UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA
(FINAL PROJECT PAPER)

VALIDACIJA PROGRAMSKE OPREME ZA MOLEKULSKO SIDRANJE CmDock (EXTENSIVE VALIDATION OF MOLECULAR DOCKING SOFTWARE CmDock)

IVANA ŠTRBAC

2022
ZAKLJUČNA NALOGA (FINAL PROJECT PAPER)
ŠTRBAC

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga
(Final project paper)

Validacija programske opreme za molekulsko sidranje CmDock

(Extensive validation of molecular docking software CmDock)

Ime in priimek: Ivana Štrbac
Študijski program: Bioinformatika
Mentor: doc. dr. Marko Jukić
Somentor: izr. prof. dr. Jure Pražnikar

Koper, september 2022

Ključna dokumentacijska informacija

Ime in PRIIMEK: Ivana ŠTRBAC

Naslov zaključne naloge: Validacija programske opreme za molekulsko sidranje CmDock

Kraj: Koper

Leto: 2022

Število listov: 34 Število slik: 11 Število tabel: 1

Število referenc: 12

Mentor: doc. dr. Marko Jukić

Somentor: izr. prof. dr Jure Pražnikar

Ključne besede: molekulsko sidranje, CmDock, vaba, receptor, roc krivulja, obogatitev

Izvleček:

V prizadevanju za identifikacijo najprimernejših ligandov poskušajo računalniki napovedati interakcije in izračunati vezavne energije z uporabo programske opreme za molekularno priklop. Cilj projektnega dokumenta je oceniti nedavno ustvarjeno priklopno programsko opremo CmDock.

Receptor adenozina A2a je bil povezan z znanim ligandom. Najprej se konstruira priklopni volumen, nato se določi receptor. Podatki se ustvarijo in dodatno potrdijo s primerjavo načinov vezave pritrjenih ligandov in rezultatov priklopa ligandov v različnih nastavitvah. Navidezni presejalni poskus je bil izveden z aktivnimi ligandi z dopolnjevanjem iz baze podatkov o vabah in končno predvidevanjem novih ligandov iz baze podatkov o vabah. Najboljši in najslabši zasidrani ligandi so bili identificirani s programom PyMol in nato raziskani s spletnim orodjem PLIP. Podobno je vrednost RMSD, pridobljena z dokončanjem poravnave v PyMol, omogočila primerjavo pritrjene konformacije z referenčno konformacijo. Protokol poti je bil ustvarjen s programsko opremo KNIME, ki je omogočila izdelavo roc krivulj, statistično analizo različnih kemijskih značilnosti molekul in rezultate priklopa aktivnih snovi in vab. Izračun faktorja obogatitve za vsak poskus je služil kot zadnji korak vrednotenja.

Glede na vse podatke lahko zaključimo, da se program CmDock odlično obnese. Faktor obogatitve, ki je okoli štiri, je bil približno enak v vseh treh testih. Poleg tega je bila opažena dobra reprodukcija poze. Vendar je bilo tudi ugotovljeno, da rezultati še zdaleč niso obsežni in nadaljnje priporočilo bi bilo oceniti celotno ciljno bazo podatkov iz DUD-E z uporabo programa CmDock.

Key document information

Name and SURNAME: Ivana ŠTRBAC

Title of the final project paper: Extensive validation of molecular docking software CmDock

Place: Koper

Year: 2022

Number of pages: 34 Number of figures: 11 Number of tables: 1

Number of references: 12

Mentor: Assist. Prof. Marko Jukić PhD

Co-Mentor: Assoc. Prof. Jure Pražnikar PhD

Keywords: molecular docking, CmDock, ligand, decoy, receptor, roc curve, enrichment

Abstract:
In an effort to identify the best-fitting ligands, the computers attempt to predict interactions and compute binding energies using molecular docking software. The project paper's objective is to evaluate the recently created docking software CmDock.
The adenosine A2a receptor was docked with the known ligand. The docking volume is constructed first, then the receptor is specified. Data is generated and further validated by comparing the docked ligands' binding modes and docking scores of the ligands in various setups. A virtual screening experiment was conducted with active ligands by supplementing from a database of decoys, and ultimately predicting new ligands from the decoy database. The best and worst docked ligands were identified using the PyMol program and then investigated using the PLIP web tool. Similarly, the RMSD value acquired by completing an alignment in PyMol permitted comparison of the docked conformation with the reference conformation. The pathway protocol was created using KNIME software, which allowed for the production of roc curves, statistical analysis of various chemical characteristics of molecules, and docking scores of actives and decoys. The calculation of the enrichment factor for each experiment served as the evaluation's last step.
Considering all of the data, we can conclude that the program CmDock performs perfectly. The enrichment factor, which is around four, was roughly the same across all three tests. Aside from that, good pose reproduction was observed. The results, however, were also determined to be far from extensive, and a further recommendation would be to evaluate the whole target database from DUD-E using the program CmDock.

ACKNOWLEDGEMENTS

Without Dr. Marko Jukić, who served as my mentor and was an inexhaustible source of patience, helpful advice, and flawless guidance, this paper would not have been possible. Acknowledgments are expressed to bioinformatics program coordinator and Docent Matjaž Hladnik at the University of Primorska for his support with this final project article. I also want to offer sincere gratitude to Dr. Jure Pražnikar for helping me with my final thesis and accepting to be my project comentor. Last but not least, my best friend and my family are my strongest supporters and the wind at my back.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

RMSD	Root mean square distance
ROC	Receiver operator characteristic
VS	Virtual screening
IVS	Inverse virtual screening
HTS	High throughput screening
LBDD	Ligand-based drug design
SBDD	Structure-based drug design
FPR	False positive rate
TPR	True positive rate
PDB	Protein data bank
DUD	Directory of useful decoys
NMR	Nuclear magnet resonance
TTD	Therapeutic target database
PDTD	Potential drug target database
PLIP	Protein Ligan Interaction Profiler
sc-PDB	Screening PDB
SMR	molecule refractivity
TPSA	topological polar surface area
AMW	average molecular weight
SlogP	log of water partition coefficient

1 INTRODUCTION

Identification of new ligands for protein or nucleic acid binding sites is usually done with software for molecular docking. The programs try to anticipate interactions and compute binding energies in order to find the best matching ligands [1]. The aim of the project paper is to analyze a newly developed docking software CmDock. The software is a collection of optimizations and updates on an open-source program called RiboDock.

The known ligand was docked into a receptor, adenosine A2a. Data is generated and further validated by comparing the docked ligands' binding modes and docking scores of the ligands in various setups. A virtual screening experiment was conducted with active ligands by supplementing from a database of decoys, and ultimately predicting new ligands from the decoy database. For the evaluation roc curves were created as well as the path protocol in Knime software..

1.1 Molecular docking

Drug research and development has a lengthy history, and it began as a random quest for chemicals with therapeutic characteristics. This strategy is increasingly being used to replace the focused rational design of active components based on a thorough understanding of a biological target's structure and function [1]. The geometric and electrostatic features of such a ligand, which is generally a tiny molecule, and its binding site on a target biological macromolecule, usually a protein, can be thought of as favoring complementarity in the molecular context of physiologically active substances [10]. A cyclic process is typically used to describe the process of rational drug design. It designs and shifts novel physiologically active compounds for hits using structural data or data on existing ligands [3, 9]. The biological activity of selected molecules is then assessed in numerous experiments, and in the best-case scenario, fresh structural data on their binding is obtained [1, 3]. The fresh cycle of optimization of the first molecules based on this information begins, both in terms of biological activity and physicochemical properties. A molecule with the required qualities, a preclinical candidate, can be created after numerous iterations and then undergo preclinical and eventually clinical testing for the active component [10].

Molecular docking is a computer-aided design approach that calculates the preferred confirmation of a selected molecule, generally, a tiny ligand, but also macromolecules, at a specified active site of a biological macromolecule (target), assuming they form a stable complex [1]. There are two types of algorithms in molecular anchoring programs. Search algorithms take care of the formation of ever-new binding conformations, it calculates the

possible orientations of the ligand in the active site of the target, the evaluation function evaluates the benefit of intermolecular interactions, and the scoring function calculates the corresponding affinity energy of the ligand, which represents the estimated intensity of interactions with the macromolecule [10]. The two methods complement each other during the computation of the complex geometry, therefore the estimate function also acts as a quantitative assessment of which conformations will be passed to the next phase in the anchoring process [9].

While the macromolecule is static, most molecular anchoring algorithms, such as AutoDock, e-HITS, DOCK, Glide, GOLD, and others, handle the ligand dynamically and modify its conformation while computing the complex conformation [10]. This unfortunately leads to inaccuracies in calculating binding poses, which some systems try to minimize at least in part by taking into account the flexibility of the side chains or a portion of the active site. As a result, before solving the molecule libraries, the molecular anchoring software is evaluated to see under what conditions it effectively reproduces the binding conformations of ligands for which we already have experimental data, which is the crystal structure of the complex [9, 10]. The majority of molecular anchoring algorithms that have been validated have been shown to correctly mimic empirically verified ligand binding geometries. To construct a molecular complex, molecular anchoring programs usually employ two search techniques. The first is based on the idea of incremental construction, in which a bigger fragment of a molecule is attached in the active site and then developed the conformation of the molecule in the active site using various limiting constraints, either energetic or geometric [1, 4]. Genetic algorithms, which strive to simulate the processes of evolution and selection, such as mutations, cross-overs, and so on, while developing new conformations, are another prominent search approach.

The affinity of the generated ligand-protein complex is determined in the second phase of molecular anchoring [3]. A non-covalent and sometimes covalently bound supramolecular complex can be formed when a small molecule and its macromolecular partner interact. However, only noncovalent bonds are supported by molecular anchoring at the molecular mechanics level. The relationship between the quantities of G, which is a change in Gibbs free energy, and Kd, which is the binding constant, is based on thermodynamics and illustrates the relationship between the amounts of ΔG, which is a change in Gibbs free energy [3, 10].

$$\Delta G = -RT\ln Kd!$$

ΔG - the difference between free energy in the binding G and in the non-binding state G
Kd - chemical equilibrium constant for the decay reaction of the ligand-receptor complex

A series of affinity-ranked computed ligand/macromolecule complexes is ultimately the final outcome of molecular anchoring.

1.2 Virtual screening

Virtual screening (VS) refers to the computational chemistry approaches employed to assist in the design and discovery of novel physiologically active compounds. The method is described as the computer-assisted exploration of virtual molecule libraries with the goal of limiting the chemical space down to a few classes of compounds that are most likely active on the target under inquiry [4]. It is comparable to the experimental version, high-throughput screening (HTS). The difference is that the chemicals in a virtual screening library are stored in databases rather than being physically constructed by the screening organization. Large numbers of molecules are computationally screened in a matter of days using the information about the target protein based on a crystal structure, homology models, or a predicted protein structure, and knowledge of ligands that may bind to the target. This is based on docking into the anticipated binding pocket of interest in 2D and 3D shape and charge [4, 6].

Building a library of 3D molecules before executing a virtual solution is a crucial step, in which each molecule is characterized by the right conformation. As a result, 3D structure generators are employed to create meaningful 3D representations of molecules [1]. The chemical relevance of the acquired conformations, the validity of protonation patterns, aromaticity, stereochemistry, and the molecule's probable tautomeric forms must all be carefully evaluated. The success rate of virtual solutions is much lower when virtual libraries are badly and carelessly constructed. Real or virtual compound libraries are used to find potentially intriguing compounds [1, 2]. When designing and developing virtual libraries, it's important to make sure the compounds can be synthesized.

The boundary requirements that the virtual molecule must fulfill in order to be identified as a possible hit in the solution process are defined in the second phase of virtual solving [2]. There are two types of virtual solution approaches that we are familiar with. Ligand-Based Drug Design (LBDD) or indirect drug design, relies on knowledge of other molecules, known active ligands, that bind to the biological target of interest [3]. These additional molecules can be utilized to create a pharmacophore model, which specifies the structural features that a molecule must have in order to bind to the target. Structure-Based DrugDesign relies on knowledge of the 3D image of the biological target or ligand-target complex [1, 3].

After the virtual screening procedure is completed, the findings are assessed, and compounds from the studied library are chosen, which are then tested for biological activity on the target of interest. The Virtual Screening process is iterative and part of the drug design cycle [6].

1.3 Inverse molecular docking

Inverse docking entails docking a single small-molecule ligand into the possible binding cavities of a collection of therapeutically important macromolecular targets, allowing for early prediction of a drug's side effects and toxicity [2]. According to the tightness of binding, detailed evaluations of the binding properties lead to a rating of the targets. The method might lead to the discovery of new molecular targets for the ligand that are important to its mechanism of action and side effect profile [3]. Another area where reverse docking might be useful is during the lead finding and optimization stages of the drug development process. In the Inverse Virtual Screening (IVS) method, a molecular docking process is employed to screen a protein database for a query ligand, and then an enriched subset including probable ligand targets is supplied [3, 9]. A protein database and a molecular docking tool are both required to conduct a docking-based IVS analysis. The target database consists of a collection of protein structures or active sites. An ideal target database for docking-based IVS may be built using the rapidly increasing number of structures deposited in the Protein Data Bank (PDB) [9]. Homology modeling approaches can also be used to expand the target database. Then, using docking software, a potentially intriguing small molecule is docked to each piece of the target database.

There are various computational approaches for target identification besides docking-based IVS, such as ligand-based methods, binding site comparisons, protein-ligand interaction fingerprints, and so on. The molecular similarity concept asserts that compounds with similar structures have comparable biological actions [9]. Ligand-based approaches are built on this notion. These approaches rely significantly on prior knowledge of the molecules in the database and need the creation of a database of small compounds with known binding sites [3, 9]. Despite the fact that ligand-based approaches are frequently utilized for target identification and have had a lot of success, they are completely worthless for dissimilar ligands. Similarly, at least one protein-ligand complex structure of the query small molecule is required for the procedures of binding site comparison and protein-ligand interaction finger-printing [9]. All of the aforementioned IVS methods are characterized as “knowledge-based”. Docking-based IVS, on the other hand, is the sole technique that does not require such background data, making it a more appealing alternative in the field of target identification [9]. A docking engine is used in docking-based IVS to dock a given small molecule to the binding site of each protein in a target

database. The binding scores generated by a scoring function are then used to rank target proteins [3, 9].

1.4 Receptor Adenosine A2A

Adenosine A2a receptor belongs to a purinergic class, a family of plasma membrane molecules found in mammalian tissues, G protein-coupled receptors that possess seven transmembrane alpha helices, as well as an extracellular N-terminus and an intracellular Cterminus form, a large group of evolutionarily related proteins, which detect molecules outside the cell and activate cellular responses and are located on a cell surface, with adenosine as endogenous ligand [11]. Its gene name is ADORA2A. The receptor structure shows that it is an alpha helix that is located close to the membrane. It has two ligand binding pockets, allosteric and orthosteric. The gene encodes a protein which is one of several receptor subtypes for adenosine. The activity of the encoded protein is mediated by G proteins which activate adenylyl cyclase, which induce synthesis of intracellular cAMP [11]. The receptor is believed to regulate myocardial oxygen demand and increase coronary circulation by vasodilation and can suppress immune cells, thereby protecting tissue from inflammation [11, 12]. It is also expressed in the brain, where it has important roles in the regulation of glutamate and dopamine release, making it a potential therapeutic target for the treatment of conditions such as insomnia, pain, depression, and Parkinson's disease [11, 12].

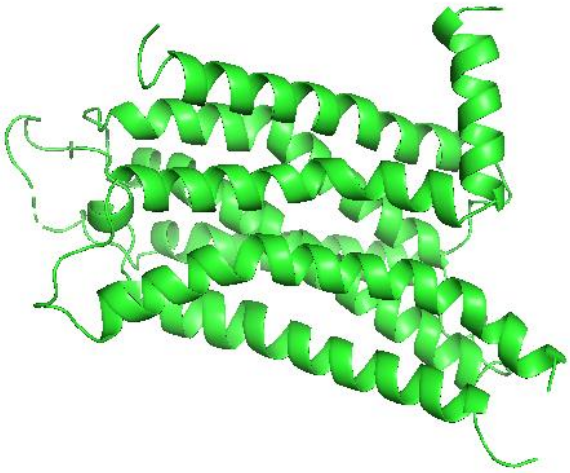

Figure 1: Visualisation of receptor molecule

1.5 Molecular docking testing

Generally, a docking program consists of two main components, the sampling algorithm, and the scoring function [3]. Scoring functions are mathematical operations used to

roughly estimate the binding affinity of two molecules or ligand positions relative to one another after they have been docked [2,3]. The equations of statistical mechanics provide the theoretical foundation for the construction of scoring functions, which make predicting the binding constant of molecules faster and easier than using a fully physically based approach. Statistical thermodynamics is a branch of physical chemistry that provides for the physical explanation of a system's thermodynamic characteristics based on the behavior of the particles that make up the molecular system under consideration [3]. Each of the known scoring functions has its own set of benefits and drawbacks, and none presently matches all of the criteria for an ideal function that is intended to be dependable, accurate, robust, globally applicable, and physically grounded [2, 3]. Only if the physical forces involved in the contact are properly defined and the three-dimensional structure of the receptor is known are scoring functions considered successful. In order to anticipate the intensity of intermolecular interactions between two proteins or between a protein and DNA, scoring methods have also been devised. In order to parametrize scoring functions a data set containing experimentally discovered binding affinities between molecular species that are comparable to the species that one seeks to predict is typically used [6].

The scoring criteria can be based on four classes: force-field, empirical, knowledge-based, and machine learning. When the force-field scoring function is used the interaction is estimated as the sum of unbound van der Waals and electrostatic components. These estimate functions make use of well-known empirical fields of MM forces, such as AMBER in CHARMM [3]. The function is often constructed using one of two techniques. The first tries to establish a thermodynamic cycle that explains the development of a complex between a protein and a ligand in a consistent manner, which takes a long period. The second technique assumes that the protein-ligand structure determined by protein crystallography is the same as the equilibrium complex structure [2, 3]. The interaction energy is then determined using the protein-ligand complex's 3D coordinates. The algorithm determines the separate contributions of the interaction energy, which may then be combined together to obtain an estimate of the overall interaction energy due to the additivity of the thermodynamic functions. The master equation was born as a result of this process. Experimentally determined protein-ligand complexes are used to create empirical scoring functions [6]. In this case, the interaction is estimated as a weighted sum of various empirical components such as H-bonds, hydrophobic interactions, and interactions with metals. Its greatest benefit is speed since it can calculate in a fraction of a second. However, the function’s flaw is that it is unclear to what degree they can be utilized to forecast the binding constants of ligands with structurally different properties than those used to determine the weighting factors. Experimentally established 3D structures of ligand-receptor complexes are used to generate knowledge-based scoring functions [3]. The background idea is that interatomic lengths are better than other distances. Simple

potentials of atomic contacts are found by statistically evaluating the distances between the ligand and protein atoms, which are then used to determine the total interaction. It's fairly usual to combine different valuation algorithms, such as consensus scoring, to produce superior results.

Scoring in docking is really made up of three separate expressions: produced configurations, ranking various ligands against proteins, and one or more ligands ranking against various proteins by their binding affinities [6]. Consensus scoring functions could not greatly increase accuracy since diverse scoring functions are generally co-linear. In a perfect scoring function, the binding free energy between the ligand and its target may be predicted. However, in practice, this objective is constrained by both computing techniques and computational resources [6]. The causes of this scoring issue have been attributed to protein flexibility and induced-fit upon ligand binding, oversimplification of energy terms commonly used in the scoring functions, such as solvation and entropy contributions, and specific non-covalent interactions, such as weak hydrogen bonds [6]. Before the research can advance further, it must first overcome the obstacles of predicting the ligands' most stable tautomeric states, water-mediated hydrogen bonding, metal coordination, and discovery of pKa-based protonation micro species. Therefore, techniques that reduce the amount of false positive and false negative ligands are often used.

1.6 ROC curve

A Receiver Operator Characteristic (ROC) curve is a graphical representation of a binary classifier's diagnostic capacity [4]. Its origins are in signal detection theory, but it is currently employed in a variety of fields including medical, radiography, natural disasters, and machine learning. Plotting the true positive rate (TPR) against the false positive rate (FPR) generates the curve. The true positive rate is the fraction of all positive observations that were properly expected to be positive, and the false-positive rate is the fraction of negative observations that are mistakenly predicted to be positive [4]. The trade-off between sensitivity or TPR and specificity, which we get when we subtract FPR from one, is depicted by the ROC curve. Classifiers with curves that are closer to the top-left corner perform better. A random classifier is supposed to offer points that are diagonal as a baseline. Diagonal represents the equal values of FPR and TPR [4, 6]. The test becomes less accurate when the curve approaches the ROC space's 45-degree diagonal. It might be helpful to summarize the performance of each classifier into a single metric when comparing multiple classifiers. One popular metric for measuring docking performance is the area under the curve (AUC). It's the chance that a randomly selected positive instance will be ranked higher than a randomly selected negative instance [4, 6]. ROC plots, on the other hand, frequently utilize a semilog modification of the x-axis to focus on early changes.

1.7 Enrichment factor

The ratio of the annotated ligand concentration in the top-scoring docking hits to the entire database is known as the enrichment factor. It is a method that is frequently used to assess the effectiveness of molecular docking experiments [3, 4]. It is, essentially, just a measurement of the number of additional actives we discover inside a certain "early recognition" proportion of the sorted list in comparison to a random distribution [8]. The following formula may be used to compute the enrichment factors:

$$E_f = \frac{E_{experimental}}{E_{active}} 100$$

$E_{experimental}$ - *the number of experimentally found active structures in the top x% of the sorted database*

E_{active} - *the total number of active structures in database*

1.8 Decoy compunds

Decoys act as natural moderators of powerful reactions including inflammation, cell proliferation stimulation, and apoptosis [4]. They are definitely an important part of the cell's regulatory system, since they play a role in fine-tuning cell proliferation and death, and can be used to generate recombinant anti-toxins, antiviral medicines, and new therapies for cancer and inflammatory illness. They also help viruses evade the immune system [8]. The underlying notion of recombinant soluble receptors as effective treatments existed before the discovery of their natural homologs. Toxin decoy neutralization looks to be quite successful and should result in actual biologic antidotes. Antiviral prospects appear to be more complicated, as the desire for sterilizing efficiency now exceeds the potency of existing decoys [3, 8]. To make this form of therapy effective, decoys' affinity for their cognate viruses may need to be significantly increased. Initially, they were hypothesized as new biologics, envisioned as receptor mimetics that would catch and detain a pathogen in solution before it could interact with its cellular target [4]. However, as gene cloning, recombinant protein production, and genomics progressed, it became clear that manufactured decoys were trailing behind what turned out to be nature's primary mode of operation [3, 4]. The capacity of virtual docking algorithms to distinguish active ligands from decoy molecules is frequently used to assess them.

1.9 DUD-E database

By adding decoys from a library of synthetically possible compounds that are physically similar but chemically distinct to the active ligands, the Directory of Useful Decoys (DUD) reduces bias and ensures that enrichment is more than just the separation of superficial characteristics which makes it less probable that they are binders [4]. The ligands were occasionally overrepresented in a few chemotypes, whereas the decoys were occasionally ligands rather than decoys, which led to significant challenges when initially chosen for DUD [2, 4]. Furthermore, the released collection has lost the mapping of certain ligands to their corresponding decoys. The ligand variety in each of the 102 ligand and decoy combinations that make up DUD-E is significantly increased, eliminating the bias that might arise from a single chemotype scoring well. DUD-E enables more representative docking screen testing thanks to at least 40 ligands for each target and a preference for maximizing chemotype diversity [7]. In turn, property-matching decoys to each ligand individually, while deleting fake decoys allows scientists to directly match particular ligands to their decoy molecules, reducing what had previously been artifactually low enrichment for some targets in DUD [8]. Adding net charge as a characteristic to match between ligands and decoys in DUD addresses a disparity between them, where the ligands tended to be more charged on average than the decoys, skewing our estimate of physical forces like desolvation [7]. Clustering the ligands for diversity decreases enrichment in isolation from other effects, as one might predict because high-performing, overrepresented sets have been mostly eliminated. In contrast to the DUD performance, the new decoys boost enrichment [4,7]. This looked odd at first, because a better-balanced, more strict decoy set would seem to present a bigger challenge to docking software. However, the elimination of what had been bogus decoys (ligands), artifactually lowered enrichment in DUD since, as ligands, they had typically scored well, but when counted as decoys, they diluted the annotated ligands, artifactually reducing enrichment in DUD [7, 8]. Ultimately, the enhanced enrichment in DUD-E should give better sensitivity for evaluating docking techniques, allowing it to respond more quickly to changes that reduce and increase enrichment [7].

1.10 RMSD

A typically used metric of similarity between two protein structures is the root mean square deviation (RMSD) between the corresponding atom orientation of two protein chains. The lower the difference in RMSD between two structures, the more similar they are [6]. The RMSD between predicted and experimental structures is required in protein structure prediction before a prediction may be judged successful. Only when the RMSD is as tiny as that of closely homologous proteins (< 3Å) is success evident [6]. In the more general

scenario, one must compare the native structure not just with the predicted one, but also with randomly picked protein-like folds to determine the quality of the prediction [6].

1.11 Ligand's crystallographic pose

Many possible poses are generated by docking algorithms. Each “snapshot” of the pair is referred to as a pose. Some can be discarded right away because they interfere with the protein [1]. The remaining are ranked one ligand relative to another and assessed using a scoring method that accepts a posture as input and provides a number indicating the chance that the pose reflects a good binding interaction. On the other side, in crystallography, the orderly configuration of atoms, ions, or molecules in a crystalline substance is referred to as the crystal structure [6]. Because of the intrinsic properties of the component particles, ordered structures develop in the matter to produce symmetrical patterns that repeat along the major axes of three-dimensional space. The unit cell of the structure is the smallest collection of particles in the substance that makes up this repeating pattern. The crystal's symmetry and structure are entirely reflected in the unit cell, which is created by repeatedly translating the unit cell along its main axes. Many physical features, including cleavage, electronic band structure, and optical transparency, are strongly influenced by the crystal structure and symmetry [6, 8].

1.12 X-Ray crystallography

Crystallography is a technique for determining the three-dimensional arrangement of atoms within a crystal that has been irradiated with x-rays. Rays disperse in certain directions when they interact with the crystal, depending on the crystal's structure and the molecule it contains [10]. Crystallography may determine a three-dimensional picture based on the known electron density of the protein in the crystal by understanding the angles of scattered rays and their phases. It has the benefit of being able to determine the structure of nearly any size biological molecule and providing information on how ligands attach to their targets.

A solution of pure protein in its original physiological form is required to begin protein crystallography. There are two ways to obtain protein. It may be separated from the source, which is generally insufficient for obtaining enough protein [10]. To obtain higher levels of protein, the protein gene is usually introduced into an appropriate expression system, such as bacteria or fungus. Purity and homogeneity, protein solubility, monodispersity, functional activity, and stability are all characteristics used to assess the resulting protein sample. The signal is amplified by an organized crystal lattice of molecules in roughly the same conformation. The diffraction picture will be poor if the protein is not evenly

organized in the crystal, and no relevant data will be gathered from it to establish the structure. Proteins may form crystals under particular circumstances, despite the fact that they are huge biological molecules [3]. The essential principle behind the purification of a protein is that the sample is first brought to supersaturation of the solution by adding precipitants under crystallization conditions which causes the production of the crystallization nuclei, which leads to more spontaneous nucleation and development of the protein crystal. Temperature, pH, precipitant concentrations, and organic solvents are all used to induce supersaturation. Protein crystallization frequently employs the hanging drop technique in which a stock solution is used as the crystallization buffer [3, 10]. The stage that is most difficult to forecast in advance is determining the appropriate conditions for crystallization. The quantity of information that may be collected in all methods of microscopy is restricted by the wavelength of the electromagnetic wave employed.

1.13 PDB database

PDB stands for Protein Data Bank, and it's a freely accessible database (*www.rcsb.org/*) of three-dimensional structural data for biological macromolecules. After determining the location of each atom relative to each other in a molecule using methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, all of the generated data must be preserved in order to be useful not only in the present but also in the future study [1]. As a result, they are annotated and publicly put into the repository by the Worldwide Protein Data Bank, wwPDB, which controls and permits the deposition [1]. Scientists are now required to contribute their structural data to the PDB by most major scientific publications and several funding sources. Protein structures deposited in the PDB are used by many other databases. The number of structures in the PDB has expanded at an essentially exponential pace [10]. In the PDB, there are now around 83900 biological macromolecular structures.

The selected required file has the PDB identification number and macromolecule name with the release date, author's name, protein classification, and the instrument used for determining the 3D structure and citation. The selected/searched macromolecules are downloadable (download PDB file), viewable (view PDB file), and view as a 3D file (view in 3D with Jmol). The tool option also has the “RCBS PDB protein comparison tool” to calculate pairwise sequence or structure alignments and “RCBS PDB widgets” a resource for web developers. PDB also has a section called “Deposition”, which has a built-in structures deposit option. Here one can upload the macromolecular data and the uploaded file is validated and released based on publisher policies. The PDB identification number and macromolecule name with the release date, authors name, protein categorization, an instrument used to calculate the 3D structure, and references are all included in the

necessary file. The macromolecules that were searched are downloadable and may be seen as a PDB file or as a 3D structure. The “RCBS PDB protein comparison tool” may be used to generate pairwise sequence or structural alignments, and “RCBS PDB widgets” is a web developer's resource. PDB also contains a “Deposition” section with a built-in structure deposit option.

The ProBiS-Dock Database is a collection of all protein binding sites that may be extracted from the Protein Data Bank (PDB) [5]. It allows for the identification of all proteins with a certain binding site and hence offers a lot of potential in drug development. The Screening-PDB (sc-PDB) is a virtual screening subset of data derived from the PDB. It contains all of the high-resolution crystal structures of protein-ligand complexes using nucleotides, peptides, cofactors, or chemical molecules as ligands [5]. The information regarding the binding sites is included in the known protein-ligand complex structures in the database, which considerably decreases the sample area for docking [5].

The therapeutic target database (TTD) is a collection of proteins and nucleic acids identified as known or prospective therapeutic targets in the literature [9]. The database contains important information such as targeted diseases, route information, and associated drugs/ligands [9]. The TTD database, on the other hand, does not give users 3D structures of the targets, which must be obtained from the PDB database. Another database concentrating on therapeutic targets is the Potential Drug–Target Database (PDTD) which only has protein targets. Users benefit from cleaned 3D structures for both protein and active sites, which reduces the complexity of docking preparation [9].

Homology modeling backs up this trend of growing data, and most pharmaceutically important protein classes have structural coverage across the proteome. For huge collections of proteins, such as the PDB, docking that considers the full protein surface remains difficult [1, 10].

2 ANALYSIS

As previously stated, the primary goal of this study was to assess the performance of molecular docking in the context of the newly created CmDock program. The task is accomplished with the assistance of several software and websites.

The data is initially acquired from the DUD-E website, which includes 102 targets. There was no need to convert the file because all the data required for the study was downloaded in an SDF format. Adenosine A2a, with the target name AA2AR and the PDB number 3eml, is the intended target for the project. This is a typical target, which is why it was chosen for the project experiment. It already has many known ligands, as well as previous experiments and studies that can be utilized for comparison of findings and simpler analysis. For the target in the DUD-E website, there are the data collected and available for three different organisms, human, guinea pig, and mouse, which are all combined in a retrieved folder.

The binaries of the CmDock software are built under Windows, Linux Ubuntu/Fedora/Red Hat, and Raspbian Stretch/Buster. For this project paper, Windows was used. The docking experiment is conducted using the simple commands we write after the software is run in the command prompt.

The command for the receptor is the following:

cmcavity –r ivana_rec_1_aa2ar.prm –W –d

The above command generates docking volume, and the receptor that will be further used in the docking analysis. The output file generates a grid with the receptor's location in it.

The explanation of the first command's parameters:
-r defines a receptor used as an input
-W writes a receptor file as an output
-d dumps grid file

The second command is used to complete the docking using the ligand databases, actives or decoys, or both. The docking experiment is repeated three times, each time using a different genetic algorithm run of 10, 50, and 100 as a parameter. Due to the fact that we are repeating the experiment three times for each dataset, we will thus receive a total of six new files: three for active ligands and three for decoys.

For the molecular docking the command that is used is the below mentioned:

cmdock –i test_5.sdf –p dock.prm –r 5u52_target.prm –n 10 –b 1 –o test_out.sdf

The explanation of the second command's parameters:
-i defines input file (ligand base/reference ligand)
-p defines parameters (internal in CmDock)
-n defines number of runs
-b saves best pose only
-o defines output's file name

The PyMol program allows us to see how the receptor molecule binds to the ligands with the best and worse score poses. Five photos of the five distinct ligands located in the receptor molecule's active binding region are shown below, three of which are thought to be in favorable positions. The best, fifth, and tenth-best poses are included in the figures. The other two figures represent the worst-case scenario, which was the 100th pose in a sorted list of reference docked ligands, and the 90th pose.

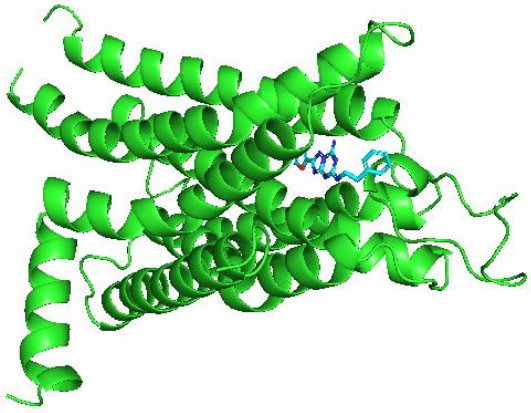

Figure 2: The best scored docked reference ligand bound to the receptor

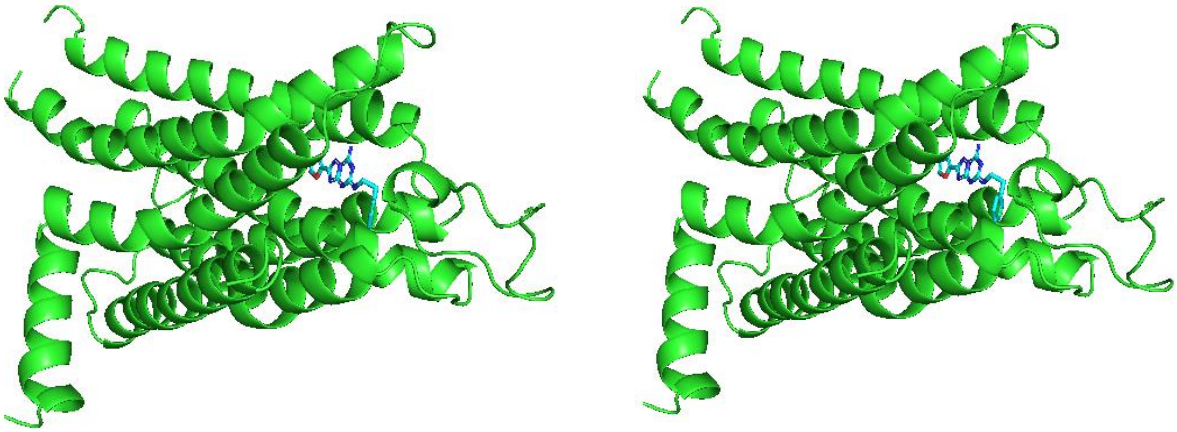

Figure 3: The fifth (left) and tenth (right) best scored docked reference ligands bound to the receptor

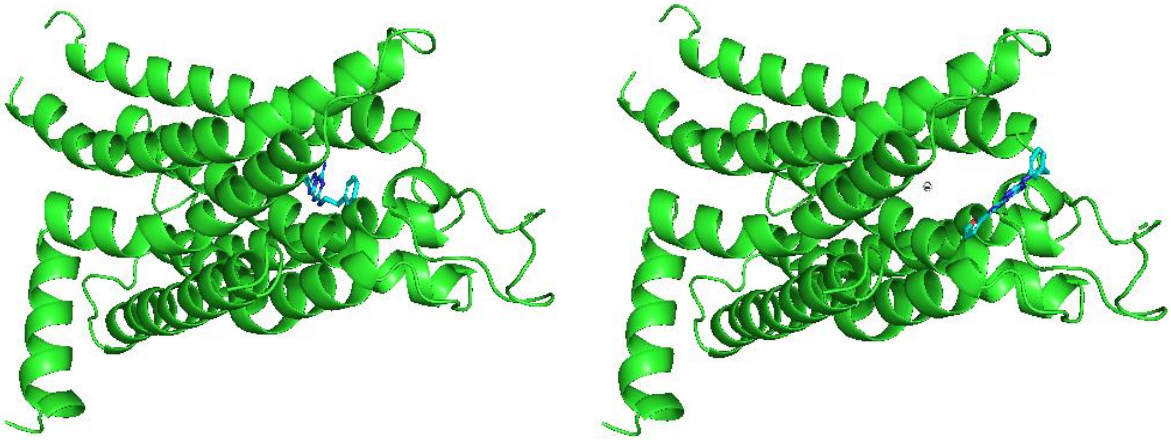

Figure 4: The worst (left)and 90th (right) scored docked reference ligands bound to the receptor

To continue the in-depth investigation of the receptor molecule's interactions with the small ligand, we may utilize the Protein Ligan Interaction Profiler, PLIP, a web tool that readily finds non-covalent interactions between biological macromolecules and their ligands. The highest and worst-scoring ligand-receptor interactions in PyMol software are stored as two new molecules in a PDB format, which are then uploaded to a PLIP web tool for additional analysis. The choice of the atom donor and acceptor was the key factor distinguishing the best poses from the poorest ones. The results gave us identified thirteen interactions, out of which seven are with (3-formyl-3-but-enyl)-phosphonic acid, CYX, four referring to 4-[6,6-dimethyl-4-oxo-3-(trifluoromethyl)-4,5,6,7-tetrahydro-1H-indazol1-yl]-2-[(cis-4-hydroxycyclohexyl)amino]benzamide, HIE ligand, nd1-phosphonohistidine (HIP), and one is unknown. Additionally, we may state that hydrogen

bonds make up the majority of interactions, with a very small number of contacts being hydrophobic.

In addition, several reference ligand docked positions might be discussed while examining the RMSD value. This research was done to see how closely the reference ligand and the docked postures of the same molecule matched up. The RMSD value is obtained by separately storing each position and then using PyMol software to align each pose with the reference ligand one at a time. The poorest reference ligand alignment with the original molecule produced an RMSD value of 3.69, whilst the best one generated a value of 1.96. The alignment of the best and worst reference ligand poses with the original reference ligand molecule is illustrated below in Figure 5. and Figure 6.

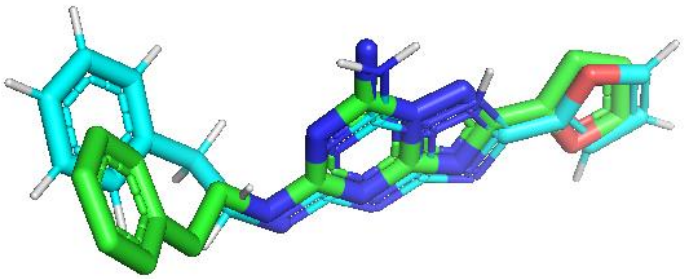

Figure 5: The best docked reference ligand pose aligned with the reference ligand

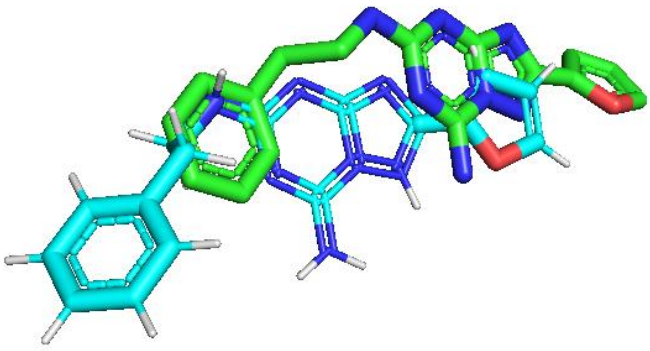

Figure 6: The worst docked reference ligand pose aligned with the reference ligand

The pathway is built using the KNIME software after generating results for both, active ligands and decoys, in several setups, specifically 10, 50, and 100 runs. There were three different settings, therefore it stands to reason that the analysis' last phase would result in three distinct roc curves. There are two inputs for each setup in the pathway, one for active ligands and the other for decoys. The stages in the analysis are the same for both inputs until the node containing the mathematical formula required to create the roc curve is reached. The first node is a file reader for SDF files. Row sampling in a file is accomplished by the addition of the second node, and the third node adds a column with a

constant value. The addition of this column will facilitate further analysis. That is, it is an additional column that, depending on whether it is a decoy or an active, only contains zeros or only ones, respectively. Additionally, two outputs are produced, one that is independent and uses statistics, and the other that dives deeper into the research. Then, the actives and decoys tables are merged. When the two tables are combined, the additional column with a constant value that was inserted previously now has two alternative values, zero or one, and makes it simple to distinguish between a decoy and an active ligand. A mathematical equation is applied to the generated table, which only negates the values of a SCORE.INTER column, that is because of a node, and the output is a roc curve. The SCORE.INTER value, which is a score relating to intermolecular interactions, including those between a small ligand and a large receptor, is employed in this computation. One pathway for the setup of 100 runs is shown below in Figure 1, and the same procedure is done for all three configurations.

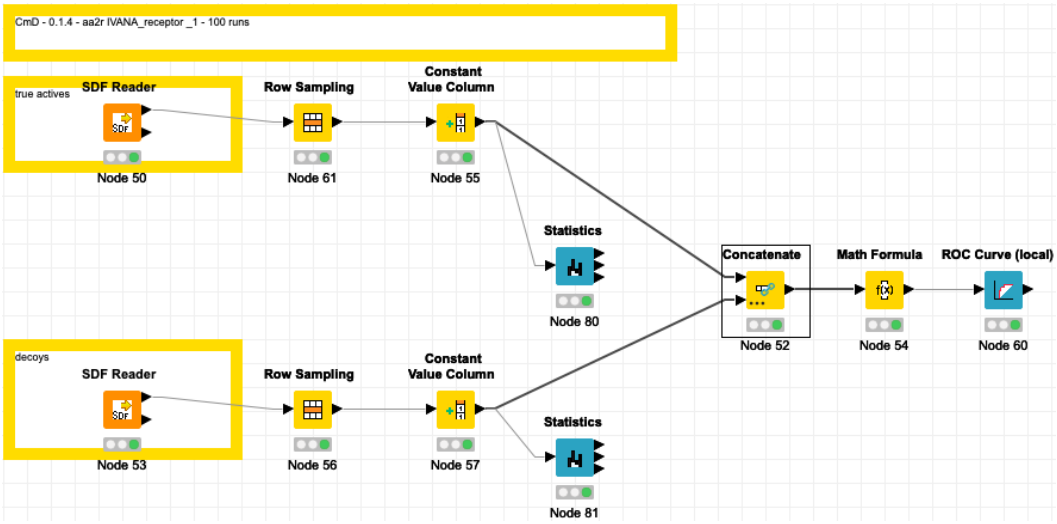

Figure 7: KNIME pathway for 100 runs

As previously stated, the three graphs below depict the roc curves for the three distinct settings. The differences between all of the graphs are not as significant, as can be observed. How we usually select the best score is by looking at the SCORE.INTER value, which should be the lowest possible to be considered as good. By calculating the statistics of the tables for the actives and decoys we can find the minimum and maximum values of the SCORE.INTER. For 100 runs, the minimum value for actives is -30.191 and for decoys -30.227, and the maximum is -10.933 for actives and -5.385 for decoys. For 50 runs, the minimum value for actives is -30.278 and -31.227 for decoys, and the maximum is -10.945 for actives and -4.614 for decoys. For 10 runs, the minimum value for actives is -27.67 and for decoys -30.07, the maximum is -6.319 for actives and 3.183 for decoys. When it comes to the roc curves, they should reflect the same findings as we could see in comparing the above minimum and maximum values. Score plus outcome, which denotes the area under

the curve, is displayed in the bottom right corner of the graph. The better the outcome, the higher the area. The score plus result is 0.7478 for 10 runs, 0.7680 for 50 runs, and 0.7625 for 100 runs. Therefore, the setup with 50 runs yields the best result, whereas the setup with 10 runs yields the worst. In addition, the results regarding the 10 runs were the only ones with having the maximum score positive, and with the greatest difference between a maximum score of decoys and actives, which made the general result the worse, since the decoys actually have a good minimum score, even though it is a little bit smaller than the one for 50 or 100 runs, but the active ligands were further away in the minimum score. However, we cannot say that the score we get for 10 runs is actually bad since it is greater than 0.5, which means that we can distinguish active ligands easily. If the AUC is 0.5, indicating that the roc curve is a diagonal line, it is designated a binary classification since we cannot discriminate between real positive and false positive results. Also, the maximum value for decoys in 50 runs setup was greater, than the maximum value for decoys in 100 runs, but the minimum value was lower, which indicated, as earlier mentioned, that what influences the most is the SCORE.INTER lowest value, which was observed in both actives and decoys in 50 runs setup.

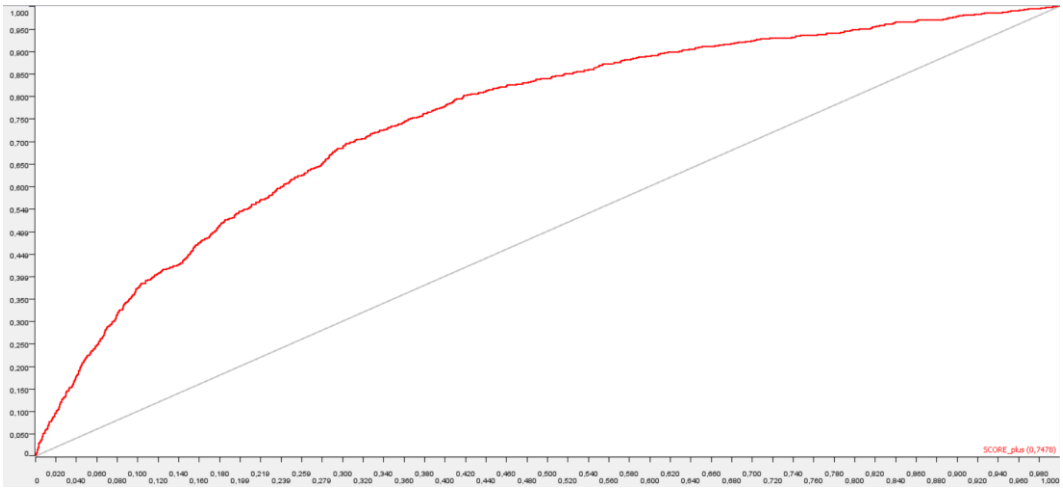

Figure 8: Roc curve for 10 runs

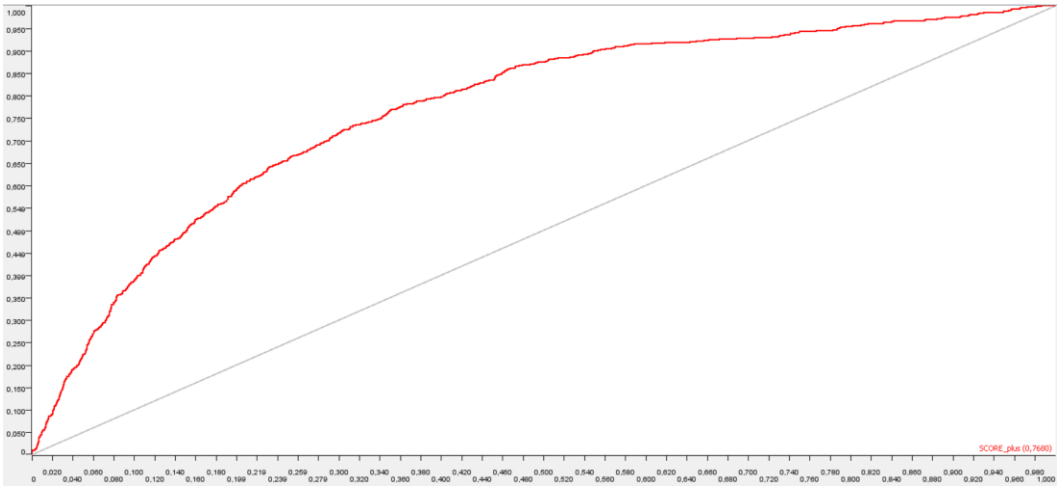

Figure 9: Roc curve for 50 runs

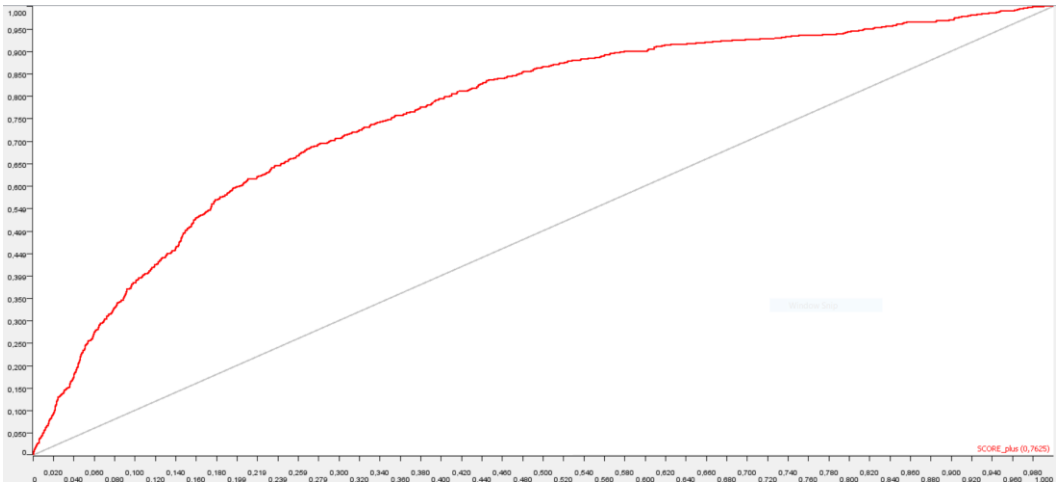

Figure 10: Roc curve for 100 runs

We may once more employ KNIME software by adding more nodes if we want to obtain a deeper knowledge of our small molecules, actives, and decoys. Installing the open-source cheminformatics toolkit KNIME extension RDKit was the first step. A table with an additional column for both decoys and actives outputs the new node as its third output. As a result, new findings regarding compounds like SlogP, SMR, TPSA, and AMW are produced. However, the statistics node should be placed after the RDKit node to provide a summary of the data in a manner similar to previously in order to make the findings more intelligible. The newly created pathway for the setup of 50 runs is shown below in Figure 6, and the same procedure is done for all three configurations.

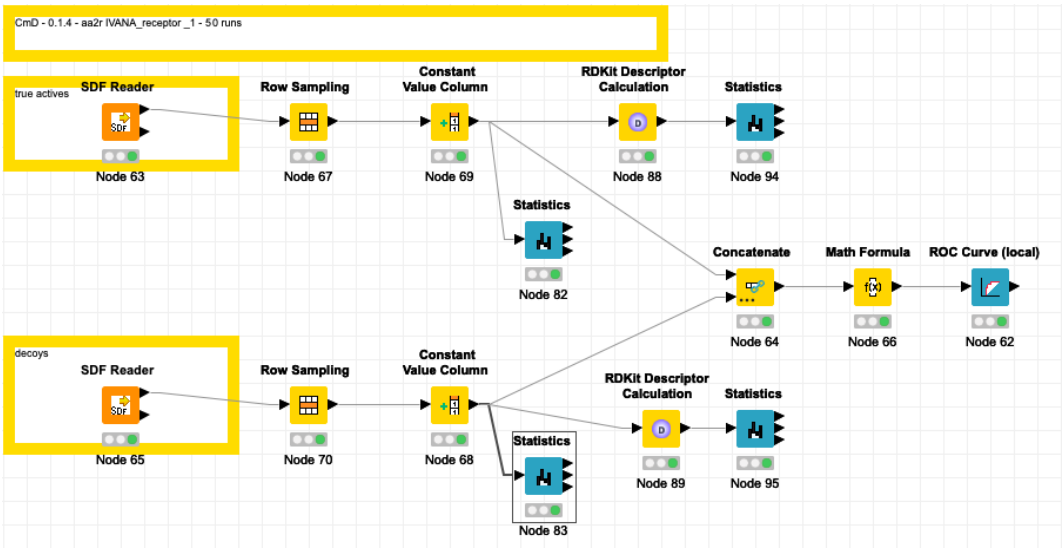

Figure 11: Edited KNIME pathway for 50 runs

Since the analysis is done on the same set of data for decoys and actives, the newly calculated statistics will be the same for all three different parameters used. Therefore we can only at the statistics of RDKit of actives and decoys from only one experiment, and compare the functionalities we are interested in.

Table 1: RDKit statistics (SlogP, SMR, TPSA, and AMW) of actives and decoys

	actives min	actives mean	actives max	decoys min	decoys mean	decoys max
SlogP	-3.6544	1.5233	6.9205	-7.1516	2.3901	7.994
SMR	46.5757	110.3301	167.2104	50.8947	114.5094	165.1797
TPSA	34.89	119.5765	237.54	24.5	92.3471	197.62
AMW	180.167	415.3873	598.688	218.216	431.4678	615.507

SlogP is mostly referred to the solubility, and the negative result indicates preferential solubility in water, while the positive indicates an affinity for octanol. As we can see the active ligands have a higher solubility in water than decoys, when all three values, minimum, maximum, and mean are compared. Next, maybe an even more important parameter is molar refractivity, SMR, which is a real volume of a molecule. This, of course, influences the binding of a ligand and a receptor, and it seems that the set of decoys has a greater volume than actives, which could be one of the reasons why decoys are not a perfect match for our receptor. The topological polar surface area is a sum of the surfaces of all polar atoms in a molecule. Again, decoys also have a greater polar surface area than actives, which implies that decoys are more oppositely charged. The last value we have listed in the table above is the average molecular weight, which is the weight of all chains

divided by the total number of chains. When it comes to the mean value, decoys and actives do not differ a lot, but the maximum value of decoy AMW is more than three times higher than the maximum AMW of actives.

The dataset contains in total 11743 ligands, out of which 844 are active ligands and 10899 are decoys. Therefore it follows that in the whole dataset there is approximately 7.2% of active ligands. The enrichment is calculated for two different cases, 1%, and 2% as well as for all three experiments. The top 1% of the whole database would be 117.43, so 117 first ligands in a database will be considered in further analysis and the top 2% is 234.86, so we will be looking at the 235 first top ligands. When considering the experiment conducted under the parameter of 100 runs, in the top 1% of scores, there are 34 actives, which is 29% of actives, and the enrichment factor is 4. For 50 runs, the results are completely the same as for 100 runs, regarding the top 1% of scores. There are 37 actives in the top 1% in the experiment conducted with 10 runs, which gives the result of the enrichment factor of 4.4 approximately, and the percentage of actives is therefore 31.6%. When completing 100 runs in an experiment, in the top 2% of the scores, there are 64 actives, which is 27%, and the enrichment factor is 7.6. For the parameter of 50 runs in an experiment, there were 62 active ligands identified, which gives the enrichment factor of 7.4, and the percentage of actives would be 26.4%. When considering the experiment conducted under the parameter of 10 runs, in the top 2% of scores, there are 65 actives, which is 27.7% of actives, and the enrichment factor is 7.7.

3 CONCLUSION

Three experiments were performed with three distinct docking parameters, the number of genetic algorithm runs, 10, 50, and 100, as part of the full validation of the CmDock program. The software operates in a stochastic, not systematic, manner, which refers to a random probability distribution or pattern that may be statistically analyzed but may not be accurately anticipated. Therefore, it is advised to experiment with various configurations as better results are typically produced with more runs. However, since the volume of the receptor active site is mostly responsible for this, it is likely that it was substantially smaller for this particular molecule, leading to findings that were essentially high for each of the three criteria employed. Therefore, we can conclude that the software works perfectly with even a small number of runs and it can easily identify active binding sites between small ligands and receptors as big molecules. On the other side, if another receptor with a higher volume would be used in the experiment, it would be expected for a roc curve to grow even more, and the AUC would be bigger, and in that case, the greater number of runs would give the higher roc curve as an outcome.

By contrasting the outcomes with the known, it was simple to verify that the experiment demonstrated how well the program performs. However, it is not believed that one receptor and three executed tests provide enough analysis for the program to be validated. Therefore, in order to confirm the accuracy of the CmDock program, it is advised that a further study should be performed using more diverse receptors and targets.

4 DALJŠI POVZETEK V SLOVENSKEM JEZIKU

Običajno se novi ligandi za vezavna mesta na beljakovinah ali nukleinskih kislinah odkrijejo s programsko opremo za molekularno priklop. V prizadevanju za identifikacijo najprimernejših ligandov poskušajo računalniki napovedati interakcije in izračunati vezne energije. Cilj projektnega dokumenta je oceniti nedavno ustvarjeno priklopno programsko opremo CmDock.

Programska oprema je niz izboljšav in posodobitev odprtokodnega programa RiboDock. Vsi podatki so povzeti s spletne strani DUD-E. Receptor adenozina A2a je bil povezan z znanim ligandom. Dva glavna ukaza sta bila uporabljena v programski opremi za priklop. Najprej se sestavi priklopni volumen, nato se določi receptor in shrani v novo datoteko, ki se nato uporabi v drugem ukazu, v katerem se uporabi podatkovna baza ligandov in definirajo eksperimentalni parametri. Ta ukaz se večkrat uporabi za dokončanje številnih poskusov z različnim številom izvajanj. Podatki se ustvarijo in dodatno potrdijo s primerjavo načinov vezave pritrjenih ligandov in rezultatov priklopa ligandov v različnih nastavitvah. Navidezni presejalni poskus je bil izveden z aktivnimi ligandi z dopolnjevanjem iz baze podatkov o vabah in končno predvidevanjem novih ligandov iz baze podatkov o vabah. Najboljši in najslabši zasidrani ligandi so bili identificirani na aktivnem mestu vezave receptorja s programom PyMol in nato raziskani s spletnim orodjem PLIP. Podobno je vrednost RMSD, pridobljena z dokončanjem poravnave v PyMol, omogočila primerjavo pritrjene konformacije z referenčno konformacijo. Protokol poti je bil ustvarjen s programsko opremo KNIME, ki je omogočila izdelavo roc krivulj, statistično analizo različnih kemijskih značilnosti molekul in rezultate priklopa aktivnih snovi in vab. Izračun faktorja obogatitve za vsak poskus je služil kot zadnji korak vrednotenja.

Glede na vse podatke lahko sklepamo, da se program CmDock odlično obnese, saj je uspel zaznati dejanske pozitivne rezultate že z 10-kratnim eksperimentom, na kar je najverjetneje vplival majhen volumen receptorske molekule. Ker program deluje stohastično in ne sistematično, je kljub temu priporočljivo pogostejše ponavljanje poskusov z različnimi nastavitvami, saj na splošno več ponovitev daje boljše rezultate. Faktor obogatitve, ki je okoli štiri, je bil približno enak v vseh treh testih. Poleg tega je bila opažena dobra reprodukcija poze. Vendar je bilo tudi ugotovljeno, da rezultati še zdaleč niso obsežni, nadaljnje priporočilo pa bi bilo, da se celotno ciljno bazo podatkov iz DUD-E oceni s programom CmDock, da bi jo lahko temeljito pregledali z uporabo različnih molekul, okolij in nastavitev.

5 REFERENCES

[1] Chaudhary KK, Mishra N (2016) *A Review on Molecular Docking: Novel Tool for Drug Discovery*. JSM Chem 4(3): 1029.

[2] Elizabeth Yuriev and Paul A. Ramsland, *Latest developments in molecular docking: 2010–2011 in review*, J. Mol. Recognit. 2013; 26: 215–239

[3] Francesca Stanzione, Ilenia Giangreco, Jason C. Cole, Chapter Four - *Use of molecular docking computational tools in drug discovery*, Editor(s): David R. Witty, Brian Cox, Progress in Medicinal Chemistry, Elsevier, Volume 60, 2021, Pages 273-343

[4] Izhar Wallach and Ryan Lilien, *Virtual Decoy Sets for Molecular Docking Benchmarks*, J. Chem. Inf. Model. 2011, 51, 2, 196–202, January 5, 2011

[5] Janez Konc, Samo Lešnik, Blaž Škrlj, and Dušanka Janežič, *ProBiS-Dock Database: A Web Server and Interactive Web Repository of Small Ligand–Protein Binding Sites for Drug Design*, Journal of Chemical Information and Modeling 2021 61 (8), 4097-4107

[6] Marco A. C. Neves, Maxim Totrov, andRuben Abagyan, *Docking and scoring with ICM: the benchmarking results and strategies for improvement*, J Comput Aided Mol Des (2012) 26:675–686, Springer Science+Business Media B.V. 2012

[7] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet, *Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands andDecoys for Better Benchmarking*, J. Med. Chem. 2012, 55, 14, 6582–6594, June 20, 2012

[8] Niu Huang, Brian K. Shoichet, and John J. Irwin, *Benchmarking Sets for Molecular Docking*, Journal of Medicinal Chemistry, 2006, Vol. 49, No. 23

[9] Xianjin Xu1, Marshal Huang, Xiaoqin Zou, *Docking-based inverse virtual screening: methods, applications, and challenges*, Biophys Rep 2018, 4(1):1–16

[10] Xuan Tao, Yukun Huang, Chong Wang, Fang Chen, Lingling Yang, Li Ling, Zhenming Che and Xianggui Chen, *Recent developments in molecular docking technology applied in food science: a review*, International Journal of Food Science and Technology 2020, 55, 33–45

[11] Mori, A.; Chen, J.-F.; Uchida, S.; Durlach, C.; King, S.M.; Jenner, P. *The Pharmacological Potential of Adenosine A2A Receptor Antagonists for Treating Parkinson's Disease*. Molecules 2022, 27, 2366. https://doi.org/10.3390/ molecules27072366

[12] Merighi, S.; Nigro, M.; Travagli, A.; Pasquini, S.; Borea, P.A.; Varani, K.; Vincenzi, F.; Gessi, S. *A2A Adenosine Receptor: A Possible Therapeutic Target for Alzheimer's Disease by Regulating NLRP3 Inflammasome Activity?* Int. J. Mol. Sci. 2022, 23, 5056. https://doi.org/ 10.3390/ijms23095056