Development of a combined database / in silico pipeline for the investigation of novel approaches for precision medicine

Henry Asomugha

Swansea University
Prifysgol Abertawe

Submitted to Swansea University in fulfilment of the requirements for the Master of Research

Swansea University 2022

Supervisor: Dr Jonathan Mullins

Abstract

This project investigates how integrated *in silico* approaches can be developed to advance precision medicine. A novel bioinformatics database was created and implemented. This database was able to output gene and variant information of all the known human protein drug targets, while also retrieving information on compounds, disorders and drugs associated with each protein. This database was created by writing code, in Linux, that was able to web-scrape multiple databases for information and store it in the pipeline database. The databases that were scraped were UniProt, ClinVar, PubChem, chEMBL, guide to pharmacology, MedGen and the therapeutic target database. These databases were selected as they provided the largest and most reliable data, that could be web-scraped, for each section they were scraped for. Once coded, the full pharmacology set (a set of 720 pharmacologically relevant proteins, whose pharmacological mechanism is known) was added to the pipeline, meaning all their information was downloaded and stored in the pipeline. This bioinformatics pipeline proved to be very effective as an investigative tool for identifying new avenues for personalised medicine as it was able to retrieve and integrate all the requested information on proteins, variants, diseases, and compounds when called upon. In the proof-of-concept study, the database was used to gather key information that allowed for an investigation into the effect of pathogenic variants on drug binding in proteins. This investigation was conducted by simulating the binding of a protein's wild type to two of its known drug ligands. 10 benign variants and 10 pathogenic variants of the protein were also bound to 2 drug ligands associated with the protein; their relative binding energies were collected. This allowed for comparisons to be made between the effect of pathogenic variants and benign variants on a protein's binding ability. Analysis from the docking simulations showed that in 3 of the 5 proteins studied (60%), more pathogenic mutations returned a binding energy with at least a 15% deviation from the wildtype binding energy than benign mutations. These results suggest that the binding interactions of a protein could be affected by polymorphic variation, especially pathogenic variation, although in this case study the difference between the two groups did not show statistical significance.

Table of Content

List of Tables

List of Figures

Acknowledgement

This project was conducted with the help of the Swansea University Genome & Structural Bioinformatics Group Computing Cluster, Dr Karl Austin-Muttitt and Dr Jonathan Mullins.

Abbreviations

B	**BRAF**	**Serine/threonine-protein kinase B-RAF**
C	**CAA**	**Carbonic anhydrases**
	CACNA1S	**Calcium Voltage-Gated Channel Subunit Alpha1 S**
D	**DNA**	**Deoxyribonucleic acid**
E	**EMR**	**Electronic medical records**
F	**FLT3**	**FMS-like tyrosine kinase 3**
G	**GAA**	**Lysosomal acid α-glucosidase**
	GMAF	**Global minor allele frequency**
H	**HPA**	**Hyperphenylalaninemia**
I	**IGNITE**	**Implementing Genomics in Practice**
J	**JAK2**	**Janus kinase 2**
K	**KCNH2**	**Potassium Voltage-Gated Channel Subfamily H Member 2**
	KIT	**Proto-oncogene c-KIT**
P	**PAH**	**Phenylalanine Hydroxylase**
	PATRIC	**PathoSystems Resource Integration Center**
	PDGFRB	**Platelet-derived growth factor receptor beta**
	PLANTS	**Protein-Ligand ANT System**
R	**RMSD**	**Root-mean-square deviation**
	RT-qPCR	**Real-time quantitative PCR**
	RYR1	**Ryanodine receptor**
S	**SMILES**	**Simplified Molecular Input Line Entry System**
V	**VMF**	**Von Willebrand Factor**

Declaration and Statement

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed . [redacted]

Date ....09/05/2022...

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ...[redacted]

Date ....09/05/2022..

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .[redacted]

Date ....09/05/2022..

1) Introduction

1.1) Overview

This project aimed to identify and develop a bioinformatics toolkit for research into precision medicine. Precision medicine is the specific tailoring of clinical treatments to individual patients based on particular genetic mutations and phenotypes. This project also looked to use *in silico* techniques to investigate the effect of single nucleotide variants, on a protein's binding ability with its known drug compounds. The main hypothesis of this project was that the development of a bioinformatic tool that links existing databases associated with the pathogenicity of genetic variation, with existing databases associated with drug targets in humans, databases associated with drug therapies and databases associated with diseases, would allow for the integration of information automatically and efficiently. Allowing research into relevant connections between several large and important datasets to be conducted. By integrating drug, disorder, and variant datasets, simultaneously, advancements in precision medicine will occur. Another hypothesis was that single nucleotide variants in proteins have an adverse effect on the binding between proteins and drug compounds, and therefore single nucleotide polymorphisms carry a lot of danger, as they can severely hinder the effectiveness of drug treatments. This would highlight the requirement for extra research to be conducted into precision medicine. This project also looks at if the use of bioinformatics can advance medicine and drug therapy. To test these hypotheses, case studies using the database platform were performed on the disease Acute Myeloid Leukaemia, the KIT protein and the Ryanodine receptor and then docking case studies were performed for GAA, SCN2A BRAF, KIT, and PAH.

1.2) Polymorphic Variants

The earth is a vast and diverse ecosystem, with many different species of animals. It is predicted that there are approximately 8.7 million eukaryotic species globally. (Mora, Tittensor, Adl, Simpson, &

Worm, 2011). Each species differs both phenotypically and/or genotypically, with some possessing huge differences and some with very little between them, in terms of genetics. Also, within species, there are variations, these inter-species variations can arise in many different ways. These differences can occur via mutations; however, variations can also occur in general within populations, these are called polymorphic variants. There have been many tremendous advancements in the technique of DNA sequencing, and these advancements have allowed for the human genome to be analysed cheaper and faster than ever before. Modern DNA sequencing has allowed for the identification of complex variants, mutations, polymorphisms, and drug responses (Karki, Pandya, Elston, & Ferlini, 2015). There is a notable difference between a mutation and a polymorphic variant. When a genetic variation occurs at an allele frequency of below one per cent, then the variation is considered a mutation, however, if the allele frequency is above one per cent, it is considered a polymorphic variant (Karki, Pandya, Elston, & Ferlini, 2015). Therefore, each polymorphic variant affects at least 77 million people across the globe, with this often-reaching hundreds of millions, even billions. This shows the considerable effect that polymorphisms have on the population of the world. The Proto-oncogene c-KIT (KIT) is an example of how polymorphic variants can have adverse effects on the carrier's health. The KIT plays a role in intracellular signalling, and it is understood that the mutated form of c-Kit plays a crucial role in the occurrence of some cancers (Babaei, Kamalidehghan, Saleem, Huri, & Ahmadipour, 2016). It is thought that variants of KIT are accountable for cancer occurrence in subsets of Acute Myeloid Leukaemia and melanoma. Inhibition of KIT has shown potential as a viable cancer treatment (Babaei, Kamalidehghan, Saleem, Huri, & Ahmadipour, 2016).

Currently, there is a lot of research being conducted on the human genome, this is essential as this can bring about a new era of clinical science. The progression of genomics has led to a greater understanding of DNA and the roles it plays in diseases and disorders. A key example of this occurs with polymorphic variants. A well-documented polymorphic variant occurs in the protein Serine/threonine-protein kinase B-RAF (BRAF), more specifically in the 600th residue. B-Raf proto-oncogene (BRAF) encodes a cytoplasmic serine/threonine kinase, this protein plays an essential role in the regulation of the mitogen-activated protein kinase signal transduction pathway. Mutations in this gene can lead to the constitutive activation of key regulators in cellular processes, and thus, is one of the most important driving factors behind the accelerated growth of cancer cells. This mutation aids the growth and survival of cancer cell signals and is present in multiple different types of cancers (Lung et al., 2020). To investigate the prevalence of BRAF V600E mutations in lung cancer patients of southern Taiwan, a real-time quantitative PCR (RT-qPCR) method was used this technique is highly sensitive and specific. The RT-qPCR technique can detect single-digit copies of mutant DNA. Results showed that the BRAF V600E mutation was present at a low frequency (0.65%, 2/306) in the studied patient group. The investigation concluded that Screening BRAF V600 mutations with the RT-qPCR and V600E-specific immunohistochemistry could help improve detection accuracy (Lung et al., 2020).

The homo sapiens wildtype of the BRAF protein has been sequenced to have a valine amino acid residue on the 600th residue on the amino acid sequence. An activating missense mutation within this residue has been documented to play a key role in the development of melanoma (Ascierto et al., 2012). The most common carcinogenic variant is when the valine residue is mutated to a glutamic acid residue, on the 600th amino acid in the sequence. Although first linked to the development of melanoma it has since been recognised as a key carcinogen in multiple types of tumours (Loo, Khalili, Beuhler, Siddiqi, & Vasef, 2018). In many experiments patients with BRAF V600E-mutated melanoma have been treated with FDA-approved BRAF inhibitors. The results of these trials showed that patients benefited from the use of BRAF inhibitor therapy. In this experiment a BRAF V600E mutation-specific antibody was used for mutation analysis multiple different types of tumours were analysed, including melanoma, colorectal carcinoma, papillary thyroid cancer, hairy cell leukaemia, and Langerhans cell histiocytosis. Although the results showed positive effects of BRAF inhibitor therapy, the research had some limitations. Due to the use of V600E mutation-specific immunohistochemistry, other V600 mutations had been missed.

Another protein with multiple disease-causing variants is the lysosomal acid α-glucosidase (GAA). GAA is a key enzyme in the breakdown of glycogen into glucose. Biological issues begin to arise when glycogen is not broken down properly as this causes it to build up in the body and can lead to many disorders (van der Ploeg & Reuser, 2008). One of the most common disorders associated with GAA deficiency is Pompe disease (also known as glycogen storage disease type 2 or acid maltase deficiency). This is an autosomal recessive disorder. Pompe disease is characterized by progressive muscle hypotonia and loss of motor, respiratory and cardiac functions which leads to respiratory failure (Roig-Zamboni et al., 2017). Due to the severity of GAA gene mutations, there has been much research conducted on the effects of mutation recognition for the gene. Mutation recognition in the GAA gene can provide very positive clinical effects, such as early disease diagnosis and a greater understanding of the genotype-phenotype relationship. In a study, blood samples were collected from patients with Pompe disease and healthy members of three families. The enzymatic activity of GAA was measured. Then, they performed a mutation detection, this was done by using a polymerase chain reaction, followed by direct sequencing of all exons in samples with decreased enzyme activity. Once identified the mutations were then investigated, using bioinformatics tools the possible side effects on the protein product were predicted. Three novel mutations (c.1966-1968delGAG, c.2011-2012delAT and c.1475-1481dupACCCCAC) were identified in the GAA gene. The results of the investigation of these mutations showed that there was a possibility of harmful effects and significant alterations in the protein structure, caused by the identified novel mutations. The researchers concluded that the three novel GAA gene mutations detected in the study could help expand knowledge of the molecular genetic mechanisms of Pompe disease. It can also be used to help with diagnostics (Gharesouran et al., 2020).

Another disease-causing variant that has been investigated is the c.158G>A variant in the PAH gene. A study was conducted to investigate the significance of the variant in patients with hyperphenylalaninemia (HPA). This mutation is known to cause decreased phenylalanine hydroxylase enzyme activity in HPA patients. In this study, seven unrelated Korean patients with HPA genotyped with the c.158G>A variant had their genetic data analysed and the researcher concluded that the variant should be classified as ‘Likely benign’ rather than ‘pathogenic’. The variant was observed to be homozygous in healthy subjects. The researchers also concluded that the variant causes a decreased enzyme activity without leading to the full pathology of phenylketonuria. (Choi et al., 2017). This is yet another example of how greater knowledge of the human genome can advance precision medicine and health care. As the c.158G>A variant in the PAH gene can be used as a genetic biomarker for hyperphenylalaninemia, allowing patients with the c.158G>A variant to be treated accordingly.

There are many scenarios in which accounting for genetic variations can improve outcomes for patients. This is why precision medicine is a research area of immense importance and is expected to only increase in importance over the coming years (Kosorok & Laber, 2019).

1.3) Drug Development and the use of Bioinformatics in its Advancement

Bioinformatics is a relatively new field of science. This interdisciplinary field of science involves molecular biology and genetics, computer science, mathematics, and statistics. Biological and clinical problems are addressed using the collection of large-scale data and statistics. In bioinformatics, biological processes are modelled, and the data from this modelling is then collected and analysed (Gauthier, Vincent, Charette and Derome, 2019). Due to biological modelling being computerised, large amounts of data can be processed at the same time, this means in many cases the use of bioinformatics is more effective and efficient than traditional research methods (Can, 2013). This is a reason why research into more reliable methods for bioinformatic approaches in medicine is essential, as it shows great potential and could revolutionise health care. The advancement of computer science has also aided the development of bioinformatics. This is one of the main concepts underpinning this project, as this project looks into the potential of bioinformatics in terms of advancing our knowledge and understanding of precision medicine.

Drug Development is an integral part of medicine, the creation and production of new methods of combating diseases have been essential to human survival, for many centuries. Thus, the methods by which disease therapies are produced have been changed and improved, as our knowledge of usable tools and biology has increased. The current traditional method of drug discovery consists of a large amount of time-consuming research and intrusive experimentation. This includes target identification, target lead identification, lead optimization, drug characterization, drug formulation, preclinical research, clinical trials, new drug application and then if successful, approval (Deore, Dhumane, Wagh, & Sonawane 2019). Currently, most of this isn't computerized. This has resulted in the pharmaceutical sector's drug development process becoming slow, inefficient, risky, and expensive (Kaitin, 2010). In a study conducted on how long it takes to translate research findings into biological drugs for rheumatoid arthritis. It was found that the average time it took to get from clinical development to clinical use was 11.13 years. The phase of development that was found to be the longest was, moving from basic research to clinical research, this took around 5 years on average (de Oliveira Lupatini, Zimmermann, Barreto & da Silva, 2022). The addition of automated systems of research and bioinformatics to drug development would exponentially increase the speed of drug development and reduce the cost. Using bioinformatic analysis, the drug target identification and drug candidate screening stages can be expeditated, via the use of high-throughput data analysis. This allows for large sets of data to be analysis at the same time, which is much quicker than human data analysis. Bioinformatic analysis can also be used to facilitate the characterization of side effects and predict drug resistance. Our increased knowledge of protein and RNA structures allows us to accurately simulate protein interactions. This coupled with all the structural databases of small molecules and metabolites, currently at our disposal, has made protein-ligand docking simulations a very realistic and informative method of clinical drug target screening (Xia, 2017).

Structural bioinformatics allows for three-dimensional protein structures to be constructed, and the data from these structures can be used to investigate protein disorders or predict secondary and tertiary structures of the protein. Studies have helped bring to light the interaction patterns and functions of proteins. Due to some proteins being very difficult to research clinically, protein modelling offers a more accessible approach for protein investigation in these situations. Currently, protein interaction prediction is not as accurate as experimental results. However, as more and more complex structures are being added to different protein databases, bioinformatic-driven approaches are becoming a good alternative (Sunny, & Jayaraj, 2022). Analysis of protein structures can be aimed at the prediction of protein function and structural alignment. Protein-protein and protein-ligand interactions can also be studied using bioinformatics tools. Molecular docking is an essential bioinformatics methodology. Many molecular docking tools allow for the simulated binding of proteins and ligands (or another protein), and the strength of the bonds formed can be calculated and then analysed. This is very important as the data collected from these programs help researchers answer many biomedical research queries. In recent years protein-protein docking has been used at high-frequency, in drug discovery research. The targeting of protein-protein interactions, in bioinformatics, has allowed for predicting protein-protein interactions and identifying 'hot spot' residues at the protein-protein interface (Huang, 2014). The relative lack of knowledge about binding sites means that the search algorithms and evaluation methods for protein-protein docking are different to the algorithms for protein-ligand docking and thus require more research strategies (Huang, 2014).

Another use of bioinformatics approaches in biomolecular research is the simulation of side-chain substitutions. This is where 3D models of proteins can have certain residues changed. This mimics the effect of mutations on proteins. This is useful as it aids the research into mutations and their effect on protein structure and function. Advancements in computer technology have allowed for the substitution of amino acids in computer-modelled protein structures, to create very accurate protein structures and similar mutated proteins (Smith, Lovell, Burke, Montalvao & Blundell, 2007) (Malathi & Ramaiah, 2018).

1.4) The Development and Advancements of Precision Medicine

Precision Medicine has increasingly emerged over the last decade, due to advancements in technology and medicine, also due to our increased knowledge of the human genome. Precision medicine is the specific tailoring of clinical treatments to individual patients based on their characteristics. This is commonly done using large-scale patient data including lifestyle, genetic, clinical, and biomarker information. This takes a more in-depth look into each patient and therefore has the potential to produce more effective results than traditional medicine (König, Fuchs, Hansen, von Mutius & Kopp, 2017). Precision medicine has the ability to classify (or stratify) individuals into subpopulations that differ in their susceptibility to a particular disease, in the pathology of those diseases, or in the way they react to specific drug treatments (Maier, 2019). This means pre-emptive or reactive therapies can be used for patients for whom the therapies would be effective, and the patients who will see no benefits to the treatments can be directed to alternative treatments, thus reducing costs, avoiding side effects, or even reducing patient mortality (Maier, 2019).

Many applications of precision medicine have been used in the health care system. A current example of this is genetic screening, where genetic screening tests may be conducted before conception to predict the risk of genetic disorders being passed down to the child. Genetic screening is becoming more efficient, this is largely due to advancements in bioinformatics and computational biology. Between weeks 8 and 12 of gestation, prenatal screening can be conducted on the foetus, to screen for trisomy in chromosomes 13, 18 and 21 (all these cause developmental abnormalities). Whole-genome sequencing of the foetus has been performed. At birth, DNA sequencing can be done, which aids in the detection of many disorders. This data can be analysed, using bioinformatics and with the aim of precision medicine, specific and effective treatment can be quickly implemented, which can lead to reduced morbidity and mortality (Ginsburg & Phillips, 2018). The sequencing data can also be stored and later in life, can be used to diagnose various diseases.

With an increase in electronic medical records (EMRs) and bioinformatics systems' ability to contribute to both research and health care, volunteers and patients who agree to provide biological data and share their data have been able to further advance precision medicine research. The data derived from biological data has helped generate new findings (Aronson & Rehm, 2015). A recent example of this is in the COVID-19 pandemic, where researchers collected and analysed EMR data and concluded that there was an increased risk of COVID-19 infection and mortality in people with mental disorders (Wang, Xu, & Volkow, 2021). There have been findings that have been derived from the use of mobile and wearable devices, also family history, and environmental exposures. Clinicians are able to use this data to grow their knowledge base. The assembly of genomic, environmental, and patient-reported data from multiple sources has created a strong basis for precision medicine and its advancement. This combined with the data from other nationals and clinical networks allows for knowledge to be shared, enabling more effective precision medicine systems, worldwide (Ginsburg & Phillips, 2018).

One of the biggest advancements in precision medicine in recent years is Implementing Genomics in Practice (IGNITE). IGNITE was formed in 2013, this was created to investigate the limitations of the clinical implementations of genomics, and to show its real-world application. IGNITE has been able to investigate the integration of EMR and genomic data and aid in the development of point-of-care decision-making tools. IGNITE has also been able to aid in the creation of novel approaches, the projects include the use of genetic markers for disease risk prediction and prevention. An example of this is the use of the ApoLI genetic variant as a marker for kidney disease in African Americans (Sperber et al., 2017). It was also able to use pharmacogenetic data to guide the use of medications.

Precision medicine has had a huge effect on drug development. The development of new drugs is a very long and complex process, it also comes with a high possibility that the drug might not succeed, the emergence of novel bioinformatics approaches has revolutionized methods to tackle the challenges of drug development (Qian, Zhu & Hoshida, 2019). The development of medicine is essential for humans to combat diseases and disorders from diabetes to cancer. Therefore, innovations within drug development are key in the fight against newly emerging diseases. After widespread use and genetic variance within pathogens, all drugs appear to be active for a short period against a particular disease platform (Swain, & Hussain, 2021). Thus, the production of newer drugs is very important. However, production within drug development is a complicated, time-consuming, and resource-consuming process. The introduction of bioinformatics tools is one of the renovated platforms in current drug discovery. The use of cost-effective throughput screening, computerised target identification, and ligand optimization has shortened the drug development process. The growth in bioinformatics has pioneered the newer drug development platform (Swain, & Hussain, 2021). Due to an increase in the development of high-throughput technologies and the collection of biological data, the transition of research discoveries to clinical applications has been accelerated. An early and well-known example is aspirin. Aspirin was initially used as a treatment for analgesia, however, with the use of information from the EHRs of patients and pharmacological analysis, the researchers were able to find the potential of aspirin to treat colorectal cancer. In September 2015 the US Preventive Services Task Force released a recommendation for aspirin to be used in colorectal cancer prevention (Bibbins-Domingo, 2016). Currently, many therapeutics companies have started to integrate gene-expression analysis, and genetic screening systems with bioinformatics software, this has aided the identification of chemical structures with properties of interest for oncology drug discovery. This is one of the promising new approaches to drug development and has highlighted the potential of bioinformatics and its clinical applications (Cha et al., 2017). This also shows why more research into novel approaches for the use of bioinformatics in association with medicine should be conducted.

With more research, much larger and more complex datasets for precision medicine will be created such as individual and longitudinal multi-omics, and direct-to-consumer datasets (Zhou et al., 2019). Integration of multiple biological datasets generated for everyone along with tailored big-data analytic techniques can aid in us achieving an effective universal precision medicine system (Qian, Zhu & Hoshida, 2019).

1.5) The Role of Proteins in Disease

Proteins are large macromolecules; they are made up of multiple long chains of amino acid residues. These chains of amino acids are called polypeptides, all proteins contain polypeptides. Proteins are one of the most important biological molecules in the body, they play an essential role in all the biochemical processes that occur. Proteins are key in processes such as DNA replication, energy conduction, regulating brain function and numerous other processes (Bonetta & Valentino, 2019). The functional and structural differences between proteins arise due to variations in the sequence of amino acids, as this can cause the proteins to fold differently, resulting in the creation of completely different proteins. This structure is dictated by the amino acid residue sequence which, itself is determined by the nucleotide sequence in the person's DNA. Humans have the genetic code to construct 20 different specific amino acids. During or shortly after protein synthesis chemical modifications occur. The modification of these proteins can alter the properties both physically and chemically, this can affect the function of the protein (Arora & Katyal, 2019). Abnormal or incorrectly modified proteins degrade more rapidly than wild-type proteins, due to them being unstable. This can have an adverse effect on the body, as proteins are essential for all processes in the body (Bollong et al., 2018). Many proteins are enzymes, proteins that catalyse a number of biochemical reactions, they are also vital for cell signalling, immune responses, and cell adhesion (Bonetta & Valentino, 2019). Proteins can also bind together and function as a complex, by creating a protein complex they are able to work together to achieve a specific function. Due to variations and mutations in DNA, many people have different amino acid sequences, which can cause the proteins to be irregular and ineffective. Mutations within proteins affect their properties in many ways. A mutation affecting only a single amino acid residue is enough to drastically hinder the function of a protein. This occurs in the protein Ryanodine receptor 1 (RYR1). When specific signals are sent to the receptor, the RYR1 channel releases calcium ions from the sarcoplasmic reticulum into the cell fluid. This causes an increase in the calcium ion concentration in the muscle cells, which simulates muscle contraction (Hernández-Ochoa, Pratt, Lovering, & Schneider, 2016). However, when a mutation occurs in RYR1, it causes an increased susceptibility to hyperthermia and some skeletal muscle disorders. This is because RYR1 mutations can cause uncontrolled calcium ion production, which causes continuous muscle contraction, which increases core temperature and can lead to death (Zhang et al., 2018).

A key example of protein malfunction leading to disease is Acute myeloid leukaemia (AML). AML is a bone marrow disease characterized by the uncontrolled proliferation of myeloid cells (Saultz & Garzon, 2016). In the innate immune system, the Toll-like receptors (TLRs), play a whole role. They are tasked with recognising pathogen-associated molecular patterns (Rybka et al., 2021). TLRs have also been known to play a role in autoimmune diseases and cancer, as they are supposed to recognise endogenous danger-associated molecular patterns. Multiple TLRs have been found to be expressed in AML cells, and their expression was found to affect growth, differentiation, and immunostimulatory capacity in AML (Rybka et al., 2021). When 90 AML patients were analysed, it was detected that there were seven single nucleotide polymorphisms, located across the genes coding for TLR3, TLR4 and TLR9, that were highly present in the patient, and it was concluded that polymorphisms could be implicated in the clinical outcome of AML and is related with increased risk of infectious complications (Rybka et al., 2021).

1.6) Objectives of the Project

This project aimed to develop a bioinformatics tool that could have practical human biology uses, and that could also aid the advancement of precision medicine. This project also looked to explore the effects of mutations on protein structure and protein-ligand binding. This project attempted to highlight the influence that variations in genomics might have on precision medicine approaches, and how these variations that lead to disease-related phenotypes can impact drug binding. The research aimed to show that data collection via bioinformatics tools can be used to identify opportunities for advancements or novel approaches to the uses of precision medicine. The completion of these objectives helped test the main hypothesis, that single nucleotide variants in proteins can have an adverse effect on the binding between proteins and drug compounds, and that when they occur, these differences can be captured and systematically evaluated using the database pipeline implemented.

2) Methods

2.1) Overview

Proteins, variants, compounds, disorders, and treatment information was collected. This was done by searching many different databases, which made it possible to gauge the amount of bioinformatics data available and how the data would be able to be incorporated together. This aided the modelling of the structural effects of variations within proteins. This was conducted using structural modelling pipelines, such as those of the Genome and Structural Bioinformatics Group at Swansea University Medical School and visualised using numerous molecular visualisation tools, such as Chimera (Pettersen et al., 2004). Also, the protein-docking analysis program, Protein-Ligand ANT System (PLANTS) (Exner, Korb & ten Brink, 2009). Proteins from the Full Pharmacology Set have been studied. The full pharmacology set is a subset of the 720 known drug targets listed in DrugBank (Wishart et al., 2017) for which their pharmacological mechanism is known. Analysis was conducted on all the pharmacologically relevant proteins to identify disease-causing variants and situations where a variation has impacted the effectiveness of drug therapy. This allowed for a defined, more concise group of key proteins to be investigated in detail. The detailed investigation into these proteins involved structural modelling of the wild-type and variant forms of the protein. Subsequently, molecular docking of drugs had been carried out on the selected proteins. This has allowed for the exploration of the effects of protein variants on the efficacy of certain drugs to be done.

2.2) Data Collection

The first stage of this project involved manually gathering proteins, variants, compounds, disorders, and treatments information, this was done to gauge the amount of data available and to see how best to gather this information at scale. It also showed how all the data from the different databases could be integrated. In this phase of research, information was taken from many bioinformatics databases and the data was stored in various Excel spreadsheets. All 720 proteins from the Full Pharmacology were put into an excel spreadsheet and then numbered. Then using a random number generator 60 proteins were selected, and an in-depth look into these proteins was conducted. The criteria for which database, would be used to gather each section of data, was based on how much data was present in the database, how reputable the source is and if it would be possible to web-scrape the database. The databases that were accessed were UniProt, ClinVar, PubChem, chEMBL, Guide to PHARMACOLOGY, therapeutic target database and MedGen. (Table 1).

The first portion of data that was retrieved was the basic protein information, this included the gene name, protein length, protein mass and protein function. This information was taken from the UniProt database (Bateman et al., 2020). The next step was to gather variant information for each protein and this data was found in the ClinVar database (Landrum et al., 2017). The information gathered from ClinVar included the location of the variation, the length of the variant, the amino acids that change, the disorders associated with the variation, the pathogenicity of the variation and the global minor allele frequency (GMAF) of the variant. After these pieces of data were collected, the following step was to acquire information on the ligands of all the proteins. This information was taken from PubChem (Kim et al., 2015), ChEMBL (Mendez et al., 2018) and Guide to PHARMACOLOGY (Harding et al., 2017). These databases provided information such as the ligand name, the type of ligand Interaction and the affinity of the protein-ligand bond. The next stage of this element of the research was to find drug-binding information, this was found in both DrugBank and the therapeutic target database (Wang et al., 2019). The information in these databases was the name of the target drug(s), the phase of development in which the drug was in, the drug's pharmacological effect, the action of the drug and the type of molecule the drug is. The following stage of data collection was to gather information on the protein-protein interaction of all the proteins. This data was found in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). In searching STRING information on the functional partner probability of the protein and the predicted protein interaction (Jensen et al., 2009). The last phase of the preliminary research was to find the diseases and disorders associated with each protein. This was found in the gene card and MedGen databases (Louden, 2020) (Figure 1). Basic disorder information was provided by these databases. Next, the PDB of each of the proteins were downloaded. These were downloaded from the RCSB protein data bank (Zardecki et al., 2021).

The following stage of the project was to implement a bioinformatics pipeline. This pipeline would gather and retrieve information in the same manner as the information was gathered in the data collection stage, but in a fully automated fashion - this is why the prior phase of research was of such importance. While creating the pipeline the first step was to structure a plan in which the different entities of the pipeline would interact with each other. An entity-relationship diagram was produced, this worked as a plan for the creation of the pipeline and how the different parts communicate. The entity-relationship diagram consisted of entities, associative entities, and attributes (Figure 2).

Database Scraped	**Section of the database where the information is stored**
UniProt	Protein
ClinVar	Variant
PubChem	Compound
chEMBL	Compound
Guide to PHARMACOLOGY	Compound
therapeutic target database	Treatment

MedGen	Disorder

Table 1, This table shows the databases were scraped, and what section of the database the information is stored in.

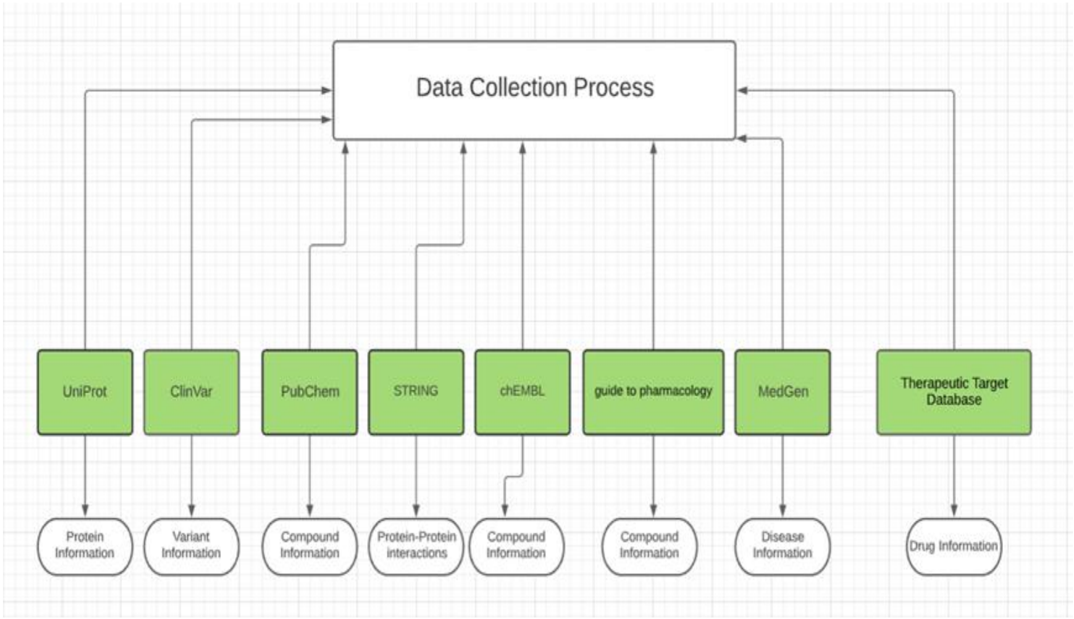


Figure 1, This figure shows the databases used in the data collection process and a basic overview of the data collected from each database

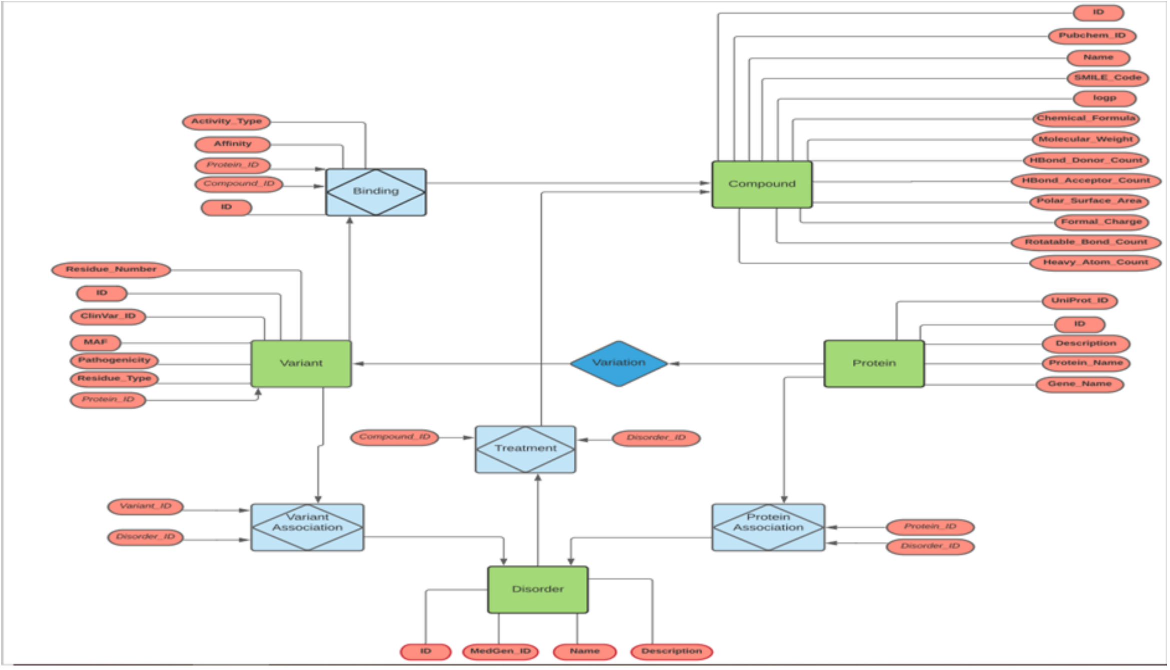


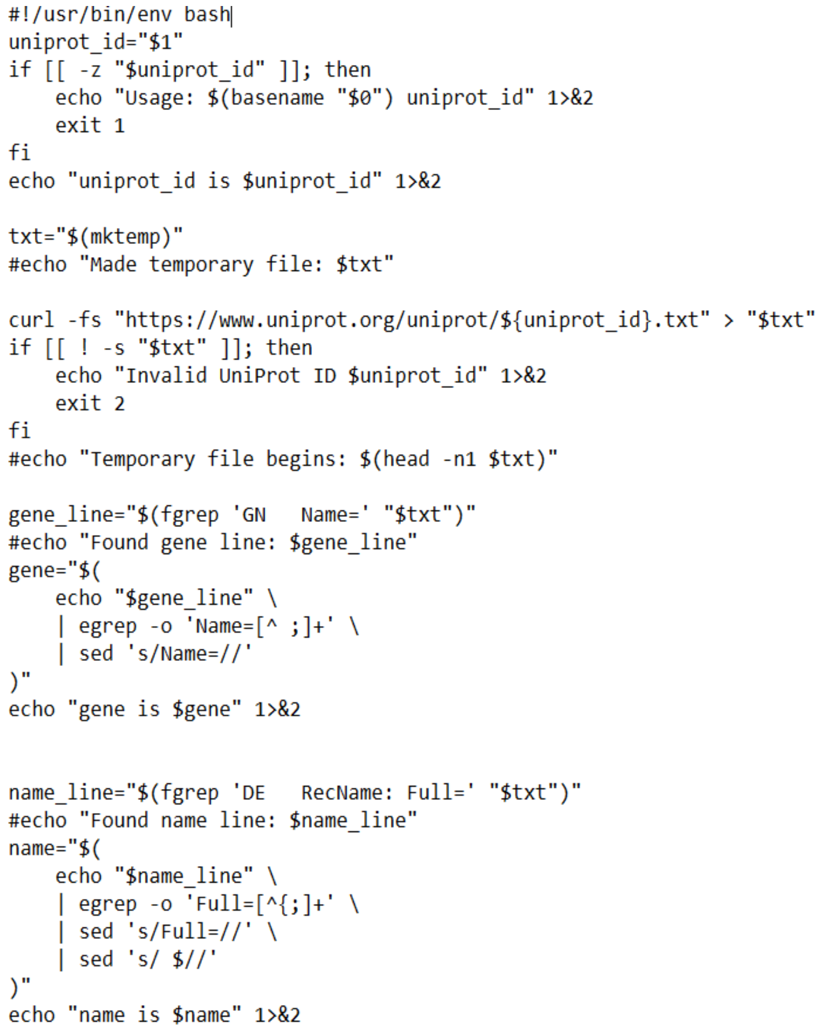
Figure 2, The figure above is the entity-relationship diagram that was used to coordinate the implementation of the pipeline. The diagram consists of entities; Compound, Variant, Disorder, Protein, and associative entities; Treatment, Binding, Protein Association and Variant Association. The attributes of the Compound entity include Compound ID, PubChem ID, Compound name, SMILES (Simplified Molecular Input Line Entry System) code, Logp, Chemical formula, molecular weight, Hydrogen donor count, Hydrogen acceptor count, polar surface area, Rotatable bond count and heavy atom count (These were gathered from PubChem). The protein entity consists of the attributes; UniProt ID, Protein ID, protein description, protein name and gene name (These were gathered from UniProt). The variant entity consists of the attributes; variant ID, ClinVar ID, Residue number, Pathogenicity, Residue type and Protein ID (These were gathered from ClinVar). The attributes of the Disorder entity include Disorder ID, MedGen ID, Disorder Name and Disorder description

(These were gathered from MedGen). The associative entity Binding is used to link the Compound entity with the Variant entity. The associative entity treatment is used to link the Disorder entity with the Compound entity. The associative entity Protein association is used to link the Disorder entity with the protein entity. The associative entity Variant association is used to link the Disorder entity with the Variant entity.

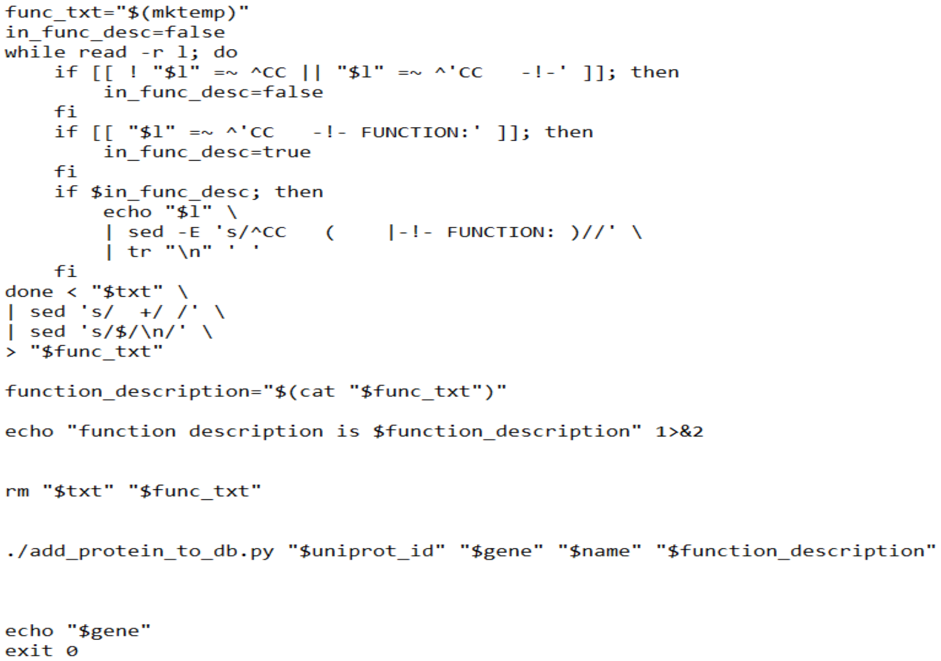
2.3) Bioinformatics Pipeline

2.3.1) Proteins

The next stage of research was to create tables in which the gathered information would be stored. These tables were coded in python, using the program PuTTY. This granted access to Linux. These coding programs were used as I am familiar with their coding language. Using the entity-relationship diagram as a plan, the different components of the tables were coded. Each table collected data from different databases and thus the method by which the tables were coded was essential. An error within the programming would severely hinder the interactions within the pipeline. Once this was complete, the tables served as an updated version of the pipeline plan. The first component of the pipeline was the protein section, the data for this section was gathered from UniProt. Therefore, a file had to be coded and produced, which was able to web scrape the UniProt site for the relevant data. The code prompted the input of a UniProt ID, when inputted this UniProt ID is used to locate a specific UniProt webpage. Once on this page, the code then retrieves the information stored under gene name, protein name and the protein description, text box in UniProt (Figure 3). This data is then stored in the protein table that was created earlier.



```
#!/usr/bin/env bash
uniprot_id="$1"
if [[ -z "$uniprot_id" ]]; then
    echo "Usage: $(basename "$0") uniprot_id" 1>&2
    exit 1
fi
echo "uniprot_id is $uniprot_id" 1>&2

txt="$(mktemp)"
#echo "Made temporary file: $txt"

curl -fs "https://www.uniprot.org/uniprot/${uniprot_id}.txt" > "$txt"
if [[ ! -s "$txt" ]]; then
    echo "Invalid UniProt ID $uniprot_id" 1>&2
    exit 2
fi
#echo "Temporary file begins: $(head -n1 $txt)"

gene_line="$(fgrep 'GN   Name=' "$txt")"
#echo "Found gene line: $gene_line"
gene="$(
    echo "$gene_line" \
    | egrep -o 'Name=[^ ;]+' \
    | sed 's/Name=//'
)"
echo "gene is $gene" 1>&2


name_line="$(fgrep 'DE   RecName: Full=' "$txt")"
#echo "Found name line: $name_line"
name="$(
    echo "$name_line" \
    | egrep -o 'Full=[^{;]+' \
    | sed 's/Full=//' \
    | sed 's/ $//'
)"
echo "name is $name" 1>&2
```

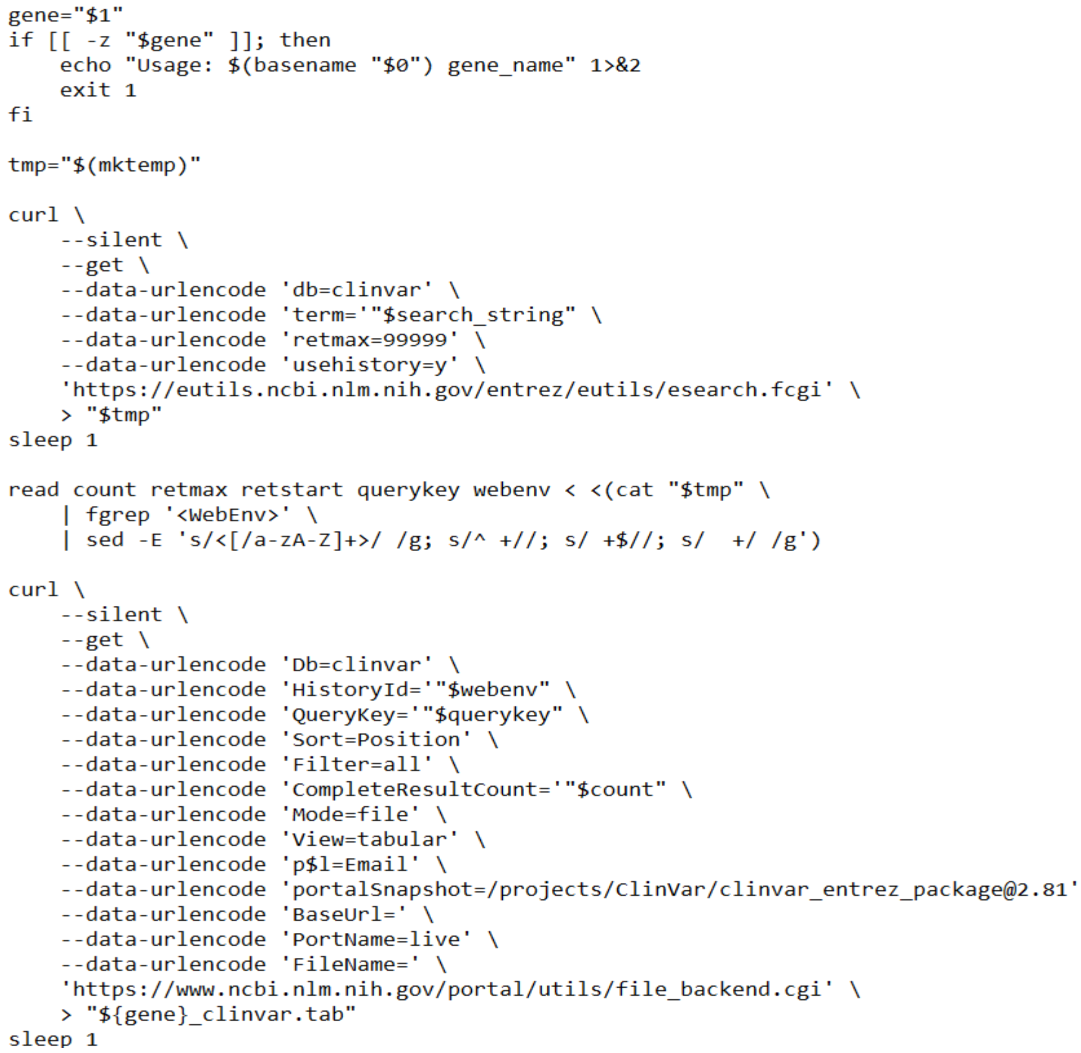


```
func_txt="$(mktemp)"
in_func_desc=false
while read -r l; do
    if [[ ! "$l" =~ ^CC || "$l" =~ ^'CC   -!-' ]]; then
        in_func_desc=false
    fi
    if [[ "$l" =~ ^'CC   -!- FUNCTION:' ]]; then
        in_func_desc=true
    fi
    if $in_func_desc; then
        echo "$l" \
        | sed -E 's/^CC   (     |-!- FUNCTION: )//' \
        | tr "\n" ' '
    fi
done < "$txt" \
| sed 's/  +/ /' \
| sed 's/$/\n/' \
> "$func_txt"

function_description="$(cat "$func_txt")"

echo "function description is $function_description" 1>&2


rm "$txt" "$func_txt"

./add_protein_to_db.py "$uniprot_id" "$gene" "$name" "$function_description"


echo "$gene"
exit 0
```

Figure 3, The figure above is the code and commands used to retrieve the data from the UniProt webpage. By entering the UniProt ID of the requested protein into the pipeline, it returns information, protein name, gene name and a functional description. This data is then stored in a table.

2.3.2) Variants

The next component of the table that needed to be filled with data was the variant table. The information in the variant table was gathered from ClinVar. A file was also created to download information from ClinVar. The initial code in the file prompted an input of a gene when inputted the command, causing the pipeline to perform a search on the ClinVar database for the requested gene. The webpage of the result is converted into an HTML format and important information was gathered. The file command then filtered out all variants that were not single gene variants and were not single nucleotide variants. The code then retrieved the ClinVar ID, residue number, the variant residue type, and the pathogenicity of the variant. A command was then written to stop the duplicate proteins for each given variant, this was done so each protein was associated with all its variants, allowing all the variants of each protein to be investigated at once. In the HTML file, the global minor allele frequency (GMAF) of the variant was not available. This meant a separate file had to be coded to retrieve the GMAF for the variants. The code for this file used the ClinVar ID from the search to access the URL page. The coded command then retrieved the GMAF (if available) and stored it in the original variant file. The protein ID of the gene was then also stored in the table. This created a direct link between the protein data and the variant data, therefore a link between UniProt and ClinVar (Figure 4).



```
gene="$1"
if [[ -z "$gene" ]]; then
    echo "Usage: $(basename "$0") gene_name" 1>&2
    exit 1
fi

tmp="$(mktemp)"

curl \
    --silent \
    --get \
    --data-urlencode 'db=clinvar' \
    --data-urlencode 'term='"$search_string" \
    --data-urlencode 'retmax=99999' \
    --data-urlencode 'usehistory=y' \
    'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi' \
    > "$tmp"
sleep 1

read count retmax retstart querykey webenv < <(cat "$tmp" \
    | fgrep '<WebEnv>' \
    | sed -E 's/<[/a-zA-Z]+>/ /g; s/^ +//; s/ +$//; s/  +/ /g')

curl \
    --silent \
    --get \
    --data-urlencode 'Db=clinvar' \
    --data-urlencode 'HistoryId='"$webenv" \
    --data-urlencode 'QueryKey='"$querykey" \
    --data-urlencode 'Sort=Position' \
    --data-urlencode 'Filter=all' \
    --data-urlencode 'CompleteResultCount='"$count" \
    --data-urlencode 'Mode=file' \
    --data-urlencode 'View=tabular' \
    --data-urlencode 'p$l=Email' \
    --data-urlencode 'portalSnapshot=/projects/ClinVar/clinvar_entrez_package@2.81'
    --data-urlencode 'BaseUrl=' \
    --data-urlencode 'PortName=live' \
    --data-urlencode 'FileName=' \
    'https://www.ncbi.nlm.nih.gov/portal/utils/file_backend.cgi' \
    > "${gene}_clinvar.tab"
sleep 1
```

Figure 4,

The figure above is the code and commands used to retrieve the data from the ClinVar webpage. By entering the ClinVar ID of the requested variant into the pipeline, it returns information, residue number, the variant residue type, and the pathogenicity of the variant. This data is then stored in a table.

2.3.3) Disorders

Coding the command to collect data for the Disorder component of the table was the next step. The data stored in the disorder section of the table was collected from MedGen. A code file was created, to instruct the pipeline on how to retrieve the disorder information required for the table. The code used the ClinVar ID, that has been inputted, to retrieve the ClinVar page of the variant, associated with the ClinVar ID. Next, the code searched for the condition tab, and the MedGen ID of these disorders is then retrieved. The MedGen ID is then used to enter the MedGen page for the associated disorders, then the command instructs the pipeline to retrieve the relevant information from the MedGen page. The disease name, description and MedGen ID are all retrieved and stored in the table. The ClinVar ID of the variant associated with the disorder was also stored in the table. This created an associative link between the variant data and the disorder data and thus linked the data of MedGen and ClinVar together. Due to the pre-existing association between protein and variants, the pipeline was able to simultaneously investigate proteins, single nucleotide polymorphisms and disorders.

2.3.4) Compounds

The compound section of the table was next to be coded. This section of the table had information collected from PubChem. A file containing the coded commands directing how the pipeline would retrieve the information from PubChem was created. The code used the PubChem ID inputted to access an HTML page file of the compound with the corresponding PubChem ID. The code then retrieved all the relevant information about the compound from its PubChem page. The compound's; name, SMILES (Simplified Molecular Input Line Entry System) code, chemical formula, molecular weight, hydrogen bond donor count, hydrogen bond acceptor count, rotatable bond count, polar surface area, heavy atom count, Logp value and formal charge were all collected and then added to the compound table. This was the last entity in the table that had to be coded.

2.3.5) Associative Entities

Next, the associative entity components of the tables had to be coded, this was essential as associative entities allow for a connection to be made between multiple entities that otherwise would not be able to be connected. 4 associative entities had to be coded, protein association, variant association, treatment, and binding. The first associative entity to be coded was protein association. Each protein, compound, variant, and disorder were given a unique ID within the table. The protein association component consisted of protein IDs and disorder IDs allowing for an association to be made between disorders and proteins. The next associative entity to be coded was variant association. This associative entity consisted of variant IDs and disorder IDs allowing for an association to be made between disorders and variants. Then the next associative entity to be coded was the treatment associative entity, which consisted of compound IDs and disorder IDs allowing for an association to be made between disorders and compounds. The last associative entity to be coded was the binding associative entity which consisted of protein IDs and compound IDs allowing for an association to be made between protein and compound. The binding associative entity also contains information on the binding affinity between the protein and the compound, this data was gathered from PubChem.

2.4) Pipeline Testing

Once all the coding had been completed the program was run. The UniProt code of all the 720 proteins in the full pharmacology set was fed through the pipeline and it began retrieving all the relevant information that it had been coded to retrieve. Once all the information had been downloaded, the database was fully functioning. Once fully functioning, the pipeline was tested using hypothetical case studies. This was done to examine the effectiveness of the pipeline. Three different hypothetical case studies were constructed in a manner that would investigate the pipeline's ability to gather multiple groups of information, starting from various parts of the pipeline. The first case study involved using the pipeline to gather enough information to investigate drug therapies for the disorder 'Acute Myeloid Leukaemia'. The second case study was to have the pipeline gather enough relevant information on the 'KIT' protein to start an investigation into possible drug therapies for the disorders caused due to KIT protein. The third case study involved gathering information on the most polymorphic proteins and most pathogenic variants. This was constructed to imitate the data collection process that would be done in a research project aimed at investigating the most pathogenic variants individually and attempt to postulate feasible therapies for the variants in question. The fourth case study involved gathering information on compounds that bind to the Ryanodine receptor. This was constructed to show if the pipeline would be able to highlight the effects of the compound on a specific protein, for those investigating a specific compound or developing a structurally similar compound. The next stage of the research project was to use the newly created pipeline to assess the effect of polymorphisms on the disorders and currently used drugs.

As a control for this pipeline testing, the databases from which the information was derived, were checked against the information returned by the pipeline to ensure that the information being returned by the pipeline was accurate and complete.

2.5) Protein Docking

Using the pipeline, highly polymorphic proteins were found, along with the variants, associated diseases, and drug therapies. The proteins selected were GAA, SCN2A BRAF, KIT, and PAH. The PDB structures of these proteins were retrieved. Then using Chimera, they were modelled. Using Chimera, side-chain substitutions were performed on the proteins. For each protein, 10 pathogenic variants of the proteins were modelled, and 10 non-pathogenic variants were modelled. The 10 non-pathogenic variants were selected as a control. The pathogenicity of the variants was dictated by the variant's pathogenicity rating on ClinVar. The selection of a control set was essential as it was important in the assessment of the complete effect of polymorphisms and any variations in drug binding. The pathogenic and non-pathogenic variants were chosen at random. For each of the side-chain substitutions, an amino acid in the sequence was substituted with a different amino acid in accordance with the residue displayed on the ClinVar variant page. When assigning the rotamer for the residue, the position with the highest-scoring rotamer probability was selected, as long as the rotamer was feasible. The Dunbrack 2010 backbone dependant rotamer library was used during the side-chain substitutions (Shapovalov & Dunbrack, 2011). Once this process was completed for all 20 variants of all 5 selected proteins, the next part of the research was to start binding the drug compounds for each protein, to the 20 different variants of each protein.

The docking process involved various stages to ensure the results were as accurate as possible. Once the protein PDB was retrieved the structure had to be denuded, this involved removing all additional structures, such as oligosaccharides and ions, from the protein, unless the structure was a known cofactor for the protein. Once denuded, the protein structure was prepared for docking using the DockPrep module in UCSF Chimera. This feature protonated the protein, which is essential as the addition of hydrogen atoms causes the addition of different forces, such as H bonds that will affect the docking of external molecules (Protein docking simulations are conducted in a vacuum and thus it is assumed that the pH of the surrounding does not affect the binding energy and the protein's protonation state). Next, the DockPrep module calculated the relative charges of all the molecules in the structure. This process is also essential in making the results of protein docking as accurate as possible. The calculation of each molecule's relative charge is important because it stimulates the electrostatic interactions that would occur during biological docking. Then the ligand had to be prepared for docking. The ligands were protonated using the OpenBabel tool (O'Boyle et al., 2011) and then had their relative charges calculated using Antechamber (Wang, Wang, Kollman & Case, 2006). Protein docking simulations are conducted in a vacuum and thus it is assumed that the pH of the surrounding does not affect the binding

The next step in the docking process was to select the volume within the protein in which the ligand would attempt to dock. The volume of the space in which the ligand could dock was kept as small as possible while still containing all parts of the protein's active site/ or active sites. Once the docking was complete the configuration with the highest binding energy was selected and documented.

The molecular docking was conducted on SSH using a program called PLANTS (Exner, Korb, & ten Brink, 2009). This program allowed for the mol2 coordinates file of a protein and ligand to be inputted. Then using the box method, the binding area in which the ligand is docked can be set. In the box method, 6 coordinates are given to the program, the first three coordinates are the x, y and z (respectively) coordinates of the binding site. The next three coordinates tell the program size of the x, y, and z (respectively) axis of the binding area. Inputting all this information into the program, allowed it to return information on the complex created by the protein-ligand binding. One of the pieces of information returned was the binding energy of the complex. The docking process was conducted on all the selected proteins of interest and their ligands. The results of this were documented in an Excel spreadsheet. The PLANTS system was then used again, as it allows for docking via another method. This second method works by using the location of a pre-bound ligand as an example to determine where the newly requested ligand would dock. For the selected proteins whose PDB's already had a ligand-bound (3; KIT, PAH, BRAF) this method was also run. This was done by isolating and extracting the ligand bound to them in the PDB, saving the ligand as a mol2 file and then inputting this to the PLANTS program, which was able to use the coordinates of the ligand as a reference. Running this method provides an extra set of docking data points for the calculation of the binding energies of these proteins and ligands. The second method of docking (via reference) did not provide any viable results and thus the results were not added to the results of this project.

Once these results were collected, statistical tests were then run on the data, to identify its relevance and significance. The statistical tests that were conducted were mean calculation, standard deviation calculation and an ANOVA (or Kruskal-Wallis H test if data is not normally distributed and therefore non-parametric(Table 2)) (Table 3). An ANOVA was conducted on the data that was normally distributed data, the input data was two independent categories (Wild-type and Variant), and the outcome data were quantitative (binding energy).

When comparing the binding energies, differences between wildtype and variant energies, of more than 15% were considered notable. The 15% threshold was applied because it identified a discrete group of proteins that could be used for the comparative analysis.

Shapiro-Wilk Test for normality significance value										
	GAA		SCN2A		BRAF		KIT		PAH	
	Acarbose	Miglitol	Tramadol	Zonisamide	Sorafenib	Vemurafenib	Fludiazepam	Ergocalciferol	Droxidopa	Sapropterin
Pathogenic	0.256	0.001	0.385	0.156	0.052	0.010	0.002	0.064	0.393	0.004
Control	0.126	0.002	0.023	0.033	0.009	0.007	0.086	0.341	0.237	0.001

Table 2, this table shows the significance value of all binding energies, when a Shapiro-Wilk Test for normality was conducted on it. The cells filled in red, signify data that were not normally distributed, and thus Kruskal-Wallis H tests were performed on their data sets as opposed to an ANOVA.

ANOVA and Kruskal-Wallis H significance value										
	GAA		SCN2A		BRAF		KIT		PAH	
	Acarbose	Miglitol	Tramadol	Zonisamide	Sorafenib	Vemurafenib	Fludiazepam	Ergocalciferol	Droxidopa	Sapropterin
Pathogenic	0.575	0.113*	0.527	0.940	0.065	0.114*	0.206*	0.920	0.314	0.112*
Control	0.406	0.113*	0.752*	0.752*	0.114*	0.114*	0.509	0.820	0.460	0.113*

Table 3, this table shows the significance value of all binding energies, when an ANOVA or Kruskal-Wallis H Test for significance was conducted on it. The test was conducted to see if there was any statistically significant difference in the binding energy of variants proteins and wild-type proteins. A score of below 0.05 would indicate statistical significance. The values returned from Kruskal-Wallis H are marked with an *.

3) Results

3.1) Overview

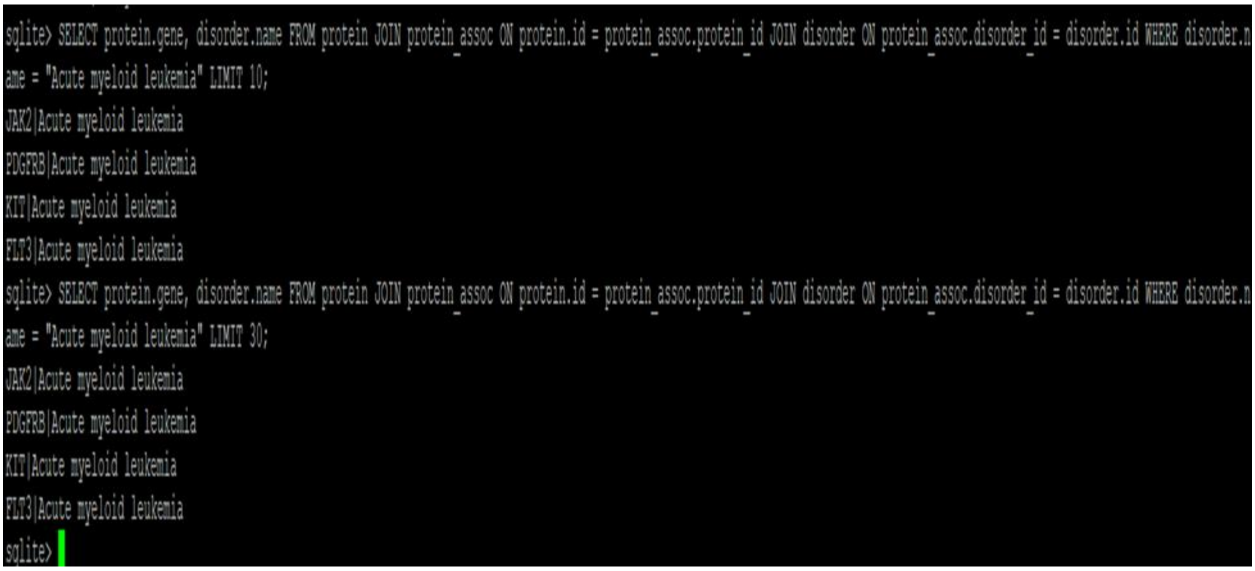
The bioinformatic pipeline was coded, and then instructed to web-scrape multiple databases for key information on proteins, variants, compounds, disorders and drugs. The pipeline was able to identify which databases to examine, due to the in-depth analysis that was conducted on 60 proteins randomly selected from the Full Pharmacology Set. After this, the pipeline was put to use. Firstly, the full pharmacology set was added to the pipeline, this is a set of 720 pharmacologically relevant proteins. The pipeline retrieved and stored the data for all the proteins in the set. Then the pipeline was used to gather information to aid the data collection process of multiple hypothetical case studies such as research into Acute Myeloid Leukaemia, the KIT gene and the ryanodine receptor. The database was able to retrieve all the relevant data required for the case studies, it showed great potential as it was able to return a multitude of key information, which would have all been useful for the case studies. Protein docking was then conducted on 5 proteins (GAA, SCN2A, BRAF, KIT and PAH), for each protein, its wildtype, 10 of its pathogenic variants and 10 of its benign variants were bound to 2 different drug therapies associated with the protein. The binding energies of the dockings were collected and analysed. The results showed that there was no statistically significant difference between pathogenic and benign variants' binding energy, however, in most cases, more pathogenic variants had binding energies that differed from the wildtype binding energy by at least 15%.

3.2) Data Collections

Prior to the construction of the bioinformatic database, 60 randomly selected proteins from the Full Pharmacology Set, were investigated in-depth. This was done so the best method of data collection was understood, which allowed for the code written in the pipeline to mimic authentic data collection, accurately. The research into these proteins provided; information on the protein's ligands (PubChem/chEMBL), key data about the genes that encode the protein (UniProt), the disorders associated with the protein (MedGen), the variants of the protein (ClinVar), the drugs that interact with the protein (Therapeutic Target Database) and the protein-protein interactions that the proteins undergo (STRING). These pieces of data were collected for all 60 proteins and the database from which these pieces of information were retrieved and stored. This provided all the key databases that needed to be searched, by the pipeline, in order to return the relevant information.

3.3) Bioinformatics Database

The first stage of this research project consisted of creating a bioinformatics database. This database was a collation of data from multiple pre-existent proteins, drug, and disease databases. The aim of this was to investigate the initial hypothesis. The hypothesis of this project is that by developing a bioinformatic tool that can link existing datasets associated with the pathogenicity of genetic variation in humans, with existing datasets associated with drug targets in humans, and existing datasets associated with drug therapies with diseases. This can allow for the integration of information from different types of datasets automatically. Therefore, allowing the automatic and rapid exploration of the relevant connections between several large and important datasets. Also, integrating drug, disorder and variant datasets, simultaneously, it can allow for advancements to be made in the field of precision medicine. In this project, the underlying contention is that single nucleotide polymorphic variants in proteins have an effect on the binding between proteins and drug therapies, and the use of bioinformatic tools can advance precision medicine by detecting these effects. Once created the bioinformatic pipeline was used in multiple hypothetical case studies. It proved to be very useful in the collection of data in all the scenarios tested. There were case studies created to test the database and, in all cases, the database was very effective in collecting data and incorporating multiple categories of data together.



```
sqlite> SELECT protein.gene, disorder.name FROM protein JOIN protein_assoc ON protein.id = protein_assoc.protein_id JOIN disorder ON protein_assoc.disorder_id = disorder.id WHERE disorder.n
ame = "Acute myeloid leukemia" LIMIT 10;
JAK2|Acute myeloid leukemia
PDGFRB|Acute myeloid leukemia
KIT|Acute myeloid leukemia
FLT3|Acute myeloid leukemia
sqlite> SELECT protein.gene, disorder.name FROM protein JOIN protein_assoc ON protein.id = protein_assoc.protein_id JOIN disorder ON protein_assoc.disorder_id = disorder.id WHERE disorder.n
ame = "Acute myeloid leukemia" LIMIT 30;
JAK2|Acute myeloid leukemia
PDGFRB|Acute myeloid leukemia
KIT|Acute myeloid leukemia
FLT3|Acute myeloid leukemia
sqlite>
```

Figure 5, This figure shows, the results returned when the pipeline is asked to provide all proteins associated with Acute Myeloid Leukaemia. The pipeline was able to do this by inputting the MedGen code of Acute Myeloid Leukaemia, into ClinVar and then retrieving the proteins that are linked with the disorder.

Genes involved with Acute Myeloid Leukaemia	
Gene	**Disorder**
JAK2	Acute Myeloid Leukaemia
PDGFRB	Acute Myeloid Leukaemia
KIT	Acute Myeloid Leukaemia
FLT3	Acute Myeloid Leukaemia

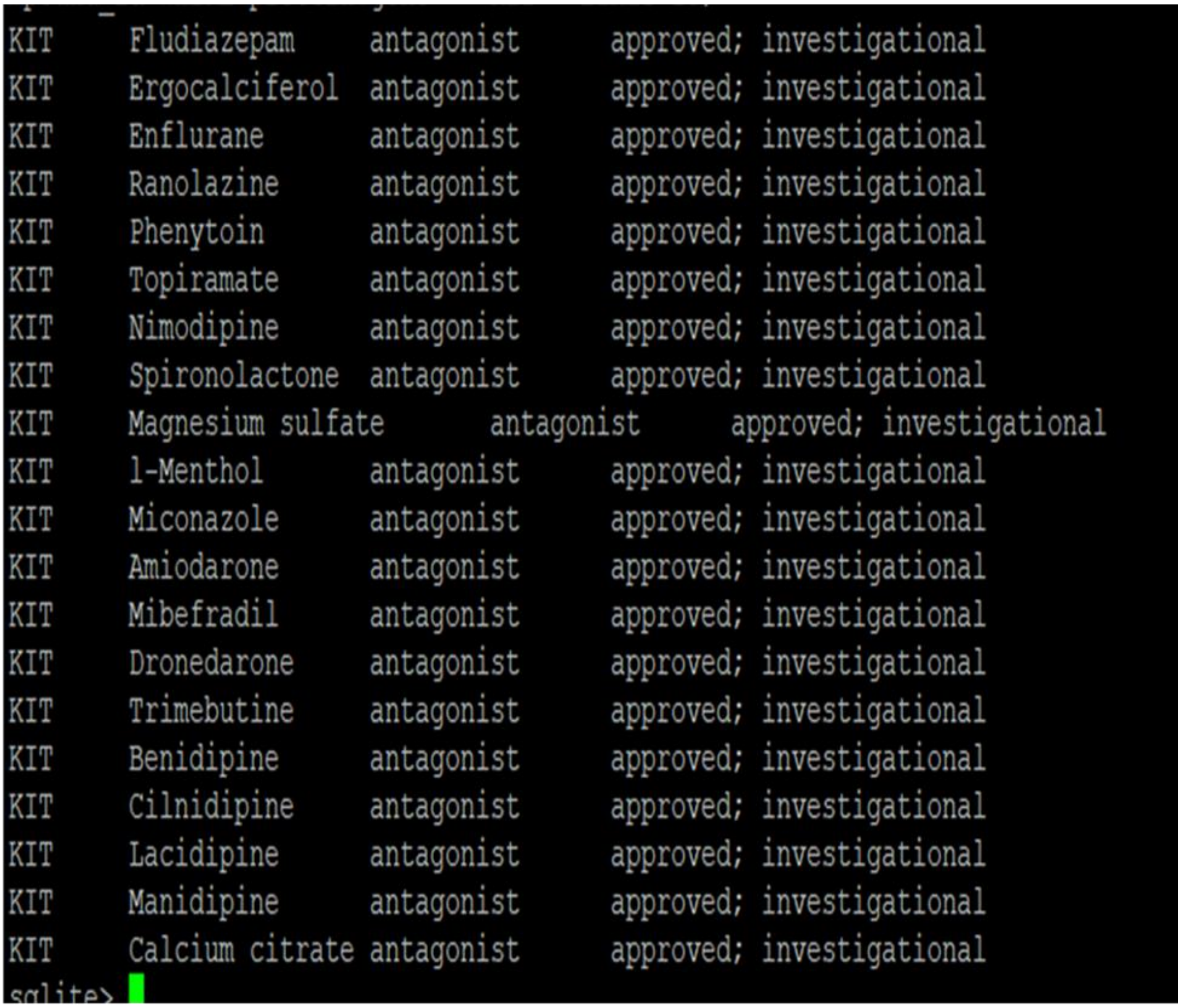
Table 4 - This table shows, the results returned when the pipeline is asked to provide all proteins associated with Acute Myeloid Leukaemia.

3.3.1) Acute Myeloid Leukaemia (Case Study)

In the first case, information was gathered in a manner that would replicate an investigation into drug therapies to combat Acute Myeloid Leukaemia. The database was able to gather all the requested information, it gathered all the proteins associated with Acute Myeloid Leukaemia. It also retrieved all compounds that interact with each protein that was associated with Acute Myeloid Leukaemia. The full details of all the compounds were provided, including the type of interaction with the protein, the compound's SMILES code and the relative polar charge of the compound. The information provided was very thorough, and with enough data to begin investigating drug therapies for Acute Myeloid Leukaemia (Figure 5). The use of the pipeline was considered successful in this case study. The results presented by the pipeline indicate that Acute Myeloid Leukaemia was associated with 4 genes (Janus kinase 2 (JAK2), Platelet-derived growth factor receptor beta (PDGFRB), Proto-oncogene c-KIT (KIT) and FMS-like tyrosine kinase 3 (FLT3) (Table 4).

3.3.2) Proto-oncogene c-KIT (Case Study)

The subsequent case study was an investigation into the Proto-oncogene c-KIT (KIT). In this case, study, the objective was to investigate the protein that the gene encodes, and then to research data into the drug therapies and disorders that are associated with the gene. The pipeline was also used to great effect in this case study. A large amount of key information was returned using the pipeline. The database was able to return information, such as all the compounds that bind to the KIT protein, the biological activity of all the compounds and the development phase of all the compounds. The pipeline was then able to return information on the single nucleotide polymorphic variants that were present in the KIT protein, along with the residue number of the variant, the pathogenicity of the variant and the globe minor allele frequency (GMAF) of the variant (If the GMAF was available). This was evidence of the strong capability that the pipeline has, as the pipeline provided all the information required to start an in-depth research project into the KIT protein. The results displayed by the pipeline showed that the KIT gene; has 20 compound ligands; all of which were antagonists to the KIT protein and also were all approved compounds (Table 5)(Table 6) (Figure 6) (Figure 7).



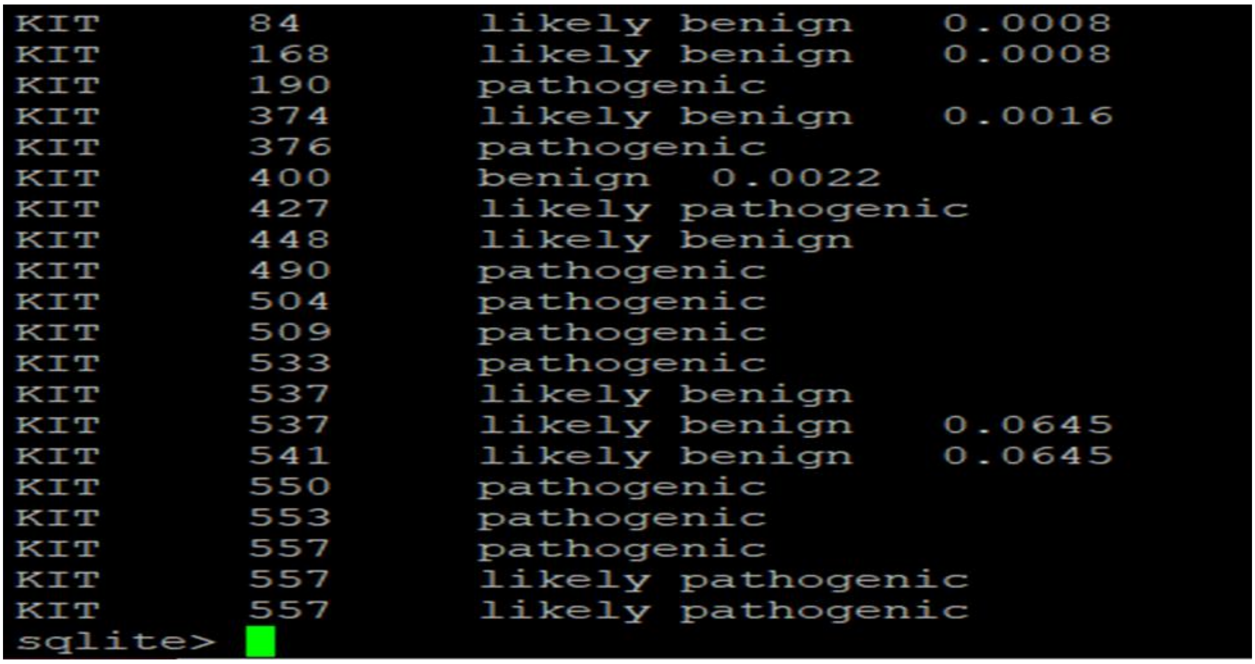
```
KIT     Fludiazepam     antagonist      approved; investigational
KIT     Ergocalciferol  antagonist      approved; investigational
KIT     Enflurane       antagonist      approved; investigational
KIT     Ranolazine      antagonist      approved; investigational
KIT     Phenytoin       antagonist      approved; investigational
KIT     Topiramate      antagonist      approved; investigational
KIT     Nimodipine      antagonist      approved; investigational
KIT     Spironolactone  antagonist      approved; investigational
KIT     Magnesium sulfate       antagonist      approved; investigational
KIT     l-Menthol       antagonist      approved; investigational
KIT     Miconazole      antagonist      approved; investigational
KIT     Amiodarone      antagonist      approved; investigational
KIT     Mibefradil      antagonist      approved; investigational
KIT     Dronedarone     antagonist      approved; investigational
KIT     Trimebutine     antagonist      approved; investigational
KIT     Benidipine      antagonist      approved; investigational
KIT     Cilnidipine     antagonist      approved; investigational
KIT     Lacidipine      antagonist      approved; investigational
KIT     Manidipine      antagonist      approved; investigational
KIT     Calcium citrate antagonist      approved; investigational
sqlite>
```

Figure 6, This figure shows, the results returned when the pipeline is asked to provide all compounds associated with the KIT protein. It also returned the mechanism of action in which the compound and protein interact. It also provided the development phase of the compound. The pipeline was able to do this by inputting the UniProt code of KIT protein, into PubChem and then retrieving the compounds and the requested key information.

Gene	Compound	Compound Action	Development Stage
KIT	Fludiazepam	potentiator	experimental; illicit
KIT	Ergocalciferol	potentiator	experimental; illicit
KIT	Enflurane	potentiator	experimental; illicit
KIT	Ranolazine	potentiator	experimental; illicit
KIT	Phenytoin	potentiator	experimental; illicit
KIT	Topiramate	potentiator	experimental; illicit
KIT	Nimodipine	potentiator	experimental; illicit
KIT	Spironolactone	potentiator	experimental; illicit
KIT	Magnesium sulfate	potentiator	experimental; illicit
KIT	l-Menthol	potentiator	experimental; illicit
KIT	Miconazole	potentiator	experimental; illicit
KIT	Amiodarone	potentiator	experimental; illicit
KIT	Mibefradil	potentiator	experimental; illicit
KIT	Dronedarone	potentiator	experimental; illicit
KIT	Trimebutine	potentiator	experimental; illicit
KIT	Benidipine	potentiator	experimental; illicit
KIT	Cilnidipine	potentiator	experimental; illicit
KIT	Lacidipine	potentiator	experimental; illicit

KIT	Manidipine	potentiator	experimental; illicit
KIT	Calcium citrate	potentiator	experimental; illicit

Table 5 - This table shows, the results returned when the pipeline is asked to provide all compounds associated with the KIT protein. It also returned the mechanism of action in which the compound and protein interact. It also provided the development phase of the compound.



```
KIT      84      likely benign    0.0008
KIT      168     likely benign    0.0008
KIT      190     pathogenic
KIT      374     likely benign    0.0016
KIT      376     pathogenic
KIT      400     benign   0.0022
KIT      427     likely pathogenic
KIT      448     likely benign
KIT      490     pathogenic
KIT      504     pathogenic
KIT      509     pathogenic
KIT      533     pathogenic
KIT      537     likely benign
KIT      537     likely benign    0.0645
KIT      541     likely benign    0.0645
KIT      550     pathogenic
KIT      553     pathogenic
KIT      557     pathogenic
KIT      557     likely pathogenic
KIT      557     likely pathogenic
sqlite>
```

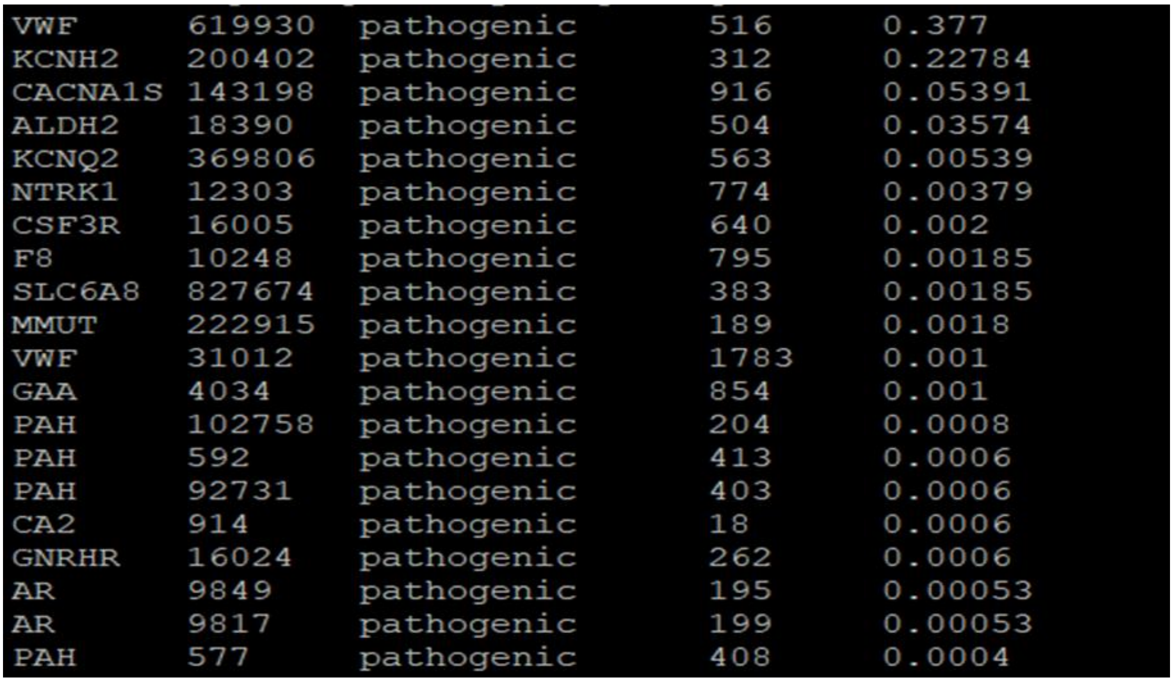
Figure 7, This figure shows the results returned when the pipeline is prompted to provide all single nucleotide polymorphic variants associated with the KIT protein. It also returned the residue number of all the polymorphic variants, the pathogenicity of the variant and the GMAF of the variant. The pipeline was able to do this by inputting the UniProt code of the KIT protein, into ClinVar and then retrieving the variant and the requested key information.

Gene	Residue Number	Variant Pathogenicity	GMAF
KIT	84	Likely benign	0.0008
KIT	168	Likely benign	0.0008
KIT	190	Pathogenic	-
KIT	374	Likely benign	0.0016
KIT	376	Pathogenic	-
KIT	400	Benign	-
KIT	427	Likely pathogenic	-
KIT	448	Likely benign	-
KIT	490	Pathogenic	-
KIT	504	Pathogenic	-
KIT	509	Pathogenic	-
KIT	533	Pathogenic	-
KIT	537	Likely benign	-
KIT	537	Likely benign	0.0645
KIT	541	Likely benign	0.0645
KIT	550	Pathogenic	-
KIT	553	Pathogenic	-
KIT	557	Pathogenic	-
KIT	557	Likely pathogenic	-
KIT	557	Likely pathogenic	-

Table 6 - This table shows the results returned when the pipeline is prompted to provide all single nucleotide polymorphic variants associated with the KIT protein. It also returned the residue number of all the polymorphic variants, the pathogenicity of the variant and the GMAF of the variant

3.3.3) Prevalent Polymorphisms (Case Study)

In the third case study created to test the database, the pipeline was instructed to retrieve information on the most pathogenic and polymorphic variants. This case study was constructed to mimic a scenario where a researcher would like to investigate the most pathogenic variants individually and attempt to find therapies for the variants in question. The pipeline was able to present information on all the pathogenic (and likely pathogenic) single nucleotide polymorphic variants and rank them in descending order of GMAF. The pipeline also presented the variant number and residue number of all the variants. The database was able to handle this scenario very well. The results displayed by the pipeline indicate that the genes; Von Willebrand Factor (VWF), Potassium Voltage-Gated Channel Subfamily H Member 2 (KCNH2) and Calcium Voltage-Gated Channel Subunit Alpha1 S (CACNA1S), had the most polymorphisms, and had the greatest number of pathogenic variants. VWF has the pathogenic variant with the highest prevalence (38% minor allele frequency) (Table 7) (Figure 8).



```
VWF      619930  pathogenic     516     0.377
KCNH2    200402  pathogenic     312     0.22784
CACNA1S  143198  pathogenic     916     0.05391
ALDH2    18390   pathogenic     504     0.03574
KCNQ2    369806  pathogenic     563     0.00539
NTRK1    12303   pathogenic     774     0.00379
CSF3R    16005   pathogenic     640     0.002
F8       10248   pathogenic     795     0.00185
SLC6A8   827674  pathogenic     383     0.00185
MMUT     222915  pathogenic     189     0.0018
VWF      31012   pathogenic     1783    0.001
GAA      4034    pathogenic     854     0.001
PAH      102758  pathogenic     204     0.0008
PAH      592     pathogenic     413     0.0006
PAH      92731   pathogenic     403     0.0006
CA2      914     pathogenic     18      0.0006
GNRHR    16024   pathogenic     262     0.0006
AR       9849    pathogenic     195     0.00053
AR       9817    pathogenic     199     0.00053
PAH      577     pathogenic     408     0.0004
```

Figure 8, This figure shows the results returned when the pipeline is prompted to provide the most prevalent pathogenic variants. It also returned the residue number and ClinVar ID of all the polymorphic variants. It also provided the pathogenicity of the variant and the GMAF of the variant. The pipeline was able to do this by inputting the UniProt code of all proteins, into ClinVar and then retrieving the variant data for all the pathogenic variants.

Gene	Clinvar ID	Pathogenicity	Residue Number	GMAF
VWF	619930	Pathogenic	516	0.377
KCNH2	200402	Pathogenic	312	0.22784
CACNA1S	143198	Pathogenic	916	0.05391
ALDH2	18390	Pathogenic	504	0.03574
KCNQ2	369806	Pathogenic	563	0.00539
NTRK1	12303	Pathogenic	774	0.00379
CSF3R	16005	Pathogenic	640	0.002
F8	10248	Pathogenic	795	0.00185

SLC6A8	827674	Pathogenic	383	0.00185
MMUT	222915	Pathogenic	189	0.0018
VWF	31012	Pathogenic	1783	0.001
GAA	4034	Pathogenic	854	0.001
PAH	102758	Pathogenic	204	0.0008
PAH	592	Pathogenic	413	0.0006
PAH	92731	Pathogenic	403	0.0006
CA2	914	Pathogenic	18	0.0006
GNRHR	16024	Pathogenic	262	0.0006
AR	9849	Pathogenic	195	0.00053
AR	9817	Pathogenic	199	0.00053
PAH	577	Pathogenic	408	0.0004

Table 7 - This table shows the results returned when the pipeline is prompted to provide the most prevalent pathogenic variants. It also returned the residue number and ClinVar ID of all the polymorphic variants. It also provided the pathogenicity of the variant and the GMAF of the variant.

3.3.4) Ryanodine receptor (Case Study)

There was also a piece of data collection research that was conducted on the pipeline. The pipeline was instructed to collect information on the gene ryanodine receptor (RYR1). This includes information on the variants of RYR1, such as the ClinVar ID, pathogenicity, residue number and the GMAF. Then also information on all the compounds that are associated with RYR1 (Table 8) (Table 9)

Gene	ClinVar ID	Pathogenicity	Residue number	MAF
RYR1	93279	benign	2060	0.05491
	93265	benign	1342	0.05411
	133011	benign	3751	0.03514
	133149	benign	1787	0.01697
	224403	benign	4501	0.01078
	133223	likely benign	2787	0.00899
	159846	benign	1352	0.00839
	256490	likely benign	1109	0.00719
	132999	benign	3578	0.00679
	29878	likely benign	3976	0.00599
	93275	likely benign	1878	0.00539
	93244	likely benign	3642	0.00339
	133157	likely pathogenic	2129	0.00319

Table 8, This table shows all the variants of RYR1 with a known minor allele frequency (MAF). The variants are in descending order of MAF. This Figure also shows the pathogenicity of the variant, the ClinVar ID and the residue number of the variant.

Gene	Drug	Development Stage	Action
RYR1	Caffeine	approved	N/A
RYR1	Dantrolene	approved; investigational	antagonist
RYR1	Suramin	investigational	agonist
RYR1	Tetracaine	approved; vet approved	modulator
RYR1	Calcium Citrate	approved; investigational	substrate
RYR1	Calcium phosphate	approved	substrate

Table 9, This table shows, all compounds associated with the RYR1 protein. It also returned the mechanism of action in which the compound and protein interact. It also provided the development phase of the compound. This information was gathered using the pipeline.

As a control measure, after all this information was returned by the bioinformatic pipeline, the results were then cross-referenced with databases from which the data was retrieved. To ensure the data retrieve is accurate. In all the case studies the data returned was complete and accurate when cross-referenced with the web-scraped database.

3.4) Protein Binding Energy

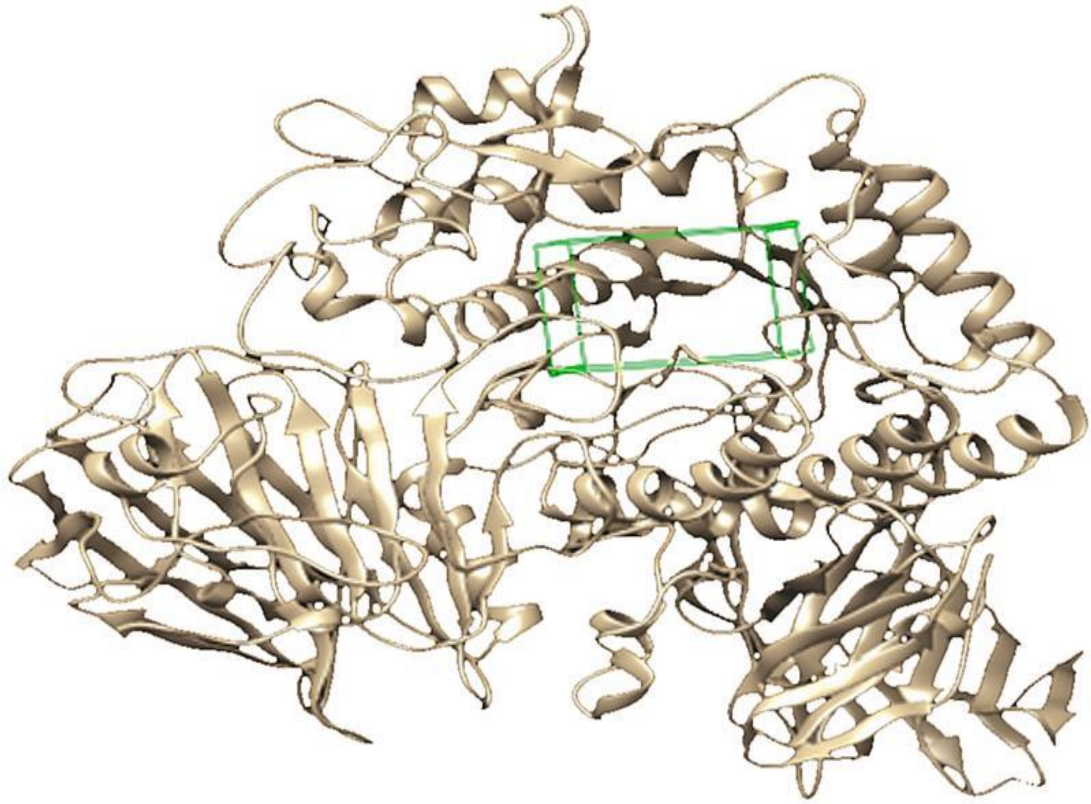
The second part of this research project consisted of using knowledge from the bioinformatics database to identify the effect of variants on protein-ligand binding. The aim of this was to investigate one of the project's hypothesises, that single nucleotide polymorphisms have a significant adverse effect on protein-ligand binding. This hypothesis was tested by calculating the protein-ligand binding energy of multiple protein-ligand complexes. The program used to perform this test was the PLANTS program, a bioinformatics tool that simulates protein-drug binding and calculates potential binding energies. The binding was conducted via two methods; the box method and also by using a reference ligand to determine the binding site. However, due to an error with the PLANTS program while using the reference method no viable results were retrieved. 5 proteins were docked using the box method, for each protein 2 drug therapies were selected as ligands. Each protein had 20 variants generated, 10 of the variants were benign and another 10 of the variants were pathogenic. The data from all these dockings were collected, analysed, and then stored in a table (Table 10).

		GAA		SCN2A		BRAF		KIT		PAH	
		Acarbose	Miglitol	Tramadol	Zonisamide	Sorafenib	Vemurafenib	Fludiazepam	Ergocalciferol	Droxidopa	Sapropterin
Average binding energy	Wildtype	-89.4	-71.59	-66.13	-54.75	-76.63	-69.28	-67.38	-85.24	-74.75	-69.99
	Control	-84.95	-67.92	-64.2	-54.43	-89.63	-89.18	-69.24	-85.56	-78	-76.84
	Pathogenic	-80.79	-67.51	-64.15	-54.42	-91.95	-88.06	-71.18	-84.42	-77.41	-76.62
Standard Deviation	Control	7.3	5.42	2.8	3.97	5.9	5.5	3.95	2.93	2.91	0.34
	Pathogenic	9.41	5.61	4.26	3.82	4.69	5.41	5.27	3.33	3.28	0.6

Table 10, The table above shows the average binding energies of all the proteins and the ligands, that were bound together. It also has the standard deviation of the binding energies.

3.4.1) GAA

3.4.1.1) GAA and Acarbose



((Figure 9, This Figure shows the protein structure of dock-prepped GAA wild-type protein and the docking volume, in which the ligands would be bound. The co-ordinate of this box is x = -12.28, y = -36.64, z = 95.29. The size of the binding area is x = 11.60, y = 23.14, and z = 16.00 (arbitrary units).

The first protein-ligand complexes that were bound were the Lysosomal alpha-glucosidase (GAA, (5nn8 PDB)) and the compound acarbose. This was done via the box method, where the active site location of the GAA protein was set to the coordinates x = -12.28, y = -36.64, z = 95.29. The size of the binding area was set to x = 11.60, y = 23.14, and z = 16.00 (arbitrary units) (Figure 9). When acarbose was bound to the GAA wildtype with those volume parameters stated, the calculated binding energy was -89.40 kcal per mol. When acarbose was then bound to the benign variants of the GAA proteins, the average calculated binding energy was -84.95 kcal per mol (p=0.235), with a standard deviation of 7.30. In the benign cohort of variant-ligand binding, only one of the docking simulations provided a binding energy that was notably different from the wild-type binding energy. This was benign variant 7 (variant number: 92477), this complex had a binding energy of -76.41 kcal per mol. which was more than 15% weaker than the wild-type binding energy. The mutation of this variant was valine to isoleucine (residue number 816). The average root-mean-square deviation (RMSD) for the acarbose ligands bound to the benign group of GAA variants was 8.95, with a standard deviation of 6.20. When acarbose was then bound to the pathogenic variants of the GAA proteins, at the parameters stated the average calculated binding energy was -80.79 kcal per mol (p=0.235), with a standard deviation of 9.41. In the pathogenic group of variant-ligand binding, (with the acarbose compound), 4 of the docked complexes had a binding energy that was considerably weaker than that of the wild type. These variants were pathogenic variants 1, 4, 5 and 8 (variant numbers: 92483, 189188, 956209 and 188797, respectively). The complex formed with the pathogenic variant 1 and acarbose had a binding energy of -67.17 kcal per mol. Pathogenic variant 4 had a binding energy of -76.68 kcal per mol. The complex formed when the pathogenic variant 5 was bound to acarbose had a binding energy of -71.21 kcal per mol. Pathogenic variant 8 had binding energy of 69.18 kcal per mol. The mutation of pathogenic variant 1 was cysteine to glycine (residue number 103). The mutation of pathogenic variant 4 was arginine to tryptophan (residue number 224). The mutation of pathogenic variant 5 was leucine to phenylalanine (residue number 291). The mutation of pathogenic variant 8 was Glycine to Arginine (residue number 309). All of these binding energies were at least 15% weaker than the wild-type binding energy. These complexes were also among those with the high RMSD values, of the pathogenic variant group (11.22, 15.52, 12.58 and 13.12 respectively). The average root-mean-square deviation (RMSD) for the acarbose ligands bound to the pathogenic group of GAA variants was 9.27, with a standard deviation of 5.13.

3.4.1.2) GAA and Miglitol

The next protein-ligand complex that was bound was GAA and the compound miglitol. This was done via the box method. The binding site location for this docking simulation was set to the same coordinates as that of the acarbose binding simulation, x = -12.28, y = -36.64, z =95.29. The size of the binding area was set to x = 11.60, y = 23.14, and z = 16.00 (arbitrary units). When miglitol was bound to the GAA wildtype at the set parameters the calculated binding energy was -71.59 kcal per mol. When miglitol was then bound to the benign variants of the GAA proteins, at the parameters stated, the average calculated binding energy was -55.88 kcal per mol (p=0.869), with a standard deviation of 41.06. In the benign group of variant-ligand binding, two of the variants returned a binding energy that was substantially weaker than the wild-type binding energy. These variants were benign variants 4 and 8 (variant numbers: 92467 and 92482 respectively). The mutation of benign variant 4 was Glycine to Serine (residue number 576). The mutation of benign variant 8 was threonine to isoleucine (residue number 927). These complexes had a binding energy of -56.97 and -60.20 kcal per mol respectively. These binding energies were at least 15% weaker than the wild-type binding energy. The average root-mean-square deviation (RMSD) for the miglitol ligands bound to the benign group of GAA variants was 4.55, with a standard deviation of 8.07. When the pathogenic group of variants were bound to miglitol the average binding energy was -67.51 (p=0.869). This group of data had a standard deviation of 5.61. In the pathogenic group of variants, 3 docked complexes had binding energy that was 15% weaker than the binding energy of the wild-type complex. These variants were pathogenic variants 1, 6 and 9 (variant numbers: 92483, 4036 and 972790 respectively). The pathogenic variant 1 complex had a binding energy of -61.18 kcal per mol. Pathogenic variant 6 had a binding energy of -61.03 kcal per mol. The complex formed when the pathogenic variant 9 was bound to miglitol, it had a binding energy of -56.98 kcal per mol. All of these binding energies were considerably weaker than the wild-type binding energy. The mutation of pathogenic variant 6 was Glycine to Arginine (residue number 293). The mutation of pathogenic variant 9 was Glycine to Arginine (residue number 335). The average root-mean-square deviation (RMSD) for the miglitol ligands bound to the pathogenic group of GAA variants was 4.51, with a standard deviation of 8.04 (Table 11).

GAA (5nn8)						
Variant Type	**Variant Number**	**Variant Residue Number**	**Acarbose Binding Energy**	**Acarbose RMSD**	**Miglitol Binding Energy**	**Miglitol RMSD**
Wildtype	-		-89.4	4.447	-71.59	
Control 1	281330	451	-81.41	15.488	-71.39	0.066
Control 2	92488	223	-78.76	16.941	-71.26	0.093
Control 3	284497	429	-94.03	2.031	-71.35	0.051
Control 4	92467	576	-78.09	8.061	-56.97	24.058
Control 5	4030	689	-97.57	1.681	-71.4	0.057
Control 6	92476	780	-90.61	4.255	-71.37	0.055
Control 7	92477	816	-76.41	5.53	-64.54	3.366
Control 8	92482	927	-88.91	4.597	-60.2	4.046
Control 9	283498	448	-81.78	15.451	-69.34	13.66
Control 10	714463	449	-81.89	15.475	-71.39	0.07
Pathogenic 1	92483	103	-67.17	11.221	-61.18	4.059
Pathogenic 2	847865	108	-90.85	4.537	-71.41	0.066
Pathogenic 3	189065	219	-88.53	4.982	-71.36	0.041
Pathogenic 4	189188	224	-76.68	15.521	-67.55	3.281
Pathogenic 5	956209	291	-71.21	12.575	-71.37	0.061
Pathogenic 6	4036	293	-78.23	17.033	-61.03	13.384
Pathogenic 7	180144	299	-88.76	4.37	-71.4	0.068
Pathogenic 8	188797	309	-69.18	13.123	-71.38	0.068
Pathogenic 9	972790	335	-91.24	4.211	-56.98	24.056
Pathogenic 10	284093	355	-86.09	5.114	-71.41	0.062

Table 11, This table shows the binding energies of the GAA wildtype and 20 of its variants when they were bound to the ligands acarbose and miglitol. This table also shows the RMSD of the ligands in the complexes created. The ClinVar variant number and the variant residue number are also presented in this table. The figures highlighted in yellow are binding energies that are at least 15% lower than the wildtype binding energy. (Pathogenic Acarbose p= 0.575, Control Acarbose p= 0.406015138, Pathogenic Miglitol p= 0.113, Control Miglitol p= 0.113)

5 variants had binding energies that were at least 15% weaker than the wild type when bound to the respective ligand. This means that 25% of the GAA dockings were affected by variants. Although there was no statically significant difference between pathogenic variant docking and benign variant docking, the fact that 25% of the variants had their docking ability considerably affected, would illustrate that variations, have a serious effect on the protein's ability to bind to a ligand, and thus could mean that they would have an effect on the binding between drugs and proteins, which could lead to a drug being less effective.

The control variant 4 and pathogenic variant 9 had the weakest binding energies when bound to miglitol, this is to be expected as the residue numbers of these mutations are close to the active site, pathogenic variants 5 and 6 are also located close to the active site and they caused the binding strength of the protein and the ligand to be reduced by more than 15%. From this, it can be inferred that the location of the variant plays a key role in the effect of a protein's binding energy.

3.4.2) SCN2A

3.4.2.1) SNC2A and Tramadol

The next protein-ligand complexes that were simulated were the Sodium Voltage-Gated Channel Alpha Subunit 2 (SCN2A, (2kav PDB)) and the compound tramadol. This was done via the box method, where the binding location of the SCN2A protein was set to the coordinates x = 5.01, y = -4.73, z = -0.12. The size of the binding area was set to x = 35.60, y = 35.35, and z = 29.48 (arbitrary units). When tramadol was bound to the SCN2A wildtype the calculated binding energy was -66.13 kcal per mol. Tramadol was then bound to the benign variants of SCN2A, the average calculated binding energy for these complexes was -64.20 kcal per mol (p=0.975), with a standard deviation of 2.79. In the benign cohort of variant-ligand complexes, none of the docking simulations returned a binding energy that was notably different to the wild-type binding energy. The average RMSD of the tramadol ligands bound to the benign group of SCN2A variants was 7.38, with a standard deviation of 4.85. Next tramadol was docked with the pathogenic variants of the SCN2A proteins, and the program returned an average binding energy of -64.15 kcal per mol (p=0.975) for these complexes, with a standard deviation of 4.26. In the pathogenic group of variant-ligand binding, there were also no docked complexes that had a binding energy that was notably different to that of the wild type. The average RMSD for the tramadol ligands bound to the pathogenic group of SCN2A variants was 12.82, with a standard deviation of 7.12.

3.4.2.2) SNC2A and Zonisamide

SCN2A protein was then bound to the compound zonisamide. This was done using the box method. The binding site location for this docking simulation was set to the same coordinates as that of the tramadol binding simulation, x = 5.01, y = -4.73, z = -0.12. The size of the binding area was set to x = 35.60, y = 35.35, and z = 29.48 (arbitrary units). When zonisamide was bound to the SCN2A wildtype the calculated binding energy was -54.75 kcal per mol. zonisamide was then bound to the benign variants of SCN2A, the average binding energy was -54.43 kcal per mol (p=0.998), this group of results had a standard deviation of 4.00. In the benign group of variant-ligand complexes, none of the docking simulations had a binding energy that was substantially lower than the wild-type binding energy. The average RMSD for the zonisamide ligands bound to the benign group of SCN2A variants was 12.09, this data had a standard deviation of 8.69. The pathogenic variants of the SCN2A were then bound to zonisamide. When the docking was conducted the average calculated binding energy was -54.42 kcal per mol (p=0.998), with a standard deviation of 3.82. In the pathogenic group of variant-ligand binding, there were also no docked complexes that had a binding energy that was substantially lower than that of the wild type. The average RMSD for the Zonisamide ligands, when bound to the pathogenic SNC2A variants was 10.15, with a standard deviation of 9.70 (Table 12).

SCN2A						
Variant Type	**Variant Number**	**Variant Residue Number**	**Tramadol Binding Energy**	**Tramadol RMSD**	**Zonisamide Binding Energy**	**Zonisamide RMSD**
Wildtype	-		-66.13	-	-54.75	-
Control 1	207026	1812	-64.81	6.284	-51.71	11.705
Control 2	916178	1813	-67.46	6.219	-55.83	11.434
Control 3	431830	1831	-62.77	12.631	-55.05	1.571
Control 4	1316180	1843	-66.04	3.361	-62.56	11.136
Control 5	808853	1844	-65.88	3.504	-51.37	13.596

Control 6	493290	1849	-67.63	0.509	-51.28	22.751
Control 7	493291	1850	-63.57	6.042	-58.88	3.56
Control 8	1318870	1860	-59.67	15.063	-50.11	29.483
Control 9	1176401	1855	-59.92	13.83	-55.62	3.404
Control 10	452500	1780	-64.26	6.343	-51.86	12.298
Pathogenic 1	801787	1778	-68.37	6.327	-56.53	11.459
Pathogenic 2	207025	1780	-60.21	15.97	-57.24	5.052
Pathogenic 3	495262	1851	-68.46	6.369	-50.62	27.692
Pathogenic 4	813765	1872	-65.67	6.158	-50.48	2.481
Pathogenic 5	207028	1882	-68.64	6.327	-49.17	27.134
Pathogenic 6	207029	1882	-59.77	19.689	-50.12	11.97
Pathogenic 7	934576	1781	-59.68	13.217	-56.16	3.841
Pathogenic 8	654341	1804	-62.81	27.536	-57.95	4.942
Pathogenic 9	533496	1809	-68.84	10.959	-57.03	5.028
Pathogenic 10	1320953	1819	-59.04	15.659	-58.93	1.883

Table 12, This table shows the binding energies of the SCN2A wildtype and 20 of its variants when they were bound to the ligands Tramadol and Zonisamide. This table also shows the RMSD of the ligands in the complexes created. The ClinVar variant number and the variant residue number are also presented in this table. (Pathogenic Tramadol p= 0.527, Control Tramadol p= 0.752, Pathogenic Zonisamide p= 0.940, Control Zonisamide p= 0.752)

From these results (Table 12), we can infer that these variants of the SCN2A protein did not have any effect on the protein's ability to bind to its ligands. This would mean that patients that possess these variants would likely not observe any differences in drug binding. Meaning a drug therapy designed to interact with SCN2A would not be any less effective on a patient with these variants.

3.4.3) BRAF

3.4.3.1) BRAF and Sorafenib

Serine/threonine-protein kinase B-RAF (BRAF (2fb8 PDB)) was then docked with sorafenib, this docking simulation was conducted using the box method. The active site of this protein was determined to have the coordinates of x = -16.07, y = 6.56, and z = -4.39. The size of the binding area was set to x = 20.24, y = 18.74, and z = 13.48 (arbitrary units). When sorafenib was bound to the BRAF wildtype, the calculated binding energy was -76.63 kcal per mol. Next, the docking simulation between sorafenib and the benign variants of BRAF was conducted, the average binding energy of these complexes was -89.63 kcal per mol (p=0.343), and this group of results had a standard deviation of 5.90. In this docking simulation, the results returned were surprising, almost all the complexes returned a binding energy that was considerably stronger than the wildtype's. The only docked complexes that provided a binding strength that was not considerably stronger than the wild-type binding energy was benign variant 8 (variant number: 40386). The mutation of this variant was Isoleucine to Threonine (residue number 632). All the other benign variants had binding strengths that were stronger than the wild-type binding energy by 15% or more. The average RMSD for the sorafenib ligands bound to the benign group of BRAF variants was 14.25, this data had a standard deviation of 1.10. Sorafenib was then bound to the pathogenic variants of BRAF, and the average calculated binding energy was -91.95 kcal per mol (p=0.343), with a standard deviation of 4.69. Similar to the benign group, in the pathogenic group of variant-ligand binding, almost all the complexes returned binding energies that were considerably stronger than the wildtype's. The only 2 complexes that did not, were pathogenic variants 3 and 7 (variant numbers: 177844 and 666569, respectively). The mutation of pathogenic variant 3 was Leucine to phenylalanine (residue number 485). The mutation of pathogenic variant 7 was Tryptophan to Serine (residue number 531). All the other binding strengths were at least 15% higher than the wild-type binding energy. The average RMSD for the sorafenib ligands bound to the pathogenic BRAF variants was 13.77, with a standard deviation of 4.79.

3.4.3.2) BRAF and Vemurafenib

BRAF was then docked to the compound vemurafenib, this was carried out using the box method. The binding site location for this docking simulation was set to the same coordinates as that of the sorafenib docking, x = -16.07, y = 6.56, z = -4.39. The volume of the binding area was set to x = 20.24, y = 18.74, and z = 13.48 (arbitrary units). When vemurafenib was bound to the BRAF wildtype binding energy was -69.28 kcal per mol. The subsequent set of docking was vemurafenib and the benign variants of BRAF, the mean binding energy was -89.18 kcal per mol (p=0.651), with a standard deviation of 5.50. This result was unexpected as almost all the complexes returned a binding energy that was considerably stronger than the wildtype's. The only variant that did not bind stronger with its ligand was benign variant 7 (variant numbers: 55794). This complex had a binding energy of -75.3 kcal per mol. This variant was Asparagine to Histidine (residue number 621). The rest of the benign variants had binding energies that were at least 15% more efficient than the wild-type binding energy. The RMSD for the vemurafenib ligands bound to the benign group of BRAF variants was 14.30, with a standard deviation of 1.37. Vemurafenib was then docked with the pathogenic group of BRAF variants. The average binding energy was -88.06 (p=0.651), and similar to the benign only 1 docked complex had a binding energy that was similar to the binding energy of the wild-type complex. This was pathogenic variant 8 (variant number: 44180). The pathogenic variant 8 complexes had a binding energy of -76.33 kcal per mol. This mutation was Histidine to Tyrosine (residue number 574). The average root-mean-square deviation (RMSD) for the vemurafenib ligands bound to the pathogenic group of BRAF variants was 13.92, with a standard deviation of 1.14 (Table 13).

BRAF						
Variant Type	Variant Number	Variant Residue Number	Sorafenib Binding Energy	Sorafenib RMSD	Vemurafenib Binding Energy	Vemurafenib RMSD
Wildtype	-		-76.63	-	-69.28	-
Control 1	802375	477	-90.01	11.662	-94.3	15.177
Control 2	1299096	705	-94.71	14.478	-87.58	12.937
Control 3	547184	501	-94.23	16.058	-92.79	13.474
Control 4	503530	536	-86.54	14.441	-89.27	14.643
Control 5	864012	581	-94.35	14.49	-93.13	13.474
Control 6	239870	594	-84.3	13.543	-86.71	14.101
Control 7	55794	621	-93.94	14.576	-75.3	15.083
Control 8	40386	632	-77.62	14.283	-88.82	12.383
Control 9	1050479	637	-85.72	14.493	-92.32	17.191
Control 10	960949	664	-94.87	14.484	-91.61	14.569
Pathogenic 1	13962	462	-94.87	14.49	-87.38	12.324
Pathogenic 2	44803	469	-94.73	14.502	-89.71	14.689
Pathogenic 3	177844	485	-82.52	1.987	-91.65	14.546
Pathogenic 4	40372	499	-92.7	14.628	-91.75	13.535

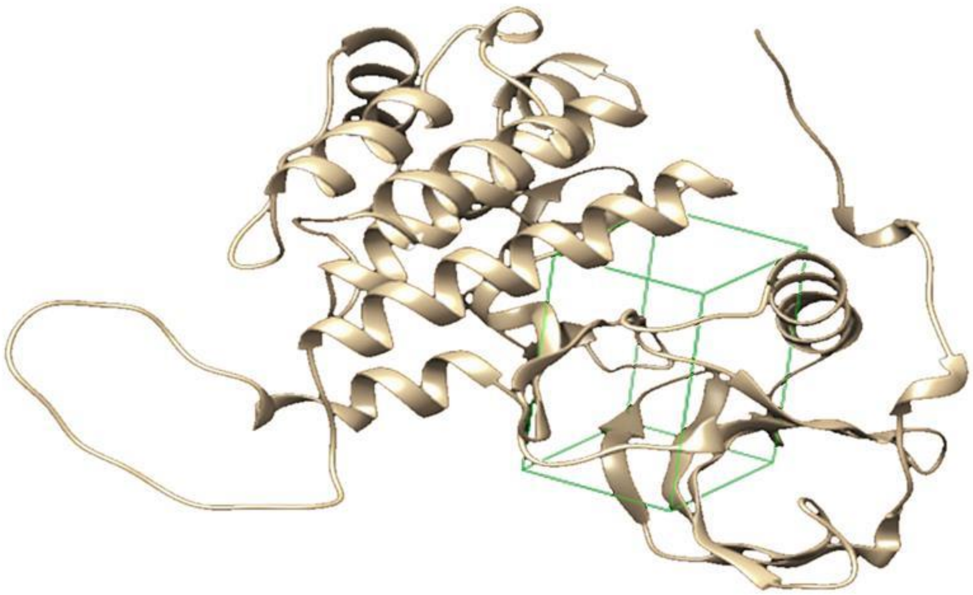
Pathogenic 5	40373	501	-96.04	13.551	-89.23	14.53
Pathogenic 6	40375	505	-95.04	16.419	-91.24	14.365
Pathogenic 7	666569	531	-85.48	14.464	-80.38	15.369
Pathogenic 8	44810	574	-94.66	14.51	-76.33	11.761
Pathogenic 9	13979	581	-94.58	14.491	-92.62	13.493
Pathogenic 10	13961	600	-88.87	18.643	-90.31	14.56

Table 13, This table shows the binding energies of the BRAF wildtype and 20 of its variants when they were bound to the ligands sorafenib and vemurafenib. This table also shows the RMSD of the ligands in the complexes created. The ClinVar variant number and variant residue number are also presented in this table. The figures highlighted in blue are binding energies that are within 15% of the wild-type binding energy. (Pathogenic sorafenib p= 0.065, Control sorafenib p= 0.114, Pathogenic Vemurafenib p= 0.114, Control Vemurafenib p= 0.114)

When BRAF was docked with sorafenib and Vemurafenib, the average binding energies of both groups of variants were higher than the binding energy of the wild type, this was very surprising. A possible explanation for these results could be due to an incorrectly sequenced protein PDB or possibly an error in the PLANTS program. This is a possible limitation of bioinformatics as in some cases, it may be unreliable or produce abnormal results.

3.4.4) KIT

3.4.4.1) KIT and Fludiazepam



(Figure 10, This Figure shows the protein structure of dock-prepped KIT wild-type protein and the docking volume, in which the ligands would be bound. The co-ordinate of this box is x = 17.19, y = -16.07, z = 3.04. The size of the binding area is x = 11.44 y = 19.60, and z = 13.36 (arbitrary units)).

The protein Receptor Tyrosine Kinase (KIT (6mob PDB)) was then docked to the compound fludiazepam, using the box method. The active site of this protein was determined to have the coordinates x = 17.19, y = -16.07, and z = 3.04. The volume of the binding site was set to x = 11.44 y = 19.60, and z = 13.36 (arbitrary units) (Figure 10). When bound to fludiazepam, the KIT wildtype had a binding energy of -67.38 kcal per mol. Fludiazepam was then docked onto the benign variants of KIT. The average binding energy of these complexes was -69.24 (p=0.364), this group of results had a standard deviation of 3.95. In the benign cohort of variant-ligand binding, there was one of the docked complexes provided a binding energy that was considerably stronger than the wild-type binding energy. This was benign variant 7 (variant number: 953798). This mutation was Aspartic acid to Alanine (residue number 820). All the other benign variants had binding energies that were within 15 % of the wild-type binding energy. The average RMSD of the fludiazepam ligands bound to the benign KIT variants was 20.18, this data had a standard deviation of 5.12. Fludiazepam was then bound to the pathogenic variants of the KIT proteins, at the same coordinates stated. The average binding energy for these complexes was -71.18 (p=0.364) and the standard deviation of this data was 5.27. In the pathogenic group of variant-ligand binding, 3 of the docked complexes returned surprising binding energies. The binding efficiency of these variants was more than 15% higher than the wild-type binding energy. These variants were pathogenic variants 2, 4 and 10 (variant numbers: 13862, 375933 and 375928 respectively). The pathogenic variant 2 complex had a binding energy of -74.25 kcal per mol. Pathogenic variant 4 had a binding energy of -79.61 kcal per mol. The complex formed when the pathogenic variant 10 was bound to fludiazepam had a binding energy of -79.62 kcal per mol. The mutation of pathogenic variant 2 was Glutamic Acid to Lysine (residue number 839). The mutation of pathogenic variant 4 was Alanine to Proline (residue number 829). The mutation of pathogenic variant 10 was Aspartic acid to Tyrosine (residue number 820). The average RMSD of the fludiazepam ligands bound to the pathogenic group of KIT variants was 19.64, this data had a standard deviation of 6.93

3.4.4.2) KIT and Ergocalciferol

The KIT was then subsequently bound to the compound ergocalciferol. The binding site location for this docking simulation was set to the same coordinates as that of the fludiazepam docking simulation. When ergocalciferol was bound to the KIT wildtype the calculated binding energy was -85.24 kcal per mol. Then Ergocalciferol was docked onto the benign variants of KIT, and the average binding energy returned was -85.56 kcal per mol (p=0.429), this group of results had a standard deviation of 2.93. In the benign group of docked complexes, none of the docking simulations had a binding energy that was considerably lower than the wild-type binding energy. The average RMSD for the ergocalciferol ligands bound to the benign group of KIT variants was 18.84, this data had a standard deviation of 6.58. Subsequently, ergocalciferol was bound to the pathogenic variants of KIT. When the docking was conducted the average calculated binding energy was -84.42 kcal per mol (p=0.429), with a standard deviation of 3.33. None of these docked complexes had a binding energy that was notably lower than that of the wild type. The average RMSD of the ergocalciferol ligands bound to the pathogenic variants of KIT was 16.59 with a standard deviation of 8.63 (Table 14).

KIT						
Variant Type	**Variant Number**	**Variant Residue Number**	**Fludiazepam Binding Energy**	**Fludiazepam RMSD**	**Ergocalciferol Binding Energy**	**Ergocalciferol RMSD**
Wildtype	-		-67.38	-	-85.24	-
Control 1	1062801	869	-68.55	20.262	-87.9	20.269
Control 2	237266	874	-69.34	20.254	-88.33	20.483
Control 3	1061062	877	-67.61	22.948	-80.62	9.42
Control 4	578876	891	-67.67	22.906	-82.12	10.036
Control 5	838460	899	-69.14	20.247	-88.29	20.471
Control 6	577278	804	-69.43	20.23	-84.75	25.813

Control 7	953798	820	-79.51	25.526	-86.45	25.738
Control 8	528546	830	-64.06	6.272	-82.28	10.253
Control 9	409781	844	-69.45	20.217	-88.27	20.488
Control 10	1037520	867	-67.67	22.904	-86.57	25.442
Pathogenic 1	13861	847	-69.39	20.211	-79.28	20.515
Pathogenic 2	13862	839	-74.25	24.874	-83.06	2.685
Pathogenic 3	13864	584	-69.42	20.198	-88.33	20.178
Pathogenic 4	375933	829	-79.61	25.544	-88.08	20.366
Pathogenic 5	13843	664	-69.31	20.259	-84.07	22.338
Pathogenic 6	13863	816	-69.47	20.209	-86.65	25.826
Pathogenic 7	13858	796	-68.23	10.979	-80.41	1.557
Pathogenic 8	13866	642	-69.99	24.433	-88.38	20.24
Pathogenic 9	375918	569	-62.52	4.262	-83.92	22.324
Pathogenic 10	375928	820	-79.62	25.499	-82.06	9.913

Table 14, This table shows the binding energies of the KIT wildtype and 20 of its variants when they were bound to the ligands Fludiazepam and Ergocalciferol. This table also shows the RMSD of the ligands in the complexes created. The ClinVar variant number and variant residue number are also presented in this table. The figures highlighted in green are binding energies that are more than 15% higher than the wildtype binding energy. (Pathogenic fludiazepam p= 0.206, Control fludiazepam p= 0.509, Pathogenic ergocalciferol p= 0.920, Control ergocalciferol p= 0.820)

In this result when fludiazepam was bound to the variants of KIT, some complexes provided bindings that were more than 15% stronger than the wildtype, binding energy (Table 14). It is possible that this is due to the mutation causing an increased affinity to the ligand. In drug metabolism, an increased affinity can be more dangerous than a decreased affinity as the drug may be upregulated or unable to be metabolised, which can lead to death (Zhang & Tang, 2018). As there are no instances of increased binding energy when KIT is bound to ergocalciferol, it can be assumed that this increase in affinity to a ligand is ligand-dependent, thus supporting the argument for precision medicine. As patients would have different responses depending on the variants they possess and the type of drug therapy.

The residue numbers of the 4 variants that had binding energies that were more than 15% stronger than the wildtype were all between 820-839, this infers that residues in that region may play a role in binding, even though these residues are not directly in the active site.

3.4.5) PAH

3.4.5.1) PAH and Droxidopa

The next protein-ligand complexes that were produced were Phenylalanine Hydroxylase (PAH, (1j8u PDB)) and the compound droxidopa. This was done via the box method, where the binding site location of the PAH protein was set to the coordinates x = -5.47, y = 24.79, and z = 6.32. The volume of the binding site was set to x = 12.65, y = 14.00, and z = 7.09 (arbitrary units). When droxidopa was bound to the PAH wildtype the calculated binding energy was -74.75 kcal per mol. Droxidopa was then docked onto the benign variants of the PAH proteins, the average calculated binding energy was -78.03 kcal per mol (p=0.671), with a standard deviation of 2.91. In the benign cohort, none of the docking simulations provided a binding energy that was considerably weaker than the wild-type binding energy. The average RMSD of the droxidopa ligands bound to the benign variants of PAH was 3.74, with a standard deviation of 1.16. droxidopa was then docked to the pathogenic variants of PAH, the average binding energy of these complexes was -77.41 kcal per mol (p=0.671), with a standard deviation of 3.28. Again, all the docked complexes had a binding energy that was within 15% of the wild-type binding energy. The average RMSD of the droxidopa ligands bound to the pathogenic variants of PAH was 3.87, with a standard deviation of 0.99.

3.4.5.2) PAH and Sapropterin

The last protein-ligand complexes that were produced using the box method were the PAH-sapropterin complexes. The binding site location for this docking simulation was set to the same coordinates as the droxidopa binding simulation. The binding energy when sapropterin was bound to the PAH wildtype was -69.99 kcal per mol. Next, sapropterin bound to the benign variants of PAH, and the average binding energy was -76.84 kcal per mol (p=0.328), this group of results had a standard deviation of 0.35. All of the docking simulations had binding energies that were within 15% of the wild-type binding energy. The average RMSD for the sapropterin ligands bound to the benign variants of PAH was 2.63, this data had a standard deviation of 1.80. Sapropterin was then bound to the pathogenic variants of the PAH proteins. When the docking was conducted the average calculated binding energy was -76.62 kcal per mol (p=0.328), with a standard deviation of 0.60. In the pathogenic group of variant-ligand binding, there were also no docked complexes that had a binding energy that was considerably lower than that of the wild type. The average RMSD of the sapropterin ligands bound to the pathogenic group of PAH variants was 2.02, with a standard deviation of 0.72 (Table 15).

PAH						
Variant Type	**Variant Number**	**Variant Residue Number**	**Droxidopa Binding Energy**	**Droxidopa RMSD**	**Sapropterin Binding Energy**	**Sapropterin RMSD**
Wildtype	-		-74.75	-	-69.99	-
Control 1	102851	274	-80.21	2.524	-77	6.036
Control 2	975467	274	-79.61	3.836	-76.09	1.773
Control 3	763076	337	-78.68	4.529	-77.25	6.021
Control 4	120290	290	-74.68	3.096	-76.91	1.789
Control 5	872846	356	-79.75	2.501	-76.95	1.81
Control 6	102469	340	-81.8	2.501	-76.95	1.816
Control 7	937913	418	-76.62	2.984	-76.94	1.814
Control 8	932266	412	-79.84	5.059	-76.98	1.815
Control 9	102544	392	-76.51	5.06	-76.36	1.538
Control 10	99161	374	-72.33	5.3	-77	1.821
Pathogenic 1	102579	417	-72.6	5.336	-76.11	1.776
Pathogenic 2	552488	419	-79.79	4.721	-77.01	1.817
Pathogenic 3	940659	413	-76.09	4.926	-76.96	1.792
Pathogenic 4	853581	392	-79.66	3.841	-76.99	1.81
Pathogenic 5	108515	225	-79.64	3.883	-76.95	1.818
Pathogenic 6	102874	282	-79.81	4.359	-76	4.065
Pathogenic 7	102866	280	-81.9	2.49	-76.96	1.805
Pathogenic 8	556296	267	-76.61	2.988	-77.03	1.815
Pathogenic 9	619150	263	-72.25	3.616	-75.31	1.686

Pathogenic 10	619163	233	-75.7	2.534	-76.91	1.811

Table 15, This table shows the binding energies of the PAH wildtype and 20 of its variants when they were bound to the ligands Droxidopa and Sapropterin. This table also shows the RMSD of the ligands in the complexes created. The ClinVar variant number is also presented in this table. (Pathogenic Droxidopa p= 0.314, Control Droxidopa p= 0.460, Pathogenic Sapropterin p= 0.112, Control Sapropterin p= 0.113)

Similar to the results from the SCN2A protein, these results (Table 15), allow for an inference to be made that patients who possess these variants would likely not observe any differences in drug binding. As the results show that none of the docked variants, had a considerable effect on the binding ability of PAH. This would mean that drug therapies created to interact with PAH should work as designed, and would not face a decrease in effectiveness due to these variants

4) Discussion and Future Works

4.1) Discussion

4.1.1) Bioinformatic Pipeline

There is a lot to take away from the results produced in this research project as the project investigated many avenues within bioinformatics, precision medicine and polymorphic variation. This project hypothesised the development of a bioinformatic tool, that retrieves and integrates vast amounts of information from multiple databases and could be used to aid many clinical research investigations. This project also had a hypothesis that single nucleotide variants in proteins have an adverse effect on the binding between proteins and drug compounds, which can severely hinder how effective a drug is. In this project, the bioinformatic pipeline provided a lot of key information, that allowed for the investigation into the effect of protein docking to take place. This information included all the known variants of each protein, along with a large amount of information on each variant. It also provided all the compounds that are associated with each protein and an abundance of information on the compounds. This showed that bioinformatics tools are very useful. The pipeline was able to show its real potential to drive forward our use of precision medicine, during the case studies. This showed the pipeline was able to gather vast amounts of information, sort the information and incorporate multiple databases. The pipeline was able to gather enough information to find the most prevalent pathogenic variants, using one database of information and then it was able to find compounds that are associated with these variants, which incorporates another database. The pipeline also demonstrated its ability to start the data collection and flow of information from any part of the database. This means the user does not have to start from the same point and is able to start the collection of information from any point. An example of this was shown in the Acute Myeloid Leukaemia case study, where the pipeline was able to start with a disorder and then incorporate other databases involved to find, the proteins associated with the disorder and then the compounds associated with the proteins. The pipeline is able to retrieve every known variant for all the proteins input, and due to the pipeline's ability to select and sort data, it is possible to display pathogenic variants of a protein, that occur in a select range of residue numbers. The sheer power of the pipeline serves as an example of the full potential of bioinformatics and the implications it can have for precision medicine.

The process of manually gathering proteins, variants, compounds, disorders, and treatment information, was very effective and very useful. This allowed for the optimisation of the pipeline, allowing it to be as efficient as possible. Due to this, the pipeline can present all the key information available from the databases that were scraped. All the entities present in the entity-relationship diagram, and later implemented in the pipeline were included as they proved a wholistic view of each protein, compound, and disorder. This makes the pipeline powerful and unique, as no other database is able to provide this combination of data simultaneously.

There are no other databases, online, that possess the specific utility of this pipeline. Unlike, many other databases such as ClinVar and UniProt, the database produced in this project allows for the seamless integration of multiple different datasets of information. This means that when searching for variant information, the user can also be informed of drug compound information for drugs associated with the proteins. Subsequently, if a user was to investigate a specific disorder, the pipeline can provide information on protein and drug compounds associated with the disorder. This is a feature currently unavailable in databases like ClinVar and MedGen, which is evidence of how powerful the database is.

4.1.2) The difference between pathogenic and benign variants

For all the groups of protein-ligand docking conducted, an ANOVA was undertaken on the data. The null hypothesis of this ANOVA was that there was no significant difference between the benign variants and pathogenic variants. When the ANOVAs were conducted on all the docked proteins none of them returned an ANOVA significance value that was lower than p=0.05, this meant that the null hypothesis had to be accepted. This result could be due to several reasons, one being that the pathogenicity of the variant is not related to the protein's binding ability, therefore meaning that the binding energy of the protein is not affected by the mutation but rather, its ability to carry out its function more widely.

In the dockings with the GAA protein, there were more variants from the pathogenic cohort than the benign cohort that had binding energies that were at least 15% lower than the wild-type docking. When an ANOVA was conducted on this data, the p values were p=0.575 and p=0.113 (when bound to Acarbose and Miglitol, respectively) meaning there was no statistically significant difference between the benign variants and pathogenic variants. These results allowed for the conclusion to be made that although as a whole the difference between the pathogenic docking and the benign docking was not large, individually more pathogenic variants carry a greater risk of having a negative effect on the binding ability of a protein and its ligand. This shows that on some occasions these approaches might be useful to implement in precision medicine strategies, as a protein's ability to bind and dock with the drug treatments can be severely hampered by polymorphic variations. Precision medicine would be very useful in this scenario, as genetic screening techniques (such as quantitative PCR) can be used, this would mean patients with the variants, that are known to possess a variant associated with impaired binding of a specific ligand would be detected, allowing for a plan of treatment that consists of drug treatments that are more likely to be effective to be developed for them. It can also stop the wasteful use of drugs on patients who would not find them effective due to polymorphic variation, this can save a lot of money (Akhmetov & Bubnov, 2015).

The results also showed that 25% of the GAA dockings were affected by variants, from this it was inferred that variations could have a serious effect on protein binding and therefore have an adverse effect on drug binding, which could lead to a drug being less effective. This finding is echoed in a research project conducted on Alpha-1-Acid Glycoprotein, where they stated age-related changes to drug binding, had caused Alpha-1-Acid Glycoprotein to bind less efficiently to drugs (Smith & Waters, 2018).

4.1.3) The effect variant location has on binding

In the GAA docking simulation, when bound to miglitol, control variant 4 and pathogenic variant 9 had the lowest binding energy, the residue numbers of these mutations are close to the active site. Pathogenic variants 5 and 6 are also located close to the active site and they caused the binding energy of the protein to be reduced by more than 15%. This implies that the location of the mutation plays a role in how much the binding site is affected. This is what was expected of the results, as changes to the amino acid near the binding site could affect the bonds produced when a ligand is docked. This is similar to what was found in the bind site analysis of human carbonic anhydrases, where it was stated that moieties needed to form H-bonds with H64 for CAA to work (Petreni et al., 2021).

In the docking simulation between the variants of KIT and fludiazepam, some complexes provided binding energies that were more than 15% higher than the wild-type (Table 14). The understanding of this is that the mutation could have caused an increased affinity to the ligand. This can be very dangerous as an increased affinity for a drug can cause its effects to be upregulated, which can cause serious problems, and sometimes even death (Zhang & Tang, 2018). The results, when KIT was bound to Ergocalciferol, showed no variant with a binding energy that was considerably different to the wild-type binding energy. From this, it was gathered that the increase in affinity was dependent on the ligand thus reinforcing the case for consideration in precision medicine, as this shows patients would have different responses to the ligands depending on the variants they possess.

The residue numbers of the four variants that had binding energies that were more than 15% higher than the wildtypes were all between 820-839. This was not a region next to the drug binding site however, this could mean, that region may play a role in binding, even though residues are not directly in the active site, as this is not unheard of. There have been some cases where outside active site mutations have increased enzyme activity, it is thought that this occurs due to changes in the enzyme's secondary structure (Ali, Azam, & Khan, 2018).

In the dockings conducted on SCN2A and PAH, none of the docked complexes returned a binding energy that was more than 15% lower than the binding energy of the wild type. From this result, it can be inferred that none of the variants studied appeared to have any effect on the protein-ligand binding. Of the 20 randomly chosen variants of PAH, only 3 of them were located near the PAH binding site, pathogenic variants 6, 7, and 10 (residue numbers: 282, 280 and 237 respectively). The lack of variant binding energies that were more than 15% lower than the wildtype binding energy could be attributed to there being, only a few variant residues located near the binding site. This supports the argument, that the location of a variant residue can heavily influence the protein's ability to bind to a ligand and therefore, would affect drug binding ability. Contrastingly, all the variants of SCN2A were in the binding site, and none of the docked variant complexes returned a binding energy that was more than 15% lower than the wild-type binding energy. A reason for this could be due to the weak binding that was present. Of all the proteins docked, SCN2A had the lowest wild type, average benign variant and average pathogenic variant binding energy. From this, it can be inferred the bonds involved in the docking between the protein and its ligands were already weaker than in the other proteins. Therefore, the introduction of a variant, to the protein sequence, perhaps has a relatively lower effect on protein binding ability, than it would in a complex with strong protein-ligand binding forces.

On the contrary, it has been documented that mutations outside the active site may still have an impact on the proteins' ability to binding to ligands. Structural studies on 45 HIV-1 protease mutants, showed that 35 of them were located outside the active site. These mutations were considered secondary or accessory mutations and were thought to indirectly impact inhibitor binding while also affecting the enzyme's fitness and stability (Ragland et al., 2014). In another study, random mutants of β-glycosidase were screened to investigate the effect it has on binding with the substrate glycone. Among the mutations selected eight occurred in the C-terminal half of β-glycosidase and only two were at the active site. Enzyme kinetics confirmed that these mutations resulted in a change in the preference for glycone. (kcat/Km fucoside)/ (kcat/Km glucoside) ratios, showed that the mutations in the active site and the mutations outside the active site had similar effects on substrate specificity. Based on the data they collected they concluded that even mutations far from the active site affected the binding of glycosidase to its substrates (Mendonça & Marana, 2011).

In some clinical scenarios, a weaker bond between the protein and its drug compound may be preferable. Such a situation would be when a patient has been treated with a drug compound of considerable strength or toxicity. In this scenario, if the proteins' ability to bind to the compound had increased, they may encounter severe side effects and thus a weaker interaction between the protein and the compound would be preferable. Using molecular medicine techniques could, however, nullify this, as a complete understanding of a patient's genome, would allow for the identification of specific proteins that would have weaker interactions, and thus their treatment could be modified to incorporate this difference.

The results gathered from the use of molecular docking, suggest that there is a relationship between the location of a variant to the drug binding site and the impact on drug binding. This supports one of the main hypotheses of this study. This has been concluded because the results show most of the variants that have binding energies that considerably differed from the wildtype's were found near the drug binding site. Also, SCN2A had no variant complexes with a binding energy that differed by at least 15% from the wild type, and a large majority of its variants were located away from the drug binding site. This conclusion was able to be reached by the use of bioinformatic tools, further illustrating that bioinformatic tools and approaches can aid the development of precision medicine approaches.

4.1.4) The effect of amino acids mutations

When the docking between GAA and acarbose was conducted, the variants that had a binding energy that was at least 15% lower than the wildtype had the mutations of cysteine to glycine (residue number 103), arginine to tryptophan (residue number 224), leucine to phenylalanine (residue number 291), glycine to arginine (residue number 309) and valine to isoleucine (residue number 816). In 3 out of 5 of these mutations, the polarity of the side chain of the amino acid had been altered (cysteine to glycine, arginine to tryptophan and glycine to arginine). From this, it can be inferred that the changes in the side chain polarity of specific residues play a role in a protein's ability to bind to drug therapies. This further demonstrates the hypothesis that polymorphic variants can adversely affect protein-drug binding, thus making the drug treatment less effective. This is because a change in polarity can affect the bonds involved in docking, thus affecting binding energy (Raschka, Wolf, Bemister-Buffington & Kuhn, 2018). This is further supported in the docking between GAA and Miglitol, where all the variants with binding energies more than 15% lower than the wildtype had mutations that cause an alteration to the polarity of the wildtype residue variants were glycine to arginine (residue number 293), glycine to arginine (residue number 335), glycine to serine (residue number 576), threonine to isoleucine (residue number 927) and cysteine to glycine (residue number 103). The substitutions of residue numbers 293, 335 and 576 are also located close to the binding site. This could be a reason why their binding energies are more than 15% lower than the wild type's. In real-world terms, this could be dangerous, if a patient were to possess a polymorphic variant that caused a change to a side chain of one of the amino acid residues, this could heavily impact its ability to bind to drugs, that are used as treatment. In a study conducted on amino acid side chains, the affinity of 3 amino acids (with different side-chain polarity) to non-stoichiometric hydroxyapatite nanoparticles was tested, and the study concluded glycine and lysine had a greater attachment to non-stoichiometric hydroxyapatite nanoparticles than aspartic acid, due to the polarity of its side chain (Comeau & Willett, 2018). This agrees with the findings that the polarity of the side chain can heavily influence its affinity for bound entities.

4.2) Future Work

There have been previous studies that have shown great promise in terms of predicting the pathogenicity of variants (Pejaver et al., 2020) (Carter, Douville, Stenson, Cooper & Karchin, 2013). However, for bioinformatics to truly have a lasting impact on clinical medicine, advancements in computer science driving the new approaches being developed are key. This will make bioinformatics tools more efficient and cost-effective. As computers get faster and datasets get larger, more complex studies can be conducted. An example of this already showing fruitful results, is in the development of the PathoSystems Resource Integration Center (PATRIC) (Davis et al., 2019). PATRIC offers web-based visualization and comparative analysis tools, for bacteria with a special emphasis on pathogens. PATRIC has been able to be achieved because of the increased cost-effectiveness of genomic and other omics-related work over the past several years.

Well-implemented pipelines can be powerful tools and, in this project, the pipeline has been shown to potentially have many real-world applications. However, as with many bioinformatic approaches, the translation of a practical scenario (in this case drug development) into a completely computerised system is very difficult and some scenarios cannot be accounted for. In this project, the pipeline is able to supplement the drug development process, especially the drug target identification phase, which can make data collection and analysis much quicker. By entering a disorder, the pipeline is able to identify proteins involved in the disorder and known compounds that bind to the specific proteins, however, the pipeline does not account for the effect of the compound on specific organs and organ systems. The pipeline also does not state the toxicity of the compounds and the level at which the compound is toxic. These are all aspects of drug development that would be investigated, either *in vivo or in vitro.* To further advance the pipeline and increase the pipeline’s ability to be applied in clinical situations, the addition of these as attributes within the compound entity should be considered. This would require the data to be present and readily available to be web-scraped and added to the database. This is one of the few disadvantages of bioinformatics, as it requires real-world practical experiments to be conducted first, in order to learn from the data and produce computed predictions and analysis. The donation of data and the development of data warehouses; for data sharing and the definition of standards for sharing phenotypic data are essential for the advancement of bioinformatics (Bellazzi et al., 2012).

Looking into the future of the pipeline, the aim would be to make constant advancements and improvements to the pipeline. The application of the pipeline into more complex clinical scenarios would require further development of the pipeline. The goal would be to expand and develop the pipeline to the stage in which it could analyse multi-variants or multi-point mutations simultaneously and then to a point where whole human genomes can be input into the pipeline. This would require upscaling of the pipeline from storing the data for 720 proteins from the full pharmacology set to over 20,000. Currently, the pipeline can be used to investigate single mutations of the proteins in the full pharmacology set. With the constant advancement in human genome sequencing, the hope for the future would be to incorporate that heavily into the pipeline allowing for the genome of a patient to be entered into the pipeline. Then, using the pipeline's capability to link variants to disorders and to treatments of the disorders, all the pathogenic variants the patient possesses would be highlighted and considered in light of treatments if necessary. As we advance into a world where genome sequencing may become standard, the aim of using such pipelines as generic screening tools looks feasible (Nurk et al., 2022).

The field of bioinformatics has great potential and can advance our skills in precision medicine. Further research would need to be done. Looking onward from this project, a similar analysis could be conducted using the pipeline established, where a larger set of proteins, variants and ligands would be investigated. This would provide greater insight into the complete association between single nucleotide variations and drug binding. It would be favourable for the database to be able to encompass the whole human proteome, this would allow for a lot more inferences to be made and increase the pipelines' capability for use as a research tool.

4.3) Conclusion

To conclude, the development of a bioinformatics pipeline showed that *in silico* approaches have great potential, both in clinical research and in precision medicine. The bioinformatics pipeline yielded very positive results as it was able to gather and integrate layers of relevant information, from multiple global sources, on a host of clinically relevant proteins and diseases. The bioinformatics pipeline is very powerful with great potential, its ability to be able to incorporate multiple databases altogether could result in it proving to be a very useful tool in research scenarios and precision medicine. The incorporation of polymorphic variants data with disorders and treatment information is one of the building blocks of precision medicine, and a tool that can seamlessly integrate databases containing a large amount of information from these fields will aid the advancement of precision medicine.

This project concluded that although there was not a statistically significant difference between the protein-ligand binding energies of all the pathogenic variants and all the benign variants, there were many individual variants that provided striking results, in being much different to the wild-type. Several individual variants returned results that were completely different to the wildtype results, illustrating a possible association between variants and impaired drug binding ability. If patients carried these variants, there would be a sizeable number of them that would have reacted very differently to the ligands bound, and therefore a customised treatment would be required for them. This project also allowed for the conclusion that the effect of the variations on protein can be caused by a multitude of factors, such as the specific residues that are mutated in the variants and the location of the mutation. These can play a large role in a protein's ability to bind to its ligands, either by affecting the bonds that are formed when bound or by a separate effect on the structure of the protein. This can cause the protein to react differently to certain ligands and drug compounds. With advancements in precision medicine, the problems caused by this would be diminished, as the identification of functionally significant variants can be detected and then considered accordingly. This saves money, time and could even save lives.

5) Bibliography

1. Akhmetov, I., & Bubnov, R. (2015). Assessing value of innovative molecular diagnostic tests in the concept of predictive, preventive, and personalized medicine. EPMA Journal, 6(1). doi: 10.1186/s13167-015-0041-3
2. Ali, A., Azam, M., & Khan, A. (2018). Non-active site mutation (Q123A) in New Delhi metallo-β-lactamase (NDM-1) enhanced its enzyme activity. International Journal Of Biological Macromolecules, 112, 1272-1277. doi: 10.1016/j.ijbiomac.2018.02.091
3. Aronson, S., & Rehm, H. (2015). Building the foundation for genomics in precision medicine. Nature, 526(7573), 336-342. doi: 10.1038/nature15816
4. Arora, S., & Katyal, A. (2019). Protein Modifications and Lifestyle Disorders. Protein Modificomics, 87-108. doi: 10.1016/b978-0-12-811913-6.00004-7
5. Ascierto, P., Kirkwood, J., Grob, J., Simeone, E., Grimaldi, A., & Maio, M. et al. (2012). The role of BRAF V600 mutation in melanoma. Journal Of Translational Medicine, 10(1). doi: 10.1186/1479-5876-10-85
6. Babaei, M. A., Kamalidehghan, B., Saleem, M., Huri, H. Z., & Ahmadipour, F. (2016). Receptor tyrosine kinase (c-Kit) inhibitors: a potential therapeutic target in cancer cells. Drug design, development and therapy, 10, 2443.
7. Bateman, A., Martin, M., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C., Arighi, C., Arminski, L., Chen, C., Chen, Y., Garavelli, J., Huang, H., Laiho, K., McGarvey, P., Natale, D., Ross, K., Vinayaka, C., Wang, Q., Wang, Y., Yeh, L., Zhang, J., Ruch, P. and Teodoro, D., 2020. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), pp.D480-D489.
8. Bellazzi, R., Masseroli, M., Murphy, S., Shabo, A., & Romano, P. (2012). Clinical Bioinformatics: challenges and opportunities. BMC bioinformatics, 13(14), 1-8.
9. Bibbins-Domingo, K. (2016). Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals Of Internal Medicine, 164(12), 836. doi: 10.7326/m16-0577
10. Bollong, M., Lee, G., Coukos, J., Yun, H., Zambaldo, C., & Chang, J. et al. (2018). A metabolite-derived protein modification integrates glycolysis with KEAP1–NRF2 signalling. Nature, 562(7728), 600-604. doi: 10.1038/s41586-018-0622-0
11. Bolós, M., Llorens-Martín, M., Jurado-Arjona, J., Hernández, F., Rábano, A., & Avila, J. (2015). Direct Evidence of Internalization of Tau by Microglia In Vitro and In Vivo. Journal of Alzheimer's Disease, 50(1), 77–87. doi:10.3233/jad-150704

12. Bolós, M., Perea, J. R., & Avila, J. (2017). Alzheimer's disease as an inflammatory disease. Biomolecular Concepts, 8(1). doi:10.1515/bmc-2016-0029
13. Bonetta, R., & Valentino, G. (2019). Machine learning techniques for protein function prediction. Proteins: Structure, Function, And Bioinformatics, 88(3), 397-413. doi: 10.1002/prot.25832
14. Can, T. (2013). Introduction to Bioinformatics. Mirnomics: Microrna Biology And Computational Analysis, 51-71. doi: 10.1007/978-1-62703-748-8_4
15. Carter, H., Douville, C., Stenson, P., Cooper, D., & Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics, 14(S3). doi: 10.1186/1471-2164-14-s3-s3
16. Cha, Y., Erez, T., Reynolds, I., Kumar, D., Ross, J., & Koytiger, G. et al. (2017). Drug repurposing from the perspective of pharmaceutical companies. British Journal Of Pharmacology, 175(2), 168-180. doi: 10.1111/bph.13798
17. Choi, R., Lee, J., Park, H., Park, J., Kim, Y., & Ki, C. et al. (2017). Reassessing the significance of the PAH c.158G>A (p.Arg53His) variant in patients with hyperphenylalaninemia. Journal Of Pediatric Endocrinology And Metabolism, 30(11). doi: 10.1515/jpem-2017-0158
18. Comeau, P., & Willett, T. (2018). Impact of Side Chain Polarity on Non-Stoichiometric Nano-Hydroxyapatite Surface Functionalization with Amino Acids. Scientific Reports, 8(1). doi: 10.1038/s41598-018-31058-5
19. Davis, J., Wattam, A., Aziz, R., Brettin, T., Butler, R., & Butler, R. et al. (2019). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Research. doi: 10.1093/nar/gkz943
20. de Oliveira Lupatini, E., Zimmermann, I. R., Barreto, J. O. M., & da Silva, E. N. (2022). How long does it take to translate research findings into routine healthcare practice?—the case of biological drugs for rheumatoid arthritis in Brazil. Annals of translational medicine, 10(13).
21. Deore, A. B., Dhumane, J. R., Wagh, R., & Sonawane, R. (2019). The stages of drug discovery and development process. Asian Journal of Pharmaceutical Research and Development, 7(6), 62-67.
22. Exner, T., Korb, O., & ten Brink, T. (2009). New and improved features of the docking software PLANTS. Chemistry Central Journal, 3(S1). doi: 10.1186/1752-153x-3-s1-p16
23. Fernandez-Gamba, A., Leal, M., Morelli, L., & Castano, E. (2009). Insulin-Degrading Enzyme: Structure-Function Relationship and its Possible Roles in Health and Disease. Current Pharmaceutical Design, 15(31), 3644–3655. doi:10.2174/138161209789271799
24. Gauthier, J., Vincent, A., Charette, S. and Derome, N., 2019. A brief history of bioinformatics. Briefings in Bioinformatics, 20(6), pp.1981-1996.
25. Gharesouran, J., Jalaiei, A., Hosseinzadeh, A., Ghafouri-Fard, S., Mokhtari, Z., & Ghahremanzadeh, K. et al. (2020). GAA gene mutation detection following clinical evaluation and enzyme activity analysis in Azeri Turkish patients with Pompe disease. Metabolic Brain Disease, 35(7), 1127-1134. doi: 10.1007/s11011-020-00586-3
26. Ginsburg, G., & Phillips, K. (2018). Precision Medicine: From Science To Value. Health Affairs, 37(5), 694-701. doi: 10.1377/hlthaff.2017.1624
27. Harding, S., Sharman, J., Faccenda, E., Southan, C., Pawson, A., Ireland, S., Gray, A., Bruce, L., Alexander, S., Anderton, S., Bryant, C., Davenport, A., Doerig, C., Fabbro, D., Levi-Schaffer, F., Spedding, M. and Davies, J., 2017. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Research, 46(D1), pp.D1091-D1106.

28. Hernández-Ochoa, E. O., Pratt, S. J., Lovering, R. M., & Schneider, M. F. (2016). Critical role of intracellular RyR1 calcium release channels in skeletal muscle function and disease. Frontiers in physiology, 6, 420.
29. Huang, S. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discovery Today, 19(8), 1081-1096. doi: 10.1016/j.drudis.2014.02.005
30. Jensen, L., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., & Muller, J. et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research, 37(Database), D412-D416. doi: 10.1093/nar/gkn760
31. Kaitin, K. I. (2010). Deconstructing the drug development process: the new face of innovation. Clinical Pharmacology & Therapeutics, 87(3), 356-361.
32. Karki, R., Pandya, D., Elston, R., & Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. BMC Medical Genomics, 8(1). doi: 10.1186/s12920-015-0115-z
33. Kim, S., Thiessen, P., Bolton, E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B., Wang, J., Yu, B., Zhang, J. and Bryant, S., 2015. PubChem Substance and Compound databases. Nucleic Acids Research, 44(D1), pp.D1202-D1213.
34. König, I., Fuchs, O., Hansen, G., von Mutius, E., & Kopp, M. (2017). What is precision medicine?. European Respiratory Journal, 50(4), 1700391. doi: 10.1183/13993003.00391-2017
35. Kosorok, M., & Laber, E. (2019). Precision Medicine. Annual Review Of Statistics And Its Application, 6(1), 263-286. doi: 10.1146/annurev-statistics-030718-105251
36. Landrum, M., Lee, J., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J., Kattman, B. and Maglott, D., 2017. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), pp.D1062-D1067.
37. Loo, E., Khalili, P., Beuhler, K., Siddiqi, I., & Vasef, M. (2018). BRAF V600E Mutation Across Multiple Tumor Types: Correlation Between DNA-based Sequencing and Mutation-specific Immunohistochemistry. Applied Immunohistochemistry & Molecular Morphology, 26(10), 709-713. doi: 10.1097/pai.0000000000000516
38. Louden, D., 2020. MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. Medical Reference Services Quarterly, 39(2), pp.183-191.
39. Lung, J., Hung, M., Lin, Y., Jiang, Y., Fang, Y., & Lu, M. et al. (2020). A highly sensitive and specific real-time quantitative PCR for BRAF V600E/K mutation screening. Scientific Reports, 10(1). doi: 10.1038/s41598-020-72809-7
40. Maier, M. (2019). Personalized medicine—a tradition in general practice!. European Journal Of General Practice, 25(2), 63-64. doi: 10.1080/13814788.2019.1589806
41. Malathi, K., & Ramaiah, S. (2018). Bioinformatics approaches for new drug discovery: a review. Biotechnology And Genetic Engineering Reviews, 34(2), 243-260. doi: 10.1080/02648725.2018.1502984
42. Maphis, N., Xu, G., Kokiko-Cochran, O. N., Jiang, S., Cardona, A., Ransohoff, R. M., Lamb, B. T. & Bhaskar, K. (2015). Reactive microglia drive tau pathology and contribute to the spreading of pathological tau in the brain. Brain, 138(6), 1738–1755. doi:10.1093/brain/awv081
43. Mendez, D., Gaulton, A., Bento, A., Chambers, J., De Veij, M., Félix, E., Magariños, M., Mosquera, J., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C., Segura-Cabrera, A.,

Hersey, A. and Leach, A., 2018. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research, 47(D1), pp.D930-D940.

44. Mendonça, L. M. F., &amp; Marana, S. R. (2011). Single mutations outside the active site affect the substrate specificity in a β-glycosidase. Biochimica Et Biophysica Acta (BBA) - Proteins and Proteomics, 1814(12), 1616–1623. https://doi.org/10.1016/j.bbapap.2011.08.012
45. Mora, C., Tittensor, D., Adl, S., Simpson, A., & Worm, B. (2011). How Many Species Are There on Earth and in the Ocean?. Plos Biology, 9(8), e1001127. doi: 10.1371/journal.pbio.1001127
46. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. Science, 376(6588), 44–53. https://doi.org/10.1126/science.abj6987
47. O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., & Hutchison, G. (2011). Open Babel: An open chemical toolbox. Journal Of Cheminformatics, 3(1). doi: 10.1186/1758-2946-3-33
48. Office of national statistics (2020). Number of deaths caused by dementia or Alzheimer's disease by place of death, England and Wales: 2018. Retrieved March 18, 2022, from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/11324numberofdeathscausedbydementiaoralzheimersdiseasebyplaceofdeathenglandanwales2018
49. Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K., Lin, G., & Nam, H. et al. (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications, 11(1). doi: 10.1038/s41467-020-19669-x
50. Petreni, A., Osman, S., Alasmary, F., Almutairi, T., Nocentini, A., & Supuran, C. (2021). Binding site comparison for coumarin inhibitors and amine/amino acid activators of human carbonic anhydrases. European Journal Of Medicinal Chemistry, 226, 113875. doi: 10.1016/j.ejmech.2021.113875
51. Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., & Ferrin, T. (2004). UCSF Chimera?A visualization system for exploratory research and analysis. Journal Of Computational Chemistry, 25(13), 1605-1612. doi: 10.1002/jcc.20084
52. Pivovarova, O., Höhn, A., Grune, T., Pfeiffer, A. F. H., & Rudovich, N. (2016). Insulin-degrading enzyme: new therapeutic target for diabetes and Alzheimer’s disease? Annals of Medicine, 48(8), 614–624. doi:10.1080/07853890.2016.1197416
53. Qian, T., Zhu, S., & Hoshida, Y. (2019). Use of big data in drug development for precision medicine: an update. Expert Review Of Precision Medicine And Drug Development, 4(3), 189-200. doi: 10.1080/23808993.2019.1617632
54. Ragland, D. A., Nalivaika, E. A., Nalam, M. N., Prachanronarong, K. L., Cao, H., Bandaranayake, R. M., Cai, Y., Kurt-Yilmaz, N., &amp; Schiffer, C. A. (2014). Drug resistance conferred by mutations outside the active site through alterations in the dynamic and structural ensemble of HIV-1 Protease. Journal of the American Chemical Society, 136(34), 11956–11963. https://doi.org/10.1021/ja504096m
55. Raschka, S., Wolf, A., Bemister-Buffington, J., & Kuhn, L. (2018). Protein–ligand interfaces are polarized: discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes. Journal Of Computer-Aided Molecular Design, 32(4), 511-528. doi: 10.1007/s10822-018-0105-2

56. Roig-Zamboni, V., Cobucci-Ponzano, B., Iacono, R., Ferrara, M., Germany, S., & Bourne, Y. et al. (2017). Structure of human lysosomal acid α-glucosidase–a guide for the treatment of Pompe disease. Nature Communications, 8(1). doi: 10.1038/s41467-017-01263-3
57. Rybka, J., Bogunia-Kubik, K., Kuszczak, B., Kalicińska, E., Łacina, P., Dratwa, M., &amp; Wróbel, T. (2021). Analysis of polymorphism in the genes TLR3, TLR4 and TLR9 in patients with acute myeloid leukemia. Blood, 138(Supplement 1), 4470–4470. https://doi.org/10.1182/blood-2021-151475
58. Saultz, J. N., & Garzon, R. (2016). Acute myeloid leukemia: a concise review. Journal of clinical medicine, 5(3), 33.
59. Shapovalov, M., & Dunbrack, R. (2011). A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. Structure, 19(6), 844-858. doi: 10.1016/j.str.2011.03.019
60. Smith, R., Lovell, S., Burke, D., Montalvao, R., & Blundell, T. (2007). Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. Bioinformatics, 23(9), 1099-1105. doi: 10.1093/bioinformatics/btm073
61. Smith, S., & Waters, N. (2018). Pharmacokinetic and Pharmacodynamic Considerations for Drugs Binding to Alpha-1-Acid Glycoprotein. Pharmaceutical Research, 36(2). doi: 10.1007/s11095-018-2551-x
62. Sperber, N., Carpenter, J., Cavallari, L., J. Damschroder, L., Cooper-DeHoff, R., & Denny, J. et al. (2017). Challenges and strategies for implementing genomic services in diverse settings: experiences from the Implementing GeNomics In pracTicE (IGNITE) network. BMC Medical Genomics, 10(1). doi: 10.1186/s12920-017-0273-2
63. Sunny, S., & Jayaraj, P. B. (2022). Protein–protein docking: Past, present, and future. The protein journal, 41(1), 1-26.
64. Swain, S. S., & Hussain, T. (2021). Role of Bioinformatics in Early Drug Discovery: An Overview and Perspective. Computation in BioInformatics: Multidisciplinary Applications, 49-67.
65. van der Ploeg, A., & Reuser, A. (2008). Pompe's disease. The Lancet, 372(9646), 1342-1353. doi: 10.1016/s0140-6736(08)61555-x
66. Wang, J., Wang, W., Kollman, P., & Case, D. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. Journal Of Molecular Graphics And Modelling, 25(2), 247-260. doi: 10.1016/j.jmgm.2005.12.005
67. Wang, Q., Xu, R., & Volkow, N. D. (2021). Increased risk of COVID-19 infection and mortality in people with mental disorders: analysis from electronic health records in the United States. World Psychiatry, 20(1), 124-130.
68. Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., Zhang, R., Zhu, J., Ren, Y., Tan, Y., Qin, C., Li, Y., Li, X., Chen, Y. and Zhu, F., 2020. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic acids research, 48(D1), D1031-D1041.
69. Wishart, D., Feunang, Y., Guo, A., Lo, E., Marcu, A., & Grant, J. et al. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1), D1074-D1082. doi: 10.1093/nar/gkx1037
70. Wittenberg, R., Hu, B., Jagger, C., Kingston, A., Knapp, M., Comas-Herrera, A., & Banerjee, S. (2020). Projections of care for older people with dementia in England: 2015 to 2040. Age and Ageing, 49(2), 264-269.
71. Xia, X. (2017). Bioinformatics and drug discovery. Current topics in medicinal chemistry, 17(15), 1709-1726.

72. Zardecki, C., Dutta, S., Goodsell, D., Lowe, R., Voigt, M., & Burley, S. (2021). PDB -101: Educational resources supporting molecular explorations through biology and medicine. Protein Science, 31(1), 129-140. doi: 10.1002/pro.4200
73. Zhang, Z., & Tang, W. (2018). Drug metabolism in drug discovery and development. Acta Pharmaceutica Sinica B, 8(5), 721-732. doi: 10.1016/j.apsb.2018.04.003
74. Zhang, H., Zhang, Z., Jia, L., Ji, W., & Li, H. (2018). Genetic polymorphism in the RYR1 C6487T is associated with severity of hypospadias in Chinese han children. BioMed Research International, 2018.
75. Zhou, W., Sailani, M., Contrepois, K., Zhou, Y., Ahadi, S., & Leopold, S. et al. (2019). Longitudinal multi-omics of host-microbe dynamics in prediabetes. Nature, 569(7758), 663-671. doi: 10.1038/s41586-019-1236-x

Appendix