal pasti sebatian novel yang berpotensi sebagai perencat CDK4. Terdapat 7 sebatian aktif yang telah dikenal pasti. Ciri farmakokinetik dan ketoksikan sebatian tersebut telah dianalisis dan menunjukkan bahawa sebatian ini mempunyai sifat seperti drug dan boleh digunakan sebagai sebatian peneraju bagi pembangunan agen teraputik antikanser yang berguna secara klinikal.

DISCOVERY OF NEW POTENTIAL PYRIDO[2,3-D]PYRIMIDIN-7-ONE BASED CDK4 INHIBITORS USING MOLECULAR MODELLING APPROACHES

ABSTRACT

Cyclin-dependent kinases (CDKs) are a class of regulatory enzymes that modulate various biochemical properties of the cell such as cell division. Different CDK enzymes have been shown to be promising biological targets for combating different types of cancer. In particular, the CDK4 enzyme has been observed to be overexpressed in several types of cancer including breast cancer. The usage of small-drug molecules that inhibits the activity of the CDK4 enzyme has proved to be clinically valid approach. However, currently available solutions suffer from various limitations such as insufficient activity. Among the most notable inhibitors of the CDK4 enzyme are the compounds based on the pyrido[2,3-d]pyrimidin-7-one scaffold. In this study, computer-aided drug design (CADD) methods including ligand-based and structure-based methods have been applied on the compounds derived from this scaffold. The applied CADD methods revealed the correlation of the physicochemical properties with the activity and gave insights into the binding process. The used ligand-based CADD methods included quantitative structure-activity relationship (QSAR) and pharmacophore modeling. Both 2D- and 3D-QSAR models were developed and rigorously validated, which yielded models that can be used for predicting the activity of compounds that are not yet synthesized or to prioritize the biological evaluation of candidate compounds. The models also provide insights regarding the structural requirements for achieving high activity. The best model's validation parameters included: coefficient of determination (R^2=0.913),

correlation coefficient of cross validation (Q^2=0.745) and coefficient of determination of the test set ($R_{2(pred)}$=0.764). The obtained pharmacophore model detailed the required pharmacophoric features for active compounds which included a hydrogen-bond acceptor, hydrophobic and two aromatic rings. Structure-based methods were also conducted including molecular docking. The docking procedure was validated by a re-docking experiment which proved that the docking results are reliable based on root mean-squared deviation (RMSD) measurement. The docking also revealed the important binding interactions which included critical hydrogen-bonds with HIS95 and VAL96 residues at the active site of the enzyme. Finally, virtual screening of 2,086,898 compounds from the ChEMBL database was carried out using the developed and validated QSAR models and the molecular docking procedure to identify novel potentially active CDK4 inhibitors. A total of 7 potentially active compounds were identified. The pharmacokinetic properties and the toxicity of the identified compounds were analyzed and showed that these compounds possess drug-like properties and can serve as lead-compounds for developing clinically useful anti-cancer therapeutic agents.

CHAPTER ONE

INTRODUCTION

Cancer is a highly prevalent disease that is characterized by rapid and abnormal cellular divisions in certain tissues (Otto and Sicinski, 2017, Hassanpour and Dehghani, 2017). Despite the relative success in chemotherapy which involves the usage of chemical substances to combat the cancerous cells, various limitations and obstacles exist. Those limitations include the ineffectiveness of the marketed therapeutics against certain types of cancers or in certain patients (Bøttcher et al., 2019, Yamamoto et al., 2019, Otto and Sicinski, 2017). Also, the emergence of resistance in the course of the therapy is one major issue in the treatment of many types of cancer (Spring et al., 2019, Pandey et al., 2019). Hence, the development of novel and effective anti-cancer agents is a continuous demand in the drug design and discovery field.

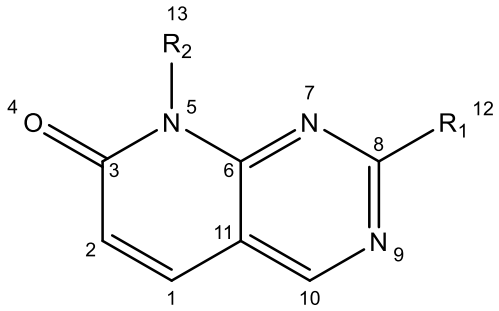
The treatment of cancer via small drug molecules mainly involves targeting certain biomolecules that are vital for the cancerous cells and inhibiting/disabling those targets leads to the death of the cell (Peyressatre et al., 2015). Different biomolecules have been identified that are present in the cancerous cells in a larger proportion compared to the normal cells. This difference makes them appealing targets for small drug molecules that are intended to act as anti-cancer agents (Otto and Sicinski, 2017). Among the most notable targets for combating different types of cancer is the cyclin-dependent kinase (CDK) family of enzymes. The CDK enzymes play important roles

in the regulation of processes related to cell division (Cicenas and Valius, 2011). It has been observed that certain CDK enzymes are overexpressed in cancerous cells. Thus, inhibiting those enzymes can halt the cancer cell division (Braal et al., 2021).

This study focuses on the cyclin-dependent kinase 4 (CDK4) enzyme. This enzyme has been overexpressed in different types of cancer, particularly breast cancer (Bøttcher et al., 2019). Also, targeting this enzyme by small drug molecules has been shown to be an effective and promising approach for the treatment of breast cancer. However, the existing CDK4 inhibitors suffer from certain issues and limitations, which mainly include the development of resistance throughout the course of the treatment as well as the lack of effectiveness in certain cases (Condorelli et al., 2018, Spring et al., 2019). Thus, the discovery of novel CDK4 inhibitors and to better characterize the structural requirements for inhibiting this target is of prime importance.

Different chemical scaffolds were discovered that possessed CDK4 inhibitory activity. Among the most notable and promising scaffolds is the **pyrido[2,3-d]pyrimidin-7-one scaffold (1)**. Derivatives based on the pyrido[2,3-d]pyrimidin-7-one scaffold were shown to have good inhibitory concentration (IC50) values against the CDK4 enzyme while exhibiting a good selectivity towards other protein kinases (Mascarenhas et al., 2010). However, limited computation studies have been conducted on the pyrido[2,3-d]pyrimidin-7-one based CDK4 inhibitors. In the modern drug design and discovery paradigm, computational molecular modeling has become

an essential and invaluable tool, particularly in early computational studies (Chandrasekaran et al., 2018).



Pyrido[2,3-d]pyrimidin-7-one scaffold (1)

In this study, various computer-aided drug design (CADD) methods were performed to analyze the current CDK4 inhibitors and characterize the essential structural and physicochemical requirements for achieving high activity. Also, quantitative structure-activity relationship (QSAR) and pharmacophore models were developed, which can be used to predict the activity quantitatively and qualitatively, respectively. Thus, this serves to accelerate the identification of novel and potentially active CDK4 inhibitors via virtual screening. Pharmacokinetic analysis and toxicity assessment are applied to the newly identified compounds to ensure that they possess drug-like properties. Hence, the newly identified structures can serve as lead compounds for developing CDK4 inhibitors anti-cancer agents.

Figure 1.1 shows a schematic diagram of the steps followed in the molecular modeling process. In the first part, QSAR and pharmacophore models were developed and validated, also the molecular docking protocol was established and validated. In

the second part, the developed models and the docking procedure were used in a virtual screening campaign to identify novel and potentially active CDK4 inhibitors. Pharmacokinetic analysis and toxicity assessment were conducted on the identified compounds to validate their drug-like properties. Also, the top identified compound was subjected to molecular dynamics (MD) simulation to further validate its activity as a CDK4 inhibitor.

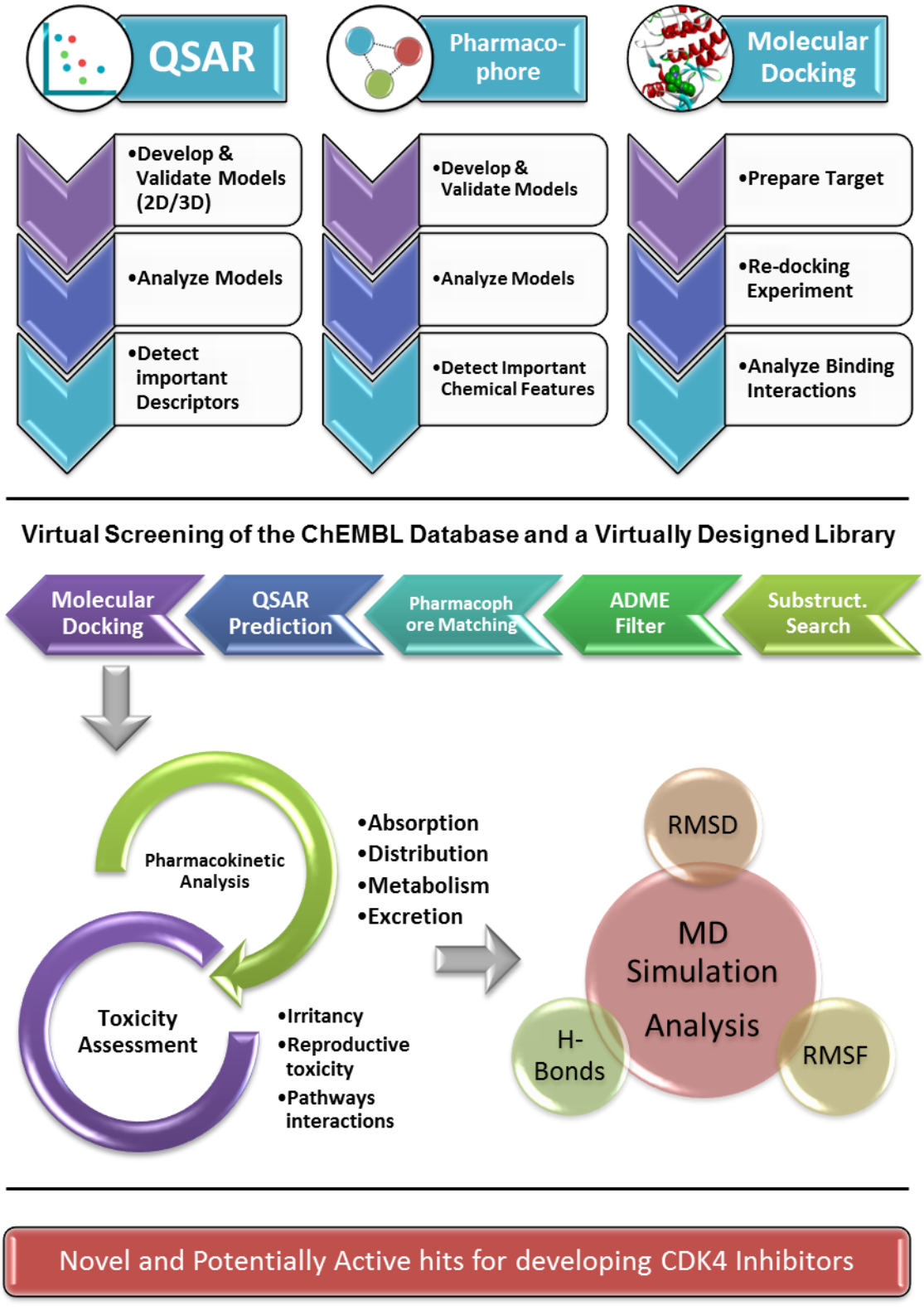


Figure 1.1: The main workflow followed in this study.

1.1 Objective of the Study

The aim of the study is to use CADD methods to identify novel and potentially active CDK4 inhibitors based on the pyrido[2,3-d]pyrimidin-7-one scaffold.

1.2 Specific Objectives

i. To develop and validate predictive QSAR models that can estimate the activity of potential CDK4 inhibitors.

ii. To develop and validate pharmacophore models that can identify potentially active CDK4 inhibitors.

iii. To predict the essential binding interactions between the CDK4 enzyme and the inhibitors by using molecular docking.

iv. To identify novel and potentially active CDK4 inhibitors via virtual screening and molecular dynamics simulation.

CHAPTER TWO

LITERATURE REVIEW

2.1 CDK Enzymes in Cancer

The CDK enzymes belong to the family of serine/threonine kinases are responsible for modulating the progression of the cell cycle (O'leary et al., 2016). This family of enzymes operates by catalyzing the phosphorylation of certain targets which would subsequently lead to the advancement of the cell cycle events. In order for the CDK enzymes to achieve their activity, they need to form heterodimeric complexes with their respective cyclin components (Dey et al., 2020). Only when a CDK enzyme is bound to its cyclin component, it can catalyze the phosphorylation of its targets. Also, some of the CDK enzymes require to be phosphorylated to become in the active state, often they are phosphorylated by the CDK-activating kinase (CAK) enzyme, which is also referred to as the CDK7 enzyme (Morgan, 1997). The transition between the active and inactive state of the CDK enzymes is strictly regulated, as improper functioning could lead to undesired division and proliferation of the cell (Morgan, 1995, Besson et al., 2008).

The various processes that occur in the cell as it grows and divides are referred to as the cell cycle (Kasten and Giordano, 1998). There are four distinct phases in each cell cycle. The first phase is known as the Gap-1 (G1) phase, in which the cell grows and increases in size. The second phase is called the Synthesis (S) phase, in which the cell synthesizes a copy of its current DNA. The third phase is the Gap-2 (G2) phase,

in which the cell prepares for the division process. The fourth and final phase is the Mitosis (M) phase, in which the cell divides into two individual cells (O'leary et al., 2016, Zhang et al., 2021). In mammalian cells, various CDK enzymes are in control of different phases of the cell cycle process, and they act as an on/off switch on the progression of the cell cycle events (Vermeulen et al., 2003). When a particular CDK enzyme is active, the cell is permitted to progress through the stage in which that CDK enzyme is modulating. On the other hand, if a particular CDK enzyme is inactive, the cell is not permitted to pass through the stage which that CDK enzyme controls and the cell cycle is halted (Zhang et al., 2021). Abnormal expression of different CDK enzymes has been observed in various types of cancer (Peyressatre et al., 2015). For example, the CDK1 enzyme was found to be overexpressed in advanced melanoma as well as B lymphoma (Abdullah et al., 2011, Zhao et al., 2009), and the CDK2 enzyme is overexpressed in laryngeal squamous cell cancer, breast cancer and advanced melanoma (Peyressatre et al., 2015).

The involvement of the CDK enzymes in the development of different cancer types gave them value in different therapeutic applications (Cicenas and Valius, 2011). For instance, the CDK1 enzyme possesses diagnostic value in breast and esophageal cancers (Hansel et al., 2005). Also, the expression and activity of the CDK2 enzyme have been used clinically to monitor the prognosis of oral and ovarian cancers (Mihara et al., 2001, Kim et al., 2008). Thus, CDK enzymes represent appealing targets in cancer therapy, both as targets for combating the cancer cell as well as for purposes of clinical diagnosis and prognosis (Cicenas and Valius, 2011).

2.1.1 CDK4 Enzyme

The CDK4 enzyme is one of the most targeted CDKs for the treatment of cancerous cells (Spring et al., 2019). Aberrant expression of the CDK4 enzyme has been observed in different types of cancer including breast, ovarian and oral cancers (Peyressatre et al., 2015). The central role of the CDK4 enzyme in the progression of the cell cycle process is the main cause of its direct involvement in the development of cancer (Sherr et al., 2016). In particular, at the early stages of the cell cycle process, the CDK4 and the CDK6 enzymes carry out the phosphorylation of the retinoblastoma family of proteins, mainly the pRb, p130 and p107. Phosphorylation of those retinoblastoma proteins leads to a subsequent cascade of events required for successful cell division (Kasten and Giordano, 1998). Thus, amplification or overexpression of the CDK4 enzyme would lead to abnormal and rapid cell divisions.

The active site of the CDK4 has been analyzed from the available crystal structures of the enzyme (in particular, PDB-ID: 2W96). It has been also compared to the active sites of other CDK enzymes such as CDK2 (PDB-ID: 1FIN) and CDK6 (PDB-ID: 1G3N). The key amino acid residues in the active site of the CDK4 enzyme include HIS95, VAL96, THR120 and ASP97 (Shafiq et al., 2012). All of these amino acid residues are often found to interact with the inhibitors of the CDK4 enzyme. For example, the carbonyl backbone of the VAL96 residue often acts as a hydrogen bond acceptor with a hydrogen bond donor of the inhibitor. Also, the ASP97 residue contains a negatively charged group in its side chain that can interact with the inhibitor present in the active site via an ionic bond (Coleman et al., 1997, Shafiq et al., 2012). The other amino acid residues of the active site of the CDK4 enzyme seem to be less significant in interacting with the inhibitors.

Various studies have shown that targeting the CDK4 enzyme by small molecule inhibitors is a validated approach against certain cancers such as breast cancer (Sherr et al., 2016). The inhibition of the CDK4 enzyme prevents the cell cycle from advancing to the G1 phase, hence no cell division occurs. Figure 2.1 illustrates the role of the CDK4 enzyme in the cell cycle flow. Currently, three clinically approved CDK4 inhibitors are used for the treatment of metastatic HR+ (hormone receptor+) breast cancer, namely, **Palbociclib (2)**, **Abemaciclib (3)** and **Ribociclib (4)**. Those CDK4 inhibitors are used usually in combination with other agents for the treatment of cancer (Sammons et al., 2017, Bøttcher et al., 2019, Braal et al., 2021). Despite the widespread of those agents, certain limitations exist and call for the development of new and improved CDK4 inhibitors. The current therapeutic agents suffer from limitations such as the lack of efficiency and the ineffectiveness in many clinical cases (Spring et al., 2019). Also, the emergence of resistant cancerous cells against the clinical CDK4 inhibitors during the course of the treatment is a major concern (Pandey et al., 2019, Condorelli et al., 2018).

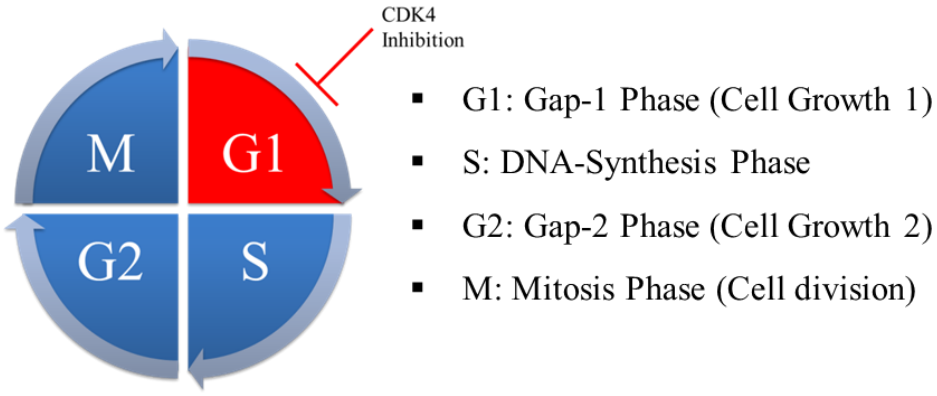
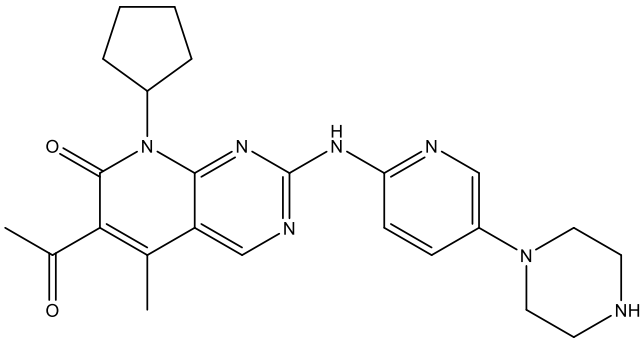
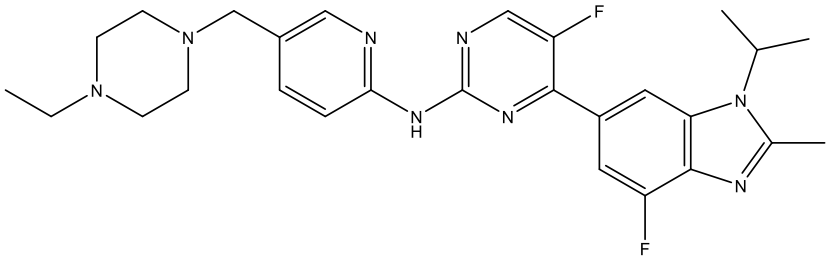


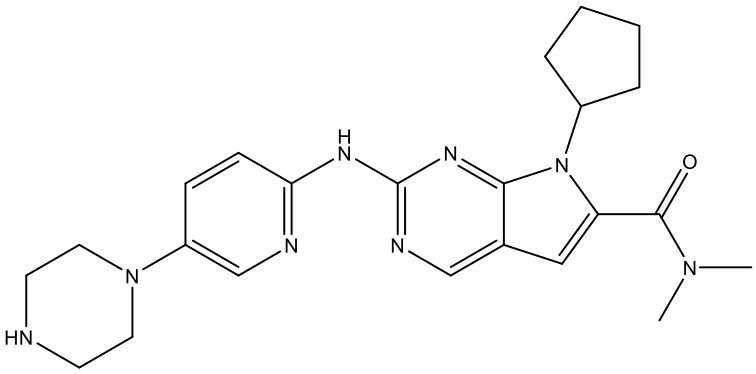
Figure 2.1: The role of the CDK4 enzyme in the cell cycle flow.



Palbociclib (2)



Abemaciclib (3)

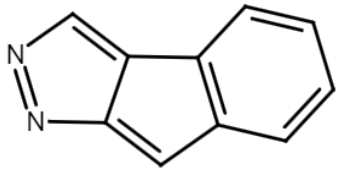


Ribociclib (4)

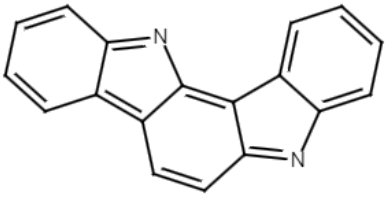
2.1.2 The pyrido[2,3-d]pyrimidin-7-ones Scaffold

There have been various scaffolds that were identified as CDK4 inhibitors such as **indenopyrazole (5)**, **indolocarbazoles (6)** and **tetrahydronaphthyridine (7)** (Muzaffar-ur-Rehman et al., 2017, O'leary et al., 2016, Zhu et al., 2003, Zha et al., 2018). Among the most notable and promising scaffolds that possess CDK4 inhibitory activity is the **Pyrido[2,3-d]pyrimidin-7-one scaffold (1)**. The main reason why this scaffold is considered to be promising is the high selectivity of its derivatives towards the CDK4 enzyme in comparison with the other kinases (Mascarenhas et al., 2010). Developing selective CDK4 inhibitors is highly desired as non-selective agents that bind to other kinases can cause a variety of non-desired adverse effects (Toogood et al., 2005, McInnes et al., 2004). Other advantages of the pyrido[2,3-d]pyrimidin-7-one scaffold include the relative ease of synthesis. Indeed, a large number of derivatives of the pyrido[2,3-d]pyrimidin-7-one scaffold have been synthesized and their biological activities have been reported in the literature as well as in databases of chemical compounds (VanderWel et al., 2005, Barvian et al., 2000). However, very limited computational studies have been performed on this class of compounds, and they have

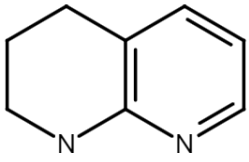
not been thoroughly studied in terms of binding modes, correlation of activity with physicochemical properties and pharmacokinetic analysis (Barvian et al., 2000). Additionally, the currently highly active compounds of this class generally do not have high drug-like properties, for example, they tend to have high molecular weight, high number of rotatable bonds and high hydrophobicity (Pollastri, 2010, Veber et al., 2002). These properties can halt the further development of these compounds into clinically useful drugs. Thus, more computational studies are required to further understand the pyrido[2,3-d]pyrimidin-7-one based CDK4 inhibitors.



Indenopyrazole (5)



Indolocarbazoles (6)



Tetrahydronaphthyridine (7)

2.2 CADD Methods in Developing CDK4 Inhibitors

Computer Aided Drug Design (CADD) methods have become essential tools in the drug discovery and development process. CADD methods mainly include structure-based and ligand-based methods (Hassan Baig, Ahmad et al. 2016). Figure 2.2 shows a general scheme of the most used CADD methods. Structure-based methods such as molecular docking are used when the crystal structure of the target is available (Śledź and Caflisch, 2018). On the other hand, ligand-based methods such as QSAR can be applied even in the absence of the target crystal structure. CADD methods can provide predictive models to predict the activity of compounds that are not yet synthesized and prioritize their biological evaluation (Roy et al., 2015a). Also, these methods can provide valuable information regarding the binding process of the inhibitors to the target as well as the structural requirements for achieving high activity (Śledź and Caflisch, 2018). In the upcoming sections, a brief description of each method and how it has been applied to analyze and develop CDK4 inhibitors are provided.

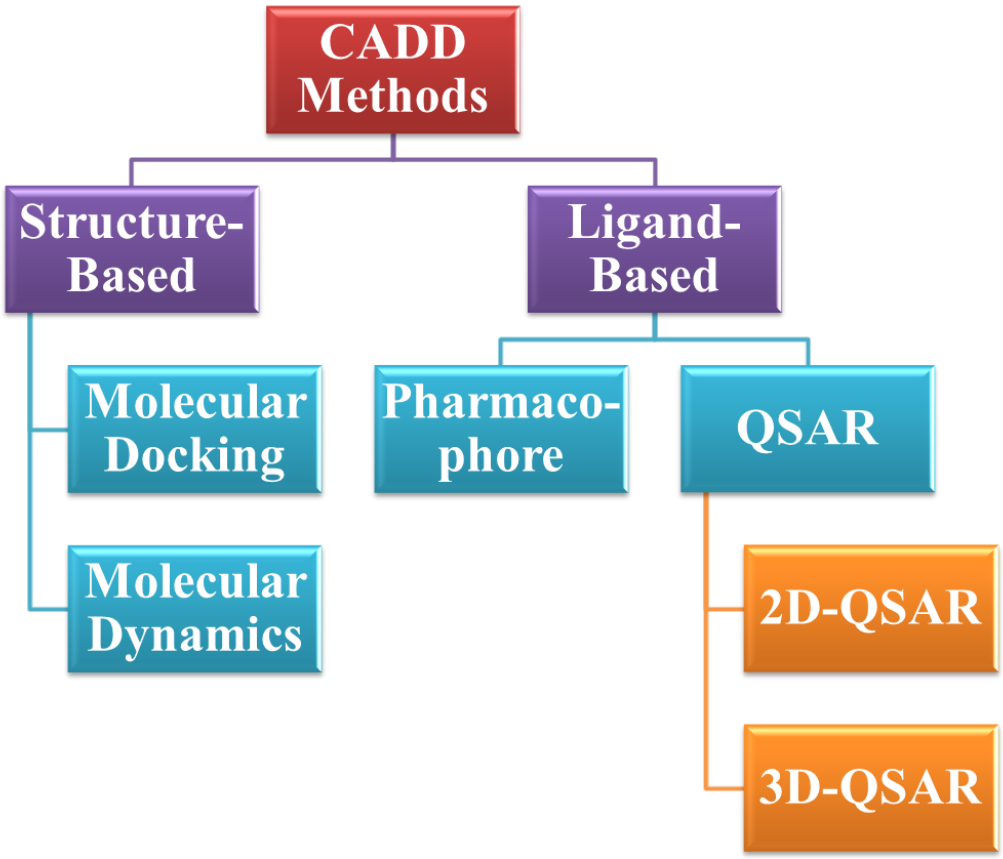


Figure 2.2: General scheme of the commonly used CADD methods based on their type.

2.2.1 QSAR Modeling

QSAR modeling is a valuable method in drug design and discovery. It is used to derive mathematical models that can be used to predict the activities of new compounds (Tropsha, 2010). Also, QSAR models provide details regarding the correlation between molecular properties and the activity (Golbraikh et al., 2016)**.** Despite the availability of many different QSAR methods, they all share a basic concept, which is to relate the physicochemical properties to biological activities. The used physicochemical properties in model development are often referred to as molecular descriptors in the context of QSAR modeling, and their types can vary widely (Yap,

2011, Prasanna and Doerksen, 2009). Figure 2.3 illustrates the general QSAR modeling workflow.

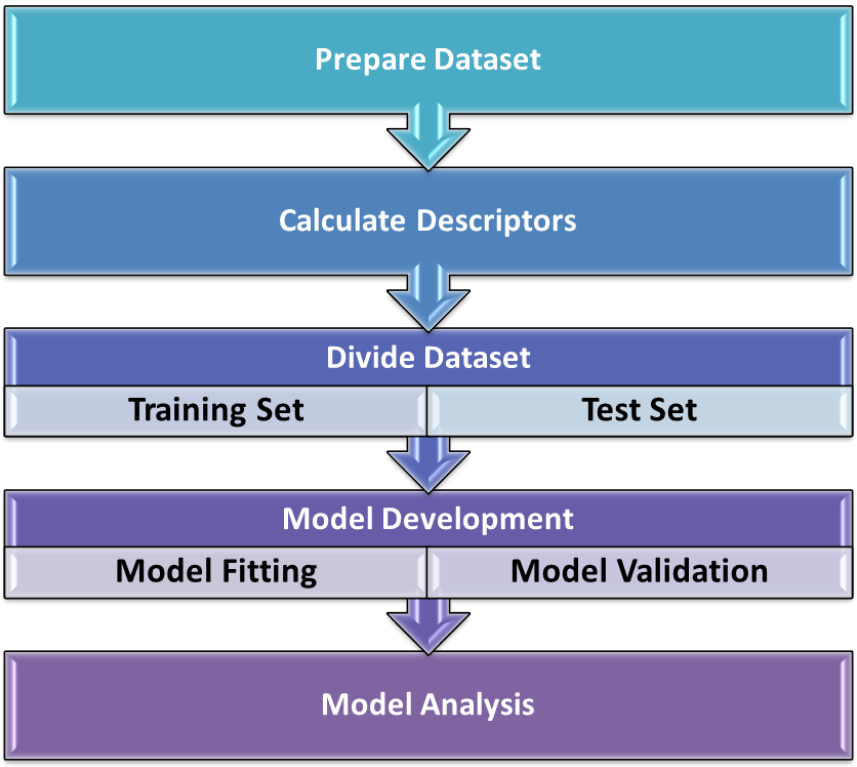


Figure 2.3: The general steps of QSAR modeling.

Another component that can vary greatly in QSAR methods is the choice of the statistical method that is used to correlate the molecular descriptors with the activity (Gramatica, 2007). Linear methods such as the multiple linear regression (MLR) method are very popular as they offer rapid model generation as well as ease of use. For instance, the MLR method does not require any specific parameters to be fined tuned when fitting the model. Also, the results of the linear methods can be easily and directly interpreted. Often, a regression coefficient is assigned to each molecular descriptor used which reflects the significance and contribution of the descriptor to the

predicted activity (Roy et al., 2016, Ambure et al., 2015, Saxena and Prathipati, 2003). On the other hand, non-linear regression methods such as artificial neural networks (ANN) require more parameter fine tuning and a relatively long time for building the model, especially when using a large dataset. Also, in most cases of QSAR model development, interpretation of the resulted model is not possible due to the non-linear nature of those methods. However; non-linear methods such as the ANN method can often provide better quantitative predictions of the activity compared to the linear methods (Devillers, 1996). Currently, there are various QSAR methods available. Since some QSAR methods may perform poorly on certain datasets it is often preferred that more than one method is applied to the available dataset (Roy et al., 2015b).

2.2.1(a) 2D-QSAR

The 2D-QSAR methods use only molecular descriptors that are computed using the 2D-graph of the structures. Thus, the information included in those descriptors is independent of the 3D conformation or orientation of the structure (Roy and Narayan Das, 2014). Although a large number of diverse molecular descriptors is available and can be computed easily, the number of descriptors used in the model building process should be kept relatively low to avoid over-fitting the model. For this purpose, a descriptor-selection method is often used that selects only a small portion of the computed descriptors for model building (Saxena and Prathipati, 2003). Also, many 2D descriptors are difficult to interpret as they are based on complex computations from the 2D graph of the structure. The main advantage of this method is that it does not require a 3D-conformation generation or alignments of the structures prior to building the model. This advantage also makes 2D-QSAR methods easier to use in the

virtual screening of large databases (Saíz-Urra et al., 2007, Roy and Narayan Das, 2014, Lewis and Wood, 2014).

Several 2D-QSAR studies on different CDK4 inhibitors have been reported, for instance, Divya and co-workers conducted a 2D-QSAR analysis on a dataset composed of 43 CDK4 inhibitors (Divya et al., 2018). The compounds were derivatives of different scaffolds and did not share a single core structure. The dataset was split into a training set (75%) for model fitting and a test set (25%) for model validation. Different combinations of physicochemical properties and statistical methods were used to generate a large number of models in a combinatorial manner. Each model was validated and scored, the top model was retained which had statistically significant parameters such as $R^2 = 0.8947$ and $Q^2 = 0.8909$.

2.2.1(b) 3D-QSAR

Compared with the 2D-QSAR approach, the 3D-QSAR method has been more widely used in the analysis of CDK4 inhibitors. Mainly, this is because of the ability of the alignment-dependent 3D-QSAR methods such as CoMFA to visualize the model onto the compounds of the dataset (Muzaffar-ur-Rehman et al., 2017, Metibemu, 2021, Cai et al., 2014). The visualization shows the regions and nature of chemical substituents that are beneficial or detrimental to the activity in terms of electrostatic and steric properties (Martin, 1998). Various 3D-QSAR models have been reported using datasets based on a variety of chemical scaffolds.

For instance, Muzaffar-ur-Rehman and co-workers used a dataset composed of 93 derivatives based on the indenopyrazole scaffold in an alignment-dependent 3D-QSAR study (Muzaffar-ur-Rehman et al., 2017). The CoMFA and CoMSIA methods were applied. Initially, all the compounds of the dataset were aligned on a template structure. The similarity of the electrostatic and steric fields was used for optimizing the compounds onto the template structure. Statistical analysis was performed via the partial least squares (PLS) method and the obtained model was validated and visualized onto the compounds for analysis. The best model had an R^2 and Q^2 values of 0.946 and 0.845, respectively.

Another similar study was conducted by Cai and co-workers where they used the CoMFA method on a dataset of thieno[2,3-d]pyrimidin-4-yl hydrazine derivatives to develop a 3D-QSAR model (Cai et al., 2014). A total of 48 compounds were used for model fitting and 9 compounds were used as an external validation test set. The statistical validation parameters indicated a reliable mode, for example, the R^2 and Q^2 had values of 0.965 and 0.724, respectively. The visualization of the contour maps of the model gave insights into the required chemical nature of substituents for activity as well as selectivity. It revealed the nature of the electrostatic and steric characteristics of the required substituents for obtaining a high activity.

In another study, Metibemu and co-workers used a different approach in which they combined alignment-independent 3D-descriptors and 2D-descriptors for developing a QSAR model (Metibemu, 2021). A total of 100 CDK4 inhibitors that are derivatives of different chemical scaffolds were used for building the model. The percentages of training and test sets were 70% and 30%, respectively. Initially, a large

number of descriptors were calculated for each compound, then a genetic algorithm was used for descriptors selection. The final model contained a total of 12 molecular descriptors. Also, the model exhibited good statistical parameters including an R^2 value of 0.872.

2.2.2 Pharmacophore Modeling

In pharmacophore modeling, the ligands are represented using their pharmacophore features such as hydrogen bond donors, acceptors and aromatic rings (Schaller et al., 2020). Pharmacophore modeling aims to generate pharmacophore models that capture the required chemical features and their spatial arrangements for binding to a biological target (Qing et al., 2014a). Generally, a set of active ligands is used to derive the models by aligning them and then extracting the necessary pharmacophore features (Schaller et al., 2020, Taminau et al., 2008). The commonly used chemical features include hydrogen bond donors, hydrogen bond acceptors, aromatic rings, positively charged centers, negatively charged centers and hydrophobic groups (Dixon et al., 2006). Once a pharmacophore model is generated, it is validated against a validation set of known active and inactive molecules (decoys) to assess the model's ability to distinguish active from inactive structures (Yang, 2010, Mysinger et al., 2012). Validated models can then be used in virtual screening as a filtering step by only retaining structures that can match the pharmacophore model. The main advantage of pharmacophore modeling in virtual screening is the relatively fast screening rate compared to other methods such as molecular docking. Hence, pharmacophore model matching is often carried out during the initial steps of the virtual screening process (Belal, 2021, Qing et al., 2014b).

In a study by Yuan and co-workers, they developed pharmacophore models that aimed to identify novel and selective CDK4 inhibitors (Yuan et al., 2020). For building the model, a total of 6 compounds that are representative and belonged to diverse chemical scaffolds were used as a training set. For model validation, a decoy set was used which contained compounds that are presumed to be inactive. Those decoy compounds were generated using the Directory of Useful Decoys-Enhanced (DUD-E) algorithm which is a dedicated algorithm for generating inactive compounds for validation purposes. The HipHop algorithm was applied for developing the pharmacophore models. The best obtained model contained seven different pharmacophore features, namely, a hydrogen bond donor, an aromatic ring and three hydrophobic groups. Subsequently, the model was used as a part of a virtual screening campaign to identify novel CDK4 inhibitors from the Specs database. Finally, the compounds filtered by the pharmacophore model were then subjected to molecular docking and visual inspection steps.

Susanti and co-workers used an interesting approach in which they relied on clinically approved CDK4 inhibitors for building a pharmacophore model (Susanti et al., 2021). Specifically, three marketed CDK4 inhibitors, namely, palbociclib, ribociclib, and abemaciclib were used as the training set. For validation, a decoy set of 650 inactive compounds was used to assess the ability of the models to distinguish active from inactive compounds. The pharmacophore models were built with the Molecular Operating Environment (MOE) software. The best model was selected based on various validation parameters including accuracy, sensitivity, and specificity.

The pharmacophore features of the top model included 5 features, namely, two hydrogen bond acceptors, a hydrogen bond donor and two aromatic rings. In the virtual screening process, the pharmacophore was used as a filtering step and was followed by structure-based methods including molecular docking and MD simulation to detect potentially active CDK4 inhibitors.

2.2.3 Molecular Docking

Molecular docking is a widely used drug design method that aims to predict the binding pose of a particular ligand towards a target. Also, the binding energy of the pose is estimated (Torres et al., 2019). Molecular docking methods use search algorithms that generates different poses of the ligand inside the binding site of the target. The generated poses are evaluated using a scoring function which aims to calculate the binding energy using certain terms such as hydrogen bonds (Trott and Olson, 2010). Molecular docking has been applied for different purposes in the design and discovery of CDK4 inhibitors. The most common uses include the analysis of binding poses of known inhibitors as well as the identifying of possibly active novel CDK4 inhibitors, often as a part of a virtual screening campaign (Belal, 2021, Susanti et al., 2021, Metibemu, 2021). Also, it has been used to understand the selectivity of the CDK4 inhibitors toward other CDK enzymes such as the CDK2 enzyme (McInnes et al., 2004).

2.2.3(a) Crystal Structures of CDK4

Although various 3D-crystal structures of the CDK4 enzyme have been crystallized and are deposited in the protein databank (PDB), those crystal structures suffer from

two main problems. The first problem is the lack of a co-crystallized ligand in the active site of the CDK4 enzyme, which could make the validation of the docking protocol more difficult (Day et al., 2009, Vaughan et al., 2006). A common solution to this problem is to align the crystal structure of the CDK4 enzyme onto a crystal structure of a CDK2 enzyme that has a co-crystallized ligand. Then the CDK2 enzyme structure (i. e., the protein part) is removed, leaving the CDK4 enzyme with the co-crystallized ligand of the CDK2 as a protein-ligand complex. Then the re-docking experiment can be performed on this created complex. This approach takes advantage of the similarity between the active sites of the CDK4 and CDK2 enzymes and has been successfully used in the literature repeatedly (Al-Warhi et al., 2020).

The other issue with the available CDK4 crystal structures is that they are in an inactive state, and they need to be in the active state prior to performing docking experiments. Various approaches have been used to tackle this problem, for instance, a homology modeling approach has been used in which a crystal structure of the CDK6 enzyme (active state) was used as a template to correct the CDK4 (inactive state) loops to make it in the active state. This produced a hybrid CDK4 model that was further refined using MD simulation. The hybrid CDK4 model has been successfully used in molecular docking and is available for researchers to use (Shafiq et al., 2012).

2.2.4 Molecular Dynamics

In drug design and discovery, MD simulations are used to investigate binding events of ligands towards a certain target (Śledź and Caflisch, 2018). By simulating the movement of the atoms of the ligand-target complex for a certain period of time, it can

be revealed whether the ligand is binding with the target or not. Also, the binding interactions can be analyzed from the simulation results to understand the binding process (Ramos et al., 2020). MD simulations use newton's laws of motion to predict the movement of atoms in a particular system. They are widely used in drug design, in particular for studying protein-ligand complexes (Huang et al., 2018, Hospital et al., 2015). The MD simulation can reveal whether a complex is stable as well as give insights regarding the non-covalent binding interactions taking place at the binding site (Śledź and Caflisch, 2018).

The main application of MD simulation with regard to the CDK4 inhibitors is in virtual screening. Commonly, MD simulation is performed on top candidate compounds identified by the previous steps of the virtual screening campaign. This is because of the computationally expensive nature of the MD simulations compared to the other drug design methods such as molecular docking; hence it is often limited to few candidate compounds (Mascarenhas et al., 2010, Sarhan et al., 2021). However, MD simulation is also used for other purposes such as analyzing the selectivity of one inhibitor towards other CDK enzymes as well as in understanding the binding interactions (Aixiao et al., 2008).

For example, Aixiao and co-workers combined molecular docking and MD simulation methods to analyze the binding and selectivity of the 2PU inhibitor on the CDK4 and CDK2 enzymes (Aixiao et al., 2008). The 2PU is a small molecule CDK4 and CDK2 inhibitor, however, it is more selective towards the CDK4 enzyme. Initially, the targets were prepared, and the ligand was docked into the active site of each