## ABSTRACT

COOK, BETHANY THERESE. Cheminformatics Modeling of Ara h 2 and CDK Binders Using 3D Docking and the Molecular Chimera Approach. (Under the direction of Dr. Denis Fourches).

Cheminformatics is the field that characterizes, models, designs, and predicts chemicals and their properties using computers. Cheminformatics (especially using 3D molecular docking) is particularly useful to better understand protein-ligand interactions.

Herein, we began our cheminformatics investigation with the analysis of Ara h 2 proteins as they are a known source for allergic reactions produced by peanuts. When an allergic individual is exposed to peanuts, the protein triggers an inflammatory response through the production of histamines. This reaction can be mild to severe (anaphylactic shock). Previous research has identified that anti-inflammatory drugs can combat the detrimental effects of histamine reactions. A natural source of compounds with anti-inflammatory activity has been discovered in blueberries and cranberries in the form of polyphenols. With the use of Schrodinger's Glide software, we were able to molecularly dock 42 known polyphenols into the identified binding site of peanut allergen Ara h 2. This computational study resulted in the identification of 10 compounds predicted to bind Ara h2 and those compounds were recommended for experimental testing. Two out of four tested polyphenols showed the ability to produce structural changes to the Ara h 2 protein. Through computational and experimental efforts, we were able to highlight and identify potential anti-inflammatory inhibitors of peanut allergens.

In another study, we investigated a large series of small molecule inhibitors of protein kinases. Cyclin dependent kinase proteins (CDK) are overexpressed in most cancer cells inhibition. Chemotherapy based on the specific inhibition of CDK proteins is thus of high interest. Previous studies have identified 316 CDK2 inhibitors for which associated crystal structures are available in the PDB. First, we analyzed this set using chemical clustering and 3D docking. Second, we used the *molecular chimera* approach, which allows for the structural fusion of two known CDK2 inhibitors through a common scaffold. The idea behind this research is to identify and combine key scaffolds and substituent groups that could result in new compounds with higher potency and selectivity towards CDK2. Third, we developed QSAR models to compute the inhibition potency ($pIC_{50}$) using 2D descriptors and machine learning

techniques. These models were used to estimate the $pIC_{50}$ values for all newly generated molecular chimera compounds. Our analysis led to the identification of over 300 newly predicted inhibitors with increased binding affinities and predicted $pIC_{50}$ values of 7.0 or higher.

# Cheminformatics Modeling of Ara h 2 and CDK Binders Using 3D Docking and the Molecular Chimera Approach

By

Bethany Therese Cook

A thesis submitted to the Graduate Faculty of
North Carolina State University
In partial fulfilment of the
Requirements for the degree of
Master of Science

Chemistry

Raleigh, North Carolina

2018

APPROVED BY:

| | |
|---|---|
| _____ | _____ |
| Denis Fourches<br>Committee Chair | Gavin Williams |
| _____ | |
| David Reif | |

## DEDICATION

I would like to dedicate this work to my parents, thank you for guiding and molding me into the young lady I am today. While I was growing up, you encouraged me to question the world and explore how things work, no matter how many times I took things apart without putting them back together.

Thank you seems like a small thing to say when compared to all the sacrifices both of you have made to raise Ashley and I. All I am, I owe to you.

I love you!

## BIOGRAPHY

Bethany Therese Cook has always had a passion for education. As a child she would wonder how the world worked and made sure to always challenged herself. Throughout school she would push for what was right and would make sure to go one step further than expected. As a freshman in high school she had the opportunity to join the Henderson County Early College (HCEC) where she ultimately became the first and sole graduate in 2012 to obtain her High School Diploma and Associate in Science Degree.

At the end of high school, Bethany knew she loved science but it was not until she enrolled at the University of North Carolina at Asheville (UNCA), that she found her passion for chemistry. Chemistry finally allowed for her to connect to and see the world in a whole new light. At UNCA, Bethany began to develop her analytical and computational skills under the guidance of Dr. Sally Wasileski. Each project she participated in not only enriched her life but allowed for her to give back to the community, through studying mineral concentrations in prenatal vitamins and by calculating the potential energy for fuel cells from a stepped catalytic surface. In 2016, Bethany graduated from UNCA with a major in chemistry and minors in math and physics, but even with her studies she still had questions about what the world had to offer.

After UNCA, Bethany jumped at the opportunity to continue learning at North Carolina State University (NCSU). She originally intended to pursue Analytical Chemistry but became interested in the computational drug discovery research of Dr. Denis Fourches. She first started working with a collaborator to discover potential compounds to inhibit peanut allergies which ultimately led to her main project of developing Python code to structurally fuse known CDK2 protein inhibitors together in the pursuit of more effective cancer treatments. Throughout her time at NCSU, Bethany has had the opportunity to learn and grow by being pushed towards computer science. Her confidence has developed into that of a true research scientist, by improving her knowledge of chemistry, her ability to effectively communicate her research, and her comfort on working on multiple projects at a time. She will be leaving NCSU with her Master of Science in Chemistry and numerous awards for her research.

## ACKNOWLEDGEMENTS

I would like to acknowledge all on my family and friends for their love and support throughout these years. This thesis would not be possible without you. I would like to specifically thank my sister, Ashley, for being an inspiration while raising my nieces and nephew. You work so hard to provide for those kids and I applaud you for that. When we were younger, we did not always see eye to eye but without our fights, I do not think we would be as close as we are today.

I would like to especially thank Karen, Jackie, Lyniesha and Quibria; without you ladies I would probably not have my sanity. All of you were there for the blood, sweat, and tears (most defiantly tears) of my graduate career and I have you to thank for keeping me going. When I was stuck you would let me talk things out and most off all when I was frustrated, we talked about being entrepreneurs by starting a table flipping business. Among all the bad, we shared tons of laughs and I look forward to more memories to come.

Also, I would like to say thank you to my mentors who have helped me become the strong and confident research scientist I am today. You believed in me when I was trying to determine who I was and helped me find my passion for Chemistry.

Lastly, I would like to acknowledge my dog Charlie (Mr. Char), Ryan and his dog Remus, for making sure I took care of myself while I pursue my dreams. When I needed 48 hours in a day, Ryan, you were there to make sure life kept moving forward. When I was so entranced by my research you made me stop and breathe, ultimately helping me become more efficient. Mr. Char, I appreciate you staying up with me until four in the morning, as a slaved over a hot keyboard, and all the comfort cuddles you and Remus gave me. Everyday, no matter what, Mr. Char and Remus were happy to see me and helped me keep going after a long day in the office.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1. INTRODUCTION TO CHEMINFORMATICS

Cheminformatics is broadly defined as the use of computers and computational techniques to solve chemical problems. More specifically, cheminformatics techniques are typically used to numerically characterize the 1D/2D/3D structures of chemicals in order to calculate their physical, chemical, and biological properties. "Cheminformatics" is a relatively new term, that was established in the late 1990s[1,2]. In fact, the field itself is so new, there is disagreement over the proper spelling, as, cheminformatics, chemoinformatics, chemical informatics, molecular informatics, and even chemiinformatics, can all refer to one thing: the representation and manipulation of chemical structures using computational methods.[1]

Although the term of cheminformatics is relatively new, and modern methods intensively rely on the most powerful computers, the ideas behind the cheminformatic methods have been around for almost 55 years. The work of four main contributors (Profs. Kennard, Wipke, Levinthal, Hansch, and Karplus) has led to the development of databases, such as the Cambridge Structural Database (CSD) and the Protein Data Base (PDB), as well as key methods such as quantitative structure-activity relationships (QSAR) and molecular dynamic simulations (MDS) were developed[2]. The CSD (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/) is a database that contains information regarding the crystal structures of organic molecules. The CSD ultimately led to the creation of the PDB (https://www.rcsb.org/), a database which contains the analysis and x-ray crystal structures of proteins. With these two databases, a researcher can mine, browse, and analyze the structures of proteins and small molecules in the hopes of determining their key characteristics to better understanding their properties and/or modes of action. The development of QSAR, key chemical and biological properties have become predictable using sets of molecules with known properties and machine learning. [1,2] MDS enables the simulation of the time-dependent motions of protein-ligand interactions for a fully-solvated, full-atom system. While it was difficult to see the utility of CSD, PDB, and QSAR during their infancy these databases and tools would one day revolutionize the way we can now browse, access, visualize, and simulate chemical biological data so easily.

The rapid development of cheminformatics is fueled by two main forces that have skyrocketed in parallel for the past twenty years:

(1) *The extreme advancement of the computational power of modern computers*: advances in CPU technology (*e.g.,* 12 core CPU), GPU technology (15,000 cores in

modern GPU workstation), and data storage (from megabytes in the early 90s to hundreds of terabytes and unlimited cloud storage in 2018) allows for accelerated parallel processing for computing the properties of hundreds of millions of virtual compounds or simulating large protein-ligand complexes over longer biological time (e.g., up to several milliseconds for a GPCR).

(2) *The growing compendium of chemical biological data available in publicly-available repository*: a vast amount of freely-accessible data is readily available. At the click of a button, one has access to over 1TB of structural data from proteins to DNA (PDB database) and close to 200 million chemical compounds (PubChem, ChEMBL, and ChemSpider). One is capable of accessing chemical biological information instantly, with over 3 million compounds in ChEMBL and over 237 million bioactivities reported for those compounds. These resources are allowing medicinal chemists and other researchers to make leaps and bounds in the advancement of chemical analysis with the help of cheminformatics.

Cheminformatics has become an essential element in the chemist's toolbox of methods. This is especially for the drug discovery pipeline by facilitating the virtual screening of very large libraries of compounds and prioritizing the ones predicted to have the most desired properties. Those "computational hits" are then advanced to experimental confirmation. Cheminformatics is also used for hit/lead optimization, especially when it comes to (1) design analogues being highly similar to the hit compounds but predicted to have higher potency/selectivity towards the target of interest, and (2) predicting the ADMET (absorption, distribution, metabolism, excretion, and toxicity of drug molecules) properties of those compounds in order to "weed-out" compounds with potentially low bioavailability and toxicity[3].

A key approach in cheminformatics is molecular docking. When performing a simple concept search on PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) for molecular docking, which is the computational binding of ligands to proteins through computer-assisted drug design, over 34,000 results were returned[4]. Over half (~20,000) of these papers shown on PubMed were published in the last five years, with many of the topics centered around inhibition studies.

The primary goal of the research presented herein is to highlight key features of molecular docking techniques that can lead to the production of hit candidates for combatting peanut allergies (**Chapter 2**) and cancers (**Chapter 3**). The research demonstrated that the

implementation of modern cheminformatics techniques yields the ability to develop and utilize chemical libraries for the inhibition of proteins. This thesis will also further discuss computational or experimental procedures that can be performed to enhance this research (potential future work).

# CHAPTER 2*

## THE ANALYSIS OF POLYPHENOLIC COMPOUNDS ON ARA H 2

Bethany Cook and Denis Fourches

Collaboration with Drs. Nathalie Plundrich (NCSU), Mary Ann Lila (NCSU),

and Soheila Maleki (USDA)

Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA.

*Currently under review for the *Journal of Food Chemistry (IF = 4.95)*

## CHAPTER 2. THE ANALYSIS OF POLYPHENOLIC COMPOUNDS ON ARA H 2

### Chapter 2 Summary

The unique structure of Ara h 2 allows for the activation of immunoglobulin E (IgE) antibodies on mast cells, resulting in an allergic reaction through the release of histamine and other inflammatory hormones. Polyphenols have been demonstrated to act as anti-inflammatory agents; to block the production of these antigens decreasing the effects of peanut allergens. Herein we analyze the computational modeling of 42 polyphenols, found in blueberries and cranberries, to the Ara h 2 protein. We utilized molecular docking and molecular dynamic simulations to determine the likelihood that the polyphenols will non-covalently interact with the protein, reducing the side effects of peanut allergens. The top five polyphenols were then recommended for experimental analysis, performed by a collaborator, to verify the computational results.

**Keywords:** Molecular Docking, Virtual Screening, Peanut Allergies, Ara h 2

## 2.1 Introduction

In the United States alone there are over 50 million people that suffer from allergies, with one of the highest causes being food related allergens. According to the Centers for Disease Control (CDC), the number of individuals afflicted with food allergies has increased by at least 50 percent in a 10-year time span. In 2010, there were more than 170 different allergenic foods with the most frequent being milk, eggs, peanut, tree nuts, gluten and seafood[5,6]. Roughly every hour 20 individuals (200,000 per year) have sought emergency medical care due to the effects of food allergens, with the highest affliction rate being in children[5]. Though uncommon, some allergies can be outgrown; unfortunately, the most severe allergens, peanuts, tree nuts and seafood are typically lifelong afflictions[5,7]. The symptoms of food allergies vary based on the severity of the allergy. The most common reaction, anaphylactic shock, is the most serious. Anaphylactic shock occurs in about 40 percent of the children afflicted with food allergies[5–7].

Currently, there are few ways to treat peanut allergies. In fact, the primary method used for treatment is avoiding the source altogether[5,8]. Methods of removal of peanut remnants are even uncertain, the most effective of which involves proper hand washing and the use of household cleaners to completely remove any peanut dust[5,6]. Recent studies show that advisory labels stating that products may contain peanuts are optional, and some labeled products have been found to contain high enough levels to induce allergic reactions[5]. Behind the scenes of this huge peanut pandemic are several protein factors, that can activate peanut allergens, but the most potent allergenic proteins are Ara h 2 and Ara h 6[9].

Ara h 2 and Ara h 6 are 2S albumin storage proteins that contain eight cysteine residues that can affect the properties of structures[10–12]. These cysteine residues allow Ara h 2 to be classified in the prolamin protein superfamily known to contain high amounts of proline and glutamine[13]. 2S albumins are known to be homologous structures, and are one of the major contributors to any anaphylactic side effects, especially in peanuts[13,14]. When compounds bind to a 2S albumin storage protein the structural conformation can be altered, resulting in conformational changes of alpha helices and beta sheets in the bound protein[11,14,15]. These different protein conformations peanut allergies have been known to activate immunoglobulin E (IgE) antibodies at IgE epitopes, located on mast cells (MC). These interactions in turn cause the allergic reaction with the release of histamine and other inflammatory hormones[10,16].

The use of anti-inflammatory agents has been demonstrated to reduce the side effects of peanut allergen exposure. Fruits and vegetables are a natural source of anti-inflammatory agents as they are rich in polyphenols[17]. Polyphenols are a class of compounds containing hydroxyl substituents on aromatic rings and are defined by several subgroups (e.g. phenolic acid, flavonoids, and stilbenes). Flavonoids are the most commonly consumed type of polyphenol, as they are abundant in plants and fruits. Polyphenols are classified as anti-inflammatory agents due to their capacity to reduce allergic reactions by attacking the mast cells, limiting the production of histamine[16,18].

There are two pathways that polyphenols can take to inhibit or alter immune responses. The first pathway occurs during exposure where polyphenols help to create insoluble structures that result in hypoallergenic proteins. Another pathway comes from directly effecting the production of dendritic cells, which are located in soft tissues, such as skin, and can cause the immune response to allergens. The polyphenols alter the efficiency of antigen production in the dendritic cells allowing for decreased inflammation[17].

In this study, structure-based-docking was implemented to determine and analyze possible interactions between the protein, Ara h 2, and polyphenols endogenous to blueberries and cranberries[15,18]. As peanut allergies can develop following ingestion of peanuts or peanut-containing foods, Ara h 2 and the polyphenols need to be studied at a physiological pH of th stomach (pH = 3.26) and neutral pH (pH = 7.00). Advanced molecular modeling techniques were implemented to identify specific target binding sites located on the Ara h 2 protein and molecular dynamic simulations (MDS) were performed to confirm binding stability between ligands and protein. Due to unfamiliar characteristics of Ara h 2, a dimer of the protein was generated and similar molecular docking protocols were implemented to identify increased protein-ligand interactions. The top five predicted binders and one predicted non-binder were recommended for experimental testing by collaborators in the Lila group at North Carolina State University.

## 2.2 Methods and Materials

### 2.2.1 Preparation of the Ara h 2 Protein Structure

The X-ray crystal structure of Ara h 2 bound to a Maltose Binding Protein (MBP) was obtained from the Protein Data Bank (PDB), with a resolution of 2.71 Å (PDB: 3OB4)[19]. The MBP protein was removed for docking and Ara h 2 was curated using the Schrodinger Suite's Protein Preparation Wizard[20,21]. All bond orders were assigned, and explicit hydrogens were added to the original structure. No missing side chains or missing loops were detected by Prime and all water molecules were removed[22–24]. The EPIK program in the Schrodinger Suite was used to determine the protonation states of Ara h 2 at both pH 3.26 and 7.00 while a restrained minimization of the protein was performed with an OPLS3 force field[21,25–28].

### 2.2.2 Binding Sties at Ara h 2

There was no apparent binding pocket located on the Ara h 2 protein[19]. To identify potential binding pockets, two different grid generation methods were conducted: SiteMap and "Blind Docking"[25,27–33]. The SiteMap receptor grid was generated with 15 Å in the X, Y and Z directions centered around the Dscore = 0.954 binding pocket. A "Blind Docking" protocol was then utilized. "Blind Docking" involves generation of a receptor grid without specifying a binding pocket. The "Blind Docking" receptor grid was formed by 25 Å in the X, Y, and Z directions centered around the entire Ara h 2 protein[33].

### 2.2.3 Preparation of the Screening Library

A chemical library of 42 polyphenols was derived from literature and previous research from our collaborator; our chemical library is recapitulated in **Table 2.1**[12,15,18,34,35]. These polyphenols were original tested in the form of blueberry and cranberry juice extracts. The compounds were processed using LigPrep from the Schrodinger Suite with an OPLS3 force field[21,26,36]. Tautomeric states of each compound was generated at pH 3.26 and 7.00 using EPIK, while retaining their specified chiralities[25,37,38].

**Table 2.1:** List of 42 polyphenols considered in this study and their respective properties.

| 42 polyphenol compounds | Molecular Weight (g/mol) | HBA | HBD | AlogP |
|---|---|---|---|---|
| Cyanidin-3-Galactoside | 449.108 | 11 | 8 | 1.11 |
| Cyanidin-3-Glucoside | 449.108 | 11 | 8 | 1.11 |
| Cyanidin-3-Arabinoside | 454.812 | 10 | 7 | 1.06 |
| Peonidin-3-Galactoside | 463.124 | 11 | 7 | 1.33 |
| Peonidin-3-Arabinoside | 468.838 | 10 | 6 | 1.09 |
| Malvidin-3-Glucoside | 493.135 | 12 | 7 | 1.32 |
| Malvidin-3-Galactoside | 493.135 | 12 | 7 | 1.32 |
| Malvidin-3-Arabinoside | 498.864 | 11 | 6 | 0.84 |
| Delphinidin-3-Galactoside | 465.103 | 12 | 9 | 0.87 |
| Delphinidin-3-Glucoside | 493.135 | 12 | 7 | 1.32 |
| Cyanidin | 287.056 | 6 | 5 | 3.04 |
| Malvidin | 331.082 | 7 | 4 | 3.25 |
| Quercetin | 302.043 | 7 | 5 | 1.63 |
| Myicetin | 318.038 | 8 | 6 | 1.39 |
| Quercetin-3-Rutinoside (rutin) | 610.153 | 16 | 10 | -1.16 |
| Quercetin-3-Arabinoside (furanoside) | 434.085 | 11 | 7 | 0.21 |
| Quercetin-3-Arabinoside (pyranoside) | 434.085 | 11 | 7 | 0.21 |
| Quercetin-3-Rhamnoside (quercetrin) | 448.101 | 11 | 7 | 0.59 |
| Quercetin-3-Galactoside (hyperoside) | 464.096 | 12 | 8 | -0.30 |
| Kaempferol | 286.048 | 6 | 4 | 1.87 |
| Quercetin-3-Xyloside | 434.085 | 11 | 7 | 0.21 |
| Quercetin-3-Glucoside (isoquercetin) | 464.096 | 12 | 8 | -0.30 |
| Chlorogenic Acid | 354.095 | 9 | 6 | -0.34 |
| Caffeic Acid | 180.042 | 4 | 3 | 1.44 |
| P-coumaric Acid | 164.047 | 3 | 2 | 1.69 |
| Ferulic Acid | 194.058 | 4 | 2 | 1.67 |
| Benzoic Acid | 122.037 | 2 | 1 | 1.46 |
| P-hydroxybenzoic Acid | 138.032 | 3 | 2 | 1.22 |
| Tannic Acid | 1700.173 | -- | -- | -- |
| Sinapic Acid | 224.069 | 5 | 2 | 1.65 |
| Vanillic Acid | 168.042 | 4 | 2 | 1.20 |
| Gallic Acid | 170.022 | 5 | 4 | 0.73 |
| Resveratrol | 228.079 | 3 | 3 | 3.09 |
| Catechin | 290.079 | 6 | 5 | 2.02 |
| Epicatechin | 290.079 | 6 | 5 | 2.02 |
| Procyanidin B2 | 578.142 | 12 | 10 | 3.57 |
| Procyanidin B1 | 578.142 | 12 | 10 | 3.57 |
| Procyanidin A2 | 576.127 | 12 | 9 | 3.76 |
| Procyanidin C1 | 866.506 | 18 | 15 | 5.11 |
| Epigallocatechin | 306.074 | 7 | 6 | 1.78 |
| Epigallocatechin Gallate | 458.085 | 11 | 8 | 3.10 |
| Epicatechin Gallate | 442.09 | 10 | 7 | 3.34 |

### 2.2.4 Molecular Docking

After protein and ligand curation, the 42 unique compounds were docked using Schrodinger's GLIDE software for both standard precision (SP) and extra precision (XP) scoring functions; while utilizing SiteMap and "Blind Docking" receptor grids[25,27–29]. Overall, this represents 168 docking calculations. All docking results were analyzed by the docking and eModel scores associated with each docking pose[25,27–29,31,32]. To consider the compound "active" in the binding site, the docking score is required to be less than or equal to -7 kcal/mol and the eModel score needs to be less than or equal to -50 kcal/mol. The DS and eM thresholds were previously discovered through virtual screening protocols of micromolar binders. These thresholds can vary based on the protein and the scoring function and are considered guidelines for what is predicted to bind[39–41] .

### 2.2.5 Molecular Dynamics and Dimers

The stability of the molecular docking poses were analyzed using Desmond to perform molecular dynamic simulations (MDS). Each MDS model was built using the molecular docked pose and was explicitly solvated with water molecules to conduct a 50 ns simulation with TIP3P solvation model, an OPLS3 force field energy minimization, and an Orthorhombic volume. The calculation interval ran at 1.0 ps with an integration output of 1.0 fs[42–45].

An Ara h 2 dimer was generated using a protein-protein interaction software, InterEvDock[46]. InterEvDock calculated 150 dimer combinations using three different criteria's: SOAP_PP, FRODOCK, and IES. SOAP_PP is determined by the statistical potentials of protein-protein docking[47]. FRODOCK is Fast Rotational Docking and is determined by different potentials of van der Waals, electrostatics and desolvation, and InterEvScore (IES) is generated from multiple sequence alignment docking scores[48,49].

### 2.2.6 Experimental

With collaboration from Dr. Plundrich, a former member of the Lila Lab at NCSU, the *in-silico* protein-ligand interactions predicted for Ara h 2 and polyphenol complexes were experimentally analyzed. Dr. Plundrich utilized spectroscopy methods to analyze binding potential and immunoblotting to test for hypoallergenic properties in the protein-ligand complexes.

Circular dichroism (CD) spectroscopy was performed for the experimental analysis of proteins on instrument Jasco J-815 spectropolarimeter. This approach used a quartz cuvette with a path length of 1 mm and implemented a buffer exchange into MilliQ water with PD Mini Trap G-25 columns. All polyphenol stock solutions were prepared in MilliQ water (1.5 mM) and were titrated onto the protein solutions to obtain various concentration ratios.

Ultraviolet-visible (UV-Vis) spectroscopy was used to evaluate changes in polyphenol absorption with a Shimadzu UV-2450 spectrometer in a 10 mm reduced volume quartz cuvette. The spectra were recorded from 900nm - 200nm at room temperature.

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting was performed to evaluate protein distribution in protein-ligand complexes and IgE binding capacity[16]. Each sample was mixed with sample loading buffer containing β-mercaptoethanol and immunoblotting was performed in quadruplets.

## 2.3 **Results and Discussion**

### 2.3.1 **Computational Analysis**

The protein, Ara h 2, is found in peanut allergens and can cause a range of inflammatory side effects, including anaphylactic shock. Currently, prevention of peanut allergies is the only method used to avoid an allergic reaction to peanut-containing foods, with no medical treatment available for this condition. However, anti-inflammatory agents can combat peanut related symptoms. Natural forms of anti-inflammatory agents are found in plants in the form of polyphenols. This study computationally examined the protein-ligand interaction of Ara h 2 and 42 naturally occurring polyphenols that are found in blueberries and cranberries.

We began our structure-based-docking analysis with the examination of the Ara h 2 protein. The structure of Ara h 2 located in the Protein Data Bank (PDB) database, PDB ID: 3OB4, contains both the Ara h 2 protein and a maltose binding protein (MBP)[19], **Figure 2.1**. Between these two structures, there are 26 alpha helices and 21 beta sheets. The Ara h 2 protein is naturally found in peanuts and was fused to the MBP to help increase crystallization for experimental studies and stability[19]. The MBP has been linked to reducing the ability of antigen binding, concluding that the MBP might protect some of the IgE receptors in the current crystallization[8,10,15]. The MBP was removed to test the natural peanut allergen, resulting in a 176 amino acid chain that contains five alpha helices and several protein loops (**Figure 2.2**). Once

separated the Ara h 2 protein was fully prepped as stated in the “Methods” section to ensure proper protonation and accurate chemical structures.

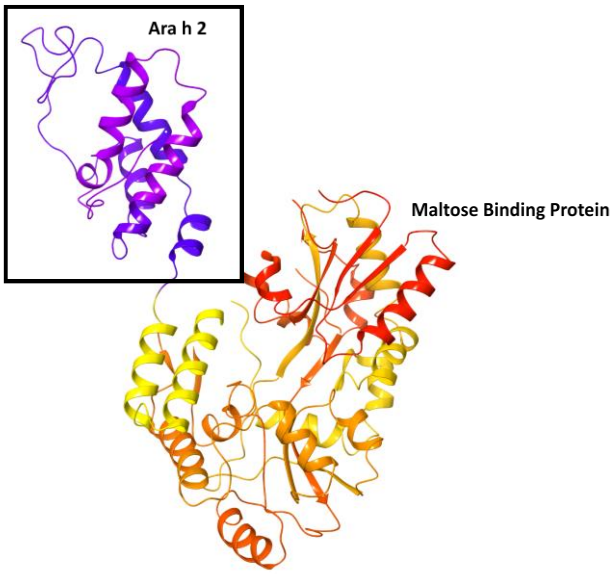



**Figure 2.1:** The protein structure of Ara h 2 bound to the Maltose Binding Protein. Obtained from the PDB database, PDB ID: 3OB4.

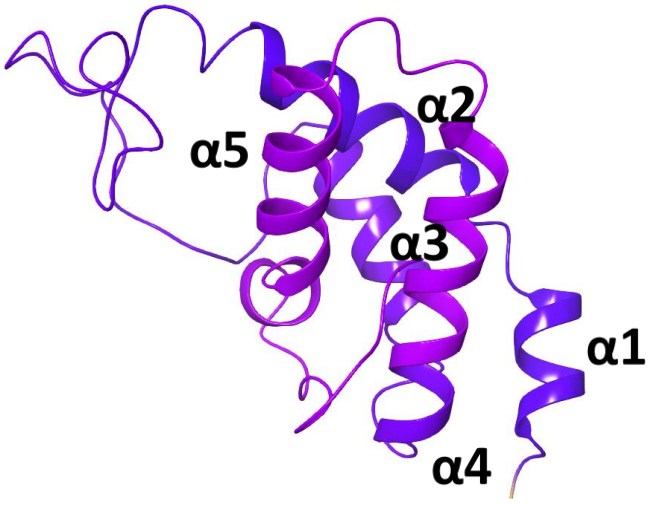



**Figure 2.2:** The refined structure of Ara h 2 after the removal of the Maltose Binding Protein. The α-helices are labeled in consideration to the published literature[19].

The chemical library considered for this study included a series of naturally occurring polyphenols that could potentially bind to Ara h 2 and block one of the IgE receptors to inhibit allergens. For the computational chemical library curation, we analyzed research previously performed by our collaborator on different polyphenols contained in blueberries and cranberries. The polyphenols were originally tested in juices and extracts and showed the potential to reduce allergic responses. These samples contained a combination of 42 different polyphenols ranging in molecular weight from 100 g/mol to over 1,500 g/mol, **Table 2.1**[15,16,18]. Each polyphenol was prepared using Schrodinger's LigPrep software as stated in the "Methods" section. This process is used to provide different tautomeric states of inhibitors to demonstrate all potential protein-ligand interactions.

When obtained the Ara h 2 protein does not noticeably contain a binding pocket (i.e. there is native ligand with crystal structure PDB ID: 3OB4). This observation led to the implication of two different grid generations. First, we utilized the SiteMap (**Figure 2.3**) program which calculates possible binding sites based on several physical characteristics (*e.g.*, exposure to solvent, volume, and hydrophobic and hydrophilic spaces)[30–32]. Using these physical

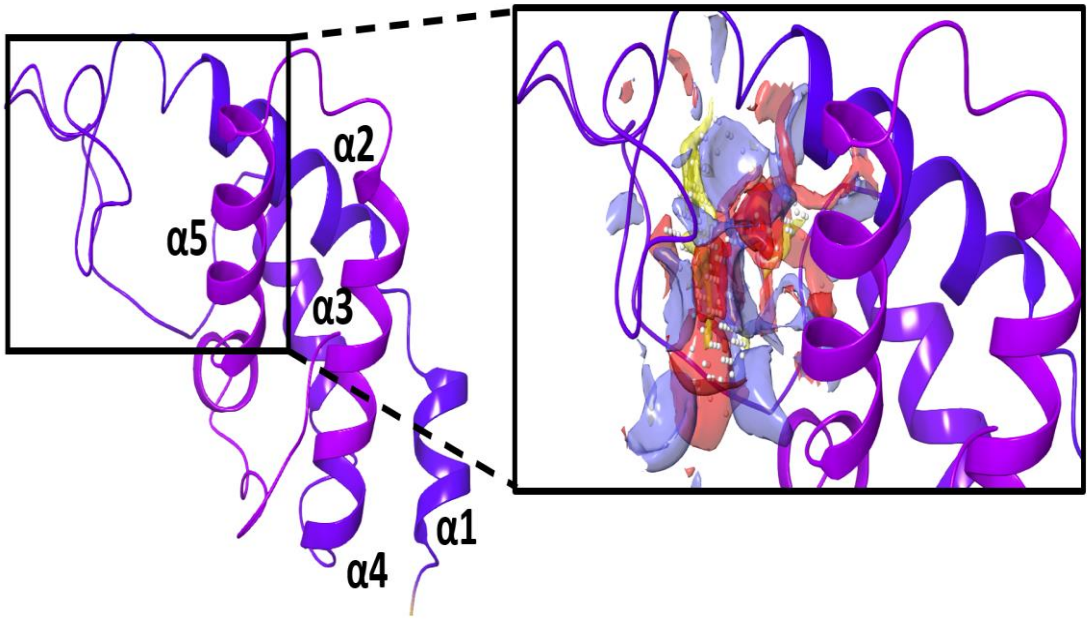



**Figure 2.3:** The refined Ara h 2 protein (left) with the predicted binding pocket determined with the use of SiteMap (right). SiteMap identifies hydrophobic areas (blue), hydrogen bond donor and acceptor positions (red) and metal site functionalities (yellow) with the help of density grid (white beads).

characteristics, a Druggability Score (Dscore) can be calculated. A Dscore greater than 0.83 indicates a ligand is likely to bind in a binding pocket[30,31]. By using the default parameters for SiteMap (5 Å buffer region, a minimum of 15 site points, restrictive hydrophobicity, and a fine grid), several possible binding pockets were discovered for Ara h 2[25,27–32]. The position with the highest Dscore of 0.954 (volume = 381.5 $Å^3$) was used to create the receptor gird for the ligand docking. The SiteMap for Ara h 2 can be seen in **Figure 2.3**, which demonstrates the different reactive sites, displayed as either blue, red, or yellow sections, depending on their physical characteristics. These colored regions represent the hydrophobic areas, hydrogen bond donor and acceptor positions and possible metal site functionalities, respectively.

Secondly, we used "Blind Docking" which is where a receptor grid is generated without specifying a binding pocket. With this type of docking the entire protein is used to calculate the optimal positioning of the ligand and can only be used on small proteins because the max grid generated is 40 Å in the X, Y and Z directions. The "Blind Docking" receptor grid analyzed the entire Ara h 2 protein to determine the optimal docking position, **Figure 2.4**. This receptor grid allowed any of the polyphenols to be docked on the entire protein. When examining the positioning of the compounds, it appears that the optimal docking location determined by both SiteMap and "Blind Docking" was towards the middle of the protein (e.g. between two α-helices).

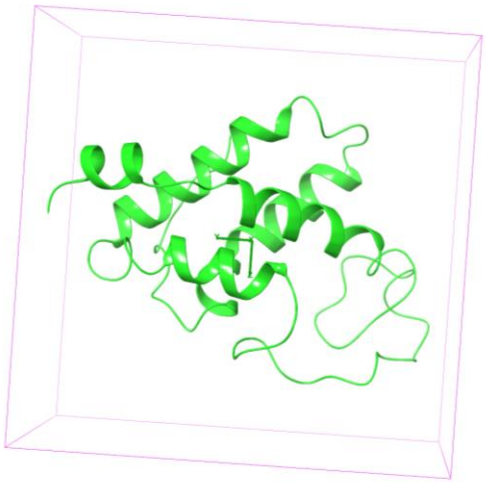


**Figure 2.4:** "Blind Docking" grid generation utilized as 25 Å box in the X, Y, and Z directions indicated by the purple box, located about Ara h 2 (green).

All 42 polyphenols were docked at pH 3.26 and 7.00 for both grids generated by SiteMap and "Blind Docking" with SP and XP scoring functions. "Active" compounds must meet threshold limits of -7 kcal/mol docking scores and -50 kcal/mol glide eModel scores. The docking score (DS) is composed of Glide Scores and represents the ligands binding affinity in the pocket; the eModel score (eM) helps represent the likelihood of the ligand conformation[25,27–29,31,32]. When the results for the Ara h 2 docking were analyzed it was determined that the DS and eM scores for both grid methods were close to identical. Upon analysis of the polyphenols docked, at different pH on the same receptor grid, it was determined that pH does not seem to have a significant impact on the binding affinities, **SI Table 2.1,** resulting in averaged docking scores. The significance in docking scores was determined based on a two-tailed T test and failed to reject the null hypothesis with the mean average of zero. The results of the compounds from both SiteMap and "Blind Docking" were also averaged due to the same binding site and close binding mode results for each method, shown in **SI Table 2.2**.

Docking results for all 42 compounds are given in **SI Table 2.3**. Out of the 42 compounds, 13 compounds received an average docking score lower than the -7 kcal/mol threshold. However, only 12 compounds were considered to be likely binders. Tannic acid has an above threshold eM score (> -50 kcal/mol), disqualifying it from the list of possible compounds. The most ideal docking was of delphinidin-3-glucoside (**Figure 2.5**) which received a docking

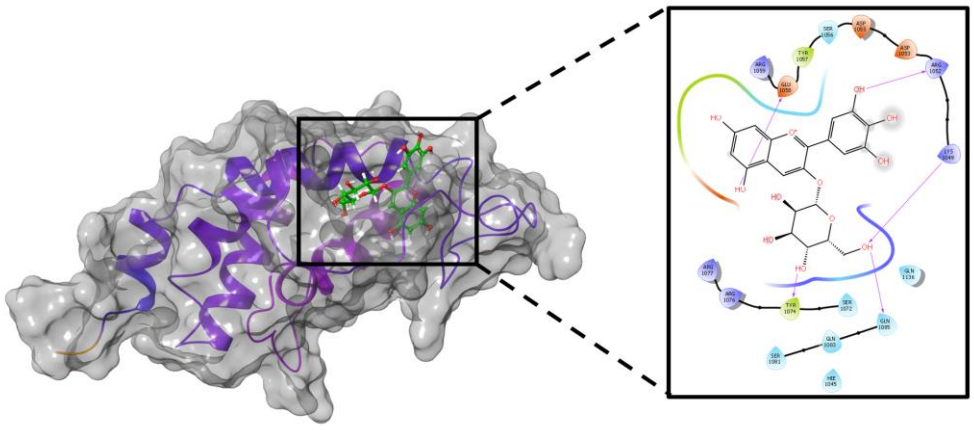


**Figure 2.5:** Through the use of SiteMap the most ideal docking score for the polyphenols was discovered, delphinidin-3-glucoside DS: -8.70 kcal/mol. The ligand is shown to demonstrate five potential hydrogen bonds with Ara h 2 (right).

score of -8.70 kcal/mol when docked with SiteMap at a pH of 7.00. In **Figure 2.5**, the interactions between the protein (Ara h 2) and the ligand (delphinidin-3-glucoside) are demonstrated to have non-covalent interactions by showing the functional groups hydrogen bound to interacting amino acids. The compound that has the lowest DS and eM score overall is delphinidin-3-glucoside, with an average DS of -8.30 kcal/mol and eM score of -68.90 kcal/mol. To narrow down the results and determine the top binders, the compounds with the lowest docking scores were selected to be the top ten compounds for molecular dynamic simulations (MD), **Table 2.2**.

**Table 2.2:** The top 10 compounds that were determined for experimental testing. Four compounds were selected for further testing based on their availability in blueberries and cranberries.

| The Top 10 Polyphenols for Experimental Testing | |
|---|---|
| Compound Name | Docking Score |
| Delphinidin-3-Glucoside | -8.7 kcal/mol |
| Delphinidin-3-Galactoside | -8.0 kcal/mol |
| Cyanidin | -7.8 kcal/mol |
| Quercetin-3-Rhamnoside | -7.7 kcal/mol |
| Cyanidin-3-Glucoside | -7.5 kcal/mol |
| Procyanidin C1 | -7.4 kcal/mol |
| Quercetin-3-Rutinoside (rutin) | -7.4 kcal/mol |
| Quercetin-3-Arabinoside (pyranoside) | -7.3 kcal/mol |
| Chlorogenic Acid | -7.2 kcal/mol |
| Cyanidin-3-Arabinoside | -7.2 kcal/mol |

Through the approach of MD simulations protein-ligand interactions can be confirmed with the addition of solvent to determine any displacement of ligands (**Figure 2.6**). When the MD results were analyzed, all of the top ten compounds appeared to remain in the binding pocket. Further examination was explored with the development of dimer proteins. The structure of Ara h 2 is relatively small (176 amino acids in length) and structural information is not quite known so it is undetermined if the Ara h 2 protein will remain a monomer in solution or will it form dimer complexes through protein-protein interactions. To test this theory Ara h 2 dimers were generated using a protein-protein interaction software, InterEvDock, which results in a larger binding pocket that could cause more variability in ligand DS and the potential increase in docking scores could help identify stronger binders[47] .

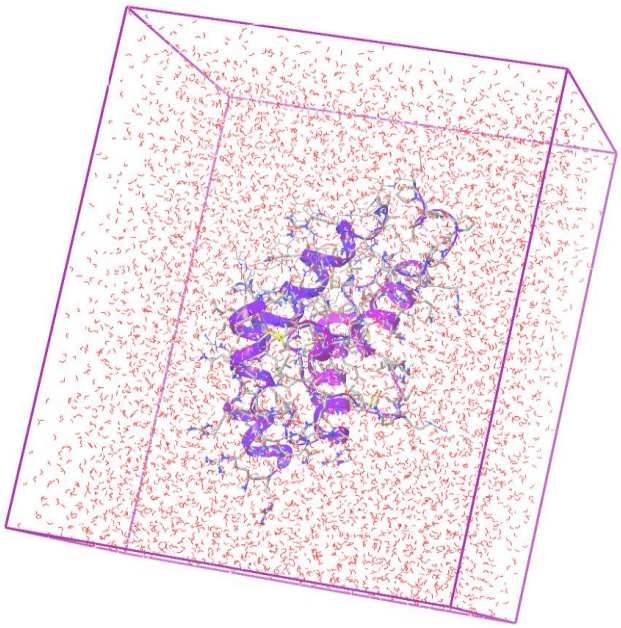


**Figure 2.6:** The molecular dynamic simulations were performed with the original protein and ligand docking pose. A solvent grid was added to account for any ligand displacement, indicated by the purple box and water molecules.

Through the use of InterEvDock software a total of 150 dimers (50 dimers from each protocol mentioned in the "Methods" section) was generated. Once the dimers were obtained each of the top ten compounds, listed in **Table 2.2**, were docked. Ten different dimer positions were obtained from the 150 dimers calculated. These ten dimers were the likeliest conformations to form based off of InterEvDock's scoring system. The top ten compounds from the previous docking and MD simulations were docked to determine if protein-ligand interactions increased. Each of the docking scores for the compounds slightly improved when docked in the binding pocket of the dimer, resulting in no changes in the overall order. A representation of the dimers is shown in **Figure 2.7**, with the two Ara h 2 chains colored green and teal and the respective ligand (PAC C1) colored purple.

The results from the monomer and dimer docking were then analyzed to determine the top five compounds to be tested. These compounds were selected based off different protein-ligand interactions and their frequency of occurrence in blueberries and/or cranberries.

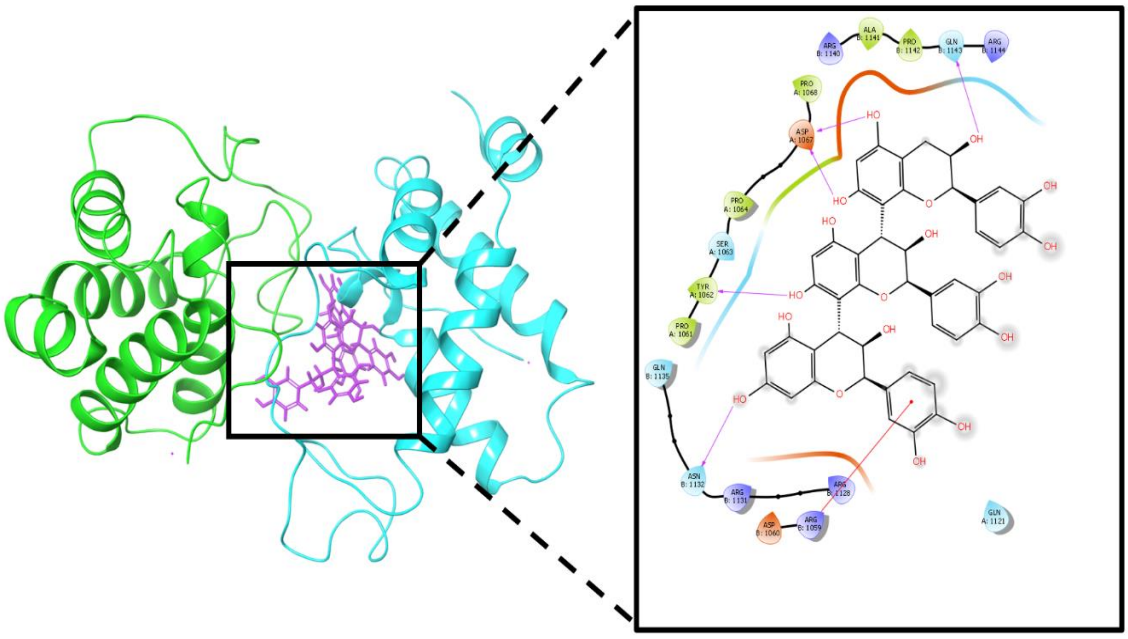


**Figure 2.7:** InterEvDock software developed dimer representations of Ara h 2 (left), as green and teal chains, for the docking of polyphenols. When docked procyanidin C1 (purple) was shown to have increase interactions between both Ara h 2 chains. DS = -8.40 kcal/mol at pH 7.00 on dimer FRODOCK3.

The top five compounds to be tested are delphinidin-3-glucoside, chlorogenic acid, procyanidin C1, cyanidin-3-glucoside, and quercetin-3-arabinoside. To further check the binding results, a lower/non-binder was recommended for experimental analysis. This compound was benzoic acid, which had an average DS of -4.6 kcal/mol and eM score of -20.3 kcal/mol. Benzoic acid's DS and eM are well above the "active" thresholds and is expected to be a low or non-binder.

### 2.3.2 Experimental Analysis

Five out of the six *in-silico* compounds were experimentally evaluated at neutral pH (pH of 7.00). Quercetin-3-arabinoside was not tested. Dr. Plundrich performed all experimental analysis of the Ara h 2 and polyphenol complexes.

CD spectroscopy was utilized to help identify any influence and interaction of polyphenols binding to the secondary structure of Ara h 2. The findings (**Figure 2.8**) demonstrate potential differences in purified Ara h 2 and polyphenol bound Ara h 2 through the

analysis of three artifacts, two minimum and one maximum shifts, located at roughly 208nm, 222nm and 190nm respectively. Each of these spectra "peaks" are identified in α-helical containing proteins, which is confirmed by previous observations[50]. Through analysis of the CD spectra it was determined that in the native, un-bound, Ara h 2 protein there is an α-helical percent of 33 ± 2 %.

Experimental results from titration and CD analysis of the five polyphenol-Ara h 2 complexes provide evidence that procyanidin C1 (PAC C1) and chlorogenic acid produce a secondary structural change. Results indicate that PAC C1-Ara h 2 lead to in an increase of 6% in α-helical concentration, whereas chlorogenic acid produced a 3% decrease. An increase in α-helical content suggests intensified protein skeletons, while decreases in α-helical content indicate potential skeletal weakening. Though the other three polyphenol-protein complexes did not show any changes in CD spectra there is no indication that interactions did not occur; rather, these results simply suggest the interactions are not secondary structural changes. The spectroscopy data for PAC C1, chlorogenic acid and benzoic acid is shown in **Figure 2.8.**

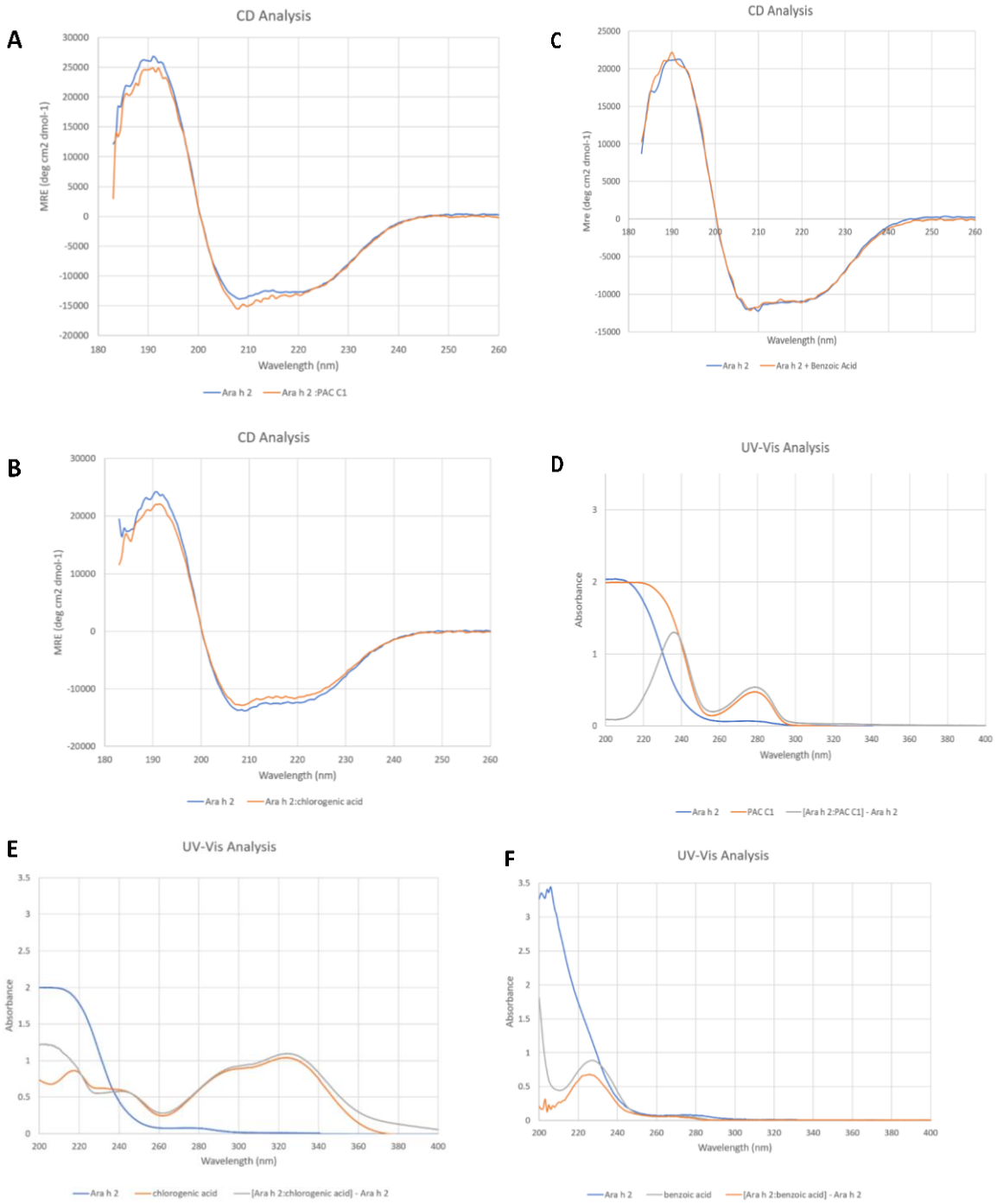



**Figure 2.8:** (A-C) demonstrate the circular dichroism spectra produced by Ara h 2 and the polyphenol complexes. (A) Spectra of Ara h 2-procyanidin C1 interactions; (B) Spectra of Ara h 2-chloro genic acid interactions; (C) spectra of Ara h 2-benzoic acid interactions. (D-F) show the UV-Vis analysis of the titrated polyphenols. (D) Absorbance of procyanidin C1 and Ara h 2; (E) absorbance of chlorogenic acid and Ara h 2; (F) absorbance of benzoic acid and Ara h 2.

A confirmation of secondary structural chances was conducted through the use of UV-Vis spectroscopy. The resulting spectra for each polyphenol-protein complex were analyzed for shifts in the maximum peak (~324nm). Results showed both PAC C1 and chlorogenic acid had small shifts in wavelengths. PAC C1-Ara h 2 produced a small increase in absorption intensity where chlorogenic acid-Ara h 2 had a slight decrease in intensity, when compared to the control Ara h 2 protein. Interestingly, the benzoic acid-Ara h 2 complex produced a maximum shift at approximately at 227nm indicating possible secondary interactions. UV-Vis spectroscopy data can be viewed in **Figure 2.8**.

SDS-PAGE was employed to identify protein interactions and abundance using weight distributions. This process is performed through the protein movement in the presence of an electrical field. On the gel electrophoresis for each polyphenol-Ara h 2 complex, a doublet band was observed at approximately 20 kDa, which is almost identical to the control Ara h 2. This finding may suggest that none of the proteins form tertiary structures through aggregation **(Figure 2.9)**. While the findings of this study elucidate whether Ara h 2 reacts as a monomer or a dimer when binding, they neither confirm nor deny the binding of polyphenols to Ara h 2. To evaluate the binding capacity of each complex, immunoblotting was performed by SDS-PAGE transferal to a polyvinylidene difluoride (PVDF) membrane for the testing of IgE binding epitopes. When compared to the pure Ara h 2 IgE binding, the polyphenol-protein complexes showed potential reduction in IgE epitope interactions. PAC C1, chlorogenic acid and benzoic acid resulted in decreased IgE binding showing that even after the application of reducing agents and heat, through the SDS-PAGE, that protein-ligand interactions are possibly still present.

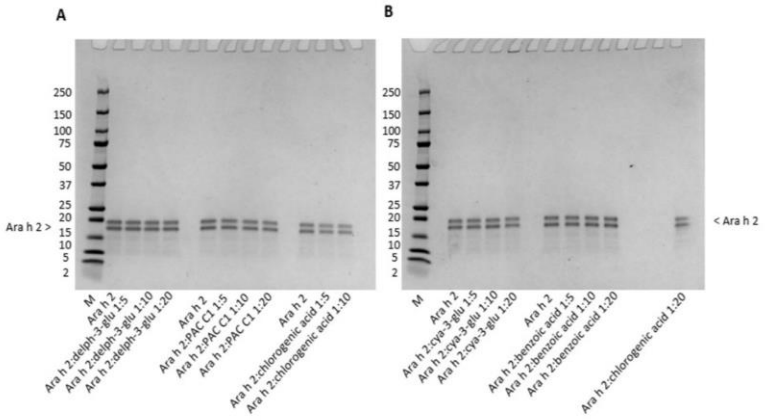



**Figure 2.9:** SDS-PAGE analysis of Ara h 2 with titrated amounts of polyphenols.

A surprising experimental result is the potential binding of benzoic acid to Ara h 2 since *in-silico* calculations predicted it as a low or non-binder. Both the UV-Vis analysis and immunoblotting show some interactions with the titrated benzoic acid sample but the CD spectra does not indicate any secondary α-helical changes. With this in mind an additional molecular docking was performed to help understand all interactions, benzoic acid could possibly have with Ara h 2. Through docking procedure demonstrated that benzoic acid has the capability to bind to 14 different binding sites with various H-bond, Pi-Pi stacking and salt bridge interactions (**Figure 2.10**). There is potential for multiple benzoic acids binding to one Ara h 2 protein, which could explain positive experimental results.

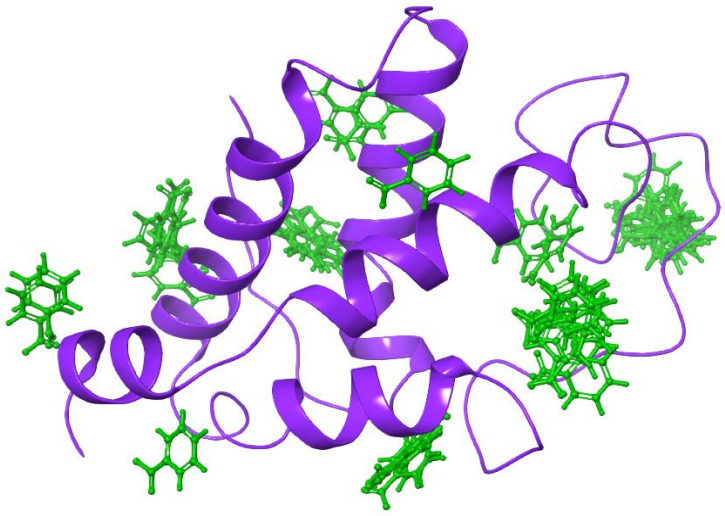


**Figure 2.10:** Further molecular docking of benzoic acid revealed 14 possible binding positions on Ara h 2.

## 2.4 Conclusion

Herein, forty-two naturally occurring polyphenols, found in blueberries and cranberries, were examined with structure-based molecular docking and were experimentally evaluated to determine likely protein-ligand interactions to Ara h 2. Computationally, the binding sites of Ara h 2 were determined with the use of SiteMap and "Blind Docking" protocols. All polyphenols were docked with Glide's SP and XP scoring functions to calculate the likelihood of binding. From this approach, it was determined that both docking grids are comparable and help identify the appropriate binding pocket. Through the additional use of MD simulations and molecular

docking of protein dimers; the top ten polyphenols were identified and recommend for further experimental evaluation.

Collaboration with the Lila group allowed us to experimentally evaluate the interactions of polyphenol-protein interactions. Four likely binders and one predicted non-binder were tested with spectroscopy (circular dichroism and UV-Vis) and immunoblotting (SDS-PAGE and Western blotting) methods to detect Ara h 2 binding potential, induced protein secondary structural changes, and the capability of IgE epitope inhibition. These experimental procedures demonstrated that three of the five polyphenols, procyanidin C1, chlorogenic acid, and benzoic acid have binding potential to Ara h 2 and have the capability of inhibiting peanut allergens through blocking IgE binding epitopes.

# CHAPTER 3*

# DESIGNING CDK2 INHIBITORS USING THE NEW MOLECULAR CHIMERA APPROACH

Bethany Cook and Denis Fourches

Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA.

*Soon to be submitted to *Molecular Informatics*

# CHAPTER 3: DESIGNING CDK2 INHIBITORS USING THE NEW MOLECULAR CHIMERA APPROACH

## Chapter 3 Summary

Cyclin dependent kinase (CDK) have been linked to be a main contributor of cancer, due to their large involvement in the cell cycle. There is a growing compendium of CDK inhibitors available in the public domain, especially for CDK2. However, more potent and selective CDK2 inhibitors are needed. Herein, we analyzed and modeled a set of >300 CDK2 inhibitors, for which crystal structures and experimental CDK2 inhibition potency are accessible. We also utilized the *molecular chimera* approach to generate new CDK2 inhibitors based on structural fusion of known potent ones. The most interesting compounds were selected based on their expected binding affinity predicted using 3D molecular docking and their expected inhibition potency predicted using QSAR models.

**Keywords:** Molecular Docking, Virtual Screening, *Molecular chimera*, CDK2, Cancer, Python

## 3.1 Introduction

Worldwide, one of the leading causes of death is cancer; with fatality rates slowly approaching that of cardiovascular disease[51,52]. Cancer is caused by the overproduction of cells that leads to the development of tumor masses. The dysregulation of cells can lead to devastating effects on the entire body, as the prolific cancer cells have a decreased rate in normal cell death[51,53]. In the United States alone, close to two-million new cases of cancer are projected to be diagnosed in 2018, resulting in over 600,000 deaths by the year's end[54]. A few of the most common types of cancer are breast, lung, prostate, colon, and skin cancer[51,54]. Nearly 40% of the population is predicted to suffer from some form of cancer in their lifetime, with a majority of those diagnosed being male[54]. In 2016, there were an estimated 15.5 million cancer survivors in the U.S., with a predicted 4.8 million increase in survivors in the next 10 years[54].

Over the past 50 years, several forms of cancer treatments have been discovered, including surgery, chemotherapy and radiation therapy. Advancements have been made in each of these treatments; however, this has primarily been the result of trial and error. Currently, the gold standard treatment for cancer is chemotherapy but unfortunately chemotherapy has been linked to apoptosis of healthy cells in addition to cancerous cell[51]. Due to negative side effects targeted inhibition through small molecules was developed[54–58].

A large portion of cell activity is regulated by cyclin dependent kinase (CDK), also referred to as cell division kinase, whose dysregulation has been implicated in the development of several different types of cancer[53,56,59]. CDKs are serine/threonine-based enzymes that are typically proline directed and are inactive in the monomeric form[60–65]. CDKs are dependent on interaction with cyclin components to become active heterodimeric complexes. Over twenty CDK complexes have been discovered, however, fewer than ten have been directly linked to cell cycle involvement[60–65]. CDK/cyclin complexes are responsible for cell growth and division along with the phosphorylation of DNA and with the regulation of dephosphorylated ATP. One of the most important groups to become phosphorylated is the retinoblastoma gene (Rb), which is a tumor suppressor. The Rb protein, when disabled, has been linked to the hyperproliferation of cancer cells[60–65].

Previous studies have demonstrated that CDKs and CDK/cyclin complexes play an important role in the success of cell production throughout the entirety of the cell cycle. CDKs are divided into two categories: those that are involved in the regulation of the cell cycle (CDK1,

CDK2, CDK3, CDK4 and CDK6), and those that are used in transcription[53,57,65–72]. In particular, the deregulation of CDK2 has been linked to very aggressive brain tumors (*e.g.*, glioblastoma), ovarian cancer, colorectal cancer and skin cancers. CDK2 is directly responsible for regulating the cell cycle through its activity in the G1-S transition, as well as centrosome duplication and DNA synthesis[53,65]. This regulation allows for CDK2 to bind to cyclin E, which facilitates phosphorylation of retinoblastoma tumor suppressor proteins. Overall, these factors make CDK2 a prime target for cancer therapy[65,70,73–75].

Small molecule CDK inhibitors (CKIs) have previously been utilized to competitively inhibit CDK protein phosphorylation. CKIs are mainly comprised of purine derivatives, butyrolactones, flavopiridols, staurosporines, and paullones[53,55,63,67,68,71,76–79]. Roscovitine (**Figure 3.1**), a purine derivative that targets CDK 1, 2, 5, and 7, was utilized in clinical trials for patients with solid tumors[77,80–84]. Unfortunately, it showed weak activity, with potency ranging from 0.16 – 100 μM for the different CDK isozymes. Meanwhile, roscovitine showed promise regarding the reduction of tumor mass and stabilization of neurodegenerative diseases. This led to the creation of other purine analogues with different levels of CDK selectivity[77,80–84]. The first CKI to reach clinical trials was a flavopiridol, named alvocidib, which afforded beneficial results when used as a single agent or when combined with chemotherapeutic drugs, depending on the cancer type that was being targeted. Alvocidib (**Figure 3.1**) has a range $IC_{50}$ from 0.04 – 0.4 μM for CDK1, CDK2, CDK4 and CDK7, but also exhibits an $IC_{50}$ of 8 nM for CDK9. Alvocidib is currently in phase 2 of clinical trials for the treatment of acute myeloid leukemia[68,81,83]. Research showed that alvocidib in combination with venetoclax helped improve patient outcomes[85]. Moreover, purvalanol A (**Figure 3.1**), a very potent inhibitor, has demonstrated to be useful only for CDK2 ($IC_{50}$ = 4-70 nM) and CDK5 ($IC_{50}$ = 75 nM)[77,81]. To date, the most potent CDK2 inhibitor is dinaciclib, with an $IC_{50}$ of 1 nM. Dinaciclib (**Figure 3.1**) is currently going through clinical trials to determine other potent inhibitions[81,86,87]. It is clear that these inhibitors have complex polypharmacology at the cost of mild-to-severe drug-induced side effects[68,77,80–83,85–87]. Therefore, we are in need of are new small molecule inhibitors with improved potent and selectivity towards CDK2.

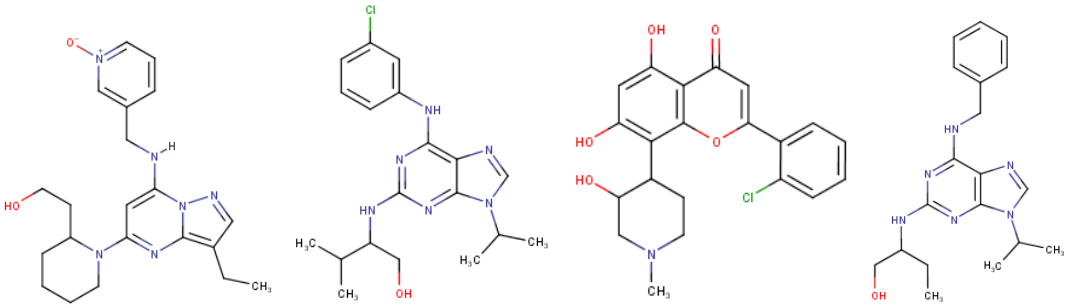


**Figure 3.1:** Chemical structures from left to right: Alvocidib, Roscovitine, Purvalanol A and Dinaciclib.

Previous cheminformatics studies of CDK2 inhibitors include the creation of pharmacophores and various homology models for structure-based virtual screening[88–92]. Molecular docking allowed Tripathi et al. to examine the non-covalent interactions of 27 known inhibitors bound to one CDK2 protein (PDB ID: 1URW)[90,93]. They concluded that there are several key sites of interest for the design of CDK2 inhibitors that includes the hinge region and polar interaction site[90,93]. Canduri et al. have demonstrated through homology studies that the CDK2 proteins contain two main lobes, a small lobe used for ATP binding and a large lobe for peptide binding and catalysis[91]. A substructure known as the molecular fork has been identified in the ATP binding lobe and has been linked to providing three potential hydrogen bonds which are ideal for CKI binding[91].

Herein, we developed the concept of *molecular chimera* to generate new small molecule candidates as potential CDK2 kinase inhibitors. This novel concept is based on the intimate fusion of two different molecular objects. In this proof-of-concept study, a workflow was developed to generate series of *molecular chimera* compounds by structurally fusing known CDK2 inhibitors and generating new analogues with increased complexity. Docking of the newly generated *molecular chimera* compounds in the CDK2 binding site led to the identification of a subset of molecules having similar binding modes to known binders but affording better docking scores. Those potential candidates are prioritized for chemical synthesis and experimental testing. This study is thus an attempt to develop a new method for generating and identifying potential CDK2 inhibitors.

## 3.2 Materials and Methods

### 3.2.1 Dataset Compilation and Preparation

The initial dataset of inhibitors was obtained from the Protein Data Bank database (PDB) and was filtered through an advanced search for two queries 1) structure title containing a kinase and 2) has ligands. This inquiry of data resulted in 6,486 unique kinase with a total of 3,390 different ligands from 1,256 protein structures[94]. To compile all known inhibitors of CDK2 the generic dataset obtained from the PDB was narrowed down to meet the criteria of macromolecule name 1) cyclin dependent kinase 2, 2) CDK2, or 3) cell division protein kinase 2. At the time of this study, the Protein Data Bank database contained 355 known CDK2 crystal structures with 316 unique ligands. To thoroughly investigate the features and similarities of the CDK2 inhibitors, the curated dataset of 316 ligands, was clustered using a hierarchical clustering algorithm constructed of 166 two-dimensional molecular descriptors and computed using the KNIME workflow software[95].

### 3.2.2 Binding Site

With each inhibitor obtained from different crystal structures there is not one binding pocket, but instead, a few hundred. To narrow down and compare each of the inhibitors it was decided that one protein with the smallest resolution, in angstroms (Å), would be used to "self" dock the original ligands. The smallest resolution value is known as a high-resolution structure with highly ordered atoms, for the case of CDK2 the crystal structure PDB ID: 4EK4 (resolution = 1.26 Å) was selected[94,96]. Protein 4EK4 was prepped using Schrödinger Suite's Protein Preparation Wizard, where missing side chains and loops were generated using PRIME, all tautomeric states were generated for pH 7.0 ± 2.0 with EPIK, and the protein's overall energy was minimized with an OPLS3 force field[20–26,37,38,97]. The position of the binding pocket for 4EK4 was identified by the original placement of the native ligand, three letter ID: 1CK. To acquire the appropriate docking position a GLIDE Receptor Grid was developed, centered around the ligand 1CK, extending 15 Å in the X, Y and Z directions[27–29,97,98].

### 3.2.3 Concept of *Molecular Chimera*

*Molecular chimera* was designed based on the structural fusion of two known inhibitors, A and B, to potentially develop a more potent and more selective inhibitor C, **Figure 3.2**. This concept was implemented through a combination of KNIME workflows and Python scripts that utilized the Pandas package[95,99]. The KNIME workflow was designed to identify key scaffolds (e.g. benzene rings) then label and remove each attached R-group. The Python script generates the new inhibitors from the output of the KNIME workflow. Each R-group from two inhibitors are simultaneously incorporated into the Simplified Molecular Input Line Entry System (SMILES) string of the original key scaffold. Due to unpredicted favoritism or priority of the R-groups the Python script took into consideration substitute groups located in the same R position of each inhibitor. In other words, this means that the script generates two almost identical SMILES strings with one substitute group difference among them, the first favoring inhibitor A and the second favoring inhibitor B. Lastly, due to the potential of offset structural fusion full rotation of all R-groups were considered and all ortho, meta and para fusion positions were generated.

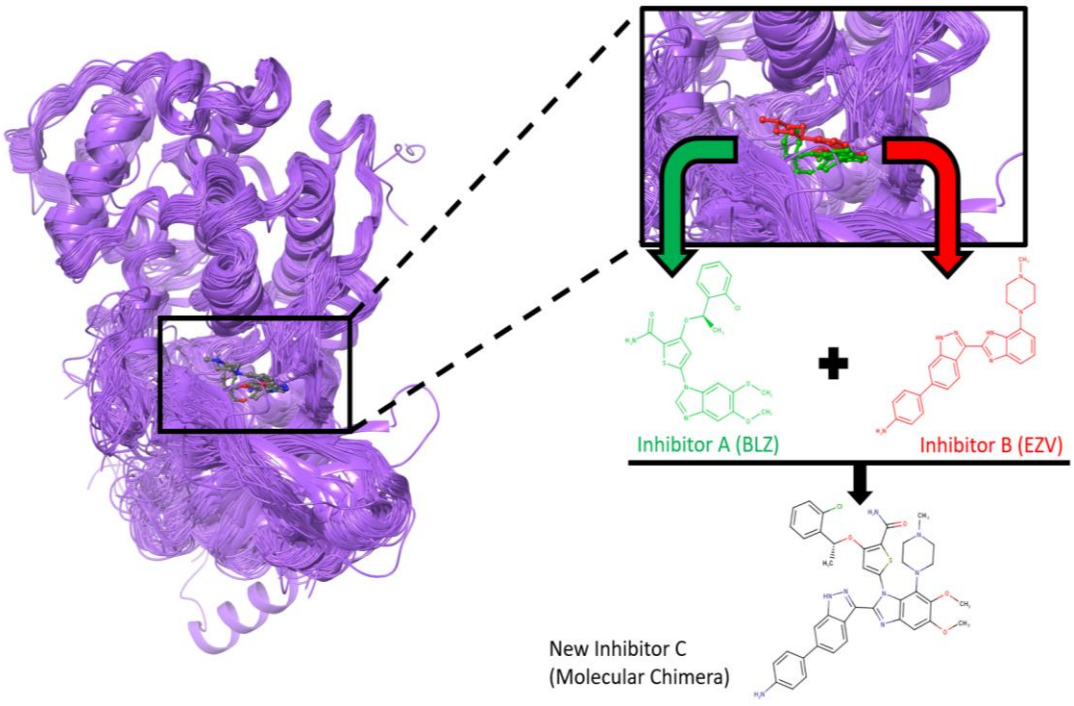



**Figure 3.2:** Concept of *molecular chimera* based on the fusion of two known CDK2 inhibitors (A- BLZ $pIC_{50}$=5.8; B- EZV $pIC_{50}$=6.0) to generate a new molecule C.

### 3.2.4 Preparation of the Screening Library

The chemical library considered for this study was the generated set of chimera C inhibitors built using the KNIME workflow and Python script described above[95,99]. Due to many compounds curated for each inhibitor C a python script was developed which resulted in unique inhibitors. Before preparation of the new compounds, all inhibitors were narrowed down with the use of Lipinski's Rule of 5 to help in druglike prediction[100]. Any compounds that failed more than one of Lipinski's Rule of 5 were eliminated from further processing. All compounds that passed with failing one or less of Lipinski's rules were processed using LigPrep from the Schrödinger Suite with an OPLS3 force field[26,36,97]. Tautomeric states of each compound were generated to a biological relevant pH (pH 7.0 ± 2.0) using EPIK, while retaining their specified chiralities[25,37,38].

### 3.2.5 Molecular Docking

After protein and ligand curation, the generated inhibitors were docked using Schrödinger's GLIDE software with three forms of scoring functions, high throughput virtual screening (HTVS), standard precision (SP), and extra precision (XP)[27–29,98]. The GLIDE software utilized the 4EK4 receptor grid previously generated, centered about the native 1CK ligand. Overall this represents 3,542,523 docking calculations. All docking results were analyzed by the docking and eModel scores associated with each docking pose and scoring function. Compounds were determined to be "active" if the empirical thresholds were met of Docking Score (DS) ≤ -7 kcal/mol and eModel Score (eM) ≤ -50 kcal/mol. Docking scores are composed of multiple Glide Scores and represent the binding affinity of each ligand; the eModel scores are predicted values that represent plausibility of the ligand conformation[39–41]. The DS and eM thresholds were determined through previous research of virtual screening protocols and are only used as guidelines for potential "active" compounds[41].

### 3.2.6 Analysis of *Molecular Chimera* Inhibitors

All molecular XP docking scores were subjected to appropriate filtration in order to resemble properties of inhibitors that matched pharmaceutical industry standards of a minimum oral bioavailability of 20%. The *molecular chimera* compounds were narrowed down by Lipinski's Rules of 5, 10 or fewer non-terminal rotatable bonds, hydrogen bond acceptors and

donors adding up to 12 or fewer, and lastly topological polar surface areas less than or equal to 140 Å[3,100–104]. Increased bioavailability has also been linked to compounds with high saturation levels and high amounts of stereocenters[101]. *Molecular chimera* inhibitors that fit these criteria were highly considered. Lastly, due to a considerable amount of *molecular chimera* inhibitors remaining "active" and ≥ 20% oral bioavailability a machine learning technique was developed to identify potential inhibition potency ($pIC_{50}$).

### 3.2.7 Multiple Protein Inhibition Conformation

Overall, the top compounds produced from the *molecular chimera* approach, underwent further examination with the docking of compounds to their respective native proteins (e.g. ligand 1QKX75 native ligands are 1QK and X75) [94]. The proteins were selected based off the two native ligands that each molecular compound is comprised of. Each of the crystal structures were prepared with Schrödinger Suite's Protein Preparation Wizard, where missing side chains and loops were generated using PRIME, all tautomeric states were generated for pH 7.0 ± 2.0 with EPIK, and the protein's overall energy was minimized with an OPLS3 force field[20–26,37,38,97]. The binding pocket for each of the proteins were identified by the placement of the native ligand. A Glide Receptor Gird was generated centered about the native ligand and extend 15 Å in the X, Y, and Z directions[27–29,98].

The top selected *molecular chimera* compounds underwent LigPrep and EPIK with a biological pH of 7.0 ± 2.0, as before, and were docked with HTVS, SP, and XP scoring functions[25,27–29,36–38,98].

## 3.3 Results and Discussion

### 3.3.1 Data Curation

This study was conducted using 316 experimental CDK2 ligands available from the PDB database[94]. These compounds were first investigated using a hierarchical clustering algorithm to develop the circular dendrogram as described above in the "Methods" section. The circular dendrogram (**Figure 3.3**) revealed interesting clusters of compounds with three master leaves. The molecular weights of these molecules range from 46 (formic acid) to 511 ({(2-Bromo-4-methylphenyl)[6-({4-[(2S)-3-(dimethylamino)-2-hydroxypropoxy]phenyl}amino)-4-pyrimidinyl]amino}acetonitrile) g/mol with 305 of the 316 molecules containing at least one ring structure[94]. One particularly interesting aspect of these inhibitors is the lack of known inhibition

potencies (either provide by the PDB or available in ChEMBL)[94,105]. Only about two-thirds (213 of 316) have reported potencies including many with two or fewer reported experimental $IC_{50}$ values. Many of the 316 compounds break zero of Lipinski's Rule of Five however, 16 ligands do violate a minimum of one rule.

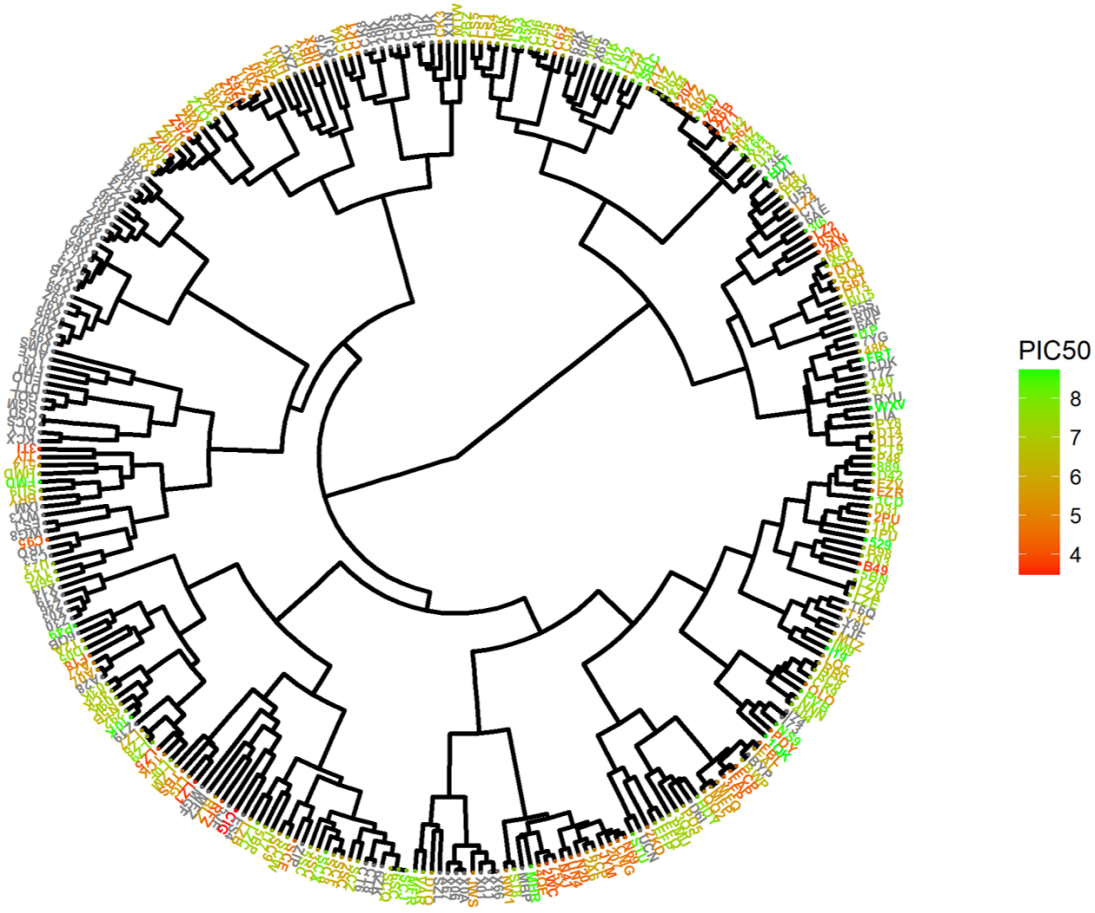



**Figure 3.3:** Circular dendrogram representing the clustering of all 316 CDK2 binders with their associated $pIC_{50}$ inhibition potency values. Unknown values are given in gray.

Of the 213 native ligands there are 132 compounds with high potencies (pIC50 ≥ 6) and 81 compounds with low potencies (pIC50 < 6). The distribution of the native $pIC_{50}$ values, **Figure 3.4**, seems to have a large variance. Between the clustering of the ligands and their respective $pIC_{50}$ there is yet to be a determination of what aspect of CDK2 inhibitors contribute to compound potencies.

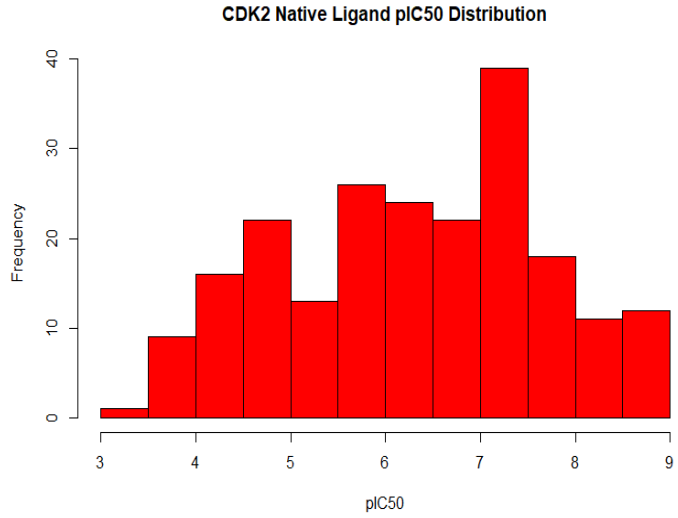



**Figure 3.4:** The distribution of known $pIC_{50}$ values retrieved from ChEMBL and PDB databases.

In this computational study, docking scores were calculated on one crystal structure PDB ID: 4EK4, as described above, to determine any similarities between known inhibitors[96]. The 316 experimental ligands were "self" docked using the HTVS, SP, and XP scoring functions, the XP

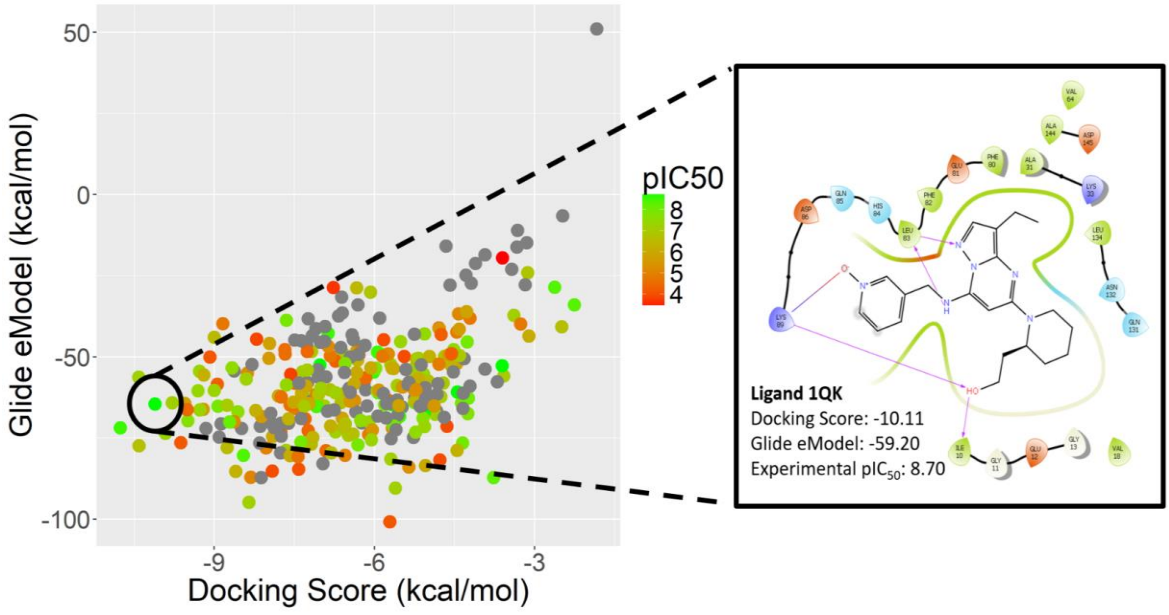



**Figure 3.5:** "Self" docking results of 316 native CDK2 ligands on protein 4EK4 (left). Ligand 1QK, Dinaciclib, is predicted to be the most potent CDK2 inhibitor and received one of the highest docking scores. (Note: 1QK is not the native ligand of protein 4EK4).

docking results can be seen in **Figure 3.5**[27–29,97,98]. Ultimately, there is no direct correlation between inhibition potencies and binding affinities so for proper comparison of inhibitors only the binding affinities were considered. Of the 316 original CDK2 inhibitors, only 95 compounds afforded docking scores below -7 kcal/mol and eModel scores below -50 kcal/mol. The potencies of the "active" compounds included 46 compounds with $pIC_{50} \geq 6.0$[39–41]. By analyzing the binding affinities, those that do obtain a high potency ($pIC_{50} \geq 6.0$) do have "active" docking scores. The most potent native ligand, Dinaciclib with a $pIC_{50}$ of 8.70, was calculated to have close to the lowest docking score, -10.11 kcal/mol (**Figure 3.5**). As stated there is no direct correlation between binding affinities and inhibition potencies however, these results do indicate that the docking protocol is able to separate many of the experimentally-confirmed actives from inactives.

Through further analysis of the clustered ligands and their respective docking scores, there are observed potential activity cliffs. Activity cliffs are defined by compounds that have similar structural properties but have a change in potency[106,107]. The actual change in potency that determines an activity cliff is undecided but is typically considered as a 100-fold difference in $IC_{50}$, 1 nM to 100 nM ($pIC_{50}$ 9 down to 7).[106,107] Many of the native CDK2 inhibitors have a wide variety of potencies ($pIC_{50}$), **figure 3.4**. By analyzing structure activity relationship (SAR) models and the docking scores of the inhibitors there is an observed correlation when considering the three leaves of the circle dendrogram[106,107]. One leaf that was considered contains ligands 1QK (Dinaciclib), I73, I74 NS9, and PDY, **figure 3.6**. Each of these ligands are clustered based on their structural properties with potencies ($pIC_{50}$) of 8.70, N/A, N/A, 8.52 and 4.64 respectively, and docking scores -10.1, -7.9, -8.5, -10.8 and -7.7 kcal/mol respectively. Each of these structures have a common scaffold of Pyrazolo[1,5-A] pyrimidine and varying size substituent groups located at the second and fourth positions. When analyzing these five ligands two have unknown potencies, so a potential activity cliff comparison could be linked to their docking scores. The difference between ligands 1QK and I73 are minor, however the 2 kcal/mol difference in docking scores can be contributed to the extra hydrogen bond generated by the hydroxyl group at the end of an ethane chain instead of directly attached to a ring structure. Ligands I73 and I74 are almost completely identical except for the R or S stereocenter of one substituent group. This small change results in the addition of two hydrogen bonds contributing to almost 0.5 kcal/mol difference. Following traditional activity cliff guidelines and analyzing

the drop in potency there is more than a 100 fold difference resulting in $pIC_{50}$ 8.70 and 4.64 for ligands 1QK and PDY respectively. By examining the binding modes of these ligands in the 4EK4 binding site there is evidence that the length in substituent groups can result in significant reduction in potency, producing an additional hydrogen bond and salt bridge interaction for ligand 1QK. The activity of ligands 1QK and PDY can also be compared with a 2 kcal/mol difference in docking scores. Based on observations from ligands 1QK and NS9 it seems that substituent groups that contain ring structures are limited in the amount of hydrogens bonds that can be formed, resulting in a 0.6 kcal/mol docking score variation and the formation of 4 additional hydrogen bonds on ligand NS9. The decrease in potency could be contributed to the different native proteins. Slight structural variations found in each of these ligands could explain the significant activity cliffs among the clustered ligands as well as the sections of unknown potencies.

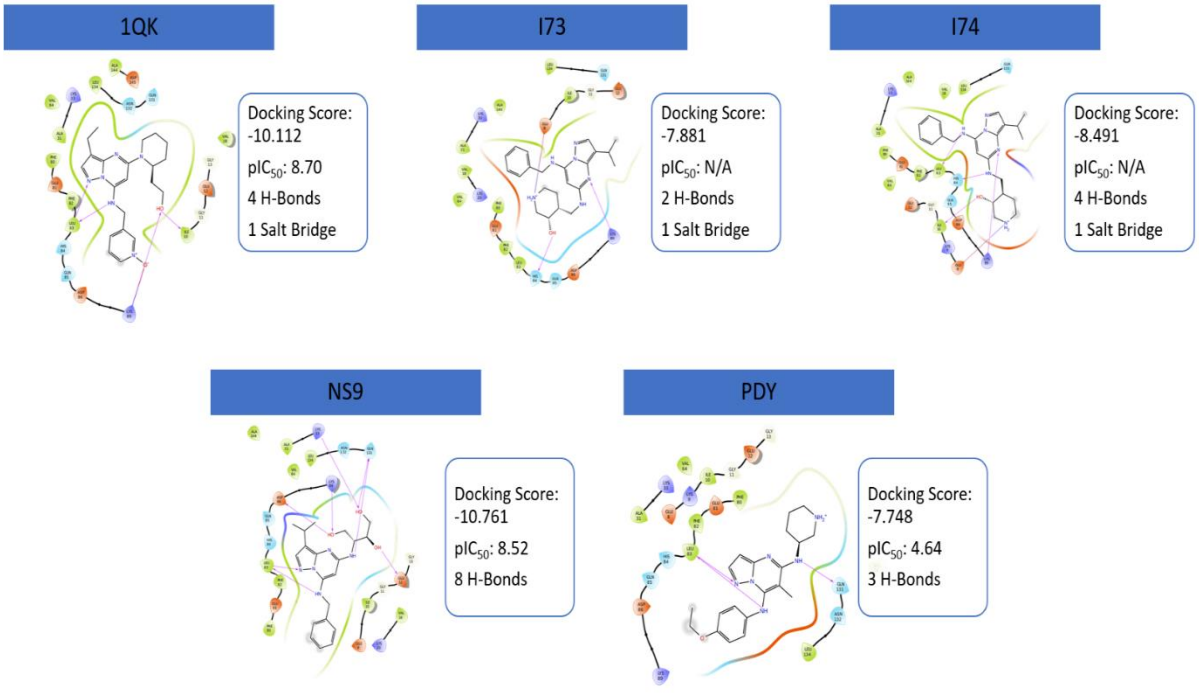



**Figure 3.6:** Demonstrates the binding affinities and potencies for five clustered CDK2 ligands.

### 3.3.2 *Molecular Chimera*

After the initial investigation, the common scaffolds of the 316 ligands were identified with the KNIME workflow, **Figure 3.7**[95]. There are several key elements to the KNIME workflow, the first is identifying the common key scaffolds. Each of the original ligands were imported as structure data files (SDF) and fully prepped to ensure that all hydrogens were added, and all structures had correct bond representation (**Figure 3.7, Box 1**)[108]. The second section, focuses on separating the key ring scaffolds from surrounding substituent groups (**Figure 3.7, Box 2**)[108]. To reconstruct the new inhibitors, it is important to know where each original substituent is positioned to identify which R-groups will compete for binding interactions. To insure proper alignment each substituent was identified and labeled as [R#*], with each pound sign (#) varying based on original location. To reinsert the substituent group onto the identified key ring structure [R#*] needed to be removed from the substituent sequence, this was performed with cell splitters in the third section of the workflow (**Figure 3.7, Box 3**).

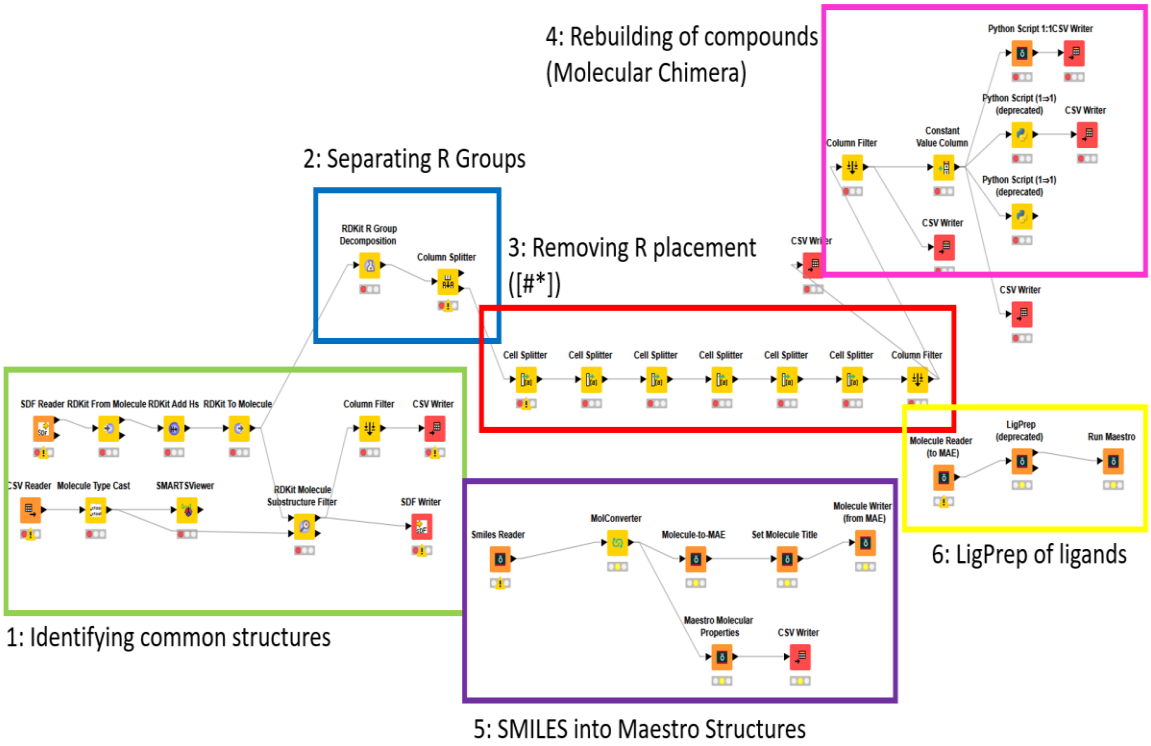



**Figure 3.7:** KNIME workflow for the creation of molecular chimera for CDK2 inhibitors.

The fourth section of the workflow is the most important step, the merging of compounds through *molecular chimera* (**Figure 3.7, Box 4**). In order to build new inhibitors a python script was written to take substituent groups from compound A and merge them with substituents from compound B around the common ring scaffold[99]. However, when the compounds merge there is no certainty to which R substituent group will bind. This means that [R1*] from compound A can compete against [R1*] from compound B for binding on the ring scaffold. Thus, every possible combination of R substituent must be accounted for. Since this code is scaffold dependent a minimum of 162 and a maximum of 508,000 C inhibitors were generated for scaffolds ranging in size from cyclopropane to naphthalene respectively.

The KNIME workflow resulted in the identification of 63 ring scaffolds that were used to separate the inhibitors and utilized as a base for the SMILES string in the Python script. Once all 63 scaffolds were processed through the Python script the curated dataset resulted in 4,259,339 new unique inhibitors, a breakdown of the dataset is shown in **SI Table 3.1**. All scaffolds that housed only one original inhibitor were processed with all R-groups from the same size scaffold (e.g. the scaffold c1nncn1 utilized all five membered rings to develop new compounds) to generate *molecular chimera* inhibitors. To identify the components that helped generate the new inhibitors each compound was given a unique name structure composed of the three letter ID for inhibitor A, inhibitor B, a number, and lastly the SMILES of the common scaffold (e.g. MTW1RO.147c1nncn1). Each inhibitor went through the "RDKit add hydrogen node" in the KNIME workflow for full and proper protonation[108]. Due to the large amount of ligands and the potential for thousands to have poor druggability, the *molecular chimera* inhibitors were narrowed down based off Lipinski's rules as stated in the "Methods" section[100]. The resulting dataset contained 1,555,708 compounds before preparation with Schrödinger's LigPrep and EPIK software. Proper biologically relevant pH (pH 7.0 ± 2.0) and tautomeric states from LigPrep and EPIK software delivered 2,782,115 initial poses for molecular docking[25,36–38,97]. Again, molecular docking was performed on a single protein crystal structure, PDB ID: 4EK4, to allow for inhibitor comparison between the 316 original inhibitors and the potential *molecular chimera* inhibitors[96]. The 2,782,115 curated ligands that reflect the biologically relevant pH and tautomeric states were all docked with HTVS scoring function. Any ligands that received an "active" DS ≤ -7 kcal/mol and eM ≤ -50 kcal/mol further preceded to the next scoring function (SP and XP respectively)[27–29,39–41,98].

### 3.3.3 Analysis of "Predicted-to-be-Active" Inhibitors

There are 31,089 potential CDK2 inhibitors developed by the *molecular chimera* approach and predicted to be "active" by molecular docking thresholds (DS ≤ -7 kcal/mol and eM ≤ -50 kcal/mol) for all XP scoring function compounds[39–41]. Due to all tautomeric states of the *molecular chimera* compounds being docked there was a potential for duplicated poses, all duplicates were removed to allow for the highest XP docking score to be represented for each pose. Out of the newly curated dataset 34 compounds have docking scores ranging -13 to -12 kcal/mol, 95 compounds ranging -12 to -11 kcal/mol, 590 compounds ranging -10 to -11 kcal/mol, 3730 compounds ranging -9 to -10 kcal/mol, 10551 compounds ranging -8 to -9 kcal/mol and 16089 compounds ranging -7 to -8 kcal/mol. A visualization of all docking score ratios can be seen in **Figure 3.8**, this figure represents a comparison of XP docking scores for the 316 original inhibitors and the *molecular chimera* inhibitors.

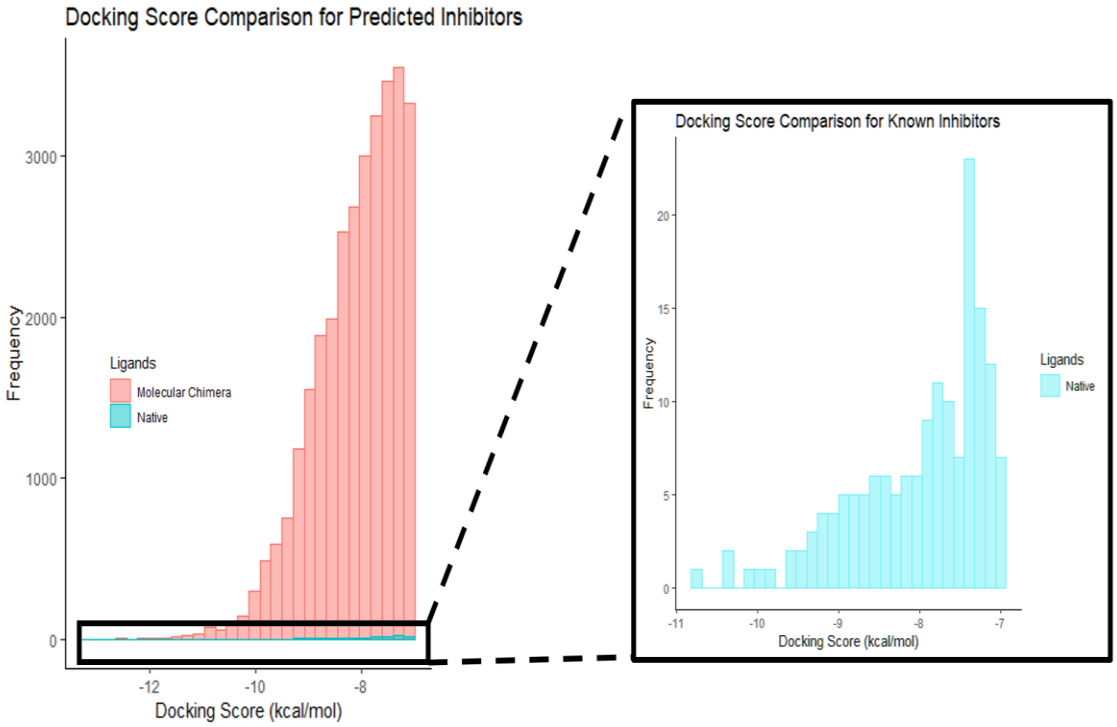



**Figure 3.8:** Plot of "active" ligands, based on docking scores, for all inhibitors, Native (blue) and Molecular Chimera (pink). The threshold for active docking scores is ≤ -7 kcal/mol.

### 3.3.4 Bioavailability

An inhibitor being "active" does not confirm druggability. The molecular docking simulation only calculates how well a compound will bind to the pocket it does not take into consideration the ligands ability to reach the binding site. One key factor in drug discovery is high oral bioavailability for therapeutic agents[102–104]. To narrow down the C inhibitors generated by the *molecular chimera* approach previous research was analyzed for key features of experimental "drug" candidates.

Previous research performed by *Nicholas Meanwell* focuses on key aspects to improve compound characteristics and safety through the collection of recent medicinal studies[101]. Lead optimization has been linked to failing one or fewer of Lipinski's Rules of 5, having fewer than 13 non-terminal rotatable bonds (NRB), high saturation ($sp^3$ hybridization) levels, and topological polar surface areas (TPSA) ≤ 140 Å. Several other characteristics of oral bioavailability, in compounds, correlates to a decrease of solubility such as, an increase of aromatic rings, high molecular weights (MW), high H-bond counts (≤ 12 H-bond acceptors (HBA) and donors (HBD)), and low amounts of stereocenters. The lipophilicity and oral absorption of compounds were also compared based off of Caco-2 permeability data and showed unfavorable results with compounds that contained high TPSA, HBD, HBA, RB and MW, confirming previous thoughts[101].

When the lipophilicity and oral absorption were considered for the 316 native CDK2 ligands it revealed 260 of the compounds were favorable according to the Caco-2 permeability data[101]. These results were then compared to the *molecular chimera* compounds that were deemed favorable and are shown in **SI Figure 3.1**.

Other researchers have also come to the same conclusions as *Meanwell* that though Lipinski's rules are a good guideline for druggability most compounds only pass 3 out of 4 criteria so, polar surface area, rotatable bonds, or H-bond donors and acceptors should be considered for increased oral bioavailability. When directly comparing MW and NRB there seems to be no set correlation of compounds with MW < 400 and the number of NRB, though it has been shown that NRB ≤ 10 is favored among all MWs. Though each property is only a guideline through the analysis of previous drug studies and the oral bioavailability of compounds on rats, each does follow the pharmaceutical industry standard of a minimum oral bioavailability of 20%[3,100,102–104].

By following the criteria of meeting 3 of the 4 Lipinski's rules and having TPSA ≤ 140 Å, H-bond count ≤ 12, NRB ≤ 10 the curated dataset was narrowed down from 31,089 potential CDK2 inhibitors to 17,512 inhibitors. When high saturation was considered, a cut off threshold for all compounds of ≥ 0.10 was set, this meant that all compounds had a minimum of 0.10 and a maximum of 0.25 (based off largest data value) for $sp^3$ hybridization. The saturation level of the compounds had a meaningful impact resulting in a dataset of 3,986 compounds that potentially have an oral bioavailability of 20% or greater. Docking and eModel scores were not considered for the analysis of *molecular chimera* compounds that met the pharmaceutical industry standard of oral bioavailability.

### 3.3.5 *Molecular chimera* Predicted Potency

Though a large portion of *molecular chimera* compounds were eliminated from further studies based on "active" docking and eModel scores and oral bioavailability about 4,000 compounds is still a considerable amount to synthesize. Based on this realization a machine learning technique, Random Forest Regression, was applied to the *molecular chimera* compounds to predict their inhibition potency ($pIC_{50}$) as stated in the "Methods" section[109]. This machine learning technique utilized a circular ECFP6 fingerprint and a random forest learner and predictor. A ten-fold cross-validation and ensemble of 1,000 regression tress was used to predicted $pIC_{50}$ values[109].

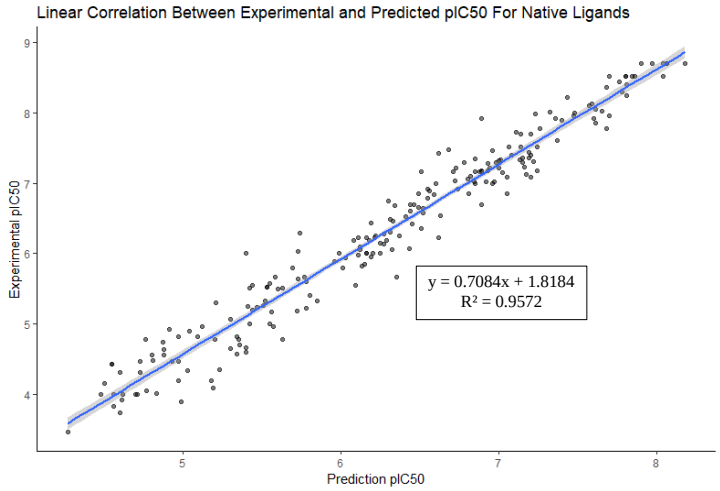



**Figure 3.9:** Linear correlation of machine learning predicted $pIC_{50}$ values versus the known experimental values retrieved from ChEMBL and the PDB databases. The linear equation is y=0.7084x + 1.81184 with and $R^2 = 0.96$.

For this technique the 213 compounds with known potencies of the original 316 inhibitors were used as a test set to develop the learned pattern of the dataset for future predictions. The predicted regression resulted in a 0.98 linear correlation between the $pIC_{50}$ and predicted $pIC_{50}$ with an average of $pIC_{50}$ ± 0.36 SD, **Figure 3.9**. With this random forest machine learning technique there is however a $pIC_{50}$ out-of-bag error of 0.45, most likely an increased amount due to inhibition of native proteins. As a small confirmation for this approach the training set was tested on the remaining 103 ligands with unknown potencies, to determine structural behavior and potency relationship. The compounds were then graphed with the hierarchical clustering algorithm as mentioned in the "Methods" section with the coloration of the predicted $pIC_{50}$, seen in **Figure 3.10**.Comparable scatter plots were graphed to represent the DS and eM scores for the experimental $pIC_{50}$ and the predicted $pIC_{50}$, **Figure 3.11**. It appears that most predicted $pIC_{50}$ values are concentrated in the moderate to low range ($pIC_{50} \leq 6$) but does follow more of a consistency when clustered, compared to the clustering with unknown potencies (**Figure 3.2**). A second test set was generated for the *molecular chimera* compounds. By taking into consideration the predicted $pIC_{50}$ the curated *molecular chimera* dataset was narrowed down to 3,522 inhibitors with potencies of $pIC_{50}$ 6.0 – 8.0 and 354 inhibitors of $pIC_{50}$ 7.0 – 8.0.

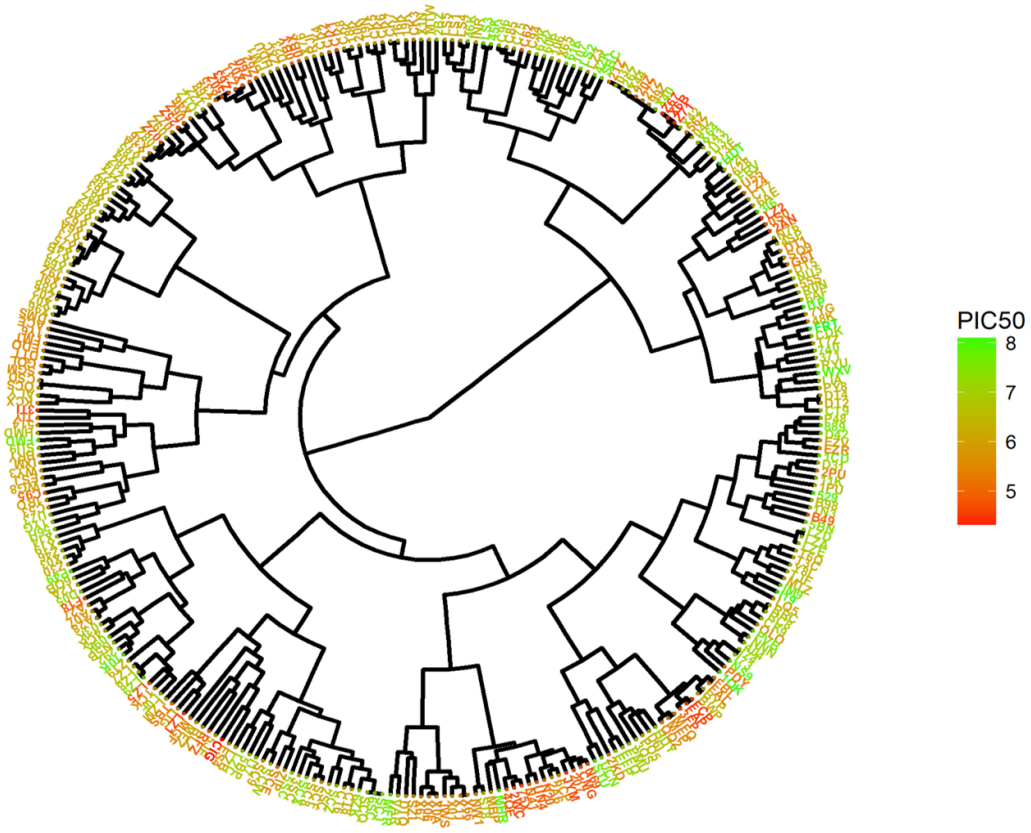



**Figure 3.10:** Circular dendrogram of all 316 CDK2 binders with their associated predicted $pIC_{50}$ inhibition potencies.

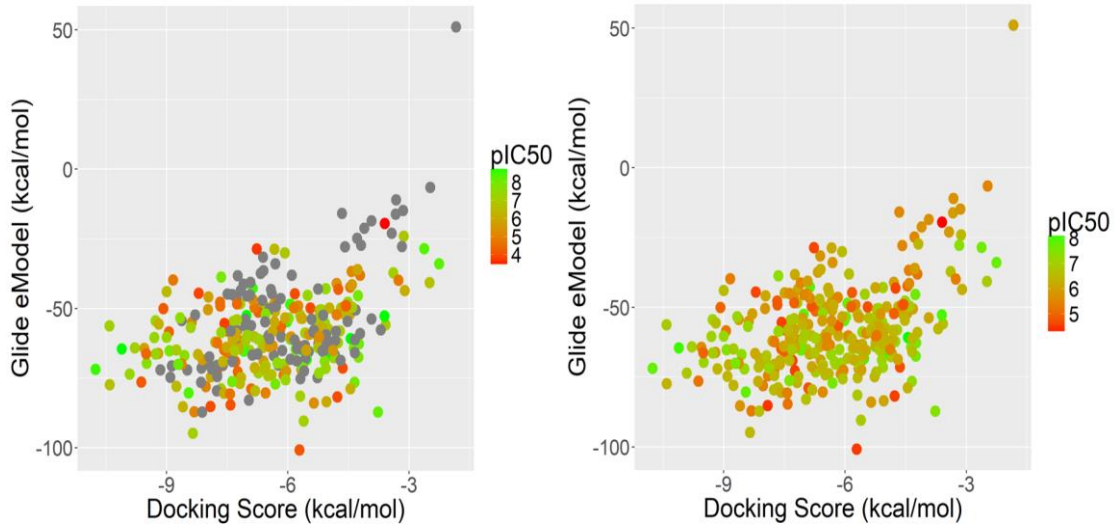



**Figure 3.11:** Distribution of 316 ligands according to their docking and glide eModel scores colored by known $pIC_{50}$ (left) and predicted $pIC_{50}$ (right).

### 3.3.6 Further Self Docking

The top 50 *molecular chimera* inhibitors were selected based off $pIC_{50} \geq 7.0$ and were narrowed downed by compounds with the lowest docking scores. The curated dataset obtained docking scores of -12 to -9 kcal/mol and were comprised of native ligands 1QK, 0S0, 5SC, NS9, PDY, X75, SCZ, SCQ, I73, I74, 2SC, RJI, 07Z, LZB, LIA, and SCX. To further examine the binding potential of the new *molecular chimera* compounds each were docked on the native ligand's protein structure. Each CDK2 crystal structure and respective ligand are represented in **Table 3.1**[94].

**Table 3.1:** The Native ligands that were most frequent in the top 50 compounds generated by *molecular chimera* and their respective protein PDB IDs.

| Native Ligand | PDB ID |
|---|---|
| 1QK | 4KD1, 5L2W |
| 0S0 | 4EZ3 |
| 5SC | 2R3Q |
| NS9 | 3NS9 |
| PDY | 3WBL |
| X75 | 3R1Q |
| SCZ | 2R3N |
| SCQ | 2R3K |
| I73 | 5JQ8 |
| I74 | 5JQ5 |
| 2SC | 2R3O |
| RJI | 5ANK |
| 07Z | 3RK5 |
| LZB | 2VTR |
| LIA | 2FVD |
| SCX | 2R3M |

Before docking all ligands were prepared using LigPrep as stated in the "Methods" section[25,36–38,97]. The prepped dataset returned 79 compounds for further docking. Each of the respective crystal structures were prepped and Glide Girds were generated about the native ligand[27–29,97,98]. Once docked with HTVS, SP and XP scoring functions, all duplicate poses were removed revealing 472 out of 850 docked ligand poses achieving docking scores ≤ -7 kcal/mol and eModel scores ≤ -50 kcal/mol[27–29,39–41,97,98]. Each of these protein-ligand interactions incorporate various amounts of hydrogen bonds but all compounds remain under the limitations

set by Lipinski's Rule of 5 and those to meet the minimum oral bioavailability standard (≥ 20%)[3,100–104].

### 3.3.7 Top Compounds

For the top five inhibitors (**Figure 3.12**) interestingly the nine-membered indole type ring structure is favorable. This scaffold seems to play an important role due to resulting in at least one hydrogen bond between nitrogen and the leucine-83 amino acid of protein 4EK4. Though when docked these top five compounds did increase from their native ligand docking scores (native DS: -10.11 to -7.75 kcal/mol, *molecular chimera* DS: -12.61 to -11.84 kcal/mol). The binding mode of both chemical libraries (native and *molecular chimera*) seems to be comparable due to similar non-covalent interactions in the region of leucine-83. However, when the binding affinities of these top five inhibitors were analyzed on multiple CDK2 proteins (**Table 3.1**) it was discovered each received a DS less than -7 kcal/mol indicating uncertain protein selectivity.

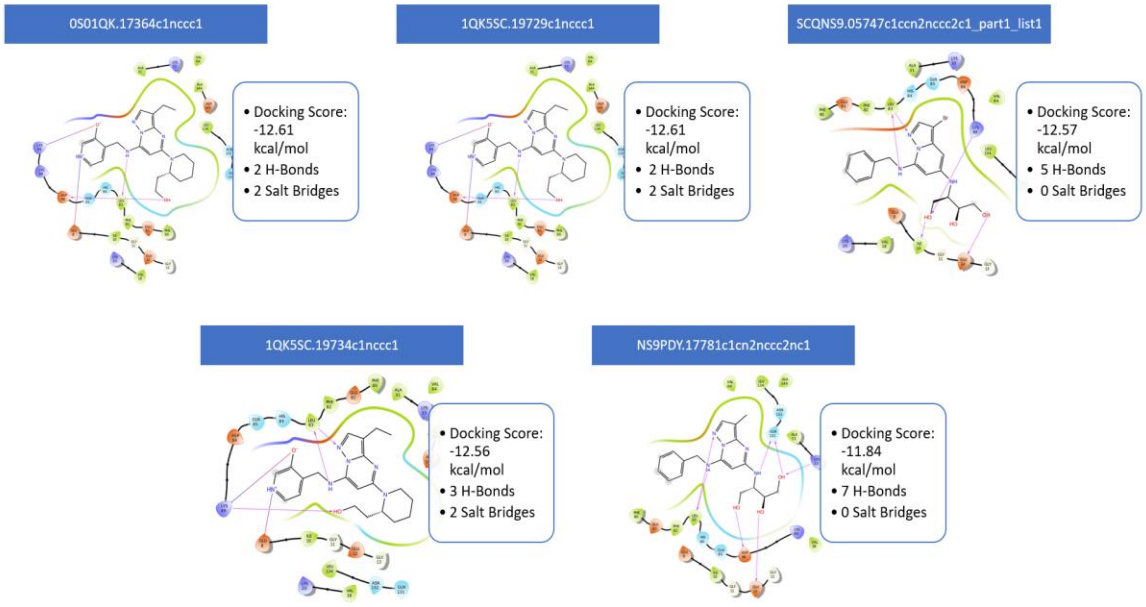



**Figure 3.12:** The top five compounds determined by the docking scores on protein 4EK4.

The chemical scaffolds, DS and predicted potencies of the top 50 compounds are provided in **SI Figure 3.2** and **SI Table 3.2**. Through the analysis of these inhibitors it was

observed that many "active" and bioavailable *molecular chimera* compounds favor an indole type scaffold and CDK2 protein inhibition. Overall, it seems with the ability to structurally fuse two known inhibitors together there is a potential to obtain inhibitors with higher potency and undermined selectivity.

### 3.4 Conclusions

Herein, we developed and applied the *molecular chimera* approach to automatically generate novel CDK2 inhibitors based on previously evaluated inhibitors. From 95 original "active" inhibitors, our technique generated over 40,000 new compounds. After filtering and molecular docking based on three scoring functions and 18 different crystal structures, 50 newly generated *molecular chimera* compounds were identified as potential CDK2 inhibitors. Interestingly, several of the *molecular chimera* inhibitors contain at least one R-group from the compound with the highest potency for CDK2, dinaciclib $pIC_{50} = 8.7$. Using the 50 potentially active *molecular chimera* compounds, we plan to collaborate with synthetic chemists for the development and testing of screening assays for CDK2 inhibition. This experimental analysis could confirm our model's predictive capabilities. Once experimental binding results for the *molecular chimera* compounds have been analyzed and compared to computational results, our structural fusion protocol will be improved as needed. These initial findings of *molecular chimera* revealed that there are potential benefits to combining two known inhibitors through the use of scaffold identification and structural fusion. *Molecular chimera* can provide new insights and guidance to the develop of future protein inhibitors to advance precision medicine.

# CHAPTER 4: FUTURE DIRECTIONS

## 4.1 Future Experiments for Ara h 2

For peanut allergens the protein, Ara h 2, has been shown to be a main contributor of severer allergic responses and polyphenols have computationally been shown to form protein-ligand interaction but still further analysis can be performed. In the previous study performed by our collaborator, Dr. Lila's Lab, polyphenols were believed to be the cause behind the reduction in allergic responses, when extracts from blueberries and cranberries were analyzed, but what if it was some other chemical reaction entirely? To begin to further understand the reactions taking place we could analyze all the chemical compounds produced by the blueberry and cranberry extracts to make sure polyphenols are indeed the contributing factor. In order to determine what compounds are interacting with Ara h 2 a similar docking protocol, to the one discussed in this document, would be performed with the compounds found in the extracts. Depending on the results of this analysis, future computational and experimental procedures can be determined

Though polyphenols have been previously shown to act as natural forms of anti-inflammatories these inhibitors may not be ideal, which could lead us to further analyze a large database of known inhibitors. The ZINC library is a collection of commercially available chemical compounds. By docking a set library of other known anti-inflammations from the ZINC database we could identify the best target compounds to combat peanut allergens. This analysis might give us a better understanding of the role polyphenols play in our protein-ligand interactions or could lead us in an entirely different direction.

With each of the additional computational studies mentioned above there is one similarity, both examine the interactions of compounds on Ara h 2, but there is an additional protein of interest concerning peanut allergens, Ara h 6. By performing similar docking studies, that have been mentioned in this document, on protein Ara h 6 key components of peanut protein inhibition can be realized. When docked, if the 42 polyphenols have stronger interactions or large variations in binding modes this could identify a target protein and compounds to reduce peanut allergens.

Depending on collaboration, further experimental studies can be conducted to determine if polyphenols are actively inhibiting Ara h 2 IgE epitopes. Since peanuts are a consumable product the lowest pH of the body (stomach acid pH = 3.26) needs to be taken into consideration

to ensure once inhibited the protein, Ara h 2, remains inactive. The reactivation of IgE binding epitopes later in the digestion system can still have repercussions due to contact with soft tissues such as the intestines.

To provide evidentiary support of protein-ligand interactions, Two-dimensional Gel Electrophoresis (2-DE) can be implemented. 2-DE is similar to the SDS-PAGE, which was already performed in this experiment, but allows for the further analysis of proteomes by not only applying a separation of molecular weight up also different isoelectric points[110]. This technique will highlight protein expression under foreign stimulation resulting in the "plotting" of protein changes due to chemical, physical or biological conditions. In previous research the study of proteomes with 2-DE determined antigens that would be ideal for Ara h 2 research[110,111].

Further analysis can also be conducted on the experimental procedure, SDS-PAGE, that was already performed. By coupling the protein separation performed by SDS-PAGE with in-gel digestion (IGD) and lastly analysis by liquid chromatography-tandem mass spectrometry (LC-MS/MS) the evaluation of each individual protein can be performed[112,113]. This process is also referred to as GeLC-MS/MS and is useful in the discovery of biomarkers, protein expression profiling, and abundant protein depletion. From the SDS-PAGE each separate band of protein is removed and analyzed for peptide sequence against a sequence database for total identification and potential quantification of each sample[112,113].

Lastly, another technique that could be performed is x-ray crystallography. This technique will allow for a visual conformation if the polyphenols are indeed binding to the Ara h 2 protein. X-ray crystallography utilizes the use of diffraction patterns to determine the molecular structure of a crystal. From previous research performed by Mueller et al. the Ara h 2 protein is shown to crystalize with the help of a maltose binding protein (MBP)[19]. If the MBP was attached to the potential polyphenol bound Ara h 2 complex a direct analysis of protein-ligand interactions could be performed. By adding the MBP; however, there could be competitive binding at some of the IgE epitopes. Any competitive binding could result in a reduction of polyphenol inhibition but might allow for conformation of interactions at still available epitopes.

Both computational and experimental techniques could provide interesting results for the further analysis of polyphenol-Ara h 2/6 complexes. With these techniques we could be one step closer to the identification of potential inhibitors and a solution to the peanut allergen pandemic.

### 4.2 Further Analysis of the CDK Kinase Family

Through the development of the *molecular chimera* approach we were able to produce potentially more potent and selective inhibitors through increased binding affinities but we are still uncertain how many proteins will have interactions with the top selected compounds of protein 4EK4. This means there is no evidence for *molecular chimera* compounds to selectively inhibit a single CDK2 protein let alone target just one CDK kinase family.

Recent research has shown that once CDK2 proteins are inhibited, CDK1 proteins have the capacity to evolve and replace CDK2 kinase in the cell cycle. This replacement can result in the restart of cells but also the possible reproduction of cancer[114,115]. This transition has only been briefly studied but does give rise to the following questions: can one inhibitor be able to effectively inhibit both CDK1 and CDK2 proteins simultaneously? CDK1 alone? If so, what are the structural characteristics that enable that selectivity? Overall, the literature is scarce when it comes to the differential, structure-focused analysis of CDK1 vs CDK2 interactions formed with known CDK inhibitors.

We believe that, through the further exploration of the CDK kinase family, we could identify structure-based characteristics for stronger and selective CDK-inhibitor interactions. By studying the non-covalent, dynamic CDK1/2-ligand interactions, we will be able to better understand the key features in driving the binding selectivity towards these kinases. We will establish a foundation for small molecule compounds that could be resilient to drug resistant mutations and new approaches to the production of chemical probes (the *molecular chimera* approach).

After compiling and curating the entries for both CDK1 and CDK2 proteins, from the PDB, we will compute and characterize all CDK-inhibitor interactions occurring for each inhibitor towards both CDK1 and CDK2. This is a unique task that has never been done in the literature and could thus be very impactful. A special focus will be given for those inhibitors for which we already have their native binding conformations and their respective PDB crystal structures. All binding affinities (pKi) and potencies ($IC_{50}$) will be extracted from ChEMBL so the relationships between computed docking scores and experimental inhibitor potencies will be studied. This analysis will highlight subsets of active ligands with various CDK interaction patterns that are target specific and thus have the potential of being less susceptible to evolved resistance. Once identified the top-10 compounds will be analyzed with MD simulations (all

atoms, 20 ns, 300K, NPT, Desmond) to analyze the dynamic interactions (e.g., frequency of occurrence of a particular H-bond) at the origin of CDK1 / CDK2 selectivity.

After the full analysis of CDK1 and CDK2 proteins, we will further explore the docking results of our newly generated *molecular chimera* compounds on both CDK1 and CDK2 proteins. We plan to generate 1M new compounds using this technique with knowns inhibitor from CDK1 and dock all compounds in the binding sites of CDK1 and CDK2 proteins with the best resolutions (lowest Å). To compare inhibitors, we will use our quantitative structure activity relationship (QSAR) models that were built using machine learning through KNIME workflows to relate binding affinities to known experimental inhibition potencies ($IC_{50}$). By combining the binding affinities from docking with the predicted potencies from QSAR, we will be able to predict experimental outcomes of those new *molecular chimera* compounds, including both their potency and selectivity. To confirm any predicted protein-ligand interactions a MD simulation will be ran to determine if any ligands will move from the binding site due to solvent involvement and will show any active inhibition. The top hits from the MD simulations will be recommended for experimental testing.

At the end of our protein analysis we plan to collaborate with an experimental synthetic chemist for the development of the top 10% of our new *molecular chimera* compounds, that effectively inhibit both CDK1 and CDK2 proteins simultaneously.

# CHAPTER 5: CONCLUDING REMARKS

Cheminformatics and molecular modeling methods have developed into essential tools for the extraction and analysis of chemical information. Through the use of in silico research bench-top chemistry can become revolutionized with the elimination of unfavorable drug candidates. We have shown in this study that the use of computational studies, especially molecular docking, can help to overcome pandemics such as peanut allergens and cancer.

Our research presented in **Chapter 2** has started the development for anti-inflammatories targeting peanut allergens. Computationally and experimentally exploring the relationship between polyphenols and protein Ara h 2 we have determined that there is a potential to reduce the anaphylactic reactions to peanuts. This *proof-of-concept* study demonstrated the molecular docking of 42 polyphenols resulting in 10 potentially "active" compounds. With the use of spectroscopy methods two of the four computationally "active" compounds demonstrated spectral shifts indicating the blocking of IgE epitopes. We further learned, based off the size of polyphenols there is a possibility for multiple ligands to bind to Ara h 2 increasing the chances of inhibition of IgE epitopes. Overall, there is a potential for polyphenols found in blueberries and cranberries to reduce the allergic reactions of peanuts but further evaluation needs to be performed.

Molecular docking has also led to the curation of the *molecular chimera* approach (**Chapter 3**). Through the exploration of CDK2 ligands we were able to structurally fuse two known inhibitors together for the opportunity to develop more potent and selective inhibitors. Once generated all *molecular chimera* compounds were docked and we saw an over 10-fold increase in the number of potential "active" CDK2 inhibitors. However, when the top 50 compounds were docked on multiple CDK2 proteins it was determined that several ligands still obtain "active" docking scores. In order to determine the selectivity of the new *molecular chimera* inhibitors further analysis will be explored.

**References:**


(1) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; 2007.

(2) *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH, 2005.

(3) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51* (4), 817–834.

(4) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *NIH Publich Access* **2011**, *7* (2), 146–157.

(5) *Food Allergy Facts and Statistics for the U.S.*; 2011.

(6) Peanut Allergy https://acaai.org/allergies/types/food-allergies/types-food-allergy/peanut-allergy.

(7) Deak, P. E.; Vrabel, M. R.; Kiziltepe, T.; Bilgicer, B. Determination of Crucial Immunogenic Epitopes in Major Peanut Allergy Protein, Ara H2, via Novel Nanoallergen Platform. *Sci. Rep.* **2017**, *7* (1), 1–13.

(8) Stanley, J. S.; King, N.; Burks, A. W.; Huang, S. K.; Sampson, H.; Cockrell, G.; Helm, R. M.; West, C. M.; Bannon, G. A. Identification and Mutational Analysis of the Immunodominant IgE Binding Epitopes of the Major Peanut AllergenAra h 2. *Arch. Biochem. Biophys.* **1997**, *342* (2), 244–253.

(9) Chen, X.; Negi, S. S.; Liao, S.; Gao, V.; Braun, W.; Dreskin, S. C. Conformational IgE Epitopes of Peanut Allergens Ara h 2 and Ara h 6. *Clin. Exp. Allergy* **2016**, *46* (8), 1120–1128.

(10) Wu, Z.; Lian, J.; Zhao, R.; Li, K.; Li, X.; Yang, A.; Tong, P.; Chen, H. Ara h 2 Cross-Linking Catalyzed by MTGase Decreases Its Allergenicity. *Food Funct.* **2017**, *8* (3), 1195–1203.

(11) Barre, A.; Borges, J. P.; Culerrier, R.; Rougé, P. Homology Modelling of the Major Peanut Allergen Ara h 2 and Surface Mapping of IgE-Binding Epitopes. *Immunol. Lett.* **2005**, *100* (2), 153–158.

(12) Grace, M. H.; Ribnicky, D. M.; Kuhn, P.; Poulev, A.; Yousef, G. G.; Raskin, I.; Ann, M. Hypoglycemic Activity of a Novel Anthocyanin-Rich Formulation from Lowbush Blueberry, Vaccinium Angustifolium Aiton. *NIH Public Access* **2009**, *16* (5), 406–415.

(13) Breiteneder, H.; Radauer, C. A Classification of Plant Food Allergens. *J. Allergy Clin. Immunol.* **2004**, *113* (5), 821–830.

(14) Vesic, J.; Stambolic, I.; Apostolovic, D.; Milcic, M.; Stanic-Vucinic, D.; Cirkovic Velickovic, T. Complexes of Green Tea Polyphenol, Epigalocatechin-3-Gallate, and 2S Albumins of Peanut. *Food Chem.* **2015**, *185*, 309–317.

(15) Plundrich, N. J.; White, B. L.; Dean, L. L.; Davis, J. P.; Foegeding, E. A.; Lila, M. A. Stability and Immunogenicity of Hypoallergenic Peanut Protein–polyphenol Complexes during in Vitro Pepsin Digestion. *Food Funct.* **2015**, *6* (7), 2145–2154.

(16) Plundrich, N. J.; White, B. L.; Dean, L. L.; Davis, J. P.; Foegeding, E. A.; Lila, M. A. Protein-Bound Vaccinium Fruit Polyphenols Decrease IgE Binding to Peanut Allergens and RBL-2H3 Mast Cell Degranulation in Vitro. *Food Funct.* **2017**, *8* (4), 1611–1621.

(17) Singh, A.; Holvoet, S.; Mercenier, A. Dietary Polyphenols in the Prevention and Treatment of Allergic Diseases. *Clin. Exp. Allergy* **2011**, *41* (10), 1346–1359.

(18) Plundrich, N. J.; Kulis, M.; White, B. L.; Grace, M. H.; Guo, R.; Burks, A. W.; Davis, J. P.; Lila, M. A. Novel Strategy to Create Hypoallergenic Peanut Protein-Polyphenol Edible Matrices for Oral Immunotherapy. *J. Agric. Food Chem.* **2014**, *62* (29), 7010–7021.

(19) Mueller, G. A.; Gosavi, R. A.; Pomés, A.; Wünschmann, S.; Moon, A. F.; London, R. E.; Pedersen, L. C. Ara h 2: Crystal Structure and IgE Binding Distinguish Two Subpopulations of Peanut Allergic Patients by Epitope Diversity. *Allergy Eur. J. Allergy Clin. Immunol.* **2011**, *66* (7), 878–885.

(20) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27* (3), 221–234.

(21) Schrödinger Release 2018-2: Schrödinger Suite 2018-2 Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2016; Impact, Schrödinger, LLC, New York, NY, 2016; Prime, Schrödinger, LLC, New York, NY, 2018.

(22) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 351–367.

(23) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. *J. Mol. Biol.* **2002**, *320* (3), 597–608.

(24) Schrödinger Release 2018-2: Prime, Schrödinger, LLC, New York, NY, 2018.

(25) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: A Software Program for PKa Prediction and Protonation State Generation for Drug-like Molecules. *J. Comput. Aided. Mol. Des.* **2007**, *21* (12), 681–691.

(26) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12* (1), 281–296.

(27) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(28) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessmnet of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(29) Schrödinger Release 2018-2: Glide, Schrödinger, LLC, New York, NY, 2018.

(30) Halgren, T. A. Identifying and Characterizing Binding Sites and Assessing Druggability. *J. Chem. Inf. Model.* **2009**, *49* (2), 377–389.

(31) Halgren, T. New Method for Fast and Accurate Binding-Site Identification and Analysis. *Chem. Biol. Drug Des.* **2007**, *69* (2), 146–148.

(32) Schrödinger Release 2018-3: SiteMap, Schrödinger, LLC, New York, NY, 2018.

(33) Can I dock ligands to the whole protein, without specifying a particular binding site? https://www.schrodinger.com/kb/599.

(34) Cunningham, D. G.; Vannozzi, S.; Shea, E. O. Analysis and Standardization of Cranberry Products. *Management* **2001**, 151–166.

(35) Grace, M. H.; Guzman, I.; Roopchand, D. E.; Moskal, K.; Cheng, D. M.; Pogrebnyak, N.; Raskin, I.; Howell, A.; Lila, M. A. Stable Binding of Alternative Protein-Enriched Food Matrices with Concentrated Cranberry Bioflavonoids for Functional Food Applications. *J. Agric. Food Chem.* **2013**, *61* (28), 6856–6864.

(36) Schrödinger Release 2018-2: LigPrep, Schrödinger, LLC, New York, NY, 2018.

(37) Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C. Towards the Comprehensive,

Rapid, and Accurate Prediction of the Favorable Tautomeric States of Drug-like Molecules in Aqueous Solution. *Journal of Computer-Aided Molecular Design*. 2010, pp 591–604.

(38) Schrödinger Release 2018-2: Epik, Schrödinger, LLC, New York, NY, 2018.

(39) Shityakov, S.; Förster, C. In Silico Structure-Based Screening of Versatile P-Glycoprotein Inhibitors Using Polynomial Empirical Scoring Functions. *Adv. Appl. Bioinforma. Chem.* **2014**, *7* (1), 1–9.

(40) Shityakov, S.; Foster, C. In Silico Predictive Model to Determine Vector-Mediated Transport Properties for the Blood-Brain Barrier Choline Transporter. *Adv. Appl. Bioinforma. Chem.* **2014**, *7* (1), 23–36.

(41) Fourches, D.; Muratov, E.; Ding, F.; Dokholyan, N. V; Tropsha, A. Predicting Binding Affinity of CSAR Ligands Using Both Structure- Based and Ligand-Based Approaches. *Chem. Inf. Model.* **2013**, 1915–1922.

(42) Shivakumar, D.; Williams, J.; Wu, Y. J.; Damm, W.; Shelley, J. C.; Sherman, W. Prediction of Absolute Solvation Free Energies Using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J. Chem. Theory Comput.* **2010**, *6* (5), 1509–1519.

(43) Guo, Z.; Mohanty, U.; Noehre, J.; Sawyer, T. K.; Sherman, W.; Krilov, G. Probing the Alpa-Helical Structural Stability of Stapled P53 Peptides: Molecular Dynamics Simulations and Analysis. *Chem. Biol. Drug Des.* **2010**, *75* (4), 348–359.

(44) Bowers, K. J.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; et al. Molecular Dynamics---Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; 2006; p 84.

(45) Schrödinger Release 2018-3: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2018. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2018.

(46) Yu, J.; Vavrusa, M.; Andreani, J.; Rey, J.; Tufféry, P.; Guerois, R. InterEvDock: A Docking Server to Predict the Structure of Protein-Protein Interactions Using Evolutionary Information. *Nucleic Acids Res.* **2016**, *44* (W1), W542–W549.

(47) Dong, G. Q.; Fan, H.; Schneidman-Duhovny, D.; Webb, B.; Sali, A.; Tramontano, A. Optimized Atomic Statistical Potentials: Assessment of Protein Interfaces and Loops.

*Bioinformatics* **2013**, *29* (24), 3158–3166.

(48) Garzon, J. I.; Lopéz-Blanco, J. R.; Pons, C.; Kovacs, J.; Abagyan, R.; Fernandez-Recio, J.; Chacon, P. FRODOCK: A New Approach for Fast Rotational Protein-Protein Docking. *Bioinformatics* **2009**, *25* (19), 2544–2551.

(49) Andreani, J.; Faure, G.; Guerois, R. InterEvScore: A Novel Coarse-Grained Interface Scoring Function Using a Multi-Body Statistical Potential Coupled to Evolution. *Bioinformatics* **2013**, *29* (14), 1742–1749.

(50) Lehmann, K.; Hoffmann, S.; Neudecker, P.; Suhr, M.; Becker, W. M.; Rösch, P. High-Yield Expression in Escherichia Coli, Purification, and Characterization of Properly Folded Major Peanut Allergen Ara h 2. *Protein Expr. Purif.* **2003**, *31* (2), 250–259.

(51) Sudhakar, a. History of Cancer, Ancient and Modern Treatment Methods Akulapalli. *J Cancer Sci Ther.* **2010**, *1* (2), 1–4.

(52) *Heart Disease and Stroke Statistics 2018*; 2018.

(53) Morgan, D. O. *The Cell Cycle: Principles of Control*; Lawrence, E., Ed.; Primers in Biology, 2007.

(54) NIH. Cancer Statistics - National Cancer Institute https://www.cancer.gov/about-cancer/understanding/statistics (accessed Aug 2, 2018).

(55) Echalier, A.; Hole, A. J.; Lolli, G.; Endicott, J. A.; Noble, M. E. M. An Inhibitor's-Eye View of the Atp-Binding Site of CDKs in Different Regulatory States. *ACS Chem. Biol.* **2014**, *9* (6), 1251–1256.

(56) Turkson, J. Cancer Drug Discovery and Anticancer Drug Development. *Mol. Basis Hum. Cancer* **2017**, 695–707.

(57) Law, M. E.; Corsino, P. E.; Narayan, S.; Law, B. K. Cyclin-Dependent Kinase Inhibitors as Anticancer Therapeutics. *Mol. Pharmacol.* **2015**, *88* (5), 846–852.

(58) Asghar, U.; Witkiewicz, A. K.; Turner, N. C.; Knudsen, E. S. The History and Future of Targeting Cyclin-Dependent Kinases in Cancer Therapy. *HHS Public Access* **2015**, *14* (2), 130–146.

(59) Neganova, I.; Vilella, F.; Atkinson, S. P.; Lloret, M.; Passos, J. F.; Von Zglinicki, T.; O'Connor, J. E.; Burks, D.; Jones, R.; Armstrong, L.; et al. An Important Role for CDK2 in G1 to S Checkpoint Activation and DNA Damage Response in Human Embryonic Stem Cells. *Stem Cells* **2011**, *29* (4), 651–659.

(60) Lim, S.; Kaldis, P. Cdks, Cyclins and CKIs: Roles beyond Cell Cycle Regulation. *Development*. 2013, pp 3079–3093.

(61) Malumbres, M.; Barbacid, M. Cell Cycle, CDKs and Cancer: A Changing Paradigm. *Nat. Rev. Cancer* **2009**, *9* (3), 153–166.

(62) Payton, M.; Chung, G.; Yakowec, P.; Wong, A.; Powers, D.; Xiong, L.; Zhang, N.; Leal, J.; Bush, T. L.; Santora, V.; et al. Discovery and Evaluation of Dual CDK1 and CDK2 Inhibitors. *Cancer Res.* **2006**, *66* (8), 4299–4308.

(63) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science (80-. ).* **2004**, *303* (5665), 1800–1805.

(64) Duronio, R. J.; Xiong, Y. Signaling Pathways That Control Cell Proliferation. *Cold Spring Harb. Perspect. Biol.* **2013**, *5* (3), 1–12.

(65) Peyressatre, M.; Prevel, C.; Pellerano, M.; Morris, M. C. Targeting Cyclin-Dependent Kinases in Human Cancers: From Small Molecules to Peptide Inhibitors. *Cancers*. 2015, pp 179–237.

(66) Mariaule, G.; Belmont, P. Cyclin-Dependent Kinase Inhibitors as Marketed Anticancer Drugs: Where Are We Now? A Short Survey. *Molecules* **2014**, *19* (9), 14366–14382.

(67) Sánchez-Martínez, C.; Gelbert, L. M.; Lallena, M. J.; De Dios, A. Cyclin Dependent Kinase (CDK) Inhibitors as Anticancer Drugs. *Bioorganic Med. Chem. Lett.* **2015**, *25* (17), 3420–3435.

(68) Filgueira, W. A. De; Mueller-Dieckmann, H.-J.; Schulze-Gahmen, U.; Worland, P. J.; Sausville, E.; Kim, S. Structural Basis for Specificity and Potency of a Flavonoid Inhibitor of Human CDK2, a Cell Cycle Kinase. *Biochemistry* **1996**, *93* (April), 2735–2740.

(69) Iwata, H. Clinical Development of CDK4 / 6 Inhibitor for Breast Cancer. *Breast Cancer* **2018**, No. 0123456789, 10–14.

(70) Balakrishnan, A.; Vyas, A.; Deshpande, K.; Vyas, Di. Pharmacological Cyclin Dependent Kinase Inhibitors: Implications for Colorectal Cancer. *World J. Gastroenterol.* **2016**, *22* (7), 2159–2164.

(71) Shapiro, G. I. Cyclin-Dependent Kinase Pathways as Targets for Cancer Treatment. *J. Clin. Oncol.* **2006**, *24* (11), 1770–1783.

(72) Harper, J. W.; Adams, P. D. Cyclin-Dependent Kinases. *Chem. Rev.* **2001**, *101* (8), 2511–2526.

(73) Cheng, C. K.; Gustafson, W. C.; Charron, E.; Houseman, B. T.; Zunder, E.; Goga, A.; Gray, N. S.; Pollok, B.; Oakes, S. A.; James, C. D.; et al. Dual Blockade of Lipid and Cyclin-Dependent Kinases Induces Synthetic Lethality in Malignant Glioma. *Proc. Natl. Acad. Sci.* **2012**, *109* (31), 12722–12727.

(74) Echavarria, I.; Jerez, Y.; Martin, M.; López-Tarruella, S. Incorporating CDK4/6 Inhibitors in the Treatment of Advanced Luminal Breast Cancer. *Breast Care* **2017**, *12* (5), 296–302.

(75) Etemadmoghadam, D.; Au-Yeung, G.; Wall, M.; Mitchell, C.; Kansara, M.; Loehrer, E.; Batzios, C.; George, J.; Ftouni, S.; Weir, B. A.; et al. Resistance to CDK2 Inhibitors Is Associated with Selection of Polyploid Cells in CCNE1-Amplified Ovarian Cancer. *Clin. Cancer Res.* **2013**, *19* (21), 5960–5971.

(76) Besson, A.; Dowdy, S. F.; Roberts, J. M. CDK Inhibitors: Cell Cycle Regulators and Beyond. *Dev. Cell* **2008**, *14* (2), 159–169.

(77) Ibrahim, D. A.; El-Metwally, A. M. Design, Synthesis, and Biological Evaluation of Novel Pyrimidine Derivatives as CDK2 Inhibitors. *Eur. J. Med. Chem.* **2010**, *45* (3), 1158–1166.

(78) Senderowicz, A. M.; Sausville, E. A. Preclinical and Clinical Development of Cyclin-Dependent Kinase Modulators. *J. Natl. Cancer Inst.* **2000**, *92* (5), 376–387.

(79) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nat. Rev. Cancer* **2009**, *9* (1), 28–39.

(80) Cicenas, J.; Kalyan, K.; Valius, M. Roscovitine in Cancer and Other Diseases. *Ann. Transl. Med.* **2015**, *3* (10), 1–12.

(81) Cicenas, J.; Kalyan, K.; Sorokinas, A.; Jatulyte, A.; Valiunas, D.; Kaupinis, A.; Valius, M. Highlights of the Latest Advances in Research on CDK Inhibitors. *Cancers (Basel).* **2014**, *6* (4), 2224–2242.

(82) Meijer, L.; Raymond, E. Roscovitine and Other Purines as Kinase Inhibitors. From Starfish Oocytes to Clinical Trials. *Acc. Chem. Res.* **2003**, *36* (6), 417–425.

(83) Johnson, L. N. Protein Kinase Inhibitors: Contributions from Structure to Clinical Compounds. *Q. Rev. Biophys.* **2009**, *42* (1), 1–40.

(84) Byth, K. F.; Thomas, A.; Hughes, G.; Forder, C.; McGregor, A.; Geh, C.; Oakes, S.; Green, C.; Walker, M.; Newcombe, N.; et al. AZD5438, a Potent Oral Inhibitor of Cyclin-Dependent Kinases 1, 2, and 9, Leads to Pharmacodynamic Changes and Potent

Antitumor Effects in Human Tumor Xenografts. *Mol. Cancer Ther.* **2009**, *8* (7), 1856–1866.

(85) Bogenberger, J.; Whatcott, C.; Hansen, N.; Delman, D.; Shi, C.-X.; Kim, W.; Haws, H.; Soh, K.; Lee, Y. S.; Peterson, P.; et al. Combined Venetoclax and Alvocidib in Acute Myeloid Leukemia. *Oncotarget* **2017**, *8* (63), 107206–107222.

(86) Paruch, K.; Dwyer, M. P.; Alvarez, C.; Brown, C.; Chan, T. Y.; Doll, R. J.; Keertikar, K.; Knutson, C.; McKittrick, B.; Rivera, J.; et al. Discovery of Dinaciclib (SCH 727965): A Potent and Selective Inhibitor of Cyclin-Dependent Kinases. *ACS Med. Chem. Lett.* **2010**, *1* (5), 204–208.

(87) Guha, M. Cyclin-Dependent Kinase Inhibitors Move into Phase III. *Nat. Rev. Drug Discov.* **2012**, *11* (12), 892–894.

(88) Ece, A.; Sevin, F. The Discovery of Potential Cyclin A/CDK2 Inhibitors: A Combination of 3D QSAR Pharmacophore Modeling, Virtual Screening, and Molecular Docking Studies. *Med. Chem. Res.* **2013**, *22* (12), 5832–5843.

(89) Samanta, S.; Debnath, B.; Basu, A.; Gayen, S.; Srikanth, K.; Jha, T. Exploring QSAR on 3-Aminopyrazoles as Antitumor Agents for Their Inhibitory Activity of CDK2/Cyclin A. *Eur. J. Med. Chem.* **2006**, *41* (10), 1190–1195.

(90) Tripathi, S. K.; Muttineni, R.; Singh, S. K. Extra Precision Docking, Free Energy Calculation and Molecular Dynamics Simulation Studies of CDK2 Inhibitors. *J. Theor. Biol.* **2013**, *334*, 87–100.

(91) Canduri, F.; Da Silveira, N. J. F.; Camera, J. C.; De Azevedo, W. F. Structural Bioinformatics Study of Cyclin-Dependent Kinases Complexed with Inhibitors. *Eclet. Quim.* **2003**, *28* (1), 45–53.

(92) Baltus, C. B.; Jorda, R.; Marot, C.; Berka, K.; Bazgier, V.; Kryštof, V.; Prié, G.; Viaud-Massuard, M. C. Synthesis, Biological Evaluation and Molecular Modeling of a Novel Series of 7-Azaindole Based Tri-Heterocyclic Compounds as Potent CDK2/Cyclin E Inhibitors. *Eur. J. Med. Chem.* **2016**, *108*, 701–719.

(93) Tripathi, S. K.; Singh, S. K. Insights into the Structural Basis of 3,5-Diaminoindazoles as CDK2 Inhibitors: Prediction of Binding Modes and Potency by QM-MM Interaction, MESP and MD Simulation. *Mol. Biosyst.* **2014**, *10* (8), 2189–2201.

(94) H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N.

Shindyalov, P. E. B. The Protein Data Bank Nucleic Acids Research https://www.rcsb.org/ (accessed Aug 2, 2018).

(95) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Otter, T. K.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; 2007.

(96) Kang, Y. N.; Stuckey, J. A. *Crystal Structure of the Cdk2 in Complex with Aminopyrazole Inhibitor*.

(97) Schrödinger Release 2018-2: Maestro, Schrödinger, LLC, New York, NY, 2018.

(98) Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J. L.; Glide. Glide: A New Approach for Rapid, Accurate Docking and Scoring. II. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *2* (47), 1750–1759.

(99) McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **2010**, *1697900* (Scipy), 51–56.

(100) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Develop Ment Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.

(101) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties as a Means of Improving Compound Disposition and Safety. *Chem. Res. Toxicol.* **2011**, *24* (9), 1420–1456.

(102) Veber, D. F.; Johnson, S. R.; Cheng, H.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(103) Curatolo, W. Physical Chemical Properties of Oral Drug Candidates in the Discovery and Exploratory Development Settings. *Pharm. Sci. Technol. Today* **1998**, *1* (9), 387–393.

(104) Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43* (21), 3867–3877.

(105) Gaulton, A.; Hersey, A.; Nowotka, M. L.; Patricia Bento, A.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954.

(106) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (7), 2932–2942.

(107) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. Structure-Based Predictions of Activity Cliffs. *J. Chem. Inf. Model.* **2015**, *55* (5), 1062–1076.

(108) RDKit: Open-Source Cheminformatics. http://www.rdkit.org.

(109) Song, J.; Tang, J.; Guo, F. Identification of Inhibitors of MMPS Enzymes via a Novel Computational Approach. *Int. J. Biol. Sci.* **2018**, *14* (8), 863–871.

(110) Tabandeh, F.; Shariati, P.; Khodabandeh, M. Application of Two-Dimensional Gel Electrophoresis to Microbial Systems. *Gel Electrophor. - Princ. Basics* **2012**, 335–366.

(111) *Proteomic of Microbial Pathogens*; Jungblut, P. R., Hecker, M., Eds.; 2007; Vol. 2.

(112) Piersma, S. R.; Warmoes, M. O.; Wit, M. de; Reus, I. de; Knol, J. C.; R., J. C. Whole Gel Processing Procedure for GeLC-MS / MS Based Proteomics. **2013**, 1–9.

(113) Dzieciatkowska, M.; Hill, R.; Hansen, K. C. GeLC-MS/MS Analysis of Complex Protein Mixtures. **2014**, *1156*, 53–66.

(114) Horiuchi, D.; E. Huskey, N.; Kusdra, L.; Wohlbold, L.; Merrick, K. A.; Zhang, C.; Creasman, K. J.; Shokat, K. M.; Fisher, R. P.; Goga, A. Chemical-Genetic Analysis of Cyclin Dependent Kinase 2 Function Reveals an Important Role in Cellular Transformation by Multiple Oncogenic Pathways. *Proc. Natl. Acad. Sci.* **2012**, *109* (17), E1019–E1027.

(115) Enserink, J. M.; Kolodner, R. D. An Overview of Cdk1-Controlled Targets and Processes. **2010**, 1–41.

# APPENDIX

## Appendix 1: Supplemental Information (Tables and Figures)

**SI Table 2.1:** Docking scores of pH 3.26 and 7.00 calculated with the SiteMap docking grid.

| pH 3.26 Results | | **pH 7.00 Results** | | |
|---|---|---|---|---|
| Variant | Docking score | Variant | Docking score | **Absolute Difference** |
| Benzoic acid-1 | -3.971 | Benzoic acid-1 | -4.540 | **0.569** |
| Caffeic acid-1 | -5.149 | Caffeic acid-1 | -5.046 | **0.103** |
| Caffeic acid-2 | -6.195 | Caffeic acid-2 | -4.136 | **2.059** |
| Catechin - D-1 | -5.759 | Catechin - D-1 | -5.759 | **0.000** |
| Chlorogenic acid-1 | -7.382 | Chlorogenic acid-1 | -7.103 | **0.279** |
| Cyanidin-1 | -7.819 | Cyanidin-1 | -7.791 | **0.028** |
| Cyanidin-3-arabinoside-1 | -7.236 | Cyanidin-3-arabinoside-1 | -7.207 | **0.029** |
| Cyanidin-3-galactoside-1 | -7.081 | Cyanidin-3-galactoside-1 | -7.051 | **0.030** |
| Cyanidin-3-glucoside-1 | -7.486 | Cyanidin-3-glucoside-1 | -7.457 | **0.029** |
| Delphinidin-3-galactoside-1 | -8.053 | Delphinidin-3-galactoside-1 | -8.023 | **0.030** |
| Delphinidin-3-glucoside-1 | -8.706 | Delphinidin-3-glucoside-1 | -8.677 | **0.029** |
| Epicatechin gallate-1 | -5.917 | Epicatechin gallate-1 | -5.890 | **0.027** |
| Epicatechin-1 | -5.779 | Epicatechin-1 | -5.779 | **0.000** |
| Epigallocatechin gallate-1 | -6.295 | Epigallocatechin gallate-1 | -6.249 | **0.046** |
| Epigallocatechin-1 | -6.566 | Epigallocatechin-1 | -6.566 | **0.000** |
| Ferulic-1 | -4.600 | Ferulic-1 | -5.336 | **0.736** |
| Ferulic-2 | -4.860 | Ferulic-2 | -5.098 | **0.238** |
| Gallic-1 | -5.915 | Gallic-1 | -5.547 | **0.368** |
| Kaempferol-1 | -5.742 | Kaempferol-1 | -5.742 | **0.000** |
| Malvidin-1 | -5.474 | Malvidin-1 | -5.446 | **0.028** |
| Malvidin-3-arabinoside-1 | -6.164 | Malvidin-3-arabinoside-1 | -6.134 | **0.030** |
| Myricetin-1 | -7.131 | Myricetin-1 | -7.131 | **0.000** |
| p-coumaric acid-1 | -4.173 | p-coumaric acid-1 | -4.542 | **0.369** |
| Peonidin-3-arabinoside-1 | -5.186 | Peonidin-3-arabinoside-1 | -5.156 | **0.030** |
| Peonidin-3-galactoside-1 | -6.362 | Peonidin-3-galactoside-1 | -6.333 | **0.029** |
| p-hydroxybenzoic acid-1 | -4.031 | p-hydroxybenzoic acid-1 | -4.718 | **0.687** |
| Procyanidin A2-1 | -5.876 | Procyanidin A2-1 | -5.876 | **0.000** |
| Procyanidin B1-1 | -6.104 | Procyanidin B1-1 | -5.657 | **0.447** |
| Procyanidin B2-1 | -6.933 | Procyanidin B2-1 | -6.136 | **0.797** |
| Procyanidin C1-1 | -7.563 | Procyanidin C1-1 | -7.218 | **0.345** |
| Quercetin-1 | -7.023 | Quercetin-1 | -7.023 | **0.000** |
| Quercetin-3-arabinoside (furanoside)-1 | -7.216 | Quercetin-3-arabinoside (furanoside)-1 | -7.216 | **0.000** |

**SI Table 2.1** (continued).

| | | | | |
|---|---|---|---|---|
| Quercetin-3-arabinoside (pyranoside)-1 | -7.278 | Quercetin-3-arabinoside (pyranoside)-1 | -7.278 | **0.000** |
| Quercetin-3-galactoside (hyperoside)-1 | -7.248 | Quercetin-3-galactoside (hyperoside)-1 | -7.248 | **0.000** |
| Quercetin-3-glucoside (isoquercetin)-1 | -6.237 | Quercetin-3-glucoside (isoquercetin)-1 | -6.237 | **0.000** |
| Quercetin-3-rhamnoside (quercetrin)-1 | -7.658 | Quercetin-3-rhamnoside (quercetrin)-1 | -7.658 | **0.000** |
| Quercetin-3-rutinoside (rutin)-1 | -7.425 | Quercetin-3-rutinoside (rutin)-1 | -7.425 | **0.000** |
| Quercetin-3-xyloside-1 | -6.336 | Quercetin-3-xyloside-1 | -6.336 | **0.000** |
| Resveratrol - trans-1 | -4.354 | Resveratrol - trans-1 | -4.354 | **0.000** |
| Sinapinic acid-1 | -4.946 | Sinapinic acid-1 | -7.607 | **2.661** |
| Vanillic-1 | -4.058 | Vanillic-1 | -4.995 | **0.937** |

**SI Table 2.2**. XP Docking and eModel scores for top 10 results at pH 7.00.

| Top 10 Polyphenols | SiteMap Docking Score | SiteMap eModel Score | "Blind Docking" Docking Score | "Blind Docking" eModel Score | Absolute Difference |
|---|---|---|---|---|---|
| Delphinidin-3-Glucoside | -8.677 | -71.314 | -7.842 | -66.502 | 0.835 |
| Delphinidin-3-Galactoside | -8.023 | -67.300 | -7.977 | -66.671 | 0.046 |
| Cyanidin | -7.791 | -59.763 | -8.101 | -54.732 | 0.310 |
| Quercetin-3-Rhamnoside | -7.658 | -56.817 | -6.108 | -59.834 | 1.550 |
| Chlorogenic Acid | -7.103 | -57.828 | -8.164 | -56.176 | 1.061 |
| Procyanidin C1 | -7.218 | -89.078 | -6.545 | -86.484 | 0.673 |
| Cyanidin Glucoside | -7.457 | -66.466 | -8.538 | -66.849 | 1.081 |
| Quercetin-3-Rutinoside | -7.425 | -84.813 | -8.766 | -91.763 | 1.341 |
| Quercetin-3-Arabinoside (pyranoside) | -7.278 | -55.639 | -7.551 | -58.218 | 0.273 |
| Cyanidin-3-Arabinoside | -7.207 | -63.371 | -7.700 | -63.058 | 0.493 |

**SI Table 2.3**. Average of docking and eModel scores for SiteMap and "Blind docking" at pH 3.26 and 7.00

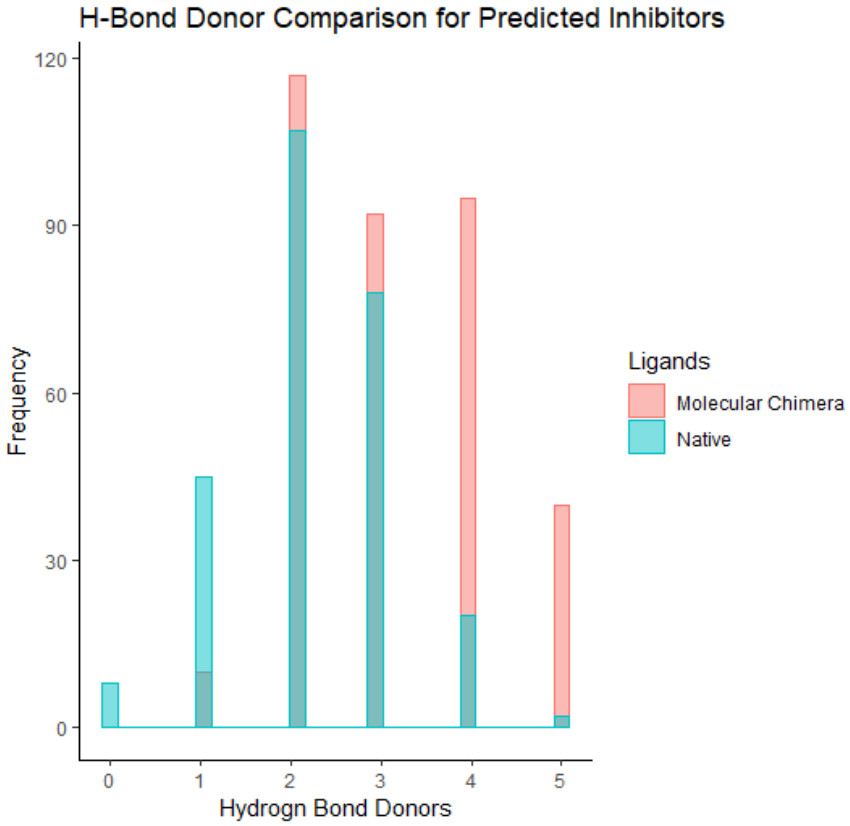
| Compound | Docking Score | eModel Score |
|---|---|---|
| Tannic acid B | -9.237 | 3229.436 |
| Delphinidin-3-glucoside | -8.274 | -68.908 |
| Quercetin-3-rutinoside (rutin) | -8.096 | -88.288 |
| Delphinidin-3-galactoside | -8.015 | -66.986 |
| Cyanidin-3-glucoside | -8.012 | -66.658 |
| Cyanidin | -7.960 | -57.248 |
| Chlorogenic acid | -7.561 | -61.352 |
| Procyanidin C1 | -7.487 | -90.985 |
| Cyanidin-3-galactoside | -7.472 | -65.826 |
| Cyanidin-3-arabinoside | -7.468 | -63.215 |
| Quercetin-3-arabinoside (pyranoside) | -7.415 | -56.929 |
| Myricetin | -7.151 | -59.326 |
| Quercetin-3-arabinoside (furanoside) | -7.050 | -66.510 |
| Quercetin-3-galactoside (hyperoside) | -6.917 | -57.817 |
| Quercetin-3-rhamnoside (quercetrin) | -6.883 | -58.326 |
| Quercetin | -6.840 | -58.629 |
| Epigallocatechin gallate | -6.701 | -68.313 |
| Epigallocatechin | -6.538 | -48.619 |
| Quercetin-3-glucoside (isoquercetin) | -6.297 | -61.148 |
| Peonidin-3-galactoside | -6.213 | -59.501 |
| Malvidin-3-arabinoside | -6.046 | -66.937 |
| Epicatechin gallate | -5.967 | -68.792 |
| Procyanidin B2 | -5.902 | -54.751 |
| Gallic | -5.832 | -36.780 |
| Epicatechin | -5.817 | -52.728 |
| Quercetin-3-xyloside | -5.743 | -51.493 |
| Kaempferol | -5.736 | -53.390 |
| Procyanidin B1 | -5.662 | -55.435 |
| Sinapic acid | -5.598 | -35.805 |
| Catechin - D | -5.560 | -48.134 |
| Malvidin-3-Glucoside | -5.514 | -65.594 |
| Procyanidin A2 | -5.379 | -62.730 |
| Peonidin-3-arabinoside | -5.333 | -57.237 |
| Malvidin-3-Galactoside | -5.253 | -62.282 |
| Malvidin | -5.234 | -60.263 |
| Caffeic acid | -5.173 | -36.803 |

**SI Table 2.3** (continued).

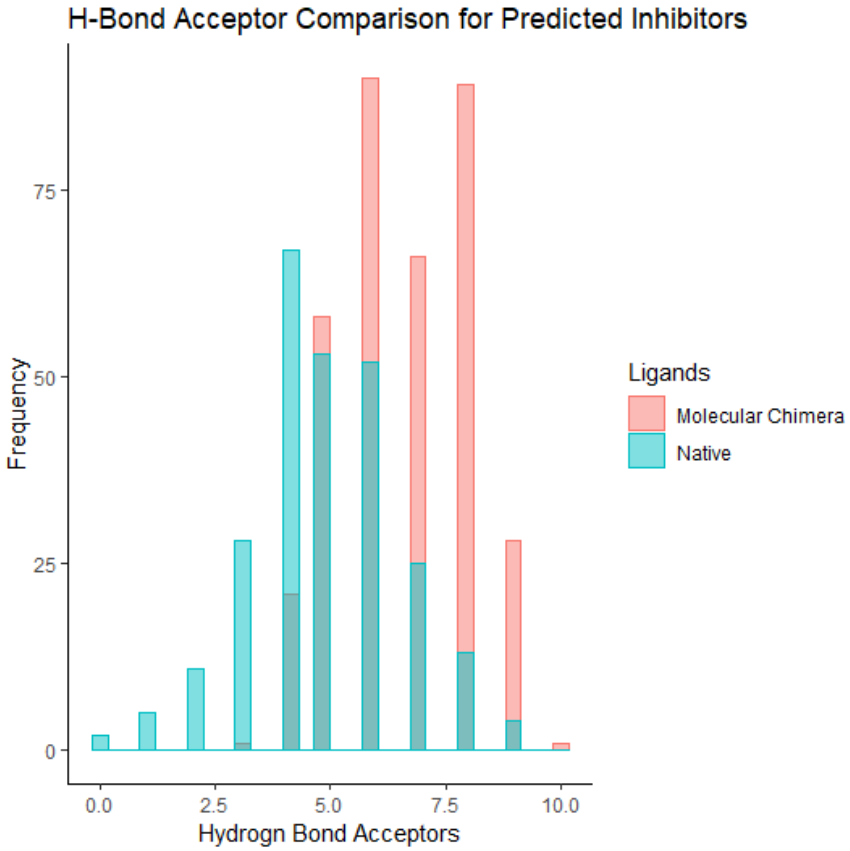
| | | |
|---|---|---|
| Ferulic | -5.084 | -33.875 |
| Vanillic | -4.885 | -29.084 |
| p-hydroxybenzoic acid | -4.750 | -30.105 |
| Resveratrol - trans | -4.516 | -42.862 |
| p-coumaric acid | -4.405 | -30.252 |
| Benzoic acid | -4.290 | -25.216 |

**SI Figure: 3.1:** (A-E) These five graphs show the comparable results for native and *molecular chimera* inhibitors according to lipophilicity, oral absorption and Caco-2 permeability. (A) Hydrogen Bond Donors acceptable < 5; (B) Hydrogen Bond Acceptors acceptable < 10; (C)Molecular Weight acceptable lower MW are preferred; (D) Rotatable Bonds acceptable ≤ 13; (F) Topological Polar Surface Area acceptable ≤ 140 Å.

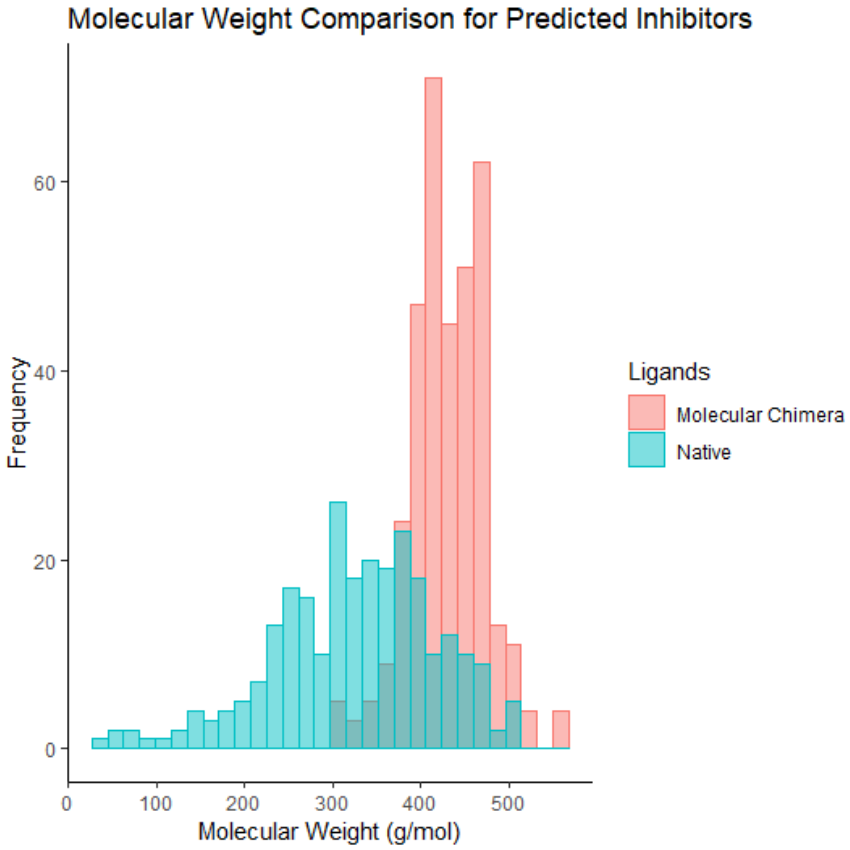
**A**


H-Bond Donor Comparison for Predicted Inhibitors
Frequency
120
90
60
30
0
0
1
2
3
4
5
Hydrogn Bond Donors
Ligands
Molecular Chimera
Native

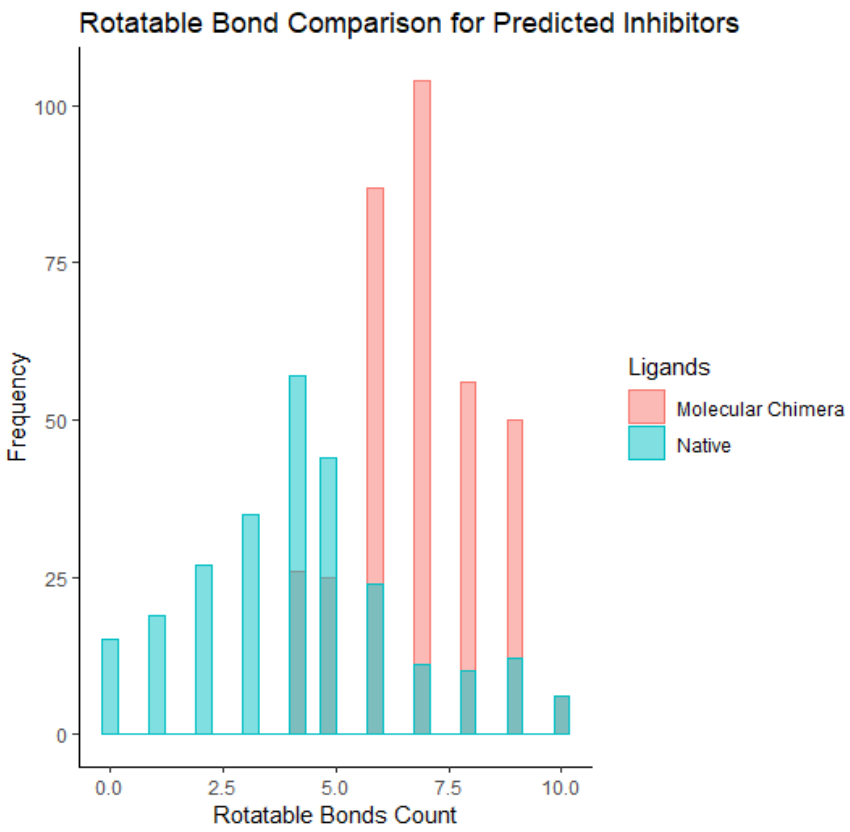
**B**


H-Bond Acceptor Comparison for Predicted Inhibitors
Frequency
75
50
25
0
0.0
2.5
5.0
7.5
10.0
Hydrogn Bond Acceptors
Ligands
Molecular Chimera
Native

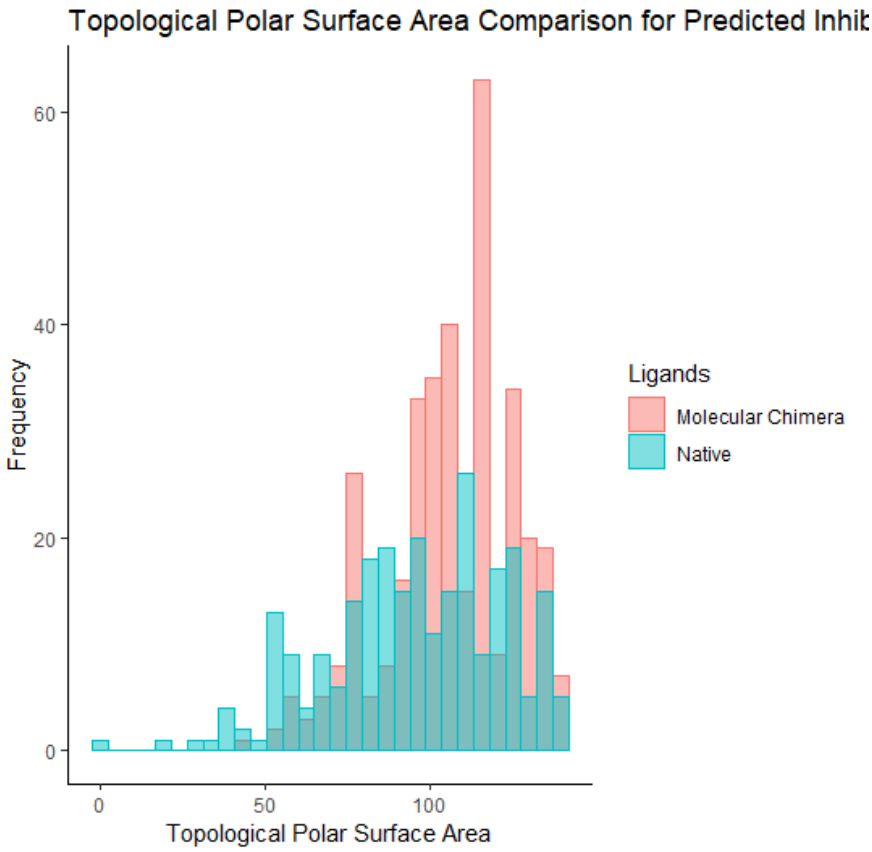
**C**


Molecular Weight Comparison for Predicted Inhibitors
Frequency
0
20
40
60
0
100
200
300
400
500
Molecular Weight (g/mol)
Ligands
Molecular Chimera
Native

**D**

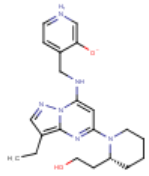

Rotatable Bond Comparison for Predicted Inhibitors
Frequency
0
25
50
75
100
0.0
2.5
5.0
7.5
10.0
Rotatable Bonds Count
Ligands
Molecular Chimera
Native

**E**

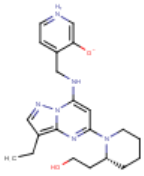

Topological Polar Surface Area Comparison for Predicted Inhib
Frequency
60
40
20
0
0
50
100
Topological Polar Surface Area
Ligands
Molecular Chimera
Native

**SI Figure 3.2**: Structures of the top 50 active *molecular chimera* inhibitors.

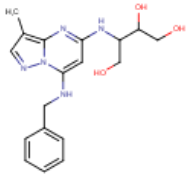
0S01QK.17364c1nccc1



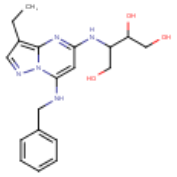
1QK5SC.19734c1nccc1



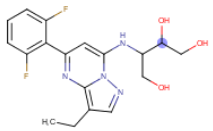
NS9PDY.17781c1cn2nccc2nc1



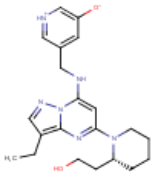
1QKNS9.13799c1cn2nccc2nc1



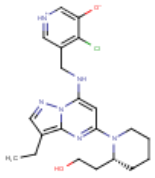
NS9SCX.17938c1cn2nccc2nc1



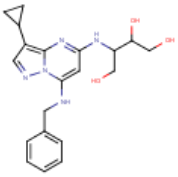
1QK5SC.19710c1nccc1



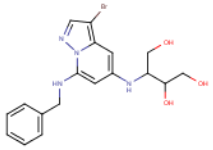
1QKX75.06592c1nccc1



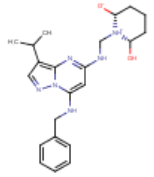
NS9SCZ.12661c1cn2nccc2nc1



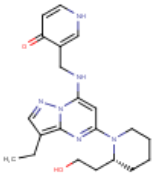
SCQNS9.05747c1ccn2nccc2c1_part1_list1



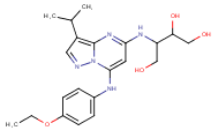
I73I74.09595N1CCCCC1



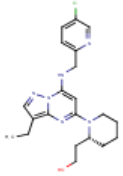
0S01QK.17383c1nccc1



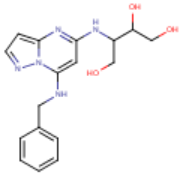
NS9PDY.17783c1cn2nccc2nc1



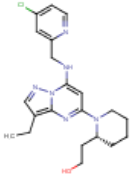
1QKX75.06617c1nccc1



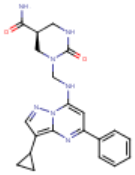
2SCNS9.03912c1cn2nccc2nc1



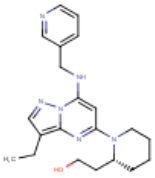
1QKX75.06606c1nccc1



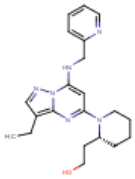
RJISCZ.078789C1CNCNC1



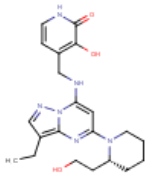
07Z1QK.00196c1nccc1



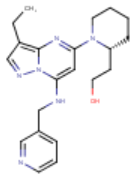
07Z1QK.00203c1nccc1



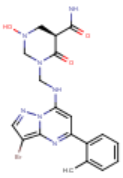
0S01QK.17369c1nccc1



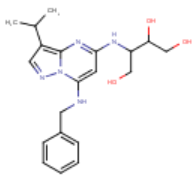
1QK2SC.00006c1cn2nccc2nc1



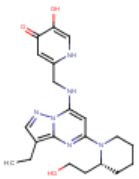
RJI5SC.021317C1CNCNC1



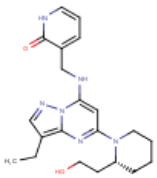
2SCNS9.03888c1cn2nccc2nc1



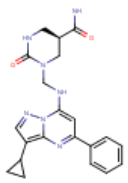
0S01QK.17366c1nccc1



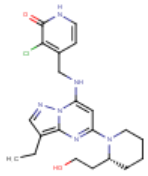
0S01QK.17344c1nccc1



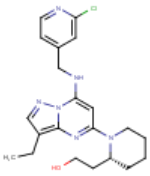
RJISCZ.078832C1CNCNC1



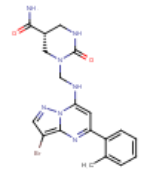
1QKX75.06573c1nccc1



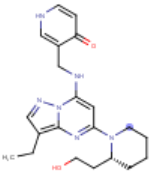
1QKX75.06583c1nccc1



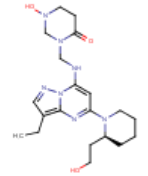
RJI5SC.021273C1CNCNC1



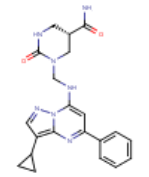
1QK5SC.19711c1nccc1



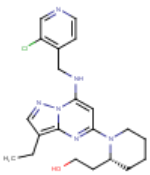
RJI1QK.009354C1CNCNC1



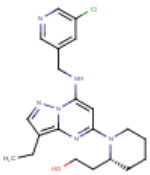
RJISCZ.078832C1CNCNC1



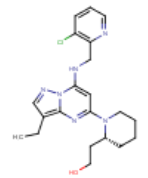
1QKX75.06578c1nccc1



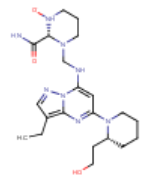
1QKX75.06562c1nccc1



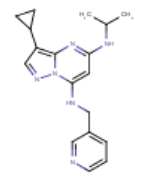
1QKX75.06612c1nccc1



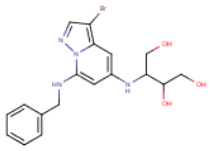
RJI1QK.009753C1CNCNC1



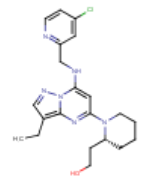
LZBSCZ.17686c1cn2nccc2nc1



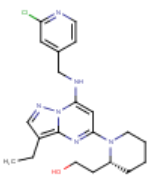
SCQNS9.05747c1ccn2nccc2c1_part1_list1



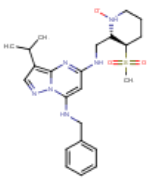
1QKX75.06637c1nccc1



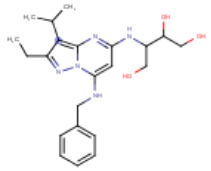
1QKX75.06570c1nccc1



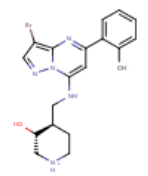
I73LIA.05220N1CCCCC1



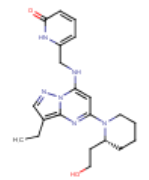
NS9SCX.17855c1cn2nccc2nc1



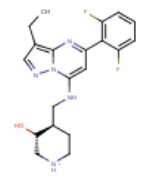
5SCI73.06146c1cn2nccc2nc1



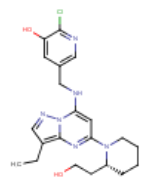
07Z1QK.00205c1nccc1



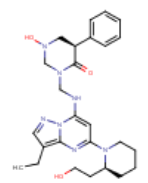
I73SCX.17521c1cn2nccc2nc1



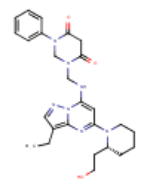
1QKX75.06590c1nccc1



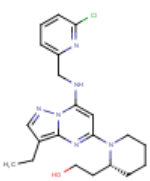
RJI1QK.009873C1CNCNC1



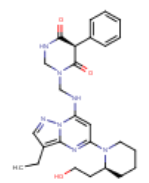
RJI1QK.009285C1CNCNC1



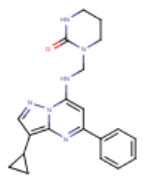
1QKX75.06619c1nccc1



RJI1QK.009785C1CNCNC1



RJISCZ.078837C1CNCNC1



**SI Table 3.1.** List of frequent scaffolds found in the 316 CDK2 ligands and the number of new *molecular chimera* created based on them.

| Key Scaffolds | Number of heavy atoms | Number of occurrences in the 316 compounds | Range of experimental $pIC_{50}$ | Number of ligands with unknown $IC_{50}$ | Unique number of compounds generated |
|---|---|---|---|---|---|
| C1CC1 | 3 | 14 | 6.5 - 8.7 | 2 | 589 |
| C1CCC1 | 4 | 2 | 7.1 - 7.4 | 0 | 95 |
| C1=CN=NC1 | 5 | 2 | 5.9 - 8.7 | 0 | 43 |
| c1nccc1 | 5 | 3 | 3.9 - 8.5 | 0 | 704 |
| C1CSCN1 | 5 | 7 | 4.3 - 7.5 | 1 | 1,849 |
| C1=CSCN1 | 5 | 4 | 7.0 | 3 | 1,572 |
| c1nncn1 | 5 | 1 | 6.6 | 0 | 732 |
| c1ncsc1 | 5 | 62 | 4.0 - 8.2 | 16 | 18,794 |
| c1nncc1 | 5 | 30 | 4.0 - 8.5 | 6 | 15,605 |
| c1sccc1 | 5 | 6 | 4.8 - 8.5 | 2 | 573 |
| C1CCNC1 | 5 | 7 | 4.5 - 8.7 | 0 | 1,808 |
| C1=NCCS1 | 5 | 3 | NA | 3 | 256 |
| c1occn1 | 5 | 3 | 7.2 - 8.1 | 0 | 65 |
| C1CNSC1 | 5 | 2 | 7.1 - 7.5 | 0 | 147 |
| C1CCCC1 | 5 | 2 | 7.0 - 7.3 | 0 | 31 |
| C1CNCN1 | 5 | 1 | 8.7 | 0 | 13,805 |
| c1nccnc1 | 6 | 4 | 3.5 - 5.3 | 2 | 318 |
| C1=NCN=CC1 | 6 | 2 | 6.1 - 6.6 | 0 | 90 |
| N1CCOCC1 | 6 | 7 | 6.1 - 6.3 | 5 | 477 |
| C1CNCNC1 | 6 | 1 | NA | 1 | 103,842 |
| C1CCOCC1 | 6 | 2 | 8.3 | 1 | 13,392 |
| C1=CCCC=C1 | 6 | 2 | NA | 2 | 241 |
| c1ccccc1 | 6 | 226 | 3.9 - 8.7 | 69 | 3,575,180 |
| c1ncccc1 | 6 | 107 | 4.3 - 8.7 | 20 | 25,736 |
| c1ncccn1 | 6 | 70 | 4.4 - 8.7 | 31 | 82,395 |
| C1=CCNC=C1 | 6 | 2 | 4.3 | 1 | 727 |
| C1CCCCC1 | 6 | 30 | 4.0 - 8.0 | 1 | 15,494 |
| C1NCCNC1 | 6 | 8 | 5.0 - 7.9 | 2 | 2,826 |
| c1ncncn1 | 6 | 7 | 5.2 | 6 | 258 |
| N1CCCCC1 | 6 | 16 | 4.6 - 8.7 | 5 | 10,561 |
| N1CCNCCC1 | 7 | 2 | 5.9 | 1 | 43,218 |
| c1nnc2c1CNC2 | 8 | 1 | 8.1 | 0 | 230 |
| C1=[NH+]CC2=C1 CN=N2 | 8 | 1 | 5.9 | 0 | 128 |
| c1cnn2nccc2c1 | 9 | 2 | 6.3 - 6.9 | 0 | 143 |
| c1ccc2ncnc2c1 | 9 | 7 | 4.2 - 7.3 | 2 | 3,339 |
| c1ccn2ccnc2c1 | 9 | 6 | 5.5 - 8.7 | 0 | 2,960 |

**SI Table 3.1** (continued).

| | | | | | |
|---|---|---|---|---|---|
| c1ccc2nncc2c1 | 9 | 9 | 5.5 - 8.4 | 2 | 14,816 |
| c1cnc2ccN[n+]2c1 | 9 | 1 | NA | 1 | 16,876 |
| c1cnc2ncN[n+]2c1 | 9 | 1 | NA | 1 | 26,506 |
| c1cnc2ncnn2c1 | 9 | 5 | 5.0 - 6.6 | 1 | 344 |
| c1ccc2scnc2c1 | 9 | 1 | NA | 1 | 9,727 |
| c1cnc2nccc2c1 | 9 | 2 | 8.0 - 8.5 | 0 | 171 |
| c1ncn2nccc2n1 | 9 | 1 | 7.4 | 0 | 8,263 |
| c1cn2ccnc2cn1 | 9 | 2 | 4.7 - 6.1 | 0 | 128 |
| c1ccn2nccc2c1 | 9 | 1 | 7.0 | 0 | 28,714 |
| c1ncc2cncn2n1 | 9 | 1 | 6.6 | 0 | 8,511 |
| c1ccc2nnnc2c1 | 9 | 2 | 4.8 - 8.0 | 0 | 194 |
| c1nnc2c1CCCC2 | 9 | 5 | NA | 5 | 887 |
| C1=c2ccccc2=NC1 | 9 | 1 | 7.2 | 0 | 28,854 |
| c1cc2ncnc2nc1 | 9 | 1 | 7.2 | 0 | 16,437 |
| c1cn2nccc2nc1 | 9 | 21 | 4.6 - 8.7 | 5 | 18,276 |
| c1cc2ncnc2cn1 | 9 | 1 | 4.9 | 0 | 9,829 |
| c1ncc2ncnc2n1 | 9 | 29 | 4.0 - 8.1 | 7 | 23,152 |
| c1cc2cnccc2n1 | 9 | 1 | 7.0 | 0 | 17,755 |
| c1ncc2nncc2n1 | 9 | 1 | 7.4 | 0 | 8,530 |
| c1cc2CSCc2cc1 | 9 | 2 | 7.9 - 8.1 | 0 | 300 |
| c1cc2NCCc2cc1 | 9 | 15 | 3.9 - 8.5 | 4 | 29,856 |
| c1cc2nccc2cc1 | 9 | 1 | NA | 1 | 41,839 |
| c1cc2ncncc2cc1 | 10 | 2 | 6.0 - 6.2 | 0 | 498 |
| c1cc2cccnc2cc1 | 10 | 3 | 5.0 | 2 | 852 |
| c1cc2ccccc2cc1 | 10 | 8 | 3.8 - 7.9 | 0 | 7,720 |
| c1Cc2ccccc2Oc1 | 10 | 2 | 7.1 | 1 | 2,081 |

**SI Table 3.2.** The top 50 *molecular chimera* inhibitors with respective docking and eModel scores as well as predicted $pIC_{50}$.

| Inhibitor Name | Docking Score | eModel Score | Predicted $pIC_{50}$ |
|---|---|---|---|
| 0S01QK.17364c1nccc1 | -12.614 | -61.272 | 7.08 |
| 1QK5SC.19729c1nccc1 | -12.614 | -61.272 | 7.08 |
| SCQNS9.05747c1ccn2nccc2c1_part1_list1 | -12.575 | -86.572 | 7.00 |
| 1QK5SC.19734c1nccc1 | -12.562 | -61.291 | 7.08 |
| NS9PDY.17781c1cn2nccc2nc1 | -11.844 | -88.795 | 7.30 |
| 1QKNS9.13799c1cn2nccc2nc1 | -11.348 | -78.107 | 7.45 |
| NS9SCX.17938c1cn2nccc2nc1 | -11.298 | -75.801 | 7.09 |
| 1QK5SC.19710c1nccc1 | -11.193 | -60.807 | 7.32 |
| 1QKX75.06592c1nccc1 | -11.147 | -63.558 | 7.03 |
| NS9SCZ.12661c1cn2nccc2nc1 | -10.746 | -83.246 | 7.31 |
| NNNN41.03876c1ncn2nccc2n1_part1_list1 | -10.647 | -63.813 | 7.00 |
| RJISCZ.078832C1CNCNC1 | -10.516 | -82.447 | 7.06 |
| I73I74.09595N1CCCCC1 | -10.378 | -62.492 | 7.02 |
| 0S01QK.17383c1nccc1 | -10.352 | -66.129 | 7.21 |
| 1QK5SC.19747c1nccc1 | -10.352 | -66.129 | 7.21 |
| NS9PDY.17783c1cn2nccc2nc1 | -10.307 | -72.852 | 7.02 |
| RJI1QK.009354C1CNCNC1 | -10.221 | -83.788 | 7.05 |
| RJI5SC.021317C1CNCNC1 | -10.184 | -90.294 | 7.04 |
| NNNCK9.03691c1ncn2nccc2n1_part1_list1 | -10.183 | -64.589 | 7.08 |
| 1QKX75.06617c1nccc1 | -10.046 | -67.887 | 7.20 |
| RJI1QK.009606C1CNCNC1 | -10.037 | -67.857 | 7.14 |
| 2SCNS9.03912c1cn2nccc2nc1 | -10.030 | -76.426 | 7.02 |
| RJI1QK.009491C1CNCNC1 | -10.020 | -75.271 | 7.02 |
| 1QKX75.06606c1nccc1 | -10.018 | -71.513 | 7.15 |
| RJII73.053870C1CNCNC1 | -10.004 | -70.348 | 7.18 |
| RJII73.053732C1CNCNC1 | -9.988 | -65.349 | 7.20 |
| RJI5SC.021273C1CNCNC1 | -9.964 | -83.086 | 7.18 |
| 07Z1QK.00196c1nccc1 | -9.957 | -66.381 | 7.28 |
| 07Z1QK.00203c1nccc1 | -9.954 | -67.930 | 7.20 |
| SCXX84.360737c1ccccc1_4_of_4 | -9.938 | -99.741 | 7.05 |
| RJI5SC.021919C1CNCNC1 | -9.921 | -73.098 | 7.01 |
| 0S01QK.17369c1nccc1 | -9.910 | -72.645 | 7.11 |
| 1QK5SC.19723c1nccc1 | -9.910 | -72.645 | 7.11 |
| RJISCZ.078817C1CNCNC1 | -9.893 | -73.442 | 7.05 |
| RJI1QK.009285C1CNCNC1 | -9.876 | -70.181 | 7.16 |
| 1QK2SC.00006c1cn2nccc2nc1 | -9.840 | -65.497 | 7.28 |
| SCXX84.360659c1ccccc1_4_of_4 | -9.839 | -99.752 | 7.05 |
| 2SCNS9.03888c1cn2nccc2nc1 | -9.829 | -70.393 | 7.80 |
| 0S01QK.17366c1nccc1 | -9.828 | -75.350 | 7.13 |
| 0S01QK.17344c1nccc1 | -9.824 | -77.017 | 7.09 |

**SI Table 3.2** (continued).

| | | | |
|---|---|---|---|
| RJISCZ.078837C1CNCNC1 | -9.813 | -73.663 | 7.05 |
| 1QKX75.06573c1nccc1 | -9.804 | -72.077 | 7.11 |
| 1QKX75.06583c1nccc1 | -9.804 | -65.888 | 7.14 |
| 1QK5SC.19711c1nccc1 | -9.769 | -73.907 | 7.21 |
| 404PVB.04256c1ncc2nncc2n1_part1_list1 | -9.755 | -65.144 | 7.02 |
| RJI5SC.021977C1CNCNC1 | -9.712 | -83.429 | 7.02 |
| RJI1QK.009753C1CNCNC1 | -9.712 | -66.523 | 7.04 |
| NNNI73.01652c1ncn2nccc2n1_part1_list1 | -9.673 | -56.630 | 7.05 |
| RJI5SC.021942C1CNCNC1 | -9.665 | -89.364 | 7.03 |
| RJI5SC.022139C1CNCNC1 | -9.661 | -77.909 | 7.14 |